Wireless Communications and Mobile Computing

Intelligent Computing for the Next-Generation Communication Networks: Emerging Trends and Challenges

Lead Guest Editor: Lisheng Fan
Guest Editors: Arumugam Nallanathan and Zhao Junhui

Intelligent Computing for the Next-Generation Communication Networks: Emerging Trends and Challenges

Wireless Communications and Mobile Computing

Intelligent Computing for the Next-Generation Communication Networks: Emerging Trends and Challenges

Lead Guest Editor: Lisheng Fan
Guest Editors: Arumugam Nallanathan and Zhao Junhui

Chief Editor

Associate Editors

Academic Editors

Contents

Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 5472324, 15 pages
https://doi.org/10.1155/2022/5472324

Research Article

Transferable Adversarial Attacks against Automatic Modulation Classifier in Wireless Communications

Lin Hu, Han Jiang, Wen Li, Hao Han, Yang Yang, Yutao Jiao, Haichao Wang, and Yuhua Xu

College of Communications Engineering, PLA Army Engineering University, Nanjing, China

Correspondence should be addressed to Han Jiang; jh_forward@126.com

Received 30 June 2022; Revised 9 August 2022; Accepted 25 August 2022; Published 27 September 2022

Academic Editor: Lisheng Fan

Deep neural network-based automatic modulation recognition (AMR) technology has become an increasingly important area due to the advantages of self-extraction of features and high identification accuracy. Based on the view of security threats to machine learning classifiers, we investigate the influence of adversarial samples on the AMR model in this paper. The traditional method is based on label gradient attack without taking advantage of the feature-level transferability, resulting in the attack effect that is not perfect. So, we exploit the feature-level transferability property that could be met to fulfill realistic imperceptibility and transfer needs. In this paper, firstly, we proposed an AMR scheme with high recognition accuracy as our attack model. Secondly, we proposed a transferable attack method based on a feature gradient-based, which increases perturbation to clean signal based on features space. Finally, we introduce a new attack strategy, in which we select two original and one adversarial target signal sample as the input of triplet loss to achieve higher attack strength and high transferability. Meanwhile, this paper proposes indicators of signal characteristics to test the effectiveness of our proposed attack method. Based on experimental results, our proposed feature gradient-based adversarial attack method outperforms the currently labeled gradient attack methods regarding attack effectiveness and transferability.

1. Introduction

Deep learning (DL) has been revealed to be successful in conducting diverse wireless communication tasks like signal recognition [1] and spectrum prediction [2]. The key technology of signal detection and demodulation is AMR. It can effectively solve the increasingly crowded and complex electromagnetic space environment, and it is also an important premise for alleviating the spectrum resources shortages. Convolutional neural network (CNN) and long-short term memory (LSTM) are two methods that have achieved good recognition accuracy. However, DL in general has been discovered to be vulnerable to attack by introducing a subtle perturbation that is imperceptible to the human eye [3]. This paper investigates the challenges in signal classification tasks, because signal classification is most widely studied in communication tasks. Actually, the developed methodology is easily transferred to all other tasks. Studying the threat posed by adversarial samples is crucial, not only to enable us to create algorithms that are resistant to interference from malicious samples but also for preventing adversaries from executing signal recognition tasks through such clever intervention. It is important to note that the assault, which is a direct access attack, is started by manipulating the receivers' signal modulation classifier. This kind of attack might not be feasible in the real world because it necessitates the penetration of a target model. Nevertheless, direct access attack methods remain helpful [4] visualizing adversarial perturbations in modulation recognition by reconstructing the waveforms, while compared to other forms of attack, they are more difficult to detect. [5] analyzed direct attack and physical attack that are closer to hardware requirements using traditional FGSM methods. Thus, research into such

direct and digital attacks is of great significance in real world applications, and direct access attacks with AMR may play the role of the foundation for more complex over-the-air attacks [6]. Above all are utilizing the based label gradient attack method. Feature gradient-based adversarial methods are already available in the field of image recognition [7].

To summarize, existing attack methods for signal classification models typically suffer from the disadvantages presented below. Firstly, the transferability of adversarial examples is imperfect in attacking the black-box model, particularly in the presence of targeted attacks. That is because the current methods mostly adopt single-layer features rather than attacking with taking the use of features space. In fact, the middle layer of the CNN representation is transferable. Normally, CNN low-level features have a lot of granular information, while its high-level features have a lot of global semantic information. Secondly, the adversarial sample is difficult to categorize into the stated target class since the standard label gradient-based assaults only limit the distance between the adversarial sample and the target class. Evaluating only the success rate of an attack does not correspond to the merely evaluation measure of the effectiveness of an attack in the field of signal recognition. In the real communication environment, we know relatively little a priori information, and it is necessary to maintain a certain degree of imperceptible to achieve the effect of the attack.

To address the abovementioned issues, we propose a feature gradient-based attack method, which relies on two basic observations. The first is a deep learning classifier model that predicts mainly on the basis of the signal samples information and differentiation regions. However, the presence of such regions weakens the models. The second conclusion is that perturbation in the middle layer features of well-trained networks is transferable [8–10]. Research [8] concluded that feature representations are universal in neural networks, and that feature representation can be transferred for learning by transferring to the target network. Furthermore, features from various levels exhibit diverse features. [9] improves the evidence, proving that adversarial examples can be produced through operating image representations under deep neural networks. The current work focuses on adding the potential representation space of adversarial ingestion to those regions of the signal sample that are informative and distinguishable. This contributions are as follows:

(i) To provide more transferable and efficient adversarial examples, this paper proposed a transferable attentive method concentrating on the informative and discriminative feature regions, adding perturbation at the feature level will be more adaptable to realistic scenarios. The proposed attack methods are more effective when compared the previous methods in the modulation recognition scenario

(ii) We have conducted experiments in all metrics of our method with a new system of indicators that better suit the signal characteristics. Our method surpasses that of the traditional label gradient method in most indexes

The remaining of this paper is arranged as follows: Section 2 presents the related work of DL in modulation signal classification and the threat of adversarial examples; Section 3 of this paper introduces the methodology of adversarial examples based on feature gradient; Section 4 develops a series of experiments from the perspectives of white-box attack and black-box attack, explores the experimental results, and verifies the effectiveness. Finally, this paper is summarized and looks forward to the future.

2. Related Work

2.1. AMR Model. The concept of AMR was first proposed in [11], as one of the pattern recognition research, and it has filled everyone's vision. Machine learning (ML) methods have been extensively used based on the constant advancement of technology. DL has been developed recently into a popular technology for breaking through the performance bottleneck of pattern recognition tasks, and this technology has also been introduced into the field of AMR. Based on their perspective of development in the field of AMR, recognition algorithms are classified into two types: classical modulation recognition methods and deep learning-based modulation recognition methods [12]. The classical methods can be divided into recognition methods based on likelihood function [13] and recognition methods based on feature extraction [14].

With increasingly complex and diverse communication systems, wireless signal data is more complex and diverse than ever, with stronger randomness and heterogeneity. Traditional modulation recognition requires manual extraction of features and relies on prior information. The workload is heavy, and the recognition accuracy is low. Therefore, the industry applies DNN to the field of signal recognition. The DNN model requires a large amount of training data, and the massive features of wireless communication signals were provided. In comparison to traditional methods, DNN can automatically extract modulated signal features, eliminating the errors that may be introduced by the manual selection of features and the dependence on expert knowledge in the identification process. The most important thing is that AMR can achieve more accurate results. The present study investigates the threats specific to the signal classification stage and is thus related to adversarial machine learning [15] which has witnessed an increase in activity in the context of CV [16]. Recently, the search for DL signal recognition has mainly been based on two perspectives: signal array and imaged-based. The texture map of the in-phase and quadrature (IQ) waveform of the communication signal is applied as the input of the DL model in the signal array recognition method. According to Rajendran et al. [17], through the transformation of IQ data into AP (amplitude/phase) information and adoption of a simple LSTM model, a perform accuracy was attainable. The model enabled the extraction of temporal signal traits from the training data,

where it is unnecessary to extract the expert traits manually. Attention mechanism (AM), which was originally adopted for machine translation [18] as a crucial concept in the DL domain, is currently applied extensively in areas like speech recognition, NLP (natural language processing), statistical learning, and computers. Chen et al. [19] put forward a new attention cooperative framework in which the input feature maps were made mutually dependent by incorporating the classifiers with a self-AM and a Squeeze-and-Excitation block [20]. The validity of AM is proved in AMR. The present study is aimed at developing an AMR model based on DL. The AP information is initially extracted from the IQ data, and then the classification outcomes are derived using an AM-based monolayer LSTM model. Our developed scheme is compared to the existing CNN-AP, LSTM-AP, and CLDNN-AP schemes. The accuracy of classification can be less influenced by the signal frequency offset when using CNN [21]. LSTM is appropriate for obtaining time-series signal traits [22]. CLDNN (convolution, LSTM, deep neural network), which integrates the benefits of DNN, CNN, and LSTM, is proven to be highly competent in classifying the modulation modes [23, 24].

2.2. Adversarial Evasion Attack. The first step in guaranteeing system security is to identify the systems challenges. This paper firstly characterizes the possible source of challenges, envisions new challenges, and describes the restrictions in adversarial attacks under the background of wireless communications. The uninterpretable DNN exposes them to a variety of security risks. Szegedy et al. discovered that by adding some carefully crafted tiny human-imperceptible perturbation to the input samples, the accuracy of DNN classifiers can be significantly reduced, and such added perturbed samples are called adversarial example [25]. Adversarial attacks can be categorized into two categories based on whether or not the adversarial sample has a target: targeted attacks and untargeted attacks. Targeted attacks are those where the adversarial sample must misclassify the input sample into a specific class to deceive the model. For example, in modulated signal classification, if the attacker specifies the target class as ASK, 8PSK, QPSK, or any other class of signals, it will be incorrectly classified as ASK after being attacked, while targetless attacks are the inverse of targeted attacks, where no specific attack signal class is required, i.e., the target can be any type of signal other than its signal.

Untargeted attacks can be classified into white-box attacks, black-box attacks, and gray-box attacks based on the knowledge level in the target model. In a white-box attack, the adversary is aware of the training data, architecture, algorithms, and optimization techniques, which enables it to fully access the trained model. A black-box attack neither knows nor has accesses to the training data and training model, making it a more realistic and practical scenario that also increases the difficulty of the attack. A gray-box attack is one in which only a limited amount of information is known ahead of time.

The majority of the adversarial sample research are currently focused on image recognition. Goodfellow et al. presented fast gradient sign method (FGSM) to attack deep network models, the core idea being to obtain the adversarial sample by computing the gradient of the loss function relative to the input sample itself [26]. Kurakin et al. put forward the iterative FGSM (basic iterative fast gradient sign method, BIM), which uses multiple iterations to generate an adversarial sample [27]. Dong et al. presented momentum into the gradient calculation process in the iterative attack and proposes the momentum iterative method (MIM) method to enhance the stability of the model at each iteration and the generalization of the adversarial samples [28]. Moosavi-dezfooli et al. proposed an algorithm called Deep Fool, which replaces the deep classification model with a linear model for attack [29]. Lin et al. introduced the Nesterov accelerated gradient into the iterative attack process and proposed PGD to increase the adversarial samples migrability [30]. Kurakin et al. presented an approach to performing adversarial training on the model to explore the impact of the adversarial samples on the model robustness [31]. Carlini and Wagner proposed three methods to generate perturbations, using three different metrics (L_1, L_2, L_∞) to avoid the robustness of the model [32]. The real-world artifacts can also be used to trick the classification model [33].

Little work has been done to apply adversarial example attacks to AMR, and Lin et al. applied the traditional adversarial method based on label computation gradient to modulated signal recognition and verified that AMR is vulnerable to adversarial sample attack [34], However, the above methods still use the alternative model approach when performing black-box attacks and do not fully utilize the features of modulated signal data samples. Moreover, the recognition accuracy of the target model itself chosen for modulated signal recognition is not high, only about 70%. Because of the disadvantage of the previous work, we first propose a target recognition model with high accuracy, which could attain a top accuracy about 91%. The adversarial example was then generated using a feature gradient. Finally, we use a new strategy in which we select two original samples and one target sample as triplet loss input.

3. Transferable Attack Methodology

This study proposes a new black-box targeted attack method for signal classification, named transferable adversarial attack, which can deceive white-box models. The current section firstly depicts the methodology of the fundamental idea of generating adversarial examples. The algorithm flow is then given. Finally, evaluate the feasibility of the proposed algorithm.

3.1. Backgrounds. The most of raw IQ signal classifiers attempt to get a signal snapshot x and output the most confident result class y. In most situations, x denotes a two-dimensional matrix (IQ, number of samples) that reflects a single channel of complicated data with little preprocessing. It employs DNN to learn a mapping from data by solving

problems, particularly in the domain of communications.

$$\underset{\theta}{\operatorname{argmin}} L(f(\theta, x), y)), \tag{1}$$

where x and y denote the training and true labels, respectively, and f denotes the network architecture used. To learn the network variable θ, a loss function is usually used in conjunction with an optimizer in DNN training. We assume that the data set is constant without data augmentation throughout model training, and that it is sampled from a distribution that is similar to that observed later in the communication system's operation. FGSM uses untargeted adversarial examples to build untargeted adversarial examples.

$$x^* = x + \varepsilon \cdot \operatorname{sign}\left(\nabla_x J(x, y, w)\right), \tag{2}$$

where y is the real input label, and ∇_x indicates the gradient of the loss function in terms of the original input x. The proposed approach in a single step can create adversarial examples x^* restricted by a distance ε, in the feature space.

The average energy per symbol (E_s) of a transmission can be calculated based on

$$[E_s] = \frac{\text{sps}}{N} \sum_{i=0}^{N} |s_i|^2, \tag{3}$$

where sps denotes samples per symbol, N is the total number of samples, and s_i denotes a particular sample in time. Without losing generality, the present study assumes the average energy per symbol of the modulated signal, $E_s = 1$. As a result, the underlying transmissions power ratio to the perturbation signal (E_j) can be derived as

$$\frac{E_s}{E_j} = \frac{1}{E_j} = 10^{-E_j(dB)/10}, \tag{4}$$

Since the input of sign (∇_x) in (2) is complicated, the output also remains complicated and is thus a vector with values $[\pm 1, \pm j]$. As a result, the magnitude of each the perturbation sample is computed as

$$|\operatorname{sign}(\nabla_x)| = |\operatorname{sign}(z)| = \sqrt{(\pm 1)^2 + (\pm 1)^2} = \sqrt{2}, \tag{5}$$

Thereby, the energy per symbol of sign (∇_x) can be calculated by plugging (5) into (6), leading to

$$E_{\operatorname{sign}(\nabla_x)} = \frac{\text{sps}}{N} \sum_{i=0}^{N} |\operatorname{sign}(\nabla_x)|^2 = 2 \times \text{sps}, \tag{6}$$

Since sps is fixed through transmission, a closed form scaling factor, ε, is deduced to obtain the desired energy ratio (E_s/E_j) by using

$$\varepsilon = \sqrt{\frac{E_j/E_s}{E_{\operatorname{sign}}(\nabla_x)}} = \sqrt{10\frac{E_j(\text{dB})}{2 \times \text{sps}}}, \tag{7}$$

Plugging ε into (2) allows the creation of adversarial examples constrained by (E_s/E_j) and can be simply expressed as

$$x^* = x + \sqrt{\frac{10^{E_j(dB)/10}}{2 \times \text{sps}}} \times \operatorname{sign}\left(\nabla_x L(f(\theta, x), y)\right). \tag{8}$$

Above constraining the power ratio in this way can be beneficial for assessing system design trade-offs.

$$\begin{aligned} &\min \|x^* - x\|_p, \\ &s.t, l(x) \neq l(x^*), \\ &x^* - x \in \varepsilon, \end{aligned} \tag{9}$$

where $\|\cdot\|_p$ suggests the L_p norm. Furthermore, the L_p of δ is defined as

$$\|\delta\|_p = \left(\sum_{i=1}^{n} \|\delta\|_p\right)^{1/p}, \tag{10}$$

where L_0, L_2, L_∞ are the three most common metrics. The L_0 is a quantitative metric for the pixel variations in an image, whereas it quantifies the nonzero vectors of perturbation in a signal. The L_2 metric quantifies the Euclidean distance between adversarial and original examples as an Euclidean norm; the L_∞ is responsible for the maximum alteration constraint of all signal vectors/pixels in the adversarial examples. The power budget of a transmitter is usually constant, and in this research, an adversarial strategy of ML that is unaware of underlying signal is considered. Hence, the power applied to the jamming signal is inapplicable to the underlying transmission.

3.2. Triplet Loss for Adversarial Attack. The traditional label gradient attack method calculates the gradient using the sample label y, incorporating the initial clean sample x and the consistent label y into the target model loss function. The attack direction can be obtained by computing the gradient and sign function and then multiplied by the perturbation size to realize the adversarial perturbation; finally, then combine with the original clean sample to form an adversarial example. Currently, the main attack methods based on label gradient are FGSM, BIM, and MIM. Obviously, failure to take advantage of fast gradient varies at the feature space and transferability.

The proposed method is a momentum iterative FGSM at the feature space level. Thus, we suggest using triple loss, which can minimizes between an anchor and a positive, both of which have the same identity, and maximize the distance between the anchor and a negative of a

different identity. As a result, the information region and the discriminative region in the sample may be perturbed through optimization of the triplet loss on the feature space. Because of extracting features, we need to truncate the target model from the L layer to obtain the truncated model, to ensure that the selected feature space is abundant enough, and this paper uniformly selects the activation layer as our target layer. Then, put x and y into f_L to obtain the original signal sample feature $f_L(x)$. The loss here uses a triplet pair $(f_L(x_i^a), f_L(x_i^p), f_L(x_i^n))$, the anchor, positive, and negative terms of the triplet loss, respectively, and signal samples from the same class should be near together in the embedding space, forming several well-separated clusters. As a result, triplet loss ensures that the attack process not only makes the original sample close to positive sample (target sample) and away from negative sample (untarget sample).

Triplet loss can be expressed as

$$L_{\text{tri}} = \sum_{i}^{N} \left[\left\| f(x_i^a) - f(x_i^p) \right\|_2^2 - \left\| f(x_i^a) - f(x_i^n) \right\|_2^2 + \alpha \right]_+, \tag{11}$$

where $\alpha \in R^+$ denotes a margin between negative and positive pairs. The triple loss is adopted for adversarial example crafting in addition to strengthen the adversarial robustness [35], and the triple loss is also exploited for the adversarial example crafting purpose. The present work is the first attempt to use the triplet loss to craft the adversarial examples, where a source sample feature is drawn closer to the target class while being propelled away from the source class. In contrast to the conventional triplet loss, the clean signal sample acts as the anchor example, while the other clean and target class samples act as the negative and positive examples, respectively. With our attack, the anchor and positive examples are reasonably separated, while the distance between the anchor and negative examples is increased. The adversarial examples are easily misclassified into the target class by our triple loss-based algorithm. Furthermore, unlike the standard triplet loss in which every element is a clean sample, our triplet loss includes an adversarial example term, which can be found in Figure 1.

3.3. Basic Ideas. This research proposes two methods based on the aforementioned motivation. This algorithm can simulate the traditional BIM and MIM attack methods. To destroy the potential representation space, we propose to optimize triplet state loss rather than crossentropy loss. Furthermore, this study proposes two methods, with more intuitive variants explained in Algorithms 1 and 2.

When AMR is attacked, it is expected to add an imperceptible slight perturbation in the clean original sample, resulting in an error recognition rate. Suppose the original signal sample is x, the classification result is y, and the perturbation is small enough to meet $\|\eta\|_\infty \le \varepsilon$. So, FGSM was described below.

$$\begin{cases} \eta = \varepsilon \cdot \text{sign}\ (\nabla_x J(x, y)), \\ x^* = x + \eta, \end{cases} \tag{12}$$

where J is the target models loss function, and $\nabla_x J(x, y)$ refers to the derivative of the loss function over sample x. Because FGSM refers to a one-step attack, it is impossible to update the adversarial example by querying the model parameters in multiple times. The basic iterative method (BIM) denotes an extended FGSM in which adversarial examples are generated in various iterations. Every iteration has a small step size, and each step should be within the perturbation neighborhood of the original input.

$$\begin{cases} x_0 = x, \\ x_{n+1} = \text{Clip}_{x,\varepsilon}\{x_n + \varepsilon\ \text{sign}\ (\nabla_x J(x_n, y))\}. \end{cases} \tag{13}$$

$\text{Clip}_{x,\varepsilon}\{\}$ means to limit it to the scope $[x - \varepsilon, x + \varepsilon]$.

MIM is a reduction algorithm iteration technology that accelerates the speed under the gradient through accumulating the velocity vector in the gradient direction of the loss function. It can be denoted as follows:

$$\begin{cases} x_0^* = x, g_0 = 0, \\ g_{n+1} = \mu \cdot g_n + \dfrac{\nabla_{x_n^*} J(x_n^*, y)}{\left\| \nabla_{x_n^*} J(x_n^*, y) \right\|_1}, \\ x_{n+1}^* = \text{Clip}_{x,\varepsilon}\{x_n^* + \beta \cdot \text{sign}\ (g_{n+1})\}. \end{cases} \tag{14}$$

g_{n+1} represents the cumulative gradient generated by the previous $n + 1$ iteration, and μ is the attenuation factor.

MIM, same as BIM, incorporates an acceleration gradient into the iterative attack process and improves the migration performance of the adversarial examples, which can be denoted as

$$\begin{cases} x_0^* = x, g_0 = 0, \\ x_n^{\text{nes}} = x_n^* + \beta \cdot \mu \cdot g_n, \\ g_{n+1} = \mu \cdot g_n + \dfrac{\nabla_{x_n^*} J(x_n^{\text{nes}}, y)}{\left\| \nabla_{x_n^*} J(x_n^{\text{nes}}, y) \right\|_1}, \\ x_{n+1}^* = \text{Clip}_{x,\varepsilon}\{x_n^* + \beta \cdot \text{sign}\ (g_{n+1})\}. \end{cases} \tag{15}$$

Among them, x_n^{nes} is a Nesterov item, which jointly participates in the calculation of gradient.

3.4. Description of Attack Method. In order to perform an attack on the feature space of the AMR model, it is first necessary to find a suitable feature space. Meanwhile, to ensure that the selected feature space is sufficiently informative, the truncation layer of the target model is chosen as the final fully connected layer of the model. In order to ensure that the selected feature space is rich enough, the truncation layer of the target model is chosen as the

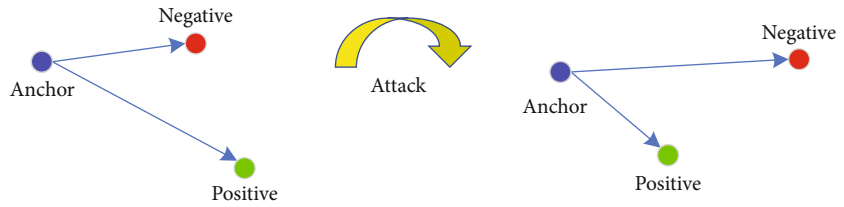


Figure 1: Schematic diagram of triple loss. The attack process not only makes the original sample close to positive sample (target sample), and away from negative sample (untarget sample), by pushing and pulling, but also to achieve a better attack effect.

Input: A classifier truncated model f_L; original signal sample to be attacked x^α; target signal sample x^p; clean signal sample x^n; Loss Function J_{AL}.
Parameter: perturbation size ε=0.001, number of iterations T.
Output:the adversarial sample x^* that satisfy
$\|x^* - x_T\|_2 \le \varepsilon$.
$\alpha = \varepsilon/N$.
put x^α into f_L,obtain feature $f_L(x^\alpha)$;
put x^p into f_L,obtain feature $f_L(x^p)$;
put x^n into f_L,obtain feature $f_L(x^n)$;
  for t=0 to T-1 **do**
put x_n^* into f_L,obtain feature $f_L(x_n^*)$
Obtain the gradient $\nabla_{x^\alpha} J_{AL}$,
where $J_{AL} = L_{tri}(f_L(x_i^a), f_L(x_i^p), f_L(x_i^n))$;
calculate the accumulated gradient,renew the g_{n+1}:
  $g_{n+1} = \mu \cdot g_n + (\nabla_{x_n^*} J_{AL} / \|\nabla_{x_n^*} J_{AL}\|_1)$
update the x_{n+1}^* with gradient method
  $x_{n+1}^* = Clipx, \varepsilon\{x_n^* + \beta \cdot \text{sign}\ (g_{n+1})\}$
  end for
return$x^* = x_N^*$

Algorithm 1: AL-BIM.

Input: A classifier truncated model f_L; original signal sample to be attacked x^α; target signal sample x^p; clean signal sample x^n; Loss Function J_{AL}.
Parameter:perturbation size ε=0.001, number of iterations T.
Output:the adversarial example x^* that satisfy
$\|x^* - x_T\|_2 \le \varepsilon$;
$\alpha = \varepsilon/N$
put x_T into f_L,obtain feature $f_L(x_T)$;
put x^p into f_L,obtain feature $f_L(x^p)$;
put x^n into f_L,obtain feature $f_L(x^n)$;
put x_n^* into f_L,obtain feature $f_L(x_n^*)$;
  for t =0 to T-1 **do**
calculate $x_n^{nes} = x_n^* + \alpha \cdot \mu \cdot g_n$
put x_n^{nes} into f_L,obtain feature $f_L(x_n^{nes})$
Obtain the gradient $\nabla_{x^\alpha} J_{AL}$,
where $J_{AL} = L_{tri}(f_L(x_i^a), f_L(x_i^p), f_L(x_i^n))$;
  $g_{n+1} = \mu \cdot g_n + (\nabla_{x_n^{nes}} J_{AL} / \|\nabla_{x_n^{nes}} J_{AL}\|_1)$
update the x_{n+1}^* with gradient method
  $x_{n+1}^* = Clipx, \varepsilon\{x_n^* + \beta \cdot \text{sign}\ (g_{n+1})\}$
  end for
return$x^* = x_N^*$

Algorithm 2: AL-MIM.

activation layer before the final fully connected layer. Therefore, For activation L layer in BIM (AL-BIM), the attack process is as follows:

$$J_{\mathrm{AL}}(x_S, x_T, x_{\mathrm{adv}}) = L_{\mathrm{tri}},$$
$$L_{\mathrm{tri}} = \sum_{i}^{N} \left[\left\| f(x_i^a) - f(x_i^p) \right\|_2^2 - \left\| f(x_i^a) - f(x_i^n) \right\|_2^2 + \alpha \right]_+, \tag{16}$$

$$\begin{cases} x_0^* = x_S, g_0 = 0, \\ g_{n+1} = \mu \cdot g_n + \dfrac{\nabla_{x_n^*} J_{\mathrm{AL}}(x_T, x_n^*, x_{\mathrm{adv}})}{\left\| \nabla_{x_n^*} J_{\mathrm{AL}}(x_T, x_n^*, x_{\mathrm{adv}}) \right\|_1}, \\ x_{n+1}^* = \mathrm{Clip}x, \varepsilon\{x_n^* + \beta \cdot \mathrm{sign}\ (g_{n+1})\}, \end{cases} \tag{17}$$

where $\|\cdot\|_2$ is the L_2 norm, a similarity measure representing adversarial sample features and original sample features. So, the workflow of the AL-BIM method could be shown in Algorithm 1.

Activation L layer in MIM (AL-MIM) is similar to the AL-BIM algorithm; before calculating the gradient, a Nesterov x_n^{nes} needs to be calculated, and its workflow is shown in Algorithm 2.

3.5. Feasibility Analysis of Attack Methods

(1) The amount of information in the spectrum signal sample is small compared with the high-dimensional data of the image. If a classification model with an AM is used to extract the effective features of the attack object, it may improve the attack precision and intensity. The maximum misclassification effect is achieved with minimum perturbation of intensity. At the same time, after training different AMR models, the feature of the samples is transferable

(2) Based on the above considerations, this paper uses signal samples to extract effective features in the model, calculates the gradient from the feature level, and then attacks the proposed AMR model, which may achieve a higher misclassification rate with less fewer disturbances. Furthermore, from the feature level, it may better reflect the migration of the attack effect. Recently, the research material of the adversarial attack method has not been seen, which is based on the AM to extract effective features, and then adds disturbances by gradient calculation from the feature level

(3) Different from the traditional label gradient attack method, we must truncate the target model from the L layer because of the extracting features to obtain the truncated model and put x, x_t, and x_{adv} into f_L to obtain the original signal sample features $f_L(x)$, $f_L(x_t)$, and $f_L(x_{\mathrm{adv}})$, and the gradient of the feature is calculated

(4) Following the perturbation imposition, the modulation signal is sent into the target CNN for identification and classification. Given the high attack susceptibility of CNN, the classifier can be deceived by crafty perturbations, resulting in highly confident misclassifications. Section 4 will investigate how different parameters like perturbation levels and SNRs influence the CNN attacks and validate the attack feasibility and effectiveness by using the waveform and accuracy assessment methodologies. Figure 2 displays the block diagram for the adversarial attack assessment in modulation identification

Based on the advantages of the above feature level and the ternary loss function to reduce the Euclidean distance, we can propose the above algorithm with better transferable and concealment.

4. Experiment and Result Analysis

To test the effectiveness of adversarial ML on raw IQ-based AMR, the models we proposed are applying the model trained on Radio-ML2016.10a.

4.1. Experimental Data Set. Radio-ML2016.10a is a publicly available modulated signal data set from Bradley University, which is a data set used during the experiments in this research that employs GNU Radio to synthesize I/Q signal samples containing 11 modulation types, with signal-to-noise ratios ranging from -20 dB to 18 dB, uniformly distributed at 2 dB intervals. There are 128 complex floating point time samples in each signal. The data set is $220{,}000 \times 128 \times 2$ in size. The I and Q paths hold the real and imaginary parts of the 128 signal points, respectively.

In this research, we selected the signal in the data set with a high SNR greater than or equal to 10 dB. Furthermore, the number of samples in the training set is 35200, and test set data were classified by the proposed model to obtain 91.01%.

4.2. AMR Model. We should value the model we want to attack. If an AMR model recognition effect is poor, the effect of the attack may not be properly reflected considering that the spectrum signal and image have different characteristics and parameters. Aiming white-box attack, in the study, we develop a LSTM-AP model with an AM that performs perfect in modulation recognition. Aiming black-attack, to confirm the transfer for the attack, we present two AMR models, one is LSTM-AP, and another is CLDNN-AP; the specific model parameters will not be described. Figure 3 depicts the LSTM-AP model with attention mechanism for AMR. The signal embedding module is covered in the first section. Besides, the data format in RML2016.10a is 2×128, and it can be used as an input to LSTM, as IQ data input to LSTM. A learnable matrix is used in the fully integrated process of signal embedding to multiply data. Signal embedding is adopted because of the quite universal features of low-dimensional data, which necessitates the strengthening of the model's robust field through the continuous rise of the data dimensions. Variations in data dimensionality are

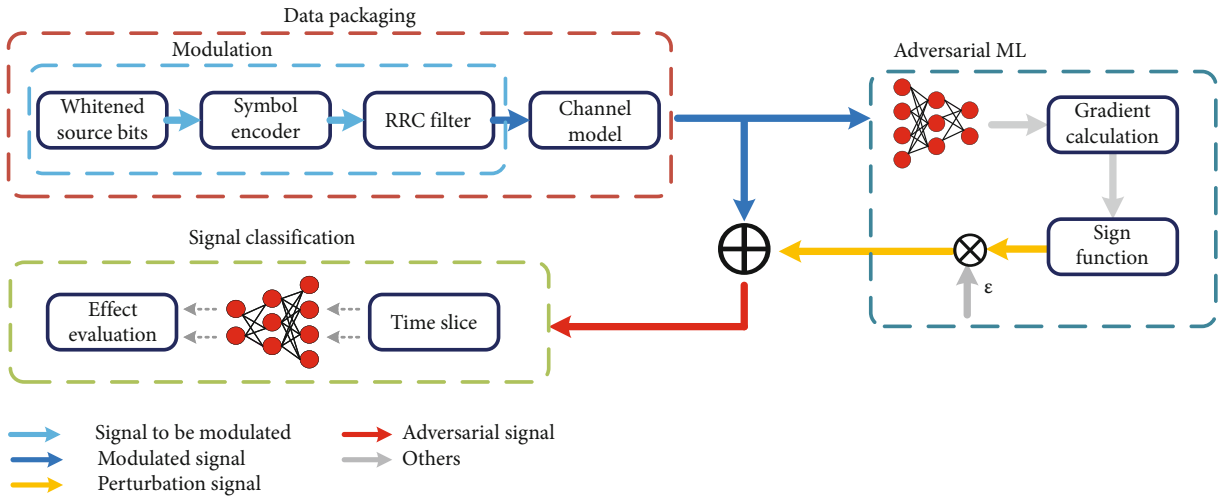


Figure 2: In this paper, the flow chart of modulation identification against attacks, the modulated signal is first data encapsulated, then the network gradient is obtained through the target network, and perturbation is added to the gradient to form an adversarial sample, which leads to the model recognition error.

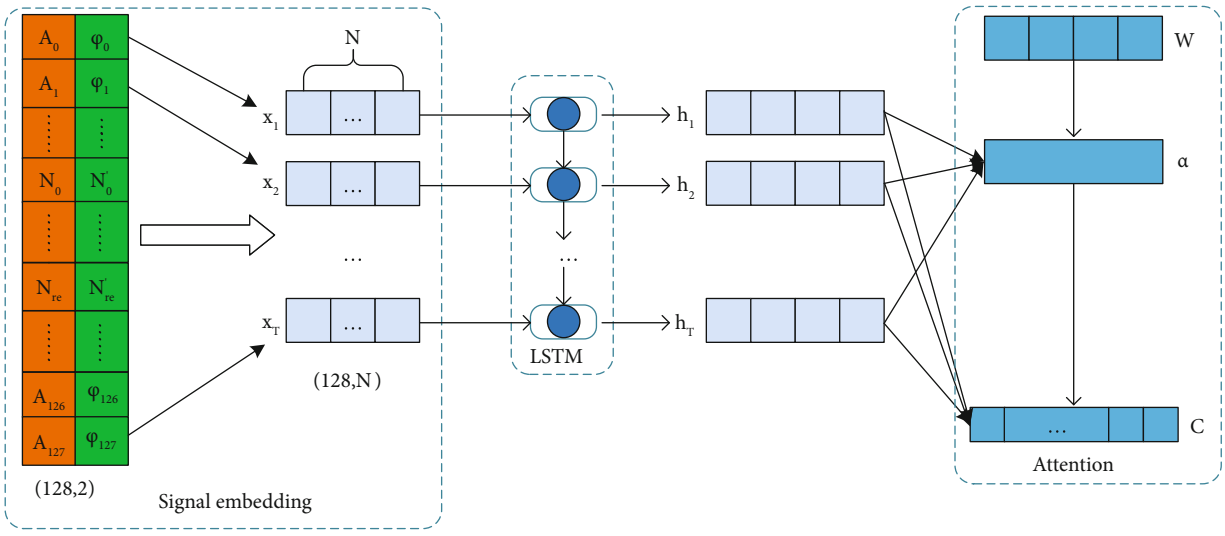


Figure 3: The proposed LSTM-AP model with attention mechanism for AMR.

derived via persistent model learning, and the best dimension for extracting features is sought ultimately. As a result of the embedding, the modulation information included by the input matrix will be larger and more accurate. The second part is the monolayer LSTM, which is excellent in acquiring temporal features like the time-series data of modulated signals and the information about phase and amplitude that varies by the mode of modulation. The final component is the AM module. The amplitude and phase information of a partial piece of data can be used by the AM to focus on the mode of modulation of a modulated signal sequence. Assume our input consists of T points of sequential signal data.

4.3. Evaluation Indicators. To assess the efficiency and transferability of the attack method in the current work, the following evaluation metrics are defined for the generated adversarial examples such as imperceptibility and signal properties.

(1) Attack success rate (ASR):

$$\mathrm{ASR} = \frac{\mathrm{ACC}_{\mathrm{ori}} - \mathrm{ACC}_{\mathrm{adv}}}{\mathrm{ACC}_{\mathrm{ori}}}. \tag{18}$$

ASR calculates the attackers percentage of misclassification, $\mathrm{ACC}_{\mathrm{ori}}$ is the classification accuracy of the original signal sample, and $\mathrm{ACC}_{\mathrm{adv}}$ is the classification accuracy obtained by the adversarial sample using the same classification model. The attack success rate can show an attack methods potential to cause misclassification.

Imperceptibility is as follows: L_0 norm and L_2 norm.

$$L_0 = \frac{C_c}{N}. \tag{19}$$

L_0 can be calculated as the proportion of the total number of points that a signal sample changes after an attack. C_c is the number of modified points in a sample (128×2 points), N refers to the number of data points in a signal sample, and the N value of the data set used in this paper is 256 (128×2).

$$L_2 = \sqrt{\sum_{i=1}^{N} |V_{\text{oi}} - V_{\text{ai}}|^2}. \tag{20}$$

L_2 calculates the numerical Euclidean distance between an original signal sample and an adversarial sample. V_{oi} indicates the value of the ith data point of the original sample, V_{ai} represents the value of the ith data point of the adversarial sample, and N is 256 (128×2).

(2) Signal characteristics: since the unique characteristics of the signal, we verify three indicators: ACR (amplitude change rate), APD (average phase difference), and PSR (perturbation signal rate)

ACR (amplitude change rate) is as follows:

$$A = \sqrt{I^2 + Q^2}, \tag{21}$$

$$\text{ACR} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{A_{\text{oi}} - A_{\text{ai}}}{A_{\text{oi}}}\right|. \tag{22}$$

ACR calculates the amplitude change rate of the signal before and after the attack. In the procedure of signal processing, different from the independent pixels in the image, there is a one-to-one correspondence between the I/Q channels in the signal data set, which are the sampling values of the real part and the imaginary part of the complex signal, if the reference is the same as the image, the calculation method ignores the correlation of the I/Q two-way. A represents the signals effective amplitude, while I and Q are the coefficients of the real and imaginary parts of the signal, respectively. A_{oi} is the effective amplitude of the ith sampling point of the original signal, A_{ai} denotes the effective amplitude of the ith sampling point of the signal after the attack, n represents the number of sampling points in a signal sample, and the value of n in the data set used in this study reaches 128. Different from each independent pixel in the image, for the 128×2 sample in the signal, it is more accurate to describe a signal sampling point by the matching I channel and Q channel than to regard it as 256 independent points.

APD (average phase difference) is as follows:

$$\text{APD} = \frac{1}{n}\sum_{i=1}^{n}\left|\arctan\frac{Q_{\text{oi}}}{I_{\text{oi}}} - \arctan\frac{Q_{\text{ai}}}{I_{\text{ai}}}\right|. \tag{23}$$

APD calculates the average phase difference at each sample point in a signal sample. The phase is an important factor to evaluate the signal attack. As an important measure to describe the change of the signal waveform, the delay of the phase can completely change a signal, thus making it impossible to extract the real message. I_{oi} is the real part coefficient of the i-th sample point of the original signal, and Q_{oi} indicates the imaginary part coefficient of the ith sample point of the original signal. I_{ai} indicates the real coefficient of the ith sampling point of the signal after the attack, and Q_{ai} refers to the ith sampling point of the original signal imaginary.

PSR (perturbation signal rate) is as follows:

$$P = \frac{\sum_{i=1}^{n} A_i^2}{n}, \tag{24}$$

$$\text{PSR} = \frac{P_P}{P_s}. \tag{25}$$

PSR analyses the power ratio of the disturbance craft adversarial sample to the signal P, where P is the signal power and A_i denotes the effective amplitude of the ith sampling point of the signal.

(3) TR (transition rate): in order to evaluate the transition of the attack, it is assumed that all signal samples are correct classification of white model f_w,and black model f_w. The original data set is $D_{\text{orig}} = \{(x^{(1)}, y^{(1)}_{\text{true}}), \cdots, (x^{(N)}, y^{(N)}_{\text{true}})\}$, and each attack method would generate an adversarial data set $D_{\text{adv}} = \{(x^1_{\text{adv}}, y^1_{\text{target}}, y^1_{\text{true}}), \cdots, (x^N_{\text{adv}}, y^N_{\text{target}}, y^N_{\text{true}})\}$. The data x_{adv} and y_{target} are obtained by the target attack performed by the original data set on the white-box model f_b. The mobility of adversarial examples refers to the number of samples that could deceive both the white-box model f_w and the black-box model f_b in the adversarial data set D_{adv} and the number of successfully deceived white box model f_w. Define the data set of successful deceiving white box model as $D_{f_w} \subseteq D_{\text{adv}}$. Then, mobility can be defined as follows:

$$\frac{1}{\left|D_{f_b}\right|}\sum_{(x_{\text{adv}}, y_{\text{true}}) \in D_{f_b}} 1[(f_b(x_{\text{adv}})) \neq y_{\text{true}}]. \tag{26}$$

This evaluation method intuitively shows the possibility that the adversarial examples generated in the white-box attack may potentially play a role in the black-box model.

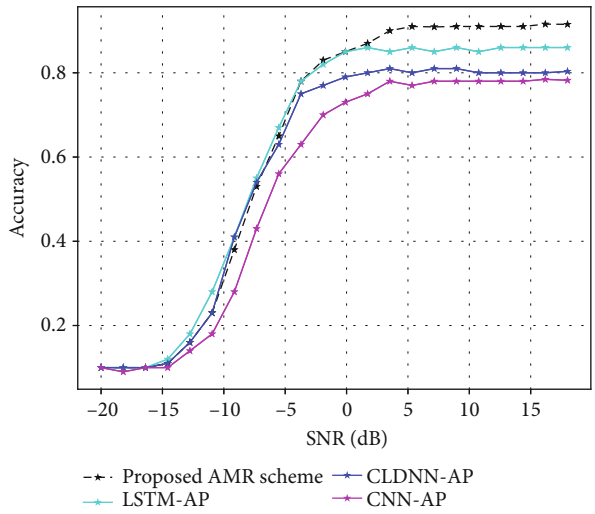


Figure 4: Recognition accuracy of different models.

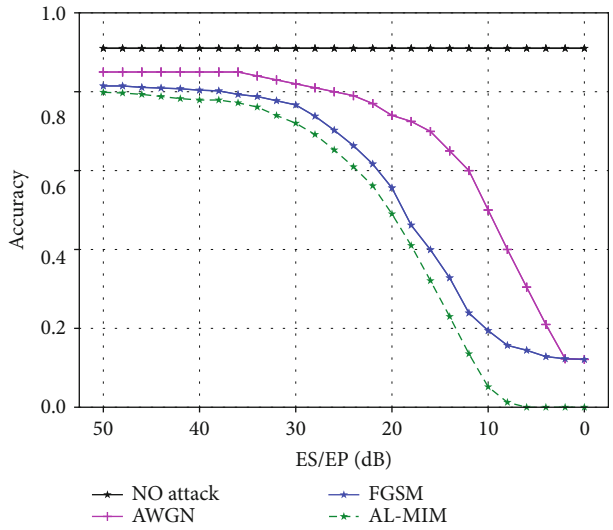


Figure 5: Changes in the recognition accuracy of different types of perturbation.

4.4. Experimental Result

4.4.1. Training and Results of Target Model. The number of iterations, learning rate, and other major parameters is consistent during the model training stage, to manage the efficiency and consistency of training. The number of iterations is set to 500, the learning rate is 0.001, and an automatic update mechanism is set: if the loss value of the test set does not drop for five consecutive times, the learning rate is decreased by halved. Additionally, considering that the modulated signal data set is composed of 20 SNRs, the data for each SNR has a different number. The characteristics suggest that the model be trained by combining all SNR data as a data set for training and then verifying its recognition accuracy on each SNR independently during verification.

Figure 4 compares the recognition accuracy of the proposed model to other three schemes: CNN-AP, CLDNN-AP, and LSTM-AP. CNN-AP has relatively low classification accuracy, demonstrating that CNN performs poorly when extracting features from time series data.. The efficiency is insignificant even when the CNN training data used is the IQ signal information about phase and amplitude, with mere maximum accuracy of 83.4% for the CNN-AP. Meanwhile, where the input of CLDNN-AP is the information about phase and amplitude, 85.2% accuracy of classification is attained. As displayed in Figure 4, when the LSTM input is the IQ data, the accuracy of classification is low. The reason is that the displayed phase and amplitude traits vary among modulation schemes, which are not reflected by the IQ data. The accuracy of classification is 87.13% at a SNR of 0 dB, and the average accuracy is 90.69% at a 0 dB SNR of 18 dB, showing a superior accuracy over the CNN-IQ design where training is accomplished based on the IQ data. The accuracy of classification with the present scheme is 89.2% at a SNR of 0 dB. Besides, the average accuracy at 0 dB SNR of 18 dB is 92.87%, and the maximum accuracy is up to 93.091%. As demonstrated by the simulations, our scheme outperforms the controls regarding classification accuracy.

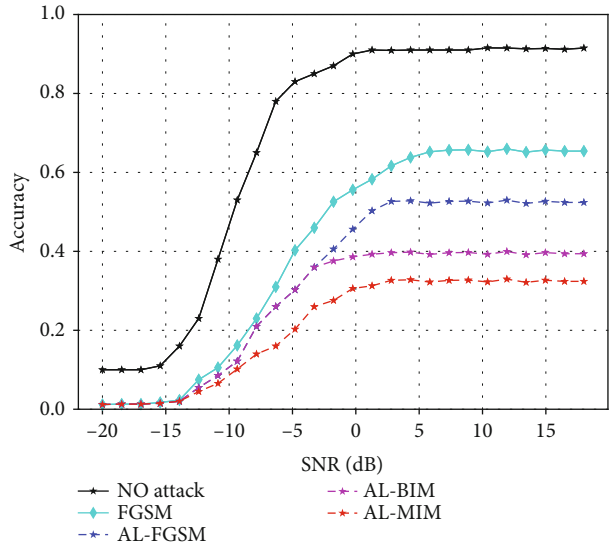


Figure 6: White-box untargeted attack.

4.4.2. White Attack. To investigate and analyze the impact of the attack on the modulation classification, this study compared the attack effect of the white-box methods in Figure 5. To demonstrate the effectiveness of our attack method, this paper selects the optimal recognition model proposed in this paper, uniformly selects the sample signal with a SNR of 18 dB, and then gives the recognition accuracy of the model under different signal-to-interference ratios (ES/E0). Figure 5 shows the results of the white-box attack. When the signal-to-interference ratio is insignificant, that is, the disturbance power is relatively large, and the accuracy of the FGSM method can be reduced to about 18% while the method proposed in this paper AL-MIM can be reduced to 0. In the range of 0-10 dB, we still attack the model to make its accuracy 0. From the overall trend, our proposed method is better than FGSM, and FGSM is superior to adding

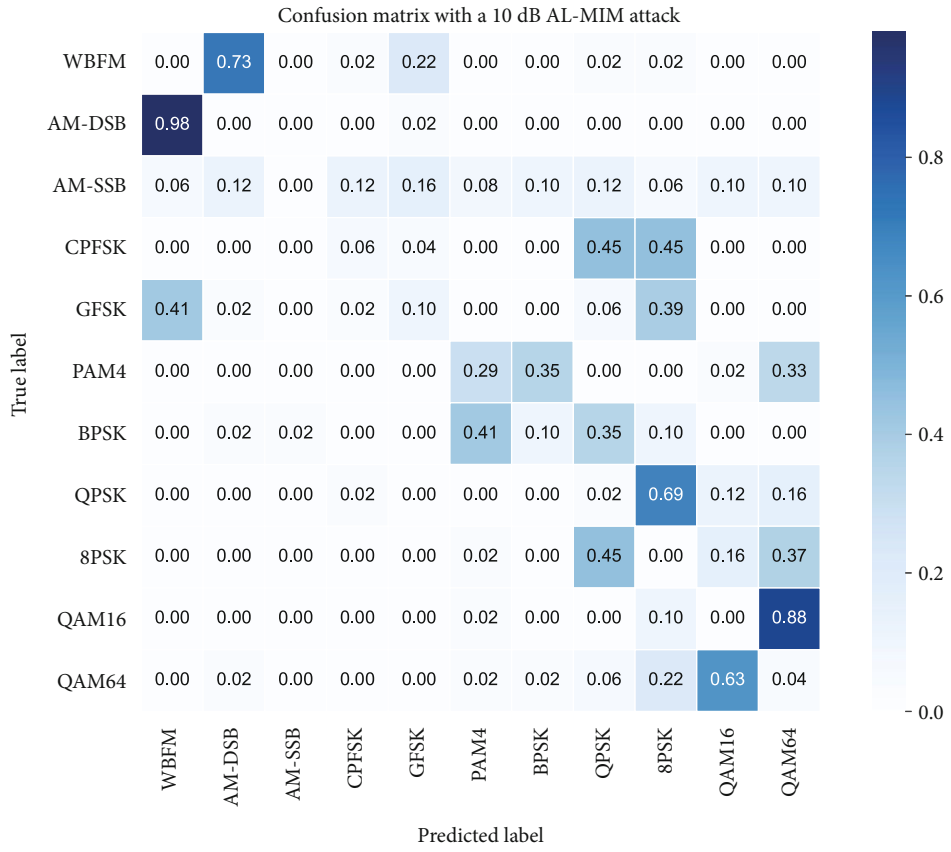


Figure 7: Confusion matrix of the AMR (SPR = 10 dB) predictions after AL-MIM attack.

ordinary white noise. The accuracy of about 90% of the training is reduced accordingly.

Furthermore, the modulation signal has the characteristic of different SNR values; so, we are carrying out the adversarial attack, and it is necessary to carry out the attack one by one for different SNR with $E_s/E_pE_p = 30$ dB. At the same time, the iterative attack is considered to achieve the best attack effect. Figure 5 shows the changes in the accuracy of the AMR scheme model based on the three attacks at -20-18 dB. According to Figure 6, with the SNR value added, the accuracy of the model's output shows an initially progressively improving trend and afterwards fluctuating around a specific value. As the only noniterative one-step attack algorithm, FGSM vary is fast, but the attack effect is not satisfactory. We could see that the two attack methods we proposed are better than FGSM and MIM. To deeply analyze the adversarial attack, Figure 7 presents a confusion matrix of the target model after yielding adversarial examples based on AL-MIM with SPR = 10 dB. It can be clearly found that there is an obvious chaotic impact on the type of modulation signals.

4.4.3. Black Attack. In contrast to the ideal experimental environment, the target model in the actual modulation signal recognition and communication adversarial environment is often invisible to the attacker, resulting in a black-box attack. That is, there are high requirements for the mobility of adversarial examples. Usually, the traditional attack uses alternative models to replace the target black-box model, and the black-box attack applied in the present work is a direct way to transfer the adversarial samples generated from the proposed scheme white-box attack to execute the attack with the purpose of better verifying the transferable of the adversarial example. Apart from that, the black-box attack is tested on two different network models, LSTM-AP and CLDNN-AP, respectively, and the experimental results are illustrated in Figures 8 and 9.

According to Figure 8, it can be observed that for the black-box model of LSTM-AP, the original adversarial samples that can bring down the target model in the white-box model have a significant decrease in the attack success rate when they are migrated to the LSTM model, especially for the label gradient-based attack method FGSM. In contrast, AL-FGSM, AL-BIM, and AL-MIM can still achieve better adversarial attack effect, reducing the accuracy rate of LSTM model drop to about 30%. A similar conclusion can be drawn from Figure 9, and the adversarial examples based on feature gradient can still maintain a good attack effect after transfer to the CLDNN-AP black box model even though it is not as efficient as the white-box attack.

Figure 10 represents the comparison of the transfer rate of the adversarial samples on the two black box models, where the FGSM method is not included in the comparison methods because of its poor attack. From Figure 10, it could

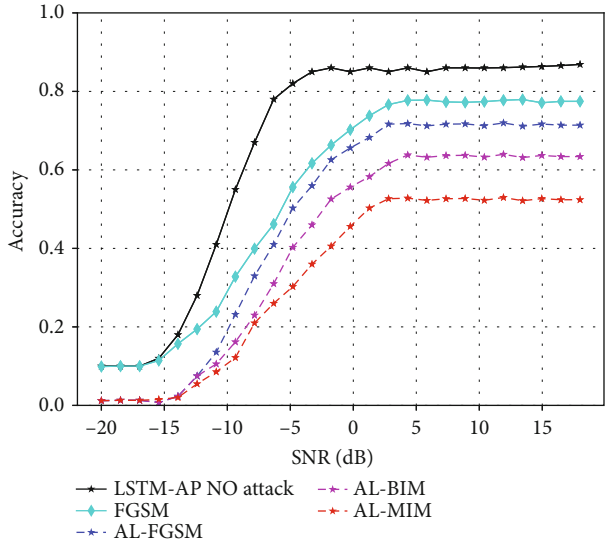


Figure 8: Black-box untargeted attack on LSTM-AP.

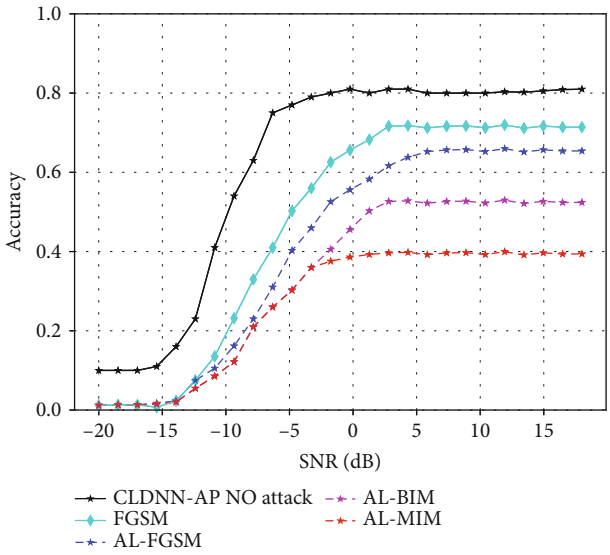


Figure 9: Black-box untarget attack on CLDNN-AP.

be seen that the black-box transfer rates of the two feature-based attack methods are higher than the traditional label-based methods for both LSTM-AP and CLDNN-AP models, which indicates that the feature-based attack methods have excellent attack transfer performance.

4.4.4. Analysis of the Effectiveness of the Attack. First, we make sure that the perturbation we introduce is small enough not to be recognized by the human eye with checking the effect of the perturbation on the signal fluctuations. The following modulation carrier formula is presented as

$$S(t) = I\cos(2\pi f t) + Q\sin(2\pi f t). \tag{27}$$

Furthermore, to the criterion of the success rate of the sample's attack on the model, the magnitude and intensity

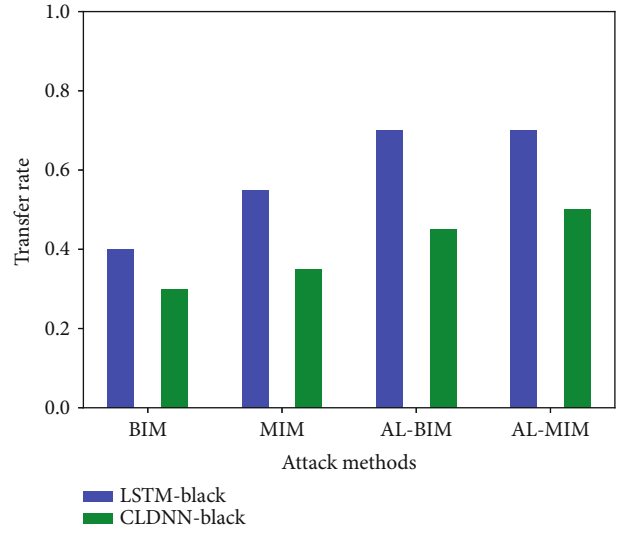
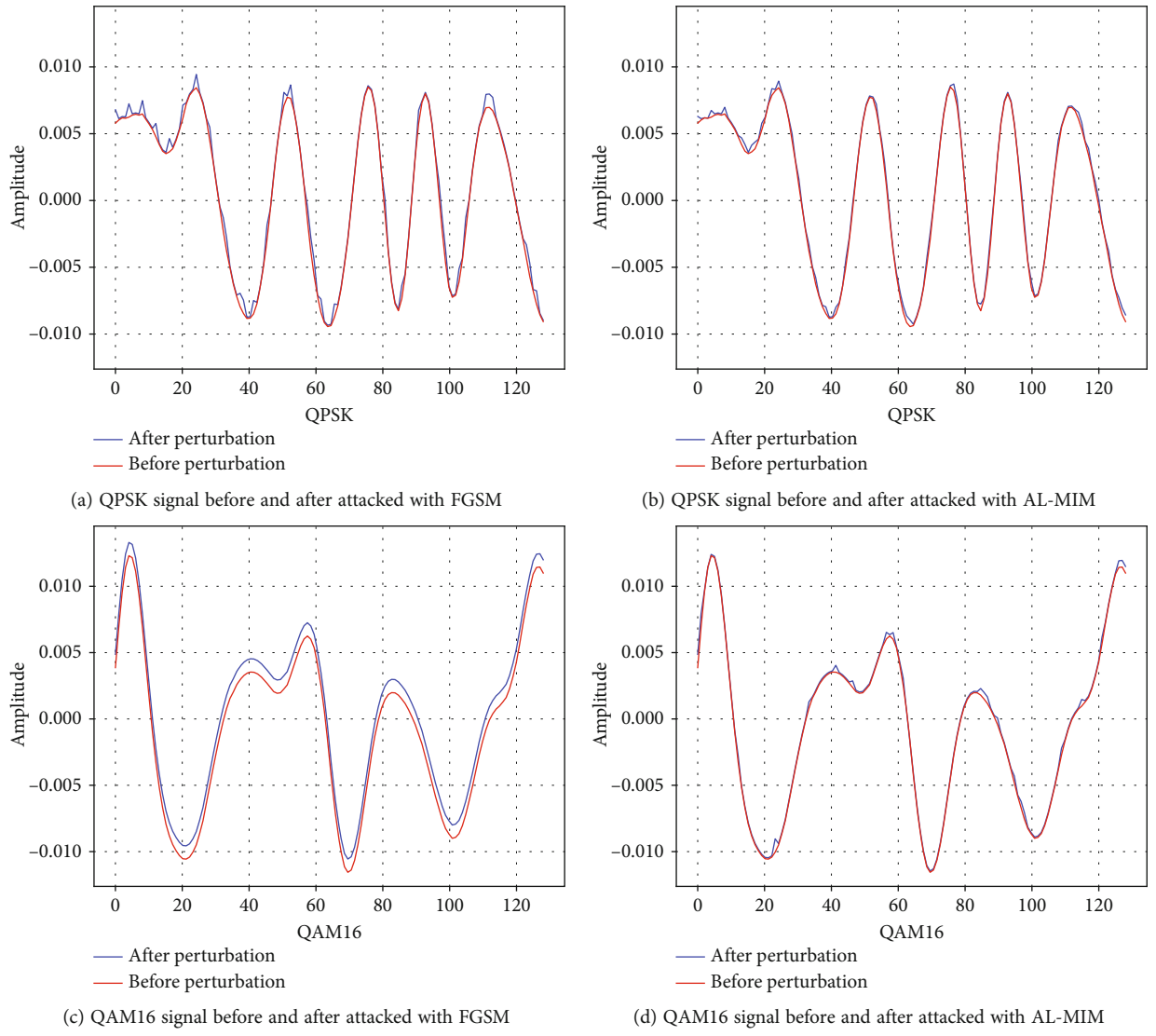


Figure 10: Black-box transfer rate.

of the perturbation of the modulated signal adversarial sample compared with the original sample are also important evaluation criteria. I represents the in-phase component, Q indicates the quadrature component, and f represents the carrier frequency. Subsequently, a primitive $S(t)$ signal can be yielded. By visualizing the $S(t)$, we could obtain the time domain waveform of the modulation signal. The time domain plots of the adversarial samples generated from the QPSK signal samples and their original signals are presented in Figures 11(a) and 11(b), while the plots of the adversarial samples generated from the QAM16 signal samples and their original signals are presented in Figures 11(c) and 11(d). Figures 11(a) and 11(c) show the signal perturbation based on the traditional label gradient method, while Figures 11(b) and 11(d) show the signal perturbation images based on the feature gradient AL-MIM method. It can be seen that for the same signal sample, the disturbance generated by the label gradient-based attack method often has continuous and violent jitter, which often does not suit to the image characteristics of a high signal-to-noise ratio modulated signal and is easily detected, and for the adversarial sample signal image of the feature gradient transferable attack, since the introduced perturbation is less in magnitude and jitter, it is more difficult to detect.

Then, to further analyze the adversarial attack approach, we selected high SNR value signals in the data set above or equal to 10 dB, with 32,000 samples in the training set. The results of the attack evaluation metrics will be presented, as shown in Table 1.

The attack method from misclassification and feature gradient attack outperform iterative attack and single step attack, and our method outperforms the traditional method from two metrics of imperceptibility. A signal is a physical quantity that representing a message, for example, an electrical signal can stand for different messages via changes in parameters such as amplitude, frequency, and phase. The signal is the carrier of the message, and in the process of signal attack, the variation of signal



(a) QPSK signal before and after attacked with FGSM

(b) QPSK signal before and after attacked with AL-MIM

(c) QAM16 signal before and after attacked with FGSM

(d) QAM16 signal before and after attacked with AL-MIM

Figure 11: Modulation samples with 18 dB before and after adding adversarial perturbation.

Table 1: Attack indicator results.

Type	Misclassification	Imperceptible		Signal characteristics			
Attack methods	ASR (%)	L_0	L_2	ACR	APD	PSR	TR
FGSM	95.46	0.85	5.14	1.41	0.30	-12.98	0.15
BIM	96.80	0.78	0.88	0.08	0.11	-23.38	0.40
MIM	97.45	0.70	1.01	0.11	0.14	-22.01	0.46
AL-BIM	98.90	0.50	0.95	0.07	0.10	-25.63	0.70
AL-MIM	99.97	0.32	0.22	0.06	0.03	-27.45	0.75

amplitude, phase, and other characteristics is extremely significant, and excessive distortion will make it difficult to extract the correct information. Four indicators (ACR, APD, PSR, TR) are used in this research to measure the distortion and migration rate of the signal. Based on Table 1, these performance indicators outperform the traditional attack methods.

5. Conclusion and Future Work

This paper addresses the security issues of the deep neural network model for AMR that is vulnerable to gradient attacks, and we propose a new adversarial attack method based on feature gradient transferability and design two attack algorithms, namely, AL-BIM and AL-MIM. These

methods aimed at feature attacks, which can extract and run regional attacks on the feature region of the original example captured by the neural network model by optimizing the triplet loss. The proposed scheme is more effective at attacking stable features in AMR-extracted signals, compared to the traditional label-based adversarial attack methods. Comprehensive experiments on public data sets show that the proposed feature gradient-based attack method in terms of attack method surpasses the traditional label gradient-based attack method in terms of attack success rate and transferability methods in both black-box attack and white-box attack scenarios. Additionally, the perturbation crafted using the feature gradient-based attack method is smoother and less perceptible. At the same time, four signal character indicators (ACR, APD, PSR, TR) are used in this research to measure the distortion and migration rate of the signal, and these performance indicators outperform the traditional attack methods. Further, decreasing the attack disturbance and narrowing the attack range are also our further research.

Data Availability

The simulation data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China (Grant No. 62101594, No. 61901520) and the Natural Science Foundation on Frontier Leading Technology Basic Research Project of Jiangsu under Grant BK20212001.

References

[1] T. J. OShea, J. Corgan, and T. C. Clancy, "Convolutional radio modulation recognition networks," in *Engineering Applications of Neural Networks*, C. Jayne and L. Iliadis, Eds., pp. 213–226, Springer International Publishing, Cham, 2016.

[2] O. Omotere, J. Fuller, L. Qian, and Z. Han, "Spectrum occupancy prediction in coexisting wireless systems using deep learning," in *IEEE Vehicular Technology Conference (VTC-Fall)*, Chicago, IL, USA, 2018.

[3] M. Sadeghi and E. G. Larsson, "Adversarial attacks on deep-learning based radio signal classification," *IEEE Wireless Communications Letters*, vol. 8, no. 1, pp. 213–216, 2019.

[4] Y. Lin, H. Zhao, X. Ma, T. Ya, and M. Wang, "Adversarial attacks in modulation recognition with convolutional neural networks," *IEEE Transactions on Reliability*, vol. 70, no. 1, pp. 389–401, 2021.

[5] R. Bryse Flowers, M. Buehrer, and W. C. Headley, "Evaluating adversarial evasion attacks in the context of wireless communications," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 1102–1113, 2020.

[6] B. Kim, Y. E. Sagduyu, K. Davaslioglu, T. Erpek, and S. Ulukus, "Over-the-air adversarial attacks on deep learning based modulation classifier over wireless channels," in *2020 54th Annual Conference on Information Sciences and Systems (CISS)*, Princeton, NJ, USA, 2020.

[7] L. Gao, Z. Huang, J. Song, Y. Yang, and H. T. Shen, "Push & pull: transferable adversarial examples with attentive attack," *IEEE Transactions on Multimedia*, vol. 24, pp. 2329–2338, 2021.

[8] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, *How transferable are features in deep neural networks?*, I. P. S. Neur, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds., pp. 3320–3328, 2014.

[9] S. Sabour, Y. Cao, F. Faghri, and D. J. Fleet, "Adversarial manipulation of deep representations," *ICLR*, Y. Bengio and Y. LeCun, Eds., 2016.

[10] N. Inkawhich, W. Wen, H. H. Li, and Y. Chen, "Feature space perturbations yield more transferable adversarial examples," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7066–7074, Long Beach, CA, USA, 2019.

[11] C. S. Weaver, C. A. Cole, R. B. Krumland, and M. L. Miller, *The Automatic Classification of Modulation Types by Pattern Recognition*, 1969.

[12] Z. T. Huang, J. Yang, X. Wang, X. Cui, and F. Y. Wang, "A survey of modulation recognition algorithms in noncooperative communication," *Science & Technology Review*, vol. 37, no. 4, pp. 55–62, 2019.

[13] J. L. Xu, W. Su, and M. Zhou, "Likelihood function-based modulation classification in bandwidth-constrained sensor networks," in *2010 International Conference on Networking, Sensing and Control (ICNSC)*, Chicago, IL, USA, 2010.

[14] P. Ghasemzadeh, S. Banerjee, M. Hempel, and H. Sharif, "Performance evaluation of feature-based automatic modulation classification," in *2018 12th International Conference on Signal Processing and Communication Systems (ICSPCS)*, Cairns, QLD, Australia, 2018.

[15] L. Huang, A. D. Joseph, B. Nelson, B. I. Rubinstein, and J. D. Tygar, "Adversarial machine learning," in *Proceedings of the 4th ACM workshop on Security and artificial intelligence*, pp. 43–58, New York, NY, USA, 2011.

[16] N. Akhtar and A. Mian, "Threat of adversarial attacks on deep learning in computer vision: a survey," *IEEE Access*, vol. 6, pp. 14410–14430, 2018.

[17] S. Rajendran, W. Meert, D. Giustiniano, V. Lenders, and S. Pollin, "Deep learning models for wireless signal classification with distributed low-cost spectrum sensors," *IEEE Transactions on Cognitive Communications and Networking*, vol. 4, no. 3, pp. 433–445, 2018.

[18] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2015, https://arxiv.org/abs/1409.0473.

[19] S. Chen, Y. Zhang, Z. He, J. Nie, and W. Zhang, "A novel attention cooperative framework for automatic modulation recognition," *IEEE Access*, vol. 8, pp. 15673–15686, 2020.

[20] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132–7141, Salt Lake City, UT, USA, 2018.

[21] Y. LeCun, K. Kavukcuoglu, and C. Farabet, "Convolutional networks and applications in vision," in *Proceedings of 2010*

IEEE international symposium on circuits and systems, pp. 253–256, Paris, France, 2010.

[22] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[23] T. N. Sainath, A. W. Senior, O. Vinyals, and H. Sak, "Convolutional, long short-term memory, fully connected deep neural networks," 2020, U.S. Patent No. 10, 783,900.

[24] Y. Chen, W. Shao, J. Liu, L. Yu, and Z. Qian, "Automatic modulation classification scheme based on LSTM with random erasing and attention mechanism," *IEEE Access*, vol. 8, pp. 154290–154300, 2020.

[25] C. Szegedy, W. Zaremba, I. Sutskever et al., "Intriguing properties of neural networks," 2013, https://arxiv.org/abs/1312.6199.

[26] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," 2014, https://arxiv.org/abs/1412.6572.

[27] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *Artificial intelligence safety and security*, pp. 99–112, Chapman and Hall/CRC, 2018.

[28] Y. Dong, F. Liao, T. Pang et al., "Boosting adversarial attacks with momentum," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, Salt Lake City, UT, USA, 2018.

[29] S. M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: a simple and accurate method to fool deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, Las Vegas, NV, USA, 2016.

[30] J. Lin, C. Song, K. He, L. Wang, and J. E. Hopcroft, "Nesterov accelerated gradient and scale invariance for adversarial attacks," 2019, https://arxiv.org/abs/1908.06281.

[31] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," 2016, https://arxiv.org/abs/1611.01236.

[32] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *2017 IEEE symposium on security and privacy (sp)*, San Jose, CA, USA, 2017.

[33] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok, "Synthesizing robust adversarial examples," in *International conference on machine learning*, Stockholm, Sweden, 2018.

[34] Y. Lin, H. Zhao, Y. Tu, S. Mao, and Z. Dou, "Threats of adversarial attacks in DNN-based modulation recognition," in *IEEE INFOCOM 2020-IEEE Conference on Computer Communications*, Toronto, ON, Canada, 2020.

[35] A. Jeddi, M. J. Shafiee, M. Karg, C. Scharfenberger, and A. Wong, "Learn2perturb: an end-to-end feature perturbation learning to improve adversarial robustness," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA, 2020.

Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 6136944, 15 pages
https://doi.org/10.1155/2022/6136944

Research Article

A Resource Allocation Scheme for Intelligent Tasks in Vehicular Networks

Jiujiu Chen,[1,2] **Caili Guo**,[1,2] **Chunyan Feng**,[1,2] **Chuanhong Liu**,[1] **Xin Sun**,[1] **and Jun Liu**[3]

[1]*Beijing Laboratory of Advanced Information Networks, Beijing University of Posts and Telecommunications, Beijing 100876, China*
[2]*Beijing Key Laboratory of Network System Architecture and Convergence, Beijing University of Posts and Telecommunications, Beijing 100876, China*
[3]*Tsinghua University, Beijing 100080, China*

Correspondence should be addressed to Jiujiu Chen; chenjiujiu@bupt.edu.cn

Received 25 May 2022; Accepted 25 August 2022; Published 27 September 2022

Academic Editor: Lisheng Fan

Lots of resource-consuming intelligent tasks need to be handled in vehicular networks, and traditional resource allocation schemes are hard to meet the intelligent demands. Therefore, this paper proposes a task-oriented resource allocation scheme for intelligent tasks in vehicular networks. First, we propose a task-oriented communication system and formulate a resource allocation problem, which is aimed at maximizing the task performance. Second, based on the system model, an intelligent task-oriented resource allocation optimization criterion is proposed, which is formulated as a mathematical model, and its parameters are solved by the proposed gradient descent-based algorithm. Third, to solve resource allocation problem, a multiagent deep Q-network-(MADQN-) based algorithm is proposed, whose convergence and complexity are further analyzed. Last, experiments on real datasets verify the performance advantages of our proposed algorithms.

1. Introduction

Recently, more and more vehicles are equipped with high-definition cameras to enhance visual perception [1]. More than 90% of the driving environment information can be collected and acquired by the cameras [2]. At the same time, the vehicles use artificial intelligence technology to fully analyze the large amount of data collected by the vehicular cameras, so as to complete the various tasks in the process of driving [3]. These intelligent tasks, based on a deep neural network, such as classification, detection, and recognition, put forward a huge demand for computing resource at the vehicle ends.

Thanks to the development of Internet of vehicles (IoV) technology in recent years [4], a feasible solution is that the vehicles transmit data to the edge server to complete the intelligent tasks. Then, the edge server feeds back the calculation results to the vehicles, so as to support the intelligent needs of various applications, such as assisted driving and automatic driving in IoV scenarios.

However, a large amount of multimedia data transmission and computing task offloading from the vehicles to the edge server bring great pressure to the communication resources. Therefore, in order to promote the integration of communication and computing processes in the IoV system, it is urgent to study efficient resource allocation schemes to improve the resource utilization rate in the vehicular networks and better serve the intelligent tasks of the vehicles.

1.1. Related Works and Motivations. Existing studies on resource allocation in the IoV are mainly oriented to network efficiency or user experience, with the purpose of maximizing quality of service (QoS) or quality of experience (QoE). Vehicle mobility and service diversity in the IoV scenarios lead to different QoS requirements, so QoS-based resource allocation schemes focus on how to build QoS models suitable for IoV scenarios [5, 6] and how to design resource allocation algorithms based on QoS models, such as channel selection [7], power control [8], and spectrum

sharing [9]. Although these QoS-based schemes can improve the traditional performance metrics such as capacity, they do not handle the subjective needs.

Compared to network efficiency, user experience-oriented demands are more subjective. QoE-based resource allocation schemes focus more on the needs of human users. Most works focus on QoE modeling to meet the diverse needs of humankind users [10]. For example, safety traffic services have higher requirements on video definition and resolution, while entertainment services have higher requirements on video smoothness. Moreover, some works pay attention to algorithm designing for resource allocation to meet QoE requirements in the IoV scenarios. The main method is to map QoE requirements to communication resource requirements such as spectrum and power [11, 12]. And the other is to establish the mapping or association between QoE and QoS and then use the multiobjective resource allocation methods to improve the quality of user experience [13, 14].

Most of the existing resource allocation schemes for the IoV scenarios mentioned above ignore the needs for intelligent tasks and seldom consider the contents or semantics of data. When the transmitted data is used for intelligent tasks such as classification, detection, and recognition, its goal is no longer network efficiency or user experience, but the understanding or analysis accuracy of visual contents or semantics, so the traditional QoS or QoE-based resource allocation schemes are no longer optimal [15]. A more suitable resource allocation scheme is needed for intelligent task-oriented communication system.

It is worth noting that the future network is becoming more and more intelligent. It is no longer simply concerned about the pursuit of transmission speed but pays more attention to the demands of intelligence [16]. The intelligent requirements of 6G make the research shift from the traditional communication based on Shannon's framework to the semantic or goal-oriented communication [17]. Recently, some works focus on intelligent end-to-end semantic communication systems based on deep learning [18, 19]. They propose that image or video features, instead of the full data, are uploaded to the servers for data analyses. Although transmitting features can save wireless resources, they are not suitable for all kinds of tasks since the full data is required to be archived in the server for future investigations. Moreover, even if they have achieved good results in solving the end-to-end intelligent tasks' requirements, they ignore the problems of resource limitation. Both sending and receiving ends require a lot of computing power, and it is difficult to support the implementation of those algorithms under the condition of resource shortage.

In the IoV scenarios where intelligence and network connectivity are highly integrated, it is urgent to explore the balance between network intelligence and efficient resource utilization. According to the above analysis, the main challenges are as follows:

(i) How to design a multimedia data transmission system for vehicles with limited computing capacity, to improve transmission efficiency and meet the needs of intelligent computing?

(ii) How to design an optimization criteria for resource allocation in intelligent scenarios of the IoV, to solve the contradiction between intelligent task requirements and traditional resource allocation methods?

(iii) How to design a resource allocation algorithm with high stability and low complexity to adapt to the dynamic changing environment of the IoV?

1.2. Contributions and Organization. To address these challenges, in our previous work [20, 21], we made preliminary exploration and proposed a single-task oriented spectrum allocation algorithm. In order to meet the requirements of multitasks and to extend the flexibility of resource allocation scheme in mobile scenarios, this paper further proposes a deep reinforcement learning-based resource allocation scheme for multiple intelligent tasks in vehicular networks, which are aimed at maximizing the performance of intelligent tasks under resource constraints. Contributions of this paper are as follows:

(i) We design a task-oriented communication system for multi-intelligent tasks in the IoV scenario. Based on the proposed system model, we construct a multivariable resource allocation optimization problem, which is aimed at maximizing the performance of intelligent tasks

(ii) We propose an intelligent task-oriented resource allocation optimization criterion, which is expressed as a mathematical model, and we design a gradient descent-based algorithm for solving model parameters

(iii) We propose a multiagent deep Q-network-(MADQN-) based algorithm to solve the resource allocation problem, and analyze the convergence and complexity of the proposed algorithm

In addition, we verify the performance advantages of the proposed algorithms based on the new dataset found by us and the existing datasets.

The rest of this paper is organized as follows. In Section 2, the system model is described and the resource allocation problem is formulated. In Section 3, a detailed description about resource allocation optimization criteria is given. In Section 4, a MADQN-based resource allocation algorithm is given. The numerical results are shown in Section 5, followed by the concluding remarks in Section 6.

2. System Model and Problem Formulation

2.1. Scenario and System Model. A typical scenario of task-oriented vehicular network is shown in Figure 1. The network consists of an edge server and multiple vehicles equipped with cameras. Vehicles are randomly distributed in the server's coverage area. Vehicles perform data collecting and preprocessing, and an edge server completes data storage and intelligent tasks, such as classification, detection, and reidentification (Re-ID).

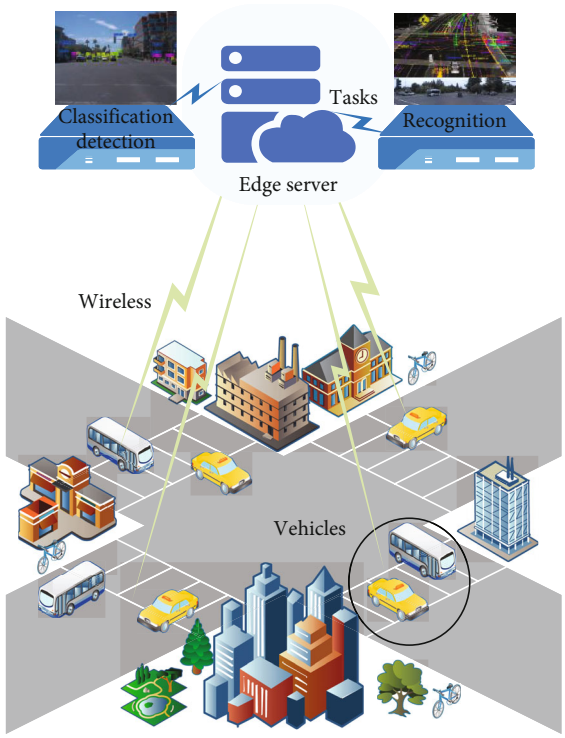


Figure 1: The task-oriented vehicular network scenario.

Specifically, vehicles perceive the surrounding environment through cameras, including traffic signals and moving objects (such as pedestrians and other vehicles). This visual environmental information, like image and video data perceived by vehicles, needs to be further analyzed. Considering the huge amount of collected multimedia data and the limited computing capacity of the vehicles, the data is transmitted to the edge server through wireless channels to complete the data storage, data analysis, and subsequent intelligent tasks.

Therefore, the goal of the communication system is to improve the tasks' performance as much as possible under the condition of limited resources. To achieve that goal, we design a task-oriented system, as shown in Figure 2.

The proposed system is divided into three parts: vehicles, wireless channel, and edge server. The vehicles contain three modules: (1) collecting module, which is used to collect original multimedia data and complete data preprocessing; (2) encoding module, completing the source coding and digital signal modulation; and (3) control module, which controls the transmission power and selects transmission frequency band and channel. A wireless channel is a mobile time-varying fading channel. The edge server consists of two modules: (1) decoding module, which completes the demodulation and decoding of signals to restore the original data, and (2) computing module, including data storage (for data analysis and model training), intelligent tasks performed by neural networks, optimization criteria modeling, and system resource allocation.

In this system, the vehicles select appropriate encoding data rate, power, and bandwidth according to the results of resource allocation and transmit the compressed data to the edge server through wireless channel. The edge server restores the data and then completes the tasks. The main purpose of the system is to achieve the best task performance by reasonable resource allocation, which is a multivariable resource allocation optimization problem.

2.2. Resource Allocation Problem Formulation. Due to the proposed communication system being task-oriented, the communication goal is to maximize the task performance. Based on the proposed system, task performance is mainly related to the quality of received multimedia data, which is mainly affected by compression rate and bit error rate. Hence, the resource optimization criteria can be formulated as

$$F_m^{\text{task}}(q_m, p_m) = \lambda_m \cdot F_m^{\text{task}}(q_m) \cdot F_m^{\text{task}}(p_m), \tag{1}$$

where $m \in \{1, 2, \cdots, M\}$ is the vehicle index, M is the maximum number of vehicles, F_m^{task} denotes the task performance such as accuracy or mAP, λ_m is the weight coefficient, and q_m and p_m are the compression rate and bit error rate, respectively.

Here, F_m^{task} denotes the task performance like accuracy and mAP, which is the metric of resource allocation schemes for intelligent tasks. From Equation (1), the task performance values are related to traditional communication performance metrics, compression rate, and bit error rate. The following is the explanation. According to Figure 2, the communication progress involves two processes, encoding and transmission, in which compression and noise are the key factors leading to data quality decline and affecting the performance of subsequent tasks. Therefore, Equation (1) describes the relationship between the performance of intelligent tasks and communication parameters. It reveals that task performance like accuracy and mAP can be improved by reasonable communication resource allocation. Besides, the accurate mathematical model of Equation (1) is the basis for resource allocation problems.

Accordingly, the key of optimization problem formulation lies in the calculation of compression rate and bit error rate.

2.2.1. Compression Rate. According to [22], the compression rate is related to the source coding scheme, data block size, and encoder packet number; hence, it can be expressed as

$$q_m = 1 - \frac{L_m \cdot G_m}{S_m}, \tag{2}$$

where L_m is the encoder packet length, G_m is the packet size, and S_m is the block size. Then, the encoded data rate is

$$R_m^B = \frac{L_m \cdot G_m}{T_m^G}, \tag{3}$$

where R_m^B is the data rate and T_m^G is the block duration.

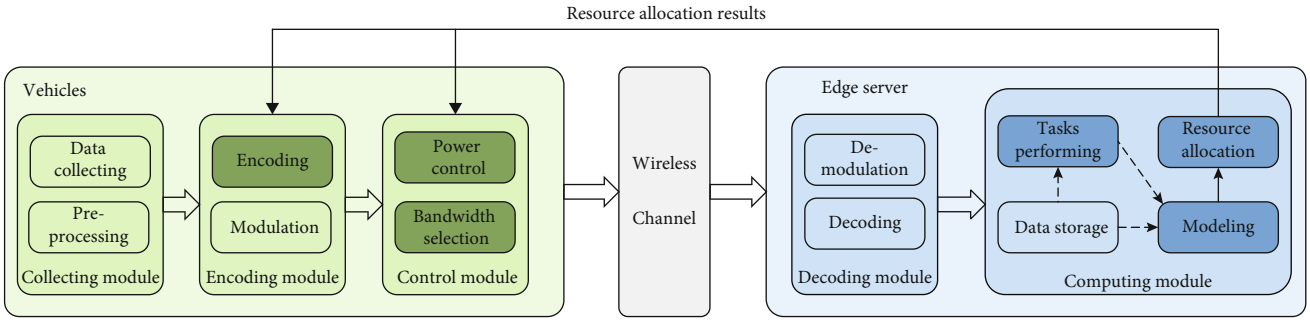


Figure 2: The structure of our proposed task-oriented communication system in a vehicular network.

2.2.2. Bit Error Rate. Based on [23], with quadrature amplitude modulation (QAM), the bit error rate is expressed as

$$p_m = \frac{2\left(1-\left(1/\sqrt{N}\right)\right)}{\log_2\left(\sqrt{N}\right)} Q\left(\sqrt{\left(\frac{3\log_2\left(\sqrt{N}\right)}{N-1}\right) 2E\left(\frac{P_m h_m}{n_0 R_m}\right)}\right), \tag{4}$$

where N is the modulation order. $Q(\cdot)$ is the Q-function, and $Q(x) = \int_x^{+\infty}(1/\sqrt{2N})e^{-t^2/2}dt$. $E(\cdot)$ is the expectation function. P_m is the transmission power, h_m is the channel gain, n_0 is the noise power spectral density, and $R_m = B_m \log_2(1 + (P_m h_m/n_0 B_m))$ is the transmission rate, where B_m is the bandwidth, P_m is the power, and h_m is the channel gain of each vehicle.

The channel gain h_m is modeled as an independent random variable, accounting for both large-scale fading h_m^L (contains path loss h_{pl} and shadowing h_{sd}) and small-scale fading effects h_m^S. Since the large-scale fading of channels is typically determined by vehicle locations, which do not change too much during transmission slots [24]. Here, the path loss is modeled as $h_{\text{pl}} = 148.1 + 37.6\log_{10}(d_m)(\text{dB})$, where d_m (in km) is the distance between the m-th vehicle and the edge server. Shadowing is modeled by using a log-normal distribution, with a standard deviation of 8 dB and zero mean [24]. However, the small-scale fading components might change. Considering the dynamic nature of the small-scale fading, we model the time-varying coefficients as independent first-order auto-regressive processes [25], given by

$$h_m^S(t) = \rho_m(t_e)h_m^S(t - t_e) + e_h, \tag{5}$$

where t_e is the time interval, e_h is the process noise sequence from a $\mathcal{CN}(0, 1-\rho_m^2(t_e))$ distribution, $\rho_m(t_e)$ is the channel autocorrelation function, and $\rho_m(t_e) = J_0(2\pi c v_m t_e/f_c)$, where $J_0(\cdot)$ is the zero-order Bessel function of the first kind, c is the velocity of light, f_c is the band mid-frequency, and v_m is the velocity of the m-th vehicle.

2.2.3. Optimization Problem. The goal of the proposed system is to transmit images and videos to the edge server, the server returns results in time. Therefore, the optimization problem is to allocate resources to transmit images or videos under constraints to achieve the best task performance, which can be formulated as

$$\begin{aligned}
\text{P1}: &\max_{L_m, B_m, P_m} \sum_{m=1}^{M} F_m^{\text{task}}(q_m, p_m) \\
\text{s.t.}\quad &\text{C1}: \sum_{m=1}^{M} \frac{L_m \cdot G_m}{T_m^G} \le R_{\max}^B \\
&\text{C2}: \sum_{m=1}^{M} B_m \le B_{\max} \\
&\text{C3}: \sum_{m=1}^{M} P_m \le P_{\max} \\
&\text{C4}: \frac{L_m \cdot G_m}{R_m(B_m, P_m)} \le \tau \\
&\text{C5}: F_m^{\text{task}}(q_m) \ge f_{\min} \\
&\text{C6}: F_m^{\text{task}}(p_m) \ge f_{\min} \\
&\text{C7}: q_m(L_m) \le q_e \\
&\text{C8}: p_m(B_m, P_m) \le p_e,
\end{aligned} \tag{6}$$

where $R_{\max}^B$, $B_{\max}$, and $P_{\max}$ denote the total data rate, power, and bandwidth of communication system, respectively. τ denotes the maximum transmission delay allowed by the system. $f_{\min}$, q_e, and p_e denote the task performance threshold, compression rate threshold and bit error rate threshold, respectively. The constraints are described in detail below.

C1 ~ C3 represents the constraints of resources, which means the sum of data rate, bandwidth, and power allocated to each vehicle are not larger than the total available data rate resource $R_{\max}^B$, the total available bandwidth $B_{\max}$, and the total available power $P_{\max}$, respectively. C4 represents the constraint of delay requirement, which means the transmission time must be shorter than the maximum delay τ allowed by the system. C5 ~ C6 represents the constraints of task performance requirements, in order to ensure that the results of intelligent tasks can meet the minimum requirements $f_{\min}$ of vehicle users. C7 ~ C8 represents the

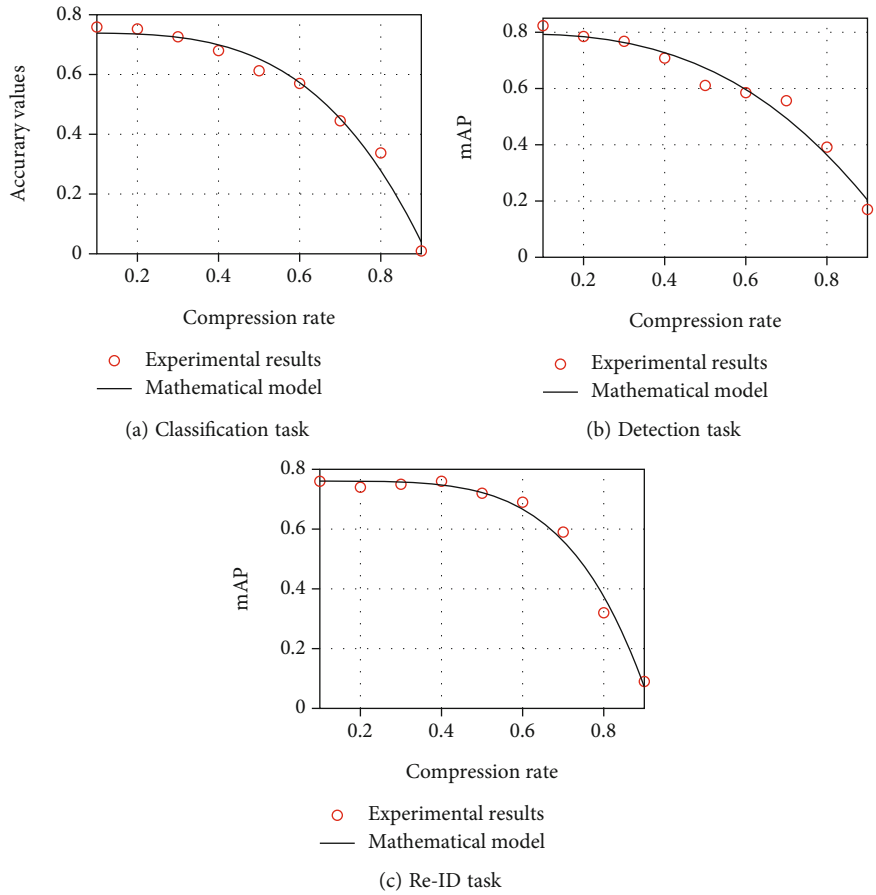


Figure 3: The relationship between task performance and compression rate with different tasks.

constraints of communication requirements. C7 represents the constraint of the compression rate, which guarantees the image and video quality. C8 represents the constraint of bit error rate, which guarantees the transmission quality.

3. Resource Allocation Criteria

In this section, we introduce the way to get the accurate mathematical models of optimization criterion F_m^{task}, which describe the mathematical relationship between the performance of intelligent tasks and communication parameters. Besides, these mathematical models are the basis for resource allocation problems. The steps of optimization criterion modeling are described as follows.

First is related data generation. In order to get the final mathematical model as $y = f(x)$, we first need to obtain a large number of data as (x, y), where x represents communication parameters and y represents the performance values of task completion. Due to communication progress involving two processes, encoding and transmission, in which compression and noise are the key factors leading to data quality decline and affecting the performance of subsequent tasks, in this paper, x represents different compression rates q_m and bit error rates p_m. Based on the proposed communication system, multimedia data is transmitted at different compression rates and bit error rates, and then, intelligent tasks are completed at the edge server to obtain the performance values such as accuracy and mAP of corresponding tasks.

Second is mathematical model selection. Based on the data obtained in the first step, scatter diagrams are drawn to analyze the trend, as shown in Figures 3 and 4. Then, according to the experimental results, the power function model is selected in this paper. Firstly, the power function model can reflect the corresponding relationship well, and secondly, it has monotonicity and is convenient for derivation.

Third is model parameter solution. Based on the large amount of data obtained in the first step and the power function model selected in the second step, mean square error (MSE) criterion is taken as the guidance to design the algorithm for solving model parameters. The flow is shown as Algorithm 1.

According to the proposed method, we obtain the models for different AI tasks, such as classification, detection, and Re-ID. The models reveal the relationship between task performance and communication parameters like compression rate or bit error rate. The mathematical formulas for the criteria models are as follows:

$$F_m^{\text{task}}(q_m) = a_1 \cdot {q_m}^{a_2} + a_3, \tag{7}$$

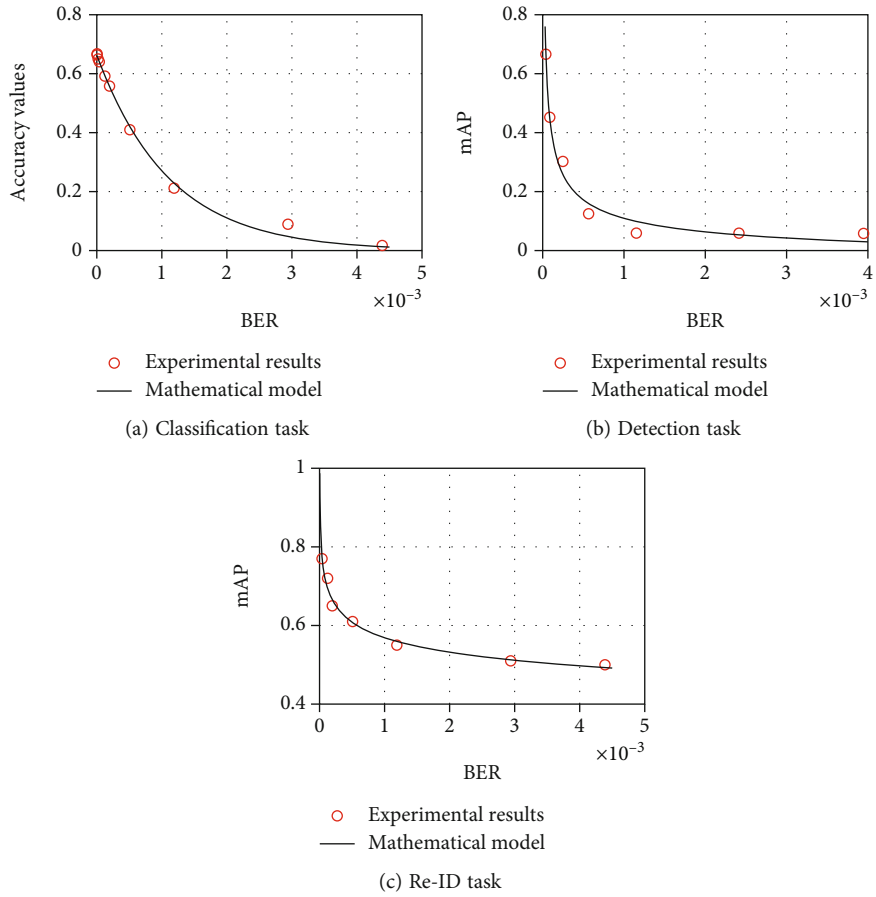


Figure 4: The relationship between task performance and bit error rate with different tasks.

Input: Initial parameters $\mathbf{a} = [a_1, a_2, a_3]$, $\mathbf{b} = [b_1, b_2, b_3]$, data set D, step length δ, and threshold L_0;
Output: Parameters $\mathbf{a}, \mathbf{b}$.
while $L_1(\boldsymbol{a}) \geq L_0$ or $L_2(\boldsymbol{b}) \geq L_0$ **do**
  $(q_m, F_m^{\text{task}}(q_m)^*), (p_m, F_m^{\text{task}}(p_m)^*) \in D$ **do**
    $F_m^{\text{task}}(q_m) = a_1 \cdot {q_m}^{a_2} + a_3$
    $F_m^{\text{task}}(p_m) = b_1 \cdot {p_m}^{b_2} + b_3$
  end for
  Compute the loss based on MSE:
    $L_1(\boldsymbol{a}) = 1/2D \sum_{d=1}^{D} (F_m^{\text{task}}(q_m) - F_m^{\text{task}}(q_m)^*)^2$
    $L_2(\boldsymbol{b}) = 1/2D \sum_{d=1}^{D} (F_m^{\text{task}}(p_m) - F_m^{\text{task}}(p_m)^*)^2$
  Compute the gradient of $\boldsymbol{a}$ and $\boldsymbol{b}$:
    $G_1(\boldsymbol{a}) = \partial L_1(\boldsymbol{a})/\partial \boldsymbol{a}$, $G_2(\boldsymbol{b}) = \partial L_2(\boldsymbol{b})/\partial \boldsymbol{b}$
  Gradient descent based update strategy:
    $\boldsymbol{a} \coloneqq \boldsymbol{a} - \delta G_1(\boldsymbol{a})$, $\boldsymbol{b} \coloneqq \boldsymbol{b} - \delta G_2(\mathbf{b})$
end while

Algorithm 1: Gradient Descent algorithm of resource allocation criterion modeling.

$$F_m^{\text{task}}(p_m) = \frac{b_1}{{p_m}^{b_2}} + b_3, \tag{8}$$

where F_m^{task} denotes the task performance like accuracy and mAP, q_m and p_m are the compression rate and bit error rate, respectively, and the other symbols are model parameters, which are solved by Algorithm 1.

In the flowchart, **a** and **b** represent the vectors of model parameters. The loss functions L_1 and L_2 represent the MSE values of $F_m^{\text{task}}(q_m)$ and $F_m^{\text{task}}(p_m)$, respectively, and L_0 is the MSE threshold. Moreover, G_1 and G_2 represent the

derivatives of loss functions L_1 and L_2, respectively, which are used to iteratively update the solutions of model parameters.

4. Resource Allocation Algorithm

According to the above analysis, we have gotten the models to describe the relationship between the AI task performances and communication metrics. To get the AI tasks done better in the edge server, reliable resource allocation should be carried out in the process of transmission, which means to maximize the accuracy or mAP performance. Hence, in this section, we first analyze the difficulties in solving the above optimization problem. Then, we propose a multiagent deep Q-network-(MADQN-) based algorithm to solve the problem.

4.1. Optimization Problem Analysis. Based on the above optimization problem, the proposed system transmission model and the scenario, the design of algorithm faces the following challenges.

4.1.1. Nonconvexity Problem. For the optimization problem, its objective function $F_m^{\text{task}}(q_m, p_m)$ is nonlinear, and its constraint conditions C4 and C8 are nonconvex and nonlinear (which can be derived by Equation (4), and its optimization variables are multidimensional. Therefore, the optimization problem P1 is essentially a NP-hard problem, which is difficult to be solved by convex optimization methods.

4.1.2. Dynamic Transmission Conditions. For the proposed transmission model, the channel gain h_m is dynamically changing, which leads to the traditional convex optimization algorithms, and heuristic algorithms are difficult to capture the dynamic characteristics of the channel.

4.1.3. High Requirements in the Scenario. For the proposed task-oriented dynamic vehicular network scenario, the complexity and stability of the algorithm are highly required.

Recently, deep reinforcement learning has shown strong advantages in solving resource allocation optimization problems in dynamic environments. Therefore, this paper considers using deep reinforcement learning to solve the challenges mentioned above, mainly based on the following.

By transforming the original nonconvex mathematical problems into sequential decision problems, deep reinforcement learning uses historical data and interaction with the environment to learn strategies and uses neural networks to approximate the optimal solution, so as to solve the NP-hard problems. One of the advantages of deep reinforcement learning is to solve the problems of dynamic environment [26], which refers to randomness under fixed state distribution, such as the channel model h_m proposed in this paper. The neural network fitting function itself can map the close set to the close domain so as to improve the robustness of dynamic learning. Another advantage of deep reinforcement learning lies in its strong generalization ability [27]. Through full offline training, online decision-making can greatly reduce the complexity, which fits well with the proposed transmission system proposed, as shown in Figure 2. In addition, introducing experience replay and iterative updating mechanism can also improve the convergence stability of deep reinforcement learning.

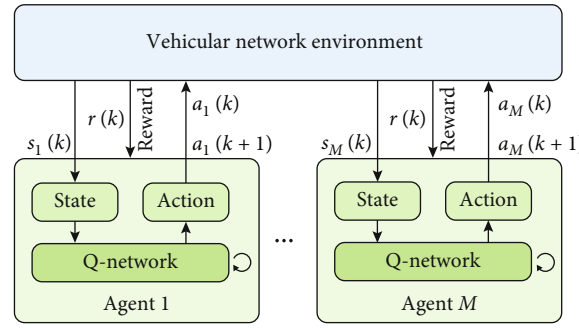


Figure 5: The proposed DQN algorithm flowchart.

4.2. Key Elements for Deep Reinforcement Learning. The DQN is a widely used deep reinforcement learning algorithm. Figure 5 is the flowchart of the proposed DQN algorithm in this paper, in which all vehicles act as agents to optimize the objective function in the optimization problem P1, namely, the performance metrics of the intelligent tasks. After performing an action $a(k)$, the agents can receive feedback information such as reward values $r(k)$ related to tasks' performance from the environment, update the current state $s(k)$, and then select the next action $a(k+1)$ according to the current state and action selection strategy by Q-network, until the convergence of the best strategy is obtained.

There are some key elements in the design process of DQN algorithm, especially to the optimization problem and system model in this paper.

4.2.1. Agent. We regard M vehicles in a vehicular network as the agents in the multiagent DQN algorithm.

4.2.2. Action. In the proposed scheme, the execution action contains three dimensions, namely, the change of the encoding packet length L_m, the bandwidth resource B_m, and the transmission power P_m of the m-th vehicle. There are three options for actions (increases, remains, or decreases) in each dimension. For example, $L_A = \{+\Delta L, 0, -\Delta L\}$ represents the change of the allocated packet length; the same goes for the changes of allocated bandwidth B_A and power P_A. Action space is represented as

$$A_m = \{\mathbf{a}_m | \mathbf{a}_m \in \{L_A \times B_A \times P_A\}\}, \tag{9}$$

where $\times$ denotes Cartesian product. Therefore, the action space size is $3 \times 3 \times 3$, and its dimension is 3. For example, $\mathbf{a}_m = (+,-,0)$ indicates that the current action is as follows: the packet length allocated to the m-th vehicle increases, the bandwidth decreases, and the power remains unchanged. Where considering that the packet length $L_m \in N^*$, the change of L_m is set as $\Delta L = 1$. In the discretization of B_m and P_m, $\Delta B = 0.001 B_{\max}$ and $\Delta P = 0.001 P_{\max}$, respectively.

4.2.3. State. We set the state of observation as the allocated resources of vehicles, as well as the channel state information

obtained from the environment, so the state space is expressed as

$$S_m = \{\mathbf{s}_m | \mathbf{s}_m \in \{L_m, B_m, P_m, h_m\}\}, \tag{10}$$

where the dimension of the state space S_m is 4.

4.2.4. Reward. The objective function is that all users execute the action strategy to maximize the total task performance under constraint conditions. To achieve this purpose, in the case of packet length, bandwidth and power changes, and channel state changes, the reward function is set as

$$r = \begin{cases} \sum_{m=1}^{M} F_m^{\text{task}}, & \text{if C1 ~ C8 are met,} \\ -10, & \text{otherwise.} \end{cases} \tag{11}$$

4.2.5. Q-Network. Figure 6 shows the Q-network structure. Each agent updates the network structure by using the data acquired by itself. It is worth noting that, considering the possible lack of computing power of the agent, the agent can upload the data to the server, and after the server completes the training, the network parameters are downlink transmitted to the agent. As shown in Figure 6, the Q-network contains three layers, namely, the input layer, the hidden layer, and the output layer.

The input dimensions are equal to the sum of the of state vector $\mathbf{s}_m$ dimensions, the hidden layer dimensions are H, and the output layer dimensions are equal to the action vector $\mathbf{a}_m$ dimensions, respectively. In addition, we use rectified linear unit (ReLU) as the activation function. For simplicity, the Q-function of agent m is expressed as Q_m. To train the DQN, we use a finite-size experience replay buffer Z to save the history transition samples $\mathbf{z}(k) = (\mathbf{s}(k), \mathbf{a}(k), r(k), \mathbf{s}(k+1))$, and old samples will be discarded when the storage is full. The two DQNs evaluation Q-network $Q_\theta(\cdot)$ and target Q-network $Q_{\theta^*}(\cdot)$ are used to approximate Q-function, with θ and θ^* being their weights, have the same structure. Here, θ is expressed as

$$\theta = (\omega_{\text{in}}, \mathbf{b}_{\text{in}}, \omega_{\text{out}}, \mathbf{b}_{\text{out}}), \tag{12}$$

where ω_{in} and $\mathbf{b}_{\text{in}}$ are parameters of the first fully connected layer from input layer to hidden layer and ω_{out} and $\mathbf{b}_{\text{out}}$ are parameters of the second fully connected layer from hidden layer to output layer, as shown in Figure 6. Moreover, the weight update of $Q_{\theta^*}(\cdot)$ is expressed as slowly approaching the weight of $Q_\theta(\cdot)$, as $\theta^* \longleftarrow \theta$, after C steps.

4.3. Multiagent DQN Algorithm. Algorithm 2 shows the process of multiagent DQN Algorithm. The algorithm mainly includes two parts: initialization process and reinforcement learning process.

Reinforcement learning is a process of repeated iteration; each iteration should solve two problems: give a strategy to obtain value function, and update the strategy according to

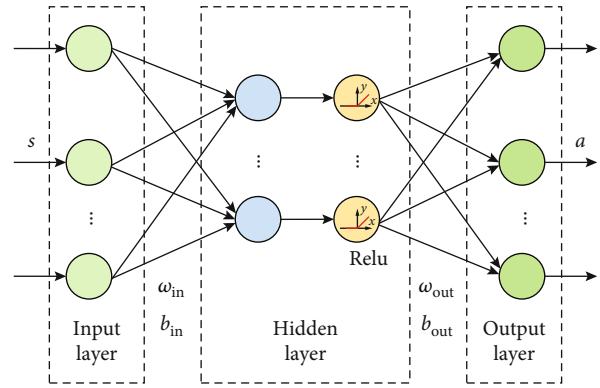


Figure 6: The proposed Q-network structure.

the value function. The iterative process is as follows: (1) The environment gives an observed state $\mathbf{s}$, and the agent obtains all $Q_\theta(\mathbf{s}, \mathbf{a})$ of state $\mathbf{s}$ based on the value function network, namely, Q-network $Q_\theta(\cdot)$. Then, the agent selects actions and makes decisions using ε-greedy strategy. (2) Upon receiving the action, the environment will give a reward and the next observation state. (3) The agent update the parameters θ of Q-network according to the loss function, and then enter the next step. (4) The cycle continues until a convergent Q-network is trained.

In the proposed DQN algorithm, experience replay mechanism and target network are introduced: (1) Experience replay is used to solve the problem of data correlation, that is, to store the experienced data in a buffer and extract a part of the data from the buffer for each parameter updating, so as to ensure that the training samples are independent and equally distributed, improve the utilization rate of data, and make the model better converge. (2) Target network is used to solve the problem of value function fluctuation in the iterative process. By using target network, the model calculating the target value will be fixed in a period of time, which can reduce the volatility of the model and make the training more stable.

In addition, the training effect of DQN algorithm is related to its main parameters. After many trials, the main parameters of the algorithm in this paper are set as follows: the learning factor μ is set at 0.001, so that the algorithm retains most of the historical training results and pays more attention to past experience, and the discount factor λ is set as 0.9, allowing the algorithm to consider 90% of the next reward and pay more attention to long-term rewards. The coefficient of ε-greedy strategy is set as $\varepsilon = 1/\sqrt{k}$, so that the algorithm strikes a balance between exploration and exploitation.

4.4. Convergence and Complexity. First, we discuss the convergence of the proposed algorithm: The convergence of DQN is difficult to prove directly theoretically, but the training of DQN is stable due to the introduction of the mechanism of experience replay and target network. It can be proved that network parameters can converge to a very small interval, so that a stable approximate optimal solution can be obtained through DQN. The proof process is shown in Appendix A.

Input: action space A, state space S, learning factor μ, discount factor λ, and other system parameters;
Onput: Target Q-network $Q_{\theta^*}(\cdot)$.
Initialization Process:
for each agent m **do**
    Initialize replay buffer Z_m;
    Initialize the weights of Q-network θ;
    Initialize the weights of target Q-network $\theta^* = \theta$.
end for
Learning Process:
for each agent m **do**
    repeat
        Initialization state $\mathbf{s}$
        for $k = 1, K$ **do**
        Generate a random number $x_k \in [0, 1]$;
        if $x_k < \varepsilon$ **then**
            Select a random action $\boldsymbol{a}$ from A;
        else
            Select the action $\boldsymbol{a} = \arg\max Q_\theta(\mathbf{s}, \boldsymbol{a})$.
        end if
        Perform action $\boldsymbol{a}$;
        Get new state $\mathbf{s}'$ and reward $r(k)$ by Equation (11);
        Store transition $\mathbf{z}(k) = (\mathbf{s}, \boldsymbol{a}, r, \mathbf{s}')$ in Z;
        Sample random transitions $(\tilde{\mathbf{s}}, \tilde{\mathbf{a}}, \tilde{r}, \widetilde{\mathbf{s}'})$ from Z;
        $$y = \begin{cases} \tilde{r}, & \text{if } \widetilde{\mathbf{s}'} \text{ is the termination state;} \\ \tilde{r} + \gamma \max \mathrm{Q}_{\theta^*}(\widetilde{\mathbf{s}'}, \widetilde{\mathbf{a}'}), & \text{otherwise.} \end{cases}$$
        Use $L(\theta) = (y - Q_\theta(\tilde{\mathbf{s}}, \tilde{\boldsymbol{a}}))^2$ as a loss function to train Q-network;
        Updata the weights: $\theta \longleftarrow \theta - \mu \cdot \nabla\theta$;
        $\mathbf{s} \longleftarrow \mathbf{s}'$;
        $\theta^* \longleftarrow \theta$, after C steps.
        end for
    until $Q_{\theta^*}(\cdot)$ converges.
end for

Algorithm 2: Multiagent DQN algorithm.

Second, we discuss the complexity of the proposed algorithm: The complexity of DQN can be divided into two parts: training complexity and inference complexity. The training complexity is related to the number of back propagation and iterations. Each iteration requires one back propagation, and one back propagation requires four parallel derivative operations. Assuming that the time complexity of a derivative operation is $O_\partial(1)$, the time complexity of the training stage is

$$O_{\text{Training}} = O_\partial(4 * K). \tag{13}$$

Inference complexity is mainly related to Q-network structure. As shown in Figure 6, each inference needs to go through two fully connected layers, so the computational complexity in the reasoning stage can be expressed by the required multiplication times, which is

$$\begin{aligned} O_{\text{Inference}} &= (d_{\text{in}} * d_{\text{h}} + d_{\text{h}}) + (d_{\text{h}} * d_{\text{out}} + d_{\text{out}}) \\ &= (d_{\text{in}} + d_{\text{out}} + 1)d_{\text{h}} + d_{\text{out}}, \end{aligned} \tag{14}$$

where $d_{\text{in}}, d_{\text{h}}, d_{\text{out}}$ are dimensions of input layer, hidden layer, and output layer, respectively, and the dimensions represent the number of neurons of each layer. The consumed time of online resource allocation depends on the inference complexity of the proposed DQN algorithm. The proposed multiagent DQN is distributed, its training complexity is only related to the number of iterations, and its inference complexity is only related to the dimensions of each layer, which are fixed. The complexity will not expand exponentially with the increase of vehicles. Therefore, the proposed algorithm is suitable for scenarios with low delay and high access.

5. Numerical Results

In this section, numerical results are provided for the performance evaluation of the proposed algorithms. Firstly, the parameter settings of communication system simulation and intelligent task experiment are introduced. Secondly, the datasets used in this paper and the related schemes are introduced. Thirdly, the performance evaluation results of Algorithm 1 are presented and analyzed. Lastly, the performance evaluation results of Algorithm 2 are presented and analyzed.

5.1. Parameter Setting. The communication simulation system is constructed based on Simulation of Urban Mobility (SUMO), Matlab R2019a, and Pycharm 2019.1.1 platform. Firstly, the simulation scenarios of vehicular networks are based on the urban intersection scene generated by SUMO. Each road includes 4 two-way lanes 3.5 m wide, and the initial positions of vehicles are randomly generated. Secondly, the Winner model in the 3GPPTR 36. 885 standard is used for the channel model, as described in Section 4. Meanwhile, the small scale fading caused by vehicle movement and construction is considered, and the time-decrement scale fading model shown in Equation (5) is adopted. The specific system simulation parameter settings are based on 3GPP TR 36.885, as shown in Table 1. Lastly, considering the time-varying unsteady characteristics of the vehicular networks in practice, 200 independent Monte Carlo simulations are used in each experiment to take the average value to eliminate errors caused by abnormal data

In the experiments of intelligent tasks, the GPU model is NVIDIA GeForce RTX 3090, and the training and testing environment is Windows 10 + CUDA 10.2. The intelligent tasks adopted in this paper include classification, object detection, and Re-ID. Deep learning based experimental parameter settings are shown in Table 2, mostly based on experience. In addition, JPEG 2000 is used for image compression and coding, and HEVC is used for video compression and coding.

5.2. Datasets and Related Scheme Introduction. The datasets used in this paper include four existing datasets and the new dataset constructed in this paper. The datasets are summarized in Table 3. The STL-10 [28] dataset mainly includes image data and is applicable to classification tasks. Caltech [29] and Waymo [30] datasets are capture by real vehicular cameras in driving scenarios, which are mainly used for object detection tasks in vehicular networks. Market-1501 [31] mainly contains image data and is suitable for Re-ID tasks. In addition, a semantic communication-oriented dataset [32], namely, SCO, containing 5100 images and 10 video clips, is also constructed for classification tasks and object detection tasks.

To verify the performance advantages of the proposed resource allocation scheme, three comparison schemes are used. Based on the main characteristics of each scheme, the scheme names are all abbreviated. The schemes are detailed as follows:

(i) TPO-JRA scheme (Task Performance-oriented Joint Resource Allocation scheme): it is the proposed scheme in this paper, which is aimed at maximizing the performance of intelligent task, and the optimization variables include data rate resource, bandwidth resource, and power resource. The optimization algorithm is based on MA-DQN

(ii) CPO-RA scheme (Content Priority Oriented Resource Allocation Scheme): in this scheme, the optimization objective is to maximize the effective information, the definition of which is shown in [20]. The optimization variables include power resource and bandwidth resource, and the optimization algorithm is based on Q-learning

(iii) QoC-RA scheme (Quality of Content-based Resource Allocation scheme): in this scheme, the optimization objective is to maximize content quality, which is defined in [15]. Under the object detection task, that is, the average detection accuracy, the optimization variable is bit-rate resource, and the adopted optimization algorithm is based on convex optimization

(iv) MRA scheme (Mean Resource Allocation scheme): in this scheme, all resources are evenly allocated to each user. This scheme serves as a baseline for other schemes to verify the performance gain of the proposed scheme

Table 1: Communication system parameters.

Parameters	Values
The edge server coverage	0.5 km
Number of vehicles M	3 ~ 18
Range of vehicle speed $v_m, \forall m$	0 ~ 60 km/h
The modulation order N	4
Noise power spectral density n_0	-174 dBm/Hz
Band mid-frequency f_c	2GHz
The time interval of small-scale fading t_e	50 ms
The packet size G_m	20 KB
The video block size S_m	400 KB
The video block duration f_c	0.1 s
The total data rate $R^B_{\max}$	200 ~ 700 Kbps
The total bandwidth $B_{\max}$	0.1 ~ 0.9 MHZ
The total power $P_{\max}$	5 ~ 45 dBm
The minimum performance limit $f_{\min}$	0.1
The delay threshold range τ	50 ~ 300 ms
The compression rate threshold q_e	0.95
The bit error rate threshold p_e	0.001

In addition, in order to ensure the fairness of comparison, the same algorithm is adopted in this paper under the same scenario and intelligent task.

5.3. Performance Results of Model Solving Method. Figures 3 and 4 show the relationship between task performance and compression rate and bit error rate, respectively. Task performance decreases with the increase of compression rate and bit error rate. This is because the semantic or content of the original data is lost due to lossy compression or noise interference, so that the machine at the edge server cannot correctly identify the semantic or content of the original data through deep learning networks. Secondly, the relationship between task performance and compression rate presents the nature of concave function, while the relationship

Table 2: Deep learning-based experiment parameters.

Parameters	Values
The learning factor μ	0.001
The discount factor γ	0.9
The update frequency of target Q-network C	5
The maximum number of iterative steps K	10^4
The size of replay buffer Z_m	200
The dimensions of input layer d_{in}	4
The dimensions of hidden layer d_{h}	10
The dimensions of output layer d_{out}	3

Table 3: The adopted datasets for intelligent tasks.

Tasks	Networks	Datasets
Classification	ResNet-18	STL-10,SCO
Re-ID	ResNet-50	Market-1501
Detection	Faster-RCNN	Caltech, Waymo, SCO

between task performance and bit error rate presents the nature of convex function. This is because they have different ways of semantic distortion. The lossy compression mainly leads to blur distortion, and the noise mainly causes the symbol errors in the decoding process. In addition, the curves of each task have differences, which is because different tasks are based on diverse deep learning networks, which have various degrees of sensitivity to compression distortion and noise.

Table 4 shows the numerical results of model parameters and RMSE metric, and all numerical results are reserved for 4 significant digits. RMSE reflects the degree of agreement between the predicted values based on the mathematical model and the actual values. For the proposed method, in the classification, detection, and Re-ID tasks, the RMSE values of model Equation (7) are 0.03438, 0.04303, and 0.02984, respectively, and the RMSE values of model Equation (8) are 0.04078, 0.03739, and 0.01633, respectively. All of them are less than 0.05, reflecting that the modeling method (Algorithm 1) has good accuracy performance and can reflect the relationship between task performance and system variables.

5.4. Performance Results of Resource Allocation Scheme. In this section, we use the task performance values related to F_m^{task} in Equation(1) to verify the advantages of the proposed scheme. The experiments include three different intelligent tasks, classification, detection, and Re-ID tasks. A different task is related to different metric, such as accuracy or mAP, but the values are between 0 and 1. Hence, we use the weighted average values of different vehicles with different kinds of tasks to present the performance metric of the whole system. The larger the values are, the better intelligent tasks of the system are completed.

5.4.1. Performance versus Resource Parameters. Figures 7–9 show the curves of task performance versus bit rate, bandwidth, and power, respectively. With the variation of resource parameters, the proposed resource allocation scheme has the best performance, which verifies the effectiveness of the proposed scheme. Combined with Figures 7–9, the following conclusions can be drawn: Firstly, optimization of resource allocation is crucial to the performance improvement of intelligent tasks, which indicates that in intelligent task-oriented communication system, resource optimization can further ensure the correct understanding of transmitted data by the server, rather than only pursuing the improvement of computing capacity of the server. Secondly, when resources are limited, it is necessary to jointly optimize different types of resource parameters, because the influence of various resource parameters on video quality is coupled in the transmission process.

In addition, within the numerical range of experimental settings, the performance gains brought by optimization of different resource parameters are inconsistent, which is due to the different mathematical relationship between resource parameters and the objective function. There is an inequality of average performance gains with resource parameters, $\Delta\bar{F}(P_m) > \Delta\bar{F}(L_m) > \Delta\bar{F}(B_m)$, which indicates that in the actual environment, the allocation of power should be given

Table 4: The results of model solution.

Parameters	Classification	Detection	Re-ID
a_1	-1.017	-0.7849	-1.155
a_2	3.551	2.699	4.908
a_3	0.7387	0.7940	0.7606
b_1	-6.420	0.008075	0.3320
b_2	-0.3953	0.4442	0.08931
b_3	0.7364	-0.06427	-0.04626
RMSE-a	0.03438	0.04303	0.02984
RMSE-b	0.04078	0.03739	0.01633

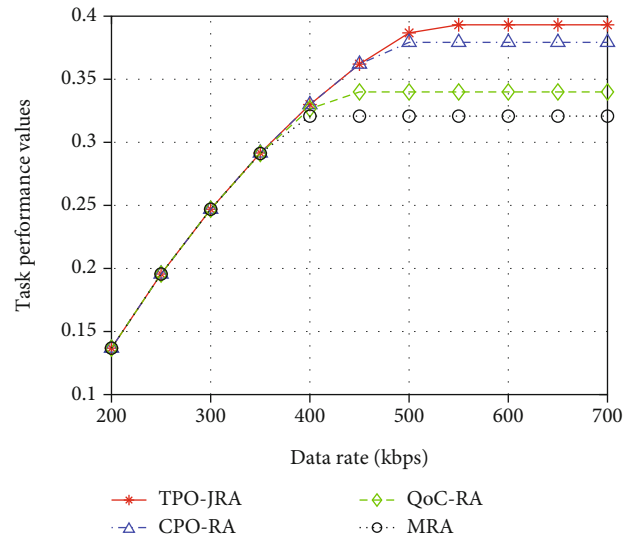


Figure 7: Task performance versus data rate resource.

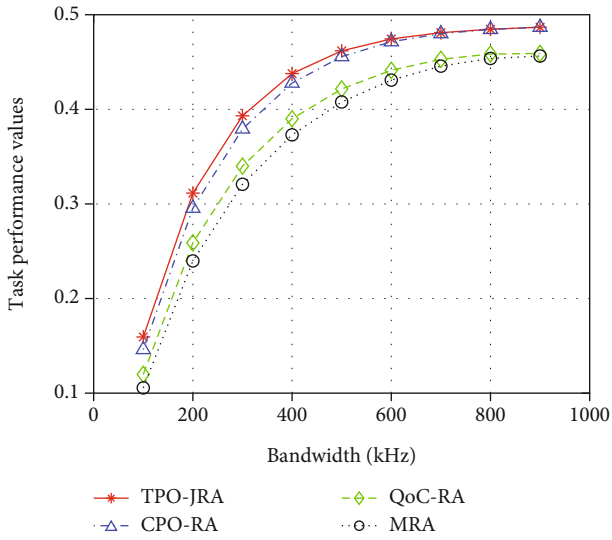


Figure 8: Task performance versus bandwidth resource.

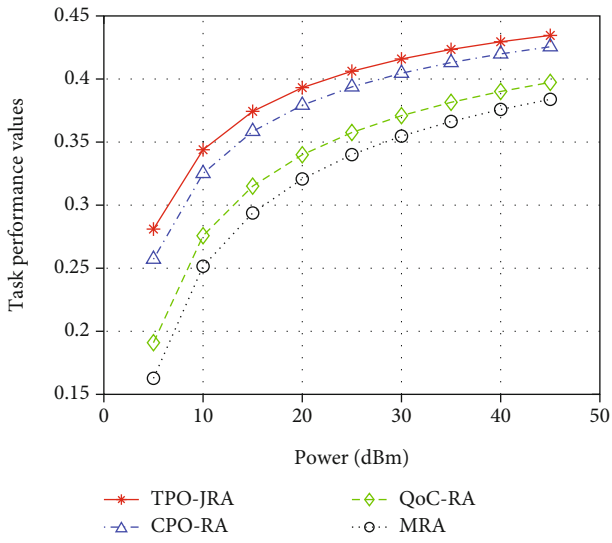


Figure 9: Task performance versus power resource.

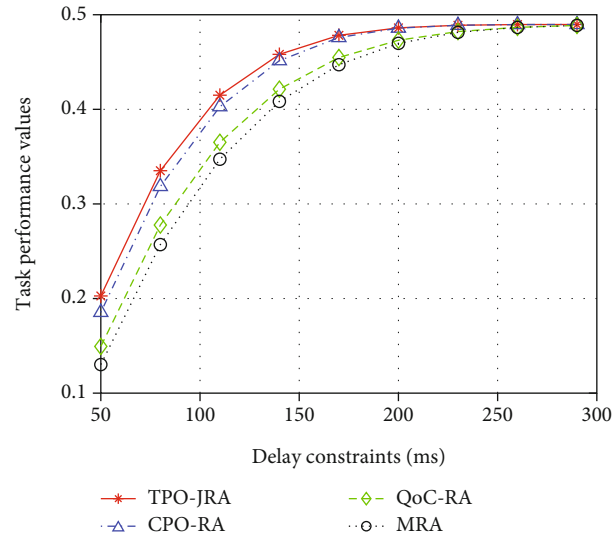


Figure 10: Task performance under different delay constraints.

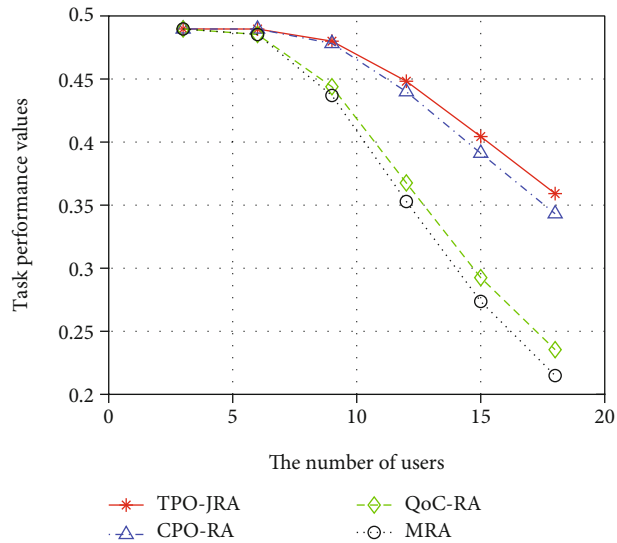


Figure 11: Task performance under different number of vehicles.

priority. Lastly, with the increase of resources, the performance gain decreases gradually, because under the condition of resource saturation, the performance tends to be stable, which is determined by the neural network structure related to intelligent tasks.

5.4.2. Performance versus Delay and Number of Vehicles. With different schemes, the change of task performance under the time-delay constraint is shown in Figure 10. Under the condition of strict delay constraints, the performance of the proposed scheme has advantages, which is due to the reasonable design of the constraints in the optimization problem. The proposed algorithm explores the best action strategy to maximize the task performance. At the same time, the delay requirements are met, according to Equation (11). Secondly, the complexity of the proposed algorithm is low. Based on the above analysis, the proposed resource allocation scheme is more suitable for time-critical scenarios.

The change of task performance with the number of vehicles is shown in Figure 11. It can be seen that with the increase of the number of vehicles connected to the edge server, the overall performance shows a downward trend, because the increase of vehicles causes more intense competition for resources, and the average resources that can be allocated to each vehicle decreases. However, with the increase of the number of vehicles, the proposed resource allocation scheme still maintains the optimal performance, which verifies the performance advantage of the proposed scheme in the scenario of large access or resource shortage. In addition, it also reveals that large-scale access systems must optimize the multidimensional parameters to ensure the accurate understanding of the data at the receiving end.

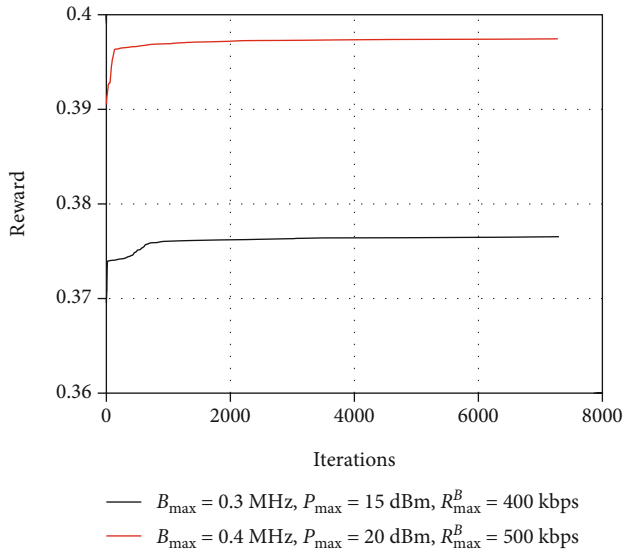


Figure 12: Convergence curve of the proposed algorithm.

5.4.3. Convergence Performance. The convergence results of the proposed resource allocation algorithm are shown in Figure 12. To get rid of burrs, the convergence curves are smoothed for better visual effect. More precisely, the value of each step is the average of the near 5 original steps. The smoothing operation does not affect the convergence performance of the algorithm. It can be seen that the proposed algorithm converges stably within a limited number of iterations. There are two reasons for this. First, the introduction of experience replay mechanism and target network enables the parameters of Q-network to converge. Second, the design of action selection strategy makes the algorithm pay more attention to learning historical experience rather than exploring when the number of iterations increases, which ensures the stability of Q-network. Moreover, according to the curves, our algorithm seems to converge very fast, that is, because the structure of Q-network we designed is relatively simple as shown in Figure 6, which only contains three layers, and the dimensions of each layer are 4, 10, and 3, respectively. This results in fast convergence of the algorithm. Because, in each update iteration, the computation of network parameters becomes smaller, as shown in Equation (13) and (14). The choice of network structure and layer dimensions are the results of experience.

6. Conclusion

This paper studied resource allocation schemes, in order to achieve efficient transmission of multimedia data and accurate completion of intelligent tasks in the IoV scenarios. In this paper, a mathematical model of resource allocation optimization criterion for intelligent tasks was constructed, and the model parameters were solved by a gradient descent algorithm. For classification, detection, and Re-ID tasks, the RMSE index of the algorithm was less than 0.05. Secondly, under the guidance of this model, this paper designed a joint allocation algorithm of data rate, bandwidth, and power resources based on MADQN and discussed the convergence and complexity of the algorithm. Experimental results showed that the proposed resource allocation scheme is more suitable for the intelligent networked environment with lots of computer vision tasks. The results revealed that to improve the performance of intelligent tasks; not only the intelligent algorithms but also the communication technologies like resource allocations should be considered.

In line with the general trend of the close combination of communication technology and artificial intelligence, the scheme proposed in this paper provides a new solution to the difficult problems existing in the era of intelligent driving, such as complex traffic environment and low transmission efficiency. In the future work, we will deduce a more universal resource allocation criterion from the perspective of information theory and analyze the performance of intelligent tasks oriented communication system from the theoretical perspective.

Appendix

A. Proof of the Convergence of Q-Network

In the proposed MADQN algorithm, parameters update and follow the rule:

$$\theta = \theta - \alpha \cdot \nabla\theta, \tag{A.1}$$

where α is the learning factor and $0 < \alpha < 1$ and $\nabla\theta$ denotes the partial derivative of the value function with respect to θ, which is expressed as

$$\begin{aligned}\nabla\theta &= \frac{\partial L(\theta)}{\partial\theta} = 2\left(Q\left(a'\right) - Q_\theta(a)\right)\left(-\frac{\partial Q_\theta(a)}{\partial\theta}\right) \\ &= -2\left(Q\left(a'\right) - Q_\theta(a)\right)a.\end{aligned} \tag{A.2}$$

Hence, by replacing the true value function with the current estimated value function, we can get the parameter update rule, which is expressed as

$$\theta = \theta + \alpha\left(r + \gamma\theta a' - \theta a\right)a. \tag{A.3}$$

After k iterations, the parameters can be expressed as

$$\theta_k = \theta_k + \alpha\left\{(1-\alpha)^k\theta_k a + + \left[1 - (1-\alpha)^k\right]\left(r + \gamma\theta_k a'\right) - \theta_k a\right\}a. \tag{A.4}$$

Due to $0 < (1-\alpha) < 1$, when k is big enough, there is $(1-\alpha)^k \longrightarrow 0$; thus,

$$\theta_k = \theta_k + \alpha\left(r + \gamma\theta_k a' - \theta_k a\right)a = \theta_k - \alpha \cdot \nabla\theta_k. \tag{A.5}$$

That is, after k iterations, the parameters' updating form is still the original form of Behrman equation. Based on the conclusion of [33], the parameters' updating of the proposed algorithm tends to converge.

Data Availability

The experimental data used to support the findings of this study are available from the corresponding author upon request.

Disclosure

A preliminary work of this paper has been presented as Arxiv in Cornell University according to the following link: https://arxiv.org/abs/2009.13379.

Conflicts of Interest

The authors declare that there is no conflict of interest.

Acknowledgments

This work was supported by the Fundamental Research Funds for the Central Universities under Grant No. 2021XD-A01-1, the Beijing Natural Science Foundation under Grant No. 4202049, and the Industrial Internet Research Institute (Jinan) of Beijing University of Posts and Telecommunications under Grant No. 201915001.

References

[1] R. Hussain and S. Zeadally, "Autonomous cars: research results, issues, and future challenges," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 2, pp. 1275–1313, 2019.

[2] J. Zhang, F. Wang, K. Wang, W. H. Lin, X. Xu, and C. Chen, "Data-driven intelligent transportation systems: a survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, no. 4, pp. 1624–1639, 2011.

[3] B. Ranft and C. Stiller, "The role of machine vision for intelligent vehicles," *IEEE Transactions on Intelligent Vehicles*, vol. 1, no. 1, pp. 8–19, 2016.

[4] J. Zhang and K. B. Letaief, "Mobile edge intelligence and computing for the Internet of vehicles," *Proceedings of the IEEE*, vol. 108, no. 2, pp. 246–261, 2020.

[5] J. W. Kim and D. K. Jeon, "A cooperative communication protocol for QoS provisioning in IEEE 802.11p/wave vehicular networks," *Sensors*, vol. 18, no. 11, p. 3622, 2018.

[6] Z. Wu, Z. Lu, P. C. Hung, S. C. Huang, Y. Tong, and Z. Wang, "QaMeC: a QoS-driven IoVs application optimizing deployment scheme in multimedia edge clouds," *Future Generation Computer Systems*, vol. 92, pp. 17–28, 2019.

[7] L. Sun, H. Shan, A. Huang, L. Cai, and H. He, "Channel allocation for adaptive video streaming in vehicular networks," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 1, pp. 734–747, 2016.

[8] H. Zhang, Y. Ma, D. Yuan, and H. H. Chen, "Quality-of-service driven power and sub-carrier allocation policy for vehicular communication networks," *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 1, pp. 197–206, 2011.

[9] L. Liang, G. Y. Li, and W. Xu, "Meeting different QoS requirements of vehicular networks: a D2D-based approach," in *IEEE international Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3734–3738, New Orleans, LA, USA, 2017.

[10] X. Tao, Y. Duan, M. Xu, Z. Meng, and J. Lu, "Learning QoE of mobile video transmission with deep neural network: a data-driven approach," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 6, pp. 1337–1348, 2019.

[11] A. K. Bairagi, S. F. Abedin, N. H. Tran, D. Niyato, and C. S. Hong, "QoE-Enabled unlicensed Spectrum sharing in 5G: a game theoretic approach," *IEEE Access*, vol. 6, pp. 50538–50554, 2018.

[12] M. Jalil Piran, N. H. Tran, D. Y. Suh, J. B. Song, C. S. Hong, and Z. Han, "QoE-driven channel allocation and handoff management for seamless multimedia in cognitive 5G cellular networks," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 7, pp. 6569–6585, 2017.

[13] H. Zhu, Y. Cao, W. Wang, B. Liu, and T. Jiang, "QoE-aware resource allocation for adaptive device-to-device video streaming," *IEEE Network*, vol. 29, no. 6, pp. 6–12, 2015.

[14] H. Gu, Y. Dong, and T. Cao, "Data driven QoE-QoS association modeling of conversational video," *IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, vol. 2019, pp. 1–4, 2019.

[15] X. Chen, J. Hwang, D. Meng, K. H. Lee, R. L. de Queiroz, and F. M. Yeh, "A quality-of-content-based joint source and channel coding for human detections in a mobile surveillance cloud," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 1, pp. 19–31, 2017.

[16] E. C. Strinati and S. Barbarossa, "B6G networks: beyond Shannon towards semantic and goal-oriented communications," *Computer Networks*, vol. 190, article 107930, 2021.

[17] G. Shi, Y. Xiao, Y. Li, and X. Xie, "From semantic communication to semantic-aware networking: model, architecture, and open problems," *IEEE Communications Magazine*, vol. 59, no. 8, pp. 44–50, 2021.

[18] H. Xie, Z. Qin, G. Y. Li, and B.-H. Juang, "Deep learning enabled semantic communication systems," *IEEE Transactions on Signal Processing*, vol. 69, pp. 2663–2675, 2021.

[19] H. Xie and Z. Qin, "A lite distributed semantic communication system for internet of things," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 1, pp. 142–153, 2021.

[20] J. Chen, C. Guo, C. Feng, M. Zhu, and Q. Sun, "Content driven and reinforcement learning based resource allocation scheme in vehicular network," in *IEEE International Conference on Communications (ICC)*, pp. 1–6, Montreal, QC, Canada, 2021.

[21] M. Zhu, C. Feng, J. Chen, C. Guo, and X. Gao, "Video semantics based resource allocation algorithm for spectrum multiplexing scenarios in vehicular networks," in *IEEE/CIC International Conference on Communications in China (ICCC Workshops)*, pp. 31–36, Xiamen, China, 2021.

[22] G. J. Sullivan, J. R. Ohm, W. J. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1649–1668, 2012.

[23] B. Sklar, *Digital Communication Fundamentals and Applications*, Prentice Hall, Upper Saddle River, New Jersey, USA., 2001.

[24] C. Guo, L. Liang, and G. Y. Li, "Resource allocation for vehicular communications with low latency and high reliability," *IEEE Transactions on Wireless Communications*, vol. 18, no. 8, pp. 3887–3902, 2019.

[25] J. Shi, Z. Yang, H. Xu, M. Chen, and B. Champagne, "Dynamic resource allocation for LTE-based vehicle-to-infrastructure

networks," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 5, pp. 5017–5030, 2019.

[26] N. C. Luong, D. T. Hoang, S. Gong et al., "Applications of deep reinforcement learning in communications and networking: a survey," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 4, pp. 3133–3174, 2019.

[27] T. T. Nguyen, N. D. Nguyen, and S. Nahavandi, "Deep reinforcement learning for multiagent systems: a review of challenges, solutions, and applications," *IEEE Transactions on Cybernetics*, vol. 50, no. 9, pp. 3826–3839, 2020.

[28] A. Coates, A. Ng, and H. Lee, "An analysis of single-layer networks in unsupervised feature learning," in *2011 International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 215–223, Ft. Lauderdale, FL, USA, 2011.

[29] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: an evaluation of the state of the art," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 4, pp. 743–761, 2012.

[30] P. Sun, H. Kretzschmar, X. Dotiwalla et al., "Scalability in perception for autonomous driving: Waymo Open Dataset," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2443–2451, Seattle, WA, USA, 2020.

[31] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: a benchmark," in *IEEE International Conference on Computer Vision (ICCV)*, pp. 1116–1124, Santiago, Chile, 2015.

[32] J. Chen, C. Guo, S. Wei, X. Sun, and C. Feng, "SCO: a dataset for semantic communications," https://github.com/WeiSY0516/SCO-dataset.git.

[33] F. S. Melo, "Convergence of Q-learning: a simple proof," *Institute of Systems and Robotics Technical Report*, pp. 1–4, 2001.

Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 9121059, 10 pages
https://doi.org/10.1155/2022/9121059

Research Article

Deep Learning-Based Nonstationary Channel Prediction in Tactical Vehicle-to-Vehicle Communication Environments

Xin Lin[1], **Aijun Liu**[1], **Chen Han**[2], **Xiaohu Liang**[3], **Wenyu Wang**[1], **and Enyu Li**[4]

[1]*College of Communications Engineering in Army Engineering University of PLA, Nanjing 210000, China*
[2]*Sixty-Third Research Institute, National University of Defense Technology, Nanjing 21007, China*
[3]*School of Information Science and Engineering in Southeast University and the College of Communications Engineering in Army Engineering University of PLA, Nanjing 210000, China*
[4]*Department of Electronic Engineering of Qingdao University, Qingdao 266520, China*

Correspondence should be addressed to Aijun Liu; liuaj.cn@163.com

Received 11 May 2022; Revised 26 May 2022; Accepted 1 August 2022; Published 26 August 2022

Academic Editor: A.H. Alamoodi

In this paper, we focus on the vehicle-to-vehicle dynamic channel in tactical communication environments, which shows time-varying and nonstationary characteristics due to the fast mobility, directional antennas, and harsh terrain. These situations present great challenges for the channel state information (CSI) acquisition. To obtain an accurate CSI and reduce pilot overhead, we propose a CSI predictor based on the long short-term memory (LSTM) network. As an improved recurrent neural network (RNN), LSTM units have an excellent learning result on both long- and short-term inputs by adding the gating mechanism. Using the outdated sampling CSI sequence as input data of LSTM units enables the predictor to extract complex data characteristics and capture the temporal law of the nonstationary channel. Simulation results are demonstrated to verify that the LSTM-based predictor has better performance than conventional algorithms in IEEE 802.11p standard. Additionally, the key factors that affect the performance of the proposed predictor are further analyzed.

1. Introduction

Tactical vehicle platforms include the armored vehicles and the unmanned ground vehicles (UGVs), which have been widely employed in the land battlefield nowadays [1]. These platforms are interconnected through the mobile ad hoc network (MANET) to perform the military missions [2]. In actual operations and maneuvers, the time-varying and nonstationary channel characteristics are caused by the fast mobility, directional antennas, and harsh terrain, which leads to poor transmission conditions and link interruption. Therefore, the reliability and validity of the vehicle-to-vehicle (V2V) wireless link are faced with great challenges. To tackle this problem, adaptive link transmission scheme and channel equalization techniques depending on the channel state information (CSI) are proposed to improve the system performance in this scenario [3–6]. Various data pilot-aided (DPA) channel estimation methods have been proposed to extract accurate CSI, which utilizes the demapped data symbols as pilot information. The constructed data pilot (CDP) scheme evaluates the reliability of estimated subcarrier value by comparing the adjacent symbols and deciding whether or not to replace the estimated value [7]. The spectral temporal averaging (STA) algorithm [8] averages the estimated values in the time-frequency domain to reduce detection error. In test frequency domain interpolation (TRFI) scheme, the unreliable estimated value is renewed by frequency domain interpolation of the reliable data pilots [9]. However, these traditional DPA channel estimation approaches have difficulties in combating the dynamic time-frequency selective channel characteristics in such high mobile and complex environment [10]. On the one hand, the shorter coherence time and bandwidth will lead to frequent pilot insertion and reduce the effective communication rates [11]. On the other hand, the demapping and detection errors incur error propagation and limit the estimation performance. Additionally, estimated CSI will quickly be outdated due to the feedback

delay and nonstationary channel characteristics in the dynamic time-varying conditions [12].

Recently, deep learning (DL) techniques show great promise due to the powerful feature extraction capability and data-driven characteristics. Extensive research has applied DL techniques to DPA channel estimation process. The authors of [13] introduce an autoencoder- (AE-) based on neural network to conventional DPA channel estimation scheme where the AE is trained for updating the estimated value and attenuate the error propagation effect. In [14], the authors proposed a DL-based MIMO-OFDM channel estimation scheme. The convolutional neural network (CNN) and bidirectional long short-term memory (LSTM) are applied for frequency-domain interpolation and time-domain prediction, respectively. A gated recurrent unit- (GRU-) based channel estimation is developed in [15]. The error propagation in DPA process is suppressed by extracting the complex time-frequency correlation features. Simulation results show that the proposed method exhibits a performance gain without increasing computation complexity. In [16], a novel LSTM-based channel estimation scheme was designed for IEEE 802.11p standard in nonstationary V2V scenario. The proposed network structure is utilized for tracking channel characteristic and mitigating noise.

Compared with channel estimation techniques, channel prediction provides us with a more effective solution to the above problems. Based on the outdated historical CSI, channel prediction technique can extract the future CSI in a given period without sacrificing the scarce pilot resources and achievable data rates. Additionally, channel prediction technique can explore the changing law of the channel state and speculate on the future CSI in advance to weaken the effect of outdated information. Therefore, the channel prediction can overcome the disadvantage of excessive pilot overhead and outdated effect in channel estimation, which is suitable for the time-varying and nonstationary V2V channel. In this paper, we propose a predictor based on LSTM neural network for the dynamic V2V channel [17]. By introducing a new internal cell state and the gating mechanism, the LSTM neural network can not only record the history information but also control the path of data transmission. Therefore, the LSTM units are sensitive to both long- and short-term inputs. The LSTM-based predictor can capture the complex law of the dynamic channel and exploit the temporal correlation to get the accuracy and real-time CSI within a period. This method can not only avoid the extra pilot overhead but also improve the prediction accuracy by virtue of the powerful feature extraction capabilities of the LSTM neural network. The main contributions of our work are summarized as follows:

(I) the system and V2V channel model in the tactical communication environments is developed, and the nonstationary channel characteristics are analyzed in detail.

(II) To mitigate the effect of the outdated information, a two-stage LSTM-based prediction scheme is proposed to explore the temporal correlation of the CSI and realize the future CSI prediction.

(III) The proposed predictor is shown to outperform the conventional algorithms by evaluating the normalized mean-squared error (NMSE) index. Besides, the impact of key factors on the propose scheme is analyzed in the end.

The remainder of this paper is organized as follows. Section 2 introduces the V2V system and nonstationary channel model for the tactical communication environments, in which the effect of fast mobility, directional antennas, and harsh terrain are taken into consideration. Section 3 proposes a two-stage LSTM-based channel prediction scheme and the architecture of the predictor. Simulation results that demonstrate the system performance are provided in Section 4. Finally, Section 5 gives the conclusions.

2. System and Nonstationary Channel Model

Figure 1 illustrates a point-to-point V2V communication system, where directional antennas are equipped in both transmit and receive platforms to improve the signal to interference plus noise ratio (SINR) and increase the communication distance in the tactical environments. Due to the complex scattering environment, the received signals can be divided into two components: the line-of-sight (LoS) component and the non-LoS (NLoS) component. Therefore, the channel impulse response can be expressed as

$$h(t,\tau) = \sqrt{\frac{K}{K+1}} h^{\mathrm{LoS}}(t)\delta\left(\tau - \tau^{\mathrm{LoS}}\right), + \sqrt{\frac{1}{K+1}} \sum_{n=1}^{N-1} h_n^{\mathrm{NLoS}}(t)\delta\left(\tau - \tau_n^{\mathrm{NLoS}}\right) \tag{1}$$

where K and N represent the Rice factor and the total number of the multipath component, respectively. $h^{\mathrm{LoS}}(t)$ and τ^{LoS} represent the channel complex coefficient and the delay of the LoS component. $h_n^{\mathrm{NLoS}}(t)$ and τ_n^{NLoS} are the channel complex coefficient and the delay of the n-th NLoS path. $h^{\mathrm{LoS}}(t)$ and $h^{\mathrm{NLoS}}(t)$ can be expressed as

$$h^{\mathrm{LoS}}(t) = e^{j\Phi^{\mathrm{LoS}}} e^{j2\pi f_{D,Tx}^{\mathrm{LoS}} \cos\left(\phi_{Tx}^{\mathrm{LoS}} - \gamma_{Tx}\right)t} \times e^{j2\pi f_{D,Rx}^{\mathrm{LoS}} \cos\left(\phi_{Rx}^{\mathrm{LoS}} - \gamma_{Rx}\right)t}, \tag{2}$$

$$h_n^{\mathrm{NLoS}}(t) = \sqrt{\frac{1}{N-1}} \lim_{M\longrightarrow\infty} \frac{1}{\sqrt{M}} \sum_{m=1}^{M} e^{j\Phi_m^{\mathrm{NLoS},n}} \times e^{j2\pi f_{D,Tx}^{\mathrm{NLoS},n} \cos\left(\phi_{Tx,m}^{\mathrm{NLoS},n} - \gamma_{Tx}\right)t} \times e^{j2\pi f_{D,Rx}^{\mathrm{NLoS},n} \cos\left(\phi_{Rx,m}^{\mathrm{NLoS},n} - \gamma_{Rx}\right)t}, \tag{3}$$

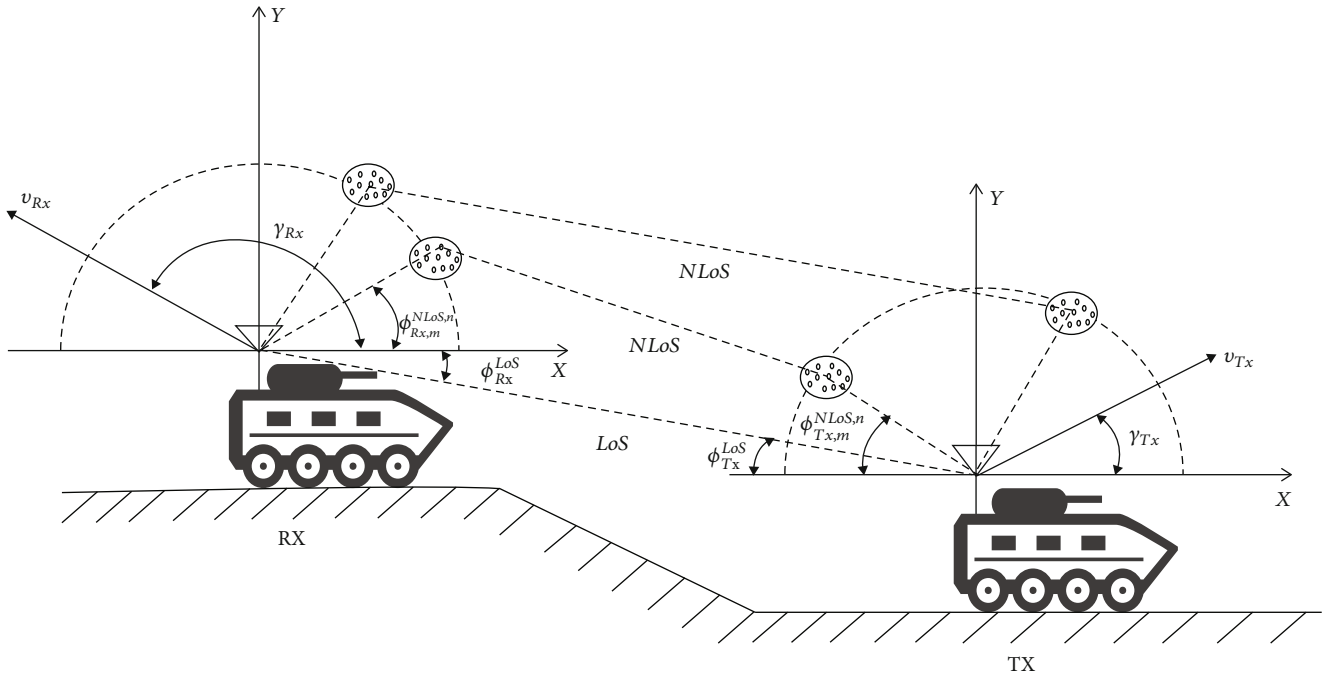


Figure 1: Point-to-point vehicle-to-vehicle communication system in the tactical environments.

where the channel variable $\{\Phi^{\mathrm{LoS}}, f_{D,Tx}^{\mathrm{LoS}}, f_{D,Rx}^{\mathrm{LoS}}, \phi_{Tx}^{\mathrm{LoS}}, \phi_{Rx}^{\mathrm{LoS}}\}$ represents the initial phase, the maximal Doppler shift caused by the motion of TX, the maximal Doppler shift caused by the motion of RX, the direction of departure, and the direction of arrival of the LoS path, respectively. $\{\Phi_m^{\mathrm{NLoS},n}, f_{D,Tx}^{\mathrm{NLoS},n}, f_{D,Rx}^{\mathrm{NLoS},n}, \phi_{Tx,m}^{\mathrm{NLoS},n}, \phi_{Rx,m}^{\mathrm{NLoS},n}\}$ are the channel parameters of the m-th scattering path in the n-th NLoS path component. γ_{Tx} (γ_{Rx}) is the movement angle of the TX (RX) and M denotes the number of the scattered path. By taking the Fourier transform of Equation (1), the time-varying transmission function can be expressed as

$$H(t,f) = \int_{-\infty}^{+\infty} h(t,\tau)e^{-j2\pi f\tau}d\tau = \sqrt{\frac{K}{K+1}}h^{\mathrm{LoS}}(t)e^{-j2\pi f\tau^{\mathrm{LoS}}} \cdot + \sqrt{\frac{1}{K+1}}\sum_{n=1}^{N-1} h_n^{\mathrm{NLoS}}(t)e^{-j2\pi f\tau_n^{\mathrm{NLoS}}}. \tag{4}$$

Due to the fast mobility and low height of transmitting and receiving antennas deployed in mobile vehicle, the number and strength of propagation paths change frequently, so the channel statistic characteristics are time-varying, resulting in the nonstationary channel feature [11]. The "birth-death" process introduced in [18] is employed to model the "lifetime" of the multipath component, and the transition between "birth" and "death" can be described using the state transition matrix. In such tactical scenario, an additional term $z(t)$ can be used to account for the frequent changing state of the multipath component [19].

The directional antenna misalignment problem of the harsh terrains should be considered in the tactical environments. As shown in Figure 2, when mobile vehicle move on the flat terrain, the maximum gain direction of the transceiver antenna beam can be aligned. However, the main beam of the transceiver antenna cannot be aligned in the undulating terrain, resulting in a time-varying power loss of the propagation path. According to [20], the antenna power pattern $G(\theta,\psi)$ is applied to represent the directional misalignment effect, where $\{\theta,\psi\}$ denotes the direction of the antenna main beam. Based on the above analysis, the time-varying and nonstationary V2V channel in tactical communication environments can be reformulated as Equation (5), where the channel parameters $\{G(\theta_{Tx}^{\mathrm{LoS}}, \psi_{Tx}^{\mathrm{LoS}}), G(\theta_{Rx}^{\mathrm{LoS}}, \psi_{Rx}^{\mathrm{LoS}}), z^{\mathrm{LoS}}(t)\}$ represent the directional antenna power pattern of TX and RX as well as the state of the LoS path, respectively. The channel parameters of the n-th path in the NLoS component are $\{G(\theta_{Tx,n}^{\mathrm{NLoS}}, \psi_{Tx,n}^{\mathrm{NLoS}}), G(\theta_{Rx,n}^{\mathrm{NLoS}}, \psi_{Rx,n}^{\mathrm{NLoS}}), z_n^{\mathrm{NLoS}}(t)\}$. Note that the propagation environment changes during the moving process; thus, the parameters in Equation (5) are all time-varying.

$$H(t,f) = \sqrt{\frac{K}{K+1}}\sqrt{G\left(\theta_{Tx}^{\mathrm{LoS}}, \psi_{Tx}^{\mathrm{LoS}}\right)G\left(\theta_{Rx}^{\mathrm{LoS}}, \psi_{Rx}^{LoS}\right)}z^{\mathrm{LoS}}(t)h^{\mathrm{LoS}}(t)e^{-j2\pi f\tau^{\mathrm{LoS}}} + \sqrt{\frac{1}{K+1}}\sum_{n=1}^{N-1}\sqrt{G\left(\theta_{Tx,n}^{\mathrm{NLoS}}, \psi_{Tx,n}^{\mathrm{NLoS}}\right)G\left(\theta_{Rx,n}^{\mathrm{NLoS}}, \psi_{Rx,n}^{\mathrm{NLoS}}\right)}z_n^{\mathrm{NLoS}}(t)h_n^{\mathrm{NLoS}}(t)e^{-j2\pi f\tau_n^{\mathrm{NLoS}}}. \tag{5}$$

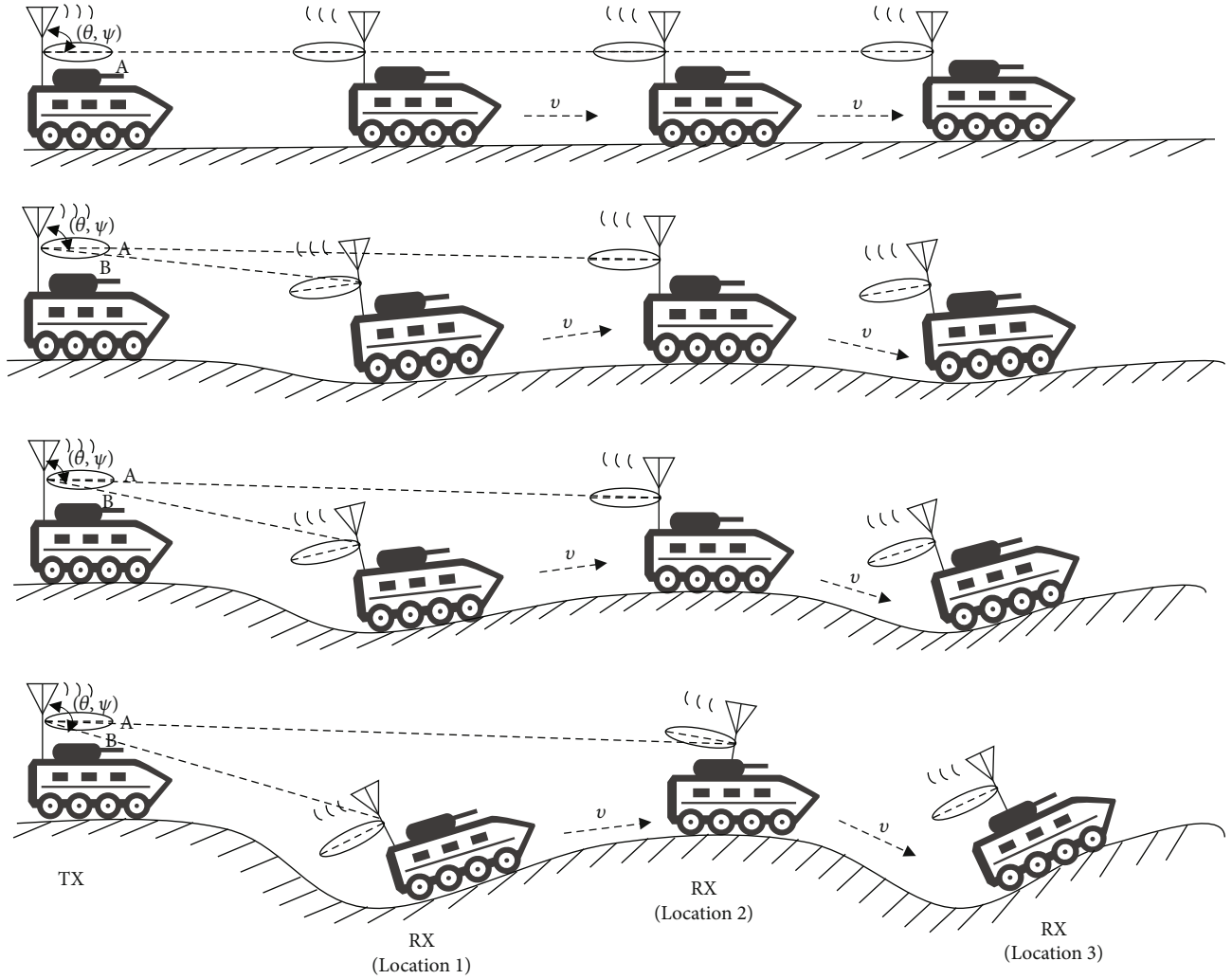


Figure 2: The misaligned angle of the transceiver antenna in different terrains.

3. Long Short-Term Memory-Based Channel Prediction

According to the above analysis, the CSI obtained using the channel estimation technique is outdated due to the significant time-varying and nonstationary channel characteristics. Here, we employ an LSTM-based predictor to explore complex channel characteristics and extract real-time CSI. The proposed model is trained and optimized by minimizing the root mean square error (RMSE) $J(\Theta)$ between the predicted CSI $\hat{H}(t,f)$ and supervision value $H(t,f)$. $J(\Theta)$ can be expressed as Â

$$J(\Theta) = \sqrt{\frac{1}{Q}\sum_{q=1}^{Q}\left[\hat{H}_q(t,f) - H_q(t,f)\right]^2}, \tag{6}$$

where Q denotes the number of channel sequence samples in the training process.

3.1. Two-Stage Prediction Scheme. The whole prediction scheme can be divided into two stages: the training and prediction stage, as shown in Figure 3.

The proposed LSTM-based predictor will learn the correlation between the historical CSI in the training stage. Then, the trained model takes the outdated CSI as the input data to extract the future moment CSI in the prediction stage. Note that the training stage is complex and time-consuming, but it can be finished in offline mode. Moreover, with the development of hardware technology, high-performance processing modules are deployed in tactical devices. Hence, the computing power and processing capacity of tactical devices has been greatly enhanced, which can support the computational requirement of the proposed approach.

3.2. Architecture of the Channel Predictor. Figure 4 illustrates the architecture of the proposed LSTM-based channel predictor. By adding the cell state and the gating mechanism, the LSTM cell addresses the problem of gradient explosion or disappearance of the RNN. The LSTM cell is mainly

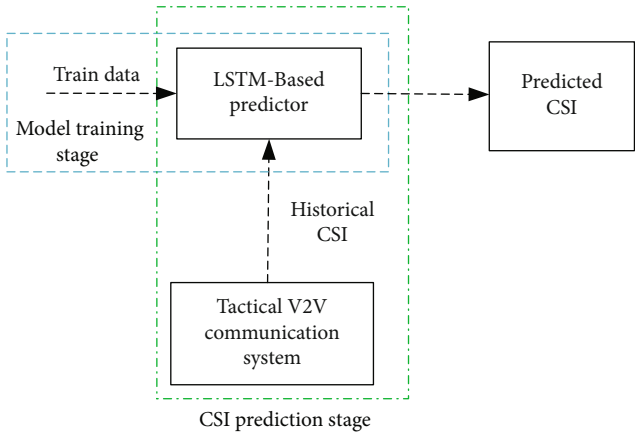


FIGURE 3: Two-stage prediction scheme for vehicle-to-vehicle channel in tactical communication environment.

composed of three gates, namely, the forget gate f_t, the input gate i_t, and the output gate o_t. The three gates and candidate states $\tilde{c}_t$ are calculated by the external state h_{t-1} at the previous moment and the input $x_t \in \mathbb{R}^{M\times 1}$, which can be expressed as

$$f_t = \sigma\left(W_f \cdot [x_t, h_{t-1}] + b_f\right), \tag{7}$$

$$i_t = \sigma(W_i \cdot [x_t, h_{t-1}] + b_i), \tag{8}$$

$$o_t = \sigma(W_o \cdot [x_t, h_{t-1}] + b_o), \tag{9}$$

$$\tilde{c}(t) = \tanh\left(W_c \cdot [x_t, h_{t-1}] + b_c\right), \tag{10}$$

where $\{W_f, W_i, W_o, W_c\}$ and $\{b_f, b_i, b_o, b_c\}$ are weight matrices and bias vectors for LSTM units, respectively. The gates f_t and i_t are then combined to update the memory unit c_t, as shown in Equation (11) and Equation (12). Finally, by calculating c_t and o_t, the internal state is transferred to the next external state h_t.

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t, \tag{11}$$

$$h_t = o_t \odot \tanh(c_t). \tag{12}$$

The above process can not only achieve the linear transmission of the cyclic information but also nonlinearly output the information to the external state. Therefore, the LSTM neural network can realize excellent learning results for time series in both long and short terms. During the training process, the LSTM structure will automatically learn and update the weight matrix W_* and bias term b_*. The LSTM Layer1 is an input layer to receive historical CSI, and the two following LSTM layers play a key role in exploring the temporal features of the nonstationary channel. Finally, the output layer is a fully connected layer, which will reshape and output the predicted CSI $\hat{H}(t, f)$. The rectified linear unit (Relu) function is selected as the activate function for all layers. It should be noted that the neural network modules are only capable of dealing with the real-valued data, but channel coefficients are complex-valued. To improve the efficiency, most articles decompose complex channel coefficients into real and imaginary parts in training stage [13, 15, 16]. In test stage, the separating parts predicted by the neural network will be recombined into the complex value as the channel coefficients. The LSTM-based V2V channel prediction algorithm for tactical communication environments can be summarized as Algorithm 1.

3.3. Computation Complexity Analysis. In this subsection, the multiply accumulate (MAC) operation is used to analyze the computation complexity of the proposed LSTM-based predictor, which comes from the matrix operation of the neural network. The CSI sample $x_t \in \mathbb{R}^{M\times 1}$ is the input data of the LSTM Layer1, whose weight matrices and bias vectors are $W_i, W_o, W_f, W_c \in \mathbb{R}^{(M+L_1)\times L_1}$ and $b_i, b_0, b_f, b_c \in \mathbb{R}^{L_1\times 1}$, respectively. Therefore, the MAC operation for the first LSTM layer is $4K\left[(M + L_1)\, L_1 + L_1\right]$ where K is the length of the input sequence. For the other LSTM layers $(2 \le j \le J)$, the input data is $h_{t-1} \in \mathbb{R}^{(L_{j-1}\times 1)}$, so the MAC operation for them is $4K \cdot [\sum_{j=2}^{J}(2L_{j-1} \cdot L_j + L_j)]$. Finally, MAC operation for the output layer can be expressed as $4K \cdot (L_J \cdot M + M)$.

4. Numerical Simulation Result

In this section, simulation results are performed to validate the performance of the proposed prediction scheme. The simulations are performed on the TensorFlow-GPU 2.0.0 platform and relevant parameters are set as follows. The total number of paths is set to 12 and the Rician factor K is 20 dB. The amplitude of the multipath component follows Rayleigh distribution. For the directional antennas, a half-wave oscillator antenna is selected, whose maximum beam angle is 78°. The neuron numbers in the three LSTM layers are 128, 256, and 512, respectively. To avoid overfitting, we set the dropout value to 0.3. To improve the training efficiency, the Adam optimizer is used. The epoch limit for training and the batch size are 500 and 128, respectively. The sizes of training and test sets are 20000 and 2000, respectively. To verify the generalization property of the proposed algorithm, the channel model parameters of test set are not the same as the training one. There are new channel parameters for the test set. For example, the test set included terrain and Doppler shift parameters that were not presented in training set. In addition, the channel prediction is performed based on the outdated CSI, which will be regarded as the input data for neural network model. In this paper, we adopt the traditional comb-type pilot pattern to estimate the outdated CSI. The pilot symbols are equally distributed on the subcarriers with a spacing of 4. The length of input sequence is set to 30. Absolutely, as the input length increases, more channel features can be captured by LSTM networks, and the changing law of channel can be extracted more accurately. Certainly, as the dimension of input data increases, the complexity of neural networks also increases greatly.

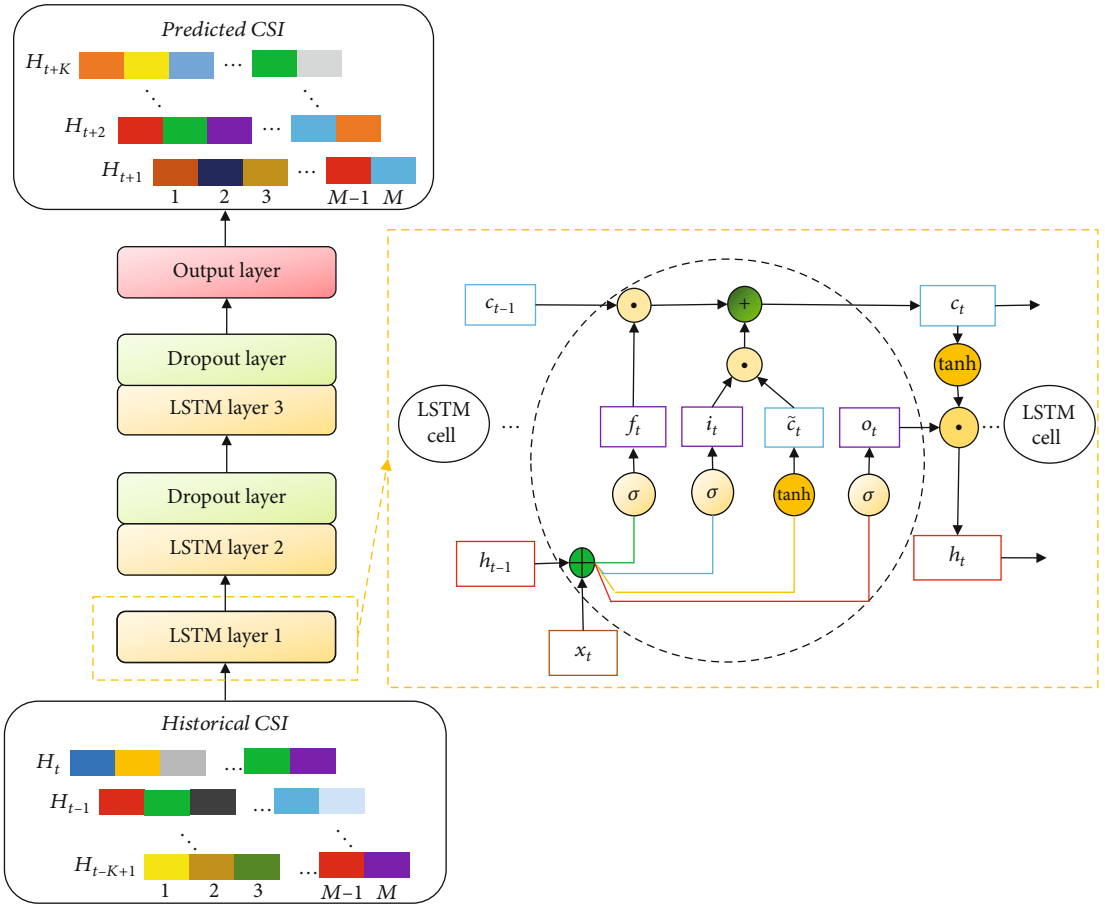


Figure 4: The architecture of the long short-term memory-based prediction framework.

input: Complex CSI sequence, the set of hyperparameters Θ.
1: **Training stage:**
2: Generate the input data of the training set by dividing the complex channel sequence into two subsequences according to the real and imaginary component.
3: **for** epoch e =1 : E**do**
4: Explore the temporal non-stationary characteristics of the V2V channel using the LSTM-based neural networks model.
5: Obtain the predicted CSI via the model.
6: The hyperparameters are optimized by minimizing the loss function $J(\Theta)$ in Equation (6).
7: **end**
8: **Prediction stage:**
9: Generate the test input data by converting the complex data to real domain.
10: Predict the target CSI via inputting the historical data into the trained LSTM-based predictor.
output: The trained LSTM-based channel predictor.

Algorithm 1: The long short-term memory-based vehicle-to-vehicle channel prediction algorithm for tactical communication environments.

To evaluate the prediction accuracy, we take the normalized mean square error (NMSE) as the performance index, which can be calculated using

$$\mathrm{NMSE} = E\left\{\frac{1}{P}\sum_{p=1}^{P}\frac{\left\|\widehat{H}_p(t,f) - H_p(t,f)\right\|_2^2}{\left\|H_p(t,f)\right\|_2^2}\right\}, \quad (13)$$

where P is the total number of the test samples and $E\{\cdot\}$ denotes the expectation operation.

We first analyze the temporal correlation of the V2V channel with different Doppler shift, as shown in Figure 5(a). As the Doppler shift decreases, the correlation coefficient gradually increases due to the longer coherence time, indicating that the channel information is predictable within a certain time interval. Figure 5(b) shows the

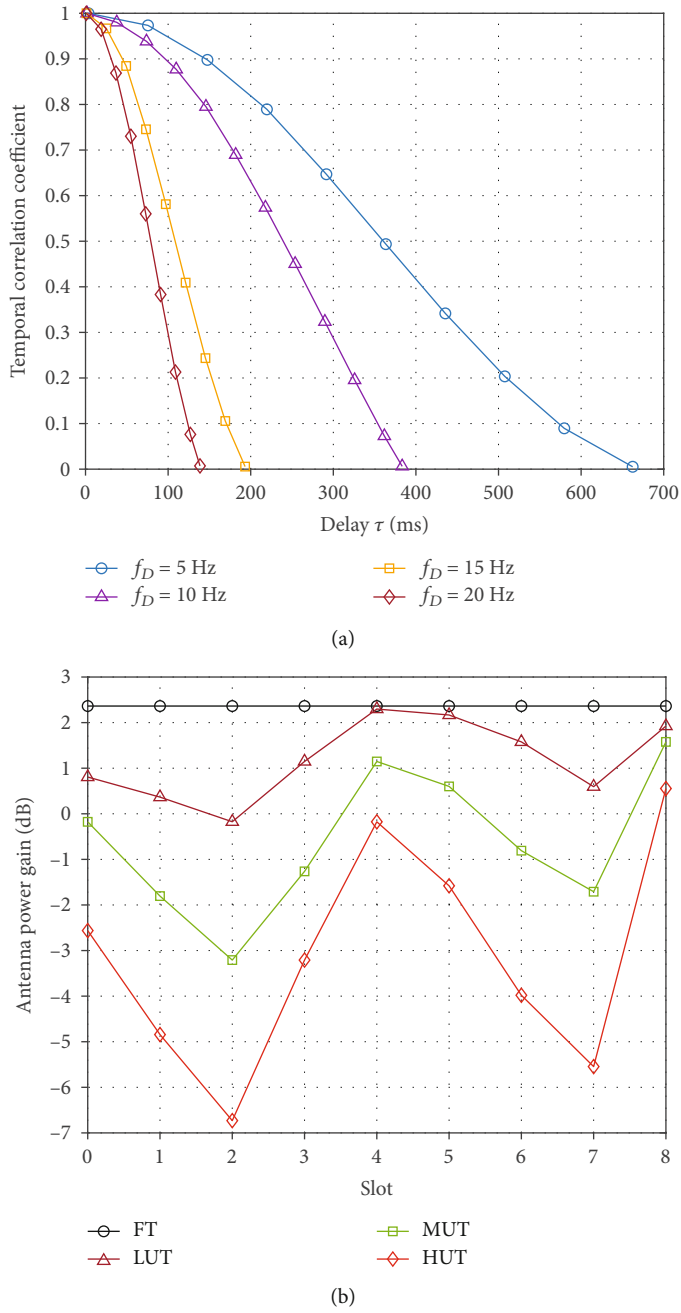


FIGURE 5: (a) Temporal correlation characteristic with different maximal Doppler shifts. (b) Comparison of the antenna power gain in different terrain conditions.

comparison results of antenna pattern gain in four terrain conditions, including flat terrain (FT), lightly undulating terrain (LUT), moderately undulating terrain (MUT), and heavily undulating terrain (HUT). It can be seen that a more undulating terrain will cause severe fluctuations of directional antenna gain and make the changing law of the V2V channel more complex, which will bring more challenges to channel prediction.

As shown in Figure 6, the conventional channel estimation algorithms in IEEE 802.11p standard, such as constructed data pilots (CDP) [7] and spectral temporal averaging (STA) [8], show bad performance due to the estimation error propagation effect caused by noise and channel variation. Compared with conventional algorithms, the proposed LSTM-based predictor realizes performance improvement due to the strong ability of the DNNs in extracting complex features and correlation relationships. For a fair comparison, the multilayer perception (MLP) in [21] is deployed with the same number of layers and neurons as the proposed LSTM-based predictor. Due to the great

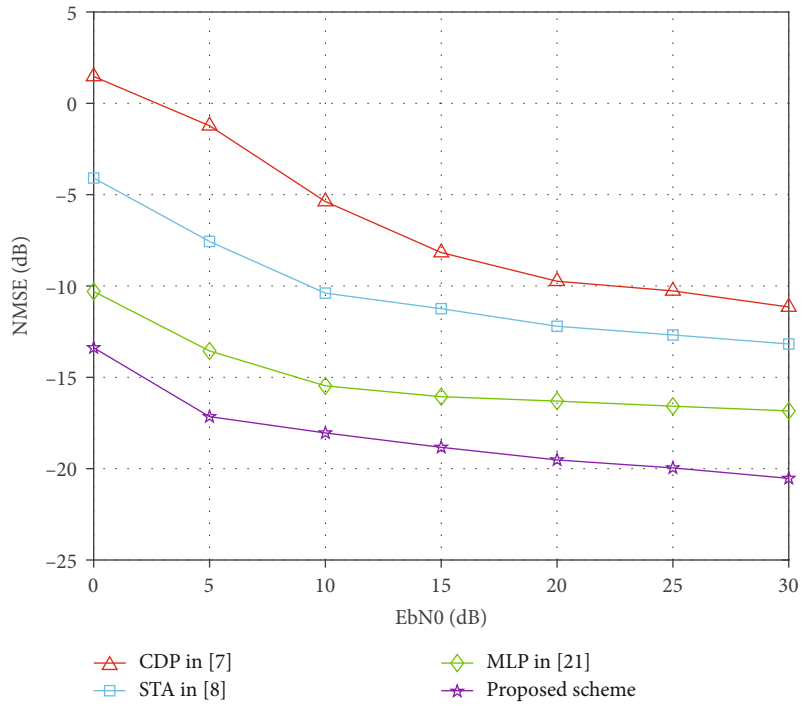


Figure 6: The NMSE of the proposed scheme and conventional algorithms for different SNRs.

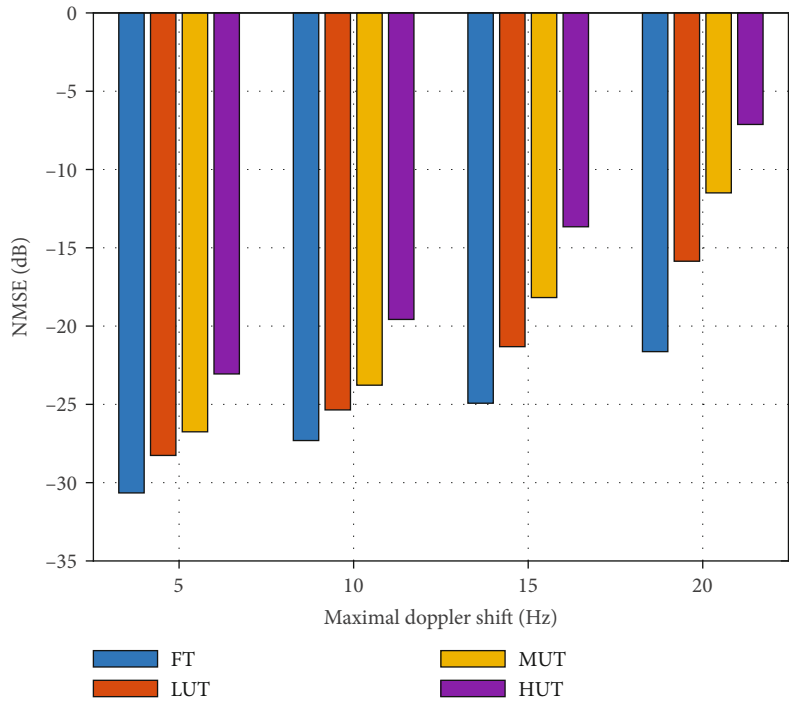


Figure 7: The NMSE of the proposed scheme for different terrain with different vehicle speeds.

advantages of the LSTM neural networks in processing time-series data, there is also a performance gain between the proposed scheme and MLP.

Figure 7 analyzes the key factors that affect the performance of the proposed scheme. The training and test datasets of the four terrains are generated according to the direction of antennas main lobe $\{\theta, \psi\}$ in corresponding terrains. The mobility scenarios are indicated by the maximal Doppler shift of 5, 10, 15, and 20 Hz, which are consistent with the vehicle speed. Therefore, the training and test

datasets of different Doppler shift are generated according to different vehicle speed in actual scenario. With the increase in the Doppler shifts, the temporal correlation of the V2V channel weakens and the prediction accuracy gradually decreases. Additionally, it is intuitive that the proposed scheme is more suitable to be applied for the less undulating terrain under the same vehicle speed. This is because the temporal changing law of antenna gain is more complex in undulating terrain, which will increase the difficulty for the LSTM-based model to extract the channel characteristics and reduce the accuracy of the performance.

5. Conclusion

In this paper, we first introduced the V2V channel model and nonstationary characteristic in tactical communication environments. Then, we proposed a LSTM-based channel predictor to reduce pilot overhead and mitigate the impacts of outdated information. The simulation results showed that the proposed method can get better prediction accuracy than other conventional algorithms in the index of NMSE. Based on above analysis, it can be concluded that the terrain conditions and the vehicle speed influence on the performance of the proposed scheme. In future work, we will design the adaptive transmission scheme according to the predicted CSI to overcome the adverse influence of the V2V channel and improve the reliability and validity of the tactical communication system.

Data Availability

The original data used to support this work is generated by MATLAB 2019, and the method of dataset generation is included within the manuscript.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFB1801103, in part by the Natural Science Foundation on Frontier Leading Technology Basic Research Project of Jiangsu Province under Grant BK20192002, in part by the National Natural Science Foundation of China under Grant 61901516, in part by the China Postdoctoral Science Foundation under Grant 2019M651648, and in part by the Natural Science Foundation of Jiangsu Province under Grant BK20180578.

References

[1] J. S. Lee, Y.-S. Yoo, H. S. Choi, T. Kim, and J. K. Choi, "Energy-efficient TDMA scheduling for UVS tactical MANET," *IEEE Communications Letters*, vol. 23, no. 11, pp. 2126–2129, 2019.

[2] B. Roh, M. H. Han, M. Hoh, K. Kim, and B. H. Roh, "Tactical MANET architecture for unmanned autonomous maneuver network," in *MILCOM 2016-2016 IEEE Military Communications Conference*, pp. 829–834, Baltimore, MD, USA, 2016.

[3] A. Duel-Hallen, "Fading channel prediction for mobile radio adaptive transmission systems," *Proceedings of the IEEE*, vol. 95, no. 12, pp. 2299–2313, 2007.

[4] W. Jiang and H. D. Schotten, "Deep learning for fading channel prediction," *IEEE Open Journal of the Communications Society*, vol. 1, pp. 320–332, 2020.

[5] M. Wen, B. Ye, E. Basar, Q. Li, and F. Ji, "Enhanced orthogonal frequency division multiplexing with index modulation," *IEEE Transactions on Wireless Communications*, vol. 16, no. 7, pp. 4786–4801, 2017.

[6] J. Li, S. Dang, Y. Yan, Y. Peng, S. Al-Rubaye, and A. Tsourdos, "Generalized quadrature spatial modulation and its application to vehicular networks with NOMA," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 7, pp. 4030–4039, 2021.

[7] J. A. Fernandez, K. Borries, L. Cheng, B. V. K. Vijaya Kumar, D. D. Stancil, and F. Bai, "Performance of the 802.11p physical layer in vehicle-to-vehicle environments," *IEEE Transactions on Vehicular Technology*, vol. 61, no. 1, pp. 3–14, 2012.

[8] Z. Zhao, X. Cheng, M. Wen, B. Jiao, and C.-X. Wang, "Channel estimation schemes for IEEE 802.11p standard," *IEEE Intelligent Transportation Systems Magazine*, vol. 5, no. 4, pp. 38–49, 2013.

[9] Y.-K. Kim, J.-M. Oh, Y.-H. Shin, and C. Mun, "Time and frequency domain channel estimation scheme for IEEE 802.11p," in *17th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, pp. 1085–1090, Qingdao, China, 2014.

[10] X. Lyu, W. Feng, N. Ge, and X. Wang, "Deep learning-based symbol detection for time-varying nonstationary channels," *China Communications*, vol. 19, no. 3, pp. 158–171, 2022.

[11] R. M. Rao, V. Marojevic, and J. H. Reed, "Adaptive pilot patterns for CA-OFDM systems in nonstationary wireless channels," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 2, pp. 1231–1244, 2018.

[12] W. Jiang and H. D. Schotten, "Neural network-based fading channel prediction: a comprehensive overview," *IEEE Access*, vol. 7, pp. 118112–118124, 2019.

[13] S. Han, Y. Oh, and C. Song, "A deep learning based channel estimation scheme for IEEE 802.11p systems," in *ICC 2019-2019 IEEE International Conference on Communications (ICC)*, pp. 1–6, Shanghai, China, 2019.

[14] Y. Liao, Y. Hua, and Y. Cai, "Deep learning based channel estimation algorithm for fast time-varying MIMO-OFDM systems," *IEEE Communications Letters*, vol. 24, no. 3, pp. 572–576, 2020.

[15] J. Hou, H. Liu, Y. Zhang, W. Wang, and J. Wang, "GRU-based deep learning channel estimation scheme for the IEEE 802.11p standard," *IEEE Wireless Communications Letters*, 2022.

[16] J. Pan, H. Shan, R. Li, Y. Wu, W. Wu, and T. Q. S. Quek, "Channel estimation based on deep learning in vehicle-to-everything environments," *IEEE Communications Letters*, vol. 25, no. 6, pp. 1891–1895, 2021.

[17] C. Luo, J. Ji, Q. Wang, X. Chen, and P. Li, "Channel state information prediction for 5G wireless communications: a deep learning approach," *IEEE Transactions on Network Science and Engineering*, vol. 7, no. 1, pp. 227–236, 2020.

[18] C. Li, L. Liu, and J. Xie, "Finite-state Markov wireless channel modeling for railway tunnel environments," *China Communications*, vol. 17, no. 2, pp. 30–39, 2020.

[19] I. Sen and D. W. Matolak, "Vehicle–vehicle channel models for the 5-GHz band," *IEEE Transactions on Intelligent Transportation Systems*, vol. 9, no. 2, pp. 235–245, 2008.

[20] H. Yang, M. H. A. J. Herben, I. J. A. G. Akkermans, and P. F. M. Smulders, "Impact analysis of directional antennas and multiantenna beamformers on radio transmission," *IEEE Transactions on Vehicular Technology*, vol. 57, no. 3, pp. 1695–1707, 2008.

[21] A. K. Gizzini, M. Chafii, A. Nimr, and G. Fettweis, "Deep learning based channel estimation schemes for IEEE 802.11p standard," *IEEE Access*, vol. 8, pp. 113751–113765, 2020.

Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 4979057, 12 pages
https://doi.org/10.1155/2022/4979057

Research Article

An Information-Centric Network Caching Method Based on Popularity Rating and Topology Weighting

Yaxin Chang,[1] Jiafei Guo,[2] Hanbo Wang,[2] Dapeng Man,[2] and Jiguang Lv[2]

[1]*China Energy, Beijing 100011, China*
[2]*Information Security Research Center, Harbin Engineering University, Harbin 150001, China*

Correspondence should be addressed to Jiguang Lv; lvjiguang@hrbeu.edu.cn

Received 6 June 2022; Revised 13 July 2022; Accepted 22 July 2022; Published 11 August 2022

Academic Editor: Lisheng Fan

Ubiquitous caching is a feature shared by all proposed information-centric network (ICN) architectures. Prioritising storage resources to popular content in the network is a proven way to guarantee hit rates, reduce the number of hops forwarded, and reduce user request latency. An ideal ICN caching mechanism should make the best use of relevant information such as content information, network state, and user requirements to achieve optimal selection and have the ability to adaptively adjust the decision cache content for dynamic scenarios. Since router nodes have limited cache space, it is then useless to accurately predict the popularity of the content with very low popularity, as this content has no chance of being cached. A more effective approach is to focus on content with high popularity that influences caching decisions. As for different nodes, they have different sets of popular content, and using this property, this paper designs a caching method based on the popularity hierarchy with topological weights. The method considers managing the cached content in nodes with a hierarchy of popularity and improving their distribution in terms of the importance of the nodes' position in the network. Finally, the scheme is simulated by changing the parameter settings under different actual topologies on the simulation platform to confirm the feasibility of the scheme.

1. Introduction

Different from the TCP/IP architecture, ICN does not need to establish a connection between the source address and the destination address before data transmission but directly finds and receives content in the way of user-initiated requests. The router can store the content forwarded through it for a period of time (this storage depends on the size of the cache space and the replacement policy as well as the timeliness of the content) and make it available to the requesting consumer on a hop-by-hop basis. The network's built-in cache is therefore an important feature of ICN networks, and it is one of the two goals of the proposed future Internet. When a router node serves a query for content, the local cache node may have a copy in its content store (henceforth CS) for the purpose that if a new request for that content arrives, it can be satisfied locally, rather than being forwarded to the source-destination server node for network resources [1]. This approach will improve content hit rates, increase bandwidth utilization, and reduce the number of hops of data forwarding and content retrieval latency, which is one of the biggest differences between ICN and TCP/IP architectures other than the protocol stack.

The idea behind caching is that the content being cached is assumed to be potentially accessible in the future, and storing them on the router reduces the average hop count and reduces the load from redundant network traffic. The primary advantage of serving content from the router to the user is that the user experiences lower latency, thanks to the fact that not all requests need to be routed to the content host; the content may be cached on the router during the return of the packet to the user, depending on the cache management policy. Secondly, because routers close to the user (i.e., at the edge of the network) can respond to a large proportion of requests, this alleviates network congestion to a certain extent. However, similar to the

CDN technology in the TCP/IP network, the ICN network cache also faces the problem of low cache utilization caused by the misuse of the cache.

In recent years, the ICN caching technology based on content popularity has been widely proposed to solve the problem of cache misuse by calculating the historical popularity as an indicator for deciding the content to be cached by nodes in the future through a statistical approach to improve the efficiency of cache usage. However, in practice, the impact of the statistical popularity method on the cache performance of router nodes has not been considered, and there is a problem of taking up a lot of computing resources and space and even pulling down the retrieval rate. The caching strategy that uses content popularity alone as a decisive indicator of whether to cache content does not take into account the issue of content redundancy and whether it is adapted to the network topology in which the node is located, as it can only significantly reduce user request latency if the cached content is sunk to an edge node that is closer to the users.

In order to improve the cache utilization and improve the user's network experience as much as possible, ICN needs an effective cache mechanism. In an actual network, different nodes are interested in different contents, so the content preferentially stored by different nodes is also different. According to the above reasons, this paper proposes a caching scheme based on content popularity and topology weighting.

The main contributions are as follows:

(i) A popularity-based cache hierarchy is proposed to divide contents with different popularity into four levels

(ii) It is proposed to combine the popularity level and topology weight as the cache priority index of nodes to make reasonable cache decisions

(iii) Through comparative experiments, the feasibility and effectiveness of the caching mechanism in the ICN environment are verified

2. Related Work

Caching mechanisms commonly used in existing architectures implementing ICN networks include LCE [2], LCD [3], MCD [4], and Prob [3], but these caching methods are generally based on the migration of web caching to ICNs and do not fit perfectly with the characteristics of ICNs. For example, LCE is a "ubiquitous" copy, which improves cache hit rates and reduces content retrieval time but does not provide efficient management of network resources. Content that will not be revisited leaves a large number of copies in the global network, and storing them in cache space is a misuse of cache resources. Although ICN maintains some caching strategies at the beginning of its design, these methods all have certain problems: Prob can be seen as an improvement on the random form of LCE, but nodes can only keep local copies with probability p. LCD reduces redundancy by sinking the content cache to the next hop node in the cache hit but still causes all nodes in the communication path to cache the same content, using bandwidth to the maximum. MCD reduces the number of identical copies between the requesting host and the server but increases request latency due to eviction operations.

Content popularity distribution is an important network parameter that affects the performance of caching policies, and caching policy design based on content popularity is becoming a common approach. Existing caching policies based on content popularity are based on the analysis of user preferences, and models are trained by collecting various metrics to distinguish or predict popular content. These methods can increase the cache hit rate of network edge nodes, but there is a problem that training the model requires a lot of data collection and resource-intensive parameter update and also fails to effectively utilize the cache resources of nodes in the upstream network.

Several studies [5–7] use request frequency, content timestamp, content quantity, content name, etc., as popularity evaluation metrics. A piece of content is only cached when its popularity exceeds a set threshold. However, if certain content is heavily accessed during a certain time period, this will lead to an increase in the threshold value, new requested content will not enter the cache, and the cache hit rate will be reduced as a result.

Zhang et al. [5] in the Optimal Cache Placement based on Content Popularity (OCPCP) policy calculate the popularity of incoming content based on the stored content and store new content based on its popularity value. OCPCP makes caching decisions by considering only the content request records on a single node. That is, if the number of content requests is high, it has higher popularity and is therefore considered for caching.

Time-aware least recent use (TLRU) [6] is a content lifecycle-aware eviction policy that improves on the LRU cache management policy, where the timestamp of arriving content is calculated locally by the cache node. If the average request time is less than the timestamp of the stored content, the arriving content is cached. If space is available in the cache, TLRU stores the content; otherwise, it applies a simple LRU to the cached content, creating space for new arrivals.

The fine-grained popularity-based cache (FGPC) [7] will cache all incoming content if the control center of the network node is not full. Otherwise, it stores only popular content and periodically modifies the content popularity threshold. When forwarding content downstream or receiving content from upstream, three statistics are updated in this policy, namely, the content counter, content name, and timestamp. If the value of the new content is greater than the predefined threshold, FGPC uses the LRU policy to cache the new content in the CS; otherwise, it is ignored.

The ProNDN scheme proposed by Pu [8] is a combination of per-network state forwarding and in-network caching policies. Its collaborative data caching consists of two schemes: CacheData and CacheFace. In short, it combines both schemes with the default in-network cache. Popular data is cached using CacheData; if the router holding the data is closer to the edge router than to the producer, then CacheFace is used to cache the interface. Otherwise, the default caching scheme is used.

Zhang and Wang [9] achieve collaborative content and space utilization between local and neighbouring nodes by caching the content replaced by the local node in the neighbouring node's cache. When a local node receives a packet of interest, the node with which it shares information can satisfy that content. Zheng et al. [10] proposed a noncooperative game theoretic-based ICN pricing model for free content to address the problem that existing ICN pricing mechanisms only study paid content and ignore free content in the network. Considering the coexistence of paid and free content in real networks, the authors analyze the impact of caching and pricing on the revenue of all entities and develop a win-win pricing strategy.

Liu and Han [11] focus on allocating cache sizes for each node within a given total cache space budget. The authors explored the impact of heterogeneous cache allocation on content dissemination under the same ICN infrastructure, quantifying the importance of nodes in terms of content dissemination and network topology. They implemented a hierarchical approach based on content dissemination, then developed a set of weight calculations for these hierarchies, and provided cache space allocation per node to assign the total cache space budget to each node in the network. Shekhawat et al. [12] proposed a heterogeneous path cache budget allocation method based on the reference location of nodes to assign caches to content stores. The experimental results were compared with a traditional on-path caching decision (the data is sent back according to the interest forwarding path) mechanism and achieved a 14% improvement in the cache hit rate.

Chiu et al. [13] investigated a two-tier caching scheme where administrative autonomy was achieved by adding nonpath collaborative caches within the management node AS to eliminate redundancy.

Alhowaidi et al. [14] devised a centralized approach to managing/mapping the current content of CS using SDN controllers. SDN controllers are effectively used to analyze the network state and redirect incoming interest to off-path routers that have cached the requested content. This approach enhances the data by allowing NDN consumers and NDN routers to fetch content from multiple off-path locations based on the network state.

While these approaches are clearly layered, real-world networks are more complex than experimental ones, and the way a piece of content is stored during delivery may be a mix of these characteristics. In some mobile network topologies, interconnected communication nodes are constantly changing, and changes to the cached content set become frequent as a result. These situations illustrate the complexity of caching in real networks, increasing the redundancy and replacement of cached copies. Therefore, business-oriented requirements need to be considered when designing caching solutions, as it is difficult to find a universal solution.

3. Analytical Methods

This paper proposes an on-path caching policy algorithm PT-Cache (popularity-topology cache) that uses packet popularity ratings to make them compete directly on caching nodes based on their potential to save forwarding hops for future packets of interest. The strategy not only improves cache hit rates but also looks to reduce packet forwarding hops by analyzing the network topology with nodes closer to the user for caching. Nodes at higher levels cache less popular content so that the content of most interest to users is cached for hits at the edge routers and the rest is stored at the upstream routers, improving the overall hit rate and cache space usage efficiency and reducing the number of hops to forward packets of interest for popular content.

3.1. Content Popularity Rating Model. In the content popularity rating model of this paper, the data content in the ICN is graded according to the popularity of the content, and the level indicates the importance of whether the node is cached, represented by the following: 0 to 3.0 means unnecessary caching, 1 means unknown level, 2 means cacheable, and 3 means high cache value. The content should be graded differently for different node popularity, because the scheme takes into account both the location of the caching node in the network topology. That is, the behavioural performance of the content on a communication path is depicted as shown in Figure 1.

Figure 1 shows part of the communication path in the network. Consumers 1 and 4 both request C_n and Consumer 1 also requests C_q. So, C_n is the more popular content and is cached by R_1. However, in R_3, the cache level of C_n drops to 1. Similarly, for R_2 and R_4, the highest levels are C_m and C_q requested by Consumers 2 and 3. In particular, Consumer 2 also requests C_p at the same time at a lower frequency, so the level of C_p is somewhat lower than C_m. $C_o$is stored at the R_5 node close to the producer, and since it is not stored at any of the edge nodes, the packet of interest will be forwarded to R_5.

3.1.1. Request Frequency Collector. Massive content access is the most important feature of ICNs. With the emergence of more and more large content providers and user-produced content applications, there is a trend towards diversification of content and applications on the network. The transparency of in-network caching in ICNs separates applications from caches, allowing different types of content from different content servers to be stored in the cache space of the same node. This makes the analysis of network caching systems very difficult. The idea of popularity hierarchy was explained in the previous section in conjunction with a diagram. Since the popularity of content varies over time, in order to achieve a classification for the popularity level, it should be dynamically adjusted according to its request frequency. Specifically, this section implements a packet-of-interest request frequency collector, which is used as a basis for differentiating the popularity of content entries stored by router nodes into different levels.

The popularity of the content reflects the user's interest in the content at the local time for a period of time. In order to ensure that the content popularity can truly

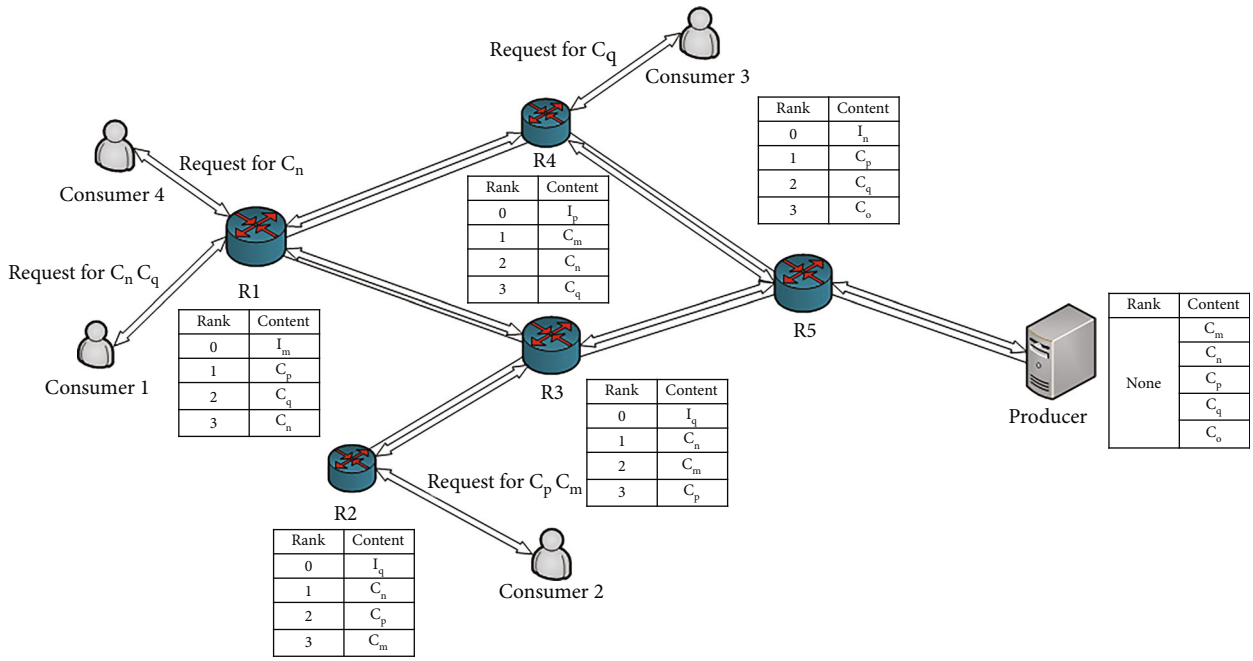


Figure 1: Interest packet forwarding process.

reflect the content request situation of the current network, each router node will regularly update the access frequency of the interest packet. Specifically, this scheme first defines a time period T. Then, each router node updates the request frequency of interest packet I_i corresponding to the requested content C_i at the end of each period, denoted by rF_i^j, as in

$$rF_{C_i}^j = \beta * MF_{C_i}^{T_j} + (1-\beta) * rF_{C_i}^{j-1}, \quad (1)$$

where j is the count index of a certain time period, indicating the jth time period. $MF_{C_i}^{T_j}$ is recorded to indicate the number of requests for packet of interest I_i collected by the router node in the jth period. $rF_{C_i}^{j-1}$ is the frequency of interest packet requests recorded in the previous time cycle, i.e., the $j-1$th period. And $\beta \in (0, 1]$ is a weighting factor. Obviously, β reflects the trend of tracking the change in request frequency. The larger the β, the faster the response to changes in the frequency of interest packet requests. However, using a large β value will also cause the observed request frequency to fluctuate more frequently as well. For this reason, it is set by default to 0.6 in this chapter, in the hope that it will reflect the trend in user preferences and thus consider future popularity.

3.1.2. Popularity-Based Cache Hierarchy. The set of stored content items is divided into 4 levels of different popularity. Consider a collection of content items M of size $|m|$ whose content popularity follows a Zip-f distribution. Thus, let rank be a random variable indicating the popularity level of the requested object, then rank (C_i) is the popularity level of content object C_i, which when it takes the values 0, 1, 2, or 3, respectively, denotes the meaning as shown in Table 1.

Table 1: Cached content collection popularity ranking.

Popularity rank	Implication
0	No cache necessity
1	Unknown prevalence
2	Consider to cache
3	High cache value available

Thus, rank $(C_i) = 3$ is the set of most popular content objects, i.e., they account for 40% of the total number of requests. Similarly, rank $(C_i) = 2$ is the set of objects that receive the next 30% of requests; while rank $(C_i) = 1$ is unknown for its popularity rating, they account for 20% and have a tendency to shift to a potentially higher popularity rating, as well as the possibility of swapping out content entries that belong to a rating of 2 or 3. Finally, the set of content entries with rank $(C_i) = 0$ accounts for 10%. It is worth noting that their state in the cache changes most erratically. Because if new content wins the storage competition when the cache fills up, they are the set that is to be replaced out first, but this does not indicate that they will never be cached by a node.

The structure of the hierarchy in the node cache is shown in Figure 2. For a set of content collections classified to the same popularity level, they will be sorted according to the observed frequency of interest requests. That is, there are actually two combinations of sorting here, the first by popularity level and the second by request frequency. Content collections that are downgraded from a higher level of popularity are stored in the next level of collections first comparing their popularity with the top entries in that level and then finding the right place to insert them, thus

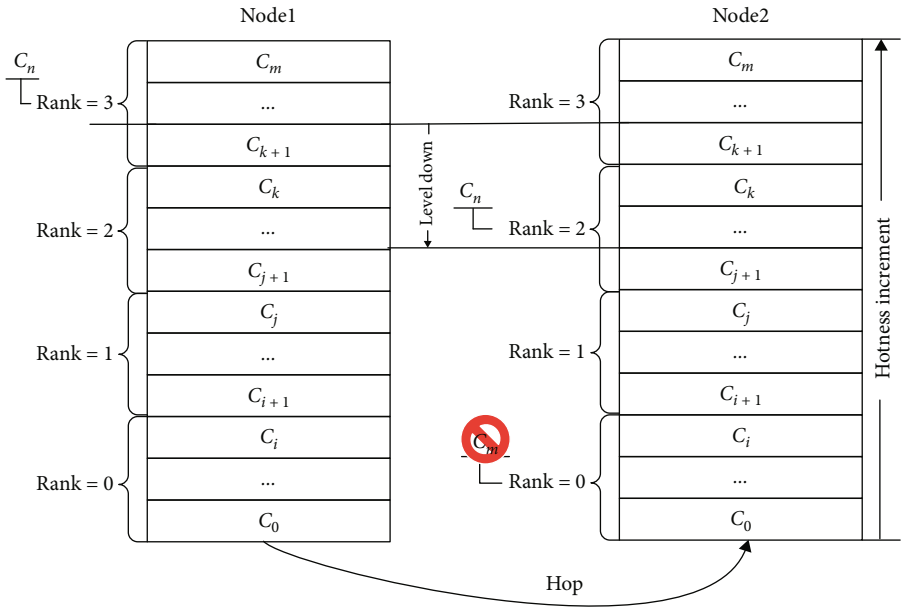


Figure 2: Distribution of popularity content segmentation.

distinguishing between frequently hit and frequently replaced content collections.

In Figure 2, the content popularity is increasing in rank. C_n being cached in the CS of node 1 has a rank value of 3, which is determined by the popularity of the content, as consumers connected to node 1 will have a higher frequency of requests for C_n, which will then have a higher popularity level according to the setting in the scenario. Obviously, the impact of this scenario for node 2 is that node 1 will not send interest packets for C_n from node 1 to node 2 when it satisfies the interest packet request for C_n, so the rank value of C_n becomes smaller in node 2. The set of contents stored in each node along the entire communication path is not identical; they are intersecting sets. So, after the CS in node 2 stores the new content, it expels C_m with a rank value of 0 from the storage space.

When a router node receives a request for content belonging to rank $(C_i) = 0$ and records its interest packet request frequency, if the cache is full, it will mark the interest packet with the topology node weight BeCentry(n) value and then forward it to the next-hop router node; the next-hop node compares this value with its own BeCentry(n) value based on this value, and if the next-hop BeCentry(n) value is higher than the source router's BeCentry(n) value, then the popularity level is raised at this next-hop node, and then, the decision to store the packet in the cache is based on the correlation analysis of its popularity level with the topology node weight.

The popularity ranking is a mapping of the consumer's interest in the content to the rank, and the individual content entries are ranked with the calculated interest packet request frequencies between them to facilitate the comparison of their popularity levels with the request frequencies. Once the interest packet request frequency is collected, if the interest packet cannot be responded to locally, the interest packet structure is reconstructed to include the request

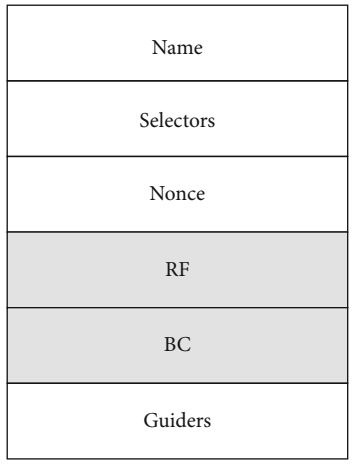


Figure 3: Modified interest package structure.

frequency rF_i^j collected at that node in the header information of the interest packet. Therefore, the interest packet needs to add the marker field RF indicating the request frequency, which is extracted for processing when the interest packet is received and added for updating when it is forwarded.

3.2. Topological Node Weights

3.2.1. Topological Node Importance Evaluation Method. The homogeneous caching scheme ignores the differences in the importance of nodes in the network topology and content distribution. Although this default scheme is simple and easy to implement, it does not make full use of the caching capacity of upstream nodes and does not distinguish the difference in caching capacity between core and edge nodes, which seriously affects the performance of the caching system.

Input: Node Received Interest packet Interest(C_i)
Output: The ranking state and processing status
if$C_i$is in CS**then**
  Parse header information from interest packages for**BC, RF**
    Calculate ΔBCaccording to (4-3)
if$\Delta BC \geq 0$**then**
$MF_{C_i} \longleftarrow MF_{C_i} + RF$
else
$MF_{C_i} \leftarrow MF_{C_i} - 1$
    Forward $Data(C_i)$ in the reverse path, according to the pit interface
return SUCCESS
    else
    rank (C_i)initialized as 1
    Parse header information from interest packages for **RF**
    if$Interest(C_i)$ is in PIT**then**
    $RF \longleftarrow new\,RF + RF$
    Add **RF** to interest package, update $Interest(C_i)$ item in PIT
    Discard $Interest(C_i)$
    return SUCCESS
    end if
    end if
    if$Interest(C_i)$ in FIB **then**
    Add **RF** to interest package, update $Interest(C_i)$ item in FIB
    Forward $Interest(C_i)$ to the next hop
    return SUCCESS
    else
      Discard $Interest(C_i)$
    return FAIL
    end if
    Update the rank layout of the content according to equation (1)

Algorithm 1: The packet of interest processing for this strategy.

Therefore, to improve the performance of the ICN caching system, a heterogeneous allocation approach is needed, where nodes that play an important role in content distribution should be allocated a larger cache size. Therefore, network topology will necessarily have a significant impact on content distribution, but in many scenarios, the importance of nodes in the topology is not exactly the same as the importance of nodes in content distribution. Data paths on router nodes connecting multiple users contain a large number of different requests, and such nodes should be allocated more cache size so that they are weighted higher than paths connecting fewer users.

As will be explained next, the centrality of a node will be a measure of how widespread cached content is when served on that node. Considering network topology information to define metric node weights, it provides support for designing hierarchical caching strategies based on topology weights and expected popularity.

Node centrality is used to measure the number of times a node appears in the content delivery path [15]. Nodes with higher centrality values can access more content on the network. Limited by the size of the cache, the priority of cache node selection is proportional to the centrality; that is, the higher the centrality, the higher the score, and the more content the node should cache. Using this graph to represent a router network, the centrality of router node n is expressed as the following:

$$\mathrm{BeCentry}(n) = \sum_{\forall s,t \in V \setminus n} \frac{\sigma_{st}(n)}{\sigma_{st}}, \tag{2}$$

where σ_{st} is the number of shortest path entries between s and t and $\sigma_{st}(n)$ is the number of shortest path entries between s and t through n.

The assumption made for the topology node weight-based part is that $\mathrm{BeCentry}(n)$ is calculated offline in advance for all routers, so that each router node knows its own $\mathrm{BeCentry}(n)$ value and marks this value when forwarding a packet of interest, i.e., by adding a BC field to the packet of interest. In the caching policy, router nodes receive requests and not only extract the RF field from the packet of interest to aggregate the content request frequency but also determine the popularity level of the requested content object. The $\mathrm{BeCentry}(n)$ value of the router node that originated the packet of interest request is also extracted from the packet of interest for comparison with the $\mathrm{BeCentry}(n)$ value of the current node, as shown in

$$\Delta \mathrm{BC}(C_i) = \mathrm{BeCentry}(n) - \mathrm{BeCentry}\left(C_i^{r1}\right), \tag{3}$$

Input: Node Received Data packet $Data(C_i)$
Output: Processing status
1: **if**C_i not in PIT **then**
2: Discard the $Data(C_i)$
3: **return** FAIL
4: **else**
5: **if**rank $(C_i) = 3$**then**
6: **if***CS*is Filled **then**
7: Execute LFU policy within current rank,delete content in rank = 1
8: Insert C_i into *CS*
9: **end if**
10: **else if**rank $(C_i) = 2$**then**
11: Execute LFU policy within current rank,delete content in rank = 2
12: **else if**rank $(C_i) = 1$**then**
13: Insert C_i into *CS*in probability according to equation (4)
14: **elseif**rank $(C_i) = 0$**then**
15: **continue**
16: Forward $Data(C_i)$ according to FIB
17: return SUCCESS
18: **end if**
19: **end if**

Algorithm 2: The packet of data processing for this strategy.

where $\Delta BC(C_i)$ represents the difference between the mesoscopic centrality of two router nodes, $\Delta BC(C_i)$ is the mesoscopic centrality of the current node n, and $\Delta BC(C_i)$ is the currently received mesoscopic centrality for content C_i from node $r1$. They may take on positive, negative, or zero values; if positive, it indicates that the current node is more important for content C_i; if negative, it indicates that node $r1$ has a stronger role in the topology; if zero, it indicates that both are of equal importance.

3.2.2. Correlation Analysis of Popularity and Topological Weights. The goal of the caching policy is to select a subset of router nodes in the transport path for specific content caching. The algorithm is based on the correlation of content popularity and topology importance. The two correlations show how well the two match from a content perspective and a topology perspective. The hierarchy of nodes for cached content should adapt to the different distributions of nodes in the network topology and be able to make dynamic adjustments on its own.

The correlation function used in the calculation of the correlation between popularity and topological weights calculates the absolute difference between the two variables. As expressed in

$$\mathrm{COV}(\mathrm{RF}, \mathrm{BC}) = \mathrm{Correlation}(\mathrm{RF}, \mathrm{BC}). \quad (4)$$

In this scenario, a popularity-based scheme is used to update the cache. As mentioned earlier, each packet carries a content popularity level rank on its transmission path. When a decision is made to add a new revenue content to a router whose cache is full, the router compares the frequency of requests for the new content with the rank of the content in the same popularity level in the cache. If the new content is requested more frequently, the content in the cache with a lower request frequency in the same popularity level will be ranked decentralized; this operation will only cause a redistribution of the content popularity level distribution in the content collection if the cache is not full; however, when the cache is full, a cache replacement operation will be triggered. Otherwise, the new content will be discarded.

Table 2: Basic experimental parameter settings.

Parameter name	Parameter values
Topology	GARR, GEANT, TISCALI, WIDE
Request rate	30 s^{-1}
Number of contents	100000
In-network cache size	[0.03,0.05,0.08,0.1]
Values of α	[0.8,1.2,1.6,2.0]
Int-domain time delay	2 ms
Out-domain time delay	30 ms
Number of prefilled caches	3000
Actual number of caches measured	6000
Branching factor	8

3.3. PT-Cache Specific Solution Design. To begin with, if a router node fails to respond to a particular packet of interest, the frequency of interest packet access to that node will be aggregated to the next-hop router node. In practice, two phenomena exist in ICN networks.

(1) The closer a router node is to the source server of a packet of interest, the better the chance that the

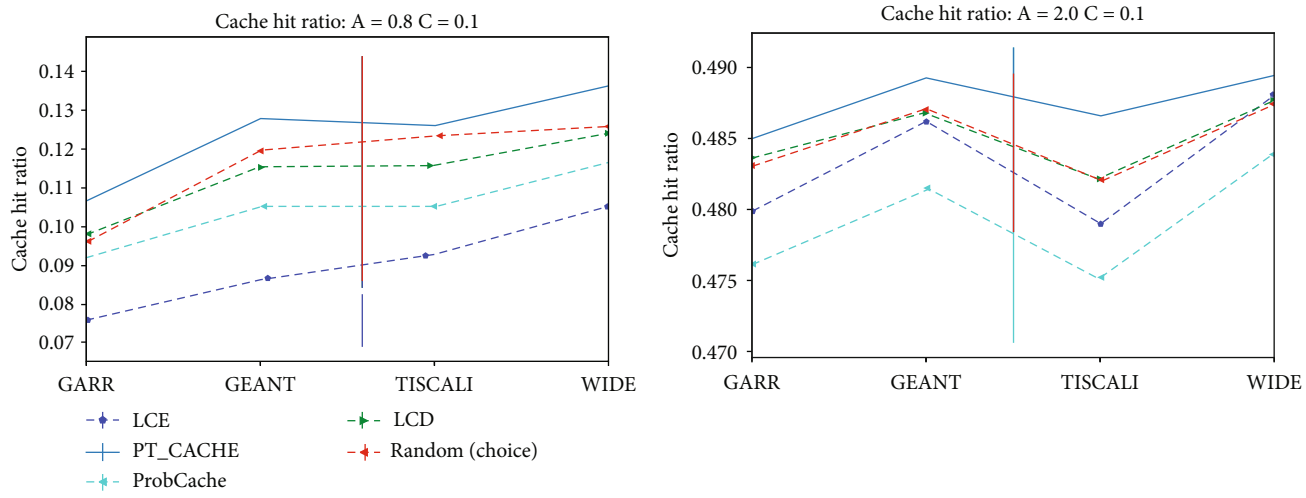


Figure 4: Cache hit rate for different α values in different topologies.

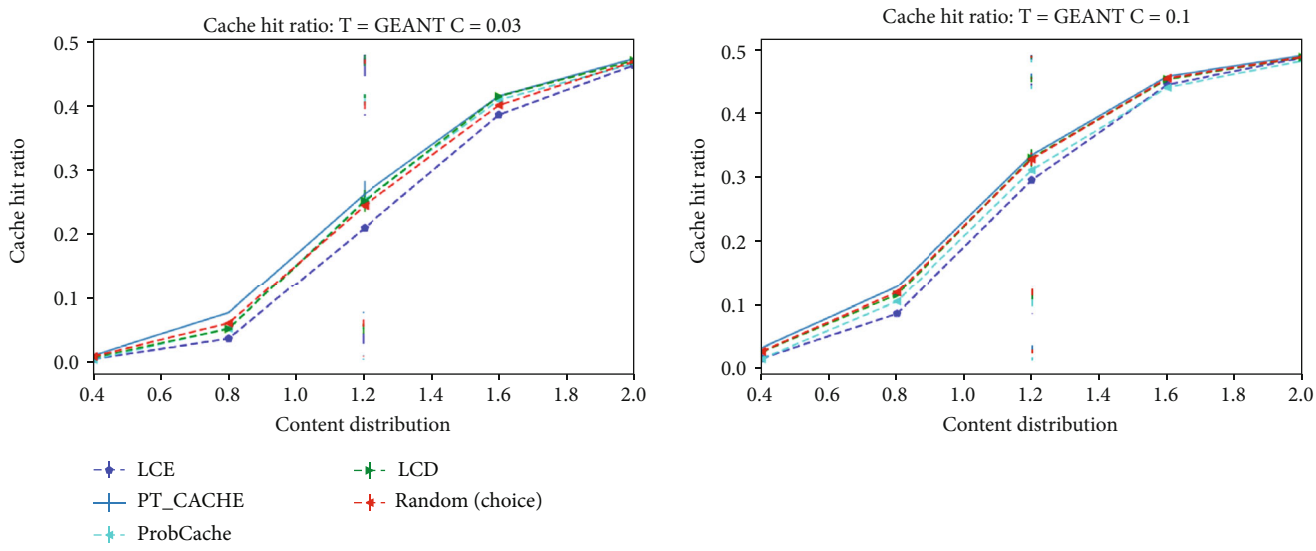


Figure 5: Cache hit rate for different in-network cache sizes under GEANT.

router node will aggregate the frequency of access from the packet of interest. In other words, the closer a router node is to the producer, the more likely it is to aggregate packets of interest sent from downstream

(2) The closer the packet is stored to the node requesting it, the fewer hops it can reduce the number of forwarding hops for the same request frequency. This is considered for the case where the packet of interest needs to be forwarded

By combining the two scenarios above, it can be seen to store popular content packets close to the consumer and less popular packets away from the consumer. In this way, packets with high popularity levels are maximised to preserve the number of forwarding hops, while more frequent requests are aggregated for content with average popularity levels. To this end, the following strategy is proposed.

(1) All packets arriving at the router compete for cache space based on popularity level and topology node weights

(2) The expected caching or noncaching of packets is calculated based on the correlation between the frequency of packet of interest requests measured at the router node and the topology node weights

In this process, packets with high popularity levels are considered first in the competition of nodes to be cached due to their higher request frequency. However, their caches at or near the edge router will directly satisfy the request. Packets of interest will be "blocked" at this node, thus

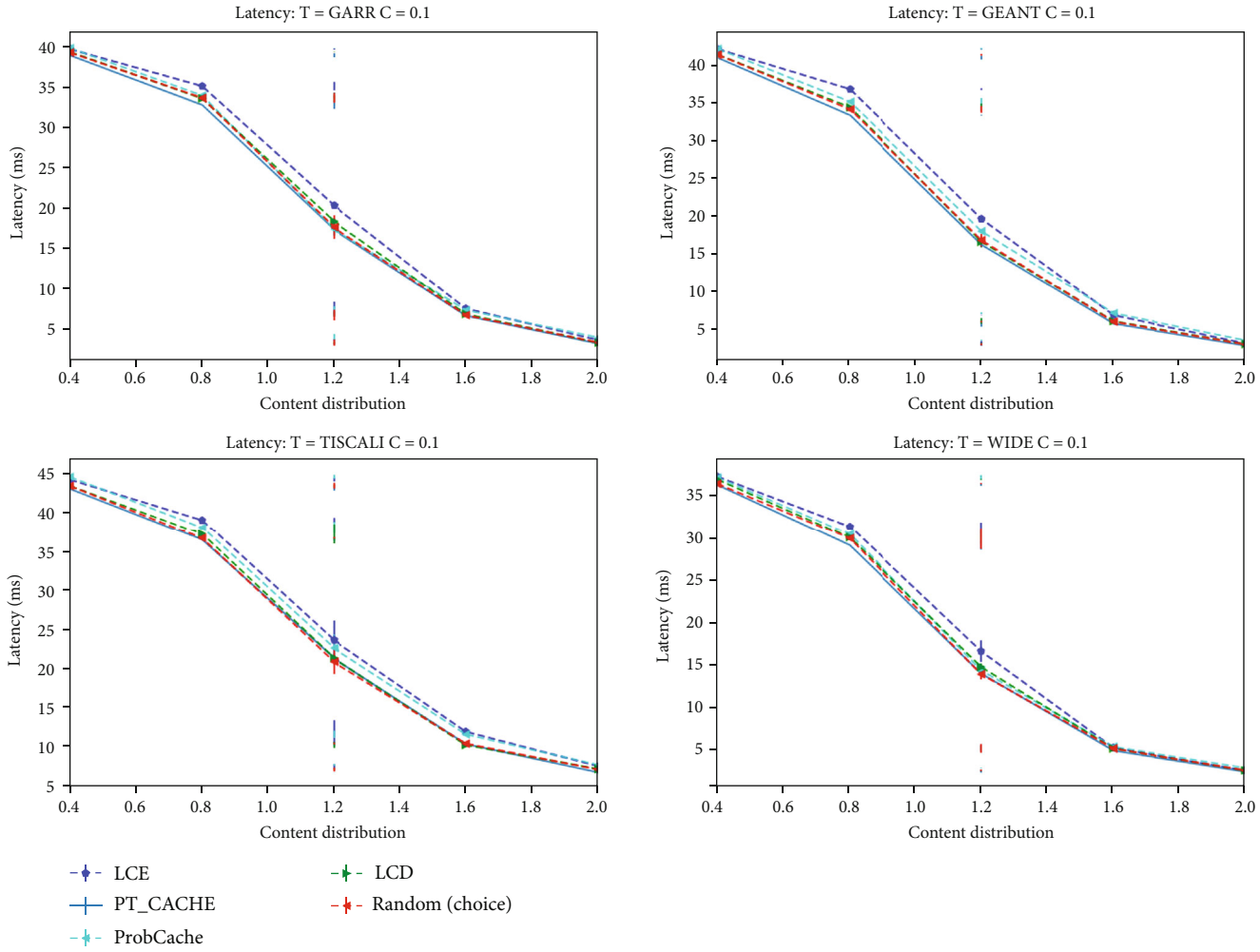


Figure 6: Average response delay for different topologies.

reducing their frequency of requests at routers close to the consumer. This allows the popularity level to change from router node to router node, so that content storage nodes are reasonably spread out in the network, rather than frequently replaced and rewritten at the same node. Packets with lower prevalence levels at edge nodes are able to boost the prevalence level height and thus gain caching opportunities when they are forwarded to upstream nodes by aggregating the request frequency. As a result, popular packets sink towards the consumer side and unpopular packets move closer to the producer side.

In order to implement the collection of interest packet request frequencies and the calculation of relative topological weights in the scheme, a modification to the interest packet structure is required, which modification is shown in Figure 3.

The RF and BC fields are added to interest packets; RF indicates the content popularity level of the interest packet on its source node and the frequency of interest packet requests, and BC is the mesocentricity of the node. The validity stems from the fact that as requested content is added to the cache, they are ranked up and down in order of content popularity. By adopting this strategy, diversity in the cache repository can be achieved, as less popular content can be cached on more distant nodes. Algorithm 1 shows the packet of interest processing for this strategy.

The solution does not change the overall structure of the packet, but does so because in an information-centric network, where content is the object of distribution, all copies of the same content in the network should be identical. Caching in different nodes is simply a copy of the content, and modifying the packet not only defeats the purpose of the clearinghouse network design but also bloats the packet and takes up valuable network resources during communication. Algorithm 2 shows the packet of data processing for this strategy.

4. Experimental Results and Analysis

4.1. Experimental Program. This paper implements the proposed PT-Caching scheme on Icarus [16]. The hardware environment on which the platform runs is the same as in Section 3 and will not be elaborated on. Icarus is a python-based caching simulator for ICN-based architectures (focusing on CCNX and NDN), using named content, a request-response model (e.g., interest and content

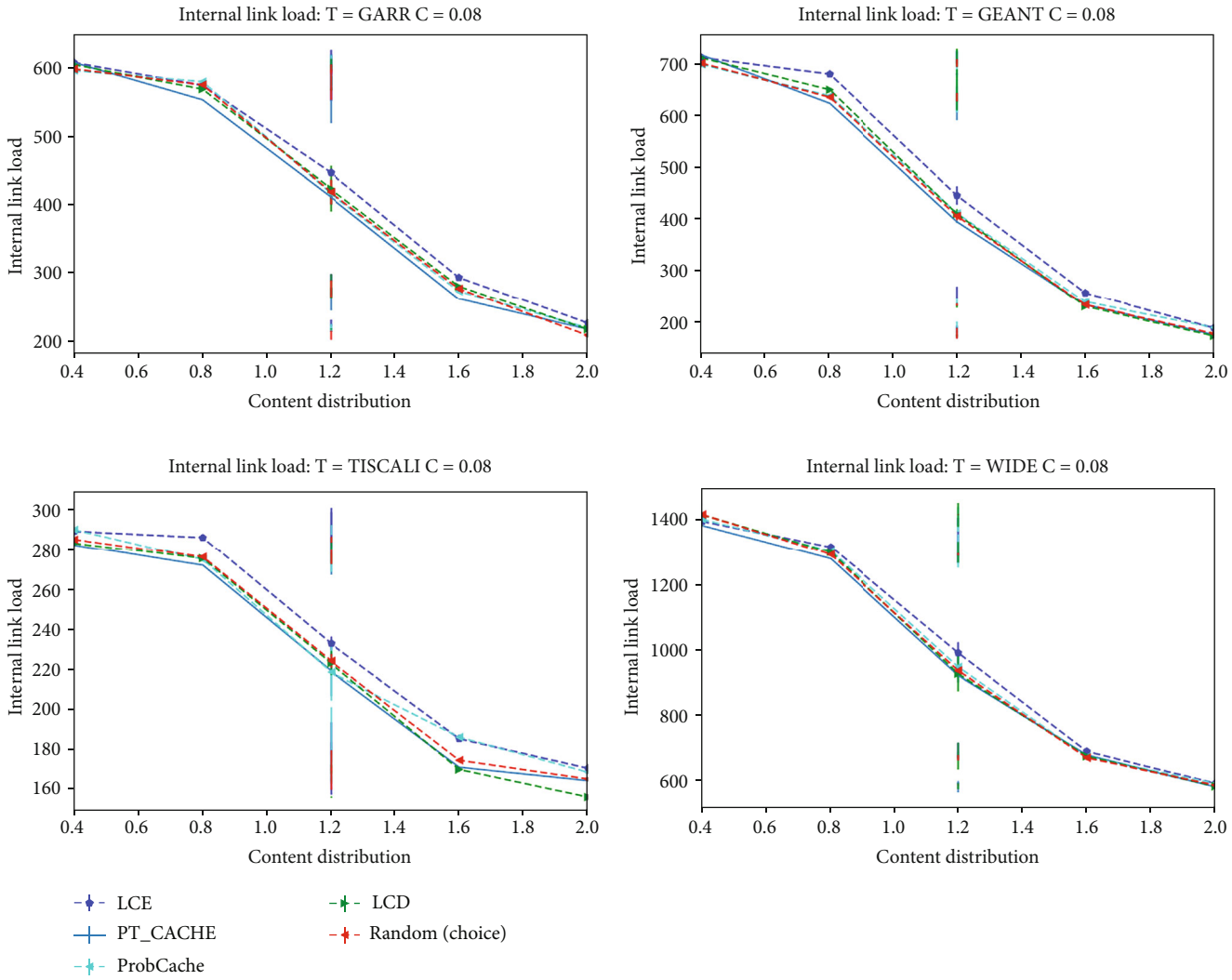


Figure 7: Link load for different topologies.

request), and supporting various evaluation protocols. The proposed solution is compared in performance with the cache placement policy (location of placed content copies) supported by Icarus.

Four real network topologies were used in the experiments, namely, GARR (Italian national university and research computer network), GEANT (European data network for the research and education community), TISCALI (Italian telecommunication network), and WIDE (from Japan). The performance of the cache hit rate, average latency, and link traffic for each scenario was evaluated for different content popularity distribution parameters α and total network cache capacity, and the basic parameters of the simulation were set as shown in Table 2.

In Table 2, $30\,s^{-1}$ is the number of requests per second (over the whole network) that belongs to the experiment parameters. That is to say, $30\,s^{-1}$ is set to thirty requests per second.

This section compares the proposed scheme with the LCE, LCD, ProbCache, and Random schemes setting different network topologies, α values to compare cache hit rates, average response latency, and link traffic. The following are the basic ideas of the four schemes.

LCE: in this method, packets are cached at each node of the path as they are forwarded downstream. This means that the content is cached at each node along the path.

ProbCache: this policy reduces the redundancy of cached content by probabilistically caching content on the way.

Random: in this caching policy, content can be cached on any of the downstream nodes. The content is cached on a particular downstream node which is chosen at random.

4.2. Analysis of Results

4.2.1. Cache Hit Rate. Figure 4 shows the hit rates for the four network topologies for different α fetching values. α values are used to generate the Zip-f distribution of content requests and must be positive. The larger the value, the more skewed the distribution of content popularity. Therefore, as the α value increases, the hit rate of all scenarios improves. PT-Cache shows a high hit rate in all topologies, and the performance advantage is especially evident in the GEANT

topology. This is since the GEANT topology has a more pronounced hierarchical feature compared to other topologies, which is suitable for the content popularity grading and extraction of node topological weights in the proposed scheme. The hit rate performance of each scheme in the GEANT topology for different network cache sizes can be observed in Figure 5. As the content distribution popularity value increases, the hit rates of all schemes improve, with LCD and PT-Cache having the best hit rates and LCE being the worst, and this scheme will have a large amount of cache redundancy in the path.

4.2.2. Average Latency. Average latency quantifies the duration from requesting a file to delivering it and is an intuitive representation of the user's network experience. As the use of streaming applications grows, response time becomes increasingly important. The average hop count is also a factor in the average latency, so the average latency is used to measure whether the average hop count has decreased. Shown in Figure 6 is the average response latency for each topology at different α values. At α values less than 0.8, each strategy exhibits a high average delay with a small difference in values. However, when the value of α is larger, within the interval $[0.8, 1.6]$, the average latency of each strategy tends to decrease substantially with the concentration of content popularity. However, when the value of α exceeds 1.6, when the content popularity is very concentrated, the proposed scheme has no obvious advantage in this case, the average latency of all the schemes decreases at a slower rate, and the performance of each strategy is relatively similar.

4.2.3. Link Load. The amount of data passing through the transmission path at each time is defined as link traffic. In an ICN, this metric is positively correlated with redundant traffic. Analyzing link occupancy helps to ensure the quality of service in the network. Figure 7 shows that the trend of link traffic under the influence of the α value is very similar to the first two metrics. On the one hand, with $\alpha = 1.6$ as a threshold, the performance advantage of the PT-Cache becomes insignificant. The PT-Cache policy has the lowest load on the link traffic in the range $[0.4, 1.6]$, due to its hierarchical treatment of the cached content in the nodes on the path, which better eliminates redundancy. On the other hand, the higher prevalence of content that can be cached to the nodes brings a high hit rate, which reduces request retransmissions.

The results of these experiments show that the PT-Cache method performs better under the three metrics of cache hit rate, average latency, and link traffic, even under different topologies, with the best performance under α values of $[0.8, 1.6]$, and its advantage is no longer obvious beyond 1.6. This is because it is difficult to distinguish clearly between the levels of content distribution, which is like that of a dedicated server, and the other solutions also show a significant performance improvement under such conditions. Therefore, the solution is better suited to be deployed in networks with a high diversity of content and a high degree of dynamism to take advantage of it.

5. Conclusions

High cache hit ratio, low data retrieval cost, and low user latency are the goals pursued in designing cache mechanisms. This section designs a caching method PT-Cache based on popularity and topology weighting. For different nodes, the content they are interested in is different, so for different nodes, the content to be stored preferentially is also different. Therefore, content popularity and packet hop-saving capability are the basis for packets competing for caching opportunities. Adopting this method helps router nodes to distribute content more reasonably and reduce cache redundancy in the network. Finally, PT-Cache is compared with other caching schemes with different topologies through simulation, which shows that PT-Cache can achieve better performance under different topologies and parameters, and the scheme can be effectively applied to various social network type web apps.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research is supported by the National Natural Science Foundation of China (Grant Nos. 61831007, 61971154, and U21B2019) and the Fundamental Research Funds for the Central Universities (Grant No. 3072022CF0601).

References

[1] M. Conti, A. Gangwal, M. Hassan, C. Lal, and E. Losiouk, "The road ahead for networking: a survey on icn-ip coexistence solutions," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 3, pp. 2104–2129, 2020.

[2] Y. Miao, Y. Wu, and W. Wei, "Co-clustering of multi-entities sparse relational data in mi-croblogging," *Journal on Communications*, vol. 37, no. 1, pp. 151–159, 2016.

[3] N. Laoutaris, H. Che, and I. Stavrakakis, "The LCD interconnection of LRU caches and its analysis," *Performance Evaluation*, vol. 63, no. 7, pp. 609–634, 2006.

[4] L. Ramaswamy and L. Liu, "An expiration age-based document placement scheme for cooperative web caching," *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 5, pp. 585–600, 2004.

[5] G. Zhang, B. Tang, X. Wang, and Y. Wu, "An optimal cache placement strategy based on content popularity in content centric network," *Journal of Information & Computational Science*, vol. 11, no. 8, pp. 2759–2769, 2014.

[6] M. Bilal and S.-G. Kang, "Time Aware Least Recent Used (TLRU) cache management policy in ICN," in *16th International Conference on Advanced Communication Technology*, pp. 528–532, Pyeongchang, Korea (South), 2014.

[7] M. D. Ong, M. Chen, T. Taleb, X. Wang, and V. C. M. Leung, "FGPC: fine-grained popularity-based caching design for content centric networking," in *Proceedings of the 17th ACM international conference on Modeling, analysis and simulation of wireless and mobile systems - MSWiM '14*, pp. 295–302, Montreal, Canada, 2014.

[8] C. Pu, "Pro^{NDN}: MCDM-based interest forwarding and cooperative data caching for named data networking," *Journal of Computer Networks and Communications*, vol. 2021, 2021.

[9] C. Zhang and H. Wang, "Cooperative caching method based on neighbor node content and space," *International Core Journal of Engineering*, vol. 7, no. 6, pp. 88–96, 2021.

[10] Q. Zheng, J. Zhang, R. Wu, H. He, X. Tan, and L. Yuan, "An ICN cache pricing mechanism based on non-cooperative game model of users and advertisers," in *2020 3rd International Conference on Hot Information-Centric Networking (HotICN)*, pp. 77–83, Hefei, China, 2020.

[11] H. Liu and R. Han, "A hierarchical cache size allocation scheme based on content dis-semination in information-centric networks," *Future Internet*, vol. 13, no. 5, p. 131, 2021.

[12] V. S. Shekhawat, A. Vineet, and A. Gautam, "Efficient content caching for named data network nodes," in *Proceedings of the 16th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*, pp. 11–19, Houston, United States, 2019.

[13] K. H. Chiu, J. M. Wang, A. M. Abdelmoniem, and B. Bensaou, "A two-tiered caching scheme for information-centric networks," in *2021 IEEE 22nd International Conference on High Performance Switching and Routing (HPSR)*, pp. 1–6, Paris, France, 2021.

[14] M. Alhowaidi, D. Nadig, B. Hu, B. Ramamurthy, and B. Bockelman, "Cache management for large data transfers and multipath forwarding strategies in named data networking," *Computer Networks*, vol. 199, article 108437, 2021.

[15] B. Nour, H. Khelifi, H. Moungla, R. Hussain, and N. Guizani, "A distributed cache placement scheme for large-scale information-centric networking," *IEEE Network*, vol. 34, no. 6, pp. 126–132, 2020.

[16] L. Saino, I. Psaras, and G. Pavlou, "Icarus: a caching simulator for information centric networking (icn)," in *Proceedings of the Seventh International Conference on Simulation Tools and Techniques*, pp. 66–75, Lisbon, Portugal, 2014.

Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 1154325, 17 pages
https://doi.org/10.1155/2022/1154325

Research Article

A Lazy Learning-Based Self-Interference Cancellation Approach for In-Band Full-Duplex Wireless Communication Systems

Ou Zhao, Wei-Shun Liao, Keren Li, Takeshi Matsumura, Fumihide Kojima, and Hiroshi Harada

National Institute of Information and Communications Technology (NICT), 3-4 Hikarino-oka, Yokosuka 239–0847, Japan

Correspondence should be addressed to Ou Zhao; zhaoou@nict.go.jp

Received 25 March 2022; Revised 31 May 2022; Accepted 20 July 2022; Published 3 August 2022

Academic Editor: Lisheng Fan

We propose a new lazy learning-based cancellation approach to improve spectral efficiency for current wireless communication systems, suppress self-interference (SI) sent from base stations, and enable in-band full-duplex (IBFD) transmissions in cellular networks. Our proposed approach consists of two phases based on traditional IBFD systems: an offline phase for database generation and an online phase for data transmission. In the offline phase, the output before a 0/1 decision is premeasured without the desired signal input and recorded in a database with self-defined feature vectors (FVs). In the online phase, a suitable result is sought from the generated database with the help of a learning method and FV for the same system architecture with the desired signal input. The result is then assigned an SI cancellation value. Regular and eager learning-based cancellation approaches are employed to evaluate the proposed method and simulate the transmission output. Computer simulation results indicated that the proposed cancellation methods could achieve about 134 dB SI suppression and achieve nearly the same transmission levels as methods with no SI effect, enabling the IBFD operations in wireless communication systems better than the regular and eager learning-based techniques.

1. Introduction

Wireless communication traffic has been rapidly increasing with the prevalence of smartphone applications and Internet of Things devices. Development of new spectrum resources and improvement of spectral efficiency (SE) to provide proliferating wireless traffic for future mobile communication systems beyond the 5th generation require time, effort, and money [1, 2].

This study concentrates on a potentially disruptive technology called in-band full-duplex (IBFD) to increase the SE in wireless communication systems. IBFD systems can double the SE under ideal conditions compared to traditional time-division duplex (TDD) and frequency-division duplex (FDD) systems. Independent transmitting and receiving are performed in TDD systems over a common frequency band (FB), and inversely simultaneous transmitting and receiving are carried out in FDD systems using an independent FB. Consequently, the protocol for simultaneous transmission and reception over the same FB is adopted. However, the quality of the desired signal seriously deteriorates because of the self-interference (SI) effect caused by employing the same FB for transmitting and receiving. Thus, the SE either improves slightly or becomes worse [3].

Currently, various approaches have been proposed in the existing works to suppress SI and enhance SE, such as antenna, analog, and digital cancellation [4–11]. Generally, antenna cancellation is aimed at increasing the isolation between transmission and reception [5]. Analog cancellation is used to suppress the SI power by combining a reference SI signal in which the phase and amplitude are adjusted [6, 7, 11]. However, because of several physical constraints upon antenna design and inaccuracies in obtaining SI signals in analog circuits, residual SI remains powerful in the desired

signal [12]. Therefore, digital cancellation (DC) is introduced to construct a replica of the SI and subtract it from the received composite signals [9].

The desired portion of the received composite signal should be successfully demodulated if its power is tens of decibels larger than the SI power through a SI canceller. Unfortunately, in most common wireless communication systems, the power of the desired signal component is not much greater than system noise power, and the noise power is difficult to suppress. Therefore, the biggest challenge for the design of the SI canceller is to decrease the SI power down to the system noise level and double the SE. For instance, suppose that the considered IBFD operation is used in a common base station (BS) with an equivalent isotopically radiated power (EIRP) of more than 40 dBm and an average noise power of less than -100 dBm at the receiver component. The combined effect of the antenna, analog, and DCs is preferably encouraged to 140 dB or more to suppress the SI power down to the noise level [12]. In fact, a cross-polarization technique yielding antenna cancellation of 50 dB was presented in [13], and a combined analog and DC of 60 dB was proposed in [14]. These exciting results showed about 110 dB combined suppression effect for SI power can be achieved in experimental environments. Unfortunately, probably because of several hardware restrictions such as maximum output power of device and noise floor, currently, there have been few results related to SI cancellation approaches with suppression capabilities over 110 dB to the best of our knowledge.

Furthermore, most previous works such as [6, 7, 10, 15, 16] using analog and digital cancellers for IBFD systems are based on regular methods. These methods estimate the SI signal effects experienced in the time and frequency domains, including nonlinearity in the transmitter power amplifier (PA) and low noise amplifier (LNA), wireless channel propagation, and circuit delay. However, since the power gap between the SI and received signals is considerable, the nonlinearity of the PA and LNA, connector return and insertion losses, and other unknown losses are always present in practice. Therefore, a small error in the estimation process may cause large residual SI, and strict measurements are required in the whole system, even though that is challenging to ensure. Consequently, a tiny fracture grows into a chasm.

Several researchers such as in [7, 17–22] inspired by big data and machine learning (i.e., two of the most popular and useful technologies [23–25]) tried to cancel the effects of SI by (including but not limited to) predicting SI signals or estimating channel state information (CSI) to reconstruct SI waveforms with the help of the deep learning method, which is representative of eager learning (EL) in machine learning techniques. For example, the work of [18] proposed a real-time nonlinear SI cancellation solution using deep learning to realize IBFD wireless communication. In this solution, SI channel is modeled by a deep neural network (NN), and the NN is trained for cancellation of SI at wireless node. The results from their software-defined radio- (SDR-) based testbed showed a performance of 17 dB in DC and yielded an average of 8.5% bit error rate (BER) over many scenarios and different modulation schemes. Similarly, to enable IBFD transmissions, the authors in [19] proposed a nonlinear DC approach by adapting support vector regression which is one of algorithms to solve regression problems in machine learning. Their tests were also performed on SDR-based platform and indicated that their proposal can provide more than 30 dB digital suppression for transmit power levels higher than 20 dBm.

Recently, a joint detection and nonlinear SI cancellation approach using EL with NN was proposed in [20]. In this work, the EL method is used to derive a function between output of desired binary data and received signal, and thus, the desired signal can be directly demodulated in the presence of SI. Although several questions need to be addressed in their future works, the preliminary experimental results showed that EL techniques can perform better comparing to conventional SI cancellation techniques. Moreover, hybrid beamforming design for IBFD millimeter wave systems using EL-based scheme was proposed in [21]. In this work, two frameworks based on extreme learning machine and convolutional NN were presented to design hybrid beamformers and further achieve SI cancellation. Their results showed that both learning-based schemes can provide more robust performance, improve spectral efficiency, and decrease computation time. In additional, an alternative application using EL for IBFD systems was proposed in [22]. The authors in this work introduced a use of EL with NN to accelerate tuning of multitap adaptive radio frequency (RF) cancellers by training the weights of in-phase and quadrature channels in adaptive cancellers. The results illustrated a fast convergence speed by using EL in RF cancellers and thereby enabled IBFD operation in dynamic interference environments.

Generally, in many cases, an eager learner abstracts away from the data during training and uses the trained model (abstraction) to make predictions. The most important benefit of using EL is solving the SI problem; time-insensitive parameters (e.g., PA and LNA nonlinearity as well as antenna return loss) in the entire system can be included in the trained model and do not have to be estimated. However, prediction errors caused by the trained model and estimated CSI for the SI calculation always exist and cannot be avoided because the model training is an essential operation [26]. Thus, errors can unexpectedly be amplified by powerful SI signals similar to the regular method described above.

Inspired by the previous discussion, it can be known that estimating or reconstructing SI signal waveforms for cancellation is unnecessary. However, all practical effects on the desired signal caused by the SI must be quantified and tagged with feature vectors (FVs), and these quantified values can be precisely found by using these FVs when required [27]. That highlights the difference with the model generation-based learning methods such as [17–22]. More specifically, the effects of the SI on wireless transmissions can be quantified by constellation values at the 0/1 decision in a signal demodulator; an FV can be defined as a set constructed with the estimated CSI between the interested user, antennas, and transmitted symbol from the BS. Contrary to EL, the lazy learning (LL) method [27, 28] does not require

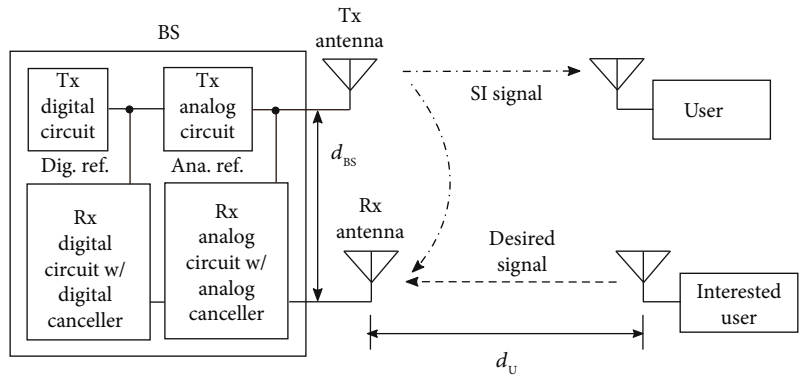


Figure 1: Unidirectional IBFD system illustrations.

model training, thereby the prediction error is not introduced in related operations for SI cancellation and may not be amplified by the powerful SI signals. Therefore, we propose an LL-based cancelation method to solve the SI problem.

We separately perform an offline phase for database generation and an online phase for data transmission. In the offline phase, the output before a 0/1 decision is premeasured without the desired signal input and recorded to a database with self-defined FVs. In the online phase, a suitable result is sought from the generated database with the help of a learning method and FV for the same system architecture with the desired signal input. The result is then assigned as the SI cancellation value. In other studies such as [22], the demodulation of composite signal, including desired signal, thermal noise, and potential residual SI, is independently performed after all of the cancellation processes are done (even the residual SI is still powerful). In our proposal, desired signals are directly demodulated in the presence of SI and thermal noise. We further provide system-level performance such as BER to evaluate the proposed approach in the IBFD systems, contrary to existing works [4–10, 12–16] where most studies only focused on the SI suppression capability.

The remainder of this paper is organized as follows. In Section 2, we describe the system architecture under consideration and formulate the problem. In Section 3, we explain three kinds of DC approaches in greater detail: the proposed technique, regular, and an EL-based method. Details regarding the employed channel models are provided in Section 4. In Section 5, we present and analyze simulated results and summarize the key findings. Discussion and concluding remarks are presented in Section 6 and Section 7, respectively.

2. System Model and Problem Formulation

2.1. System Model. We consider a unidirectional IBFD system in which an interested user equipped with a single antenna sends desired signals to a BS receiving antenna from a distance of d_U, while the BS simultaneously sends signals to other users via the same FB. The BS is equipped with a pair of transmitting and receiving antennas with an interantenna distance of d_{BS}, and the latter one is used to receive a composite signal sent by the BS transmitting antenna (i.e., SI) and the interested user. Analog and digital cancellers are commonly employed in the IBFD BS structure for suppression of SI signals, and references from analog and digital transmitter circuits can provide supports for cancellation processes. An illustration of the considered unidirectional IBFD systems is shown in Figure 1.

In this study, we design a valid frame structure for transmitted signals, including the SI and desired signal, to evaluate the considered systems using the proposed cancellation approach. The structural design includes downlink and uplink aspects used for BS and an interested user, respectively. The downlink and uplink transmissions with N_{frame} frames indexed with i are performed, and there are N_{symbol} modulated symbols indexed with j in the data part in each frame. We assume that frame synchronization at receive side is perfect and pilot signals with N_{pilot} symbols are deployed at the head of the frame to obtain CSI at an arbitrary receiving terminal. An illustration of the frame structure is given in Figure 2.

2.2. Problem Formulation. In this study, the considered IBFD system is employed with the uses of antenna, analog, and DCs. An IBFD BS structure is shown in Figure 3. For jth data symbol in ith frame, the received waveform at BS which consists of desired signal, SI signal, and noise can be expressed as

$$\boldsymbol{y}(i,j) = G_{\text{ant,rx}}^{1/2}\boldsymbol{h}_{\text{U}}(i)\boldsymbol{x}_{\text{U}}(i,j) + \widehat{C}_{\text{ant}}^{1/2}G_{\text{ant,rx}}^{1/2}\boldsymbol{h}_{\text{BS}}(i)\boldsymbol{x}_{\text{BS}}(i,j) + \boldsymbol{w}_{\text{rx}}(i,j), \tag{1}$$

where $G_{\text{ant,rx}}$ is antenna gain of BS receive antenna. $\boldsymbol{h}_{\Psi}(i)$ for $\Psi = \text{U}$ and $\Psi = \text{BS}$ are defined as the channel gain coefficient of user-BS and the coefficient of BS's interantenna corresponding to the ith frame, respectively. We further use $\rho_{\Psi}(i)$ and $\phi_{\Psi}(i)$ to represent the amplitude coefficient and phase shift of the complex variable $\boldsymbol{h}_{\Psi}(i)$ so that

$$\boldsymbol{h}_{\Psi}(i) = \rho_{\Psi}(i)e^{j\phi_{\Psi}(i)}. \tag{2}$$

The length of frames is limited by the channel coherence time, and wireless channels are considered static

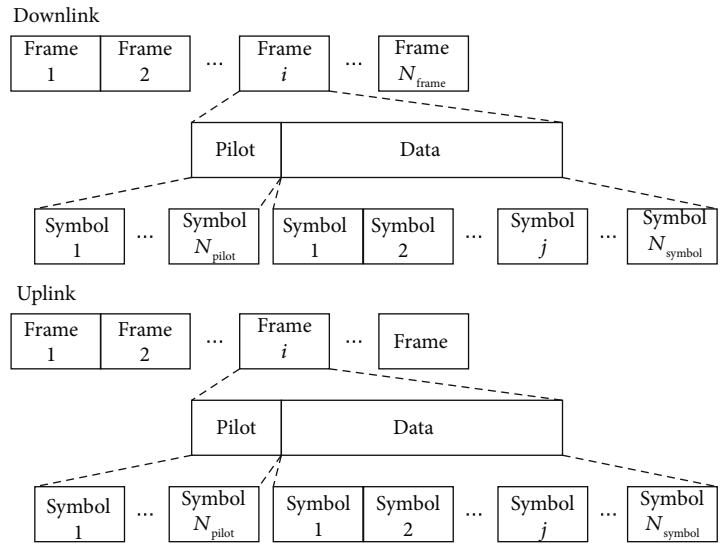


Figure 2: Frame structure illustration of IBFD systems.

during this coherence time [29]. Thus, effects due to the channel attenuation on all of the symbols in the same frame are identical and can be estimated using pilot signals. $\boldsymbol{w}_{\mathrm{rx}}(i,j)$ is the complex additive white Gaussian noise (AWGN) at the receive antenna of BS with an average power of Ω_{rx}. $\widehat{C}_{\mathrm{ant}}$ represents the antenna cancellation without path loss effect of interantenna at BS and further expressed as

$$\widehat{C}_{\mathrm{ant}} = C_{\mathrm{ant}} L_{\mathrm{BS}}^{-1}, \tag{3}$$

where L_{BS} represents the path loss between the antennas at BS and C_{ant} denotes the employed antenna cancellation, which can suppress the power of the SI signals.

$\boldsymbol{x}_{\Psi}(i,j)$for $\Psi = \mathrm{U}$ or $\Psi = \mathrm{BS}$ is complex signal output from the transmit antenna of user or BS and can be expressed as

$$\boldsymbol{x}_{\Psi}(i,j) = G_{\mathrm{ant,tx},\Psi}^{1/2} Q_{\mathrm{PA},\Psi}(\boldsymbol{m}_{\Psi}(i,j) + \boldsymbol{w}_{\mathrm{tx},\Psi}(i,j)) + G_{\mathrm{ant,tx},\Psi}^{1/2} \boldsymbol{w}_{\mathrm{PA},\Psi}(i,j), \tag{4}$$

where $G_{\mathrm{ant,tx},\Psi}$ denotes antenna gain for user or BS. $Q_{\mathrm{PA},\Psi}(z)$ is PA function for user or BS with third-order intermodulation distortion. According to [30], $Q_{\mathrm{PA},\Psi}(\boldsymbol{z})$ can be expressed as

$$Q_{\mathrm{PA},\Psi}(\boldsymbol{z}) = G_{\mathrm{PA},\Psi}^{1/2} \boldsymbol{z} - \frac{4 G_{\mathrm{PA},\Psi}^{1/2}}{3 O_{3,\Psi}} \boldsymbol{z}^3, \tag{5}$$

where $G_{\mathrm{PA},\Psi}$ denotes PA gain of user or BS, while $O_{3,\Psi}$ is the third-order intercept point (OIP3) of user or BS PA.

The terms of $\boldsymbol{w}_{\mathrm{tx},\Psi}(i,j)$ are the complex AWGN existing in the user or BS modulator with an average power of $\Omega_{\mathrm{tx},\Psi}$, while $\boldsymbol{w}_{\mathrm{PA},\Psi}(i,j)$ is denoted as the additional output noises caused by PA of user or BS with an average power of $\Omega_{\mathrm{PA},\Psi}$ that can be easily calculated by

$$\Omega_{\mathrm{PA},\Psi} = (F_{\Psi} - 1) \Omega_{\mathrm{tx},\Psi} G_{\mathrm{PA},\Psi}^{1/2}, \tag{6}$$

where F_{Ψ} is noise factors in user or BS PA [31].

The terms of $\boldsymbol{m}_{\Psi}(i,j)$ in (4) represent the output complex signal from the user or BS modulator. For the use of a common modulation scheme, $\boldsymbol{m}_{\Psi}(i,j)$ can be written as

$$\boldsymbol{m}_{\Psi}(i,j) = A_{\Psi}(i,j) \cos\left(\theta_{\Psi}(i,j)\right) + j A_{\Psi}(i,j) \sin\left(\theta_{\Psi}(i,j)\right), \tag{7}$$

where $A_{\Psi}(i,j) = \sqrt{2 E_{\mathrm{symbol}}(i,j) f_{\mathrm{symbol}}}$ and f_{symbol} is symbol rate for the considered modulator, while $E_{\mathrm{symbol}}(i,j)$ denotes energy of jth symbol in ith frame. $\theta_{\Psi}(i,j)$ are the corresponding phase angles after symbol mapping processed on binary data which the user or BS needs to transmit.

Refer to Figure 3 with switches ON, after an analog cancellation C_{ana} for the SI signals in receive waveform $\boldsymbol{y}(i,j)$ is done, the CSI of antennas at BS and the CSI of user-BS part are estimated as

$$\tilde{\boldsymbol{h}}_{\Psi}(i) = \tilde{\rho}_{\Psi}(i) e^{j\widetilde{\phi}_{\Psi}(i)}, \tag{8}$$

for $\Psi = \mathrm{U}$ and $\Psi = \mathrm{BS}$, respectively; then, a running of search algorithm for DC is employed even though it is not necessary in common wireless communication systems without IBFD operations. Thereafter, a signal detection method $\mathscr{D}(\boldsymbol{z})$ in the detector should be adopted and performed. The output of the detector as well as the input of demodulator is written as

$$\boldsymbol{y}_{\mathrm{det}}(i,j) = \mathscr{D}\Big(G_{\mathrm{ant,rx}}^{1/2} \boldsymbol{h}_{\mathrm{U}}(i) \boldsymbol{x}_{\mathrm{U}}(i,j) + C_{\mathrm{ana}}^{1/2} \widehat{C}_{\mathrm{ant}}^{1/2} G_{\mathrm{ant,rx}}^{1/2} \boldsymbol{h}_{\mathrm{BS}}(i) \boldsymbol{x}_{\mathrm{BS}}(i,j) + \boldsymbol{w}_{\mathrm{rx}}(i,j) \Big). \tag{9}$$

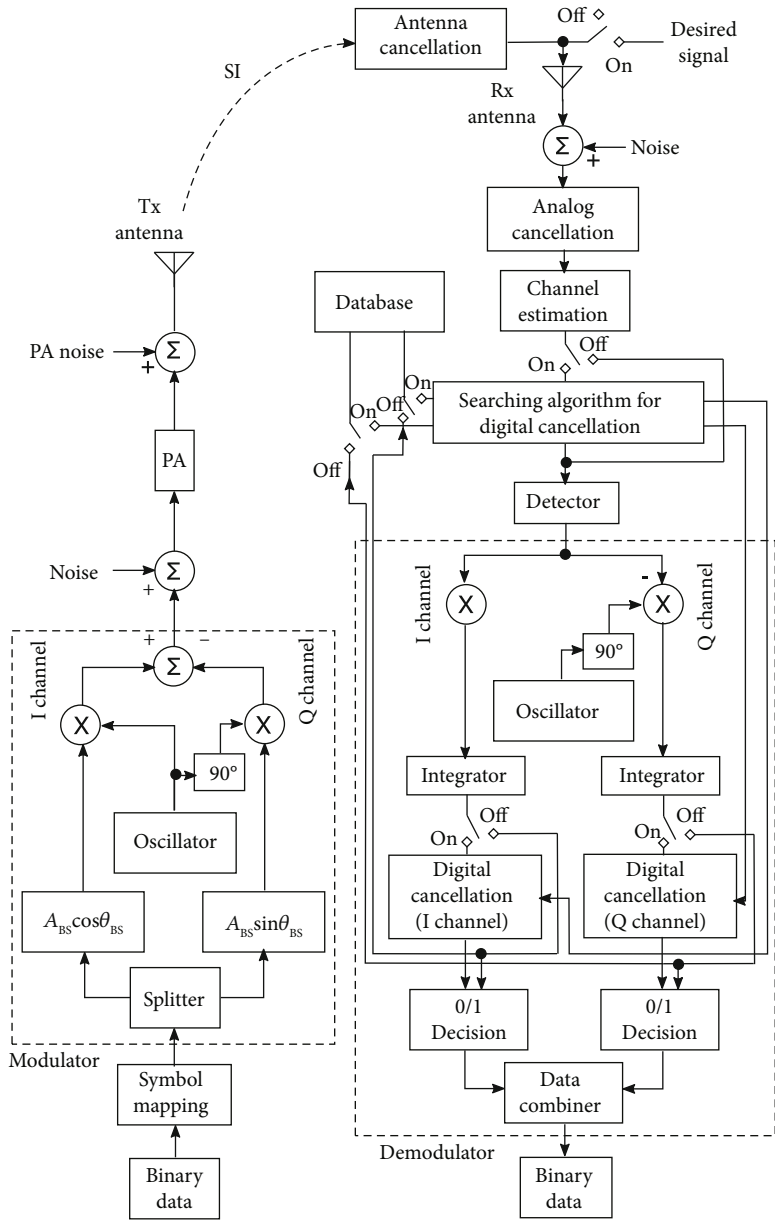


Figure 3: An IBFD BS structure with switches OFF for the offline phase and ON for the online phase. The offline phase is used for an offline-generated database in which constellation values of SI have been previously measured and recorded. The online phase is used for data transmission in which the desired signals are separated by subtracting the recorded SI from the received composite signals.

After a series of operations in demodulation with the uses of DCs $\boldsymbol{C}_{\text{dig}}$ which formed as a complex number, the inputs of 0/1 decision can be expressed as

$$\boldsymbol{y}_{\text{dec}}(i,j) = \boldsymbol{y}_{\text{det}}(i,j) - \boldsymbol{C}_{\text{dig}}(i,j,\boldsymbol{\alpha}), \tag{10}$$

where the real part and imaginary part of $\boldsymbol{y}_{\text{dec}}(i,j)$, i.e., $\Re(\boldsymbol{y}_{\text{dec}}(i,j))$ and $\Im(\boldsymbol{y}_{\text{dec}}(i,j))$, are used for I and Q channels in 0/1 decision block, respectively. $\boldsymbol{\alpha}$ is an FV constructed by some input parameters for DC that is explained in the following section.

Substitute (9) and the related equations given in this section into (10), we rewrite (10) through the necessary mathematical simplification as

$$\boldsymbol{y}_{\text{dec}}(i,j) = \boldsymbol{S}(i,j) + \boldsymbol{I}(i,j) - \boldsymbol{C}_{\text{dig}}(i,j,\boldsymbol{\alpha}) + \boldsymbol{W}(i,j), \tag{11}$$

where $\boldsymbol{S}(i,j)$ and $\boldsymbol{I}(i,j)$ represent the complex constellation values of the desired and SI signals, respectively, and can be calculated by

$$\boldsymbol{S}(i,j) = G_{\text{ant,rx}}^{1/2} G_{\text{ant,tx,U}}^{1/2} G_{\text{PA,U}}^{1/2} \lambda^{-1} \boldsymbol{h}_{\text{U}}(i) \left(\boldsymbol{m}_{\text{U}}(i,j) - \frac{4}{3O_{3,\text{U}}} \boldsymbol{m}_{\text{U}}^{3}(i,j) \right), \tag{12}$$

$$\boldsymbol{I}(i,j) = C_{\text{ana}}^{1/2}\widehat{C}_{\text{ant}}^{1/2}G_{\text{ant,rx}}^{1/2}G_{\text{ant,tx,BS}}^{1/2}G_{\text{PA,BS}}^{1/2}\,\lambda^{-1}\boldsymbol{h}_{\text{BS}}(i)\left(\boldsymbol{m}_{\text{BS}}(i,j) - \frac{4}{3O_{3,\text{BS}}}\boldsymbol{m}_{\text{BS}}^{3}(i,j)\right). \tag{13}$$

In the current study, a simple matched filter (MF) detection method [32] for $\mathscr{D}(\boldsymbol{z})$ is used and can roughly be expressed as

$$\mathscr{D}(\boldsymbol{z}) = \boldsymbol{\lambda}^{-1}z. \tag{14}$$

For the most common wireless communication systems, $\boldsymbol{\lambda}$ is assigned by $\tilde{\boldsymbol{h}}_{\text{U}}(i)$; thus, the desired signals can be detected because of $\tilde{\boldsymbol{h}}_{\text{U}}^{-1}(i)\boldsymbol{h}_{\text{U}}(i) \approx 1$ [32, 33]. The complex value $\boldsymbol{W}(i,j)$ caused by various noises can be further extended as

$$\boldsymbol{W}(i,j) = \boldsymbol{W}_{\text{tx,U}}(i,j) + \boldsymbol{W}_{\text{PA,U}}(i,j) + \boldsymbol{W}_{\text{tx,BS}}(i,j) + \boldsymbol{W}_{\text{PA,BS}}(i,j) + \boldsymbol{W}_{\text{rx}}(i,j), \tag{15}$$

where $\boldsymbol{W}_{\text{tx,U}}(i,j)$, $\boldsymbol{W}_{\text{PA,U}}(i,j)$, $\boldsymbol{W}_{\text{tx,BS}}(i,j)$, $\boldsymbol{W}_{\text{PA,BS}}(i,j)$, and $\boldsymbol{W}_{\text{rx}}(i,j)$ are defined as the complex values of noises generated from user's modulation and PA, BS's modulation and its PA, and the receive antenna at BS side, respectively, and can be easily calculated by

$$\begin{aligned}\boldsymbol{W}_{\text{tx,U}}(i,j) = \big(&G_{\text{ant,rx}}^{1/2}G_{\text{ant,tx,U}}^{1/2}G_{\text{PA,U}}^{1/2}\boldsymbol{\lambda}^{-1}\,\boldsymbol{h}_{\text{U}}(i)\\ &\cdot\left(1 - \frac{4}{O_{3,\text{U}}}\boldsymbol{m}_{\text{U}}^{2}(i,j)\right)\boldsymbol{w}_{\text{tx,U}}(i,j)\\ &- \frac{4}{O_{3,\text{U}}}\boldsymbol{m}_{\text{U}}(i,j)\boldsymbol{w}_{\text{tx,U}}^{2}(i,j) - \frac{4}{3O_{3,\text{U}}}\boldsymbol{w}_{\text{tx,U}}^{3}(i,j),),\end{aligned} \tag{16}$$

$$\boldsymbol{W}_{\text{PA,U}}(i,j) = G_{\text{ant,rx}}^{1/2}G_{\text{ant,tx,U}}^{1/2}\boldsymbol{\lambda}^{-1}\boldsymbol{h}_{\text{U}}(i)\boldsymbol{w}_{\text{PA,U}}(i,j), \tag{17}$$

$$\begin{aligned}\boldsymbol{W}_{\text{tx,BS}}(i,j) = &C_{\text{ana}}^{1/2}\widehat{C}_{\text{ant}}^{1/2}G_{\text{ant,rx}}^{1/2}G_{\text{ant,tx,BS}}^{1/2}G_{\text{PA,BS}}^{1/2}\boldsymbol{\lambda}^{-1}\boldsymbol{h}_{\text{BS}}(i)\\ &\cdot\left(\left(1 - \frac{4}{O_{3,\text{BS}}}\boldsymbol{m}_{\text{BS}}^{2}(i,j)\right)\times\boldsymbol{w}_{\text{tx,BS}}(i,j)\right.\\ &\left.- \frac{4}{O_{3,\text{BS}}}\boldsymbol{m}_{\text{BS}}(i,j)\boldsymbol{w}_{\text{tx,BS}}^{2}(i,j) - \frac{4}{3O_{3,\text{BS}}}\boldsymbol{w}_{\text{tx,BS}}^{3}(i,j)\right),\end{aligned} \tag{18}$$

$$\boldsymbol{W}_{\text{PA,BS}}(i,j) = C_{\text{ana}}^{1/2}\widehat{C}_{\text{ant}}^{1/2}G_{\text{ant,rx}}^{1/2}G_{\text{ant,tx,BS}}^{1/2}\boldsymbol{\lambda}^{-1}\boldsymbol{h}_{\text{BS}}(i)\boldsymbol{w}_{\text{PA,BS}}(i,j), \tag{19}$$

$$\boldsymbol{W}_{\text{rx}}(i,j) = \boldsymbol{\lambda}^{-1}\boldsymbol{w}_{\text{rx}}(i,j), \tag{20}$$

respectively.

Actually, to get the desired binary data sent by the user successfully, we strongly hope that the input values of 0/1 decision, i.e., $\boldsymbol{y}_{\text{dec}}(i,j)$, are purely decided by the desired part $\boldsymbol{S}(i,j)$. Unfortunately, in IBFD systems, because the strength of SI signal at the BS's receive antenna is very huge compared to the desired signal, $\boldsymbol{y}_{\text{dec}}(i,j)$ presented in (11) is mainly dominated by $\boldsymbol{I}(i,j)$. That is the reason that we further propose the DC $\boldsymbol{C}_{\text{dig}}$ to suppress the influence of SI signals. Define the residual SI as

$$\widehat{\boldsymbol{I}}(i,j) = \boldsymbol{I}(i,j) - \boldsymbol{C}_{\text{dig}}(i,j,\boldsymbol{\alpha}), \tag{21}$$

naturally our target is to make $\widehat{\boldsymbol{I}}(i,j)$ close to zero for arbitrary frame i and symbol j as much as possible.

3. Digital Cancellation Approaches

In this study, we propose an LL-based method to set DC and further enable the IBFD operations in the considered wireless communication systems. Several existing works such as [7, 16] have been studied on SI suppression with the help of deep learning, which is a typical representative of EL in machine learning techniques. Other works such as [6] employed a regular method in which the SI signal is approximately calculated using estimated CSI. In this section, we give more explanations for the DCs mentioned above.

3.1. Proposed LL-Based Cancellation Approach. Generally, LL is a learning method where generalization of the trained model is not necessary, and searching of the needed data is delayed until a query is made to the system, as opposed to in EL, where the system tries to generalize the model using training data before receiving queries [28]. The core of the proposed LL-based cancellation approach in our study is constructed by (1) an offline-generated database in which the constellation values of SI at the input of 0/1 decision block are premeasured and recorded and (2) an online data transmission in which desired signals are separated by subtracting the recorded SI from the received composite signals. The details are provided as follows.

3.1.1. Offline Database Generation. We first define a database $\mathscr{B}$ which consists of an input space named $\mathscr{B}_{\text{in}}$ and an output space named $\mathscr{B}_{\text{out}}$. According to (13), because SI is further affected by $\boldsymbol{\lambda}$ and $\boldsymbol{\lambda}$ is related to the coefficient of user channel for the considered MF detection method; in the input space $\mathscr{B}_{\text{in}}$, we define three independent parameters $\bar{A}_{\text{BS}}$, $\overline{\theta}_{\text{BS}}$, and $\boldsymbol{h}_{\text{U}}$ and form an FV. The first and second parameters $\bar{A}_{\text{BS}}$ and $\overline{\theta}_{\text{BS}}$ are assigned with amplitude and phase angle used in BS's modulation and can yield discrete values corresponding to the used modulation method. The last parameters $\bar{\boldsymbol{h}}_{\text{U}}$ are assigned with the channel coefficient $\boldsymbol{h}_{\text{U}}$ between the interested user and BS, and in theory, its absolute value and phase can range from 0 to ∞ and 0 to 2π, respectively. The values in the input space $\mathscr{B}_{\text{in}}$ are previously designed to take over all possible combinations of these parameters as much as possible under an acceptable level of computational cost.

To reduce the database size for our learning systems, in the output space $\mathscr{B}_{\text{out}}$, we define two parameters which are $\bar{\boldsymbol{h}}_{\text{BS}}$ and $\bar{\boldsymbol{I}}$. $\bar{\boldsymbol{h}}_{\text{BS}}$ is used to record the estimated values of channel coefficient $\boldsymbol{h}_{\text{BS}}$ and can be considered as a tag of $\bar{\boldsymbol{I}}$ when we need to locate $\bar{\boldsymbol{I}}$ from the database $\mathscr{B}_{\text{out}}$. $\bar{\boldsymbol{I}}$ is used to record the measured constellation values of SI. To get the values for the output space $\mathscr{B}_{\text{out}}$, a series of provisional transmissions using the designed input space $\mathscr{B}_{\text{in}}$ needs to be performed, where the provisional transmission corresponding to Figure 3 with switches OFF means

Input: Modulation, Possible channel gain for $\boldsymbol{h}_{\mathrm{U}}$;
Output: $\mathscr{B}_{\mathrm{in}}$, $\mathscr{B}_{\mathrm{out}}$;
Initialization: $\mathscr{B}_{\mathrm{in}} = \varnothing$, $\mathscr{B}_{\mathrm{out}} = \varnothing$;
%Generation of input space for database $\mathscr{B}$;
foreach $A_{\mathrm{BS}} \in$ adopted modulation **do**
  foreach $\theta_{\mathrm{BS}} \in$ adopted modulation **do**
    foreach designed $\boldsymbol{h}_{\mathrm{U}}$ **do**
      $\bar{A}_{\mathrm{BS}}(b) = A_{\mathrm{BS}}$, $\overline{\theta}_{\mathrm{BS}}(b) = \theta_{\mathrm{BS}}$, $\bar{\boldsymbol{h}}_{\mathrm{U}}(b) = \boldsymbol{h}_{\mathrm{U}}$;
      $\mathscr{B}_{\mathrm{in}}(b) = [\bar{A}_{\mathrm{BS}}(b)\,; \overline{\theta}_{\mathrm{BS}}(b)\,; \bar{\boldsymbol{h}}_{\mathrm{U}}(b)]$;
    end
  end
end
foreach b **do**
  $\mathscr{B}_{\mathrm{in}} = \mathscr{B}_{\mathrm{in}} \cup \mathscr{B}_{\mathrm{in}}(b)$;
end
%Generation of output space for database $\mathscr{B}$;
foreach $\mathscr{B}_{\mathrm{in}}(b)$ **do**
  foreach b' **do**
    Measure $\boldsymbol{h}_{\mathrm{BS}}$ as $\tilde{\boldsymbol{h}}_{\mathrm{BS}}$
    Record the results to $\tilde{\boldsymbol{h}}_{\mathrm{BS}}(b, b')$;
    Measure SI using (23) w/o desired signal input;
    Record the results to $\bar{\boldsymbol{I}}(b, b')$;
    $\mathscr{B}_{\mathrm{out}}(b, b') = [\tilde{\boldsymbol{h}}_{\mathrm{BS}}(b, b')\,; \bar{\boldsymbol{I}}(b, b')]$;
  end
end
foreach b and b' **do**
  $\mathscr{B}_{\mathrm{out}} = \mathscr{B}_{\mathrm{out}} \cup \mathscr{B}_{\mathrm{out}}(b, b')$
end

Algorithm 1: Offline database generation.

transmissions without desired signal inputs, i.e., $\boldsymbol{m}_{\mathrm{U}}(i, j) = 0 \forall i, j$, and DC is not performed at BS. Note that the outputs of channel estimation block in Figure 3 are the estimated CSI of user to BS and the estimated CSI of interantenna of BS, and the value of the former one is replaced by the designed $\bar{\boldsymbol{h}}_{\mathrm{U}}$ in the offline database generation.

Intuitively, we first chose a designed FV of input space as

$$\mathscr{B}_{\mathrm{in}}(b) = \left[\bar{A}_{\mathrm{BS}}(b)\,; \overline{\theta}_{\mathrm{BS}}(b)\,; \bar{\boldsymbol{h}}_{\mathrm{U}}(b)\right], \tag{22}$$

where $b = 1, \cdots, |\mathscr{B}_{\mathrm{in}}|$, and then use $\mathscr{B}_{\mathrm{in}}(b)$ to perform the provisional transmission following the considered IBFD system. For this case of b, one of the possible channel gain $\bar{\boldsymbol{h}}_{\mathrm{BS}}$ measured by the channel estimation block is recorded into $\bar{\boldsymbol{h}}_{\mathrm{BS}}(b, b')$, and at the same time, $\bar{\boldsymbol{I}}(b, b')$ is used to recorded the input of 0/1 decision, i.e., the constellation value of SI. According to (13) and (15), after averaging the noise effects by pilot signals, in mathematics, the recorded value of $\bar{\boldsymbol{I}}(b, b')$ by measurement can be written as

$$\bar{\boldsymbol{I}}\left(b, b'\right) = \bar{\boldsymbol{I}}\left(b, b'\right) \left| \boldsymbol{\lambda} = \bar{\boldsymbol{h}}_{\mathrm{U}}(b), A_{\mathrm{BS}}(i, j) = \bar{A}_{\mathrm{BS}}(b), \theta_{\mathrm{BS}}(i, j) = \overline{\theta}_{\mathrm{BS}}(b). \right. \tag{23}$$

A summary for this generation process can be found in Algorithm 1.

3.1.2. Online Data Transmission. Once the database $\mathscr{B}$ is created, we can use it into online transmission for the considered IBFD systems as shown in Figure 3 with switches ON. When the BS prepares to demodulate the desired signals $\boldsymbol{S}(i, j)$ transmitted by the user under the effect of SI $\boldsymbol{I}(i, j)$ and noise $\boldsymbol{W}(i, j)$; in the block of searching algorithm for DC, the BS first searches an "optimal" FV from input space $\mathscr{B}_{\mathrm{in}}$ using estimated CSI of $\tilde{\boldsymbol{h}}_{\mathrm{U}}(i)$ and the related modulation information $A_{\mathrm{BS}}(i, j)$ and $\theta_{\mathrm{BS}}(i, j)$ of the signals the BS sent.

The "optimal" FV is decided by the vector $[A_{\mathrm{BS}}(i, j)\,; \theta_{\mathrm{BS}}(i, j)\,; \tilde{\boldsymbol{h}}_{\mathrm{U}}(i)]$ and written as

$$\mathscr{B}_{\mathrm{in}}(b^*) = \left[\bar{A}_{\mathrm{BS}}(b^*)\,; \overline{\theta}_{\mathrm{BS}}(b^*)\,; \tilde{\boldsymbol{h}}_{\mathrm{U}}(b^*)\right], \tag{24}$$

where

$$\begin{aligned} b^* = &\left\{\arg_{b \in \{1, \cdots, |\mathscr{B}_{\mathrm{in}}|\}} \bar{A}_{\mathrm{BS}}(b) = A_{\mathrm{BS}}(i, j)\right\} \cap \\ &\cdot \left\{\arg_{b \in \{1, \cdots, |\mathscr{B}_{\mathrm{in}}|\}} \overline{\theta}_{\mathrm{BS}}(b) = \theta_{\mathrm{BS}}(i, j)\right\} \cap \\ &\cdot \left\{\arg \min_{b \in \{1, \cdots, |\mathscr{B}_{\mathrm{in}}|\}} \left\| \bar{\boldsymbol{h}}_{\mathrm{U}}(b) - \bar{\boldsymbol{h}}_{\mathrm{U}}(i) \right\|\right\}. \end{aligned} \tag{25}$$

Based on the fact that $A_{\mathrm{BS}}(i, j)$ and $\theta_{\mathrm{BS}}(i, j)$ for all of i

and j are known at BS and have been decided by modulation method, the result of (25) is dominated by the difference between $\bar{\boldsymbol{h}}_{\mathrm{U}}(b)$ and $\bar{\boldsymbol{h}}_{\mathrm{U}}(i)$. In fact, the last part in (25) is the well-known optimization problem of nearest neighbor search (NNS) in LL. The simplest solution to the NNS problem is the so-called linear search which computes the Euclidean distance taking over all of candidates. Other methods such as space partitioning like k-dimensional tree [34] or greedy search [35] can give an exact or approximate solution at a lower computational cost. Some analysis and performance study for NNS problems can be found in [36].

After the "optimal" FV $\mathscr{B}_{\mathrm{in}}(b^*)$ is found, the corresponding set of the potential SI effects in the output space $\mathscr{B}_{\mathrm{out}}$ is then located as $\mathscr{B}_{\mathrm{out}}(b^*, b') = [\bar{\boldsymbol{h}}_{\mathrm{BS}}(b^*, b'); \bar{\boldsymbol{I}}(b^*, b')]$. To assign suitable values to the DC $\boldsymbol{C}_{\mathrm{dig}}(i, j, \boldsymbol{\alpha})$, we also consider the estimated CSI of $\tilde{\boldsymbol{h}}_{\mathrm{BS}}(i)$. The value of $\bar{\boldsymbol{I}}(b^*, b'^*)$ is finally assigned to $\boldsymbol{C}_{\mathrm{dig}}(i, j, \boldsymbol{\alpha})$ where b'^* is given by

$$b'^* = \arg \min_{b' \in \{1, \cdots, |\mathscr{B}_{\mathrm{out}}(b^*)|\}} \left\| \bar{\boldsymbol{h}}_{\mathrm{BS}}\left(b^*, b'\right) - \tilde{\boldsymbol{h}}_{\mathrm{BS}}(i) \right\|, \tag{26}$$

where $|\mathscr{B}_{\mathrm{out}}(b^*)|$ suggests the number of measured CSI between interantenna of BS. Notably, because effects from noise in $\bar{\boldsymbol{h}}_{\mathrm{BS}}(b^*, b')$ and $\tilde{\boldsymbol{h}}_{\mathrm{BS}}(i)$ have been averaged by using multiple pilot signals, the result of (26) ensured that $\bar{\boldsymbol{h}}_{\mathrm{BS}}(b^*, b') \approx \tilde{\boldsymbol{h}}_{\mathrm{BS}}(i)$. In addition, our proposed approaches work with an assumption that a channel realization between the BS antennas $\boldsymbol{h}_{\mathrm{BS}}(i)$ can only be tagged by one estimated CSI $\tilde{\boldsymbol{h}}_{\mathrm{BS}}(i)$; in other words, there has to be a one-to-one correspondence between $\boldsymbol{h}_{\mathrm{BS}}(i)$ and $\tilde{\boldsymbol{h}}_{\mathrm{BS}}(i)$. Generally, this one-to-one correspondence can be guaranteed. Thus, based on the above explanation, the parameter of $\boldsymbol{\alpha}$ in the proposed DC $\boldsymbol{C}_{\mathrm{dig}}(i, j, \boldsymbol{\alpha})$ is written as

$$\boldsymbol{\alpha} = \left[A_{\mathrm{BS}}(i, j); \theta_{\mathrm{BS}}(i, j); \tilde{\boldsymbol{h}}_{\mathrm{U}}(i); \tilde{\boldsymbol{h}}_{\mathrm{BS}}(i) \right]. \tag{27}$$

Notably, according to (23), because the interference term $\bar{\boldsymbol{I}}(b, b')$ in the output space of database is premeasured using $\boldsymbol{\lambda} = \bar{\boldsymbol{h}}_{\mathrm{U}}(b)$ in offline database generation, to suppress the SI effects, in online transmission, the parameter $\boldsymbol{\lambda}$ in the detector process should be set to $\bar{\boldsymbol{h}}_{\mathrm{U}}(b^*)$ rather than the common $\tilde{\boldsymbol{h}}_{\mathrm{U}}(i)$ in the proposed DC. So, corresponding to the parameters of $A_{\mathrm{BS}}(i, j)$, $\theta_{\mathrm{BS}}(i, j)$, $\boldsymbol{h}_{\mathrm{U}}(i)$, and $\boldsymbol{h}_{\mathrm{BS}}(i)$, the constellation value of SI, excluding the noise part after using the proposed method in online transmissions, is written as

$$\boldsymbol{I}(i, j) = \boldsymbol{I}(i, j) | \boldsymbol{\lambda} = \bar{\boldsymbol{h}}_{\mathrm{U}}(b^*), \tag{28}$$

tagged by $\tilde{\boldsymbol{h}}_{\mathrm{BS}}(i)$, whereas the DC value $\boldsymbol{C}_{\mathrm{dig}}(i, j, \boldsymbol{\alpha})$ for this case is decided by database $\mathscr{B}_{\mathrm{out}}$ and is written as

$$\begin{aligned} \boldsymbol{C}_{\mathrm{dig}}(i, j, \boldsymbol{\alpha}) &= \bar{\boldsymbol{I}}\left(b^*, b'^*\right) = \boldsymbol{I}(i, j) | \lambda = \bar{\boldsymbol{h}}_{\mathrm{U}}(b^*), A_{\mathrm{BS}}(i, j) \\ &= \bar{A}_{\mathrm{BS}}(b^*), \theta_{\mathrm{BS}}(i, j) = \bar{\theta}_{\mathrm{BS}}(b^*), \end{aligned} \tag{29}$$

1 **Input**: $\mathscr{B}_{\mathrm{in}}$, $\mathscr{B}_{\mathrm{out}}$, $A_{\mathrm{BS}}(i, j)$, $\theta_{\mathrm{BS}}(i, j)$, $\tilde{\boldsymbol{h}}_{\mathrm{U}}(i)$, $\tilde{\boldsymbol{h}}_{\mathrm{BS}}(i)$;
2 **Output**: $\mathbf{C}_{\mathrm{dig}}(i, j, \boldsymbol{\alpha})$, $\boldsymbol{\lambda}$;
3 **Initialization**: $\boldsymbol{\alpha} = [A_{\mathrm{BS}}(i, j); \theta_{\mathrm{BS}}(i, j); \tilde{\boldsymbol{h}}_{\mathrm{U}}(i); \tilde{\boldsymbol{h}}_{\mathrm{BS}}(i)]$;
4 **Calculate** b^* using (25);
5 **Calculate** b'^* using (26);
6 $\mathbf{C}_{\mathrm{dig}}(i, j, \boldsymbol{\alpha}) = \bar{\boldsymbol{I}}(b^*, b'^*)$; %For DC blocks;
7 $\boldsymbol{\lambda} = \bar{\boldsymbol{h}}_{\mathrm{U}}(b^*)$; %For detection block;

Algorithm 2: Value assignment for the proposed DC.

tagged by $\bar{\boldsymbol{h}}_{\mathrm{BS}}(b^*, b'^*)$. Because (26) ensured that tags $\bar{\boldsymbol{h}}_{\mathrm{BS}}(b^*, b'^*) \approx \tilde{\boldsymbol{h}}_{\mathrm{BS}}(i)$ and $A_{\mathrm{BS}}(i, j)$, $\theta_{\mathrm{BS}}(i, j)$ are known at BS; further, the one-to-one correspondence between $\tilde{\boldsymbol{h}}_{\mathrm{BS}}(i)$ and $\boldsymbol{h}_{\mathrm{BS}}(i)$ can be generally guaranteed, and the results of (28) and (29) are almost equal. Consequently, the residual SI of nonnoise part $\widehat{\boldsymbol{I}}(i, j)$ can be substantially suppressed. A summary of this process can be found in Algorithm 2.

Moreover, according to (12), the desired signal part in the online transmissions thereby becomes to

$$\boldsymbol{S}(i, j) = \boldsymbol{S}(i, j) | \boldsymbol{\lambda} = \bar{\boldsymbol{h}}_{\mathrm{U}}(b^*), \tag{30}$$

and thus introduces an error caused by $|\bar{\boldsymbol{h}}_{\mathrm{U}}(b^*) - \tilde{\boldsymbol{h}}_{\mathrm{U}}(i)|$ in the demodulation process because $\boldsymbol{\lambda}$ is commonly assigned by $\tilde{\boldsymbol{h}}_{\mathrm{U}}(i)$. Fortunately, the differences between $\bar{\boldsymbol{h}}_{\mathrm{U}}(b^*)$ and $\tilde{\boldsymbol{h}}_{\mathrm{U}}(i)$ decrease with an increase in database size. Naturally, the residual signal-to-interference-plus-noise ratio (RSINR) for the ith frame and the jth symbol after using the proposed DC is written as

$$\begin{aligned} \eta_{\mathrm{pro}}(i, j) \approx\ & G_{\mathrm{PA,U}} \left\| \tilde{\boldsymbol{h}}_{\mathrm{U}}^{-1}(i) \boldsymbol{h}_{\mathrm{U}}(i) \boldsymbol{m}_{\mathrm{U}}(i, j) \right\|^2 \\ & \cdot \left(G_{\mathrm{PA,U}} \left\| \left(\bar{\boldsymbol{h}}_{\mathrm{U}}^{-1}(b^*) - \tilde{\boldsymbol{h}}_{\mathrm{U}}^{-1}(i) \right) \times \boldsymbol{h}_{\mathrm{U}}(i) \boldsymbol{m}_{\mathrm{U}}(i, j) \right\|^2 \right. \\ & + C_{\mathrm{ana}} \widehat{C}_{\mathrm{ant}} \left(G_{\mathrm{PA,BS}} \left\| \bar{\boldsymbol{h}}_{\mathrm{U}}^{-1}(b^*) \boldsymbol{h}_{\mathrm{BS}}(i) \mathbf{w}_{\mathrm{tx,BS}}(i, j) \right\|^2 \right. \\ & \left. \left. + \left\| \bar{\boldsymbol{h}}_{\mathrm{U}}^{-1}(b^*) \boldsymbol{h}_{\mathrm{BS}}(i) \boldsymbol{w}_{\mathrm{PA,BS}}(i, j) \right\|^2 \right) + \left\| \bar{\boldsymbol{h}}_{\mathrm{U}}^{-1}(b^*) \boldsymbol{w}_{\mathrm{rx}}(i, j) \right\|^2 \right)^{-1}, \end{aligned} \tag{31}$$

where we considered the results of (28) and (29) as approximate, assumed gains of all of antennas are 0 dBi, all of PAs are linear, and ignored noise effects from the interested user for simplicity of expressions.

The expression of (31) suggests that the SI effect has been suppressed to the first and second parts in the denominator of (31) by the proposed DC. The first part has to be left behind if $\bar{\boldsymbol{h}}_{\mathrm{U}}(b^*) \neq \tilde{\boldsymbol{h}}_{\mathrm{U}}(i)$ and results to a limitation in RSINR for high desired signal power. The second part caused by the noises in BS's modulation and PA can only be decreased depending on the antenna and analog cancellations (AACs). Our simulation results can provide some evidence for the analysis.

3.2. EL-Based Cancellation Approach. As a comparison object to evaluate our proposed approach, the EL-based cancellation approach should be taken into consideration. The EL used approach needs to train a model using a prepared database, such as the one generated following the description in Section 3.1.1, to predict the constellation value of SI in the 0/1 decision process. Assuming that the generation of database $\mathscr{B}$ is complete, we can construct a training set using parameters in $\mathscr{B}$ to train the model.

The training set includes a part of FVs $\mathscr{V}$ and a part of target values $\mathscr{T}$ which are designed by

$$\mathscr{V} = \left\{ \left[\bar{A}_{\mathrm{BS}}(z) ; \overline{\theta}_{\mathrm{BS}}(z) ; \bar{\boldsymbol{h}}_{\mathrm{U}}(z) ; \bar{\boldsymbol{h}}_{\mathrm{BS}}(z) \right] \right\}_{z=1}^{\sum_{b=1}^{|\mathscr{B}_{\mathrm{in}}|} |\mathscr{B}_{\mathrm{out}}(b)|}, \tag{32}$$

$$\mathscr{T} = \left\{ \bar{\boldsymbol{I}}(z) \right\}_{z=1}^{\sum_{b=1}^{|\mathscr{B}_{\mathrm{in}}|} |\mathscr{B}_{\mathrm{out}}(b)|}, \tag{33}$$

respectively. Afterward, a suitable model (or function) $\mathbb{M}^*$ can be trained by substituting $\mathscr{V}$ and $\mathscr{T}$ into a given network architecture and performing learning algorithms. Without loss of generality and noting that $\mathscr{T}$ are numerical output, we employ deep NN architecture and adopt a regression learning algorithm where the former is one of the most popular networks in EL methods, and the latter is used to train the model $\mathbb{M}^*$ [37]. Mathematically, $\mathbb{M}^*$ can be written as

$$\mathbb{M}^* = \arg \min_{\mathbb{M}} \mathscr{L}(\mathscr{T} - \mathbb{M}(\mathscr{V})), \tag{34}$$

where $\mathscr{L}(\boldsymbol{z})$ denotes loss function for training process in EL methods.

In fact, the trained model $\mathbb{M}^*$ is used to estimate constellation values of SI in 0/1 decision, and obviously, the outputs of trained model should be assigned to DC for SI suppression. Therefore, the DC $\boldsymbol{C}_{\mathrm{dig}}(i, j, \boldsymbol{\alpha})$ is calculated by

$$\boldsymbol{C}_{\mathrm{dig}}(i, j, \boldsymbol{\alpha}) = \mathbb{M}^*(\boldsymbol{\alpha}), \tag{35}$$

where the same $\boldsymbol{\alpha}$ in (27) in the proposed LL-based cancellation approach is used.

According to (35), the introduction of error between the real and predicted SI in the 0/1 decision cannot be avoided since both real CSI of user-BS and interantenna of BS cannot be acquired in practice, and there are always errors in the model training process [26]. Moreover, as shown in

$$\begin{aligned}\eta_{\mathrm{EL}}(i, j) = {} & G_{\mathrm{PA,U}} \left\| \tilde{\boldsymbol{h}}_{\mathrm{U}}^{-1}(i) \mathbf{h}_{\mathrm{U}}(i) \boldsymbol{m}_{\mathrm{U}}(i, j) \right\|^2 \\ & \cdot \left(\left\| C_{\mathrm{ana}}^{1/2} \widehat{C}_{\mathrm{ant}}^{1/2} G_{\mathrm{PA,BS}}^{1/2} \tilde{\boldsymbol{h}}_{\mathrm{U}}^{-1}(i) \boldsymbol{h}_{\mathrm{BS}}(i) \boldsymbol{m}_{\mathrm{BS}}(i, j) - \mathbb{M}^* \right. \right. \\ & \cdot \left. \left(\left[A_{\mathrm{BS}}(i, j) ; \theta_{\mathrm{BS}}(i, j) ; \tilde{\boldsymbol{h}}_{\mathrm{U}}(i) ; \tilde{\boldsymbol{h}}_{\mathrm{BS}}(i) \right] \right) \right\|^2 \\ & + C_{\mathrm{ana}} \widehat{C}_{\mathrm{ant}} \left(G_{\mathrm{PA,BS}} \left\| \tilde{\boldsymbol{h}}_{\mathrm{U}}^{-1}(b^*) \times \boldsymbol{h}_{\mathrm{BS}}(i) \boldsymbol{w}_{\mathrm{tx,BS}}(i, j) \right\|^2 \right. \\ & \left. + \left\| \tilde{\boldsymbol{h}}_{\mathrm{U}}^{-1}(b^*) \boldsymbol{h}_{\mathrm{BS}}(i) \boldsymbol{w}_{\mathrm{PA,BS}}(i, j) \right\|^2 \right) + \left\| \tilde{\boldsymbol{h}}_{\mathrm{U}}^{-1}(b^*) \boldsymbol{w}_{\mathrm{rx}}(i, j) \right\|^2 \Big)^{-1}, \end{aligned} \tag{36}$$

for RSINR of the EL methods with the same conditions as the proposed method, suppose the residual SI, i.e., the first component in the denominator, still significantly dominates the composite signal power because of amplification effect from $G_{\mathrm{PA,BS}}$. In that case, the demodulation for the desired signals is not optimistic. In the last section, we perform more simulations to evaluate this cancellation approach.

3.3. Regular Cancellation Approach. Since pilot signals are commonly used in each frames sent by users and BS in modern wireless communication system; the real CSI between the interested user and BS $\boldsymbol{h}_{\mathrm{U}}(i)$ and the real CSI between the interantenna of BS $\boldsymbol{h}_{\mathrm{BS}}(i)$ are estimated by $\tilde{\boldsymbol{h}}_{\mathrm{U}}(i)$ and $\tilde{\boldsymbol{h}}_{\mathrm{BS}}(i)$, respectively. According to the SI expression in (13), an estimation-based regular method for assignment of DC $\boldsymbol{C}_{\mathrm{dig}}(i, j, \boldsymbol{\alpha})$ is directly calculating the constellation value of SI $\boldsymbol{I}(i, j)$ using the estimated CSI $\tilde{\boldsymbol{h}}_{\mathrm{U}}(i)$ and $\tilde{\boldsymbol{h}}_{\mathrm{BS}}(i)$. Mathematically, $\boldsymbol{C}_{\mathrm{dig}}(i, j, \boldsymbol{\alpha})$ in this regular cancellation approach is assigned as

$$\boldsymbol{C}_{\mathrm{dig}}(i, j, \boldsymbol{\alpha}) = \boldsymbol{I}(i, j) | \boldsymbol{\lambda} = \tilde{\boldsymbol{h}}_{\mathrm{U}}(i), \boldsymbol{h}_{\mathrm{BS}}(i) = \tilde{\boldsymbol{h}}_{\mathrm{BS}}(i), \tag{37}$$

where $\boldsymbol{\alpha}$ is decided by (27).

The RSINR of the regular approach can be expressed as

$$\begin{aligned}\eta_{\mathrm{reg}}(i, j) = {} & G_{\mathrm{PA,U}} \left\| \tilde{\boldsymbol{h}}_{\mathrm{U}}^{-1}(i) \boldsymbol{h}_{\mathrm{U}}(i) \boldsymbol{m}_{\mathrm{U}}(i, j) \right\|^2 \\ & \cdot \left(C_{\mathrm{ana}} \widehat{C}_{\mathrm{ant}} G_{\mathrm{PA,BS}} \left\| \tilde{\boldsymbol{h}}_{\mathrm{U}}^{-1}(i) \left(\boldsymbol{h}_{\mathrm{BS}}(i) - \tilde{\boldsymbol{h}}_{\mathrm{BS}}(i) \right) \boldsymbol{m}_{\mathrm{BS}}(i, j) \right\|^2 \right. \\ & + C_{\mathrm{ana}} \widehat{C}_{\mathrm{ant}} \left(G_{\mathrm{PA,BS}} \left\| \tilde{\boldsymbol{h}}_{\mathrm{U}}^{-1}(b^*) \boldsymbol{h}_{\mathrm{BS}}(i) \boldsymbol{w}_{\mathrm{tx,BS}}(i, j) \right\|^2 \right. \\ & \left. + \left\| \tilde{\boldsymbol{h}}_{\mathrm{U}}^{-1}(b^*) \boldsymbol{h}_{\mathrm{BS}}(i) \boldsymbol{w}_{\mathrm{PA,BS}}(i, j) \right\|^2 \right) + \left\| \tilde{\boldsymbol{h}}_{\mathrm{U}}^{-1}(b^*) \boldsymbol{w}_{\mathrm{rx}}(i, j) \right\|^2 \Big)^{-1}, \end{aligned} \tag{38}$$

based on the same conditions as the proposed approach. This expression indicates that the SI effect, i.e., the first part in denominator, cannot be completely removed because an error of $\boldsymbol{h}_{\mathrm{BS}}(i) - \tilde{\boldsymbol{h}}_{\mathrm{BS}}(i) \neq 0$ has to be introduced, and this error is further multiplied by $G_{\mathrm{PA,BS}} \| \tilde{\boldsymbol{h}}_{\mathrm{U}}^{-1}(i) \|^2$ times. This fact suggests that, basically, only an excellent performance for estimation (e.g., estimation of wireless channel and PA nonlinearity) and good channel condition for the desired signal (i.e., bigger $\boldsymbol{h}_{\mathrm{U}}(i)$) can make sure that IBFD systems

with the regular DC work. Surely, we perform more simulations to evaluate this cancellation approach.

4. Channel Model

Without loss of generality and for simplicity, in this study, we consider composite fading channels with path loss and Rayleigh fading to simulate the attenuation between the interested user and BS, i.e., $\boldsymbol{h}_{\mathrm{U}}(i) = \rho_{\mathrm{U}}(i)e^{j\phi_{\mathrm{U}}(i)}\forall i$. The phase shift $\phi_{\mathrm{U}}(i)\forall i$ is modeled as independent and identically distributed (i.i.d.) random variable (RV) and follows uniform distribution between 0 and 2π radians. The amplitude gain $\rho_{\mathrm{U}}(i)$ for ith frame can be expressed as

$$\rho_{\mathrm{U}}(i) = \frac{c_o}{4\pi f_0}\sqrt{d_{\mathrm{U}}^{-\zeta_{\mathrm{U}}}(i)}r_{\mathrm{U}}(i), \tag{39}$$

where c_o is the velocity of light, $d_{\mathrm{U}}(i)$ denotes the distance between the interested user and BS for the transmitting of i th frame, and ζ_{U} is the path loss exponent around user [38, 39]. The term of $r_{\mathrm{U}}(i)\forall i$ is also modeled as i.i.d. RV and follows the Rayleigh distribution with the same cumulative distribution function (CDF) expressed as

$$\mathrm{CDF}_{r_{\mathrm{U}}(i)}(z) = 1 - \exp\left(-\frac{z^2}{2\sigma^2}\right), \forall i, \tag{40}$$

where σ is the scale parameter of the Rayleigh distribution [32].

For the modeling of channel attenuation between the BS's antennas $\boldsymbol{h}_{\mathrm{BS}}(i)$, considering two facts of not complicated surroundings and low building density around BS, further, since a centralized antenna deployment with a limited interantenna distance [40] is usually employed on the traditional BS, the channel varying between transceiver antennas can be assumed to be static. Based on the above assumption and description, the channel gains of $\boldsymbol{h}_{\mathrm{BS}}(i)$ for all of frames are thus modeled by the path loss model and expressed as

$$\boldsymbol{h}_{\mathrm{BS}}(i) = \frac{c_o}{4\pi f_0}\sqrt{d_{\mathrm{BS}}^{-\zeta_{\mathrm{BS}}}}\forall i, \tag{41}$$

where d_{BS} is the distance between the BS's antennas and ζ_{BS} denotes the path loss exponent around BS.

5. Computer Simulations

In this section, we evaluate and analyze the proposed LL-based cancellation approach for the IBFD systems with the help of computer simulations. To do that, we first explain how to generate the offline database based on the designed format, which is introduced in Section 3.1. Then, the use of the generated database for online transmission in the considered IBFD systems is described. At last, we present and analyze simulated results with comparisons of the EL-based cancellation approach described in Section 3.2 and the regular cancellation approach described in Section 3.3 and finally summarize the key findings.

5.1. Offline Database Generation for Simulations. In simulations, since quadrature phase shift keying (QPSK) modulations are planned to be used into both user and BS, the dimension for each FV of input space $\mathscr{B}_{\mathrm{in}}$ in database $\mathscr{B}$ can be reduced to $\overline{\theta}_{\mathrm{BS}}(b)$ and $\boldsymbol{h}_{\mathrm{U}}(b) = \bar{\rho}_{\mathrm{U}}(b)e^{j\overline{\phi}_{\mathrm{U}}(b)}$, where $\overline{\theta}_{\mathrm{BS}}(b)$ can be assigned as four possible values in the common QPSK modulation and is written as

$$\overline{\theta}_{\mathrm{BS}}(b) \in \left\{\frac{\pi}{4}, \frac{3\pi}{4}, \frac{5\pi}{4}, \frac{7\pi}{4}\right\}. \tag{42}$$

The term of $\bar{\rho}_{\mathrm{U}}(b)$ is the amplitude gain of the channel between the user and BS, and it can range from 0 to ∞ caused by the Rayleigh fading. In the simulations, a variable range for $\bar{\rho}_{\mathrm{U}}(b)$ is assumed as $[\bar{\rho}_{\min}, \bar{\rho}_{\max}]$, and it is calculated by

$$\bar{\rho}_{\min} = \frac{c_o}{4\pi f_0}\sqrt{-2\sigma^2\ln(1-\delta)d_{\mathrm{U}}^{-\zeta_{\mathrm{U}}}}, \tag{43}$$

$$\bar{\rho}_{\max} = \frac{c_o}{4\pi f_0}\sqrt{-2\sigma^2\ln\delta d_{\mathrm{U}}^{-\zeta_{\mathrm{U}}}}, \tag{44}$$

according to Rayleigh CDF, where δ denotes a given probability in the CDF of Rayleigh distribution and can range from 0 to 1. Thereafter, we define Δ_{am} as the resolution of amplitude coefficient $\bar{\rho}_{\mathrm{U}}(b)$, and thus, there are $\lceil(\bar{\rho}_{\max} - \bar{\rho}_{\min})/\Delta_{\mathrm{am}}\rceil$ possible values to be assigned, and

$$\bar{\rho}_{\mathrm{U}}(b) \in \left\{\bar{\rho}_{\min}, \bar{\rho}_{\min} + \Delta_{\mathrm{am}}, \bar{\rho}_{\min} + 2\Delta_{\mathrm{am}}, \cdots, \bar{\rho}_{\min} + \left(\left\lceil\frac{\bar{\rho}_{\max} - \bar{\rho}_{\min}}{\Delta_{\mathrm{am}}}\right\rceil - 1\right)\Delta_{\mathrm{am}}\right\}. \tag{45}$$

The term of $\overline{\phi}_{\mathrm{U}}(b)$ is the phase shift caused by the Rayleigh fading, and it can range from 0 to 2π radians. For the database generation, we define Δ_{ph} as the resolution of phase shift $\overline{\phi}_{\mathrm{U}}(b)$, and hence, there are $\lceil 2\pi/\Delta_{\mathrm{ph}}\rceil$ possible values to be assigned. Mathematically, $\overline{\phi}_{\mathrm{U}}(b)$ can be written as

$$\overline{\phi}_{\mathrm{U}}(b) \in \left\{0, \Delta_{\mathrm{ph}}, 2\Delta_{\mathrm{ph}}, \cdots, \left(\left\lceil\frac{2\pi}{\Delta_{\mathrm{ph}}}\right\rceil - 1\right)\Delta_{\mathrm{ph}}\right\}. \tag{46}$$

Based on the above configurations, an input space $\mathscr{B}_{\mathrm{in}}$ with

$$|\mathscr{B}_{\mathrm{in}}| = 4 \times \left\lceil\frac{\bar{\rho}_{\max} - \bar{\rho}_{\min}}{\Delta_{\mathrm{am}}}\right\rceil \times \left\lceil\frac{2\pi}{\Delta_{\mathrm{ph}}}\right\rceil, \tag{47}$$

unduplicated FVs can be generated for our simulations. For measurement and record of the output space $\mathscr{B}_{\mathrm{out}}$, based on our channel assumptions, the output space $\mathscr{B}_{\mathrm{out}}$ is, thus, reduced to $\bar{\boldsymbol{I}}$, and can create a one-to-one correspondence with input space $\mathscr{B}_{\mathrm{in}}$. The size of $\mathscr{B}_{\mathrm{in}}$ as well as $\mathscr{B}_{\mathrm{out}}$ becomes larger with decreases of Δ_{am}, Δ_{ph}, and δ, and thus, more storage for saving and higher compute capability for searching are needed. Certainly, the parameters of $\bar{\rho}_{\min}$,

$\bar{\rho}_{\max}$, Δ_{am}, and Δ_{ph} are not restricted by the considered channel models. Actually, these parameters can be set freely and independently to fit various hard- or software environments. In addition, in the database generation process, we generate the same database $N_{\mathscr{B}}$ times and average them to eliminate the effects caused by the various noises as much as possible.

5.2. Model Training for EL-Based Cancellation Approach. According to Section 3.2, once the above database is generated, we can format a training set which consists of FV

$$\mathscr{V} = \left\{ \left[\overline{\theta}_{\text{BS}}(z) ; \bar{\boldsymbol{h}}_{\text{U}}(z) \right] \right\}_{z=1}^{\sum_{b=1}^{|\mathscr{B}_{\text{in}}|} |\mathscr{B}_{\text{out}}(b)|}, \tag{48}$$

and the corresponding target value

$$\mathscr{T} = \left\{ \bar{\boldsymbol{I}}(z) \right\}_{z=1}^{\sum_{b=1}^{|\mathscr{B}_{\text{in}}|} |\mathscr{B}_{\text{out}}(b)|}, \tag{49}$$

for model training in the EL-based cancellation approach. A deep NN architecture is employed to train model $\mathbb{M}^*$ and is shown in Figure 4 in which the input layer and the output layer are used to accept the FV $\mathscr{V}$ and target value $\mathscr{T}$, respectively. A Levenberg-Marquardt back-propagation algorithm [41] is used to train node coefficients of all of layers in our feedforward NNs, and an introduction study of deep learning for wireless physical layer can be found in [37]. With a consideration of limited computational cost, we configure a maximum value of epoch for model training in which an epoch can be described as one complete cycle through the entire training dataset. Some essential parameters for model training in our current simulations are listed in Table 1.

5.3. Online Data Transmission for Simulations. Considering 4.6 ~ 4.9 GHz is assigned to a 5G-based private network called local 5G in Japan [11], we employ 4.6 GHz as carrier frequency in our simulations. After the database $\mathscr{B}$ and model $\mathbb{M}^*$ are generated, they can be used in online transmissions for the considered IBFD systems. When the desired symbol is received under the effects of SI and noises, the BS first attenuates the received composite signal waveform excluding the desired portion using the AACs C_{ant} and C_{ana}. Then, the searching algorithm is implemented for the proposed cancellation approach following the instructions described in Section 3.1.2 with the help of the generated database $\mathscr{B}$. The SI can also be estimated through the trained model $\mathbb{M}^*$ for the EL-based cancellation approach, or the estimated CSI can predict the SI for the regular cancellation approach. Subsequently, constellation values of SI found from the database $\mathscr{B}$, predicted by the model $\mathbb{M}^*$, or estimated by the regular approach are taken out and subtracted independently before the 0/1 decisions in BS's demodulator. Finally, binary data sent by the user is recovered, and some transmission performances are evaluated and analyzed. Simulation parameters are listed in Table 2.

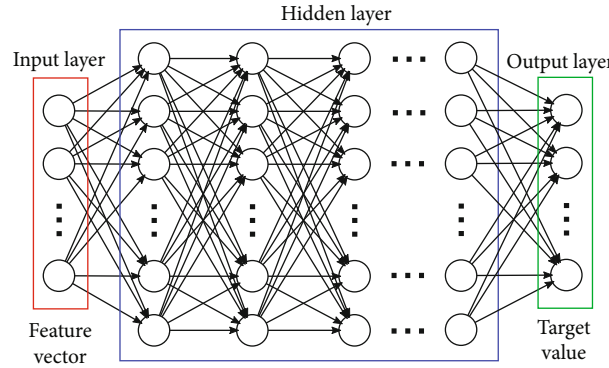


FIGURE 4: An illustration of fully connected deep NN architecture employed in EL-based cancellation approach for the IBFD systems.

5.4. Simulation Results. In this subsection, we first list three factors which may affect the performances of the above-mentioned DC approaches. They are (1) PA additional noise at BS side, (2) errors in channel estimation, and (3) database size. The additional noise is assumed as generated noise in all PAs, and its existence results noise figures are not one. The errors in channel estimation process occurred if the BS cannot obtain perfect CSI. In our simulations, the estimated CSI $\tilde{\boldsymbol{h}}_{\Psi}(i)$ for $\Psi = \text{U}$ or $\Psi = \text{BS}$ is generated by

$$\tilde{\boldsymbol{h}}_{\Psi}(i) = \boldsymbol{h}_{\Psi}(i) + \Delta_{\text{err},\Psi} |\boldsymbol{h}_{\Psi}(i)| e^{j\tau}, \tag{50}$$

where $\boldsymbol{h}_{\Psi}(i)$ denotes the real CSI of the interested user to BS or interantenna at BS side. $\Delta_{\text{err},\Psi}$ is defined as channel estimation error and can range from 0 to 1. τ is a random phase ranged from 0 to 2π and follows uniform distribution. The database size can be calculated by (47). Thereafter, we demonstrate simulation results under the effects of these factors, respectively. Finally, we evaluate transmission performances with all of the mentioned factors for the different DCs and show some key findings.

Notably, in our current systems, although all of the signals passing through the PAs are somewhat distorted owing to the PA nonlinearity, estimation errors on PA nonlinearity were not considered in this study. Inaccurate estimation of the PA nonlinearity may substantially degrade transmission performances for the estimation-based regular cancellation approach. However, considering the time-invariant property in all PAs, it seems not to be a major issue for the learning- and EL-based cancellation approaches. Nevertheless, our future work may concentrate on estimating PA nonlinearity.

In Figure 5, we demonstrate some comparisons of BER simulated by no SI, the proposed, the regular, and the EL-based DC approaches with variable AACs representing by $C_{\text{ant}}C_{\text{ana}}$ and fixed average receive signal-to-noise ratio (SNR). The average receive SNR is defined as the ratio of the average power of the desired signal to the average power of noise at the receiver antenna and is mainly dominated by the large-scale fading between the interested user and BS. The channel between the interested user and BS is assumed to be static to evaluate the effect caused by the factor (1), and channel estimation errors for all channels are ignored.

Table 1: Parameters for model training.

Parameters	Values	Parameters	Values
Number of hidden layers	3	Loss functions	Mean squared error
Number of hidden layer nodes	16, 8, 4	Train algorithm	Levenberg-Marquardt
Activation functions	Sigmoid (i.e., $(1+e^{-z})^{-1}$)	Max epoch	1000

Table 2: Simulation parameters.

Parameters	Values	Parameters	Values
Num. of frames, N_{frame}	1000	Carrier frequency, f_0	4.6 GHz
Num. of generated databases, $N_{\mathscr{B}}$	20	Modulation method for BS and user	QPSK
Num. of symbols in data part, N_{symbol}	8	Energy per symbol, E_{symbol}	2×10^{-6} mJ
Num. of symbols in pilot part, N_{pilot}	4	Symbol rate, f_{symbol}	5×10^{6}/s
Antenna gain at BS receive side, $G_{\text{ant,rx}}$	0 dBi	Antenna cancellations, C_{ant}	Variable
Antenna gain at BS transmit side, $G_{\text{ant,tx,BS}}$	0 dBi	Analog cancellations, C_{ana}	Variable
Antenna gain at user transmit side, $G_{\text{ant,tx,U}}$	0 dBi	PA gain for user, $G_{\text{PA,U}}$	13 dB
PA gain for BS, $G_{\text{PA,BS}}$	30 dB	PA OIP3 for user, $O_{3,\text{U}}$	18 dBm
PA OIP3 for BS, $O_{3,\text{BS}}$	35 dBm	Noise factor in user's PA, F_{U}	2.5
Noise factor in BS's PA, F_{BS}	2.5	Inter-antenna distance at BS, d_{BS}	0.2 m
Detection method, $\mathcal{D}(\boldsymbol{z})$	MF	Path loss exponent around BS, ζ_{BS}	2
Distance of user-BS, $d_{\text{U}}(i)\forall i$	Variable	Phase shift resolution, Δ_{ph}	Variable
Path loss exponent around user, ζ_{U}	4	Amplitude resolution, Δ_{am}	$\bar{\rho}_{\min}/10^2$
Probability in the CDF of Rayleigh distribution, δ	0.01	Noise power at BS receive antenna, Ω_{rx}	−100 dBm
Scale parameter of Rayleigh distribution, σ	$1/\sqrt{2}$	Noise power at user's modulator, $\Omega_{\text{tx,U}}$	−100 dBm
Noise power at BS's modulator, $\Omega_{\text{tx,BS}}$	−100 dBm	Estimation error for user-BS, $\Delta_{\text{err,U}}$	Variable
Estimation error for BS's antennas, $\Delta_{\text{err,BS}}$	Variable		

This figure indicates that the DC mentioned above approaches achieved similar BER performances over the entire range of $C_{\text{ant}}C_{\text{ana}}$ for a given SNR. The results clarify that factor (1) is not the major reason for the difference among the proposed, the regular, and the EL-based DC approaches. Moreover, there are gaps between BER performances of no SI and that of using DCs when $C_{\text{ant}}C_{\text{ana}} > -50$ dB. This phenomenon occurs because various noises at the BS side generated by PA and modulator are added into SI signals. Because these noises are random and there is nothing to do about them by the DC approaches, AACs had to be used to limit their power strength that can be confirmed by the second term in the denominator in (31), (36), and (38). Consequently, using a larger AACs with $C_{\text{ant}}C_{\text{ana}} \leq -50$ dB decreased the effects from the noises and resulted the similar BER performances as that of no SI effect. In fact, numerous authors have reported that they achieved more than 45 dB SI suppression capability using antenna cancellation (i.e., $C_{\text{ant}} < -45$ dB) [5, 13] and more than 30 dB suppression capability using analog cancellation (i.e., $C_{\text{ana}} < -30$ dB) [6, 11]. Considering feasibility, in our further simulations, we refer to the results of previous works and conservatively set $C_{\text{ant}} = -45$ dB and $C_{\text{ana}} = -15$ dB to suppress the effects of these noises.

In Figure 6(a), we present comparisons of RSINR versus SNR to demonstrate how factor (2) affects the mentioned DC approaches. The comparisons were obtained employing the proposed, regular, and EL-based DC approaches with variable channel estimation errors, assumption of static channels, and 60 dB of AACs, which suppressed the effect from factor (1). The BER performances of the mentioned DC approaches corresponding to Figure 6(a), with the same simulation conditions, are shown in Figure 6(b) to further verify the effect caused by the factor (2).

Figure 6(a) indicates that the RSINR of the regular and EL-based DCs is significantly decreased with increasing of channel estimation error from 10^{-5} to 10^{-3} over the entire range of SNR; on the contrary, the RSINR of the proposed approach is not affected by the channel estimation error. One of the major reasons for this result is whether the estimated CSI $\tilde{\boldsymbol{h}}_{\text{BS}}(i)$ is directly adopted in DC approaches or not. For the proposed DC, $\tilde{\boldsymbol{h}}_{\text{BS}}(i)$ works as a tag of SI and is used to help record quantized SI effect into database in offline phase or is used to help search target SI value from database in online phase. In other words, $\tilde{\boldsymbol{h}}_{\text{BS}}(i)$, actually, is not designed to participate in any SI estimation-related computations in the proposal. Naturally, it does not appear in the

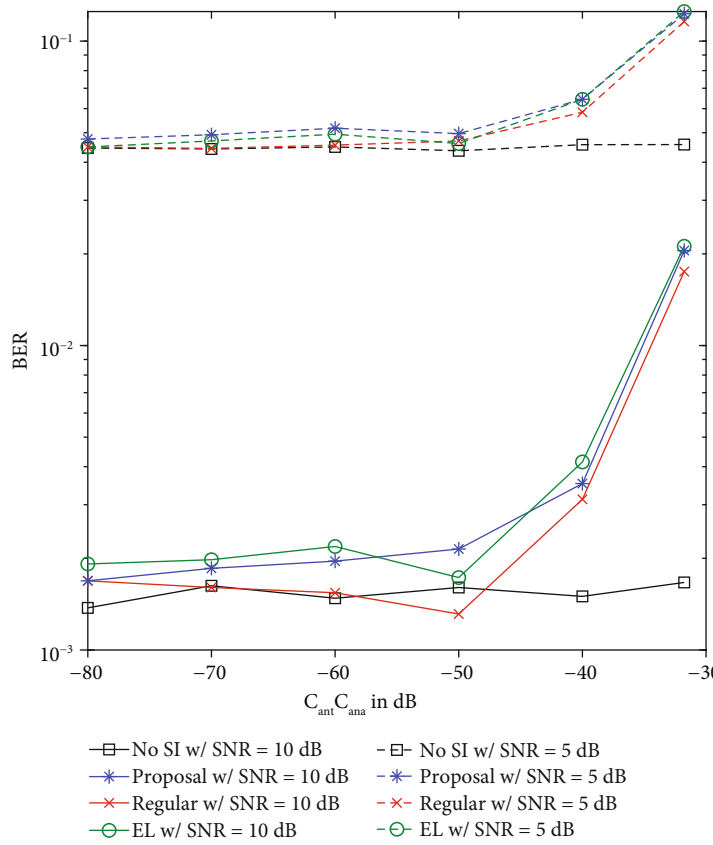


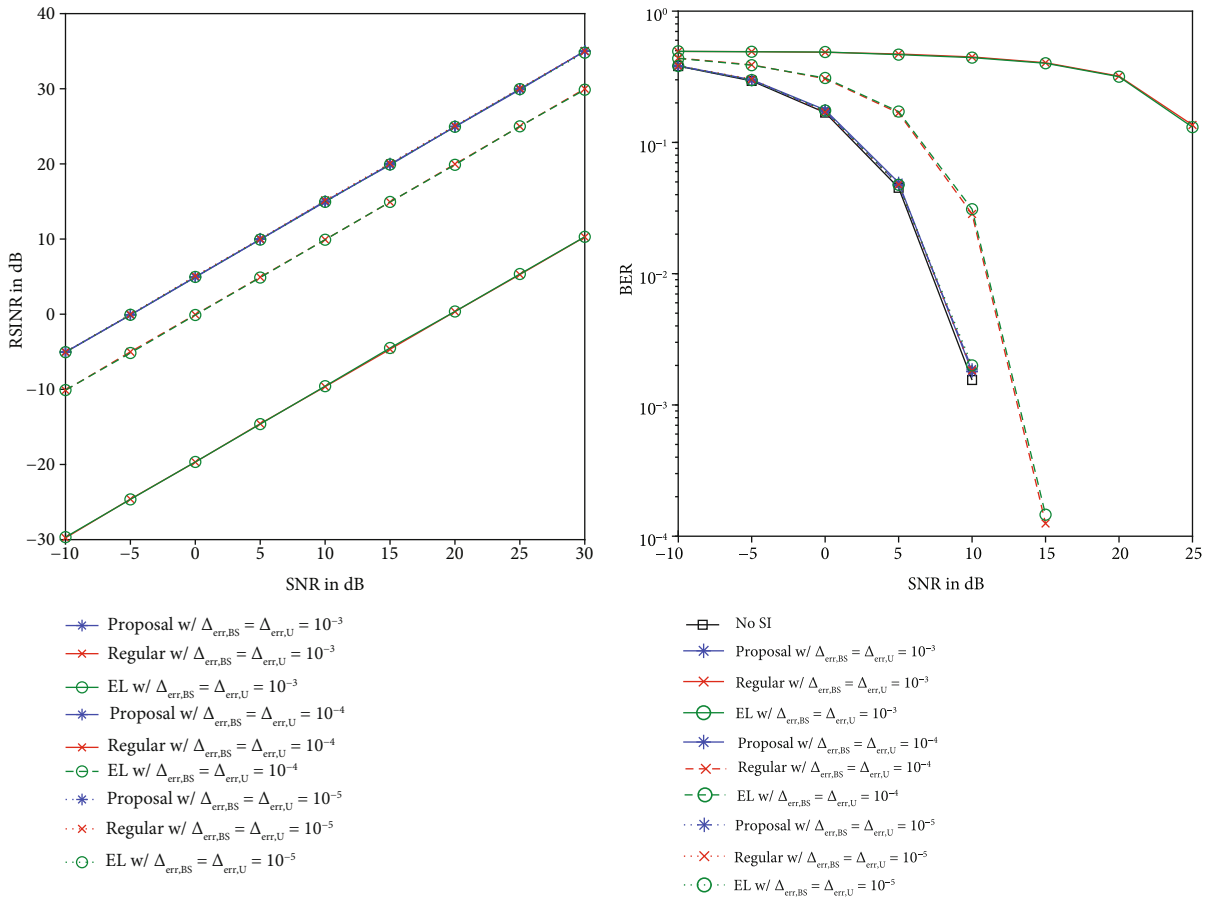
FIGURE 5: Analysis for factor (1): comparisons of BER simulated by no SI, the proposed, the regular, and the EL-based DC approaches with variable $C_{ant}C_{ana}$ and fixed average receive SNR. Channels are assumed to be static, and channel estimation errors are ignored.

expression of RSINR (31) and hardly affects the transmission performances such as BER. This also highlights the difference from other DC approaches. Conversely, in the other two DC approaches, $\tilde{\boldsymbol{h}}_{BS}(i)$ is a necessary variable to calculate the estimated value of SI and thereby directly affects the RSINR values in expressions (36) and (38). In practice, 10^{-3} of channel estimation error is a typical setting and acceptable. However, the RSINR of the regular and EL-based DCs with 10^{-3} estimation error cannot meet the general communication needs, and consequently, their BER performance shown in Figure 6(b) is terrible, unless improving measurement accuracy on channel estimation process to 10^{-5}.

Figures 7(a) and 7(b) exhibit RSINR and BER versus SNR plots simulated by no SI and the three DCs mentioned above under the effects of factor (3) database size. Rayleigh fading is considered in the channel between the interested user and BS to evaluate the effect caused by the factor (3), and perfect channel estimation and 60 dB AACs are assumed. Facing the limited computational cost and storage, we only vary the resolution of phase shift Δ_{ph} to control the database size. Note that since the regular DC approach does not require a database, there is only one curve for the regular approach in figures.

Both Figures 7(a) and 7(b) demonstrate that transmission performance (RSINR and BER) when using the proposed approach and EL-based DC approaches can be improved by increasing the database size. The results can be well explained by the big data technology because more potential SI values can be stored in a larger database and thus the probability of finding out an estimated SI value that is more closer to the real value of SI becomes higher. Interestingly, for a given database size, the RSINR of the proposed approach gradually approaches a fixed value with increasing SNR and finally results in a floor in BER. In fact, (31) also indicated that in a large SNR range, for example, a large $G_{PA,U}$, RSINR of the proposed approach is not linear growth with SNR unless $\bar{\boldsymbol{h}}_U(b^*) = \tilde{\boldsymbol{h}}_U(i)$. However, the RSINR and BER performances are much better than the EL-based DC approach because the latter introduced an unwanted error caused by the model training, and this error was amplified by BS's PA according to the first term in the denominator in (36). In addition, the estimation-based regular approach shows better performances than the proposed approach and the EL-based DCs because channel estimation errors, i.e., factor (2), in these simulations are not considered.

Finally, to evaluate a mixed effect of factors (1), (2), and (3), in Figures 8(a) and 8(b), we present some comparisons of RSINR and BER simulated by the proposed, the regular, and the EL-based DC approaches over fading channel with a common channel estimation error $\Delta_{err,BS} = \Delta_{err,U} = 10^{-3}$, a given database size $\Delta_{ph} = 0.1\pi$, and 60 dB of AACs. Under the consideration of the mixed effects of factors (1), (2), and (3), in Figure 8(a), for the case of SNR = 30 dB (i.e., a



(a) Comparisons of RSINR of the mentioned DC approaches

(b) Comparisons of BER of the mentioned DC approaches

FIGURE 6: Analysis for factor (2) with variable channel estimation errors, assumption of static channels, and 60 dB of AACs.

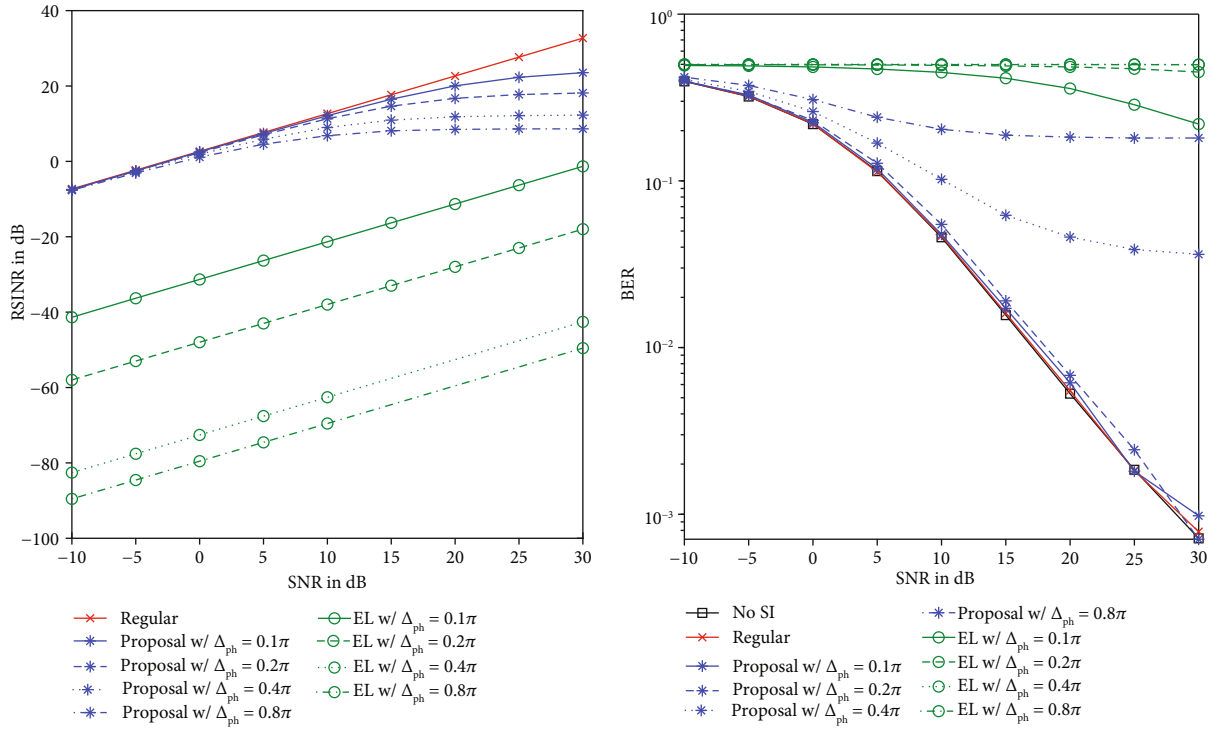
desired signal power of −70 dBm compared to noise power of −100 dBm), the proposed approach with $\Delta_{ph} = 0.1\pi$ approximately improved the RSINR to 24 dB with −94 dBm of residual SI power. Considering 40 dBm of EIRP at the BS, the figure clearly indicates that about 134 dB of SI suppression capability is achieved by the proposed approach with 60 dB of AACs, and Figure 8(b) thereby claims that our proposed approach can enable IBFD operations in the considered wireless communication systems and, consequently, can double the SE approximately.

Conversely, there are considerable RSINR and BER gaps between no SI and the regular or EL-based DC approach. The main reasons may be rough channel estimation and insufficient number of hidden layers and nodes in the network architecture. In fact, enhanced transmission performance for the regular DC can be achieved through strong AACs with highly accurate and comprehensive estimation operations. Similarly, applying strong AACs with highly extensive databases or nodes and layers in the NN architecture may improve EL-based DC performance. Therefore, although the performances of the regular and EL-based DC are inadequate under the current simulation setting, we do not deny the two potential DC approaches for IBFD systems. Instead, we strongly encourage exploring them, especially, for the EL-based DC approach, by, for example, using more faster learning algorithms, reducing dimension of the database, and upgrading hardware devices. Because the proposed LL-based DC approach for IBFD wireless systems can also benefit a lot from technological advance in machine learning-related research fields. Certainly, any type of DC approach should be combined with the outstanding AAC techniques; thereby, improving the existing AACs and solving related issues for the above-mentioned three DC approaches should be studied.

6. Discussion

Although our simulation results for SI cancellation are promising, the proposed LL-based DC approach is still needs to be evaluated using more hard- and/or software environments for the practical applications, and there are several works that need to be addressed.

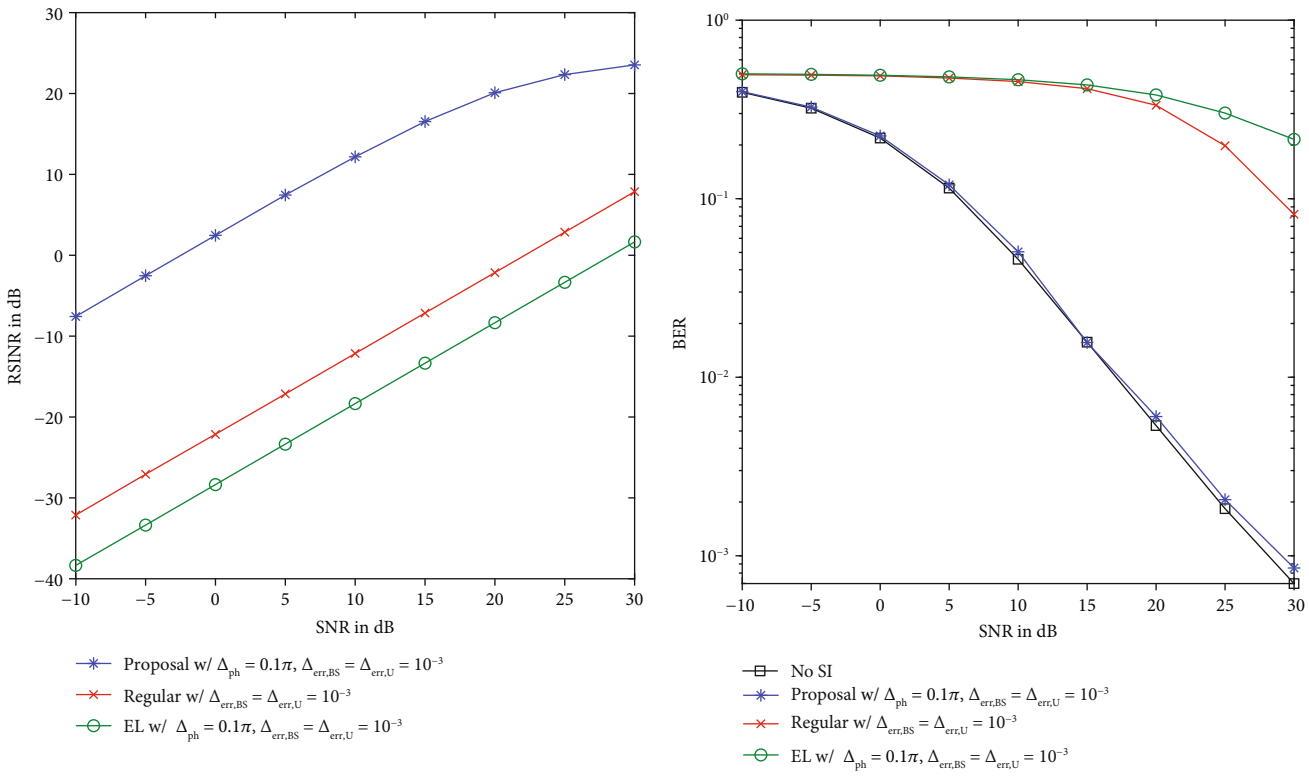
For example, the proposed approach in the current simulations is compatible with a signal carrier transmission system. For the popular orthogonal frequency division multiplexing (OFDM) which is adopted in most modern communications standards, multiple databases should be independently generated for each subcarrier because of frequency selective characteristics in wireless channels. Since creating and searching operations on a large number of



(a) Comparisons of RSINR of the mentioned DC approaches

(b) Comparisons of BER of the mentioned DC approaches

Figure 7: Analysis for factor (3) over fading channel with perfect channel estimation and 60 dB of AACs.



(a) Comparisons of RSINR of the mentioned DC approaches

(b) Comparisons of BER of the mentioned DC approaches

Figure 8: Analysis for mixed effect of factors (1), (2), and (3) over fading channel with a given channel estimation error, a given database size, and 60 dB of AACs.

databases must be considered, a lots of challenges on learning algorithm and calculation ability still need to be overcome before widespread use of the proposal.

Moreover, it seems straightforward to extend our proposal to the high-order constellations such as 256 quadrature amplitude modulation (QAM) by completely using the designed amplitude term $\bar{A}_{\mathrm{BS}}$ and phase term $\overline{\theta}_{\mathrm{BS}}$ in database (actually, $\bar{A}_{\mathrm{BS}}$ in the database is not used for the current QPSK modulation-based simulations). However, more advanced modulation means more combinations of amplitude and phase. Hence, a very large database also introduced a huge challenges on learning algorithm and calculation ability in this case.

Finally, in the present simulations, we assumed static channel between transceiver antennas on BS side. In fact, in a real system, the channel is time-varying. Similarly, the proposed approach can work for this situation by activating $\bar{\boldsymbol{h}}_{\mathrm{BS}}$ in the designed database (actually, $\bar{\boldsymbol{h}}_{\mathrm{BS}}$ is not used in the current simulations because of static channel). However, a large database caused by the dimension increase is still one of the unavoidable problems, and it is unclear how the complexity will increase in this case.

Based on the above description, the biggest challenge on the learning-based proposal is how to handle the huge amounts of data generated from the systems with OFDM operations, high-order constellations, and complex propagation environment. In the future, one of our major works is to solve this problem by introducing more advanced machine learning-related algorithms and technologies. Furthermore, we focus on conducting multiple evaluations for the proposal based on hard- and software environments, for example, using a SDR platform.

7. Conclusion

In this study, we proposed an LL-based DC approach for SI suppression and enabled IBFD transmissions to improve SE for current wireless communication systems. An offline and online phases for database generation and data transmission, respectively, are performed separately. In the offline phase, the output before the 0/1 decision is previously measured without the desired signal input and is recorded to a database with self-defined FV. In the online phase, a suitable result is located from the generated database with the help of the learning method and the FV usage for the same system architecture with desired signal input; the result is then assigned as the value of SI cancellation. An estimation-based regular and an EL-based DC approach are employed to simulate the transmission performance and evaluate the proposed approach. The simulation results signify that the proposed method could achieve about 134 dB SI suppression capability, BER performance comparable to zero SI, and enabled IBFD operations in wireless communication systems, superior to the aforementioned approaches.

Data Availability

The data including simulation configurations, parameters, and results used to support the findings of this study are included within the article.

Disclosure

Partial content of this study has been previously presented in [27] as conference.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was partly conducted under a contract of the R&D for Expansion of Radio Wave Resources (JPJ00254), organized by the Ministry of Internal Affairs and Communications (MIC), Japan.

References

[1] F. Boccardi, R. W. Heath, A. Lozano, T. L. Marzetta, and P. Popovski, "Five disruptive technology directions for 5G," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 74–80, 2014.

[2] M. Agiwal, A. Roy, and N. Saxena, "Next generation 5G wireless networks: a comprehensive survey," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 3, pp. 1617–1655, 2016.

[3] A. Sabharwal, P. Schniter, D. Guo, D. W. Bliss, S. Rangarajan, and R. Wichman, "In-band full-duplex wireless: challenges and opportunities," *IEEE Transactions on Slected Areas in Communications*, vol. 32, no. 9, pp. 1637–1652, 2014.

[4] D. Bharadia, E. McMilin, and S. Katti, "Full duplex radios," in *In The 13th ACM Special Interest Group on Data Communication (SIGCOMM)*, pp. 1–12, Hong Kong, China, August 2013.

[5] J.-H. Xun, L.-F. Shi, W.-R. Liu, and D.-J. Xin, "A self-interference suppression structure for collinear dipoles," *IEEE Antennas and Wireless Propagation Letters*, vol. 18, no. 10, pp. 2100–2104, 2019.

[6] Y. Liu, W. Ma, X. Quan, W. Pan, K. Kang, and Y. Tang, "An architecture for capturing the nonlinear distortion of analog self-interference cancellers in full-duplex radios," *IEEE Microwave and Wireless Components Letters*, vol. 27, no. 9, pp. 845–847, 2017.

[7] H. Guo, J. Xu, S. Zhu, and S. Wu, "Realtime software defined self-interference cancellation based on machine learning for in-band full duplex wireless communications," in *In 2018 International Conference on Computing, Networking and Communications (ICNC)*, pp. 1–5, Maui, HI, USA, July 2018.

[8] L. Shen, B. Henson, Y. Zakharov, and P. Mitchell, "Digital self-interference cancellation for full-duplex underwater acoustic systems," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 67, no. 1, pp. 192–196, 2020.

[9] C. Zhang and X. Luo, "Adaptive digital self-interference cancellation for millimeter-wave full-duplex backhaul systems," *IEEE Access*, vol. 7, pp. 175542–175553, 2019.

[10] K. Komatsu, Y. Miyaji, and H. Uehara, "Basis function selection of frequency-domain Hammerstein self-interference canceller for in-band full-duplex wireless communications," *IEEE Transactions on Wireless Communications*, vol. 17, no. 6, pp. 3768–3780, 2018.

[11] T. Matsumura and F. Kojima, "Prototype of analog self-interference cancellation based on super-heterodyne architecture for in-band full-duplex cellular system," in *In 2020 23rd International Symposium on Wireless Personal Multimedia*

Communications (WPMC), pp. 1–5, Okayama, Japan, October 2020.

[12] C. D. Nwankwo, L. Zhang, A. Quddus, M. A. Imran, and R. Tafazolli, "A survey of self-interference management techniques for single frequency full duplex systems," *IEEE Access*, vol. 6, pp. 30242–30268, 2018.

[13] C. Psomas, C. Skouroumounis, I. Krikidis, A. Kalis, Z. Theodosiou, and A. Kounoudes, "Performance gains from directional antennas in full-duplex systems," in *In 2015 IEEE International Conference on Microwaves, Communications, Antennas and Electronic Systems (COMCAS)*, pp. 1–5, Tel Aviv, Israel, November 2015.

[14] A. K. Khandani, "Two-way (true full-duplex) wireless," in *In 2013 13th Canadian Workshop on Information Theory*, pp. 33–38, Toronto, ON, Canada, October 2013.

[15] J. Mirza, G. Zheng, S. Saleem, and K.-K. Wong, "Optimization of uplink CSI training for full-duplex multiuser MIMO systems," *IEEE Communications Letters*, vol. 23, no. 12, pp. 2325–2329, 2019.

[16] W. Zhang, J. Yin, D. Wu, G. Guo, and Z. Lai, "A self-interference cancellation method based on deep learning for beyond 5G full-duplex system," in *2018 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC)*, pp. 1–5, Qingdao, China, December 2018.

[17] K. Satyanarayana, M. El-Hajjar, A. A. M. Mourad, and L. Hanzo, "Multi-user full duplex transceiver design for mmWave systems using learning-aided channel prediction," *IEEE Access*, vol. 7, pp. 66068–66083, 2019.

[18] H. Guo, S. Wu, H. Wang, and M. Daneshmand, "DSIC: deep learning based self-interference cancellation for in-band full duplex wireless," in *In 2019 IEEE global communications conference (GLOBECOM)*, pp. 1–6, Waikoloa, HI, USA, December 2019.

[19] M. Erdem, H. Ozkan, and O. Gurbuz, "Nonlinear digital self-interference cancellation with SVR for full duplex communication," in *In Wireless Communications and Networking Conference (WCNC)*, pp. 1–6, Seoul, Korea (South), May 2020.

[20] A. Balatsoukas-Stimming, "Joint detection and self-interference cancellation in full-duplex systems using machine learning," in *In 2021 55th Asilomar conference on signals, Systems, and Computers*, pp. 989–992, Pacific Grove, CA, USA, November 2021.

[21] S. Huang, Y. Ye, and M. Xiao, "Learning-based hybrid beamforming design for full-duplex millimeter wave systems," *IEEE Transactions on Cognitive Communications and Networking*, vol. 7, no. 1, pp. 120–132, 2021.

[22] K. E. Kolodziej, A. U. Cookson, and B. T. Perry, "RF canceller tuning acceleration using neural network machine learning for in-band full-duplex systems," *IEEE Open Journal of the Communications Society*, vol. 2, pp. 1158–1170, 2021.

[23] L. Shan, O. Zhao, K. Temma, K. Hattori, F. Kojima, and F. Adachi, "Evaluation of machine learnable bandwidth allocation strategy for user cooperative traffic forwarding," *IEEE Access*, vol. 7, pp. 85213–85225, 2019.

[24] M. G. Kibria, K. Nguyen, G. P. Villardi, O. Zhao, K. Ishizu, and F. Kojima, "Big data analytics, machine learning, and artificial intelligence in next-generation wireless networks," *IEEE Access*, vol. 6, pp. 32328–32338, 2018.

[25] W.-S. Liao, O. Zhao, M. G. Kibria, G. P. Villardi, K. Ishizu, and F. Kojima, "Machine learning based signal detection for comp downlink in ultra-dense small cell networks," *IEEE Access*, vol. 8, pp. 17454–17463, 2020.

[26] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, MIT Press, 2016.

[27] O. Zhao, W.-S. Liao, K. Li, T. Matsumura, F. Kojima, and H. Harada, "Lazy learning-based self-interference cancellation for wireless communication systems with in-band full-duplex operations," in *In 2021 IEEE 32th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, pp. 1589–1594, Helsinki, Finland, October 2021.

[28] E. K. Garcia, S. Feldman, M. R. Gupta, and S. Srivastava, "Completely lazy learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 9, pp. 1274–1285, 2010.

[29] F. A. P. de Figueiredo, F. A. C. M. Cardoso, I. Moerman, and G. Fraidenraich, "Channel estimation for massive MIMO TDD systems assuming pilot contamination and flat fading," *EURASIP Journal on Wireless Communications and Networking*, vol. 2018, 11 pages, 2018.

[30] J. P. Dunsmore, *Handbook of Microwave Component Measurements*, John Wiley & Sons, 2012.

[31] Agilent, "Fundamentals of RF and microwave noise figure measurements," *Application Note*, pp. 1–34, Keysight Technologies, 2010.

[32] A. Goldsmith, *Wireless Communications*, Stanford University, 2012.

[33] O. Zhao, L. Shan, W. S. Liao et al., "A device-centric clustering approach for large-scale distributed antenna systems using user cooperation," *IEICE Transactions on Communications*, vol. 7, no. 2, pp. 359–372, 2019.

[34] J. L. Bentley, "Multidimensional binary search trees used for associative searching," *Communications of the ACM*, vol. 18, no. 9, pp. 509–517, 1975.

[35] Y. A. Malkov and D. A. Yashunin, "Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs," *IEEE Transactions on Pattern Analysis and Machine Intelligence, Early Access Article*, vol. 42, no. 4, pp. 824–836, 2018.

[36] R. Weber, H. J. Schek, and S. Blott, "A quantitative analysis and performance study for similarity search methods in high dimensional spaces," in *In VLDB'98 Proceedings of the 24rd International Conference on Very Large Data Bases*, pp. 194–205, New York City, New York, USA, 1998.

[37] T. Wang, C.-K. Wen, H. Wang, F. Gao, T. Jiang, and S. Jin, "Deep learning for wireless physical layer: opportunities and challenges," *China Communications*, vol. 14, pp. 92–111, 2017.

[38] O. Zhao and H. Murata, "A study on dynamic clustering for large-scale multi-user MIMO distributed antenna systems with spatial correlation," *IEICE Transactions on Communications*, vol. 99, no. 4, pp. 928–938, 2016.

[39] O. Zhao, W.-S. Liao, K. Ishizu, and F. Kojima, "Dynamic and non-centric networking approach using virtual gateway platforms for low power wide area systems," *IEEE Access*, vol. 7, pp. 186078–186090, 2019.

[40] O. Zhao and H. Murata, "Sum-rate analysis for centralized and distributed antenna systems with spatial correlation and inter-cell interference," *IEICE Transactions on Communications*, vol. E98.B, no. 3, pp. 449–455, 2015.

[41] H. P. Gavin, "The Levenberg-Marquardt method for nonlinear least squares curve-fitting problems," *Department of Civil and Environmental Engineering, Duke University*, vol. 19, 2019.

Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 3918207, 15 pages
https://doi.org/10.1155/2022/3918207

Research Article

An Optimized Approach for Industrial IoT Based on Edge Computing

Hongyang Huang,[1] Mohammed Dauwed,[2] Morched Derbali,[3] Imran Khan,[4] Sun Li,[5] Kai Chen,[6] and Sangsoon Lim[7]

[1]*Graduate School of Business, Segi University, Jalan Teknologi, Kota Damansara, 47810 Petaling Jaya, Selangor, Malaysia*
[2]*Department of Medical Instrumentation Techniques Engineering, Dijlah University College, Baghdad, Iraq*
[3]*King Abdulaziz University (KAU), Faculty of Computing and Information Technology (FCIT), Jeddah, Saudi Arabia*
[4]*Department of Electrical Engineering, University of Engineering & Technology, Peshawar 814, Pakistan*
[5]*Advanced Information Research Center of Xi'an Jiaotong University, China*
[6]*Huawei Technologies, Stockholm, Sweden*
[7]*Department of Computer Engineering, Sungkyul University, Anyang 430010, Republic of Korea*

Correspondence should be addressed to Sangsoon Lim; slim@sungkyul.ac.kr

Received 17 May 2022; Accepted 24 June 2022; Published 9 July 2022

Academic Editor: Lisheng Fan

The Internet of Things (IoT) is an information network that connects gadgets and sensors to allow new autonomous tasks. The Industrial Internet of Things (IIoT) refers to the integration of IoT with industrial applications. Some vital infrastructures, such as water delivery networks, use IIoT. The scattered topology of IIoT and resource limits of edge computing provide new difficulties to traditional data storage, transport, and security protection with the rapid expansion of the IIoT. In this paper, a recovery mechanism to recover the edge network failure is proposed by considering repair cost and computational demands. The NP-hard problem was divided into interdependent major and minor problems that could be solved in polynomial time by using the Benders decomposition technique and cutting plane approximation. To ensure the nonincreasing character of the Benders upper limit, a local branching method was also added to improve the convergence. Simulation results indicated that the proposed method is superior to the existing method and has better overall performance.

1. Introduction

The Industrial Internet of Things (IIoT) is regarded as an important driver of the intelligent transformation of the global industrial system. Relying on hundreds of millions of seamlessly deployed sensors, collectors, and controllers, the Industrial Internet of Things can simulate, predict, and control the full cycle of the manufacturing process [1]. As the "brain" of the IIoT, the edge computing network provides more sufficient computing processing capabilities for wireless acquisition devices, effectively reducing processing and transmission delays, and is useful for digital twin (DT) [2], virtual reality (VR), etc. Enterprise high-level applications have laid a solid foundation [3]. At the same time, the wired link connection between edge computing nodes also makes the migration of computing tasks between nodes smoother, effectively alleviating the problem of unbalanced space-time allocation of computing resources caused by the space-time fluctuations of computing demands of the IIoT.

The normal and stable operation of the edge computing network is the key to the efficient operation of the IIoT. Once the "brain" is damaged, the IIoT system will lose effective control over the "limbs" (such as supply chain monitoring, data visualization, and analysis), bringing countless economic losses, even life-threatening. However, the stability requirements of edge computing networks face internal and external challenges. On the one hand, the edge computing network is coupled with the power grid, control network,

and other subnetworks in the IIoT, forming a highly vulnerable interdependent network [4]. Any slight fluctuations in other subnets may be transmitted to the edge computing network, causing large-scale system cascading failures. On the other hand, unpredictable events such as natural disasters and man-made attacks also test the robustness of edge computing networks at any time [5]. In order to meet the above challenges, on the one hand, the reliability of the edge computing network can be strengthened so that it can adaptively cope with various network fluctuations and prevent network failures. On the other hand, and more importantly, it is necessary to explore the rapid repair mechanism after the network is damaged, so that the network performance can be restored to the level close to before the damage as soon as possible.

Although the design of the repair mechanism is crucial for the sustainable and stable operation of the network, there is currently no research literature specifically targeting edge computing network scenarios in the IIoT. In view of the similarity of network topology, network dynamics, and other characteristics, some existing network repair strategies can still provide some reference for the design of edge computing network repair mechanisms in the IIoT. The research on the existing network repair mechanism mainly focuses on the rapid repair after the local network is damaged. Reference [6] constructs the problem of single node or link damage in the network as an integer linear programming problem and proposes a data migration-aware repair model, which achieves an effective balance between service interruption rate and repair cost. Further, in the literature [7, 8], considering the problem of network connectivity damage in multinode failure scenarios, it is proposed that users can be used as transit nodes between disconnected edge nodes in a device-to-device (D2D) way [7] or mobile devices can be used. The access node realizes that the data between network nodes can be reached everywhere [8], ensuring the connectivity of the damaged network. Different from the above studies, the reference [9] considers the continuous damage state in the case of network attack, transforms the network dynamic repair problem into a differential game theory problem, and enhances the network repair ability through Nash equilibrium necessary conditions and competitive strategy sets. However, the above-mentioned research on local network repair often ignores global network nodes, dynamic characteristics between links (such as flow migration), and practical scene constraints (such as link space layout without changes); it is difficult to effectively extend to the large-scale network damage scenarios that are more likely to occur in the IIoT.

After the network is damaged on a large scale, the initial available repair resources (such as the number of repair personnel and the number of replaceable devices) are often limited. How to effectively balance the limited system repair resources at the initial stage of repair and the urgent need for system performance recovery is an urgent problem to be solved in the current IIoT. Current research mainly focuses on the analysis of network topology. Reference [10] believes that large-degree nodes, that is, nodes with large node degrees, play a more important role in network connectivity and need to be repaired first. Similarly, in [11], considering that the link is damaged, the link with large betweenness centrality (BC) should be repaired first. Reference [12] found through the analysis of actual network data that weakly connected nodes in the network, that is, nodes with low degrees of themselves but connected to several large-degree nodes, play the most critical role in network connectivity, and their repairs are prioritized. The level should be higher than the large-degree node. However, the above schemes based on the growing maximum connected subgraph of network connectivity are all static network architecture analysis, ignoring the dynamic characteristics of the network. The large-scale network repair in the real environment often forms independent subgraphs first and then connects multiple subgraphs to form a maximum connected graph [13]. Therefore, in the engineering analysis, it is necessary to integrate more network equipment details and actual transmission dynamic analysis [14]. In order to solve the above problems, a heuristic algorithm that is easier to solve is proposed in [15]. However, it lacks the macroscopic analysis of the network, and the performance variance of the algorithm is large, so it cannot provide reliable performance guarantee. Similarly, both the simulated annealing algorithm [16] and the genetic algorithm [17] face the same dilemma as general heuristic algorithms and are easily trapped in local optimal solutions. Although the hill-climbing algorithm [18] or the gradient descent algorithm [19] can easily jump out of the local optimum and greatly reduce the computational complexity of the problem, the optimality of the solution cannot be guaranteed. This type of problem is also known as a network design problem (NDP). Due to the addition of the dynamic characteristics of the network, the complexity of the network design problem is extremely high (at least the NP-complete problem), and the traditional dynamic programming algorithm will cause the "dimension disaster" [20]. In reference [21], the problem of multihop computing task offloading in the hybrid edge cloud computing environment is studied, and the offloading method that meets the quality of service requirements is realized through the game method. In reference [22], optimization is carried out in terms of time cost and energy consumption cost to achieve the optimal allocation of large-scale green energy-saving computing resources. In reference [23], in order to meet the real-time requirements, an intelligent resource planning strategy under the hybrid computing structure is proposed. In summary, edge computing can provide more sufficient computing resources for field devices and reduce network load by deploying at the network edge closer to field devices, thereby meeting the requirements of task service quality and reducing system overhead in different scenarios.

In view of the practical dilemma faced by the current research, this paper proposes a repair mechanism for damaged edge computing networks in the IIoT scenario. Different from the existing literature, this paper deeply excavates the structural characteristics (topological relationship, link capacity) and dynamic characteristics (node computing requirements), and the link priority repair set decision and network computing migration issues are jointly considered,

in order to achieve an efficient balance between the initial computing requirements and repair costs of network repair. The main contributions of this paper are summarized as follows.

A network repair mechanism is proposed in the case of large-scale damage to the edge computing network. Combined with the network structure and dynamic characteristics of edge computing, an analysis framework of priority repair set decision and resource scheduling in the early stage of network repair is provided.

Based on the Benders decomposition algorithm, the complex mixed-integer problem is decomposed into two parts, the main problem and the subproblem, which are easier to solve. For the subproblems of multivariable groups, by adding virtual source nodes and destination nodes, the problem is transformed into a network maximum flow problem to solve.

A Benders decomposition acceleration algorithm based on local branching method is designed. The trust region based on the Hamming distance is used to shrink the search range of the feasible region and accelerate the convergence speed of the algorithm.

Simulation results show that the proposed algorithm has better convergence performance and lower system overhead. Compared with the existing random repair, maximum connected graph repair, and betweenness centrality sorting repair and other topology-based repair algorithms, the proposed algorithm has better performance in multiple scenarios and has good scalability and adaptability.

2. System Model

For the edge computing network in the IIoT, consider an edge computing network with N nodes, which is represented by the set $\forall i \in N$. The edge nodes are connected by wired links; the link set is represented by E. Since wired links are usually reliable, and only a small amount of channel coding complexity is required for computing tasks relative to edge nodes, the link between two edge nodes can be considered error-free [24]. The system parameters are shown in Table 1. It is worth noting that the edge computing network in the actual IIoT is often a hybrid link transmission network composed of wired links and wireless links. Since there is almost no damage to the wireless link, and its repair cost is negligible compared to the wired link, this paper only considers the case where all transmission links are wired links. Nevertheless, if the deep weakening of the short-term wireless channel is not considered, the wireless link in the static scenario (without considering the spectrum reallocation in the repair process) can be regarded as a nondestructive link with constant capacity, and the hybrid link transmission network can be considered equivalent to a pure wired link network, and the proposed algorithm will still be applicable.

The topology of the damaged edge computing network is shown in Figure 1. In the actual network, because edge nodes often have complex self-protection mechanisms (such as overheating protection), they are usually not prone to damage. Therefore, this paper does not consider node damage and focuses on link damage. In order to simulate the state of large-scale network failure caused by factors such as natural disasters, it is assumed that there are $q|E|$ wired links in the network at the initial moment, where q represents the percentage of damaged links in all links, and the set E_0 represents the damaged link. The set $E^1 = E \setminus E^0$ represents the link set that can still work normally. The damage of the link will cause the balanced computing migration flow between edge nodes to be broken and even form a computing island (computing migration cannot be performed, as shown in node 1 in Figure 1), resulting in a mismatch between computing requirements and computing capabilities, affecting the network computing performance. Constrained by limited repair resources at the initial stage of network repair (e.g., the number of repair personnel and replacement equipment inventory), it is impractical to repair all damaged links at the same time. In order to restore the network state as soon as possible, it can be distributed according to the network computing requirements, and some damaged links can be repaired preferentially, which is represented by the set E^r ($E^r \subseteq E^0$). For any link $ij \in E^0$, due to the different degree of damage, the cost c_{ij} of maintenance, repair, and replacement is also different.

Considering that the local data computing requirement of the edge node $i \in N$ is r_i, the actual local computing amount is p_i. It should be noted that r_i is the total calculated arrival amount of the ith edge node in the initial stage of network repair. For a given scenario, r_i is a fixed value that does not change with time and can be estimated more accurately based on historical computing needs. For any edge node i, the p_i satisfies

$$0 \le p_i \le \bar{p}_i, \forall i \in N. \tag{1}$$

Among them, $\bar{p}_i$ is the maximum computing power of node i.

Let the computational cost of migration between node i and node $j \in N\{i\}$ be f_{ij}. If $f_{ij} > 0$, the data computing task is migrated from node i to j. If $f_{ij} < 0$, the data computing task is migrated from node j to i. Limited by the wired link capacity $\bar{f}_{ij}$, the actual calculated migration of link ij satisfies

$$\left| f_{ij} \right| \le \bar{f}_{ij}, \forall ij \in E. \tag{2}$$

Further, for damaged links, the actual computational migration amount is not only determined by the wired link capacity, but also affected by the link repair decision, i.e.

$$\left| f_{ij} \right| \le \bar{f}_{ij} e_{ij}, \forall ij \in E^0, \tag{3}$$

where

$$e_{ij} \in \{0, 1\}, \forall ij \in E^0. \tag{4}$$

Among them, $e_{ij} = 1$ represents the priority to repair the link ij, that is, $ij \in E^r$; when $e_{ij} = 0$ means that the link ij is not repaired, it actually calculates the migration

Table 1: System parameters.

Parameter	Description
q	Damaged links as a percentage of all links
E^1	The set of all normal links in the edge computing network
E^0	The set of all damaged links in the edge computing network
E	The set of all links in the edge computing network
N	Number of edge computing network nodes
c_{ij}	Repair cost of link ij
c_i	The cost per unit of data discarded by node i
p_i	The amount of local computing data calculation of node i
$\bar{p}_i$	The maximum computing power of node i
r_i	The local data computation requirement of node i
d_i	The amount of computational data discarded by node i
f_{ij}	Computational migration from node i to node j
$\bar{f}_{ij}$	Capacity of link ij
e_{ij}	Whether to fix the decision variable of link ij
t	Number of iterations for Benders decomposition
θ^t	The optimal value of the objective function of the subproblem at iteration t
Δ^t	Left-branch problem threshold at iteration t

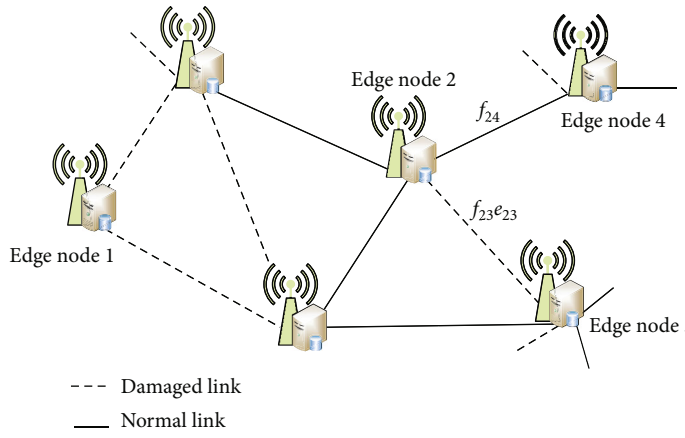


Figure 1: Damaged edge computing network topology.

amount $f_{ij} = 0$. In addition, for a single link ij, it can be known from the migration flow symmetry:

$$f_{ij} = -f_{ji}, \forall ij \in E. \tag{5}$$

When the links in the set E^r are repaired, it can be known from the law of the conservation of computation of nodes:

$$P_i + \sum_{ij \in E^0} f_{ij} + \sum_{ij \in E^1} f_{ij} + d_i = r_i, i \in N. \tag{6}$$

Among them, r_i represents the local data computation requirement of node i; d_i represents the amount of computing tasks that edge node i has to give up due to data backlog due to limited computing power.

$$d_i \geq 0, \forall i \in N. \tag{7}$$

In order to repair as many links as possible, the amount of data discarded can be reduced, and the system performance can be improved, but it will cause a large system repair overhead. If there are too few repair links, the system repair overhead will be reduced, but it may cause the data to be discarded because it cannot be processed locally, which will damage the network performance. For quantitative analysis, the network performance in this paper is measured by the total cost of system data discarding. The greater the amount of data discarded, the higher

the total cost of data discarding, and the worse the network performance. Therefore, in order to balance the repair cost and network performance at the initial stage of network repair (total cost of data discarding), the following system cost minimization problem can be constructed

$$P: \min_{e,d,f,p} \varnothing = \sum_{ij \in E^0} c_{ij} e_{ij} + \sum_{i \in N} c_i d_i, \quad \text{s.t.Eqs.}(1) \sim (7). \tag{8}$$

Among them, c_{ij} represents the repair cost required to repair the link $ij \in E^0$, c_i denotes the cost of discarding each unit of data for node $i \in N$. The link repair decision vector $\boldsymbol{e} = (e_{ij}, \forall ij \in E^0)$, data discarding decision vector $\boldsymbol{d} = (d_i, \forall i \in N)$, the flow allocation vector $\boldsymbol{f} = (f_{ij}, \forall ij \in E)$, and the local actual computation vector $\boldsymbol{p} = (p_i, \forall i \in N)$. In problem P, the total system overhead consists of two parts: the total cost of system repair overhead and data discarding. As mentioned above, the system repair cost $\sum_{j \in E^0} c_{ij} e_{ij}$ and the total cost of data discarding $\sum_{i \in N} c_i d_i$ are a pair of contradictory quantities, an increase in one value will lead to a decrease in the other value, and minimizing their sum can effectively balance their effects. When the link repair cost c_{ij} is large, it indicates that the importance of computing data is relatively low, and the system tends to temporarily repair less damaged links and discard data that cannot be processed. When the cost per unit of data discarding is large, the system tends to repair more damaged links to reduce the discarding of computing data.

Problem P is a mixed-integer problem, NP-hard problem, and its data scale is large, and there is no known polynomial-time algorithm to solve the above problem. The current methods for solving the above problems can be divided into three categories: Heuristic algorithm [25], approximate algorithm [26], and exact algorithm [27]. The Heuristic algorithms are fast and easy to apply, but they lack rigorous theoretical proofs, and the results often deviate significantly from the optimal solution. Approximate algorithms, such as the slack variable method, have a limited solution scale, and the solution to the slack problem cannot accurately describe the optimal solution to the original problem. Different from the above two methods, the exact algorithm, such as the cut plane algorithm, can well explore the optimal solution of the problem through the iterative update of the cut plane and is widely used in the process of solving mixed-integer problems.

3. Algorithm Design

Benders decomposition algorithm [28] is a classical cutting plane algorithm, which is widely used to deal with real mixed-integer programming problems (such as locomotive scheduling and aviation route planning). This algorithm does not significantly increase the number of iterations with the increase of operating variables like the branch and bound method, nor does it produce dimensional disaster like dynamic programming, and does not appear heuristic, simulated annealing, and other algorithms have huge variance [29]. This section will give an efficient solution to problem P based on the Benders decomposition algorithm.

3.1. Subproblem Description and Transformation. In the Benders decomposition algorithm, the original problem can be decomposed into two parts, the main problem and the subproblem, and the optimization variables in the main problem are called complex variables. When the complex variables are fixed, the remaining optimization problems (i.e., subproblems) in the original problem become relatively easy to solve. For the problem P, if the value of the complex variable e_{ij}^t in the tth iteration process is given, the subproblem can be expressed as

$$S: \min_{d,f,p} \theta = \sum_{i \in N} c_i d_i, \quad \text{s.t.Eqs.}(1) \sim (2), (5) \sim (7), \tag{9}$$

$$\left| f_{ij} \right| \le \bar{f}_{ij} e_{ij}^t, \forall ij \in E^0. \tag{10}$$

Since the 0-1 variable e_{ij} in the original problem P is decomposed into the main problem, and the optimization variables in the above subproblems are all continuous variables, this problem can be equivalent to a minimum cost flow problem. This paper considers a typical environmental monitoring scenario in the IIoT. Each edge node in the edge network is responsible for processing the environmental data collected by wireless sensors. In the environment monitoring scenario, the computing tasks at each edge node have the same computing priority [3, 20], that is, the node discards the same cost per unit of data $(c_i = c_j, \forall ij \in N)$. Thus, the problem can be further transformed into the network maximum flow problem [30]. Figure 2 illustrates the above equivalence relationship with an edge network with four nodes.

In Figure 2, the connection line between nodes 2 and 3 is the damaged link that needs to be repaired determined in the iterative process, namely $\{ij, ij \in E^r,\ E^r \subseteq E^0\}$. The source node s and the destination node z are newly added virtual nodes, the virtual link capacity of the source node s and the four edge nodes is the maximum computing power of the four edge nodes, and the virtual link between the four edge nodes and the destination node z represents the local data computation requirement of each edge node. The numerical values on each edge in Figure 2 represent the link capacity of the link, so that maximizing the total arrival flow to the destination node z is equivalent to minimizing the total data discarded. The above transformed maximum flow problem can be efficiently solved by existing algorithms, such as the Ford-Fulkerson algorithm [31].

3.2. Cut Plane Generation and Main Problem Construction. In the Benders decomposition algorithm, the solutions of the subproblems are substituted into the main problem to generate linear constraints in the main problem, namely Benders cuts. Since this paper considers the perfect resource [32], that is, for any feasible solution to the main problem, there are always feasible subproblem solutions, and there is no need to generate feasible cuts; only the optimal cut needs

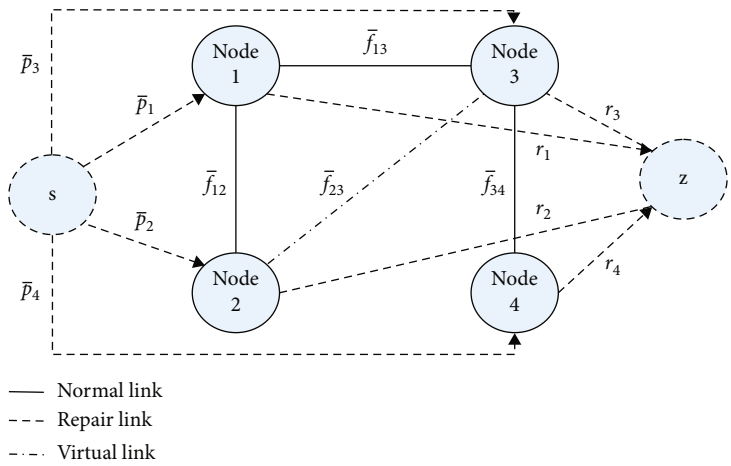


Figure 2: Subproblem equivalent maximum flow problem.

to be constructed. In order to form the optimal cut plane of Benders, it is necessary to use the complementary relaxation principle of the dual problem [33]. According to the solution of the above subproblems, extract the dual variable $\mu_{ij}^t, ij \in E^0$, corresponding to the constraint in Eq. (3), which can be repaired system increment of link ij.

Theorem 1. *The optimal cut plane of Benders in the tth iteration can be expressed as*

$$\eta \geq \theta^t + \sum_{ij \in E^0} \mu_{ij}^t \bar{f}_{ij} \left(e_{ij} - e_{ij}^t \right). \tag{11}$$

Among them, η is the upper bound of the optimal solution of the objective function of the subproblem, and θ^t is the optimal value of the objective function obtained in the tth iteration of the subproblem, that is, the minimum total cost of data discarding in the tth iteration.

Proof. See Appendix A.1. □

Substituting the above Benders optimal cutting plane into the constraints of solving the main problem, the main problem can be obtained as

$$\begin{aligned} M: \min_{e,\eta} \underline{\varnothing} &= \sum_{ij \in E^0} c_{ij} e_{ij} + \eta. \\ \text{s.t.Eqs.}(4), (9). \end{aligned} \tag{12}$$

Among them, $\underline{\varnothing}$ represents the lower bound of the optimal value of the original problem P, because the main problem only considers part of the constraints and is a relaxation problem of the original problem.

Different from the subproblem that determines the data migration, calculation, and discarding of the edge network, the main problem is responsible for determining the link set E^r that needs to be repaired preferentially in the damaged link set E^0. In the loop iteration process of the main and subproblems, the solution of the main problem is carried out given the amount of data migration, calculation, and discarding. The E^r obtained by solving the main problem is in turn used for further solving of the subproblems, thereby gradually approximating the optimal data migration, calculation, discarding, and link repair strategies of the system.

3.3. Iterative Path Repair and Computational Migration Algorithm Based on Benders Decomposition Theory. The above main problem is initialized into subproblems, and the loop iteration starts to find the optimal solution. If the decision variables of the main problem cannot satisfy all the constraints in the iterative process, the algorithm is terminated, and the original problem has no solution. Otherwise, the iterative process continues until the optimal configuration of the network is found. The specific steps of the above process are shown in Algorithm 1.

When the algorithm converges, the system will repair the link whose value is 1 according to the e^t calculated by the current iteration. After that, the solution f^t of the subproblem is used to determine the amount of data flow migration between nodes, the actual computing power of each node is adjusted according to p^t, and the data d^t that exceeds the computing power is discarded. It should be noted that in the main problem, η is a continuous variable and e is a discrete variable, and the problem is still a mixed-integer problem. Although it can be solved by algorithms such as genetic ant colony algorithm, the complexity of the algorithm is still very high due to its large search domain space. In addition, in the process of each iteration, the introduction of a new cut plane in the main problem keeps the lower bound $\underline{\varnothing}$ of the algorithm nondecreasing, but there is no similar mechanism to guarantee the monotonicity of the upper bound $\overline{\varnothing}$ of the algorithm. This nonmonotonic constraint bound property will further aggravate the computational time overhead of the above algorithm.

4. Benders Decomposition Acceleration Algorithm Design

In order to solve the problems existing in the iterative path repair and computational migration algorithm based on

Define: $t = 1$, given network parameters $\{c_{ij}, \bar{f}_{ij} | \forall ij \in E\}$, $\{r_i, \bar{p}_i, c_i | \forall i \in N\}$.
Initialization: Upper bound $\overline{\varnothing} = +\infty$ and algorithm lower bound $\underline{\varnothing} = -\infty$, iteration accuracy parameter ε.
1: Cycle
2: Solve the main problem M, if the main problem has no solution, terminate the algorithm, the original problem has no solution. Otherwise, $e^t = (e_{ij}^t, \forall ij \in E^0)$ and the algorithm lower bound $\underline{\varnothing}$
3: Solve the subproblem S, get $\boldsymbol{d}^t = (d_i^t, \forall i \in N)$, $\boldsymbol{p}^t = (p_i^t, \forall i \in N)$, $\boldsymbol{f}^t = (f_{ij}^t, \forall ij \in E)$, $\boldsymbol{\mu}^t = (\mu_{ij}^t, \forall ij \in E^0)$ and the minimum objective function value θ^t
4: Update algorithm upper bound $\overline{\varnothing} = \min\{\overline{\varnothing}, \sum_{ij \in E^0} c_{ij} e_{ij}^t + \theta^t\}$
5: $t = t + 1$
6: Until the algorithm converges, the convergence condition is $(\overline{\varnothing} - \underline{\varnothing})/\overline{\varnothing} < \varepsilon$

Algorithm 1: Bender iterative path repair and computational migration.

Benders decomposition theory proposed in Algorithm 1, this paper further introduces the local branching technique in the iterative solution process [34]. Its main purpose is to find a better upper bound of the problem in each iteration process, in order to realize the inward clamping of the upper and lower bounds and reduce the computational complexity of the main problem.

4.1. Trust Region and Hamming Distance. As mentioned earlier, the Benders decomposition algorithm based on cut planes is not a stable algorithm, and in the early stages of the iteration, the solution of the problem fluctuates widely in different feasible regions, resulting in a slow convergence rate. In the edge computing scenario in the IIoT considered in this paper, a large number of optimization variables introduced by large-scale network damage, especially the introduction of the repair link decision vector e with an initial search space of $2^{|E^0|}$, will make this convergence speed problems are further exacerbated. The trust region is an excellent strategy to address the above-mentioned large-scale fluctuation characteristics. Considering that e in the main problem is a set of 0-1 variables, the Hamming distance can be used to limit the distance between the two iterative solutions.

Assuming that (e^t, d^t, p^t, f^t) is the feasible solution obtained by the tth iteration of the original problem, the set of all 0-1 optimization variables with a value of 1 can be expressed as $E^t = \{e_{ij}^t | e_{ij}^t = 1, e_{ij}^t \in E^0\}$, and then the Hamming distance between the $(t+1)$th iteration and the tth iteration is

$$D\left(e_{ij}^{t+1}, e_{ij}^t\right) = \sum_{ij \in E^t} \left(1 - e_{ij}^{t+1}\right) + \sum_{ij \in E^0 \setminus E^t} e_{ij}^{t+1}. \tag{13}$$

That is, the number of binary variables e changes in the $(t+1)$th iteration relative to the tth iteration.

To speed up the convergence, the solution space can be decomposed into two independent trust regions:

$$D\left(e_{ij}^{t+1}, e_{ij}^t\right) \le \Delta^{t+1}, \tag{14}$$

$$D\left(e_{ij}^{t+1}, e_{ij}^t\right) \ge \Delta^{t+1} + 1. \tag{15}$$

Among them, Δ^{t+1} represents the size of the trust region in the $t+1$)th iteration, and the selection of its value depends on the complexity of the main problem and the volatility requirements of the search range. In the above way, the original problem is naturally divided into two subsolution spaces, and the local branching method is based on this. In the local branching method, Equations (14) and (15) are called the left and right branches, respectively.

4.2. Local Branching. Based on the Hamming distance, the solution space of the original problem P can be divided into two closely connected neighborhood spaces according to the feasible solutions (e^t, d^t, p^t, f^t) obtained by the tth iteration of the Benders decomposition. Let $e^k, k \in K^t$ be the solutions of all e (including feasible solutions and nonfeasible solutions) calculated by the local branch method iteration during the tth iteration of Benders decomposition, where the set of feasible solutions is denoted as L^t $(L^t \subseteq K^t)$. According to the Hamming distance, the original problem P can be branched into two independent left and right branch problems. The left branch problem is

$$P_k : \min_{e,d,f,p} \varnothing = \sum_{ij \in E^0} c_{ij} e_{ij} + \sum_{i \in N} c_i d_i, \quad \text{s.t. Eqs.}(1) \sim (7), \tag{16}$$

$$D\left(e_{ij}, e_{ij}^k\right) \ge 1, \forall ij \in E^0, k \in K^t, \tag{17}$$

$$D\left(e_{ij}, e_{ij}^l\right) \ge \Delta^t + 1, \forall ij \in E^0, l \in L^t, \tag{18}$$

$$D\left(e_{ij}, e_{ij}^m\right) \le \Delta^t, \forall ij \in E^0. \tag{19}$$

Among them, Equation (17) indicates that the e value that has been compared before is not repeatedly compared, Equation (18) indicates that the current left branch should be the branch of the right branch in the previous branch, and Equation (19) indicates the left branch restriction, and e_{ij}^m is the currently obtained optimal solution. Correspondingly,

the right branch problem can be obtained as

$$\bar{P}_k : \min_{e,d,f,p} \varnothing = \sum_{ij \in E^0} c_{ij} e_{ij} + \sum_{i \in N} c_i d_i, \quad \text{s.t.Eqs.}(1) \sim (7), (15) \sim (17), \tag{20}$$

$$D\left(e_{ij}, e_{ij}^m\right) \geq \Delta^t + 1, \forall ij \in E^0. \tag{21}$$

Here, Equation (18) represents the right branch restriction.

Let $(e^{k+1}, d^{k+1}, p^{k+1}, f^{k+1})$ be the optimal solution of the left branch problem P_k, and the corresponding objective function value is $\varnothing_{k+1}$, and then there is a local branching method algorithm flow as shown in Algorithm 2.

The $T_{\max}$ in the loop condition of Algorithm 2 is to avoid the situation that the branching problem cannot be solved because the Hamming distance is set too large during the calculation process. In a large-scale damaged network, the solution space of the repair link decision e is very large, and choosing a smaller trust region size Δ^t will be more conducive to improving the calculation speed of the left branch problem.

In the iterative process, the strict upper bound $\overline{\varnothing}_k = \min_{k \in K^t}\{\varnothing_k\}$ of the original problem can be obtained. At the same time, since the main difficulty of the original problem P lies in the acquisition of the lower bound, a series of optimal cutting planes generated in each iteration process can also speed up the search speed, so that the Benders decomposition can simultaneously enhance the optimal solution of the problem in the iterative process. Search for upper and lower bounds [35].

4.3. Benders Decomposition Acceleration Algorithm Based on Local Branching Method. The Benders decomposition algorithm accelerated by the local branching method obtained by integrating the local branching algorithm into the Benders decomposition is shown in Algorithm 3.

Similar to Algorithm 1, when Algorithm 3 converges, the system will determine the set of repair links based on the et calculated by the current iteration. Then, according to the calculated values of f^t, p^t, and d^t, the migration, calculation, and discarding of data in the network are determined. Different from Algorithm 1, the Algorithm 3 ensures the strict decrease of the upper bound of the original problem P in the iterative process through the local branching technique. For large-scale damaged networks, this means that in each iteration process, the search space of the link repair decision vector e^t to be optimized for the main problem M is gradually reduced, and the search speed of the solution is gradually accelerated with the increase of the number of iterations.

In the new cuts obtained by each iteration of the above algorithm, not all cuts need to be added to the main problem solving process, and only the deepest cuts can be added (even the cuts with the smallest feasible region). Considering the convergence requirements of the algorithm and the poor performance of the initial stage of the Benders decomposition, it is possible to add trust region constraints only in the initial iteration stage. When the iterations become stable, this restriction is lifted.

It is worth noting that in each branch process, the branch problem $P_1, P_2, \cdots$ has the same structure as the original problem P. Therefore, for each branch problem, Algorithm 1 can be used to solve it. The cut plane obtained from the previous branching problem can also be used to solve the subsequent branching problem.

5. Simulation Results

In this section, the proposed iterative path repair and computational migration algorithm based on Benders decomposition theory (hereinafter referred to as Benders decomposition algorithm) and the iterative path repair and computational migration algorithm based on Benders decomposition acceleration theory based on local branching method (hereinafter referred to as Benders decomposition acceleration performance of the proposed algorithm) are simulated and tested. The performance advantages of the proposed algorithm compared with other benchmark algorithms are verified and analyzed.

5.1. Simulation Parameters and Comparison Algorithms. This paper considers an edge computing network with $N = 50$ nodes, and the maximum computing power $\bar{p}_i$ of each node in the initial stage of network repair is independent of each other and evenly distributed in (10, 25) Gbit. Considering the matching of computing power and computing requirements, the computing requirements r_i of nodes in the initial stage of network repair also obey the uniform distribution on (10, 25) Gbit. Without loss of generality, it is assumed that the number of wired links in the original topology of the edge network is $|E| = 2N$, the network topology is the same as the random network [4], and the data transmission between nodes is reachable everywhere. Considering the fluctuation range of processing power of computing nodes, the link capacity $\bar{f}_{ij}$ obeys a uniform distribution on (5, 15) Gbit. In order to simulate the large-scale damage of the network, in the following, unless otherwise specified, the link damage ratio $q = 75\%$. Assuming that the cost of repairing each damaged link c_{ij} is evenly distributed in $(1 \times 10^4, 2 \times 10^4)$ pesos, the cost c_i of discarding data due to insufficient computing power is 10^4 peso/Gbit.

In order to verify the effect of the damaged network repair mechanism, the proposed algorithm is compared with the following benchmark algorithms.

(1) Random repair algorithm. The damaged links in the network are randomly selected for repair, regardless of the specific topology of the network and the dynamic characteristics of network computing flow migration

(2) Maximum connected graph repair algorithm [15]. Repair all damaged links in the maximum connected graph of the network to ensure that all nodes in the maximum connected graph of the network can reach a strong connected state

Define: $k = 1$, let the initial feasible solution of local branch iteration $(e^k, d^k, p^k, f^k) = (e^t, d^t, p^t, f^t)$, and the corresponding objective function value is $\varnothing_k$. Given Δ^t, the maximum allowable computation time $T_{\max}$ and the iteration accuracy parameter ε
1: Cycle
2: Divide the current branch into left branch P_k and right branch $\bar{P}_k$, calculate $\varnothing_{k+1}$ and $(e^{k+1}, d^{k+1}, p^{k+1}, f^{k+1})$
3: If $\varnothing_{k+1} < \varnothing_k$ holds
4: Then update $L^t = L^t \bigcup \{k\}$, and jump to the right branch
5: Otherwise, $\varnothing_{k+1} \geq \varnothing_k$ or P_k has no solution
6: Let $\Delta^t = \Delta^t + 1$, add $D(e_{ij}, e_{ij}^k) \geq 1, \forall ij \in E^0$ to the constraint Eq. (15), and update $K^t = K^t \bigcup \{k\}$. Recalculate $(e^{k+1}, d^{k+1}, p^{k+1}, f^{k+1})$ and $\varnothing_{k+1}$ of the left branch problem P_k, go back to step 3)
7: $k = k + 1$
8: Until the algorithm computation time reaches $T_{\max}$ or meets the accuracy requirement $(\varnothing_k - \varnothing_{k+1})/\varnothing_k < \bar{\varepsilon}$

Algorithm 2: Local branch scheme.

Define: $t = 1$, given network parameters $\{c_{ij}, \bar{f}_{ij} | \forall ij \in E\}$, $\{r_i, \bar{p}_i, c_i | \forall i \in N\}$.
Initialization: Upper bound $\overline{\varnothing} = +\infty$ and algorithm lower bound $\underline{\varnothing} = -\infty$, iteration accuracy parameter ε
1: Cycle
2: Solve the main problem M, if the problem has no solution, terminate the algorithm, and the original problem has no solution. Otherwise, $e^t = (e_{ij}^t, \forall ij \in E^0)$ and the algorithm lower bound $\underline{\varnothing}$
3: Solve the subproblem S, and get $d^t = (d_i^t, \forall i \in N)$, $p^t = (p_i^t, \forall i \in N)$, $f^t = (f_{ij}^t, \forall ij \in E)$, the dual variable $\mu^t = (\mu_{ij}^t, \forall ij \in E^0)$ and the minimum objective function value θ^t
4: Update upper bound $\overline{\varnothing} = \min\{\overline{\varnothing}, \sum_{ij \in E^0} c_{ij} e_{ij}^t + \theta^t\}$
5: if $(\overline{\varnothing} - \underline{\varnothing})/\overline{\varnothing} < \varepsilon$
6: Determine the optimal solution, then terminate the algorithm
7: Otherwise, generate a new Benders optimal cutting plane
8: Run Algorithm 2 to obtain an enhanced upper bound $\overline{\varnothing}_k$, and a series of Benders optimal cuts generated by different feasible solutions
9: $t = t + 1$
10: Until the algorithm converges

Algorithm 3: Iterative path repair and computational migration.

(3) Betweenness centrality sorting repair algorithm [11]. The topological structure of the edge network before the damage is analyzed, the betweenness centrality of each wired link is sorted, and the damaged link with large betweenness centrality is preferentially repaired. The number of repaired links is the same as the proposed algorithm

5.2. Algorithm Convergence. Figure 3 shows the convergence comparison between the proposed Benders decomposition algorithm and the Benders decomposition acceleration algorithm. It can be seen from Figure 3 that the two algorithms can achieve fast iteration within a limited number of times and have good convergence. It should be noted that the number of iterations represents the total number of loop iterations of the main problem and subproblems in Algorithm 1 and Algorithm 3. In each iteration process, the computational complexity of the subproblem is $\mathcal{O}(|N||E|^2)$, and the computational complexity of the main problem is $\mathcal{O}(|N||E^0|^3)$, where n is the inner loop iteration for solving the main problem frequency. Although the Benders decomposition acceleration algorithm reduces the total number of iterations only once compared to the Benders decomposition algorithm, its actual computational time complexity is reduced by $(|N||E|^2 + n|E^0|^3)$. In a large-scale compromised network environment, the improvement of the computational time complexity is still considerable. From the comparison of the upper and lower bounds of the algorithm under the same number of iterations, it can be seen that, compared with the Benders decomposition algorithm, the local branch method-assisted Benders decomposition acceleration algorithm has a significantly improved convergence speed due to the introduction of deeper cutting planes and can be better adapt to the application requirements of large-scale networks. In addition, the total system cost of the Benders decomposition algorithm and the Benders decomposition acceleration algorithm tend to be consistent after convergence. Therefore, in the following, for the convenience of comparison with the benchmark algorithm, the Benders decomposition algorithm and the Benders decomposition acceleration algorithm are collectively referred to as the proposed algorithm.

5.3. Algorithm Performance. Figures 4 and 5 are the total system cost curves of the proposed algorithm, the random repair algorithm, the maximum connected graph repair

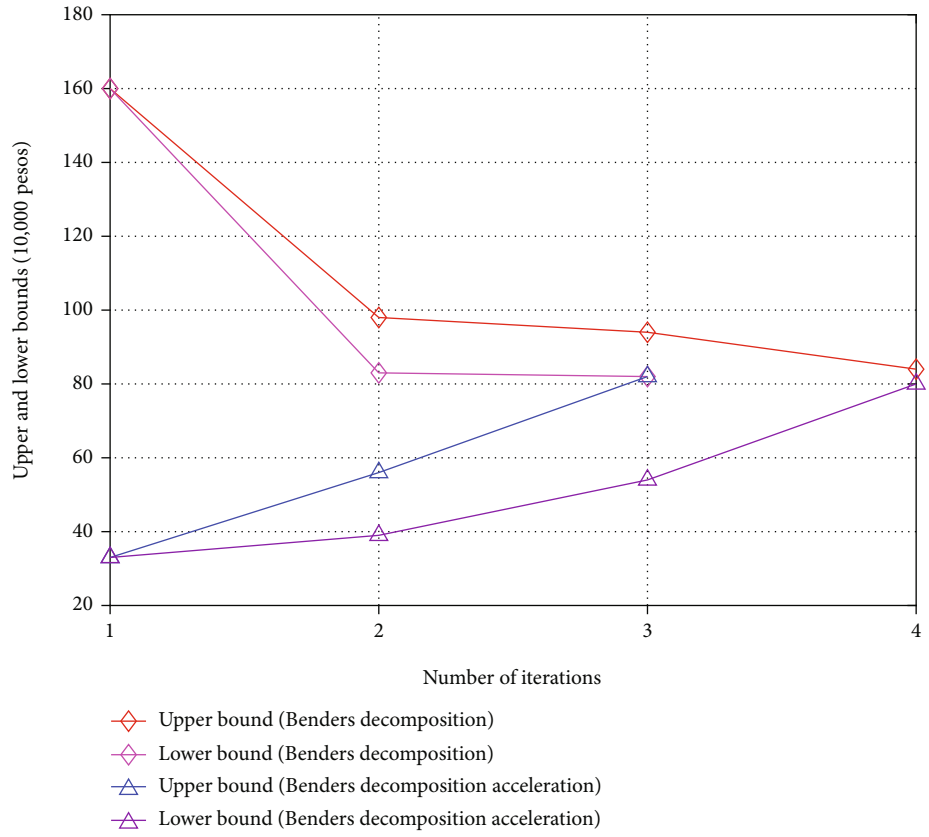


Figure 3: Convergence comparison between the proposed Benders decomposition algorithm and the Benders decomposition acceleration algorithm.

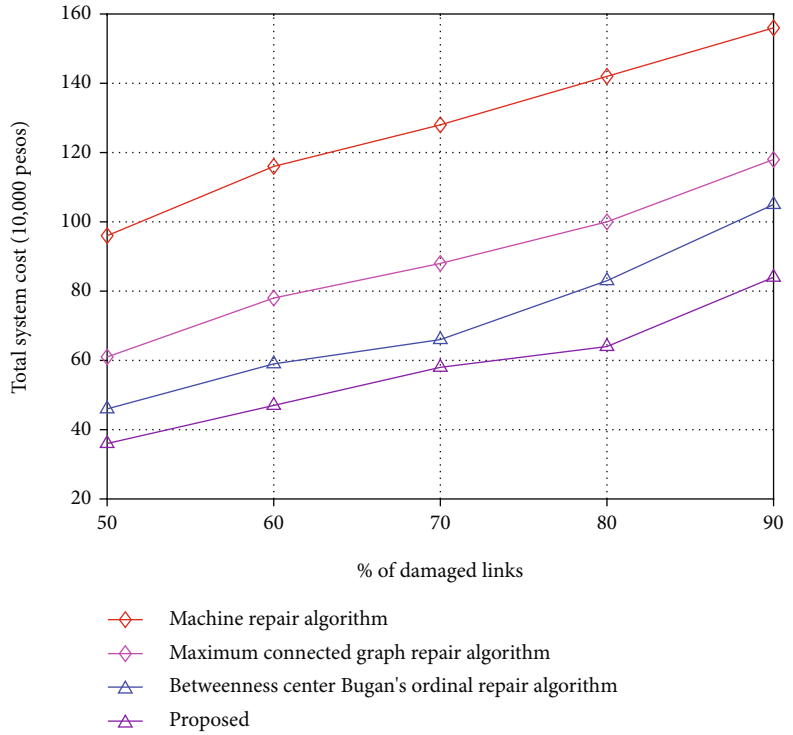


Figure 4: Algorithm performance comparison under different network damage levels.

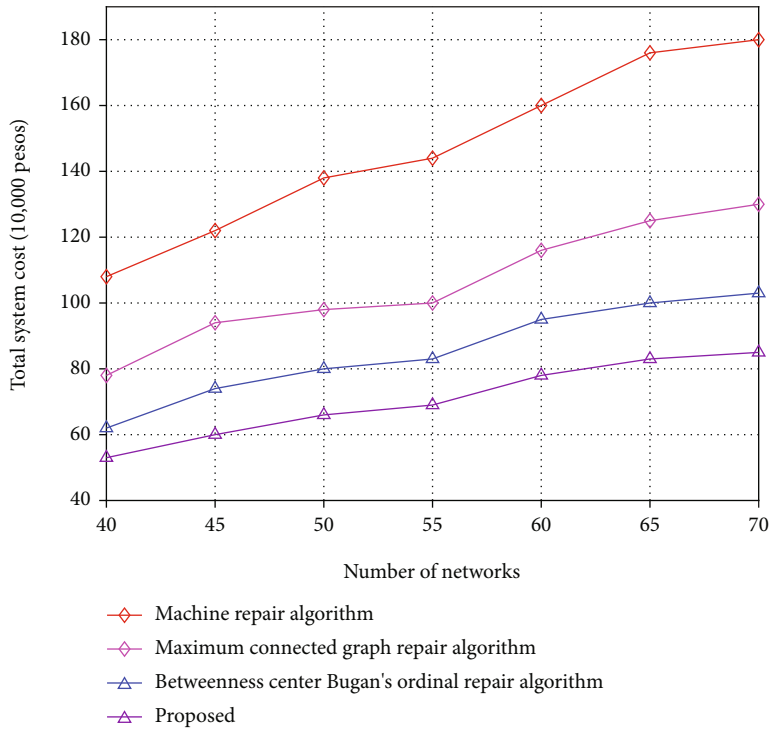


Figure 5: Algorithm performance comparison under different network scales.

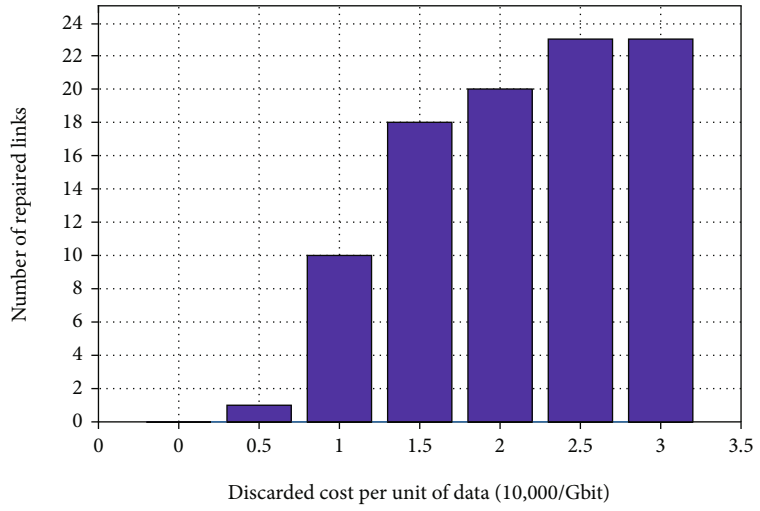


Figure 6: The number of damaged links repaired by the proposed algorithm under different unit data discarding costs.

algorithm, and the betweenness centrality sorting repair algorithm under different network damage degree q and different network scales N, respectively. It can be seen from Figures 4 and 5 that under different network damage degrees and network scales, the performance of the proposed algorithm is excellent due to the consideration of the network dynamic characteristics such as the computing demand and actual computing power of the nodes in the network and link capacity limitations on the benchmark algorithm. It is worth noting that the system overhead performance of the maximum connected graph repair algorithm is the worst among all algorithms, even weaker than the random repair algorithm, which is in good agreement with the state of network repair in real networks. In the real network, the repair process always makes the large and small independent clusters form in the network first, and at the end of the repair, the clusters are connected to form the maximum connected graph [13].

5.4. Multiscenario Deployment. Figures 6 and 7, respectively, show the number of damaged links repaired by the proposed algorithm and the amount of data discarded under different

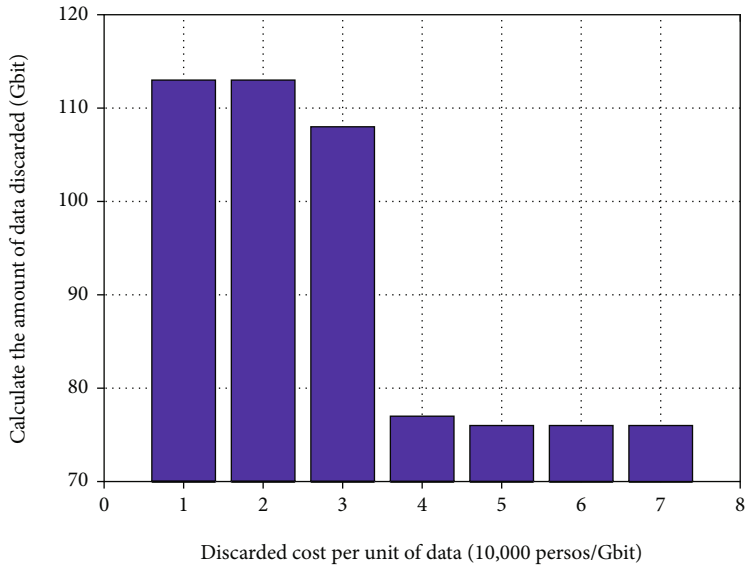


Figure 7: The amount of data discarded under different unit data discarding costs of the proposed algorithm.

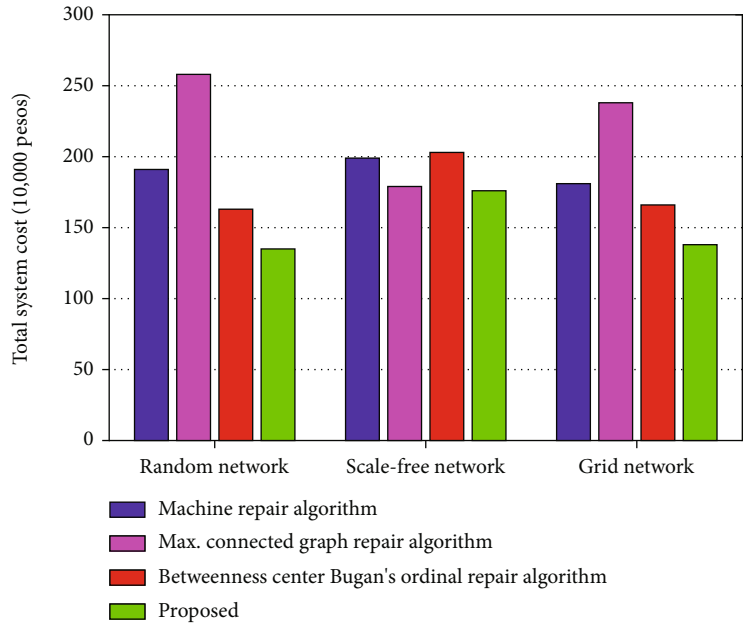


Figure 8: Comparison of algorithms under different network topologies with $N = 100$.

unit data discarding cost scenarios. Considering the double consideration of the total system overhead of the proposed algorithm for link repair cost and data discarding, with the increase of unit data discarding cost c_i, the number of repaired links and the amount of data discarding show an opposite increase or decrease relationship. Consistent with experience, as c_i increases, i.e., the cost of data discarding increases, the network tends to repair more links so that the computational demands can be met. When the value of c_i is (1, 1.5) million/Gbit, the amount of data discarded in Figure 7 begins to drop significantly, because within this range, the system is just near the sensitive critical point. Around this value, the system has the strongest ability to balance the cost of repairing with the overhead of discarding data. Any small change in the c_i value may bring about a larger change in equilibrium decision-making. After the c_i exceeds 20,000 pesos/Gbit, the marginal benefit brought by continuing to repair damaged links is reduced, and the number of network repaired links and the amount of computing data discarded remain stable. It can be seen that the proposed algorithm can be well applied to a variety of different data discarding cost scenarios and has good scalability and adaptability.

In order to further analyze the performance of the algorithm in this paper in different network topology scenarios, this paper extends the network form from a single random network scenario to a grid network and a scale-free network [36]. Different from the random network, the degree of each node of the grid network is exactly the same, and the degree of the nodes of the scale-free network obeys a power-law distribution. As can be seen from Figure 8, under the three types of networks, the performance of each algorithm will fluctuate to varying degrees, but the overall system overhead performance of the algorithm in this paper is better than the benchmark algorithm. Thanks to the joint analysis of network topology and dynamic features, the algorithm in this paper has good adaptability to multiple scenarios.

In summary, it can be seen from the simulation results that the proposed method outperforms the existing methods in various parameters evaluation and different scenarios. This provides solid basis that the proposed approach has practical value in IIoT deployment.

6. Conclusion

Aiming at the vulnerable characteristics of edge computing networks in the IIoT, this paper proposes a repair mechanism after large-scale damage to the network and gives a joint optimization method of priority repair link set decision and computing migration configuration, which effectively alleviates the problem of the initial stage of network repair; there is a conflict between limited repair resources and a large amount of data computing requirements. Considering the difficulty of solving the original problem, the Benders decomposition algorithm is used to transform it into a main problem and subproblems that are easier to solve. Furthermore, combined with the local branching method, a Benders decomposition acceleration algorithm is designed, which effectively improves the convergence speed of the algorithm. The simulation results show that the proposed algorithm has better system repair performance compared to the existing topology-based repair algorithms.

Although for the sake of data fairness, this paper considers that the computing data has the same processing priority, that is, the cost per unit of data discarded by different nodes is the same. The subproblem solution can be replaced with the existing minimum cost flow algorithm (such as Dinic's algorithm).

The proposed algorithm is mainly aimed at the situation where the nodes are connected by wired links. For scenarios with dynamic wireless links, such as the scenario where the UAV acts as a mobile edge node, the channel capacity changes with the location of the UAV, and issues such as the UAV's trajectory and wireless spectrum allocation need to be further considered jointly. This will make the already complex network repair problem more difficult to solve, which is beyond the scope of this paper and is reserved for follow-up research.

Appendix

A.1. Benders Optimal Cut Proof

When $e_{ij}^t, ij \in E^0$ is given, from the original problem P, we can get:

$$\theta\left(e_{ij}^t\right) = \min_{d,f,p}\left\{\sum_{i\in N} c_i d_i \mid (1) \text{ to } (3), (5) \text{ to } (7)\right\}. \tag{A.1}$$

The objective function of the equivalent dual problem of the above problem can be written as:

$$\theta\left(e_{ij}^t\right) = \max_{\alpha,\beta,\lambda,\mu}\left\{\sum_{i\in N}(\alpha_i r_i + \beta_i \bar{p}_i) + \sum_{ij\in E^1}\lambda_{ij}\bar{f}_{ij} + \sum_{ij\in E^0}\mu_{ij}\bar{f}_{ij}e_{ij}^t\right\} \tag{A.2}$$

Among them, $\alpha = (\alpha_i, \forall i \in N), \beta = (\beta_i, \forall i \in N), \lambda = (\lambda_{ij}, \forall i\, j \in E^1)$ optimizes variables (dual variables) for the four sets of dual problems. In the tth iteration process, it is assumed that the solution of the above dual variable set is $\alpha = (\alpha_i^t, \forall i \in N), \beta^t = (\beta_i^t, \forall i \in N), \lambda^t = (\lambda_{ij}^t, \forall ij \in E^1)$, the optimal solution of the problem $\theta^t = \theta(e_{ij}^t)$, and then we have

$$\theta(e_{ij}) \geq \sum_{i\in N}\left(\alpha_i^t r_i + \beta_i^t \bar{p}_i\right) + \sum_{iij\in E^1}\lambda_{ij}^t \bar{f}_{ij} + \sum_{ij\in E^0}\mu_{ij}^t \bar{f}_{ij}e_{ij}. \tag{A.3}$$

From linearization operation around $e_{ij}^t, ij \in E^0$, we get:

$$\begin{aligned}\eta \geq \theta(e_{ij}) &\geq \sum_{i\in N}\left(\alpha_i^t r_i + \beta_i^t \bar{p}_i\right) + \sum_{iij\in E^1}\lambda_{ij}^t \bar{f}_{ij} + \sum_{ij\in E^0}\mu_{ij}^t \bar{f}_{ij}e_{ij} \\ &+ \sum_{ij\in E^0}\mu_{ij}^t \bar{f}_{ij}\left(e_{ij} - e_{ij}^t\right) = \theta^t + \sum_{ij\in E^0}\mu_{ij}^t \bar{f}_{ij}\left(e_{ij} - e_{ij}^t\right).\end{aligned} \tag{A.4}$$

Data Availability

The data used for the findings of this study is available upon from the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest to report regarding the present study.

Authors' Contributions

Conceptualization was contributed by Hongyang Huang and Imran Khan; data curation was contributed by M. Dauwed and M. Derbali; formal analysis was contributed by Sun Li and Kai Chen; funding acquisition was performed by Sangsoon Lim; supervision was performed by Imran Khan and Sangsoon Lim.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea Government (MSIT) (No. 2021R1F1A1063319).

References

[1] J. Lu, L. Chen, J. Xia et al., "Analytical offloading design for mobile edge computing-based smart internet of vehicle," *EURASIP Journal on Advances in Signal Processing*, vol. 2022, no. 1, p. 10, 2022.

[2] L. Zhang, W. Zhou, J. Xia et al., "DQN-based mobile edge computing for smart internet of vehicle," *EURASIP Journal on Advances in Signal Processing*, vol. 3, 10 pages, 2022.

[3] C. Feng, P. Han, X. Zhang, B. Yang, Y. Liu, and L. Guo, "Computation offloading in mobile edge computing networks: a survey," *Journal of Network and Computer Applications*, vol. 202, no. 5, article 103366, 2022.

[4] L. Xing, "Cascading failures in internet of things: review and perspectives on reliability and resilience," *IEEE Internet of Things Journal*, vol. 8, no. 1, pp. 44–64, 2021.

[5] S. Ayoubi, C. Assi, Y. Chen, T. Khalifa, and K. B. Shaban, "Restoration methods for cloud multicast virtual networks," *Journal of Network and Computer Applications*, vol. 78, no. 3, pp. 180–190, 2017.

[6] D. Satria, D. Park, and M. Jo, "Recovery for overloaded mobile edge computing," *Future Generation Compute Systems*, vol. 70, no. 8, pp. 138–147, 2017.

[7] R. Teng, H. Li, and R. Miura, "Dynamic recovery of wireless multi-hop infrastructure with the autonomous mobile base station," *IEEE Access*, vol. 4, pp. 627–638, 2016.

[8] P. Li and X. Yang, "On dynamic recovery of cloud storage system under advanced persistent threats," *IEEE Access*, vol. 7, pp. 103556–103569, 2019.

[9] G. J. Baxter, G. Timár, and J. F. F. Mendes, "Targeted damage to interdependent networks," *Physical Review E*, vol. 98, no. 3, pp. 3328–3339, 2018.

[10] C. D. Brummitt, R. M. D'Souza, and E. A. Leicht, "Suppressing cascades of load in interdependent networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 109, no. 12, pp. 680–689, 2012.

[11] F. Morone and H. Makse, "Influence maximization in complex networks through optimal percolation," *Nature*, vol. 524, no. 7563, pp. 65–68, 2015.

[12] H. Rudnick, S. Mocarquer, E. Andrade, E. Vuchetich, and P. Miquel, "Disaster management," *IEEE Power and Energy Magazine*, vol. 9, no. 2, pp. 37–45, 2011.

[13] G. Punzo, A. Tewari, E. Butans et al., "Engineering resilient complex systems: the necessary shift toward complexity science," *IEEE Systems Journal*, vol. 14, no. 3, pp. 3865–3874, 2020.

[14] J. Moon, M. Yang, and J. Jeong, "A novel approach to the job shop scheduling problem based on the deep Q-network in a cooperative multi-access edge computing ecosystems," *Sensors*, vol. 21, no. 4, pp. 1–17, 2021.

[15] M. Oudani, "A simulated annealing algorithm for intermodal transportation on incomplete networks," *Applied Sciences*, vol. 11, no. 10, article 4467, 2021.

[16] A. Hamed, M. Alkinani, and M. Hassan, "A genetic algorithm to solve capacity assignment problem in a flow network," *Computers, Materials & Continua*, vol. 64, no. 3, pp. 1579–1586, 2020.

[17] A. Ranjbari, M. Hickman, and Y. Chiu, "A network design problem formulation and solution procedure for intercity transit services," *Transportmetrica A: Transport Science*, vol. 16, no. 3, pp. 1156–1175, 2017.

[18] D. Li, Q. Zhang, E. Zio, S. Havlin, and R. Kang, "Network reliability analysis based on percolation theory," *Reliability Engineering & Systems Safety*, vol. 142, no. 5, pp. 556–562, 2015.

[19] X. Lyu, C. Ren, W. Ni, H. Tian, and R. P. Liu, "Distributed optimization of collaborative regions in large-scale inhomogeneous fog computing," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 3, pp. 574–586, 2018.

[20] A. Smith, M. Posfai, and M. Rohden, "Competitive percolation strategies for network recovery," *Scientific Reports*, vol. 9, no. 1, pp. 965–983, 2019.

[21] Z. Hong, W. Chen, W. Huang, S. Guo, and Z. Zheng, "Multi-Hop cooperative computation offloading for industrial IoT-edge-cloud computing environments," *IEEE Transactions on Parallel and Distributed Systems*, vol. 30, no. 12, pp. 2759–2774, 2019.

[22] Y. Yu, X. Bu, K. Yang, Z. Wu, and Z. Han, "Green large-scale fog computing resource allocation using joint benders decomposition, Dinkelbach algorithm, ADMM, and branch-and-bound," *IEEE Internet of Things Journal*, vol. 6, no. 3, pp. 4106–4117, 2019.

[23] X. Li, J. Wan, H. Dai, M. Imran, M. Xia, and A. Celesti, "A hybrid computing solution and resource scheduling strategy for edge computing in smart manufacturing," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 7, pp. 4225–4234, 2019.

[24] Z. H. Abbas, Z. Ali, G. Abbas et al., "Computational offloading in mobile edge with comprehensive and energy efficient cost function: a deep learning approach," *Sensors*, vol. 21, no. 10, article 3523, 2021.

[25] M. Avgeris, D. Spatharakis, D. Dechouniotis, A. Leivadeas, V. Karyotis, and S. Papavassiliou, "ENERDGE: distributed energy-aware resource allocation at the edge," *Sensors*, vol. 22, no. 2, p. 660, 2022.

[26] G. Codato and M. Fischetti, "Combinatorial Benders' cuts for mixed-integer linear programming," *Operations Research*, vol. 54, no. 4, pp. 756–766, 2006.

[27] R. Rahmaniani, T. Crainic, M. Gendreau, and W. Rei, "The Benders decomposition algorithm: a literature review," *European Journal of Operational Research*, vol. 259, no. 3, pp. 801–817, 2017.

[28] J. Tanveer, A. Haider, R. Ali, and A. Kim, "An overview of reinforcement learning algorithms for handover management in 5G ultra-dense small cell networks," *Applied Sciences*, vol. 21, no. 1, pp. 1–14, 2022.

[29] M. F. Pereira, L. V. G. Pinto, S. F. Cunha, and G. Oliveira, "A decomposition approach to automated generation/transmission expansion planning," *IEEE Transactions on Power Apparatus and Systems*, vol. PAS-104, no. 11, pp. 3074–3083, 1985.

[30] S. Latif, M. Driss, W. Boulila et al., "Deep learning for the industrial internet of things (IIoT): a comprehensive survey of techniques, implementation frameworks, potential applications, and future directions," *Sensors*, vol. 21, no. 22, article 7518, 2021.

[31] F. Oliveira, I. Grossmann, and S. Hamacher, "Accelerating Benders stochastic decomposition for the optimization under

uncertainty of the petroleum product supply chain," *Computers & Operations Research*, vol. 49, no. 6, pp. 47–58, 2014.

[32] E. Vega, R. Soto, B. Crawford, J. Peña, and C. Castro, "A learning-based hybrid framework for dynamic balancing of exploration-exploitation: combining regression analysis and metaheuristics," *Mathematics*, vol. 9, no. 16, article 1976, 2021.

[33] M. Fischetti and A. Lodi, "Local branching," *Mathematical Programming*, vol. 98, no. 1-3, pp. 23–47, 2003.

[34] T. Santoso, S. Ahmed, M. Goetschalckx, and A. Shapiro, "A stochastic programming approach for supply chain network design under uncertainty," *European Journal of Operational Research*, vol. 167, no. 1, pp. 96–115, 2005.

[35] M.-X. Lu, G.-Z. Du, and Z.-F. Li, "Multimode gesture recognition algorithm based on convolutional long short-term memory network," *Computational Intelligence and Neuroscience*, vol. 2022, Article ID 4068414, 9 pages, 2022.

[36] A. Yu, N. Wang, and N. Wu, "Scale-free networks: characteristics of the time-variant robustness and vulnerability," *IEEE Systems Journal*, vol. 3, no. 99, pp. 1–11, 2020.

Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 5659979, 12 pages
https://doi.org/10.1155/2022/5659979

Research Article

Integrated Classification Algorithm for Unbalanced Data Streams Based on Joint Nonnegative Matrix Factorization

Jin Li and Ruibo Zhao

Tencent Technology Company Limited, Beijing 100086, China

Correspondence should be addressed to Ruibo Zhao; zhaor3@cuc.edu.cn

Received 14 February 2022; Revised 15 March 2022; Accepted 13 April 2022; Published 13 June 2022

Academic Editor: Lisheng Fan

The purpose of this paper is to study the unbalanced data flow integration classification algorithm based on joint nonnegative matrix factorization, in order to solve the problem that the basic clustering results obtained from the original data set have some information loss, thereby reducing the effective information in the integration stage. In this paper, the accuracy of the unbalanced data and the detection time consumption are selected as the research object. Six data sets with imbalanced proportions of minority and majority samples are selected for experiments. Mathematical statistical analysis is first used to observe text classification, disease diagnosis, and network intrusion detection and the classification accuracy of majority class and minority class; the commonly used algorithm for unbalanced data is statistical analysis method. Comparing the univariate method for comprehensive classification of unbalanced data flow based on nonnegative matrix factorization with the unbalanced data algorithm, the observation has accurate rate and detects time-consuming changes. Among them, the comprehensive classification algorithm of unbalanced data flow is based on the classification of data, classifying the data, judging whether two data points belong to the same category, and determining their degree of balance. The research data shows that the unbalanced data flow integrated classification algorithm based on joint nonnegative matrix decomposition can reasonably evaluate the classification performance of the classifier for a few classes, and the detection speed is faster and saves more time. The experimental research shows that the algorithm combines the relationship matrix and information matrix from the original data set into a consensus function, uses NMF technology to obtain the membership matrix, effectively uses potential information, improves the accuracy rate of 69.73%, and shortens 71.65% of the time consumed.

1. Introduction

The number is huge, and the dynamically changing incoming data is called the data stream. The classification of data streams is widely used for e-commerce and real-time monitoring of sensor networks and networks. However, the distribution by class in these applications is often uneven. This kind of data flow characterized by unbalanced distribution is called unbalanced. The data related to the unbalanced distribution of these categories gives traditional data extraction and classification algorithms and even poses serious problems for the existing data flow classification. Unbalanced mixed data processing is an important application in real life, especially in medical treatment, transportation, fault handling, and so on. Therefore, using various classification algorithms to process unbalanced mixed data has become an important research content in data mining.

With the rapid development of information technology, these data contain a lot of information in a series of application fields (such as wireless sensor networks, real-time traffic systems, network traffic monitoring, and credit card fraud detection), which prompt us to mine urgently. At present, there are few classification algorithms that can process unbalanced mixed data at the same time. This paper explores the comprehensive classification algorithm of unbalanced data flow based on joint nonnegative matrix factorization, in order to provide a significant contribution to data mining research.

Lu and Miao's decomposition of data into a small number of basic components is usually an effective strategy for data exploration, analysis, and interpretation [1]. Various working methods, such as principal component analysis (PCA) and nonnegative matrix factorization (NMF), are developed along this line of thought. These methods impose different constraints (e.g., orthogonality of PCA) to obtain compact or physically meaningful basis. Ying-Ying et al. discuss the molecular typing and prognosis prediction of gastric cancer based on nonnegative matrix factorization (NMF) [2]. The gene expression spectrum (GEO) was detected in patients with gastric cancer. The expression profile of INc RNA was analyzed using INC RNA mining method. The NMF model was established using consistent clustering +software package.

In order to improve the performance of traditional data stream integration algorithms in big data mining applications, a parallel data level integration algorithm was designed and implemented with the help of cloud computing-related technologies and nonnegative matrix factorization methods [3]. The purpose of this paper is to study the unbalanced data flow integration classification algorithm based on joint nonnegative matrix factorization, in order to solve the problem that the basic clustering results obtained from the original data set have some information loss, thereby reducing the effective information in the integration stage.

2. Programs Method

2.1. Basic Content of Nonnegative Matrix Factorization. Nonnegative matrix factorization makes all components after factorization nonnegative (requiring a purely additive description) and at the same time achieves nonlinear dimensionality reduction. This nonnegativity restriction leads to a certain degree of sparsity in the corresponding descriptions, and sparsity representations have been shown to be an efficient form of data descriptions between fully distributed descriptions and those of a single active component. Nonnegative matrix factorization (NMF) method has been widely used in multidimensional data similarity data clustering, text clustering, and social network clustering, but its serial calculation is the most difficult. The time is for big data processing operations [4, 5]. Previously, in the field of parallel data cluster processing for multidimensional data parallelization, there were cluster computer and shared memory computing methods, as well as grid computing, peer-to-peer computing, and widely distributed computing model spectra, all with excellent results. However, in the era of cloud computing, predistributed distributed computing models used for large amounts of PB often appear to be insufficient, so proper attention should be paid to cloud-based data classification groups. Optimization of traditional data aggregation methods is based on nonnegative matrix factorization. And NMF has gradually become one of the most popular multidimensional data processing tools in research fields such as signal processing, biomedical engineering, pattern recognition, computer vision, and image engineering.

(1) Unbalanced data stream integrated classification algorithm and nonnegative matrix factorization

The so-called unbalanced data refers to the fact that there are more samples of some classes than other classes in a data set. The class with more samples is generally called the majority class, and the class with fewer samples is called the minority class. The unbalanced data flow integrated classification algorithm is based on the classification of the data to classify the data, whether the two data points belong to the same category, and determine how balanced they are [6–8]. When the balance between them is greater than a certain value, they belong to the same cluster; otherwise, the two data points belong to different clusters. However, there are still some difficulties. Due to the serious skew in quantity, the performance of classification algorithms for classifying unbalanced data sets is not satisfactory. Because minority class samples are usually more difficult to identify than common samples, most data mining classification algorithms have great difficulty in dealing with minority class samples.

There are large-scale data in the presence of practical problems, which makes the matrix that stores these large data very large, and the stored information is unevenly distributed, which means that existing methods cannot process the data contained in the matrix efficiently and quickly [9]. In order to better deal with such data, the effective method category is the decomposition of the matrix, which can greatly reduce the size of the description problem and can compress and summarize the data. To this end, there are many methods of factorization matrix, such as the decomposition of exogenous values, the analysis of independent components, and the analysis of principal components [10]. Matrix factorization is a method of reducing a matrix to its constituent parts. This approach simplifies more complex matrix operations that can be performed on the decomposed matrix rather than the original matrix itself. The decomposition results obtained from cluster analysis are based on the decomposition of nonnegative matrices, which can ensure that its elements are not negative and represent their actual physical meaning. Therefore, in recent years, they have become the object of special attention.

The clustering method based on nonnegative matrix factorization NMF is as follows: considering that the data set can be represented as a vector set $X = \{X_1, X_2, \cdots, X_n\}$, and each vector represents the m-dimensional data point $X_j = \{x_{j1}, x_{j2}, \cdots, x_{\mathrm{jm}}\}$, NMF method is to divide X into two nonnegative low-rank matrices W and H, which can be achieved by optimizing the following formula as much as possible:

$$\min_{x>0} \left\| G - \hat{X}\hat{X}^T \right\|_F^2, \tag{1}$$

where $X^{\wedge}$ can be obtained by the following multiplication update rule:

$$\hat{X}_{ik} \longleftarrow \hat{X}_{ik}\left[\frac{1}{2} + \frac{\left(G\hat{X}\right)_{ik}}{\left(2\hat{X}\hat{X}^{T}\hat{X}\right)_{ik}}\right]. \tag{2}$$

(2) After decomposition, NMF can retain more information reflected by the original sample. The result obtained after decomposition is nonnegative and has good physical meaning, and the implementation process is simple and fast. As the name implies, NMF decomposes a nonnegative matrix into two nonnegative matrices, and the result of multiplying these two matrices is equal to the original matrix before decomposition [11]. The objective function is shown in

$$\min \|X - WH\|_F^2 \tag{3}$$

Among them, the nonnegative data set $X \in R^{m*n}n$ is the original matrix, $X_{m*n} = (x_1, x_2, \cdots, x_n)$, and x_i represents an m -dimensional column vector; that is, the information of a sample [12]. The basis matrix $W \in R^{m*r}$, $W_{m*r} = (w_1, w_2, \cdots, w_r)$, and w_i represents an m-dimensional column vector, representing a basis vector. The coefficient matrix $H \in R^{r*n}$; $H_{r*n} = (h_1, h_2, \cdots, h_n)$, where h_i is the column vector of r dimension, which can be regarded as the coordinates of the projection of the x_i vector in the new space defined by the W-based matrix, satisfying $x_i = W * h_i$, where x_i is the projection coefficient. r satisfies the condition $(m + n) * r < m * n$, that is to decompose a high-dimensional nonnegative matrix into the product of two low-rank nonnegative matrices. The iteration rules are shown in equations (4) and (5); Θ represents the Hada code program.

$$W \longleftarrow W\Theta\frac{\left(XH^{T}\right)_{ij}}{\left(WHH^{T}\right)_{ij}}, \tag{4}$$

$$H \longleftarrow H\Theta\frac{\left(W^{T}X\right)_{ij}}{\left(W^{T}WX\right)_{ij}}. \tag{5}$$

In the NMF iteration process, the base matrix has no constraints, and there is a lot of redundancy between the data [13]. Therefore, in recent years, many improved algorithms for NMF have been proposed.

(3) Joint nonnegative matrix initialization method

As with other models based on iterative optimization, since the local minimum is not unique, the result of nonnegative matrix decomposition is usually more sensitive to the initial value of the factor matrix, which means that the initialization of W and H will affect the convergence speed and final result of the algorithm [14–16]. For the sake of simplicity, random initialization is often used to assign initial values to W and H in many studies, which often makes the algorithm's convergence rate slower. To this end, some researchers have proposed some other methods to initialize NMF. Common initialization methods include the following categories:

(1) Multiple initialization

The core idea of this type of method is to perform multiple random initializations on the factor matrix, run the NMF algorithm once for each initialization, and then select the best estimate as the final decomposition result [17]. Due to the need to perform NMF decomposition multiple times, the computational overhead of such algorithms is often relatively large.

(2) Initialization based on matrix factorization

Nonnegative matrix factorization is actually a low-rank factorization technique with constraints, so we can use the results of other low-rank factorization algorithms as NMF initialization. Typical examples include SVD-based initialization and CUR decomposition-based initialization.

(3) Cluster-based initialization

Based on the characteristics of nonnegative matrix factorization, we can regard nonnegative matrix factorization as a clustering process, so the results of other clustering algorithms, such as k-means and fuzzy clustering, are used as NMF initialization. Compared with the initialization method based on matrix decomposition, using this kind of method as the preprocessing process is often too complicated and may cause the algorithm to terminate at a poor local solution. In practical applications, the selection of NMF initialization methods cannot be generalized, and the initialization method that is effective for one data set may not be applicable to another data set. Therefore, it is often necessary to select a suitable initialization method based on practical problems and certain prior knowledge.

(4) Common nonnegative matrix factorization constraints

(a) Sparseness constraint

The sparsity constraint helps to improve the uniqueness of the calculation results of nonnegative matrix factorization, and at the same time, it helps to strengthen the characteristics based on the partial representation. If W is regarded as a base matrix and H is regarded as a coefficient matrix, then applying a sparsity constraint to each column of W will make each base vector only affect a small part of the original observations: column sparsity constraints; then, each observation will only be represented by a linear combination of a few base vectors; and if sparsity constraints are imposed on each row of H, then each base vector will only be used to approximate some of the training data, or it can be understood that each basis vector is derived from part of the training data, which has a strong correlation with clustering.

(b) Orthogonality constraint

The addition of orthogonality constraints in nonnegative matrix factorization is to minimize the redundancy between basis vectors [18]. If the orthogonality constraint is applied to each column of W, that is, $W^TW = I$, then it will make the basis vectors have the greatest discrimination; and the orthogonality constraint is applied to each row of H, that is, $VV^T = I$, which will improve the accuracy of clustering. It is worth noting that applying orthogonality constraints to W and H, respectively, is actually equivalent to clustering the rows and columns of the input data matrix, respectively. If one factor matrix in NMF is regarded as a clustering center, the other is equivalent to an indicator vector. Therefore, orthogonality constraints have also been applied in clustering research.

(c) Discriminant constraints

From the perspective of pattern recognition, the traditional NMF algorithm can be regarded as an unsupervised learning process. By combining discriminative information and decomposition process, the basic NMF algorithm can be extended to supervised learning, and the model generation and classification tasks can be integrated into a framework. This method has been successfully applied to classification applications such as face recognition and expression recognition.

2.2. Components of an Integrated Classification Algorithm for Unbalanced Data Streams

2.2.1. Characteristics of Unbalanced Data. The data imbalance problem mainly exists in supervised machine learning tasks. When encountering unbalanced data, traditional classification algorithms with overall classification accuracy as the learning objective will pay too much attention to the majority class, thus degrading the classification performance of minority class samples. Unbalanced data is mainly composed of two types of interclass imbalance and intraclass imbalance [19, 20]. The imbalance between classes leads to uneven data distribution between classes, as shown in Figures 1(a) and 1(b). In some practical applications, the data shows that the data between the classes is extremely unbalanced, and the unbalance rate can reach 1000: 1 or greater in some cases. The imbalance in a category refers to the imbalance in the sample size of a category and its subcategories, or the data of a category has multiple different terms that are not so important, as shown in Figures 1(c) and 1(d). A large number of studies have shown that the imbalance of data between categories is not the only factor affecting classification learning, and the imbalance of data within categories is a key factor affecting the effect of classification [21]. Therefore, the classification problem of unbalanced data is mainly due to the complexity of the data distribution, as shown in Figures 1(b) and 1(c) (data overlap) and Figure 1(d) (small fragmentation problem); all of these problems will directly affect the classifier's learning result. Unbalanced data scenarios appear in all aspects of Internet applications, such as click prediction of search engines (clicked web pages often occupy a small proportion), product recommendation in the field of e-commerce (the proportion of recommended products being purchased is very low), credit card fraud detection, network attack identification, and cancer detection.

2.2.2. Integrated Classification Technology for Unbalanced Data. The integrated classification algorithm was aimed at improving the accuracy of the overall learning and cannot be directly used to deal with the classification learning of unbalanced data. Based on the currently available results, the imbalanced data can be classified through integrated algorithms at two levels: algorithm or data [22, 23]. Algorithm processing includes introducing a cost factor in the formation process of the comprehensive classification algorithm. According to whether the cost of heterogeneous samples is different from the cost of incorrect classification, the different cost factors are attributed to the cost factor to form an integrated cost-sensitive type classification algorithm. Since the AdaBoost algorithm is a series of trainings for different basic classifiers, they are obtained by changing the weight values of the training samples, so an integrated cost-sensitive classification algorithm is usually introduced to form a cost factor in the update. The legal values of the training champion. Data processing refers to the technique of rebalancing the data sampling during the establishment of the basic classifier, so that the integrated algorithm can build the classifier on the balanced training data that does not affect the learning performance. The combination of different data balancing strategies for resampling and integrated classification algorithms has led to integrated classification algorithms based on data processing boosting, integrated classification based on data processing bagging, and integrated classification based on mixed data processing.

2.2.3. Unbalanced Data Classification Algorithm Evaluation System. The classification of the decomposed data requires the classifier to achieve a high degree of classification accuracy for a limited number of classified samples without harming most of the classified samples. The evaluation criteria commonly used in learning machines are indicators of global classification accuracy and are not suitable for classification algorithms that evaluate classified data [24, 25]. Evaluation criteria that can provide more information are usually adopted in existing studies, such as single evaluation indicators based on confusion matrix, precision curve before recall, ROC curve, and cost curve.

2.2.4. Unbalanced Data Processing. Maintain the function of the original sample distribution; on the other hand, in order to make better use of most of the information in the sample (useful for the computer imbalance phase), first provide a random sample of the cyclic feature subset and then a small percentage relative to the number of these samples. The number of samples in each weight category is calculated comprehensively with the majority shared sample type. The base and composite weights for each sample category plus the number of sample types make up the percentage of sample formation. At last, the processed training sample

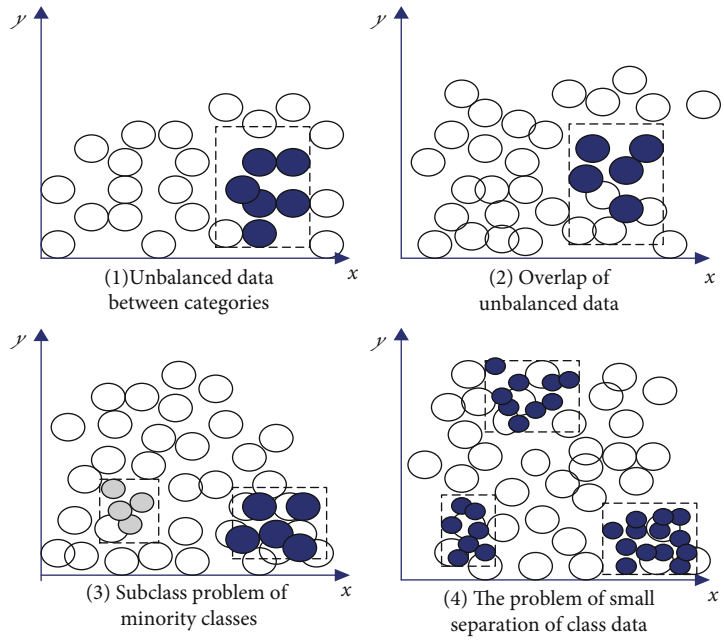


Figure 1: Characteristics of unbalanced data.

subset and feature subset are finally obtained. Among them, for the unbalanced data sampling method, the basic idea is to eliminate or reduce the imbalance of the data by changing the distribution of the training data. The specific process is shown in Figure 2. The way to deal with data imbalance is as follows: in the case of very few positive and negative samples, data synthesis should be used. In the case where there are enough negative samples and very few positive samples and the proportion is very disparate, the classification method should be considered. In the case where there are enough positive and negative samples and the proportions are not particularly disparate, sampling or weighting methods should be considered.

3. The Experiments

3.1. Experimental Data Set. This experiment selects 10 data from two sources: artificial data set and UCIE data set, of which 2d4c is a randomly generated artificial data set based on Gaussian distribution, and the rest are all from UCI's real data set, among which balance, heart, liver. They are the abbreviations of data set balance-scal, heart-statlog, liver disorders, and contraceptivc-method-choice. The relevant statistical information of all test data sets is listed in Table 1.

3.2. Experimental Design. Set the number of runs of the imbalanced data flow integrated classification algorithm $M = 10$, and combine the different result sets into an information matrix for experiments. Set the relationship matrix weight parameter δ to 0.0001, which is empirically obtained.

Compare the algorithm in this paper with the traditional algorithm, observe the use of the integrated nonnegative matrix decomposition-based unbalanced data stream integrated classification algorithm and common algorithm, and observe the classification of text classification, disease diagnosis, network intrusion detection, and the majority and minority categories rate. At the same time, compare the classification accuracy of the two algorithms in various scenes and the speed of the time consumption.

3.3. Evaluation Criteria. This experiment will use $F1$ and RI (Rand index) to evaluate the experimental results.

The definition of $F1$ is as follows:

$$\mathrm{F1} = \frac{2 * \mathrm{PR}}{\mathrm{P} + \mathrm{R}}, \tag{6}$$

where P is the precision rate, which represents the proportion of extracted correct objects in the extracted objects, and R is the recall rate, which represents the proportion of extracted correct objects in the samples.

The definition of RI is as follows:

$$\mathrm{RI}\left(\prod, \pi\right) = \frac{n_{11} + n_{00}}{n_{11} + n_{10} + n_{01} + n_{00}} \tag{7}$$

where II is the real data set, π is the clustering result label, n_{11} represents the number of data objects in a cluster in both the II and π sets, n_{01} represents the number of different clusters in the π set that are in the same cluster but in II, and the meanings of n_{00} and n_{10} are the same.

According to the above definition, the larger the values of $F1$ and RI, the better the clustering effect.

4. Discussion

4.1. Effectiveness of Using Integrated Nonmatrix Decomposition-Based Unbalanced Data Flow Integrated Classification Algorithm

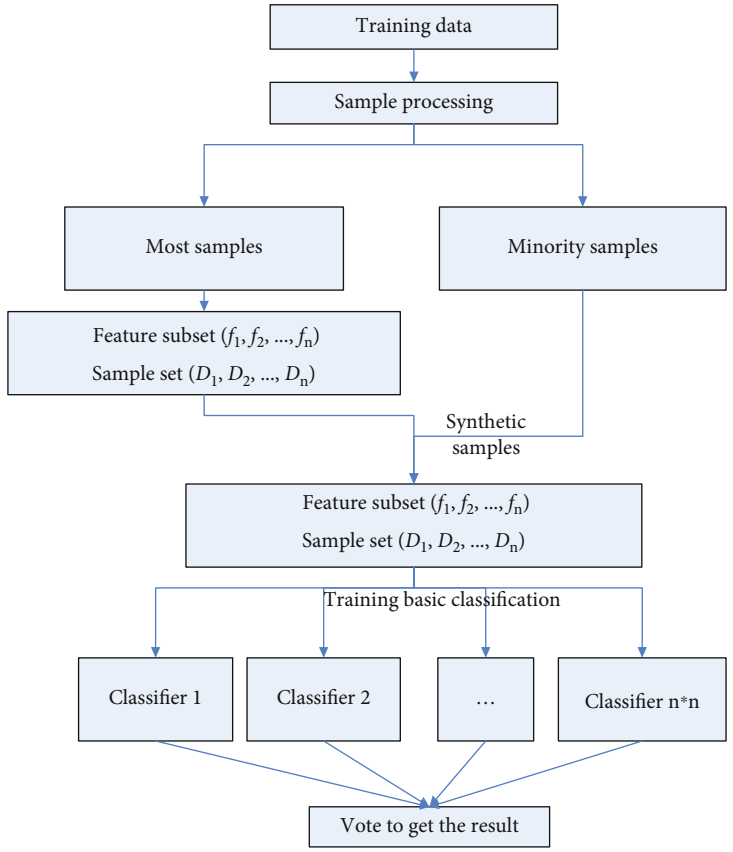


Figure 2: Method flow.

Table 1: Relevant information description of experimental test data set.

Data set	Number of samples	Attributes	Number of categories	Source
2d4c	200	2	4	Artificial
Wine	190	13	3	UCI
Iris	150	4	3	UCI
Glass	239	9	6	UCI
Segment	2319	19	7	UCI
Balance	572	2	3	UCI
Diabetes	721	9	2	UCI
Heart	281	13	3	UCI
Liver	273	6	2	UCI
Emc	1252	9	3	UCI

(1) In this experiment, the unbalanced data stream integrated classification algorithm based on joint nonnegative matrix decomposition is used to observe the classification accuracy of the majority and minority categories of text classification, disease diagnosis, and network intrusion detection. The data shows that after five tests, the text classification, disease diagnosis classification, and network intrusion detection classification have obtained obvious correct rates. The text classification accuracy rate is 78.45% on average, the disease diagnosis classification is 65.72% on average, the average detection classification of network intrusion is 83.23%. Based on the comparison of classification results based on joint nonnegative matrix factorization, only when the recall and precision rates are large, the nonnegative matrix factorization will be correspondingly large. Therefore, nonnegative matrix factorization can reasonably evaluate the classification performance of the classifier for minority classes. The data collection table is shown in Table 2 and Figure 3

Table 2: Effects of using nonnegative matrix factorization (unit: %).

Test	Text categorization	Disease diagnosis	Network intrusion detection
Test 1	78.34	63.34	80.34
Test 2	79.84	65.24	82.38
Test 3	77.65	64.92	84.23
Test 4	78.69	65.86	82.47
Test 5	78.82	64.38	83.23

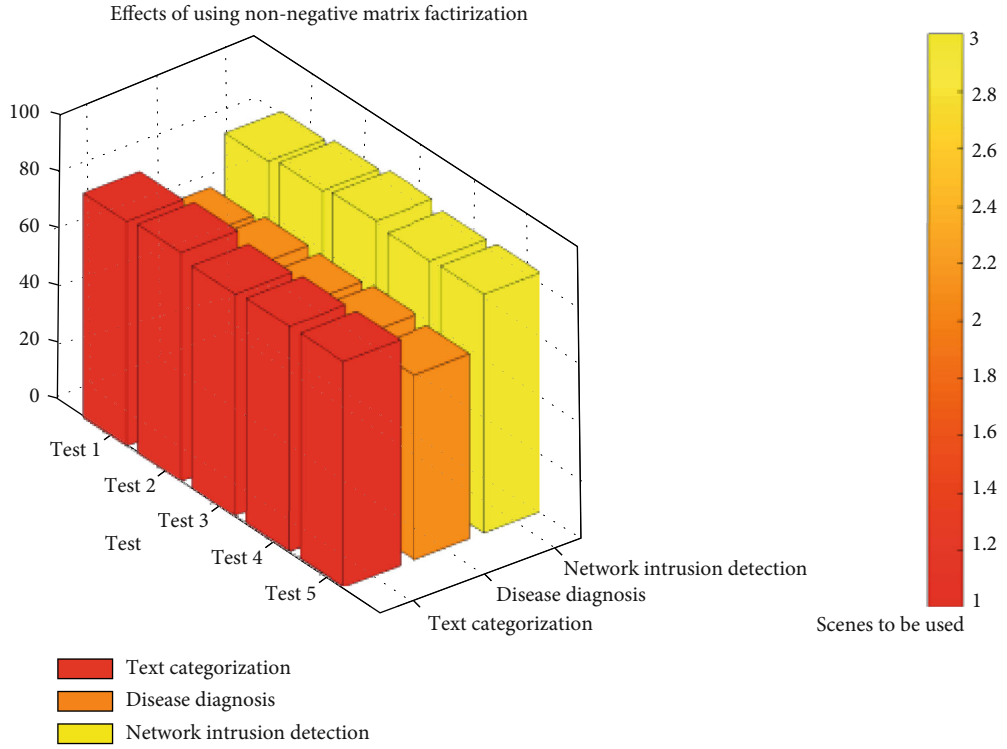


Figure 3: Effects of using nonnegative matrix factorization (unit: %).

(2) After using the general unbalanced data flow integration classification algorithm in this experiment, observe the classification accuracy of the majority and minority categories of text classification, disease diagnosis, and network intrusion detection. The data shows that after 5 tests, the text classification, disease diagnosis classification, and network intrusion detection classification can only get a lower correct rate. The average accuracy rate of the text classification is 22.45%, and the disease diagnosis classification is 24.64%. The average network intrusion detection classification is 19.54%. Using ordinary unbalanced data flow integrated classification algorithms, the recall and precision are small, and it is difficult to improve the accuracy of minority and majority classification. Therefore, the general unbalanced data flow integrated classification algorithm is not suitable for the classification of unbalanced data. The data collection table is shown in Table 3 and Figure 4

4.2. Convenience of Using Integrated Nonmatrix Decomposition-Based Unbalanced Data Flow Integration Classification Algorithm

(1) In this experiment, the unbalanced data flow integrated classification algorithm based on joint nonnegative matrix decomposition and the general unbalanced data flow integrated classification algorithm are used to classify the majority and minority categories in text classification, disease diagnosis, and network intrusion detection. The data shows that after five tests, the text classification based on the unbalanced data flow integrated classification algorithm under the joint nonnegative matrix decomposition, the disease diagnosis classification, and the network intrusion detection classification has obtained obvious correct rates; and based on the general classification, the accuracy of the

Table 3: Classification accuracy of common algorithms (unit: %).

Test	Text categorization	Disease diagnosis	Network intrusion detection
Test 1	20.12	25.45	19.45
Test 2	22.33	26.05	18.64
Test 3	23.74	25.84	19.56
Test 4	23.58	25.38	20.18
Test 5	21.83	23.12	20.23

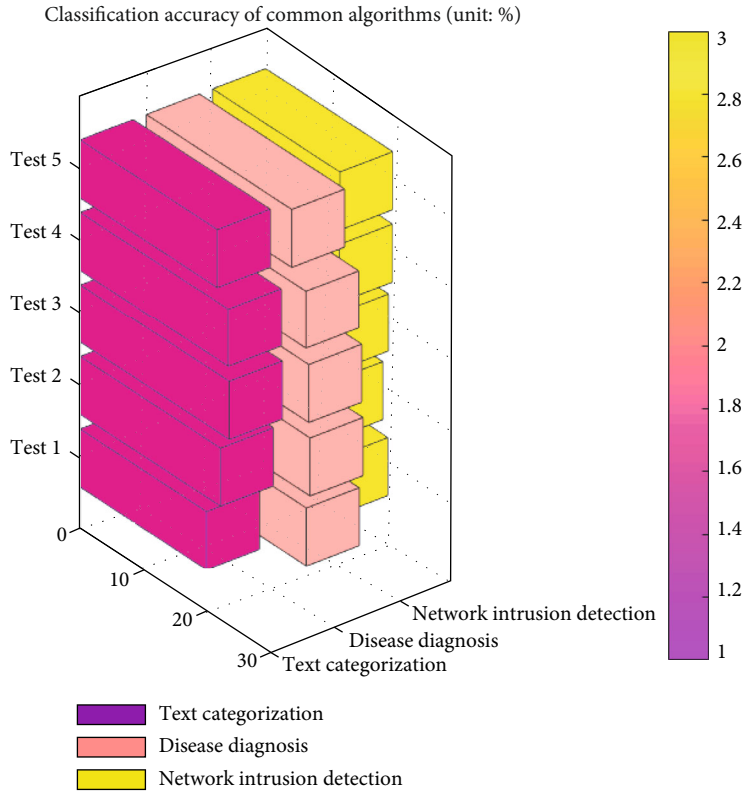


Figure 4: Classification accuracy of common algorithms (unit: %).

Table 4: The accuracy of the two algorithms compared (unit: %).

Test	Text categorization		Disease diagnosis		Network intrusion detection	
	Nonnegative matrix factorization	Ordinary decomposition	Nonnegative matrix factorization	Ordinary decomposition	Nonnegative matrix factorization	Ordinary decomposition
Test 1	78.34	20.12	63.34	25.45	80.34	19.45
Test 2	79.84	22.33	65.24	26.05	82.38	18.64
Test 3	77.65	23.74	64.92	25.84	84.23	19.56
Test 4	78.69	23.58	65.86	25.38	82.47	20.18
Test 5	78.82	21.83	64.38	23.12	83.23	20.23

unbalanced data obtained under the algorithm is very low, mainly because the recall and precision are small. Therefore, it can be seen that the unbalanced data flow integrated classification algorithm based on joint nonnegative matrix decomposition is more suitable for the data classification of

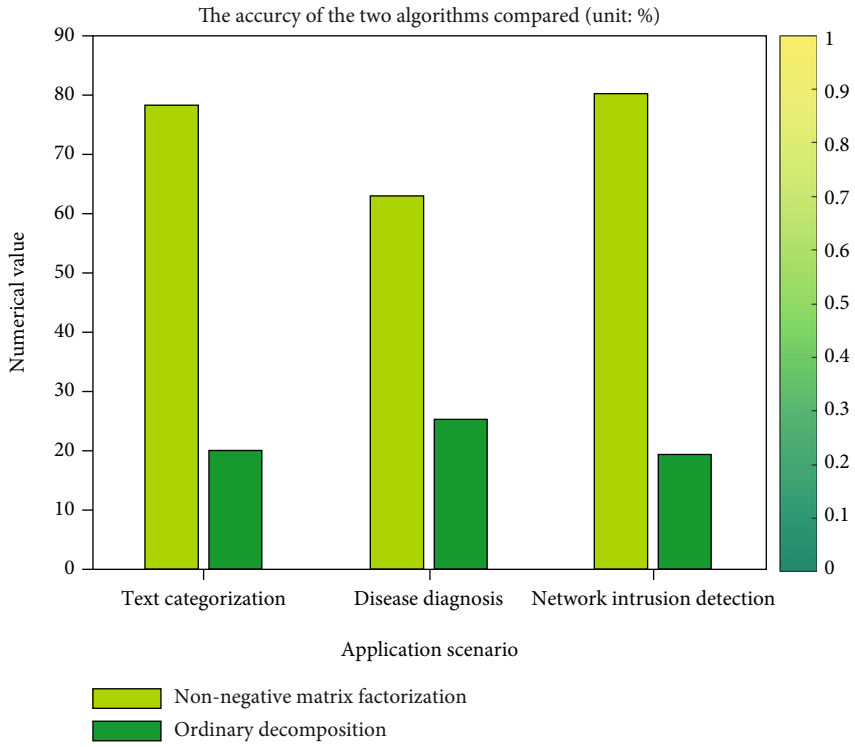


Figure 5: The accuracy of the two algorithms compared (unit: %).

Table 5: Comparison of the detection time consumption of the two algorithms (unit: h).

Test	Text categorization		Disease diagnosis		Network intrusion detection	
	Nonnegative matrix factorization	Ordinary decomposition	Nonnegative matrix factorization	Ordinary decomposition	Nonnegative matrix factorization	Ordinary decomposition
Test 1	1.50	5.3	1.84	5.6	1.32	4.84
Test 2	1.34	5.1	1.85	5.83	1.33	5.02
Test 3	1.24	5.22	1.78	5.89	1.23	4.99
Test 4	1.54	5.63	1.58	5.84	1.07	4.85
Test 5	1.50	5.23	1.63	6.09	1.13	5.12

unbalanced data. The data collection table is shown in Table 4 and Figure 5

(2) This experiment compares the time between the majority class and the minority class in text classification, disease diagnosis, and network intrusion detection using the unbalanced data flow integrated classification algorithm based on joint nonnegative matrix decomposition and the general unbalanced data flow integrated classification algorithm. Consume quickly. The data shows that after five tests, the text classification in the integrated classification algorithm based on unbalanced data flow under the joint nonnegative matrix decomposition, disease diagnosis classification, and network intrusion detection classification detection time consumption is shorter; and based on the general classification, under the algorithm, the detection time consumption of unbalanced data is shorter. The high accuracy of the integrated classification algorithm of unbalanced data flow based on joint nonnegative matrix decomposition reduces unnecessary errors, improves the consumption of detection time and speed, and obtains faster accurate classification results. The data collection table is shown in Table 5 and Figure 6

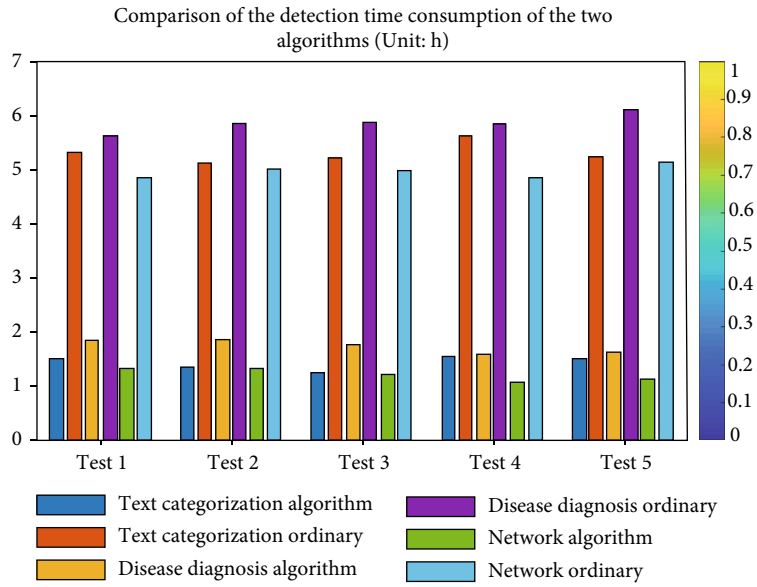


Figure 6: Comparison of the detection time consumption of the two algorithms (unit: h).

5. Conclusions

(1) In recent years, with the continuous deepening of research on class imbalance, the problem of class imbalance in data flow has attracted a large number of researchers. This paper proposes an ensemble classification framework to address the class imbalance problem in data streams. Its adaptive algorithm uses sampling techniques to deal with imbalance problems. The study led to the definition of a method for dealing with unbalanced data streams, which not only added examples of positive classes but also added classification errors in negative classes and also proposed a new method for defining class boundaries and negative to improve the integration effect of the classifier. An integrated classifier model for handling unbalanced data streams has been proposed, which combines weighted based integrated classifiers and sampling techniques. The data imbalance problem mainly exists in supervised machine learning tasks. When encountering unbalanced data, traditional classification algorithms with overall classification accuracy as the learning objective will pay too much attention to the majority class, thus degrading the classification performance of minority class samples. The vast majority of common machine learning algorithms do not work well with imbalanced data sets

(2) Classification is one of the main means of acquiring knowledge in the field of machine learning and data extraction. The most common classification algorithms, such as decision trees, tree networks, support vectors, and neural networks, have been used on a large scale. Existing classification algorithms usually assume that the data used for training is balanced; that is, the number of samples present in each type is approximately equal. In the case of imbalanced class data, the traditional classification algorithm (taking the accuracy of population classification as the learning target) pays too much attention to most classes, thereby reducing the ability to classify a few samples. But in fact, the cost of classification errors for a limited number of category samples is higher than most categories. For example, when predicting software defects, the size of defective samples is much smaller than the size of nondefective samples, but the purpose of classification is to identify a limited number of samples of defect categories. Other areas include medical diagnosis, oil spill control, cyber conspiracy control, and credit card fraud. The classification of unbalanced data is related to the performance of the learning algorithm when the class data is unbalanced or underexpressed. Based on the results of existing research, cost-sensitive techniques or sampling techniques can be used to reclassify data to solve the classification problem of classified data

(3) The classification learning of unbalanced data has a wide range of applications in many fields, such as software defect prediction and network intrusion detection. Due to the advantages of integration technology in dealing with unbalanced data learning, it is a research hotspot in the field of machine learning in recent years. The purpose of this paper is to study the unbalanced data flow integration classification algorithm based on joint nonnegative matrix factorization, in order to solve the problem that the basic clustering results obtained from the original data set have some information loss, thereby reducing the effective information in the integration stage. In this paper, the accuracy of the unbalanced data and the detection time consumption are selected as the

research object. Six data sets with imbalanced proportions of minority and majority samples are selected for experiments. Mathematical statistical analysis is first used to observe text classification, disease diagnosis, and network intrusion detection and the classification accuracy of majority class and minority class; the commonly used algorithm for unbalanced data is statistical analysis method. Comparing the univariate method for comprehensive classification of unbalanced data flow based on nonnegative matrix factorization with the unbalanced data algorithm, the observation has accurate rate and detects time-consuming changes. Experimental data shows that the unbalanced data flow integrated classification algorithm based on joint nonnegative matrix decomposition can reasonably evaluate the classification performance of the classifier for a few classes, and the detection speed is faster and saves more time. The experimental research shows that the algorithm combines the relationship matrix and information matrix from the original data set into a consensus function, uses NMF technology to obtain the membership matrix, effectively uses potential information, improves the accuracy rate of 69.73%, and shortens 71.65% of the time consumed. With the development of artificial intelligence deep learning, the advantages of deep network structure are becoming more and more obvious. In order to further study deep nonnegative matrix factorization, I think that with the research and development of deep nonnegative matrix factorization, more algorithms for optimizing deep model can be proposed, which will further improve the clustering performance of the model

Data Availability

No data were used to support this study.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this article.

References

[1] N. Lu and H. Miao, "Structure constrained nonnegative matrix factorization for pattern clustering and classification," *Neurocomputing*, vol. 171, pp. 400–411, 2016.

[2] C. Ying-Ying, Z. Xiao-Qiang, and C. Hao-Yan, "Case study of the molecular classification and prognostic prediction of gastric cancer based on nonnegative matrix factorization," *Journal of Shanghai Jiaotong University(Medical Science)*, vol. 37, no. 9, pp. 1187–1194, 2017.

[3] L. I. Xu, T. U. Ming, and W. Xiaofei, "Single-Channel speech separation based on non-negative matrix factorization and factorial conditional random field," *Acta Electronica Sinica*, vol. 27, no. 5, pp. 1063–1070, 2018.

[4] F. Segovia, J. M. Górriz, J. Ramírez, F. J. Martinez-Murcia, and M. García-Pérez, "Using deep neural networks along with dimensionality reduction techniques to assist the diagnosis of neurodegenerative disorders," *Logic Journal of the IGPL*, vol. 6, p. 6, 2018.

[5] F. Zhuang, P. Luo, and D. Changying, "Triplex transfer learning: exploiting both shared and distinct concepts for text classification," *IEEE Transactions on Cybernetics*, vol. 44, no. 7, pp. 1191–1203, 2014.

[6] T. Afzal, K. Iqbal, and G. White, "A method for locomotion mode identification using muscle synergies," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 6, pp. 1–1, 2016.

[7] W. Sun, M. Jiang, and W. Li, "Band selection using sparse self-representation for hyperspectral imagery," *Geomatics & Information Science of Wuhan University*, vol. 42, no. 4, pp. 441–448, 2017.

[8] H. Rajaguru and S. K. Prabhakar, "Variational Bayesian matrix factorization and certain post classifiers for classification of epilepsy from EEG signals," *Research Journal of Pharmacy & Technology*, vol. 9, no. 6, p. 750, 2016.

[9] C.-H. Yeh, C.-Y. Lin, K. Muchtar, and P.-H. Liu, "Rain streak removal based on non-negative matrix factorization," *Multimedia Tools & Applications*, vol. 77, no. 15, pp. 20001–20020, 2018.

[10] S.-S. Wang, A. Chern, Y. Tsao et al., "Wavelet speech enhancement based on nonnegative matrix factorization," *IEEE Signal Processing Letters*, vol. 23, no. 8, pp. 1101–1105, 2016.

[11] L. Sun, G. Zhao, and D. Xinpeng, "CUR based initialization strategy for non-negative matrix factorization in application to hyperspectral unmixing," *Journal of Applied Mathematics & Physics*, vol. 4, no. 4, pp. 614–617, 2016.

[12] L. Tong, J. Zhou, Y. Qian, X. Bai, and Y. Gao, "Nonnegative matrix factorization based hyperspectral unmixing with partially known endmembers," *IEEE Transactions on Geoscience & Remote Sensing*, vol. 54, no. 11, pp. 6531–6544, 2016.

[13] Z. Zhang and Y. Liu, "A list-wise matrix factorization based POI recommendation by fusing multi-tag, social and geographical influences," *Journal of Internet Technology*, vol. 19, no. 1, pp. 127–136, 2018.

[14] W. Pak and Y. J. Choi, "High performance and high scalable packet classification algorithm for network security systems," *IEEE Transactions on Dependable & Secure Computing*, vol. 14, no. 1, pp. 37–49, 2017.

[15] T. Kim, B. Do Chung, and J.-S. Lee, "Incorporating receiver operating characteristics into naive Bayes for unbalanced data classification," *Computing*, vol. 99, no. 3, pp. 1–16, 2016.

[16] M. Nakata and T. Hamagami, "Revisit of rule-deletion strategy for XCSAM classifier system on classification," *Transactions of the Institute of Systems Control & Information Engineers*, vol. 30, no. 7, pp. 273–285, 2017.

[17] A. Care, F. A. Ramponi, and M. C. Campi, "A new classification algorithm with guaranteed sensitivity and specificity for medical applications," *IEEE Control Systems Letters*, vol. 2, no. 3, pp. 393–398, 2018.

[18] L. F. Gao, S. J. Zhao, and Y. U. Dong-Mei, "Unbalanced support vector machine coupling negative-samples cutting with asymmetric misclassification cost," *Acta Electronica Sinica*, vol. 45, no. 12, pp. 2978–2986, 2017.

[19] B. Zou, X. Chen, and L. Yang, "k-Times Markov sampling for SVMC," *IEEE Transactions on Neural Networks & Learning Systems*, vol. 29, no. 4, pp. 1328–1341, 2018.

[20] S. K. Mishra, S. C. Swain, and L. N. Tripathy, "Fault detection & classification in UPFC integrated transmission line using DWT," *International Journal of Power Electronics & Drive Systems*, vol. 8, no. 4, pp. 1793–1803, 2017.

[21] S. K. Mishra, S. C. Swain, and L. N. Tripathy, "A time-frequency transform based fault detection and classification of STATCOM integrated single circuit transmission," *International Journal of Power Electronics & Drive Systems*, vol. 8, no. 4, p. 1804, 2017.

[22] B. Hu, X. W. Li, S. T. Sun, and M. Ratcliffe, "Attention recognition in EEG-based affective learning research using CFS +KNN algorithm," *IEEE/ACM Transactions on Computational Biology & Bioinformatics*, vol. 15, no. 1, pp. 38–45, 2018.

[23] K. Ding and P. Y. Jiang, "Social sensors (S2ensors): a kind of hardware-software-integrated mediators for social manufacturing systems under mass individualization," *Chinese Journal of Mechanical Engineering*, vol. 30, no. 5, pp. 1150–1161, 2017.

[24] M. Marbac, C. Biernacki, and V. Vandewalle, "Latent class model with conditional dependency per modes to cluster categorical data," *Advances in Data Analysis and Classification*, vol. 10, no. 2, pp. 183–207, 2016.

[25] A. Cano, D. T. Nguyen, S. Ventura, and K. J. Cios, "Ur-CAIM: improved CAIM discretization for unbalanced and balanced data," *Soft Computing*, vol. 20, no. 1, pp. 173–188, 2016.

Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 9129436, 8 pages
https://doi.org/10.1155/2022/9129436

Research Article

Cooperative RIS and Relaying IoV Networks: A Deep Study on Position Analysis

Jiaxing Zhu,[1] Guoan Zhang,[1] Yan Jiang,[2] Wei Duan,[1] Jianghong Ou,[3] and Dahua Fan[3]

[1]*School of Information Science and Technology, Nantong University, Nantong 226019, China*
[2]*Engineering Training Center, Nantong University, Nantong 226019, China*
[3]*Starway Communication, Guangzhou Science City, Guangzhou 510663, China*

Correspondence should be addressed to Guoan Zhang; gzhang@ntu.edu.cn

Received 17 March 2022; Revised 20 April 2022; Accepted 13 May 2022; Published 9 June 2022

Academic Editor: Lisheng Fan

In this paper, a hybrid relay- and reconfigurable intelligent surface- (RIS-) assisted cooperative communication system is proposed. Considering the overall user position referring to the reflective RIS, the downlink propagation can be summarized as two cases: for the first case, the mobile user can only be assisted by the relay (on the back of RIS); for the second case, the mobile user will be assisted by both of the RIS and relay. In our proposed system, a novel concept named "balance position" is investigated, which can be used to resolve specific deployment issue of the RIS. Due to the difficulty of obtaining the closed-form expressions of the achievable capacity, we derive the tight upper bound for the channel gain through observing the central limit theorem (CLM) and Jensen's inequality and determine its trend for the mobile user. Numerical results verify the correctness of our analysis and superiority of our proposed scheme. Moreover, for an increasing number of RIS elements, the system capacity will be significantly improved, and the balance position will be far away from RIS.

1. Introduction

With the proposal of the concept of smart city and the development of intelligent transportation system (ITS), people's daily travel has become more convenient and safer. As the cornerstone of the future ITS [1], Internet of Vehicles (IoV) enables the vehicles to maintain robust connection with their surroundings and remote entities, while provides extensive and convenient services for vehicles [2, 3]. Therefore, IoV has higher requirements for delay and throughput. However, due to the complexity of the environment and the mobility of vehicles, how to improve the quality of vehicle communication in the IoV networks has become an enduring problem in the field of wireless transmission research.

Recently, with the continuous development of 5G and B5G, a great number of emerging technologies are proposed to meet the increasing demand for the data traffic, such as massive multiple-input multiple-output (MIMO), deep learning, and mmwave [4, 5]. However, most services deal with the huge transmission by employing large active antennas, which is too expensive and deployed hardly in practical environment. Reconfigurable intelligent surface (RIS) technology receives considerable attentions for its passive reflecting character, moderate price, and flexible deployment. Meanwhile, RIS possesses superiority spectrum efficiency (SE) and great communication coverage, by intelligently reconfiguring the propagation environment of wireless channels [6]. In addition, RIS consists of plenty of reflective elements, which can take advantage of the ultrathin planar structure to independently manage the amplitude and phase of the incoming signals. With deploying the RIS in wireless networks, the traditional random channel state information (CSI) becomes controllable [7, 8].

In the existing IoV networks, relay has been widely deployed and applied, while the technology is relatively mature. Comparing with the conventional relay technology, the RIS embraces significant advantages, and a considerable amount of theoretical researches and applications have been studied recently. In order to compare the performance of the RIS-aided and amplify-and-forward relaying systems, the authors provided a theoretical framework in [9]. Through analyzing the characteristics of the RIS and relay, the author

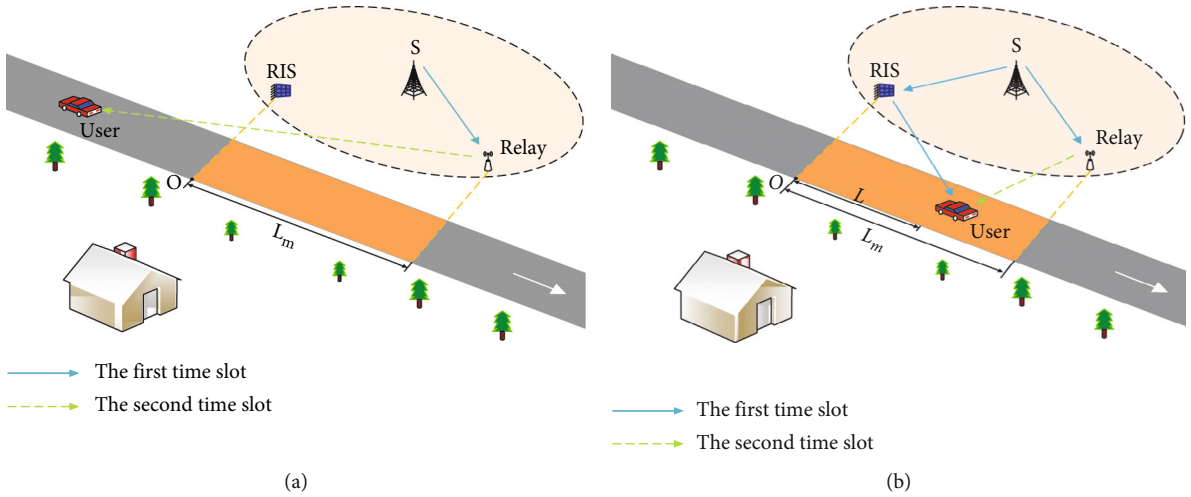


Figure 1: System model: (a) without assistance from RIS, (b) with assistance from RIS.

in [10] elaborated the major differences and similarities. Moreover, when either the relay or RIS is selected to maximize signal-to-noise ratio (SNR), a comparison between the RIS and decode-and-forward (DF) relaying is provided in [11]. Besides, the authors in [12] proposed a novel hybrid relay- and RIS-assisted system to enhance system performance. In [13], a transmission RIS was investigated, which has been a research hotspot recently.

Most of these literatures only focus on the applications of the RIS in a specific environment, e.g., fixed user locations. However, focusing on the current smart city, the capacity of information interaction for mobile devices and vehicular applications is breathtakingly increasing among IoV networks, and a great number of novel technologies, such as mobile edge computing (MEC), are proposed to resolve the huge data traffic [14, 15]. Motivated by above observations, it is promising to study the cooperative RIS and relay jointly assisted IoV networks. Considering the actual situation that the RIS should assist with the existing relay in the initial application, this paper proposes a cooperative reflection RIS and relaying communication system with mobile users, where two scenarios are considered: (i) the mobile user is only assisted by the relay (on the back of RIS); (ii) the mobile user is assisted by both of the RIS and relay. The main contributions of this paper are summarized as follows.

(i) A novel reflective RIS- and relay-assisted cooperative system is investigated, in which the overall position of the mobile user is considered. To clarify the deployment of RIS, a concept of "balance position" is provided, i.e., the equivalent achievable capacity by the RIS and relay

(ii) For the ideal RIS case with optimal reflection, we derive closed-form expressions of the channel gain by exploiting the Jensen's inequality and central limit theorem. The tight upper bound on the achievable capacity is also studied with its trend for mobile user

(iii) Numerical results verify the accuracy of the theoretical analysis revealing the achievable capacity increases firstly and then decreases with the movement of users in the system. Then, the effect of the deployment distance between the RIS and relay on the system is discussed in limited resource environment. The results show that if the deployment distance is too close, the performance of relay will be always better than the RIS

2. System Model

In this section, a hybrid RIS- and relay-assisted communication system is proposed, which is including a source (S), a relay (R), an RIS (I) with N elements, as well as a mobile user (D), as shown in Figure 1. The half-duplex relay adopting the DF protocol and reflective RIS assist the information transfer from S to D, due to the limited coverage of S [16]. In the actual scenes, without loss of generality, the user will enter the communication range from the left or right side. Since these two scenes are similar, for simplicity, we only consider that the activity direction of D is from the back of the RIS towards to the relay. Clearly, there are two possible cases: (i) the user is only assisted by the relay (on the back of RIS); (ii) the user is assisted by both of the RIS and relay (including scenarios of that the user is near to the RIS but far from the relay, and the user is closed to the relay but far from the RIS). Specifically, as shown in Figure 1, for the first case, D enters the coverage of R from the blind sight of I, resulting in that D only receives the signals from the relay. Then, when D moves into the orientation of I, it simultaneously receives the signals from the RIS and relay. For simplicity, we further consider that the RIS and relay will not forward the signals to each other. In the following sections, we denote the channels $S \longrightarrow I$, $S \longrightarrow R$, $I \longrightarrow D$, and $R \longrightarrow D$ as $h_{\mathrm{SI}} \in \mathbb{C}^N$, $h_{\mathrm{SR}} \in \mathbb{C}$, $h_{\mathrm{ID}} \in \mathbb{C}^N$, and $h_{\mathrm{RD}} \in \mathbb{C}$, satisfying independent Rayleigh distributions.

2.1. Only Relay-Assisted Transmission. In this case, S transmits signals to R and I. Due to the blind region of I, D can only receive signals from R, as shown in Figure 1(a). Specifically, according to the half-duplex DF protocol, the transmission from S to D involves two time slots. During the first time slot, S transmits the signal to R, and the received signal at R can be expressed as

$$y_{1,R} = h_{\mathrm{SR}}\sqrt{P}s + n_{1,R}, \tag{1}$$

where $n_{1,R} \sim \mathscr{CN}(0, \sigma^2)$ means the received additive white Gaussian noise (AWGN) at R, P means the transmit power, and s means the transmitted signal with $E[|s|^2] = 1$.

During the second time slot, R decodes and forwards s to D. Therefore, the received signal at D can be given as

$$y_{2,D} = h_{\mathrm{RD}}\sqrt{P}s + n_{2,D}, \tag{2}$$

where $n_{2,D} \sim \mathscr{CN}(0, \sigma^2)$ means the AWGN at D. Denoting $P/\sigma^2 = \rho$ and adopting the maximum ratio combining (MRC), the achievable rate at D can be obtained from [17]

$$\mathscr{R}_{\mathrm{DF}} = \frac{1}{2}\log_2\left(1 + \rho \min\left(|h_{\mathrm{SR}}|^2, |h_{\mathrm{RD}}|^2\right)\right), \tag{3}$$

where 1/2 means the transmission involving two time slots.

2.2. Relay- and RIS-Assisted Transmission. In this case, D enters the reflection orientation of I, in which it can receive the signals reflected by I and forwarded from R. Note that I is embedded with N discrete elements, and the size of each element is incomparable with the wavelength. Therefore, it can flexibly scatter the incident signal with almost constant gain in all directions of interest. In this manner, the signal reflected by I can be expressed as

$$y_{\mathrm{RIS}} = \sqrt{P}\left(\mathbf{h}_{\mathrm{SI}}^{\mathrm{T}}\boldsymbol{\Theta}\mathbf{h}_{\mathrm{IR}}\right)s + n_I, \tag{4}$$

where $n_I \sim \mathscr{CN}(0, \sigma^2)$ means the AWGN at I, and $\boldsymbol{\Theta} = \mathrm{diag}\left(\eta_1 e^{j\theta_1}, \cdots, \eta_N e^{j\theta_N}\right)$ denotes the reflection matrix of the RIS, where $\eta_i \in [0, 1]$ and $\theta_i \in [0, 2\pi]$, respectively, denote the amplitude attenuation and phase-shift of the ith element, for $i = [1, 2, \cdots, N]$. Therefore, the received SNR at I can be obtained from [12]

$$\gamma_{\mathrm{RIS}} = \rho\left|\sum_{i=1}^{N}\eta_i e^{j\theta_i}[\mathbf{h}_{\mathrm{SI}}]_i[\mathbf{h}_{\mathrm{ID}}]_i\right|^2. \tag{5}$$

Based on Shannon theory, the capacity of RIS-assisted scenario is calculated by

$$\mathscr{R}_{\mathrm{RIS}} = \log_2(1 + \gamma_{\mathrm{RIS}}) = \log_2\left(1 + \rho\left|\sum_{i=1}^{N}\eta_i e^{j\theta_i}[\mathbf{h}_{\mathrm{SI}}]_i[\mathbf{h}_{\mathrm{ID}}]_i\right|^2\right). \tag{6}$$

Since the received signal at D from R is similar to the only relay-assisted transmission case previous subsection, we omit it here. Combining Equations (3) and (6), the achievable sum-rate at D is finally expressed as

$$\mathscr{R}_{\mathrm{sum}} = \mathscr{R}_{\mathrm{RIS}} + \mathscr{R}_{\mathrm{DF}}. \tag{7}$$

The practical transmission of the relay should consume two time slots. During the second time slot, the relay decodes and forwards the signal to D. Meanwhile, the RIS reflects the signal received from S again. Therefore, the achievable sum-rate at D of two time slots can be expressed as $2\mathscr{R}_{\mathrm{sum}}$. Since the sum-rate has the linear relationship with the time slot, and the subsequent analysis focuses on the trend of $\mathscr{R}_{\mathrm{sum}}$, the sum-rate of one time slot is analyzed in the following.

3. Performance Analysis for Proposed Scheme

In this section, considering the number of the RIS elements and distances from D to I and R, the influence of D in different positions on its achievable rate is discussed.

3.1. Maximal Achievable Rates for Proposed System. For simplicity, we assume that the perfect CSI is obtained to realize the ideal passive beamforming at I [18]. It is the optimal phase shift that align the signal reflected by RIS with the user in the proposed system [7]. Therefore, it is clear that the optimal γ_{RIS} can be obtained when $\theta_i = 0$. Meanwhile, it is assumed that the phase shifts can be controlled to change continuously to obtain the optimal phase shift, while the reflection amplitude of all elements is assumed to be η. Accordingly, optimal γ_{RIS} can be equivalently simplified as

$$\gamma_{\mathrm{RIS}} = \rho\left|\sum_{i=1}^{N}\eta_i e^{j\theta_i}[\mathbf{h}_{\mathrm{SI}}]_i[\mathbf{h}_{\mathrm{ID}}]_i\right|^2 = \rho\left|\sum_{i=1}^{N}\eta[\mathbf{h}_{\mathrm{SI}}]_i[\mathbf{h}_{\mathrm{ID}}]_i\right|^2. \tag{8}$$

Through observing the statistical characteristics of the CSI, we derive the ergodic capacity of our scheme, following the Jensen's inequality as

$$\mathbb{E}[\log_2(1 + v)] \le \log_2(1 + \mathbb{E}[v]). \tag{9}$$

The channels are denoted as $h_j = g_j d_j^{-\alpha/2}, j \in \{\mathrm{SR}, \mathrm{RD}\}$, and $\mathbf{h}_k = \mathbf{g}_k d_k^{-\alpha/2}, k \in \{\mathrm{SI}, \mathrm{ID}\}$, in which $g_j \in \mathbb{C}$ and $\mathbf{g}_k \in \mathbb{C}^N$ stand for the complex Gaussian fading factors following $\mathscr{CN}(0, 1)$, and d and α, respectively, denote the distance and path-loss exponent for the corresponding channels.

Note $\sqrt{\beta_j} = |h_j|$, $\sqrt{\beta_{\mathrm{SID}}} = \eta/N\sum_{i=1}^{N}|[\mathbf{h}_{\mathrm{SI}}]_i[\mathbf{h}_{\mathrm{ID}}]_i|$, where β_j means exponentially distributed with $\mathbb{E}[\beta_j] = d_j^{-\alpha}$. Morever, for the RIS with plenty of elements, β_{SID} follows a noncentral chi-square distribution with $\mathbb{E}[\beta_{\mathrm{SID}}] = \eta^2[\pi^2 + (1/N)(16 - \pi^2)]/16 d_{\mathrm{SI}}^{\alpha} d_{\mathrm{ID}}^{\alpha}$ [19]. With these observations, the expectation of the upper bound for γ_{RIS} can be derived as

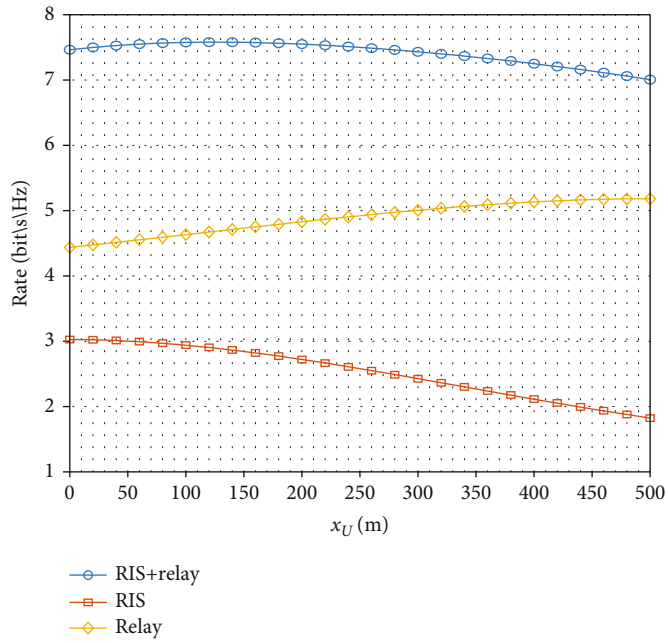


Figure 2: Achievable rate of user versus user position in our proposed system.

$$\mathbb{E}[\tilde{\gamma}_{\mathrm{RIS}}] = \mathbb{E}\left[\rho\left(N\sqrt{\beta_{\mathrm{SID}}}\right)^2\right] = \rho N^2 \mathbb{E}[\beta_{\mathrm{SID}}] = \frac{1}{16}\rho d_{\mathrm{SI}}^{-a} d_{\mathrm{IR}}^{-a}\eta^2 N\left(16 + (N-1)\pi^2\right). \tag{10}$$

Finally, the upper bound ergodic capacity of our scheme can be obtained from

$$\tilde{\mathscr{R}}_{\mathrm{sum}} = \tilde{\mathscr{R}}_{\mathrm{DF}} + \tilde{\mathscr{R}}_{\mathrm{RIS}} \le \frac{1}{2}\log_2\left(1 + \min\left(\mathbb{E}[\beta_{\mathrm{SR}}], \mathbb{E}[\beta_{\mathrm{RD}}]\right)\right) + \log_2\left(1 + \mathbb{E}[\tilde{\gamma}_{\mathrm{RIS}}]\right). \tag{11}$$

As shown in Figures 1(a) and 1(b), the intersection of I projection and D motion route is represented as O, the distance between the location of I and the point O is denoted by d, the distance relative to point O is set to L, and the distance between I and R is denoted to L_m. By this way, the distances from D to I and R can be, respectively, given as $d_{\mathrm{ID}} = \left(L^2 + d^2\right)^{1/2}$ and $d_{\mathrm{RD}} = \left((L_m - L)^2 + d^2\right)^{1/2}$. With above results and Equation (11), $\tilde{\mathscr{R}}_{\mathrm{sum}}$ can be derived as

$$\tilde{\mathscr{R}}_{\mathrm{sum}}(L) = \frac{1}{2}\log_2\left[1 + \rho\min\left(d_{\mathrm{SR}}^{-\alpha}, \left((L_m - L)^2 + d^2\right)^{-\alpha/2}\right)\right] + \log_2\left[1 + \frac{1}{16}\rho\eta^2 d_{\mathrm{SI}}^{-\alpha} N\left(16 + (N-1)\pi^2\right)\left(L^2 + d^2\right)^{-\alpha/2}\right]. \tag{12}$$

In order to discuss the characteristics of $\tilde{\mathscr{R}}_{\mathrm{sum}}$, the first-order derivation of $\tilde{\mathscr{R}}_{\mathrm{sum}}$ is obtained, with respect to L. Due to the complexity of mathematical manipulation, the function is divided into two cases for better analysis. For $d_{\mathrm{RD}} \ge d_{\mathrm{SR}}$, the first-order derivation of $\tilde{\mathscr{R}}_{\mathrm{sum_a}}$ can be derived as

$$\widetilde{\mathscr{R}}'_{\mathrm{sum_a}} = \frac{\alpha\rho\left((L_m - L)^2 + d^2\right)^{-(\alpha/2)-1}(L_m - L)}{2\ln 2\left(1 + \rho\left((L_m - L)^2 + d^2\right)^{-\alpha/2}\right)} + \underbrace{\frac{-\alpha\rho\eta^2 d_{\mathrm{SI}}^{-\alpha} N\left(16 + (N-1)\pi^2\right)L\left(L^2 + d^2\right)^{-(\alpha/2)-1}}{16\ln 2\left[1 + 1/16\rho\eta^2 d_{\mathrm{SI}}^{-\alpha} N(16 + (N-1)\pi^2)\left(L^2 + d^2\right)^{-\alpha/2}\right]}}_{A}. \tag{13}$$

For $d_{\mathrm{RD}} \le d_{\mathrm{SR}}$, due to that the relay forwarded signal is not related to L, the first-order derivation derivative of $\tilde{\mathscr{R}}_{\mathrm{sum_b}}$ can be simply written as

$$\widetilde{\mathscr{R}}'_{\mathrm{sum_b}} = A. \tag{14}$$

From Equations (13) and (14), it is clear that, for $L = 0$, the first derivation of $\tilde{\mathscr{R}}_{\mathrm{sum}_a}$ will always be greater than 0, and the first derivation of $\tilde{\mathscr{R}}_{\mathrm{sum}_b}$ is a negative result. Therefore, it can be concluded that $\mathscr{R}_{\mathrm{RIS}}$ is a monotonically increasing function, while $\mathscr{R}_{\mathrm{DF}}$ is a monotonically subtraction function. Hence, according to the existence theorem of zero points, it is possible to find an optimal L^* maximizing the system capacity. Due to the difficulty of obtaining the closed-form solution of this optimal value, the numerical solution of the optimal value under specific conditions is obtained in following simulation results.

3.2. Balance Position of the Relay and RIS. Considering some actual scenarios, e.g., limited resources, the RIS and relay are not allowed to serve users simultaneously. Balance position

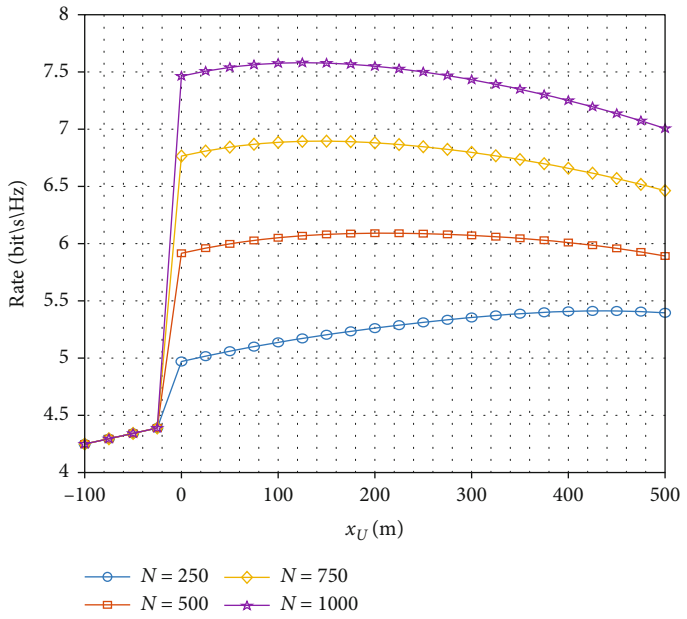


Figure 3: Achievable rate of user versus user position in our proposed system.

denotes the position where the equivalent achievable capacity can be achieved by the RIS and relay. The significance of the balance position is that a more efficient communication mode can be selected according to position of the user relative to the balance position.

From Equation (10), it is clear that $\tilde{\gamma}_{\mathrm{RIS}}$ is an increasing function of N. If there is a balance position, the reflection element N should satisfy the following constraints:

$$\mathscr{R}_{\mathrm{DF}_{\min}} \le \mathscr{R}_{\mathrm{RIS}}(N) \le \mathscr{R}_{\mathrm{DF}_{\max}}. \tag{15}$$

As mentioned above, $\mathscr{R}_{\mathrm{DF}}$ is an increasing function of L. Therefore, from Equation (6), when $L = 0$, $\mathscr{R}_{\mathrm{DF}_{\min}} = 1/2 \log_2(1 + \rho \min(d_{\mathrm{SR}}^{-\alpha}, (L_m^2 + d^2)^{-\alpha/2}))$; meanwhile, for $L = L_m$, $\mathscr{R}_{\mathrm{DF}_{\max}} = 1/2 \log_2(1 + \rho \min(d_{\mathrm{SR}}^{-\alpha}, d^{-\alpha}))$. Therefore, Equation (15) can be expressed as

$$\begin{aligned}
&\frac{1}{2}\log_2(1+\rho\min(d_{\mathrm{SR}}^{-\alpha}, d^{-\alpha}))\\
&\ge \log_2\left[1+\frac{1}{16}\rho\eta^2 d_{\mathrm{SI}}^{-\alpha} N\left(16+(N-1)\pi^2\right)\left(L^2+d^2\right)^{-\alpha/2}\right]\\
&\ge \frac{1}{2}\log_2\left(1+\rho\min\left(d_{\mathrm{SR}}^{-\alpha}, \left(L_m^2+d^2\right)^{-\alpha/2}\right)\right).
\end{aligned} \tag{16}$$

Through inequality transformation, Equation (16) is simplified to the inequality about N as

$$\begin{aligned}
N \ge\ & \frac{1}{2d_{\mathrm{SI}}^{-\alpha}\rho\eta^2\pi^2}\sqrt{d_{\mathrm{SI}}^{-\alpha}\rho\eta^2\left(256d_{\mathrm{SI}}^{-\alpha}\rho\eta^2 + d_{\mathrm{SI}}^{-\alpha}\rho\eta^2\pi^4 + 32\pi^2\left(2d^{\alpha}\left(\left(1+\rho\min\left(d_{\mathrm{SR}}^{-\alpha}, \left(L_m^2+d^2\right)^{-\alpha/2}\right)\right)^{1/2}-1\right) - d_{\mathrm{SI}}^{-\alpha}\rho\eta^2\right)\right)}\\
&+ \frac{d_{\mathrm{SI}}^{-\alpha}\left(-16+\pi^2\right)\rho\eta^2}{2d_{\mathrm{SI}}^{-\alpha}\rho\eta^2\pi^2},
\end{aligned} \tag{17}$$

$$\begin{aligned}
N \le\ & \frac{1}{2d_{\mathrm{SI}}^{-\alpha}\rho\eta^2\pi^2}\sqrt{d_{\mathrm{SI}}^{-\alpha}\rho\eta^2\left(256d_{\mathrm{SI}}^{-\alpha}\rho\eta^2 + d_{\mathrm{SI}}^{-\alpha}\rho\eta^2\pi^4 + 32\pi^2\left(2\left(d^2+L_m^2\right)^{\alpha/2}\left(\left(1+\rho\min\left(d_{\mathrm{SR}}^{-\alpha}, d^{-\alpha}\right)\right)^{1/2}-1\right) - d_{\mathrm{SI}}^{-\alpha}\rho\eta^2\right)\right)}\\
&+ \frac{d_{\mathrm{SI}}^{-\alpha}\left(-16+\pi^2\right)\rho\eta^2}{2d_{\mathrm{SI}}^{-\alpha}\rho\eta^2\pi^2}.
\end{aligned} \tag{18}$$

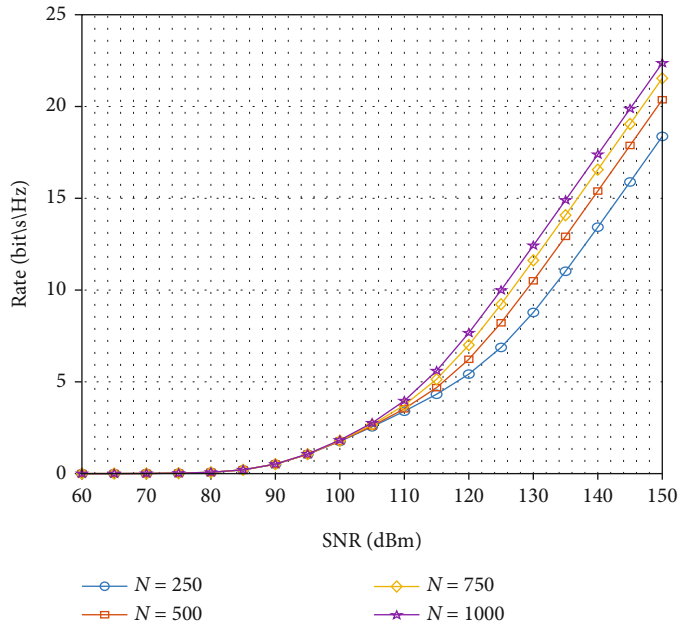


Figure 4: Achievable rate of user versus SNR in our proposed system.

To figure out where the balance position is, the problem about L is proposed as

$$\log_2\left[1+\frac{1}{16}\rho\eta^2 d_{\mathrm{SI}}^{-\alpha}N\left(16+(N-1)\pi^2\right)\left(L^2+d^2\right)^{-\alpha/2}\right] = \frac{1}{2}\log_2\left[1+\rho\min\left(d_{\mathrm{SR}}^{-\alpha},\left((L_m-L)^2+d^2\right)^{-\alpha/2}\right)\right]. \tag{19}$$

According to the knowledge of information theory, the achievable rate of the relay forwarding is limited by the less value of the two-stage channel gain. Since the position of the user is changing, and the channel relayed to user changes with the user's movement, the balance position is discussed in two cases.

Discussion 1. If $d_{\mathrm{SR}} \geq d_{\mathrm{RD}}$, Equation (19) can be further expressed as

$$\log_2\left[1+\frac{1}{16}\rho\eta^2 d_{\mathrm{SI}}^{-\alpha}N\left(16+(N-1)\pi^2\right)\left(L^2+d^2\right)^{-\alpha/2}\right] = \frac{1}{2}\log_2\left(1+\rho d_{\mathrm{SR}}^{-\alpha}\right). \tag{20}$$

It is easy to see that these exists two solutions for Equation (20). Omitting the negative result, the optimal result of L for Equation (20) is

$$L^* = \left(\left(\frac{16\left(\left(1+\rho d_{\mathrm{SR}}^{-\alpha}\right)^{1/2}-1\right)}{\rho\eta^2 d_{\mathrm{SI}}^{-\alpha}N(16+(N-1)\pi^2)}\right)^{-2/\alpha}-d^2\right)^{1/2}. \tag{21}$$

Discussion 2. If $d_{\mathrm{SR}} \leq d_{\mathrm{RD}}$, Equation (19) can be expressed as

$$\log_2\left[1+\frac{1}{16}\rho\eta^2 d_{\mathrm{SI}}^{-\alpha}N\left(16+(N-1)\pi^2\right)\left(L^2+d^2\right)^{-\alpha/2}\right] = \frac{1}{2}\log_2\left(1+\rho\left((L_m-L)^2+d^2\right)^{-\alpha/2}\right). \tag{22}$$

Since the optimal result of L in Equation (22) is quite difficult to obtain, we turn to verify it by Monte Carlo simulation in following numerical results.

4. Numerical Results and Analysis

In this section, we perform simulation to verify the achievable rate with the mobile user and the influence of N for our proposed system. By means of the simulation results, the effect of the deployment distance between the RIS and relay on the balance position is also analyzed.

Similar to [20], the 3GPP urban micro (UMi) under 3 GHz operating frequency is used to model the channel gains [21], which is the typical channel gain with distance as the parameter. Hence, the path loss model can be given as

$$\beta(L)[\mathrm{dB}] = G_{\mathrm{t}} + G_{\mathrm{r}} - 37.5 - 22\log_{10}(L/1\mathrm{m}), \tag{23}$$

where G_{t} and G_{r}, respectively, stand for the antenna gains at the transmitter and receiver. The simulation setup in Figure 1 is considered. For simplicity, following locations of each node is considered: $(x_S, y_S) = (250,600)$, $(x_I, y_I) = (0,400)$, and $(x_R, y_R) = (500,400)$. Meanwhile, the user moves right along the x-axis; the coordinates of the mobile user are expressed as $(x_D, 0)$. We assume that the mobile user is equipped with a 0 dBi omnidirectional antenna; other nodes have 5 dBi antenna gain. Moreover, it is assumed that $P = 500$ W, $B = 10$ MHz, and the noise power is −94 dBm.

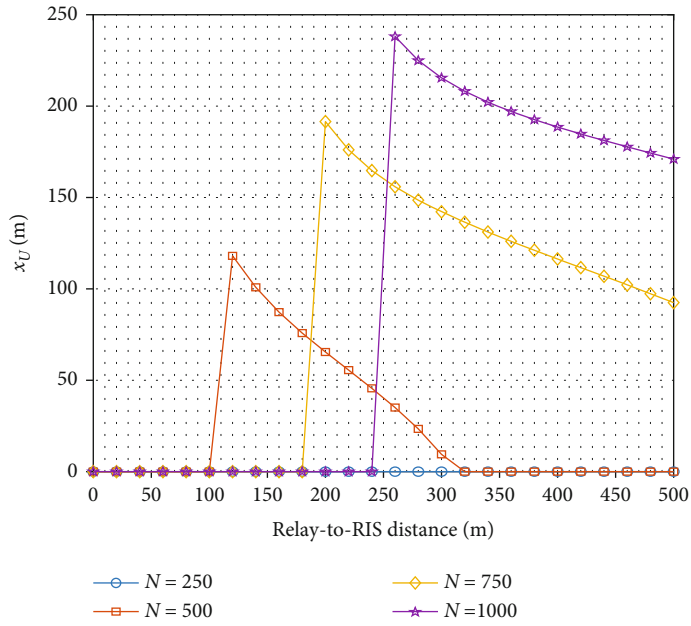


Figure 5: Balance position versus distance between the RIS and relay.

4.1. Achievable Rates of the Proposed Hybrid Scheme. Figure 2 demonstrates the capacity change of each link in the proposed system for $N = 1000$. $d_{IR} = 50$ m, and $N = 50$. It is easy to see that, with the movement of the user, the capacity of the RIS-assisted link decreases monotonically, while that the capacity corresponding to relay-assisted link is improved. Specifically, there exists one point maximizing the system capacity, which is consistent with our theoretical analysis. Figure 3 demonstrates the change of achievable rate under different N when the user enters the system from $(-100,0)$, for $N \in \{250,500,750,1000\}$. It is clear that the achievable rate increases before user enters the reflection region of the RIS. Once entering the RIS coverage, with the reflected signals from the RIS, the achievable rate will be significantly increased verifying the trend in Figure 2. Specifically, the position where the maximum rate occurs is also affected by N. Figure 4 demonstrates the change of achievable rate at different SNR under different N, for x_D=100 m. It is clear that when the channel fading is sufficient large, a greater N results in a more superior performance for high SNR. In conclusion, on the premise that the actual technical conditions permit, a greater N results in a better performance in terms of capacity.

4.2. Balance Position. In the proposed system, when the mobile user moves to the left of the balance position and only one node can be selected to maximize the SNR, the performance of the selecting relay is better than that of the RIS. Otherwise, the RIS should be selected. The influence of changing the distance between the RIS and relay (d_{IR}) on the balance position is discussed. We set the ordinate of the RIS and relay as 200 m, and the source ordinate as 300 m, while adjusting d_{IR} from 0 m to 500 m, the relationship between the balance positions and d_{IR} is illustrated in Figure 5. Clearly, the balance position will be close to the RIS due to the increasing d_{IR} and decreasing N. The points falling on the x-axis indicates that there is no balance position in these conditions. For $N = 250$, there is no balance position, since N is not enough to satisfy the conditions of Equations (17) and (18). Therefore, when considering the practical application scenario, especially for the one with limited transmit power, it is also necessary to select the appropriate deployment distance. Otherwise, the effect of the RIS may always be worse than that of relay due the passive reflection for RIS.

5. Conclusion

Focusing on the reflective RIS, a hybrid relay- and RIS-assisted cooperative IoV communication system was proposed, with the considering of overall mobile user positions. In addition, a novel balance position concept for such system is investigated to resolve specific deployment issue of the RIS. The tight upper bound of the capacity for the mobile user is derived. By means of numerical results, our proposed system show its superiority, especially for an increasing number of RIS elements. Moreover, the balance position will be far away from the RIS for a greater N.

Data Availability

No data were used to support this study.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 61801249 and Grant 61971245 and the Postgraduate Research & Practice Innovation Program of Jiangsu Province under Grant KYCX21-3086.

References

[1] J. Wang, C. Jiang, K. Zhang, T. Q. S. Quek, Y. Ren, and L. Hanzo, "Vehicular sensing networks in a smart city: principles, technologies and applications," *IEEE Wireless Communications*, vol. 25, no. 1, pp. 122–132, 2018.

[2] A. Al-Hilo, M. Samir, M. Elhattab, C. Assi, and S. Sharafeddine, "Reconfigurable intelligent surface enabled vehicular communication: joint user scheduling and passive beamforming," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 3, pp. 2333–2345, 2022.

[3] J. Wang, C. Jiang, Z. Han, Y. Ren, and L. Hanzo, "Internet of Vehicles: sensing-aided transportation information collection and diffusion," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 5, pp. 3813–3825, 2018.

[4] B. Lin, F. Gao, S. Zhang, T. Zhou, and A. Alkhateeb, "Deep learning based antenna selection and CSI extrapolation in massive MIMO systems," *IEEE Transactions on Wireless Communications*, vol. 20, no. 11, pp. 7669–7681, 2021.

[5] F. Gao, B. Lin, C. Bian, T. Zhou, J. Qian, and H. Wang, "FusionNet: enhanced beam prediction for mmWave communications using sub-6 GHz channel and a few pilots," *IEEE Transactions on Communications*, vol. 69, no. 12, pp. 8488–8500, 2021.

[6] Q. Wu and R. Zhang, "Towards smart and reconfigurable environment: intelligent reflecting surface aided wireless network," *IEEE Communications Magazine*, vol. 58, no. 1, pp. 106–112, 2020.

[7] Q. Wu, S. Zhang, B. Zheng, C. You, and R. Zhang, "Intelligent reflecting surface-aided wireless communications: a tutorial," *IEEE Transactions on Communications*, vol. 69, no. 5, pp. 3313–3351, 2021.

[8] Q. Wu and R. Zhang, "Beamforming optimization for wireless network aided by intelligent reflecting surface with discrete phase shifts," *IEEE Transactions on Communications*, vol. 68, no. 3, pp. 1838–1851, 2020.

[9] A. A. Boulogeorgos and A. Alexiou, "Performance analysis of reconfigurable intelligent surface-assisted wireless systems and comparison with relaying," *IEEE Access*, vol. 8, pp. 94463–94483, 2020.

[10] M. Di Renzo, K. Ntontin, J. Song et al., "Reconfigurable intelligent surfaces vs. relaying: differences, similarities, and performance comparison," *IEEE Open Journal of the Communications Society*, vol. 1, pp. 798–807, 2020.

[11] J. Ye, A. Kammoun, and M.-S. Alouini, "Spatially-distributed RISs vs relay-assisted systems: a fair comparison," *IEEE Open Journal of the Communications Society*, vol. 2, pp. 799–817, 2021.

[12] Z. Abdullah, G. Chen, S. Lambotharan, and J. A. Chambers, "A hybrid relay and intelligent reflecting surface network and its ergodic performance analysis," *IEEE Wireless Communications Letters*, vol. 9, no. 10, pp. 1653–1657, 2020.

[13] S. Zhang, H. Zhang, B. Di, Y. Tan, Z. Han, and L. Song, "Beyond intelligent reflecting surfaces: reflective-transmissive metasurface aided communications for full-dimensional coverage extension," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 11, pp. 13905–13909, 2020.

[14] W. Duan, J. Gu, M. Wen, G. Zhang, Y. Ji, and S. Mumtaz, "Emerging technologies for 5G-IoV networks: applications, trends and opportunities," *IEEE Network*, vol. 34, no. 5, pp. 283–289, 2020.

[15] W. Duan, X. Gu, M. Wen, Y. Ji, J. Ge, and G. Zhang, "Resource management for intelligent vehicular edge computing networks," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–12, 2021.

[16] D. Lee and J. H. Lee, "Outage probability of decode-and-forward opportunistic relaying in a multicell environment," *IEEE Transactions on Vehicular Technology*, vol. 60, no. 4, pp. 1925–1930, 2011.

[17] J. N. Laneman, D. N. C. Tse, and G. W. Wornell, "Cooperative diversity in wireless networks: efficient protocols and outage behavior," *IEEE Transactions on Information Theory*, vol. 50, no. 12, pp. 3062–3080, 2004.

[18] Q. Wu and R. Zhang, "Intelligent reflecting surface enhanced wireless network via joint active and passive beamforming," *IEEE Transactions on Wireless Communications*, vol. 18, no. 11, pp. 5394–5409, 2019.

[19] E. Basar, M. D. Renzo, J. D. Rosny, M. Debbah, M.-S. Alouini, and R. Zhang, "Wireless communications through reconfigurable intelligent surfaces," *IEEE Access*, vol. 7, pp. 116753–116773, 2019.

[20] E. Bjornson, O. Ozdogan, and E. G. Larsson, "Intelligent reflecting surface versus decode-and-forward: how large surfaces are needed to beat relaying?," *IEEE Wireless Communications Letters*, vol. 9, no. 2, pp. 244–248, 2020.

[21] "Further Advancements for E-UTRA Physical Layer Aspects(-Release 9)," *3GPP Technical Specification, TR 36.814*, vol. 4, no. 1, 2009http://www.3gpp.org.

Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 9560806, 9 pages
https://doi.org/10.1155/2022/9560806

Research Article

Joint Deployment and Power Optimization for UAV Relay in Multiuser Networks

Ang Ji and Jianjun Wu

School of Electronics, Peking University, Beijing, China

Correspondence should be addressed to Ang Ji; ji.ang@pku.edu.cn

Received 16 March 2022; Accepted 23 May 2022; Published 8 June 2022

Academic Editor: Lisheng Fan

Within the UAV network, the UAV first receives signals from multiple remote mobile devices (MDs) and then amplifies and forwards the transmitted signals to the base station (BS) with different amplification coefficients to form a UAV relay multiuser network. In this paper, we propose a new method to solve the problem of maximizing the throughput of the relay network. The proposed problem is decoupled into two subproblems, UAV deployment design and amplification coefficient optimization, to be solved iteratively, respectively. We solve the UAV deployment problem by adjusting its trajectory with a gradient descent-based method and solve the amplification coefficient subproblem with a convex optimization-based method iteratively. Simulation shows that the proposed UAV deployment and amplification coefficient of the UAV optimization design algorithm significantly improves the sum-rate compared with existing fixed relay and equal power allocation schemes. Finally, we discuss future potential performance enhancing methods including multiple UAV cooperation, massive multi-input multioutput (MIMO) communications, and nonorthogonal multiple access (NOMA) communications.

1. Introduction

The unmanned aerial vehicle (UAV) is becoming a promising technique that can be widely used in various scenarios, such as civil, emergency rescue, and military services [1]. Owning to its flexible deployment, high mobility, and low operational costs, the UAV has gained increasing popularity in the use of wireless networks [2] and has been formally discussed in the Third Generation Partnership Project (3GPP) specifications [3]. Applications in which UAVs significantly outperform terrestrial facilities include wireless coverage extension, remote sensing, and search and rescue [4].

UAVs are drawing significant attention in wireless networks to provide low-latency and easy-access wireless services by working as mobile relays in the above applications. Relevant reports indicate that, by the end of 2021, more than 29 million UAVs have been deployed [5]. In UAV relay networks, UAVs can be dynamically deployed at the optimal locations for serving high-mobility users [6]. UAV relays can also be deployed in emergency networks to keep user nodes in service in disasters [7].

Several recent works have explored the network performance improvement brought by the high mobility in UAV relay networks. In [8], the authors studied the downlink sum-rate maximization problem of a UAV relay network and designed the UAV trajectory and power allocation. In [9], the outage probability of UAV relay network was derived, and a trajectory design and power allocation method was proposed. In [10], the deployment and routing of the ad hoc-based UAV relay network were studied to reduce the transmission latency. In [11], the authors proposed a hybrid network architecture by leveraging the use of UAV as an aerial mobile base station (BS) to offload traffic from ground BSs; a joint UAV trajectory, bandwidth allocation, and user partitioning algorithm was proposed to maximize the sum-rate of the network. A UAV that works as a relay to collect data from ground sensor nodes was studied in [12], and a joint design of UAV trajectory and communication scheduling method was proposed. A UAV-assisted URLLC service system where the blocklength of channel codes is finite was studied in [13] with the constraint of uplink energy of the sensor nodes. However, the above works either consider the UAV as a BS and omit the performance of

the fronthaul link to the core network or consider the UAV as a data collector that receives data from users one by one, which is different from the working scheme of cellular relays. As a result, the above solutions and simulations may have a large gap to the realistic UAV relay networks [14, 15].

In this paper, we consider an uplink UAV amplify-and-forward (AF) relay network with a UAV, a BS, and multiple mobile devices (MDs). Unlike the existing works [8–13], we study the sum-rate maximization problem by adopting the unique air-to-ground model and considering the performance of the MD-UAV link and the UAV-BS link jointly. We formulate this performance maximum problem to a joint UAV deployment and amplification coefficient optimization problem and design an iterative algorithm to solve the nonconvex problems effectively. The analyses on the proposed algorithm in terms of convergency and complexity are also studied. The proposed algorithm can also be applied in multi-UAV scenarios, where the interference management should also be considered. Since the interference caused by the UAV relay is similar to that of the UAV working as BSs, the interference management can be solved by existing works as proposed in [16]. In addition, some extensive scenarios of UAV relay communications are introduced in this paper, and some corresponding open problems and potential solutions are discussed, as illustrated below:

(1) Cooperative UAV relay network: cooperative UAVs are capable to extend the coverage of the relay network owing to their flexible deployment. To reduce power consumption and interference, the trajectory and radio resource management of the UAVs can be designed jointly

(2) Millimeter-wave (mmWave) UAV relay network: the emerging mmWave technique enables narrow beam transmission for the air-to-ground communications, which not only enhances the strength of receive signal but also avoids severe interference caused by the high probability of LoS links between the UAVs and MDs

(3) Nonorthogonal multiple access (NOMA) for UAV relay network: NOMA can be utilized in the mmWave UAV relay network to provide high throughput and massive connectivity. With proper beamforming in the mmWave communication system, MD pairs with significant channel gain differences can obtain significant uplink transmission rate gain when compared with orthogonal multiple access (OMA) communications

The rest of this paper is organized as follows. In Section 2, the system model and the sum-rate maximization problem are described and formulated, respectively. In Section 3, a joint UAV deployment and amplification coefficient optimization algorithm is given to solve the formulated problem. Simulation results are presented in Section 4, and the extensive scenarios of the UAV relay communications are described in Section 5. Finally, the conclusion of this paper is summarized in Section 6.

The notations used in this paper are listed in Table 1 for ease of reference.

2. System Model and Problem Formulation

As shown in Figure 1, we consider a cellular network with one BS and K MDs. The MDs, marked as $K = \{1, 2, \cdots, K\}$, are acquired to perform uplink transmissions to the BS; however, these MDs are beyond the coverage of the BS. To provide service coverage for the K MDs, a UAV is used as an AF relay between the BS and MDs. The high mobility of the UAV enables that it can dynamically adjust its location according to the distribution of the MDs to improve the quality of services (QoS) for the MDs. We assume that the transmission process contains T time slots, and the uplink transmission is performed in every two consecutive time slots. In the first time slot, the UAV receives uplink data from the K MDs. In the second time slot, the UAV amplifies the received signals and forwards them to the BS. We denote the positions of BS and MD i by B and M_i, respectively.

In time slot t, we denote the location of the BS by $(0, 0, H)$, the location of MD i by $l_i(t) = (x_i(t), y_i(t), 0)$, and the location of the UAV by $L(t) = (X(t), Y(t), H(t))$. The distance between the UAV and BS and the distance between the MD i and UAV are given by $d_{\mathrm{B}}^t = |L(t) - (0, 0, H)|$ and $d_i^t = |l_i(t) - L(t)|$, respectively. We denote the UAV speed in time slot t by $v(t)$, which is no more than the maximum UAV speed $v_{\max}$.

The air-to-ground propagation model proposed in [17, 18] is utilized to describe the MD-UAV and UAV-BS transmissions. To study the average performance of the network, this paper only focuses on the large-scale fading of the transmission links, while the small-scale fading can be omitted.

The channel model contains two parts: line-of-sight (LoS) path loss and non-line-of-sight (NLoS) path loss. In time slot t, the LoS and NLoS path loss models between the UAV and MD i in dB are given by

$$P_{\mathrm{L}}^{i,t} = L_{\mathrm{FS}} + 20 \log \left(d_i^t\right) + \eta_{\mathrm{LoS}}, \tag{1}$$

$$\mathrm{PL}_{\mathrm{N}}^{i,t} = L_{\mathrm{FS}} + 20 \log \left(d_i^t\right) + \eta_{\mathrm{NLoS}}, \tag{2}$$

where L_{FS} is the free space path loss given by $L_{\mathrm{FS}} = 20 \log (f) + 20 \log (4\pi/c)$, and f is the system carrier frequency. η_{LoS} and η_{NLoS} are additional attenuation factors due to the LoS and NLoS connections. This model assumes that all the antennas on the BS, UAV, and MDs are vertically deployed. Based on these assumption, the LoS connection probability is as follows:

$$\mathrm{Pr}_{\mathrm{L}}^{i,t} = \left(1 + \alpha \exp\left(-\beta\left(\phi^{i,t} - \alpha\right)\right)\right)^{-1}, \tag{3}$$

In the equation, α and β are environmental parameters, and $\phi^{i,t} = \sin^{-1}(H/d_i^t)$ is the elevation angle. The average large-scale path loss can then be expressed as

$$\mathrm{PL}_i^t = 10^{\left(\mathrm{Pr}_{\mathrm{L}}^{i,t} \times \mathrm{PL}_{\mathrm{L}}^{i,t} + \mathrm{Pr}_{\mathrm{N}}^{i,t} \times \mathrm{PL}_{\mathrm{N}}^{i,t}\right)/10}, \tag{4}$$

where $\mathrm{Pr}_{\mathrm{N}}^{i,t} = 1 - \mathrm{Pr}_{\mathrm{L}}^{i,t}$. In this paper, We assume the transmission power of each MD is fixed as a constant PT, and there is

Table 1: Notations.

Symbol	Description
$L(t)$	Location of UAV in time slot t
d_i^t	Distance between MD i and UAV
$v_{\max}$	Maximum UAV speed
d_{B}^t	Distance between UAV and BS
$P_{\mathrm{L}}^{i,t}$	LoS path loss
$\mathrm{PL}_{\mathrm{N}}^{i,t}$	NLoS path loss
$\mathrm{Pr}_{\mathrm{L}}^{i,t}$	Probability of LoS connection
$\mathrm{Pr}_{\mathrm{N}}^{i,t}$	Probability of NLoS connection
$P_{i,\mathrm{U}}^t$	Received power of the UAV from MD i
P_{T}	MD transmission power
N_0	Noise variance
G_i^t	UAV amplification coefficient for MD i's signal
$P_{i,\mathrm{B}}^t$	Transmission power of UAV for MD i's signal
P_{UAV}	Maximum transmit power of the UAV
$\mathrm{PL}_{\mathrm{B}}^t$	Average path loss of UAV-BS transmission
PL_i^t	Average path loss of MD i-UAV transmission
$P_{\mathrm{B},i}^t$	Received power of the BS of MD i's signal
γ_i^t	SNR of the MD i-BS link
R_i^t	Data rate of the MD i-BS link
W	Bandwidth

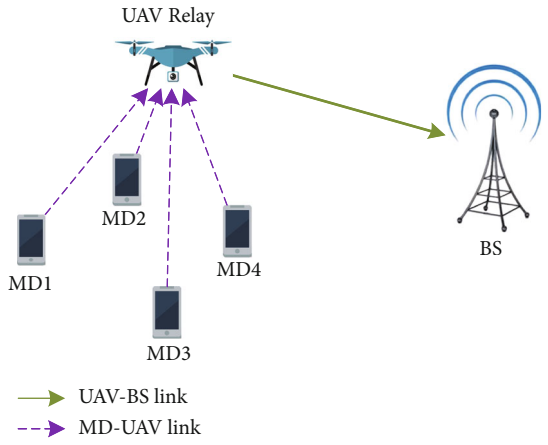


Figure 1: System model for AF-UAV relay-assisted multi-MD uplink communication.

no power control of MDs. The average received power of the UAV from MD i is given by

$$P_{i,\mathrm{U}}^t = \frac{P_{\mathrm{T}}}{\mathrm{PL}_i^t}, \tag{5}$$

where P_{T} is the transmission power of each MD, which can be considered as a content. The received signal of UAV relay from MD i is expressed as

$$y_{i,\mathrm{U}}^t = \sqrt{P_{\mathrm{T}}\mathrm{PL}_i^t}X_i^t + n_{\mathrm{U}}^t, \tag{6}$$

where X_i^t is the signal of unit energy from MD i and n_{U}^t is the additive white Gaussian noise (AWGN) received at the UAV relay, which satisfies Gaussian distribution with zero mean and N_0 as variance.

After receiving the signals from the MDs, the signals are amplified and forwarded to the BS by the UAV relay. We assume that the communication link for each MD occupies an independent channel, and there is no interchannel interference. In time slot t, let G_i^t be the signal amplification coefficient of the UAV relay for MD i's signal and $P_{i,\mathrm{B}}^t$ be the transmission power of the UAV relay for MD i's signal. The following relation holds:

$$P_{i,\mathrm{B}}^t = \left(G_i^t\right)^2\left(P_{i,\mathrm{U}}^t + N_0\right). \tag{7}$$

We assume that the maximum transmit power of the UAV is P_{UAV}. Therefore, the signal amplification coefficients for all the MDs satisfy

$$\sum_{i=1}^{K}\left(G_i^t\right)^2\left(\frac{P_{\mathrm{T}}}{\mathrm{PL}_i^t} + N_0\right) \le P_{\mathrm{UAV}}, \tag{8}$$

where N_0 is the AWGN power.

Similar to the MD-UAV transmission, the UAV-BS transmission also follows the air-to-ground channel model. We denote the average large-scale path loss of the UAV-BS transmission by $\mathrm{PL}_{\mathrm{B}}^t$, which can also be obtained by equations (1)–(4). The received power of the BS of MD i's signal can be expressed as

$$P_{\mathrm{B},i}^t = \frac{P_{i,\mathrm{B}}^t}{\mathrm{PL}_{\mathrm{B}}^t}. \tag{9}$$

The received signal from MD i to the BS can be expressed as

$$y_{i,\mathrm{B}}^t = G_i^t\sqrt{\frac{P_{\mathrm{T}}}{\mathrm{PL}_i^t\mathrm{PL}_{\mathrm{B}}^t}}X_i^t + \frac{G_i^t n_{\mathrm{U}}^t}{\sqrt{\mathrm{PL}_{\mathrm{B}}^t}} + n_{\mathrm{B}}^t, \tag{10}$$

where n_{B}^t is the noise received at the BS. According to (10), the joint signal-to-noise ratio (SNR) of the two-step uplink transmission between MD i and the BS is given by

$$\gamma_i^t = \frac{P_{\mathrm{T}}\left(G_i^t\right)^2/\mathrm{PL}_i^t\mathrm{PL}_{\mathrm{B}}^t}{\left(G_i^t\right)^2 N_0/\mathrm{PL}_{\mathrm{B}}^t + N_0}. \tag{11}$$

The data rate of the uplink transmission from MD i to the BS can be expressed as

$$R_i^t = W\log_2\left(1+\gamma_i^t\right), \tag{12}$$

where W is the bandwidth that can be considered as fixed value.

Our objective is to maximize the sum-rate of all the MDs by optimizing both the UAV deployment $\mathscr{L} = \{L(t)\}$ and

the amplification coefficients $\mathcal{G} = \{G_i^t\}, i = 1, \cdots K$ for K MDs in T time slots. The problem can be formulated by

$$\min_{\mathcal{L},\mathcal{G}} \quad \sum_{t=1}^{T}\sum_{i=1}^{K} R_i^t \tag{13a}$$

$$\text{s.t.} \quad \sum_{i=1}^{K} \left(G_i^t\right)^2 \left(\frac{P_{\mathrm{T}}}{\mathrm{PL}_i^t} + N_0\right) \le P_{\mathrm{UAV}}, \quad t = 1, \cdots, T \tag{13b}$$

$$G_i^t \ge 0, \quad t = 1, \cdots, T, i = 1, \cdots K \tag{13c}$$

$$v(t) \le v_{\max}, \tag{13d}$$

where (13b) and (13c) are the power constraints for UAV and MD and (13d) shows the UAV mobility constraint.

3. UAV Deployment and Amplification Coefficient Optimization

This section proposes a solution to the aforementioned problem of optimizing the deployment of UAV relay and amplification coefficients for the MDs jointly. Problem (13) is nonconvex with respect to $\mathcal{L}$ and $\mathcal{G}$. To solve this problem, we decouple (13) into two subproblems: UAV deployment and amplification coefficient optimization, and propose an iterative algorithm to solve them jointly. The convergency and complexity analyses are then followed.

3.1. UAV Deployment. In this part, we consider the amplification coefficients to be fixed and design the deployment of the UAV. The subproblem can be shown as

$$\min_{\mathcal{L}} \quad \sum_{t=1}^{T}\sum_{i=1}^{K} R_i^t \tag{14a}$$

$$\text{s.t.} \quad v(t) \le v_{\max} \tag{14b}$$

To solve problem (14), we first discuss the convexity of the sum-rate with respect to the location of the UAV.

Theorem 1. *The sum-rate of all the MDs is a concave function with respect to the location of the UAV approximately.*

Proof. See the appendix. □

To achieve the optimal UAV deployment, we propose a gradient ascent method as follows. Since the maximum UAV velocity in each time slot $v_{\max}$ is much shorter than the transmission distance d_i^t and d_{B}^t, we first set the UAV velocity as $v(t) = v_{\max}$. We assume that the UAV is at a random location $L^0(t) = (X^0(t), Y^0(t), H^0(t))$ initially and adjusts its location in a sequence of time slots. In each time slot, the UAV is moved along the direction with the maximum sum-rate ascent velocity, i.e.,$\nabla\sum_{i=1}^{K} R_i^t = ((\sum_{i=1}^{K}\partial R_i^t)/\partial x|_{y=Y^0(t),h=H^0(t)}, (\sum_{i=1}^{K}\partial R_i^t)/\partial y|_{x=X^0(t),h=H^0(t)}, (\sum_{i=1}^{K}\partial R_i^t)/\partial h|_{x=X^0(t),y=Y^0(t)})$. The location of the UAV is then adjusted to $L^1(t) = (X^1(t), Y^1(t), H^1(t))$, with $X^1(t) = X^0(t) + v|(\sum_{i=1}^{K}\partial R_i^t)/\partial x|^2/(|(\sum_{i=1}^{K}\partial R_i^t)/\partial x|^2 + |(\sum_{i=1}^{K}\partial R_i^t)/\partial y|^2 + |(\sum_{i=1}^{K}\partial R_i^t)/\partial h|^2)$, $Y^1(t) = Y^0(t) + v|(\sum_{i=1}^{K}\partial R_i^t)/\partial y|^2/(|(\sum_{i=1}^{K}\partial R_i^t)/\partial x|^2 + |(\sum_{i=1}^{K}\partial R_i^t)/\partial y|^2 + |(\sum_{i=1}^{K}\partial R_i^t)/\partial h|^2)$, and $H^1(t) = H^0(t) + v|(\sum_{i=1}^{K}\partial R_i^t)/\partial h|^2/(|(\sum_{i=1}^{K}\partial R_i^t)/\partial x|^2 + |(\sum_{i=1}^{K}\partial R_i^t)/\partial y|^2 + |(\sum_{i=1}^{K}\partial R_i^t)/\partial h|^2)$. We also set a minimum gradient threshold $\delta \longrightarrow 0^+$. When $\nabla\sum_{i=1}^{K} R_i^t < \delta$, it is regarded that the maximum sum-rate is achieved, and the UAV hovers at the optimal location.

Since $\sum_{i=1}^{K} R_i^t$ is a concave function with respect to the location of the UAV, when the locations of the MDs are assumed to be fixed, the UAV can approach the optimal location in finite time slots, with the error being no larger than v. If the locations of the MDs are not fixed, the UAV adjusts the direction of the gradient dynamically according to the locations of the MDs in the current time slot.

3.2. Amplification Coefficient Optimization. In this part, we design the amplification coefficients in each time slot, with the location of the UAV $L(t)$ given. The amplification coefficient subproblem can be given as

$$\min_{\mathcal{G}} \quad \sum_{i=1}^{K} R_i^t \tag{15a}$$

$$\text{s.t.} \quad \sum_{i=1}^{K} \left(G_i^t\right)^2 \left(\frac{P_{\mathrm{T}}}{\mathrm{PL}_i^t} + N_0\right) \le P_{\mathrm{UAV}}, \quad t = 1, \cdots, T \tag{15b}$$

$$G_i^t \ge 0, \quad t = 1, \cdots, T, i = 1, \cdots K \tag{15c}$$

As shown in (7), variable G_i^t is a function of $P_{\mathrm{U},i}^t$ when the locations between the UAV and MD i are given. Therefore, problem (15) can be converted to the following UAV transmission power optimization problem

$$\min_{P_{\mathrm{U},i}^t} \quad \sum_{i=1}^{K} R_i^t \tag{16a}$$

$$\text{s.t.} \quad \sum_{i=1}^{K} P_{i,\mathrm{B}}^t \le P_{\mathrm{UAV}}, \quad t = 1, \cdots, T \tag{16b}$$

$$P_{i,\mathrm{B}}^t \ge 0, \quad t = 1, \cdots, T, i = 1, \cdots K \tag{16c}$$

Problem (16) is convex and can be solved with water filling algorithm proposed in [19]. The optimal power allocation strategy can be expressed as

$$P_{i,\mathrm{B}}^{t,\mathrm{opt}} = \left[\lambda - \frac{1}{\mathcal{H}_i/N_0}\right]^+, \tag{17}$$

where

$$\mathcal{H}_i = \frac{P_{\mathrm{T}}/\mathrm{PL}_i^t\mathrm{PL}_{\mathrm{B}}^t}{N_0/\mathrm{PL}_{\mathrm{B}}^t + N_0\left(P_{\mathrm{T}}/\mathrm{PL}_i^t + N_0\right)} \tag{18}$$

is the equivalent channel gain of the MD-UAV-BS link, and

$$\lambda = \frac{1}{K}\left(P_{\mathrm{UAV}} + \sum_{i=1}^{K}\frac{1}{\mathcal{H}_i/N_0}\right) \tag{19}$$

is the water-filling level. The optimal amplification coefficient is solved as

$$G_i^{t,\mathrm{opt}} = \sqrt{\frac{P_{i,\mathrm{B}}^{t,\mathrm{opt}}}{P_{\mathrm{T}}/\mathrm{PL}_i^t + N_0}} \tag{20}$$

3.3. Algorithm Summary. In this part, the proposed algorithm that jointly optimizes UAV deployment and amplification coefficients is summarized as follows. In each iteration, give the initial location of the UAV, and the optimal amplification coefficients can be solved as proposed in Section 3.2. Afterwards, the UAV deployment is adjusted by moving along its trajectory as proposed in Section 3.1 with the amplification coefficients given. We then update the location of the UAV for the amplification coefficient optimization accordingly. Since the transmission distance is much larger than the moving distance of the UAV in one iteration, the performance degradation of the amplification coefficient optimization caused by the change of the UAV location in one iteration can be neglected. Iterations of amplification coefficient optimization and UAV deployment design are proposed until the performance gain of an iteration is less than a threshold ω. The joint UAV deployment and amplification coefficient optimization algorithm is summarized in Algorithm 1. We denote the sum-rate of the network after the rth iteration by $\mathscr{R}(\mathscr{L}^r, \mathscr{G}^r)$.

3.4. Algorithm Analysis. In this part, we analyse the convergency and complexity of the proposed algorithm.

Theorem 2. *The proposed UAV deployment and amplification coefficient optimization algorithm is convergent.*

Proof. In the $(r+1)$th iteration, we first find the optimal solution to the amplification coefficients with the location of the UAV being $\mathscr{L}^r$. Therefore, we have

$$\mathscr{R}\left(\mathscr{L}^r, \mathscr{G}^{r+1}\right) \geq \mathscr{R}(\mathscr{L}^r, \mathscr{G}^r), \tag{21}$$

i.e., the sum-rate does not decrease with amplification coefficient optimization in the $(r+1)$th iteration. When designing the UAV deployment, we give the optimal UAV moving trajectory $\mathscr{L}^{r+1}$ with the amplification coefficients being $\mathscr{G}^{r+1}$, and thus, we have

$$\mathscr{R}\left(\mathscr{L}^{r+1}, \mathscr{G}^{r+1}\right) \geq \mathscr{R}\left(\mathscr{L}^r, \mathscr{G}^{r+1}\right). \tag{22}$$

□

Combining (2) and (2), we have the following inequation:

$$\mathscr{R}\left(\mathscr{L}^{r+1}, \mathscr{G}^{r+1}\right) \geq \mathscr{R}\left(\mathscr{L}^r, \mathscr{G}^{r+1}\right) \geq \mathscr{R}(\mathscr{L}^r, \mathscr{G}^r). \tag{23}$$

As shown in (2), in each iteration, the objective function does not decrease. In the meanwhile, such a network has a capacity bound, and the uplink sum-rate cannot increase unlimitedly with iterations of deployment design and amplification coefficient optimization. Therefore, the objective function is upper-bounded and will converge to a stable solution in limited iterations; i.e., the proposed UAV deployment and amplification coefficient optimization algorithm is convergent.

Theorem 3. *The complexity of the proposed UAV deployment and amplification coefficient optimization algorithm is* $O(K^2T)$.

Proof. In each time slot, the proposed UAV deployment and amplification coefficient optimization algorithm contains iterations of UAV deployment design and amplification coefficient optimization. In each iteration, the complexity of UAV deployment design, i.e., finding the gradient descent direction, is $O(1)$, while the complexity of the amplification coefficient optimization, i.e., the convex optimization progress, can be minimized to $O(K)$ [20]. Therefore, the complexity of each iteration is $O(K)$. The number of iterations in each time slot is determined by the sum-rate improvement in each iteration and the total sum-rate improvement. The sum-rate improvement in each iteration is no less than the given threshold δ, while the total sum-rate improvement is no more than linear, i.e., $O(K)$, with respect to the number of the MDs. In summary, the number of iterations increases no more than linear, i.e., $O(K)$, with respect to the number of the MDs. Therefore, the complexity of the proposed UAV deployment and amplification coefficient optimization algorithm in T time slots is $O(K) \times O(K) \times O(T) = O(K^2T)$. □

4. Simulation Results

In this section, the performance of Algorithm 1 is evaluated with simulations. We select the simulation parameters based on the 3GPP specification basis [3] and related existing works. The values of the key parameters in the simulation are listed in Table 2. The simulation is performed in Monte Carlo scheme, with each curve generated by averaging the results of 10^5 instances.

We provide three schemes in comparison with the proposed scheme: fixed location relay scheme, circular trajectory relay scheme, and equal power scheme. In the fixed location relay scheme, the UAV stays at the initial location in every time slot. In the circular trajectory relay scheme, the trajectory is a circle whose center is (250, 0, 0) and radium is 100. The initial location of UAV is a random point on this circle and the moving distance is 1 meter for a time slot. The amplification coefficient design is the same as our proposed algorithm. In the equal power scheme, the transmit power of the UAV is equally allocated to every user, regardless of the channel quality, and the UAV deployment is designed the same as the proposed scheme.

Figure 2 depicts the average sum-rate of different schemes with time axis. The sum-rate of the network increases with the location adjustment of the UAV and converges to the maximum value in about 350 time slots. The proposed UAV deployment can improve the sum-rate for about 80% when compared with the fixed location relay and is about 15 bit/s/Hz higher than the equal power scheme. When compared with the circular trajectory scheme, the proposed algorithm

1: **Initialize** Obtain the initial location of the UAV and the MDs
2: **While** $\mathscr{R}(\mathscr{L}^r, \mathscr{G}^r) - \mathscr{R}(\mathscr{L}^{r-1}, \mathscr{G}^{r-1}) > \omega$
3: **If** $\nabla \sum_{i=1}^{K} R_i^t \geq \delta$
4: Solve amplification coefficient optimization subproblem (15) for time slot t;
5: Solve UAV deployment optimization subproblem (15) for time slot t
6: Update the UAV location and the MDs' locations;

Algorithm 1: Joint UAV deployment and amplification coefficient optimization algorithm.

Table 2: Simulation parameters.

Variable	Value
Total time slots T	500
Number of MDs K	10
MD distribution range	$100 \times 100\,\mathrm{m}^2$
Average distance between the MDs and BS	500 m
BS location	0, 0, 50
Initial location of the UAV	0, 0, 100
Maximum UAV speed $v_{\max}$	1 m per time slot
Algorithm convergency threshold δ	-10^{-2}
Maximum UAV transmission power P_{UAV}	26 dBm
Noise variance N_0	-76 dBm
Path loss parameter η_{LoS}	1
Path loss parameter η_{NLoS}	20
Path loss parameter α	12
Path loss parameter β	0.135

improves the sum-rate for about 40% on average. The sum-rate performance of the proposed algorithm converges faster than the equal power one, which shows that the UAV can be more rapidly deployed with the proposed algorithm.

In Figure 3, we study the scenario with MDs moving randomly on the ground. The sum-rate is illustrated with different average MD speeds v_{MD}. It is shown that when the MD mobility is much lower than that of the UAV, the sum-rate of the network tends to converge to that of the static MDs, but with a longer time. When the mobility of the MDs is comparable to the UAV, the sum-rate of the network cannot converge to that of the static MDs because of the rapid variation of the optimal UAV location. However, the sum-rate can still be 60% higher than that of the fixed location relay, due to the proposed UAV deployment design method.

5. Extensions of UAV Relay Networks

After discussing the design of the UAV deployment and amplification coefficients, we present several promising study directions of UAV relay networks in this section. Three extensive scenarios, together with the corresponding open problems and potential solutions are listed below.

5.1. Cooperative UAV Relay Network. One promising study of the UAV relay network is the UAV cooperative relay communication design, in which multiple relay UAVs perform transmission cooperatively in a multihop mode. A cooperative UAV relay enables long distance transmission with high QoS requirements. In the rapid developing Internet of Things (IoT) networks, various applications with large data rate and long transmission distance requirements are emerging, e.g., live video streaming and extended reality. Such applications raise challenges on the conventional terrestrial cellular network for two reasons. First, the severe shadowing leads to a high probability of NLoS transmissions in terrestrial network. It is more difficult for the terrestrial relays to find a LoS transmission path than the UAV relay network, thus leading to higher large-scale fading. Second, the terrestrial relays are fixed or with low mobility, which are incapable to improve the transmission QoS by adjusting their locations dynamically. The UAV relay communication provides more flexible deployment and high LoS transmission possibility, thus improving the service coverage and transmission rate of the network.

By bringing in cooperative UAV relay communications into IoT networks, some new study aspects need to be further studied. Unlike the terrestrial communications in wireless sensing networks, the topology of the cooperative UAV relay network changes rapidly. Therefore, works on designing the routing protocol that suits the rapid changing topology of the cooperative UAVs should be further discussed. Recently, the designs of multihop transmission routing protocols for the cooperative UAV relays are emerging [21, 22]. However, most of the proposed protocols in existing works are heuristic, and the deep analysis on the routing protocol that considers the UAV buffer and onboard energy jointly has not been well discussed, which has the potential to significantly affect the performance in practical systems. In future works, the protocol that jointly considers the physical constraints and UAV deployment can be designed.

5.2. mmWave UAV Relay Network. mmWave has been considered as one of the most important evolution directions in 5G and the upcoming 6G networks [23]. It offers a high integration of massive antennas that enables electronically steerable and highly directional beamforming, thus mitigating the interference for UAV communications [24], as multiple MDs can access the channel concurrently and be separated by spacial beams [25]. In this way, the interference can be eliminated, and the sum-rate can be significantly improved given the ultrawide mmWave bandwidth. To be specific, to obtain spacial diversity gain in multiantenna system, the distance between two antennas should be no less than half of the wavelength. In mmWave communications,

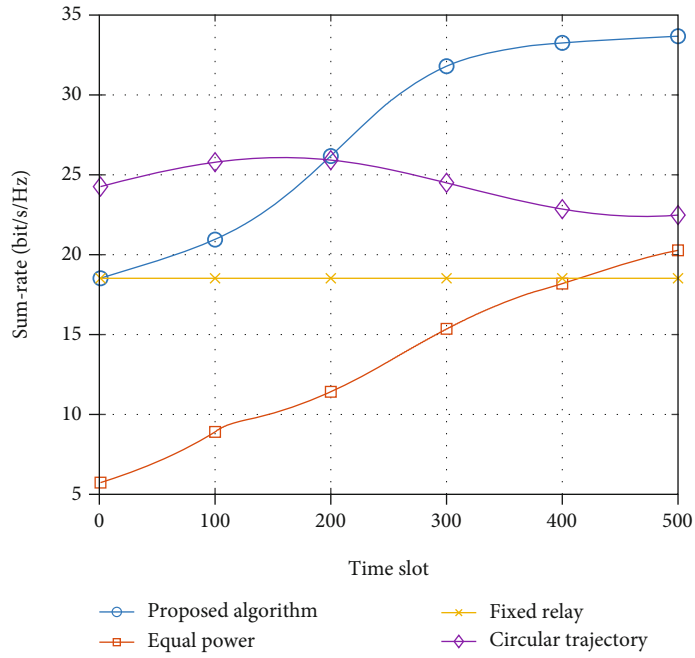


Figure 2: Sum-rate of different schemes.

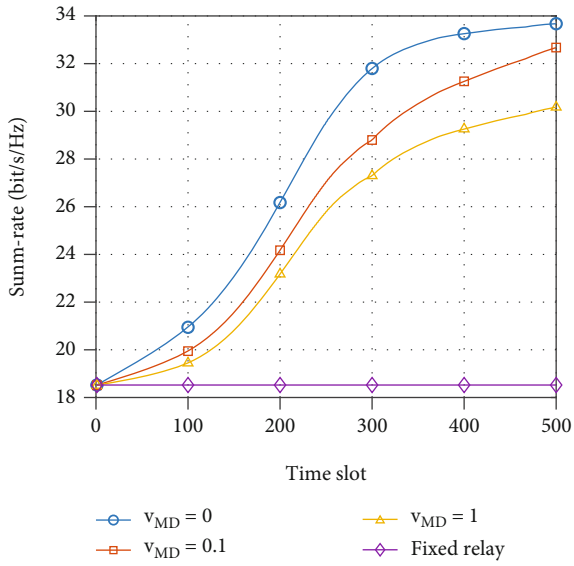


Figure 3: Sum-rate of different MD speeds.

the antennas can be much closer to each other than those in the conventional sub-6G systems, owning to a much shorter transmission wavelength. On this condition, a massive antenna array can be integrated in a small area, which is appropriate for UAV relays with strict space limitation. The massive antenna array is capable to achieve high array gain and reduce the propagation loss of the transmission links. For the above reasons, mmWave communications can strongly support the UAV relay communications with beamforming technique [26].

In addition to the beamforming technique and ultra-wide-spectrum resources, mmWave UAV communications have a few additional advantages. Due to the characteristic of weak scattering in the mmWave band, the mmWave channel has better performance of sparsity and directivity when compared with sub-6G communications. In particular, the LoS path is longstanding for the UAV relays with high altitude and can be actively created on demand via the movement of UAVs. The LoS component of the transmission links can be over 20 dB higher than that of the NLoS ones. Thus, the mmWave communications with directional beamforming give full play to the advantages of the LoS transmission paths for UAV relays. Moreover, the dynamic beam direction of the mmWave communications enables the highly mobile UAVs to adjust the transmission and reception timely, in order to obtain higher channel gain than that of the conventional full coverage scheme. Thus, the mmWave UAV relay communications have the capability to increase spectrum efficiency of the network. The above-mentioned beamforming, interference management, and spectrum efficiency improvement problems are promising studies in the mmWave UAV relay networks, which can be further studied in the upcoming researches.

5.3. NOMA for UAV Relay Network. The ultrahigh MD density poses huge pressure on the limited number of subbands in mmWave communications and the sparsity of UAV relay deployment. As a result, the OMA scheme may suffer severe congestion risks when massive MDs intend to perform data upload simultaneously. To tackle the challenges of access collision reduction and massive connectivity, NOMA has been raised as a promising solution, which allows the MDs to access the radio resources nonorthogonally. It is especially helpful in the uplink transmission in mmWave UAV relay communication system. As introduced in [27], NOMA achieves considerable performance gain when the channel gains of a paired MDs differs significantly. In the mmWave UAV relay network, UAV relays can adjust their reception

beams for a pair of uplink MDs to construct such a channel gain difference, thus fully exploring the potential performance gain of NOMA technique.

NOMA can be utilized in the UAV relay communications in massive connectivity scenarios, where multiple MDs access the channel nonorthogonally by either code domain [28] or power domain [29] multiplexing. Multiple MDs can improve the spectrum efficiency and sum-rate of the network by performing concurrent transmissions on the same channel. In order to cope with the cochannel interference in the nonorthogonal scheme, multiuser detection techniques such as successive interference cancellation can be utilized in the receivers, with which the superposed signals can be decoupled and demodulated, thus making this system practical. Due to the air-to-ground communication properties and the high mobility of the UAVs, the study of NOMA for UAV relay networks is different from that of the terrestrial ones in many aspects, such as power control, spectrum management, and signaling control, which should be further studied in future works.

6. Conclusions

In this paper, we consider that a multi-MD uplink network with a relay UAV amplifies and forwards the signals from the MDs to the BS with different amplification coefficients. With UAV speed and power constraints, a joint optimization to the UAV deployment design and amplification coefficient design is given. This paper analysed the convergence and complexity of the proposed algorithm. Simulation result shows that the sum-rate of the proposed solution outperforms fixed location relay and equal power allocation schemes significantly and can improve the sum-rate of mobile MDs. This paper also discussed the extensions and open problems of UAV relay communications based on the model of this paper, including UAV cooperation, mmWave, and NOMA.

Appendix

A.1. Proof of Theorem 1

Proof. According to (1) and (2), the path loss of the air-to-ground communication is negatively quadratic related to the transmission distance. Therefore, we consider the change of the path loss as a negatively quadratical function of the transmission distances, i.e., $\mathrm{PL}_i^t \propto (d_i^t)^{-2}, (d_{\mathrm{B}}^t)^{-2}$. We then substitute (1) into (1), and the uplink data rate of MD i can be approximated as

$$R_i^t = W \log_2\left(1 + \frac{P_{\mathrm{T}}\left(d_i^t\right)^{-2}\left(G_i^t\right)^2\left(d_{\mathrm{B}}^t\right)^{-2}}{N_0\left(G_i^t\right)^2\left(d_{\mathrm{B}}^t\right)^{-2} + N_0}\right). \tag{24}$$

□

It is shown in (2) that R_i^t is negatively related with both d_i^t and d_{B}^t. Therefore, the maximum value of R_i^t is achieved when $d_i^t + d_{\mathrm{B}}^t$ is minimized. Otherwise, the value of R_i^t can be improved by reducing the value of d_i^t or d_{B}^t. Let L be the minimum value of $d_i^t + d_{\mathrm{B}}^t$; when R_i^t is maximized, we have $d_{\mathrm{B}}^t = L - d_i^t$. When we substitute $d_{\mathrm{B}}^t = L - d_i^t$ into (2), R_i^t becomes a univariate function of d_i^t, which can be expressed as

$$R_i^t = W \log_2\left(1 + \frac{P_{\mathrm{T}}\left(G_i^t\right)^2}{N_0\left(G_i^t\right)^2\left(d_i^t\right)^2 + N_0\left(d_i^t\right)^2\left(L - d_i^t\right)^2}\right). \tag{25}$$

The derivative of R_i^t with respect to d_i^t is given as

$$\frac{dR_i^t}{d\left(d_i^t\right)} = \frac{-WA'}{\ln 2\left(A + P_{\mathrm{T}}\left(G_i^t\right)^2\right)A}, \tag{26}$$

where

$$A = N_0 d_i^t\left(\left(G_i^t\right)^2 + \left(L - d_i^t\right)^2\right), \tag{27}$$

$$A' = \frac{dA}{d\left(d_i^t\right)} = 4N_0\left(d_i^t\right)^3 - 6N_0\left(d_i^t\right)^2 L + 2N_0\left(L^2 + \left(G_i^t\right)^2\right)d_i^t. \tag{28}$$

The second-order derivative of R_i^t with respect to d_i^t is given as

$$\frac{d^2R_i^t}{d\left(d_i^t\right)^2} = \frac{-WP_{\mathrm{T}}\left(G_i^t\right)^2}{\ln 2} \times \frac{A''\left(A + P_{\mathrm{T}}\left(G_i^t\right)^2\right)A - A'\left(A'A + A'P_T\left(G_i^t\right)^2 + A'A\right)}{\left(A + P_{\mathrm{T}}\left(G_i^t\right)^2\right)^2 A^2}, \tag{29}$$

where

$$A'' = \frac{d^2A}{d\left(d_i^t\right)^2} = 12N_0\left(d_i^t\right)^2 - 12N_0 d_i^t L + 2N_0\left(L^2 + \left(G_i^t\right)^2\right). \tag{30}$$

We then substitute (27), (28), and (3) into (29), and it can be proved that $d^2R_i^t/d\left(d_i^t\right)^2 < 0$; i.e., the data rate of MD i is a concave function of the location of the UAV. According to the properties of concave function, the sum of all the uplink data rate $\sum_{i=1}^{K} R_i^t$ is also a concave function of the UAV deployment.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] L. Gupta, R. Jain, and G. Vaszkun, "Survey of important issues in UAV communication networks," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 2, pp. 1123–1152, 2015.

[2] M. Erdelj, E. Natalizio, K. R. Chowdhury, and I. F. Akyildiz, "Help from the sky: leveraging UAVs for disaster management," *IEEE Pervasive Computing*, vol. 16, no. 1, pp. 24–32, 2017.

[3] *Enhanced LTE support for aerial vehicles, Release 15, document 3GPP*, vol. 36, p. 777, 2017.

[4] Y. Zeng, R. Zhang, and T. J. Lim, "Wireless communications with unmanned aerial vehicles: opportunities and challenges," *IEEE Communications Magazine*, vol. 54, no. 5, pp. 36–42, 2016.

[5] L. Wood, *Global UAV payload and subsystems market analysis of growth, trends & Forecast 2018-2023*, 2018, https://www.researchandmarkets.com/.

[6] Y. Zhou, N. Cheng, N. Lu, and X. S. Shen, "Multi-UAV-aided networks: aerial-ground cooperative vehicular networking architecture," *Ieee Vehicular Technology Magazine*, vol. 10, no. 4, pp. 36–44, 2015.

[7] N. Zhao, W. Lu, M. Sheng et al., "UAV-assisted emergency networks in disasters," *IEEE Wireless Communications*, vol. 26, no. 1, pp. 45–51, 2019.

[8] Z. Xue, Q. Wu, Z. Feng, C. Zhong, and G. Ding, *Sum rate maximization in UAV-enabled mobile relay networks*, IEEE WCSP, Hangzhou, China, 2018.

[9] S. Zhang, H. Zhang, Q. He, K. Bian, and L. Song, "Joint trajectory and power optimization for UAV relay networks," *IEEE Communications Letters*, vol. 22, no. 1, pp. 161–164, 2018.

[10] S. Park, C. S. Shin, D. Jeong, and H. Lee, "DroneNetX: network reconstruction through connectivity probing and relay deployment by multiple UAVs in ad hoc networks," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 11, pp. 11192–11207, 2018.

[11] J. Lyu, Y. Zeng, and R. Zhang, "Spectrum sharing and cyclical multiple access in UAV-aided cellular offloading," in *GLOBECOM 2017-2017 IEEE Global Communications Conference*, pp. 1–6, Singapore, 2017.

[12] C. You and R. Zhang, "Hybrid offline-online design for UAV-enabled data harvesting in probabilistic LoS channels," *IEEE Transactions on Wireless Communications*, vol. 19, no. 6, pp. 3753–3768, 2020.

[13] K. Chen, Y. Wang, J. Zhao, X. Wang, and Z. Fei, "URLLC-oriented joint power control and resource allocation in UAV-assisted networks," *IEEE Internet of Things Journal*, vol. 8, no. 12, pp. 10103–10116, 2021.

[14] H. Zhang, L. Song, and Z. Han, *Unmanned Aerial Vehicle Applications over Cellular Networks for 5G and Beyond*, Switzerland: Springer, 2020.

[15] Y. Kawamoto, H. Nishiyama, N. Kato, F. Ono, and R. Miura, "Toward future unmanned aerial vehicle networks: architecture, resource allocation and field experiments," *IEEE Wireless Communications*, vol. 26, no. 1, pp. 94–99, 2019.

[16] A. A. Khuwaja, G. Zheng, Y. Chen, and W. Feng, "Optimum deployment of multiple UAVs for coverage area maximization in the presence of co-channel," *IEEE Access*, vol. 7, pp. 85203–85212, 2019.

[17] A. Al-Hourani, S. Kandeepan, and A. Jamalipour, "Modeling air-toground path loss for low altitude platforms in urban environments," in *IEEE global communications conference*, Austin, TX, USA, 2014.

[18] D. Athukoralage, I. Guvenc, W. Saad, and M. Bennis, "Regret based learning for UAV asisted LTE-U/WiFi public safety networks," in *IEEE Global Communications Conference (GLOBECOM)*, Washington, DC, USA, 2016.

[19] W. Yu, W. Yhee, S. Boyd, and J. Cioffi, "Iterative water-filling for Gaussian vector multiple-access channels," *IEEE Transactions on Information Theory*, vol. 50, no. 1, pp. 145–152, 2004.

[20] S. Bubeck, "Convex optimization: algorithms and complexity," *Foundations and Trends® in Machine Learning*, vol. 8, 2015.

[21] Z. Zheng, A. K. Sangaiah, and T. Wang, "Adaptive communication protocols in flying ad hoc network," *IEEE Transactions on Information Theory*, vol. 56, no. 1, pp. 136–142, 2018.

[22] Q. Zhang, M. Jiang, Z. Feng, W. Li, W. Zhang, and M. Pan, "IoT enabled UAV: network architecture and routing algorithm," *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 3727–3742, 2019.

[23] A. Ghosh, A. Maeder, M. Baker, and D. Chandramouli, "5G evolution: a view on 5G cellular technology beyond 3GPP release 15," *IEEE Access*, vol. 7, pp. 127639–127651, 2019.

[24] L. Wang, Y. L. Che, J. Long, L. Duan, and K. Wu, "Multiple access mmWave design for UAV-aided 5G communications," *IEEE Wireless Communications*, vol. 26, no. 1, pp. 64–71, 2019.

[25] C. Sun, X. Gao, S. Jin, M. Matthaiou, Z. Ding, and C. Xiao, "Beam division multiple access transmission for massive MIMO communications," *IEEE Transactions on Communication*, vol. 63, no. 6, pp. 2170–2184, 2015.

[26] Z. Xiao, L. Zhu, and X. G. Xia, "UAV communications with millimeter-wave beamforming: potentials, scenarios, and challenges," *China Communications*, vol. 17, no. 9, pp. 147–166, 2020.

[27] L. Dai, B. Wang, Y. Yuan, S. Han, I. Chih-lin, and Z. Wang, "Non-orthogonal multiple access for 5G: solutions, challenges, opportunities, and future research trends," *IEEE Communications Magazine*, vol. 53, no. 9, pp. 74–81, 2015.

[28] B. Di, L. Song, Y. Li, and G. Y. Li, "TCM-NOMA: joint multi-user codeword design and detection in trellis-coded modulation-based NOMA for beyond 5G," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 3, pp. 766–780, 2019.

[29] S. Zhang, B. Di, L. Song, and Y. Li, "Sub-channel and power allocation for non-orthogonal multiple access relay networks with amplify-and-forward protocol," *IEEE Transactions on Wireless Communications*, vol. 16, no. 4, pp. 2249–2261, 2017.

Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 7476225, 11 pages
https://doi.org/10.1155/2022/7476225

Research Article

The Application of Wireless Sensor Technology of Internet of Things in Korean Language Teaching

Yajie Bi

School of Humanities and Education, Xijing University, Xi'an, 710123 Shaanxi, China

Correspondence should be addressed to Yajie Bi; 20130083@xijing.edu.cn

Received 2 March 2022; Revised 14 April 2022; Accepted 25 April 2022; Published 26 May 2022

Academic Editor: Lisheng Fan

The Internet of Things sensor technology network can be regarded as consisting of three parts: data acquisition network, data distribution network, and control management center. Sensor technology has the comprehensive processing ability to obtain various information signals, and the technology is integrated into the sensor equipment, so that the sensor technology is connected to form a sensor network with comprehensive information processing capability, which has been widely used in foreign language teaching in recent years. This paper analyzes the shortcomings of traditional Korean language teaching methods in the new situation in detail. It starts from the actual situation and focuses on stimulating students' interest and proposes a Korean language teaching method based on the Internet of Things wireless sensor technology. This article focuses on analyzing the application of the wireless sensor technology of the Internet of Things in Korean language teaching and applies this method in practical teaching. It verifies the feasibility and effectiveness of the technology for Korean language teaching through the analysis and comparison of the collected data. In this paper, the experiment shows that the application of the wireless sensor technology of the Internet of Things to the Korean language teaching can improve the students' grades by about 30%. This stimulates students' interest in learning very well and is of great help to improve students' learning ability, oral communication ability, and Korean thinking ability, which stimulates students' interest in learning and improves students' Korean listening, speaking, reading, writing, and comprehensive learning abilities.

1. Introduction

The rise and development of Internet of Things technology have made it more and more widely used in other fields. As one of the technologies in the Internet of Things, which has a huge effect on education and teaching, wireless sensor technology is one of the modern sciences and technologies to promote educational reform, and its application in the field of education has become an irresistible trend [1]. And as the Korean trend has gradually penetrated into the lives of the people in China, major universities have successively opened Korean language learning courses in order to conform to the trend of the times [2]. However, there are still big problems in traditional Korean language teaching, such as lack of Korean language teaching resources, poor teaching effect, and difficult teaching environment to meet the standards of modern education. In addition, the evaluation of Korean language teaching is also a big problem. It is difficult for teachers to accurately evaluate students' Korean learning effect and overall Korean teaching level, so they cannot formulate accurate teaching plans. Therefore, it is imperative to promote the reform of Korean language teaching with the help of the Internet of Things and wireless sensor technology.

By applying the Internet of Things to Korean language teaching, students and teachers can search for abundant teaching and learning resources through the Internet of Things, to promote the sharing of Korean language teaching and learning resources, and improve the problem of lack of teaching resources in Korean language teaching. It promotes the reform of Korean teaching mode, changes the teacher-led teaching mode in traditional Korean teaching, and promotes the innovation of Korean teaching methods [3]. The application of wireless sensor technology to the Korean language teaching can effectively promote the formation of the Korean language learning environment, improve the single

phenomenon of the Korean language teaching curriculum, and improve the rationality of the teaching curriculum. It can also enable students to speak Korean with their mouths open and improve their oral Korean proficiency. It not only focuses on cultivating students' listening, speaking, reading, and writing skills but also improves students' driving ability and improves classroom teaching efficiency. It changes the Korean language teaching mode to improve students' motivation to learn Korean and guide students to set correct learning goals so that they can have enough motivation to learn the language.

In order to promote the change of the teaching mode of Korean and improve the enthusiasm of students to learn Korean, many scholars have discussed and researched the current Korean teaching and learning. Among them, Tong is exploring the possibility of applying flipped learning, which has recently attracted attention in the world of education, to oral language education in order to improve the communication skills of Korean language learners in Chinese universities. He proposed the problems of oral Korean education in Chinese universities and the actual teaching model of flipped learning oral teaching [4]. In order to improve the Korean language ability of learners from multicultural backgrounds, Park has studied effective reading education and teaching programs and confirmed the learners' self-esteem and learning motivation induction and other meaningful theories for Korean language learning [5]. Lee believes that the myth of Tangun is a practical colloquial literary text that is useful for teaching Korean language and culture. But in practice, most textbooks fail to activate the potential power of myth as oral literature. Therefore, beneficial features such as dynamic communication of communicative competence and active dissemination of cultural knowledge of texts as oral literature rarely appear in language classrooms [6]. Kwon examines the teaching and assessment methods for learning Korean music in elementary schools and details the activities used in and out of school and students' expected learning outcomes by increasing the number of achievement standards [7]. Although these studies have certain implications for Korean language teaching and learning, these studies only discuss Korean language teaching and learning in theory, without specific experimental support and data support, and we cannot know whether it has significantly improved the traditional Korean teaching mode and whether it can promote the improvement of students' comprehensive ability.

The research on Korean language teaching in this paper has the following innovations: (1) this paper summarizes and analyzes the existing problems of Korean language teaching by consulting relevant literature and materials and evaluates the existing Chinese language teaching mode in combination with the actual situation; (2) bring the Internet of Things into Korean teaching, and use the Internet of Things to promote the sharing of Korean teaching resources and the change of Korean teaching mode; (3) use the wireless sensor technology to promote the generation of the Korean language environment and introduce the mother tongue-like learning method into the Korean language learning, in order to improve the students' enthusiasm for Korean language learning; and (4) carry out experiments on the application of wireless sensor technology in Korean language teaching to test whether it can promote Korean language teaching.

2. Discuss the Application Method of Wireless Sensor Technology in Korean Language Teaching

2.1. Wireless Sensor Technology. With the continuous development and progress of related disciplines, the sensor network also has the comprehensive processing ability to obtain a variety of information signals. And it is associated with sensor control to form a sensor network with information synthesis and processing capabilities [8]. The wireless sensor network structure is shown in Figure 1.

Wireless sensor network is a distributed network, which is a network form formed by freely conducting and combining a large number of sensor nodes through radio technology. It can be connected to the Internet through a multihop self-organizing network formed by wireless communication [9]. With the help of wireless sensor technology, Korean language teaching can expand the scope of acquisition of Korean language teaching resources and obtain more abundant learning resources. Because the sensor technology needs to be connected to the wireless network, it is necessary to test the distance between the network and the wireless sensor device. Our traditional ranging method uses a three-dimensional centroid positioning algorithm, which is also to allow the sensor to sense the rule of the wireless network so that the two can be connected, so that the sensor device can obtain the required Korean teaching resources through the wireless network [10]. Then, the principle of the 3D centroid positioning algorithm is as follows.

We assume that the centroid is the node at the center of the 2D sensing area, and the 2D coordinate of this node is (s_i, y_i). It is assumed that there are w secondary network nodes in the sensor network, and it is assumed that the two-dimensional coordinate of a secondary network node is (s_n, y_n), and the coordinate of other unknown nodes is (s_0, y_0). Then, the following algorithm is obtained by the position coordinates of the secondary network nodes:

$$s_0 = \sum_{i=2}^{N} \frac{s_i}{w}, \tag{1}$$

$$y_0 = \sum_{i=1}^{N} \frac{y_i}{w}. \tag{2}$$

Then, the traditional two-dimensional space is brought into the three-dimensional space to calculate the position coordinate (s_i, y_i, t_i) of the three-dimensional center of mass; similarly, it can be concluded that the three-dimensional coordinates of other unknown nodes are (s_0, y_0, t_0), and the three-dimensional coordinates are obtained as follows:

$$s_0 = \sum_{i=1}^{N} \frac{s_i}{N}, \tag{3}$$

$$y_0 = \sum_{i=1}^{N} \frac{y_i}{N}, \tag{4}$$

$$t_0 = \sum_{i=1}^{N} \frac{t_i}{N}. \tag{5}$$

Although this method can quickly calculate the location of network nodes, the signal recovery of wireless communication network is affected by many other factors. Therefore, there will be some errors in the calculation of these unknown nodes, which will cause network delays and make the acquisition of teaching resources slower. Therefore, when the sensor device is connected to the Internet, it is necessary to carefully check whether a network node is normal, which will generate a lot of cumbersome procedures. Wireless sensor network is a multihop self-organizing network that can be formed through the Internet; then, we can use this multihop feature to improve the accuracy of its calculation of unknown node coordinates [11]. The 3D DV-hop algorithm is a positioning method based on distance vector calculation of hops. When the wireless sensor device cannot connect to the wireless network node for a long time, it will automatically jump to find the next network node. It keeps jumping until it finds the best network node to connect to the wireless communication network. In the three-dimensional DV-hop algorithm, let the number of hops be a and the hop distance be l. When we calculate the average hop distance between network nodes, the blood medicine relies on the minimum hop distance in the network node and the location information of one of the network nodes. The algorithm of the average hop distance R is as follows:

$$R = \frac{\sum_{i=n} \sqrt{(s_i - s_n)^2 - (y_i - y_n)^2 - (t_i - t_n)^2}}{a * l}. \tag{6}$$

After the average hop distance of the sensing device, it is necessary to consider how to reduce the error. Because the more the number of jumps, the error will gradually accumulate and become larger, so it is necessary to use the PSO algorithm to reduce the error and improve the accuracy of the network connection. The PSO algorithm can help the sensing device find the optimal location of the network node. It can be seen that the simplified mathematical model of the network node is as follows:

$$a(s_0 + 1) = wf(s) + n(s) - f(s), \tag{7}$$

$$l(y_0 + 1) = wf(y) + n(y) - f(y), \tag{8}$$

$$(t_0 + 1) = wf(t) - f(t). \tag{9}$$

In the above formula, f represents the speed of the jump. We set the position of the optimal network node P to be (s_p, y_p, t_p); then, the position coordinate (s_0, y_0, t_0) of the unknown node needs to gradually approach P by reducing the error; then, the error reduction process is as follows:

$$\text{Error}_{s_0} = [wf(s) + n(s) - f(s)] * (w_0 * c_0) \longrightarrow s_p, \tag{10}$$

$$\text{Error}_{y_0} = [wf(y) + n(y) - f(y)] * (w_0 * c_0) \longrightarrow y_p, \tag{11}$$

$$\text{Error}_{t_0} = [wf(t) - f(t)] * (w_0 * c_0) \longrightarrow t_p. \tag{12}$$

In the above formula, w_0 1 is the inertia coefficient, and c_0 is the acceleration coefficient, so that the position of the node at this position is close to the optimal position, and the sensing device can find a better network connection secondary node position. The nodes in the sensing area must be connected with the communication nodes of the wireless network, and there is a distance between the two nodes, as shown in Figure 2.

Figure 2(a) shows the TOA ranging principle, and (b) shows the time difference of arrival ranging method. If the TOA ranging method is used, we assume that the connection time of the two nodes is t, that is, the time from S4 to S3; the formula for calculating the distance g to be measured is as follows:

$$g_1 = \frac{[(s_1 - s_0)] * v}{2}, \tag{13}$$

$$g_1 = \frac{[(s_4 - s_3] * v}{2}. \tag{14}$$

Then,

$$g = \frac{(g_1 + g_2) * v}{2t}. \tag{15}$$

In the above formula, v is a certain value, which expresses the propagation speed of wireless signal in the case of quasielastic collision of elementary particles of matter. Although the TOA ranging method can be used to measure the distance, this method ignores a certain equipment response time, so there will be a certain error, and we can use the time difference of arrival ranging method to subdivide these. We can see that the speed of ultrasonic transmission is v_1 and the time is t_1 in Figure 2(a); while the transmission speed of radio communication is v_2 and the time is t_2, the calculation method of the distance to be measured is as follows:

$$g_1 = (s_1 - s_0) * \frac{v_2}{2} * t_2, \tag{16}$$

$$g_2 = (s_1 - s_0) * \frac{v_1}{2} * t_1. \tag{17}$$

Then,

$$g = (g_1 + g_2) * \frac{t_1 t_2}{v_1 - v_2}. \tag{18}$$

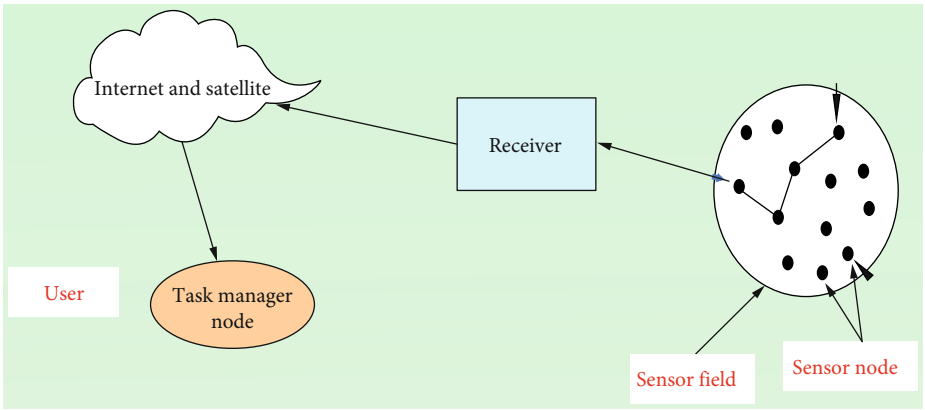


Figure 1: Wireless sensor network structure diagram.

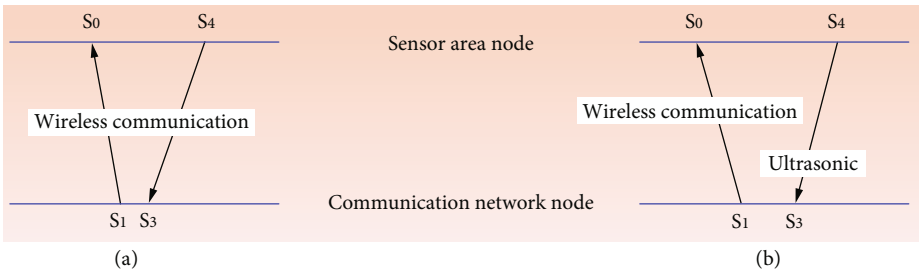


Figure 2: Ranging schematic.

The distance and position of the connection between the nodes in the sensing area and the nodes of the wireless network are obtained, which is helpful for us to adjust the connection between the sensing equipment and the wireless network. It can also better obtain online Korean learning resources and teaching resources and at the same time change the Korean teaching mode in colleges and universities.

2.2. Existing Korean Language Teaching in Colleges and Universities. Korean is not as popular as English, so there is a lack of teaching resources in Korean teaching in colleges and universities. Most of the teaching resources of students and teachers come from the textbooks provided by the school, and they continue to learn and teach repeatedly, and the teaching form is single. The traditional Korean teaching mode is shown in Figure 3.

In Figure 3, it can clearly be seen that teachers and students use textbooks issued by the school for after-school learning, so the teaching resources of Korean language teaching in domestic colleges and universities are very limited. In addition, the Korean textbooks of various colleges and universities are the same, which is the same every year, and there is no innovation in the textbooks, nor does it take into account whether the textbooks are suitable for the education of current college students [12]. Since Korean majors enter colleges and universities, they have been bound by the traditional education model, blindly adopting rigid learning methods to learn Korean and rote memorization, and plunge into language learning, regardless of other learning methods. Moreover, students cannot determine in advance what the ultimate goal of learning Korean is and what their future employment direction is [13]. The lack of flexibility makes learning Korean even more difficult. Some colleges and universities only set up majors for Korean language learning without more detailed subprofessional directions, which will also lead to students learning only some basic language knowledge and no knowledge in other fields except the language itself. Secondly, the level of Chinese and Korean teachers in colleges and universities is insufficient. Teachers are specialized personnel who cultivate the new generation of society and improve the quality of the nation. The social function of education is realized through teachers. With the so-called "famous teachers produce high apprentices," teachers' teaching level and orientation of teaching materials are directly related to students' professional development and learning effect [14, 15]. In colleges and universities, the number of Korean language teachers is sufficient, but the teachers' own teaching level and the ability of Korean language knowledge are not much different. In teaching, one-size-fits-all teaching methods are basically used, and individual teachers cannot form their own teaching directions, and at the same time, they generally lack professional knowledge other than Korean language, such as speaking, reading, and other teaching expertise, but also generally lack professional knowledge other than Korean language [16]. There is also a problem with teachers' teaching methods. At present, the teaching methods of Korean in colleges and universities are all teacher-based teaching methods. Teachers input a large amount of knowledge into students' brains, but what students have learned cannot be effectively output; teachers are

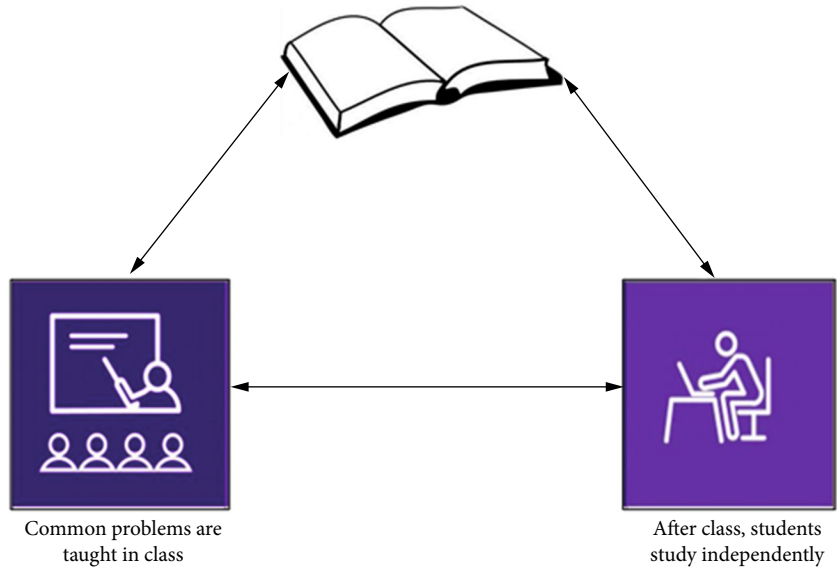


Figure 3: Traditional Korean teaching mode.

unable to achieve teaching goals and achieve teaching quality. The problems in Korean teaching are shown in Figure 4.

In addition, in colleges and universities, students' enthusiasm for learning is not high. The temptation of the outside world is more attractive to students, which will gradually wear down the willpower of students to learn, and only effective efforts will make people ignite continuous learning motivation. However, there is a big problem in the curriculum setting of the university. Students cannot see their own obvious progress in their learning, and their motivation to study is gradually wiped out. And when learning Korean, students will think about what aspects of Korean language can be linked to life after graduation, and whether it can play a role in future work and study is unclear. Students lack overall goals and are more lazy [17]. In addition, there is no specific Korean language environment, which makes the learning of Korean even more difficult. Students blindly receive the knowledge imparted by teachers, but there is no effective output, so that the learned knowledge will be gradually forgotten. Therefore, it is difficult for Korean language teaching in colleges and universities to achieve teaching goals and achieve teaching level.

2.3. Application of Wireless Sensor Technology in Korean Language Teaching. Korean language learning is all about listening, repeating, and communicating. The Korean immersion environment can promote the efficiency of students' Korean learning, and wireless sensor technology can improve students' listening, reading, and speaking skills in Korean. Learning a new language requires understanding its auditory and written forms, as well as the ability to communicate ideas through sound and speech. It sounds simple, but it is difficult to implement because learning Korean in a non-Korean-speaking country makes the immersive environment of Korean extremely lacking [18]. But IoT wireless sensing technology plays a big role in creating an immersive environment in Korean, as the sensor device system can simulate an immersive experience by using connected objects, as shown in Figure 5.

As shown in Figure 5, it is a smart device using wireless sensing technology. Students need to practice Korean and need to drive the sensor device system, and then, they can have intelligent Korean conversations. Students start speaking in Korean, and what they hear is relayed to the foreign media device. When the foreign media device receives the student's speech, it will be quickly transmitted to the smart device through the wireless network, and the smart device will transmit the sound back to the sensor device to realize Korean communication. If colleges and universities create classrooms with sensing devices, they can create a Korean learning environment for students. In addition, it can improve the motivation of students to learn Korean and greatly promote their comprehensive ability of listening, speaking, reading, and writing in Korean. It creates a simulated learning environment so that students can improve their Korean listening and speaking skills. At the same time, it can also improve students' interest in Chinese learning from the side, and the use of connected hardware in the simulation of the Korean environment helps guide learners.

The sensor equipment system can promote teachers to change their teaching methods, so that teachers can play an auxiliary role in the process of students' Korean learning, and promote students to learn autonomous learning, ensuring that students are learning Korean smoothly. In addition, teachers can give specific help to students in the process of autonomous learning and effectively teach students in accordance with their aptitude, so as to change the traditional teaching mode of teachers and make students become the center of the classroom. Under the simulation of the intelligent environment, the sensing device helps to monitor the students and promote their Korean language learning, thereby changing their learning inertia, and strengthen the online communication between teachers and students, so as to provide more influential learning guidance to students

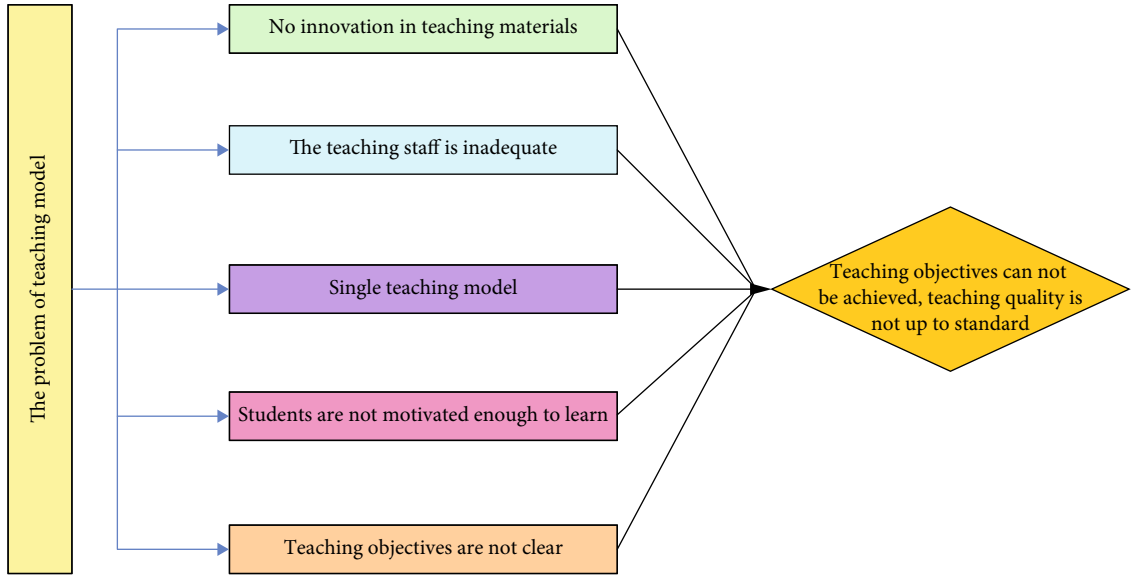


Figure 4: Problems in Korean language teaching.

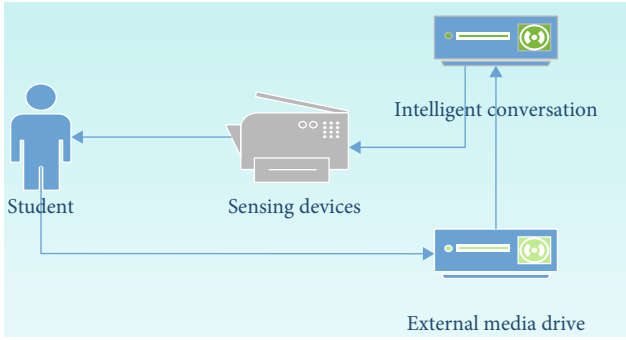


Figure 5: Sensing device system.

[19]. And the sensor equipment system has other functions as shown in Figure 6.

Shown in Figure 6 is the internal platform of the sensor equipment system, including the student's usual learning data and the recording platform of the learning situation, the teaching and learning resource platform, and the teacher-student exchange forum platform. In traditional Korean language teaching, teachers will first speak in Korean and then explain in Chinese. This practice is very unfavorable for students to master knowledge points and improve their learning ability and should be stopped. If the bilingual class tastes the same and cannot arouse students' interest, inertia will occur. Therefore, the sensor equipment technology can enable teachers to adjust the teaching method according to the students' specific Chinese learning situation, so as to improve the students' enthusiasm for learning. At the same time, it also improves the teaching ability and level of teachers and improves the teaching quality of Korean teaching [20]. It can also guide students to set correct goals and comprehensively improve their Korean learning ability and Korean language level, so that students can learn in an immersive Korean atmosphere and improve their comprehensive level of Korean [21].

3. Experiment and Analysis of the Teaching Effect of Wireless Sensor Technology in Korean Language Teaching

3.1. Experiment and Analysis of Teaching Effect

3.1.1. Experiment 1. This experiment will compare the scores of various aspects in a Korean major in a university. This time, the Korean class 1 and class 2 with the same teacher were selected. The number of students in the two classes is 25. Then, divide the two classes into groups of five, and ensure that the level between each group is basically the same. The first class of Korean uses the traditional teaching method, while the second class of Korean will use the wireless sensor equipment system for teaching, and then, compare the changes of the students in the two classes. Table 1 is the current Korean learning situation of the two classes.

At present, the average Korean learning performance of each group in the two classes is shown in Figure 7.

From Table 1 and Figure 7, the current level of the two classes is basically the same, and the level of each class grouping is basically the same. In the next period of time, the second class will use wireless sensor equipment to conduct classes and Korean language learning, while the first class will still use the usual teaching methods. We usually record the learning situation of the two classes and compare the data recorded before to see the effect of wireless sensor technology in Korean language teaching. The two-class learning situation table is recorded in Table 2.

Comparing Tables 1 and 2, we can see that the students of class 2 Korean language learning using sensing technology have improved by leaps and bounds in all aspects, and their performance in all aspects has basically reached a good or

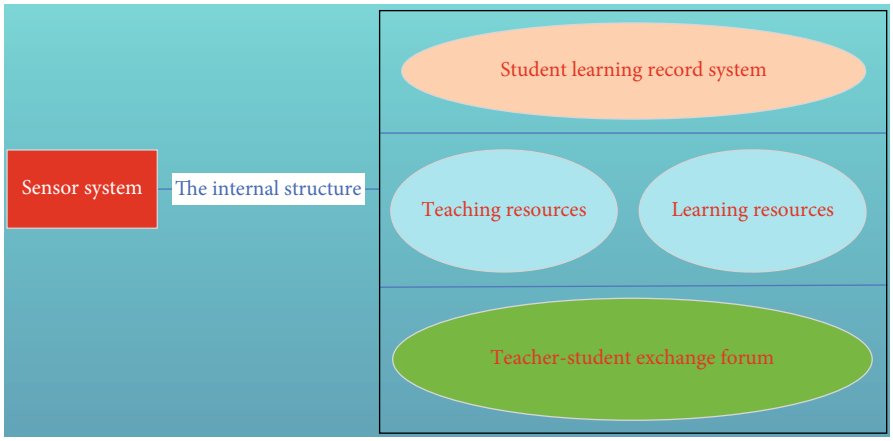


Figure 6: Sensing device background.

Table 1: The current situation of Korean language learning in the two classes.

Class	Groups	Independent learning capability	Enthusiasm	Teaching effectiveness	Teaching quality
First class	1	40	45	50	54
	2	46	45	45	54
	3	45	46	53	56
	4	43	43	46	55
	5	44	46	47	54
Second class	1	42	42	56	57
	2	45	46	55	56
	3	46	47	53	54
	4	44	45	52	53
	5	43	44	53	56

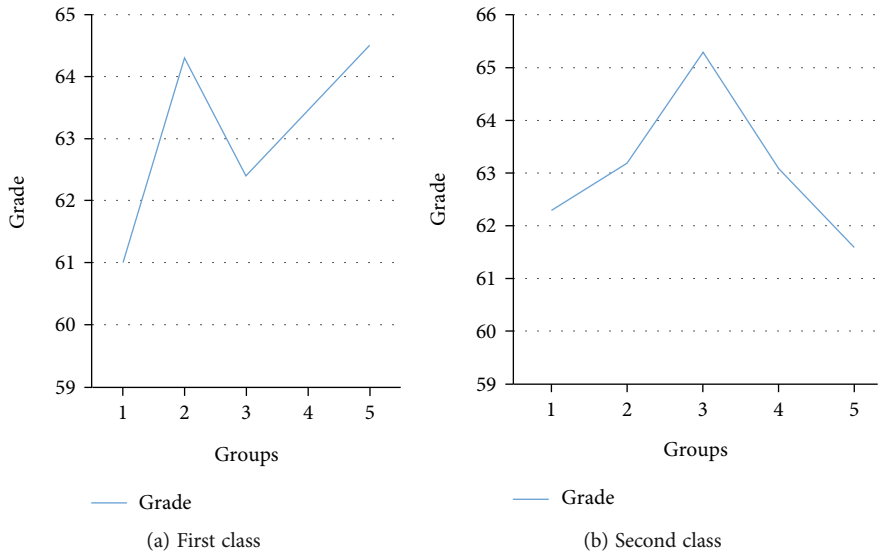


Figure 7: The current average scores of each group in the two classes.

even excellent situation, while the situation in all aspects of the first class remained basically unchanged. To this end, we also arranged tests for the two classes to compare with the previously recorded scores, as shown in Figure 8.

It can be seen from Figure 8(a) that the results of the traditional teaching method are basically the same as before, and there is not much improvement, but the average scores of each group in the second class have improved significantly.

Table 2: Two-class study situation table.

Class	Groups	Independent learning capability	Enthusiasm	Teaching effectiveness	Teaching quality
First class	1	44	45	55	50
	2	47	50	47	54
	3	44	46	53	57
	4	43	44	46	55
	5	44	47	47	56
Second class	1	66	77	69	69
	2	70	80	67	78
	3	71	78	70	73
	4	78	67	78	75
	5	68	81	75	65

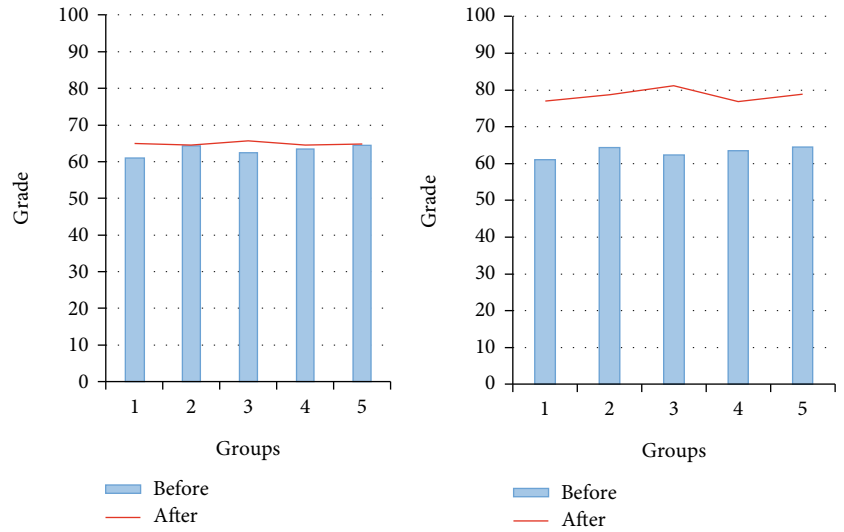


Figure 8: Score record.

It can be seen from Figure 8(b) that it is obvious that each group achieved good and excellent average scores.

3.2. The Improvement of Students' Listening, Speaking, Reading, and Writing

3.2.1. Experiment 2. In this experiment, 8 students were selected from the second class to verify the changes of these students' listening, speaking, reading, and writing abilities after using the wireless sensor device system to study [22]. Table 3 is the test score record table arranged before using the sensory equipment system to study.

After learning Korean using the sensor equipment system, we arranged a test again to study if the 8 students' listening, speaking, reading, and writing skills have improved. The comparison chart is shown in Figure 9.

From Figure 9, it can be seen that after learning Korean with the help of wireless sensor technology, these eight students can basically achieve good scores in listening, speaking, reading, and writing, indicating that their ability to learn Korean has improved in all aspects. And the changes in the comprehensive ability of these eight students are shown in Figure 10.

Table 3: Test scores.

Student	Listening	Reading	Speaking	Writing	Average
1	60	56	56	45	54.25
2	55	46	53	43	49.25
3	65	60	52	65	60.5
4	70	68	45	54	59.25
5	56	67	54	55	58
6	54	56	45	45	50
7	56	67	48	65	59
8	68	77	61	56	65.5

From Figure 10, it can be clearly seen that the comprehensive Chinese learning ability of the eight students has improved. The learning ability of each student has increased

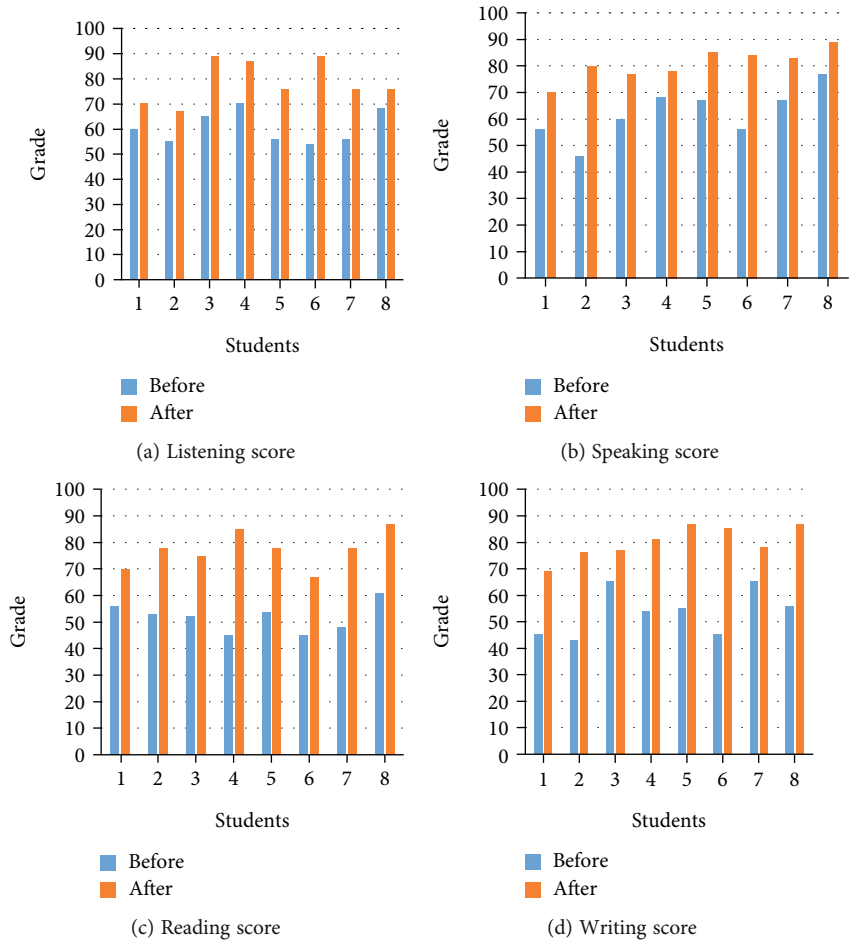


Figure 9: Grades.

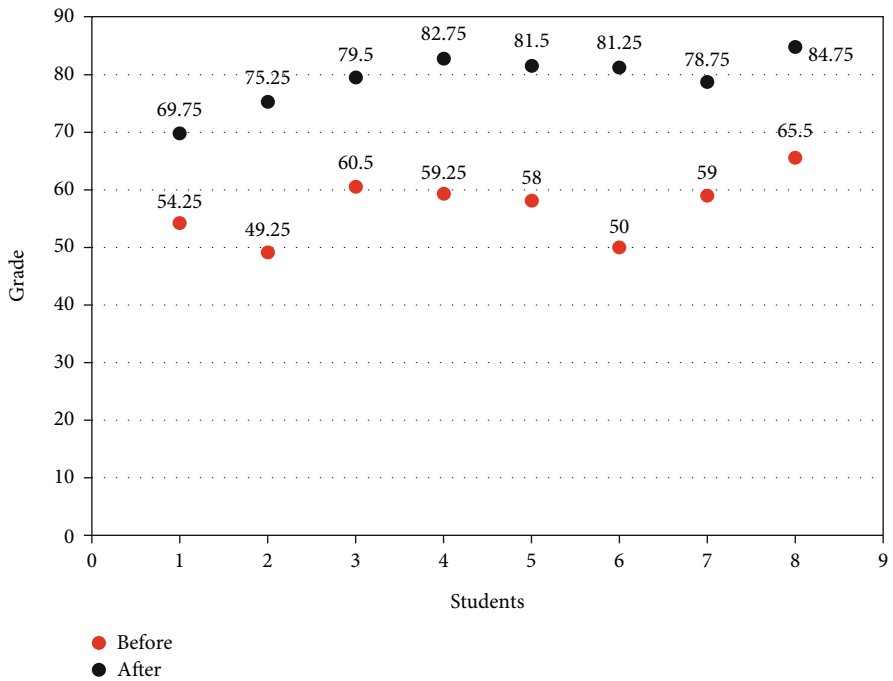


Figure 10: Changes in comprehensive capability.

by at least 30%, and the ability can basically reach a good stage.

3.3. Experimental Summary. From Experiment 1, the comparison of two different teaching methods after teaching, it can be found that the students of Korean class 2 who use the sensor equipment system improved faster, indicating that the use of sensor technology in Chinese teaching can improve students' ability and enthusiasm for learning Korean. In Experiment 2, the Korean listening, speaking, reading, and writing skills of the eight students we selected were compared before and after, and it was found that the eight students' Korean listening, speaking, reading, and writing skills had improved, and their comprehensive ability had also improved rapidly. It shows that the sensor device system can stimulate the motivation of students to learn Korean and improve their enthusiasm and initiative.

4. Discussion

This paper expounds the wireless sensing technology and finds that the sensing technology can not only be used in Korean language teaching but can also be widely used in the whole field of education and other fields. And as one of the Internet of Things technologies, it must rely on wireless communication network technology to achieve the transmission of data and information in the heroic process [23]. The sensor equipment system that needs to be involved in the application of Korean language teaching also needs to use multimedia technology to be more efficient and intelligent. This creates a more realistic language immersion environment for students, enabling students to learn in a Korean language environment that promotes knowledge and language assimilation. Students can communicate with the intelligent sensor device to achieve the language output process and improve the students' language knowledge application ability. Moreover, abundant teaching and learning resources can be obtained in the sensor equipment system, which can promote the reform of Korean language teaching.

This paper also discusses the existing Korean teaching mode and finds that the existing Korean teaching materials are not updated, resulting in very limited teaching resources in schools and limited students' vision, and it is difficult to achieve good teaching quality. To this end, the application of sensor technology can not only improve the role of wireless network in Korean language learning but can also help students change their previous rigid learning methods. At the same time, it can also practice listening, speaking, reading, and writing skills with the help of the sensor equipment system, which has a huge role in promoting the cultivation of students' comprehensive Korean ability. And students use the sensor device system to learn; the background of the device will record the students' learning situation so that teachers can find the students' problems in time. At the same time, it can also improve its own teaching plan based on this data monitoring and students can make common progress. The resources in the background of the equipment system can be used for teachers to teach and students to learn, which can improve the problem of lack of resources. Teachers can communicate with students online, which enables them to go deep into the student group to obtain the learning needs of students, to prescribe the right medicine, to more effectively improve the quality of education, and to achieve teaching goals.

The experiment in this paper also verifies the effect of sensing technology in Korean language teaching. From the comparison of the class, it can be found that the sensing technology can effectively promote the improvement of students' Korean learning ability and application ability. Continuing to use the sensory equipment system to assist teaching can make students' Korean language ability improve by leaps and bounds. Regarding the application effect of sensing technology in Korean language teaching, which can be used as a reference for the improvement of teachers' teaching plans and promote the transformation of teachers' teaching thinking, it can also promote the learning of other majors. This technology can be vigorously promoted in the campus to promote students' autonomy and change the inertia of students' learning and improve students' interest in learning.

5. Conclusions

This paper discusses the sensing technology and Korean language teaching in the Internet of Things and finds that the sensing technology has high application value. As an advanced modern technology, it has made great contributions to Korean language teaching. In the traditional Korean language teaching, it is found that students have a lot of inertia in learning Korean, and the teachers' teaching lectures are boring and just blindly input knowledge to the students, but the students cannot get effective output, which makes the students' learning efficiency low. Sensing technology can create a Korean language environment, so that students can effectively output the knowledge they have learned and promote their learning efficiency. The experiments in this paper show that the use of wireless sensor technology in Korean language teaching can promote the change of teachers' teaching plans. It can strengthen the communication between teachers and students, at the same time, it can also promote the improvement of students' comprehensive ability of Korean, such as listening, speaking, reading, writing, etc., to promote the enthusiasm of students to learn Korean, and cultivate students' interest. All in all, the application of wireless sensor technology in Korean language teaching is very effective, and the discussion in this paper has great practical value and reference significance. However, the application of wireless sensor technology knowledge in the field of Korean language teaching discussed in this paper still has many deficiencies. It is hoped that future research can make up for these deficiencies and make wireless sensor technology applied in a wider field.

Data Availability

No data were used to support this study.

Conflicts of Interest

There is no potential conflict of interest in this study.

References

[1] P. Constantinos, "Backscatter communications for wireless powered sensor networks with collision resolution," *IEEE Wireless Communications Letters*, vol. 6, no. 5, pp. 650–653, 2017.

[2] J. Jeoung, "Understanding the reasons for loss to follow-up in patients with glaucoma at a tertiary referral teaching hospital in Korea," *British Journal of Ophthalmology*, vol. 101, no. 8, pp. 1059–1065, 2017.

[3] Q. Yi, "Security and wireless communication networks," *IEEE Wireless Communications*, vol. 27, no. 3, pp. 4-5, 2020.

[4] X. Tong, "A study on the applicability of flipped learning in oral Korean teaching in China -based on the current oral Korean teaching in China's higher institutions-," *Ratio et Oratio*, vol. 10, no. 1, pp. 165–193, 2017.

[5] S.-T. Park, "A study on teaching Korean reading to learners of multicultural background," *Contemparary Society and Multiculture*, vol. 7, no. 1, pp. 161–186, 2017.

[6] E.-J. Lee, "Teaching Korean language and culture using the myth of Dangun," *Culture and Convergence*, vol. 39, no. 4, pp. 285–312, 2017.

[7] H.-J. Kwon, "Teaching-learning and evaluation methods of primary school Korean music for the 'music in daily life' domain under the 2015 revised music curriculum," *The Journal of Korean Music Education Research*, vol. 11, no. 1, pp. 5–35, 2017.

[8] F. Bahlke, O. D. Ramos-Cantor, S. Henneberger, and M. Pesavento, "Optimized cell planning for network slicing in heterogeneous wireless communication networks," *IEEE Communications Letters*, vol. 22, no. 8, pp. 1676–1679, 2018.

[9] T. Xu, L. Gong, W. Zhang, X. Li, X. Wang, and W. Pan, "Application of wireless sensor network technology in logistics information system," *AIP Conference Proceedings*, vol. 1834, no. 1, pp. 1–5, 2017.

[10] P. W. Kim, "Real-time bio-signal-processing of students based on an Intelligent algorithm for Internet of Things to assess engagement levels in a classroom," *Future Generation Computer Systems*, vol. 86, pp. 716–722, 2018.

[11] P. Wang and S. Qiao, "Emerging applications of blockchain technology on a virtual platform for English teaching and learning," *Wireless Communications and Mobile Computing*, vol. 2020, Article ID 6623466, 10 pages, 2020.

[12] F. Fan, S. C. Chu, J. S. Pan, C. Lin, and H. Zhao, "An optimized machine learning technology scheme and its application in fault detection in wireless sensor networks," *Journal of Applied Statistics*, vol. 34, no. 1, pp. 1–18, 2021.

[13] H. Hamidi and K. Fazeli, "Using Internet of Things and biosensors technology for health applications," *IET Wireless Sensor Systems*, vol. 8, no. 6, pp. 260–267, 2018.

[14] Z. M. Yuldashev, A. M. Sergeev, and N. S. Nastueva, "Perspectives for the use of the Internet of Things in portable online cardiac monitors," *Biomedical Engineering*, vol. 55, no. 3, pp. 210–214, 2021.

[15] A. Karimnia and M. Khosravani, "A comparative study of form-focused and communicative methods of language teaching in ESP courses," *Sustainable Multilingualism*, vol. 12, no. 1, pp. 152–165, 2018.

[16] J. Park, "A study on the development of the basic major Korean teaching materials for KSAP by analyzing Korean textbooks," *Korean Linguistics*, vol. 75, no. 4, pp. 129–160, 2017.

[17] H. Yoo, "Method for teaching Korean past tense to foreigners," *The Journal of Language & Literature*, vol. 71, no. 7, pp. 325–342, 2017.

[18] I. You, G. Pau, W. Wei, and C. Fung, "IEEE access special section editorial: Green communications on wireless networks," *IEEE Access*, vol. 8, no. 27, pp. 187140–187145, 2020.

[19] J.-m. Park, "Exploration on elementary school Korean traditional music teaching method for STEAM education," *Journal of the Korean Music and Education*, vol. 43, no. 78, pp. 117–142, 2017.

[20] C. Park, "The design of class introuduction in Korean languae teaching," *Korean Language in China*, vol. 87, no. 3, pp. 90-91, 2017.

[21] Y. Hyesoo, "Teaching the Korean folk song (Arirang) through performing, creating, and responding," *General Music Today*, vol. 31, no. 1, pp. 16–25, 2017.

[22] M. Lyu, "Exploring preliminary Teachers' competence through overseas educational service program and teaching practice in Korean middle school as pre-service physical education teacher education," *The Korean Journal of Physical Education*, vol. 56, no. 2, pp. 243–257, 2017.

[23] J. Lee and J. B. Kim, "Learning assistant (LA) instead of teaching assistant (TA) in Korea?," *Journal-Korean Physical Society*, vol. 73, no. 4, pp. 414–421, 2018.

Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 5905809, 15 pages
https://doi.org/10.1155/2022/5905809

Research Article

Time-Efficient Coverage Path Planning for Energy-Constrained UAV

Yanxi Huang, Jiankang Xu, Mengting Shi, and Liang Liu

College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, China

Correspondence should be addressed to Liang Liu; liangliu@nuaa.edu.cn

Received 19 March 2022; Revised 6 April 2022; Accepted 22 April 2022; Published 19 May 2022

Academic Editor: Lisheng Fan

Unmanned aerial vehicles (UAVs) have the characteristics of high mobility and wide coverage, making them widely used in coverage, search, and other fields. In these applications, UAV often has limited energy. Therefore, planning a time-efficient coverage path for energy-constrained UAV to cover the area of interest is the core issue. The existing coverage path planning algorithms assume that the UAV moves at a constant speed, without taking into account the cost of turns (including deceleration, turning, and acceleration), which is unrealistic. To solve the above problem, we propose a time-efficient coverage path planning (TECPP) algorithm for the energy-constrained UAV. We build a novel gadget-based graph model, which formalizes the time and energy costs of the flight path including straight flights and making turns (deceleration, turning, and acceleration). Moreover, our graph model is suitable for irregular-shaped areas with multiple obstacles. Finally, we transform the above problem into a generalized traveling salesman problem (GTSP) and use the generalized large neighborhood search (GLNS) solver to solve it. The experimental results show that TECPP outperforms the existing coverage path planning algorithms, and TECPP saves at least 21.6% of time.

1. Introduction

Unmanned aerial vehicles (UAVs) have high mobility and wide coverage, which are often applied to disaster monitoring [1, 2], photogrammetry [3, 4], precision agriculture [5, 6], search and rescue [7, 8], Internet of Things [9–11], etc. By dispatching UAVs equipped with onboard sensors (cameras, radars, etc.) to perform some specific tasks such as coverage and search, we can not only complete tasks conveniently and efficiently but also save human resources greatly.

In practical applications, most UAVs have limited energy, which restricts deployments to less than 30 minutes [12]. Therefore, it is crucial to plan a time-efficient coverage path for energy-constrained UAV. The existing coverage path planning algorithms, such as [13–16], mainly take the energy consumption of UAVs as the optimization goal and ignore the time cost of completing tasks. Furthermore, they assume that UAV moves at a constant speed, without taking into account the cost of turns (deceleration, turning, and acceleration). Besides, most of them ignore the impact of obstacles. For example, the area of interest may be irregular-shaped with multiple obstacles, and UAV must bypass all obstacles to complete coverage or search missions successfully.

In this paper, we study the time-efficient coverage path planning problem, which takes into account the cost of turns and the impact of obstacles. We call this problem the complex multifactor coverage path planning (CMFCPP) problem. To solve the CMFCPP problem, we propose a time-efficient coverage path planning (TECPP) algorithm for energy-constrained UAV, which takes the time consumption as the optimization goal and the energy consumption as the constraint. We build a novel gadget-based graph model, which formalizes the time and energy costs of the flight path including straight flights and making turns (deceleration, turning, and acceleration). Moreover, our graph model takes into account the impact of obstacles and is suitable for irregular-shaped areas with multiple obstacles. Finally, we transform the CMFCPP problem into

a generalized traveling salesman problem (GTSP) and use the generalized large neighborhood search (GLNS) solver [17] to solve it.

We summarize the major contributions as follows:

(i) We build a novel gadget-based graph model considering the impact of obstacles, which formalizes the time and energy costs of the flight path including straight flights and making turns (deceleration, turning, and acceleration)

(ii) Based on this graph model, we formulate the CMFCPP problem and transform the problem into a generalized traveling salesman problem (GTSP)

(iii) We propose TECPP for solving the CMFCPP problem, which can save at least 21.6% of time compared to the existing coverage path planning algorithms

The remainder of the paper is organized as follows. In Section 2, we discuss the related work. In Section 3, we introduce the preliminaries. In Section 4, we give the UAV system model, including the UAV coverage model and UAV movement model. In Section 5, we build a gadget-based graph model to formalize the time and energy costs of the flight path including straight flights and making turns (deceleration, turning, and acceleration). In Section 6, we formulate the CMFCPP problem and transform the problem into GTSP. In Section 7, we present our experimental results comparing TECPP to the traditional coverage path planning algorithm (baseline) and Lin-Kernighan heuristic algorithm for drones (LKH-D). We summarize our main results in Section 8.

2. Related Work

Recently, there have been many studies on coverage path planning for UAVs. We classify the existing coverage path planning algorithms involved in these studies as follows.

2.1. Geometric Algorithms. Torres et al. [13] propose the back-and-forth (BF) algorithm for concave or multiple polygon coverage. They calculate the optimal line sweep direction and decompose a complex region into several regular regions to minimize the number of turns in the process of performing coverage tasks. Finally, they plan a flight path for the UAV to move back and forth to reduce the energy consumption of the UAV as much as possible. In [14], the authors propose an E-spiral algorithm to solve the coverage path planning problem. They establish a new energy model to set the optimal speeds for different stages of straight flights to reduce energy consumption. Besides, they also improve the established energy model to predict the total energy consumption of completing coverage tasks. However, the above two geometric algorithms do not take into account the time cost of completing the tasks and the influence of obstacles that may be included in the area of interest.

2.2. Grid-Based Algorithms. In [18], the authors propose a grid-based path planning algorithm for irregular-shaped areas, which is mainly based on depth-limited search with a backtracking algorithm. This paper uses approximate cell decomposition to discretize the covered target area into regular square grids and converts them into a regular graph. What is more, they use a simple cost function to minimize the number of turns to reduce the energy consumption of UAV. However, some important factors that affect energy consumption, such as acceleration and deceleration in the specific process of turn, are not considered. Based on their work, [19] improves the algorithm and proposes an energy-aware grid-based algorithm to minimize the energy consumption of completing the search tasks in irregular-shaped areas. Moreover, this paper also applies two pruning techniques to the original algorithm and improves the speed of the algorithm greatly.

2.3. Column Generation Algorithms. Choi et al. [20] introduce a novel coverage path planning (CPP) algorithm for a unmanned aerial system (UAS) imagery mission. To mitigate the limitation of the conventional vehicle-routing-based approaches for the CPP not capturing a turning motion of the vehicle, they propose a vehicle-routing-based approach using a column generation algorithm. Based on [20, 21], it presents a new coverage path planning (CPP) algorithm for an aerial imaging mission with multiple unmanned aerial vehicles (UAVs). To solve a CPP problem with multiple UAVs, they divide the coverage mission into five mission segments: take-off, cruise, hovering, turning, and landing. They introduce a new route-based optimization model with column generation that can trace the amount of energy required for all different mission phases to solve the limitation of the traditional approaches.

2.4. Heuristic Algorithms. The Lin-Kernighan heuristic algorithm for drones (LKH-D) [22] improves the traditional Lin-Kernighan heuristic (LKH) algorithm to minimize the total energy consumption of covering the target areas. Piao et al. [23] first explore the use of unmanned aerial vehicles to realize the automatic construction of CSI maps for indoor positioning. They propose an energy optimization problem based on the coverage path planning problem, which is eventually transformed into the generalized traveling salesman problem (GTSP). However, these two algorithms consider the problem of turn but ignore the specific process of turn (deceleration, turning, and acceleration).

Yu et al. [5] consider that UAV can land on an unmanned ground vehicle (UGV). The UGV can also charge UAV while it is being transported to the next take-off location. In the scene of precision agriculture, a boustrophedon cell corresponds to a row of crops, and they use a UAV equipped with a camera sensor to monitor all crops. They propose a new regional coverage path planning algorithm to minimize the time cost and take into account the symbiotic relationship between UGV and UAV. Finally, they convert the problem into a generalized traveling salesman problem (GTSP), which can be solved using the GLNS solver. However, the algorithm ignores the influence of obstacles and the cost of the turning process.

2.5. Machine Learning Algorithms. Theile et al. [24] train a double deep Q-network to make control decisions for the UAV and balance limited power budget and coverage task. Steiger et al. [25] consider a scenario in which a UAV acts as an aerial base station to provide emergency communications services over an area of unknown and uneven user distribution, and they propose an online algorithm that simultaneously solves the path planning and coverage mapping problems using a deep learning model. To solve this problem of coverage path planning in cellular unmanned aerial vehicle networks, Challita et al. [26] propose a deep reinforcement learning algorithm based on echo state network (ESN) cells.

In summary, most of the above algorithms plan to optimize the energy consumption of UAVs and ignore the time cost and the process of turn to complete coverage tasks. However, the optimization goal of our work is the time consumption of completing coverage tasks. In addition, we take into account the time and energy costs of UAV in the process of turn (deceleration, turning, and acceleration) and the influence of irregular-shaped areas with multiple obstacles.

3. Preliminaries

In Section 6, we transform the CMFCPP problem into GTSP and use the GLNS solver to solve it. In Section 7, we take the Lin-Kernighan heuristic algorithm for drones (LKH-D) as one of the experimental comparison algorithms. For the readability of this paper, we briefly introduce GTSP, GLNS, and LKH-D in this section.

3.1. GTSP. The generalized traveling salesman problem (GTSP) [27–29] is a promotion of the classic traveling salesman problem (TSP). GTSP can be expressed as finding a special Hamiltonian loop on a fully weighted graph $G = (V, E, W)$. V is the set of all vertices, representing all city sets. E is the set of all arcs, denoting the set of edges connected between two cities. W is a set of weight, representing the distance or cost between any two cities in graph G.

GTSP is to find a Hamiltonian loop with the smallest sum of weights in the above graph G. This loop does not need to pass through all cities but must pass through each city group once and only once. Moreover, GTSP can be divided into two types. The first type of GTSP is that the Hamiltonian circuit corresponding to the optimal solution passes through each city group once and only passes through one city in each city group. The second type of GTSP is that the Hamiltonian corresponding to the optimal solution passes through each city group once but can pass through multiple cities in each city group. At present, common problems such as coverage path planning, random vehicle scheduling, and mailbox fetching can all be transformed into GTSP.

3.2. GLNS. Generalized large neighborhood search (GLNS) [17], based on adaptive large neighborhood search framework, is an efficient solver for the first type of generalized traveling salesman problem (GTSP). GLNS is proposed by Smith et al., who present a novel insertion mechanism that contains special cases nearest, farthest, and random insertions. The mechanism allows for greater randomization when exploring neighbors of a given GTSP tour.

Besides, they provide extensive benchmarking results for the GLNS solver in comparison to the state of the art on a wide range of existing and new problem libraries. They show that, on the one hand, GLNS is competitive with the most famous algorithms on the GTSP-LIB library. And on the other hand, given the same amount of time, GLNS can find higher-quality solutions than existing approaches on several other libraries.

3.3. LKH-D. The Lin-Kernighan heuristic (LKH) algorithm [30] achieves local optimization through the iterative improvement of random solutions, which is mainly used to solve the classic TSP. However, LKH is not suitable for solving the coverage path planning problem, because it does not take into account the cost of turns.

Based on LKH, Modares et al. [22] propose a Lin-Kernighan heuristic algorithm for drones (LKH-D). LKH-D uses a complex cost function to account for the cost of UAV in the specific process of turn, which is calculated as a weighted sum of the length of the tour and the sum of the turn angles within the tour. Finally, they transform the UAV coverage path planning problem into a variant TSP problem and solve it.

4. System Model

4.1. UAV Coverage Model. The UAVs can be classified into two main categories: fixed-wing and rotary-wing UAVs [31]. The fixed-wing UAV has greater endurance to support longer flights and high-speed motion. However, the fixed-wing UAV cannot perform hovering tasks, since it needs to constantly move during missions. The rotary-wing UAV presents maneuverability advantages using rotary blades. Therefore, the rotary-wing UAV can perform vertical take-off and landing, low-altitude flight, and hovering tasks. In summary, the rotary-wing UAV is more suitable and has been widely used for coverage tasks. As shown in Figure 1, the Crazyflie UAV is a typical rotary-wing UAV, which is used for indoor tasks.

The area of interest may be an irregular-shaped area with multiple obstacles. We draw lessons from the work of Cabreira et al. [19] and divide the area into several square grids. Figure 2 shows that a quadrotor UAV equipped with a camera sensor is dispatched to perform coverage tasks. By adjusting the flying height of the UAV and the relevant parameters of the onboard camera (angle of view, image resolution, etc.), the camera footprint of the UAV just overlaps the grid when the UAV passes through the center of the grid as shown in Figure 2.

4.2. UAV Movement Model. This paper mainly considers the following two choices of actions:

(1) *Straight flight*: without loss of generality, we assume that UAV can only move in eight directions (the angle between adjacent directions is 45°) on the same horizontal plane as shown in Figure 3. During

Figure 1: The Crazyflie UAV.

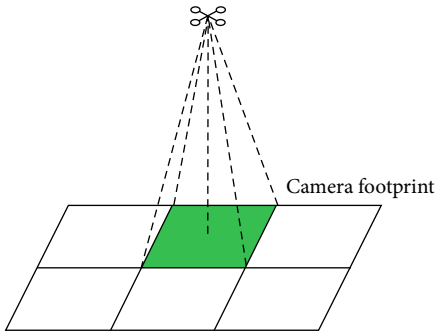


Figure 2: UAV coverage model.

straight flights, UAV has to accelerate before reaching a constant speed of v_0 and decelerate before hovering. In other words, the process of straight flights may be accompanied by acceleration, deceleration, and constant-speed phases. To formalize the time and energy costs of straight flight, we represent the time cost to pass through the adjacent grid by T_s and the energy cost by E_s

(2) *Making a turn*: for a turn, the UAV needs to decelerate in one direction until hovering, then turn, and finally accelerate in another direction until reaching the constant speed of v_0. In other words, the specific process of turn is composed of three phases (deceleration, turning, and acceleration). Without loss of generality, we assume that UAV has only four possible turn angles: 45°, 90°, 135°, and 180°. In addition, the deceleration distance and the acceleration distance are both constant denoted by S_{dec} and S_{acc}. To formalize the time and energy costs of the turning process, we denote the time and energy costs for making a turn by T_t and E_t

5. Problem Modeling

5.1. Graph Representation of Coverage Tasks. To approach the CMFCPP problem, we divide the given area with multiple obstacles into several square grids as shown in Figure 4(a). The gray grids represent no-fly zones with obstacles that the UAV cannot pass, and the nongray grids

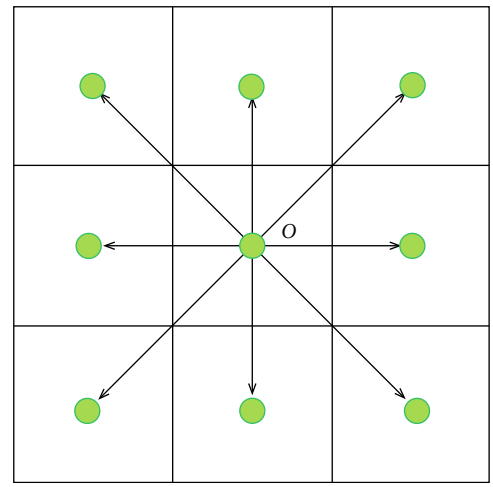


Figure 3: The movement directions of UAV.

represent free zones that the UAV can pass. By connecting the centers of adjacent nongray grids, we model the irregular-shaped area by a graph $G = (V, E)$ as shown in Figure 4(b), where V represents the set of all vertices and each vertex denotes a nongray grid. E represents the set of all edges, and each edge e_{ij} denotes the path that the UAV can pass through. To complete coverage tasks, the UAV needs to visit all the vertices.

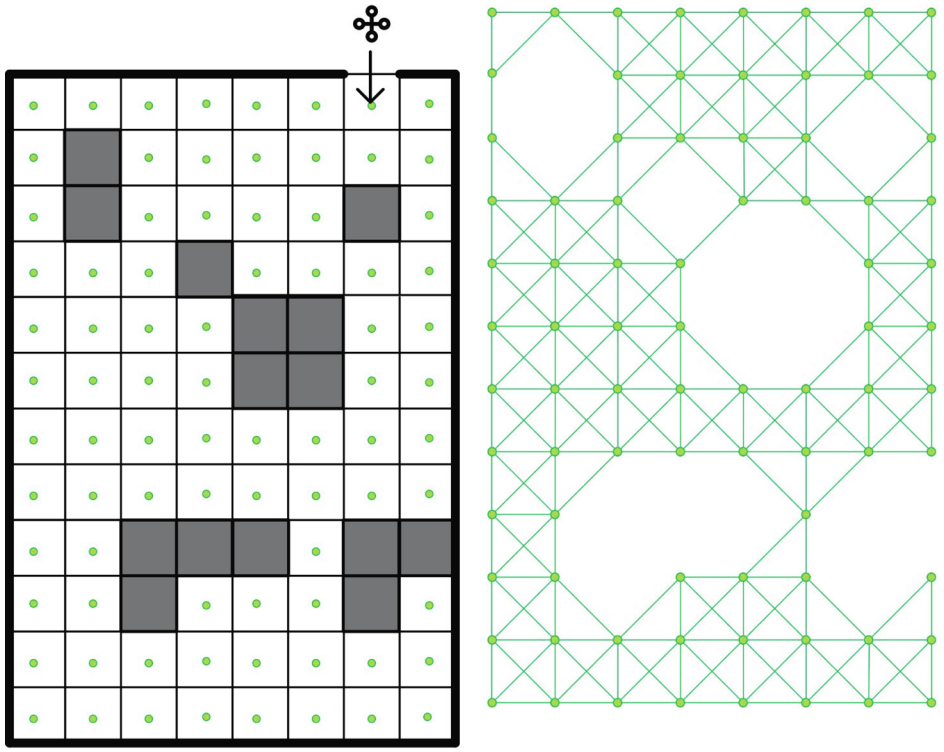
As described above, the UAV has two choices of actions, i.e., straight flight and making a turn. To model the actions, we assign the costs of the actions as the weights of corresponding edges in the graph. Then, by finding a loop on the graph, we can obtain a sequence of actions for the UAV to perform. Finally, the minimum cost is attained by minimizing the summed weights of the visited edges. To achieve this, we should ensure that the cost of UAV's actions is precisely modeled by the weight of edges. We next present the modeling of the cost for a straight flight and making a turn in detail.

5.2. The Cost Model of Straight Flight and Making a Turn. As shown in Figure 5, the UAV has a flight path $a \longrightarrow k \longrightarrow m \longrightarrow b \longrightarrow n \longrightarrow c$. The speed of constant-speed phases is v_0. Let (x_i, y_i) denote the coordinates of vertex i, v_i denote the speed of UAV at vertex i, and $d_{i,j}$ denote the distance between any two vertices i, j. Moreover, d_{ak} and d_{bn} are equal to S_{acc}, and d_{mb} is equal to S_{dec}.

The Cartesian coordinate distance between any two vertices i and j can be obtained by the following equation:

$$d_{ij} = \sqrt{\left(x_i - x_j\right)^2 + \left(y_i - y_j\right)^2}. \tag{1}$$

5.2.1. The Cost Model of Straight Flight. During the straight flight of the UAV from vertex a to vertex b, we need to confirm the flight speed of the UAV at vertex a and vertex b to determine whether this process is accompanied by deceleration or acceleration. We use γ and δ to represent the time and energy costs of the UAV moving one meter. Moreover,



(a) The area of interest (b) Graph representation

Figure 4: Graph representation of coverage tasks.

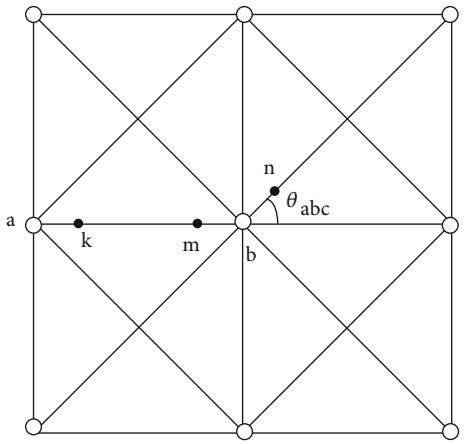


Figure 5: The cost model of straight flight and making a turn.

we denote the time and energy costs of acceleration by T_{acc} and E_{acc} and use T_{dec} and E_{dec} to denote the time and energy costs of deceleration.

We get the time cost T_{ab} and energy cost E_{ab} of the UAV in this process as follows:

$$T_{ab} = \begin{cases} \gamma d_{am}, & \text{if } v_a = v_0, \\ T_{acc} + \gamma d_{km}, & \text{if } v_a = 0 \end{cases}, \tag{2}$$

$$E_{ab} = \begin{cases} \delta d_{am}, & \text{if } v_a = v_0, \\ E_{\text{acc}} + \delta d_{km}, & \text{if } v_a = 0. \end{cases} \tag{3}$$

Finally, we denote the time cost T_s and energy cost E_s of the UAV passing through the adjacent grids by

$$T_s = \begin{cases} T_{ab}, & \text{if } e_{ab} \text{ is traversed}, \\ \infty, & \text{otherwise}, \end{cases} \tag{4}$$

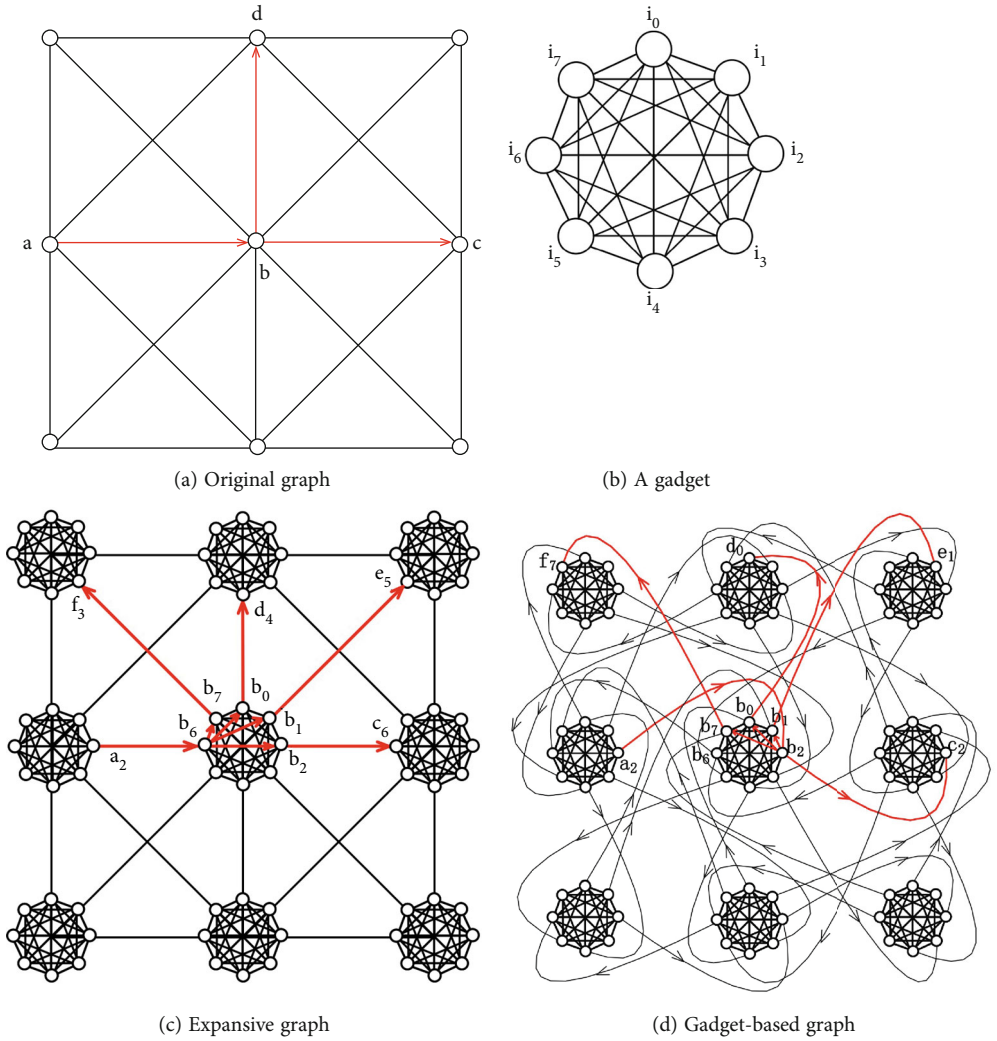
$$E_s = \begin{cases} E_{ab}, & \text{if } e_{ab} \text{ is traversed}, \\ \infty, & \text{otherwise}. \end{cases} \tag{5}$$

5.2.2. The Cost Model of Making a Turn. During the process of making a turn, we represent the time and energy costs of the UAV turning one degree by α and β. By calculating the lengths of the sides e_{ab}, e_{bc}, and e_{ac} and then using the Law of Cosines, the angle of turn at vertex b can be calculated by θ_{abc}.

We get the time cost T_{abc} and energy cost E_{abc} of the turning process $m \longrightarrow b \longrightarrow n$ as follows:

$$T_{abc} = T_{\text{dec}} + 180 \times \frac{\theta_{abc}}{\pi} \times \alpha + T_{\text{acc}}, \tag{6}$$

$$E_{\text{abc}} = E_{\text{dec}} + 180 \times \frac{\theta_{abc}}{\pi} \times \beta + E_{\text{acc}}, \tag{7}$$



(a) Original graph

(b) A gadget

(c) Expansive graph

(d) Gadget-based graph

FIGURE 6: The improvements of modeling.

$$\theta_{\text{abc}} = \pi - \cos^{-1}\left[\frac{\left(d_{ab}^2 + d_{bc}^2 - d_{ac}^2\right)}{2d_{ab}d_{bc}}\right]. \tag{8}$$

Therefore, we denote the time cost T_t and energy cost E_t of the UAV making a turn by

$$T_t = \begin{cases} T_{abc}, & \text{if } e_{ab}, e_{bc} \text{ is traversed,} \\ \infty, & \text{otherwise,} \end{cases} \tag{9}$$

$$E_t = \begin{cases} E_{abc}, & \text{if } e_{ab}, e_{bc} \text{ is traversed,} \\ \infty, & \text{otherwise.} \end{cases} \tag{10}$$

5.3. The Improvements of Modeling. In the previous subsection, we have formalized the cost of straight flight and making a turn. However, the cost of the turning process cannot be expressed by the weight of edges in the graph G. As shown in Figure 6(a), the weight of the edge does not reflect the turning cost. For example, the cost for path $a \longrightarrow b \longrightarrow c$ is equal to the cost for path $a \longrightarrow b \longrightarrow d$, and the cost of turning at vertex b is not considered.

To accurately express the turning cost of the UAV, we improve the original graph as follows.

We expand the vertices in the original graph as shown in Figure 6(b), and a vertex is expanded into 8 vertices, representing 8 different directions, respectively. We call the 8 vertices as a gadget and establish a weighted edge between any two vertices in a gadget. The weight of each edge in a gadget means that the UAV takes corresponding costs at different turn angles. The time cost weight $T_t(i_j, i_k)$ and energy cost weight $E_t(i_j, i_k)$ of any two vertices i_j, i_k in a gadget are defined as

$$T_t(i_j, i_k) = T_{\text{acc}} + \min\{n, 8-n\} \times 45\alpha + T_{\text{dec}}, \quad \text{if } |j-k| = n \text{ or } 8-n, n \in \{1,2,3,4\}, \tag{11}$$

$$E_t(i_j, i_k) = E_{\text{acc}} + \min\{n, 8-n\} \times 45\beta + E_{\text{dec}}, \quad \text{if } |j-k| = n \text{ or } 8-n, n \in \{1,2,3,4\}. \tag{12}$$

By introducing the concept of gadget, we establish an expansive graph as shown in Figure 6(c). However, the cost for paths $a_2 \longrightarrow b_6 \longrightarrow b_7 \longrightarrow f_3$ and $a_2 \longrightarrow b_6 \longrightarrow b_1 \longrightarrow$

Input: Two-dimensional map of the interest area *map*
Coverage of all obstruction-free areas *Coverage*
Output: An efficient flight path of UAV *path*
1: initialize *map* and divide *map* into square grids set S
2: set the obstacle-free grids set as F, $F \subseteq S$
3: **for** each grid node g in F**do**
4: expend g into 8 vertices v_1-v_8
5: add v_1-v_8 to V
6: **end for**
7: **for** each vertex v in V**do**
8: **for** each neighbor n of g**do**
9: $e_{vn} \longleftarrow$ an *edge*
10: add e_{vn} to E
11: $w_{vn} \longleftarrow$ *energy* and *time cost* for *one motion*
12: add w_{vn} to W
13: **end for**
14: **end for**
15: Directed Weighted Graph $G(V, E, W) \longleftarrow V, E, W$
16: $V_{num} \longleftarrow |V|$
17: **for** each vertex v in V**do**
18: **Dijkstra**(v) to get all minimum cost array $c[V_{num}]$
19: $w_v \longleftarrow c[v]$
20: **for** each vertex v' in V**do**
21: add $e_{vv'}$ to E
22: **end for**
23: **end for**
24: **for** each vertex v in V**do**
25: **if**!(v in *Coverage*)**then**
26: remove vertex v from G
27: **end if**
28: **end for**
29: *coordPoints* $\longleftarrow$ **GLNS**$(G, vertSet)$
30: *path* $\longleftarrow$ **ConvertToPath**$(G, coordPoints)$
31: **return***path*

Algorithm 1: TECPP.

e_5 is unreasonable. It cannot reflect the fact that the UAV will spend more time and energy costs as the turn angles increase.

To solve the above problem, we transform Figures 6(c) and 6(d). As shown in Figure 6(d), we establish weighted directed edges between two vertices in the adjacent gadgets that represent the same direction. The weights of these edges represent the cost of straight flights. In this way, the cost for paths $a_2 \longrightarrow b_2 \longrightarrow b_7 \longrightarrow f_7$ and $a_2 \longrightarrow b_2 \longrightarrow b_1 \longrightarrow e_1$ consists with the fact. Thus, the cost of making turns can be expressed accurately by weighted directed edges.

5.4. Gadget-Based Graph Model. The gadget-based graph is a weighted directed graph $G = (V, E, W)$. V represents the set of all vertices (a gadget contains 8 vertices), and we use $\mathcal{G}_\eta \subset V$ to denote the gadget. E represents the set of all edges, including the weighted directed edges between adjacent gadgets and the 28 weighted edges inside each gadget. W (including W_T, W_E) represent the weights on each edge, W_T represents the time cost on each edge, and W_E represents the energy cost on each edge. Among them, $W_T \in \{T_s, T_t\}$, $W_E \in \{E_s, E_t\}$.

On the above graph G, a loop represents a flight path for UAV, and the sum of its weights is equal to the total time or energy cost. In the next section, we will present our solution to find the shortest loop passing all gadgets on the graph.

6. Problem Solving

In this section, we formulate the CMFCPP problem based on the graph G. Then, we transform the problem to GTSP, which can be solved efficiently using the GLNS solver.

6.1. CMFCPP Formulation. Let a nonnegative integer variable M_{ab} denote the number of times UAV moves from vertex a to vertex b, and let the binary decision variable $\mathcal{D}_a \in \{0, 1\}$ denote whether vertex a is visited and it is defined for each vertex a as

$$\mathcal{D}_a = \begin{cases} 0, & \text{if } \sum_{b \in V} M_{ab} + \sum_{c \in V} M_{ca} = 0, \\ 1, & \text{if } \sum_{b \in V} M_{ab} + \sum_{c \in V} M_{ca} \geq 1. \end{cases} \tag{13}$$

Our goal is to minimize the time cost of the energy-constrained UAV to complete coverage tasks. In other words, we need to find a loop passing all widgets with minimum time cost and less energy cost than $E_{\max}$. Since a loop on the graph can be represented by a sequence of edges between vertices, the total cost of the loop is as follows:

$$\sum_{a,b\in V} W_{T_{ab}} M_{ab}, \text{for} \sum_{a,b\in V} W_{E_{ab}} M_{ab} E_{\max}. \tag{14}$$

Next, we discuss the constraints that need to be met to solve the above problem.

(1) *Gadget coverage constraint*: we require each gadget in the gadget-based graph G to be visited at least once. Thus, the gadget coverage constraint is expressed as

$$\sum_{a\in\mathcal{G}_\eta} \mathcal{D}_a \geq 1, \quad \forall \mathcal{G}_\eta \subset V \tag{15}$$

(2) *Flow conservation constraint*: for each vertex in the graph, the inflow should be consistent with the outflow, so the flow conservation constraint is as follows:

$$\sum_{a\in V\setminus\{b\}} M_{ab} = \sum_{c\in V\setminus\{a,b\}} M_{bc}, \quad \forall b \in V \tag{16}$$

(3) *Subtour elimination constraint*: to avoid the generation of subtours, we set the following constraint to eliminate subtours:

$$\sum_{a\in\mathcal{G}_\eta, b\in V\setminus\mathcal{G}_\eta} M_{ab} + M_{ba} \geq 1, \quad \forall \mathcal{G}_\eta \subset V \tag{17}$$

6.2. Transform CMFCPP into GTSP. However, the above constraints cannot make exactly one vertex to be visited for each gadget. To solve this problem, we propose to transform the graph G into a complete graph by assigning the weight between any two vertices, which can be calculated using the Dijkstra algorithm. Then, the UAV can find the minimum weight of an edge between any two vertices with the total cost unchanged. Now, we require exactly one vertex to be visited for each gadget:

$$\sum_{a\in\mathcal{G}_\eta} \mathcal{D}_a = 1, \quad \forall \mathcal{G}_\eta \subset V. \tag{18}$$

In this way, we ensure that the solution visits each gadget once and there always exists a feasible solution. The resulted objective function and constraints now exactly model our CMFCPP problem, summarized as follows:

$$\begin{aligned} \min \quad & \sum_{a,b\in V} W_{T_{ab}} M_{ab}, \quad \text{for} \sum_{a,b\in V} W_{E_{ab}} M_{ab} < E_{\max} \\ \text{s.t.} \quad & \begin{cases} \sum_{a\in\mathcal{G}_\eta} \mathcal{D}_a = 1, & \forall \mathcal{G}_\eta \subset V \\ \sum_{a\in V\setminus\{b\}} M_{ab} = \sum_{c\in V\setminus\{a,b\}} M_{bc}, & \forall b \in V \\ \sum_{a\in\mathcal{G}_\eta, b\in V\setminus\mathcal{G}_\eta} M_{ab} + M_{ba} \geq 1, & \forall \mathcal{G}_\eta \subset V \\ M_{ab} \in \mathbb{N}, \quad \mathcal{D}_a \in \{0,1\} \end{cases} \end{aligned} \tag{19}$$

Our goal is to find a path that traverses one vertex in each gadget and spends the minimum time cost for UAV to complete coverage tasks along this path without exhausting all of the energy as shown in Equation (19). Therefore, the CMFCPP problem is equivalent to the first type of generalized travelling salesman problem (GTSP), which is obviously a NP-hard problem.

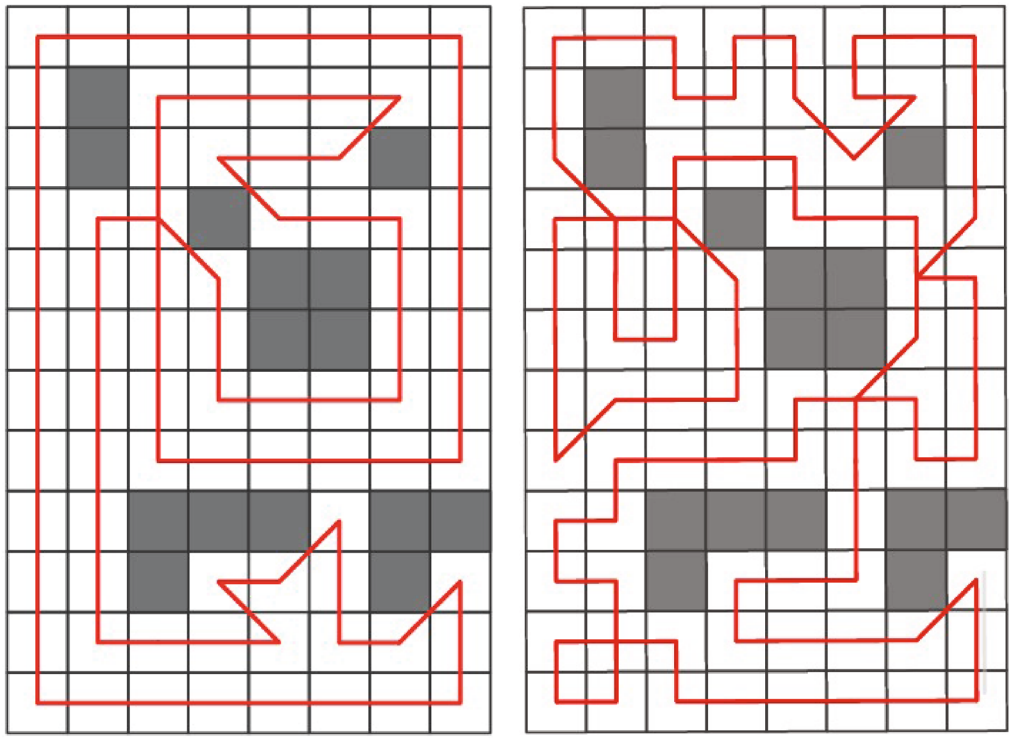
We propose a time-efficient coverage path planning (TECPP) algorithm to solve the CMFCPP problem as shown in Algorithm 1. TECPP can give a feasible solution for an instance of GTSP at an acceptable cost (i.e., computation time and space). We can find an efficient flight path that takes the minimum time cost by traversing a vertex in each gadget, which is the solution to the problem. Finally, we use the GLNS [17] solver to solve the generalized traveling salesman problem (GTSP) in this algorithm.

Table 1: The cost parameters of UAV.

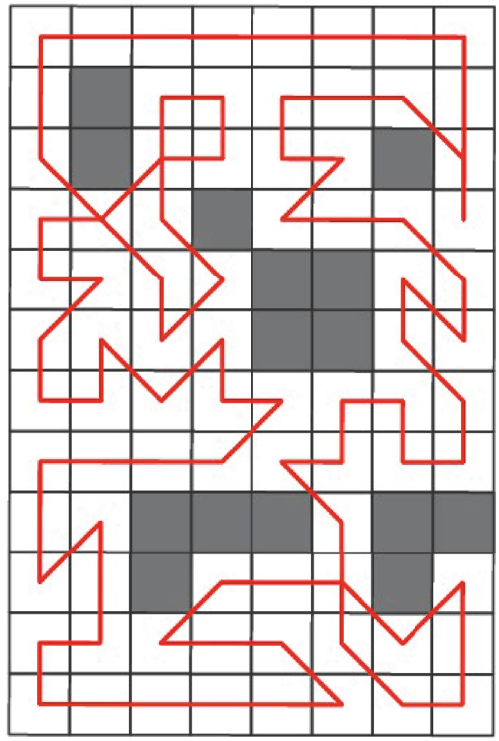
Parameter	Explanation	Value
v_0	The speed of constant-speed phases	2 m/s
γ	Time cost parameter (straight flight)	0.5 s/m
δ	Energy cost parameter (straight flight)	0.12 KJ/m
α	Time cost parameter (making a turn)	0.02 s/deg
β	Energy cost parameter (making a turn)	0.018 KJ/deg
S_{acc}	Acceleration distance	0.31 m
S_{dec}	Deceleration distance	0.22 m
T_{acc}	Time cost of acceleration	0.4 s
T_{dec}	Time cost of deceleration	0.3 s
E_{acc}	Energy cost of acceleration	0.04 KJ
E_{dec}	Energy cost of deceleration	0.03 KJ

7. Simulations

We perform three sets of experiments to evaluate the performance of TECPP for the CMFCPP problem. In terms of time cost, energy cost, and calculation time, we



(a) TECPP (b) LKH-D

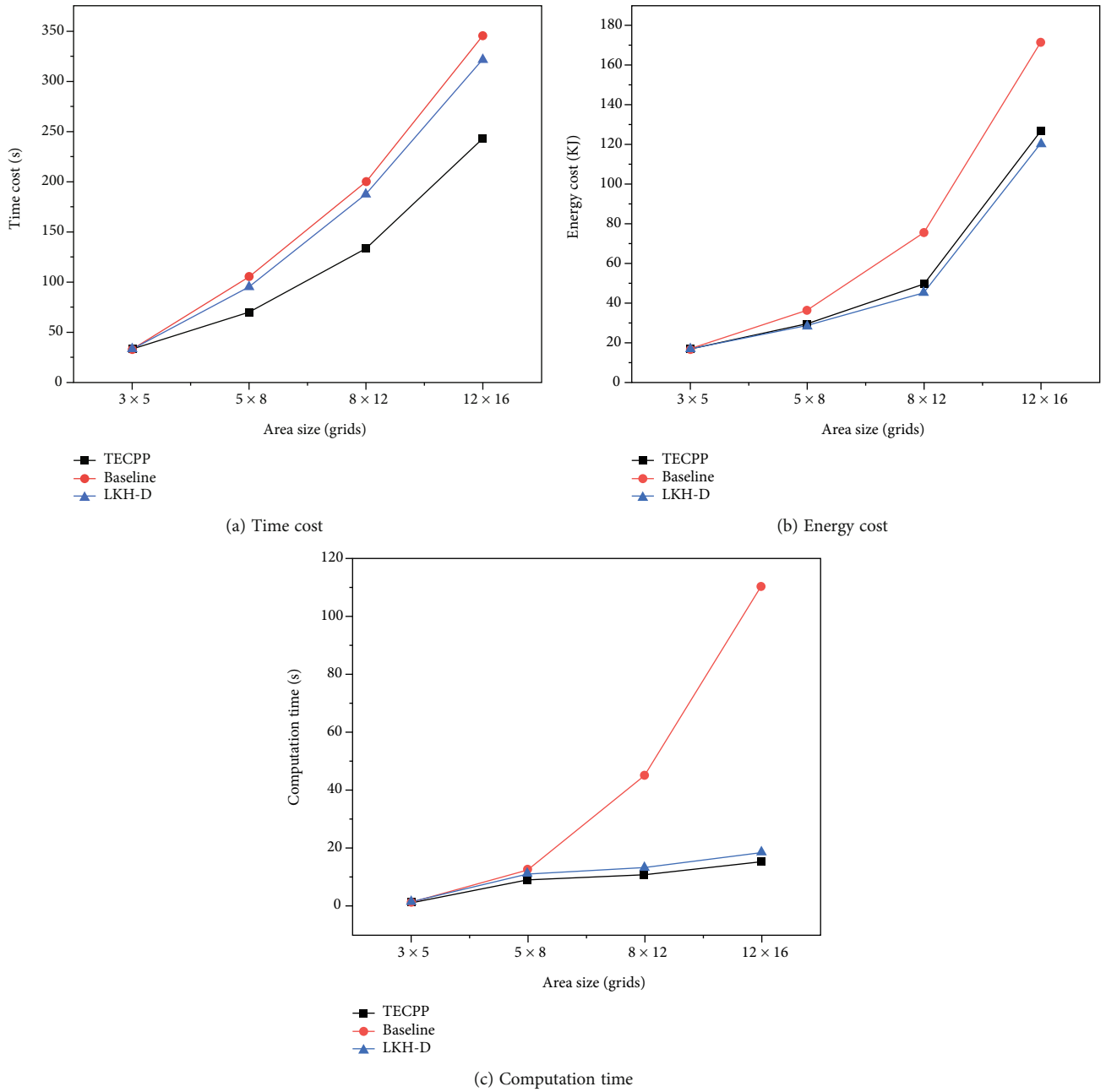


(c) Baseline

Figure 7: Optimal paths planned by TCC-CPP, LKH-D, and baseline algorithms.

compare TECPP with a baseline algorithm and the Lin-Kernighan heuristic algorithm for drones (LKH-D) in several scenarios.

In the baseline, we use the traditional coverage path planning algorithm [13] as a baseline approach. This baseline algorithm considers the number of turns and obtains a



(a) Time cost

(b) Energy cost

(c) Computation time

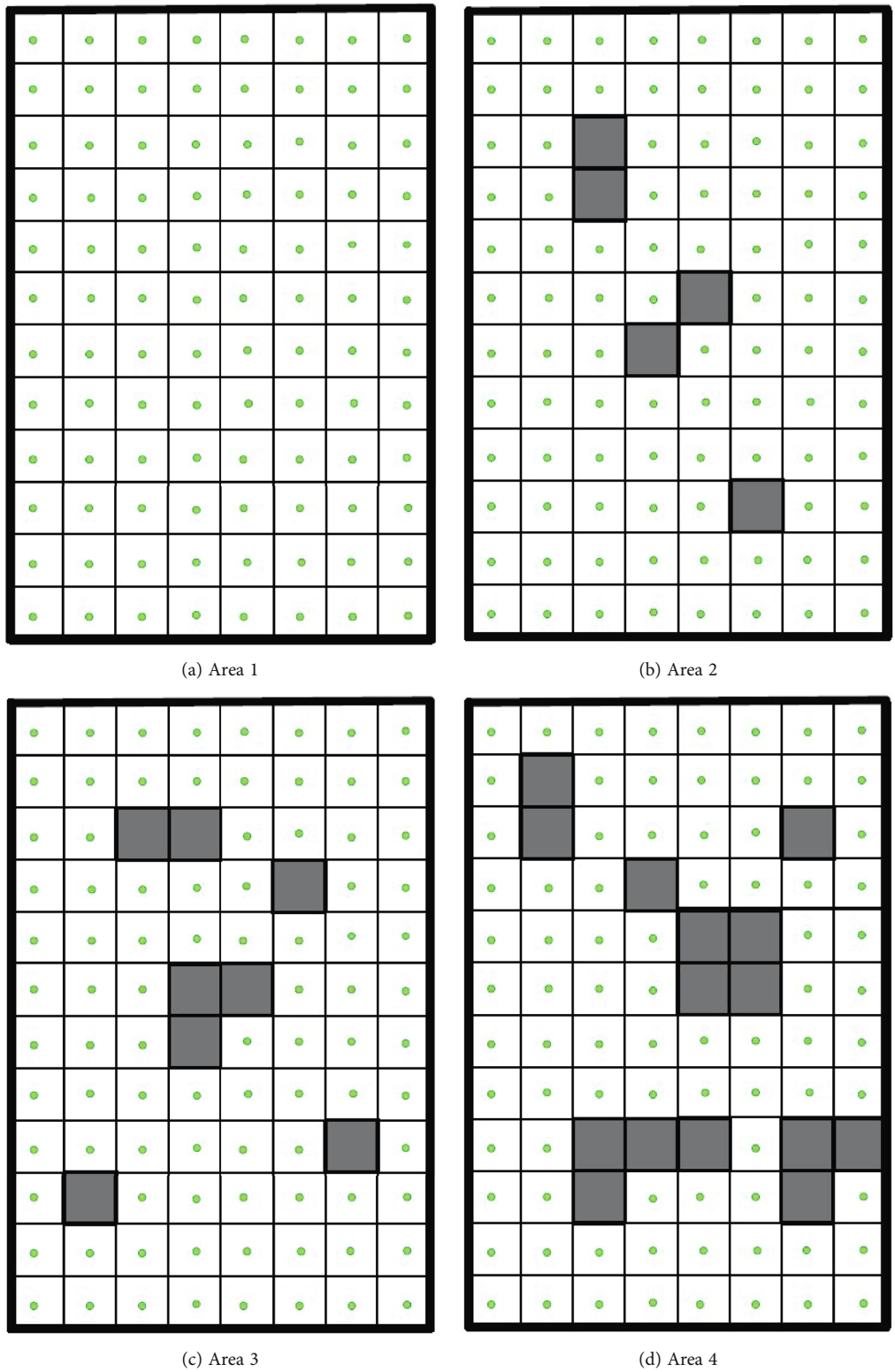
FIGURE 8: The impact of area size on the time cost, energy cost, and computation time.

path that reduces energy consumption by minimizing the number of turns.

In these experiments, we set the appropriate simulation parameters for our gadget-based graph model as shown in Table 1. It is assumed that the UAV with limited energy performs coverage task in the region of interest shown in Figure 4(a), as shown in Figure 7, and the optimal coverage path of each algorithm is given.

7.1. The Impact of Area Size. In this subsection, we study the impact of area size on time cost, energy cost, and computation time. We show the performances of the compared algorithms for simple rectangular areas with dimensions 3×5, 5×8, 8×12, and 12×16 in Figure 8. The grids in each of these areas are $2\,\text{m} \times 2\,\text{m}$ squares without obstacles.

Figure 8(a) shows how the time cost scales with the number of grids for each algorithm, and Figure 8(b) shows the energy cost under each algorithm. As would be expected, TECPP saves more time than the baseline and LKH-D as the number of grids increases. In addition, TECPP achieves a comparable performance to LKH-D and does better than the baseline in terms of the energy cost. The reason is that TECPP greatly optimizes the time cost under the constraint of energy and has high scalability. Although the baseline and



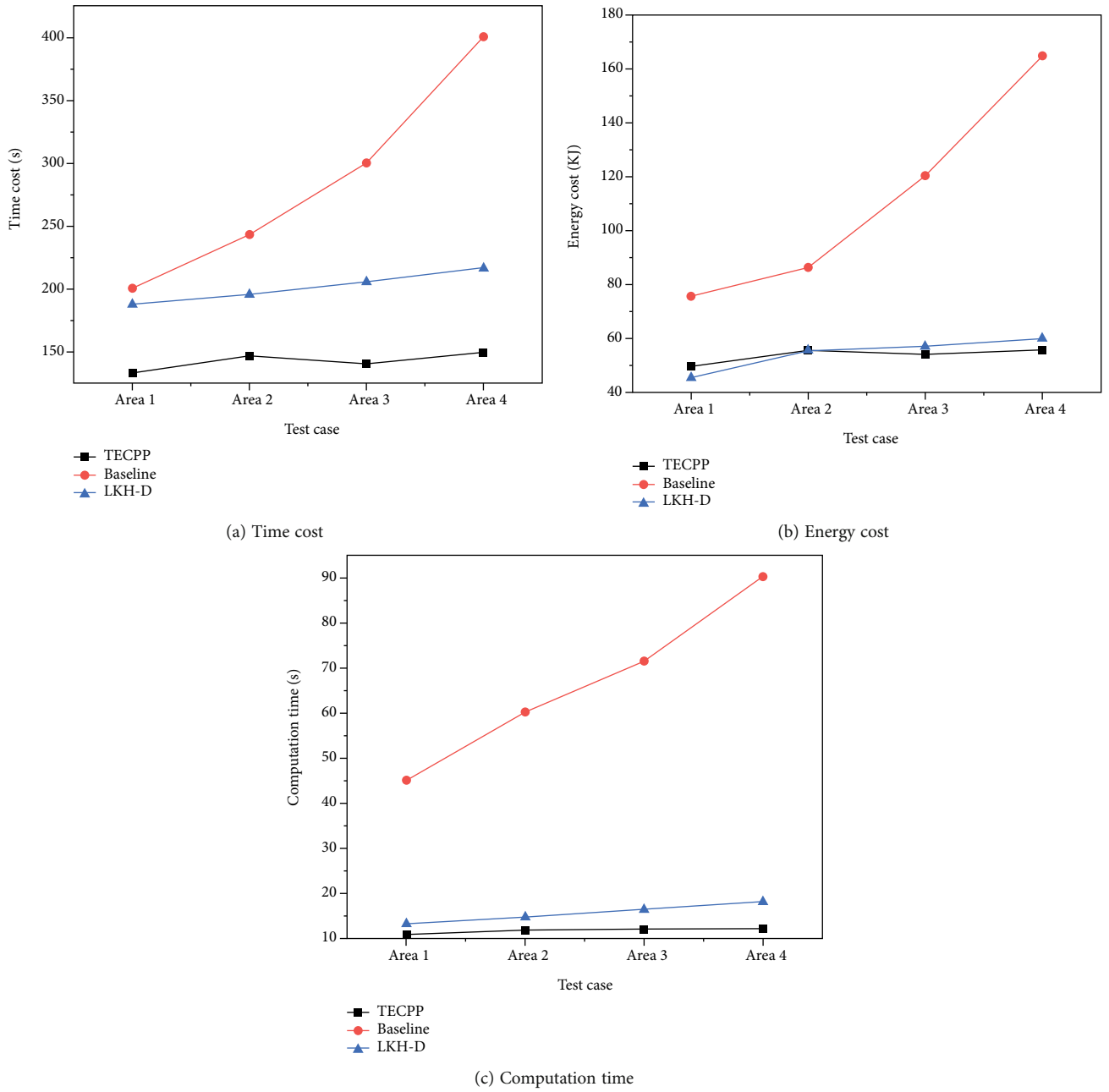
(a) Area 1 (b) Area 2

(c) Area 3 (d) Area 4

Figure 9: Rectangular grid areas used to evaluate TECPP.

LKH-D achieve comparable performance to TECPP for small areas, TECPP saves at least 24.3% of time as the increase of area size. Figure 8(c) shows TECPP is one order of magnitude faster than the baseline for large area.

7.2. The Impact of Obstacles. To elucidate the impact of obstacles on time cost, energy cost, and computation time, we generate four 8×12 ($2\,\text{m} \times 2\,\text{m}$) rectangular areas with different numbers of obstacles as illustrated in Figure 9.



(a) Time cost

(b) Energy cost

(c) Computation time

Figure 10: The impact of obstacles on the time cost, energy cost, and computation time.

Figure 10(a) shows the time cost for the compared algorithms when they are applied to the four areas defined in Figure 9, and Figure 10(b) shows the corresponding energy cost. In the four scenarios, TECPP saves at least 24.9% of time compared to LKH-D and 33.4% of time compared to the baseline. Figure 10(b) shows that TECPP is also better at saving energy than LKH-D as the number of obstacles increases. The reason is that our gadget-based graph model takes the impact of obstacles into account. As shown in Figure 10(c), LKH-D achieves comparable performance to TECPP and these two algorithms save more calculation time than baseline as the number of obstacles increases.

7.3. The Impact of Grid Size. In this subsection, to understand the impact of grid size on the time cost, energy cost, and computation time, we set the size of all grids in Area 4 of Figure 9(d) as $1\,m \times 1\,m$, $2\,m \times 2\,m$, $3\,m \times 3\,m$, and $4\,m \times 4\,m$, respectively. As shown in Figure 11, we compare the performance of TECPP, baseline, and LKH-D in the above four cases. Figure 11(a) shows how the time cost scales with grid size for the compared algorithms, and Figure 11(b) shows the energy cost under each algorithm.

Figure 11(a) shows that TECPP saves at least 21.6% of time than LKH-D and 49.4% of time than the baseline as the grid size increases. Although LKH-D spends the least energy cost compared to the baseline and TECPP as

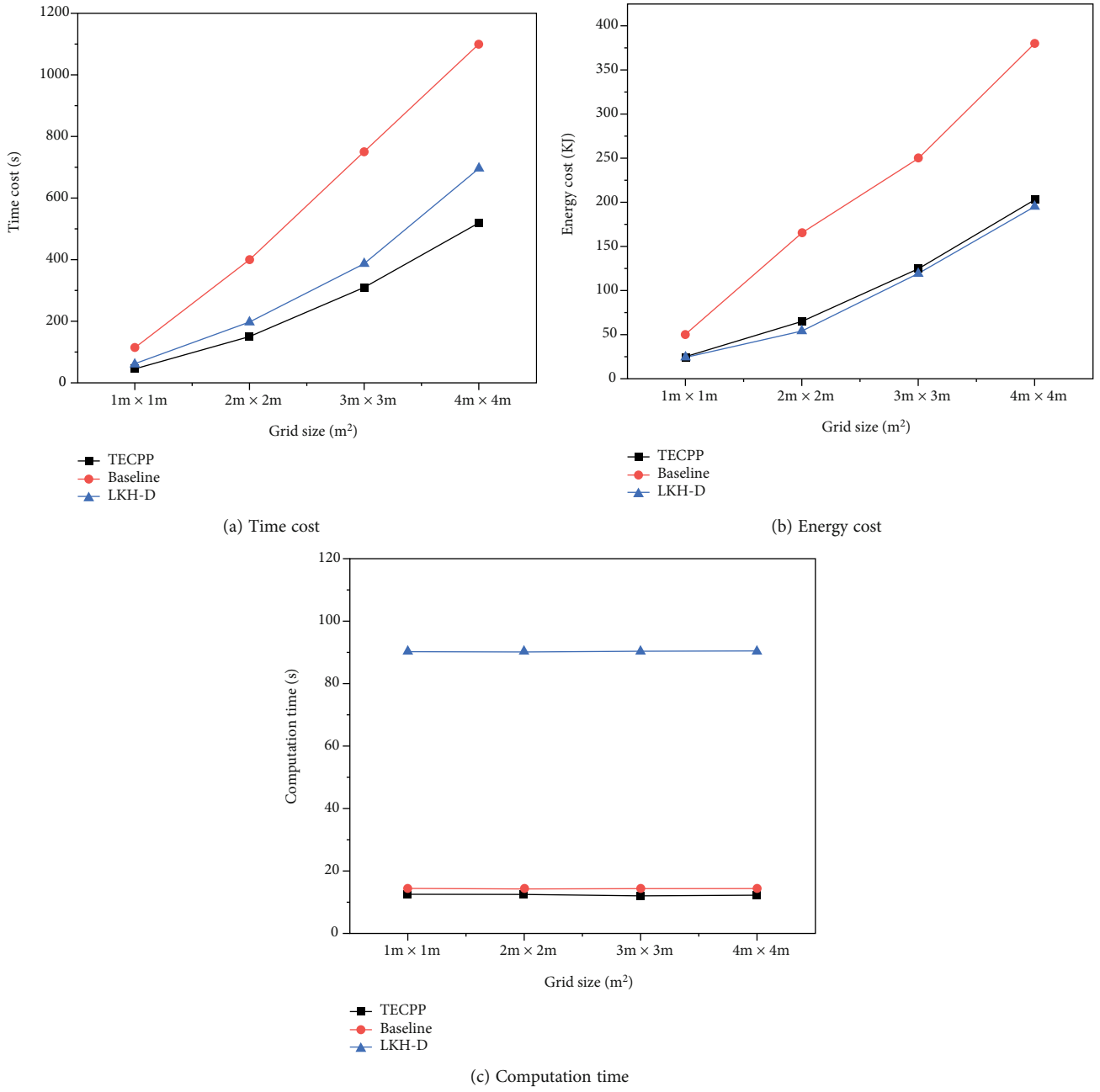


Figure 11: The impact of grid size on the time cost, energy cost, and computation time.

shown in Figure 11(b), TECPP saves more time at a low energy cost as the grid size increases. Figure 11(c) shows that TECPP has a slightly better performance than LKH-D in speed, and the baseline is much slower than the other two algorithms.

In summary, our TECPP takes into account the impact of obstacles and the cost of turns for the CMFCPP problem, which saves at least 21.6% of time and has a faster speed compared to the existing coverage path planning algorithms.

8. Conclusion

A time-efficient coverage path planning (TECPP) algorithm is proposed to solve the CMFCPP problem in this paper. We build a novel gadget-based graph model that takes into account the impact of obstacles, which formalizes the time and energy costs of the flight path including straight flights and making turns (deceleration, turning, and acceleration). Finally, we transform the CMFCPP problem to GTSP and use the efficient GLNS solver to solve the problem. Experimental results show that compared with the existing coverage path planning algorithms, TECPP saves at least 21.6% of time.

As a future work, we intend to explore the path planning problem of UAV covering multiple nonadjacent target areas.

Data Availability

The data mainly come from simulation experiments.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] Z. Hongwei, C. Huailiang, Y. Weidong, and L. Zhongyang, "The application of shortwave infrared perpendicular water stress index in drought monitoring under normal vegetation coverage," in *2011 International Conference on System science, Engineering design and Manufacturing informatization*, pp. 187–190, Guiyang, Oct. 2011.

[2] R. Ma, X. Li, M. Sun, and Z. Kuang, "Experiment of meteorological disaster monitoring on unmanned aerial vehicle," in *2018 7th International Conference on Agro-geoinformatics (Agro-geoinformatics)*, pp. 1–6, Hangzhou, China, Aug. 2018.

[3] R. Q. Ismael and Q. Z. Henari, "Accuracy assessment of UAV photogrammetry for large scale topographic mapping," in *International Engineering Conference (IEC)*, pp. 1–5, Erbil Iraq, June 2019.

[4] E. A. Mitishita and N. L. S. de Salles Graça, "The influence of redundant images in UAV photogrammetry application," in *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*, pp. 7894–7897, Valencia, Spain, July 2018.

[5] K. Yu, J. M. O'Kane, and P. Tokekar, "Coverage of an environment using energy-constrained unmanned aerial vehicles," in *2019 International Conference on Robotics and Automation (ICRA)*, pp. 3259–3265, Montreal, QC, Canada, May 2019.

[6] P. Tokekar, J. Vander Hook, D. Mulla, and V. Isler, "Sensor planning for a symbiotic UAV and UGV system for precision agriculture," *IEEE Transactions on Robotics*, vol. 32, no. 6, pp. 1498–1511, 2016.

[7] P. Yao, Z. Xie, and P. Ren, "Optimal UAV route planning for coverage search of stationary target in river," *Technology*, vol. 27, no. 2, pp. 822–829, 2019.

[8] T. Sherman, J. Tellez, T. Cady et al., "Cooperative search and rescue using autonomous unmanned aerial vehicles," in *AIAA Information Systems-AIAA Infotech@Aerospace*, p. 1490, Kissimmee, Florida, January 2018.

[9] H. Yang, Y. Ye, X. Chu, and S. Sun, "Energy efficiency maximization for UAV-enabled hybrid backscatter-harvest-then-transmit communications," *IEEE Transactions on Wireless Communications*, vol. 21, no. 5, p. 1, 2021.

[10] Z. Na, C. Ji, B. Lin, and N. Zhang, "Joint optimization of trajectory and resource allocation in secure UAV relaying communications for Internet of Things," *IEEE Internet of Things Journal*, 2022.

[11] Z. Na, Y. Liu, J. Shi, C. Liu, and Z. Gao, "UAV-supported clustered NOMA for 6G-enabled Internet of Things: trajectory planning and resource allocation," *IEEE Internet of Things Journal*, vol. 8, no. 20, pp. 15041–15048, 2021.

[12] "10 best long flight time drones: fantastic battery life -3d insider," 2018, https://3dinsider.com/long-flight-time-drones/.

[13] M. Torres, D. A. Pelta, J. L. Verdegay, and J. C. Torres, "Coverage path planning with unmanned aerial vehicles for 3D terrain reconstruction," *Expert Systems With Applications*, vol. 55, pp. 441–451, 2016.

[14] T. M. Cabreira, C. Di Franco, P. R. Ferreira, and G. C. Buttazzo, "Energy-aware spiral coverage path planning for UAV photogrammetric applications," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 3662–3668, 2018.

[15] M. Coombes, W. H. Chen, and C. Liu, "Boustrophedon coverage path planning for UAV aerial surveys in wind," in *2017 International Conference on Unmanned Aircraft Systems*, pp. 1563–1571, Miami, FL, USA, June 2017.

[16] L. Ding, D. Zhao, H. Ma, H. Wang, and L. Liu, "Energy-efficient min-max planning of heterogeneous tasks with multiple UAVs," in *2018 IEEE 24th International Conference on Parallel and Distributed Systems*, pp. 339–346, Singapore, February 2018.

[17] S. L. Smith and F. Imeson, "GLNS: an effective large neighborhood search heuristic for the generalized traveling salesman problem," *Computers & Operations Research*, vol. 87, pp. 1–19, 2017.

[18] J. Valente, D. Sanz, J. Del Cerro, A. Barrientos, and M. Á. de Frutos, "Near-optimal coverage trajectories for image mosaicing using a mini quad-rotor over irregular-shaped fields," *Precision Agriculture*, vol. 14, no. 1, pp. 115–132, 2013.

[19] T. M. Cabreira, P. R. Ferreira, C. Di Franco, and G. C. Buttazzo, "Grid-based coverage path planning with minimum energy over irregular-shaped areas with UAVs," in *International Conference on Unmanned Aircraft Systems*, Atlanta, GA, USA, June 2019.

[20] Y. Choi, Y. Choi, S. Briceno, and D. N. Mavris, "Coverage path planning for a UAS imagery mission using column generation with a turn penalty," in *2018 International Conference on Unmanned Aircraft Systems*, Dallas, TX, USA, June 2018.

[21] Y. Choi, Y. Choi, S. Briceno, and D. N. Mavris, "Energy-constrained multi-UAV coverage path planning for an aerial imagery mission using column generation," *Journal of Intelligent and Robotic Systems*, vol. 97, no. 1, pp. 125–139, 2020.

[22] J. Modares, F. Ghanei, N. Mastronarde, and K. Dantu, "UB-ANC planner: energy efficient coverage path planning with multiple UAVs," in *2017 IEEE International Conference on Robotics and Automation*, pp. 6182–6189, Singapore, June 2017.

[23] S. Piao, Z. Ba, L. Su, D. Koutsonikolas, S. Li, and K. Ren, "Automating CSI measurement with UAVs: from problem formulation to energy-optimal solution," in *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*, pp. 2404–2412, Paris, France, May 2019.

[24] M. Theile, H. Bayerlein, R. Nai, D. Gesbert, and M. Caccamo, "UAV coverage path planning under varying power constraints using deep reinforcement learning," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1444–1449, Las Vegas, NV, USA, January 2020.

[25] J. Steiger, N. Lu, and S. Sorour, "Learning for path planning and coverage mapping in UAV-assisted emergency communications," in *GLOBECOM 2020-2020 IEEE Global Communications Conference*, pp. 1–6, Taipei, Taiwan, Dec. 2020.

[26] U. Challita, W. Saad, and C. Bettstetter, "Interference management for cellular-connected UAVs: a deep reinforcement learning approach," *in IEEE Transactions on Wireless Communications*, vol. 18, no. 4, pp. 2125–2140, 2019.

[27] A. L. Henry-Labordere, "The record balancing problem: a dynamic programming solution of a generalized traveling salesman problem," *RAIRO Operations Research*, vol. 2, pp. 43–49, 1969.

[28] J. P. Saksena, "Mathematical model of scheduling clients through welfare agencies," *Canadian Operational Research Society Journal*, vol. 8, no. 3, pp. 185–200, 1970.

[29] S. S. Srivastava, S. Kumar, R. C. Garg, and P. Sen, "Generalized traveling salesman problem through n sets of nodes,"

Canadian Operational Research Society Journal, vol. 7, no. 2, pp. 97–101, 1969.

[30] S. Lin and B. W. Kernighan, "An effective heuristic algorithm for the traveling-salesman problem," *Operations Research*, vol. 21, no. 2, pp. 498–516, 1973.

[31] J. Alvarenga, N. I. Vitzilaios, K. P. Valavanis, and M. J. Rutherford, "Survey of unmanned helicopter model-based navigation and control techniques," *Journal of Intelligent and Robotic Systems*, vol. 80, no. 1, pp. 87–138, 2015.

Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 4553805, 10 pages
https://doi.org/10.1155/2022/4553805

Research Article

IRS Backscatter-Assisted Security Transmission against Proactive Eavesdropping

Jianling Wang

School of Electronic Information Engineering, Henan Institute of Technology, Xinxiang 453003, China

Correspondence should be addressed to Jianling Wang; 15137372785@hait.edu.cn

Received 16 March 2022; Revised 10 April 2022; Accepted 21 April 2022; Published 5 May 2022

Academic Editor: Lisheng Fan

In this paper, we consider the IRS backscatter-assisted physical layer security, aimed at countering smart eavesdroppers capable of sending jamming signals. Specifically, the eavesdropper increases the eavesdropping rate by sending jamming signals and is able to adjust the transmission strategy according to the received beamforming. By using IRS backscatter communication, the jamming signal sent by the eavesdropper is converted into a useful signal and transmitted. By designing the beamforming of the base station and the IRS phase shift matrix, we established the original optimization problem. Since the eavesdropper's sending strategy is adjusted according to the received signal, we transform the original problem into two subproblems. In the first subproblem, we obtain the eavesdropper's transmit beamforming; in the second subproblem, we optimize the transmit beam alternately with the IRS phase shift matrix. Simulation results demonstrate the superiority of our proposed scheme.

1. Introduction

With the widespread popularity of 5G technology, more and more smart devices are flooding every aspect of people's lives [1, 2]. These smart devices usually require high-speed communication rates to ensure user experience [3–5]. At this time, an unavoidable problem is to ensure the user's communication security [6, 7].

The secure communication in the usual sense focuses on the encryption and encoding of the signal, and the security of the communication is ensured by the key and the complex encryption algorithm [8, 9]. However, the encoding and decoding of encrypted communication will take up extra information, and it is impossible to judge whether the encryption algorithm is decrypted by eavesdroppers [10]. Therefore, physical layer security has received more attention [11].

Early research on user security focused on potential eavesdroppers, increasing the security rate by adding redundant information to the signal [12–15]. However, adding redundant noise will also bring more communication burden to users. Therefore, in the literature [16], etc., it is proposed to use the helper to interfere with the eavesdropper without consuming the transmission power of the user [17, 18].

In recent years, the research on active eavesdropping has also developed gradually [19, 20]. The main purpose of active eavesdropping is not to protect the communication process but to eavesdropper the communication of illegal users. The eavesdropper can send jamming beams to reduce the transmission rate of illegal users, so as to achieve the purpose of successful eavesdropping [21, 22].

Fighting an eavesdropper that uses active eavesdropping mode is a tough job because eavesdroppers can adjust the transmit beam to slow down the communication rate [23]. In this paper, we consider the use of IRS backscatter communication to convert the jamming signal sent by the eavesdropper into a useful signal, thereby increasing the security rate for the user [24, 25].

IRS backscatter is a technique that combines IRS with backscatter communications [26]. Specifically, IRS is a programmable smart material that can reconfigure the input signal to achieve system goals by rewriting the phase and channel [27]. It is worth noting that IRS is a passive device, so it does not need continuous function, which solves the

problem of excessive energy consumption in traditional security communication.

There has been some progress in research on secure communications using IRS backscatter. The authors of [28] propose to use the IRS to assist secure communication, using the jamming beam sent by the eavesdropper to reduce the eavesdropper's signal-to-noise ratio, thereby increasing the security rate. The authors of [29] consider reencoding the received signal by backscattering the IRS to improve the user's acceptance rate. In [30], it further considers the multiuser case. However, none of these works consider the situation where the eavesdropper is actively listening.

In this paper, we consider eavesdroppers to intelligently send jamming signals based on received signals. Through the joint design of the transmit beam of the base station and the IRS phase shift matrix, the maximization of the user security rate is realized. The contributions of this paper are summarized as follows:

(1) We consider the security communication problem of eavesdroppers under active eavesdropping and maximize the security rate by designing the transmit beam of the base station and the phase shift matrix of the IRS

(2) We first pay attention to the design of the eavesdropper's transmit beam under proactive eavesdropping and design corresponding strategies according to the eavesdropper's transmit beam

(3) We use the alternate optimization method to jointly optimize the transmit beam of the base station and the phase shift matrix of the IRS

(4) The simulation results show that our proposed optimization method has a great improvement compared with the existing schemes

Notations: in this paper, we use uppercase letters for matrices, lowercase letters for scalars, and lowercase bold letters for vectors. C stands for the set of real numbers. For the matrix A, A^H represents its conjugate transpose. For the vector a, $\|a\|$ represents its norm.

The structure of this paper is as follows: In Materials and Methods, we first introduce the system model of IRS-assisted secure communication against intelligent listeners and then formulate the optimization problem that maximizes the secure rate. Then, we explore the transmission strategy of the intelligent listener and design the transmission strategy of the base station and the phase shift matrix of the IRS accordingly. In Results and Discussion, we conduct simulation experiments and discuss future work. Finally, we conclude this paper.

2. Materials and Methods

In this section, we first introduce the system model and then analyze the working patterns of eavesdroppers and users to establish an optimization problem. Next, we try to obtain the eavesdropper's transmit beam and then design the phase shift matrix of the base station's transmit beam and IRS based on the beamforming of the eavesdropper.

2.1. System Model. The system model is shown in Figure 1, including base station, legitimate receiver, and eavesdropper. We consider that the eavesdropper is a function that can carry out active eavesdropping; i.e., it can actively send interference signals to reduce the reachable rate of legitimate recipients, so as to achieve the purpose of eavesdropping. It is worth noting that the eavesdropper can obtain the channel state information between the emergency and legitimate receivers but cannot obtain the transmission strategy of the base station. We use IRS scatter communication to convert the jamming signal sent by the eavesdropper into a gain signal.

We set the base station to configure N antennas, both legitimate receivers and the illegitimate eavesdroppers are single antennas to receive, and the illegitimate eavesdropper takes K antennas to send jamming signals. Further, we assume that the elements of the IRS is M. At the same time, the eavesdropper configures multiple antennas to transmit interference information.

The accepted signal at the receiver is

$$y_t = h_{st}^H w s + h_{rt}^H \Phi (h_{sr} w + h_{sr} v) s + h_{et} v z + n_t, \tag{1}$$

where $h_{st} \in \mathbb{C}^{N\times 1}$ is the channel from the base station to the user, $w \in \mathbb{C}^{N\times 1}$ is the beamforming vector sent by base station, $s \in \mathbb{C}$ is the symbol from the base station, $h_{rt} \in \mathbb{C}^{M\times 1}$ is the channel from the IRS to the user, $\Phi \in \mathbb{C}^{M\times M}$ is the IRS phase shift matrix, $v \in \mathbb{C}^{K\times 1}$ is the beamforming vector sent by the eavesdropper, $z \in \mathbb{C}$ is the symbol from the eavesdropper, $h_{et} \in C^{K\times 1}$ is the channel from the eavesdropper to the user, and $n_t \in \mathbb{C}$ is the additive noise with zero mean and variance σ_t^2.

Similarly, the accepted signal at the eavesdropper is

$$y_e = h_{se}^H w s + h_{re}^H \Phi (H_{sr} w + H_{sr} v) s + \rho h_{ee} v z + n_e, \tag{2}$$

where $h_{se} \in \mathbb{C}^{N\times 1}$ is the channel from the base station to the eavesdropper, $h_{re} \in \mathbb{C}^{M\times 1}$ is the channel from the IRS to the eavesdropper, $\Phi \in \mathbb{C}^{M\times M}$ is the IRS phase shift matrix, $h_{ee} \in \mathbb{C}^{K\times 1}$ is the self-interference channel, and $n_e \in \mathbb{C}$ is the additive noise with zero mean and variance σ_e^2.

According to (1) and (2), we calculate the signal-to-interference-noise ratio of the receiver and the eavesdropper, respectively, as

$$\begin{aligned} \mathrm{SINR}_t &= \frac{\left| h_{st}^H w + h_{rt}^H \Phi (h_{sr} w + h_{sr} v) \right|^2}{\sigma_t^2 + \left| \rho h_{et} v \right|^2}, \\ \mathrm{SINR}_e &= \frac{\left| h_{se}^H w + h_{re}^H \Phi (h_{sr} w + h_{sr} v) \right|^2}{\sigma_e^2 + \left| \rho h_{ee} v \right|^2}. \end{aligned} \tag{3}$$

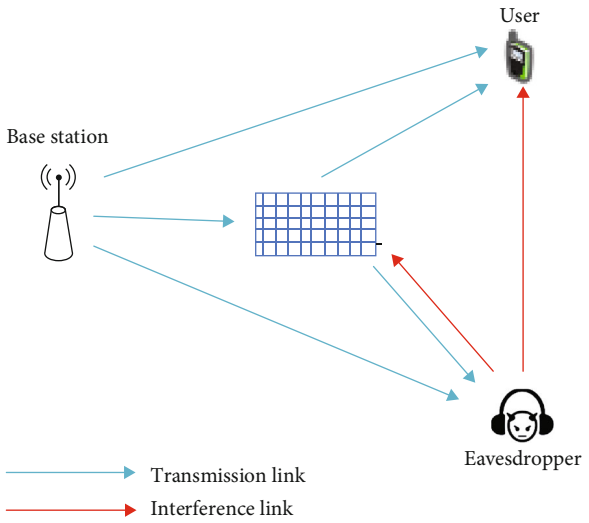


FIGURE 1: System model with IRS backscatter.

Then, we obtain the rate of the user and the eavesdropper as

$$\begin{aligned} R_t &= \log_2(1 + \mathrm{SINR}_t), \\ R_e &= \log_2(1 + \mathrm{SINR}_e). \end{aligned} \tag{4}$$

Then, we formulate the original problem in the following subsection.

2.2. Problem Formulation. Our goal is to maximize the security rate, which is defined as follows:

$$R_s = |R_t - R_e|^+, \tag{5}$$

where $|x|^+ = \max(x, 0)$ for any x.

Meanwhile, the sum energy is limited in the base station and the eavesdropper. The original problem is formulated as

$$\begin{aligned} &\max_{\{\boldsymbol{w},\boldsymbol{\Phi}\}} R_s \\ &\text{s.t.} \|\boldsymbol{w}\|^2 \le P_w \\ &|\theta_m| \le 1, m \in \mathscr{M} \end{aligned} \tag{6}$$

where $\mathscr{M}$ is the set of IRS elements. We assume that MIRS elements in this set.

It is worth noting that the original problem is strongly nonconvex since the optimization variables w and Φ are coupled in the objective of this problem. Further, the beamforming vector sent by the eavesdropper would be changed if we determine to send the beamforming in the base station and the IRS phase matrix. Hence, we need to obtain the relationship between the beamforming $\boldsymbol{v}$ and $\{w, \Phi\}$. In the next subsection, we would like to obtain the strategy of the eavesdropper.

2.3. Beamforming Acquisition for Eavesdropper. In this subsection, we will obtain the eavesdropper's transmit beamforming. It is worth noting that our purpose is not to design the eavesdropper's transmit beam but to simulate the eavesdropper's behavior pattern. We assume that the eavesdropper can select the corresponding transmission beam according to the transmission signal of the base station, and its purpose is to maximize the eavesdropping rate R_a, which is defined as

$$R_a = \begin{cases} R_t, & \text{if } R_t \le R_e, \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

The goal of the eavesdropper is to maximize the eavesdropping rate R_a according to the received signal y_e. Hence, the problem need to be solved by the eavesdropper is proposed as follows

$$\max_{v} R_a \tag{8}$$

$$\text{s.t.} \|v\|^2 \le P_e \tag{9}$$

By substituting the definition of R_a into the objective of (8), we obtain the following problem:

$$\max_{\boldsymbol{v}} R_t \tag{10}$$

$$\text{s.t.} \|v\|^2 \le P_e \tag{11}$$

$$R_t \le R_e \tag{12}$$

The objective of (10) is replaced as the transmission rate of the user with an additional constraint $R_t \le R_e$. For any given $\{w, \Phi\}$, (10) can be solved via Lagrangian dual method. In specific, we define that the optimal v^* in (10) can be decomposed as

$$v^* = \alpha h_{ee} + \beta h_{e,0}, \tag{13}$$

where $h_{e,0}$ is orthogonal to h_{ee} and maximum ratio to H_{sr}. (13) reveals that when the transmission rate is less than the eavesdropping rate. The beamforming vector in the eavesdropper would be zero. When the transmission rate is larger than the eavesdropping rate, the eavesdropper would firstly send jamming signals in the zeros mean of the self-interference channel to improve the eavesdropping rate as possible. When the transmission rate is too large or the maximum power of the eavesdropper is much too small, the eavesdropping rate would be decreased to reduce the transmission rate as much as possible. The results in (13) would be applied in the eavesdropper. Further, when the beamforming vector sent by base station is changed, the corresponding parameters in (13) would be also changed. But the structure of (13) is fixed.

However, since our goal is not to design the beamforming vector in the eavesdropper, we thus transform (10) into a feasibility analysis problem.

$$\text{find } w, \Phi \tag{14}$$

$$\text{s.t.} \|v\|^2 \le P_e \tag{15}$$

$$R_t \geq R_e \tag{16}$$

When (14) is solvable, the eavesdropping rate would be zero. Hence, we can use the results in (13) and substitute it into the original problem.

2.4. Proposed Solution of Original Problem. According to the beamforming design of the eavesdropper, we can obtain that the beamforming vector in (13) is the sending beamforming vector of the eavesdropper. Then, we obtain the following problem:

$$\max_{\{w,\Phi\}} R_s \tag{17}$$

$$\text{s.t.} \|w\|^2 \leq P_w \tag{18}$$

$$|\theta_m| \leq 1, m \in \mathcal{M} \tag{19}$$

$$= \alpha h_{ee} + \beta h_{e,0} \tag{20}$$

Problem (17) is still a nonconvex problem since w and Φ are coupled. Then, we first try to obtain the optimal beamforming vector for any given Φ.

We transform (17) into the following problem:

$$\max_{w} \frac{\left|h_{st}^H w + h_{rt}^H \Phi (h_{sr} w + h_{sr} v)\right|^2}{\left|h_{se}^H w + h_{re}^H \Phi (h_{sr} w + h_{sr} v)\right|^2} \tag{21}$$

$$\text{s.t.} \|w\|^2 \leq P_w \tag{22}$$

$$v = \alpha h_{ee} + \beta h_{e,0} \tag{23}$$

(21) is still nonconvex; we need to apply semidefinite relaxation method to solve it. In specific, we define a new variable $W = ww^H$ with a rank-1 constraint. Then, we define other parameters in the eavesdropper as

$$H_e = \text{diag}\left\{h_{re}^H\right\} h_{sr}, \tag{24}$$

$$f_e = \text{diag}\left\{h_{re}^H\right\} h_{jr} v, \tag{25}$$

$$G_e^{(1)} = \left[H_e^H ; h_{se}\right], \tag{26}$$

$$f_e^{(1)} = \left[f_e^{(1)}, 0\right], \tag{27}$$

$$G_e = \left[G_e^{(1)} ; f_e^{(1)}\right]. \tag{28}$$

Similarly, we define the parameters in the user as

$$H_u = \text{diag}\left\{h_{ru}^H\right\} h_{sr}, \tag{29}$$

$$f_u = \text{diag}\left\{h_{ru}^H\right\} h_{jr} v, \tag{30}$$

$$G_u^{(1)} = \left[H_u^H ; h_{su}\right], \tag{31}$$

$$f_u^{(1)} = \left[f_u^{(1)}, 0\right], \tag{32}$$

Input: $k = 0$, $\Phi^{(0)} = I$, $w^{(0)} = 1/2P_u l$
Output: Φ^*, w^*
Repeat:
Obtain the beamforming vector v;
For fixed $\Phi^{(k)}$, obtain the optimal $W^{(k)}$ in (34);
Recover rank-1 approximation solutions $w^{(k)}$;
For fixed $w^{(k)}$, obtain the optimal $\Phi^{(k+1)}$ in (38);
Set $k = k + 1$;
If norm $(w^{(k+1)} - w^{(k)}) \leq \varepsilon$:
Break;

Algorithm 1: Beamforming and phase matrix design.

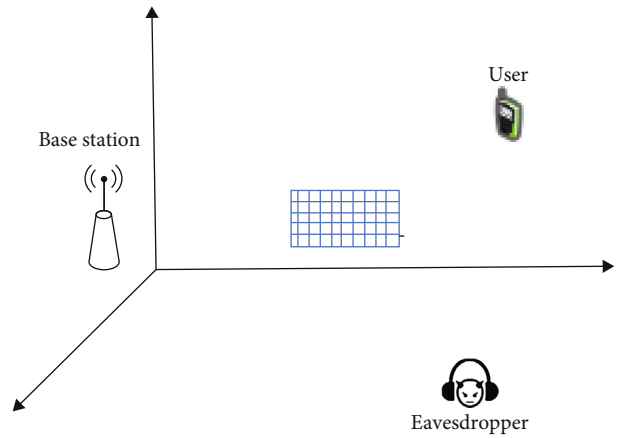


Figure 2: Simulation setup.

$$G_u = \left[G_e^{(1)} ; f_e^{(1)}\right]. \tag{33}$$

It is worth noting that G_e and G_u are both the nonlinear function with respect to the matrix W. By substituting (24) and (29) into (21), we obtain the following expression:

$$\max_{W} \frac{Tr(WG_u)}{\sigma_u^2} - \frac{Tr(WG_e)}{\sigma_e^2} \tag{34}$$

$$\text{s.t.} Tr(W) \leq P_w \tag{35}$$

$$v = \alpha h_{ee} + \beta h_{e,0} \tag{36}$$

$$\text{Rank}(W) = 1 \tag{37}$$

We ignore the rank-1 constraint in (34) and use successive convex approximation (SCA) method to obtain the suboptimal solution of (34). Finally, we apply the rank-1 constraint by Gaussian random method.

Then, we solve the IRS phase matrix for fixed beamforming vector. We first reformulate the following problem:

$$\max_{W} \frac{Tr\left(\Phi G_u^{(1)}\right)}{\sigma_u^2} - \frac{Tr\left(\Phi G_e^{(1)}\right)}{\sigma_e^2} \tag{38}$$

Table 1: Temperature and wildlife count in the three areas covered by the study.

Location	Variable	Value
Location of base station	L_b	(0, 0, 10)
Location of user	L_u	(200, 100, 0)
Location of eavesdropper	L_e	(100, -50, 0)
The channel power gain at a reference distance of $d_0 = 1m$	ρ_0	-30 dB
Maximum power in user	P_u	20 dB
Maximum power in eavesdropper	P_e	20 dB
Power of noise	σ_e	-60 dB
Number of antennas in base station	N	20
Number of elements in IRS	M	30
Number of antennas in eavesdropper	K	50

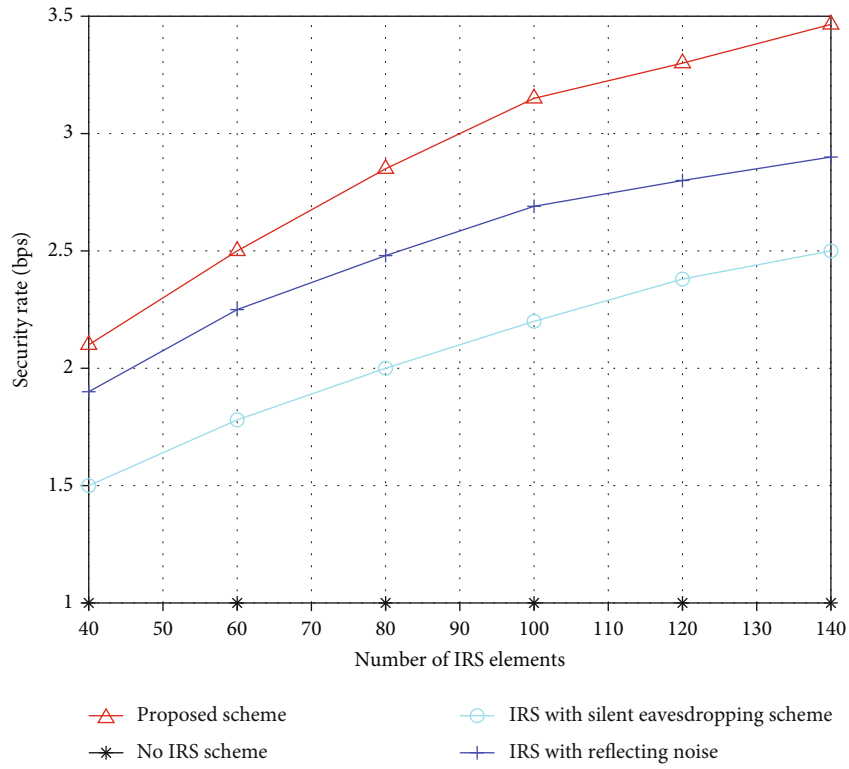


Figure 3: Security rate versus number of IRS element.

$$\text{s.t.} Tr(\Phi) \leq 1 \tag{39}$$

$$v = \alpha h_{ee} + \beta h_{e,0} \tag{40}$$

$$\text{Rank}(\Phi) = 1 \tag{41}$$

Similar as the solution of (34), (38) can be solved by SCA method with ignoring the rank-1 constraint.

The overall algorithm is proposed in Algorithm 1.

Algorithm 1 reveals that the objective function in (34) and (38) does not increase in each iteration, which means that the proposed method will converge to a local optimal solution.

3. Results and Discussion

In this section, we first show the numerical results in the first subsection. We apply the simulation in MATLAB. All results are obtained via Monte Carlo method for 1000 times. We provide the specific parameters and show the superiority of the proposed scheme. Further, we discussed about the finished work and outlook for future work.

3.1. Numerical Results. The specific scenario of our simulation is shown in Figure 2. We assume that the base station is located at the origin (0, 0, 10), the user is located at (200, 100, 0), the IRS is located at (100, 50, 5), and the eavesdropper is located at (100, -50, 0). All signals are empirical

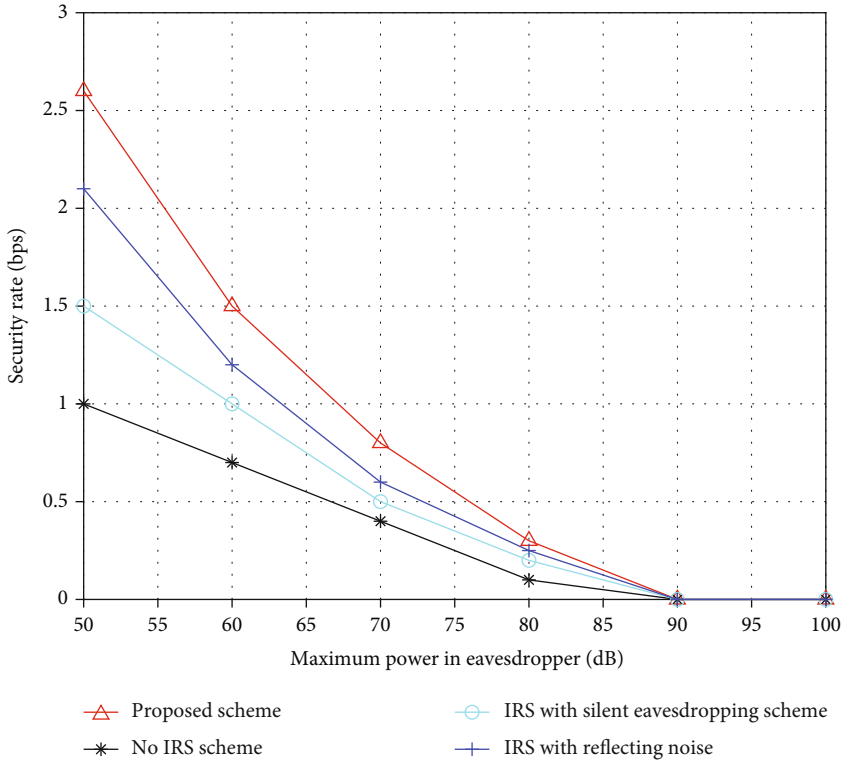


Figure 4: Security rate versus maximum power in eavesdropper.

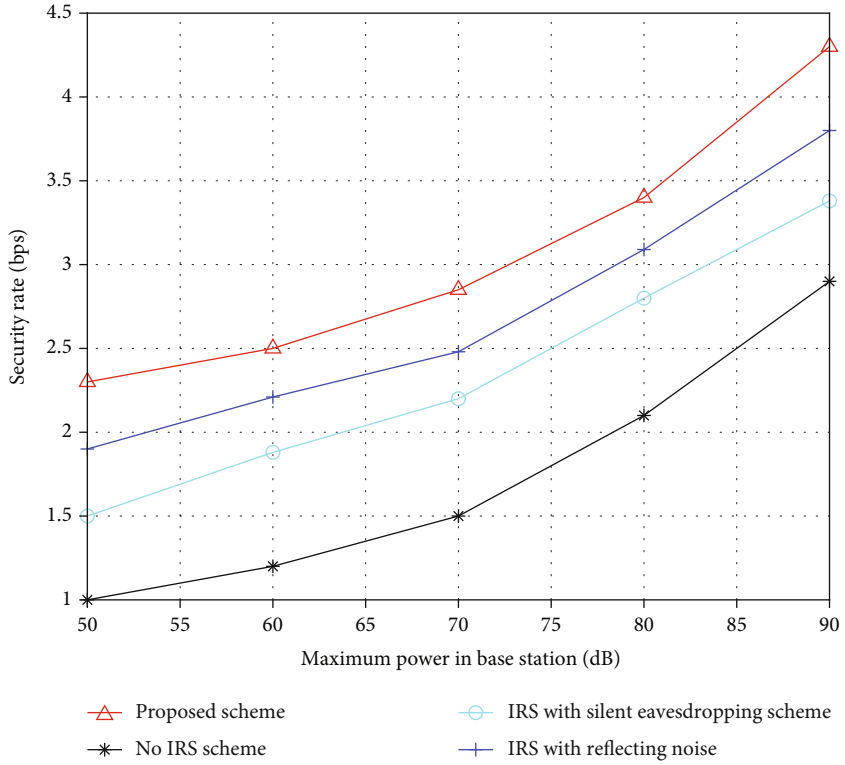


Figure 5: Security rate versus maximum power in base station.

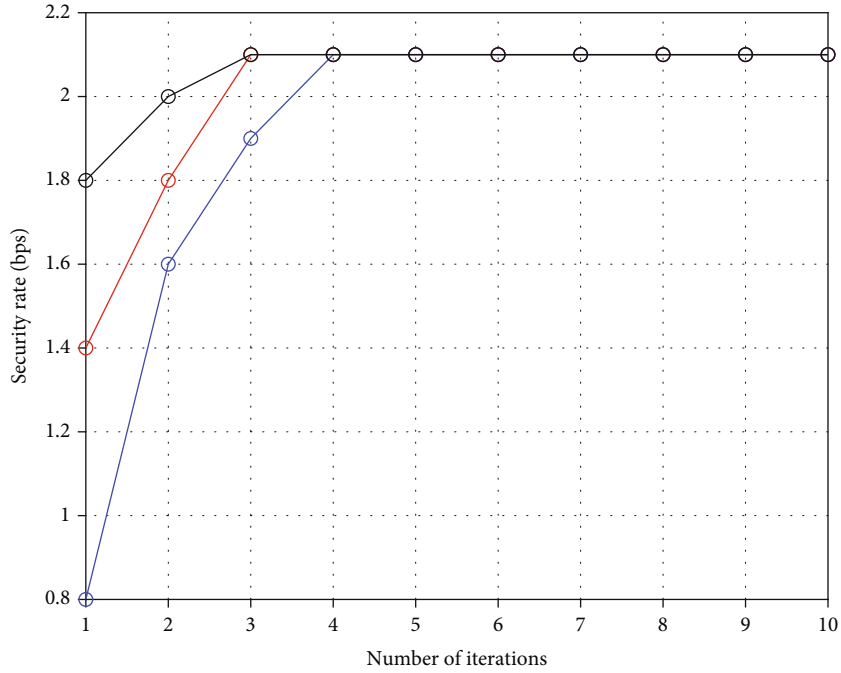


Figure 6: Security rate versus number of iterations.

fading channels, and we use Rayleigh fading model. The specific setting of Rayleigh fading channel is formulated as follows:

$$f(v) = \frac{v}{\sigma_v} \exp\left(-\frac{v^2}{2\sigma_v^2}\right). \tag{42}$$

All parameters are reflected in Table 1.

In this section, we provide four schemes for comparison. The first one is our proposed scheme, the second one is the security transmission without IRS, the third one is the IRS with silent eavesdropping, and the fourth one is the IRS with reflecting noise.

In Figure 3, we show the curve of the security rate versus the number of IRS elements. It can be seen that our proposed scheme outperforms existing schemes. When IRS is not used, its performance does not change with the number of IRS elements. For the other three schemes, an increase in the number of IRS elements brings a significant performance improvement. Further, it can be observed that IRS with reflecting noise would get better performance with respect to IRS with silent eavesdropping since reflecting noise would not only increase the transmission rate but also decrease the eavesdropping rate. For our proposed scheme, the transmission rate would be much larger than the transmission rate in other scheme, which explain the superiority.

We further show the relationship between the security rate and the maximum power at the eavesdropper in Figure 4. It can be observed that when the maximum power at eavesdropper is limited, the security rate for our proposed scheme is better than other schemes, since we transform the jamming noise into useful information for user, which increase the transmission rate. In the scheme of IRS with reflecting noise, the transmission rate is limited by the maximum power in base station. If the eavesdropper's rate is large enough, it must be able to eavesdrop. A possible situation is that the communication rate at this time is close to 0, but this is obviously not what we want to see. Therefore, in order to increase the secure communication rate, a common means is to increase the maximum power of the base station, as we show in the next figure.

Figure 5 shows the variation of the security rate with the power of the base station. When the power of the base station increases, the security rate also increases. In an extreme case, we can send the signal in the null space of the eavesdropping channel and set the corresponding phase of the IRS to the null space as well. At this time, the eavesdropping ability of the eavesdropper can be completely eliminated. However, it is more common that we add some redundant information to improve the overall security performance. When the maximum power in the base station is quite small, our proposed scheme would obtain better performance since the IRS would tend to transform the signal to user but not to the eavesdropper. When the maximum power in the base station is large, the security rate would be better.

In Figure 6, we simulate the performance of the proposed iterative algorithm. It can be found that the convergence speed of the algorithm is very fast for different initial point, and the convergence is achieved in the second or the third iteration. It has been provided that the convergence of the proposed algorithm is not related to the original point selection. Further, although we have proved the convergence of the algorithm in the article, the simulation results intuitively show the convergence speed of our proposed algorithm, which will greatly improve the efficiency of the algorithm.

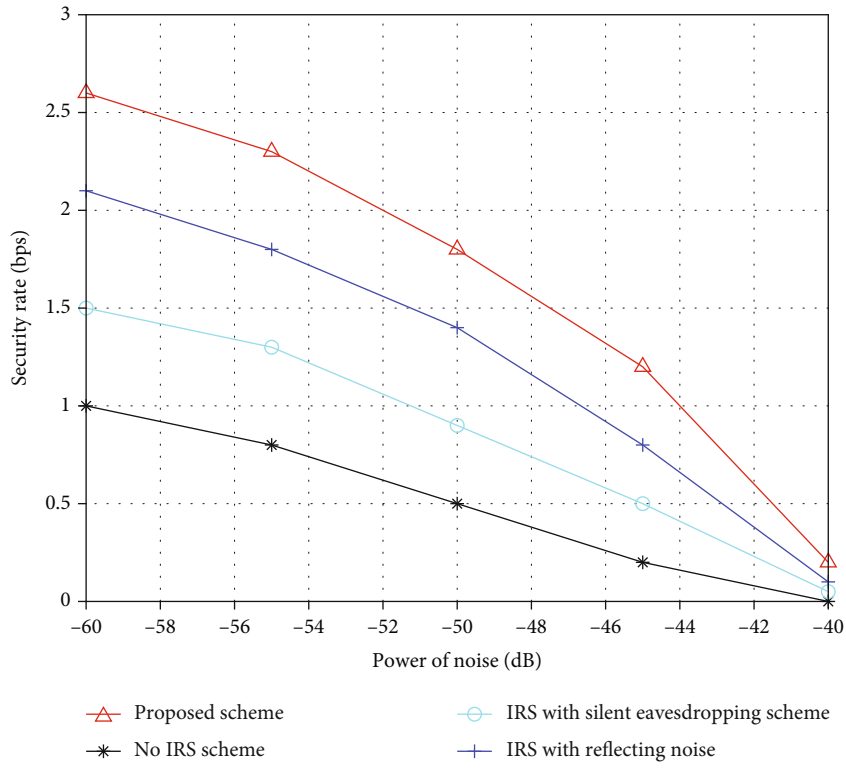


Figure 7: Security rate versus power of noise.

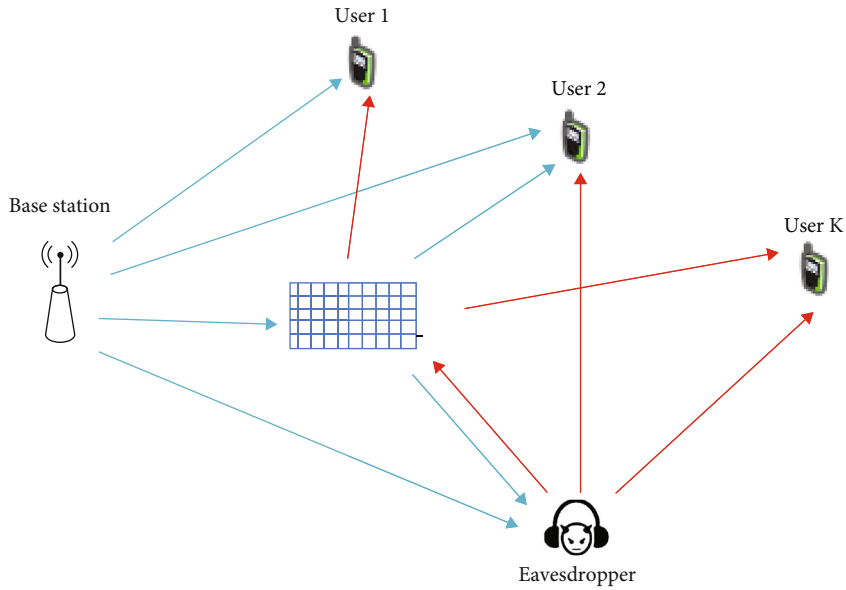


Figure 8: System model with multiuser.

Figure 7 shows the curve of the security rate as a function of noise power. As the noise power increases, the security rate decreases. Although the increase of noise will also reduce the eavesdropping ability of the eavesdropper, we set the background noise power are the same. The increasing of the power of noise would decrease the eavesdropping rate. It can be obtained from the definition of the security rate, and the security rate depends on the transmission capacity of the channel. When the transmission rate is quite small, the security rate would be very small. However, the proposed algorithm can still maintain good performance when the noise power is large, which shows its robustness.

4. Discussion

In this paper, we consider the use of IRS backscatter to enhance secure communications. Our innovations focus on the way eavesdroppers use active eavesdropping and the ability to modify the transmit beam based on environmental factors. We suppress the eavesdropping ability of eavesdroppers through the joint design of the transmit beam of the base station and the phase shift matrix of the IRS.

It is worth noting that we consider the single-user scenario, where it is feasible for the IRS to reencode the information into useful information, but the single-user scenario is relatively rare in practical applications. If there are multiple pairs of users in the same communication area, as shown in Figure 8, the IRS recoding will cause interference to other users, thereby reducing the overall communication efficiency. How to apply IRS backscatter communication to multiuser scenarios requires our further research.

On the other hand, IRS backscatter communication has better performance compared to IRS reflection noise. However, this result cannot be verified theoretically; i.e., we cannot formulate a theorem about this. Therefore, we will also try to prove the optimality condition of backscattering in the follow-up work.

5. Conclusions

In this paper, we use IRS to assist secure communication, and we propose a method for the joint design of base station beam vectors and IRS phase shift matrices, aiming to improve the security rate for users. For the proposed optimization problem, we first consider the case where the eavesdropper sends beams and substitutes its results into the original problem. Further, we carry out an alternate optimization design of the beam vector of the base station and the IRS phase shift matrix. The simulation results show that our proposed algorithm has a better performance improvement compared to the existing benchmark algorithms.

Data Availability

All synthetic data are available on MATLAB.

Conflicts of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by Zigbee-based intelligent wireless monitoring system for urban lighting, Funding Project for Young Key Teachers in Henan Province (grant number 12A510011).

References

[1] S. Han, S. Xu, W.-X. Meng, and L. He, "Channel-correlation-enabled transmission optimization for MISO wiretap channels," *IEEE Transactions on Wireless Communications*, vol. 20, no. 2, pp. 858–870, 2021.

[2] J. Du, C. Jiang, J. Wang, Y. Ren, and M. Debbah, "Machine learning for 6G wireless networks: carrying forward enhanced bandwidth, massive access, and ultrareliable/low-latency service," *IEEE Vehicular Technology Magazine*, vol. 15, no. 4, pp. 122–134, 2020.

[3] X. Yu, D. Xu, Y. Sun, D. W. Ng, and R. Schober, "Robust and secure wireless communications via intelligent reflecting surfaces," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 11, pp. 2637–2652, 2020.

[4] S. Xu, S. Han, W.-X. Meng, Y. Du, and L. He, "Multiple-Jammer-aided secure transmission with receiver-side correlation," *IEEE Transactions on Wireless Communications*, vol. 18, no. 6, pp. 3093–3103, 2019.

[5] W. Yan, X. Yuan, and X. Kuai, "Passive beamforming and information transfer via large intelligent surface," *IEEE Wireless Communications Letters*, vol. 9, no. 4, pp. 533–537, 2020.

[6] R. Long, Y. C. Liang, H. Guo, G. Yang, and R. Zhang, "Symbiotic radio: a new communication paradigm for passive Internet of Things," *IEEE Internet of Things Journal*, vol. 7, no. 2, pp. 1350–1363, 2020.

[7] Q. Wu and R. Zhang, "Joint active and passive beamforming optimization for intelligent reflecting surface assisted SWIPT under QoS constraints," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 8, pp. 1735–1748, 2020.

[8] J. Hu, Y. C. Liang, and Y. Pei, "Reconfigurable intelligent surface enhanced multi-user MISO symbiotic radio system," *IEEE Transactions on Communications*, vol. 69, no. 4, pp. 2359–2371, 2021.

[9] M. Hua, L. Yang, Q. Wu, C. Pan, C. Li, and A. L. Swindlehurst, "UAV-Assisted intelligent reflecting surface symbiotic radio system," *IEEE Transactions on Wireless Communications*, vol. 20, no. 9, pp. 5769–5785, 2021.

[10] Z. Wang, P. Babu, and D. P. Palomar, "Design of PAR-constrained sequences for MIMO channel estimation via majorization-minimization," *IEEE Transactions on Signal Processing*, vol. 64, no. 23, pp. 6132–6144, 2016.

[11] X. Guan, Q. Wu, and R. Zhang, "Joint power control and passive beamforming in IRS-assisted spectrum sharing," *IEEE Communications Letters*, vol. 24, no. 7, pp. 1553–1557, 2020.

[12] B. Di, H. Zhang, L. Li, L. Song, Y. Li, and Z. Han, "Practical hybrid beamforming with finite-resolution phase shifters for reconfigurable intelligent surface based multi-user communications," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 4, pp. 4565–4570, 2020.

[13] L. Dong and H.-M. Wang, "Secure MIMO transmission via intelligent reflecting surface," *IEEE Wireless Communications Letters*, vol. 9, no. 6, pp. 787–790, 2020.

[14] Z. Q. Luo, W. K. Ma, A. M. So, Y. Ye, and S. Zhang, "Semidefinite relaxation of quadratic optimization problems," *IEEE Signal Processing Magazine*, vol. 27, no. 3, pp. 20–34, 2010.

[15] J. Xu, L. Duan, and R. Zhang, "Surveillance and intervention of infrastructure-free mobile communications: a new wireless security paradigm," *IEEE Wireless Communications*, vol. 24, no. 4, pp. 152–159, 2017.

[16] Y. Zeng and R. Zhang, "Wireless information surveillance via proactive eavesdropping with spoofing relay," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 8, pp. 1449–1461, 2016.

[17] D. Hu, Q. Zhang, P. Yang, and J. Qin, "Proactive monitoring via jamming in amplify-and-forward relay networks," *IEEE Signal Processing Letters*, vol. 24, no. 11, pp. 1714–1718, 2017.

[18] Q. Wu and R. Zhang, "Intelligent reflecting surface enhanced wireless network via joint active and passive beamforming," *IEEE Transactions on Wireless Communications*, vol. 18, no. 11, pp. 5394–5409, 2019.

[19] M. Cui, G. Zhang, and R. Zhang, "Secure wireless communication via intelligent reflecting surface," *IEEE Wireless Communications Letters*, vol. 8, no. 5, pp. 1410–1414, 2019.

[20] C. Huang, A. Zappone, G. C. Alexandropoulos, M. Debbah, and C. Yuen, "Reconfigurable intelligent surfaces for energy efficiency in wireless communication," *IEEE Transactions on Wireless Communications*, vol. 18, no. 8, pp. 4157–4170, 2019.

[21] C. Huang, S. Hu, G. C. Alexandropoulos et al., "Holographic MIMO surfaces for 6G wireless networks: opportunities, challenges, and trends," *IEEE Wireless Communications*, vol. 27, no. 5, pp. 118–125, 2020.

[22] L. Subrt and P. Pechac, "Intelligent walls as autonomous parts of smart indoor environments," *IET Communications*, vol. 6, no. 8, pp. 1004–1010, 2012.

[23] C. Liaskos, S. Nie, A. Tsioliaridou, A. Pitsillides, S. Ioannidis, and I. Akyildiz, "A new wireless communication paradigm through software-controlled metasurfaces," *IEEE Communications Magazine*, vol. 56, no. 9, pp. 162–169, 2018.

[24] S. Nie, J. M. Jornet, and I. F. Akyildiz, "Intelligent environments based on ultra-massive MIMO platforms for wireless communication in millimeter wave and terahertz bands," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7849–7853, Brighton, UK, 2019.

[25] P. del Hougne, M. Fink, and G. Lerosey, "Optimally diverse communication channels in disordered environments with tuned randomness," *Nature Electronics*, vol. 2, no. 1, pp. 36–41, 2019.

[26] S. Y. Park and D. I. Kim, "Intelligent reflecting surface-aided phaseshift backscatter communication," in *2020 14th International Conference on Ubiquitous Information Management and Communication (IMCOM)*, pp. 1–5, Taichung, Taiwan, 2020.

[27] C. L. Holloway, E. F. Kuester, J. A. Gordon, J. O'Hara, J. Booth, and D. R. Smith, "An overview of the theory and applications of metasurfaces: the two-dimensional equivalents of metamaterials," *IEEE Antennas and Propagation Magazine*, vol. 54, no. 2, pp. 10–35, 2012.

[28] L. Dai, B. Wang, M. Wang et al., "Reconfigurable intelligent surface-based wireless communications: antenna design, prototyping, and experimental results," *IEEE Access*, vol. 8, pp. 45913–45923, 2020.

[29] A. Pors and S. I. Bozhevolnyi, "Plasmonic metasurfaces for efficient phase control in reflection," *Optics Express*, vol. 21, no. 22, pp. 27438–27451, 2013.

[30] A. S. da Silva, F. Monticone, G. Castaldi, V. Galdi, A. Alú, and N. Engheta, "Performing mathematical operations with metamaterials," *Science*, vol. 343, no. 6167, pp. 160–163, 2014.

Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 7191410, 11 pages
https://doi.org/10.1155/2022/7191410

Research Article

Research on Product Design Strategy Based on User Preference and Machine Learning Intelligent Recommendation

Jie Wu

Design Department, Taiyuan Normal University, Taiyuan 030619, China

Correspondence should be addressed to Jie Wu; wujie@tynu.edu.cn

Received 15 February 2022; Revised 18 March 2022; Accepted 9 April 2022; Published 28 April 2022

Academic Editor: Lisheng Fan

In the machine learning model, intelligent recommendation system can select valuable information from a lot of data to help users find the products or services they need, which has been more and more widely used in recent years. However, there are still many problems in machine learning recommender systems, such as data sparsity, natural noise, and cold start, which leads to the fact that machine learning recommender systems cannot obtain accurate user preferences. When a project is rated, the quality of the recommendation is greatly affected. In order to solve the problem that the existing recommendation algorithms have poor recommendation results in sparse data sets, this paper proposes a machine learning method for recommendation rating prediction based on user interest concept lattice. Firstly, the nearest neighbors are divided into direct nearest neighbors and indirect nearest neighbors by user interest concept lattice. Then, different methods are used to calculate the similarity between the direct "nearest neighbor" and the target user, and the similarity between the indirect "nearest neighbor" and the target user. Finally, the invisible item score of the target user is calculated by the similarity value. Experiments are carried out on real data sets, and the experimental results show that the CFCNN-CL algorithm and RRP-UI CL algorithm proposed in this paper have high recommendation accuracy and still have good performance in the case of sparse data.

1. Introduction

With the development of the Internet, users can access it through various devices and services. Users are more involved in the project selection process by directly controlling the items to be accessed (such as film and television dramas, music, clothing, websites, travel, accommodation, e-learning materials, gadgets, and applications), and there are many different items to choose from around each user. With the increase of information and data scale in the Internet, it is difficult for users to find interesting projects in a reasonable time, and the project selection process may become tedious and complicated [1]. In order to prevent users from choosing items among tens of millions of items and recommending items to people according to their preferences, recommender system for machine learning is introduced [2]. The recommendation system tracks the interaction information between users and their selected items and then uses this information to process into a user model through recommendation algorithm, which is used to filter out the items that users are interested in and recommend the results to users in the form of personalized list [3]. According to user's needs, interests, etc., create a list of items that users are interested in, without a lot of interaction with users [4]. Recommendation system helps users to solve the problem of too many products and difficult to choose and provides them with personalized services. Users can make appropriate purchase decisions and explore new products from the best product evaluation through the minimum online search cost. Now, recommendation system has been fully mined, it has appeared in any services that require users to make decisions, including e-commerce, information retrieval, navigation information services, social networks, and other fields [5].

The two most commonly used technologies in the development of recommendation system are content-based technology and collaborative filtering technology. Among them, content-based technology extracts the features of items first and then can provide items with similar features selected by users in the past [6]. The technology based on

collaborative filtering mainly relies on the historical records provided by users to predict the items they are interested in and mainly depends on the scoring data, which is easy to implement and has high recommendation accuracy [7]. Collaborative filtering has become the most popular recommendation algorithm at present [8]. It uses user scores to build user-user or item-item similarity index and identifies the "nearest neighbor" of users or items to generate recommendations. Collaborative filtering mainly includes neighborhood-based and model-based methods, both of which have their own advantages and disadvantages. Neighborhood-based recommendation has high accuracy, but if new users join, it will reduce performance. The model-based model has better scalability and makes up for the shortcomings of the neighborhood-based model, but the recommendation accuracy is low [9]. Compared with the traditional recommendation method, this paper adopts the nearest neighbor similarity comparison method. Firstly, the nearest neighbor is divided into direct nearest neighbor and indirect nearest neighbor by user interest concept lattice. Then, different methods are used to calculate the similarity between the direct "nearest neighbor" and the target user, and the similarity between the indirect "nearest neighbor" and the target user. Finally, the invisible item score of the target user is calculated by the similarity value. On the basis of direct nearest neighbor, an indirect nearest neighbor similarity comparison method is proposed to further obtain the optimal recommended value. Compared with traditional methods, the recommended methods in this paper are better in integrity.

2. Recommendation System Theory

The purpose of researching recommendation system is to retrieve the most relevant products and services from a large amount of data, so as to reduce information overload and provide personalized services [10]. In 1990s, recommendation system was first applied to e-commerce and Web services. In recent years, people have developed various recommendation system software for social networks, digital libraries, e-commerce, and online advertising [11]. This section mainly summarizes the commonly used recommendation algorithms and common problems in the current recommendation system.

2.1. Overview of Intelligent Recommendation Algorithms. Recommendation system can be defined as a program, which predicts users' interest in projects based on projects, users, and interaction information between projects and users, so as to recommend the most suitable projects (products or services) to specific users (target users). In recommendation system, the quality of recommendation has a great relationship with the performance of recommendation algorithm. The following will introduce the common intelligent recommendation algorithms [12, 13].

2.1.1. Collaborative Filtering (CF) Recommendation Algorithm. CF is to recommend target users by analyzing the scoring information of other users or other items, and the recommendation accuracy is higher. Two main recommendation algorithms will be introduced below.

Table 1: User-item scoring matrix.

	I_1	I_2	I_3	I_4	I_5
U_1	3	0	0	0	1
U_2	0	4	0	5	0
U_3	0	2	4	0	0

*(1) **CF** Based on Neighborhood*. In the neighborhood-based collaborative filtering recommendation algorithm, finding similar users is a key step, and the main goal of similar users is to get the most suitable recommendation items for the target users. The user-based algorithm is mainly divided into three steps: first, calculating similarity; the second is to choose the "nearest neighbor" according to the similarity; the third is to calculate the score value and make prediction and recommendation. Next, we will introduce the most used methods to calculate similarity.

Adjusted Cosine (ACOS) similarity in user u and v is calculated using Equation (1).

$$\mathrm{sim}(u,v)=\frac{\sum_{i\in I_u\cap I_v}(r_{u,i}-\overline{r_u})(r_{v,i}-\overline{r_v})}{\sqrt{\sum_{i\in I_u}(r_{u,i}-\overline{r_u})^2}\sqrt{\sum_{i\in I_v}(r_{v,i}-\overline{r_v})^2}}. \quad (1)$$

Pearson's Correlation (PC) is used to calculate the similarity between u and v.

$$\mathrm{sim}(u,v)=\frac{\sum_{i\in l_u\cap l_v}(r_{u,i}-\overline{r_u})(r_{v,i}-\overline{r_v})}{\sqrt{\sum_{i\in l_u\cap l_v}(r_u,i-\overline{r_u})^2}\sqrt{\sum_{i\in l_u\cap l_v}(r_{vi}-\overline{r_v})^2}}. \quad (2)$$

Constrained Pearson's Correlation (CPC) is using Equation (3)to calculate the similarity between u and v.

$$\mathrm{sim}(u,v)=\frac{\sum_{i\in I_u\cap I_v}(r_{u,i}-r_{\mathrm{med}})(r_{v,i}-r_{\mathrm{med}})}{\sqrt{\sum_{i\in I_u\cap I_v}(r_{u,i}-r_{\mathrm{med}})^2}\sqrt{\sum_{i\in I_u\cap I_v}(r_{v,i}-r_{\mathrm{med}})^2}}, \quad (3)$$

where r_{med} is the median of the grade.

The Jaccard similarity between u and v is calculated by using Equation (4).

$$\mathrm{sim}(u,v)=\frac{|I_u\cap I_v|}{|I_u\cup I_v|}, \quad (4)$$

where $|I_u\cap I_v|$ is the same number evaluated by u and v together.

$$\mathrm{R_{u,i}}=\overline{r_u}+\frac{\sum_{v\in N_u}sim(u,v)(r_{v,i}-\overline{r_v})}{\sum_{v\in N_u}sim(u,v)}, \quad (5)$$

where $\mathrm{sim}(u,v)$ is the similarity of user u and v, and n_u is the

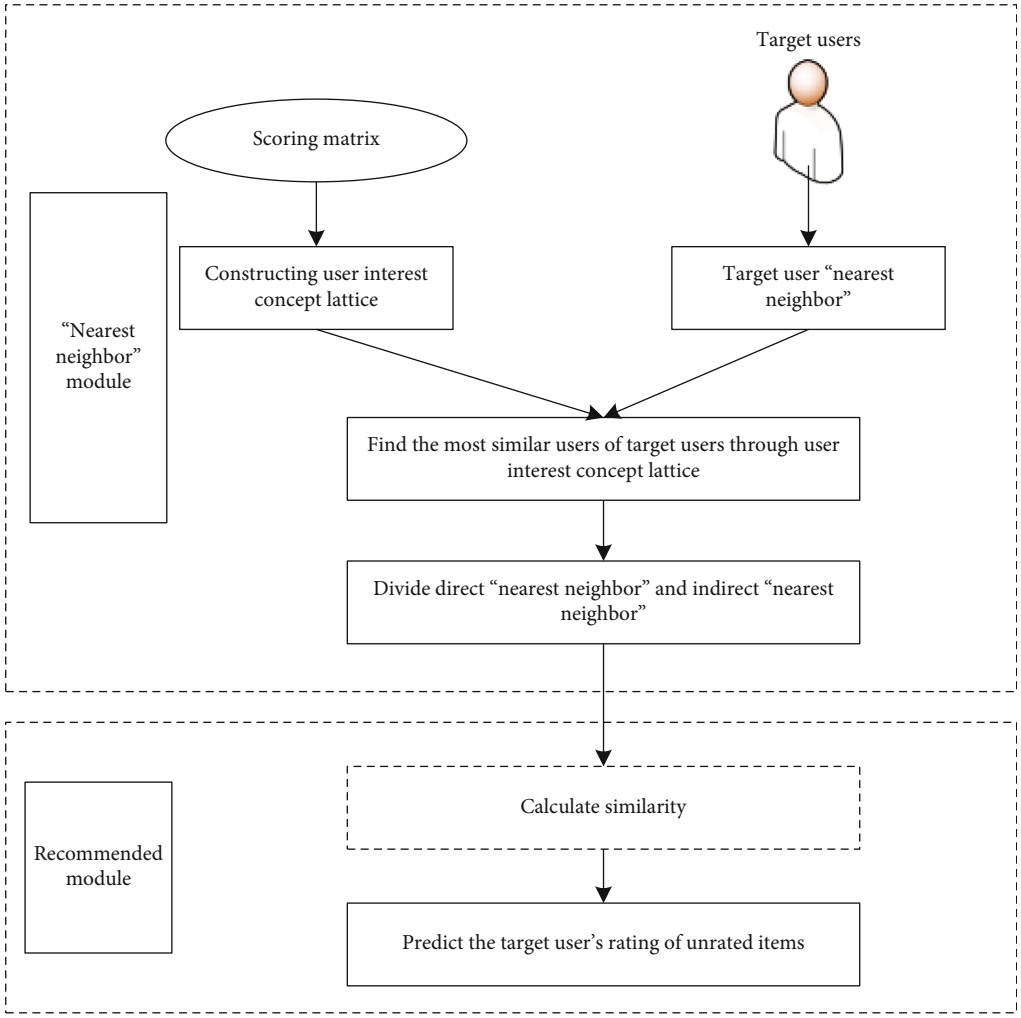


Figure 1: RRP-UICL algorithm model.

Table 2: Scoring matrix.

	I_1	I_2	I_3	I_4	I_5	I_6	I_7
U_1	0	0	0	5	0	4	1
U_2	0	4	4	5	2	0	4
U_3	0	4	4	0	5	1	2
U_4	1	2	5	4	0	3	4
U_5	4	1	0	5	0	5	0
U_6	5	3	4	5	2	0	4

"nearest neighbor" set of user u. Project-based algorithm and user-based algorithm have the same calculation principle, but the calculation objects are different, and users need to be exchanged for projects.

(2) Collaborative Filtering Based on Model. The model-based algorithm is mainly divided into two main stages. In the first stage, we need to deal with the original scoring matrix and construct an effective model representing the original matrix. In the second stage, we use the generated model as an input matrix to predict the scoring of target users. The core of this algorithm is the establishment of model, which needs to use historical information to create and generate recommended models, among which singular value decomposition (SVD) is the most widely used model [14].

In the SVD model, the original scoring matrix R is decomposed into three matrices, and the decomposition form is

$$R_K = USV^T, \tag{6}$$

where u and v are two the orthogonal matrices, s is a diagonal matrix of size $r^* r$, and r is the rank of matrix R, which is composed of singular values of scoring matrix. The matrix can be reduced by discarding the minimum value, and finally, the matrix s is obtained, where $k < r$; then, the decomposition form of the reconstructed matrix is

$$R_K = U_K S_K V_K^T. \tag{7}$$

Table 3: Background of user interest form.

	I_1	I_2	I_3	I_4	I_5	I_6	I_7
U_1	0	0	0	1	0	1	0
U_2	0	1	1		0	0	1
U_3	0	1	1	0	1	0	0
U_4	0	0	1	1	0	0	1
U_5	1	0	0	1	0	1	0
U_6	1	0	1	1	0	0	1

The scoring prediction formula is

$$\mathrm{R_{u,i}} = \overline{r_u} + \mathrm{U_K}\sqrt{\mathrm{S_k^T}(u)}\sqrt{\mathrm{S_K}}\mathrm{V_k^T}(i). \tag{8}$$

The recommendation algorithm based on collaborative filtering does not need detailed content. When the details of content cannot be accessed or it is difficult to collect or analyze the details, collaborative filtering method is very effective, and this method can find the items that target users want in a large number of items [15]. However, it will also face the problem of rating sparsity, and there will also be the problem of cold start of new users and new projects.

2.1.2. Content-Based. Content-based is the earliest recommendation algorithm, by comparing the characteristic information contained in the project with the characteristic information interested by the target user, the project is recommended to the target user, and the foundation is to find similar items by the target users before. For providing appropriate recommendations to the target users, accurate user characteristics, preferences, and demand models are needed. Firstly, the system extracts the feature information contained in each item, then classifies the items used by the target users before, extracts the feature information of the items, and then learns the feature information to obtain the user's preference characteristics. Finally, compare the user's preference characteristics with the feature information contained in the items and recommend the users through the correlation.

At present, *TF-IDF* is the most commonly used computational method in information retrieval, which is used to develop vector space model in content-based recommendation algorithm. In this method, the project content is regarded as a document D, and then, the keyword T is extracted from it, and the calculation formula of the *TF* value of the keyword T in the document D is shown in (9).

$$\mathrm{TF_{t.d}} = \frac{N_{t.d}}{\sum_k N_{k.d}}, \tag{9}$$

where $N_{t,d}$ represents the number of times the keyword t appears in the document d, and the calculation formula of *IDF* value corresponding to the keyword is shown in Equation (10).

$$\mathrm{IDF_t} = \log \frac{|D|}{1 + |d \in D : t \in d|}, \tag{10}$$

where D denotes the set of documents, and $1 + |d \in D : t \in d|$ denotes the number of keywords t contained in document d.

The Rocchio algorithm is usually used to deal with the relevance feedback in the process of information retrieval and extract the interesting feature information of the target users. Decision tree algorithm, linear classification algorithm, and Naive Bayes method are used to classify documents, and documents are interested or uninterested.

Content-based recommendation algorithm does not have the problem of data sparsity, and new items can be recommended immediately. However, because the algorithm needs to extract the feature information of the project, At the same time, the algorithm only relies on the behavior information of the target users to recommend and does not involve the behavior information of other users. There are many problems in diversity. When new users enter the recommendation system, they also face the cold start problem when selecting items.

2.1.3. Hybrid. Hybrid combines two or more recommendation algorithms to predict and recommend and improve the recommendation accuracy.

In the recommendation algorithm based on the combination of content and collaborative filtering, the prediction value based on content algorithm can be used to supplement the user's historical scoring data, adding data to form a pseudoscore matrix, in which the observed scores remain unchanged, and then, using collaborative filtering algorithm based on weighted Pearson's correlation to predict the pseudoscore matrix, the recommendation algorithm has better prediction performance and also overcomes the cold start problem and data sparsity problem.

2.2. Frequently Asked Questions on Recommendation Systems. There are some problems in the current recommendation system. At present, the common problems in the recommendation system are as follows:

2.2.1. Data Sparsity Problem. Data sparse refers to the lack of useful scoring data when recommending items to target users, which leads to the error between recommended items and users' needs.

In most recommendation systems, each user only evaluates a part of the available items, so most of the evaluation information is empty. When users only grade a few items, there will be great errors in the similarity between different users or items, and at this time, the recommendation quality of recommendation algorithm will be greatly affected.

Sparsity is related to the scoring data hidden in the recommendation system, the number of scores can be measured by sparsity, which indicates the ratio of the number of unscored data to the whole matrix space in a scoring matrix. Assuming that a scoring matrix has U users, I items and R scores in total, and S is used to represent the sparsity

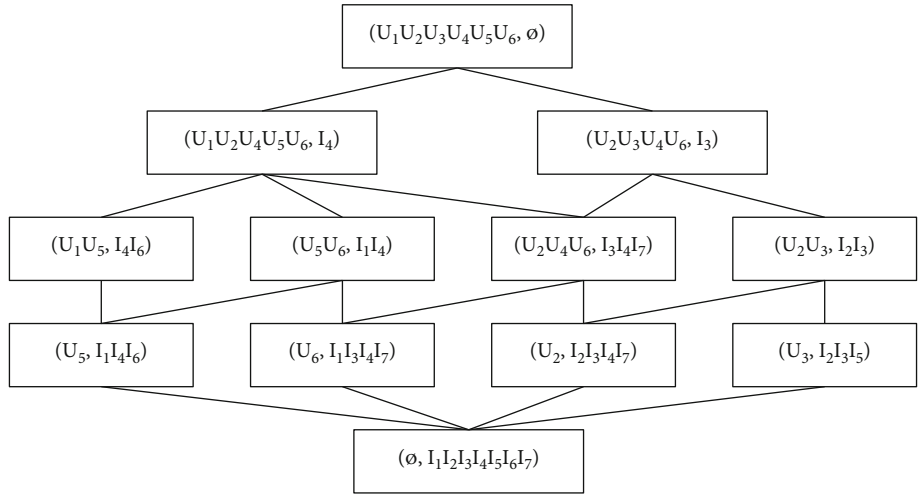


Figure 2: User interest concept lattice.

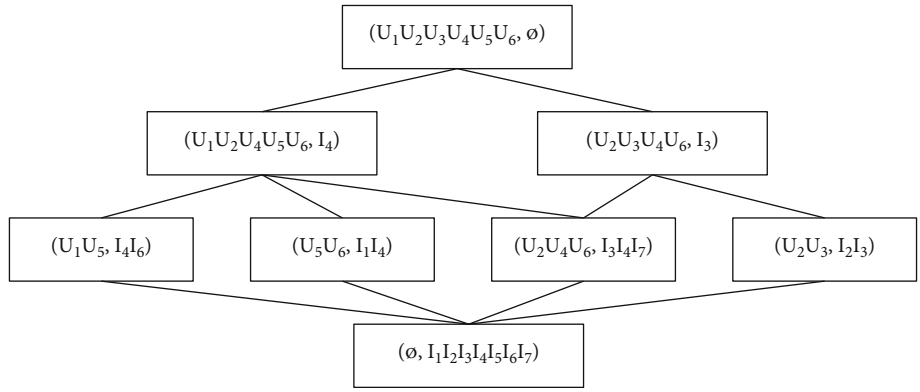


Figure 3: Final user interest concept lattice.

of the data set, the calculation formula of sparsity S is

$$S = \frac{R}{\mathrm{UI}}. \tag{11}$$

Generally, neighborhood-based collaborative filtering algorithms use similarity to find users similar to recommended users or items similar to candidate items. The similarity between items is also calculated using the scores provided by users. However, if there are few or no common scoring items in the given scoring data, these methods become inapplicable.

2.2.2. Noise Issues. Noise in recommendation system refers to the data that will affect the score prediction in the data set. Noise in recommendation system data set can be divided into malicious noise and nonmalicious noise (natural noise), both of which are very important and will have adverse effects on recommendation performance.

Malicious noise refers to the behavior that some biased data are intentionally added to the system, which is intentionally introduced by external agents and intentionally deviates the output of the system in a specific way, which has a great impact on the recommendation performance. Foreign agents will maliciously attack the recommendation system in order to have significant advantage in the recommendation system. Because many recommendation systems run in the business environment, some people will use the recommendation system to seize the advantage in the business competition. For example, if authors hope to promote their work by exporting artificially high reviews for their publications through the recommendation system, and at the same time reduce the recommendations for other similar works, they will find some people to improve the false scores, resulting in biased recommendation results.

Natural noise is the output data of users' real evaluation, which is produced by users' activity errors. This kind of noise is related to the method of collecting or inferring users' preferences in recommendation system. Because all human activities are error-prone, and the user's preference output is usually a heavy process, some errors will naturally appear in the data. In the data set noise of recommendation system, this paper mainly studies the natural noise.

2.2.3. Cold Start Problems. In the recommendation system, the cold start user problem refers to the fact that the system

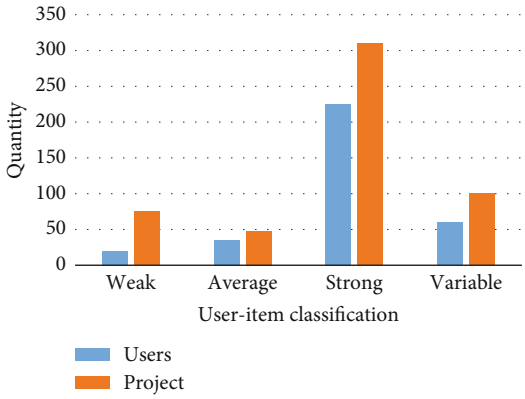


Figure 4: Sample data set user-item classification.

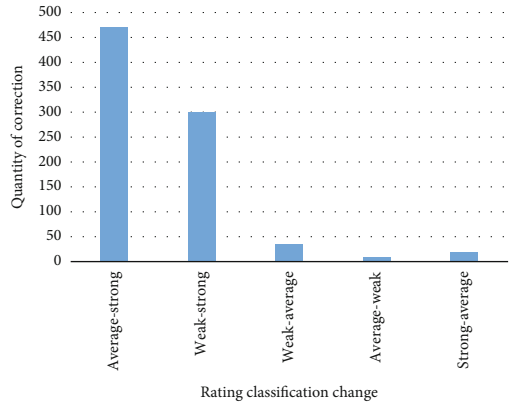


Figure 5: Number of revised ratings in sample data set.

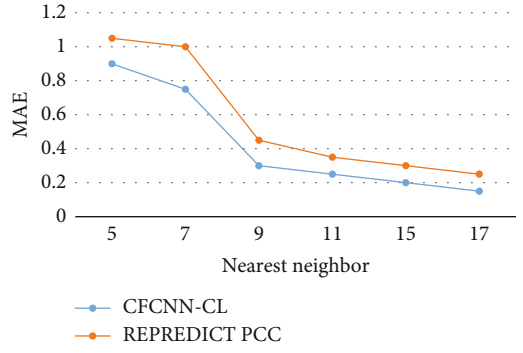


Figure 6: MAE comparison of different nearest neighbors.

cannot recommend related items for the user when the user is a new user, because there is a lack of item scoring history information to help determine the user's interest. Similarly, an item can only be recommended after a large number of users have rated it. For an item that has never been evaluated by users, the system usually cannot make high-quality suggestions. This problem is called cold start project problem.

The cold start problem is caused by the lack of user data and project scoring history. Cold start problems can be mitigated by adding information about user items, and valuable data can be provided to determine users' interest in items by identifying trust relationships between users and the influence of one user on another, which is very useful for making suggestions to users more accurately and objectively.

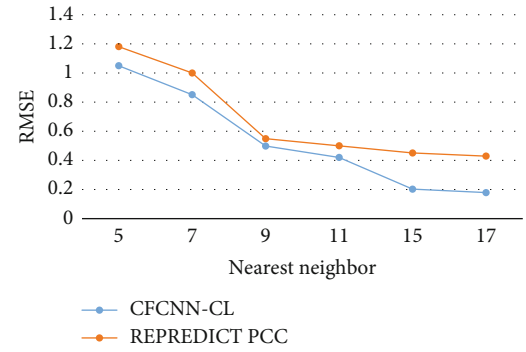


Figure 7: RMSE comparison of different nearest neighbors.

2.2.4. Scalability Issues. As the number of users and projects gradually increases, scalability problems arise. The recommendation system not only needs to deal with the interaction between the original users and projects but also needs to respond to the interaction information between new users and projects. Therefore, the recommendation system needs to deal with a large amount of data, which requires powerful computing power to execute and quick response to the needs of online users. In the recommendation system, the scalability of the system also needs to be considered. A recommendation system with good scalability can quickly deal with the needs of a large number of users and recommend accurate items.

3. Recommendation Score Prediction Algorithm Based on User Interest Concept Lattice

3.1. Problem Description and Analysis. Recommendation system mainly depends on the information left by users after browsing. Among this information, the explicit feedback information between users and items is very important. Among them, the user's rating data is the most commonly used explicit feedback information. The higher the user's rating on an item, the more the user likes and interests the item.

In the recommendation system, collaborative filtering algorithm can achieve good results when there are more score data. However, because some users have no habit of scoring after using the project, they cannot give the system a clear feedback on their love for the project, and the scoring data in most system databases will become very few, which leads to the recommendation system cannot recommend satisfactory projects to the target users well. Among them, the "nearest neighbor" selection is based on the assumption that if two users have similar scores for common items, they can be regarded as having similar preferences, and the services received by one user may be recommended to another user. In the implementation of the algorithm, the most commonly used similarity calculation method is Pearson's Correlation (PC) coefficient, and the similarity between the

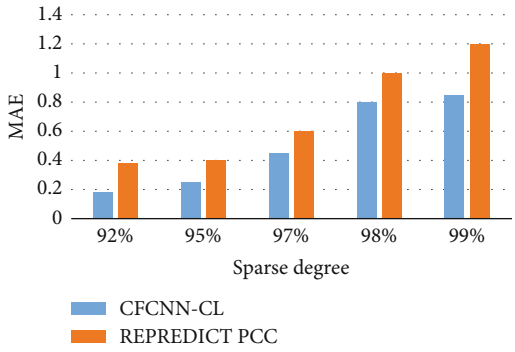


Figure 8: MAE comparison with different sparsity.

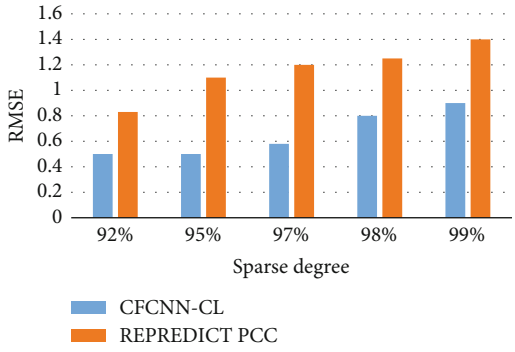


Figure 9: Comparison of RMSE with different sparsity.

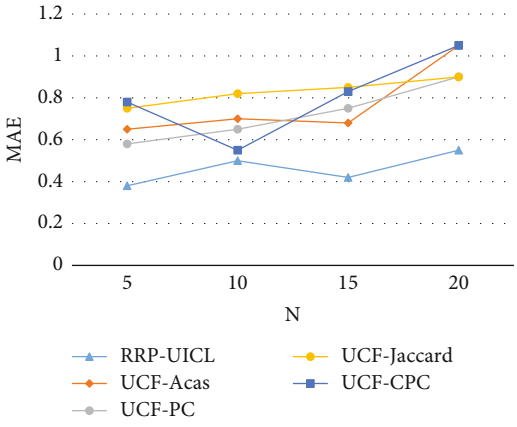


Figure 10: Comparison of MAE values of different methods in data set-1.

target user u and the neighbor user v is calculated by formula (12).

$$\mathrm{sim}(u,v)=\frac{\sum_{i\in I_u\cap I_v}(r_{u,i}-\overline{r_u})(r_{v,i}-\overline{r_v})}{\sqrt{\sum_{i\in I_u\cap I_v}(r_{u,i}-\overline{r_u})^2}\sqrt{\sum_{i\in I_u\cap I_v}(r_{v,i}-\overline{r_v})^2}}. \quad (12)$$

It can be seen from the expression of calculating similarity that the calculation of similarity mainly depends on $I_u \cap I_v$, that is, the common item set of target user u and neighbor user v. However, in the actual user-item scoring data set, the scoring data is very few, and the common item set that can be provided will be very few, which will affect the calculation of similarity. After the similarity degree is calculated, the

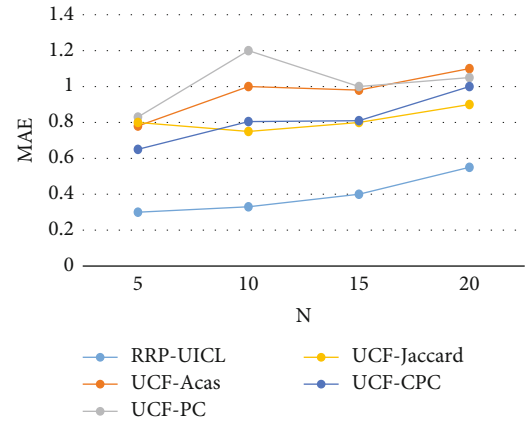


Figure 11: Data set-2 comparison of MAE values for different methods.

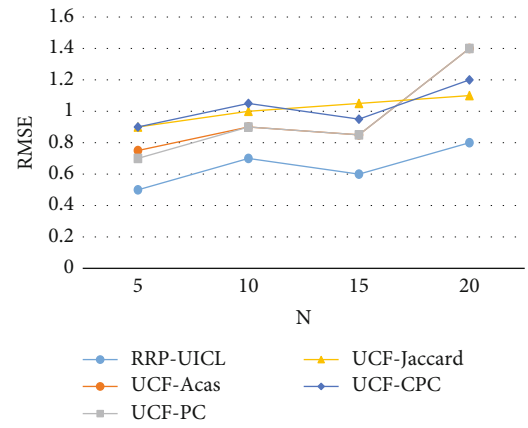


Figure 12: Comparison of RMSE values of different methods in data set-1.

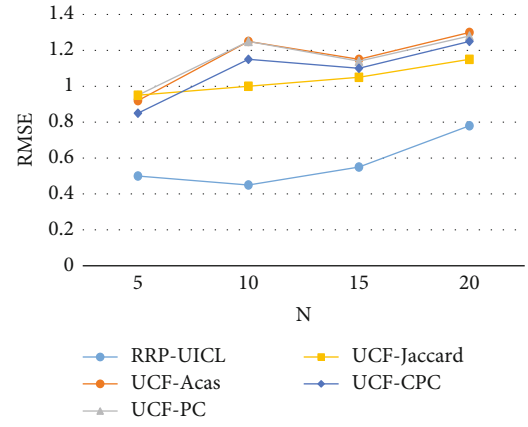


Figure 13: Data set-2 comparison of RMSE values of different methods.

Table 4: Data set-1 comparison of RMSE values of different methods before and after noise correction.

Method	N	RRP-UICL	UCF-ACos	UCF-PC	UCF-CPC	UCF-jaccard
Noise data set-1	5	0.5496	0.7835	0.7156	0.9378	0.9156
	10	0.7689	0.9845	0.9478	1.0856	1.0707
	15	0.6914	0.9989	1.0002	1.0818	1.1403
	20	0.8756	1.4956	1.4956	1.2945	1.2315
CFCNN-CL correction Noise data set-1	5	0.5471	0.7756	0.7056	0.9316	0.9118
	10	0.7051	0.9646	0.9289	1.0256	1.0239
	15	0.6845	0.9695	0.9565	1.0789	1.0956
	20	0.8535	1.2813	1.4209	1.2656	1.1489

Table 5: Data set-2 comparison of RMSE values of different methods before and after noise correction.

Method	N	RRP-UICL	UCF-ACos	UCF-PC	UCF-CPC	UCF-jaccard
Noise data set-2	5	0.4530	0.9456	1.0045	0.8812	0.9817
	10	0.4756	1.3617	1.3727	1.2233	1.0589
	15	0.5515	1.2902	1.2986	1.1945	1.0666
	20	0.8956	1.4795	1.4789	1.4569	1.3303
CFCNN-CL correction noise data set-2	5	0.4389	0.8964	1.0012	0.8759	0.9002
	10	0.4739	1.1807	1.2554	1.1854	1.0424
	15	0.4995	1.1088	1.1422	1.0867	0.9653
	20	0.8035	1.3256	1.2597	1.4068	1.3197

score value of the target user for the unscored items can be calculated, and the score value can be predicted by using Equation (13).

$$\mathrm{R_{u,i}} = \overline{r_u} + \frac{\sum_{v \in N_u} sim(u, v)(r_{v,i} - \overline{r_v})}{\sum_{v \in N_u} sim(u, v)}. \tag{13}$$

It can be seen from the expression of the predicted score value that the predicted score value mainly depends on the neighbor user set N_{u} and the similarity $\mathrm{sim}\,(u, v)$ of the target user u. The small neighbor user set and the large similarity error will affect the prediction of the score value and lead to the decrease of the recommendation accuracy.

For example, Table 1 is a simple user-item scoring matrix. It can be seen from the table that in the whole scoring matrix, only users U_2 and U_3 have a common scoring item I_2, and there are no common scoring items of U_1 and U_2, U_1 and U_3. At this time, when the similarity is calculated using the correlation-based method, since there is no common scoring item of U_1 and U_2, U_1 and U_3, the similarity of U_1 and U_2, U_1 and U_3 cannot be calculated. On the other hand, except when recommending item I_2 to user U_1, the target user has two neighbor users U_2 and U_3, and in other cases, there is only one neighbor user, and the collection of neighbor users is very small, which will affect the prediction of score value and lead to the degradation of recommendation quality of recommendation system.

3.2. Overview of Algorithm Model. In this paper, the data structure of user interest concept lattice is introduced, and a recommendation score prediction algorithm based on user interest concept lattice is proposed. The main steps of RRP-UICL algorithm proposed are shown in Figure 1.

As can be seen from Figure 1, the proposed method includes two main stages: one is the "nearest neighbor" module, and the other is the recommendation module.

In the recommendation module, when recommending to the target user, the similarity between the target user and the "nearest neighbor" should be calculated first. In this paper, different methods are used to calculate the similarity between the target user and the direct "nearest neighbor" and the indirect "nearest neighbor", and then, based on the similarity, the weighted average method is used to predict the scoring value of the target user for the unscored items. The algorithm will be described in detail below.

3.3. Constructing User Interest Concept Lattice. The binary matrix must be represented by a list of items of interest to each user, and in the rating matrix, the value of the item with the higher rating is set to <1>, and the values of all other items are set to <0>. In the reference scale, the items of score 4 and 5 are the items that users are interested in, and their values are set as <L>, while the values of other items are set as <0>.

After the scoring matrix is converted into a binary matrix, if a user marks an item as L, it means that the user has the attribute of an item. The binary matrix can be regarded as a user interest formal background $K = (U, I, R)$, where U is the set of all users, which is equivalent to the set of objects, I is the set of all items, which can be regarded as the attribute set, and R is a relationship between U and 1. After obtaining the formal background of user interest, the concept lattice structure model is established according to

the binary relationship between objects and attributes (users and items) in the formal background of user interest K, and the concept lattice of K is represented by L. After constructing the concept lattice, the recommendation algorithm can be analyzed based on the concept lattice theory, and the concept lattice theory can be applied to the recommendation algorithm.

Table 2 shows the scoring matrix of 6 users for 7 items, in this scoring matrix, according to the conversion principle of binary matrix, the items with scores of 4 and 5 are defined as items of interest to users, their values are set to <1 >, and other values are set to <0 >. The results are shown in Table 3, and the binary matrix can be regarded as a background of user interest form $K = (U, I, R)$, where user set $U = \{U_1, U_2, U_3, U_4, U_5, U_6\}$ and item set $I = \{I_1, I_2, I_3, I_4, I_5, I_6, I_7\}$. The user interest concept lattice constructed based on the user interest formal background in Table 3 is shown in Figure 2.

Because the traversal time is very complex, it needs to speed up the recommendation. It is necessary to delete some redundant L of user interest. This paper defines two conditions to delete formal concepts:

For a formal concept Z in the user interest concept lattice L_k,

(1) Z can be deleted if $\exists Z \in L_K$ is such that $|Ext(Z)| = 1$

(2) Z can be deleted if $\forall Z \in L_K$ is so that $Int(Z) \in Int(Z')$

According to the deletion condition of redundant formal concepts, the user interest concept lattice in Figure 2 is deleted, and the obtained end user interest concept lattice L_k is shown in Figure 3.

3.4. Partitioning "Nearest Neighbor." In this stage, the existing methods are mainly used to find the most similar users, and the "nearest neighbors" are divided by the most similar users.

The immediate "nearest neighbor" N_u^d of the target user u is represented as follows:

$$N_u^d = \{x | x \in N_u \text{ 且 } x \in MN_u\}. \quad (14)$$

Similarly, the indirect "nearest neighbor" N_u^{id} of the target user u is represented as follows:

$$N_u^{id} = \{x | x \in N_u \text{ 且 } x \notin N_u \cap MN_u\}. \quad (15)$$

Among the other "nearest neighbors" users, these users are just similar to the target users but do not show great interest in the recommended items. This paper classifies these users as indirect "nearest neighbors". For example, if a "nearest neighbor" Nu of a target user u is $\{U_1, U_2, U_3, U_4, U_5, U_6\}$, and the most similar user MNu is $\{U_3, U_4, U_6\}$. According to the above definition, the direct "nearest neighbor" N_u^d is $\{U_2, U_4, U_6\}$, and the indirect "nearest neighbor" N_u^{id} is $\{U_1, U_5\}$.

3.5. User Interest Forecast. According to the obtained direct "nearest neighbor" and indirect "nearest neighbor" of the target user, items can be recommended to the target user. There are two main methods of project recommendation: prediction method and list method. In the prediction method, in the list method, all items of interest to the "nearest neighbor" user are recommended to the target user.

3.5.1. Calculation of Correlation Coefficient between Users. For indirect "nearest neighbor" users, this paper uses Equation (16) to calculate the similarity between indirect "nearest neighbor" users and target users, and the similarity calculation formula between user U and user V is defined [16]:

$$\text{sim}(u, v) = \frac{\max(1, |I_u \cap I_v|) \sum_{i \in I_u} \sum_{j \in I_v} (r_{u,i}/r_{v,j})}{|I_u| \cdot |I_v| \cdot |I_u \cup I_v|}. \quad (16)$$

Weighted average forecast unscored items.

The prediction of score value is the last important step in the recommendation algorithm. Use the weight to obtain the final prediction score of each item. The steps of the prediction method are as follows:

First, you need to calculate the average score of recommended items. For recommended item I, use Equation (17) to calculate the average score.

$$\overline{r_i} = \frac{\sum_{v \in N_u^d \cup} N_u^{id} r_{v,i}}{\left| N_u^d \bigcup N_u^{id} \right|}. \quad (17)$$

In the scoring matrix of Figure 2, it is necessary to recommend the item I_4 to the target user U_3. It can be obtained that the "nearest neighbor" Nu of the target user U_3 is $\{U_1, U_2, U_3, U_4, U_5, U_6\}$, the most similar user MNu is $\{U_2, U_4, U_6\}$, the direct "nearest neighbor" N_u^d is $\{U_2, U_4, U_6\}$, and the indirect "nearest neighbor" N_u^{id} is $\{U_1, U_5\}$. According to the scores of the direct "nearest neighbor" and the indirect "nearest neighbor", the average score of the recommended item I_4 is calculated as $\overline{r_i} = (5 + 5 + 4 + 5 + 5)/5 = 4.8$.

Then, the score of the recommended item is predicted, and the score R_u, i of the target user u for the item i is predicted using formula (18)

$$R_{u,i} = \frac{\sum_{v \in N_u^d} r_{v,i} \left(a - \left|r_{v,i} - \overline{r_i}\right| - 1\right)^2 + \sum_{v \in N_u^{id}} r_{v,i} \text{sim}(u, v) \left(a - \left|r_{v,i} - \overline{r_i}\right| - 1\right)^2}{\sum_{v \in N_u^d} \left(a - \left|r_{v,i} - \overline{r_i}\right| - 1\right)^2 + \sum_{v \in N_u^{id}} \text{sim}(u, v) \left(a - \left|r_{v,i} - \overline{r_i}\right| - 1\right)^2}. \quad (18)$$

4. Experimental Design and Result Analysis

4.1. Experimental Design. In the experiment part, firstly, we validate the effectiveness of CFCNN-CL algorithm to solve the natural noise in the recommendation system, and then, we validate the effectiveness of RRP-UICL algorithm to solve the data sparse problem in the recommendation system. Finally, we combine CFCNN-CL algorithm and RRP-UICL algorithm to recommend, and validate the effectiveness through experiments. Three parts of the experiment are

using the recommendation algorithm to evaluate the average absolute error and mean root error for comparative analysis.

4.2. Experimental Data Set. In this paper, the data set MOVIELENS 100 K is used to verify the effectiveness of the algorithm. MOVIELENS data set is one of the most commonly used data sets to evaluate the effectiveness of the recommendation algorithm. A score of 4 means that the user likes the movie, and a score of 5 means that the user likes the movie very much. In the whole data set, each user evaluates at least 20 scores after watching the movie.

4.3. Performance Evaluation Indicators. Average absolute error and mean root error are used to evaluate the accuracy of our method. Under normal circumstances, the smaller the MAE, the higher the prediction accuracy, and the calculation formula is

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |r_i - p_i|}{n}. \tag{19}$$

The RMSE is calculated by dividing the sum of squares of the difference between the actual score value and the predicted score value by the score set in the test set. The calculation formula is

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (r_i - p_i)^2}{n}}. \tag{20}$$

4.4. Method Analysis

4.4.1. Performance Analysis of CFCNN-CL Algorithm. According to the algorithm, the users and items of the sample data set are classified, as shown in Figure 4.

See Figure 5 for the number of natural noise correction in sample data set by collaborative filtering method based on concept lattice.

For verifying the effectiveness of CFCNN-CL algorithm, CFCNN-CL algorithm and existing PCC reprediction methods are tested on data sets, and the effects of nearest neighbor and sparsity on MAE and RMSE values are compared.

(1) The Influence of Nearest Neighbor. With the change of the nearest neighbor number, the changes of MAE and RMSE values in the corresponding two methods are shown in Figures 6 and 7, respectively. By analyzing Figures 6 and 7, it can be concluded that the MAE and RMSE of CFCNN-CL algorithm and PCC reprediction method decrease with the increase of nearest neighbors, and the increase of nearest neighbors improves the accuracy of the two methods. However, the MAE and RMSE values of the proposed method are lower than those of PCC reprediction method, so the noise correction method proposed in this paper has better performance and better prediction accuracy than PCC reprediction method.

(2) The Effect of Sparsity. For verifying the effectiveness in sparse scenarios, the available ratings in sample data sets are randomly changed to zero to form five data sets with different sparseness, which are 92%, 95%, 97%, 98%, and 99%, respectively. Then, the CFCNN-CL algorithm and the existing PCC reprediction method are tested on five data sets with different sparseness. Finally, the MAE and RMSE values of different methods are compared, respectively.

Figures 8 and 9 show the experimental results of MAE and RMSE value changes in different sparse scenarios with the proposed method and PCC reprediction method. The natural noise correction method proposed in this paper is superior to PCC reprediction method.

4.4.2. Performance Analysis of RRP-UICL Algorithm. When the recommended items $N = 5$, 10, 15, and 20, for data set-1 and data set-2 with different sparsity, the experimental results of MAE of the five methods varying with the recommended items are shown in Figures 10 and 11.

As can be seen from Figures 10 and 11, of the five methods, the MAE values of the RRP-UICL method in both data sets are smaller than those of the other four methods. It can be seen from the analysis chart that in sparse scenes, this method has better prediction accuracy than the commonly used collaborative filtering methods.

Similarly, under different recommended items $N = 5$, 10, 15 and 20, the experimental results of RMSE values of five methods in data set-l and data set-2. RRP-UICL method has better prediction accuracy than commonly used collaborative filtering methods in Figures 12 and 13.

4.4.3. Recommended Performance Analysis of Fusion CFCNN-CL and RRP-UICL. In this section, CFCNN-CL algorithm to solve natural noise and RRP-UICL algorithm to solve data sparsity are recommended, and data set-1 and data set-2 with different sparsity in the previous section are used for experimental analysis. At the beginning of the experiment, CFCNN-CL algorithm is used to correct the natural noise in data set-1 and data set-2. Then, the recommended score prediction algorithm RRP-UICL and the four commonly used methods UCF-ACos, UCF-PC, UCF-CPC, and UCF-Jaccard are tested on data set-1 and data set-2 with corrected natural noise, respectively. Finally, the experimental results are compared with those on data set-1 and data set-2 without corrected natural noise, and the results are analyzed.

When the recommended item $N = 5$, 10, 15, and 20, the experimental results of RMSE values of the five methods varying with the recommended items are shown in Table 4 for the data set-1 without correcting the natural noise and the data set-1 with correcting the natural noise, and the experimental results of RMSE values of the five methods varying with the recommended items are shown in Table 5 for the data set-2 without correcting the natural noise and the data set-2 with correcting the natural noise.

The recommendation combining CFCNN-CL and RRP-UICL also has the smallest RMSE value and the highest recommendation accuracy in data set-1 and data set-2 in Table 4 and Table 5. For the comparison before and after noise correction, the RMSE value of the five methods after

noise correction is lower than that before noise correction, and the performance has been improved accordingly. However, UCF-ACos, UCF-PC, UCF-CPC, and UCF-Jaccard are all affected by data sparsity to varying degrees, so the recommendation performance will decrease in sparse data, and with the increase of sparsity, the recommendation performance will become worse. The recommendation method based on CFCNN-CL and RRP-UICL avoids the influence of natural noise and data sparsity, and the recommendation accuracy is kept in good condition.

5. Conclusion

With the continuous development of recommendation system and the increasing demand of people, the performance and accuracy of recommendation algorithm are required to be higher and higher. Firstly, this paper analyzes the development status of recommendation system and concept lattice and explains the research background and significance of this paper. After that, the related theories of recommendation system and concept lattice are introduced, which provides theoretical support for the following methods.

The existing recommendation algorithms cannot recommend accurately due to the influence of sparse data, this paper proposes a recommendation rating prediction algorithm based on user interest concept lattice, considering the different influence degree of the "nearest neighbor" users of the target users in the rating prediction process. The prediction method proposed in this paper not only solves the problem of sparse data in recommendation system but also has high performance and prediction accuracy.

In the experimental part, the experimental settings are introduced firstly, and then, the effectiveness of CFCNN-CL algorithm and RRP-UICL algorithm and the fusion of CFCNN-CL and RRP-UICL recommendation are verified by using sample data sets. In the experimental results of CFCNN-CL algorithm, under the influence of nearest neighbor or sparsity, the noise correction method proposed in this paper has better performance and better prediction accuracy than PCC reprediction method. In the experimental results of RRP-UICL algorithm, in sparse scenarios, the proposed method has better performance and prediction accuracy than four commonly used methods: modified cosine similarity measure, Pearson's correlation measure, constrained Pearson's correlation measure, and Jaccard measure. In the final experimental results, under the influence of the number and sparsity of recommended items, the MAE value and RMSE value of CFCNN-CL and RRP-UICL recommendation method are the smallest, and the recommendation accuracy is the highest.

Data Availability

The experimental data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The author declares no conflicts of interest regarding this work.

References

[1] Y. Chen, "Mining of instant messaging data in the Internet of Things based on support vector machine," *Computer Communications*, vol. 154, pp. 278–287, 2020.

[2] A. Suresh and M. J. Carmel Mary Belinda, "Online product recommendation system using gated recurrent unit with Broyden Fletcher Goldfarb Shanno algorithm," *Evolutionary Intelligence*, 2021.

[3] Z. Ali, I. Ullah, A. Khan, A. Ullah Jan, and K. Muhammad, "An overview and evaluation of citation recommendation models," *Scientometrics*, vol. 126, no. 5, pp. 4083–4119, 2021.

[4] F. He and P. Wei, "Research on comprehensive point of interest (POI) recommendation based on spark," *Cluster Computing*, vol. 22, no. S4, pp. 9049–9057, 2019.

[5] A. Da'U, N. Salim, and R. Idris, "An adaptive deep learning method for item recommendation system," *Knowledge-Based Systems*, vol. 213, no. 8, article 106681, 2021.

[6] R. Gerbaudo, R. Gaspar, and R. G. Lins, "Novel online video model for learning information technology based on micro learning and multimedia micro content," *Education and Information Technologies*, vol. 26, no. 5, pp. 5637–5665, 2021.

[7] X. Han, Z. Wang, and H. J. Xu, "Time-weighted collaborative filtering algorithm based on improved mini batch K-means clustering," *Advances in Science and Technology*, vol. 105, pp. 309–317, 2021.

[8] Y. Gao and L. Ran, "Collaborative filtering recommendation algorithm for heterogeneous data mining in the Internet of Things," *IEEE Access*, vol. 7, pp. 123583–123591, 2019.

[9] Y. Wu, S. Zhao, and R. Guo, "A novel community answer matching approach based on phrase fusion heterogeneous information network," *Information Processing & Management*, vol. 58, no. 1, article 102408, 2021.

[10] B. Kaya, "A hotel recommendation system based on customer location: a link prediction approach," *Multimedia Tools and Applications*, vol. 79, no. 3-4, pp. 1745–1758, 2020.

[11] H. Zhang and F. Ye, *A personalized recommendation algorithm for user-preference similarity through the semantic analysis*, vol. 2, Springer International Publishing, 2016.

[12] X. Ye and D. Liu, "An interpretable sequential three-way recommendation based on collaborative topic regression," *Expert Systems with Applications*, vol. 168, no. 2, article 114454, 2021.

[13] Y. T. Song and S. Wu, "Slope one recommendation algorithm based on user clustering and scoring preferences," *Procedia Computer Science*, vol. 166, pp. 539–545, 2020.

[14] Y. Chen, "Research on personalized recommendation algorithm based on user preference in mobile e-commerce," *Information Systems and e-Business Management*, vol. 18, no. 4, pp. 837–850, 2020.

[15] Q. Zhou, F. Liao, C. Chen, and L. Ge, "Job recommendation algorithm for graduates based on personalized preference," *CCF Transactions on Pervasive Computing and Interaction*, vol. 1, no. 4, pp. 260–274, 2019.

[16] C. Zou, D. Zhang, J. Wan, M. M. Hassan, and J. Lloret, "Using concept lattice for personalized recommendation system design," *IEEE Systems Journal*, vol. 11, no. 1, pp. 305–314, 2017.

Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 9356725, 10 pages
https://doi.org/10.1155/2022/9356725

Research Article

The Structural Features and Translation Skills of English in the Era of Radio Communication Networks

Tiantian Wu

School of Foreign Languages, Xinyang Agriculture and Forestry University, Xinyang, 464000 Henan, China

Correspondence should be addressed to Tiantian Wu; 2012280001@xyafu.edu.cn

Received 13 February 2022; Revised 8 March 2022; Accepted 23 March 2022; Published 20 April 2022

Academic Editor: Lisheng Fan

A wireless communication network using an embedded microprocessor is a communication network method that uses radio waves to transmit the sound, text, pictures, data, and other information that the sender needs to transmit to the receiver through space and ground. With the rapid development of world science and technology, the application of radio communication network technology and international exchanges and cooperation have become increasingly active. In the era of radio communication networks, through radio communication English and radio communication English translation, we can receive real-time information about the development of radio communication technology. It can be seen that radio communication English plays an important role in promoting international radio technology exchanges and cooperation. Therefore, more and more researches on the structural features and translation skills of radio communication English have appeared in the academic circles. This article is aimed at studying the structural features and translation skills of English in the era of radio communication networks. The article first briefly introduces radio communication and then introduces the structural features of radio communication English, including lexical features and syntactic features. It also introduced two common radio communication spectrum detection algorithms. Finally, it explores the translation skills of radio communication English based on case analysis and provides some method reference for radio communication English translation.

1. Introduction

Radio communication is an advanced communication science and technology born with the development of social science and technology. It uses humans' extensive use of radio wave transmission to transmit information in the air. Transmitting the sound, text, image, electronic data, and other information that the transmitter needs to transmit to the receiver through radio wave debugging and help the sender and receiver to exchange and transmit information as required. Compared with traditional wired communication methods, radio communication has advantages such as no need to set up transmission lines, long communication distance, and good mobility [1]. But at the same time, it has disadvantages such as susceptibility to information dissemination and the instability of information transmission quality due to the influence of natural causes. However, radio communication has become the main contemporary space communication method relying on its advantages of fast information transmission, convenient and fast communication, and better information interactive transmission performance [2]. With the development of today's society, the use of radio communication technology, communication, and collaboration has become more and more active [3]. It is hoped that people can learn more about the research results of radio communication technology in China in real time and keep up with the development of radio communication technology. The translation study of radio communication English has become an important topic in the field of translation. Radio communication English combines technical professional English and ordinary English and at the same time possesses strong professionalism and practicality. In

addition, radio communication English is widely used, related terms are updated quickly, and the standard is getting stronger and stronger, which puts higher requirements on communication English translators. In summary, the research on radio communication English translation skills is very meaningful. With the continuous development of scientific research at home and abroad, many studies on radio communication English translation techniques have emerged.

Among them, Wang analyzed the structural features and translation skills of radio communication English through his own research and put forward his personal opinions on the translation of communication English [4]. Qiang took communication professional literature as an example to analyze the word formation characteristics of compound words in communication English. And through example words, it specifically elaborates several translation skills of communication English compound words, including the positive order method, the adjustment method, and the augmented translation method [5]. In Wenting's research on the characteristics of communication English terminology and translation strategies, he first analyzed the characteristics of communication technology English terminology and at the same time discussed the translation strategies of communication English with examples. In order to help translators to hold a clear translation strategy when carrying out translation activities, accurately convey information, and play a communicative role [6], Dan and Yong mainly studied the passive voice translation in radio communication English. They combined a large number of examples to explore the passive voice translation skills by analyzing the examples in professional English in the communication field [7]. Yutao et al. jointly studied the characteristics and laws of English abbreviations for radio communication. They started from five aspects: basic characteristics, classification, abbreviation, choice of translation, and spelling, and conducted a more comprehensive analysis and research on the abbreviations of communication English. The characteristics and laws of communication English abbreviations are sorted out, and a preliminary foundation is laid for the exploration of translation skills [8]. Tong has studied the application of the Hypotaxis Parataxis Theory in the translation of communication English. In his research, he combined many specific translation practices to show how the Hypotaxis Parataxis Theory can play a guiding role in the translation of communication English [9]. Tingting specifically studied the translation methods of nonpredicate verbs in communication English texts and explored the translation strategies of nonpredicate verbs in communication English based on translation examples [10].

The above studies are closely related to the translation skills of communication English, and the research is more specific, which can be used as a reference for the follow-up research on the translation skills of radio communication English. The innovations of this article are as follows: (1) On the basis of previous studies, the research on the translation skills of radio communication English has carried out content and method innovations. (2) This article introduces the structural characteristics of radio communication and representative spectrum detection algorithms and combines relevant translation examples to study its translation skills.

2. Radio Communication

2.1. Introduction to Radio. Simply put, radio communication is an advanced communication technology that uses radio waves to achieve spatial information transmission [11]. The information that can be transmitted includes audio, text, data, and pictures, and the transmission of information has the characteristics of real time and rapidity. In 1887, German physicist Hertz [12] accidentally discovered electromagnetic waves in one of his experiments. The establishment and improvement of electromagnetic theory laid a theoretical foundation for the generation of radio communication [13]. Finally, in 1895, Russian physicist A. C. Popov [14] and Italian physicist G. Marconi [15] successfully carried out radio communication experiments, and radio communication technology was born. The biggest advantage of radio communication lies in its function of transmitting information by means of the fluctuation of radio waves, which eliminates the problem of laying wires and helps people achieve faster, more convenient, and barrier-free information exchange and communication. The wavelengths used in radio communication can be roughly divided into 4 bands, as shown in Table 1.

A simple display of radio communication is shown in Figure 1.

After the birth of radio communication, it brought great changes to people's communication life. Radio communication technology makes people's communication methods and channels more flexible and convenient. With radio communication technology, people can conduct cross-space interactive communication anytime, anywhere. After more than a hundred years of development, radio communication has been applied in more and more various industries: for example, satellite mobile, space operation, radio navigation, radio determination, and other industries.

Radio can be divided into four categories according to the range of wireless connection; they are wireless personal area network (WPAN), wireless local area network (WLAN), wireless metropolitan area network (WMAN), and wireless wide area network (WWAN). The wireless personal area network refers to the form of wireless local area network formed in the air with high privacy, and it refers to the form of short-distance wireless local area network within a radius of 100 meters. A wireless metropolitan area network is a form of wireless network that connects multiple wireless local area networks, and the connection distance is generally within several kilometers. A wireless wide area network is a wireless communication technology that uses wireless network technology to connect scattered local area networks [16]. The transmission distance is generally within a radius of 15 kilometers.

The classification of radio communication systems is shown in Figure 2.

2.2. Radio Pulse Signal Design. The radio pulse signal system transmits information by sending a series of narrow pulses.

Table 1: Radio communication band table.

Segment number	Band name	Frequency range	Band name	Wavelength range
1	Low frequency	30-300 Hz	Long wave	10-100 km
2	Middle frequency	300-3000 Hz	Middle wave	10-100 m
3	High frequency	3-30Z Hz	Short wave	100-10 m
4	Super high frequency	30-300 Hz	Super short wave	10-1 dm

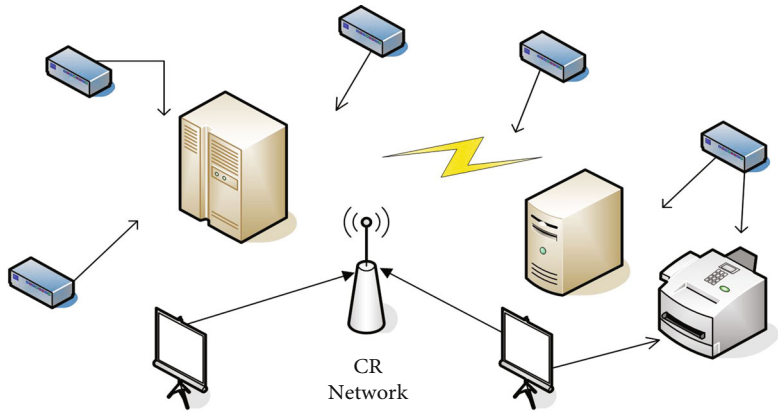


Figure 1: Radio communication.

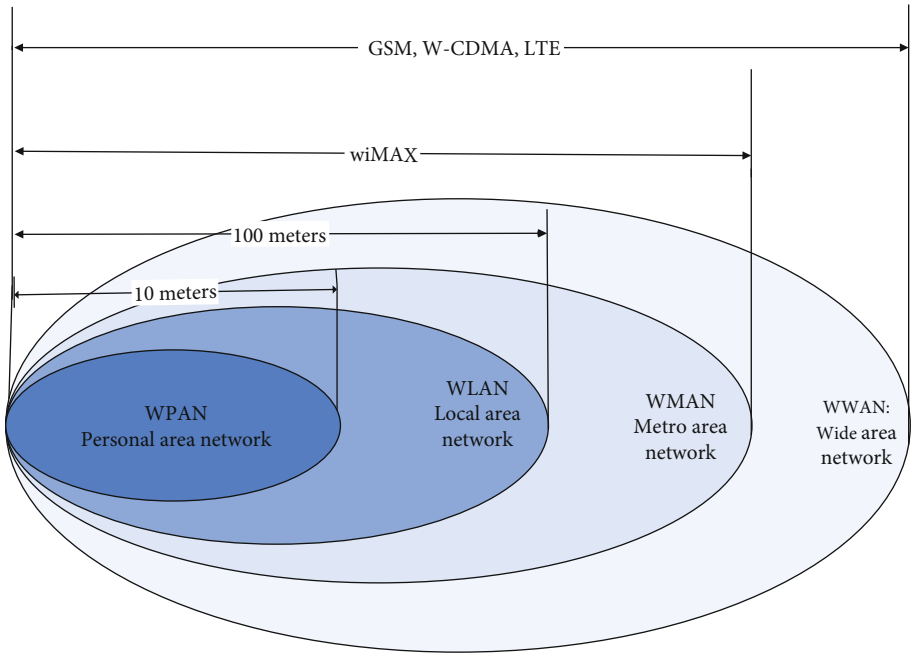


Figure 2: Classification of radio systems.

The simplest and most common one is a single-period pulse signal, such as a Gaussian pulse. Because the antenna attenuates and deforms the pulse more severely than other narrow-band system signals, many studies use Gaussian functions to analyze the system. Gaussian waveforms are named because their mathematical definition is similar to Gaussian functions. A general Gaussian pulse can be expressed as

$$p_t(b) = \frac{1}{\sqrt{3\pi\sigma}} \exp\left[-\frac{1}{2}\left(\frac{b-v}{\sigma}\right)\right]. \quad (1)$$

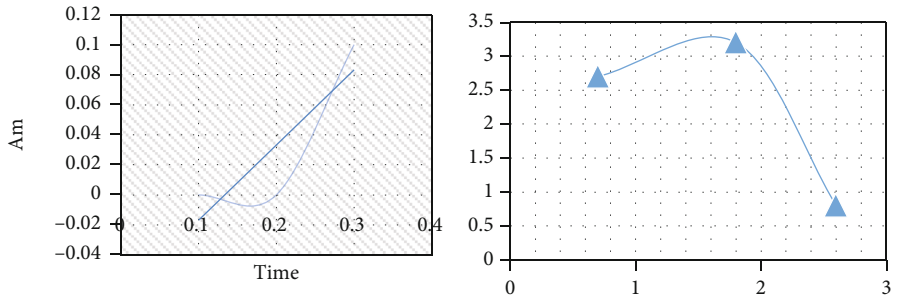


Figure 3: Gaussian waveform and its power spectral density.

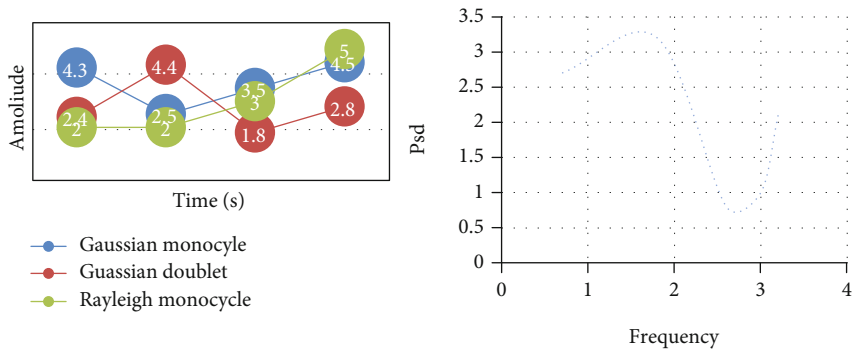


Figure 4: Gaussian pulse and its power spectral density.

Among them, μ is the pulse center, and σ determines the intensity of the pulse signal. The Gaussian waveform and its power spectral density are shown in Figure 3.

Some single-cycle pulses have evolved from Gaussian pulses. The Gaussian single-period pulse is the second derivative of the Gaussian pulse:

$$p(t) = C\left[1 - \frac{t-u}{\sigma}\right] \exp\left[-\frac{1}{4}\left(\frac{u}{\sigma}\right)\right]. \quad (2)$$

Among them, σ determines the single-cycle pulse width t, but the effective pulse duration $T = 7\sigma$, and the pulse waveform contains 99% of the total cycle pulse energy. B_g is the introduction of energy normalization. The bipolar Gaussian pulse is another improved pulse waveform of the Gaussian pulse, and it is also a bipolar signal. It contains two Gaussian pulses with opposite amplitudes separated by a time gap Tw [17]. The formula for a single-cycle bipolar Gaussian pulse is

$$B_g(t) = C_g \exp\left[-\frac{1}{2}(u)\right]. \quad (3)$$

Among them, the pulse width is determined by parameter μ. When $\mu = 14\sigma$ and $T = 7\sigma$, the pulse contains 99% of the total single-cycle pulse energy.

The Rayleigh single-cycle pulse waveform is derived from the first derivative of the Gaussian pulse, and its mathematical expression is

$$p_t = A\{\exp(2t - \mu) - \exp\sigma\}. \quad (4)$$

Like the Gauss single-cycle pulse, when the effective pulse time is $t = 7\sigma$ and the pulse center is at $\mu = 2\sigma$, the pulse waveform contains 99% of the total single-cycle pulse energy. Among these pulse waveforms, the Gaussian pulse waveform has the same DC component as the rectangular pulse waveform, but other single-cycle pulse waveforms do not have the same DC component, so that the radio signal is transmitted more effectively [18]. The Gaussian pulse and its power spectrum are shown in Figure 4.

2.3. Development Status and Trends of Radio Communication Technology. At this stage, with its advantages, radio communication technology is more and more widely used worldwide. From the establishment of the magnetic field theory to the birth of radio communication technology, it has been a long time. Radio communication technology has now become an important part of people's lives and has played an irreplaceable role in all aspects of people's social life; for example, it has played a role in real-time monitoring and feedback of weather changes, space station monitoring, and production technical guidance. With the continuous development of information technology, in the future, radio communication technology will inevitably enter one new technological stage after another and obtain new technological development and improvement [19]. But at the same time, the current radio communication technology also has some undeniable problems; that is, there are still certain technical defects, which leads to insufficient communication stability and signals susceptible to interference [20]. Therefore, the current development trend of radio communication technology is to continuously improve technical defects and improve communication stability to adapt to

the development of the times and meet people's requirements for communication quality. Since there is still a broad space for development of radio communication technology, it is of certain significance to improve the technical defects and carry out continuous development and promotion of radio technology [21]. To ensure that the radio communication technology develops towards a positive trend, it is necessary to ensure the technological innovation and improvement of radio communication, which requires the following points. First of all, we must formulate corresponding policies and measures to think of ways to improve the resource efficiency of the radio electronic metrology spectrum, so as to ensure the stability of information, not only to prevent interference and affect the quality of communication but also to ensure the safety of users. Second, we must strive to achieve broadbandization of radio electronic metrology information, because broadbandization of information is a key measure to improve information transmission rate and communication quality. With the popularization and advancement of information broadband on a global scale, radio communication technology is also moving towards broadband development of wireless access. Therefore, striving to achieve and promote broadband communication technology has great positive significance for improving signal strength and ensuring communication quality. Third, while actively promoting the broadbandization of radio communication technology, we must also try to introduce personal information technology, so as to reduce the limitation of information transmission time and increase the information transmission rate. Finally, for the radio communication technology itself, effective management must be taken to ensure its standardized, safe, and effective development. Therefore, this also means strengthening radio control and ensuring the legalization, standardization, and scientific operation of radio management. In addition, we must earnestly do a good job in the construction of the radio monitoring system, improve the utilization rate of spectrum resources, and ensure the normal and effective operation of radio communication services [22]. All in all, the radio communication technology is developing well at this stage. Although there are still some technical problems, the general trend is steadily moving forward. If the technical defects can be resolved as soon as possible, radio communication technology will inevitably be further developed faster and enter the next stage of development. Of course, all these are naturally inseparable from the efforts of all technicians and researchers.

3. Embedded Radio Communication Spectrum Detection Algorithm and Structural Features of Radio Communication English

3.1. Embedded Radio Spectrum Detection Algorithm. Common embedded radio communication spectrum detection algorithms are shown in Table 2.

This article mainly briefly introduces two spectrum detection methods: energy detection method and cycle detection method.

3.1.1. Energy Detection Algorithm. The energy detection algorithm is one of the most commonly used radio spectrum detection algorithms. Based on the existing useful signal plus the energy of the noise signal, it calculates the energy greater than the energy of the noise signal alone [23]. It can be expressed by

$$E\{(s(t)+n(t)\} = E\{pt(n)\} > E[nt]. \tag{5}$$

Among them, $E\{(s(t)+n(t)\}$ represents the energy of the existing signal plus the noise signal, and 2 represents the final total energy obtained by the energy detection algorithm, and $E[nt]$ represents the energy when the noise signal exists alone. The principle is to first input the sampled digital signal into the square algorithm module to obtain the signal energy and then use the comprehensive calculation module to average the signal energy and record it as energy statistics, namely, T. Finally, compare it with the preset decision upper limit L, and draw the final verdict. Comparing the energy statistics T with the upper decision limit L, it can be expressed by

$$T < L, H_o, \tag{6}$$

$$T \geq L, H_1. \tag{7}$$

Among them, L is the fixed decision upper limit of the spectrum state at a certain moment. If at a certain moment, the main user has no signal input; only noise signals can be received. Through the judgment, it can be known that the spectrum is free at this moment and can be allocated to secondary users.

Assuming that the average power of the signal X transmitted by the primary user is S, the mean value of the noise signal is 0, and the variance i is Gaussian white noise with S. That is, $X(n) = N(0, x)$, and the transmitted signal and noise signal exist independently. Therefore, when the sample size is N, the variance of the energy statistic T can be expressed by

$$E(T_{\text{ed}}) = Nx^2, H_0, \tag{8}$$

$$E = N(1+\gamma)x, H_1, \tag{9}$$

$$\text{Var}(T) = (2Nx, H_o), \tag{10}$$

$$\text{Var} = 2N(1+\gamma), X, H_1. \tag{11}$$

Among them is the signal-to-noise ratio. From the knowledge of probability theory and statistics, if there are N independent random variables that obey a normal distribution, the sum of the squares of the random variables obeys the chi-square distribution with N degrees of freedom. And when the mean value of the random variables is nonzero, the random variable formed by their sum of squares obeys the noncentral chi-square distribution with N degrees of freedom, as shown in

$$T \sim X_m^2, H_1. \tag{12}$$

Table 2: Common spectrum detection methods.

Spectrum detection	
Single cognitive user detection	
Matched filter detection	Centralized detection method
Energy detection method	Distributed detection method
Loop detection method	Hybrid detection method

According to the study of polynomial complexity and non-determinism, first of all, a false alarm probability P needs to be given. Secondly, the judgment valve is calculated according to the given P. Then, the detection probability K can be calculated based on calculation. According to the false alarm probability P, when $P = a$, that is,

$$P = \left\{ \frac{1}{\sqrt{2N}} - \frac{1}{2N\alpha} \exp\left(-\frac{1}{2}k\right) dT \right\}. \tag{13}$$

The detection performance of the energy detection algorithm is shown in Figure 5.

It can be seen from Figure 5 that when the signal-to-noise ratio remains the same, the detection probability increases as the false alarm probability increases. Energy detection is a kind of blind detection algorithm, mainly the incoherent detection of signals. The calculation of energy statistics is mainly based on the signal energy received in the frequency domain. In the time domain, the energy detection algorithm approximates the received signal energy through the accumulation of the modulo square of the signal amplitude in the sensing period. In the frequency domain, the energy detection algorithm accumulates approximately the received signal energy by sensing the power spectrum of the signal in the frequency band [24].

3.1.2. Loop Detection Algorithm. Generally speaking, in a communication system, due to the huge amount of modulation, sampling and coding of the signal, and the statistical characteristics of the signal have periodic changes in time, it can be regarded as a periodic steady signal in a macroscopic view. The cyclostationary feature detection algorithm is an algorithm that uses signal cycle stationarity to determine whether the main user exists, and through simple calculation steps, the signal judgment result can be obtained [25].

Analyzing the cyclostationary signal and the $F(t)$ characteristic mainly use two functions, the cyclic autocorrelation and the cyclic spectrum correlation. The definition formula of its function is as follows:

$$F(t) = \lim \frac{1}{T} \left\{ x\left(t + \frac{\pi}{2}\right) x \right\}, \tag{14}$$

where t is the cycle frequency and X represents the spectral component of the signal $X(t)$ at the center frequency of m. The principle diagram of the loop detection algorithm is shown in Figure 6.

Suppose the cyclic power spectral density of the signal $X(t)$ transmitted by the authorized user is f; the cyclic power spectral density of Gaussian white noise t is $S(f)$, and the cyclic power spectral density of the received signal y is the sum of the above two. In this case, the decision result of the signal passing the loop detection algorithm is

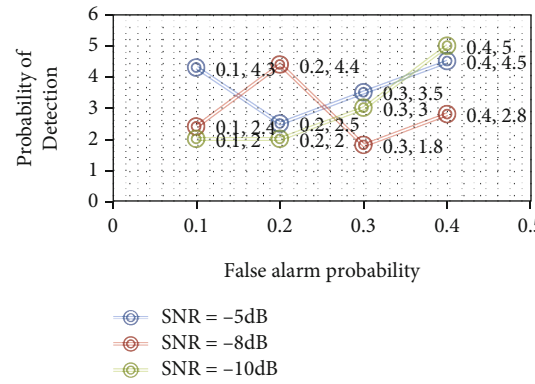


Figure 5: Energy algorithm detection performance.

$$S(f) = H(f)S_x^0 + S_n^0, H_1, \tag{15}$$

$$S^a_{r(f)=H(f+2a)H(f-a)H_1}. \tag{16}$$

Among them, $S(f)$ represents the decision result, $H(f)$ represents the Fourier change of the signal impulse response, and S is the original signal, and the judgment standard of the cycle detection can be obtained from the following formula:

$$S_r^a(f) > n, H_1, \tag{17}$$

$$s_r^a(f) < n, H_0. \tag{18}$$

Among them, n is the decision threshold, similar to the energy detection algorithm, which can derive the false alarm probability and detection probability of loop detection. The false alarm probability P and the detection probability Q are expressed by formula (17) and formula (18), respectively:

$$P = Q\left(\frac{\mu - N}{\sqrt{2N}}\right), \tag{19}$$

$$Q = P\left(\frac{\alpha - N(1+\gamma)}{\sqrt{2N}}\right). \tag{20}$$

Among them are the signal-to-noise ratio and the average power of the signal transmitted by the main user. The loop detection steps are shown in Figure 7.

It can be seen from Figure 7 that the loop detection step is not complicated, and sampling data storage and signal processing are two key steps. The loop detection performance is shown in Figure 8.

When the given false alarm probability is 0.2, when the signal-to-noise ratio of loop detection increases, using this algorithm makes the detection performance stronger. The greater the signal length, the better the detection performance.

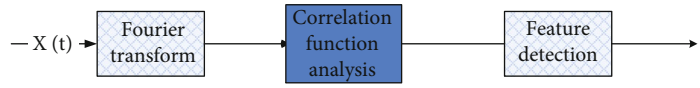


Figure 6: Schematic diagram of loop detection algorithm.

3.2. Structural Features of Radio Communication English

3.2.1. Vocabulary Features. Radio communication English belongs to a category of English for Science and Technology, and one of the notable lexical features is as follows: More professional terms and fixed terms are used, and the format is more rigorous and formal: for example, Software Defined Radio software radio and Digital signal Processor signal processor. The second lexical feature of radio communication English is the use of derivative words. The main derivation methods are prefix and suffix: for example, anti-, anti-missile (anti-missile), and anticatalyst (anti-catalyst) insulation. Derivative usage is mainly reflected in professional terminology. Another significant vocabulary feature of radio communication English is Acronym. Acronyms are very common in radio communication English. A phrase consisting of multiple words, usually only the first letter of each word, is intercepted: for example, SISO single input single output (Single Input Single Output); VHDL Very-High-SpeedIntegratedCircuit Hardware Description Language (Very-High-SpeedIntegratedCircuit Hardware Description Language); and ASIC (Application Specific Integrated Circuit). There is also a lexical feature of radio communication, that is, vocabulary synthesis. Vocabulary synthesis is to combine two or more words according to a certain order and rules to form a new word. It is one of the important methods for the generation and development of English terminology in the field of radio communication, and it is also the most commonly used word-building method in the field of scientific and technological English. In radio communication English, the composition method is mainly composed of compound words, then compound adjectives, with or without hyphens and with hyphens: such as benchmarking, payload, and serial-input-out-put.

In a word, radio communication English vocabulary has many professional terms, the format is fixed and formal, and more derived vocabulary, acronyms, and compound words are used [26].

3.2.2. Syntactic Features. The most notable syntactic feature in communication English is the use of passive sentences. The main reason is that the passive structure is more objective than the active structure; emphasizing objective facts is in line with the logical and rigorous characteristics of technical English. Secondly, under normal circumstances, passive sentences will be more concise than active sentences, making the content more eye-catching and beautiful to attract attention [27]. It widely uses complex and long sentences such as attributive clauses. As we all know, communication English is a language used to explain the content of the field of communication technology or describe its regular characteristics, and it faces a wide range of groups. According to Nida's functional equivalence theory, radio communication English translation must be combined with the content and functions of radio communication technology for translation. This requires correct expression, strict structure, and strong logic. Therefore, in order to meet the above characteristics, communication English often adopts complex and long sentences with multiple modifiers, multiple components, and multiple levels. Such a long and complex sentence with many structures is a very common sentence pattern in communication English. In addition to the form of attributive clauses, there are also various forms of complex long sentences in communication English. Such sentence pattern features cause people to be good at splitting and understanding structure and sentence meaning when reading or translating communication English [28].

All in all, in order to express more objectively and rigorously, especially when describing related important technical concepts, radio communication English usually has more passive sentences and long sentences in its syntactic structure. This puts forward high demands for radio communication English translators.

4. Discussion on Translation Skills of Radio Communication English

Combining the vocabulary and syntactic features of radio communication English introduced above, it has explored some communication English translation techniques based on Nida's functional equivalence translation theory.

4.1. Skills in Vocabulary. We all know that radio communication English has multiple technical terms, multiple derivatives, acronyms, and compound words. To translate every communication English vocabulary accurately and appropriately, it will inevitably be inseparable from a certain professional knowledge base. Because if there is no relevant professional knowledge base, the translator is easy to misinterpret certain professional vocabulary. For radio communication English vocabulary, based on a certain professional knowledge, translators also need to be good at using translation methods such as literal translation, free translation, transliteration, and retention of original abbreviations. In communication English, vocabularies such as compound nouns and derivative words are mostly used literal translation and free translation to ensure accurate interpretation. In addition, in radio communication English, most of the new foreign words and unit of measurement words brought about by technological development can be accurately and quickly translated using transliteration. Of course, in order to make the translation more concise, the unit of measurement may not be translated, such as 500 MHz~900 MHz, without affecting the understanding. Similarly, because some acronyms translated into words are too long, it will affect the coherence of the original text and the need for translation is

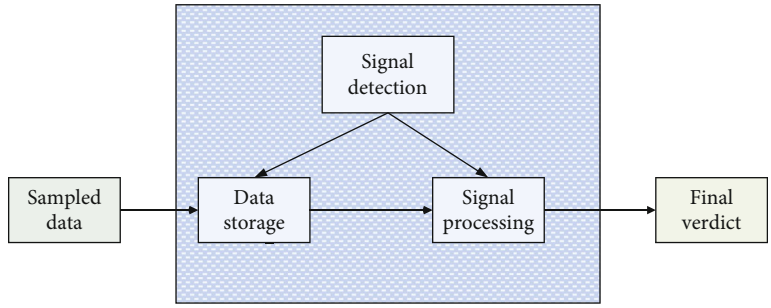


Figure 7: Cycle detection step diagram.

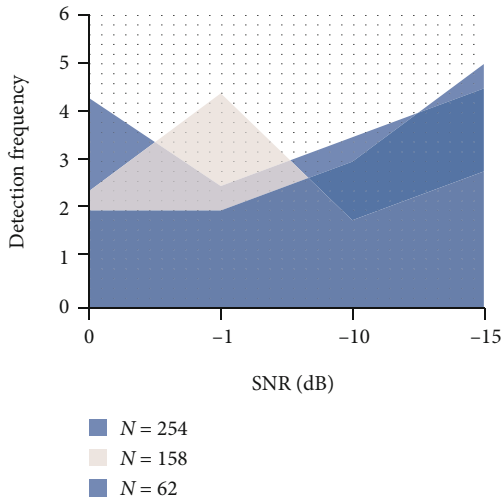


Figure 8: Cycle detection performance graph.

low; it can save the translation. Common acronyms such as GSO and MEO can also be untranslated. In short, the vocabulary translation of radio communication English seems simple, but in fact, each vocabulary needs to be carefully distinguished and considered before the appropriate translation method can be selected. Only in this way can the translation of the vocabulary be accurate [29].

4.2. Syntactic Skills. Combining with the above-mentioned general English syntactic structure features, according to Nida's functional equivalence translation theory, if technical concepts are to be understood by different audiences, we can analyze the syntactic translation skills in turn. First of all, according to the feature that passive sentences are often used in communication English, combined with translation practice, the passive structure is extracted into Chinese unsubjected sentences; passive structures are translated into Chinese active sentences by these two translation methods. Under normal circumstances, when the passive structure in English is unnecessary or unable to describe the performer of the action, it can be translated into a Chinese without subject sentence. This translation not only is accurate but also makes the sentences fluent, in line with Chinese expression habits, and easy for people to understand and accept. In general, in the translation of communication English sentences, due to the complexity of the long sentence, translators need to flexibly choose and use appropriate translation methods in accordance with the actual situation. Only the proper translation method can ensure the quality of translation, ensure that people can correctly understand the relevant content, and ensure the sound development of radio communication English translation [30].

5. Conclusions

Today's society is an information society, with rapid development and innovation of science and technology, and international exchanges and cooperation are becoming more frequent. The close exchanges and exchanges between countries have promoted the development of science and technology and have also given birth to the demand for science and technology translation. With the development of science and technology, people need to understand the latest technological development information through science and technology translation. Radio communication English translation belongs to a category of scientific translation. As the representative of the current advanced communication technology, radio communication has been used more and more widely all over the world. International exchanges and cooperation on radio communication technology are becoming more frequent. Therefore, the translation of radio communication English has become an important translation topic, and it has also become an important research field of practical significance. The research of radio communication English has strong practical significance and has a positive effect on promoting the development of radio communication technology [31]. This article briefly introduces radio communication technology, discusses the structural features and translation skills of radio communication English, and provides some methodological references and references for the translation of communication English. Radio communication English has unique characteristics of vocabulary and syntax. In terms of vocabulary, there are many acronyms, compound words, and professional vocabulary; in terms of syntax, long sentences with multiple structures, such as passive sentences and attributive clauses, are often used. Combining the characteristics of both the vocabulary and syntactic structure of Radio English, adopting corresponding translation strategies and methods, and prescribing the right medicine can improve the quality of radio communication English translation, ensure technical

exchanges, and promote technological development. As a field of technical English translation, the accuracy and logic of the translation of communication English are relatively high. The translator of every word and sentence should not take it lightly. This requires translators to pay attention to the accumulation of professional vocabulary and professional knowledge. Secondly, we must be good at using some appropriate translation methods and techniques flexibly, and methods such as literal translation, free translation, phonetic translation, provincial translation, passive translation, active, passive to no-owner are used to improve translation quality, so as to achieve the purpose of effectively conveying information to readers, realize the "faith" and "reach" of English translation of communication technology [32], so as to ensure the technical exchange and development of radio communication technology. In order to continuously improve the level of translation of communication English, translators also need to proactively seize every opportunity and have the courage to undertake the translation task of radio communication English and continuously reflect, summarize, and improve in translation practice. In addition, with the continuous development of radio communication technology, translators should keep up with the technological trend and always pay attention to radio communication information to ensure that they can keep abreast of the latest developments in radio communication technology and understand the technical content. Update related terminology translations at any time, and flexibly change translation strategies based on actual conditions, so as to ensure the correct delivery of information, better perform their own translation duties and obligations, and contribute to technological development. In the future, radio communication technology will continue to climb peaks and reach new technological heights one after another. The radio communication English translation will also be updated and developed continuously to ensure that technology is not caught in communication barriers and that people can keep abreast of the latest developments in the field of radio communication. Ensure good communication and cooperation between people and countries in communication technology. Due to the limited research level, this article still has some shortcomings, but the research on radio communication English translation skills will never stop. Academia will inevitably emerge more and more researches on the structural features and translation skills of radio communication English and will get more scientific and effective communication English translation skills, so as to promote the progress and development of radio communication English translation.

Data Availability

No data were used to support this study.

Conflicts of Interest

The author declares that there are no conflicts of interest regarding the publication of this article.

References

[1] S. Wan, Z. Gu, and Q. Ni, "Cognitive computing and wireless communications on the edge for healthcare service robots," *Computer Communications*, vol. 149, pp. 99–106, 2020.

[2] W. Li and H. Song, "ART: an attack-resistant trust management scheme for securing vehicular ad hoc networks," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 4, pp. 960–969, 2016.

[3] Z. Lv, "The security of Internet of drones," *Computer Communications*, vol. 148, pp. 208–214, 2019.

[4] W. Chunhui, "Structural features and translation of communication technology English," *China New Telecommunications*, vol. 18, no. 19, pp. 1–12, 2016.

[5] W. Qiang, "Composition and translation of compound words in communication English," vol. 12, no. 26, pp. 138–141, 2021.

[6] H. Wenting, "The characteristics of communication technology English terminology and translation strategies," vol. 7, no. 218, pp. 115–120, 2021.

[7] L. Dan and W. Yong, "Translation of passive voice in English for communication majors," *Examination and Evaluation (University English Teaching and Research Edition)*, vol. 96, no. 5, pp. 32–35, 2018.

[8] F. Yutao, M. Qingxun, and W. Haiyang, "Features and laws of English abbreviations for communication technology," *Chinese Science and Technology Translation*, vol. 14, no. 2, pp. 86–95, 2021.

[9] X. Tong, "Hypotaxis parataxis and Chinese translation of communication in English," *A Comparative Study of Cultural Innovation*, vol. 2, no. 10, pp. 3–15, 2019.

[10] X. Tingting, "Translation of non-predicate verbs in communication texts," *Comparative Research on Cultural Innovation*, vol. 3, no. 23, pp. 3–11, 2019.

[11] C. Li, P. Liu, C. Zou, F. Sun, J. M. Cioffi, and L. Yang, "Spectral-efficient cellular communications with coexistent one- and two-hop transmissions," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 8, pp. 6765–6772, 2016.

[12] L. Chuang, "The treatment of unsubjected sentences in Chinese-English translation of communication patents," *China Science and Technology Translation*, vol. 34, no. 3, pp. 4–23, 2021.

[13] H. Nishizawa, "Radio communication device," *Physics Experimentation*, vol. 5, no. 290, pp. 443–450, 2018.

[14] T. Cooklev, L. Wilhelmsson, and M. Ariyoshi, "Wireless and radio communications," *IEEE Communications Standards Magazine*, vol. 2, no. 4, pp. 42–42, 2018.

[15] Y. Yang, M. J. Crisp, R. V. Penty, and I. H. White, "Low-cost MIMO radio over fiber system for multiservice DAS using double sideband frequency translation," *Journal of Lightwave Technology*, vol. 34, no. 16, pp. 3818–3824, 2016.

[16] B. Han, J. Li, J. Su, and J. Cao, "Self-supported cooperative networking for emergency services in multi-hop wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 30, no. 2, pp. 450–457, 2012.

[17] H. Tian, Z. Liu, W. Xi, G. Nie, L. Liu, and H. Jiang, "Beam axis detection and alignment for uniform circular array-based orbital angular momentum wireless communication," *IET Communications*, vol. 10, no. 1, pp. 44–49, 2016.

[18] T. Cooklev, L. Wilhelmsson, and P. Zhu, "Wireless and radio communications," *IEEE Communications Standards Magazine*, vol. 3, no. 3, pp. 18–18, 2019.

[19] I. Kitouni, D. Benmerzoug, and F. Lezzar, "Smart agricultural enterprise system based on integration of internet of things and agent technology," *Journal of Organizational and End User Computing*, vol. 30, no. 4, pp. 64–82, 2018.

[20] S. Xu, G. Zhu, B. Ai, and Z. Zhong, "A survey on high-speed railway communications: a radio resource management perspective," *Computer Communications*, vol. 86, no. 2, pp. 12–28, 2016.

[21] Z. Lv, D. Chen, H. Feng, R. Lou, and H. Wang, "Beyond 5G for digital twins of UAVs," *Computer Networks*, vol. 197, article 108366, 2021.

[22] G. Kaddoum, "Wireless chaos-based communication systems: a comprehensive survey," *IEEE Access*, vol. 4, no. 27, pp. 2621–2648, 2016.

[23] G. Yong, "Application of full duplex guarantees secure wireless communication," *Journal of Communications and Networks*, vol. 19, no. 2, pp. 105–113, 2017.

[24] Z. Liu, "Modern security launch in wireless communication," *Advances in computational sciences and technology*, vol. 10, no. 12, pp. 3233–3238, 2017.

[25] W. Peng, D. Chen, W. Sun, C. Li, and G. Zhang, "Communication delay analysis under constrained condition for multi-radio WSNs," *Ad-hoc & sensor wireless networks*, vol. 42, no. 1, pp. 125–144, 2018.

[26] P. Jacob, R. P. Sirigina, A. S. Madhukumar, and V. A. Prasad, "Cognitive radio for aeronautical communications: a survey," *IEEE Access*, vol. 4, pp. 3417–3443, 2016.

[27] T.-Y. Kim and H.-J. Lee, "Vulnerability analysis of Bluetooth communication based on GNU radio," *Journal of the Korea Institute of Information and Communication Engineering*, vol. 20, no. 11, pp. 2014–2020, 2016.

[28] M. Condoluci, M. Dohler, G. Araniti, A. Molinaro, and J. Sachs, "Enhanced radio access and data transmission procedures facilitating industry-compliant machine-type communications over LTE-based 5G networks," *IEEE Wireless Communications*, vol. 23, no. 1, pp. 56–63, 2016.

[29] M. A. Abd-Elmagid, T. Elbatt, and K. G. Seddik, "A generalized optimization framework for wireless powered communication networks," *Wireless Networks*, vol. 9, pp. 1–18, 2016.

[30] J. C. Lin, "Interaction of wireless communication fields with blood-brain barrier of laboratory animals," *URSI Radio Science Bulletin*, vol. 315, pp. 33–38, 2017.

[31] S. Fukumoto, "Wireless communication system, base station device, mobile station device, and wireless communication method," *Fujifilm Corporation*, vol. 18, no. 6, pp. 3–9, 2017.

[32] K. Xu, B. Jiang, Z. Su et al., "High frequency communication network with diversity: system structure and key enabling techniques," *China Communications*, vol. 15, no. 9, pp. 46–59, 2018.