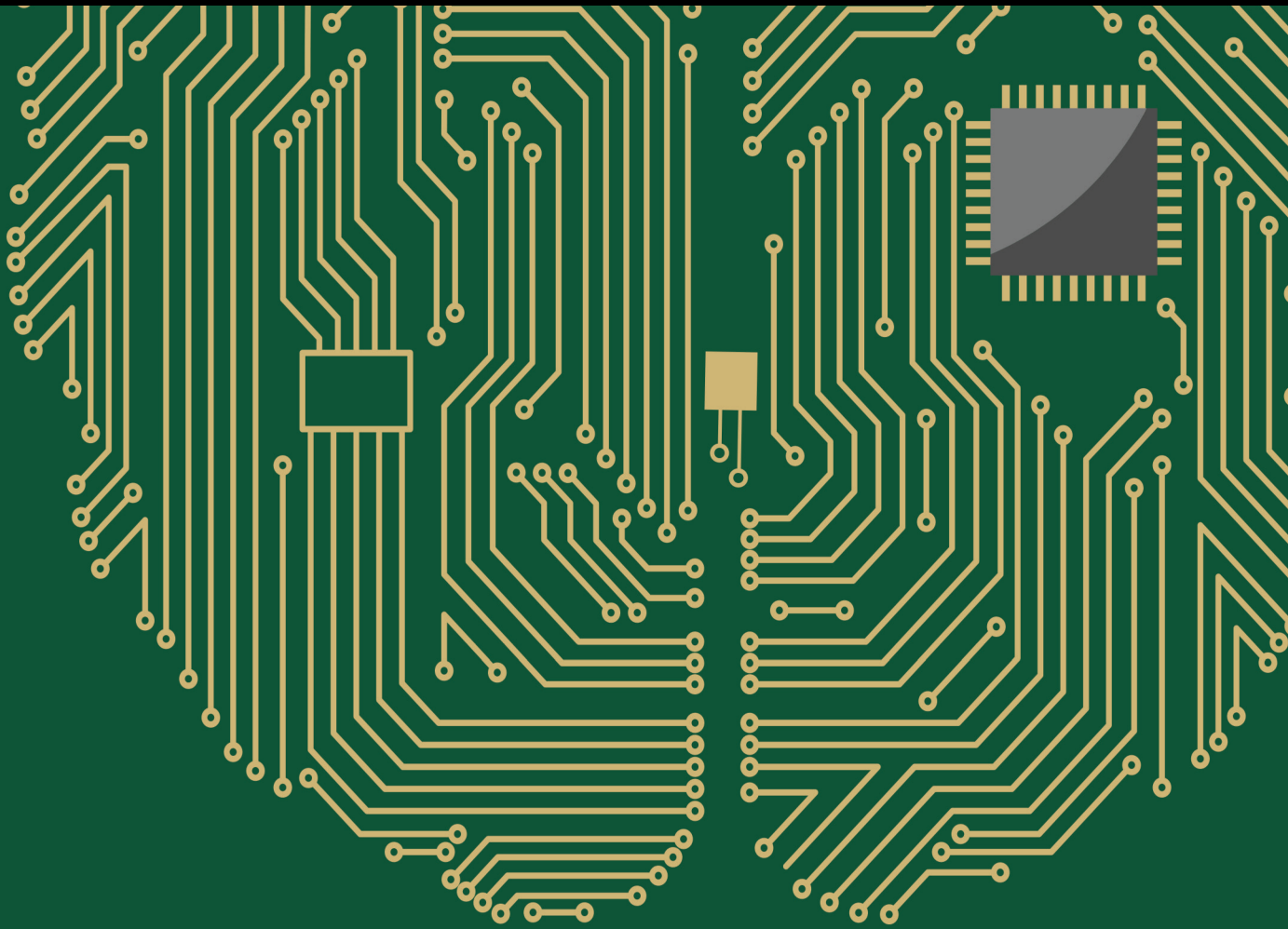


# Interpretation of Machine Learning: Prediction, Representation, Modeling, and Visualization 2022

Lead Guest Editor: Nian Zhang

Guest Editors: Zhishan Guo and Yide Zhang





---

**Interpretation of Machine Learning:  
Prediction, Representation, Modeling, and  
Visualization 2022**

Computational Intelligence and Neuroscience

---

**Interpretation of Machine Learning:  
Prediction, Representation, Modeling,  
and Visualization 2022**

Lead Guest Editor: Nian Zhang

Guest Editors: Zhishan Guo and Yide Zhang



---


Copyright © 2023 Hindawi Limited. All rights reserved.

This is a special issue published in "Computational Intelligence and Neuroscience." All articles are open access articles distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

# Chief Editor

Andrzej Cichocki, Poland

## Associate Editors

Arnaud Delorme, France  
Cheng-Jian Lin , Taiwan  
Saeid Sanei, United Kingdom

## Academic Editors



Mohamed Abd Elaziz , Egypt  
Tariq Ahanger , Saudi Arabia  
Muhammad Ahmad, Pakistan  
Ricardo Aler , Spain  
Nouman Ali, Pakistan  
Pietro Aricò , Italy  
Lerina Aversano , Italy  
Ümit Ağbulut , Turkey  
Najib Ben Aoun , Saudi Arabia  
Surbhi Bhatia , Saudi Arabia  
Daniele Bibbo , Italy  
Vince D. Calhoun , USA  
Francesco Camastra, Italy  
Zhicheng Cao, China  
Hubert Cecotti , USA  
Jyotir Moy Chatterjee , Nepal  
Rupesh Chikara, USA  
Marta Cimitile, Italy  
Silvia Conforto , Italy  
Paolo Crippa , Italy  
Christian W. Dawson, United Kingdom  
Carmen De Maio , Italy  
Thomas DeMarse , USA  
Maria Jose Del Jesus, Spain  
Arnaud Delorme , France  
Anastasios D. Doulamis, Greece  
António Dourado , Portugal  
Sheng Du , China  
Said El Kafhali , Morocco  
Mohammad Reza Feizi Derakhshi , Iran  
Quanxi Feng, China  
Zhong-kai Feng, China  
Steven L. Fernandes, USA  
Agostino Forestiero , Italy  
Piotr Franaszczuk , USA  
Thippa Reddy Gadekallu , India  
Paolo Gastaldo , Italy  
Samanwoy Ghosh-Dastidar, USA

Manuel Graña , Spain  
Alberto Guillén , Spain  
Gaurav Gupta, India  
Rodolfo E. Haber , Spain  
Usman Habib , Pakistan  
Anandakumar Haldorai , India  
José Alfredo Hernández-Pérez , Mexico  
Luis Javier Herrera , Spain  
Alexander Hošovský , Slovakia  
Etienne Hugues, USA  
Nadeem Iqbal , Pakistan  
Sajad Jafari, Iran  
Abdul Rehman Javed , Pakistan  
Jing Jin , China  
Li Jin, United Kingdom  
Kanak Kalita, India  
Ryotaro Kamimura , Japan  
Pasi A. Karjalainen , Finland  
Anitha Karthikeyan, Saint Vincent and the  
Grenadines  
Elpida Keravnou , Cyprus  
Asif Irshad Khan , Saudi Arabia  
Muhammad Adnan Khan , Republic of  
Korea  
Abbas Khosravi, Australia  
Tai-hoon Kim, Republic of Korea  
Li-Wei Ko , Taiwan  
Raşit Köker , Turkey  
Deepika Koundal , India  
Sunil Kumar , India  
Fabio La Foresta, Italy  
Kuruva Lakshmana , India  
Maciej Lawrynczuk , Poland  
Jianli Liu , China  
Giosuè Lo Bosco , Italy  
Andrea Loddo , Italy  
Kezhi Mao, Singapore  
Paolo Massobrio , Italy  
Gerard McKee, Nigeria  
Mohit Mittal , France  
Paulo Moura Oliveira , Portugal  
Debajyoti Mukhopadhyay , India  
Xin Ning , China  
Nasimul Noman , Australia  
Fivos Panetsos , Spain




Evgeniya Pankratova , Russia  
Rocío Pérez de Prado , Spain  
Francesco Pistolesi , Italy  
Alessandro Sebastian Podda , Italy  
David M Powers, Australia  
Radu-Emil Precup, Romania  
Lorenzo Putzu, Italy  
S P Raja, India  
Dr.Anand Singh Rajawat , India  
Simone Ranaldi , Italy  
Upaka Rathnayake, Sri Lanka  
Navid Razmjooy, Iran  
Carlo Ricciardi, Italy  
Jatinderkumar R. Saini , India  
Sandhya Samarasinghe , New Zealand  
Friedhelm Schwenker, Germany  
Mijanur Rahaman Seikh, India  
Tapan Senapati , China  
Mohammed Shuaib , Malaysia  
Kamran Siddique , USA  
Gaurav Singal, India  
Akansha Singh , India  
Chiranjibi Sitaula , Australia  
Neelakandan Subramani, India  
Le Sun, China  
Rawia Tahrir , Iraq  
Binhua Tang , China  
Carlos M. Travieso-González , Spain  
Vinh Truong Hoang , Vietnam  
Fath U Min Ullah , Republic of Korea  
Pablo Varona , Spain  
Roberto A. Vazquez , Mexico  
Mario Versaci, Italy  
Gennaro Vessio , Italy  
Ivan Volosyak , Germany  
Leyi Wei , China  
Jianghui Wen, China  
Lingwei Xu , China  
Cornelio Yáñez-Márquez, Mexico  
Zaher Mundher Yaseen, Iraq  
Yugen Yi , China  
Qiangqiang Yuan , China  
Miaolei Zhou , China  
Michal Zochowski, USA  
Rodolfo Zunino, Italy

# Contents




## **HAZMAT Vehicle Reidentification in Road Tunnels Based on the Fusion of Appearance and Spatiotemporal Information**

Lei Jia , Xiaobao Li, Wen Wang, Jianzhu Wang, Haomin Yu, Tianyuan Wang, and Qingyong Li   
Research Article (10 pages), Article ID 3677387, Volume 2023 (2023)

## **CAW: A Remote-Sensing Scene Classification Network Aided by Local Window Attention**

Wei Wang , Xiaowei Wen , Xin Wang , Chen Tang, and Jiwei Deng  
Research Article (10 pages), Article ID 2661231, Volume 2022 (2022)







## **Semisupervised Semantic Segmentation with Mutual Correction Learning**

Yifan Xiao, Jing Dong , Dongsheng Zhou , Pengfei Yi, Rui Liu, and Xiaopeng Wei   
Research Article (9 pages), Article ID 8653692, Volume 2022 (2022)




## **Fast Detection of Defective Insulator Based on Improved YOLOv5s**

Zhao Liquan , Zou Mengjun , Cui Ying , and Jia Yanfei   
Research Article (12 pages), Article ID 8955292, Volume 2022 (2022)




## **Intelligent Detection Method of Gearbox Based on Adaptive Hierarchical Clustering and Subset**

Huimiao Yuan , Yongwei Tang , Huijuan Hao , Yuanyuan Zhao , Yu Zhang , and Yu Chen   
Research Article (10 pages), Article ID 6464516, Volume 2022 (2022)


## **Feature Selection Based on Adaptive Particle Swarm Optimization with Leadership Learning**

Zhiwei Ye , Yi Xu , Qiyi He , Mingwei Wang, Wanfang Bai, and Hongwei Xiao  
Research Article (18 pages), Article ID 1825341, Volume 2022 (2022)




## **A Variable Radius Side Window Direct SLAM Method Based on Semantic Information**

Yan Chen , Jianjun Ni , Emmanuel Mutabazi, Weidong Cao , and Simon X. Yang  
Research Article (18 pages), Article ID 4075910, Volume 2022 (2022)

## **PointTransformer: Encoding Human Local Features for Small Target Detection**

Yudi Tang , Bing Wang, Wangli He, Feng Qian, and Zhen Liu  
Research Article (10 pages), Article ID 9640673, Volume 2022 (2022)

## **SR-DSFF and FENet-ReID: A Two-Stage Approach for Cross Resolution Person Re-Identification**

Zongzong Wu, Xiangchun Yu , Donglin Zhu , Qingwei Pang, Shitao Shen, Teng Ma, and Jian Zheng   
Research Article (11 pages), Article ID 4398727, Volume 2022 (2022)

## **Spatial-Temporal Change Trend Analysis of Second-Hand House Price in Hefei Based on Spatial Network**

Zheng Yin , Rui Sun , and Yuqing Bi   
Research Article (10 pages), Article ID 6848038, Volume 2022 (2022)

## **A Model for Surface Defect Detection of Industrial Products Based on Attention Augmentation**

Gang Li , Rui Shao , Honglin Wan , Mingle Zhou , and Min Li   
Research Article (12 pages), Article ID 9577096, Volume 2022 (2022)

## Research Article

# HAZMAT Vehicle Reidentification in Road Tunnels Based on the Fusion of Appearance and Spatiotemporal Information

Lei Jia <sup>1,2</sup> Xiaobao Li,<sup>1,2</sup> Wen Wang,<sup>1,2</sup> Jianzhu Wang,<sup>1,2</sup> Haomin Yu,<sup>1,2</sup> Tianyuan Wang,<sup>3</sup> and Qingyong Li <sup>1,2</sup>

<sup>1</sup>Beijing Key Lab of Traffic Data Analysis and Mining, Beijing Jiaotong University, Beijing 100044, China

<sup>2</sup>Frontiers Science Center for Smart High-Speed Railway System, Beijing Jiaotong University, Beijing 100044, China

<sup>3</sup>Shenzhen Urban Transport Planning Center Co. Ltd., Shenzhen 518000, China

Correspondence should be addressed to Qingyong Li; liqy@bjtu.edu.cn

Received 12 June 2022; Revised 9 October 2022; Accepted 11 October 2022; Published 14 February 2023

Academic Editor: Nian Zhang

Copyright © 2023 Lei Jia et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Vehicles transporting hazardous material (HAZMAT) pose a severe threat to highway safety, especially in road tunnels. Vehicle reidentification is essential for identifying and warning abnormal states of HAZMAT vehicles in road tunnels. However, there is still no public dataset for benchmarking this task. To this end, this work releases a real-world tunnel HAZMAT vehicle reidentification dataset, VisInt-THV-ReID, including 10,048 images with 865 HAZMAT vehicles and their spatiotemporal information. A method based on multimodal information fusion is proposed to realize vehicle reidentification by fusing vehicle appearance and spatiotemporal information. We design a spatiotemporal similarity determination method for vehicles based on the spatiotemporal law of vehicles in tunnels. Compared with other reidentification methods based on multimodal information fusion, i.e., PROVID, Visual + ST, and Siamese-CNN, experimental results show that our approach significantly improves the vehicle reidentification recognition precision.

## 1. Introduction

Hazardous materials (HAZMAT) could endanger the health and safety of people, environment, and property. With the increasing demand of HAZMAT, traffic accidents occurred frequently during HAZMAT transportation, and especially, a risk increase is generally observed in the presence of tunnels [1–3], which makes it of great importance to tighten regulation for vehicles transporting HAZMAT in tunnels.

HAZMAT vehicle reidentification (ReID) methods face the following challenges in tunnel scenes: (1) the strong reflection of the tank of a HAZMAT vehicle can cause large differences in its appearance under the uneven lighting conditions of a tunnel; (2) it is difficult to distinguish the HAZMAT vehicles with the same vehicle type effectively, due to their close appearance. However, there still remains a research gap both in HAZMAT vehicle data and in specialized algorithms. This motivates us to focus on the study of HAZMAT vehicle reidentification in tunnels.

Vehicle ReID aims to determine whether a vehicle image captured in nonoverlapping cameras belongs to the same vehicle in traffic monitoring scenarios. Existing methods mainly perform research on vehicle ReID based on the vehicle appearance [4]. However, due to the special and complex tunnel environment containing dim illumination and limited viewing field, it is more challenging for the tunnel vehicle ReID problem than that in open road scenes [5, 6]. Thus, large fluctuation can be seen by merely conducting tunnel vehicle ReID based on the appearance information. As shown in Figure 1, the red, green, and blue lines in each subfigure are RGB channel color histograms for each image. Vehicles for the second and third images may have similar appearance features, whereas they are actually two different IDs. From such instance, we can see that in real-world applications, it is extremely sensitive to environmental changes to merely perform vehicle ReID via appearance information.

To address the above problem, except for appearance information, the spatiotemporal information is further leveraged



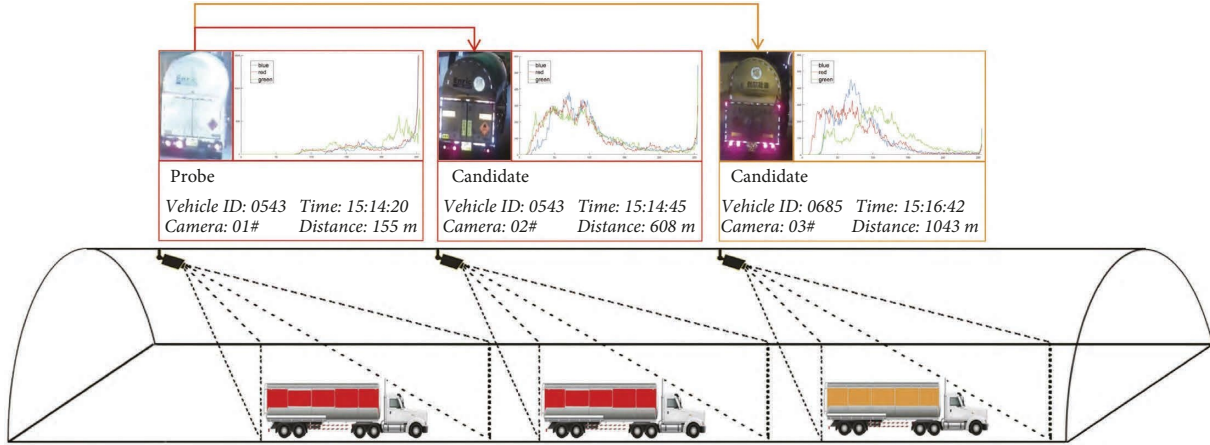


FIGURE 1: The HAZMAT vehicles are difficult to distinguish due to their close appearance. The reflection of the tank causes significant differences in its appearance under the variable lighting conditions in the tunnel.

to improve vehicle ReID performance in recent works [7–9]. This is inspired by the fact that the vehicle movements follow some implicit motion pattern according to the traffic rules. However, due to the randomness of vehicle motion, it is difficult to accurately model the spatiotemporal motion laws of vehicles in the open road. But the traffic rules of vehicles in tunnels are more distinct than in the open road, such as vehicles are expected to move in one fixed direction within limited speed, and no U-turns. It leads to the urgent need for a special spatiotemporal model tailored to the tunnel scene.

Therefore, to realize HAZMAT vehicle ReID in tunnel scenes, this work proposes a vehicle ReID method based on the fusion of vehicle appearance and tunnel spatiotemporal information. For vehicle appearance modeling, a deep residual network (i.e., Resnet50 [10]) is chosen as a feature extractor to model the complex appearance variation of tunnel vehicle. Meanwhile, to capture the spatiotemporal cues between cameras and vehicles, we develop a novel spatiotemporal similarity metric to model the between-vehicle structure correlation as well as the camera-vehicle topological relationship.

Furthermore, the extracted appearance representation and the spatiotemporal model are combined to efficiently encode the appearance variation and movement pattern for the tunnel vehicles. Moreover, to evaluate the HAZMAT vehicle ReID problem in the tunnel scenes, we construct and release a real-world HAZMAT Vehicle ReID dataset, named by VisInt-THV-ReID, containing 10,048 images of 865 HAZMAT vehicles collected from four high-resolution cameras. These images were captured by 4 cameras in the tunnel. Each camera monitors a space with a range of 150 meters and takes around 3 pictures of vehicles with far, middle, and near distances, respectively. Each vehicle is attached by the camera mileage and the picture shooting time. According to the spatial coordinate transformation method [11], we infer the spatial positions of vehicles in tunnel from the perspective of camera monitoring and obtain their temporal information by comparing timestamps of monitoring cameras. We use the vehicle ReID to determine whether the HAZMAT vehicles are exiting the tunnel within a normal time. If one vehicle passes the tunnel more than once, we identify the HAZMAT vehicle with a

different vehicle ID for each time in the dataset. More attention is paid to the driving condition of the HAZMAT vehicle each time when it passes through the tunnel. The proposed method is evaluated to be effective through exhaustive experiments on the VisInt-THV-ReID dataset.

The main contributions of this work are summarized as follows:

- (i) We extend the scenarios of vehicle ReID task to the challenging problem of HAZMAT vehicle ReID in tunnel scenes and propose a method that fuses both appearance modeling and spatiotemporal mining for more precise vehicle ReID.
- (ii) We design a spatiotemporal metric approach based on the movement law of vehicles in road tunnels which brings in the description of between-vehicle structure correlation as well as the camera-vehicle topological relationship.
- (iii) We build a real-world tunnel HAZMAT vehicle ReID dataset, named as VisInt-THV-ReID. As far as we know, the released VisInt-THV-ReID is the first HAZMAT vehicle ReID dataset captured in the tunnel scenes, which is crucial for the promotion of automatic regulation of HAZMAT transportation. Exhaustive experiments demonstrate that the proposed method can generate a state-of-the-art performance.

The rest of this work is organized as follows: The review related works are presented in Section 2. Section 3 details the proposed HAZMAT vehicle ReID method. In Section 4, we execute experiments for the evaluation of the proposed approach on VisInt-THV-ReID. Finally, we conclude this work in Section 5.

## 2. Related Work

Vehicle ReID in traffic monitoring scenarios can be seen as a part of multicamera tracking. Given an image of a vehicle in a specific area, the task is to find its image as captured under

other cameras. This work studies vehicle ReID with spatiotemporal information fusion in tunnel scenes. We introduce related work from the aspects of vehicle ReID in tunnel scenes and multimodal information fusion.

*2.1. Vehicle ReID Methods in Tunnels.* Vehicle ReID in tunnel scenes is challenging due to low resolution, dim light, and dramatic changes in vehicle appearance. A vehicle is detected and tracked by each camera in road tunnels, and a detected vehicle is matched with the previous camera.

Frías-Velázquez et al. [6] proposed a probabilistic framework based on a two-step strategy that reidentifies vehicles in road tunnels. They built a Bayesian model that finds the optimal assignment between vehicles of connected groups based on descriptors such as trace transform signatures, lane changes, and motion discrepancies. Rios-Cabrera et al. [12] presented an integrated solution to detect, track, and identify vehicles in a tunnel surveillance application, taking into account practical constraints, such as real-time operation, imaging conditions, and decentralized architecture. AdaBoost [13] cascade is used for vehicle detection, and a comprehensive confidence score integrates the information of all stages of the cascade. Jelača et al. [14] proposed a real-time tracking method of multiple non-overlapping cameras in a road tunnel monitoring scene, using AdaBoost for vehicle detection. The vehicle detector and a Kalman filter of average optical flow are used for tracking. The ReID algorithm applies the projection feature similarity of a radon transform between vehicle images. Chen et al. [15] proposed a spatiotemporal successive dynamic programming algorithm to identify vehicles between pairs of cameras. They extracted features based on Harris corner detection and OpponentSIFT descriptors, considering color information [16]. Zhu et al. [5] proposed a synergistically cascaded forest model to gradually construct the linking relationships between vehicle samples with increasing alternative random forest and extremely randomized forest layers.

The abovementioned methods generally focus on the extraction of hand-designed features of vehicle images, which can only show good performance in specific scenes. These manual features are susceptible to the interference of a complex tunnel environment, and they are difficult to improve the precision of ReID.

*2.2. Methods Using Multimodal Information.* As a vehicle is far from cameras and the illumination is insufficient, the image resolution is low. Due to their similarity, it is impractical to effectively identify HAZMAT vehicles without special markings only by appearance. Recent work on vehicle ReID has improved the model by combining multi-dimensional information of vehicle attributes such as type, color, time, and space information with appearance features.

To reidentify vehicles based on fusion different appearance information, Liu et al. [17] designed a network using BOW-SIFT [18], BOW-CN [19], and GoogLeNet [20] to extract texture, color, and semantic features, respectively. Handmade features are fused with the vehicle type and color

features obtained through deep learning. Liu et al. [21] proposed PROVID, which makes full use of appearance features, license plates, camera locations, and semantic information to carry out a progressive search from coarse to fine in the feature domain and from near to far in physical space.

To reidentify vehicles based on spatiotemporal information, Zhong et al. [7] proposed a vehicle pose guide model using a spatiotemporal probability model based on the Gaussian distribution to predict the spatiotemporal motion of vehicles. A convolution neural network (CNN) was used to predict the driving direction of a vehicle and the results of visual appearance, and then, the driving direction and spatiotemporal models were fused. Shen et al. [8] proposed a two-stage framework incorporating complex spatiotemporal information to effectively regularize ReID results. A candidate visual-spatiotemporal path was generated by a chain Markov random field model with a deeply learned potential function. A Siamese-CNN + Path-LSTM model takes the candidate path and pairwise queries to generate a similarity score. Jiang et al. [9] proposed an approach with a multi-branch architecture and a reranking strategy using the spatiotemporal relationship among vehicles from multiple cameras.

### 3. Method

*3.1. Overview.* Typically, a tunnel surveillance system consists of a series of cameras  $C = \{C_0, C_1, C_2, \dots, C_M\}$ , with nonoverlapping visual receptive fields.  $\vec{A}_i$  denotes the 2048-dimensional appearance feature vector obtained from the  $i$ -th vehicle image through the image appearance feature extraction network, and  $\vec{S}_i$  denotes the spatiotemporal feature vector of the  $i$ -th vehicle collected by the camera. The spatiotemporal features involved are velocity  $v_i$ , timestamp  $t_i$ , and space position  $l_i$  of the tunnel.

We use  $P_a(i, j)$  to represent the similarity of the appearance feature vectors of vehicles  $i$  and  $j$  from upstream and downstream cameras and  $P_{st}(i, j)$  to represent the similarity of the spatiotemporal features of the vehicle pairs.  $P(i, j)$  is the probability that vehicle pairs are identical after fusing multimodal information. The inputs of the proposed model are vehicle image pairs  $(i, j)$  and their spatiotemporal features  $(\vec{S}_i, \vec{S}_j)$  involved velocity, timestamp, and space position in the tunnel. The output is the probability  $P(i, j)$  of whether the pair of vehicle images is the same vehicle.

The framework of the proposed method has three parts, as shown in Figure 2.

- (1) Similarity calculation of vehicle appearance features. Resnet50 [10] is used as the feature extractor to obtain a 2048-dimensional appearance feature vector of a vehicle.
- (2) Based on the spatiotemporal movement law of HAZMAT vehicles, we calculate the theoretical distance and the actual distance of the vehicle pairs. The tunnel spatial discrepancy  $\varepsilon_{ij}$  is used to evaluate

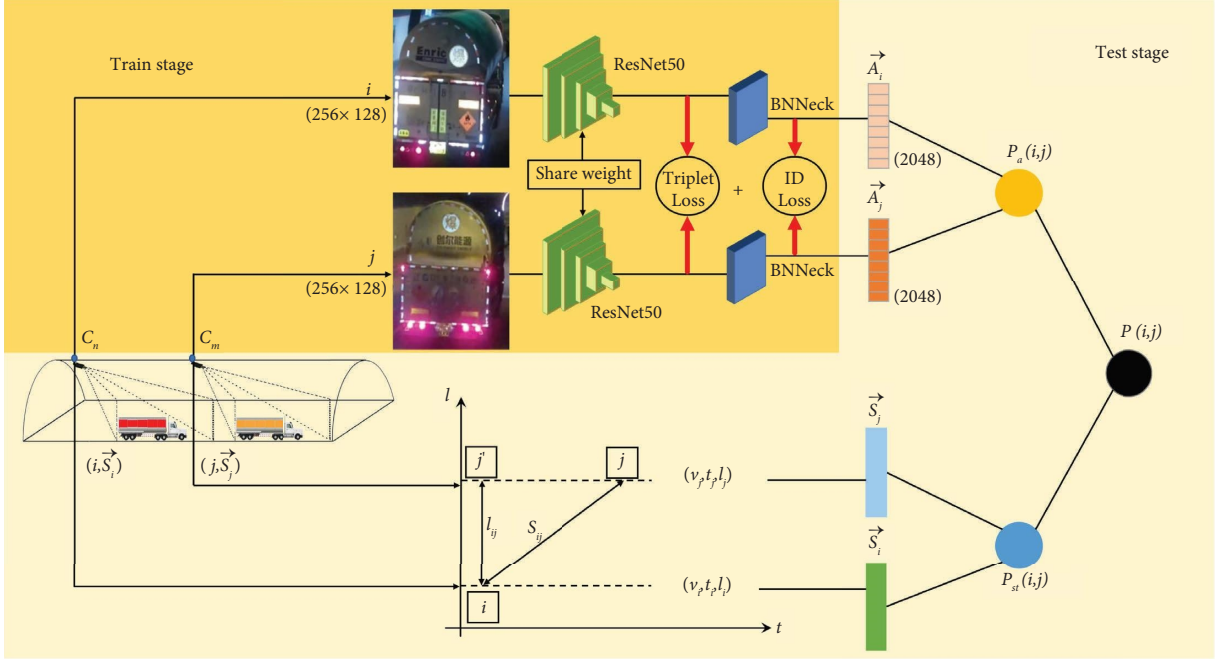


FIGURE 2: Vehicle ReID pipeline based on the fusion of appearance and spatiotemporal information.

the diversity between the theoretical distance and the actual distance.

- (3) Similarity calculation of multimodal information fusion. Based on parts 1 and 2, the spatiotemporal and appearance similarity of the input vehicle image pairs are summed with a weight. We rerank the vehicle similarity of fusion information.

**3.2. Appearance Features of Vehicle ReID.** The vehicle appearance feature extraction network is shown in Figure 3. We use Resnet50 as the feature extraction backbone network and adjust each image to  $256 \times 128$  pixels. Given an input image  $x_i$  with label  $y_i$ , the predicted probability of  $x_i$  being recognized as class  $y_i$  is encoded with a softmax function, represented by  $p(y_i | x_i)$ . ID prediction  $p(y_i | x_i)$  is used to calculate ID loss [22]. The model outputs ReID feature  $\vec{A}_i$  which is used to calculate triplet loss [23]. The output dimension of the full connection layer is changed to the number of vehicle IDs in the training dataset.

The ID loss treats the training process of vehicle ReID as an image classification problem [24], i.e., each identity is a distinct class. In the testing phase, the output of the pooling layer or embedding layer is adopted as the feature extractor. The identity loss is then computed by the cross-entropy.

$$L_{ID} = -\frac{1}{N} \sum_{i=1}^N \log(p(y_i | x_i)), \quad (1)$$

where  $N$  represents the number of training samples within each batch.

The triple loss for feature extraction can reduce the intraclass distance of positive pairs and increase the inter-class distance of negative pairs. Given a triplet  $(x^a, x^p, x^n)$ , including an anchor image  $x^a$ , a positive  $x^p$ , and negative  $x^n$ , the triplet loss is formulated as follows:

$$L_{Tri} = \sum_{i=1}^N \left[ \|f(x_i^a) - f(x_i^p)\|_2^2 - \|f(x_i^a) - f(x_i^n)\|_2^2 + \alpha \right], \quad (2)$$

where  $\alpha$  is a margin and usually set to 0.3.  $N$  is the number of training samples within each batch.  $f(\bullet)$  stands for the appearance feature extractor.

In this work, we use ID loss and triplet loss together for optimizing the model. For image pairs in the embedding space, ID loss mainly optimizes the cosine distances while triplet loss focuses on the Euclidean distances. The feature vectors of the two losses are inconsistent in the embedding space. To address this problem, the BNNNeck [22] is applied for more effective loss computation. BNNNeck adds a batch normalization (BN) layer before the classifier FC layers of the model. The feature before the BN layer is denoted as  $\vec{A}_i$ . We let  $\vec{A}_i$  pass through the BN layer to acquire a normalized feature  $\vec{a}_i$ . In the training stage, the feature  $\vec{A}_i$  is used to compute the triplet loss. The feature  $\vec{a}_i$  is used to compute the ID loss. Finally, the triplet loss and ID loss are combined to optimize the model. To train the ReID model, we combine ID loss and triplet loss as follows:

$$L = L_{ID} + L_{Tri}. \quad (3)$$

In the test stage, the appearance features  $(\vec{A}_i, \vec{A}_j)$  for input image pairs  $(i, j)$  are generated using the vehicle

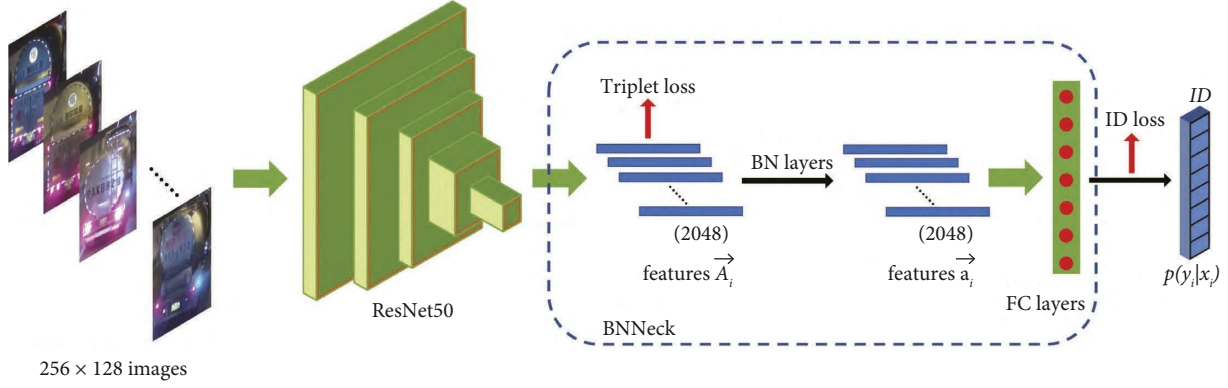


FIGURE 3: The framework of vehicle appearance modeling.

appearance feature extraction network. We use the cosine distance to measure the similarity between features and is expressed as follows:

$$P_a(i, j) = \frac{\vec{A}_i \cdot \vec{A}_j}{\|\vec{A}_i\| \|\vec{A}_j\|}. \quad (4)$$

**3.3. Vehicle Spatiotemporal Features.** The motion of the vehicle is limited by its speed and spatiotemporal motion. The time that the vehicle travels through a pair of cameras should be within a reasonable range. In a highway tunnel monitoring system, the driving speed of a vehicle is within the range of 10–80 km/h. The time interval of vehicle movement is affected by the camera installation position and the topological relationship of the tunnel and cameras. We analyze the motion law of the vehicle time interval between cameras in the VisInt-THV-ReID dataset. For each pair of cameras, the vehicle space interval can be modeled as a random variable that follows a certain distribution [6, 7].

In order to derive the spatiotemporal similarity probability distribution of the vehicle, we propose a feature called spatial discrepancy. We introduce the spatial discrepancy by considering Figure 4(a). This figure shows the spatiotemporal graph that relates vehicle  $i$  observed in upstream camera with another vehicle  $j$  observed in downstream camera. The motion variables involved are velocity  $v_i$  of vehicle  $i$ , timestamp  $t_i$ , and space position  $l_i$  of the tunnel. The state vector  $\vec{S}_i$  expresses the spatiotemporal state of vehicle  $i$ .

To construct the spatiotemporal similarity relationship between the vehicle pairs, we calculate the theoretical distance and the actual distance of the vehicle pairs and define the indicator  $\varepsilon_{ij}$  to calculate the diversity of the distances. According to the constant acceleration model, the theoretical distance of the vehicle is calculated as follows according to the upstream and downstream cameras of the tunnel:

$$s_{ij} = \frac{v_i + v_j}{2} \cdot (t_j - t_i). \quad (5)$$

The actual distance between the current position of the vehicle collected by the upstream and downstream cameras is expressed as follows:

$$l_{ij} = |l_j - l_i|. \quad (6)$$

The spatial discrepancy  $\varepsilon_{ij}$  evaluates the fitness between the displacement estimate  $s_{ij}$  and the actual distance  $l_{ij}$  as stated in Figure 4(a). The tunnel spatial discrepancy is expressed as follows:

$$\varepsilon_{ij} = \frac{(s_{ij} - l_{ij})}{|s_{ij}| + |l_{ij}|} \in (-1, 1), \quad (7)$$

which is used to evaluate the diversity between the theoretical distance and the actual distance. The spatial discrepancy  $\varepsilon_{ij}$  is evaluated by the vehicle spatiotemporal features involving velocity, timestamp, and space position.

To maintain the consistency of the data structure of the multimodal data fusion, we maintain the consistency of the spatiotemporal similarity discriminant method with the appearance feature discriminant method and use the chord function to represent the spatiotemporal similarity probability distribution of the vehicle. The  $P_{st}(i, j)$  is defined as follows:

$$P_{st}(i, j) = \cos\left(\varepsilon_{ij}^2 \cdot \frac{\pi}{2}\right). \quad (8)$$

As shown in Figure 4(b),  $P_{st}(i, j)$  increases as  $\varepsilon_{ij}$  tends to 0. Based on  $P_{st}(i, j)$ , we can determine candidate matching vehicles according to the spatiotemporal similarity in tunnels.

**3.4. Vehicle ReID by Fusing Image and Tunnel Spatiotemporal Information.** To make full use of the vehicle appearance and spatiotemporal information, we establish a multimodal information strategy. The vehicle ReID probability is defined as follows:

$$P(i, j) = \lambda P_a(i, j) + (1 - \lambda) P_{st}(i, j), \quad (9)$$

where the weight coefficient,  $\lambda \in (0, 1)$ , is used to fuse the spatiotemporal and appearance similarity.

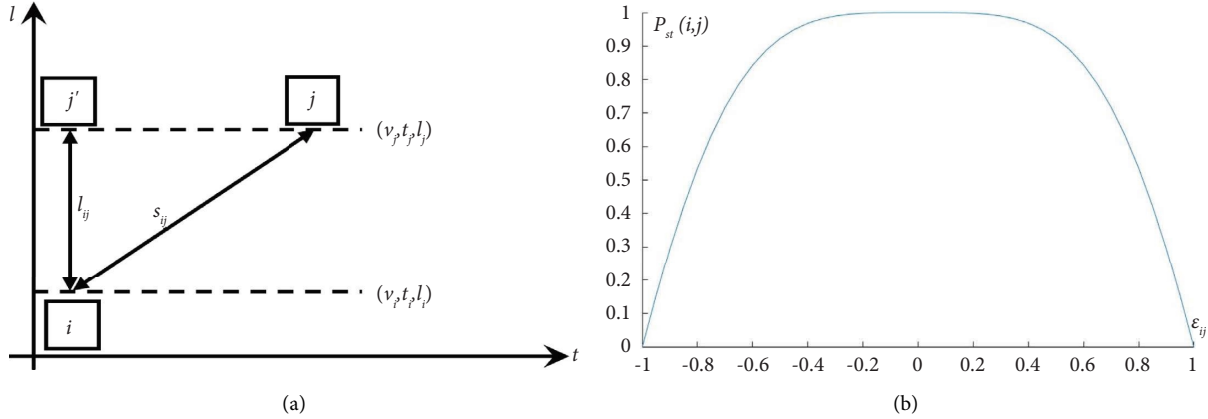


FIGURE 4: (a) Motion states of vehicles  $i$  and  $j$ . (b) Spatiotemporal similarity distribution in tunnels.

## 4. Experiments

**4.1. VisInt-THV-ReID Dataset.** We verified the effectiveness of the proposed method on the VisInt-THV-ReID (The dataset is open-sourced at the following website: <https://github.com/jialei-bjtu/VisInt-THV-ReID>) dataset, which is collected from four cameras deployed in Taijia Expressway Linxian No. 3 tunnel in Shanxi province, China, providing high-definition video data of 6 million pixels and spaced at 300 meters. We collected video data for 10 hours daily over 3 days, from November 26 to 28, 2020, from 10:00 to 20:00. We annotated 10,048 pictures of 865 HAZMAT vehicles with their spatial position, speed, and timestamp information. To the best of our knowledge, this is the first open-source HAZMAT vehicle ReID dataset. The sample dataset is shown in Figure 5.

To mark the spatiotemporal and speed information of a vehicle, we must transform its spatial coordinates. Perspective transformation is used to transform the vehicle driving area under the camera vision to a fixed-size rectangle [11], as shown in Figure 6.

The position  $(x_i, y_i)$  of a vehicle in the camera field of view in the tunnel is calculated as follows:

$$\left\{ \begin{array}{l} \left[ \begin{array}{l} x' \\ y' \\ \omega' \end{array} \right] = \left[ \begin{array}{l} x^o \\ y^o \\ 1 \end{array} \right] \cdot T, \\ T = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}, \\ \left[ \begin{array}{l} x_i \\ y_i \end{array} \right] = \left[ \begin{array}{l} \frac{x'}{\omega} \\ \frac{y'}{\omega} \end{array} \right], \end{array} \right. \quad (10)$$

where  $x_i$  is the lateral distance of the vehicle from the left wall of the tunnel,  $y_i$  is its longitudinal distance from the current camera installation position,  $(x^o, y^o)$  is the lower midpoint of the vehicle object detection box in the image,

and  $T$  is the transformation matrix defining the mapping between the original region and the transformation region. Using the image sequence taken by the surveillance camera, the speed of vehicle  $i$  in the tunnel can be obtained as follows:

$$v_i = \left( \sqrt{x_i^2 + y_i^2} - \sqrt{x_{i-1}^2 + y_{i-1}^2} \right) \cdot f, \quad (11)$$

where  $f$  is the frame rate of the monitoring camera, the spatial position vector  $l_i$  obtained by the camera at time  $t_i$  is  $(x_i, y_i)$ , and the spatiotemporal vector of vehicle  $i$  is  $S_i(v_i, t_i, l_i)$ .

We trained and tested the model on the VisInt-THV-ReID dataset, whose 10,048 images of 865 HAZMAT vehicles were divided into training, query, and test sets at a 10:1:9 ratio. The training set had 433 HAZMAT vehicles and 4980 images. There were 432 HAZMAT vehicles in the query and test sets, with 432 vehicle images in the query set and 4636 in the test set.

**4.2. Experimental Settings.** The mAP [21] and cumulative matching characteristic (CMC) curve [25] were used to evaluate the performance of the proposed method on the VisInt-THV-ReID dataset. The average precision for a query image is calculated as follows:

$$AP = \frac{\sum_{k=1}^n P(k) \cdot \text{gt}(k)}{N_{\text{gt}}}, \quad (12)$$

where  $n$  is the number of images in the test set,  $N_{\text{gt}}$  is the number of ground truths,  $P(k)$  is the current precision result of the  $k$ -th query image, and  $\text{gt}(k)$  is an indicator function. When the matching result of the  $k$ -th query image is correct,  $\text{gt}(k) = 1$ , and  $\text{gt}(k) = 0$  when it is incorrect.

The mAP is calculated as follows:

$$\text{mAP} = \frac{\sum_{q=1}^Q AP(q)}{Q}, \quad (13)$$

where  $Q$  is the number of pictures in the query dataset. The CMC curve shows the probability that the correct matching image of the vehicle appears in the candidate lists. The CMC of the  $k$ -th position is as follows:

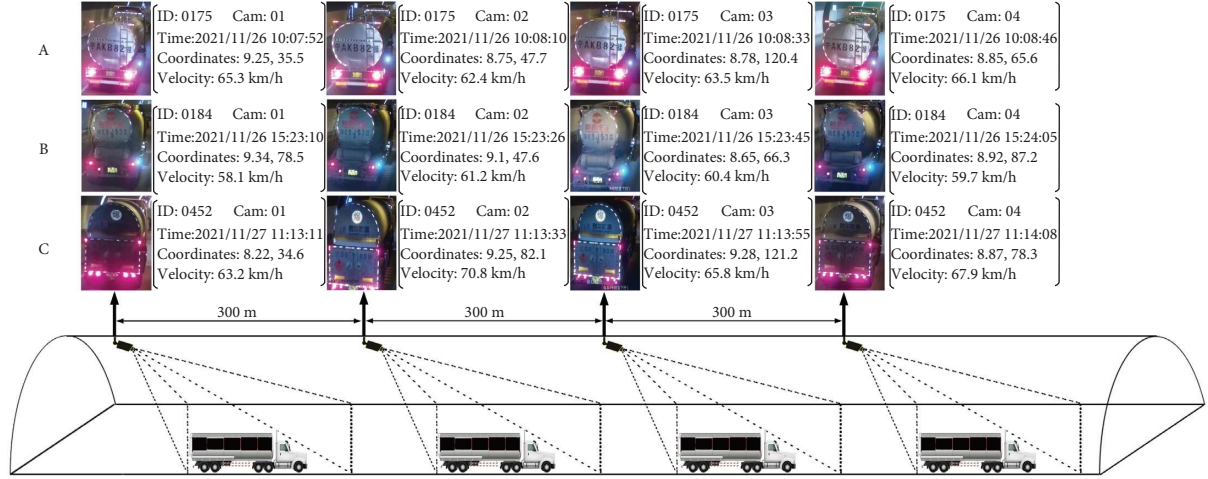


FIGURE 5: VisInt-THV-ReID dataset.

$$\text{CMC}(k) = \frac{\sum_{q=1}^Q \text{gt}(q, k)}{Q}, \quad (14)$$

where  $\text{gt}(q, k)$  is an indicator function, which equals 1 when the ground truth of the  $q$  query image appears before the  $k$  position. We also used Rank-1, Rank-5, Rank-10, and Rank-20 in the field of ReID to evaluate the model.

**4.3. Ablation Study.** Table 1 compares the experimental results of the multimodal fusion ReID method with those of Visual and ST-COS, which are appearance-based and spatiotemporal-based, respectively.

The method of Visual achieved 89.7% mAP and 96.3% Rank-1. The method of ST-COS achieved 85.5% mAP and 71.3% Rank-1. The fusion method Visual + ST-COS achieved 99.7% mAP and 99.8% Rank-1. The mAP of the fusion method increases by 142% and 10% compared to Visual and ST-COS and the Rank-1 rises by 3.5% and 28.5%.

The above results show that the multimodal information fusion method is superior to the use of appearance or spatiotemporal information alone and verify the effectiveness of the proposed multimodal information fusion method.

**4.4. Comparison with Baselines.** Table 2 shows the recognition precision of three baseline methods, PROVID [21], Visual + ST [7], and Siamese-CNN [8], comparing to that of Visual + ST-COS on the VisInt-THV-ReID dataset.

**4.4.1. Appearance Feature Extraction and STR Spatiotemporal Fusion (PROVID).** The method of PROVID extracts the appearance features of HAZMAT vehicles by the Resnet50 network and uses the STR method to measure the spatiotemporal relationship [21]. The STR is defined as follows:

$$\text{STR}(i, j) = \frac{T_i - T_j}{T_{\max}} \cdot \frac{\delta(C_i, C_j)}{D_{\max}}, \quad (15)$$

where  $T_i$  and  $T_j$  are the timestamps for the vehicles  $i$  and  $j$  captured by the cameras.  $T_{\max}$  is the maximum time interval of vehicles passing through the tunnel.  $\delta(C_i, C_j)$  is the actual distance between the current position of the vehicles collected by the upstream and downstream cameras, and  $D_{\max}$  is the global maximum distance between any vehicles. We set  $D_{\max}$  as the length of the tunnel.

**4.4.2. Visual + ST.** The method of Visual + ST extracts the appearance features of HAZMAT vehicles with the Resnet50 network and uses a spatiotemporal model based on the Gaussian distribution to predict the probability of vehicles [7].  $P_{\text{stG}}(i, j)$  presents the similarity of the spatiotemporal features of vehicle pairs, and it is defined as follows:

$$P_{\text{stG}}(i, j) = e^{-10\epsilon_{ij}^2}, \quad (16)$$

where  $\epsilon_{ij}$  is the tunnel spatial discrepancy as defined in equation (7).

**4.4.3. Siamese-CNN.** The method of Siamese-CNN uses a Resnet50 network to extract the appearance features of HAZMAT vehicles, and a multilayer perceptron network is applied to obtain their spatial and temporal relationships [8]. The spatiotemporal branch computes the spatiotemporal compatibility. Given the timestamps  $(t^i, t^j)$  and the positions  $(l^i, l^j)$  of vehicles, the input features of the branch are calculated as their time difference  $\Delta t(t^i, t^j)$  and spatial difference  $\Delta d(l^i, l^j)$ . The scalar spatiotemporal compatibility is obtained by feeding the concatenated features,  $[\Delta t(t^i, t^j), \Delta d(l^i, l^j)]^T$ , into a multilayer perceptron with two fully connected layers. The outputs of the two branches are concatenated and input into a  $2 \times 1$  fully connected layer with a sigmoid function to obtain the final compatibility

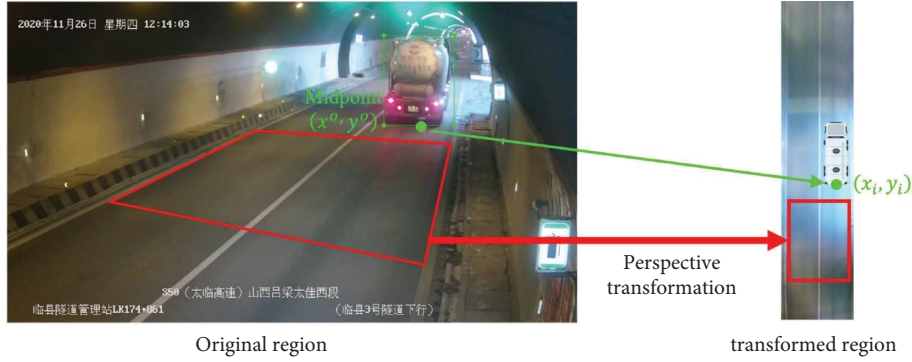


FIGURE 6: Coordinate transformation of vehicle position in tunnel space based on surveillance video.

TABLE 1: Results of ablation experiment.

Methods	mAP (%)	Rank-1 (%)	Rank-5 (%)	Rank-10 (%)	Rank-20 (%)
Visual	89.7	96.3	99.5	99.5	99.8
ST-COS	85.5	71.3	85.9	98.8	100
Visual + ST-COS	99.7	<b>99.8</b>	<b>100</b>	<b>100</b>	<b>100</b>

The bold values in Table 1 are the best values from the same column of data.

TABLE 2: Results of comparative experiments.

Methods	mAP (%)	Rank-1 (%)	Rank-5 (%)	Rank-10 (%)	Rank-20 (%)
PROVID	90.0	95.6	99.5	99.8	99.8
Visual + ST	90.8	96.1	99.5	99.8	99.8
Siamese-CNN	82.2	96.8	98.4	99.1	99.3
Our method	<b>99.7</b>	<b>99.8</b>	<b>100</b>	<b>100</b>	<b>100</b>

The bold values in Table 2 are the best values from the same column of data.

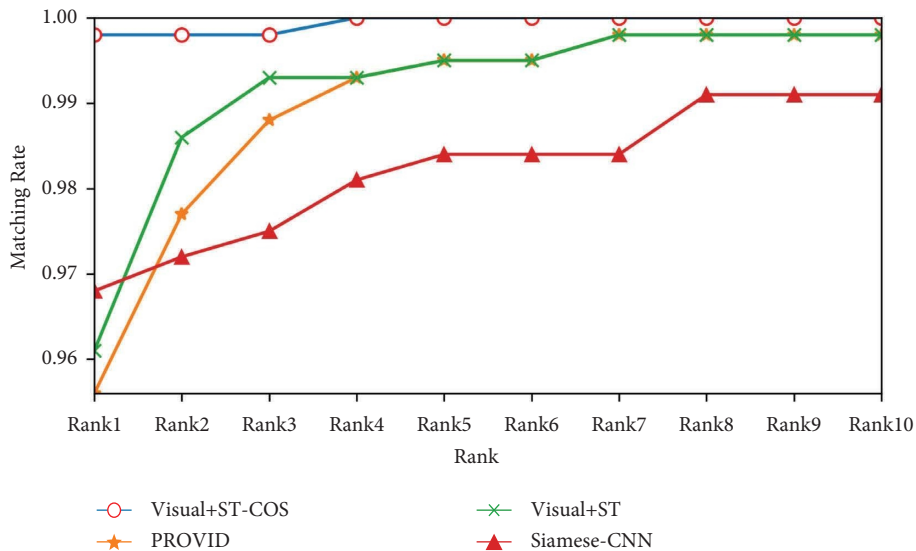


FIGURE 7: CMC curves on VisInt-THV-ReID dataset.

between the two states. Siamese-CNN takes all visual, spatial, and temporal information into consideration.

The results show that the proposed method achieves the best performance. It improves *mAP* and Rank-1 by 9.7% and

4.2%, respectively, compared with PROVID. This indicates that the STR spatiotemporal measurement method is not accurate enough to express the spatiotemporal information of vehicles in road tunnels. Compared with Siamese-CNN,

TABLE 3: Experimental results of coefficient  $\lambda$  under different values.

Results	$\lambda = 0.1$	$\lambda = 0.2$	$\lambda = 0.3$	$\lambda = 0.35$	$\lambda = 0.4$	$\lambda = 0.5$	$\lambda = 0.6$	$\lambda = 0.7$	$\lambda = 0.8$	$\lambda = 0.9$
mAP	88.2	97.3	99.3	<b>99.7</b>	99.7	99.7	99.7	99.5	98.6	96.2
Rank-1	81.0	98.8	99.8	<b>99.8</b>	99.8	99.7	99.8	99.8	99.5	99.1
Rank-5	94.7	100	100	<b>100</b>	99.8	99.8	99.8	99.8	99.8	99.8
Rank-10	99.8	100	100	<b>100</b>	100	100	99.8	99.8	99.8	99.8
Rank-20	100	100	100	<b>100</b>	100	100	100	99.8	99.8	99.8

The bold values in Table 3 are the best values from the same column of data.

the proposed method improves mAP and Rank-1 by 17.5% and 3.0%. Since Siamese-CNN uses a multilayer perception network to train the spatial and temporal information of vehicles, the difficulty of model training is decreased and the precision is not ideal. Compared with Visual+ST, the proposed method improves mAP and Rank-1 by 8.9% and 3.7%, respectively. This shows that the proposed cosine spatiotemporal model can more accurately express the spatiotemporal state of a tunnel compared with Gaussian distribution. The CMC curves of all methods are shown in Figure 7.

**4.5. Parameter Analysis.** We experimented with the parameters of  $\lambda$  in the interval of 0.1–0.9. The best fusion result is achieved when  $\lambda$  equals 0.35. The comparison results of the parametric experiments are shown in Table 3. It can be observed from the table that a larger  $\lambda$  would cause appearance features to dominate vehicle identification, while a smaller  $\lambda$  causes spatiotemporal information to dominate. Table 3 shows that  $\lambda$  can have an important effect on the fusion results, and  $\lambda$  is relatively insensitive to the results in the interval 0.3–0.7.

## 5. Conclusion and Future Work

In this study, we presented a vehicle ReID method based on the fusion of vehicle appearance and tunnel spatiotemporal information for the task of HAZMAT vehicle ReID in road tunnels. The proposed method was evaluated on the VisInt-THV-ReID dataset. This study could play a role in promoting HAZMAT vehicle monitoring and traffic safety management in road tunnels.

Our future work has two aspects. Based on vehicle ReID research, we will study multicamera vehicle tracking technology to collect vehicle trajectories. In addition, we will use the time-to-collision (TTC) to indirectly evaluate safety and study a tunnel accident risk prediction model based on the traffic flow state.

## Data Availability

The data that support the findings of this study are openly available in GitHub at <https://github.com/jialei-bjtu/VisInt-THV-ReID>.

## Conflicts of Interest

The authors declare that they have no conflicts of interest.

## Authors' Contributions

Jia L. and Li X. conceived and designed the experiments; Li X. and Wang J. performed the experiments; Tianyuan W. and Haomin Y. analyzed the data; Jia L. and Wang W. wrote the paper; Li Q. reviewed and edited the paper. All authors have read and agreed to the published version of the manuscript.

## Acknowledgments

This work was supported in part by the Fundamental Research Funds for the Central Universities (Science and Technology Leading Talent Team project, grant no. 2022JBQY007), NIM R&D Project under grant no. 35-AKYZD2116-1, and China Academy of Railway Sciences Co., Ltd R&D Project under grant no. 2021IMXM04.

## References

- [1] R. Bubbico, S. Di Cave, B. Mazzarotta, and B. Silveti, "Preliminary study on the transport of hazardous materials through tunnels," *Accident Analysis & Prevention*, vol. 41, no. 6, pp. 1199–1205, 2009.
- [2] B. Fabiano and E. Palazzi, "HazMat transportation by heavy vehicles and road tunnels: a simplified modelling procedure to risk assessment and mitigation applied to an Italian case study," *International Journal of Heavy Vehicle Systems*, vol. 17, no. 3/4, p. 216, 2010.
- [3] L. Jia, J. Wang, T. Wang, X. Li, H. Yu, and Q. H. M. D.-N. Li, "HMD-net: a vehicle hazmat marker detection benchmark," *Entropy*, vol. 24, no. 4, p. 466, 2022.
- [4] J. Deng, Y. Hao, M. S. Khokhar et al., "Trends in vehicle Re-identification past, present, and future: a comprehensive review," *Mathematics*, vol. 9, no. 24, p. 3162, 2021.
- [5] R. Zhu, J. Fang, S. Li et al., "Vehicle re-identification in tunnel scenes via synergistically cascade forests," *Neurocomputing*, vol. 381, pp. 227–239, 2020.
- [6] A. Frias-Velázquez, P. Van Hese, A. Pižurica, and W. Philips, "Split-and-match: a Bayesian framework for vehicle re-identification in road tunnels," *Engineering Applications of Artificial Intelligence*, vol. 45, pp. 220–233, 2015.
- [7] X. Zhong, M. Feng, W. Huang, Z. Wang, and S. Satoh, "Poses guide spatiotemporal model for vehicle Re-identification," in *MultiMedia Modeling* Springer International Publishing, Berlin, Germany, 2018.
- [8] Y. Shen, T. Xiao, H. Li, S. Yi, and X. Wang, "Learning deep neural networks for vehicle Re-ID with visual-spatio-temporal path proposals," in *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*, IEEE, Venice, Italy, October 2017.



- [9] N. Jiang, Y. Xu, Z. Zhou, and W. Wu, "Multi-attribute driven vehicle Re-identification with spatial-temporal Re-ranking," in *Proceedings of the 2018 25th IEEE International Conference on Image Processing (ICIP)*, IEEE, Athens, Greece, October 2018.
- [10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Las Vegas, NV, USA, June 2016.
- [11] Z. Xiong, M. Li, Y. Ma, and X. Wu, "Vehicle Re-identification with image processing and car-following model using multiple surveillance cameras from urban arterials," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 12, pp. 7619–7630, 2021.
- [12] R. Rios-Cabrera, T. Tuytelaars, and L. Van Gool, "Efficient multi-camera vehicle detection, tracking, and identification in a tunnel surveillance application," *Computer Vision and Image Understanding*, vol. 116, no. 6, pp. 742–753, 2012.
- [13] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119–139, 1997.
- [14] V. Jelača, J. O. N. Castañeda, A. Frías-Velázquez, A. Pižurica, and W. Philips, "Real-time Vehicle Matching for Multi-Camera Tunnel Surveillance," *Real-Time Image and Video Processing 2011*, pp. 232–239, Society of Photographic Instrumentation Engineers, Cergy-Pontoise, France, 2011.
- [15] H. T. Chen, M. C. Chu, C. L. Chou, S. Y. Lee, and B. S. Lin, "Multi-camera vehicle identification in tunnel surveillance system," in *Proceedings of the 2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, IEEE, Turin, Italy, June 2015.
- [16] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek, "Evaluating color descriptors for object and scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1582–1596, 2010.
- [17] X. Liu, W. Liu, H. Ma, and H. Fu, "Large-scale vehicle re-identification in urban surveillance videos," in *Proceedings of the 2016 IEEE International Conference on Multimedia and Expo (ICME)*, IEEE, Seattle, WA, USA, July 2016.
- [18] L. Zheng, S. Wang, W. Zhou, and Q. Tian, "Bayes merging of multiple vocabularies for scalable image retrieval," in *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Columbus, OH, USA, June 2014.
- [19] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person Re-identification: a benchmark," in *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, IEEE, Santiago, Chile, December 2015.
- [20] C. Szegedy, W. Liu, Y. Jia et al., "Going deeper with convolutions," in *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Boston, MA, USA, June 2015.
- [21] X. Liu, W. Liu, T. Mei, and H. Ma, "PROVID: progressive and multimodal vehicle reidentification for large-scale urban surveillance," *IEEE Transactions on Multimedia*, vol. 20, no. 3, pp. 645–658, 2018.
- [22] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, "Bag of tricks and a strong baseline for deep person Re-identification," in *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, IEEE, Long Beach, USA, March 2019.
- [23] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: a unified embedding for face recognition and clustering," in *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Boston, MA, USA, June 2015.
- [24] L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang, and Q. Tian, "Person Re-identification in the wild," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3346–3355, Honolulu, HI, USA, July 2017.
- [25] H. Liu, Y. Tian, Y. Wang, L. Pang, and T. Huang, "Deep relative distance learning: tell the difference between similar vehicles," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Las Vegas, NV, USA, June 2016.

## Research Article

# CAW: A Remote-Sensing Scene Classification Network Aided by Local Window Attention

Wei Wang , Xiaowei Wen , Xin Wang , Chen Tang, and Jiwei Deng

School of Computer and Communication Engineering, Changsha University of Science and Technology, Changsha 410114, China

Correspondence should be addressed to Xin Wang; wangxin@csust.edu.cn

Received 27 August 2022; Accepted 23 September 2022; Published 11 October 2022

Academic Editor: Nian Zhang

Copyright © 2022 Wei Wang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Remote-sensing image scene data contain a large number of scene images with different scales. Traditional scene classification algorithms based on convolutional neural networks are difficult to extract complex spatial distribution and texture information in images, resulting in poor classification results. In response to the above problems, we introduce the vision transformer network structure with strong global modeling ability into the remote-sensing image scene classification task. In this paper, the parallel network structure of the local-window self-attention mechanism and the equivalent large convolution kernel is used to realize the spatial-channel modeling of the network so that the network has better local and global feature extraction performance. Experiments on the RSSCN7 dataset and the WHU-RS19 dataset show that the proposed network can improve the accuracy of scene classification. At the same time, the effectiveness of the network structure in remote-sensing image classification tasks is verified through ablation experiments, confusion matrix, and heat map results comparison.

## 1. Introduction

With the development of satellite remote-sensing technology and unmanned aerial vehicle technology, the intersection of remote sensing and computer vision provides a new research area for remote-sensing image processing. Compared to terrestrial imagery, remote-sensing imagery provides a different perspective to describe the Earth's surface and facilitate a range of Earth observation missions [1]. Remote-sensing image scene classification is the fundamental work for understanding remote-sensing imagery and plays an important role in remote-sensing imagery applications such as Land Use/Land Cover (LULC) classification [2–4] and urban planning [5].

Remote-sensing image scene classification refers to the classification of different remote-sensing images in a dataset according to certain classification features, so the key to scene classification lies in the extraction of image features. The following are three types of methods for image feature extraction: First, the feature descriptors are directly extracted from the image, such as color histogram, scale-invariant feature transform SIFT [6], directional gradient

histogram HOG [7], and local binary pattern LBP ; the second is to continue feature extraction based on some underlying features extracted from image blocks, such as the bag-of-words model BOVW and sparse coding [8]; and the third is to automatically extract features from images through deep learning methods. Each of the three methods has its own advantages and disadvantages, while the deep learning method does not need to manually extract feature descriptors, and it possesses excellent classification effect, so the trend of using deep learning methods for remote-sensing image scene classification is increasing [9, 10] at present. Among these deep learning methods, traditional convolutional neural network (CNN) is the most widely used one. Compared with traditional handcrafted feature extraction methods, its multistage feature extraction architecture can extract more discriminative semantic features and provides an end-to-end framework. Deep learning techniques for remote-sensing image scene classification are mainly divided into three types, namely unsupervised image classification, supervised image classification, and object-based image analysis [11]. In this paper, the technique of supervised classification is used to classify remote-sensing images.

The difficulty of remote-sensing scene classification is that when determining the scene scheme, (1) the size of key objects varies greatly, (2) many objects unrelated to the scene scheme are often submerged in the image, and (3) compared with natural images, remote-sensing scenes are more complex in terms of spatial arrangement and object distribution [12, 13]. Therefore, how to effectively perceive regions of interest of different sizes and build more discriminative representations from complex object distributions is crucial for understanding remote-sensing scenes. Figure 1 below shows the changes in the size and number of objects in the aerial images selected in this paper.

In recent years, the transformer has achieved great success in the fields of natural language processing (NLP) and speech processing (SP). Due to its powerful global feature extraction capability, this structure was introduced into the field of computer vision [14]. The dominant model in the field of computer vision is the CNN network. As the transformer structure becomes more and more efficient, the use of the vision transformer to complete visual tasks has become a new research direction. Vision transformer has powerful global modeling capabilities, but there are some limitations, such as the lack of information exchange in the local area, the large amount of parameters and calculation, getting extremely prone to over-fitting, and the internal structure information of the image block getting destroyed in the process of image patching. In response to the above problems, researchers have redesigned the vision transformer network model. One of the design solutions is to combine the vision transformer and CNN network structure. This network can fuse the global modeling ability of vision transformer and the local feature extraction ability of CNN to improve the model efficiency and performance to a certain extent, such as the conformer [15], CoAtNet [16], visual attention network (VAN [17]), twins [18], and LocalViT [19]. Another method is to control the model capacity by dividing the input feature map into small windows for local-window self-attention. This method can enhance the capture efficiency of local relationships and greatly reduce the computational complexity of the model, such as the Swin transformer [20]. However, it should be noted that in this method, there will be the problem of window limitation. The information of the image only interacts in each small window, and there will be a lack of information interaction between the windows. A Swin transformer uses a shifted window attention to construct the global input image, but it is not constructed in overlapping local windows, so weights can only be shared on channel dimensions and not including global weight sharing on space, and in the form of shifted window attention, it does not really override the relationship between global objects.

For remote-sensing scene classification tasks, it is extremely important to design a network that can learn local and global features to solve the problem of the size change of key targets in each pixel area. The contributions of this paper mainly include the following three points:

- (1) A parallel model structure is proposed, which spatially solves the problem of limited receptive field of

small window self-attention and enhances the spatial-channel modeling capability of the network

- (2) According to some lightweight vision transformer structures, the computational efficiency has been improved
- (3) The enhanced classification module is introduced to enhance the feature representation capability of high-level feature remote-sensing image scenes and enhance the expressive capability of the network

Compared with other network structures, this network has higher classification accuracy. Validated on the RSSCN7 dataset and WHU-RS19 dataset, it achieved good results.

The rest of the chapter is structured as follows. The second section is related work, including the research status and analysis of some lightweight convolutional neural network structures and vision transformer structures, as well as the role of parallel structures in feature extraction. Section 3 provides the method of this paper, including the overall framework of the network and the introduction of each module. Section 4 shows the experiments of our method on two remote-sensing scene classification datasets. Finally, a conclusion is drawn in Section 5.

## 2. Related Work

*2.1. Scene Classification Lightweight Network.* For the traditional convolutional neural network, the core of the lightweight network is to lighten the network in terms of volume and speed under the premise of maintaining the accuracy as much as possible. For example, the classic convolutional neural network SqueezeNet [21] uses model compression to replace  $3 \times 3$  convolution with  $1 \times 1$  convolution to reduce the amount of parameters and calculation and ShuffleNet [22] proposes pointwise group convolution and channel shuffle to maintain accuracy and reduce the parameters and calculation. MobileNet [23] proposes a depthwise separable convolution structure instead of ordinary convolution, which greatly reduces the model volume and improves the calculation speed. These network structures are widely used in scene classification tasks due to their low computational cost [24].

The introduction of the traditional vision transformer structure into the remote-sensing scene classification task will inevitably introduce a large amount of parameters and calculations. In the existing research, the work of reducing the parameters and calculations of the vision transformer model while maintaining the network accuracy are as follows: The Swin transformer divides the feature map into multiple small windows, adopts the local-window self-attention mechanism in the small windows to reduce the computational complexity, and realizes the global modeling of the image on the channel through the shifted window attention operation and obtains good results; MPViT [25] uses multiscale patch and multipath structure, while reducing the number of channels and reducing model parameters to achieve good performance; CMT [26] introduced depthwise separable convolution in the self-



FIGURE 1: (a) and (b) Examples of object size and number variation in remote-sensing images.

attention module to downsample the feature map, by which computational resources are saved effectively.

This paper refers to the lightweight structure and principles of the convolutional neural network and the vision transformer structure and designs a lightweight network architecture that can combine the advantages of the vision transformer and the convolutional neural network feature extraction.

**2.2. Transformer Parallel Structure.** Parallel structures in neural networks, such as GoogLeNet, [27] improve network performance by paralleling convolution kernels of different sizes (different receptive fields) and Big-Little Net [28] obtains multiscale features by fusing two branches at different scales. According to the structural characteristics of the convolutional network structure and transformer structure, iFormer [29] applies the frequency ramp structure to trade off the high and low frequency components and improves the efficiency through the channel splitting mechanism. In order to be able to learn key objects of different sizes within remote-sensing images and use less amount of parameters and calculation, this paper parallelizes equivalent large convolution kernels with local-window self-attention capturing local relations and global feature extraction.

The channel assignment in the parallel network structure can be divided into two types: one is to compress the channel to a specified number by point convolution, and the other is to divide the channel into a specified number by channel split [30]. Compared with channel split, the method of applying point convolution for channel compression has more parameters. Finally, we split the feature map output by patch merging into two equal parts by channel split and then use channel concatenating and shuffling. The method integrates different features in the branch to realize the construction of global features in the network space-channel range.

### 3. Methodology

**3.1. Framework Overview.** The overall framework of this network structure is shown in Figure 2(a), which consists of three parts: stem, stage, and enhanced classification. Stem consists of convolutional layers and pooling layers, which downscale an input image of size  $256 \times 256$  to  $64 \times 64$ . Each stage consists of the patch merging module and CAW module. Patch merging mainly plays the role of downsampling the image, and CAW block is the main feature extraction module. The patch merging module changes the size of the feature map to 1/2 times the original size by selecting elements in the row and column directions of the

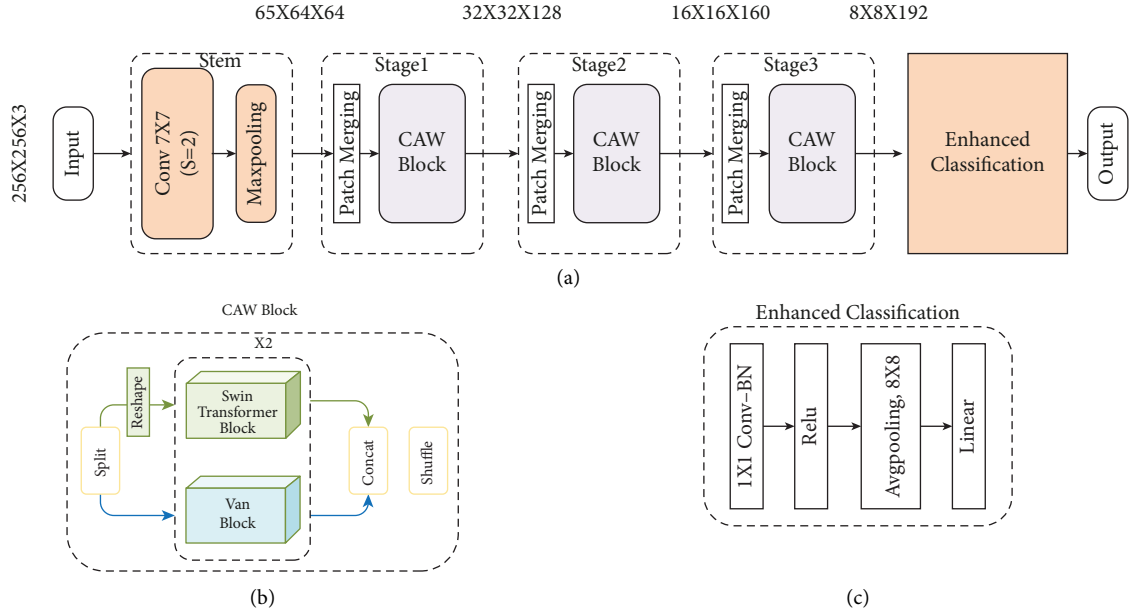


FIGURE 2: Network structure diagram.

feature map at intervals of 2 and then stitching them together as a whole tensor. At this time, the channel dimension will become original four times, and a fully connected layer is used to adjust the channel dimension to twice the original to achieve downsampling. After the feature extraction of the three-layer stage, the feature map is input into the enhanced classification layer to obtain the final classification result.

**3.2. CAW Block.** In the task of remote-sensing scene classification, it is of great significance to the classification of remote-sensing scenes to better capture the characteristics of target objects of different sizes and make the features more representative. The concatenation of local-window self-attention mechanism and shifted window self-attention can realize the global modeling of the image in the channel direction. For general image classification tasks, images are generally localized, and this structure can learn most of the content in the image. However, in the scene classification image, there are changes in the size of the target object, so it is particularly important to introduce a global modeling in the space. In the process of using vision transformer to patch the feature map, the internal structure information of the image block will be destroyed, and the feature map is not patched when the convolutional neural network is used to extract the features of the image, which can ensure the integrity of the internal features of the image. Therefore, we consider adding a convolution kernel to the parallel branch for feature extraction. In the convolutional neural network, a larger convolution kernel can achieve more global feature extraction, but a large convolution kernel will bring a huge amount of parameters and calculation, so we introduce the VAN module. The VAN network mainly consists of two parts which are the large kernel attention (LKA) structure and the multilayer perceptron (MLP) structure, where the LKA structure uses a

$5 \times 5$  depthwise convolution, a  $7 \times 7$  depthwise convolution (with a dilation rate of 3), and a  $1 \times 1$  convolution to approximate a  $21 \times 21$  convolution kernel, which can be used in the image with a slight compute costs and parameters to capture long-range relationships.

The CAW block proposed in this paper is a parallel structure module of vision transformer. The Swin transformer divides the feature map into several small windows and then uses the self-attention mechanism for feature extraction for each small window, while the VAN mainly is composed of LKA and MLP. LKA stacks depthwise convolution (DW-Conv), depthwise dilated convolution (DW-D-Conv), and  $1 \times 1$  convolution ( $1 \times 1$  Conv) to make LKA equivalent for larger convolutional neural networks. In this paper, a Swin transformer with a window size of  $4 \times 4$  and a VAN network with an equivalent window of  $21 \times 21$  are used to form a parallel structure. This parallel mechanism not only retains the feature extraction advantages of the Swin transformer's local-window self-attention but also makes up the window limit problem for the Swin transformer. The CAW block module diagram is shown in Figure 2(b), and the input feature map channel is divided into two equal parts. The operation description and expressions of the entire network structure are as follows, where  $X, Y \in R^{h \times w \times c/2}$  are the feature maps obtained by patch merging and channel split.

The feature map of the upper branch converts the feature map of size  $H \times W \times C/2$  into the feature vector of  $HW \times C/2$  through reshape operation and then uses layer normalization (LN) to normalize the feature vector, and inputs the Swin transformer module for feature extraction; the Swin transformer module is mainly composed of windowed multihead self-attention (W-MSA), moving window multihead self-attention (SW-MSA), MLP, and skip connections. The output formula of the local-window self-attention branch is expressed as

$$\begin{aligned}
X' &= W - \text{MSA}(\text{LN}(\text{Reshape}(X))) + \text{Reshape}(X), \\
X_1 &= \text{MLP}(\text{LN}(X')) + X', \\
X'_1 &= \text{SW} - \text{MSA}(\text{LN}(X_1)) + X_1, \\
X_2 &= \text{MLP}(\text{LN}(X'_1)) + X'_1.
\end{aligned} \tag{1}$$

The feature map of the lower branch enters the VAN for global feature enhancement. In the VAN module, the feature map is first normalized through batch normalization (BN), then through a  $1 \times 1$  convolution kernel, then nonlinearly activated with Gaussian Error Linear Unit (GELU), then through LKA and a  $1 \times 1$  convolution kernel, and finally passes through the MLP structure. The output formula of the global feature supplementary branch is expressed as

$$\begin{aligned}
Y' &= \text{Conv1} \times 1 (\text{LKA}(\text{GELU}(\text{Conv1} \times 1 (\text{BN}(Y)))) + Y, \\
Y_1 &= \text{MLP}(\text{BN}(Y')) + Y', \\
Y'_1 &= \text{Conv1} \times 1 (\text{LKA}(\text{GELU}(\text{Conv1} \times 1 (\text{BN}(Y_1)))) + Y_1, \\
Y_2 &= \text{MLP}(\text{BN}(Y'_1)) + Y'_1.
\end{aligned} \tag{2}$$

Finally, merge the feature maps of the two branches and then perform the Shuffle operation to shuffle the feature maps in the two channels so that the feature maps of the two channels are fused. The final output of the module is

$$\text{OUTPUT} = \text{Shuffle}(\text{Concat}(X_2, Y_2)). \tag{3}$$

**3.3. Enhanced Classification.** Current CNNs usually take the final downsampling operation, the fully connected layer, and the softmax classifier as a whole, treating it as a classification layer. Some salient features of this classification layer include those as follows: It usually does not have any convolutional layers, the number of parameters is small, and it is usually a linear feature representation structure. For remote-sensing image scenes, owing to interclass similarity and intraclass variation, it is necessary to highlight local semantics and more discriminative features. Therefore, it is particularly important to optimize the classification layer to have stronger feature representation capabilities. To enhance the feature representation of high-level feature remote-sensing imagery scenes, an additional  $1 \times 1$  convolutional layer and a ReLU activation function are added before the classifier. As shown in Figure 2(c), adding a  $1 \times 1$  convolutional layer before the classifier can increase the nonlinearity of the network and enhance the expressive ability of the network to a certain extent.

## 4. Experiments and Results

**4.1. Network Complexity.** This network is designed based on the vision transformer structure. In order to ensure the accuracy of the network and reduce the amount of parameters and calculation of the network structure, this paper refers to some vision transformer network structures with less parameters and less calculation in the design of the network structure. In order to prove the effectiveness of the

network structure proposed in this paper in remote-sensing image classification tasks, this paper selects some classic convolutional neural networks and vision transformer structures for comparative experiments. The comparison table of parameters and calculation is shown in Table 1:

**4.2. Dataset.** This paper conducts experiments on two widely used remote-sensing image classification datasets: RSSCN7 dataset and WHU-RS19 dataset.

The RSSCN7 dataset [34] was released by Qin Zou of Wuhan University in 2015. It contains 2800 remote-sensing images and a total of seven typical scene categories including grassland, forest, farmland, parking, residential, industrial, river, and lake. Each category contains 400 images with a pixel size of  $400 \times 400$ , and the diversity of scene images makes it more challenging. In the experiment, we divide the dataset into training sets and test sets in an 8:2 ratio by random selection.

The WHU-RS19 dataset [12] was released by Wuhan University in 2011, containing 1005 remote-sensing images and a total of 19 typical scene categories including airports, beaches, bridges, business districts, deserts, farmland, football fields, forests, factories, grassland, mountains, parks, parking, ponds, ports, railway stations, residential, rivers, and viaducts, each of which contains 50 images with a pixel size of  $600 \times 600$ . Compared with the RSSCN7 dataset, this dataset is more diverse and has fewer training samples, so it is more challenging. The distribution ratio of training set and test set of this dataset is the same as that of RSSCN7 dataset.

**4.3. Evaluation Criteria.** In this section, we explain the evaluation metrics used to quantify the classification performance of network models: accuracy, precision, sensitivity, specificity, and F1-score. To represent the above metrics, we also need to count four quantities in the confusion matrix: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). The evaluation index formula is expressed as follows:

$$\begin{aligned}
\text{Accuracy} &= \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \\
\text{Precision} &= \frac{\text{TP}}{\text{TP} + \text{FP}}, \\
\text{Recall} &= \frac{\text{TP}}{\text{TP} + \text{FN}}, \\
\text{Specificity} &= \frac{\text{TN}}{\text{TN} + \text{FP}}, \\
\text{F1 - score} &= 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}
\end{aligned} \tag{4}$$

Confusion matrices are often used to measure model classification performance. This matrix can intuitively reflect the difference between the predicted value and the true

TABLE 1: Comparison of parameters and calculations of the model.

Model	FLOPs (G)	Parameter (M)
ResNet50 [31]	5367.48	23.52
Vgg16 [32]	20185.96	138.36
DenseNet [33]	3742.61	7.98
GoogleNet [27]	2071.13	6.99
ViT-Ti [14]	21980.16	86.38
Swin transformer [20]	7078.50	28.24
VAN [17]	1149.08	41.00
Conformer-Ti [15]	3241.03	11.31
CMT [26]	1580.74	8.17
MPViT [25]	4212.25	6.08
CAW	566.61	1.27

value. It consists of four quantities: TP, TN, FP, and FN, which are specifically expressed as follows:

$$\begin{bmatrix} \text{True Positive (TP)} & \text{False Negative (FN)} \\ \text{False Positive (FP)} & \text{True Negative (TN)} \end{bmatrix}. \quad (5)$$

**4.4. Preprocessing and Experimental Set-Up.** In order to obtain a better training effect, the pictures in the experiment are all subjected to the same preprocessing. First, the pictures in the dataset are scaled and adjusted to  $256 \times 256$ , and then the pictures are digitized and normalized. The normalized means set is  $[0.485, 0.456, 0.406]$ , and standard deviation is set to  $[0.229, 0.224, 0.225]$ .

The experimental environment of this paper is shown in the following table, including software and hardware information, and the same experimental environment and experimental platform are applied to ensure the fairness and feasibility of the experiment. The training set and test set use the batchsize of 16, and the optimizer uses AdamW, the weight decay coefficient is  $5e-2$ , and the learning rate is 0.0001. The experimental platform data is shown in Table 2.

In the training process, in order to make the network get better convergence effect, a total of 500 epochs were trained in each experiment. We take the highest value of the recognition accuracy of the experimental test set as the final classification accuracy and use the accuracy, sensitivity, precision, specificity, and F1 value as evaluation indicators.

**4.5. Experimental Results and Discussion.** In order to verify that the introduction of VAN based on the structure of Swin transformer can solve the problem of limited receptive field of the Swin transformer and improve the classification effect of remote-sensing scene images, this paper conducts experiments on the RSSCN7 dataset and the WHU-RS19 dataset. Among them, 4 sets of ablation experiments and 10 sets of comparison experiments are set on the RSSCN7 dataset, and 10 sets of comparison experiments are set on the WHU-RS19 dataset. The comparative experiments in this paper include 4 groups of classic convolutional neural networks and 6 groups of transformer structure-related network structures. In order to ensure the accuracy of the experimental results, all experiments in this paper use the

TABLE 2: Experimental platform data.

Attributes	Configuration information
Operating system	Windows 10
CPU	Intel(R) Core (TM) i5-10300H CPU @ 2.50 GHz
GPU	GeForce RTX 2060
CUDA	CUDA 11.6.110
Frame	PyTorch 3.7

same experimental environment, learning rate, loss function, optimizer, batchsize, etc.

In order to study the influence of the depth of CAW on the classification performance of remote-sensing images, we increased the number of module layers at different stages, and compared the accuracy, parameter amount, and computation amount of CAW-Net with different depths, where brackets represent the number of CAW blocks at different stages. The experimental data are shown in Tables 3 and 4:

From the experimental results, with the increase of the number of network layers, the amount of parameters and the amount of calculation increase, the model appears overfitting, which leads to a decrease in the accuracy rate. Considering both the classification performance and model complexity, we believe that CAW (1, 1, 1) has the best price-performance ratio.

In order to prove the complementarity of the two vision transformer structures and achieve the effect of improving the performance of remote-sensing scene image classification, in the ablation experiment, we split and replace the two branches into four different combined structures to conduct experiments on RSSCN7. The maximum value in the 500 epochs is used as the experimental result, and the experimental results are shown in Table 5. Among them, the Swin transformer-only and VAN-only models are network models obtained by paralleling the same module with other structures unchanged; No Shuffle is the network model obtained by removing the Shuffle structure in the original network structure; and point convolution is a network structure model that replaces the channel segmentation structure in the original network structure with point convolution for channel compression.

It can be seen from Table 5 that the parallel connection of Swin transformer and VAN can solve the problem of limited receptive field of local-windows self-attention and further improve the performance of the network. Compared with using the two modules alone, the accuracy is increased by 0.54% and 1.56%, respectively; Adding Shuffle after the two branches which are connected in parallel can better integrate the features of the two branches, and the network accuracy is increased by 0.89%. In the channel allocation, the spilt operation is better than channel compression, which improves the network performance by 0.72%. Considering the classification performance and model complexity, we believe that this network structure has the best cost performance. In Figure 3, we give the seven-category confusion matrix of the RSSCN7 dataset of this network, and Figure 4 shows the 19-category confusion matrix of this network.

TABLE 3: Comparison results of CAW-Net networks with different depths.

Model	Accuracy (%)	Precision (%)	Recall (%)	Specificity (%)	F1-score (%)
CAW (1, 2, 1)	95.18	95.24	95.17	99.21	95.19
CAW (1, 2, 2)	95.54	95.39	95.54	99.29	95.53
CAW (1, 2, 3)	95.35	95.40	95.34	99.24	95.37
CAW (1, 1, 1)	<b>96.25</b>	<b>96.27</b>	<b>96.24</b>	<b>99.40</b>	<b>96.24</b>

TABLE 4: Comparison of parameters and calculations of CAW-Net with different depths.

Model	FLOPs (G)	Parameter (M)
CAW (1, 2, 1)	656.68	1.58
CAW (1, 2, 2)	695.12	2.03
CAW (1, 2, 3)	733.55	2.48
CAW (1, 1, 1)	566.61	1.27

TABLE 5: Comparison results of parallel networks with different structures.

Model	Accuracy (%)	Precision (%)	Recall (%)	Specificity (%)	F1-score (%)
Swin transformer-only	95.71	95.84	95.73	99.30	95.71
VAN-only	94.69	94.69	94.64	99.14	94.63
No shuffle	95.36	95.39	95.34	99.24	95.36
Point convolution	95.53	95.64	95.53	99.27	95.56
CAW	<b>96.25</b>	<b>96.27</b>	<b>96.24</b>	<b>99.40</b>	<b>96.24</b>

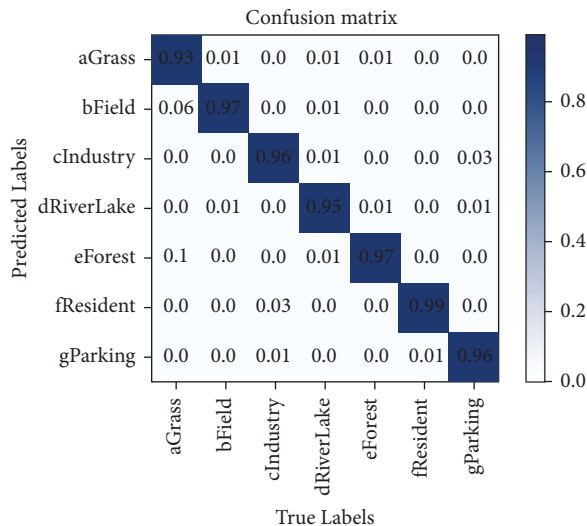


FIGURE 3: RSSCN7 dataset classification confusion matrix.

In order to reflect the recognition effect of this network structure on remote-sensing datasets, this paper uses some classic convolutional neural network models and vision transformer network models that perform well in computer vision for comparative experiments. The experiments are performed on the RSSCN7 dataset and the WHU-RS19 dataset under the same environment. The experimental results are shown in Tables 6 and 7. The experimental results with the best effect are marked in bold, and the results are kept to two decimal places.

From the results in Tables 5–7, we can see that compared with other network structures, the network structure proposed in this paper achieves good results on

remote-sensing datasets with exponentially reduced parameters and calculation. The parameters of this network are 9.4 times that of ResNet50, 38.8 times that of ViT-Ti, and 12.5 times that of Swin transformer. Compared with these networks, on the RSSCN7 dataset, the accuracy rates of the networks proposed in this paper have increased by 1.79%, 5.36%, and 2.32%, respectively, and the accuracy rates on the WHU-RS19 dataset have increased by 1.46%, 14.08%, and 4.86%.

We apply Gradient-weighted Class Activation Mapping (Grad-CAM) [35] to a different network, using images from the RSSCN7 validation set. Grad-CAM is a recently proposed visualization method, which highlights the feature map in the form of a heat map in order to visualize the feature representation learned by the neural network from an intuitive effect.

As shown in Figure 5, we compare the visualization results of Swin transformer, VAN, and our network. Both the Swin transformer and VAN can capture the area where the target object is located, but it is not accurate enough and there is a certain misjudgment; for example, in the factory scene, the field next to the factory with a similar color is misjudged as a factory. Although VAN can identify the scene area in these scenes, it is more divergent. For example, in the grass and industry scenes, the VAN network can capture the area where the grass and the industry are located, but the range is small and not accurate enough. Our model captures details representing semantic features in complex background images, and it has higher confidence than baseline models in the classification of some difficult objects. We can infer that our model has stronger feature extraction ability and can learn more discriminative features.



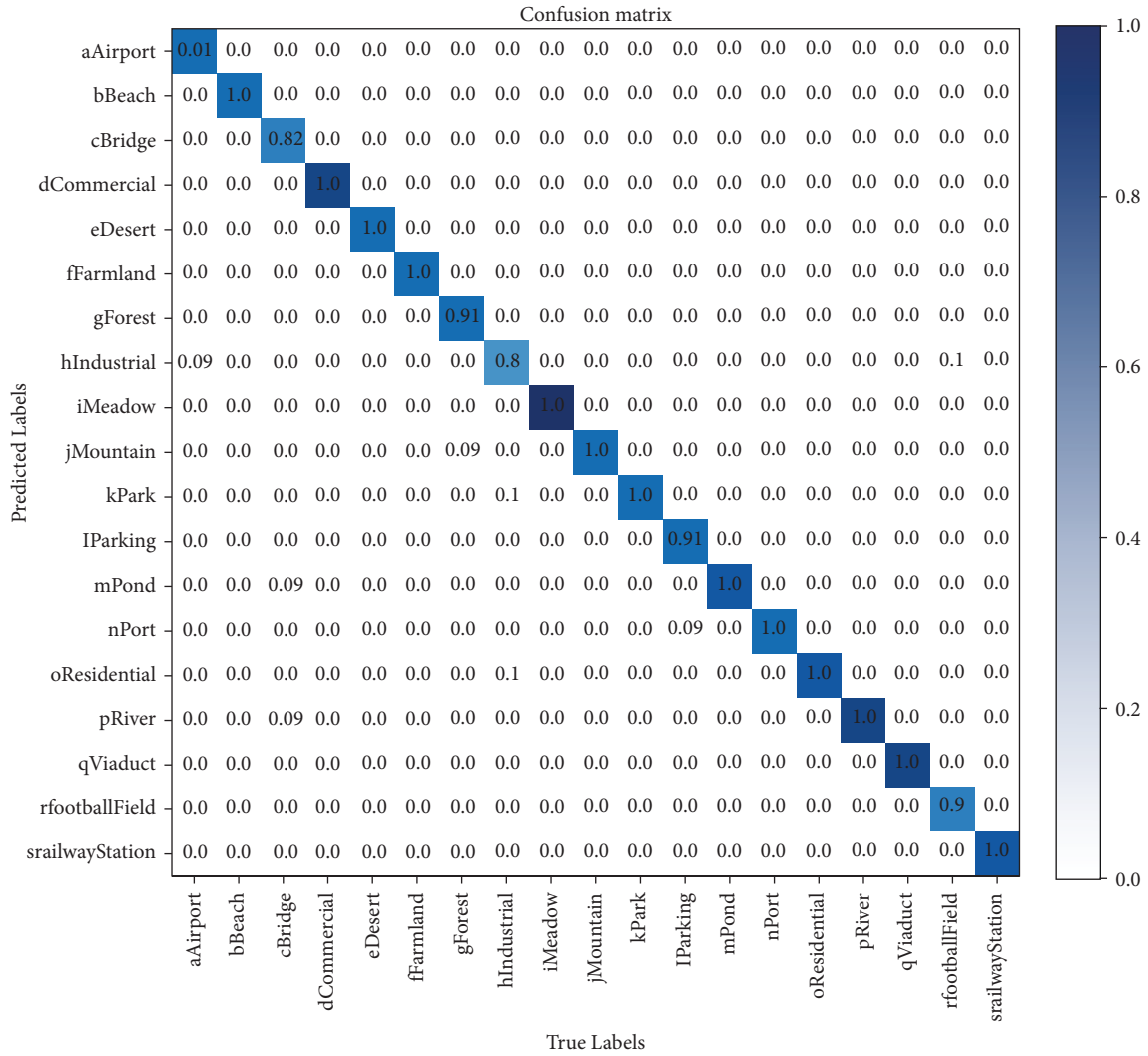


FIGURE 4: WHU-RS19 dataset classification confusion matrix.

TABLE 6: Overall accuracy and other parameters of the method on the RSSCN7 dataset.

Model	Accuracy (%)	Precision (%)	Recall (%)	Specificity (%)	F1-score (%)
ResNet50 [31]	94.46	94.59	94.09	99.09	94.49
Vgg16 [32]	93.75	93.79	93.76	98.99	93.71
GoogleNet [27]	93.57	93.61	93.57	98.93	93.56
DenseNet [33]	93.21	93.34	93.21	98.89	93.21
ViT-Ti [14]	90.89	90.89	90.89	98.49	90.89
Swin transformer [20]	93.93	93.96	93.91	99.00	93.93
VAN [17]	94.11	94.17	94.11	99.03	94.11
Conformer-Ti [15]	95.00	95.06	95.00	99.20	95.00
CMT [26]	94.82	95.06	94.83	99.14	94.81
MPViT [25]	95.00	95.03	95.00	99.19	95.00
CAW	<b>96.25</b>	<b>96.27</b>	<b>96.24</b>	<b>99.40</b>	<b>96.24</b>

TABLE 7: Overall accuracy and other parameters of the method on the WHU-RS19 dataset.

Model	Accuracy (%)	Precision (%)	Recall (%)	Specificity (%)	F1-score (%)
ResNet50 [31]	94.66	95.15	94.62	99.71	94.62
Vgg16 [32]	94.66	95.44	94.61	99.71	94.78
GoogleNet [27]	90.29	90.89	90.50	99.47	90.27
DenseNet [33]	95.15	96.08	95.14	99.73	95.34
ViT-Ti [14]	82.04	83.74	82.11	99.01	82.28
Swin transformer [20]	91.26	92.25	91.35	99.52	91.25
VAN [17]	93.67	94.44	93.72	99.66	93.59
Conformer-Ti [15]	95.63	95.75	95.54	99.76	95.55
CMT [26]	95.63	96.18	95.68	99.76	95.77
MPViT [25]	95.63	95.93	95.75	99.76	95.69
CAW	<b>96.12</b>	<b>96.23</b>	<b>96.03</b>	<b>99.79</b>	<b>95.96</b>

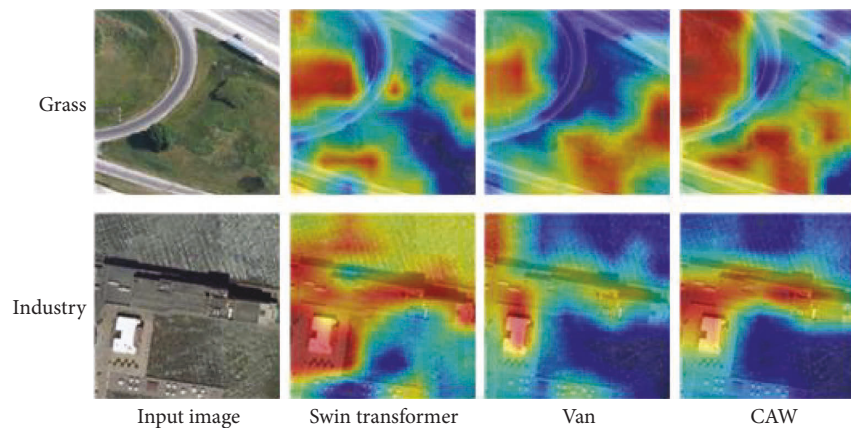


FIGURE 5: The visualization results (heat maps) of the CAW, Swin transformer, and Van models.

## 5. Conclusions

Aiming at the problems of large size changes of key objects, complex spatial arrangement, and object distribution in remote-sensing scene classification tasks, this paper proposes a parallel network model combining the local-window self-attention mechanism and equivalent large convolution kernel. The complementary parallel structure of Swin transformer and VAN realizes the space-channel modeling of transformer network structure with a small amount of parameters and calculation. A series of experiments on two challenging remote-sensing image scene classification datasets show that the network proposed in this paper has good remote-sensing image scene classification results.

In the follow-up work, we will further simplify the network structure and try to optimize the network performance by introducing some other attention mechanisms that can improve the network performance.

## 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 conflicts of interest.

## Acknowledgments

Key Research and Development Projects of Hunan Province (2020SK2134), Natural Science Foundation of Hunan Province (2019JJ80105, 2022JJ30625), Science and Technology Plan Project of Changsha (kq2004071), and Scientific research project of Hunan Provincial Department of Education (20C1249).

## References

- [1] W. Wang, Y. J. Yang, J. Li, Y. Hu, Y. Luo, and X. Wang, "Woodland Labeling in Chenzhou, China via deep learning approach," *International Journal of Computational Intelligence Systems*, vol. 13, no. 1, pp. 1393–1403, 2020.
- [2] X. Y. Tong, G. S. Xia, Q. Lu et al., "Land-cover classification with high-resolution remote sensing images using transferable deep models," *Remote Sensing of Environment*, vol. 237, Article ID 111322.
- [3] W. Wang, C. Tang, X. Wang, and B. Zheng, "A ViT-based multiscale feature fusion approach for remote sensing image segmentation," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, Article ID 4510305, 2022.
- [4] W. Wang, Y. Kang, G. Liu, and X. Wang, "SCU-net: semantic segmentation network for learning channel information on remote sensing images," *Computational Intelligence and Neuroscience*, vol. 2022, p. 11, Article ID 8469415, 2022.
- [5] Y. Zhang, K. Qin, Q. Bi, W. Cui, and G. Li, "Landscape patterns and building functions for urban land-use classification from remote sensing images at the block level: a case

- study of Wuchang District, Wuhan, China,” *Remote Sensing*, vol. 12, no. 11, p. 1831, 2020.
- [6] Y. Yang and S. Newsam, “Bag-of-visual-words and spatial extensions for land-use classification,” in *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pp. 270–279, San Jose, CA, USA, November 2010.
  - [7] S. Chaib, H. Liu, Y. Gu, and H. Yao, “Deep feature fusion for VHR remote sensing scene classification,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 8, pp. 4775–4784, 2017.
  - [8] Q. Zhu, Y. Zhong, B. Zhao, G. S. Xia, and L. Zhang, “Bag-of-visual-words scene classifier with local and global features for high spatial resolution remote sensing imagery,” *IEEE Geoscience and Remote Sensing Letters*, vol. 13, no. 6, pp. 747–751, 2016.
  - [9] X. X. Zhu, D. Tuia, L. Mou et al., “Deep learning in remote sensing: a comprehensive review and list of resources,” *IEEE Geoscience and Remote Sensing Magazine*, vol. 5, no. 4, pp. 8–36, 2017.
  - [10] G. S. Xia, J. Hu, F. Hu et al., “AID: a benchmark data set for performance evaluation of aerial scene classification,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 7, pp. 3965–3981, 2017.
  - [11] G. I. S. Geography, “Image Classification Techniques in Remote Sensing,” *GIS Geography Website*, <http://gisgeography.com/image-classification-techniques-remote-sensing>, 2020.
  - [12] F. Hu, G. S. Xia, J. Hu, and L. Zhang, “Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery,” *Remote Sensing*, vol. 7, no. 11, Article ID 14680, 2015.
  - [13] E. Li, J. Xia, P. Du, C. Lin, and A. Samat, “Integrating multilayer features of convolutional neural networks for remote sensing scene classification,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 10, pp. 5653–5665, 2017.
  - [14] A. Dosovitskiy, L. Beyer, A. Kolesnikov et al., “An Image Is worth 16x16 Words: Transformers for Image Recognition at Scale,” October 2020, <https://arxiv.org/abs/2010.11929>.
  - [15] Z. Peng, W. Huang, S. Gu et al., “Conformer: local features coupling global representations for visual recognition,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 367–376, IEEE, Montreal, QC, Canada, October 2021.
  - [16] Z. Dai, H. Liu, and Q. V. Le, “Coatnet: marrying convolution and attention for all data sizes,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 3965–3977, 2021.
  - [17] M. H. Guo, C. Z. Lu, and Z. N. Liu, “Visual Attention Network,” February 2022, <https://arxiv.org/abs/2202.09741>.
  - [18] X. Chu, Z. Tian, and Y. Wang, “Twins: revisiting the design of spatial attention in vision transformers,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 9355–9366, 2021.
  - [19] Y. Li, K. Zhang, and J. Cao, “Localvit: Bringing Locality to Vision Transformers,” April 2021, <https://arxiv.org/abs/2104.05707>.
  - [20] Z. Liu, Y. Lin, Y. Cao et al., “Swin transformer: hierarchical vision transformer using shifted windows,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10012–10022, IEEE, Montreal, QC, Canada, October 2021.
  - [21] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, “SqueezeNet: AlexNet-Level Accuracy with 50x Fewer Parameters And < 0.5 MB Model Size,” February 2016, <https://arxiv.org/abs/1602.07360>.
  - [22] X. Zhang, X. Zhou, M. Lin, and J. Sun, “Shufflenet: an extremely efficient convolutional neural network for mobile devices,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6848–6856, IEEE, Salt Lake City, USA, June 2018.
  - [23] A. G. Howard, M. Zhu, B. Chen et al., “Mobilenets: Efficient Convolutional Neural Networks for mobile Vision Applications,” April 2017, <https://arxiv.org/abs/1704.04861>.
  - [24] W. Wang, Y. Li, T. Zou, X. Wang, J. You, and Y. Luo, “A novel image classification approach via dense-MobileNet models,” *Mobile Information Systems*, vol. 2020, Article ID 7602384, 8 pages, 2020.
  - [25] Y. Lee, J. Kim, J. Willette, and S. J. Hwang, “MPViT: multi-path vision transformer for dense prediction,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7287–7296, IEEE, New Orleans, US, June 2022.
  - [26] J. Guo, K. Han, H. Wu et al., “Cmt: convolutional neural networks meet vision transformers,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, Article ID 12175.
  - [27] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, and A. Rabinovich, “Going deeper with convolutions,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition IEEE Computer Society (CVPR)*, pp. 1–9, IEEE, Boston, USA, June 2015.
  - [28] C. F. Chen, Q. Fan, N. Mallinar, T. Sercu, and R. Feris, “Big-little Net: An Efficient Multi-Scale Feature Representation for Visual and Speech Recognition,” July 2018, <https://arxiv.org/abs/1807.03848>.
  - [29] C. Si, W. Yu, P. Zhou, X. Wang, and S. Yan, “Inception Transformer,” May 2022, <https://arxiv.org/abs/2205.12956>.
  - [30] W. Wang, W. Huang, X. Wang, P. Zhang, and N. Zhang, “A COVID-19 CXR image recognition method based on MSA-DDCovNet,” *IET Image Processing*, vol. 16, no. 8, pp. 2101–2113, 2022.
  - [31] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, IEEE, Las Vegas, USA, June 2016.
  - [32] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” September 2014, <https://arxiv.org/abs/1409.1556>.
  - [33] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4700–4708, IEEE, Hawaii, USA, July 2017.
  - [34] Q. Zou, L. Ni, T. Zhang, and Q. Wang, “Deep learning based feature selection for remote sensing scene classification,” *IEEE Geoscience and Remote Sensing Letters*, vol. 12, no. 11, pp. 2321–2325, 2015.
  - [35] R. R. Selvaraju, M. Cogswell, and A. Das, “Grad-cam: visual explanations from deep networks via gradient-based localization,” in *Proceedings of the IEEE International Conference on Computer Vision (CVPR)*, pp. 618–626, Hawaii, USA, July 2017.
  - [36] W. Wang, X. Tan, P. Zhang, and X. Wang, “A CBAM based multiscale transformer fusion approach for remote sensing image change detection,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 6817–6825, 2022.

## Research Article

# Semisupervised Semantic Segmentation with Mutual Correction Learning

Yifan Xiao,<sup>1</sup> Jing Dong ,<sup>1</sup> Dongsheng Zhou ,<sup>1,2</sup> Pengfei Yi,<sup>1</sup> Rui Liu,<sup>1</sup>  
and Xiaopeng Wei <sup>2</sup>

<sup>1</sup>Key Laboratory of Advanced Design and Intelligent Computing, Ministry of Education, Dalian University, Dalian 116622, China

<sup>2</sup>School of Computer Science and Technology, Dalian University of Technology, Dalian 116024, China

Correspondence should be addressed to Jing Dong; [dongjing@dlu.edu.cn](mailto:dongjing@dlu.edu.cn) and Xiaopeng Wei; [xpwei@dlut.edu.cn](mailto:xpwei@dlut.edu.cn)

Received 11 July 2022; Revised 24 August 2022; Accepted 1 September 2022; Published 3 October 2022

Academic Editor: Nian Zhang

Copyright © 2022 Yifan Xiao et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The semisupervised semantic segmentation method uses unlabeled data to effectively reduce the required labeled data, and the pseudo supervision performance is greatly influenced by pseudo labels. Therefore, we propose a semisupervised semantic segmentation method based on mutual correction learning, which effectively corrects the wrong convergence direction of pseudo supervision. The well-calibrated segmentation confidence maps are generated through the multiscale feature fusion attention mechanism module. More importantly, using internal knowledge, a mutual correction mechanism based on consistency regularization is proposed to correct the convergence direction of pseudo labels during cross pseudo supervision. The multiscale feature fusion attention mechanism module and mutual correction learning improve the accuracy of the entire learning process. Experiments show that the MIoU (mean intersection over union) reaches 75.32%, 77.80%, 78.95%, and 79.16% using 1/16, 1/8, 1/4, and 1/2 labeled data on PASCAL VOC 2012. The results show that the new approach achieves an advanced level.

## 1. Introduction

As a fundamental task, semantic segmentation is widely used in medical image diagnosis [1], automatic driving [2], and other fields, which is the process of defining the boundaries between the various semantic entities in an image. From a technical point of view, each pixel in the image is assigned a category or semantic label. With the development of deep learning, fully supervised semantic segmentations [3–7] achieve success, but they all need enough pixel-level labels to complete the representation learning, which requires a lot of manpower.

Weakly supervised and semisupervised semantic segmentation effectively reduces the annotation burden. Weakly supervised methods use weak annotations as labels to train segmentation models. Semisupervised methods combine additional unlabeled data with a small amount of labeled data to improve segmentation model performance and close the gap with supervised models trained from fully pixel-labeled data. How to use unlabeled data for training

models to get good segmentation performance is a problem we need to solve.

In semisupervised semantic segmentation, the methods are mainly based on adversarial learning [8–10] and consistency regularization [11, 12]. The generative adversarial network (GAN)-based approach [8] proposed a full convolution discriminator, which can learn to distinguish the ground truth and the output of the generator, enhancing the consistency between the predicted maps of the segmentation network and the ground truth. Consistency regularization enforces the prediction consistency of perturbations by increasing the input image perturbation [11, 12], the feature perturbation [13], and the network perturbation [14] to make the prediction consistent among the output of multiple perturbations. Chen et al. [15] proposed the cross pseudo supervision loss, in which unlabeled data were input into two segmentation networks with different initializations to generate pseudo labels for cross supervision and strengthen the consistency of the model.

However, the cross pseudo supervision still has two drawbacks. First, the segmentation network generates inaccurate pseudo labels to guide model learning, which damages the model accuracy, and pseudo labels are directly generated by the confidence segmentation maps of unlabeled images, completely ignoring the ability of the network itself to improve pseudo labels. Second, the cross pseudo supervision is plagued by confirmation bias and tends to overfitting pseudo labels that are incorrectly predicted. After one segmentation network predicts the wrong label output, the cross pseudo supervision trains the other model with wrong knowledge, thus hindering the cross learning of the model.

To address the above two problems, we propose a new semisupervised semantic segmentation method based on cross pseudo supervision. Many works combine consistency regularization with pseudo labels, our proposed method also includes pseudo labels [16–18] and utilizes pseudo segmentation maps to enhance consistency. To address the first problem, we introduce the multiscale feature fusion attention mechanism module [19] to generate well-calibrated segmentation confidence maps, and the multiscale feature fusion attention mechanism mode fuses high-level feature maps and low-level feature maps to generate segmentation confidence maps with higher quality. To address the second problem, we propose mutual correction learning to improve the model convergence in the wrong direction caused by pseudo labels. The mutual correction loss uses the internal knowledge of pseudo labels for mutual correction, which not only strengthens the consistency of the network but also corrects the learning direction of the model. In this way, the segmentation performance of consistency training is greatly improved. To sum up, our two-fold contributions are as follows:

- (i) We propose an effective module to generate better quality segmentation confidence maps by fusing low-level texture information and high-level semantic information of the features.
- (ii) We propose mutual correction learning for semisupervised semantic segmentation, which uses the intrinsic knowledge to correct the convergence direction of the model and effectively ameliorates the problem of model performance degradation by erroneous cross pseudo supervision.

The rest of this article is arranged as follows: The second section introduces the related work of semisupervised semantic segmentation. In the approach section, we describe the details of mutual correction learning with pseudo labels. The experimental details and results are presented in the experiment section. In the conclusion section, we summarize this paper.

## 1.1. Related Work

*1.1.1. Fully Supervised Semantic Segmentation.* Fully convolutional networks (FCNs) [3] can accept input images of any size, and the deconvolution layer is used to perform upsampling of the feature map of the last convolution layer and predict each pixel. Although high-level features contain rich semantic information, they cannot capture long-term

relationships well. Therefore, global pooling [4], dilated convolution [5], pyramid pooling [6], and attention mechanisms [7] are used to better aggregate context. Deeplabv3+ [20] fuses features of different scales to refine the object boundaries of the segmentation results. However, training supervised segmentation networks requires a large amount of labeled data, which is expensive to collect. Our work alleviates the constraints of annotated data by making efficient use of unlabeled data. To make a fair comparison with previous works, we use Deeplabv3+ as the backbone architecture.

*1.1.2. Weakly Supervised Semantic Segmentation.* Weakly supervision is to further reduce the cost of data annotation based on full supervision. Some early works use weak annotations such as bounding boxes [21–23], scribbles [24], and image-level labels [25–28]. The recent methods use object location information to generate pseudo pixel annotations and train the segmentation network, and their segmentation performance is significantly improved. Al-Huda et al. [26] fused activation maps and saliency maps to guide the model to generate initial pixel-level annotations and generate more accurate pixel labeling through iteration. Although promising results have been obtained using the above methods, most of them require additional training strategies. Al-Huda et al. [28] proposed a new postprocessing method, which learned the concept of the object scale from the intermediate features of hierarchical structure through dynamic programming and further improved the segmentation accuracy.

*1.1.3. Semisupervised Semantic Segmentation.* The semisupervised method is based on incomplete supervised learning, using partially labeled data and unlabeled data for model training. The semisupervised semantic segmentation method is mainly based on the idea of consistent regularization and pseudo labeling.

Consistency regularization enforces the model to make consistent predictions concerning various perturbations. Its effectiveness is based on the smoothing assumption or the cluster assumption. These assumptions consider that data pointing close to each other are likely from the same class, which are often used in classification tasks [29, 30]. As for semantic segmentation tasks, French and Ouali found that semantic segmentation tasks do not fully comply with the clustering assumption in [11, 13]. Therefore, Ouali et al. [13] proposed to perturb the output of the encoder while maintaining the clustering assumption and used multiple auxiliary decoders to obtain a consistent prediction. French et al. [11] found that mask-based enhancement strategies were effective and introduced data enhancement technology CutMix [31]. The idea of CutMix is to mix samples by replacing part region of the image with a patch from another image and treat it as an extension of Cutout [32] and Mixup [33]. Cross consistency training (CCT) [13] used shared encoders and multiple decoders as segmentation networks, and the prediction using different decoders enhanced consistency. Mittal et al. [9] proposed a dual-branch method for semisupervised semantic segmentation, the GAN-based model solved the inaccuracy of low-level details, and the

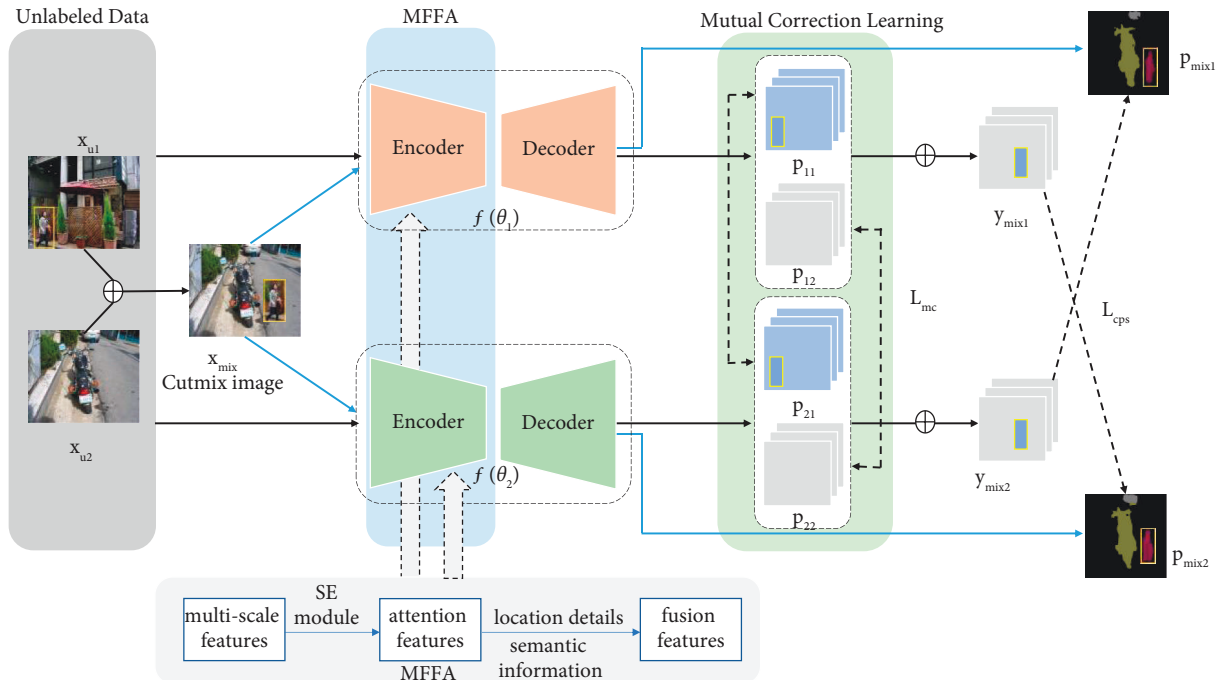


FIGURE 1: Overview of mutual correction learning. Two images  $x_{u1}$  and  $x_{u2}$  are sampled from the unlabeled dataset. The CutMix images are generated by two source images, and they are all inputted into each segmentation network.  $p_{i1}$  and  $p_{i2}$  are mixed as pseudo segmentation maps  $y_{mixi}$  to supervise the other segmentation network.  $\oplus$ : CutMix, MFFA: multiscale feature fusion attention mechanism module,  $\mathcal{L}_{mc}$ : mutual correction loss,  $\mathcal{L}_{cps}$ : cross pseudo supervision loss,  $p$ : segmentation confidence map,  $y_{mixi}$ : predicted one-hot label map, and SE module: squeeze-and-excitation module.

semisupervised multilabel classification model corrected the misunderstanding of high-level information. Lai et al. [34] proposed different contexts in the same area to enhance the consistency of context awareness. Guided collaborative training (GCT) [35] further used different initialization segmentation networks to enhance the consistency of disturbed network prediction. Our approach combines the ideas of CutMix [31] and cross pseudo supervision (CPS) [15] to enhance the consistency between mixed output and mixed input prediction.

Pseudo labeling is a technique that utilizes unlabeled data through feature learning and alternating pseudo label prediction [16–18]. Its main goal is entropy minimization, and it encourages the network to make confident predictions of unlabeled images and prevents features from being trained to the wrong class. Chen et al. [17] proposed a new two-branch network in which the pseudo network extracted the correct pseudo labels as auxiliary supervised information for the training segmentation network. Zhou et al. [18] proposed a pseudo label enhancement strategy to improve the quality of pseudo labels. The key to pseudo labeling is the quality of pseudo labels. Most models [36, 37] refine pseudo labels from external guidance, such as teachers. However, the teacher model is often fixed, making the student inherit some inaccurate predictions from the teacher. In order to generate better pseudo labels, the recent approach is to update both the teacher and student models, such as coteaching [38], dual students [14], and metapseudo labels [39]. Furthermore, it is essential that the model converges in the right direction at

the beginning of training. In the third section, mutual correction learning is used to correct the convergence direction of the model.

*1.2. Approach.* Semisupervised semantic segmentation uses labeled images  $D_l = \{x_l, y^*\}$  and unlabeled images  $D_u = \{x_u\}$  to learn a segmentation network.  $x \in R^{H \times W \times 3}$  denotes the images with a resolution of  $H \times W$ ,  $y^* \in R^{H \times W \times K}$  is the ground truth corresponding to  $x_l$  with pixels labeled by  $K$  classes, and  $f$  is a segmentation network with a weight of  $\theta$ .

The approach proposed in the paper is shown in Figure 1. The mutual correction learning model consists of two parallel segmentation networks.  $f(\theta_1)$  and  $f(\theta_2)$  are the same segmentation networks with different initialization. The network inputs are  $x_{u1}$ ,  $x_{u2}$ , and  $x_{mix}$ , unlabeled images  $x_{u1}$  and  $x_{u2}$  are with the same augmentation, and  $x_{mix}$  is obtained through CutMix [31] by (1), where  $M \in (0, 1)^{W \times H}$  is binary coding and represents the position of removing and filling from two images:

$$x_{mix} = M \odot x_{u1} + (1 - M) \odot x_{u2}. \quad (1)$$

$p$  is the segmentation confidence map obtained after softmax normalization. The output structure with a weight  $\theta_2$  is the same as the output with  $\theta_1$ .

$$p_{11} = f(x_{u1}; \theta_1), \quad (2)$$

$$p_{12} = f(x_{u2}; \theta_1), \quad (3)$$

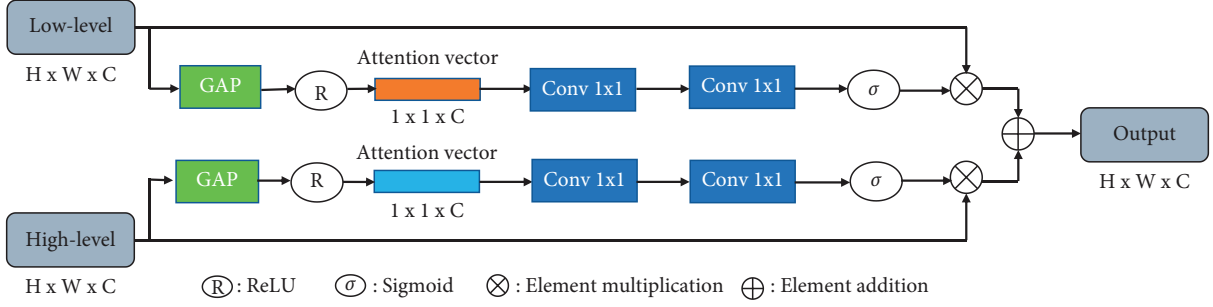


FIGURE 2: MFFA: multiscale feature fusion attention mechanism module.  $\sigma$ : sigmoid,  $\otimes$ : element multiplication, and  $\oplus$ : element addition.

$$p_{\text{mix1}} = f(x_{\text{mix}}; \theta_1), \quad (4)$$

$y$  is the predicted pseudo label. At each position  $i$ , the pseudo label  $y$  is the one-hot map computed by the segmentation confidence map  $p$ , and the value of  $M$  in (5) is the same as that in Eq. (1).

$$y_{\text{mix1}} = M \odot p_{11} + (1 - M) \odot p_{12}. \quad (5)$$

**1.2.1. Multiscale Feature Fusion Attention Mechanism Module.** Since generating pseudo labels with rich semantic information requires multiple convolution operations to continuously extract features, the dimension of features continues to expand, resulting in serious high-dimensional information redundancy. When all channel features are fused, the importance of features in each channel is not considered. Hence, Hu et al. [19] proposed the squeeze-and-excitation (SE) module for the adaptive fusion of channel features to reduce the redundancy of high-dimensional features.

This paper introduces the multiscale feature fusion attention mechanism module to fuse high-level and low-level feature maps. The attention mechanism uses two SE modules to extract different attention features from low-level features to high-level features, as shown in Figure 2. The module contains location details in low-level features and semantic information in high-level features to improve the accuracy of the prediction of different target boundaries.

In (6), the role of the global mean pooling (GAP) layer is to integrate global spatial information. It takes the feature map as input to obtain a feature vector containing semantic correlation. The attention vector is obtained by (7), and the output  $\hat{x}$  of the encoder is generated by eq (8).

$$g(x_k) = \frac{1}{W \times H} \sum_{i=1}^H \sum_{j=1}^W x_k(i, j), \quad (6)$$

where  $k = 1, 2, 3 \dots d$ ,  $d$  is channel dimensions, and  $x_k$  is the channel input of the module.

$$A_c = \delta_2 [\delta_1 [g(x) + b_\alpha] + b_\beta]. \quad (7)$$

$x = [x_1, x_2, \dots, x_d]$ ,  $g$  is the GAP layer,  $\delta_1$  and  $\delta_2$  are activation functions ReLU and sigmoid, respectively, and  $b_\alpha$  and  $b_\beta$  are the bias.

$$\hat{x} = A_c \otimes x. \quad (8)$$

The output of the encoder is the sum of low-level  $\hat{x}_l$  and high-level  $\hat{x}_h$ , which is decoded to obtain the segmentation confidence map.

**1.2.2. Mutual Correction Learning.** The two segmentation networks have different learning capabilities after different initialization, and they can learn online from the pseudo labels generated by each other. In the training process, if the segmentation network  $f(\theta_1)$  generates poor quality one-hot labels  $y_{\text{mix1}}$ , the segmentation network  $f(\theta_2)$  produces a good quality confidence map  $p_{\text{mix2}}$ , and the model may converge in the wrong direction guided by the poor quality label; the self-correction ability of the cross pseudo supervision is limited, which may degrade the performance of the model.

In order to prevent the model from converging in the wrong direction, we propose the mutual correction loss to correct this problem, and the training objectives include three losses: supervision loss  $\mathcal{L}_s$ , mutual correction loss  $\mathcal{L}_{mc}$ , and cross pseudo supervision loss  $\mathcal{L}_{cps}$ . The supervision loss is not marked in the network structure diagram.

$\mathcal{L}_s$ : the labeled image  $x_l$  does not require CutMix and is input into the two networks for supervised learning. The supervision loss  $\mathcal{L}_s$  can be written as follows:

$$\mathcal{L}_s = \frac{1}{|N_l|} \sum_{D_l} (1/W \times H) \sum_{i=1}^{W \times H} (\ell_{ce}(p_1^i, y_1^{*i}) + \ell_{ce}(p_2^i, y_2^{*i})). \quad (9)$$

$N_l$  represents the number of labeled images in a batch, and  $W$  and  $H$  represent the width and height of the input image.  $\ell_{ce}$  is the standard cross entropy loss function and  $y_1^{*i}(y_2^{*i})$  is the ground truth.

$\mathcal{L}_{mc}$ : we propose a mutual correction loss  $\mathcal{L}_{mc}$  to make the model have the ability to self-correct. Unlabeled images  $x_{u1}$  and  $x_{u2}$  are input to the network  $f(\theta_1)$  and  $f(\theta_2)$ , respectively, to produce the corresponding confidence maps  $p_{11}, p_{12}$  and  $p_{21}, p_{22}$ . Cross entropy describes the difficulty of expressing probability distributions  $p_{11}$  ( $p_{12}$ ) through probability distributions  $p_{21}$  ( $p_{22}$ ). The smaller the value of cross entropy is, the closer the two probability distributions are. According to the consistency principle, the confidence map similarity of  $p_{11}$  and  $p_{21}$  should be higher. In other

words, the loss between  $(p_{11}, p_{21})$  and  $(p_{12}, p_{22})$  should be as small as possible, so the mutual correction loss  $\mathcal{L}_{mc}$  can be written in the following form:

$$\mathcal{L}_{mc} = \frac{1}{|N_u|} \sum_{D_u} (1/W \times H) \sum_{i=1}^{W \times H} (\ell_{ce}(p_{11}^i, p_{21}^i) + \ell_{ce}(p_{12}^i, p_{22}^i)). \quad (10)$$

$\mathcal{L}_{cps}$  [15]: The cross pseudo supervision loss is symmetric, and the pseudo label  $y_{mix1}$  is used to supervise the confidence map  $p_{mix2}$  generated by another network, and the other one uses the pseudo label  $y_{mix2}$  to supervise the confidence map  $p_{mix1}$ . The cross pseudo supervision loss  $\mathcal{L}_{cps}$  can be written in the following form:

$$\mathcal{L}_{cps} = \frac{1}{|N_u|} \sum_{D_u} (1/W \times H) \sum_{i=1}^{W \times H} (\ell_{ce}(p_{mix1}^i, y_{mix2}^i) + \ell_{ce}(p_{mix2}^i, y_{mix1}^i)). \quad (11)$$

When training the segmentation network, we use multiple loss constraints on the segmentation network and minimize them for tuning.  $\gamma$  and  $\lambda$  are the hyperparameters set by the experiment, and the loss function of the whole training can be written as follows:

$$\mathcal{L} = \mathcal{L}_s + \gamma \mathcal{L}_{mc} + \lambda \mathcal{L}_{cps}. \quad (12)$$

### 1.3. Experiments

**1.3.1. Datasets.** PASCAL VOC 2012 [40] is the most widely used benchmark dataset for semantic segmentation tasks. Pascal VOC 2012 training set used in this paper contains 10,582 images and annotations, and the validation set contains 1449 images and annotations. PASCAL VOC has a total of 20 categories, such as aircraft, bicycles, birds, and boats.

Cityscapes [41] contains tagged images of urban street scenes taken from vehicles driven in European cities, specifically for semantic understanding of urban street scenes. It has 19 category tags and contains 5000 finely labeled images, including 2975 images for network training, 500 images for network verification, and 1525 images for testing. In addition, we only used the fine annotation graph for training.

Following the division protocol of GCT [35], the entire training set was randomly divided into two groups, with 1/2 (5291), 1/4 (2646), 1/8 (1323), and 1/16 (662) of the whole training set as the labeled set.

**1.3.2. Evaluation Metrics.** Mean intersection over union (MIoU) is a common evaluation metric in semantic segmentation. In (13), where  $TP_c$ ,  $FP_c$ , and  $FN_c$  represent the prediction results of true positive, false positive, and false negative of category  $c$ ,  $C$  represents the total number of categories.

$$\text{MIoU} = \frac{1}{C} \sum_{c=1}^C (TP_c / TP_c + FP_c + FN_c). \quad (13)$$

For all experiments, we used only one network for inferential prediction, testing the results of the 1456 PASCAL VOC 2012 value set (or 500 Cityscapes value set).

**1.3.3. Implementation Details.** The PyTorch deep learning framework was used to complete the proposed method and related experiments. We used ResNet-101 pretrained on ImageNet as backbone and SyncBN [42] for training. Our method set weight decay as 0.0005 and momentum as 0.9. The loss weights  $\gamma$  and  $\lambda$  are 1 and 1.5 on PASCAL VOC and 1.5 and 6 on cityscapes. We used a multiple learning rate strategy, and the initial learning rate values were set to 0.0025 for PASCAL VOC, while 0.02 for Cityscapes.

**1.3.4. Comparison with Other Methods.** In Figure 3, the improvements of this method are shown under different label proportions. All methods are based on DeepLabv3+.

Figure 3(a) shows that our approach using ResNet-50 consistently outperforms the supervised baseline approach on PASCAL VOC 2012. The improvements of our method over the baseline method are 8.28%, 6.80%, 4.23% and 3.28% under 1/16, 1/8, 1/4, and 1/2 scale settings separately. Figure 3(b) shows that our method uses ResNet-101 for 8.45%, 6.26%, 5.26%, and 4.94% lift at 1/16, 1/8, 1/4, and 1/2 scale settings, respectively.

We compared our method with some recent semi-supervised segmentation methods, including cross consistency training (CCT) [13], guided collaborative training (GCT) [35], context-aware consistency (CAC) [34], and cross pseudo supervision (CPS) [15] under different segmentation protocols. Table 1 shows the experimental comparison results on PASCAL VOC 2012. In different scale settings, our method is superior to other methods, whether ResNet-50 or ResNet-101. Especially in 1/8 and 1/4 proportions, it was 1.43% and 1.20% and 1.36% and 1.27% higher than CPS, respectively.

We further verified the effectiveness of the proposed method by comparing with other methods on Cityscapes in Table 2. Compared with CCT and GCT, the accuracy of our method is greatly improved with a small number of labeled images, especially in the case of 1/16. The main reason for the low improvement on Cityscapes results compared to PASCAL VOC 2012 is that PASCAL VOC 2012 is an object-centered semantically segmentation dataset with an average of three objects per image. Cityscapes is a highly complex urban street scene, and the resolution and scene complexity of each picture are much higher than those of PASCAL VOC 2012, which will lead to the inclusion of more complex information in the mutual correction learning and weaken the ability of mutual correction. Therefore, our method is more suitable for each dataset with fewer graph object instances.



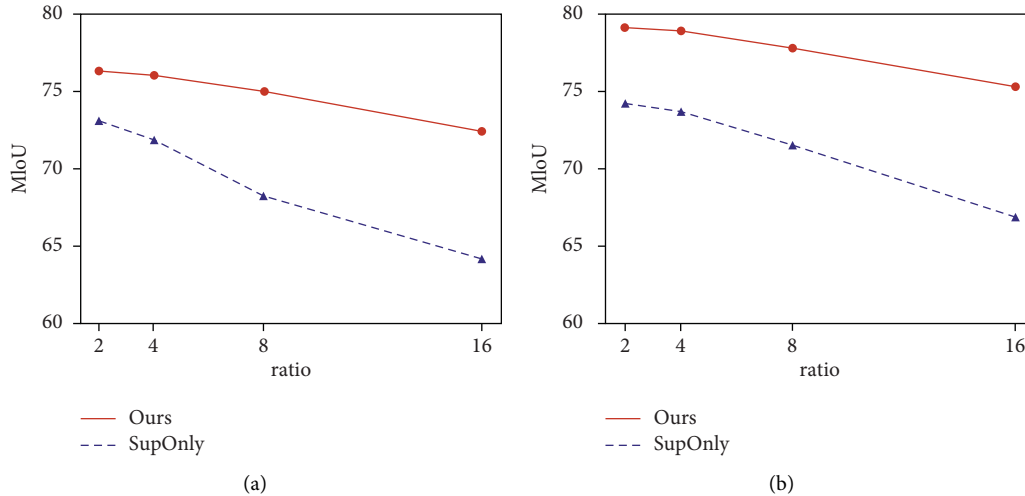


FIGURE 3: Comparison with SupOnly in PASCAL VOC 2012 (1/2, 1/4, 1/8, 1/16). (a) ResNet-50. (b) ResNet-101.

TABLE 1: Comparison with other methods on PASCAL VOC 2012 under different partition protocols. The segmentation network is DeepLabv3+. SupOnly represents supervised training, using only labeled data.

Method	ResNet-50			
	1/16 (662)	1/8 (1323)	1/4 (2646)	1/2 (5291)
SupOnly	64.20	68.30	71.87	73.12
CCT [13]	65.22	70.87	73.43	74.75
GCT [35]	64.05	70.47	73.45	75.20
CAC [34]	70.10	72.40	74.00	76.50
CPS [15]	71.98	73.67	74.90	76.15
Ours	<b>72.48</b>	<b>75.10</b>	<b>76.10</b>	<b>76.40</b>

Method	ResNet-101			
	1/16 (662)	1/8 (1323)	1/4 (2646)	1/2 (5291)
SupOnly	66.87	71.54	73.69	74.22
CCT [13]	67.94	73.00	76.17	77.56
GCT [35]	69.77	73.30	75.25	77.14
CAC [34]	72.40	74.60	76.30	78.20
CPS [15]	74.48	76.44	77.68	78.64
Ours	<b>75.32</b>	<b>77.80</b>	<b>78.95</b>	<b>79.16</b>

The meaning of the bold values represent the best results.

TABLE 2: Comparison with other methods on Cityscapes under different partition protocols. The segmentation network is DeepLabv3+, and the backbone is ResNet-50. SupOnly represents supervised training, using only labeled data.

Method	ResNet-50			
	1/16 (186)	1/8 (372)	1/4 (744)	1/2 (1488)
CCT [13]	66.35	72.46	75.68	76.78
GCT [35]	65.81	71.33	75.30	77.09
CAC [34]	—	69.70	72.70	—
CPS [15]	74.47	76.61	77.83	78.77
Ours	<b>74.47</b>	<b>76.75</b>	<b>78.03</b>	<b>78.89</b>

The meaning of the bold values represent the best results.

**1.3.5. Ablation Study.** The ablation study in Table 3 shows the contribution of each function. The ablation study was based on PASCAL VOC 2012 with 1/8 labeled data.

TABLE 3: Ablation study.  $\mathcal{L}_s$ : supervised loss.  $\mathcal{L}_{cps}$ : cross pseudo supervised loss. *MFFA*: multiscale feature fusion attention mechanism module.  $\mathcal{L}_{mc}$ : mutual correction loss.

ID	$\mathcal{L}_s$	$\mathcal{L}_{cps}$	<i>MFFA</i>	$\mathcal{L}_{mc}$	MIoU
1	✓				68.30
2	✓	✓			73.67
3	✓	✓	✓		74.03
4	✓	✓		✓	74.28
5	✓	✓	✓	✓	75.10

DeepLabv3+ and ResNet-50 were the segmentation networks. The supervised loss training (SupOnly) model was used as the benchmark of our work.

In Table 3, ID 2 shows the performance improvement with cross pseudo supervised losses, with 5.37% MIoU improvement on PASCAL VOC 2012 compared to ID 1 with supervised losses alone.

In order to prove the validity of the multiscale feature fusion attention mechanism module, we made a comparison between the model with MFFA and the model with the cross pseudo supervised loss. Features of different scales combine rich localization and semantic information to generate accurate segmentation maps of boundary information. ID 2 and ID 3 showed that the model with the MFFA module improved by 0.36% compared with the model with cross pseudo supervision loss. In addition, ID 4 and ID 5 found that MFFA improved by 0.82%.

In ID 2 and ID 4, the effectiveness of the mutual correction loss was compared with that of the supervised loss and cross pseudo supervised loss. The cross pseudo supervision learns the error information and corrects it effectively through the mutual correction loss, and MIoU shows an increase of 0.61%. ID 3 and ID 5 found that the mutual correction loss increased MIoU by 1.07% while using the MFFA module. According to ID 5, MIoU improved by 1.43% with the multiscale feature fusion attention mechanism module and mutual correction loss.



FIGURE 4: Example qualitative results from PASCAL VOC 2012. All the approaches are trained under 1/8 with ResNet-101 as the backbone: (a) input; (b) ground truth; (c) CPS; (d) ours.

**1.3.6. Qualitative Results.** Figure 4 shows the results of different methods on PASCAL VOC 2012. The actual labels are shown in column (b), CPS (column (c)), and predicted boundary errors, and our method corrects these problems in column (d). Obviously, mutual correction learning can more accurately predict the edges and categories of objects, thus improving the feature representation of the model.

**1.3.7. Limitations.** When the output predictions of the two segmentation networks are both wrong, the error correction of the mutual correction learning is limited. The results also show that our approach is influenced by the distribution of long-tail classes on semantic segmentation datasets, which

makes pseudo labels biased towards majority classes, and we will continue to study it and improve further.

## 2. Conclusion

We propose a semisupervised semantic segmentation approach based on mutual correction learning. The MFFA module is introduced to generate confidence maps, which in turn yield well-calibrated pseudo labels. To alleviate the problem of poor quality pseudo labels guiding the model to learn misinformation, we propose a mutual correction loss, utilizing the internal knowledge to correct the convergence direction of the model. Experiments show our approach further narrows the gap between fully supervised and semisupervised semantic segmentation.

## Data Availability

Previously reported PASCAL VOC 2012 and Cityscapes datasets were used to support this study and are available at DOI: <https://doi.org/10.1007/s11263-014-0733-5> and DOI: <https://doi.org/10.1109/cvpr.2016.350>. These prior studies (and datasets) are cited at relevant places within the text as references [40, 41].

## Conflicts of Interest

The authors declare that there are no conflicts of interest.

## Acknowledgments

This work was supported in part by the Key Program of NSFC (Grant no. U1908214), the Program for Innovative Research Team in University of Liaoning Province (LT2020015), the Support Plan for Key Field Innovation Team of Dalian (2021RT06), the Science and Technology Innovation Fund of Dalian (Grant no. 2020JJ25CY001), and the Dalian University Scientific Research Platform Project (No. 202101YB03).

## References

- [1] Z. Zhang, T. Zhao, H. Gay, W. Zhang, and B. Sun, "Semi-supervised semantic segmentation of prostate and organs-at-risk on 3d pelvic ct images," *Biomedical Physics & Engineering Express*, vol. 7, no. 6, Article ID 065023, 2021.
- [2] Q. Sellat, S. Bisoy, R. Priyadarshini, A. Vidyarthi, S. Kautish, and R. K. Barik, "Intelligent semantic segmentation for self-driving vehicles using deep learning," *Computational Intelligence and Neuroscience*, vol. 2022, Article ID 6390260, 10 pages, 2022.
- [3] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 39, pp. 3431–3440, 2015.
- [4] X. Lian, Y. Pang, J. Han, and J. Pan, "Cascaded hierarchical atrous spatial pyramid pooling module for semantic segmentation," *Pattern Recognition*, vol. 110, pp. 1–13, 2021.
- [5] S. Tang, J. Xia, L. Fan, X. Lei, W. Xu, and A. Nallanathan, "Dilated Convolution Based Csi Feedback Compression for Massive MIMO Systems," *IEEE Transactions on Vehicular Technology*, 2022.
- [6] C. Dewi, R.-C. Chen, H. Yu, and X. Jiang, "Robust detection method for improving small traffic sign recognition based on spatial pyramid pooling," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–18, 2021.
- [7] S. Chen, M. Zhang, X. Yang, Z. Zhao, T. Zou, and X. Sun, "The impact of attention mechanisms on speech emotion recognition," *Sensors*, vol. 21, no. 22, pp. 7530–7550, 2021.
- [8] W. Hung, Y. Tsai, Y. Liou, L. Yen-Yu, and Y. Ming-Hsuan, "Adversarial learning for semi-supervised semantic segmentation," in *British Machine Vision Conference 2018BMVC 2018*, Newcastle, UK, 2018.
- [9] S. Mittal, M. Tatarchenko, and T. Brox, "Semi-supervised semantic segmentation with high- and low-level consistency," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 4, pp. 1369–1379, 2021.
- [10] M. Qi, Y. Wang, J. Qin, and A. Li, "Ke-gan: knowledge embedded generative adversarial networks for semi-supervised scene parsing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5237–5246, Long Beach, CA, USA, 2019.
- [11] G. French, S. Laine, T. Aila, and M. Michal, "Semi-supervised semantic segmentation needs strong, varied perturbations," in *31st British Machine Vision Conference 2020Virtual Event*, UK, 2020.
- [12] J. Kim, J. Jang, and H. Park, "Structured Consistency Loss for Semi-supervised Semantic Segmentation," 2020, <https://arxiv.org/abs/2001.04647>.
- [13] Y. Ouali, C. Hudelot, and M. Tami, "Semi-supervised semantic segmentation with cross-consistency training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12674–12684, Gif-sur-Yvette, France, 2020.
- [14] Z. Ke, D. Wang, Q. Yan, J. Ren, and R. Lau, "Dual student: breaking the limits of the teacher in semi-supervised learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6728–6736, Seoul, Korea, October 2019.
- [15] X. Chen, Y. Yuan, G. Zeng, and J. Wang, "Semi-supervised semantic segmentation with cross pseudo supervision," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2613–2622, Nashville, TN, USA, June 2021.
- [16] Z. Feng, Q. Zhou, G. Cheng, T. Xin, S. Jianping, and M. Lizhuang, "Semi-supervised semantic segmentation via dynamic self-training and classbalanced curriculum," 2020, <https://arXiv.org/abs/2004.08514>.
- [17] Z. Chen, R. Zhang, G. Zhang, Z. Ma, and T. Lei, "Digging into pseudo label: a low-budget approach for semi-supervised semantic segmentation," *IEEE Access*, vol. 8, 2020.
- [18] Y. Zhou, R. Jiao, D. Wang, J. Mu, and J. Li, "Catastrophic forgetting problem in semi-supervised semantic segmentation," *IEEE Access*, vol. 10, pp. 48855–48864, 2022.
- [19] 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, June 2018.
- [20] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *European Conference on Computer Vision* Springer, Cham, New York, NY, USA, 2018.
- [21] J. Dai, K. He, and J. Sun, "Boxsup: exploiting bounding boxes to supervise convolutional networks for semantic segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1635–1643, Santiago, Chile, December 2015.
- [22] J. Wang and B. Xia, "Bounding box tightness prior for weakly supervised image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* Springer, Cham, New York, NY, USA, 2021.
- [23] G. Papandreou, L.-C. Chen, K. P. Murphy, and A. L. Yuille, "Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1742–1750, Santiago, Chile, December 2015.
- [24] Z. Al-Huda, D. Zhai, Y. Yang, and R. N. A. Algburi, "Optimal scale of hierarchical image segmentation with scribbles guidance for weakly supervised semantic segmentation," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 35, no. 10, Article ID 2154026, 2021.
- [25] Z. Jiang, W. He, M. S. Kirby et al., "Weakly supervised spatial deep learning for earth image segmentation based on

- imperfect polyline labels,” *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 13, no. 2, pp. 1–20, 2022.
- [26] Z. Al-Huda, B. Peng, Y. Yang et al., “Weakly supervised semantic segmentation by iteratively refining optimal segmentation with deep cues guidance,” *Neural Computing & Applications*, vol. 33, no. 15, pp. 9035–9060, 2021.
- [27] R. Dorent, S. Joutard, J. Shapey, A. Kujawa, M. Modat, and S. Ourselin, “Inter extreme points geodesics for end-to-end weakly supervised image segmentation,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention* Springer, Cham, New York, NY, USA, 2021.
- [28] Z. Al-Huda, B. Peng, Y. Yang, and R. N. A. Algburi, “Object scale selection of hierarchical image segmentation with deep seeds,” *IET Image Processing*, vol. 15, no. 1, pp. 191–205, 2021.
- [29] Q. Xie, Z. Dai, E. H. Hovy, L. Minh-Thang, and V. L. Quoc, “Unsupervised data augmentation for consistency training,” in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Vancouver, BC, Canada, December 2020.
- [30] K. Sohn, D. Berthelot, N. Carlini, D. C. Ekin, K. Alex, and Z. Han, “Fixmatch: simplifying semi-supervised learning with consistency and confidence,” in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Vancouver, BC, Canada, December 2020.
- [31] S. Yun, D. Han, S. J. Oh, Y. Yoo, and J. Choe, “Cutmix: regularization strategy to train strong classifiers with localizable features,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6023–6032, Seoul, Korea, October 2019.
- [32] T. Devries and G. W. Taylor, “Improved regularization of convolutional neural networks with cutout,” 2017, <https://arxiv.org/abs/1708.04552>.
- [33] H. Zhang, M. Cissé, Y. N. Dauphin, and D. Lopez-Paz, “mixup: beyond empirical risk minimization,” in *Proceedings of the 6th International Conference on Learning Representations, ICLR 2018*, Conference Track Proceedings, Vancouver, BC, Canada, Apr, 2018.
- [34] X. Lai, Z. Tian, L. Jiang et al., “Semi-supervised semantic segmentation with directional context-aware consistency,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1205–1214, Nashville, Tennessee, USA, 2021.
- [35] Z. Ke, D. Qiu, K. Li, Q. Yan, and R. W. H. Lau, “Guided collaborative training for pixel-wise semi-supervised learning,” in *Computer Vision—ECCV 2020: 16th European Conference* Springer, Glasgow, UK, 2020.
- [36] D. H. Lee, “Pseudo-label: the simple and efficient semi-supervised learning method for deep neural networks,” *Workshop on challenges in representation learning ICML*, vol. 3, no. 2, pp. 896–902, 2013.
- [37] A. Tarvainen and H. Valpola, “Mean teachers are better role models: weight-averaged consistency targets improve semi-supervised deep learning results,” in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 1195–1204, Long Beach, CA, USA, December 2017.
- [38] X. Yu, B. Han, J. Yao, N. Gang, I. Tsang, and S. Masashi, “How does disagreement help generalization against label corruption?” in *International Conference On Machine Learning* PMLR, Venue, 2019.
- [39] H. Pham, Z. Dai, Q. Xie, and Q. V. Le, “Meta pseudo labels,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11557–11568, Nashville, TN, USA, June 2021.
- [40] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, “The pascal visual object classes challenge: a retrospective,” *International Journal of Computer Vision*, vol. 111, no. 1, pp. 98–136, 2015.
- [41] M. Cordts, M. Omran, S. Ramos et al., “The cityscapes dataset for semantic urban scene understanding,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3213–3223, Las Vegas, NV, USA, June 2016.
- [42] S. Ioffe and C. Szegedy, “Batch normalization: accelerating deep network training by reducing internal covariate shift,” in *International Conference on Machine Learning* PMLR, Venue, 2015.

## Research Article

# Fast Detection of Defective Insulator Based on Improved YOLOv5s

Zhao Liqun <sup>1</sup>, Zou Mengjun <sup>1</sup>, Cui Ying <sup>2</sup>, and Jia Yanfei <sup>3</sup>

<sup>1</sup>Key Laboratory of Modern Power System Simulation and Control & Renewable Energy Technology, Ministry of Education (Northeast Electric Power University), Jilin 132013, China

<sup>2</sup>Guangdong Electric Power Corporation, Zhuhai Power Supply Bureau, Zhuhai 519000, China

<sup>3</sup>College of Electrical and Information Engineering, Beihua University, Jilin 132013, China

Correspondence should be addressed to Zhao Liqun; zhaoliqun@163.com

Received 29 July 2022; Accepted 16 August 2022; Published 3 September 2022

Academic Editor: Nian Zhang

Copyright © 2022 Zhao Liqun et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Defective insulator detection is an essential part of transmission line inspections based on unmanned aerial vehicles. It can timely discover insulator defects and repair them to avoid a power transmission accident. The detection speed of defective insulators based on artificial intelligence directly affects inspection efficiency. To improve the detection speed of defective insulators based on YOLOv5s, an improved detection method with faster detection speed and acceptable precision is proposed. First, a new ResNet unit with three branches is designed based on depthwise separable convolution with kernel three and average pooling. To reduce parameters, the new ResNet unit is used to replace the original ResNet unit used in the CSP1\_X module in YOLOv5s. Besides, we also introduce channel shuffle in the CSP1\_X module to facilitate the flow of feature information from different channels. Second, a new residual CBL module is designed based on depthwise separable and standard convolution. The new residual CBL module is used to replace the two CBL modules used in the CSP2\_X module in YOLOv5s to reduce parameters and extract more useful features. Third, we design a separate, coordinated attention module by introducing location information into channel attention. The new attention module is added to the end of the CSP2\_X module to improve the ability to extract insulator location information. Besides, we also use convolution to replace the focus model to reduce computation. Compared with defective insulator detection methods, the proposed method has smaller parameters, floating-point operations per second, and higher frames per second. Although it has lower mean precision, it has a faster detection speed. Besides, the increase in detection speed is greater than the decrease in mean precision.

## 1. Introduction

Unmanned aerial vehicle has been widely used in transmission line inspection to improve inspection efficiency and reduce the workload for inspectors [1, 2]. In the traditional transmission line inspection based on an unmanned aerial vehicle, the inspector detects the defective transmission devices by watching the screen. It takes a lot of time to detect faulty devices, which is affected by the screen size and the light intensity. With the development of artificial intelligence technology, the deep learning method has been widely used in object tracking [3, 4], image super-resolution reconstruction [5], image dehazing [6], and defective transmission device detection [7]. The images of transmission line devices captured by the unmanned aerial vehicle are transmitted via a wireless mobile communication network to

a server with high computing power. The captured images are detected by artificial intelligence methods deployed on the server. It consumes much time to transmit images to servers, affecting transmission devices' detection speed. It cannot be used in detection areas without wireless mobile network coverage.

With the development of portable edge computing devices, the captured images can be detected by artificial intelligence methods deployed on the mobile edge computing device [8]. The captured images are transmitted to portable edge computing devices carried by inspectors with short-range wireless communication. It is not dependent on the mobile network and has a greater range of applications. The computation power of portable edge computing devices is limited, which affects the detection speed. In order to improve detection speed, one is to improve the computation

power of portable edge computing devices, and the other is to reduce the computation of the artificial intelligence methods. The cost of reducing computation is lower. Therefore, we mainly focus on how to reduce computation.

Object detection based on deep learning can be divided into two categories: one-stage detection method and two-stage detection method. In the one-stage method, it directly generates detection boxes and classifies the objects without generating region proposals. The two-stage method divides the detection process into two stages. It generates region proposals in the first stage, regresses the bounding box and candidate regions, and classifies the objects in the second stage. The one-stage method focus on improving detection speed. The two-stage method focus on improving detection precision. The two-stage method is mainly developed on the servers with high computing power. It is more suitable for offline object detection. The one-stage method has lower computation than two-stage method. It is more suitable for online object detection. Although it has been widely used in defective insulator detection, the detection is still required to be improved to meet the practical requirements.

YOLOv5 is one of the one-stage methods [9]. Although it is named YOLOv5, it is designed based on YOLOv3 [10] and is unrelated to YOLOv4 [11]. The YOLOv5 has four versions that are YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x. The YOLOv5s has a smaller computation and lower precision. The YOLOv5x has a larger computation and higher precision. The YOLOv5s is more suitable for deploying on the portable edge computing device. To improve the defective insulator detection speed of YOLOv5s with acceptable precision, we propose an improved YOLOv5s.

For this paper, the main contributions are as follows:

- (1) To reduce computation and improve the detection speed of defective insulators, we design a new ResNet unit with two branches, and use it to replace the original ResNet unit used in CSP1\_X module in YOLOv5s. The new ResNet is based on depthwise separable convolution with kernel three and average pooling. It has smaller computation than the original ResNet unit In YOLOv5s. Besides, to extract more useful features from different channels, we also introduce the channel shuffle operation into CSP1\_X module in YOLOv5s.
- (2) To further reduce the computation, we also design a residual CBL module and use it instead of two original CBL modules in CSP2\_X module in YOLOv5s. The residual CBL module has smaller computation than the original two CBL modules. Besides, to extract more useful features from different channels, we also introduce the channel shuffle operation into CSP2\_X module in YOLOv5s to facilitate the flow of feature information from different channels.
- (3) To reduce background interference and make the network pay more attention to useful information, we design a separate embedded coordinated attention module and introduce it into the CSP2\_X module to extract more useful features. Besides, to

reduce computation, we also use convolution with kernel 3 to replace the focus.

The traditional transmission line inspection based on unmanned aerial vehicle requires a human eye view of the defective insulators. To improve the inspection efficiency based on unmanned aerial vehicle, artificial intelligence methods are deployed on servers or ground service stations to automatically detect defective insulators from the captured image by unmanned aerial vehicles. Although many artificial intelligence methods have higher precision, they have slower detection speed that directly affects inspection efficiency. To improve defective insulator detection speed with acceptable detection precision, we propose an improved YOLOv5s with lower computation for defective insulator detection to improve the inspection efficiency. The structure of the rest of this article is organized as follows. Related work is provided in Section 2. Our proposed method are introduced in Section 3. The experimental results and discussions are reported in Section 4. Finally, conclusions are given in Section 5.

## 2. Related Work

Object detection method based on deep learning has been widely used in defective insulator and transmission device detection. There are two categories of object detection models. The first category is two-stage detection models, such as Faster R-CNN [12], Mask R-CNN [13], Cascade R-CNN [14], and Sparse R-CNN [15]. The second category is one-stage detection models, such as the SSD series method [16, 17] and the YOLO series method [9–11]. The two-stage has a larger computation and is more suitable for offline detection. The one-stage has smaller computation and is more suitable for online detection. Therefore, the defective insulator detection method based on a one-stage method is more suitable for developing a portable edge computing device with limited computation power.

The YOLO series method is the most representative method of the one-stage detection method. The YOLOv1-method was first proposed by Redmon et al. [18]. It creatively combined candidate area and object recognition into one stage. It transformed the object recognition problem into a regression problem and directly predicted the location and class of the object using a depth convolution neural network. Compared with the two-stage method, it has lower computation and acceptable accuracy. Due to these advantages, the YOLO method attracts much attention from many scholars and has become an important branch of object detection research based on deep convolution neural networks. To improve the detection speed and precision of YOLOv1, Redmon et al. [19] proposed YOLOv2. In YOLOv2, batch normalization is introduced into the convolution layer. The new convolution layer consists of convolution, batch normalization, and LeakyReLU activating function. In the VOC2007 dataset, the mean average precision is improved from 63.4% to 65.8%. In order to solve the problem caused by the different training and detection image sizes, YOLOv2 fine-tuned the classification network that had been trained on  $244 \times 244$  low-resolution images on

448 × 448 high-resolution images. After fine-tuning, the final global average pooling and softmax layers are removed as the final backbone network. Then mean average precision is improved from 65.8% to 69.5%. It also introduced anchor boxes inspired by Faster R-CNN to improve precision. The size of the output feature map is 13 × 13. Each cell contains five anchor boxes to predict five bound boxes. Besides, YOLOv2 designed a new network named Darknet-19 and used it as a backbone. In 2018, Redmon proposed YOLOv3, which is also the last version proposed by Redmon [10]. In order to extract more useful features, YOLOv3 used the Darknet-53 as the backbone and introduced feature pyramid networks to fuse more features. The number of network layers increased and contained many residual networks in Darknet-53. Besides, it also proposed binary cross-entropy loss for classification. Although Redmon withdraws from research in artificial intelligence, many improved YOLOv3 are proposed by scholars.

Yang et al. proposed GC-YOLOv3 based on YOLOv3 to improve the mean average precision [20]. They designed a cascading network that consisted of learnable semantic fusion and used a global self-attention mechanism to extract more useful information. Qu et al. proposed an improved YOLOv3 with auxiliary networks for remote sensing image detection to improve detection precision and detection speed [21]. They used an image blocking module to feed fixed images and DIOU to replace IOU in YOLOv3. Besides, they used a convolutional block attention module to connect the backbone network and designed an auxiliary network. In order to improve detection speed, the adaptive feature fusion method was also introduced into the improved YOLOv3. In order to improve detection speed, Yin et al. proposed Faster-YOLO [22]. They used ELELEM-AE joint network and DRKCELM network to design a feature extractor. The feature extractor integrates the advantages of ELM-EA and ELM-LRF. The detection speed of Faster-YOLO is two times faster than YOLOv3. In order to improve detection precision, Cai et al. proposed a modified YOLOv3 [23] based on MobileNetV1. They used the MobileNetV1 to replace the backbone of YOLOv3 and optimized the feature map size according to the detection results.

In 2020, Alexey et al. [11] proposed the YOLOv4 method base on YOLOv3. They used the cross-stage partial connections network to replace the residual block in YOLOv3 to design the backbone of YOLOv4. The path aggregation network also was used to fuse more features. They used spatial pyramid pooling to realize multi-scale features fusion and mish function to replace the LeakyReLU function as a new activate function. Besides, they used the mosaic data argument method to improve detection precision. The YOLOv4 has a faster detection speed and precision. To improve the detection speed of YOLOv4, Deng et al. [24] proposed an improved YOLOv4. They used feature pyramid networks and atrous spatial pyramid pooling to modify the MobileNetV3 to improve real-time and feature extraction ability. The improved MobileNetV3 is used as the backbone of YOLOv4 to reduce computation. They also introduced a convolutional block attention module to YOLOv4 to extract more useful features. The original team of YOLOv4

proposed scaled-YOLOv4 based on YOLOv4-CSP to make the model can be developed on different devices [25]. To further improve the performance of YOLOv4, some improved methods based on YOLOv4 have also been proposed [26–28]. To improve the one-stage method performance, Glenn proposed YOLOv5 based on YOLOv3 [10]. YOLOv5 designed two types of cross-stage partial (CSP) Networks that are CSP1\_X and CSP2\_X. The CSP1\_X modules were used in the backbone, and the CSP2\_X were used in the Neck part of YOLOv5. YOLOv5 has four different version networks that are YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x. The different version of YOLOv5 has different depth CSP modules. YOLOv5 used the LeakyReLU function to activate the function in middle/hidden layers and the sigmoid function as an activated function in detection layers. Besides, YOLOv5 used GIOU loss as the loss function of the bounding box. YOLOv5s has a faster detection than YOLOv4 and YOLOv3. It is more suitable for developing portable edge computing devices.

Due to the advantage of the YOLO series method, they have been widely used in transmission devices and defective insulator detection. Liu et al. [29] proposed an aerial insulator image detection method based on YOLOv3 for high-voltage transmission lines. A cross-stage partially densely connected module was proposed to solve the feature reuse and propagation of feature layers in low-resolution images. It had higher detection precision defective insulator detection in complex transmission line backgrounds than YOLOv3 and YOLOv4. Qiu et al. [30] proposed a defective insulator detection algorithm based on YOLOv4. They used the Graph Cut data enhancement method to produce a new dataset, and the Laplace sharpening method was used to preprocess the insulator dataset images. To make the algorithm more lightweight, they used MobileNet as the backbone network of YOLOv4. It had a faster detection speed than YOLOv4. He et al. [31] proposed a self-exploding insulator detection algorithm based on YOLOv4. They proposed a new feature fusion structure and an improved SE attention mechanism, which effectively suppressed useless features and achieved higher detection accuracy than YOLOv4. Feng et al. [32] verified in detail the performance of YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x on the public dataset China Power Line Insulation Dataset (CPLID), and the experimental results showed that YOLOv5 performed well on this dataset, especially YOLOv5x, with a detection accuracy of 95.5%, which can effectively identify defective insulators. Since the YOLOv5x model has a large number of parameters, Lan et al. [33] selected YOLOv5s with a smaller number of parameters as the baseline model, making it lighter by replacing the original CSP structure with the Ghost module and adding CBAM attention to improve the detection accuracy. To further improve the detection accuracy of defective insulators, some other methods based on YOLO have also been proposed [34–36].

### 3. Improved YOLOv5s

*3.1. Improved CSP Module.* Two types of CSP modules are given as CSP1\_X and CSP2\_X in YOLOv5s. The CSP1\_X

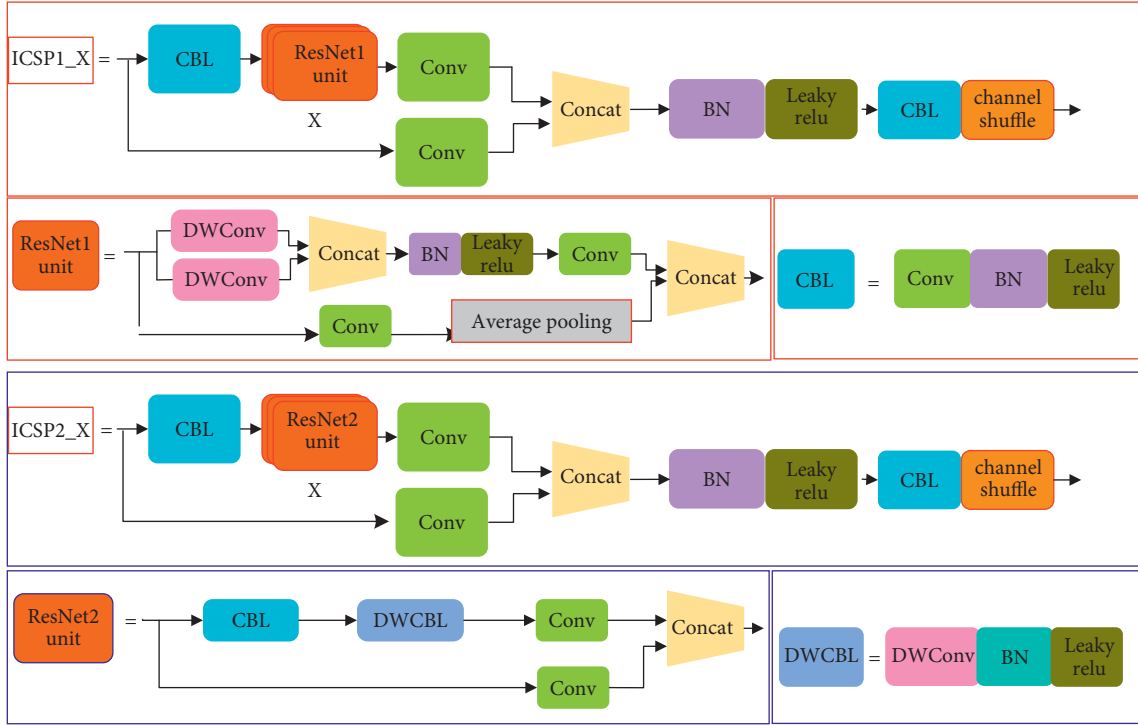


FIGURE 1: Improved CSP1\_X and CSP2\_X.

module is used in the backbone to extract features and the CSP2\_X module is used in the Neck to fuse features. The CSP1\_X consists of two branches. The first branch mainly consists of CBL module, ResNet units, and a convolution layer. The second branch only consists of a convolution layer. The output features of two branches are fused by concatenate operation. The ResNet unit computation directly affects detection speed.

To reduce the computation of CPS1\_X, we designed a new ResNet unit and replaced the original ResNet unit in CSP1\_X. The proposed ResNet unit is named as ResNet1 unit and is shown in Figure 1. The proposed ResNet1 unit consists of two branches that we name the upper branch and lower branch, respectively. The upper branch contains two parallel depthwise separable convolutions (DWConv). The kernel sizes of the two depthwise separable convolutions are 3, and the number of output channels is half the number of input channels. We use the two parallel depthwise separable convolutions to extract different features. Therefore, we use concatenate operation to connect the output feature maps of two DWConv to keep the number of channels constant. To avoid gradient explosion or gradient disappearance and improve training speed, we add batch normalization (BN) and LeakyReLU activating function in the upper branch. Besides, we use a  $1 \times 1$  convolution to reduce the number of output channels to halve the original number. The lower branches contain a  $1 \times 1$  convolution and an average pooling. The  $1 \times 1$  convolution is used to reduce the number of channels to make the lower branch have the same number of channels as the upper branch.

The average pooling is used to reduce the value of the background feature, which is useful for separating the

defective insulator and background information. In the end, the upper and lower branches are connected by concatenating operation to construct the complete ResNet1 unit. We use the proposed ResNet1 unit to replace the original ResNet unit to reduce the computation of CSP1\_X. In the end, we use the BN module, LeakyReLU activating function, and CBL module to extract features from the concatenated feature map. The CBL module consists of a  $1 \times 1$  standard convolution, a BN module, and a LeakyReLU activating function. They are used to increase network depth to make the network extract more detailed information about the defective insulator. The improved CPS1\_X is named ICSP1\_X which is shown in Figure 1. The X is the number of ResNet1 unit used in ICSP1\_X. The X is different in different modules. In YOLOv5s, two different ICSP1\_X modules are present, i.e., ICSP1\_1 and ICSP1\_3. The numbers of ResNet1 unit of ICSP1\_1 and ICSP1\_3 is 1 and 3, respectively. The ICSP1\_X module does not change the number of feature map channels. They are only used in the backbone of the network to extract features.

To further reduce the computation of the network, we also design a new residual module to replace the two CBL modules in CSP2\_X. We name the proposed residual unit and improved CSP2\_X as ResNet2 unit and ICSP2\_X, respectively. The ResNet2 unit and ICSP2\_X are shown in Figure 1. The proposed ResNet2 unit also contains two branches. The upper branch includes a CBL module, a DWConv module, and a  $1 \times 1$  convolution. The CBL module consists of a  $1 \times 1$  standard convolution, a BN module, and a LeakyReLU activating function. The CBL module is used to increase the depth of the network, which is useful for extracting more detailed information about the defective



insulator. The number of output channels for the CBL module is half the number of input channels in the ResNet2 unit. The DWCBL module consists of  $3 \times 3$  depthwise separable convolutions, a BN module and a LeakyReLU activating function. The input feature map of DWCBL module has the same number of channels as the output feature map of DWCBL module. The last  $1 \times 1$  standard convolution in ResNet2 unit is also used to increase the network depth. The lower branch only consists of a  $1 \times 1$  standard convolution. The number of output channels for the  $1 \times 1$  standard convolution is half the number of input channels. In the end, the upper and lower branches are connected by concatenating operation to construct the complete ResNet2 unit. The input feature map of the proposed ResNet2 unit module has the same number of channels as the output feature map of the ResNet2 unit. We use the proposed ResNet2 unit to replace two CBL modules in the original CSP2\_X to obtain the ICSP2\_X module. The ICSP2\_X also consists two branches that are named upper branch and lower branch, respectively. The upper branch consists of a CBL module,  $X$  ResNet2 unit and a  $1 \times 1$  standard convolution. The number of output channels for the upper branch is also the is half the number of input channels. The lower branch of ICSP2\_X is only a  $1 \times 1$  standard convolution to reduce the number of input channels. The concatenating operation is also used to connect upper and lower branches. The same BN module, LeakyReLU activating function, and CBL module are also used to extract features from the concatenated feature map in the ICSP2\_X module. Besides, we also introduce the channel shuffle into ICSP2\_X to facilitate the flow of feature information from different channels.

In the ICSP1\_X and ICSP2\_X, we use depthwise separable convolution to replace the standard  $3 \times 3$  convolution. The FLOPs of standard convolution can be expressed as:

$$N_c = 2 \times K^2 \times C_{in} \times C_{out} \times H \times W, \quad (1)$$

where  $K$  is the kernel size of standard convolution,  $C_{in}$  and  $C_{out}$  are the number of channels for input and output, respectively, and  $H$  and  $W$  are the width and height of the feature map. The FLOPs of depthwise separable convolution can be expressed as:

$$D_c = 2 \times (K^2 \times C_{in} + C_{out} \times C_{in}) \times H \times W. \quad (2)$$

The FLOPs ratio of standard convolution and depthwise separable convolution is:

$$\begin{aligned} \frac{D_c}{N_c} &= \frac{2 \times (K^2 \times C_{in} + 1^2 \times C_{in} \times C_{out}) \times H \times W}{2 \times K^2 \times C_{in} \times C_{out} \times H \times W} \\ &= \frac{K^2 + C_{out}}{K^2 * C_{out}}. \end{aligned} \quad (3)$$

In YOLO5s, the numbers of output channels are 128, 256, and 512 for different modules. The kernel sizes are 1 or 3 for other convolutions. The number of output channels is much larger than kernel size, so the (3) can be simplified as:

$$\frac{D_c}{N_c} = \frac{K^2 + C_{out}}{K^2 * C_{out}} \approx \frac{C_{out}}{K^2 * C_{out}} = \frac{1}{K^2}. \quad (4)$$

The original ResNet unit of CSP1\_X contains a  $1 \times 1$  convolution and a  $3 \times 3$  convolution. The proposed ResNet unit of ICSP1\_X contains two  $1 \times 1$  convolutions and two  $3 \times 3$  depthwise separable convolutions. For the CSP1\_X, the computation of the proposed ResNet1 unit and original ResNet unit is about  $2/9$ . For the CSP2\_X, the computation of the proposed ResNet2 unit and CBL model is about  $1/9$ . Our proposed two networks that are the ResNet1 unit and ResNet2 unit reduce the computation of CSP1\_X and CSP2\_X. The proposed ICSP1\_X and ICSP2\_X have smaller computations than the original CSP1\_X and CSP2\_X.

### 3.2. Separate Embedded Coordinated Attention Module.

The backbone of our improved YOLOv5s based on our proposed ICSP1\_X and ICSP2\_X has few parameters and computation but also reduces feature extraction ability. Inspired by [37], to balance the detection speed and precision, we designed a separated embedded coordinated attention module named SCA module and introduce it into ICSP1\_X module and ICSP2\_X module to extract more useful defective insulator features. The proposed SCA module is shown in Figure 2. It consists of two branches: the upper branch consists of short cut to reserve input feature map. The lower branch is used to compute the weight of useful features; the lower branch has two average pooling operations: adaptive average pooling (H) and adaptive average pooling (W). The kernel sizes of adaptive average pooling (H) and adaptive average pooling (W) are  $H \times 1$  and  $1 \times W$ , respectively. They encode each channel along the horizontal and vertical coordinates, respectively. We can obtain two separate position-aware feature maps from the two adaptive average pooling operations. Therefore, we can capture long-range dependencies of feature information along one spatial direction while retaining precise location information in the other spatial direction. To make the network have better expression ability. We use CBH modules consisting of a convolution layer, a BN layer and a  $H\_sigmoid$  activate function. Without changing the size of the direction-aware feature map, we use the  $1 \times 1$  convolution layer to increase the depth of the network, use BN layer to improve the generalization ability of the network and use the  $H\_sigmoid$  active function to increase the nonlinear representation ability of the network. Then two  $1 \times 1$  convolutions are used to reduce the number of channels to half the original number. The output feature of two branches that contain two adaptive average pooling are connected by concatenate operation. The CBH module is also used to fuse the feature obtained by concatenate operation. The split method is used to separate the feature map from the horizontal coordinate and vertical coordinate. In the end, we can obtain the weight of the horizontal coordinate, and the weight of the vertical coordinate of the feature map by the CBH module and sigmoid activate function. We multiply the input feature map by two weights to get the output feature map.

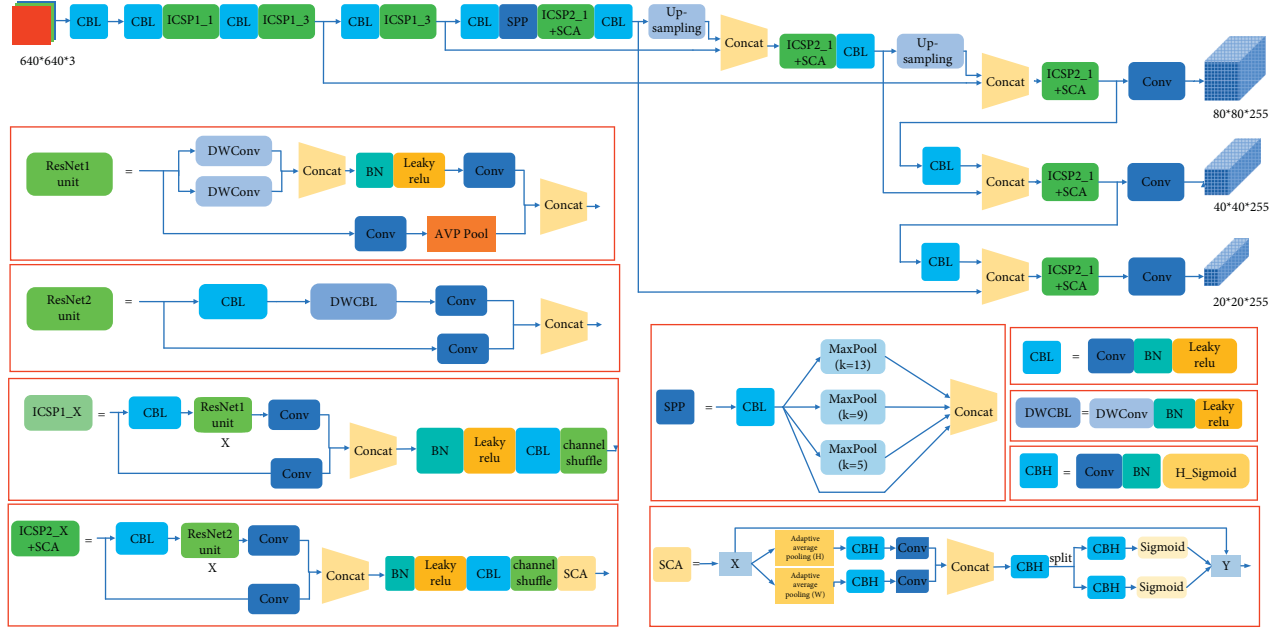


FIGURE 2: Improved YOLOv5s.

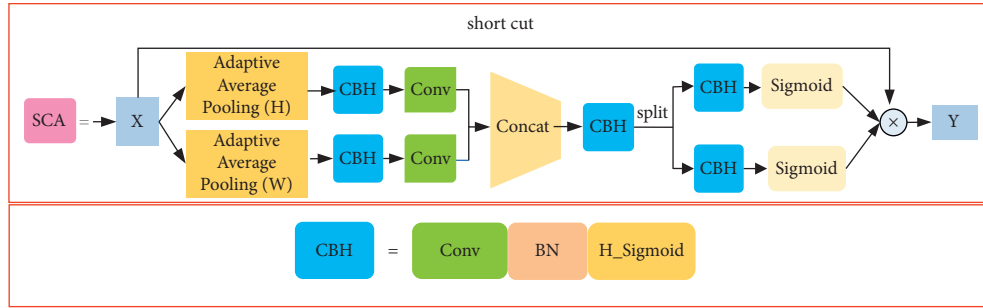


FIGURE 3: Proposed separate embedded coordinated attention module.

To better explain the model, we suppose that the input feature map  $X \in R^{C \times H \times W}$ . The number of channels is  $C$ , and the size of the feature map is  $H \times W$ . Output feature map of adaptive average pooling (H) and adaptive average pooling (W) in the  $c$ th channel can be expressed as  $Z_c(H)$  and  $Z_c(W)$ . They can be expressed as followings:

$$Z_c(H) = \frac{1}{W} \sum_{0 \leq i < W} X_c(H, i) \quad (5)$$

$$Z_c(W) = \frac{1}{H} \sum_{0 \leq i < H} X_c(i, W),$$

where  $X_c$  is the input feature map in the  $c$ th channel. The sizes of  $Z_c(H)$  and  $Z_c(W)$  are  $H \times 1$  and  $1 \times W$ , respectively. The  $1 \times 1$  convolution is used to halve the number of output channels of CBH in Figure 3. The concat operation connects the two output feature maps obtained from CBH and a  $1 \times 1$  convolution. It can be expressed as follows.

$$X_1 = [f^{1 \times 1}(CBH(Z(H))), f^{1 \times 1}(CBH(Z(W)))], \quad (6)$$

where  $f^{1 \times 1}$  is the  $1 \times 1$  convolution,  $[, ]$  is the concatenating operation.  $X_1$  is the output feature map of concat operation in Figure 2. Next, we use the split function to separate the output feature of the CBH module that is connected with concat operation module into two parts. In the end, we obtain the weights of two parts. The sizes of weight are  $H \times 1$  and  $1 \times W$ , respectively. The size of the final weight obtained by multiplying two weight vectors is  $H \times W$ . Finally, we multiply the input feature map  $X$  with the final weight to get the output feature map.

**3.3. Improved Focus Module.** In YOLOv5s, it uses focus to realize down-sampling without losing information. In focus, it firstly uses a slicing operation to expand the input channel four times. Second, it uses a  $3 \times 3$  convolution to obtain a down-sampling feature map. There is a large computation for the slicing operation. To reduce computation, we use a  $3 \times 3$  convolution with a down-sampling function to replace the focus module. We suppose the input image size is  $640 \times 640 \times 3$  and the size of the output feature map size is

TABLE 1: Pseudo code of our proposed method.

---

```

 $\theta \leftarrow$  Improved YOLOv5s Training ( $z_1: N_{data}, x_1: N_{data}, \theta$ )
/* Training network model
Input:  $\{(z_n, x_n)\}_{n=1}^{N_{data}}$ , a dataset of sequence pairs
Input:  $\theta$ , initial model parameters
Output:  $\theta$ , the trained parameters
Hyperparameters:  $N_{epochs} \in \mathbb{N}$ 
For  $i = 1, 2, \dots, N_{epochs}$  do
For  $n = 1, 2, \dots, N_{data}$  do
 $\theta \leftarrow$  Improved YOLOv5s ( $(z_n, x_n, \theta)$ )
 $loss(\theta) \leftarrow l_{con}(\theta) + 0.5 \times l_{cls}(\theta) + 0.05 \times l_{box}(\theta)$ 

 $m_n \leftarrow \beta_1 m_{n-1} + (1 - \beta_1) \nabla loss(\theta)$ 
 $v_n \leftarrow \beta_2 v_{n-1} + (1 - \beta_2) \nabla loss(\theta)^2$ 
 $\hat{m}_n \leftarrow \frac{m_n}{1 - \beta_1}, \hat{v}_n \leftarrow \frac{v_n}{1 - \beta_2}$ 
 $\theta \leftarrow \theta - \frac{\mu}{\sqrt{\hat{v}_n} + \epsilon} \times \hat{m}_n$ 
End
End
Return  $\hat{\theta} = \theta$ 

```

---

$320 \times 320 \times 32$ . The FLOPs of focus and convolution without considering the BN and activate function are shown in (7) and (8), respectively.

$$Focus = 3 \times 3 \times 12 \times 32 \times 320 \times 320 = 353894400. \quad (7)$$

$$Cov = 3 \times 3 \times 3 \times 32 \times 320 \times 320 = 88473600. \quad (8)$$

Based on (7) and (8), the computation of the focus model is about four times larger than convolution. Therefore, using convolution to replace focus can reduce computation.

Based on our proposed ICSP1\_X, ICSP2\_X, SCA, and improved focus modules, the improved YOLOv5s are shown in Figure 2. The pseudo-code of our proposed method is shown in Table 1. The Adam optimizer method is used to optimize the parameters of the network.

## 4. Simulation and Discussion

The experimental environment is configured as follows: the operating system is Ubuntu 18.04. The deep learning framework is Pytorch. Six detection methods that our proposed method, YOLOv3 [10], YOLOv4 [11], YOLOv5s [9], Fast R-Transformer [38], and Mina-Net [31] method are set with the same parameters and pre-training weights are used. The initial learning rate was set to 0.001, and the learning rate was dynamically adjusted using the cosine annealing learning rate. To prevent overfitting, the Adam optimizer is used to adjust the parameters. Pre-training is used in the first five epochs to accelerate the convergence of the model. Pre-training weights are used for all the above methods.

We sourced 195 insulator images from the Power supply and the Internet. The number of defective insulator images is

too small to train the network. Therefore, we use the same method used in the CPLID dataset [39] to expand the data set. We extract defective insulators and fuse them with different backgrounds. The parts of generated defective insulator images are shown in Figure 4. Our proposed defective dataset contains 2300 images. We use 1800 images as training images and 500 images as test images.

**4.1. Ablation Study.** We design a new ResNet1 unit and ResNet2 unit and introduce them into CSP1\_X and CSP2\_X modules to reduce the computation. Besides, we also introduce the channel shuffle operation into improved CSP1\_X and CSP2\_X modules. To simplify, we name the YOLOv5s based on improved CSP1\_X and CSP2\_X without channel shuffle operation as YOLOv5s + Res, YOLOv5s based on improved CSP1\_X and CSP2\_X with channel shuffle operation as YOLOv5s + Res + Cs. Besides, we also design a separate embedded coordinate attention module and introduce it into the improved and CSP2\_X modules. We name the YOLOv5s + Res + Cs based on a separate embedded coordinate attention module as YOLOv5s + Res + Cs + SA. We also name the YOLOv5s + Res + Cs based on the convolutional block attention module [40] as YOLOv5s + Res + Cs + CBAM, and based on Squeeze-and-Excitation [41] as YOLOv5s + Res + Cs + SE. In addition, we also use convolution to replace focus in YOLOv5s. The YOLOv5s + Res + Cs + SA based on a convolution without using focus is named YOLOv5s + Res + Cs + SA-Foc, which is also our complete improved YOLOv5s. The results are shown in Table 2.

The YOLOv5s have the most parameters, FLOPs, and the highest mean average precision (mAP). The YOLOv5s + Res that YOLOv5s based on improved CSP CSP1\_X and CSP2\_X have the smallest parameters, FLOPs, and mAP. As the number of parameters decreases, the mAP also decreases. Although the YOLOv5s + Res + cs that YOLOv5s based on improved CSP CSP1\_X and CSP2\_X with channel shuffle operation have the same number of parameters and FLOPs, but it has larger mAP than YOLOv5s + Res. The channel shuffle operation improves the precision without increasing parameters and FLOPs. Compared with YOLOv5s + Res + Cs, the YOLOv5s + Res + Cs + SE, YOLOv5s + Res + Cs + CBAM and YOLOv5s + Res + Cs + SA that are introduced different attention modules into YOLOv5s + Res + Cs have larger parameters, FLOPs, and mAP. The attention mechanism improves detection precision while increasing parameters and FLOPs. Compared with YOLOv5s + Res + Cs + SE and YOLOv5s + Res + Cs + CBAM, YOLOv5s + Res + Cs + SA has the larger mAP. The SA module is our proposed attention module. This shows that our proposed attention module performs better in precision than SE and CBAM attention modules. Compared with YOLOv5s + Res + Cs + SA, the YOLOv5s + Res + Cs + SA-Foc that uses convolution to replace the focus module to reduce computation has smaller parameters, FLOPs, and mAP.

We propose the improved CSP1\_X module and CSP2\_X module and modify the focus module to reduce computation. We propose an attention module and introduce it into



FIGURE 4: Parts of generated defective insulator images.

TABLE 2: YOLOv5s with different modules.

Methods	Params (M)	FLOPs (G)	mAP (%)
YOLOv5s	7.06	16.3	94.5
YOLOv5s + Res	5.19	12.2	91.7
YOLOv5s + Res + Cs	5.19	12.2	93.2
YOLOv5s + Res + Cs + SE	5.38	12.3	93.5
YOLOv5s + Res + Cs + CBAM	5.40	12.4	93.7
YOLOv5s + Res + Cs + SA	5.47	12.8	94.1
YOLOv5s + Res + Cs + SA-foc(ours)	5.45	12.5	93.6

the network to improve precision. Based on the analysis of Table 2, the improved CSP1\_X module, CSP2\_X module, and modified focus reduce the parameters and FLOPs, and the proposed attention module increases precision.

*4.2. Comparisons for Different Methods.* We compare our complete proposed method with YOLOv3, YOLOv4, and YOLOv5s, Fast R-Transformer, and [38] Mina-Net [31]. We randomly select three different insulators that contain defective insulators. The detection results are shown in Figures 5–7. In each figure, figure (a), figure (b), figure (c), figure (d), figure (e), and figure (f) are obtained by Fast

R-Transformer, Mina-Net YOLOv3, YOLOv4, YOLOv5s, and our proposed method, respectively. In Figure 5, there are two defective insulators, and the YOLOv3 only detects one defective insulator. The YOLOv4, YOLOv5s, Fast R-Transformer, Mina-Net, and our proposed method successfully detect all defective insulators. In Figures 6 and 7, all methods successfully detect all defective insulators. This shows that the proposed method is valid for defective insulator detection.

To verify the performance of the proposed method from a statistical point of view, we use YOLOv3, YOLOv4, YOLOv5s, Fast R-Transformer, Mina-Net, and our proposed to detect all images in the test dataset. The detection results

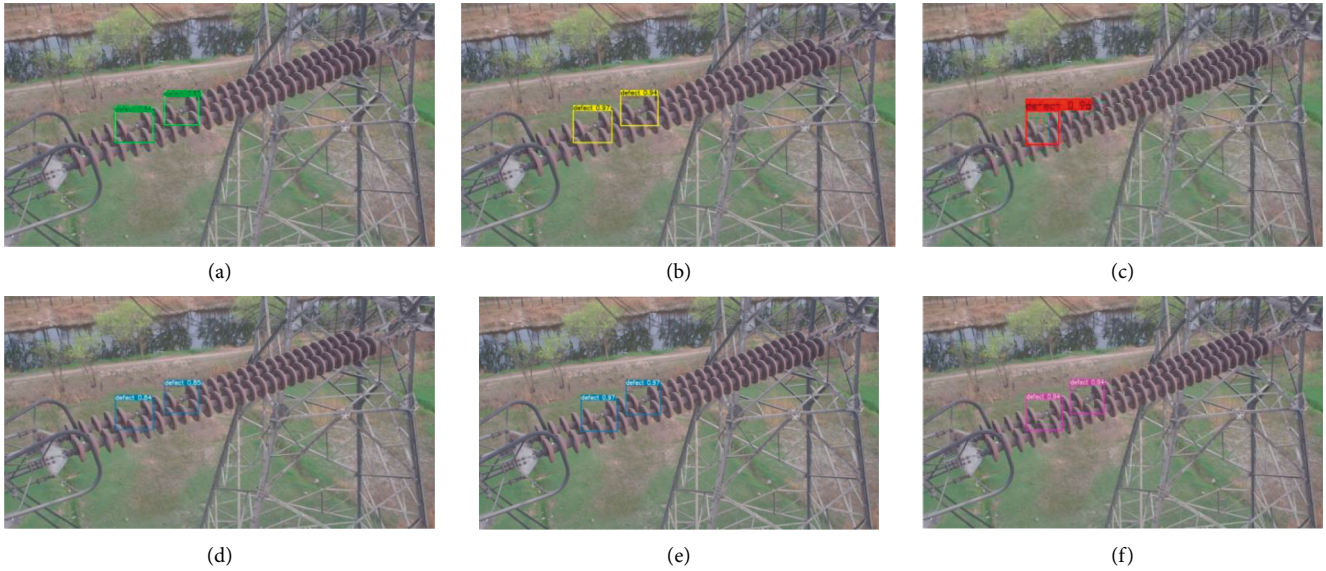


FIGURE 5: Defective porcelain insulator detection.

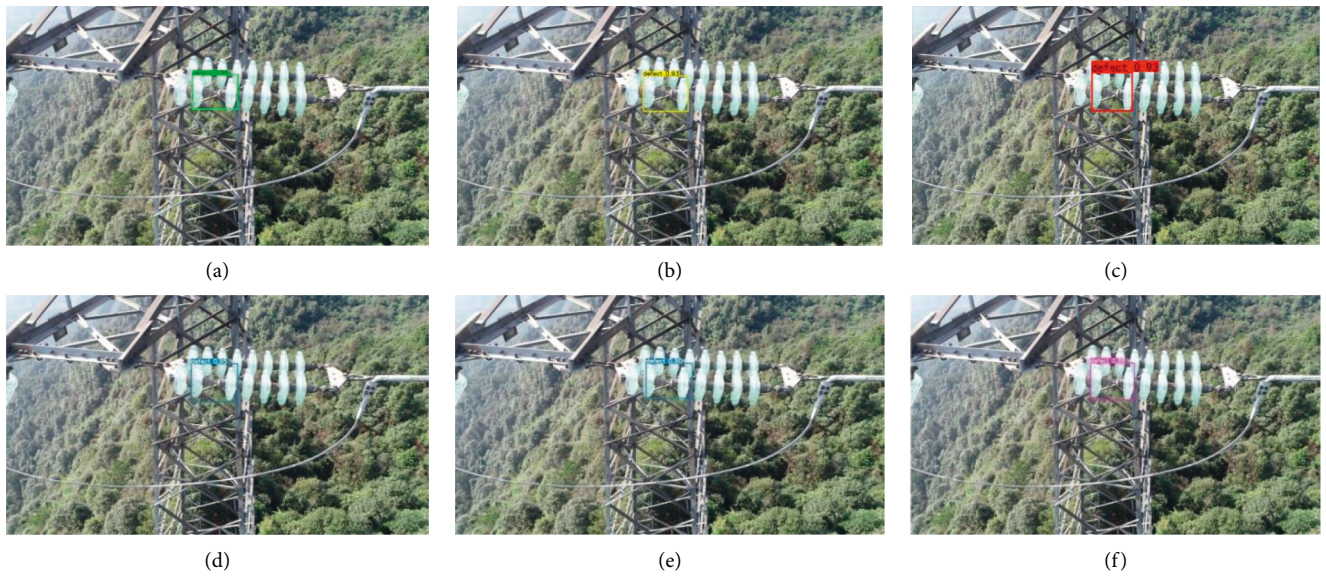


FIGURE 6: Defective toughened glass insulator detection.

are shown in Table 3. The parameters of Fast R-Transformer, Mina-Net, YOLOv3, YOLOv4, YOLOv5s, and our proposed method are 116.10 M, 67.94 M, 61.90 M, 63.94 M, 7.06 M, and 5.45 M, respectively. Our proposed method has the smallest parameters, followed by YOLOv5s. The FLOPs of Fast R-Transformer, Mina-Net, YOLOv3, YOLOv4, YOLOv5s, and our proposed method are 257.5 G, 168.2 G, 155.1 G, 157.1 G, 16.3 G, and 12.5 G, respectively. Our proposed method still has the smallest parameters, followed by YOLOv5s. The mAPs of Fast R-Transformer, Mina-Net, YOLOv3, YOLOv4, YOLOv5s, and our proposed method are 99.7%, 99.2%, 91.5%, 98.7%, 94.5%, and 93.6%, respectively. The Fast R-Transformer has the largest mAP, followed by Mina-Net and YOLOv4. The FPS of Fast

R-Transformer, Mina-Net, YOLOv3, YOLOv4, YOLOv5s, and our proposed method are 38.6, 47.5, 54.6, 52.7, 157.4, and 197.8, respectively. Our proposed method has the largest FPS, followed by YOLOv5s.

Based on the analysis of Table 3, our proposed method has the smallest parameters, FLOPs, and the largest FPS. The FLOPs of different models are computed by model\_into() function in touch\_utils.py. Compared with YOLOv5s, although the mAP of our proposed method is reduced by 0.9%, FPS of our proposed method is increased by about 20%. Although our proposed method has a smaller mAP than other methods, our proposed method has a faster detection speed. Besides, the increase in detection speed is greater than the decrease in precision.



FIGURE 7: Defective suspended glass insulator detection.

TABLE 3: Comparisons for different methods.

Methods	Params (M)	FLOPs (G)	mAP (%)	FPS
Fast R-Transformer	116.1	257.5	99.7	38.6
Mina-net	67.94	168.2	99.2	47.5
YOLOv3	61.90	155.1	91.5	54.6
YOLOv4	63.94	157.1	98.7	52.7
YOLOv5s	7.06	16.3	94.5	157.4
Ours	5.45	12.5	93.6	197.8

## 5. Conclusions

This study proposed a faster defective insulator detection method based on YOLOv5s. It designs two different ResNet units for CSP1\_X module and CSP2\_X module in YOLOv5s to reduce computation, respectively. It also introduced channel shuffle into CSP1\_X module and CSP2\_X module to extract more effective features from different channels without increasing computation. Besides, it designed a separate embedded coordinated attention module and introduced it into the CSP2\_X module to make the network pay more attention to useful information. To reduce the computation, it replaced the focus module using a convolution with stride 2. Compared with defective insulator detection methods based on Fast R-Transformer, Mina-Net, YOLOv3, YOLOv4, and YOLOv5s, our proposed method has the smallest parameters and FLOPs and the largest FPS. This shows that the proposed method has the fastest defective insulator detection speed. Although the other methods except YOLOv3 have larger mAP than our method, the difference in mAP is not large. The mAP of Fast R-Transformer is the largest. Although the mAP of Fast R-Transformer is 6.1% higher than that of our method, the detection speed of Fast R-Transformer is only 1/5 of our

method detection speed. Our method and YOLOv5s have a faster detection speed than others. Compared with YOLOv5s, although the mAP of our proposed method is reduced by 0.9%, the detection speed of our proposed method is increased by about 20%. The increase in detection speed is greater than the decrease in precision. The proposed method has a faster detection speed and an acceptable detection precision.

## Data Availability

The labeled dataset used to support the findings of this study is available from the corresponding author upon request.

## Conflicts of Interest

The authors declare that they have no conflicts of interest.

## Acknowledgments

This work was supported by the National Natural Science Foundation of China (61271115) and Science and Technology Research Project of Jilin Provincial Department of Education (JJKH20220054KJ).

## References

- [1] W. Zhao, Q. Dong, and Z. Zuo, "A method combining line detection and semantic segmentation for power line extraction from unmanned aerial vehicle images," *Remote Sensing*, vol. 14, no. 6, Article ID 1367, 2022.
- [2] X. Li, Z. Li, H. Wang, and W. Li, "Unmanned aerial vehicle for transmission line inspection: status, standardization, and perspectives," *Frontiers in Energy Research*, vol. 9, p. 336, 2021.

- [3] R. Xia, Y. Chen, and B. Ren, "Improved Anti-occlusion Object Tracking Algorithm Using Unscented Rauch-Tung-Striebel Smoother and Kernel Correlation Filter," *Journal of King Saud University - Computer and Information Sciences*, 2022.
- [4] J. Zhang, W. Feng, T. Yuan, J. Wang, and A. K. Sangaiyah, "SCSTCF: spatial-channel selection and temporal regularized correlation filters for visual tracking," *Applied Soft Computing*, vol. 118, Article ID 108485, 2022.
- [5] Y. Chen, L. Liu, V. Phonevilay et al., "Image super-resolution reconstruction based on feature map attention mechanism," *Applied Intelligence*, vol. 51, no. 7, pp. 4367–4380, 2021.
- [6] L. Zhao, Y. Zhang, and Y. Cui, "An attention encoder-decoder network based on generative adversarial network for remote sensing image dehazing," *IEEE Sensors Journal*, vol. 22, no. 11, pp. 10890–10900, 2022.
- [7] W. Rahmaniari and A. Hernawan, "Real-time human detection using deep learning on embedded platforms: a review," *Journal of Robotics and Control (JRC)*, vol. 2, no. 6, pp. 462–468, 2021.
- [8] H. Mei, H. Jiang, F. Yin, L. Wang, and M. Farzaneh, "Terahertz imaging method for composite insulator defects based on edge detection algorithm," *IEEE Transactions On Instrumentation And Measurement*, vol. 70, pp. 1–10, 2021.
- [9] J. Glenn, "YOLOv5," 2020, <https://github.com/ultralytics/YOLOv5>.
- [10] J. Redmon and A. Farhadi, "Yolov3: an incremental improvement," 2018, <http://arxiv.org/abs/1804.02767>.
- [11] A. Bochkovskiy, C. Y. Wang, and H. Y. M. Liao, "YOLOv4: optimal speed and accuracy of object detection," 2020, <http://arxiv.org/abs/2004.10934>.
- [12] Z. Zhang, Z. Xiong, B. Zhang, Y. Yang, and E. Fu, "Detection for small target ship in remote sensing image based on super resolution reconstruction technology," *Journal of Northeast Electric Power University*, vol. 42, no. 2, pp. 33–40, 2022.
- [13] K. He, G. Gkioxari, P. Dollár, R. Girshick, and R.-C. N. N. Mask, *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2961–2969, 2017.
- [14] Z. Cai and N. Vasconcelos, "Cascade R-CNN: delving into high quality object detection," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6154–6162, 2018.
- [15] P. Sun, R. Zhang, Y. Jiang et al., "End-to-End object detection with learnable proposals," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14454–14463, Nashville, TN, USA, 2021.
- [16] W. Liu, D. Anguelov, D. Erhan et al., *SSD: Single Shot Multibox Detector, European Conference on Computer Vision*, pp. 21–37, Springer, Berlin/Heidelberg, Germany, 2016, <http://arxiv.org/abs/1512.02325>.
- [17] J. Jeong, H. Park, and N. Kwak, "Enhancement of SSD by concatenating feature maps for object detection," 2017, <http://arxiv.org/abs/1705.09587> arXiv preprint arXiv:1705.09587.
- [18] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 779–788, Las Vegas, NV, USA, June 2016.
- [19] J. Redmon and A. Farhadi, "YOLO9000: better, faster, stronger," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7263–7271, Honolulu, HI, USA, July 2017.
- [20] Y. Yang and H. Deng, "GC-YOLOv3: you only look once with global context block," *Electronics*, vol. 9, no. 8, p. 1235, 2020.
- [21] Z. Qu, F. Zhu, and C. Qi, "Remote sensing image target detection: improvement of the YOLOv3 model with auxiliary networks," *Remote Sensing*, vol. 13, no. 19, p. 3908, 2021.
- [22] Y. Yin, H. Li, and W. Fu, "Faster-YOLO: an accurate and faster object detection method," *Digital Signal Processing*, vol. 102, Article ID 102756, 2020.
- [23] K. Cai, X. Miao, W. Wang, H. Pang, Y. Liu, and J. Song, "A modified YOLOv3 model for fish detection based on MobileNetv1 as backbone," *Aquacultural Engineering*, vol. 91, Article ID 102117, 2020.
- [24] T. Deng and Y. Wu, "Simultaneous vehicle and lane detection via MobileNetV3 in car following scene," *Plos one*, vol. 17, no. 3, Article ID e0264551, 2022.
- [25] C. Y. Wang, A. Bochkovskiy, and H. Y. M. Liao, "Scaled-yolov4: scaling cross stage partial network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13029–13038, 2021.
- [26] F. Guo, Y. Qian, and Y. Shi, "Real-time railroad track components inspection based on the improved YOLOv4 framework," *Automation in Construction*, vol. 125, Article ID 103596, 2021.
- [27] J. Yu and W. Zhang, "Face Mask wearing detection algorithm based on improved YOLO-v4," *Sensors*, vol. 21, no. 9, p. 3263, 2021.
- [28] Z. Ma, Y. Li, M. Huang, Q. Huang, J. Cheng, and S. Tang, "A lightweight detector based on attention mechanism for aluminum strip surface defect detection," *Computers in Industry*, vol. 136, 103585.
- [29] C. Liu, Y. Wu, J. Liu, Z. Sun, and H. Xu, "Insulator faults detection in aerial images from high-voltage transmission lines based on deep learning model," *Applied Sciences*, vol. 11, no. 10, p. 4647, 2021.
- [30] Z. Qiu, X. Zhu, C. Liao, D. Shi, and W. Qu, "Detection of transmission line insulator defects based on an improved lightweight YOLOv4 model," *Applied Sciences*, vol. 12, no. 3, p. 1207, 2022.
- [31] H. He, X. Huang, Y. Song et al., "An insulator self-blast detection method based on YOLOv4 with aerial images," *Energy Reports*, vol. 8, pp. 448–454, 2022.
- [32] Z. Feng, L. Guo, D. Huang, and R. Li, "Electrical Insulator Defects Detection Method Based on YOLOv5," in *Proceedings of the 2021 IEEE 10th Data Driven Control and Learning Systems Conference (DDCLS)*, pp. 979–984, Suzhou, China, May 2021.
- [33] Y. Lan and W. Xu, "Insulator defect detection algorithm based on a lightweight network," *Journal of Physics: Conference Series*, vol. 2181, no. 1, Article ID 012007, 2022.
- [34] Y. Wang, P. Cao, X. Wang, and X. Yan, "Research on insulator self explosion detection method based on deep learning," *Journal of Northeast Electric Power University*, vol. 40, no. 3, pp. 33–40, 2020.
- [35] M. F. Palangar, S. Mohseni, M. Mirzaie, and A. Mahmoudi, "Designing an automatic detector device to diagnose insulator state on overhead distribution lines," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 2, pp. 1072–1082, 2022.
- [36] B. Wang, M. Dong, M. Ren et al., "Automatic fault diagnosis of infrared insulator images based on image instance segmentation and temperature analysis," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 8, pp. 5345–5355, 2020.
- [37] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13713–13722, 2021.

- [38] S. Dian, X. Zhong, and Y. Zhong, "Faster R-Transformer: an efficient method for insulator detection in complex aerial environments," *Measurement*, vol. 199, Article ID 111238, 2022.
- [39] X. Tao, D. Zhang, Z. Wang, X. Liu, H. Zhang, and D. Xu, "Detection of power line insulator defects using aerial images analyzed with convolutional neural networks," *IEEE Transactions On Systems, Man, and Cybernetics: Systems*, vol. 50, no. 4, pp. 1486–1498, 2020.
- [40] S. Woo, J. Park, J. K. Leeand, and S. Kweon, "CBAM: convolutional block Attention module," *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 3–19, 2018.
- [41] 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, 2018.



## Research Article

# Intelligent Detection Method of Gearbox Based on Adaptive Hierarchical Clustering and Subset

Huimiao Yuan <sup>1</sup>, Yongwei Tang <sup>1,2</sup>, Huijuan Hao <sup>1</sup>, Yuanyuan Zhao <sup>1</sup>, Yu Zhang <sup>1</sup>,  
and Yu Chen <sup>1</sup>

<sup>1</sup>Qilu University of Technology (Shandong Academy of Sciences),  
Shandong Computer Science Center (National Supercomputer Center in Jinan),  
Shandong Key Laboratory of Computer Networks,  
Jinan 250014, China

<sup>2</sup>School of Mechanical Engineering, Shandong University, Key Laboratory of High Efficiency and Clean Mechanical Manufacture,  
Jinan 250100, China

Correspondence should be addressed to Huijuan Hao; haohj@sdas.org

Received 21 June 2022; Revised 1 August 2022; Accepted 6 August 2022; Published 30 August 2022

Academic Editor: Nian Zhang

Copyright © 2022 Huimiao Yuan et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Deep learning uses mechanical time-frequency signals to train deep neural networks, which realizes automatic feature extraction and intelligent diagnosis of fault features and gets rid of the dependence on a large number of signal processing technology and experience. Aiming at the problem of misclassification of similar samples, a fault diagnosis algorithm based on adaptive hierarchical clustering and subset (AHC-SFD) is proposed to extract features and applied to gearbox fault diagnosis. Firstly, the adaptive hierarchical clustering algorithm is used to analyze the characteristics of different data, and then the data set is clustered into multiple feature groups; finally, according to the feature group, the SubCNN model is established for multiscale feature extraction, so as to carry out fault diagnosis. The test results show that the fault recognition rate achieved by the proposed method is more than 99.7% on the gearbox dataset, and the method has better generalization ability.

## 1. Introduction

Major accidents caused by mechanical equipment failure [1] constantly alert people to ensure the safe and reliable operation of equipment, especially the mechanical equipment failure at the key core of the production line will bring significant shutdown losses to the whole production line, not only causing huge economic losses, but also endangering personal safety in serious cases. The online monitoring, fault diagnosis, and prediction of mechanical equipment [2, 3] play an important role in improving equipment operation reliability, optimizing operation and maintenance strategies, and are crucial to the maintenance of mechanical equipment. Traditional intelligent fault diagnosis methods need to master a large number of signal processing techniques to extract relatively accurate feature parameters. At the same time, if the shallow model is used to characterize the

relationship between signal and fault, and the diagnosis ability and generalization ability are insufficient, it is difficult to meet the actual needs of fault diagnosis under big data.

In recent years, the application of deep learning in fault diagnosis of complex industrial systems has begun to take shape [4]. Lei et al. [5, 6] proposed a big data health monitoring method based on denoising self-encoder (DAE) for mechanical equipment, which has realized a variety of fault diagnosis for planetary gears, reflecting the powerful ability of deep learning to extract mechanical vibration signal characteristics. Yu and Zhao [7–9] effectively integrated DAE and EN to solve the problem of noise interference in fault diagnosis, effectively detect abnormal samples in industrial processes, and isolate fault variables from normal variables. Nguyen et al. [10–12] proposed a deep learning network composed of automatic encoder and softmax classifier to identify bearing faults of different degrees. DBN is more

combined with other technologies to solve the problem of fault diagnosis. Since CNN was used to identify bearing faults in 2016, fault diagnosis performance and scope of application have been continuously improved. Hoang and Kang [13–16] proposed a new method based on CNN for rolling bearing fault diagnosis. By using the effectiveness of CNN in image classification, the CWRU bearing data set can achieve 100% diagnosis accuracy. Based on resnet-50, a transfer learning convolution neural network TCNN is proposed by Wen et al. [17, 18] for fault diagnosis, and the prediction accuracy is significantly better than other DL models and traditional diagnosis methods. The application of RNN in fault diagnosis began to recover in 2015. Abed et al. [19, 20] used RNN for bearing fault diagnosis and realized accurate detection and classification of bearing faults under non-stationary conditions. Pan et al. [21–23] proposed a method for bearing fault classification by combining one-dimensional CNN and LSTM, and the experimental test accuracy is 99.6%.

Although the above algorithm has been applied in mechanical equipment fault diagnosis, there is still a lot of room to improve the fault recognition rate. Feature extraction is a key part of fault diagnosis. It is found that for samples with similar features and belonging to different patterns, a single model will extract similar features, resulting in false recognition [24] and a reduction in the accuracy of fault diagnosis. In view of the above problems, referring to the idea of subset [25, 26], this study proposes a multiscale feature extraction fault diagnosis algorithm model AHC-SFD based on adaptive hierarchical clustering and applied to gearbox fault diagnosis. The test results show that the proposed method can achieve the fault recognition rate achieved by the proposed method is more than 99.7% on the gearbox dataset and has better generalization ability.

## 2. Gear Fault Diagnosis Algorithm Based on Adaptive Hierarchical Clustering and Subset

Gear boxes generally work in the environment with strong noise and complex structure, and the collected vibration signals are easily affected by external factors. To fully develop the feature extraction ability of the CNN network, this study proposes a fault diagnosis algorithm based on adaptive hierarchical clustering and subset. First, all data obtained the optimal clustering results through adaptive hierarchical clustering, and a multiscale feature extraction module is designed according to the clustering results to realize the classification of fault data.

*2.1. Adaptive Hierarchical Clustering.* The number of clusters is an important parameter that affects the clustering effect, but before clustering, it is often necessary to set the number of clusters to take a fixed value. As the amount of data changes, the original parameter values cannot optimize the clustering result of the algorithm. Combined with the characteristics of vibration signals, an adaptive hierarchical clustering (DIANA) algorithm is proposed in this study. The clustering contour

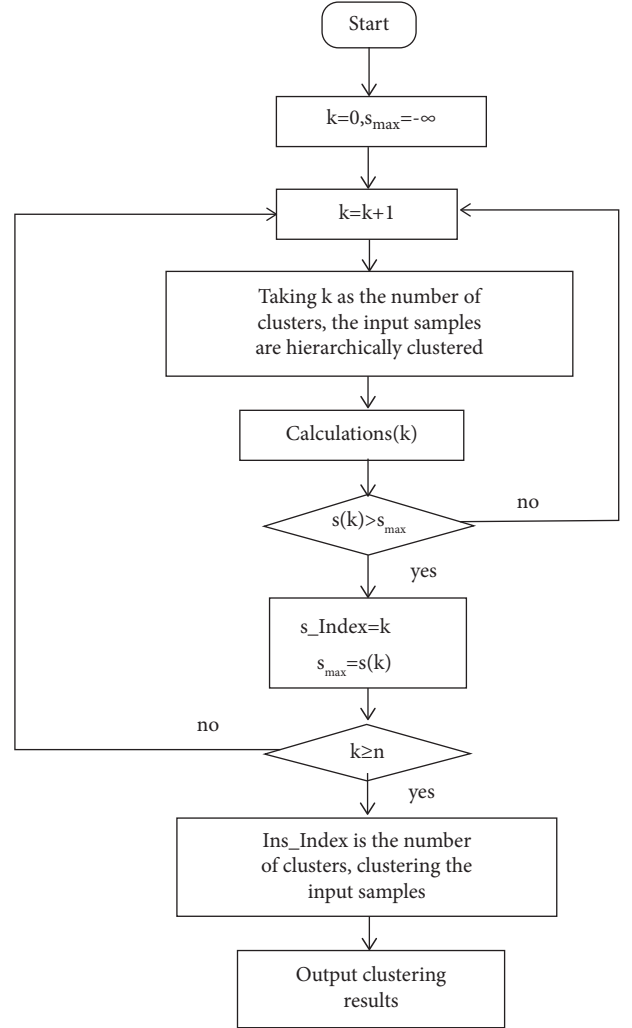


FIGURE 1: Adaptive hierarchical clustering flow chart.

coefficient is used as the index of clustering effectiveness evaluation, so that it can adaptively determine the number of clusters according to the value of self-defined discriminant function. The process is shown in Figure 1.

The specific algorithm flow chart is as follows:

- (1) Extract the average value of each original vibration signal to form a feature sample set  $X = \{x_1, x_2, \dots, x_{num}\}$ ,  $U = \{u_1, u_2, \dots, u_c\}$  indicates fault type set
- (2) Start clustering, make  $k = 0$ ,  $s_{max} = -\infty$ ;
- (3) Let  $k = k + 1$ , take  $k$  as the number of clusters, and perform hierarchical clustering on the input training samples (DIANA);
- (4) Calculate the contour coefficient  $s(k)$ ,

$$a(i) = \frac{1}{n_c - 1} \sum_{j \in C_c, i \neq j} d(i, j). \quad (1)$$

In equation (1),  $n_c$  represents the number of samples of class  $c$ ,  $C_c$  represents the samples of class  $c$ , and  $d(i, j)$  represents the absolute distance between samples  $i$  and  $j$ ;

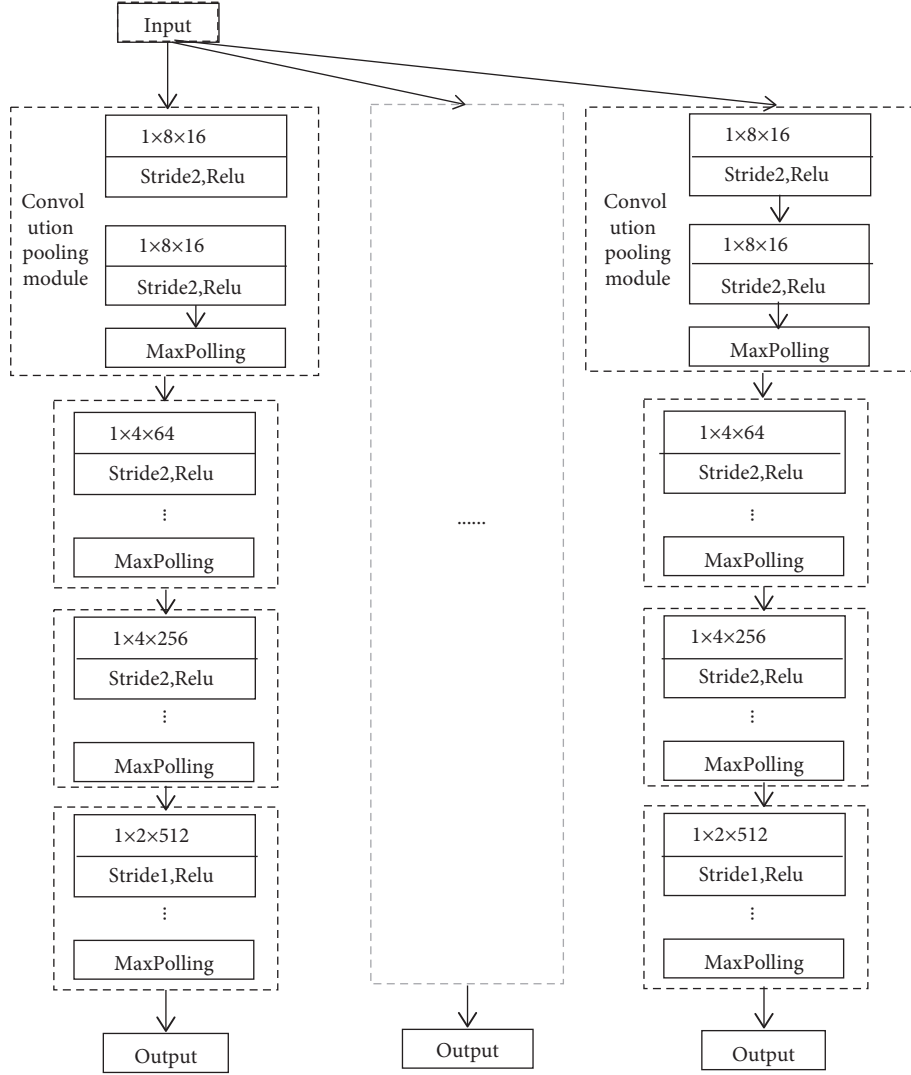


FIGURE 2: Multiscale feature extraction module.

$$b(i) = \min_{p, p \neq c} \left[ \frac{1}{n_p} \sum_{j \in C_p, i \in C_c} d(i, j) \right]. \quad (2)$$

In equation (2),  $p$  denotes a mark other than Class  $c$ ,  $n_p$  represents the number of samples not of class  $c$ ,  $C_p$  represents a sample that is not class  $c$ ,  $C_c$  is the sample of class  $c$ , and  $d(i, j)$  is the absolute distance between samples  $i$  and  $j$ ;

$$s(i) = \frac{b(i) - a(i)}{\max[a(i), b(i)]}. \quad (3)$$

In equation (3),  $a(i)$  represents the average distance between sample  $i$  and all other samples belonging to the same type of fault, and  $b(i)$  represents the minimum value of the average distance between sample  $i$  and all samples in each class of nonclass  $i$  fault;

$$s(k) = \frac{1}{\text{num}} \sum_{i=1}^{\text{num}} s_i. \quad (4)$$

In equation (4),  $s_i$  is the contour coefficient of the sample individual, num is the number of samples in the feature sample set, and  $k$  is the number of clusters;

- (5) When  $s(k) > s_{\max}$ , then  $s\_Index = k$  and  $s_{\max} = s(k)$ , perform step 7;
- (6) When  $s(k) \leq s_{\max}$ , return to step 3;
- (7) Judge whether  $k$  is less than  $n$ , where  $n$  indicates the number of dataset types:

When  $k \geq n$ ,  $s\_Index$  is the number of clusters and the clustering results are output;

When  $k < n$ , repeat step 3.

**2.2. Multiscale (Subset) Feature Extraction.** In order to maximize the extraction of feature information from

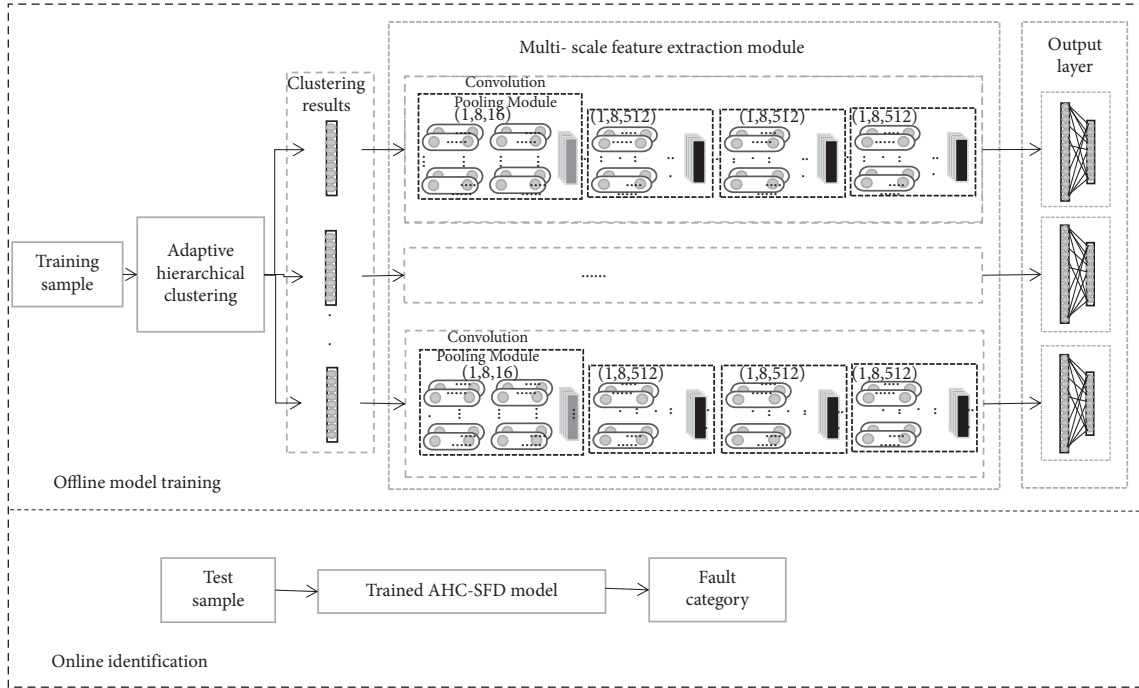


FIGURE 3: Flow chart of adaptive hierarchical clustering and subset fault diagnosis.

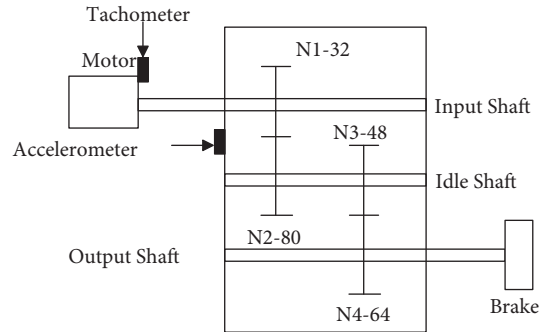


FIGURE 4: Gearbox experimental system.

training data and quickly realize iteration, this study designs a multilayer and multichannel multiscale feature extraction module based on the CNN. The structure is shown in Figure 2. The branch structure of each subset (12 layers in total) is the same, in which the convolution kernel sizes of the 8-layer convolution layers are  $1 * 8$ ,  $1 * 8$ ,  $1 * 4$ ,  $1 * 4$ ,  $1 * 4$ ,  $1 * 2$ , and  $1 * 2$ , the number of channels is set to 16, 16, 64, 64, 256, 256, 512, and 512, and the step size is set to 2, 2, 2, 2, 1, and 1. The relu activation function is used behind each convolution layer, and the max pool layer of 4 adopts the  $1 * 2$  structure. Finally, the extracted feature information is output.

**2.3. AHC-SFD Diagnostic Algorithm.** The flow chart of adaptive hierarchical clustering and subset fault diagnosis proposed in this study is shown in Figure 3. The mean value of each vibration signal is used as the input of adaptive hierarchical clustering to obtain the optimal clustering

results. The labeled samples corresponding to the results are input to the multiscale feature extraction module to obtain more effective fault data features. Finally, the features extracted by the multifeature extraction module are transformed into one-dimensional data through the fully connected layer. Output the fault diagnosis result through softmax function.

### 3. Experimental Verification and Analysis

In order to evaluate the effectiveness and accuracy of fault diagnosis of the AHC-SFD network model, the gearbox dataset is used for experimental verification. The data are collected from a reference two-stage gearbox, the gear speed is controlled by a motor, and the torque is provided by a magnetic brake, which can be adjusted by changing its input voltage. A 32-tooth pinion and an 80-tooth pinion are installed on the first stage input shaft, the second stage consists

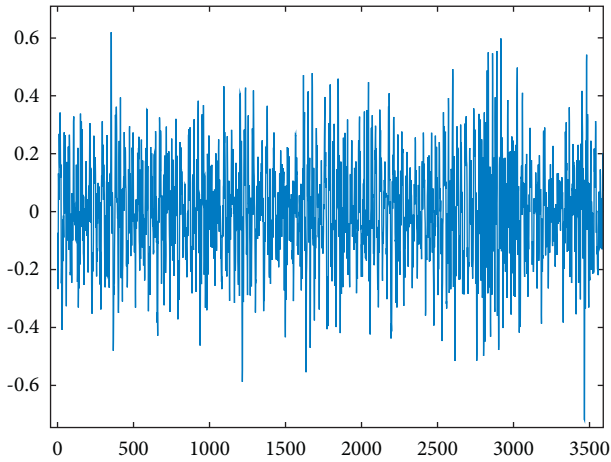


FIGURE 5: Health.

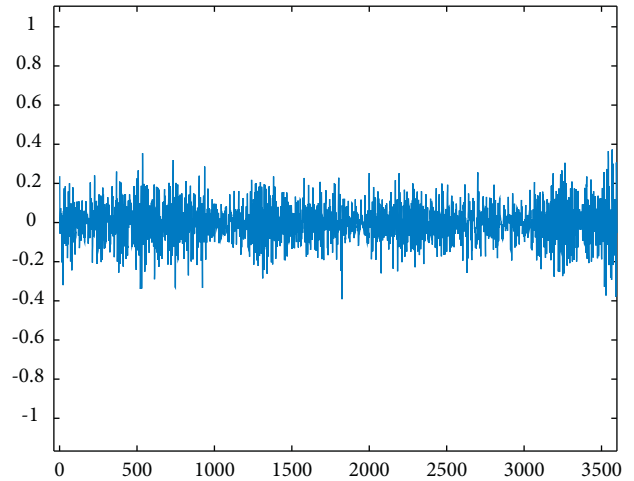


FIGURE 7: Root cracking.

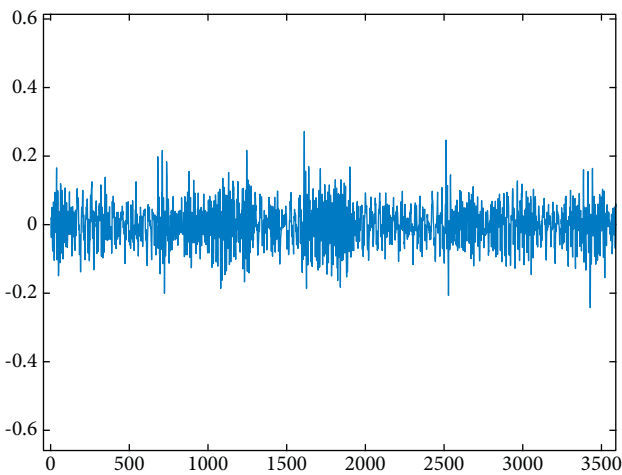


FIGURE 6: Missing teeth.

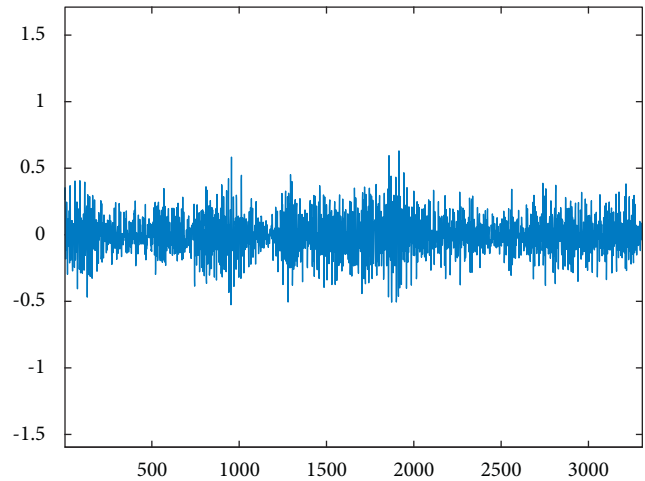


FIGURE 8: Peeling.

of a 48-tooth pinion and a 64-tooth pinion. Input shaft speed is measured by tachometer, and gear vibration signal is measured by accelerometer, as shown in Figure 4.

**3.1. Fault Dataset Description and Processing.** The pinion on the input shaft introduces 9 different gear conditions, including five different severity labels, such as health, missing teeth, root cracking, peeling, and tip cutting. The number of samples in each status tag is the same. The collected data are roughly divided into training samples and test samples in the proportion of 4:1. Each sampling sample is set to 3600 points. The dataset is described in Figures 5–13 and Table 1.

### 3.2. Adaptive Hierarchical Clustering

**3.2.1. Refactoring Input Data Format.** The dataset collected by the test-bed is a one-dimensional vibration signal sequence. In order to reduce the clustering time and carry out the adaptive hierarchical clustering operation quickly and

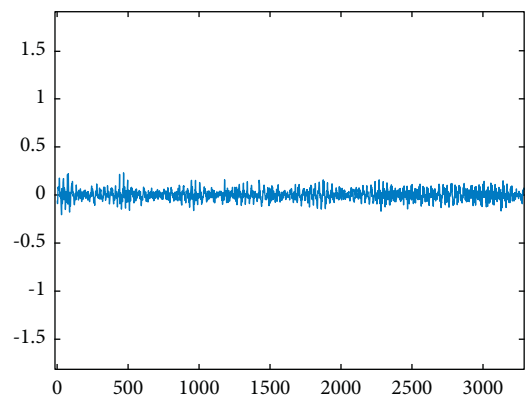


FIGURE 9: Tip cutting5.

effectively, this study takes the one-dimensional vibration signal with 3600 sampling points as the average value and takes the average value as the input value of the adaptive hierarchical clustering. The specific operation is as follows:

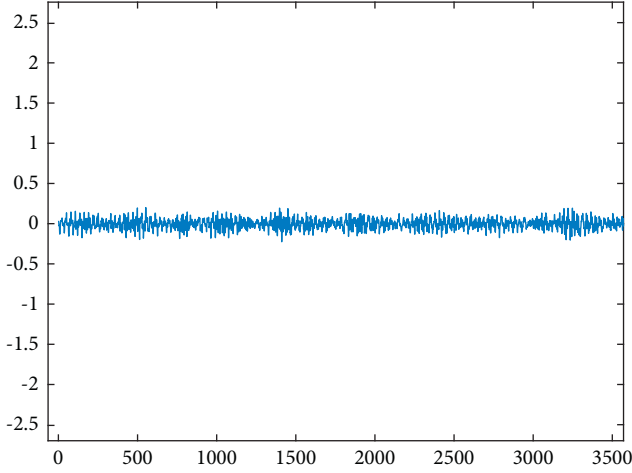


FIGURE 10: Tip cutting4.

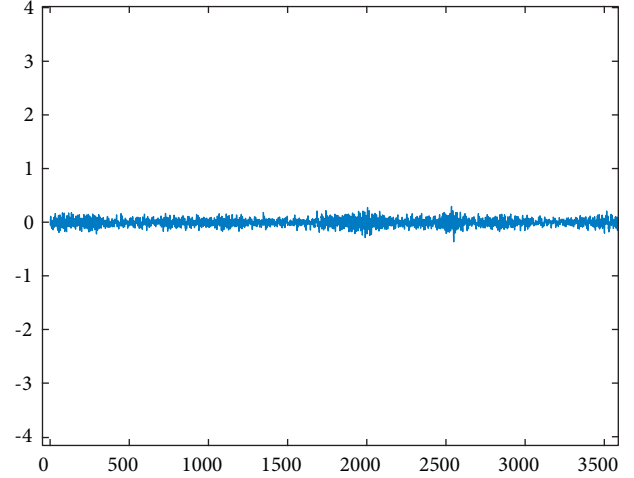


FIGURE 13: Tip cutting1.

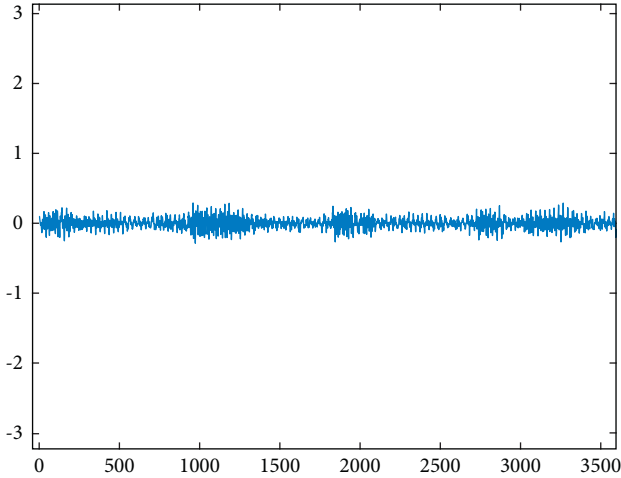


FIGURE 11: Tip cutting3.

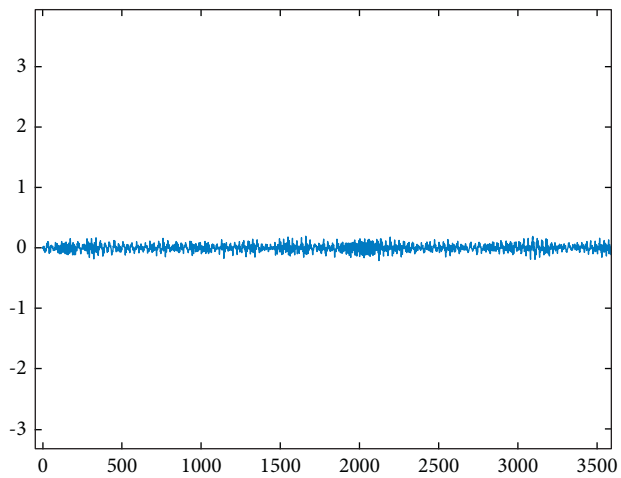


FIGURE 12: Tip cutting2.

$$X = \frac{\sum_{i=1}^{3600} x_i}{3600}. \quad (5)$$

In equation (5),  $x_i$  represents the  $i$ -th eigen value of a sample and  $X$  represents the average value of a sample.

**3.2.2. Result Output.** The principle of adaptive clustering is to obtain a certain clustering result, so that the distance between classes is as large as possible, the distance within a class is as small as possible, and the classes have good separability. It can be seen from 2.1 that the cluster contour coefficient is used as the index for cluster effectiveness evaluation in this study. The closer the cluster contour coefficient is to 1, the better the clustering result is. The closer it is to  $-1$ , the worse the clustering result is. In this study, the number of clusters is set between  $[1, 9]$ . During clustering, the cluster contour coefficients obtained with the change of the number of clusters is shown in Figure 14. It can be clearly seen that when the number of clusters are 2, the cluster contour coefficient ( $Sk$ ) is the largest. Therefore, the branch of the multiscale feature extraction module is set to 2.

### 3.3. Improved CNN Network

**3.3.1. Grouping Label Data According to Clustering Results.** Use labeled data; the labeled data samples are  $(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \dots, (x^{(m)}, y^{(m)})$ ,  $x^{(i)}$  represents the feature vector, and  $y^{(i)} \in \{1, 2, \dots, t\}$  represents the fault type. According to the clustering results in 3.2.2, the label data (one-dimensional vibration signal) is divided into two groups. The two groups are divided into training samples and test samples according to the ratio of 39:11 and 19:6, respectively. The description of the training and testing datasets is shown in Table 2.

**3.3.2. Data Standardization Operation.** In order to better speed up the network model training, make the data easy to calculate and obtain more generalized results, the input data are standardized, and the vibration signal data are mapped to the  $(0,1)$  interval by using the normalization equation. The mathematical expression is as follows:

TABLE 1: Gearbox dataset.

Fault information		Sample information		Category information
Fault type	Fault degree	Sample length	Number of samples	Category tag
Health	0	3600	104	0
Missing tooth	0	3600	104	1
Root crack	0	3600	104	2
Spalling	0	3600	104	3
Tip cutting	5 (lightest)	3600	104	4
	4	3600	104	5
	3	3600	104	6
	2	3600	104	7
	1 (most serious)	3600	104	8

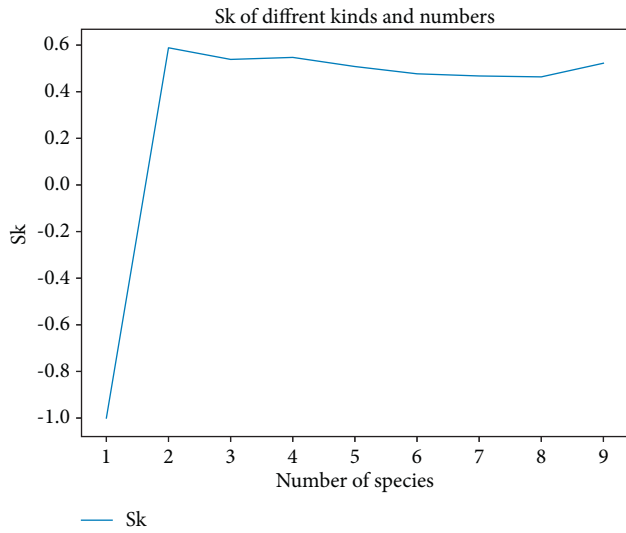


FIGURE 14: Cluster contour coefficients of different cluster numbers.

TABLE 2: Training test dataset.

Grouping information	The number of training samples	The number of test samples
Group I	360	102
Group II	360	104

$$z_i = \frac{x_i - \min_{1 \leq f \leq F} \{x_f\}}{\max_{1 \leq f \leq F} \{x_f\} - \min_{1 \leq f \leq F} \{x_f\}}. \quad (6)$$

In equation (6),  $z_i$  represents the preprocessed data,  $x_i$  represents the frequency value of the vibration signal,  $\min_{1 \leq f \leq F} \{x_f\}$  and  $\max_{1 \leq f \leq F} \{x_f\}$  represent the minimum and maximum values of frequency in each group of vibration signals, and  $f$  represents the number of each vibration signal.

3.3.3. *Diagnostic Result Output.* In order to evaluate the difference between the normalized prediction result and the corresponding sample label, the cross entropy function is used to calculate the error loss value. The mathematical expression is as follows:

$$J(\theta) = -\frac{1}{m} \sum_{i=1}^m \sum_{r=1}^t \left( I\{y^{(i)} = r\} \times \log \frac{e^{x_i^t \Delta c_k}}{\sum_{k=1}^m e^{x_i^t \Delta c_k}} \right). \quad (7)$$

In equation (7),  $J(\theta)$  represents the loss function,  $I\{\Delta\}$  represents the logical indication function (when the value is true,  $I=1$ , otherwise  $I=0$ ), and  $y^{(i)}$  represents the  $i$ -th real label of the fault.

The weight matrix  $\theta$  is iteratively updated by means of gradient descent. The iterative equation is as follows:

$$\theta_j = \theta_j - \alpha \frac{\partial J(\theta)}{\partial \theta_j}. \quad (8)$$

In equation (8),  $\theta_j$  represents the weight matrix of the  $j$ -th update.

3.3.4. *Model Parameter Structure.* The experiment was implemented on a Linux computer using Pycharm platform, Python as the programming language, and PyTorch deep learning framework.

During network training based on stochastic gradient descent, the multilayer back-propagation of the error signal can easily lead to “gradient dispersion” (too small gradient will make the returned training error signal extremely weak) or “gradient explosion” (too large gradient will lead to Nan in the model). With the increase of network depth, training becomes more and more difficult. Considering the network lightweight, during the experiment, the Adam optimizer is used to continuously update the network training parameters. The batch size is set to 30 and the number of iterations is 200. This study introduces the early stopping mechanism. By monitoring the changing value of the training set loss function between adjacent iterations during the training process, early stopping can terminate the model training in time to prevent the model from overfitting. The learning rate is 0.0005. The model is built on the basis of convolutional neural network model, so the parameter design is similar to the convolutional neural network, and the parameter design is shown in Table 3.

TABLE 3: Parameter design.

The number of layers	Structure name	Structural parameters	The number of channels	Output size
	Input	(1,3600)	1	(1,3600)
1	Convolution layer	(1,8,2)	16	(1,1797)
2	Convolution layer	(1,8,2)	16	(1,895)
3	Pool layer	(1,2)		(1,447)
4	Convolution layer	(1,4,2)	64	(1,222)
5	Convolution layer	(1,4,2)	64	(1,110)
6	Pool layer	(1,2)		(1,55)
7	Convolution layer	(1,4,2)	256	(1,26)
8	Convolution layer	(1,4,2)	256	(1,12)
9	Pool layer	(1,2)		(1,6)
10	Convolution layer	(1,2,1)	512	(1,5)
11	Convolution layer	(1,2,1)	512	(1,4)
12	Pool layer	(1,2)		(1,2)
13	Full connection layer	(1024)		(1024)
14	Full connection layer	(50)		(50)
15	Output layer	(9)		(9)

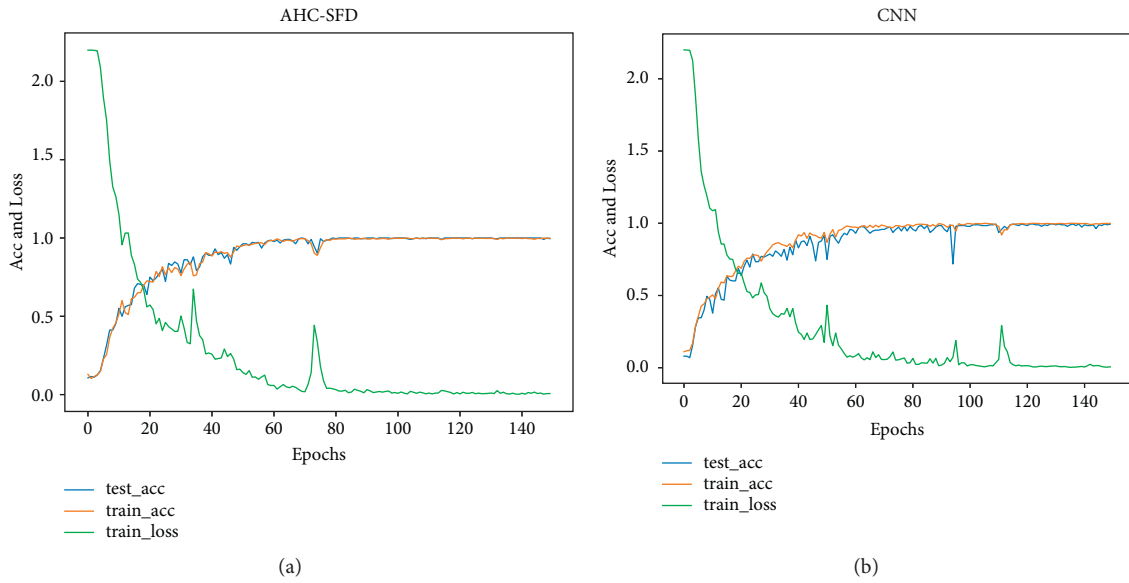


FIGURE 15: AHC-SFD and CNN experimental results.

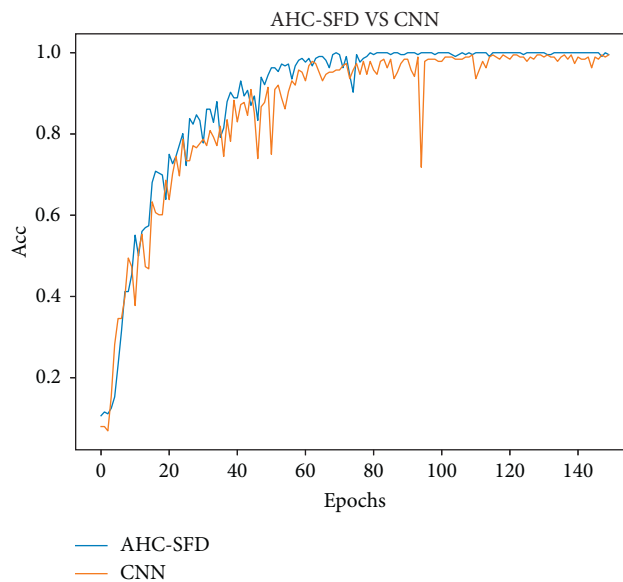


FIGURE 16: AHC-SFD Vs CNN.



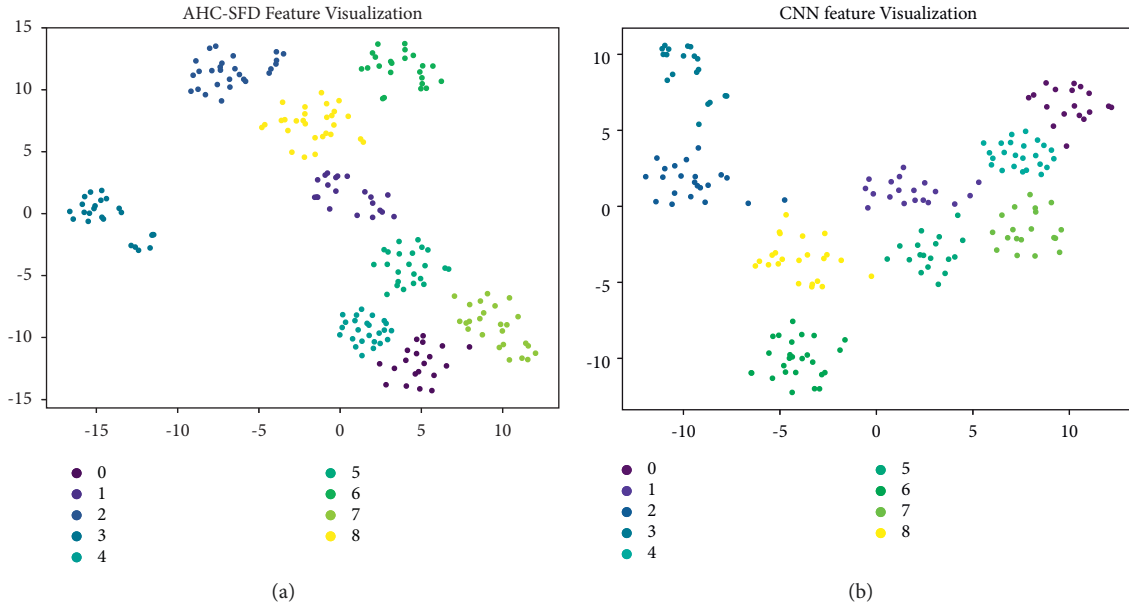


FIGURE 17: AHC-SFD and CNN feature visualization.

**3.4. Result Analysis.** To verify whether the method has a high diagnostic rate and good generalization ability, the experimental results in this study are compared with those using only the CNN. The experimental results are shown in Figure 15.

The comparison results of AHC-SFD and CNN on the test set are shown in Figure 16.

It can be seen from the comparison results in Figures 15 and 16 that after 140 epochs, the accuracy of AHC-SFD algorithm on the test set reaches 99.7%, while the accuracy of the CNN algorithm on the test set is only 98.9%. Therefore, the diagnostic methods in this study tend to be faster, more stable, with higher accuracy and stronger generalization ability.

In order to further demonstrate the learning ability of the model for different categories of features, the t-SNE dimension reduction algorithm in flow pattern learning is introduced to visualize the features learned by the full connected layer. The experimental results are shown in Figure 17.

It can be seen from the scatter plot Figure 17 that the method AHC-SFD in this study has identification errors in the samples of class 0 and class 7, and the other samples are gathered at the corresponding positions. However, CNN features have recognition errors in class 1, class 2, class 5, and class 8 samples, and there are many overlaps in class 1 and class 5 samples. It can be seen that AHC-SFD has stronger feature learning ability than the CNN.

## 4. Conclusion

The AHC-SFD algorithm established in this study is a diagnosis algorithm based on adaptive hierarchical clustering and subset, which has the following three advantages: (1) the AHC-SFD algorithm directly takes the

original vibration signal as the input of 1D-CNN, which can obtain the characteristics of vibration signal to the greatest extent. (2) A grouping method based on adaptive hierarchical clustering is proposed, which analyzes the characteristics of different data and then clusters the dataset into multiple feature groups. (3) A multiscale feature extraction module is proposed to reduce the misclassification of similar samples, thus ensuring the maximum extraction of effective information into the data. It is verified on the gearbox dataset that the diagnostic accuracy is better than the single-channel CNN model.

## Data Availability

The data set used in this article can be obtained from the corresponding author upon request.

## Conflicts of Interest

The authors declare that they have no conflicts of interest regarding this work.

## Acknowledgments

This work was supported by Innovation Ability Improvement Project of Scientific and Technological Small and Medium-Sized Enterprises in Shandong Province (grant no. 2021TSGC1089), “20 New Colleges and Universities” Funded Project in Jinan (grant no. 2021GXRC074), Major Scientific and Technological Innovation Projects in Shandong Province (grant no. 2019JZZY010117), and 2020 Industrial Internet Innovation and Development Project, Solution Application and Promotion Public Service Platform (grant no. TC200802C).

## References

- [1] Z. Hou, "Research status and development prospect of rotating machinery fault diagnosis," *Forging equipment and manufacturing technology*, vol. 56, no. 5, pp. 33–37, 2021.
- [2] X. Zhao, "Automatic on-line monitoring and fault diagnosis system for mine electromechanical equipment," *Mining equipment*, vol. 11, no. 6, pp. 246–247, 2021.
- [3] G. Fan, "Research on on-line monitoring and fault diagnosis of secondary circuit in intelligent substation," *Light source and lighting*, vol. 45, no. 2, pp. 228–230, 2022.
- [4] B. Shen, B. Chen, C. Zhao, F. Chen, W. Xiao, and N. Xiao, "A review of research on deep learning in mechanical equipment fault prediction and health management," *Machine tools and hydraulics*, vol. 49, no. 19, pp. 162–171, 2021.
- [5] Y. Lei, F. Jia, and X. Zhou, "A deep learning-based method for machinery health monitoring with big data," *Journal of Mechanical Engineering*, vol. 51, no. 21, pp. 49–56, 2015.
- [6] F. Jia, Y. Lei, J. Lin, X. Zhou, and N. Lu, "Deep neural networks: a promising tool for fault characteristic mining and intelligent diagnosis of rotating machinery with massive data," *Mechanical Systems and Signal Processing*, vol. 72–73, pp. 303–315, 2016.
- [7] W. Yu and C. Zhao, "Robust monitoring and fault isolation of nonlinear industrial processes using denoising autoencoder and Elastic Net," *IEEE Transactions on Control Systems Technology*, vol. 27, pp. 1–9, 2019.
- [8] H. Mu, *Rolling Bearing Fault Diagnosis Method Based on Integrated Soft Competition Yu Norm Art*, Wuhan University of science and technology, Wuhan, China, 2017.
- [9] Y. Tang, "Application of integrated diagnosis method in transformer fault diagnosis," *Coal mine machinery*, vol. 33, no. 5, pp. 264–266, 2012.
- [10] V. H. Nguyen, J. S. Cheng, Y. Yu, and V. T. Thai, "An architecture of deep learning network based on ensemble empirical mode decomposition in precise identification of bearing vibration signal," *Journal of Mechanical Science and Technology*, vol. 33, no. 1, pp. 41–50, 2019.
- [11] G. Chen, J. Zhang, and G. Kan, "Intelligent fault diagnosis method of bearing based on improved superposition automatic encoder," *Noise and vibration control*, vol. 42, no. 1, pp. 156–161, 2022.
- [12] S. Liu, *Research on Bearing Fault Diagnosis Based on Stack Automatic Encoder*, Taiyuan University of science and technology, Taiyuan, China, 2020.
- [13] D. T. Hoang and H. J. Kang, "Rolling element bearing fault diagnosis using convolutional neural network and vibration image," *Cognitive Systems Research*, vol. 53, pp. 42–50, 2019.
- [14] C. Wei, J. Zhou, and J. Zhang, "FDM 3D printing fault diagnosis method based on," *Agricultural equipment and vehicle engineering*, vol. 60, no. 2, pp. 149–153, 2022.
- [15] Ke Zhang, J. Wang, H. Shi, X. Zhang, and L. Fu, "Research on fault diagnosis of rolling bearing under variable working conditions based on," *Control engineering*, vol. 29, no. 2, pp. 254–262, 2022.
- [16] Y. Ye and Y. Li, "Multi wind turbine fault diagnosis based on CNN ensemble learning," *Journal of Industrial Engineering*, vol. 25, no. 1, pp. 136–143, 2022.
- [17] L. Wen, X. Li, and L. Gao, "A transfer convolutional neural network for fault diagnosis based on ResNet- 50," *Neural Computing & Applications*, vol. 31, pp. 1–14, 2019.
- [18] J. Ding, Q. Shao, Z. Qi, M. Xie, Bo Gao, and Yu Yang, "Convolution neural network fault diagnosis based on transfer learning," *Science, technology and engineering*, vol. 22, no. 14, pp. 5653–5658, 2022.
- [19] W. Abed, S. Sharma, R. Sutton, and A. Motwani, "A robust bearing fault detection and diagnosis technique for brushless DC motors under non-stationary operating conditions," *Journal of Control, Automation and Electrical Systems Automation and Electrical Systems*, vol. 26, no. 3, pp. 241–254, 2015.
- [20] M. Chang, *Fault Diagnosis and Prediction of Wind Power Rolling Bearing Based on Deep Learning*, Jiangnan University, Wuxi, China, 2021.
- [21] H. Pan, X. He, and S. Tang, "An improved bearing fault diagnosis method using one-dimensional CNN and LSTM," *Journal of Mechanical Engineering*, vol. 64, no. 7/8, pp. 443–452, 2018.
- [22] P. Zhang, X. Shu, X. Li, J. Hang, S. Ding, and Q. Wang, "Research on fault diagnosis method of AC motor system based on LSTM," *Journal of electrical machinery and control*, vol. 26, no. 3, pp. 109–116, 2022.
- [23] Y. Li, J. Hu, J. Lai, W. Wang, Y. Zhao, and Y. Fan, "Fault diagnosis of wind turbine planetary gearbox based on 1d-cnn-lstm hybrid neural network model," *Electrical automation*, vol. 43, no. 5, pp. 20–22+26, 2021.
- [24] J. Bai, Y. Wu, J. Zhang, and F. Chen, "Subset based deep learning for RGB-D object recognition," *Neurocomputing*, vol. 165, pp. 280–292, 2015.
- [25] A. T. Duong, H. T. Phan, and N. D. H. Le, *A Hierarchical Approach for Handwritten Digit Recognition Using Sparse Autoencoder. Issues and Challenges of Intelligent Systems and Computational Intelligence*, Springer, Newyork, NY, USA, 2014.
- [26] Y. Zhang, X. Li, L. Gao, and P. Li, "A new subset based deep feature learning method for intelligent fault diagnosis of bearing," *Expert Systems with Applications*, vol. 110, pp. 125–142, 2018.

## Research Article

# Feature Selection Based on Adaptive Particle Swarm Optimization with Leadership Learning

Zhiwei Ye <sup>1</sup>, Yi Xu <sup>1</sup>, Qiyi He <sup>1</sup>, Mingwei Wang,<sup>1</sup> Wanfang Bai,<sup>2</sup> and Hongwei Xiao<sup>3</sup>

<sup>1</sup>School of Computer Science, Hubei University of Technology, Wuhan 430070, China

<sup>2</sup>Xining Big Data Service Administration, Xining 810000, China

<sup>3</sup>Xining Zhiyun Digital Economy Research Institute, Xining 810000, China

Correspondence should be addressed to Qiyi He; [qiyi.he@hbut.edu.cn](mailto:qiyi.he@hbut.edu.cn)

Received 29 June 2022; Revised 7 August 2022; Accepted 9 August 2022; Published 28 August 2022

Academic Editor: Nian Zhang

Copyright © 2022 Zhiwei Ye et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

With the rapid development of the Internet of Things (IoT), the curse of dimensionality becomes increasingly common. Feature selection (FS) is to eliminate irrelevant and redundant features in the datasets. Particle swarm optimization (PSO) is an efficient metaheuristic algorithm that has been successfully applied to obtain the optimal feature subset with essential information in an acceptable time. However, it is easy to fall into the local optima when dealing with high-dimensional datasets due to constant parameter values and insufficient population diversity. In the paper, an FS method is proposed by utilizing adaptive PSO with leadership learning (APSOLL). An adaptive updating strategy for parameters is used to replace the constant parameters, and the leadership learning strategy is utilized to provide valid population diversity. Experimental results on 10 UCI datasets show that APSOLL has better exploration and exploitation capabilities through comparison with PSO, grey wolf optimizer (GWO), Harris hawks optimization (HHO), flower pollination algorithm (FPA), salp swarm algorithm (SSA), linear PSO (LPSO), and hybrid PSO and differential evolution (HPSO-DE). Moreover, less than 8% of features in the original datasets are selected on average, and the feature subsets are more effective in most cases compared to those generated by 6 traditional FS methods (analysis of variance (ANOVA), Chi-Squared (CHI2), Pearson, Spearman, Kendall, and Mutual Information (MI)).

## 1. Introduction

Large amounts of data have been generated in various fields such as social media, healthcare, cybersecurity, and education in the past decades, and edge computing provides an effective solution for data storage and transmission. However, as the dimensionality of the data increases, the curse of dimensionality problem becomes common, which has a negative impact on the stability, security, and computational efficiency of edge computing. Feature selection (FS) is a data preprocessing technique in machine learning and data mining that has been applied to improve the performance of edge computing by eliminating irrelevant and redundant features in the datasets [1–3]. In general, it is a combinatorial optimization problem [4, 5] that tries to find the optimal feature subsets with essential information from the original datasets. Given a dataset with  $N$  features, there will be  $2^N$

possible feature subsets, and the search space rises exponentially as the number of features increases [6, 7]. Hence, some traditional FS methods have received considerable interest due to their ability to evaluate feature importance and select a certain number of top-ranked features. These methods include statistical test (e.g., analysis of variance (ANOVA) [8, 9] and Chi-Squared (CHI2) [10, 11]), correlation criteria (e.g., Pearson [12], Spearman [13, 14], Kendall [15, 16]), and information theory (e.g., symmetrical uncertainty (SU) [17], mutual information (MI) [18, 19], and entropy [20]). However, the statistical test and correlation criteria techniques only consider the correlation between features and labels, and the feature subsets are not appropriate because some highly correlated but redundant features are selected. As a result, information theory techniques are applied to FS problems owing to their consideration of redundancy between features as well. Moreover, the

redundancy calculation only focuses on the interaction between two features and fails to identify those of multiple features [21], which may ignore some important features. Therefore, how to find suitable feature subsets efficiently needs to be further investigated.

Metaheuristic algorithms such as monarch butterfly optimization (MBO) [22], slime mold algorithm (SMA) [23], moth search algorithm (MSA) [24], hunger games search (HGS) [25], hybrid rice optimization (HRO) [26], colony predation algorithm (CPA) [27], weighted mean of vectors (INFO) [28], grey wolf optimizer (GWO) [29], clonal flower pollination algorithm (FPA) [30], salp swarm algorithm (SSA) [31], Harris hawks optimization (HHO) [32], and particle swarm optimization (PSO), have been used to solve combinatorial optimization problems because of their dynamic exploration and exploitation capabilities in the search space, some of which have shown to be successful in FS problems [33, 34]. For instance, Shen and Zhang [29] proposed a two-stage GWO for processing biomedical datasets, which showed better performance in terms of time consumption and classification accuracy by removing more than 95.7% of the redundant features. Hussain et al. [32] developed an FS method based on HHO, which removed 87% of features and achieved 92% of classification accuracy. Yan et al. [30] presented a binary clonal FPA for some biomedical datasets, which enhanced population diversity and selected fewer features with strong robustness. Balakrishnan et al. [31] designed an FS method based on salp SSA, which increased the ability of particles to explore different regions by randomly updating their position and improved the confidence level by 0.1033% on 6 datasets. However, a series of parameters need to be set by users in these metaheuristic algorithms, and unsuitable parameters may lead to slow convergence and local stagnation. A lot of experiments and extensive experience are needed to find the appropriate parameter settings.

Compared with the above metaheuristic algorithms, PSO is applied to solve FS problem of its fast convergence and few parameters. However, the exploration and exploitation capabilities are influenced by parameter setting and population diversity as the number of features increases. Therefore, some improved PSO based on parameter updating and population diversity updating strategies have been proposed for FS. For example, Song et al. [35] developed a three-phase hybrid FS algorithm, which reduced the computational cost by using correlation-guided clustering and an improved integer PSO. Tran et al. [36] used a bare-bones PSO for FS, which reduced the search space of the problem and improve the search efficiency. Song et al. [37] also introduced a variable-size cooperative coevolutionary PSO for high-dimensional datasets, which divided a high-dimensional FS problem into multiple low-dimensional subproblems with a low computational cost. Hu et al. [38] presented a multi-objective PSO for FS, which achieved superior performances in approximation, diversity, and feature cost by introducing a tolerance coefficient. Hosseini Bamakan et al. [39] proposed a time-varying PSO-based FS method to deal with the network intrusion detection problem, which obtained a higher detection rate and lower

false alarm rate by introducing a chaotic concept and time-varying parameters. Mafarja et al. [40] proposed a binary PSO-based FS method, which adopted a time-varying inertia weighting strategy and showed a superior convergence rate on some datasets. Huang et al. [41] utilized cut-point and feature discretization to expand the searching scope of PSO for gene expression datasets, which selected fewer features and maintained similar classification accuracy. Xue et al. [42] introduced adaptive parameters in PSO for high-dimensional datasets, which allowed particles to automatically adjust parameters during the search process and decreased time consumption. Moradi and Gholampour [43] used a PSO with the local search strategy for high-dimensional datasets, which adjusted the search process by considering the correlation information between distinct features. Chen et al. [44] introduced an FS method based on hybrid PSO and differential evolution (HPSO-DE), which enhanced population diversity by adopting mutation, crossover, and selection operators. Although the optimization ability of PSO is improved to some extent by the above techniques, the randomness of the search process may be increased and they lack consideration for jumping out of the local optima.

In the paper, an FS method based on adaptive PSO with leadership learning (APSOLL) is proposed, which combines parameter updating and population diversity updating strategies to compensate for the shortcomings of PSO. The adaptive updating strategy for parameters is used to guide particles to search in a more reasonable scope, and the leadership learning strategy is utilized to enhance population diversity. Overall, the main contributions of our work are as follows:

- (1) Based on the population state, an adaptive updating strategy for parameters is proposed to replace the constant parameters which guide particles to search in a more reasonable scope.
- (2) Adopting leadership learning strategies to provide valid population diversity by learning from the first three leaders in the population that enhances the exploration and exploitation capabilities of PSO.
- (3) The effectiveness of the proposed method is verified by comparing it with six traditional methods (ANOVA, CHI2, Pearson, Spearman, Kendall, and MI) and seven metaheuristic algorithms-based FS methods (GWO, HHO, FPA, SSA, LPSO, and HPSO-DE).

## 2. Background and Related Work

*2.1. Overview of PSO.* PSO is a population-based metaheuristic algorithm for simulating the predatory activities of bird and fish populations [45, 46], and each particle in the population has two properties: velocity vector  $v_i = (v_{i1}, v_{i2}, \dots, v_{id})$  and position vector  $x_i = (x_{i1}, x_{i2}, \dots, x_{id})$ , where  $d$  denotes the dimension. In the search process of PSO, the velocity vectors are dynamically adjusted by the personal best position ( $pbest_i$ ) and the global best position ( $gbest$ ) at the current stage, and the position vectors are the candidate solutions to the optimization problems, all of which are updated by equations (1)–(2).

$$v_i(t+1) = \omega \times v_i(t) + c_1 r_1 (p \text{ best}_i - x_i(t)) + c_2 r_2 (g \text{ best} - x_i(t)), \quad (1)$$

$$x_i(t+1) = x_i(t) + v_i(t+1). \quad (2)$$

where  $v_i$  and  $x_i$  represent the velocity and position vectors of the  $i$ -th ( $i=1, 2, \dots, N$ ) particle, and the upper and lower limits of each dimension are set to 1 and 0, respectively.  $\omega$  is defined as the inertia parameter, and it is a non-negative number.  $c_1$  and  $c_2$  are acceleration parameters, and the former represents the personal learning parameter and the latter represents the global learning parameter, which is used to control the search scope of particles and set by users.  $r_1$  and  $r_2$  are random numbers in  $[0, 1]$ .

**2.2. The Leadership Learning Strategy.** Leadership learning strategy is a management concept that describes the dynamic process of feed-forward and feedback in a living system. Hirst et al. [47] suggested that learning activities of individuals will affect the decisions of leaders, and it is called feed-forward learning flow. Moreover, effective leaders may quickly identify key information in group development and have a lasting impact on the individuals and group activities through their decisions in turn, which is regarded as feedback learning flow. In the model of leadership learning strategy, feed-forward and feedback learning flow among individuals, groups, and leaders together determine the scope of the system development, and the framework is shown in Figure 1.

Based on the leadership learning strategy, GWO was proposed with effective exploration capability and acceptable time consumption by learning from the first three best solutions (leaders) of each iteration [48–51]. In the search process, the population is divided into four levels, sequentially  $\alpha$ ,  $\beta$ ,  $\delta$ , and  $\omega$ , where  $\alpha$ ,  $\beta$ , and  $\delta$  are regarded as leaders, the remaining particles  $\omega$  are considered as individuals, and the population is considered group. Moreover, the particles and leaders learning from each other are considered as the leadership learning strategy, and it is shown in Equation (3).

$$\begin{aligned} \vec{X}_1 &= \vec{X}_\alpha - \vec{A}_1 \times \vec{D}_\alpha, \vec{X}_2 \\ &= \vec{X}_\beta - \vec{A}_2 \times \vec{D}_\beta, \vec{X}_3 \\ &= \vec{X}_\delta - \vec{A}_3 \times \vec{D}_\delta. \end{aligned} \quad (3)$$

where  $\vec{X}_\alpha$ ,  $\vec{X}_\beta$ , and  $\vec{X}_\delta$  are position vectors of  $\alpha$ ,  $\beta$ , and  $\delta$ .  $\vec{D}_\alpha = |\vec{C}_1 \times \vec{X}_\alpha - \vec{X}|$ ,  $\vec{D}_\beta = |\vec{C}_2 \times \vec{X}_\beta - \vec{X}|$ ,  $\vec{D}_\delta = |\vec{C}_3 \times \vec{X}_\delta - \vec{X}|$  denote the distance between particles and leaders.  $\vec{C}_1$ ,  $\vec{C}_2$ , and  $\vec{C}_3$  are random numbers from 0 to 2. The search scope of particles is controlled by the convergence factor  $\vec{A}$ , which is computed as Equation (4).

$$\vec{A} = 2a \times r_3 - a, \quad (4)$$

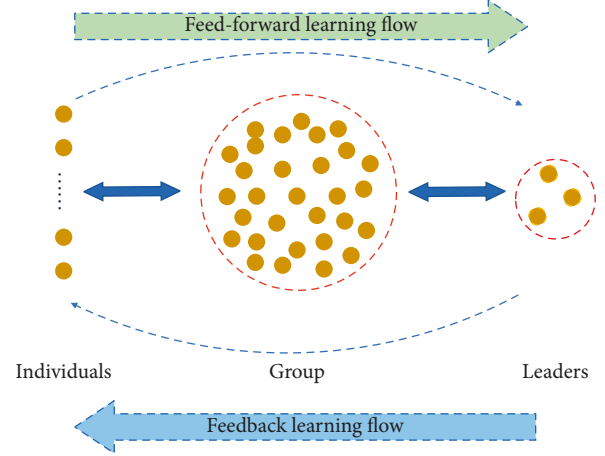


FIGURE 1: The framework of leadership learning.

where the variable  $a = 2(1 - t/T)$  is the control coefficient ( $T$  denotes the maximum number of iterations), and it decreases linearly from 2 to 0 during the search process.

### 3. The Proposed Method

In this section, an FS method based on APSOLL is presented to conduct classification on 10 UCI datasets. The corresponding techniques for the proposed method are described as follows:

**3.1. Adaptive Updating Strategy for Parameters.** During the search process of PSO, the search scope of particles is affected by convergence factor  $c_1$  and  $c_2$ . In general, they are usually less than 2 and set to constant values by users [52–54]. However, the population is dynamically changed according to the optimal fitness value, it is appropriate to adaptively adjust  $c_1$  and  $c_2$  for better exploration and exploitation. Moreover, the change of fitness value during the iteration reflects the state of the population, thus the adaptive updating strategy is proposed based on this case, and it is used to replace the convergence factor, which is shown in equations (5)–(6).

$$m = \begin{cases} m + 1, & \text{if fitness}(t) = \text{fitness}(t-1), 0, \text{ otherwise,} \\ 0, & \end{cases} \quad (5)$$

$$c = \left(\frac{m}{T}\right)^{2/3} + 1, \quad (6)$$

where  $m$  is a variable and initially set to 0, and it is increased by 1 if the fitness value is improved in the next iteration, otherwise the value of which is always 0. Thus,  $c$  is dynamically changed between 1 and 2 during the search process, and it is gradually increased if the algorithm falls into the local optima.

**3.2. The Search Process of Leadership Learning Strategy.** The population diversity of PSO may be inadequate due to the strategy learned from  $p \text{ best}_i$  and  $g \text{ best}$ . Smith [55] proposed that the more leaders of individuals engage feed-forward and feedback in a living system, the more possible it is for the

group to change, innovate, and cooperate. However, the time consumption will increase as the number of leaders increases during the process. Therefore, inspired by GWO, the leadership learning strategy from 3 leaders is used to reconstruct the velocity vectors of PSO, which will increase population diversity and provide more accurate information for better exploration and exploitation. In addition, an adaptive parameter  $c$  is combined to guide the particles to search in a more reasonable scope, and the process is shown in Equation (7).

$$\begin{aligned} v_i(t+1) &= \omega \Delta v_i(t) + \frac{c}{2} \times r_4 (\vec{X}_1 - x_i(t)) \\ &+ \frac{c}{3} \times r_4 (\vec{X}_2 - x_i(t)) \\ &+ \frac{c}{4} \times r_4 (\vec{X}_3 - x_i(t)), \end{aligned} \quad (7)$$

where  $\vec{X}_1$ ,  $\vec{X}_2$  and  $\vec{X}_3$  represent the leadership learning strategy.  $r_4$  is a random number between 0 and 1.  $c$  is updated by (6), it is dynamically changed between 1 and 2 during the search process, and it is gradually increased if the algorithm falls into the local optima. The cooperation of  $c/2$ ,  $c/3$  and  $c/4$  will allow particles to search in a more reasonable scope with higher possibilities.

As for the leadership learning strategy, Hu et al. [50] proposed that the convergence factor  $|\vec{A}|$  greater than 1 shows better exploration capability and less than 1 shows better exploitation capability. However, it can be seen from (4) that  $|\vec{A}|$  is linearly decreased and always less than 1 in the last 50% of iterations, and the exploration capability is insufficient when the algorithm is trapped in the local optima in this case. Hence, it is considered to increase the possibility that  $|\vec{A}|$  is greater than 1 at this stage and it is modified as shown in Equation (8).

$$\vec{A} = 2^c a \times r_5 - a. \quad (8)$$

where  $r_5$  is a random number in  $[0, 1]$ , and  $|\vec{A}|$  is adaptively changed during the search process. It will be greater than 1 with a higher possibility and thus enhance the exploration capability when the algorithm falls into the local optima.

**3.3. The Encoding Schema.** The core object of the proposed method is to select a suitable expression form for FS and establish a reasonable mapping between the solutions and the feature subsets. The candidate solutions that are binarized are used to represent the features, where “1” denotes the feature is selected and “0” illustrates the feature is abandoned. For instance, there is a feature dataset with 10 features, and the candidate solution is coded as 1010000011, which means the 1st, 3rd, 9th, and 10th features are selected and the others are abandoned. The position vector of each particle is binarized according to Equation (9).

$$Xb_{id} = \begin{cases} 1, & \text{if } x_{id} > 0.5, \\ 0, & \text{otherwise,} \end{cases} \quad (9)$$

where  $Xb_i = (Xb_{i1}, Xb_{i2}, \dots, Xb_{id})$ ,  $i$  and  $d$  denote the number of particles and the number of features, respectively.

**3.4. The Definition of Objective Function.** The feature subsets generated by FS methods for classification have two main goals, which are maximizing the classification accuracy (minimizing the classification error) and minimizing the number of selected features. As a mainstream classifier, K nearest neighbor (KNN) [56–58] is utilized for FS due to its advantages of simplicity and insensitivity to noisy data. Furthermore, how to reduce the number of selected features is considered another core issue. The ultimate goal is to obtain the optimal feature subsets with essential information from the original datasets while achieving higher classification accuracy with fewer features. Hence, the objective function that combines the classification accuracy and the number of selected features is adopted and it is defined as Equation (10).

$$\text{Fitness}(X) = \theta \times \text{acc}(X) + (1 - \theta) \times \left(1 - \frac{\#X}{N}\right). \quad (10)$$

where  $\text{acc}(X)$  denotes the classification accuracy of the feature subsets,  $\#X$  and  $N$  represent the number of features in the feature subset and the original dataset.  $\theta$  is a weighting factor to balance the classification accuracy and the number of selected features, and it is set to 0.7.

**3.5. Implementation of the Proposed Method.** The main process of APSOLL is to search for the optimal feature subsets with essential information from the original datasets and apply it for classification, and the pseudocode is shown in Algorithm 1. Among these, the particles are binarized to determine the corresponding feature subsets in each iteration, and the leaders are determined by computing the fitness function, which is used to guide the search process. Figure 2 shows the flowchart of APSOLL. When the algorithm starts running, it randomly initializes the velocity vector  $v_i$ , position vector  $x_i$ ,  $pbest_i$ ,  $gbest$ , and sets  $m=0$  and  $t=0$ . In each iteration, the fitness value of each particle is calculated in order to find the optimal three solutions (leaders). Based on the information provided by the leader, the velocity of the particles and the position of the population are updated. In this process, if the optimal fitness value is not changed, the adaptive parameter  $m$  is added by 1. The algorithm run is ended and the optimal solution is binarized when the maximum number of iterations is reached.

## 4. Experimental Design

All experimental procedures are implemented using *Python* 3.8 in a PC with Intel(R) Core (TM) i5-9400 @ 2.9 GHz CPU, and 16 GB DDR4 of RAM under Windows 10 Operating System. 10 public datasets are used to assess the quality of the proposed method. APSOLL is compared with 7 meta-heuristic algorithms to evaluate the optimization ability, and 6 traditional FS methods such as ANOVA, CHI2, Pearson, Spearman, Kendall, and MI are used to analyze the effectiveness of the feature subsets selected by the proposed method.

Input: the number of iterations  $T$ , population size  $N$   
 Output: The classification accuracy and the number of features among the feature subsets  
 Initialization:  $x_i = (x_{i1}, x_{i2}, \dots, x_{id})$   
 Set  $ub = 1$ ,  $lb = 0$ ,  $m = 0$ , initial iteration  $t = 0$   
 while  $t < T$  do  
   Binarize each particle by using Equation (9)  
   Compute the fitness value of each particle by using Equation (10)  
   Update  $x_\alpha$ ,  $x_\beta$ , and  $x_\delta$   
   Update  $c$  by Equation (6)  
   Update  $|\bar{A}|$  by Equation (8)  
   Compute  $\bar{X}_1$ ,  $\bar{X}_2$ , and  $\bar{X}_3$  by using Equation (3)  
   Update the velocity of each particle by using Equation (7)  
   Update the population position by using Equation (2)  
    $t = t + 1$   
 end while  
   Binarize  $x_\alpha$  by using Equation (9)  
 return the fitness value and the feature subset

ALGORITHM 1: FS based on APSOLL.

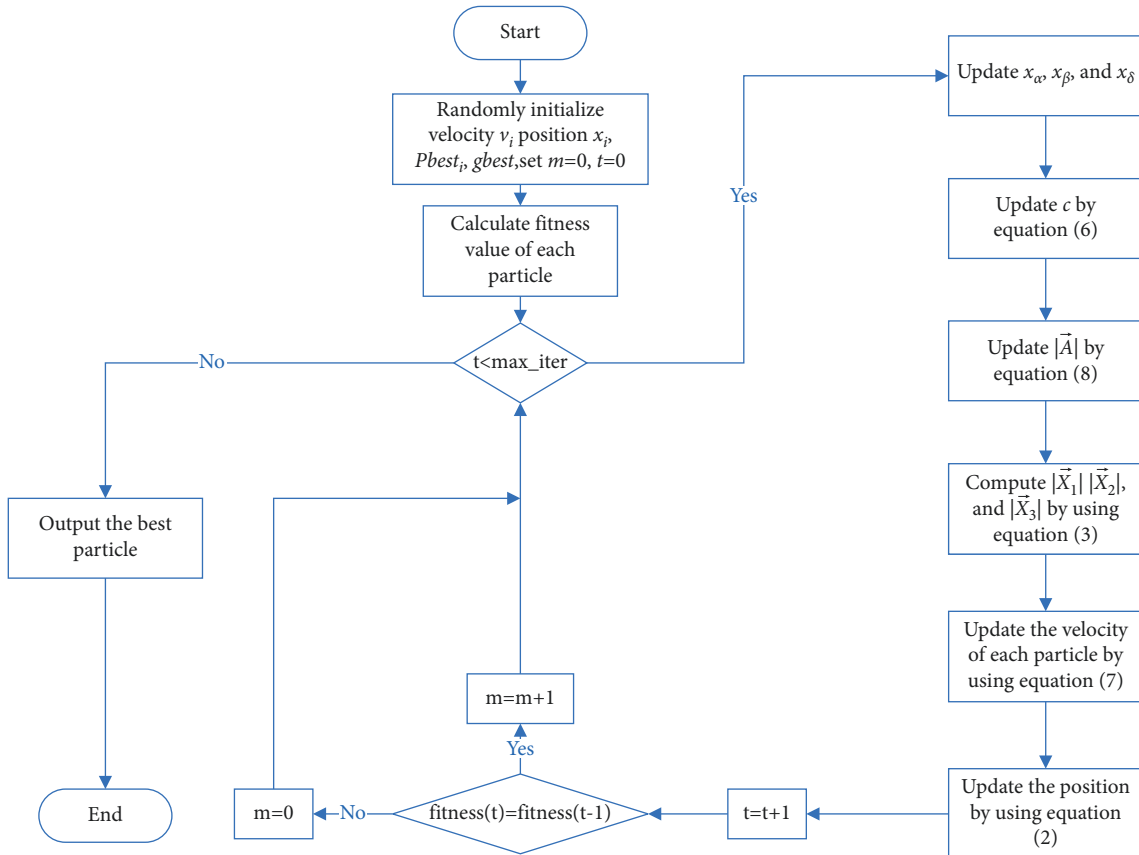


FIGURE 2: The flowchart of APSOLL.

**4.1. Datasets Description.** 10 datasets from the UCI machine learning database are used to evaluate the performance of the proposed method, including myocardial infarction complications (MIC), urban, SCADI, arrhythmia, madelon, isolet5, multiple features (MF), Parkinson's disease (PD), CNAE-9, and QSAR, all of which have more than 100 features, with the number of classes ranging from 2 to 26

and instances ranging from 69 to 2600, and the details of datasets are shown in Table 1. In the experiments, each dataset is randomly divided into two parts: a total of 70% of the instances are chosen as the training data, and the remaining 30% are used as the testing data. Li et al. [54] described in detail why the dataset dividing approach was adopted.

TABLE 1: Details of datasets.

Dataset	Number of features	Number of instances	Number of classes
MIC	124	1700	7
Urban	147	507	9
SCADI	205	69	6
Arrhythmia	279	452	13
Madelon	500	2600	2
Isolet5	617	1559	26
MF	649	2000	10
PD	754	756	2
CNAE-9	857	1080	9
QSAR	1024	1687	2

TABLE 2: Parameters Setting of different metaheuristic algorithms.

Algorithms	Parameters	Values
Common settings	Number of iterations	$T = 100$
	Population size	$N = 30$
	The upper limit of particle position	$ub = 1$
	The lower limit of particle position	$lb = 0$
GWO	Correlation coefficient	$a$ decreases linearly from 2 to 0
PSO	Acceleration factor	$c_1 = 2, c_2 = 2$
	Inertia weight	$w = 0.9$
HHO	Levy component	$\beta = 0.8$
FPA	Acceleration factor	$c_1 = 2, c_2 = 2$
	Levy component	$\beta = 1.5$
	Switch probability	$P = 0.8$
SSA	Convergence factor	$C$ decreases linearly from 2 to 0
LPSO	Acceleration factor	$c_1 = 2, c_2 = 2$
	Upper limit of inertia weight	$wmax = 0.9$
	Lower limit of inertia weight	$wmin = 0.4$
HPSO-DE	Acceleration factor	$c_1 = 2, c_2 = 2$
	Crossover rate	$CR = 0.2$
	Scaling factor	$F = 0.5$
	Predefined generation	$G = 5$
APSOLL	Inertia weight	$w = 0.9$

4.2. *Parameters Setting for Metaheuristic Algorithms.* As for APSOLL, the search process requires only one inertia weight parameter  $\omega$  to be set. In addition, some commonly used FS methods based on metaheuristic algorithms are adopted to evaluate the optimization ability, such as GWO, PSO, HHO, FPA, SSA, LPSO, and HPSO-DE. Among them, LPSO [40] and HPSO-DE [44] are classical benchmark PSO-based FS methods by adopting parameter updating and population diversity updating strategies, respectively. The parameters of each metaheuristic algorithm are set based on the published literature, which is shown in Table 2. Furthermore, the binary encoding scheme is utilized for each metaheuristic algorithm and it is run independently 30 times to take the average as the result in order to eliminate the influence of randomness.

## 5. Results and Discussion

5.1. *Experimental Results of Different Metaheuristic Algorithms.* The optimization ability of APSOLL is evaluated from the fitness value, classification accuracy, number of selected features, and CPU time. The average convergence

curves of the fitness value are shown in Figures 3-4, and the number of selected features in the search process is shown in Figures 5-6. In the experiment, the  $t$ -test with a significance level of 0.05 is used to determine whether the results obtained from the proposed algorithm are statistically significantly different from other metaheuristic algorithms, and the experimental results are presented in Tables 3-4, where  $Fit$ ,  $Acc$ , and  $\#F$  denote the fitness values, classification accuracy and number of selected features after 30 independent runs, and  $Time$  presents the CPU time of the whole process (in seconds).  $S_{fit}$ ,  $S_{acc}$ , and  $S_f$  display the  $t$ -test results, where “+” or “-” means the result is worse or better than the proposed method and “=” means they are similar in the  $t$ -test. In other words, the more “+”, the better the proposed methods.

From the variation curves of the fitness value, it is shown that APSOLL has achieved better fitness values on all datasets, which means the optimization ability of APSOLL is better than other metaheuristic algorithms by adopting the adaptive updating and leadership learning strategy. From



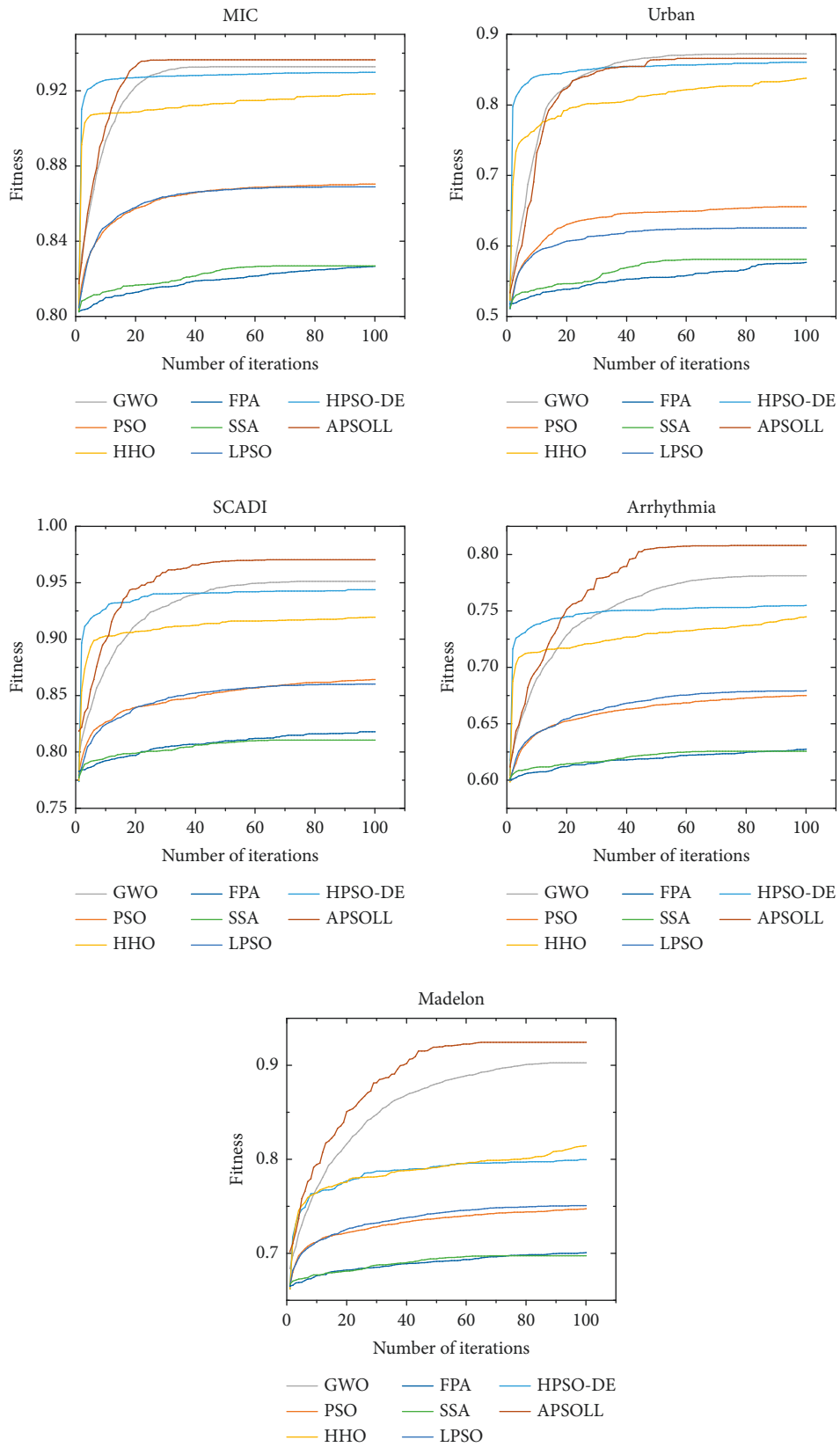


FIGURE 3: The average convergence curves of different metaheuristic algorithms for datasets below 500 dimensions.

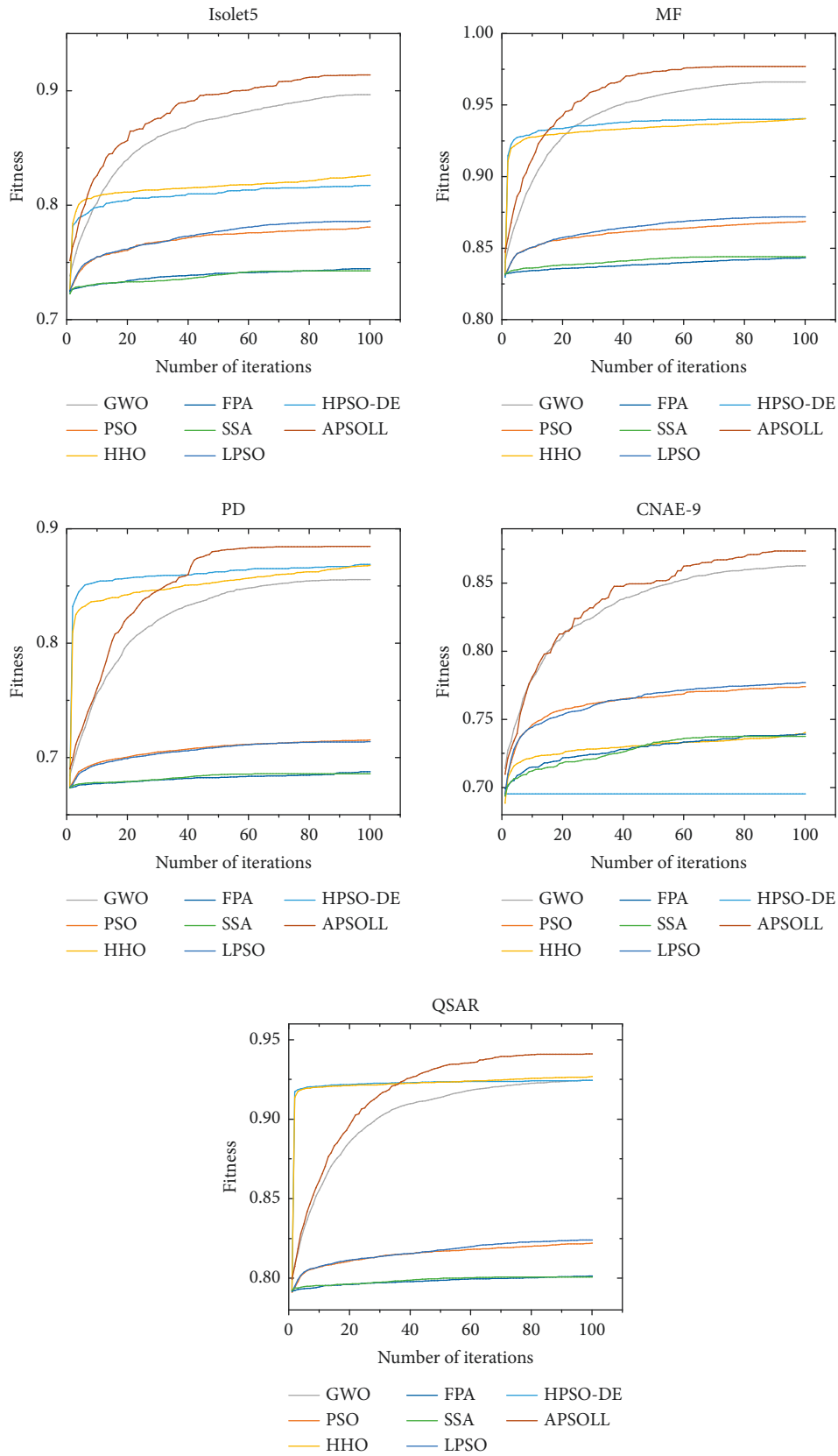


FIGURE 4: The average convergence curves of different metaheuristic algorithms for datasets above 500 dimensions.

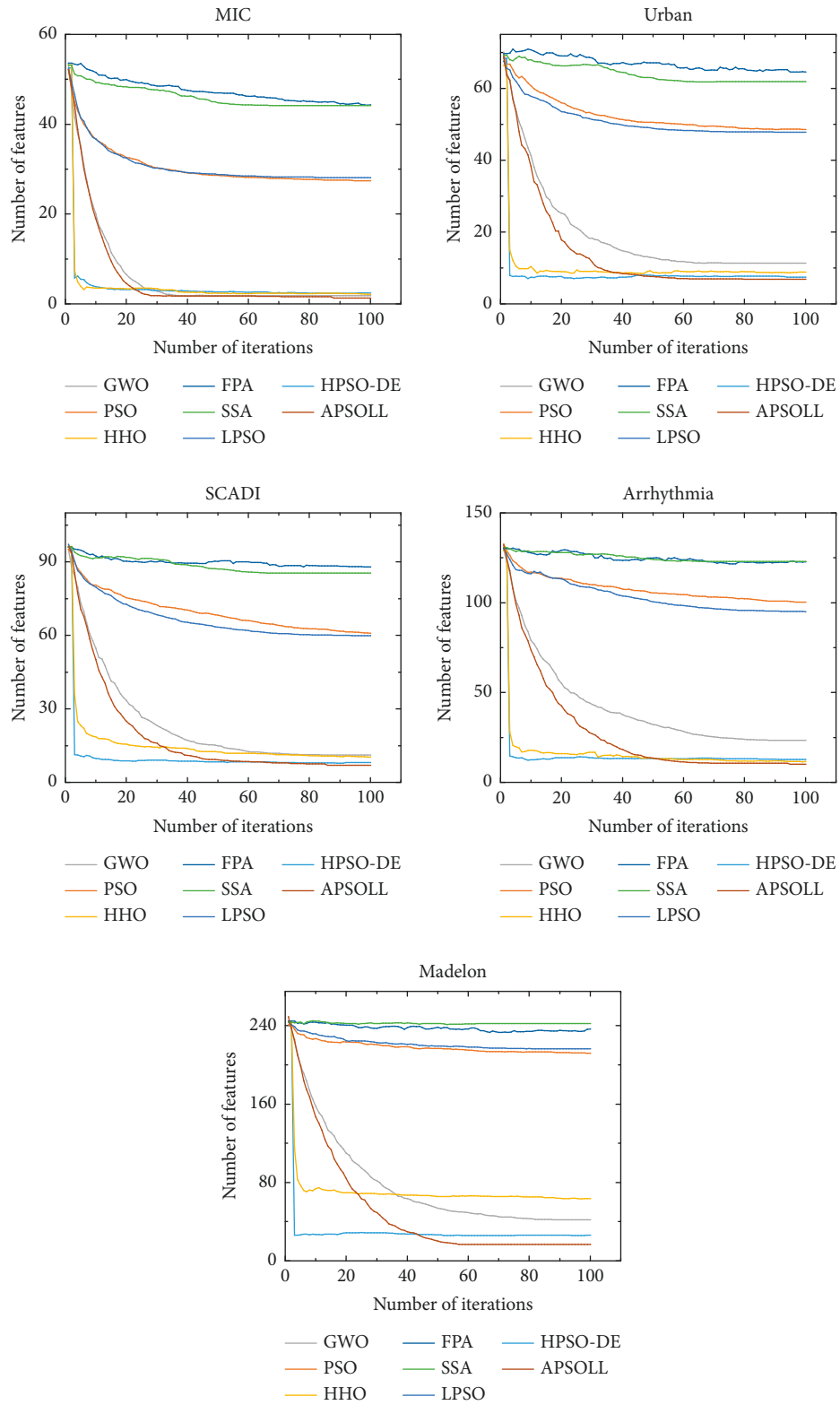


FIGURE 5: The average number of selected features for datasets below 500 dimensions by different FS methods based on metaheuristic algorithms.

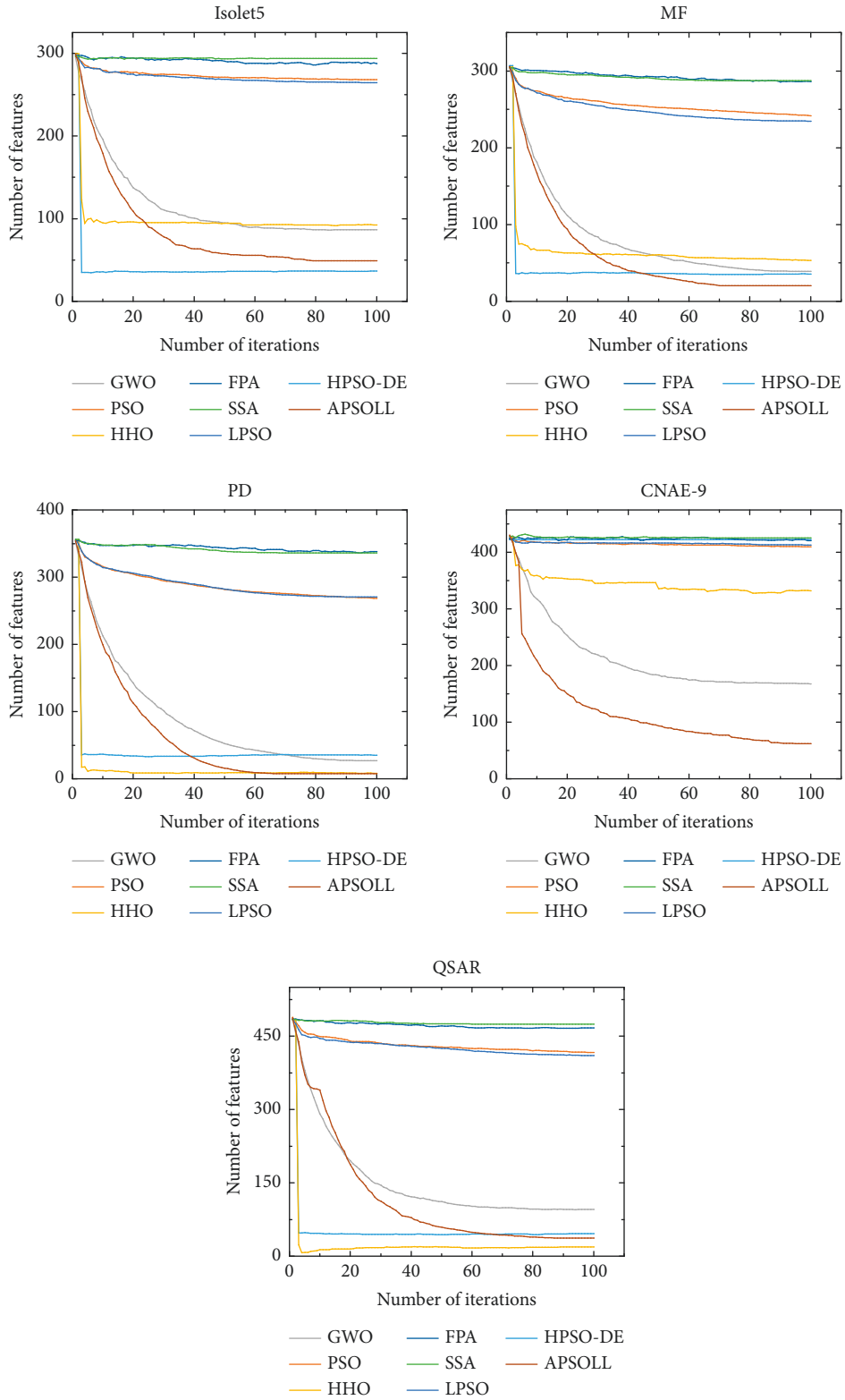


FIGURE 6: The average number of selected features for datasets above 500 dimensions by different FS methods based on metaheuristic algorithms.

TABLE 3: Comparisons between APSOLL and other metaheuristic algorithms for datasets below 500 dimensions.

Datasets	Method	Fit (std.)	$S_{fit}$	Acc (std.)	$S_{acc}$	#F (std.)	$S_f$	Time
MIC	GWO	93.28 (0.27)	+	91.03 (0.40)	+	1.80 (0.65)	=	125.37
	PSO	87.04 (0.96)	+	91.03 (0.48)	+	27.40 (3.49)	+	220.42
	HHO	91.83 (1.76)	+	89.08 (2.60)	+	2.13 (2.42)	=	162.52
	FPA	82.66 (0.53)	+	90.63 (0.75)	+	44.2 (2.50)	+	217.76
	SSA	82.68 (0.87)	+	90.65 (0.78)	+	44.2 (3.23)	+	122.10
	LPSO	86.89 (0.89)	+	91.08 (0.54)	+	28.13 (3.48)	+	220.95
	HPSO-DE	92.99 (0.38)	+	90.83 (0.45)	+	2.43 (1.09)	+	133.59
	APSOLL	93.65 (0.35)	*	91.40 (0.56)	*	1.33 (0.47)	*	122.10
Urban	GWO	87.21 (4.12)	=	85.03 (16.03)	=	11.33 (2.70)	+	96.00
	PSO	65.57 (5.57)	+	64.97 (13.90)	+	48.53 (5.89)	+	160.65
	HHO	83.78 (3.18)	+	79.43 (14.26)	+	8.93 (5.41)	=	94.06
	FPA	57.64 (2.87)	+	58.26 (10.28)	+	64.40 (6.52)	+	163.10
	SSA	58.10 (3.92)	+	58.17 (10.39)	+	61.83 (5.88)	+	162.84
	LPSO	62.53 (4.90)	+	60.41 (11.97)	+	47.80 (5.17)	+	163.89
	HPSO-DE	86.06 (1.20)	=	82.24 (14.30)	=	7.40 (2.11)	=	47.19
	APSOLL	86.60 (2.18)	*	82.84 (20.46)	*	6.83 (1.91)	*	75.32
SCADI	GWO	95.13 (2.04)	=	95.40 (2.88)	=	11.23 (7.49)	=	29.19
	PSO	86.43 (3.22)	+	93.33 (4.19)	+	60.87 (8.10)	+	124.32
	HHO	91.95 (3.61)	+	90.63 (4.51)	+	10.23 (7.14)	=	24.98
	FPA	81.80 (3.48)	+	92.38 (4.19)	+	87.90 (7.17)	+	147.73
	SSA	81.05 (3.59)	+	90.79 (4.59)	+	85.47 (8.11)	+	152.42
	LPSO	86.01 (3.12)	+	92.54 (4.20)	+	59.90 (6.65)	+	100.95
	HPSO-DE	94.38 (2.38)	+	93.65 (3.33)	+	8.07 (2.89)	=	23.31
	APSOLL	97.04 (1.63)	*	97.22 (2.35)	*	6.92 (2.75)	*	33.66
Arrhythmia	GWO	78.11 (1.31)	+	72.33 (1.70)	+	23.48 (5.65)	+	161.93
	PSO	67.50 (1.28)	+	68.97 (1.98)	+	100.23 (8.12)	+	164.50
	HHO	74.48 (1.94)	+	65.29 (3.20)	+	11.40 (10.72)	=	127.69
	FPA	62.73 (1.07)	+	65.59 (1.97)	+	122.57 (7.68)	+	160.16
	SSA	62.56 (1.30)	+	65.39 (1.77)	+	122.93 (6.44)	+	159.65
	LPSO	67.92 (1.39)	+	68.77 (1.80)	+	95.03 (6.31)	+	167.47
	HPSO-DE	75.49 (0.86)	+	66.96 (1.39)	+	12.87 (2.50)	=	80.80
	APSOLL	80.82 (1.45)	*	74.14 (1.75)	*	10.08 (3.95)	*	113.36
Madelon	GWO	90.28 (1.00)	+	89.71 (1.17)	+	42.00 (7.01)	+	310.38
	PSO	74.72 (1.12)	+	82.04 (1.18)	+	211.73 (12.36)	+	327.80
	HHO	81.44 (3.82)	+	78.95 (3.48)	+	216.47 (10.49)	+	399.71
	FPA	75.08 (0.86)	+	77.52 (1.16)	+	236.67 (9.62)	+	320.63
	SSA	70.06 (1.21)	+	77.53 (1.57)	+	242.17 (8.97)	+	322.84
	LPSO	75.07 (1.18)	+	82.94 (1.58)	+	63.70 (37.22)	+	325.15
	HPSO-DE	79.98 (1.72)	+	73.64 (2.56)	+	26.13 (5.06)	+	301.25
	APSOLL	92.44 (0.44)	*	90.65 (0.62)	*	16.92 (4.75)	*	259.51

Figures 3–4, it can be observed that HHO and HPSO-DE converge prematurely on most datasets, and PSO, SSA, FPA, and LPSO converge slower and have poor overall performance. In contrast, APSOLL achieves a balance in convergence speed and performance. In terms of classification accuracy, APSOLL-based FS method exceeds 80% on average in 9 of the 10 datasets, especially on MF, which has reached 98.07%. As it can be seen in Figures 5–6, PSO, SSA, FPA, and LPSO have limited performance in reducing the size of feature subsets, while APSO performs better than other methods on most datasets during the iterative process. In Tables 3–4, the number of selected features by APSOLL is less than those of other metaheuristic algorithms in most cases. A total of 30%–50% of features in the original datasets are selected by FPA and SSA, while less than 8% of features are selected by APSOLL. In particular, only 7.58 features are selected on average from the original 754 features on PD. As for CPU time, APSOLL consumes less time on MIC and

madelon compared to other metaheuristic algorithms. Moreover, although it consumes slightly more time on other datasets, it performs better in the two main aims of the classification accuracy and the number of selected features.

In summary, the optimization ability of APSOLL is better than other metaheuristic algorithms, and the suitable feature subsets are selected with higher classification accuracy and fewer features at an acceptable time.

*5.2. Experimental Results of Traditional Methods.* To demonstrate the effectiveness of APSOLL-based FS method, the performance is compared with that of 6 traditional methods. Figures 7–8 show the classification accuracy of 6 traditional FS methods for different numbers of selected features, and the optimal solutions of the proposed and traditional methods are presented in Table 5.

TABLE 4: Comparisons between APSOLL and other metaheuristic algorithms for datasets above 500 dimensions.

Datasets	Method	Fit (std.)	$S_{\text{fit}}$	Acc (std.)	$S_{\text{acc}}$	#F (Std.)	$S_f$	Time
Isolet5	GWO	89.66 (1.01)	+	91.23 (1.38)	=	86.53 (9.14)	+	212.36
	PSO	78.10 (1.04)	+	87.31 (1.44)	+	268.07 (10.71)	+	219.31
	HHO	82.61 (1.71)	+	81.60 (1.83)	+	92.57 (26.83)	+	283.60
	FPA	74.45 (0.89)	+	83.50 (1.34)	+	287.73 (10.48)	+	211.13
	SSA	74.25 (1.02)	+	83.60 (1.45)	+	293.53 (8.69)	+	207.08
	LPSO	78.61 (0.98)	+	87.79 (1.40)	+	264.42 (10.19)	+	215.30
	HPSO-DE	81.72 (1.08)	+	76.42 (1.68)	+	36.53 (5.12)	-	215.85
	APSOLL	91.37 (0.49)	*	91.08 (0.55)	*	48.92 (2.36)	*	219.14
MF	GWO	96.63 (0.54)	+	97.77 (0.54)	=	39.27 (6.89)	+	225.82
	PSO	86.86 (0.72)	+	97.19 (0.54)	+	241.73 (13.51)	+	274.84
	HHO	94.04 (0.95)	+	94.98 (0.98)	+	52.93 (13.66)	+	303.36
	FPA	84.31 (0.53)	+	96.47 (0.60)	+	286.13 (6.77)	+	281.36
	SSA	84.39 (0.53)	+	96.67 (0.71)	+	287.33 (9.16)	+	275.93
	LPSO	87.18 (0.63)	+	97.19 (0.61)	+	234.7 (8.29)	+	267.64
	HPSO-DE	94.05 (0.53)	+	93.84 (0.77)	+	35.57 (4.65)	+	224.65
	APSOLL	97.71 (0.33)	*	98.07 (0.53)	*	20.25 (1.23)	*	228.91
PD	GWO	85.54 (2.25)	+	80.88 (3.54)	=	27.00 (9.84)	+	187.67
	PSO	71.54 (1.62)	+	74.60 (2.11)	+	268.4 (11.07)	+	185.17
	HHO	86.78 (1.28)	+	81.60 (1.90)	+	8.43 (6.53)	=	127.75
	FPA	68.77 (1.31)	+	74.60 (2.24)	+	338.03 (12.81)	+	174.27
	SSA	68.59 (1.44)	+	74.23 (1.99)	+	336 (13.24)	+	173.38
	LPSO	71.38 (1.96)	+	74.48 (2.60)	+	270.43 (14.95)	+	185.49
	HPSO-DE	86.88 (0.87)	+	83.26 (1.39)	=	35.20 (4.53)	+	197.66
	APSOLL	88.44 (0.85)	*	83.92 (1.24)	*	7.58 (2.22)	*	152.82
CNAE-9	GWO	86.28 (1.22)	+	88.80 (1.53)	-	167.83 (20.93)	+	203.26
	PSO	77.41 (1.75)	+	88.25 (2.47)	-	409.80 (14.03)	+	197.09
	HHO	74.04 (1.83)	+	79.55 (4.12)	+	332.23 (80.68)	+	269.84
	FPA	73.91 (1.36)	+	83.79 (1.99)	+	420.70 (15.22)	+	185.77
	SSA	73.74 (1.85)	+	83.80 (2.51)	+	425.57 (12.44)	+	183.18
	LPSO	77.69 (1.27)	+	88.79 (1.92)	-	412.63 (14.06)	+	195.26
	HPSO-DE	69.52 (1.87)	+	77.60 (2.43)	+	422.40 (13.83)	+	200.46
	APSOLL	87.35 (0.55)	*	85.03 (0.93)	*	61.83 (5.38)	*	210.71
QSAR	GWO	92.45 (0.54)	+	93.21 (0.66)	=	95.57 (9.18)	+	236.35
	PSO	82.20 (0.62)	+	92.01 (0.68)	+	416.70 (17.62)	+	327.26
	HHO	92.68 (0.45)	+	90.35 (0.83)	+	19.10 (12.23)	-	227.42
	FPA	80.13 (0.49)	+	91.16 (0.80)	+	466.93 (10.49)	+	323.93
	SSA	80.06 (0.48)	+	91.37 (0.71)	+	474.40 (11.67)	+	320.06
	LPSO	82.40 (0.55)	+	92.02 (0.59)	+	410.13 (14.90)	+	323.23
	HPSO-DE	92.44 (0.27)	+	91.14 (0.41)	+	46.30 (6.15)	+	207.01
	APSOLL	94.10 (0.55)	*	93.11 (0.74)	*	36.83 (7.84)	*	231.28

It is observed from Figures 7–8 that the traditional methods are difficult to improve the classification accuracy by sequentially increasing the number of features when a certain level is reached. In comparison, more suitable feature subsets are obtained by the metaheuristic algorithm-based FS method, among these, APSOLL has better performance. In addition, it is not the case that the more features selected, the higher the classification accuracy is, which indicates that the redundancy among features affects the classification performance on most datasets.

As can be seen from Table 5, it is clear that the classification accuracy is improved by at least 1.28% on average via the proposed method on 5 datasets, especially on arrhythmia and isolet5, with 11.77% and 4.26%, respectively. Although the classification accuracy of the proposed method is about 2% on average lower than traditional methods on myocardial, MF, PD, and CNAE-9, the number of selected

features is lower than that of these methods, only 2, 21, 9, and 64 features are selected, respectively. To further analyze the number of selected features, fewer features are selected by the proposed method on 6 datasets. Among them, it is noticed that more than 30% of the features are selected by traditional methods on Isolet5 and MF, while only 7.46% and 3.24% of the features are selected by the proposed method, respectively. In terms of time consumption, traditional methods are affected by the number of features due to the sequential addition of features to the feature subsets, and its time consumption increase dramatically as the number of features increases, while APSOLL performs more stability on most datasets because its dynamic exploration and exploitation capabilities, and the CPU time is still acceptable. In brief, the proposed method is dependable and effective for solving FS problems compared with traditional methods.

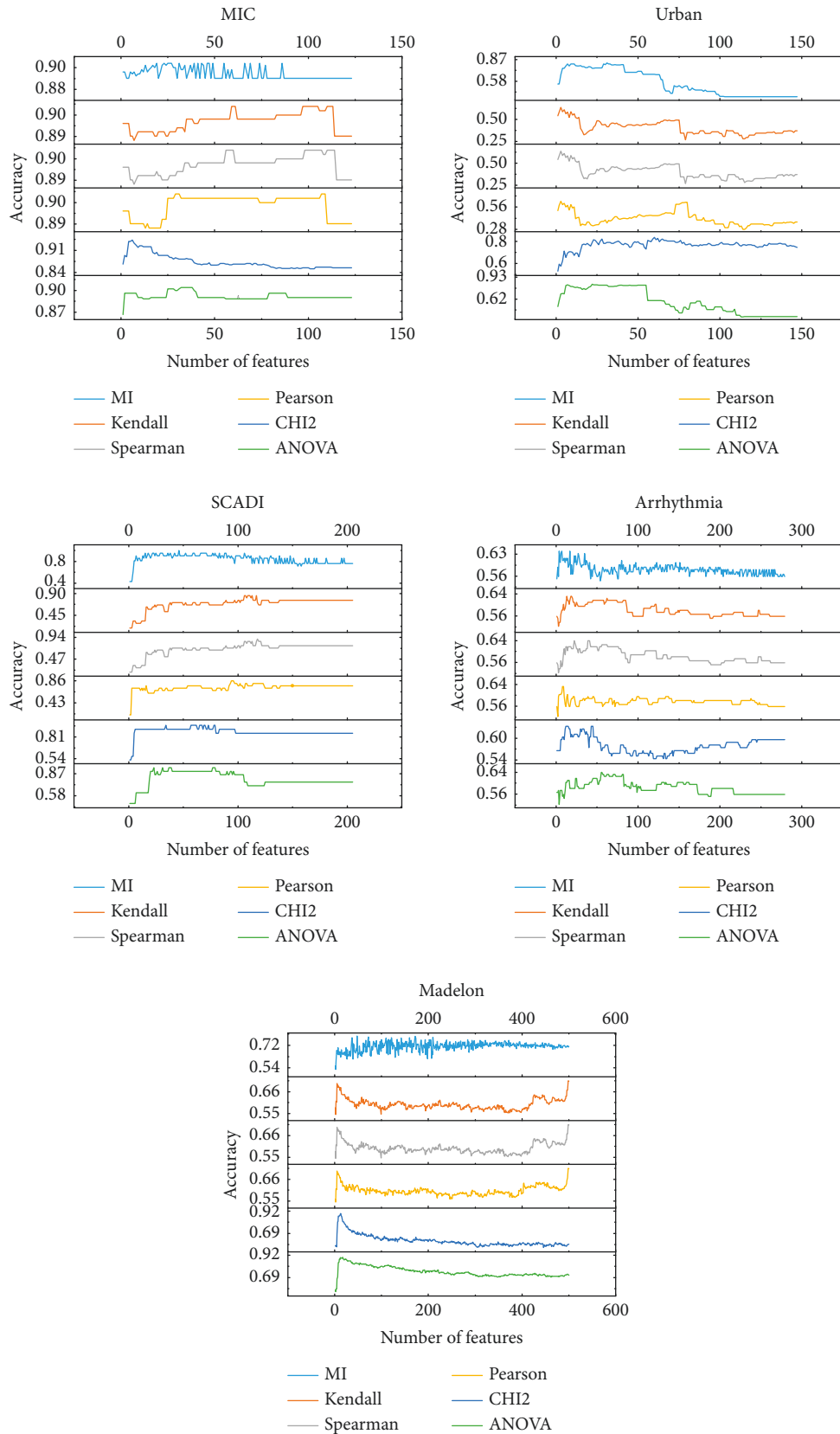


FIGURE 7: The classification accuracy of 6 traditional FS methods in selecting different numbers of features for datasets below 500 dimensions.

TABLE 5: The optimal classification accuracy, number of selected features, and CPU time in comparison to traditional methods.

Datasets	ANOVA	CH12	Pearson	Spearman	Kendall	MI	APSOLL
MIC	Acc (%)	90.39	90.39	90.39	90.39	90.39	92.55
	#F	31	6	29	56	19	2
	Time	3.38	3.24	4.03	13.04	8.15	164.07
Urban	Acc (%)	83.01	83.66	63.40	63.40	82.35	85.62
	#F	22	62	3	3	31	4
	Time	2.12	3.36	2.96	15.32	8.61	161.59
SCADI	Acc (%)	95.24	95.24	85.71	90.47	100	100
	#F	23	34	94	118	107	7
	Time	0.79	2.19	2.02	19.56	12.49	143.46
Arrhythmia	Acc (%)	63.97	63.24	63.24	63.97	63.24	75.74
	#F	55	12	8	17	13	6
	Time	2.94	2.34	8.09	56.24	40.11	675.94
Madelon	Acc (%)	89.87	89.36	71.41	71.41	79.36	91.15
	#F	17	13	499	499	48	17
	Time	38.03	40.01	49.72	223.64	137.44	2833.54
Isolet5	Acc (%)	86.97	85.47	84.83	85.26	87.82	92.08
	#F	245	289	378	351	223	46
	Time	62.04	61.61	57.49	561.71	197.98	8484.77
MF	Acc (%)	98.33	98.83	94.83	94.00	93.83	98.50
	#F	622	402	482	386	411	21
	Time	43.79	44.31	50.45	312.71	188.62	5533.33
PD	Acc (%)	79.30	87.67	80.18	81.06	78.41	85.90
	#F	4	140	16	24	25	9
	Time	61.75	62.94	62.16	395.49	234.74	2488.38
CNAE-9	Acc (%)	89.81	88.27	85.49	85.49	85.49	85.80
	#F	213	142	855	855	855	64
	Time	66.91	84.32	63.66	462.64	237.18	7087.23
QSAR	Acc (%)	91.52	91.72	91.72	91.72	91.72	93.88
	#F	110	105	984	984	984	33
	Time	126.37	135.70	125.28	1058.36	350.85	8411.24



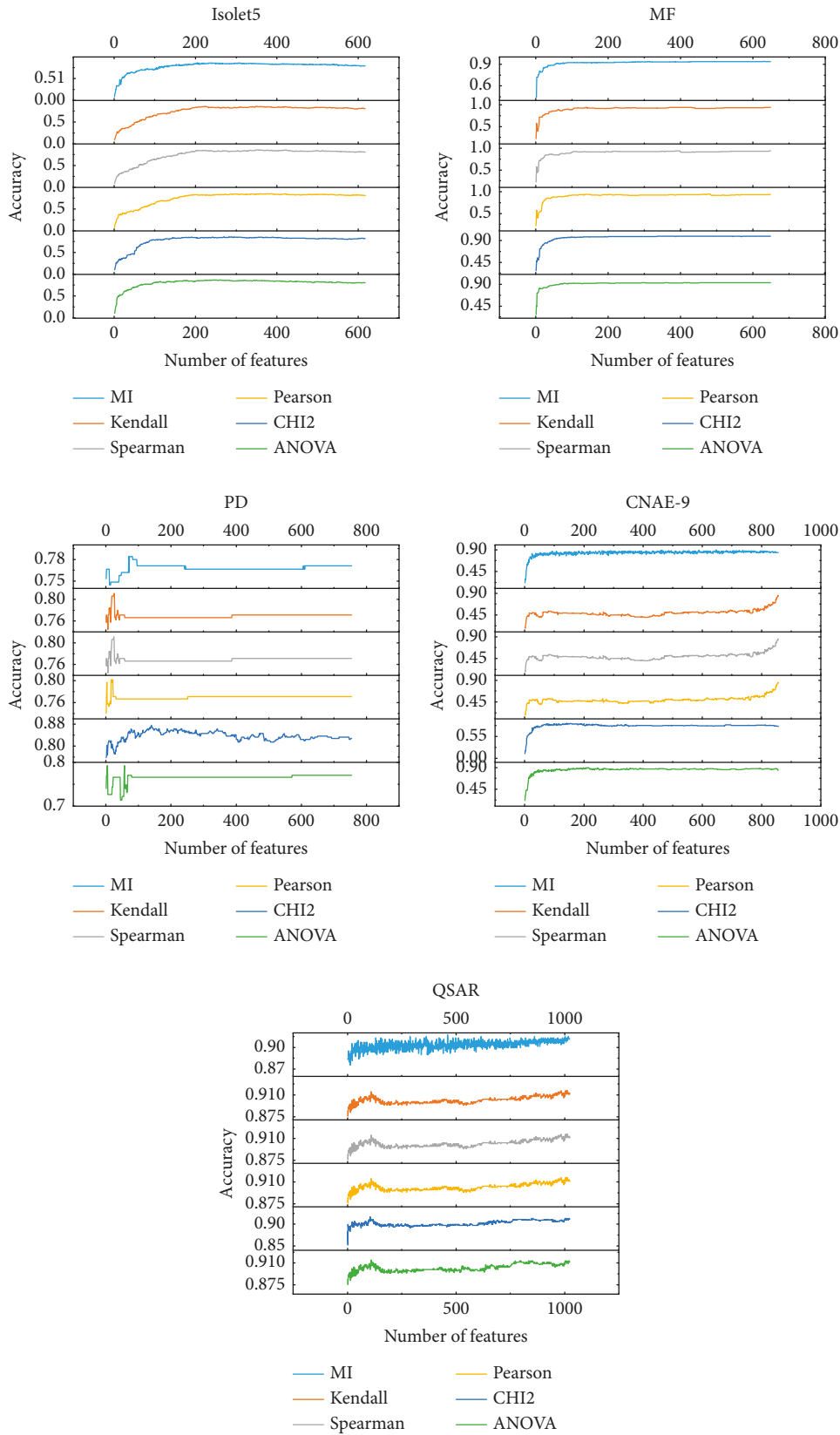


FIGURE 8: The classification accuracy of 6 traditional FS methods in selecting different numbers of features for datasets above 500 dimensions.

## 6. Conclusions and Future Work

In the paper, APSOLL is proposed for FS, which enhances exploration and exploitation capabilities by utilizing an adaptive updating strategy to guide the population search in a more reasonable scope and the leadership learning strategy to increase population diversity. Experimental results in comparison with other FS methods based on metaheuristic algorithms reveal that APSOLL offers better optimization ability and selects the suitable feature subsets within an acceptable time. Moreover, APSOLL-based FS method achieves better or approximate classification accuracy by selecting less than 8% of features from the original datasets compared to other traditional methods. In conclusion, the suitable feature subsets are selected by the proposed method while ensuring a proper balance between the classification accuracy and the number of selected features. In the future, it is interesting to decrease the CPU time of APSOLL by combining the feature ranking and applying it to process ultrahigh dimensional datasets.

### Data Availability

The data used to support the findings of this study are openly available in the UCI archive.

### Conflicts of Interest

The authors declare that they have no conflicts of interest.

### Acknowledgments

This work has been supported by the Wuhan Science and Technology Bureau 2022 Knowledge Innovation Dawning Plan Project: Detection and Optimization Method of GNSS Hybrid Attacks for Connected and Autonomous Vehicles (no. 2022010801020270).

### References

- [1] L. Zhang, "A feature selection algorithm integrating maximum classification information and minimum interaction feature dependency information," *Computational Intelligence and Neuroscience*, vol. 2021, Article ID 3569632, 10 pages, 2021.
- [2] P. P. Kundu and S. Mitra, "Feature selection through message passing," *IEEE Transactions on Cybernetics*, vol. 47, no. 12, pp. 4356–4366, 2017.
- [3] A. D. Li, Z. He, Q. Wang, and Y. Zhang, "Key quality characteristics selection for imbalanced production data using a two-phase bi-objective feature selection method," *European Journal of Operational Research*, vol. 274, no. 3, pp. 978–989, 2019.
- [4] I. Aljarah, M. Habib, H. Faris et al., "A dynamic locality multi-objective salp swarm algorithm for feature selection," *Computers & Industrial Engineering*, vol. 147, Article ID 106628, 2020.
- [5] W. Guendouzi and A. Boukra, "GAB-BBO: adaptive biogeography based feature selection approach for intrusion detection," *International Journal of Computational Intelligence Systems*, vol. 10, no. 1, pp. 914–935, 2017.
- [6] R. Diao and Q. Shen, "Nature inspired feature selection metaheuristics," *Artificial Intelligence Review*, vol. 44, no. 3, pp. 311–340, 2015.
- [7] Y. Wan, M. Wang, Z. Ye, and X. Lai, "A feature selection method based on modified binary coded ant colony optimization algorithm," *Applied Soft Computing*, vol. 49, pp. 248–258, 2016.
- [8] H. Zhuang, X. Liu, H. Wang et al., "Diagnosis of early stage Parkinson's disease on quantitative susceptibility mapping using complex network with one-way anova f-test feature selection," *Journal of Mechanics in Medicine and Biology*, vol. 21, Article ID 2140026, 2021.
- [9] U. Moorthy and U. D. Gandhi, "A novel optimal feature selection technique for medical data classification using ANOVA based whale optimization," *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, no. 3, pp. 3527–3538, 2021.
- [10] H. N. Alshaer, M. A. Otair, L. Abualigah, M. Alshinwan, and A. M. Khasawneh, "Feature selection method using improved CHI Square on Arabic text classifiers: analysis and application," *Multimedia Tools and Applications*, vol. 80, no. 7, pp. 10373–10390, 2021.
- [11] S. Solorio-Fernández, J. F. Martínez-Trinidad, and J. A. Carrasco-Ochoa, "A supervised filter feature selection method for mixed data based on spectral feature selection and information-theory redundancy analysis," *Pattern Recognition Letters*, vol. 138, pp. 321–328, 2020.
- [12] G. Fang, W. Liu, and L. Wang, "A machine learning approach to select features important to stroke prognosis," *Computational Biology and Chemistry*, vol. 88, Article ID 107316, 2020.
- [13] R. Alaiz-Rodríguez and A. C. Parnell, "An information theoretic approach to quantify the stability of feature selection and ranking algorithms," *Knowledge-Based Systems*, vol. 195, Article ID 105745, 2020.
- [14] J. Jiarpakdee, C. Tantithamthavorn, and C. Treude, "The impact of automated feature selection techniques on the interpretation of defect models," *Empirical Software Engineering*, vol. 25, no. 5, pp. 3590–3638, 2020.
- [15] U. Kaya and M. Fidan, "Parametric and nonparametric correlation ranking based supervised feature selection methods for skin segmentation," *Journal of Ambient Intelligence and Humanized Computing*, vol. 13, no. 2, pp. 821–833, 2021.
- [16] X. Zhang, J. Wang, and Y. Gao, "A hybrid short-term electricity price forecasting framework: cuckoo search-based feature selection with singular spectrum analysis and SVM," *Energy Economics*, vol. 81, pp. 899–913, 2019.
- [17] G. Karakaya, S. Galelli, S. D. Ahpaşaoğlu, and R. Taormina, "Identifying (quasi) equally informative subsets in feature selection problems for classification: a max-relevance min-redundancy approach," *IEEE Transactions on Cybernetics*, vol. 46, no. 6, pp. 1424–1437, 2016.
- [18] B. A. S. Al-Rimy, M. A. Maarof, M. Alazab et al., "Redundancy coefficient gradual up-weighting-based mutual information feature selection technique for crypto-ransomware early detection," *Future Generation Computer Systems*, vol. 115, pp. 641–658, 2021.
- [19] S. Liu, L. Liu, Y. Fan et al., "An integrated scheme for online dynamic security assessment based on partial mutual information and iterated random forest," *IEEE Transactions on Smart Grid*, vol. 11, no. 4, pp. 3606–3619, 2020.
- [20] X. Yan and M. Jia, "Intelligent fault diagnosis of rotating machinery using improved multiscale dispersion entropy and

- mRMR feature selection,” *Knowledge-Based Systems*, vol. 163, pp. 450–471, 2019.
- [21] L. Wang, S. Jiang, and S. Jiang, “A feature selection method via analysis of relevance, redundancy, and interaction,” *Expert Systems with Applications*, vol. 183, Article ID 115365, 2021.
- [22] M. Alweshah, S. A. Khalailah, B. B. Gupta, A. Almomani, A. I. Hammouri, and M. A. Al-Betar, “The monarch butterfly optimization algorithm for solving feature selection problems,” *Neural Computing & Applications*, vol. 34, no. 14, pp. 11267–11281, 2020.
- [23] M. Abdel-Basset, R. Mohamed, R. K. Chakraborty, M. J. Ryan, and S. Mirjalili, “An efficient binary slime mould algorithm integrated with a novel attacking-feeding strategy for feature selection,” *Computers & Industrial Engineering*, vol. 153, Article ID 107078, 2021.
- [24] M. Yarlagadda, K. Gangadhara Rao, and A. Srikrishna, “Frequent itemset-based feature selection and Rider Moth Search Algorithm for document clustering,” *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 4, pp. 1098–1109, 2022.
- [25] B. J. Ma, S. Liu, and A. A. Heidari, “Multi-strategy ensemble binary hunger games search for feature selection,” *Knowledge-Based Systems*, vol. 248, Article ID 108787, 2022.
- [26] Z. Ye, W. Cai, S. Liu, K. Liu, M. Wang, and W. Zhou, “A band selection approach for hyperspectral image based on a modified hybrid rice optimization algorithm,” *Symmetry*, vol. 14, no. 7, p. 1293, 2022.
- [27] B. Shi, H. Ye, L. Zheng et al., “Evolutionary warning system for COVID-19 severity: colony predation algorithm enhanced extreme learning machine,” *Computers in Biology and Medicine*, vol. 136, Article ID 104698, 2021.
- [28] A. Y. Hassan, A. A. K. Ismael, M. Said, R. M. Ghoniem, S. Deb, and A. G. Elsayed, “Evaluation of weighted mean of vectors algorithm for identification of solar cell parameters,” *Processes*, vol. 10, no. 6, p. 1072, 2022.
- [29] C. Shen and K. Zhang, “Two-stage Improved Gray Wolf Optimization Algorithm for Feature Selection on High-Dimensional Classification,” *Complex and Intelligent Systems*, vol. 8, 2022.
- [30] C. Yan, J. Ma, H. Luo, G. Zhang, and J. Luo, “A novel feature selection method for high-dimensional biomedical data based on an improved binary clonal flower pollination algorithm,” *Human Heredity*, vol. 84, no. 1, pp. 34–46, 2019.
- [31] K. Balakrishnan, R. Dhanalakshmi, and U. M. Khaire, “Improved salp swarm algorithm based on the levy flight for feature selection,” *The Journal of Supercomputing*, vol. 77, no. 11, pp. 12399–12419, 2021.
- [32] K. Hussain, N. Neggaz, W. Zhu, and E. H. Houssein, “An efficient hybrid sine-cosine Harris hawks optimization for low and high-dimensional feature selection,” *Expert Systems with Applications*, vol. 176, Article ID 114778, 2021.
- [33] O. Tarkhaneh, T. T. Nguyen, and S. Mazaheri, “A novel wrapper-based feature subset selection method using modified binary differential evolution algorithm,” *Information Sciences*, vol. 565, pp. 278–305, 2021.
- [34] M. M. Mafarja and S. Mirjalili, “Hybrid whale optimization algorithm with simulated annealing for feature selection,” *Neurocomputing*, vol. 260, pp. 302–312, 2017.
- [35] X. F. Song, Y. Zhang, D. W. Gong, and X. Z. Gao, “A fast hybrid feature selection based on correlation-guided clustering and particle swarm optimization for high-dimensional data,” *IEEE Transactions on Cybernetics*, vol. 99, pp. 1–14, 2021.
- [36] B. Tran, B. Xue, and M. Zhang, “A new representation in PSO for discretization-based feature selection,” *IEEE Transactions on Cybernetics*, vol. 48, no. 6, pp. 1733–1746, 2018.
- [37] X.-F. Song, Y. Zhang, Y.-N. Guo, X.-Y. Sun, and Y.-L. Wang, “Variable-size cooperative coevolutionary particle swarm optimization for feature selection on high-dimensional data,” *IEEE Transactions on Evolutionary Computation*, vol. 24, no. 5, pp. 882–895, 2020.
- [38] Y. Hu, Y. Zhang, and D. Gong, “Multiobjective particle swarm optimization for feature selection with fuzzy cost,” *IEEE Transactions on Cybernetics*, vol. 51, no. 2, pp. 874–888, 2021.
- [39] S. M. Hosseini Bamakan, H. Wang, T. Yingjie, and Y. Shi, “An effective intrusion detection framework based on MCLP/SVM optimized by time-varying chaos particle swarm optimization,” *Neurocomputing*, vol. 199, pp. 90–102, 2016.
- [40] M. Mafarja, R. Jarrar, S. Ahmad, and A. A. Abusnaina, “Feature selection using binary particle swarm optimization with time varying inertia weight strategies,” in *Proceedings of the International Conference on Future Networks & Distributed Systems*, pp. 1–9, New York, NY, USA, June 2018.
- [41] X. Huang, Y. Chi, and Y. Zhou, “Feature Selection of High Dimensional Data by Adaptive Potential Particle Swarm Optimization,” in *Proceedings of the IEEE Congress on Evolutionary Computation (CEC)*, pp. 1052–1059, Wellington, New Zealand, June 2019.
- [42] Y. Xue, T. Tang, W. Pang, and A. X. Liu, “Self-adaptive parameter and strategy based particle swarm optimization for large-scale feature selection problems with multiple classifiers,” *Applied Soft Computing*, vol. 88, Article ID 106031, 2020.
- [43] P. Moradi and M. Gholampour, “A hybrid particle swarm optimization for feature subset selection by integrating a novel local search strategy,” *Applied Soft Computing*, vol. 43, pp. 117–130, 2016.
- [44] K. Chen, B. Xue, M. Zhang, and F. Zhou, “Hybridising particle swarm optimisation with differential evolution for feature selection in classification,” in *Proceedings of the IEEE Congress on Evolutionary Computation*, pp. 1–8, Glasgow, UK, July 2020.
- [45] K. Zou, Y. Liu, S. Wang, N. Li, and Y. Wu, “A multiobjective particle swarm optimization algorithm based on grid technique and multistrategy,” *Journal of Mathematics*, vol. 2021, Article ID 1626457, 17 pages, 2021.
- [46] E. Cuevas and M. González, “An optimization algorithm for multimodal functions inspired by collective animal behavior,” *Soft Computing*, vol. 17, no. 3, pp. 489–502, 2013.
- [47] G. Hirst, L. Mann, P. Bain, A. Pirola-Merlo, and A. Richver, “Learning to lead: the development and testing of a model of leadership learning,” *The Leadership Quarterly*, vol. 15, no. 3, pp. 311–327, 2004.
- [48] E. Emary, W. Yamany, A. E. Hassanien, and V. Snasel, “Multi-objective gray-wolf optimization for attribute reduction,” *Procedia Computer Science*, vol. 65, pp. 623–632, 2015.
- [49] E. Emary, H. M. Zawbaa, and A. E. Hassanien, “Binary gray wolf optimization approaches for feature selection,” *Neurocomputing*, vol. 172, pp. 371–381, 2016.
- [50] P. Hu, J. S. Pan, and S. C. Chu, “Improved binary grey wolf optimizer and its application for feature selection,” *Knowledge-Based Systems*, vol. 195, Article ID 105746, 2020.
- [51] S. Khalilpourazari, H. Hashemi Doulabi, A. Özyüksel Çiftçioğlu, and G. W. Weber, “Gradient-based grey wolf optimizer with Gaussian walk: application in modelling and prediction of the COVID-19 pandemic,” *Expert Systems with Applications*, vol. 177, Article ID 114920, 2021.

- [52] Y. Zhang, D. Gong, Y. Hu, and W. Zhang, "Feature selection algorithm based on bare bones particle swarm optimization," *Neurocomputing*, vol. 148, pp. 150–157, 2015.
- [53] L. Y. Chuang, C. S. Yang, K. C. Wu, and C. H. Yang, "Gene selection and classification using Taguchi chaotic binary particle swarm optimization," *Expert Systems with Applications*, vol. 38, no. 10, pp. 13367–13377, 2011.
- [54] A. D. Li, B. Xue, and M. Zhang, "Improved binary particle swarm optimization for feature selection with new initialization and search space reduction strategies," *Applied Soft Computing*, vol. 106, Article ID 107302, 2021.
- [55] A. W. Smith, "Leadership is a living system: learning leaders and organizations," *Human Systems Management*, vol. 16, no. 4, pp. 277–284, 1997.
- [56] Z. Dawood and C. K. Babulal, "Power quality disturbance classification based on efficient adaptive Arrhenius artificial bee colony feature selection," *International Transactions on Electrical Energy Systems*, vol. 31, no. 5, Article ID e12868, 2021.
- [57] D. R. Munirathinam and M. Ranganadhan, "A new improved filter-based feature selection model for high-dimensional data," *The Journal of Supercomputing*, vol. 76, no. 8, pp. 5745–5762, 2020.
- [58] S. Zhang, F. Zhu, Q. Yu, and X. Zhu, "Identifying DNA-binding proteins based on multi-features and LASSO feature selection," *Biopolymers*, vol. 112, no. 2, Article ID e23419, 2021.

## Research Article

# A Variable Radius Side Window Direct SLAM Method Based on Semantic Information

Yan Chen <sup>1</sup>, Jianjun Ni <sup>1,2</sup>, Emmanuel Mutabazi<sup>1</sup>, Weidong Cao <sup>1,2</sup> and Simon X. Yang<sup>3</sup>

<sup>1</sup>College of Internet of Things Engineering, Hohai University, Changzhou 213022, China

<sup>2</sup>Jiangsu Key Laboratory of Power Transmission & Distribution Equipment Technology, Hohai University, Changzhou 213022, China

<sup>3</sup>Advanced Robotics and Intelligent Systems (ARIS) Laboratory, School of Engineering, University of Guelph, Guelph, ON, Canada

Correspondence should be addressed to Jianjun Ni; njhhuc@gmail.com

Received 5 May 2022; Accepted 28 June 2022; Published 22 August 2022

Academic Editor: Nian Zhang

Copyright © 2022 Yan Chen et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Simultaneous Localization and Mapping (SLAM) is a challenging and key issue in the mobile robotic fields. In terms of the visual SLAM problem, the direct methods are more suitable for more expansive scenes with many repetitive features or less texture in contrast with the feature-based methods. However, the robustness of the direct methods is weaker than that of the feature-based methods. To deal with this problem, an improved direct sparse odometry with loop closure (LDSO) is proposed, where the performance of the SLAM system under the influence of different imaging disturbances of the camera is focused on. In the proposed method, a method based on the side window strategy is proposed for preprocessing the input images with a multilayer stacked pixel blender. Then, a variable radius side window strategy based on semantic information is proposed to reduce the weight of selected points on semistatic objects, which can reduce the computation and improve the accuracy of the SLAM system based on the direct method. Various experiments are conducted on the KITTI dataset and TUM RGB-D dataset to test the performance of the proposed method under different camera imaging disturbances. The quantitative and qualitative evaluations show that the proposed method has better robustness than the state-of-the-art direct methods in the literature. Finally, a real-world experiment is conducted, and the results prove the effectiveness of the proposed method.

## 1. Introduction

Simultaneous Localization and Mapping (SLAM) plays essential roles in robotic and other related fields [1–3]. In the robotic field, SLAM systems are used to solve the problem of robots about where they are. Based on the acquisition of its pose and surrounding environment, a robot can further solve where to go or what to do [4].

Many kinds of sensors are used in SLAM systems, such as LiDAR, camera, and inertial measurement unit [5, 6]. Commonly, SLAM algorithms are divided into laser SLAM and visual SLAM according to the sensor used [7, 8]. Due to the low cost of the camera, the large amount of information it carries, and the ease of use, visual SLAM has become more popular among researchers in recent years. Visual SLAM

usually uses monocular cameras, binocular cameras, or RGB-D cameras to obtain environmental information. Compared with other types of cameras, the monocular camera is cheap and common. In addition, there are the most abundant data sources of the monocular camera. So the monocular SLAM plays an important role in the visual SLAM field and has been widely studied and applied [9, 10]. However, the monocular SLAM can obtain only image information without scale information, so it is more dependent on the quality of the image. Therefore, how to improve the robustness of monocular SLAM under different disturbances is a very challenging and important task in this field [11, 12].

There are three main implementation schemes in visual SLAM, namely feature-based method, direct method, and

semidirect method. The feature-based method finds feature points, matches them, calculates the pose, and constructs a map through geometric relations. The most commonly used methods for feature extraction are Scale Invariant Feature Transform (SIFT) [13], Speeded Up Robust Features (SURF) [14], and Oriented Fast and Rotated BRIEF (ORB) [15]. ORB is one of the best methods, which improves the speed and accuracy of FAST [16], and uses BRIEF [17] for the efficient computation of features. Accordingly, ORB-SLAM is currently the most popular visual SLAM solution [18, 19].

Unlike the feature-based method, the direct approach does not rely on the one-to-one matching of points. It optimizes the interframe pose by extracting pixels with apparent gradients and minimizing the photometric error function of the pixels, such as the large-scale direct monocular SLAM (LSD-SLAM) [20] and the direct sparse odometry (DSO) [21]. The semidirect method, such as the semidirect visual odometry (SVO) [22], uses a similar structure to the feature-based method and combines the tracking of the direct method and the motion optimization of the feature-based method. The feature-based method and the semidirect method both rely on low-level geometric feature extractors with high repeatability. They are not suitable for surfaces with many repetitive features or less texture. In contrast, the direct method can be used in a broader range of scenarios. In this paper, we focus on direct method solutions for the monocular SLAM. The main purpose of this paper is to improve the robustness of the direct methods under different disturbances.

The robustness of the direct method-based SLAM system is challenged by photometric calibration, dynamic objects, rolling shutter effect, camera imaging disturbances, and so on [23]. There have been many excellent works to improve the robustness of the direct method-based SLAM systems. For example, Zhu et al. [24] proposed a photometric transfer net (PTNet), which is trained to pixel-wisely remove brightness discrepancies between two frames without ruining the context information, to overcome the problem of brightness discrepancies. Liu et al. [25] proposed an enhanced visual SLAM algorithm based on the sparse direct method to solve the illumination sensitivity problem. Sheng et al. [26] filtered out the dynamic objects based on the semantic information to improve the positioning accuracy and robustness of DSO [21]. Zhou et al. [27] jointly optimized the 3D lines, points, and poses within a sliding window to consider the collinear constraint among the points to improve the robustness of the direct method.

The works introduced above can improve the robustness of the direct method to some extent. However, the research focusing on the influence of different camera imaging disturbances and semistatic objects is relatively lacking. During the long-term operation of the monocular SLAM system, the image quality of the camera will be affected by different disturbances from the external environment and internal sensors. In this paper, two main types of imaging disturbances are studied, namely, different noise on the camera and the brightness influence on the imaging process. The main noises on the camera include Gaussian noise and Salt-and-Pepper noise. Gaussian noise is often caused by the high

temperature of the camera sensor running for a long time and mutual interference of internal circuit components [28]. Salt-and-Pepper noise is often caused by the faulty of the camera sensor, the wear of the camera lens, and the adsorption of dust in the air [29, 30]. The brightness influence on the imaging is a very common problem of the vision-based SLAM. For example, the accumulated irradiance exceeding the camera's dynamic range can cause the camera overexposure interference when the ambient brightness is not uniform [31, 32]. Another important influence on the robustness of the direct methods in the vision-based SLAM is the semistatic objects, which refer to objects that are static most of the time but will change at a certain moment, such as the cars parked on the side of the road. Semistatic objects are not suitable for being directly filtered out like dynamic objects because most of them are rich in texture and are suitable for estimating pose when they are static [33]. Thus, the main motivation of this paper is to study how to improve the robustness of the direct method-based SLAM system in different camera imaging disturbances and reduce the specific gravity of semistatic objects.

The main contributions of this paper are as follows: (1) A regional pixel information fusion method based on multiple average calculations is proposed to improve the robustness of the direct sparse odometry with loop closure- (LDSO-) based SLAM. (2) A side window strategy is introduced into the framework of the LDSO-based SLAM to enhance the edge-preserving property. (3) A method based on semantic information is presented to reduce the effects of nonstatic objects on the LDSO-based SLAM. So there are three main improvements of the proposed method, namely, a regional pixel information fusion method for robustness, a side window strategy for edge preserving, and the semantic-based strategy for the nonstatic objects. Compared with the existing methods, the proposed method improves the robustness of the direct method-based SLAM against multiple camera imaging disturbances, including Gaussian noise, Salt-and-Pepper noise, and camera overexposure, rather than just against a single disturbance.

The rest of this paper is organized as follows. Section 2 gives out an overview of the background. The proposed algorithm is presented in Section 3. In Section 4, detailed quantitative and qualitative experimental results are provided. The discussions of the proposed algorithm are carried out in Section 5. Finally, Section 6 concludes this paper and gives out the future work.

## 2. Background

Direct method-based SLAM systems jointly estimate the position and posture changes of the camera by minimizing the photometric error in the image alignment. It makes direct methods more accurate and robust than feature-based methods in scenes that lack texture or are full of repetitive textures. However, the monocular direct methods suffer from the accumulated drift of global translation, rotation, and scale without closed-loop detection. This leads to inaccurate long-term trajectory estimation and mapping. In this paper, Direct Sparse Odometry with Loop closure

(LDSO) [34] is focused on, which adds closed-loop detection to DSO for global optimization. The main process of LDSO is reviewed in this section.

**2.1. Framework of LDSO.** The algorithm framework of LDSO is shown in Figure 1. When a new frame of image is acquired, all the active 3D points in the current sliding window of the local bundle adjustment module are projected into this frame. The initial pose of this frame is estimated by direct image alignment. This frame is added to the local windowed bundle adjustment if it is judged as a new keyframe. Old or redundant keyframes and points are marginalized. The active keyframes and the marginalized keyframes rely on bag-of-words (BoW) for closed-loop detection and verification. If the closed-loop candidate is verified, it is added to the global pose graph for optimization.

**2.2. Local Bundle Adjustment.** In the local bundle adjustment module based on sliding window, 5–7 keyframes are maintained. Their parameters are jointly optimized by minimizing the photometric error. The photometric error is defined as

$$\min_{\mathbf{T}_i, \mathbf{T}_j, \mathbf{p}_k \in W} E_{i,j,k}, \quad (1)$$

where  $W = \{\mathbf{T}_1, \dots, \mathbf{T}_m, \mathbf{p}_1, \dots, \mathbf{p}_n\}$  is the  $m$  keyframe poses represented as Euclidean transformation and  $n$  points of inverse depth parameterization in the sliding window.  $E_{i,j,k}$  is calculated by

$$E_{i,j,k} = \sum_{\mathbf{p} \in N_{\mathbf{p}_k}} w_{\mathbf{p}} \left\| \left( I_j[\mathbf{p}'] - b_j \right) - \frac{t_j e^{a_j}}{t_i e^{a_i}} \left( I_i[\mathbf{p}] - b_i \right) \right\|_y, \quad (2)$$

where  $N_{\mathbf{p}_k}$  denotes the neighborhood pattern of  $\mathbf{p}_k$ ;  $a$  and  $b$  are the affine light transform parameters;  $t$  denotes the exposure time;  $I$  is an image;  $w_{\mathbf{p}}$  is a heuristic weighting factor;  $\|\cdot\|_y$  is the Huber norm; and  $\mathbf{p}'$  denotes the reprojected pixel of  $\mathbf{p}$  on  $I_j$ , which is calculated by

$$E_{loop} = \sum_{\mathbf{q}_i \in Q_1} w_1 \left\| \mathbf{S}_{cr} \Pi^{-1}(\mathbf{p}_i, d_{\mathbf{p}_i}) - \Pi^{-1}(\mathbf{q}_i, d_{\mathbf{q}_i}) \right\|_2 + \sum_{\mathbf{q}_j \in Q_2} w_2 \left\| \Pi \left( \mathbf{S}_{cr} \Pi^{-1}(\mathbf{p}_j, d_{\mathbf{p}_j}) \right) - \mathbf{q}_j \right\|_2, \quad (4)$$

where  $Q_1$  and  $Q_2$  are the matched features in the current keyframe without and with depth, respectively;  $\mathbf{p}_i$  denotes the reconstructed feature in the closed-loop candidates;  $d_{\mathbf{q}_i}$  is the inverse depth of the feature  $\mathbf{q}_i$ ; and  $w_1$  and  $w_2$  are the weights to balance the different measurement units.

It can be noticed from equation (2) that the pose estimation of LDSO relies on minimizing the photometric error of the selected points. If the selected points are disturbed by imaging disturbances, equation (2) is converted into

$$E_{i,j,k} = \sum_{\mathbf{p} \in N_{\mathbf{p}_k}} w_{\mathbf{p}} \left\| \left( I_j[\mathbf{p}'] - b_j \right) - \frac{t_j e^{a_j}}{t_i e^{a_i}} \left( I_i[\mathbf{p}] - b_i \right) \right\|_y + E_n, \quad (5)$$

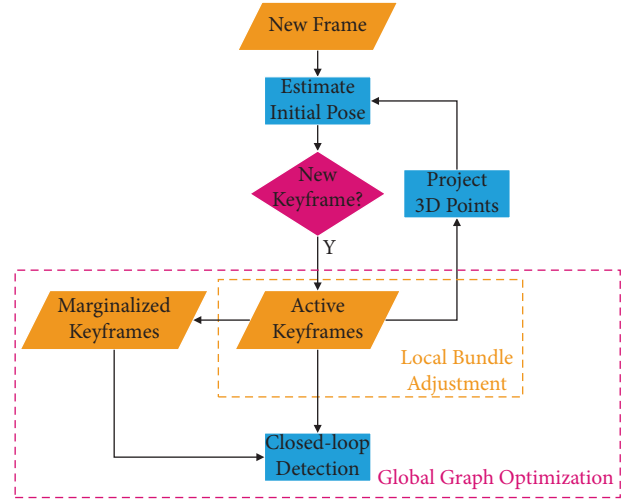


FIGURE 1: The framework of the LDSO method.

$$\mathbf{p}' = \prod \left( \mathbf{R} \Pi^{-1}(\mathbf{p}, d_{\mathbf{p}_k}) + \mathbf{t} \right), \quad (3)$$

where  $\Pi$  is the projection function from  $\mathbb{R}^3$  to  $\Omega$ ;  $\mathbf{R}$  and  $\mathbf{t}$  are the relative rotation and translation between the two frames; and  $d_{\mathbf{p}}$  is the inverse depth of point  $\mathbf{p}$ .

**2.3. Closed-Loop Detection and Verification.** In the LDSO SLAM, the DSO's point selection strategy has been modified to be more sensitive to corner points. The selected corner points are calculated as their ORB descriptors and packed into BoW. When the ORB descriptor of each keyframe is calculated, the closed-loop candidates of the keyframe are proposed by querying the BoW database. The similarity transformation from the closed-loop candidate to the current keyframe  $\mathbf{S}_{cr}$  is optimized by minimizing 3D and 2D geometric constraints:

where  $E_n$  is the error due to imaging disturbances. As the intensity of the camera imaging disturbance increases, the optimization direction for minimizing the photometric error is more inclined to the error caused by the imaging disturbances rather than the estimated pose. Therefore, the robustness of LDSO in camera imaging disturbances is not strong enough.

### 3. Proposed Method

To enhance the robustness of the direct SLAM method, the points are fused with the surrounding pixels' information. The overview of the proposed method for obtaining and using fusion points is shown in Figure 2.

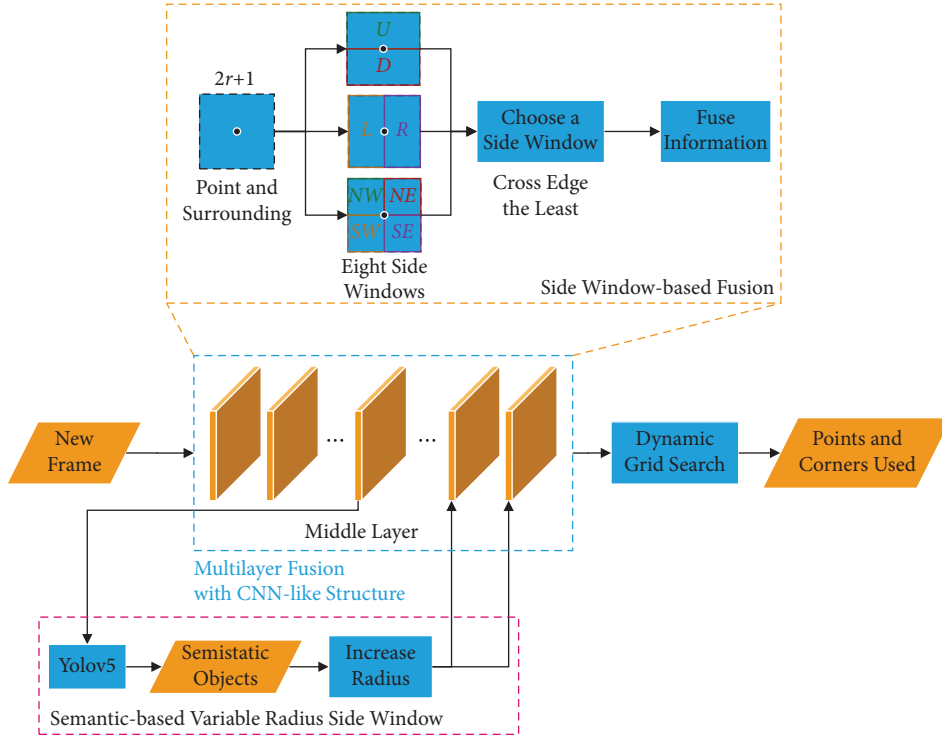


FIGURE 2: The overview of the proposed method for obtaining and using fusion points.

As shown in Figure 2, the area around each pixel is divided into blocks according to the side window strategy when a new frame arrives. The area block that crosses the image’s edge the fewest times is chosen. This region block’s pixel information is averaged into a single point. Multilayers of such pixel information fusion are superimposed to form a convolutional neural network (CNN) like structure [35, 36]. In the middle layer, semistatic objects are detected. The radiuses of the side window of the pixels belonging to the semistatic objects are increased in the back layers. The fusion points form the fused image. The points with sufficient gradient intensity and corners are selected using a dynamic grid search. These points are used in direct SLAM to improve the robustness of the system. The details of the proposed method are introduced as follows. The regional pixel information fusion is realized by multilayer fusion with a CNN-like structure. Then, the side window strategy is added to the fusion method for edge preservation. Finally, the radius of the side window is adjusted based on semantic information to reduce the weights of the semistatic objects.

**3.1. Regional Pixel Information Fusion Method.** As we know, the main reason why the robustness of feature-based SLAM is better than that of direct SLAM is that the feature carries the general information of pixels in a local area instead of a single pixel [37]. Therefore, to improve the robustness of LDSO in different camera imaging disturbances, a regional pixel information fusion method is introduced into the LDSO algorithm. Namely, each pixel can fuse the information of surrounding pixels, and the fusion intensity decreases as the distance between the pixels increases.

The mean filter is one of the most common methods of fusing pixels. Unlike other filters such as the median, max, and min filters, which select one pixel and discard others, the mean filter considers information from all pixels. In addition, the mean filter is simple to implement. So, a  $3 \times 3$  mean filter is used to fuse eight neighborhood pixels into one pixel in this study. At the same time, referring to the characteristics of the classic convolutional neural networks (CNN) [38], the mean filters are stacked in the structure of CNN. In CNN, the stacked convolutional layers are considered to extract high-level features of the image so that these feature points can be used for object classification operations. Each feature point obtained contains information about a local area. The CNN-like structure of the multilayer fusion used in this study is shown in Figure 2.

*Remark 1.* The main reason for using the  $3 \times 3$  mean filter in this paper is that it is the minimum size that can cover eight neighborhood information. Using the stacking structure, the  $3 \times 3$  receptive field can be easily expanded to  $5 \times 5$ ,  $7 \times 7$ , and other larger receptive fields. By this stacking structure, the closer the points in this area are to the edge, the fewer times they are repeatedly used and the less they affect the obtained feature points. These characteristics are precisely in line with our needs for fusing regional pixel information.

**3.2. Side Window Strategy for Pixel Fusion Area Selection.** The consistent use of the square area as the pixel fusion range can conveniently improve the overall robustness of the visual odometry, but it will also cause a certain degree of damage to the edges of the image. The more the layers are stacked, the



greater the degree of damage. In image processing, this is called nonedge preservation [39]. As mentioned earlier, the points/features selected by LDSO are pixels with sufficient intensity gradients and corner features. Pixel fusion across the edges will reduce the gradient intensity of the pixels and blur the corner features. Since it makes the selected points difficult to gather at the edge, the point cloud map constructed is very unclear. Since the corner features are blurred and difficult to be extracted, it is difficult for LDSO to detect the closed-loop effectively.

To solve the above problems caused by the nonedge preservation of pixel fusion in the fixed square area, the side window strategy is introduced into LDSO [40]. The side window strategy treats each pixel as a potential edge point. Unlike the traditional pixel fusion method that takes the pixel's position as the center of the filter window, the side window strategy aligns the edge of the filter window with the pixel. Different from nonlinear anisotropic weightings such as the spatial weighting and gray value weighting of bilateral filters, which only reduce the diffusion of pixels along the edge normal direction, the side window strategy can cut off all the normal diffusion [41].

The details of the side window strategy proposed in our multilayer fusion are as follows:

- (1) Each pixel and its surroundings are divided into eight side windows, as shown in Figure 2. They are the side windows in eight directions: up ( $U$ ), down ( $D$ ), left ( $L$ ), right ( $R$ ), northwest ( $NW$ ), northeast ( $NE$ ), southwest ( $SW$ ), and southeast ( $SE$ ). The center point  $p_i$  of the pixel fusion is located on the side or corner of the window. The radius  $r$  of the side window determines the range of the pixel fusion.
- (2) The average value of the pixels in each side window is calculated as the output  $q_n$  of the side window, where  $n \in \{U, D, L, R, NW, NE, SW, SE\}$ .
- (3) Compare the distance measured by  $L_1$  norm between the output  $q_n$  of the eight side windows and the center point  $p_i$ . The fusion output  $p^{\text{fusion}}$  of the center point  $p_i$  and its surrounding pixels is  $p^{\text{fusion}} = q_s$ , where

$$s = \arg \min_{n \in \{U, D, L, R, NW, NE, SW, SE\}} \{|q_n - p_i|\}. \quad (6)$$

*Remark 2.* In the proposed multilayer superimposed pixel fusion strategy, the diffusion of pixels along the normal edge direction will be further amplified. And the side window strategy cuts off the possibility of pixels spreading along the normal direction of the edge, which is more suitable for our multilayer fusion.

The pseudocode of the proposed side window-based multilayer fusion method is summarized in Algorithm 1.

*3.3. Semantic-Based Variable Radius Side Window Strategy.* When humans use their eyes to estimate their position and remember the environment, they do not take all the objects

they see into consideration. Instead, they focus on static objects such as walls and pillars and use semistatic objects that are stationary most of the time, such as cars parked on the side of the road, as a reference. Inspired by this, a semantic-based variable radius side window strategy is proposed to assign weights to static and semistatic objects.

First, in the first half of the stacked structure of pixel fusion, a smaller radius for the side window is used. In multilayer pixel fusion, due to the smaller coverage area, the side window with a smaller radius can make the image retain more details such as edges while reducing the impact of camera imaging disturbances. Subsequent object detection in a camera imaging disturbed environment is carried out on this basis.

Second, Yolov5 (one of the popular object detection deep networks) is used to distinguish static and semistatic objects in the input images. Yolov5 is the latest version of the Yolo object detection algorithm [42, 43]. The main reason for using the Yolov5 network is that Yolov5 can also maintain a higher processing frame rate under lower hardware conditions while achieving the accuracy of the current state-of-the-art technology. In this study, the pretrained Yolov5 model on the Microsoft COCO (Common Objects in Context) dataset is used to extract object location and category semantic information [44]. Common movable categories such as bicycles, cars, motorcycles, buses, and trucks in the COCO dataset are marked as semistatic objects.

Third, in the second half of the stacked structure of the pixel fusion, a slightly larger radius is used for the side windows of the regions where the semistatic objects are detected. A side window with a larger radius is more likely to contain more image edges. The selection principle of the side window is to select the side window whose output is most similar to the center pixel. The larger the edge gradient of the image within the coverage of the side window, the more dissimilar the output is from the center pixel. Therefore, the side window strategy is more inclined to retain the image edges with large gradients. Edges with smaller gradients in the side window will be blurred. With repeated pixel fusion, the obvious image edges in the semistatic object area will be preserved, while the pixel gradients inside will be reduced.

*Remark 3.* The specific gravity of the point in the semistatic object area selected by the LDSO with a high gradient intensity will decrease. The preserved obvious image edges can provide enough corner features for LDSO. In this way, a static object-based and semistatic object-assisted approach similar to the human positioning strategy is achieved.

A summary of the proposed points selection strategy based on the side window with semantic-based variable radius is given in Algorithm 2.

Overall, the workflow of the proposed variable radius side window direct SLAM method is summarized as follows:

*Step 1.* The radius parameters applicable to different regions are selected based on semantic information.

```

Input: Image  $I$ , Layer number  $L$ , Radius of side window  $r$ 
Output: Set of fusion points
(1) for  $\forall l \in L$  do
(2)   for  $\forall \{x_i, y_i\} \in I$  do
(3)      $S = \{(x_i - r): (x_i + r), (y_i - r): (y_i + r)\}$ ;
(4)     %  $S$  is the surrounding of the pixel  $p_i$ 
(5)     Divide  $S$  into  $\{U, D, L, R, NW, NE, SW, SE\}$ ;
(6)     for  $n \in \{U, D, L, R, NW, NE, SW, SE\}$  do
(7)        $q_n = \text{mean}(p_j), j \in n$ ;
(8)     end for
(9)      $s = \arg \min\{|q_n - p_i|\}$ ;
(10)    % Select the side window  $s$ ;
(11)     $p^{\text{fusion}} = q_s$ ;
(12)  end for
(13) end for

```

ALGORITHM 1: Side window-based multilayer fusion.

```

Input: Number of layers  $L$ , Desired number of points  $N_{\text{des}}$ 
Output: Selected points
(1) for  $\forall l \in L$  do
(2)   if  $l < 1/2L$  then
(3)     Use small radius side windows for multilayer fusion;
(4)   end if
(5)   if  $l \geq 1/2L$  then
(6)     Use Yolov5 to distinguish static and semistatic objects;
(7)     Increase the radius of the side windows of the regions where the semistatic objects are detected;
(8)   end if
(9) end for
(10) Split the image composed of fusion points into patches;
(11) while  $N_{\text{sel}} < N_{\text{des}}$  do
(12)   Randomly select a patch  $M$ 
(13)   Compute the median of gradient as the region-adaptive threshold;
(14)   Split  $M$  into  $d \times d$  blocks;
(15)   Select a point with the highest gradient which surpasses the gradient threshold from  $d \times d, 2d \times 2d, 4d \times 4d$  blocks separately;
(16) end while

```

ALGORITHM 2: Semantic variable radius side window-based points selection.

*Step 2.* The different radius parameters are applied to the side window strategy to form a variable radius side window strategy.

*Step 3.* The semantic information-based variable radius side window strategy is applied to a multilayer stacked pixel blender to fuse regional pixel information.

*Step 4.* The points are selected according to Algorithm 2 on the points fused with local information.

*Step 5.* The selected points are used to estimate the camera pose by minimizing equation (2) and perform global optimization by minimizing equation (4) when loop closures are detected.

## 4. Experimental Results and Analysis

In this section, the proposed method is comprehensively evaluated on outdoor datasets (KITTI dataset) and indoor datasets (TUM RGB-D dataset), which are introduced as follows:

- (1) KITTI dataset [45, 46]: this dataset is currently the most extensive dataset in the world for evaluating computer vision algorithms in autonomous driving scenarios. It contains real image data collected in outdoor scenes such as urban areas, villages, and highways. The “00–10” sequences in this dataset provide ground truth, which are used in this study.
- (2) TUM RGB-D dataset [47, 48]: this dataset provides RGB-D data and ground-truth data intending to establish a novel benchmark for the evaluation of



FIGURE 3: Comparison of the example scene before and after adding noise. (a) The original image. (b) The image after adding Gaussian noise. (c) The image after adding Salt-and-Pepper noise. Note that the effect of the added noise is noticeable.

visual odometry and visual SLAM systems. In this paper, the sequences “freiburg1\_xyz,” “freiburg2\_xyz,” “freiburg2\_rpy,” “freiburg1\_desk,” and “freiburg1\_desk2” are selected, which were all acquired in the office interior scene with rich texture.

The main reason for using the two datasets is that both of them provide ground truth, which is required for the quantitative evaluation. Because there is a certain natural camera overexposure problem in the two datasets [49], they are used directly to test the proposed method under the disturbance of camera overexposure. In addition, Gaussian noise and Salt-and-Pepper noise are added to the two datasets in these experiments to further test the proposed method under different camera sensor noises. In this paper, the variance of Gaussian noise added is 0.003, and the rate of Salt-and-Pepper noise added is 10%. The noise addition operation and the noise-adding parameters in this study are relatively common in the literature [50, 51]. Figure 3 shows an example scene before and after adding two kinds of noise.

**4.1. Quantitative Evaluation.** In this study, the proposed method is based on the side window fusion strategy on the direct method-based SLAM. Here, it is compared with the general direct sparse odometry method (DSO) and the general direct sparse odometry with loop closure (LDSO). In this paper, the large-scale direct monocular SLAM (LSD-SLAM) is not compared because its tracking robustness is not as good as DSO [52]. To further discuss the performance of our method, ORB-SLAM3 is also added for comparison, which is one of the state-of-the-art methods based on the feature-based method [53, 54]. The root mean squared error of absolute trajectory error (RMSE<sub>ATE</sub>) is used to evaluate the performance of these methods [55].

**4.1.1. On the KITTI Dataset.** Firstly, some comparison experiments are conducted on the KITTI dataset to show the robustness of the proposed method in the face of different camera imaging disturbances. The results with no noise added, Gaussian noise, and Salt-and-Pepper noise are listed

in Tables 1–3, respectively. The missing values in the tables mean tracking failures.

The results in Table 1 show that our method can achieve similar or better performance compared with the other direct methods in the sequences without added noise. The results on the sequences without added noise show that the performance of the proposed method is obviously better than the general LDSO method on the sequences “KITTI\_00” and “KITTI\_02,” where the RMSE values of the proposed method are 32.42% and 51.91% less than the general LDSO method. The main reason is that the sequences “KITTI\_00” and “KITTI\_02” have a large number of scenes in the shade of trees (see Figures 4(a) and 4(c)), and frequent changes in ambient light bring more frequent camera overexposure problems to the images. The results show that the proposed method can deal with the camera overexposure interference on the direct methods effectively.

In the sequences with Gaussian noise, we can see that the performance of the general direct methods decreases obviously on all of the sequences in the KITTI dataset, but the proposed method is not seriously affected by the Gaussian noise (see Table 2). In particular, the other direct methods fail to track in sequence “KITTI\_03,” “KITTI\_04,” and “KITTI\_09” while our method still works. The results in Table 2 show that the proposed method outperforms the general LDSO method by more than 13.7% on all of the sequences in the KITTI dataset. In the sequences with Salt-and-Pepper noise, DSO and LDSO are entirely inoperable, while our method obtains good performance (see Table 3).

Compared with ORB-SLAM3, our method obtains slightly better performance on the sequences without added noise, except sequences “KITTI\_08,” “KITTI\_09,” and “KITTI\_10.” The main reason is that these sequences contain very rich textures that are more suitable for feature-based methods. In particular, ORB-SLAM3 will track failure in the sequence “KITTI\_01,” whether the noise is added or not. This is due to the fact that the sequence “KITTI\_01” is a very texture-deficient highway scene and is not suitable for feature-based SLAM methods (see Figure 4(b)).

TABLE 1: RMSE<sub>ATE</sub> on KITTI dataset with no noise added.

Method	No noise added										Average	
	KITTI_00	KITTI_01	KITTI_02	KITTI_03	KITTI_04	KITTI_05	KITTI_06	KITTI_07	KITTI_08	KITTI_09		KITTI_10
DSO [21]	115.035	31.811	152.463	2.030	0.755	49.981	54.004	17.576	114.391	70.534	14.661	56.658
LDSO [34]	7.360	9.972	47.245	2.342	0.800	4.166	12.805	1.691	114.739	69.803	14.815	25.976
ORB-SLAM3 [54]	9.265	—	22.025	2.117	1.223	4.034	16.196	1.688	38.114	7.243	7.771	10.968
Ours	4.974	9.710	22.722	2.183	0.857	3.540	12.798	1.789	99.579	52.469	14.210	20.439

Note. “—” means tracking failure. The average value is calculated based on the number of successes.

TABLE 2: RMSE<sub>ATE</sub> on KITTI dataset with Gaussian noise.

Method	Gaussian noise										Average	
	KITTI_00	KITTI_01	KITTI_02	KITTI_03	KITTI_04	KITTI_05	KITTI_06	KITTI_07	KITTI_08	KITTI_09		KITTI_10
DSO [21]	115.771	56.143	185.187	—	—	50.185	59.382	38.812	127.674	—	15.287	81.055
LDSO [34]	22.543	23.052	169.247	—	—	44.010	58.729	53.481	130.993	—	16.277	64.792
ORB-SLAM3 [54]	10.645	—	59.868	2.860	1.911	9.250	19.249	1.932	42.931	8.223	8.776	16.565
Ours	17.772	13.023	120.380	2.133	1.093	5.740	13.491	1.973	102.206	52.664	14.042	31.320

Note. “—” means tracking failure. The average value is calculated based on the number of successes.

TABLE 3: RMSE<sub>ATE</sub> on KITTI dataset with Salt-and-Pepper noise.

Method	Salt-and-Pepper noise										Average	
	KITTI_00	KITTI_01	KITTI_02	KITTI_03	KITTI_04	KITTI_05	KITTI_06	KITTI_07	KITTI_08	KITTI_09		KITTI_10
DSO [21]	--	--	--	--	--	--	--	--	--	--	--	--
LDSO [34]	--	--	--	--	--	--	--	--	--	--	--	--
ORB-SLAM3 [54]	--	--	--	--	--	--	--	--	--	--	--	--
Ours	19.798	10.464	108.448	2.252	0.806	11.581	12.463	2.238	101.590	52.177	14.937	30.614

Note. "--" means tracking failure.

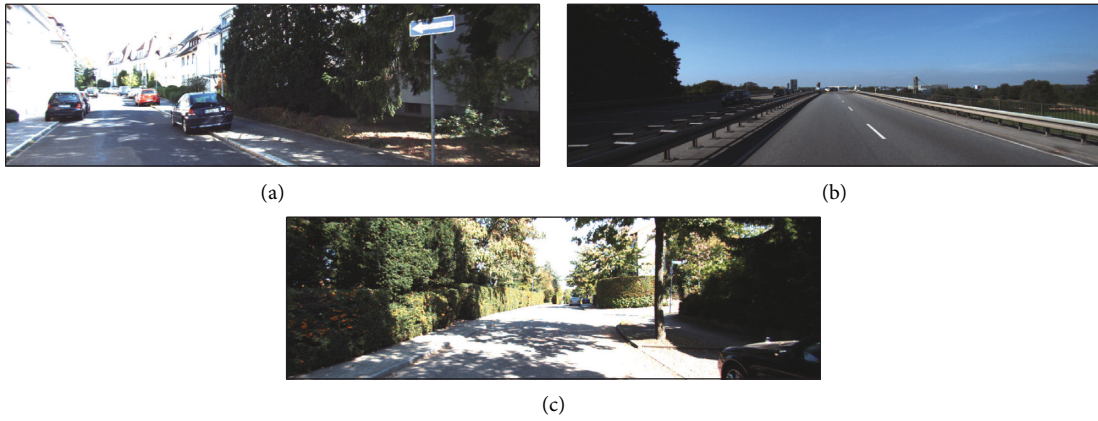


FIGURE 4: Example scenes for sequences “KITTI\_00,” “KITTI\_01,” and “KITTI\_02” in the KITTI dataset: (a) is from the sequence “KITTI\_00”; (b) is from the sequence “KITTI\_01”; (c) is from the sequence “KITTI\_02.” The sequences “KITTI\_00” and “KITTI\_02” are the sequences with more camera overexposure interference, while “KITTI\_01” is the sequence with little camera overexposure interference.

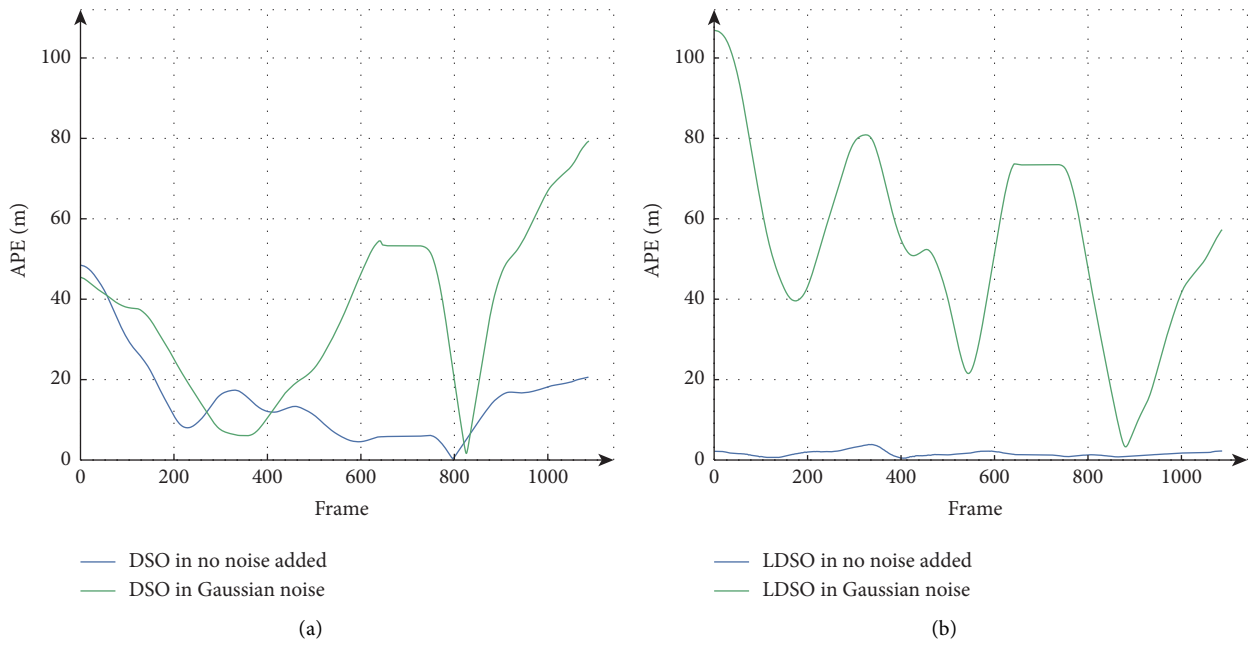


FIGURE 5: Continued.

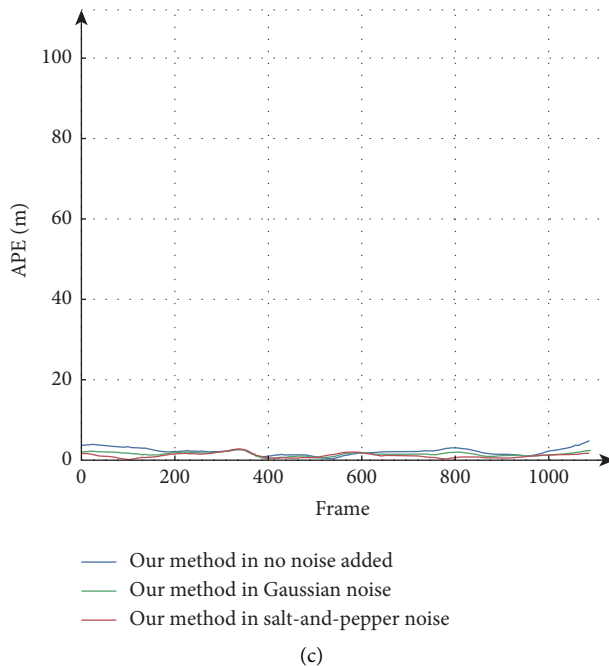


FIGURE 5: Comparison of APE with respect to translation in different noises on the sequence “KITTI\_07”: (a) APE of DSO. (b) APE of LDSO. (c) APE of our method. Note that the performance gap of our method is significantly smaller than that of DSO and LDSO. In particular, DSO and LDSO do not work in the sequence with Salt-and-Pepper noise, and the results in this case cannot be added for comparison. Besides, the performance of our method is better than the others overall.



FIGURE 6: Blurred image with smears in TUM RGB-D dataset.

TABLE 4: RMSE<sub>ATE</sub> on TUM RGB-D dataset with no noise added.

Method	No noise added				
	fr1_xyz	fr2_xyz	fr2_rpy	fr1_desk	fr1_desk2
LDSO [34]	0.061	0.011	0.046	0.774	0.904
Ours	0.063	0.012	0.043	0.780	0.905

Although ORB-SLAM3 performs better in the face of Gaussian noise (see Table 2), ORB-SLAM3 is unavailable under the influence of Salt-and-Pepper noise (see Table 3). By contrast, the results show that our method performs more consistently in different camera imaging disturbances than other methods (see Tables 2 and 3).

To compare the robustness in different camera imaging disturbances more clearly, the absolute pose errors (APE)

TABLE 5: RMSE<sub>ATE</sub> on TUM RGB-D dataset with Gaussian noise.

Method	Gaussian noise				
	fr1_xyz	fr2_xyz	fr2_rpy	fr1_desk	fr1_desk2
LDSO [34]	—	0.096	—	0.518	—
Ours	0.156	0.010	0.060	0.801	0.756

Note. “—” means tracking failure.

TABLE 6: RMSE<sub>ATE</sub> on TUM RGB-D dataset with Salt-and-Pepper noise.

Method	Salt-and-Pepper noise				
	fr1_xyz	fr2_xyz	fr2_rpy	fr1_desk	fr1_desk2
LDSO [34]	—	—	—	0.841	—
Ours	0.129	0.011	0.058	0.796	0.871

Note. “—” means tracking failure.

with respect to translation on the example sequence “KITTI\_07” in different noises are shown in Figure 5. Here, the main reason for using the sequence “KITTI\_07” as the example is that this sequence has a medium sequence length in the KITTI dataset. In the next part of this paper, the sequence “KITTI\_07” is also used as the study object, where the reason is not further explained. The lack of the APE curves of DSO and LDSO in the sequence with Salt-and-Pepper noise is due to their inability to work. Notice that our method has more consistent APE curves in different noises, and all the APEs of our method are less than 5%. This experiment highlights that our strategy effectively improves



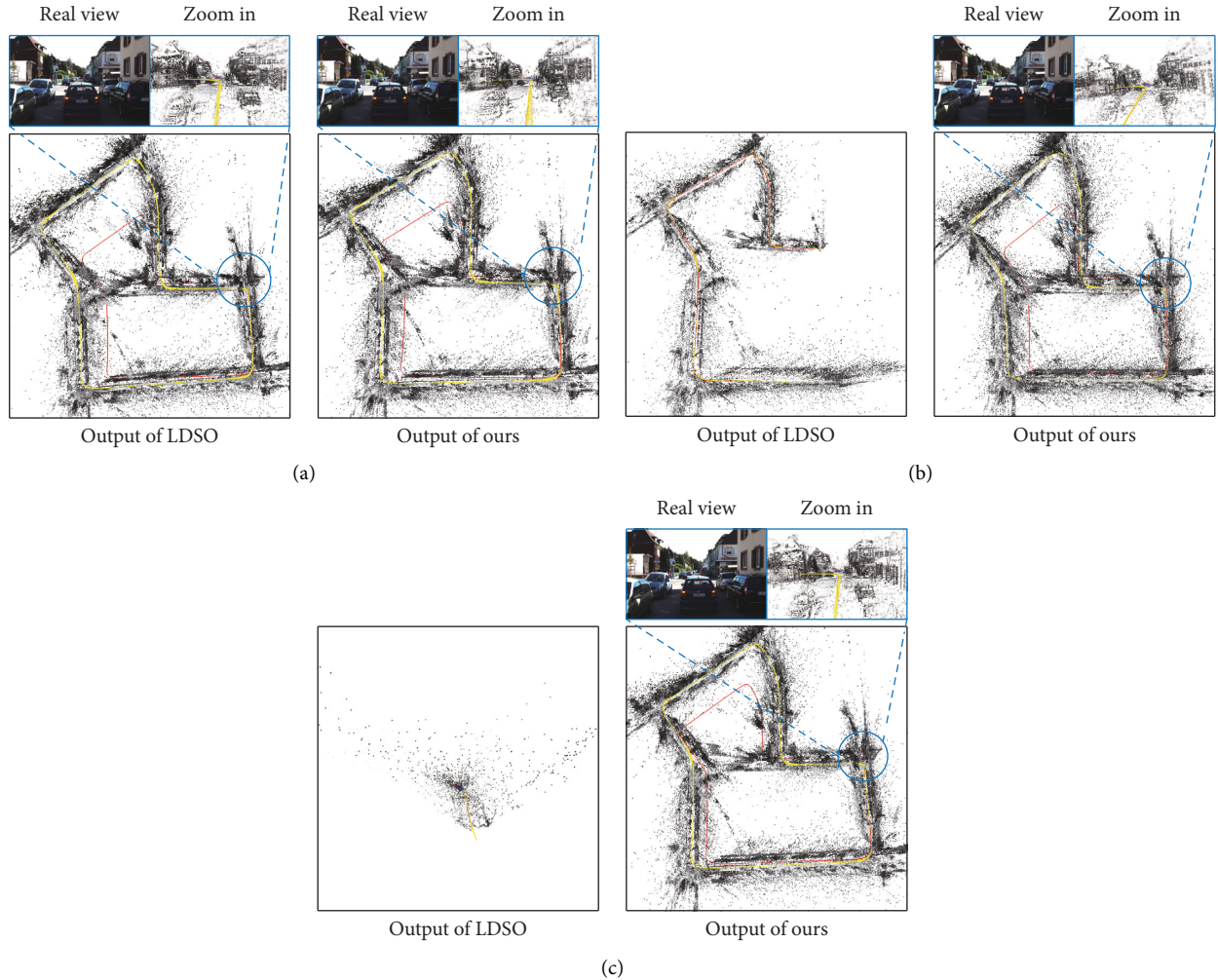


FIGURE 7: Sample outputs of the sequence “KITTI\_07”: (a), (b), and (c) are the outputs on the sequence with no added noise, Gaussian noise, and Salt-and-Pepper noise, respectively. Left: LDSO’s outputs. Right: our method’s outputs. Note that, in the sequence without adding noise, the quality of our method’s trajectory estimation and point cloud map construction is similar to that of LDSO. In the sequence with Gaussian noise added to LDSO, the closed-loop cannot be detected, and the trajectory estimation in the second half is wrong. LDSO does not work in the sequence with Salt-and-Pepper noise added. Our strategy achieves a more robust performance under different noise interferences.

the robustness of direct SLAM when facing different camera imaging disturbances outdoors.

**4.1.2. On the TUM RGB-D Dataset.** Secondly, some experiments are conducted on the TUM RGB-D dataset to verify whether our strategy has the effect of improving robustness in indoor environments. Since DSO and LDSO perform very similarly in this case, our method is only compared with LDSO. In this dataset, there are many blurred images with smears, as shown in Figure 6. In this experiment, because the selected sequences are relatively short, the difference in RMSE is not apparent. Thus, we mainly compare whether the tracking of the SLAM system based on different methods is successful. The results are shown in Tables 4–6.

The results in this experiment show that the SLAM system will fail easily after adding noise to the images. Note that, in the sequences in which no noise is added, both our

method and LDSO can track successfully (see Table 4). After adding different noises to the sequences, LDSO becomes more prone to failure tracking, while our method still tracks successfully (see Tables 5 and 6). This experiment highlights that our approach can still improve the robustness of direct SLAM under different camera imaging disturbances when faced with a poor indoor image input.

**4.2. Qualitative Evaluation.** This section mainly conducts a qualitative evaluation of the completeness and the clarity of the predicted trajectory map and the constructed point cloud map in the camera imaging disturbances. Examples of the point cloud map constructed on the sequence “KITTI\_07” are shown in Figure 7. The results show that our method is similar to LDSO in the absence of noise interference. When disturbed by Gaussian noise and Salt-and-Pepper noise, LDSO is negatively affected to varying degrees, while our method has a better and more stable

performance in the trajectory prediction and point cloud map construction. The main reason is that our method uses the multilayer pixel fusion features based on the side window strategy instead of directly using the original pixels, which can improve the robustness of the direct method-based SLAM in different camera imaging disturbances.

## 5. Discussion

The total performances of the proposed method have been proved on different datasets by some comparison experiments in Section 4. In this section, some additional comparison experiments are conducted to discuss the performance of our method in different intensities of camera imaging disturbances. In addition, the performance of the key improvement of the proposed method, namely, the points selection strategy, is further discussed. At last, the proposed method is tested in real-world applications to demonstrate the effectiveness of the proposed method.

*5.1. Performance in Camera Imaging Disturbances of Different Intensities.* Firstly, the performance of our method in the camera imaging disturbances of different intensities is discussed, where some expanded comparison experiments are conducted under the sensor noise of different intensities and the camera overexposure with different frequencies.

*5.1.1. About Different Noise Intensities.* The performance of our method in the camera sensor noise of different intensities is discussed on the sequence “KITTI\_07.” The comparison experiments are carried out separately in Gaussian noise and Salt-and-Pepper noise with different intensities. The variance of Gaussian noise ranges from 0.001 to 0.009 and is incremented by a step size of 0.002. The rate of Salt-and-Pepper noise added ranges from 2% to 10%, and the step size is 2%. The results are shown in Tables 7 and 8. For Gaussian noise, DSO tracking fails when the variance is greater than 0.005. LDSO tracking fails when the variance is greater than 0.003. Our method tracks successfully at all noise intensities and performs stably when the variance is below 0.005. This reflects that our method is more robust than other direct methods in different intensities of Gaussian noise. For Salt-and-Pepper noise, both DSO and LDSO fail to track when the noise addition rate is greater than 2%. Our method can track successfully and perform stably at all noise addition rates. It can be seen that our method is more robust than other direct methods in different intensities of Salt-and-Pepper noise. ORB-SLAM3 can also track successfully in all intensities of Gaussian noise and performs stably when the variance is below 0.007. While ORB-SLAM3 outperforms our method in robustness under different intensities of Gaussian noise, it fails to track at all addition rates of Salt-and-Pepper noise.

TABLE 7: RMSE<sub>ATE</sub> comparison in Gaussian noise of different intensities.

Variance	DSO [21]	LDSO [34]	ORB-SLAM3 [54]	Ours
0.001	24.396	2.504	1.872	1.655
0.003	38.812	53.481	1.932	1.973
0.005	45.968	—	2.101	2.946
0.007	—	—	2.566	12.343
0.009	—	—	10.242	13.816

Note. “—” means tracking failure.

TABLE 8: RMSE<sub>ATE</sub> comparison in Salt-and-Pepper noise of different intensities.

Addition rate (%)	DSO [21]	LDSO [34]	ORB-SLAM3 [54]	Ours
2	35.500	35.525	—	1.409
4	—	—	—	1.602
6	—	—	—	1.755
8	—	—	—	2.225
10	—	—	—	2.238

Note. “—” means tracking failure.

TABLE 9: RMSE<sub>ATE</sub> comparison under interference of camera overexposure at different frequencies.

Interval frames	DSO [21]	LDSO [34]	ORB-SLAM3 [54]	Ours
30	32.946	11.866	—	11.522
25	34.257	12.248	—	11.836
20	—	22.240	—	11.885
15	—	—	—	13.333
10	—	—	—	—

Note. “—” means tracking failure.

*5.1.2. About Different Overexposure Frequencies.* To discuss the performance of our method under the interference of camera overexposure, the sequence “KITTI\_01,” which suffers little from camera overexposure, is experimented with adding simulated camera overexposure disturbance at different frequencies. The camera overexposure addition operation in this study is similar to other pieces of literature [56]. The number of interval frames at which overexposure interference is added ranges from 30 to 10 and is decreased by a step size of 5. The results are shown in Table 9.

The results in Table 9 show that our method performs close to LDSO when the camera overexposure interference is not very serious. However, when the overexposure interference interval is 20 frames, the proposed method outperforms the general LDSO method by more than 46%. In addition, LDSO starts to fail to track when the overexposure interference interval is lower than 15 frames, while our method can still work when the overexposure interference interval is bigger than 10 frames. ORB-SLAM3 fails to track in the sequence “KITTI\_01” under the added camera overexposure interference. The results of this experiment show that the proposed method has better performance under the camera overexposure interference.

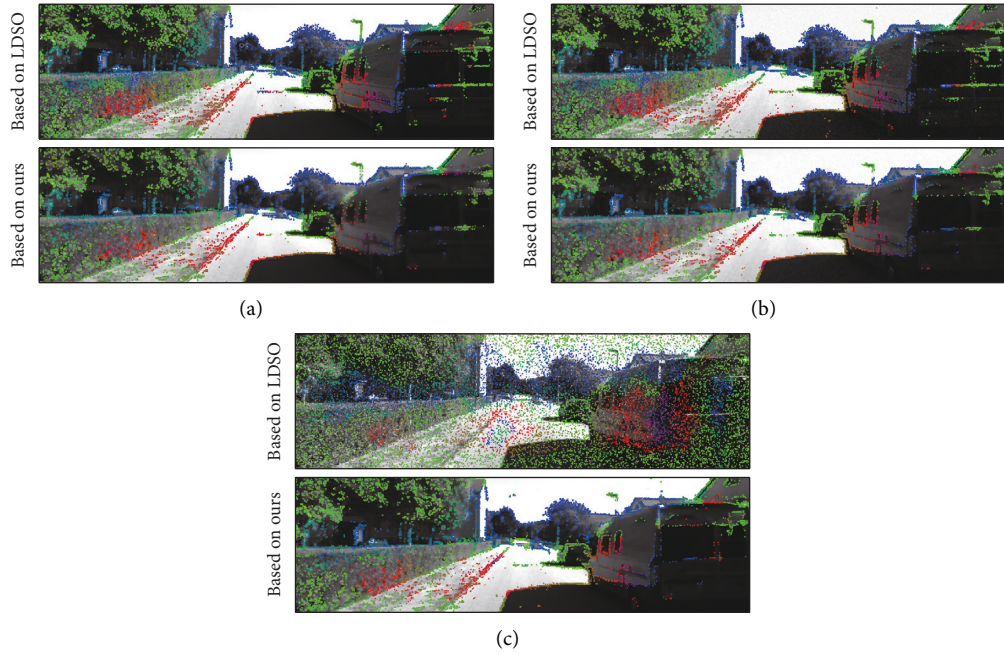


FIGURE 8: Point selection results of our strategy and LDSO in different noises: (a), (b), and (c) are the images from the KITTI dataset with no added noise, Gaussian noise, and Salt-and-Pepper noise, respectively. Top rows: point selection results of LDSO. Bottom rows: point selection results of our strategy. Note that the points selected by our strategy are more consistent in different noises. Moreover, on semistatic objects such as cars parked on the side of the road, the points selected by our approach are significantly less than those by LDSO and are mainly distributed on the apparent edges.

TABLE 10: RMSE<sub>ATE</sub> comparison of whether using semantic-based variable radius side window.

Method	No noise added		Gaussian noise		Salt-and-Pepper noise	
	KITTI_07	KITTI_08	KITTI_07	KITTI_08	KITTI_07	KITTI_08
FR-SW	2.256	106.652	2.794	112.754	2.471	106.093
SVR-SW	1.789	99.579	1.973	102.206	2.238	101.590



FIGURE 9: Some images in the real scene added with noise. (a) Added with Gaussian noise. (b) Added with Salt-and-Pepper noise.

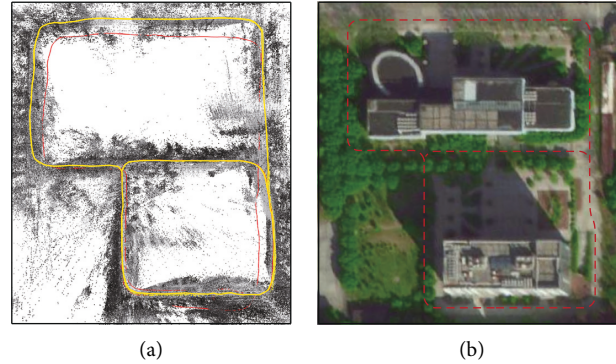


FIGURE 10: Result of experiment in real scene. (a) The trajectory estimated by our method, which is marked with a yellow curve. (b) The approximate trajectory on the satellite map, which is marked with a red dashed line.

**5.2. About Points Selection Strategy.** Secondly, the effect of the points selection strategy of our method to improve the robustness of direct method-based SLAM is discussed. Figure 8 shows the selection of points in an example scene with different types of noise. Here, our points selection strategy is compared with that of the general LDSO. It is easy to notice that the points selected by our strategy are more consistent in different noises. It is not easy for LDSO to detect closed loops under the influence of Gaussian noise. Gaussian noise creates texture in untextured areas. These textures are selected as the basis for closed-loop detection, which easily leads to the failure of closed-loop detection. In Salt-and-Pepper noise, LDSO is entirely inoperable. The reason is that the image-gradient-based features selected by LDSO are easily located at the position of the Salt-and-Pepper noise (see Figure 8(c)). These randomly generated noise positions cannot be used as the basis for estimating camera pose. As shown in Figure 8(a), the points selected by our method are significantly less than that by LDSO and are mainly distributed on the apparent edges of the semistatic objects such as cars parked on the side of the road. The consistent selection of points of our method improves the robustness of direct method-based SLAM.

The comparison results of  $RMSE_{ATE}$  based on the proposed semantic-based variable radius side window (SVR-SW) and the fixed radius side window in the general LDSO (FR-SW) are shown in Table 10. Here, the sequences “07” and “08” of the KITTI dataset are used, which contain more semistatic objects. It can be noticed that the proposed SVR-SW strategy achieves better performance on different noises. The main reason is that the semantic-based variable radius side window can reduce the weight of selected points of semistatic objects to improve the performance of direct method-based SLAM in scenes with more semistatic objects.

**5.3. Experiment in Real Scene.** Thirdly, to discuss the performance of our method in real scenes, an experiment is conducted on a real-world dataset collected outdoors by the Zenmuse X5S camera mounted on the DJI Inspire 2 drone [57]. In reality, the camera imaging disturbances often do not exist all the time but are sudden and random. For

simulation of this situation, Gaussian noise and Salt-and-Pepper noise are artificially added to parts of this dataset. Some images added with noise are shown in Figure 9, which have obvious brightness changes due to the shade of trees and lots of semistatic objects in the real scene, such as bicycles and cars. The real-world dataset is collected along the road to easily judge whether our method estimates the correct trajectory using the satellite map. The experimental result of this self-collected real dataset is shown in Figure 10. It can be seen that the trajectory estimated by our method does not deviate from the road due to the camera imaging disturbances, including the artificially added noise and the natural brightness changes. Our method performs good robustness on different camera imaging disturbances in real scenes.

## 6. Conclusion

The robustness in the camera imaging disturbances of the direct method-based SLAM is studied in this paper, and a concept of side windows is introduced into this visual SLAM system. Based on this concept, a multilayer stacked pixel blender is used to process the input images, which can significantly reduce the blurring effects on the edges of the images. In addition, the size of the fusion window can be adjusted based on semantic information to reduce the proportion of selected points on semistatic objects. At last, to more clearly evaluate the robustness of the proposed method under different camera imaging disturbances, the public datasets enhanced with different camera imaging disturbances are used to perform detailed quantitative and qualitative experiments. The results demonstrate that our strategy can improve the robustness of the direct method-based SLAM against the different camera imaging disturbances, including various sensor noises and camera over-exposure. Furthermore, the results of the real-world experiment show that the proposed method can work efficiently in real-world applications. In the future, how to further improve the robustness of the visual SLAM method while improving efficiency by using different fusion methods should be studied, such as deep neural networks.

## Data Availability

Publicly available datasets were analyzed in this study. These data can be found at [http://www.cvlibs.net/datasets/kitti/eval\\_odometry.php](http://www.cvlibs.net/datasets/kitti/eval_odometry.php) and <https://vision.in.tum.de/data/datasets/rgbd-dataset/download>.

## Conflicts of Interest

The authors declared that they have no conflicts of interest in this work.

## Acknowledgments

This work was supported by the National Natural Science Foundation of China (61873086) and the Science and Technology Support Program of Changzhou (CE20215022).

## References

- [1] J. Chang, N. Dong, and D. Li, "A real-time dynamic object segmentation framework for SLAM system in dynamic scenes," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, 2021.
- [2] J. Ni, L. Wu, X. Yang, and X. Y. Simon, "Bioinspired intelligent algorithm and its applications for mobile robot control: a survey," *Computational Intelligence and Neuroscience*, vol. 2016, Article ID 3810903, 16 pages, 2016.
- [3] Y. Ying, H. Yan, Z. Li, K. Feng, and X. Feng, "Loop closure detection based on image covariance matrix matching for visual SLAM," *International Journal of Control, Automation and Systems*, vol. 19, no. 11, pp. 3708–3719, 2021.
- [4] J. Ni, C. Wang, X. Fan, and X. Y. Simon, "A bioinspired neural model based extended Kalman filter for robot SLAM," *Mathematical Problems in Engineering*, 905826, vol. 2014, 11 pages, 2014.
- [5] H. Deilamsalehy and C. H. Timothy, "Sensor fused three-dimensional localization using IMU, camera and LiDAR," in *Proceedings of the IEEE Sensors*, Orlando, FL, USA, October 2016.
- [6] W. Xie, P. Xiaoping Liu, and M. Zheng, "Moving object segmentation and detection for robust RGBD-SLAM in dynamic environments," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, 2021.
- [7] J. Zhao, T. Li, Y. Tong, L. Zhao, and S. Huang, "2D laser SLAM with closed shape features: fourier series parameterization and submap joining," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 1527–1534, 2021.
- [8] J. Ni, Y. Chen, K. Wang, and X.Y. Simon, "An improved vision-based SLAM approach inspired from animal spatial cognition," *International Journal of Robotics and Automation*, vol. 34, no. 5, pp. 491–502, 2019.
- [9] T. H. Nguyen, T.-M. Xie, and L. Xie, "Tightly-coupled ultra-wideband-aided monocular visual SLAM with degenerate anchor configurations," *Autonomous Robots*, vol. 44, no. 8, pp. 1519–1534, 2020.
- [10] Z. Liang and C. Wang, "A semi-direct monocular visual SLAM algorithm in complex environments," *Journal of Intelligent and Robotic Systems: Theory and Applications*, vol. 101, no. 1, 2021.
- [11] H.-J. Liang, J. Sanket, C. Aloimonos, and Y. Aloimonos, "Salientdso: bringing attention to direct sparse odometry," *IEEE Transactions on Automation Science and Engineering*, vol. 16, no. 4, pp. 1619–1626, 2019.
- [12] J.-W. Kam, H.-S. Kim, S.-J. Lee, and S.-S. Hwang, "Robust and fast collaborative augmented reality framework based on monocular SLAM," *IEIE Transactions on Smart Processing and Computing*, vol. 9, no. 4, pp. 325–335, 2020.
- [13] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.
- [14] S. Sarhan, A. A. Nasr, and M. Y. Shams, "Multipose face recognition-based combined adaptive deep learning vector quantization," *Computational Intelligence and Neuroscience*, vol. 2020, Article ID 8821868, 11 pages, 2020.
- [15] E. Rublee, R. Vincent, K. Kurt, and B. Gary, "ORB: an efficient alternative to SIFT or SURF," in *Proceedings of the 2011 International conference on computer vision*, pp. 2564–2571, IEEE, Barcelona, Spain, November 2011.
- [16] E. Rosten, R. Porter, and T. Drummond, "Faster and better: a machine learning approach to corner detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 1, pp. 105–119, 2008.
- [17] M. Calonder, V. Lepetit, and M. Ozuysal, "BRIEF: computing a local binary descriptor very fast," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 7, pp. 1281–1298, 2011.
- [18] Ke Wang, S. Ma, R. Fan, and J. Lu, "SBAS: salient bundle adjustment for visual SLAM," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, 2021.
- [19] J. Ni, T. Gong, Y. Gu, J. Fan, and X. Fan, "An improved deep residual network-based semantic simultaneous localization and mapping method for monocular vision robot," *Computational Intelligence and Neuroscience*, vol. 2020, Article ID 7490840, 14 pages, 2020.
- [20] J. Engel, T. Schöps, and D. Cremers, "LSD-SLAM: large-scale direct monocular SLAM," in *Proceedings of the 13th European Conference on Computer Vision, ECCV 2014*, pp. 834–849, Springer, Zurich, Switzerland, September 2014.
- [21] J. Engel, V. Koltun, and D. Cremers, "Direct sparse odometry," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 3, pp. 611–625, 2017.
- [22] C. Forster, Z. Zhang, M. Gassner, M. Werlberger, and D. Scaramuzza, "SVO: semidirect visual odometry for monocular and multicamera systems," *IEEE Transactions on Robotics*, vol. 33, no. 2, pp. 249–265, 2016.
- [23] N. Yang, R. Wang, X. Cremers, and D. Cremers, "Challenges in monocular visual odometry: photometric calibration, motion bias, and rolling shutter effect," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 2878–2885, 2018.
- [24] K. Zhu, X. Jiang, Z. Gao, H. Fujita, and J.-N. Hwang, "Photometric transfer for direct visual odometry," *Knowledge-Based Systems*, vol. 213, Article ID 106671, 2021.
- [25] P. Liu, X. Yuan, C. Song, C. Liu, and Z. Li, "Real-time photometric calibrated monocular direct visual SLAM," *Sensors*, vol. 19, no. 16, p. 3604, 2019.
- [26] C. Sheng, S. Pan, W. Gao, Y. Tan, and T. Zhao, "Dynamic-DSO: direct sparse odometry using objects semantic information for dynamic environments," *Applied Sciences*, vol. 10, no. 4, p. 1467, 2020.
- [27] L. Zhou, S. Kaess, and M. Kaess, "DPLVO: direct point-line monocular visual odometry," *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 7113–7120, 2021.
- [28] C. Li, W. Li, Z. Wang, and Y. Wan, "Research on image feature extraction and matching algorithms for simultaneous localization and mapping," in *Proceedings of the 2021 IEEE*

- Third International Conference on Communications, Information System and Computer Engineering, CISCE*, pp. 370–376, Beijing, China, May 2021.
- [29] H. M. Bruno and E. Colombini, “LIFT-SLAM: a deep-learning feature-based monocular visual SLAM method,” *Neuro-computing*, vol. 455, pp. 97–110, 2021.
- [30] C. Rafael Steffens, L. R. Vieira Messias, P. Lilles, J. Drews, and S. S. da Costa Botelho, “Can exposure, noise and compression affect image recognition? An assessment of the impacts on state-of-the-art ConvNets,” in *Proceedings of the 2019 Latin American Robotics Symposium (LARS), 2019 Brazilian Symposium on Robotics (SBR) and 2019 Workshop on Robotics in Education (WRE)*, pp. 61–66, Rio Grande, Brazil, October 2019.
- [31] B. Paul, R. Wang, and D. Cremers, “Online photometric calibration of auto exposure video for realtime visual odometry and SLAM,” *IEEE Robotics and Automation Letters*, vol. 3, no. 2, pp. 627–634, 2018.
- [32] J. Kim and A. Kim, “Light condition invariant visual SLAM via entropy based image fusion,” in *Proceedings of the 2017 14th International Conference on Ubiquitous Robots and Ambient Intelligence, URAI*, pp. 529–533, Jeju, Republic of Korea, July 2017.
- [33] Y. Kim, W. Chung, and D. Hong, “Indoor parking localization based on dual weighted particle filter,” *International Journal of Precision Engineering and Manufacturing*, vol. 19, no. 2, pp. 293–298, 2018.
- [34] X. Gao, R. Wang, N. Demmel, and D. Cremers, “LDSO: direct sparse odometry with loop closure,” in *Proceedings of the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS*, pp. 2198–2204, Madrid, Spain, October 2018.
- [35] L. Liu, “Image classification in htp test based on convolutional neural network model,” *Computational Intelligence and Neuroscience*, vol. 2021, Article ID 6370509, 8 pages, 2021.
- [36] J. Ni, Y. Chen, and Y. Chen, “A survey on theories and applications for self-driving cars based on deep learning methods,” *Applied Sciences-Basel*, vol. 10, no. 8, 2020.
- [37] X. Zhao, L. Liu, R. Zheng, W. Ye, and Y. Liu, “A robust stereo feature-aided semi-direct SLAM system,” *Robotics and Autonomous Systems*, vol. 132, Article ID 103597, 2020.
- [38] A. Krizhevsky, I. Hinton, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.
- [39] D. Xu, X. Wang, G. Sun, and H. Li, “Towards a novel image denoising method with edge-preserving sparse representation based on laplacian of B-spline edge-detection,” *Multimedia Tools and Applications*, vol. 76, no. 17, Article ID 17839, 2017.
- [40] H. Yin, Y. Qiu, and G. Qiu, “Side window guided filtering,” *Signal Processing*, vol. 165, pp. 315–330, 2019.
- [41] L. Xiao, C. Fan, H. Ouyang, A. F. Abate, and S. Wan, “Adaptive trapezoid region intercept histogram based Otsu method for brain MR image segmentation,” *Journal of Ambient Intelligence and Humanized Computing*, vol. 13, no. 4, pp. 2161–2176, 2022.
- [42] D. Zheng, L. Li, S. Chai et al., “A defect detection method for rail surface and fasteners based on deep convolutional neural network,” *Computational Intelligence and Neuroscience*, vol. 2021, Article ID 2565500, 15 pages, 2021.
- [43] J. Ni, K. Shen, Y. Chen, W. Cao, and S. X. Yang, “An improved deep network-based scene classification method for self-driving cars,” *IEEE Transactions on Instrumentation and Measurement*, vol. 71, 2022.
- [44] T.-Y. Lin, M. Maire, and S. Belongie, “Microsoft COCO: common objects in context,” in *Proceedings of the 13th European Conference on Computer Vision, ECCV 2014*, pp. 740–755, Zurich, Switzerland, September 2014.
- [45] R. Jiang, H. Zhou, H. Wang, and S. S. Ge, “Road-constrained geometric pose estimation for ground vehicles,” *IEEE Transactions on Automation Science and Engineering*, vol. 17, no. 2, pp. 748–760, 2020.
- [46] A. Geiger, P. Lenz, C. Urtasun, and R. Urtasun, “Vision meets robotics: the kitti dataset,” *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.
- [47] S. Yang and S. Scherer, “Monocular object and plane SLAM in structured environments,” *IEEE Robotics and Automation Letters*, vol. 4, no. 4, pp. 3145–3152, 2019.
- [48] J. Cheng, H. Zhang, and Q.-H. Meng, “Improving visual localization accuracy in dynamic environments based on dynamic region removal,” *IEEE Transactions on Automation Science and Engineering*, vol. 17, no. 3, pp. 1585–1596, 2020.
- [49] X. Zhao, L. Liu, R. Zheng, W. Ye, and Y. Liu, “A robust stereo feature-aided semi-direct SLAM system,” *Robotics and Autonomous Systems*, vol. 132, 2020.
- [50] S.-J. Ryu, M.-J. Lee, and H.-K. Lee, “Detection of copy-rotate-move forgery using zernike moments, Information Hiding,” in *Proceedings of the International workshop on information hiding*, pp. 51–65, Springer, Calgary, Canada, June 2010.
- [51] L. Calatroni and K. Papafitsoros, “Analysis and automatic parameter selection of a variational model for mixed Gaussian and salt-and-pepper noise removal,” *Inverse Problems*, vol. 35, no. 11, 2019.
- [52] G. Chahine and C. Pradalier, “Survey of monocular slam algorithms in natural environments,” in *Proceedings of the 2018 15th Conference on Computer and Robot Vision, CRV*, pp. 345–352, Toronto, Canada, May 2018.
- [53] Y. Miura and J. Miura, “RDS-SLAM: real-time dynamic SLAM using semantic segmentation methods,” *IEEE Access*, vol. 9, Article ID 23772, 2021.
- [54] C. Campos, R. Elvira, and J. J. G. Rodríguez José, M. M. Montiel and D. T. Juan, Orb-Slam3: An accurate open-source library for visual, visual-inertial, and multimap SLAM,” *IEEE Transactions on Robotics*, vol. 37, no. 6, pp. 1874–1890, 2021.
- [55] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, “A benchmark for the evaluation of rgb-d slam systems,” in *Proceedings of the International Conference on Intelligent Robot Systems (IROS)*, Vilamoura-Algarve, Portugal, October 2012.
- [56] C. Hu, B. B. Sapkota, J. Alex Thomasson, and M. V. Bagavathiannan, “Influence of image quality and light consistency on the performance of convolutional neural networks for weed mapping,” *Remote Sensing*, vol. 13, no. 11, 2021.
- [57] S. Hasan, M. Dighan, and D. L. Brian, “Utility of a commercial unmanned aerial vehicle for in-field localization of biomass bales,” *Computers and Electronics in Agriculture*, vol. 180, 2021.

## Research Article

# PointTransformer: Encoding Human Local Features for Small Target Detection

Yudi Tang<sup>1</sup>, Bing Wang<sup>1</sup>, Wangli He<sup>1</sup>, Feng Qian<sup>1</sup>, and Zhen Liu<sup>2</sup>

<sup>1</sup>Key Laboratory of Smart Manufacturing in Energy Chemical Process, Ministry of Education, East China University of Science and Technology, No. 130 Meilong Road, Shanghai 200237, China

<sup>2</sup>Sinopec Shanghai Petrochemical Co., Ltd., No. 48 Jinyi Road, Shanghai 201512, China

Correspondence should be addressed to Yudi Tang; [y10190112@mail.ecust.edu.cn](mailto:y10190112@mail.ecust.edu.cn)

Received 26 April 2022; Revised 12 July 2022; Accepted 16 July 2022; Published 21 August 2022

Academic Editor: Nian Zhang

Copyright © 2022 Yudi Tang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The improvement of small target detection and occlusion handling is the key problem to be solved in the object detection task. In the field operation of chemical plant, due to the occlusion of construction workers and the long distance of surveillance shooting, it often leads to the phenomenon of missed detection. Most of the existing work uses multiple feature fusion strategies to extract different levels of features and then aggregate them into global features, which does not utilize local features and makes it difficult to improve the performance of small target detection. To address this issue, this paper introduces Point Transformer, a transformer encoder, as the core backbone of the object detection framework that first uses a priori information of human skeletal points to obtain local features and then uses both self-attention and cross-attention mechanisms to reconstruct the local features corresponding to each key point. In addition, since the target to be detected is highly correlated with the position of human skeletal points, to further boost Point Transformer's performance, a learnable positional encoding method is proposed by us to highlight the position characteristics of each skeletal point. The proposed model is evaluated on the dataset of field operation in a chemical plant. The results are significantly better than the classical algorithms. It also outperforms state-of-the-art by 12 percent of map points in the small target detection task.

## 1. Introduction

In recent years, the application of computer vision in chemical safety has developed rapidly. In the field operation of chemical plant, the most important element is safety, which often leads to very serious consequences due to the illegal construction by workers. With the development of deep learning, using this method to solve the safety problems in the field operation of chemical plant has become popular research nowadays. In surveillance video analysis, object detection algorithms, such as YOLO [1] and SSD [2], are often used to detect and identify construction sites using a large amount of training data, which can significantly improve the on-site safety protection level, as well as providing timely warnings for detected violations. However, in the field operation of chemical plant, the application scenario is very different from the traditional object detection task, where the

equipment worn by workers needs more attention from the model because a large number of targets to be detected are highly relevant. It is difficult to solve this problem using classical object detection algorithms such as YOLO.

Many recent studies have introduced feature fusion modules [3] to improve the recognition rate of object detection algorithms in small targets and occlusion phenomena. By merging shallow local features and deep global features [4], the model can focus on both local features and global semantic information. These strategies have been widely used in object detection algorithms and have the potential to significantly improve the performance of algorithm on dataset, such as COCO [5]. In order to further improve the detection performance, many studies have introduced attention mechanism modules [6, 7] to reconstruct local features at occluded locations. By designing the attention mechanism, the model can make better use of local

features in reasoning and determining the occluded regions. However, these strategies and improvements are only for the general scene application. They do not consider the special characteristics in the field operation of chemical plant. The objects to be detected are mainly focused on the construction workers, and how to extract the local features of the construction workers is the key to improving the recognition performance of our algorithm.

As shown in Figure 1, most of the targets to be detected in our research are highly related to construction workers and show a clear dependence on the skeletal point locations of workers, e.g., helmets are always worn on the head and gloves are always worn on the hand, which can be used as a priori knowledge for the detection task. Based on this, we use the trained OpenPose [8] model to extract 25 skeletal point positions of the human body as a priori information for the local features of subsequent model reconstruction. This local feature extraction method has been used to reconstruct human local features in many ReID [9, 10] studies. For example, Wang et al. [11] used human skeletal point features to solve the partial occlusion phenomenon, and inspired by this, human skeletal point local feature extraction will also be applied to our network structure.

First of all, in the backbone design, feature extraction methods such as traditional ResNet [12] and EfficientNet [13] are not used. Although these backbones have achieved excellent results in many classical challenges, the relatively deep network also impacts the construction of local features, which makes it difficult to improve the detection performance of small targets. We chose the popular transformer [14] architecture as the core feature extraction module to address this problem. Although the attention mechanism module can be used to reconstruct each local feature area better, most of the areas in our task are background, and we want the workers themselves to be given more attention by the model. Consequently, when the transformer module was designed, the method of gated positional encoding was introduced to focus on extracting local features in the human skeletal point region. Compared to the classic transformer architecture, we designed the module to focus on reconstructing features in the human skeletal points while downplaying irrelevant features such as the background.

Although the traditional self-attention [15] approach can reconstruct each part of the features by weight calculation when the attention mechanism module is designed, it is difficult to capture the interrelationship between the local features. In LoFTR [16], self-attention is used to reconstruct local features, and cross-attention is used to highlight the relationship between different key points. Inspired by this, the cross-attention method is also introduced to highlight the relational properties of different skeletal point regions when the human skeletal point region features are reconstructed. When construction workers work together, the tools and equipment they use are nearly the same, and the cross-attention approach also allows the characteristics of the construction scene and the collaborative work to be learned by the model.

Since transformer architectures are inherently insensitive to position information, it is often necessary to introduce positional encoding [17] features. Transformer architectures often use a fixed positional encoding method to highlight the characteristics of different regions, but these methods only give a unique identifier to each local feature region and do not have learning capabilities. In our study, most of the targets to be detected show obvious positional relationships. For example, the helmet must be at the top of the protective goggles. Based on this, a learnable positional encoding method is proposed by us. On one hand, the importance of position information is highlighted so that the model can better learn the position relationship between different objects. On the other hand, due to the different importance of human skeletal point features at different positions, for example, a large number of targets to be detected are concentrated on the hands and head, and a few on the human torso. Therefore, it is also possible to differentiate depending on the position of the target to be detected.

Based on our knowledge, there is currently no research on applying human skeletal points as local features in object detection algorithms for the field operation chemical plant scene. To solve the problem of small target detection and covering in this scene, we propose a novel end-to-end object detection framework with a transformer as the core backbone for feature extraction, and an improved attention mechanism is designed to highlight the relationship between local features. Additionally, since location information is particularly important in our research scenario, a learnable positional encoding method is also introduced to highlight the location relationship properties between the targets to be detected. The main contributions of our research are summarized as follows:

- (i) A new type of end-to-end object detection backbone is proposed that optimizes the local feature extraction through the features of human skeletal points while designing and improving the attention module to improve the model's detection performance.
- (ii) Multiple attentional mechanisms are proposed to reconstruct the local features of human skeletal points and their interdependence information by using self-attention and cross-attention, respectively.
- (iii) In the transformer structure, a learnable positional encoding method is proposed to optimize the feature reconstruction of each local skeletal point by utilizing a weighting mechanism of the local features.

## 2. Related Works

*2.1. Object Detection Models.* This section introduces some fundamental concepts in the field of object detection and then elaborates and illustrates several popular attention mechanism modules and positional encoding methods.



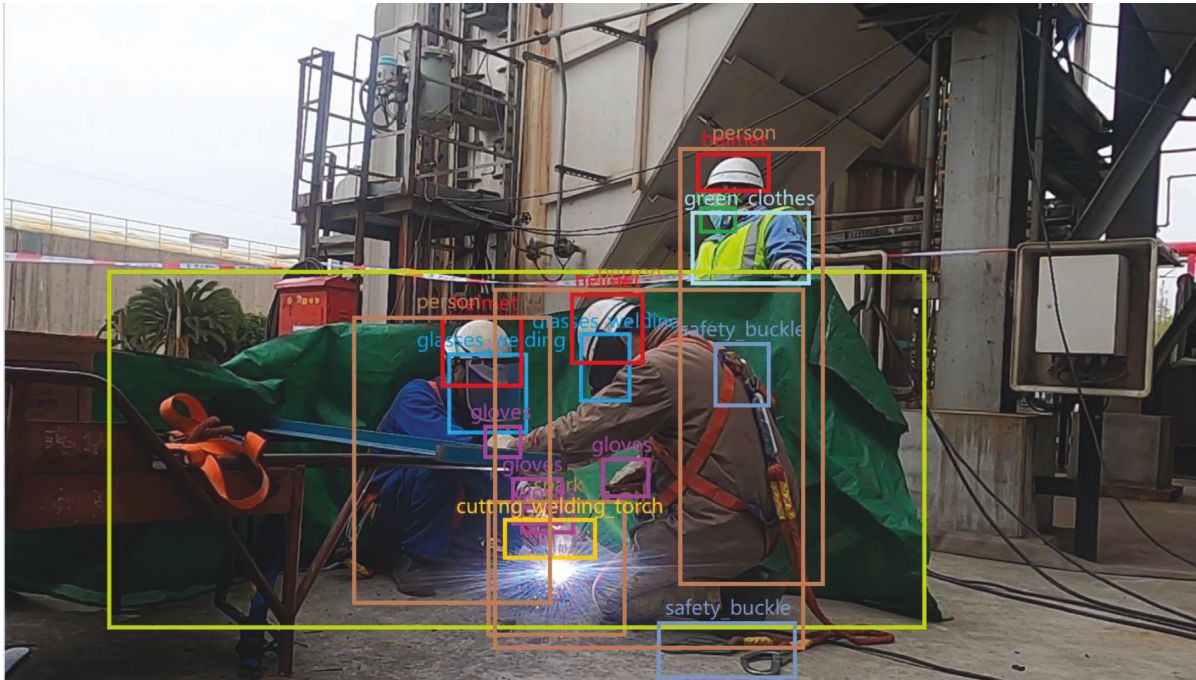


FIGURE 1: An example of labeling data, it can be seen from the figure that a large number of targets to be detected are highly correlated with construction personnel.

**2.2. Back Bone Design.** The object detection algorithm is composed of four primary modules that were developed during the process. (1) Operations for data augmentation and preprocessing. (2) Design of backbone in feature extraction module. (3) Feature fusion module. (4) Output module. The data augmentation module is primarily used to increase the amount of training data and enhance the model's generalizability. The backbone is generally trained by classical classification models, such as those obtained by using ResNet on the ImageNet dataset, and the feature fusion module is mainly used to increase the diversity of features, such as the SPP [18] layer in YOLOV5. The output layer mainly uses the learned features to get the prediction results.

Almost all object detection algorithms perform data augmentation [19, 20] operations on the training data to expand the amount of data. For example, by applying CutMix operations to the data, the data is rotated and scaled in EfficientDet [21], and the overall Map is significantly improved. In YOLOV5 [19] Mosaic Data Augmentation operations are also used to increase the amount of training data. Utilizing data augmentation can significantly improve the model's generalizability and minimize the risk of overfitting.

Numerous object detection algorithms choose DenseNet [22, 23] and EfficientNet [13] as core feature extraction modules for backbone design. On one hand, these models perform well across all classical datasets. On the other hand, due to the abundance of pre-trained models, different pre-trained models can be selected based on the difference of application scenarios. However, because these models require more convolutional layers to achieve a larger receptive field, they tend to focus on global features and ignore some local features. As a result, the traditional backbone design method is better suited to large target

detection tasks and will be significantly less effective at detecting small targets. Given that the majority of the targets in our task are related to construction workers and fall under the category of small target detection tasks, we will design and implement a new end-to-end network structure for the backbone selection.

**2.3. Transformer Encoder.** Both feature concatenation and fusion methods are widely used in the design of feature fusion modules, for example, the FPN [21, 24] method is used in mask R-CNN [25] to extract multiple layers of features, and SPP [18] is used in YOLOV5 to obtain richer features. The advantage of these methods is that they allow for the simultaneous use of deep and shallow features, which improves the model's detection performance of targets of various sizes. However, because these methods do not take into account the application scenario and do not select the appropriate features based on the characteristics of the detected target, we chose the transformer architecture, which is better suited for local feature reconstruction. The transformer architecture has demonstrated excellent performance in a variety of computer vision tasks, including object detection in DETR [26] using the transformer's encoder and decoder, and as the backbone of the Swin transformer [27, 28] in detection and classification scenarios, significantly improving model performance. The attention mechanism module is at the heart of the transformer architecture, as it recombines the features of each region by calculating the weight relationship between each local feature. The advantage of this approach is that the model can pay more attention to local features and also learn more about the relational properties between regions.

*2.4. Positional Encoding Method.* Due to the insensitivity of the transformer architecture to position information, additional position feature is typically added to highlight local features. BERT [29] uses a fixed positional encoding method to emphasize contextual information, while ViT [15] uses absolute positional encoding method to improve the model’s classification performance. Typically, the obtained positional features must be fused with local features, which introduces position information into each local feature. However, because positional features are fixed, they cannot be updated for learning purposes, limiting their usefulness. In this task, we will enhance the positional encoding method in order to highlight the positional characteristics of various local features.

### 3. Methodology

Our proposed framework is illustrated in Figure 2 and consists of the following points: a transformer encoder-based feature extraction backbone (A), which is used to extract features from the input image, mainly involving two attention mechanism modules, self-attention and cross-attention; a gated position encoding computation module (G), which is used to highlight the positional characteristics between different skeletal points; and a Head-Attention module (H), which uses the positional characteristics of skeletal points to enhance the detection effect of the output layer.

#### 3.1. Revisiting Transformers and Small Target Detection Task.

Classical object detection algorithms use ResNet, EfficientNet, etc. as the backbone [30] of feature extraction, which uses numerous layers of convolution in order to obtain a larger receptive field and then extracts features at different levels for the obtained different levels of feature maps. Although global features can be extracted at different scales by using operations such as FPN [24], the backbone with convolutional layers as the core is still insensitive to local features and it is difficult to obtain the relational properties between different regions.

The advantage of designing a backbone based on self-attention is that it can extract features for each local location; in turn, the problem of insensitivity to small targets in convolution is improved. The traditional transformer architecture first divides the input image according to a given region, for example,  $16 * 16$  as the base unit for local feature reconstruction in ViT [15]. In order to use the transformer module to process the input image ( $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$ ), we reshape it and get the sequence input ( $\mathbf{x}_l \in \mathbb{R}^{N \times (L^2 \cdot C)}$ ),  $N$  represents the length of the sequence, and  $L$  represents the size of each token. But the problem of doing this is that if the input image is large and the selected window is small, it makes the computation inefficient, but if a larger window is set, it is difficult to mention the fine-grained extraction of local features, which has become one of the main problems of transformer framework nowadays. It is difficult to handle more fine-grained local feature extraction due to the limitation of computational magnitude. In our task, some large

targets such as fire extinguishers and scaffolds are easily detected, which makes our main research problem focus on the construction workers’ bodies, and these objects to be detected are usually highly correlated with human skeletal point locations. Based on this, the human skeletal point information will be used to highlight the characteristics of local features, thus allowing the model to focus more on small targets in the human body.

In the field operation chemical plant scene, if the application scenario involves only a single construction job, there is usually not much occlusion, which also makes the detection task relatively easy. However, in our task, it is almost always a multi-person collaborative work scenario of multi-people collaborative construction, which makes many local features easily obscured from each other. To solve the problem, inspired by LoFTR [16], both self-attention and cross-attention feature extraction methods are chosen to be applied to local feature computation, which can reconstruct local features by self-attention on one hand and extract relational properties between skeletal points by cross-attention on the other hand. For self-attention layers, the input features are key points at different locations of the same person. For cross-attention layers, the input features are key points that differ from person to person. All attention mechanism calculation methods are calculated by

$$\text{Attention}(Q, K, V) = \text{softmax}(QK^T)V, \quad (1)$$

where  $Q$  indexes query vectors,  $K$  indexes key vectors, and  $V$  is the value vectors.

*3.2. Local Feature Extraction.* We first obtain the skeletal point positions of all construction workers by a trained pose estimation model [31, 32], 25 keypoints are obtained, all consisting of 2D coordinates. The feature map obtained after the backbone is expanded into a sequence for subsequent calculation of the attention mechanism module. The traditional transformer calculates self-attention on the entire sequence. However, in our task, we need to pay more attention to local features, that is, the regions corresponding to the key points. Based on this, we map the key points to the expanded sequence (downsampling ratio consistent with the backbone), which corresponds to part of the token in the corresponding sequence. In the calculation of self-attention, except for the tokens where the key points are located, the weights calculated from other positions are truncated, and the maximum is not over 0.05. The reason for this is that we do not want the model to consider too many background features. In the calculation of cross-attention, we design a mask mechanism, only the tokens corresponding to the key points will be updated. After the attention mechanism, we reshape the entire sequence to get its feature map (consistent with the size of the feature map in the last layer of the backbone).

A set of learnable weight parameters is designed by us to weight each local feature corresponding to each skeletal point. The reason for this is that most of the small targets to be detected in our dataset are concentrated on the hands and head, while the large targets to be detected are mainly on the

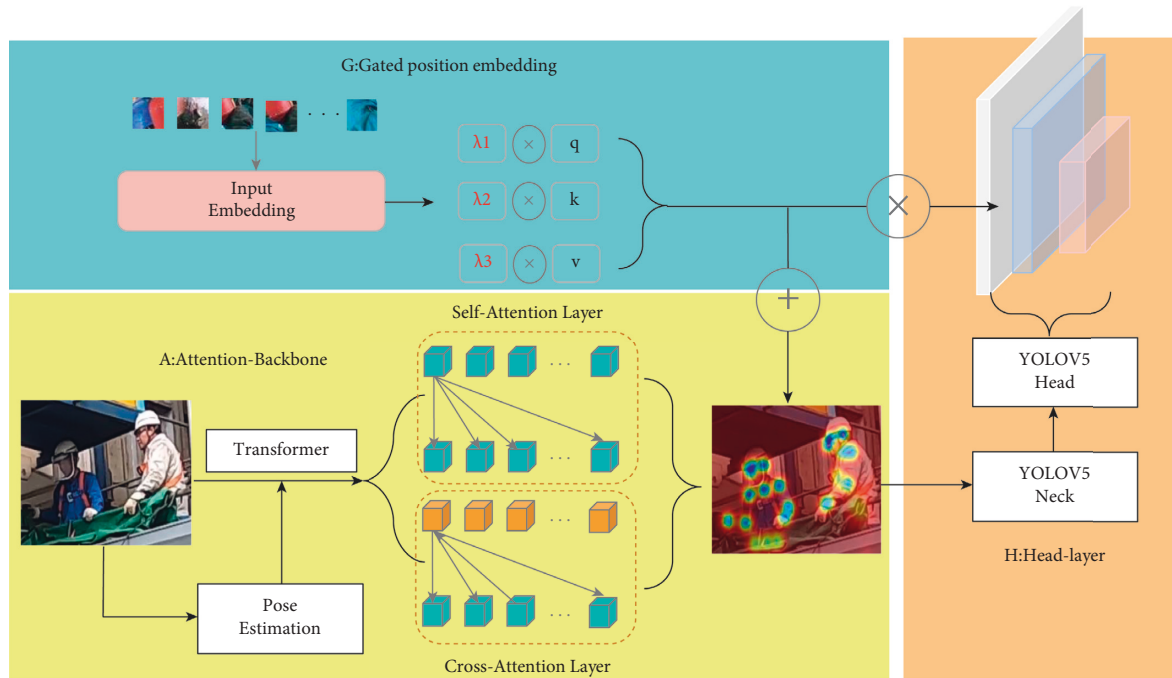


FIGURE 2: The overall architecture of our proposed model. Attention backbone (A) utilize a trained pose estimation model to reconstruct local features based on transformer encoder. Gated position embedding module (G) uses human skeletal point location information to enhance local feature learning. Head-layer module (H) reconstructs the output layer by weighting the positional encoding feature maps.

torso. In order to improve the detection of small targets, we want the model to focus more on the hand and head locations and slightly on the body and leg locations. Based on this, we designed an additional set of learnable gated parameters to combine the local features and weighted the features at the location of skeletal points before calculating attention with other local features.

In the process of calculating self-attention, the difference with the original ViT method is that we weight the features at the locations of skeletal points and the weight parameters are learnable, which has the advantage of making the model more focused on the areas where small target objects exist, which is the core of our research problem. We do not use the same or random weights for the initialization of all skeletal points, but rather give larger weights for the hands and head, initialized to 10, and smaller weights for the body and leg key points, initialized to 2. For the location of other non-human skeletal points, it is consistent with the traditional transformer architecture. The self-attention method based on gated parameters allows the model to utilize more prior knowledge and focus on local feature extraction of the human body.

When constructing local features, it is difficult to highlight the location relationship between skeletal points if only the self-attention method is used; for example, the helmet is always located above the glove location, and if one worker in the current construction scene is wearing gloves and helmet, all other workers should also be required to wear gloves and helmet. In chemical scenes, usually all workers in an area wear the same work equipment, but due to the small target and the existence of obscuration and other problems, there are frequently some missed tests phenomenon. In

order to make full use of the positional features between objects, we additionally add the cross-attention module to optimize local feature extraction. As shown in Figure 3, for each skeletal point region of the construction worker, the attention between it and other construction workers' skeletal points is calculated in the same way as the traditional self-attention, and the superimposed features are averaged if there are multiple people in the figure. In the experiment, we will discuss the effects brought by these two attention mechanism modules separately.

**3.3. The Prominent Role of Positional Encoding.** The advantage of using a transformer as a backbone for feature extraction is that it has strong reconstruction ability for local features, but such methods as self-attention are insensitive to positional information which only gives a unique identifier to each region and does not have an actual feature representation. In the field operation chemical plant scene, positional information is particularly important, e.g., tools are always held in the hands and safety buckles are always tied on the body, and there are obvious location characteristics between these objects and human skeletal points. Based on this, we introduced an additional learnable positional encoding method to highlight the importance of local features when designing the transformer architecture. This module is only for self-attention calculation and initializes the position encoding of the sequence expanded by the output feature map of the backbone. Different from the initialization method in ViT, the position encoding we designed is learnable, not a fixed parameter. In addition, it is not only related to its absolute position, but also needs to

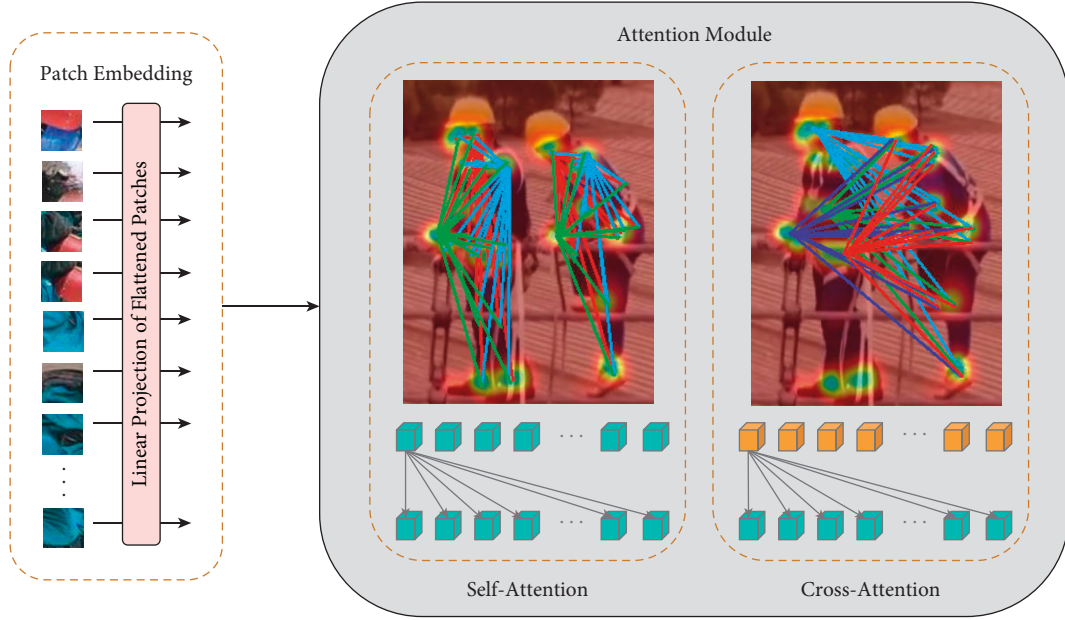


FIGURE 3: For each local feature, both self-attention and cross-attention mechanisms are used to reconstruct the local features of the human body.

consider the characteristics of  $K$  and  $V$  corresponding to its token. Inspired by Wang et al. [17], for the positional encoding features as shown in Figure 4, we learn positional information for  $Q$ ,  $K$ , and  $V$ , respectively, and its learned positional features are weighted together with the reconstructed features computed in self-attention, and the positional encoding method is computed in the same way as in Axial-DeepLab [17].

$$y = \sum_{p \in N_{max}(o)} \text{softmax}_p(q_o^T k_p + q_o^T r_{p-o}^q + k_p^T r_{p-o}^k)(v_p + r_{p-o}^v), \quad (2)$$

where  $r_{p-o}^k \in \mathbb{R}^{d_q}$  represents the learnable positional encoding for  $K$ , and  $r_{p-o}^v \in \mathbb{R}^{d_{out}}$  is the same for  $V$ .

In Medical Transformer [33], the positional encoding method is initialized randomly because the features at each location do not have a priori knowledge, but in our task, it is obvious that the location of human skeletal points has a more important role. Based on this, we do not choose a random approach when initializing the positional features, but perform a Gaussian initialization centered on each key point, which will result in a larger weight value for the region where the skeletal points are located and a smaller weight value for the other locations, which also matches the distribution of the objects to be detected in our task. Since the targets to be detected are highly concentrated in the hands and heads, we also give larger weight values when initializing their positional feature and the rest of the skeletal point locations are initialized with the same Gaussian initialization method.

**3.4. Improvement of the Output Layer.** Since the position encoding feature is very sensitive to the features corresponding to the keypoints, we use a fully connected layer to

its probability map to weight the output layer. In the object detection task, multiple anchor sizes and multiple output layers are usually designed to make the model adaptable to different size targets. Though the network structure is designed to focus on the attention method and emphasize the importance of positional information, the local features corresponding to the skeletal points cannot be well utilized if the output layer is still chosen similar to the YOLOV5 head-layer, which only predicts the features at different levels separately, so we perform an additional weighting calculation for the output features. As presented in 3.3, the learnable positional encoding features are multiplied with each output layer feature in YOLOV5 as shown in Figure 5. This enables more attention to be paid to the human skeletal point area; thus, improving the detection performance.

## 4. Experiments

In this chapter, we evaluate our proposed method in the field operation chemical plant scene. We set up several sets of ablation experiments and analyze them in comparison with the corresponding performance of the mainstream object detection algorithms presently. We will present the experimental setup and results in the following sections.

**4.1. Datasets Description.** The data we selected came from the scene of field operation chemical plant, and because the surveillance video was blurred, so we chose to shoot the construction site in person. All videos were shot with 1080p explosion-proof equipment. To make the data more diverse, we chose different angles and distances for the same construction scene. All video data were cut into images at 100 frames intervals to build a dataset and annotated it, and all data were manually annotated using the LabelMe toolkit. The overall dataset consists

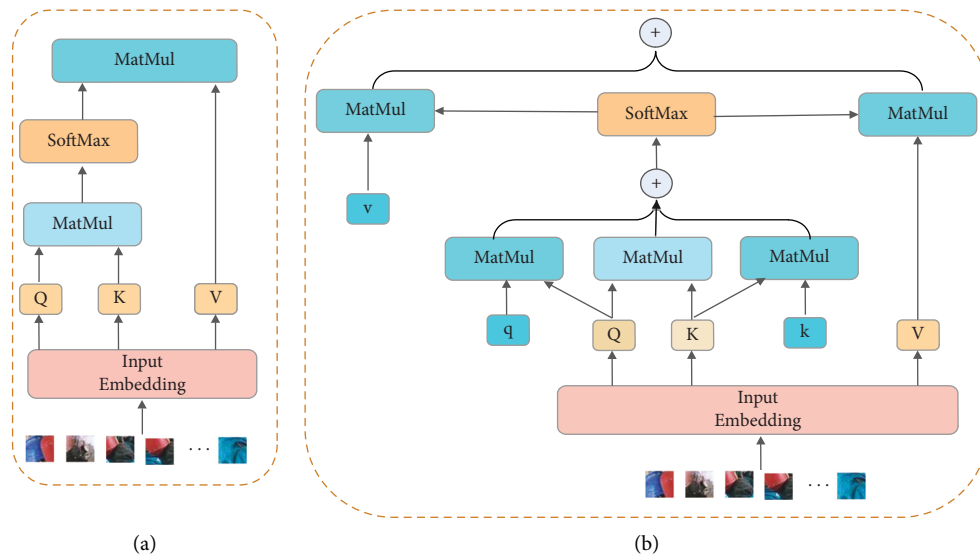


FIGURE 4: (a) The left figure shows the self-attention structure of the traditional transformer. (b) The right figure shows our proposed gated position-attention, which incorporates the influence of position information on the reconstructed features.

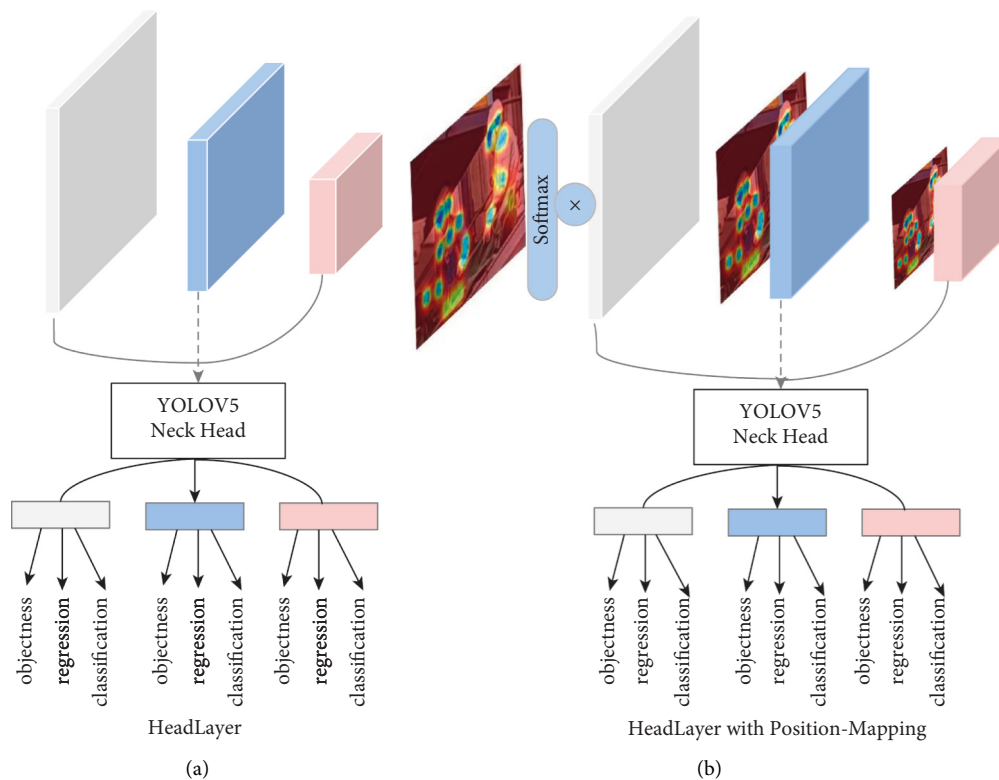


FIGURE 5: (a) The original YOLOV5 head-layer is shown on the left. (b) Our proposed head-layer with positional features mapping is shown on the right. The dependence of the model on local features is further enhanced by mapping the positional features to the output layer.

of 2400 annotated images, but since our research focuses on the presence of small targets and occlusions, some easy-to-detect data samples were excluded from the original dataset. The dataset was randomly split where the training set consisted of 1281 images and the test set consisted of 200 images. For the selection of labels, there are 19 categories of labeled objects in total, including helmets, goggles, gloves, construction equipment, fire extinguishers, and signs. However, since objects such

as fire extinguishers and scaffolds are usually easier to be detected, we only kept the objects related to construction workers in the labels, and the total number of labels is 13.

**4.2. Influence of Different Attention Methods.** In this part, we conduct a comparative experimental analysis of different design approaches for the attention mechanism module. In

the next experiments, we choose the evaluation method consistent with the COCO benchmark [5]. Firstly, all human skeletal point locations are obtained using the OpenPose model. Next in the backbone selection, we conducted the following experiments, respectively: (1) directly using the native YOLOV5 model; (2) using only the basic transformer encoder as the feature extraction backbone; (3) using both self-attention and cross-attention as the feature extraction backbone. In the above experiments, all attention mechanism modules do not use positional encoding features. Table 1 shows that if only the traditional object detection algorithm is used, it is difficult to get better performance in the dataset with mostly small targets, and when using transformer as the backbone, although the detection effect can be slightly improved on small target objects, the overall map value is not significantly improved. When both self-attention and cross-attention modules are used, the detection performance is not only improved by 5.6 percentage points on small targets, but also the overall map value is improved by 3.1 percentage points. This can be attributed to the fact that by using multiple attention mechanism strategies allows the model to learn richer and more reliable local features, and by cross-attention also allows the model to learn the relational properties between different local features.

*4.3. Influence of Position Encoding Method.* Since the targets to be detected shows an obvious dependence on the location of human skeletal points, we designed a learnable positional encoding method. In the next experiments, we will analyze the performance of different positional encoding methods on the results which are used with self-attention and cross-attention as the base backbone. First, we used the traditional transformer positional encoding method that is consistent with ViT, and Table 2 shows that the introduction of positional encoding method can increase the overall Ap value of the model by up to 1.3 percentage points, which has a significant impact on the detection performance. Furthermore, the traditional positional encoding method was replaced with a learnable gated positional encoding method, and the gated values (weights) were initialized to 0.05 for  $Q$ ,  $K$ , and  $V$ . Since the method of positional encoding is randomly initialized at the beginning of training, which may lead to instability occurred during model training, on the basis of that, a smaller initial value was chosen for this parameter. From Table 2, it can be concluded that the use of our proposed gated positional encoding method can improve the overall detection performance by 2.4 percentage points. On the effect of detection for small target objects, the improvement is 1.9 percentage points relative to the traditional positional encoding method. This can be attributed to the fact that, for the detection task of the equipment worn by the construction personnel, a large number of targets to be detected show obvious characteristics of positional relationships, and by training the learnable positional encoding method, the model can better learn the positional dependencies between different objects.

TABLE 1: Influence of different backbone design on feature encoding.

Backbone design	AP	AP50	AP75	APs	APm	API
YOLOV5-X	31.4	40.1	33.2	17.3	38.7	53.5
Transformer encoder	31.6	40.4	33.7	18.6	39.6	49.8
Self-attention and cross-attention	34.5	42.1	37.1	24.2	42.1	45.6

TABLE 2: Ablation studies of positional encoding method.

Positional-encoding method	AP	AP50	AP75	APs	APm	API
Without positional encoding	34.5	42.1	37.1	24.2	42.1	45.6
2D positional encoding	35.8	43.6	38.9	25.7	42.2	45.1
Learnable gated positional encoding	36.9	45.1	40.3	27.6	44.1	44.4

*4.4. Influence of the Output Layer.* In this part, we will compare and analyze the effect of the output layer of the object detection algorithm on the results. Three output layers are selected in YOLOV5 for regression and classification tasks after concat features from different layers, respectively, using different receptive field features for the prediction of different size targets. In our design, positional-encoded weight mapping is additionally introduced on top of it to further highlight the degree of influence of different skeletal point locations on the results in the output layer. From Table 3, it can be seen that the output layer with the positional-encoded weight mapping improves the detection of small targets by 1.7 percentage points. This can be attributed to the fact that, although multiple attention and location encoding strategies are used in the backbone module, some features and information are lost if not emphasized in the output stage. The combination of positional-encoded weight mapping with the output layer can significantly improve the detection performance of our model for small target detection tasks.

*4.5. Comparison with the SOTA Model.* To highlight the effectiveness of our proposed method, in this chapter, we will compare and analyze our method with the current SOTA object detection algorithms. In order to analyze the effectiveness of the transformer module in the field operation of chemical plant, we conduct an experimental comparison with EfficientDet and FCOS. Since the detection performance of EfficientDet is directly related to the levels of EfficientNet, we select EfficientNet-B0 and EfficientNet-B3 as backbones to observe their effect on small target detection task. In order to prove the importance of local features in the small target detection task, we choose to compare with the Transformer-based Deformable DETR method and select ResNet50 as its backbone. In the data preprocessing stage, all models use the same data augmentation strategy, and for the fairness of the experiment, all models do not use multi-scale input, and all input sizes are  $640 * 640$ . Due to the instability

TABLE 3: Influence of positional-feature-mapping on head-layer.

Head-layer type	AP	AP50	AP75	APs	APm	API
YOLOV5-X head-layer	36.9	45.1	40.3	27.6	44.1	44.4
Positional-feature-mapping head-layer	37.2	45.5	40.7	29.3	42.1	43.2

TABLE 4: Comparison with the SOTA model.

Method	AP	AP50	AP75	APs	APm	API
FCOS	28.1	39.5	31.3	16.5	35.7	48.1
EfficientNet-B0-based EfficientDet	27.9	38.6	30.8	16.8	34.5	46.6
EfficientNet-B3-based EfficientDet	30.8	40.8	33.6	17.0	36.2	49.1
Transformer-based Deformable DETR	33.5	42.7	36.1	19.4	41.5	51.7
YOLOV5-X	31.4	40.1	33.2	17.3	38.7	53.5
Our proposed model	37.2	45.5	40.7	29.3	42.1	43.2

of the label balancing method during training, we did not use this method for all models. From Table 4, it can be seen that although the EfficientDet model has a good performance in large target detection, it cannot effectively identify small targets. In addition, when the levels of backbone layers increased, the small target detection performance is not improved. Although the transformer is used as the entire encoder and decoder modules in Deformable DETR, there are still problems in the small target detection task in the field operation of chemical plant. Even if two-stage training is performed on Deformable DETR, it is difficult to improve its small target detection performance. Through the above comparison experiments, it can be found that for the small target detection problem in the field operation of chemical plant, not only the feature relationship between regions needs to be considered in the selection of network structure, but also the local feature extraction module is required to strengthen the model’s local perception ability.

**4.6. Training Details.** The overall training process of the model is consistent with YOLOV5, using ADAM [34] as the optimizer and choosing a moment value of 0.9, an initial learning rate of 0.01, and a learning rate decay and early stop strategy. All network structures are used in YOLOV5-X structure except backbone design. All experiments are based on the same evaluation criteria used in the COCO dataset after 300 epochs of iterations of RTX3090.

## 5. Conclusions

In the field operation of chemical plant, there are often small target detection tasks and construction workers obscure each other. How to perform local feature extraction becomes the key to improve the detection performance. To solve this problem, we propose the point transformer, which first uses self-attention and cross-attention for local feature reconstruction of human skeletal points. In addition, since the target to be detected in our task is highly correlated with the location of the skeletal points of the construction workers, we designed a learnable positional encoding method to highlight the importance of location information in order to make better use of this priori information. It is shown in

experiments on the scene of field operation chemical plant dataset that the proposed point transformer outperforms present-day classical object detection algorithms. Our approach can be seen as an application of optimizing the performance of small target detection tasks using local features of the human body. However, this has not yet been exploited due to the obvious synergistic relationship between the movement changes of the skeletal points during the construction work, which exhibits graph structural properties. Our future work will aim at using the graph model to construct local features of the human body to further improve the detection performance [35].

## Data Availability

The data set was taken on site during the construction of Sinopec Shanghai Petrochemical. Due to its confidentiality policy, this dataset cannot be made public.

## Conflicts of Interest

On behalf of all authors, the corresponding author states that there are no conflicts of interest. The authors declare that they have no conflicts of interest.

## Acknowledgments

This work was supported by National Key Research and Development Program of China under Grant 2018AAA0101602 and National Natural Science Foundation of China (61922030).

## References

- [1] J. Redmon and A. Farhadi, “Yolov3: an incremental improvement,” arXiv:1804.02767, 2018.
- [2] W. Liu, D. Anguelov, D. Erhan et al., “Ssd: single shot multibox detector,” in *Proceedings of the European Conference on Computer Vision*, pp. 21–37, Springer, Amsterdam, The Netherlands, October 2016.
- [3] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature pyramid networks for object detection,” in *Proceedings of the IEEE Conference on Computer Vision and*

- Pattern Recognition*, pp. 2117–2125, Honolulu, HI, USA, July 2017.
- [4] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal loss for dense object detection,” in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2980–2988, Honolulu, HI, USA, July 2017.
  - [5] T.-Y. Lin, M. Maire, S. Belongie et al., “Microsoft coco: common objects in context,” in *Proceedings of the European conference on Computer Vision*, pp. 740–755, Springer, Zurich, Switzerland, September 2014.
  - [6] P. Shaw, J. Uszkoreit, and A. Vaswani, “Self-attention with relative position representations,” arXiv:1803.02155, 2018.
  - [7] H. Zhao, J. Jia, and V. Koltun, “Exploring self-attention for image recognition,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10076–10085, Seattle, WA, USA, June 2020.
  - [8] Z. Cao, T. Simon, and S. E. Wei, “Realtime multi-person 2d pose estimation using part affinity fields,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7291–7299, Honolulu, HI, USA, July 2017.
  - [9] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. H. Hoi, “Deep learning for person re-identification: a survey and outlook,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 6, pp. 2872–2893, 2022.
  - [10] W. Li, X. Zhu, and S. Gong, “Harmonious attention network for person re-identification,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2285–2294, Salt Lake City, UT, USA, June 2018.
  - [11] G. Wang, S. Yang, and H. Liu, “High-order information matters: learning relation and topology for occluded person re-identification,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6449–6458, Seattle, WA, USA, June 2020.
  - [12] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, Las Vegas, NV, USA, June 2016.
  - [13] M. Tan and Q. Le, “Efficientnet: rethinking model scaling for convolutional neural networks,” in *Proceedings of the International Conference on Machine Learning*, pp. 6105–6114, PMLR, Long Beach, United States, 2019.
  - [14] A. Vaswani, N. Shazeer, N. Parmar et al., “Attention is all you need,” in *Advances in Neural Information Processing Systems*, pp. 5998–6008, Morgan Kaufmann Publishers Inc, San Francisco; CA, USA, 2017.
  - [15] A. Dosovitskiy, L. Beyer, and A. Kolesnikov, “An image is worth 16x16 words: transformers for image recognition at scale,” arXiv:2010.11929, 2020.
  - [16] J. Sun, Z. Shen, and Y. Wang, “LoFTR: detector-free local feature matching with transformers,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8922–8931, Nashville, TN, USA, June 2021.
  - [17] H. Wang, Y. Zhu, and B. Green, “Axial-deeplab: stand-alone axial-attention for panoptic segmentation,” in *Proceedings of the European Conference on Computer Vision*, pp. 108–126, Springer, Cham, 2020.
  - [18] K. He, X. Zhang, S. Ren, and J. Sun, “Spatial pyramid pooling in deep convolutional networks for visual recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.
  - [19] C. Shorten and T. M. Khoshgoftaar, “A survey on image data augmentation for deep learning,” *Journal of big data*, vol. 6, no. 1, pp. 60–48, 2019.
  - [20] A. Mikołajczyk and M. Grochowski, “Data augmentation for improving deep learning in image classification problem,” in *Proceedings of the 2018 international interdisciplinary PhD workshop (IIPhDW)*, pp. 117–122, IEEE, Poland, Europe, May 2018.
  - [21] M. Tan, R. Pang, and Q. V. Le, “Efficientdet: scalable and efficient object detection,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10781–810790, Seattle, WA, USA, June 2020.
  - [22] G. Huang, Z. Liu, and L. Van Der Maaten, “Densely connected convolutional networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4700–4708, Honolulu, HI, USA, July 2017.
  - [23] C. Y. Wang, H. Y. M. Liao, and Y. H. Wu, “CSPNet: a new backbone that can enhance learning capability of CNN,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 390–391, Seattle, WA, USA, June 2020.
  - [24] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, “Path aggregation network for instance segmentation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8759–8768, Salt Lake City, UT, USA, June 2018.
  - [25] K. He, G. Gkioxari, and P. Dollár, “Mask r-cnn,” in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2961–2969, Honolulu, HI, USA, July 2017.
  - [26] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, “End-to-end object detection with transformers,” in *Proceedings of the European Conference on Computer Vision*, pp. 213–229, Springer, Cham, 2020.
  - [27] Z. Liu, Y. Lin, Y. Cao et al., “Swin transformer: hierarchical vision transformer using shifted windows,” arXiv:2103.14030, 2021.
  - [28] K. Han, A. Xiao, E. Wu, J. Guo, C. Xu, and Y. Wang, “Transformer in transformer,” arXiv:2103.00112, 2021.
  - [29] J. Devlin, M. W. Chang, and K. Lee, “Bert: pre-training of deep bidirectional transformers for language understanding,” arXiv:1810.04805, 2018.
  - [30] W. Wang, E. Xie, and X. Li, “Pyramid vision transformer: a versatile backbone for dense prediction without convolutions,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 568–578, Montreal, BC, Canada, October 2021.
  - [31] K. Sun, B. Xiao, and D. Liu, “Deep high-resolution representation learning for human pose estimation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5693–5703, Long Beach, CA, USA, June 2019.
  - [32] R. A. Güler, N. Neverova, and I. Kokkinos, “Densepose: dense human pose estimation in the wild,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7297–7306, Salt Lake City, UT, USA, June 2018.
  - [33] J. M. J. Valanarasu, P. Oza, and I. Hacihaliloglu, “Medical transformer: gated axial-attention for medical image segmentation,” in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 36–46, Springer, Montreal, BC, Canada, October 2021.
  - [34] D. P. Kingma and J. Ba, “Adam: a method for stochastic optimization,” arXiv:1412.6980, 2014.
  - [35] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, “Yolov4: optimal speed and accuracy of object detection,” arXiv:2004.10934, 2020.



## Research Article

# SR-DSFF and FENet-ReID: A Two-Stage Approach for Cross Resolution Person Re-Identification

Zongzong Wu,<sup>1</sup> Xiangchun Yu ,<sup>1</sup> Donglin Zhu ,<sup>1</sup> Qingwei Pang,<sup>1</sup> Shitao Shen,<sup>1</sup> Teng Ma,<sup>2</sup> and Jian Zheng <sup>1</sup>

<sup>1</sup>Jiangxi University of Science and Technology, Ganzhou, Jiangxi 341000, China

<sup>2</sup>Xi'an Zhongtie Rail Transit Co., Ltd., Xian, Shaanxi 710000, China

Correspondence should be addressed to Jian Zheng; [gzzj\\_yanjiu@163.com](mailto:gzzj_yanjiu@163.com)

Received 28 March 2022; Accepted 18 May 2022; Published 5 July 2022

Academic Editor: Nian Zhang

Copyright © 2022 Zongzong Wu et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

In real-life scenarios, the accuracy of person re-identification (Re-ID) is subject to the limitation of camera hardware conditions and the change of image resolution caused by factors such as camera focusing errors. People call this problem cross-resolution person Re-ID. In this paper, we improve the recognition accuracy of cross-resolution person Re-ID by enhancing the image enhancement network and feature extraction network. Specifically, we treat cross-resolution person Re-ID as a two-stage task: the first stage is the image enhancement stage, and we propose a Super-Resolution Dual-Stream Feature Fusion sub-network, named SR-DSFF, which contains SR module and DSFF module. The SR-DSFF utilizes the SR module to recover the resolution of the low-resolution (LR) images and then obtains the feature maps of the LR images and super-resolution (SR) images, respectively, through the dual-stream feature fusion with learned weights extracts and fuses feature maps from LR and SR images in the DSFF module. At the end of SR-DSFF, we set a transposed convolution to visualize the feature maps into images. The second stage is the feature acquisition stage. We design a global-local feature extraction network guided by human pose estimation, named FENet-ReID. The FENet-ReID obtains the final features through multistage feature extraction and multiscale feature fusion for the Re-ID task. The two stages complement each other, making the final pedestrian feature representation have the advantage of accurate identification compared with other methods. Experimental results show that our method improves significantly compared with some state-of-the-art methods.

## 1. Introduction

The purpose of person Re-ID is to match the target person of interest across the images under multiple cameras. Due to its wide range of applications, such as intelligent surveillance, person tracking, and criminal case forensics, it has been widely used in computer vision in recent years. With the development of deep learning, many deep feature extraction networks have been designed for person Re-ID to improve the matching accuracy. However, in practical applications, person Re-ID still presents enormous challenges due to factors such as different low-resolution images [1], illumination changes [2], occlusions [3], and weather changes [4].

Some deep learning based person Re-ID methods [5, 6] perform well on the premise of ensuring that the resolutions

of gallery images and query images are consistent and high-resolution (HR). However, this premise is usually not guaranteed because the resolution of the query images is usually low, but the gallery images are all filtered HR images, which resulted in a mismatch between the resolution of the query images and gallery images. At this time, traditional person Re-ID methods cannot extract discriminative person features for target matching, so more and more people begin to focus on cross-resolution person Re-ID [7–12]. Cross-resolution person Re-ID works aim to address the problem of resolution mismatch between query images and gallery images.

Cross-resolution person Re-ID was first proposed by Li et al. [13] in 2015, which opened a precedent for cross-resolution person Re-ID research. Subsequent research on

cross-resolution person Re-ID can be divided into two stages in time. In some early works, dictionary learning or metric learning are used to learn pedestrians between images of different resolutions. The common feature representations are as shown in work [7–10]. However, the feature maps extracted by these methods based on LR images are unreliable, so the early cross-resolution person Re-ID matching accuracy is not satisfactory. Subsequently, with the proposal of some SR models [14–17], some researchers began to apply SR models to cross-resolution person Re-ID, which is the development of cross-resolution person Re-ID second stage. Jiao et al. [12] were the first to use SRCNN [18] to recover the resolution of LR images and proposed a method to train the SR sub-network and the Re-ID sub-network jointly. Since then, more and more works have begun introducing SR modules into cross-resolution person Re-ID, which further improves the matching accuracy of cross-resolution person Re-ID. For example, MMSR [19] designed a mixed-space super-resolution model to recover the resolution of LR images with variable resolution. Recently, many new methods represented by PRI [11] have improved the detection accuracy of cross-resolution person Re-ID to a new level. However, there are still some gaps in practical application.

Through the study of numerous cross-resolution person Re-ID methods in recent years, we found some of their disadvantages. Most of the current research ideas are to use the SR module to recover the query images resolution to the high-resolution displayed by the gallery images. The use of the SR modules significantly improves the matching accuracy of cross-resolution person Re-ID, but in fact, we found through experiments that the SR images generated after the SR modules will inevitably lose some original details [20]. We believe that this will bring hidden dangers to subsequent Re-ID tasks. Although Zhuang et al. [21] proposed CAD-NET to jointly learn the feature maps of the SR images and the LR images to alleviate the loss of feature details; however, there are still significant problems in directly fusing feature maps from images of different resolutions. Furthermore, most researchers use deep neural networks to capture low-level details by extracting local features [22] of images, which are likely to bring semantic ambiguity. For example, a man with a woman’s suitcase is mistaken for a woman. Therefore, we believe that it is necessary to devise better methods in reducing the loss of original details brought by the SR module and extracting image feature extraction.

In this paper, we propose a person Re-ID method that jointly optimizes the feature details of person images and the extraction of features. Specifically, we propose a deep network consisting of the SR-DSFF sub-network and the FENet-ReID sub-network. Firstly, the SR-DSFF uses a dynamic upscale module to learn the weights in the convolution kernel; these weights are then used to generate SR images. Different from other methods that utilize SR models, we treat SR-DSFF as an image enhancement model rather than a single SR model. Therefore, we added the DSFF module after the SR module. The DSFF module clearly distinguishes high- and low-resolution inputs during feature learning, so that the feature information of different

resolution images complement each other to ensure its robustness to resolution variance. Subsequently, the global-local feature extraction network (FENet-ReID) extracts person representations for person Re-ID. The FENet-ReID consists of two convolution stages (FE-C1 and FE-C2) and three feature fusion units. The two convolution stages consist of four CNN sub-networks, and each feature fusion unit sequentially fuses two equal-sized feature maps to obtain a more discriminative final feature representation of a person. The main contributions of this paper are as follows:

- (i) We propose an image enhancement sub-network named SR-DSFF. Unlike other methods, we do not rely on a single SR module to recover image resolution. Instead, the DSFF module is added after the SR module to reduce the loss of image details.
- (ii) We propose a feature extraction network based on human pose estimation named FENet-ReID, using the final features from multistage feature extraction and multiscale feature fusion to perform cross-resolution person Re-ID.
- (iii) We have done a lot of experiments on three cross-resolution person Re-ID datasets, all of which have reached the industry-leading level. Compared with other state-of-the-art methods, our proposed method achieves 2.7%, 5.4%, and 3.7% improvement on Rank-1 on MLR-Market1501, MLR-CUHK03, and CAVIAR datasets, respectively.

The rest of this article is organized as follows: Section 2 introduces the related work and Section 3 mainly introduces the proposed method. Section 4 evaluates the model’s performance through extensive experiments and concludes with a conclusion in Section 5.

## 2. Related Work

*2.1. Person Re-ID.* Person Re-ID has been studied by academia for many years since 2005. Still, it was not until 2014 that deep learning began to be applied to person Re-ID, that person Re-ID achieved a huge breakthrough. Many current methods [23–27] have achieved outstanding results in closed-world [28], and even some state-of-the-art methods have achieved accuracy close to or surpassing the human level. For example, Zheng et al. [29] proposed a method that combines the similarity of intraclass data in high-dimensional space and the difference between classes and achieves complementary effects by fusing the two loss functions. And Chen et al. [30] proposed a method on transfer learning in unsupervised situations, for the two models with the same network to fill the unlabeled part of each other and it can be further replaced by two different networks. As a result, traditional person Re-ID has entered a bottleneck period and many methods have been developed to deal with various challenges, such as pedestrians with different poses, different styles of cameras, and occlusion. For example, Wei et al. [31] proposed a GLAD that exploits both local and global features of the human body to generate a representation with strong discriminativeness to handle significant variations in human poses. Liu et al. [32]

proposed a method for uniform style, that is, to deal with the style changes caused by different cameras by generating images with a unified camera style through Unity GAN. Qian et al. [33] proposed a generative adversarial network (PNGAN) specially designed for pose normalization in Re-ID. Some methods [34–36] use human pose information to reduce background noise to solve the problem of occlusion. However, the above methods usually assume that the resolutions of query images and gallery images are similar and high enough, which will bring significant problems to applying these methods in open-world [28].

**2.2. Cross-Resolution Person Re-ID.** In order to solve the problem that the resolution span is too large, many methods have also been proposed in recent years. Traditional methods [37, 38] process images employing metric learning or dictionary learning, but the details of LR images are not obvious, so the performance of these methods is limited. With the development of super-resolution technology, some SR-based cross-resolution person Re-ID methods have been proposed in later studies. SR-based cross-resolution person Re-ID usually relies on SR modules to recover the resolution of LR images. Since Ledig et al. [16] first proposed SRGAN, SR modules have been widely used in the resolution recovery stage of cross-resolution person Re-ID. And Jiao et al. [12] jointly trained SRCNN and Re-ID networks for the first time. Mao et al. [39] proposed a Foreground-Focus Super-Resolution (FFSR) module and Resolution-Invariant Feature Extractor (RIFE). Unlike other SR-based methods, FFSR combines Re-ID loss and foreground attention loss during training and suppresses irrelevant background while restoring pedestrian image resolution. Some other SR models are also widely used in cross-resolution person Re-ID, such as Meta-SR [17] and VDSR [40].

**2.3. Feature Representation Learning in Person Re-ID.** In the field of person Re-ID, most deep learning based works [41–43] are used to extract feature maps from the entire pedestrian images, so simply extracting global features is likely to lose key information about pedestrians. Subsequent works [44–46] tried to horizontally divide pedestrian images into several fixed-length blocks to extract more detailed local features. The experimental results show that the matching accuracy of person Re-ID after adding local features is much better than those methods that use global features. However, dividing the pedestrian image into fixed-length blocks to extract local features is not sensitive to the change of the pedestrian’s posture. The pedestrians captured by the surveillance cameras often have posture changes. Therefore, it is necessary to design a better feature extractor for pedestrian pose changes.

**2.4. Discussion.** Cross-resolution person Re-ID is only a branch of the field of person Re-ID. There are still many issues to be resolved. For example, to make the person Re-ID technology applicable on a large scale, we need to design a lighter network while ensuring the accuracy so that the hardware device can accept it. In addition, in the research of

person Re-ID, I found that some techniques can also be applied to building retrieval [47] or drone-based geo-localization [48] etc.

At present, most SR-based cross-resolution person Re-ID methods focus on the reconstruction of SR images so the reconstructed SR images can be closer to the original HR images. However, these methods ignore the loss of image detail in the reconstruction process and the distribution difference between high- and low-resolution image features. Different from current methods, our network learns and fuses features from LR and HR images through dual-streams of attention-weighted feature extraction while recovering the image resolution. Compared with the way current methods deal with LR images, our method preserves richer image feature details.

### 3. Proposed Methods

Our network structure diagram is shown in Figure 1. This section introduces the SR-DSFF and FENet-ReID, respectively.

**3.1. SR-DSFF.** As the first stage of cross-resolution person Re-ID, we first consider restoring the resolution of the query images. For open-world, we often face the problem that the resolution span of query images is too large, so we cannot predict a suitable scale factor to handle query images of arbitrary resolutions. For open-world needs and improving cross-resolution person Re-ID methods, it is crucial to design a method that can handle query images at arbitrary resolutions. Inspired by some work [49], we employ a dynamic Meta-Upscale module to learn the weights in the convolution kernels, which are then used to generate SR images. Our SR module is different from some existing SR models such as FSRCNN [14], SRDenseNet [15], and SRGAN [16]. Inspired by meta-learning [50], we divide the SR module into two modules, the feature learning module and the Meta-Upscale module [17]. We choose RDN [51] as the feature learning module, and it is worth noting that we replace the ordinary upscale module with an improved Meta-Upscale module.

SR-DSFF takes a set of LR images  $I_L^n = \{I_L^1, I_L^2, I_L^3, \dots, I_L^N\}$  as input. In the training phase, we obtain images  $I_L^n$  from a set of original HR images  $I_H^n = \{I_H^1, I_H^2, I_H^3, \dots, I_H^N\}$  by down-sampling. In the SR module, our goal is to predict the SR images  $I_S^n = \{I_S^1, I_S^2, I_S^3, \dots, I_S^N\}$  from images  $I_L^n$ . Assuming that the scale factor of each pixel  $(i_1, j_1)$  of the images  $I_L^n$  is  $s$  during the enlargement process, in the prediction stage, the features  $F_{IL}^n$  of the images  $I_L^n$  is extracted by the feature learning module in the SR module. The features of the images  $I_L^n$  on its pixel  $(i_1, j_1)$  and the corresponding filter weights determine each pixel  $(i, j)$  in the generated images  $I_S^n$ .

For each pixel  $(i, j)$  in the images  $I_S^n$ , it is determined by the feature of the images  $I_L^n$  on its pixel  $(i_1, j_1)$  and the corresponding filter weights. So we can think of the Meta-Upscale module as a mapping function from  $F_{IL}$  to  $I_S^n$ . The mapping function is as follows:

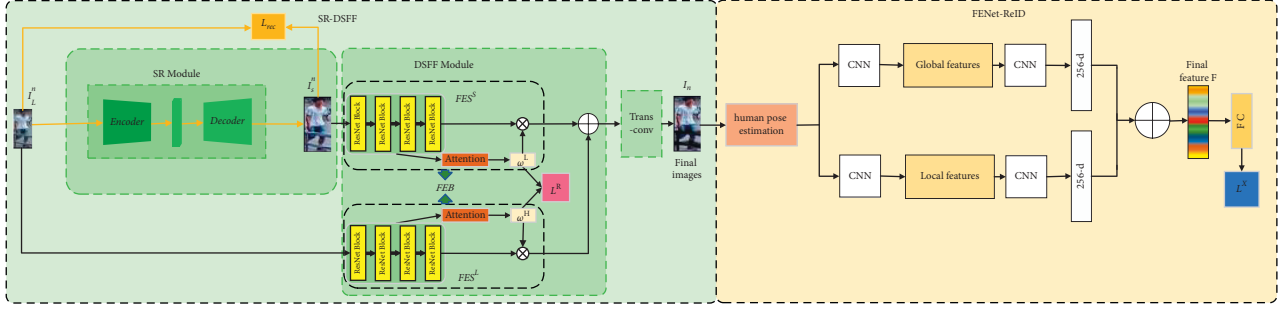


FIGURE 1: The network consists of SR-DSFF sub-network and FENet-ReID. The query images first enter the SR-DSFF, and the SR images are output through the feature extractor and the upscale module in the SR module. Then, the feature maps of the query images and the SR images are jointly learned and fused through the DSFF module, and the final images are output into FENet-ReID through transposed convolution. FENet-ReID extracts the global and local features of the images that are obtained in the SR-DSFF and fuses them to obtain the final feature maps. Finally, a fully connected (FC) layer is used on the final feature maps to predict the ID labels of pedestrians. Our network is divided into two training stages: (1) Update the SR module with the SR loss  $\mathcal{L}_{\text{rec}}$  (equation (6)); and (2) jointly train the DSFF and the FENet-ReID with the total loss  $\mathcal{L}_{\text{TOTAL}}$  (equation (12)). These two stages are represented by yellow and black arrows on the figure, respectively.

$$I_S^n(i, j) = f(F_{IL}^n(i_1, j_1), w(i, j)), \quad (1)$$

where  $I_S^n(i, j)$  is the pixel value of the images  $I_S^n$  at  $(i, j)$ ,  $f(\cdot)$  represents the feature mapping function for calculating pixel values, and  $w(i, j)$  is the weight prediction module of the pixel point  $(i, j)$  (corresponding to equation (3)).

For each pixel  $(i, j)$  in the images  $I_S^n$ , we consider the pixel  $(i, j)$  to be determined by the features of  $(i_1, j_1)$  on the LR images. We map these two pixels through a projection transformation function  $T$ :

$$(i_1, j_1) = T(i, j) = \left(\frac{i}{s}, \frac{j}{s}\right). \quad (2)$$

Specifically, we can think of the resolution recovery process as a variable fractional stride mechanism that enables convolution to use an arbitrary scale factor  $s$  (not limited to integer multiples of scale factors) to upscale feature maps. For example, when the scale factor  $s = 2$ , one pixel  $(i_1, j_1)$  determines two pixels on the images  $I_S^n$ . If the scale factor is a non-integer, taking  $s = 1.5$  as an example, some pixels determine two pixels, and some pixels determine one pixel. All in all, each pixel  $(i, j)$  on the images  $I_S^n$  can find a most relevant pixel  $(i_1, j_1)$  on the images  $I_L^n$ .

After determining the positional relationship between the images  $I_L^n$  and the images  $I_S^n$ , it is also necessary to learn the weights and offset between the two. Different from the traditional upscale module, our Meta-Upscale module predicts the corresponding number of filter weights for any scale factors employing two fully connected layers. In order to train multiple scale factors simultaneously, it is better to add the scale factors to  $v_{ij}$  to distinguish the weights of different scale factors. We can express the weight prediction and  $v_{ij}$  as follows:

$$W(i, j) = \varphi(v_{ij}; \theta), \quad (3)$$

$$v_{ij} = \left(\frac{i}{s} - \frac{i}{s}, \frac{j}{s} - \frac{j}{s}\right), \quad (4)$$

where  $W(i, j)$  is the convolution kernel weight corresponding to the pixel  $(i, j)$  on the images  $I_S^n$ ,  $v_{ij}$  is the vector associated with  $(i, j)$ ,  $\varphi$  is the weight prediction network, and  $\theta$  is the weight of the weight prediction network. Then obtain the pixel value of the pixel  $(i_1, j_1)$ . Its feature mapping function is expressed as follows:

$$\Phi(F_{IL}(i_1, j_1), W(i, j)) = F_{IL}(i_1, j_1)W(i, j). \quad (5)$$

Finally, in order to ensure that the images  $I_S^n$  have high-resolution, we define a SR loss  $\mathcal{L}_{\text{rec}}$  between the SR images and its original HR images, and the SR loss  $\mathcal{L}_{\text{rec}}$  is expressed as follows:

$$\mathcal{L}_{\text{rec}} = \mathbb{E}[\|I_S^n - I_H^n\|], \quad (6)$$

where  $I_H^n$  and  $I_S^n$  represent original HR images and SR images, respectively. As shown in Figure 2, the effect of the SR module on the resolution recovery of LR images is pronounced.

It is worth noting that although we use the SR loss  $\mathcal{L}_{\text{rec}}$  to make the images  $I_S^n$  to reduce the loss of pedestrian features during the resolution recovery process. However, in the process of resolution recovery, the loss of features is still inevitable. In addition, the visual cues contained between different resolution images are different, so it is not reliable to rely on the SR images for the Re-ID task. To sum up, we added a DSFF module after the SR module to learn the features in different resolution images  $I_L^n$  and  $I_S^n$  and fuse the learned feature maps. Since SR images and LR images contain other visual cues, different feature extractors should be used to extract image feature maps of different resolution images.

The DSFF module consists of two feature extraction branches. We denote these two branches named  $FES^L$  and  $FES^S$ , respectively. In each branch, we take ResNet101 [52] as the backbone, and ResNet101 is modified to be a Feature Extraction Block named FEB to extract the feature maps of the input images by duplicating its convolutional layers as  $FES^L$  and  $FES^S$ . And we introduce spatial attention and channel attention in  $FES^L$  and  $FES^S$ . As shown in Figure 1,



FIGURE 2: Shows the performance of our SR module on the dataset Market1501. The effect is evident by comparing it with LR images.

there is always a feature extraction branch in FEB corresponding to the images  $I_L^n$  and images  $I_S^n$ , respectively. Among them,  $FES^L$  and  $FES^S$  have the same structure. However, the training purposes of the two branches are different. For example, for the SR images, we choose a more appropriate  $FES^S$  for feature extraction, so the  $m^S$  is fused with larger weights. As shown in Figure 3, in the spatial attention, we utilize softmax to transform the learned feature vectors into weight  $\omega^{L1}$  or  $\omega^{S1}$ , in the channel attention, we use one global average pooling (GAP) layer and two fully connected (FC) layers to predict  $\omega^{L2}$  or  $\omega^{S2}$ , and the feature maps  $m^L$  and  $m^S$  obtained by each branch can be expressed as follows:

$$m^L = \omega^{L1} \times m^{L1} + \omega^{L2} \times m^{L2}, \quad (7)$$

$$m^S = \omega^{S1} \times m^{S1} + \omega^{S2} \times m^{S2}, \quad (8)$$

where  $m^L$  and  $m^S$  represent feature maps obtained by  $FES^L$  and  $FES^S$ , respectively.  $m^{L1}$  and  $m^{L2}$  represent the feature maps obtained by  $FES^L$  through spatial attention and channel attention, respectively.  $m^{S1}$  and  $m^{S2}$  represent the feature maps obtained by  $FES^S$  through spatial attention and channel attention, respectively. The resolution of the input images determines the size of  $\omega^L$  and  $\omega^S$ . For LR images,  $\omega^L$  will be larger than  $\omega^S$ , and vice versa. In order to learn  $\omega^L$  and  $\omega^S$ , we introduce resolution weighting loss  $\mathcal{L}^R$ . According to the training images  $I_L^n$  and  $I_S^n$  can be expressed as follows:

$$\mathcal{L}^R(I_r^n) = \|\omega^L - (1-r)\|_2^2 + \|\omega^S - r\|_2^2, \quad (9)$$

where  $\omega^L = (\omega^{L1}, \omega^{L2})$ ,  $\omega^S = (\omega^{S1}, \omega^{S2})$ ,  $I_r^n$  represents  $I_L^n$  or  $I_S^n$ , and  $r$  represents the resolution of  $I_r^n$ . Finally, we denote the output feature  $m$  as follows:

$$\mathbf{m} = m^L + m^S. \quad (10)$$

Finally, the feature  $\mathbf{m}$  is put into the last transposed convolutional layer of the SR-DSFF to get the final image with richer semantic information.

**3.2. FENet-ReID.** After obtaining the final images, our ultimate goal is to obtain a discriminative pedestrian feature representation for the Re-ID task. To get this feature map as shown in Figure 1, we extract global and local features from the final images and fuse them.

We utilize human pose estimation [53]. Unlike Spindie Net [54], we only select four key points on pedestrians to make our model robust to a wider variety of pedestrian poses and camera views. Based on these four key points, we get three key regions of pedestrians: the head, upper body, and lower body.

Our FENet-ReID process consists of two modules, the Feature Extraction Module (FEM) and the Feature Fusion Module (FFM). The FEM and FFM are introduced separately below.

**3.2.1. FEM.** We design a Convolutional Neural Networks (CNNs) consisting of four sub-networks in FEM. As shown in Figure 4, the FEM consists of two convolution stages (FE-C1 and FE-C2). Using the FEM, we obtain four 256-dimensional feature vectors from the pedestrian image global and three key regions. In FE-C1, there are three convolutional layers and one Inception module [55] in each CNN. First, convolve the input image to obtain a feature map with a spatial size of  $24 \times 24$ . At the same time, the same operation is performed on the three key regions of pedestrian and a ROI Pooling operation is performed to keep the feature maps obtained by FE-C1 of equal size. In FE-C2, the four feature maps obtained in the previous stage are input, and the spatial size is reduced to  $12 \times 12$  through an initial module first, then we use a global pooling layer and a fully connected layer to convert into 256-dimensional feature vectors, that is, the output of FE-C2 is four 256-dimensional feature vectors, which correspond to the global image and three human key regions images, respectively.

**3.2.2. FFM.** To make the final feature representation of pedestrians more discriminative, next we fuse together the four 256-dimensional feature vectors obtained earlier to generate a compact 256-dimensional feature vector. We adopt a feature fusion unit to fuse two feature vectors of equal size. Specifically, as shown in the right part of Figure 4, we use three such feature fusion units, where two primary operations are performed in each feature fusion unit: (1) Use the element-wise maximization operation to delete the features of the smaller value, and only keep the features of the maximum value. (2) An inner product layer is used for feature transformation, and its output can be used for subsequent feature fusion units. The three feature fusion units from left to right sequentially fuse the pedestrian's lower body and upper body into the main body, fuse the main body and head into the whole body, and finally fuse the whole body and feature vector of the full image into the final 256-dimensional feature  $F$ . Finally, we use a fully connected layer on the feature  $F$  to predict the ID labels of pedestrians. It can be expressed as person Re-ID loss by a cross-entropy loss  $\mathcal{L}^X$ , and the expression is as follows:

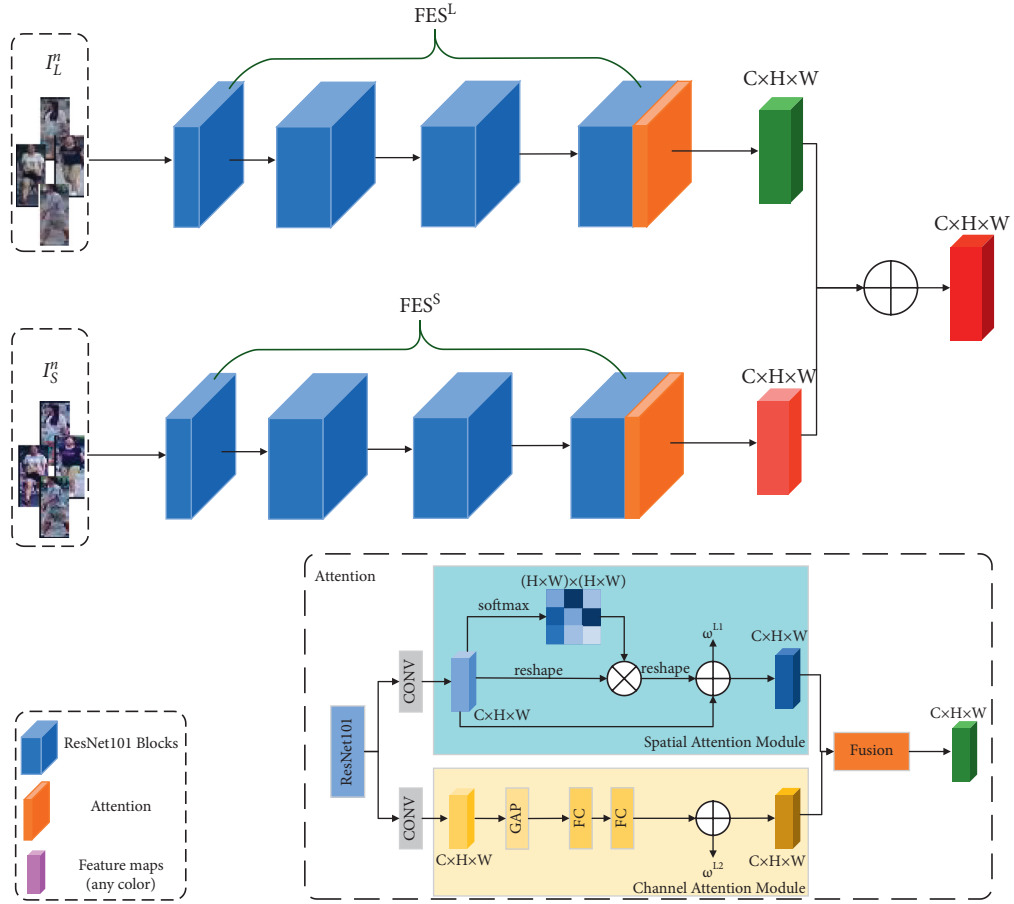


FIGURE 3: We add spatial attention and channel attention to the last ResNet101 Block. The lower right corner of the figure takes the branch  $FES^L$  as an example to give a detailed attention diagram, which is in  $FES^S$  has the same structure.

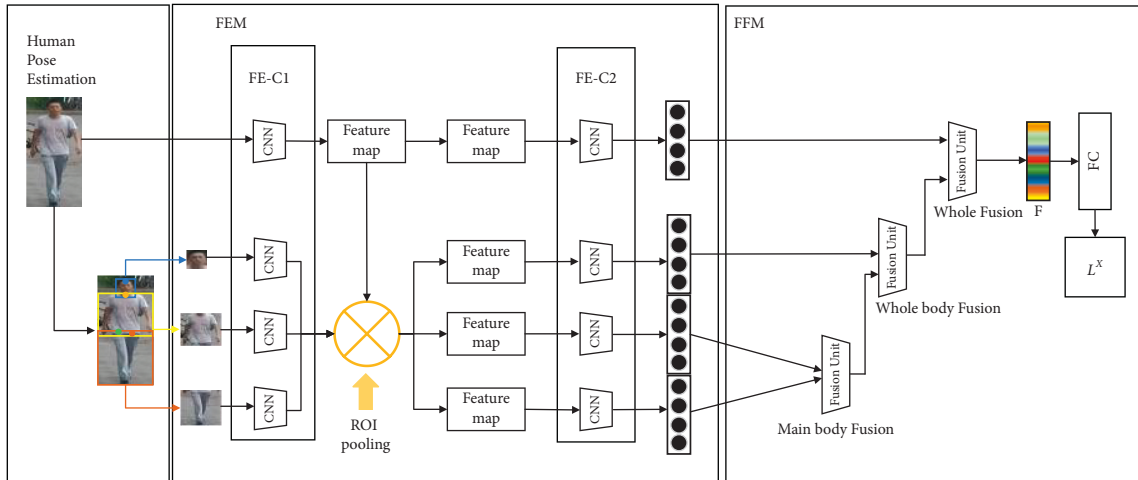


FIGURE 4: Flowchart of FENet-ReID. The full image and three human key regions images are extracted by FE-C1 and FE-C2, respectively, and the obtained 256-dimensional features are fused by three fusion units in FFM.

$$\mathcal{L}^X(I_n) = \text{CrossEntropy}(FC(F_n), P_n), \quad (11)$$

where  $I_n$  represents the final images obtained by transposed convolution in SR-DSFF and  $P_n$  represents the person ID labels of the training images  $I_n$ .

Through a training set  $Z = \{(I_H^n, I_r^n, I_n, P_n)\}, n = 1, \dots, N$ , where  $I_H^n$  represents the original HR images,  $I_r^n$  represents  $I_L^n$  or  $I_S^n$ , and  $P_n$  is the person ID label. The total loss  $\mathcal{L}_{\text{TOTAL}}$  of the DSFF module and FENet-ReID can be expressed as follows:

TABLE 1: The proposed method is compared with the current state-of-the-art methods on the dataset MLR-Market1501.

Methods	MLR-Market1501		
	Rank-1	Rank-5	Rank-10
SING [12]	74.4	87.8	91.6
SPreID [60]	77.4	89	93.9
CamStyle [61]	74.5	88.5	92.2
CAD-net [62]	83.7	92.7	95.8
FFSR + RIFE [39]	66.9	84.7	-
CRGAN [63]	83.7	92.7	95.8
INTACT [22]	<b>88.1</b>	<b>95</b>	96.9
PRI [11]	84.9	93.5	96.1
LA-transformer [64]	86.7	96.4	<b>97.4</b>
Ours	<b>90.9</b>	<b>96.4</b>	<b>97.6</b>

The best and second-best results are in bold and italics, respectively.

$$\mathcal{L}_{\text{TOTAL}} = \mathcal{L}^X(I_n) + \alpha \sum_{t=1:4} \mathcal{L}_t^R(I_r^n), \quad (12)$$

where  $I_r^n$  means  $I_L^n$  or  $I_S^n$  input into FES<sup>L</sup> or FES<sup>S</sup>.

## 4. Experiment

*4.1. Dataset.* We evaluate our method on three datasets, all of which are most commonly used for person Re-ID.

MLR-Market1501 [56]: Market1501 dataset was captured by six cameras, five of which were high-resolution cameras, and one was low-resolution. Market1501 contains 1501 different pedestrian categories with 32668 detected pedestrian bounding boxes. Among them, pedestrians of each category are captured by at least two cameras. We follow SING [12] that the images captured by one of the cameras are processed at the same down-sampling rate and the resolutions of the images captured by the other cameras remain unchanged to create the MLR-Market1501. Based on the person ID labels, we split the dataset into a training set containing 751 pedestrians and a test set containing 750 pedestrians.

CAVIAR [57] was collected in the real world, including 1220 images of 72 pedestrians captured by two cameras. According to [12], we discarded 22 identities of pedestrian images so that only HR images are included in the dataset. We randomly split the dataset into two training and test sets containing 25 pedestrian identities.

MLR-CUHK03: CUHK03 [58] is the first large-scale person Re-ID dataset, and its colossal data volume is enough to support it for deep learning. The dataset contains 632 different pedestrian categories and is photographed by five pairs of cameras. Also, according to [12], we randomly down-sample the images captured by one of the cameras of each team with the down-sampling rate of  $r \in \{2, 3, 4\}$  to create the MLR-CUHK03 dataset. We use the same number of pedestrian identities (316/316) as training/testing sets.

*4.2. Implementation Details.* Our model training is divided into two steps: (1) Firstly train the SR module separately and (2) then jointly train the DSFF and FENet-ReID.

TABLE 2: The proposed method is compared with the current state-of-the-art methods on the dataset CAVIAR.

Methods	CAVIAR		
	Rank-1	Rank-5	Rank-10
SING [12]	33.5	72.7	89
SPreID [60]	36.2	71.9	88.7
CamStyle [61]	32.1	72.3	85.9
CAD-net [62]	42.8	76.2	91.5
FFSR + RIFE [39]	36.4	72	—
CRGAN [63]	42.8	76.2	91.5
INTACT [22]	<b>44</b>	<b>81.8</b>	<b>93.9</b>
PRI [11]	43.2	78.5	91.9
LA-transformer [64]	42.1	80.7	92.4
Ours	<b>47.9</b>	<b>84.6</b>	<b>96.2</b>

The best and second-best results are in bold and italics, respectively.

In the SR module, the widely used loss function is  $L2$  loss, but according to work [59], we use  $L1$  loss to make the network better convergence. In the network training, in order to construct the LR image training set, we conduct the down-sampling operation on the images in several data sets and then adjust the image obtained by down-sampling to the same size as the original HR images. It is worth noting that we use a unified down-sampling factor  $r = 4$  to down-sample original HR images. For each batch, we randomly selected 16 LR images of  $96 * 96$  size as training images. We use Adam as the optimizer. During the training process, the training scale factor of the SR module varies from 1 to 4 with a step of 0.1, and these scale factors are uniformly distributed. Initialize the learning rate of all layers to  $10^{-4}$  and perform  $10^6$  update iterations.

DSFF and FENet-ReID are trained by Stochastic Gradient Descent (SGD), and the training is done in two steps: (1) Use  $\mathcal{L}^R$  to initialize on the target dataset and adjust the DSFF module. (2) Under the guidance of the loss function in equation (12), the DSFF and FEF are jointly trained. According to the experiment, we fix the hyper parameter in equation (12) as  $\alpha = 1$ , and each step has 60 epochs, the batch size is set to 32. The initial learning rate is set to  $10^{-2}$  in the first 30 epochs, and  $10^{-3}$  after 30 epochs. The final 256-dimensional feature is used for Re-ID with Euclidean distance.

Our network is trained on Pytorch, and all experiments are implemented with NVIDIA RTX3080Ti GPU, Intel i9 CPU, and 64 GB memory.

*4.3. Comparison with State-of-the-Art.* Tables 1–3 shows the results of our method on three datasets, as well as the comparison with other state-of-the-art methods in the last three years. The methods we choose cover two broad categories: (1) Traditional person re-id methods: SpreID [60], CamStyle [61], LA-Transformer [64]; (2) Advanced methods for cross-resolution person Re-ID (other methods in Tables 1–3). It can be seen from the comparison results that the performance of our method has improved significantly.

On the MLR-Market1501 dataset, the Rank-1 accuracy of our method improves by 2.7% over the current state-of-the-

TABLE 3: The proposed method is compared with the current state-of-the-art methods on the dataset MLR-CUHK03.

Methods	MLR-CUHK03		
	Rank-1	Rank-5	Rank-10
SING [12]	67.7	90.7	94.7
SPreID [60]	76.5	92.5	98.3
CamStyle [61]	69.1	89.6	93.9
CAD-net [62]	82.1	<b>97.4</b>	<b>98.8</b>
FFSR + RIFE [39]	73.3	92.6	—
INTACT [22]	<b>86.4</b>	<b>97.4</b>	<b>98.5</b>
PRI [11]	85.2	97.5	<b>98.8</b>
LA-transformer [64]	<b>86.3</b>	97.1	<b>98.6</b>
Ours	<b>91.8</b>	<b>97.5</b>	<b>99.2</b>

The best and second-best results are in bold and italics, respectively.

TABLE 4: Performance of different feature extractors on MLR-Market1501.

Structure	Weight learning	Rank-1	Rank-5
ResNet101	—	76.9	82.4
Two ResNet101	—	80.4	90.9
Two ResNet101	√	86.6	95.7
SR-DSFF (ours)	√	89.2	95.9

TABLE 5: The influence of different loss functions on recognition accuracy.

Loss functions	Rank-1	Rank-5	Dataset
Circle loss	88.4	95.7	MLR-Market1501
Triplet loss	88.7	94.9	MLR-Market1501
Sphere loss	89.3	96.1	MLR-Market1501
Ours	90.9	96.4	MLR-Market1501

art methods. On the MLR-CUHK03 dataset, compared with other methods, the accuracy is improved by 5.4% relative to second place in Rank-1. On the CAVIAR dataset, our Rank-1 accuracy is also 3.9% better than the current state-of-the-art. It can be seen that SR-DSFF and FENet-ReID outperform the vast majority of existing methods compared with existing cross-resolution person re-id methods. Only on dataset MLR-Market1501 and MLR-CUHK03, our method is on par with LA-Transformer [64] and PRI [11] in Rank-5 accuracy comparison.

#### 4.4. Ablation Study

**4.4.1. Validity of DSFF and FENet-ReID.** To verify the effectiveness of our SR-DSFF and FENet-ReID, as shown in Table 4, we fixed the DSFF as ResNet101 and compared it with other different SR models as shown in Table 5. It is worth noting that we use the entire SR-DSFF as an image enhancement model, because the purpose of our SR-DSFF is to obtain images that are more suitable for person Re-ID. Experiments are performed on the dataset MLR-Market1501.

In Table 4, we fix the DSFF as bilinear interpolation and compare it with three feature extractors, namely, (1) ResNet101 baseline, (2) two ResNet101 with the same

TABLE 6: Performance of different feature extractors on MLR-Market1501.

Models	DS	Weight	Rank-1	Rank-5
CycleGAN [65]	—	—	62.6	76.2
SING [12]	—	—	74.4	87.8
CSR-GAN [66]	—	√	74.3	87.7
FFSR + RIFE [39]	√	√	66.9	84.7
CAD-NET [21]	—	—	83.7	92.7
SR-DSFF (ours)	√	—	86.1	92.6
SR-DSFF (ours)	√	√	90.3	96.4

“DS” represents whether dual-stream feature fusion is performed and “Weight” indicates whether weighting loss was added during feature extraction.

weights, and (3) two ResNet101, and dual-stream feature fusion with the learned weights learned by equation (9). From Table 4, we can see that using two ResNet101 improves the model significantly. After further assigning weights to the two ResNet101, the effect also enhances. Finally, our SR-DSFF shows the best results with dual-stream feature fusion and learned weights. The accuracy of Rank-1 is enhanced by 12.3% compared to the baseline.

In Table 5, we discuss the effect of different loss functions on the recognition accuracy of the network. In addition to the loss function we adopted, we also selected three other commonly used loss functions (Circle loss, Triplet loss, and Sphere loss). In the experimental design, we use the exact same SR-DSFF and FENet-ReID and only replace the person Re-ID loss (equation (11)) with other loss functions during network training. Experimental results show that our loss function has the best performance on FENet-ReID guided by human pose estimation.

In Table 6, our method and variants of our method (trained with/without and “weighting loss”) are compared with other SR-based person Re-ID methods. It can be seen from Table 5 that adding weighting loss during training greatly improves the accuracy of Re-ID. At the same time, our method significantly improves the performance of other SR-based cross-resolution person Re-ID methods.

## 5. Conclusion

In this paper, a deep network composed of SR-DSFF and FENet-ReID is proposed to solve the cross-resolution person Re-ID problem. That is a new idea for solving cross-resolution person Re-ID problem, that is, in SR-DSFF, the dynamic Meta-Upscale module is used to recover the LR images to SR images in the SR module, and through the dual-weighted feature extraction stream in the DSFF, the fusion feature maps with more effective pedestrian information are obtained, and the final images is recovered through the transposed convolution. Then, the FENet-ReID is used to segment the three key regions of the person based on the human posture estimation, and the feature extraction is carried out combined with the full images and key region images for person Re-ID. We conducted extensive experiments on three datasets to verify the effectiveness of the proposed method.



## Data Availability

The datasets used and analyzed during the current study available from the corresponding author on reasonable request.

## Conflicts of Interest

The authors declare that they have no conflicts of interests.

## Authors' Contributions

Z.Z.W. was responsible for writing articles, proposing innovations, and conducting experiments. D.L.Z., Q.W.P., S.T.S., and T.M did background research. J.Z. and X.C.Y. supervised the whole project. All authors reviewed the manuscript.

## References

- [1] Y. Wang, L. Wang, Y. You et al., "Resource Aware Person Re-identification across Multiple Resolutions," in *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8042–8051, Salt Lake City, UT, USA, June 2018.
- [2] Y. Huang, Z.-J. Zha, X. Fu, and W. Zhang, "Illumination-invariant Person Re-identification," in *Proceedings of the 27th ACM International Conference on Multimedia (MM'19)*, pp. 365–373, Nice, France, October 2019.
- [3] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, and X. Chen, "Vrsc: Occlusion-free Video Person Re-identification," in *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7183–7192, Long Beach, CA, USA, June 2019.
- [4] J. Pang, D. Zhang, H. Li, W. Liu, and Z. Yu, "Hazy Re-ID: An Interference Suppression Model for Domain Adaptation Person Re-identification under Inclement Weather Condition," in *Proceedings of the 2021 IEEE International Conference on Multimedia and Expo (ICME)*, Shenzhen, China, July 2021.
- [5] L. Zheng, L. Shen, T. Lu, S. Wang, J. Wang, and Q. Tian, "Scalable Person Re-identification: A Benchmark," in *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, December 2015.
- [6] L. Wei, Z. Rui, X. Tong, and X. Wang, "DeepReID: Deep Filter Pairing Neural Network for Person Re-identification," in *Proceedings of the Computer Vision & Pattern Recognition*, June 2014.
- [7] X. Y. Jing, X. Zhu, F. Wu et al., "Super-resolution person re-identification with semi-coupled low-rank discriminant dictionary learning," *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 26, pp. 695–704, 2015.
- [8] Z. Wang, R. Hu, Y. Yu, J. Junjun, L. Chao, and W. Jinqiao, "Scale-adaptive low-resolution person re-identification via learning a discriminating surface," in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16)*, New York, NY, USA, July 2016.
- [9] H. Liu, Z. Xiao, B. Fan, Z. Hui, Z. Yifan, and J. Guoquan, "PrGCN: probability prediction with graph convolutional network for person re-identification," *Neurocomputing*, vol. 423, pp. 57–70, 2021.
- [10] Y. Shen, H. Li, S. Yi, C. Dapeng, and W. Xiaogang, "Person Re-identification with Deep Similarity-Guided Graph Neural network," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 486–504, Berlin, Germany, July 2018.
- [11] K. Han, Y. Huang, Z. Chen, L. Wang, and T. Tan, "Prediction and recovery for adaptive low-resolution person re-identification," in *Computer Vision-ECCV 2020* Springer, Switzerland, Europe, 2020.
- [12] J. Jiao, W. S. Zheng, and A. Wu, "Deep low-resolution person re-identification," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018, <https://www.semanticscholar.org/author/Xiatian-Zhu/2171228https://www.semanticscholar.org/author/S.-Gong/144784813>.
- [13] X. Li, W.-S. Zheng, X. Wang, T. Xiang, and S. Gong, "Multi-scale Learning for Low-Resolution Person Re-identification," in *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 3765–3773, IEEE, Santiago, Chile, December 2015.
- [14] S. Espejo, A. Rodríguez-Vázquez, R. Domínguez-Castro, and R. Carmona, "Convergence and Stability of the FSR CNN Model," in *Proceedings of the Third IEEE International Workshop on Cellular Neural Networks and Their Applications (CNNA-94)*, pp. 411–416, IEEE, Rome, Italy, December 1994.
- [15] T. Tong, G. Li, X. Liu, and Q. Gao, "Image Super-resolution Using Dense Skip Connections," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4799–4807, Venice, Italy, October 2017.
- [16] C. Ledig, L. Theis, F. Huszár et al., "Photo-realistic Single Image Super-resolution Using a Generative Adversarial Network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4681–4690, Honolulu, HI, USA, July 2017.
- [17] X. Hu, H. Mu, X. Zhang, Z. Wang, T. Tan, and J. Sun, "MetaSR: a magnification-arbitrary network for super-resolution," in *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1575–1584, Long Beach, CA, USA, June 2019.
- [18] C. Dong, C. L. Chen, K. He, and T. Xiaoou, "Image Super-resolution Using Deep Convolutional Networks," 2015, <https://arxiv.org/abs/1501.00092>.
- [19] L. Xia, J. Zhu, and Z. Yu, "Real-World Person Re-Identification via Super-Resolution and Semi-Supervised Methods," *IEEE Access*, vol. 9, Article ID 35834, 2021.
- [20] Z. Wang, J. Chen, and S. C. H. Hoi, "Deep Learning for Image Super-resolution: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, 2020.
- [21] Z. Zhuang, H. Ai, L. Chen, and C. Shang, "Cross-resolution Person Re-identification with Deep Antithetical learning," in *Proceedings of the Asian Conference on Computer Vision*, 2018.
- [22] Z. Cheng, Q. Dong, S. Gong, and X. Zhu, "Inter-task association critic for cross-resolution person re-identification," in *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2605–2615, Seattle, WA, USA, June 2020.
- [23] L. Zhang, T. Xiang, and S. Gong, "Learning a Discriminative Null Space for Person Re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1239–1248, Las Vegas, NV, USA, June 2016.
- [24] L. Wu, C. Shen, and A. Hengel, "Personnet: Person Re-identification with Deep Convolutional Neural Networks," 2016, <https://arxiv.org/abs/1601.07255>.
- [25] R. R. Varior, M. Haloi, and G. Wang, "Gated Siamese Convolutional Neural Network Architecture for Human Re-identification," in *Proceedings of the European Conference on*

- Computer Vision*, pp. 791–808, Springer, Berlin, Germany, 2016.
- [26] T. Xiao, H. Li, W. Ouyang, and X. Wang, “Learning Deep Feature Representations with Domain Guided Dropout for Person Re-identification,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1249–1258, Silver Spring, MD, USA, April 2016.
- [27] E. Ahmed, M. Jones, and K. M. Tim, “An improved deep learning architecture for person re-identification,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA, USA, June 2015.
- [28] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. H. Hoi, “Deep learning for person Re-identification: a survey and outlook,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 99, 1 page, 2021.
- [29] Z. Zheng, L. Zheng, and Y. Yang, “A discriminatively learned CNN embedding for person re-identification,” *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 14, no. 1, pp. 1–20, 2017.
- [30] H. Chen, Y. Wang, Y. Shi et al., “Deep Transfer Learning for Person re-identification,” in *Proceedings of the 2018 IEEE Fourth International Conference on Multimedia Big Data (Big MM)*, pp. 1–5, IEEE, Xi’an, China, September 2018.
- [31] L. Wei, S. Zhang, H. Yao, W. Gao, and Q. Tian, “Glad: global-local-alignment descriptor for pedestrian retrieval,” in *Proceedings of the 25th ACM International Conference on Multimedia*, pp. 420–428, Los Cabos, Mexico, September 2017.
- [32] C. Liu, X. Chang, and Y. D. Shen, “Unity Style Transfer for Person Re-identification,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6887–6896, Seattle, WA, USA, June 2020.
- [33] X. Qian, Y. Fu, X. Tao et al., “Pose-Normalized Image Generation for Person Re-identification Part IX,” in *Proceedings of the 15th European Conference*, Munich, Germany, September 2018.
- [34] J. Miao, Y. Wu, P. Liu, Y. Ding, and Y. Yang, “Pose-guided feature alignment for occluded person re-identification,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 542–551, Seoul, Korea, October 2019.
- [35] S. Gao, J. Wang, H. Lu, and Z. Liu, “Pose-guided visible part matching for occluded person ReID,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11744–11752, Seattle, WA, USA, June 2020.
- [36] T. Wang, H. Liu, P. Song, G. Tianyu, and S. Wei, “Pose-guided Feature Disentangling for Occluded Person Re-identification Based on Transformer,” 2021, <https://arxiv.org/abs/2112.02466>.
- [37] W. S. Zheng, S. Gong, and X. Tao, “Person Re-identification by Probabilistic Relative Distance Comparison,” in *Proceedings of the Computer Vision & Pattern Recognition*, June 2011.
- [38] L. Xiao, M. Song, D. Tao, X. Zhou, C. Chen, and J. Bu, “Semi-supervised Coupled Dictionary Learning for Person Re-identification,” in *Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition*, June 2014.
- [39] S. Mao, S. Zhang, and M. Yang, “Resolution-invariant Person re-identification,” 2019, <https://arxiv.org/abs/1906.09748>.
- [40] G. Zhang, Y. Ge, Z. Dong, W. Hao, Z. Yuhui, and C. Shengyong, “Deep High-Resolution Representation Learning for Cross-Resolution Person Re-identification,” *IEEE Transactions on Image Processing*, vol. 30, 2021 <https://arxiv.org/abs/2105.11722>.
- [41] A. Torralba, A. Oliva, M. S. Castelhana, and J. M. Henderson, “Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search,” *Psychological Review*, vol. 113, no. 4, p. 766, 2006.
- [42] A. Oliva and A. Torralba, “Building the gist of a scene: the role of global image features in recognition,” *Progress in Brain Research*, vol. 155, pp. 23–36, 2006.
- [43] B. Cao, A. Araujo, and J. Sim, “Unifying Deep Local and Global Features for Image Search,” *European Conference on Computer Vision*, Springer, Switzerland, Europe, pp. 726–743, 2020.
- [44] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, “Person re-identification by multi-channel parts-based cnn with improved triplet loss function,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1335–1344, Las Vegas, NV, USA, June 2016.
- [45] S. Wu, Y. C. Chen, X. Li, A. -C. Wu, J. -J. You, and W. -S. Zheng, “An Enhanced Deep Feature Representation for Person Re-identification,” in *Proceedings of the 2016 IEEE winter Conference on Applications of Computer Vision (WACV)*, pp. 1–8, IEEE, Lake Placid, NY, USA, March 2016.
- [46] D. Yi, Z. Lei, S. Liao, and S. Z. Li, “Deep metric learning for person re-identification,” in *Proceedings of the 2014 22nd international conference on pattern recognition*, pp. 34–39, IEEE, Stockholm, Sweden, August 2014.
- [47] Y. Ge, H. Wang, F. Zhu, R. Zhao, and H. Li, “Self-supervising fine-grained region similarities for large-scale image localization,” *European Conference on Computer Vision*, Springer, Switzerland, Europe, 2020.
- [48] Z. Zheng, Y. Wei, and Y. Yang, “University-1652: A Multi-view Multi-source Benchmark for Drone-based Geo-localization,” in *Proceedings of the 28th ACM international conference on Multimedia*, pp. 1395–1403, Seattle, WA, USA, February 2020.
- [49] Y. Jo, S. W. Oh, J. Kang, and J. K. Seon, “Deep Video Super-resolution Network Using Dynamic up-Sampling Filters without Explicit Motion compensation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3224–3232, Silver Spring, MD, USA, June 2018.
- [50] C. Finn, P. Abbeel, and S. Levine, “Model-agnostic Meta-Learning for Fast Adaptation of Deep Networks,” in *Proceedings of the 34th International Conference on Machine Learning*, pp. 1126–1135, PMLR, New York, NY, USA, 2017.
- [51] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, “Residual Dense Network for Image Super-resolution,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2472–2481, Salt Lake, UT, USA, June 2018.
- [52] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, Las Vegas, NV, USA, June 2016.
- [53] A. Toshev and C. Szegedy, “DeepPose: Human Pose Estimation via Deep Neural Networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1653–1660, Columbus, OH, USA, June 2014.
- [54] H. Zhao, M. Tian, S. Sun et al., “Spindle net: person re-identification with human body region guided feature decomposition and fusion,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1077–1085, Honolulu, HI, USA, July 2017.
- [55] C. Szegedy, W. Liu, Y. Jia et al., “Going Deeper with Convolutions,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–9, Boston, MA, USA, June 2015.
- [56] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, “Scalable Person Re-identification: A Benchmark,” in

- Proceedings of the IEEE International Conference on Computer Vision*, pp. 1116–1124, Santiago, Chile, December 2015.
- [57] Z. Zheng, L. Zheng, and Y. Yang, “Unlabeled samples generated by gan improve the person re-identification baseline in vitro,” in *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 3754–3762, Venice, Italy, October 2017.
- [58] W. Li, R. Zhao, T. Xiao, and X. Wang, “Deepreid: Deep Filter Pairing Neural Network for Person re-identification,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 152–159, Columbus, OH, USA, June 2014.
- [59] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, “Enhanced Deep Residual Networks for Single Image Super-resolution,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 136–144, Honolulu, HI, USA, July 2017.
- [60] M. M. Kalayeh, E. Basaran, M. Gokmen, M. E. Kamasak, and M. Shah, “Human Semantic Parsing for Person Re-identification,” in *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2018.
- [61] Z. Zhong, Z. Liang, Z. Zheng, S. Li, and Y. Yang, “Camera Style Adaptation for Person Re-identification,” in *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2018.
- [62] Y.-J. Li, Y.-C. Chen, Y.-Y. Lin, X. Du, and Y.-C. F. Wang, “Recover and identify: a generative dual model for cross-resolution person re-identification,” in *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8090–8099, Seoul, Korea, October 2019.
- [63] T. Yu, P. Xi, Z. Long, Z. Shaoting, and N. M. Dimitris, “Cr-Gan: Learning Complete Representations for Multi-View Generation (IJCAI 2018),” in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)*, Montreal, Canada, 2018.
- [64] C. Sharma, S. R. Kapil, and D. Chapman, “Person Re-identification with a locally aware transformer,” 2021, <https://arxiv.org/abs/2106.03720>.
- [65] J. Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired Image-To-Image Translation Using Cycle-Consistent Adversarial Networks,” in *Proceedings of the IEEE International Conference on Computer Vision*, October 2017.
- [66] Z. Wang, M. Ye, F. Yang, and B. Xiang, “Cascaded SR-GAN for scale-adaptive low resolution person re-identification,” in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)*, Montreal, Canada, 2018.

## Research Article

# Spatial-Temporal Change Trend Analysis of Second-Hand House Price in Hefei Based on Spatial Network

Zheng Yin , Rui Sun , and Yuqing Bi 

*School of Economics and Management, Anhui Jianzhu University, Hefei, Anhui, China*

Correspondence should be addressed to Zheng Yin; 313282807@qq.com

Received 13 April 2022; Accepted 4 May 2022; Published 23 May 2022

Academic Editor: Nian Zhang

Copyright © 2022 Zheng Yin et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Spatial Markov chain can effectively explore the spatial evolution trend of housing price under the influence of lag factor. This paper uses spatial autocorrelation and spatial Markov to study 353 second-hand houses in Hefei. The results show that (1) the housing price of Hefei urban area presents a situation of “two points and one side,” the high housing price is concentrated in the south and southwest of the urban area, and the price level gradually weakens from south to north, and the housing development shows a north-south differentiation. (2) There is a significant spatial autocorrelation between second-hand housing prices in Hefei. The “high-high” residential price clusters are mainly distributed in Shushan District and Binhu New Area, while the “low-low” residential price clusters are mostly in Yaohai district and its surrounding areas. The number of “low-high” agglomeration and “high-low” agglomeration is small, and the degree of change is not big. (3) Under the influence of different neighborhood environments, the housing prices in urban Area of Hefei show club convergence overall. At the same time, under the short-term influence of the policy, the housing prices of low level and middle and low level are promoted in the same neighborhood environment, while the housing prices of high level and middle and high level are negatively affected.

## 1. Introduction

Housing is a major issue concerning people’s livelihood, affecting the basic livelihood of millions of families. With the continuous improvement of economic life, the demand of house buyers is getting higher and higher. However, in order to meet the needs of residents, real estate developers actively launch differentiated housing products, so as to broaden the range of housing options for residents, but also produce some negative effects: urban living space misallocation, obvious spatial differentiation, housing prices, housing prices causing “speculation,” eventually leading to excessive real estate investment, forming the “bubble economy,” but really needing just to be difficult to meet the residents of housing, affecting the well-being of the people, which is not conducive to social stability and harmonious [1].

Hefei is a large city with a population of 10 million. Compared with 2010, its permanent population increased by 1.91 million, and the proportion of urban population increased by 20 percentage points [2]. Rapid urbanization and population agglomeration bring about increased demand

and supply pressure for housing. The exuberance of demand and the relative reduction of supply cause the distortion of the market and the relationship between supply and demand. Under the combined action of multiple factors, such as the rise of land price and the influx of investors into the market, the real estate market in Hefei appears irrational overheating and fails to meet the demands of consumers with real rigid demand for housing [3]. In order to promote the steady and healthy development of Hefei real estate market, Hefei city issued new real estate policies in early April 2021, which began to strictly manage the land market and real estate market, strictly regulate the price of commercial housing, and severely crack down on all kinds of real estate market disorder.

Taking 353 second-hand residential houses in Hefei as the research object, this paper analyzes the spatial distribution difference of housing prices in different regions by means of mathematical statistics and spatial analysis (see Figure 1) and explores the rule of housing price type transformation under the influence of policies. It aims to promote the benign development of Hefei housing market,

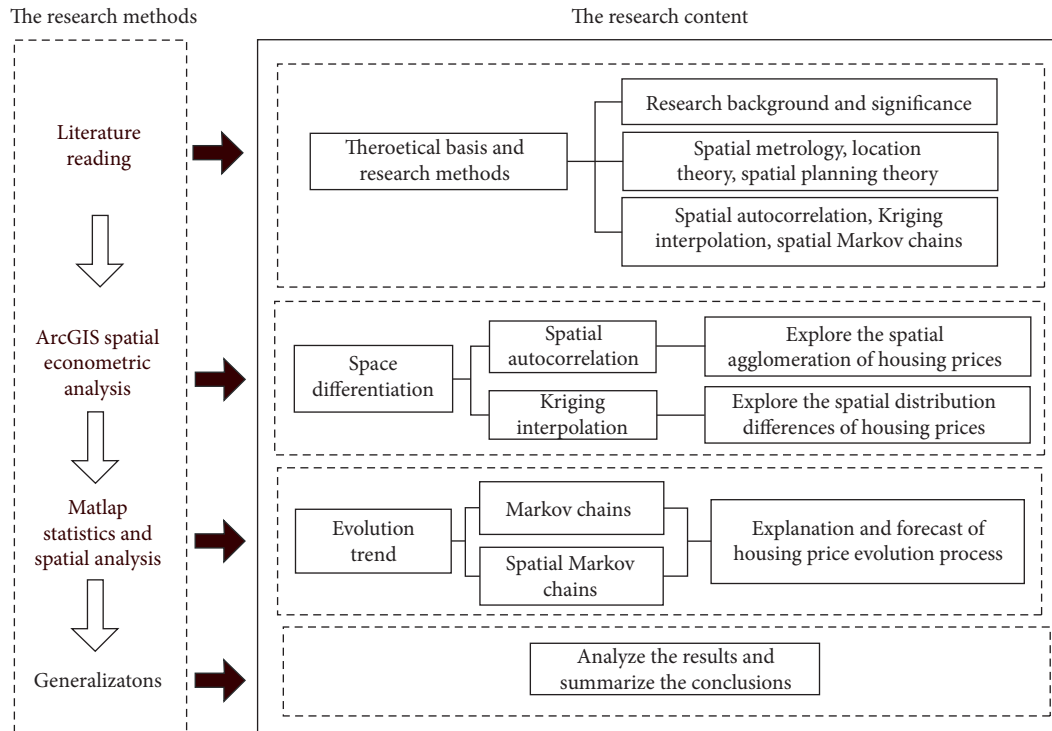


FIGURE 1: Organizational flow chart.

solve the real estate market disorder and supply and demand contradiction, and provide reference for Hefei municipal government and planning departments to formulate regional housing price regulation policies and public facilities planning.

## 2. Literature Review

As immovable property, a house has not only the basic property of living, but also high investment value. Different lots of real estate, the future appreciation space has also different size, because its high purchase cost limits the consumer group's choice range. Therefore, the change trend of housing price is one of the hot issues with the highest social attention. In recent years, many domestic scholars have introduced spatial data to study its spatial correlation and structure based on the use of traditional econometric models. The spatial distribution of urban residential prices has been studied from a spatially macroscopic perspective, and the patterns and causes of residential price changes in space have been quantified and analyzed by constructing a mathematical analysis model.

Kriging interpolation method can study the spatial differentiation of housing prices and the spatial distribution characteristics and evolution rules of housing prices, and explore the evolution trend of the outward diffusion of regional housing spatial development in recent years [4–7]. Compared with other interpolation methods, Kriging method not only considers spatial correlation, but also results are more reliable when there are more data points [4]. Kriging is an accurate local interpolation technique, which takes into account the spatial orientation of sample points

and the spatial position relationship with unknown sample points [5].

Moran's I index is often used as a tool to analyze the spatial correlation and heterogeneity of housing prices, and reflects the spatial agglomeration and dispersion characteristics of housing through significance [8–13]. Global spatial autocorrelation is an assessment of the degree of spatial autocorrelation, which reflects the overall trend of spatial correlation of observed variables in the whole research area [8]. Moran's I measures the relationship between spatial elements, which is similar to the correlation coefficient in statistics. Its value ranges from  $-1$  to  $1$ . If it is greater than zero, it indicates a positive correlation; if it is less than zero, it indicates a negative correlation; if it is equal to zero, it indicates no spatial correlation [9]. Moran's I index is greatly influenced by the spatial weight matrix, and the global space can explore the change rule of the spatial correlation of housing price under different spatial weights [10]. Significant Moran's I value indicates that the price of new residential buildings in urban areas has spatial agglomeration, that is, plots with high prices gather together, and plots with low prices gather together [11]. Luo and Wei analyzed the land value of Milwaukee by geostatistical method, and found that urban land price has significant spatial correlation, and there are differences among different locations and land use properties [12]. Liu et al. used the global autocorrelation model to find that China's real estate prices present a positive spatial correlation on the whole. For each region, the spatial spillover effect of real estate prices in the eastern and central regions is significant, and the real estate prices in the eastern region have a stronger spatial positive correlation [13].

Space Markov chain method is the traditional method of Markov chain and regional condition. Markov chain method optimized combination, giving full consideration to the space between the time series of regional interaction, through spatial weight matrix that can solve the problem of the spatial relations between areas, and with the aid of lagging behind the concept of space, define each field of spatial neighborhood. Thus, the spatial effects of geographical environment on regional development are quantitatively analyzed [14–17]. After analyzing the change of housing price in Britain, Sean Holly found that the change of housing price in London would lead to the change of housing price in other areas, and such influence had a certain lag [14]. Xue Liang used spatial Markov chain model to quantitatively study the ecological security and economic level of Guanzhong region and summarized its spatial-temporal evolution characteristics [15, 18]. Zhou Li used traditional Markov chain and spatial Markov chain methods to construct nonspatial and spatial Markov transfer probability matrices of rural economic development level, respectively, and analyzed and predicted the evolution characteristics of spatial-temporal pattern of rural economic development level in the research period [16]. Yan Tao et al. used spatial data statistical analysis model and spatial Markov model to analyze regional differences and spatiotemporal evolution characteristics of Urban economic development in China from 2001 to 2016 [17].

### 3. Data Sources and Research Methods

**3.1. Data Sources.** Four districts, one city, and four counties are under the jurisdiction of Hefei city. This paper mainly selects four municipal districts of Hefei (Yaohai district, Luyang District, Shushan District and Baohe District) as the research scope (As shown in Figure 2), and collects the price data of 353 second-hand housing in Hefei from 2020 to 2021. Among them, the sample panel data comes from Anjuke, and the spatial vector data of the base map comes from the national basic data of the National geographic Information Resource Catalogue Service system, through the use of Baidu map to pick up and edit coordinate information, combined with ArcGIS, GeoDa, and Matlab software for data processing and analysis.

#### 3.2. Research Methods

##### 3.2.1. Spatial Autocorrelation

- (1) Global autocorrelation model. Global autocorrelation is used to quantitatively describe the average degree of association of all spatial units with neighboring regions over the whole region, so as to determine whether the phenomenon exists in spatial agglomeration. In this paper, Global Moran's I is selected to reflect the overall distribution of commodity housing prices in each neighborhood, and its calculation formula is as follows.

$$\text{Global Moran's } I = \frac{n \sum_{i=1}^n \sum_{j=1}^n W_{ij} (y_i - \bar{y})(y_j - \bar{y})}{\left(\sum_{i=1}^n \sum_{j=1}^n W_{ij}\right) \sum_{i=1}^n W_{ij} (y_i - \bar{y})^2}, \quad (1)$$

where  $y_i y_j$  is the house price of the  $i$ th and  $j$ th cell, respectively,  $\bar{y}$  is the mean value of house prices of all cells,  $W_{ij}$  is the value of spatial weights between cell  $i$  and cell  $j$  (the spatial weight matrix of distances is constructed in this paper), and  $n$  is the total number of cells studied.

- (2) Local autocorrelation model. Local autocorrelation can be used to reflect the degree of spatial correlation between local study units in the study area and the values of similar attributes in the surrounding area, and the calculation formula is

$$\text{Local Moran's } I = \frac{n(y_i - \bar{y}) \sum_{j=1}^m W_{ij} (y_j - \bar{y})}{\sum_{i=1}^n (y_i - \bar{y})^2}, \quad (2)$$

where  $y_i$  is the house price of the  $i$ th,  $W_{ij}$  is the spatial weight value between cell  $i$  and cell  $j$ ,  $n$  is the total number of cells studied, and  $m$  is the number of cells adjacent to cell  $i$ .

**3.2.2. Kriging Interpolation Method.** The Kriging interpolation method is a method for unbiased optimal estimation of regionalized variables within a certain region based on the theory of variance function and structural analysis [6, 7]. The Kriging interpolation method considers the correlation between cells, and the test results are more informative in the case of multiple sample points. In this paper, the general kriging interpolation method, which has a wide range of applications, is used to interpolate housing prices, and its formula is as follows:

$$\hat{Z}(x_0) = \sum_{i=1}^n \lambda_i Z(x_i). \quad (3)$$

$\hat{Z}(x_0)$  denotes the predicted value of the unknown point,  $Z(x_i)$  denotes the value of the surrounding known points,  $\lambda_i$  denotes the weight of the  $i$ th known point on the unknown point, and  $n$  is the amount of sample data.

**3.2.3. Markov Chain.** Markov chain is a stochastic process in which both time and state are in a discrete state. In the process of analysis, continuous values are discretized and divided into  $k$  types by numerical rank, and then the probability distribution of each type and its interannual variation are calculated to approximate the evolution of things. The expressions are as follows:

$$m_{ij} = \frac{n_{ij}}{n_i}, \quad (4)$$

where  $n_i$  is the number of dwellings belonging to type  $i$  in the study time period, and  $n_{ij}$  refers to the number of residential buildings that changed from type  $i$  in  $t$  years to type  $j$  in  $t + 1$  years during the study period.

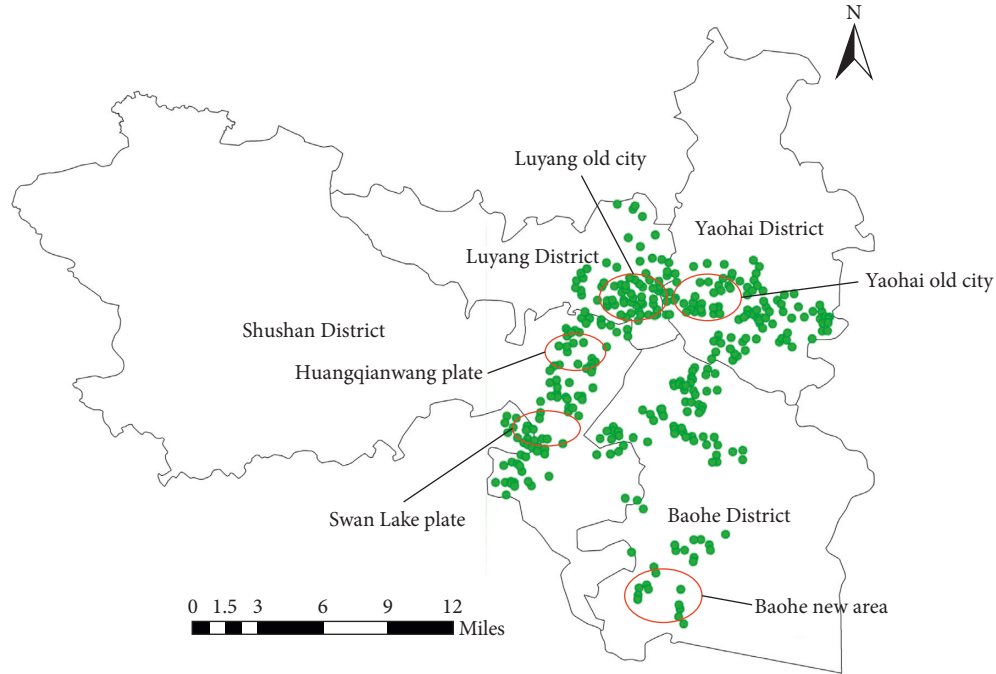


FIGURE 2: Distribution of Hefei city districts.

**3.2.4. Spatial Markov Chain.** Traditional Markov chain can count the spatiotemporal evolution of second-hand housing prices, unable to analyze the influence of the economic development in the region of neighborhood region, and thus on the basis of traditional Markov chain, it is introduced into the spatial lag conditions to build the space of the Markov chain, which is an effective analysis of the regional economic development and its surrounding residential environment for residential type change in the price. The spatial lag operator corresponding to the spatial lag operator is often used in spatial autocorrelation analysis. The spatial lag of a local area is the weighted average value of the observation values around the location; that is, the product ( $WX$ ) of the regional observation vector ( $X$ ) and the spatial weight matrix ( $W$ ) is used to determine the neighborhood state of the region. It provides a method basis for quantitative analysis of regional spatial distribution pattern [17].

$$\text{Lag} = \sum_{i=1}^n x_i w_{ij}, \quad (5)$$

where Lag is the spatial lag operator,  $x_i$  is the attribute value of the regional cell, and  $w_{ij}$  is the weight of the observation of domain  $j$  for the spatial lag operator at location  $i$ . The traditional  $k \times k$  Markov matrix is decomposed into  $k \times k \times k$  conditional transfer probability matrices conditional on the spatial lag according to the economic state or type of the adjacent region.  $m_{ij}(k)$  denotes a spatial transfer probability conditional on the spatial lag type of the cell at moment  $t$ , which is transferred from type  $i$  to type  $j$ . By comparing the traditional Markov matrix with the spatial Markov transfer matrix to explore the probability of upward or downward transferring of a study unit, the transformation of different used residential price types in different neighborhood

environments can be analyzed, and the degree of influence of the neighborhood environment on the price transfer of used residential units can be studied.

## 4. Empirical Analysis

### 4.1. Spatially Divergent Characteristics of Second-Hand House Prices in Hefei City

**4.1.1. Global Autocorrelation.** In this paper, the global Moran's I index of second-hand residential prices in Hefei city from 2020 to 2021 are calculated based on the spatial weight matrix of Euclidean distance (Table 1). It can be seen from the table that all the global Moran's I indices are positive, and are between 0.6037 and 0.6896, with a confidence of 99%. The results show that the second-hand houses in Hefei urban area have significant spatial correlation effect and show agglomeration phenomenon in space. Among them, under the influence of the new real estate policy, Moran's I index of second-hand housing price in Hefei decreased from 0.6883 to 0.6791 in a short period after April. It shows that the agglomeration effect is slightly weakened, and the trend of housing price growth has been temporarily controlled.

**4.1.2. Local Autocorrelation.** The global Moran's I index shows that there are different levels of spatial clustering of second-hand residential prices in Hefei city, which does not show the spatial clustering characteristics of second-hand residential prices in Hefei city. Therefore, this paper adopts the LISA diagram of local autocorrelation to analyze the spatial distribution and spatial clustering characteristics of

TABLE 1: Global Moran index of second-hand residential prices in Hefei city, 2020–2021.

	Moran's I	Z score	P value
Oct.	0.6037	25.6593	<0.001
Nov.	0.6079	25.8290	<0.001
Dec.	0.6305	26.7703	<0.001
Jan.	0.6430	27.2898	<0.001
Feb.	0.6450	27.3593	<0.001
Mar.	0.6848	29.0196	<0.001
Apr.	0.6883	29.1532	<0.001
May.	0.6825	28.8979	<0.001
Jun.	0.6791	28.7499	<0.001
Jul.	0.6850	28.9959	<0.001
Aug.	0.6896	29.1880	<0.001
Sept.	0.6890	29.1619	<0.001

residential prices in Hefei city. Figure 3 shows the spatial agglomeration of housing prices in the four phases.

As shown in Figure 3, the “high-high” agglomeration is mainly distributed in the area of Shushan District, such as the governmental affairs district board, Huang Qianwang board, Economic Development District board, and Binhu New District board. The “low-low” agglomeration is mainly distributed in the northeastern area, such as Yaohai Old Town board, Xinzhan District board, and partial Luyang District board. The residential houses in the “low-high” agglomeration are mainly scattered around the “high-high” agglomeration, and the number is gradually decreasing, while the residential houses in the “high-low” agglomeration are more spatially dispersed, and the number of “high-low” clusters is spatially dispersed and remains low for a long time.

Figure 3(a) shows: in October 2020, “high-high” concentrated in the government affairs area of Shushan District, closely surrounding Swan Lake and municipal government affairs center, such as The Arch of Triumph, Ink Orchid Pavilion, Swan Lake bank and Sansheng Yi Garden. “Low-low” residential cluster is mainly distributed in Yaohai district and Yaohai district and Luyang district junction; “high-low” cluster houses are small in number and scattered around “low-low” cluster houses, located at the intersection of Luyang and Yaohai old city; “low-high” clustered houses are mainly distributed around “high-high” clustered houses, between government district and economic district, and the rest are scattered in Huangqianwang plate.

Figure 3(b) shows the following: compared with October 2020, in January 2021, “high-high” clustered residences are still mainly concentrated in the government affairs district board and Binhu New District board in Shushan District, among which Binhu New District expands 5 “high-high” clustered objects; the number of “low-low” clustered residences increases, mainly in Baohe District, and the location north of the junction of Luyang District and Yaohai District; “high-low” clustered and “low-high” clustered residences do not change significantly in spatial location, and increase or decrease in quantity.

Figure 3(c) shows that, compared to January 2021, the “high-high” agglomeration in April of that year expanded in

the governmental district section of Shushan District and the Binhu New District section, and the degree of expansion was not obvious; the distribution pattern of the “low-low” agglomeration changed more, except for the local areas along the Banqiao River in Yaohai District and Luyang District, and the increase in the number in Baohe District was more obvious, mainly in the location north of Taihu Road, such as Chengjian Century Garden and the Youth District. The “high-low” agglomeration and the “low-high” agglomeration do not have significant changes in the space and number of residences.

Figure 3(d) shows, that compared with April 2021, the location of “high-high” agglomeration in July of that year has not changed significantly, but the number of “high-high” agglomeration has increased slightly compared with the previous one, mainly in Baohe District, such as Edinburgh of World Jincheng, Windsor City of World Jincheng and Oriental Residence in Baohe District; the number of residences in “low-low” agglomeration has decreased in Baohe District, but they are still concentrated in Yaohai District and its surroundings; “high-low” agglomeration is scattered around the residences in “low-low” agglomeration and the number has decreased. There is no significant change in the number of “low-high” clusters, and only one place in Shushan District, Newspaper Park (East), is influenced by the surrounding “high-high” clusters to change from “low-high” clusters to “high-high” clusters.

#### 4.2. Overall Divergent Characteristics of Second-Hand House Prices in Hefei City

4.2.1. *Kriging Interpolation Method.* The Kriging interpolation analysis method using the spatial distribution tool of ArcGIS was used to locally interpolate the residential price data to generate a continuous price surface, as shown in Figure 4.

According to the results of Kriging interpolation analysis, the housing prices in hefei urban area decrease from west to east and from south to north, which can be summarized as the distribution of “two points and one side”. Two “points” are Shushan district and Binhu New District, which are two areas with high housing prices, while “one side” is Luyang District and Yaohai District, which are relatively low and evenly distributed. At the same time, the areas with high housing prices gradually decrease from the circle to the periphery and the grading is obvious. There is no obvious leap-over phenomenon, and the housing prices show a certain agglomeration phenomenon in the region. Since the implementation of the New Deal, Yaohai district and its surrounding old city in the long-term low value level, only a few local areas of the price growth trend. There may be several reasons for the long-term low value of Yaohai District and its surroundings:

- (1) Luyang District and Yaohai District, as the old urban areas of Hefei, carry the history of urban development, and their old planning and old buildings lead



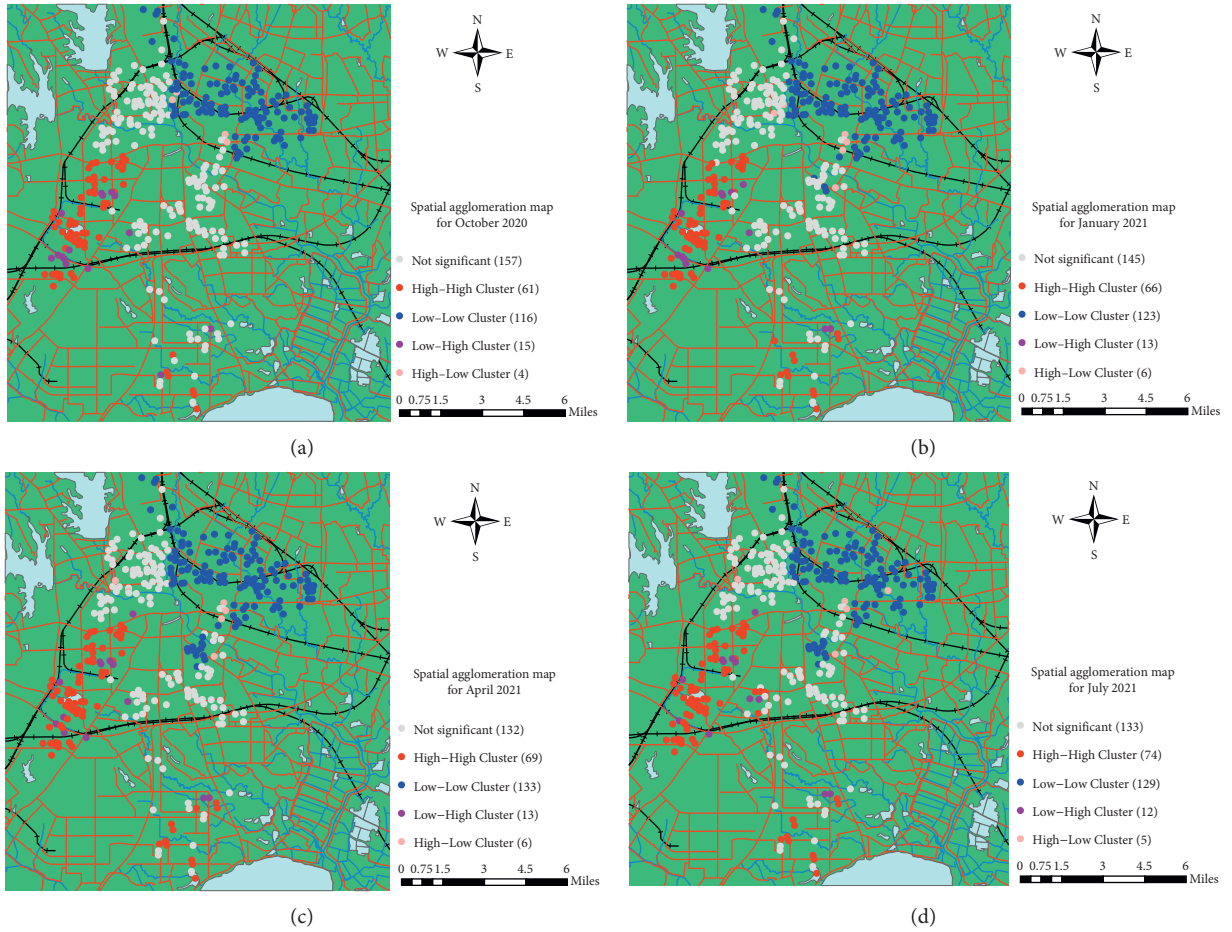


FIGURE 3: Local spatial distribution of second-hand residential prices in Hefei city, 2020–2021.

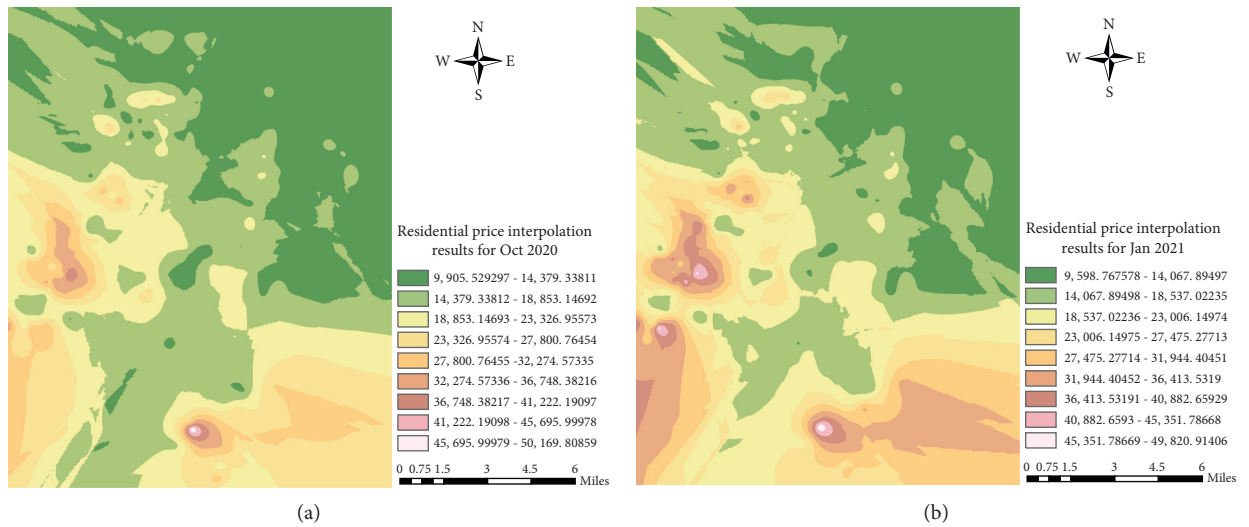


FIGURE 4: Continued.

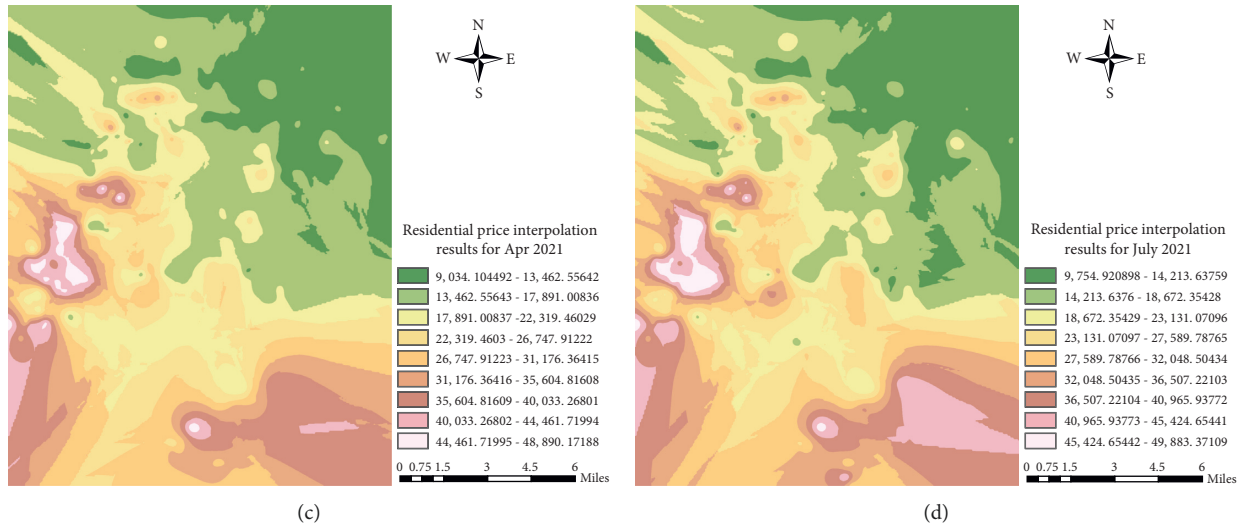


FIGURE 4: Distribution of kriging space of second-hand residential prices in Hefei city in April and September 2021.

to land scarcity to limit their real estate development potential.

- (2) For a long time, Luyang district and Yaohai District have been the economic center and industrial center of Hefei respectively. The government and enterprises have built low-income housing such as demolition houses and public rental houses to solve the housing problem of the working population. As a result, residents in the old city have no more demand for housing.
- (3) The excessive development and use of the old city makes the blocks appear to be “old and small, dirty and messy,” but the poor living environment limits the rise of housing prices. From the point of view of spatial distribution, the local area of Baohe River region changed from high to low, and then remained stable. It shows that there is a certain “inflated price” problem in the reduced regional price, and the “inflated” regional price drops to the “real” level. Shushan district and Binhu New District have obviously slowed down the trend of outward diffusion, price growth has slowed down. This shows that housing prices in Shushan district and Binhu New Area have gradually returned to a state of stable growth after the impact of policy adjustment.

### 4.3. Time Evolution Characteristics of Second-Hand House Prices in Hefei City

4.3.1. *Markov Shift Matrix.* With the support of ArcGIS software, the Markov shift probability matrix of the second-hand residential property prices in Hefei city was obtained by overlaying the data of previous years (as shown in Table 2).

- (1) As can be seen from Table 2, the larger values in the matrix of residential price types in the two different

time periods are concentrated on the main diagonal, which indicates that the residential prices in the second-hand residential market in Hefei city have a high stability in the process of development. From the values on the diagonal, it can be seen that the probability of residential prices maintaining their original level type in the first six months is at least 88.9%, and the probability of residential price levels maintaining their original state in the second six months is at least 87.6%. Compared with the first six months, the ability of residential prices to maintain their original level in the later period affected by policy regulation has slightly decreased, and policy regulation has played a certain effect.

- (2) Different types of residential prices are affected by policy adjustments to different degrees. As can be seen from Table 2, the probability of low level, low and medium level, and high level types of residential to maintain their original state has decreased by the impact of policy regulation. Among them, the probability of upward shift of low and lower level increases, while the probability of upward shift of medium and high level types of residential houses decreases, and only the probability of maintaining their original level is high. This indicates that the medium and high level types of housing are more influenced by the policy, and the target of policy regulation is more clear.
- (3) Each stage still presents upward development trend. Overall, the probability sum of the upper triangle and the lower triangle in stage 1 is 0.205 and 0.065, respectively, and the probability sum of the upper triangle and the lower triangle in stage 2 is 0.205 and 0.064, respectively. The probability sum of the upper triangle of the two stages is larger than that of the lower triangle, which shows that the probability of upward transfer of the housing price type is larger

TABLE 2: Markov shift probability matrix of second-hand residential prices in Hefei city, 2020–2021

Local status $t/t+1$	2020.10–2021.3 (stage 1)					2021.4–2021.9 (stage 2)				
	$n$	Low <25%	Lower 25%~50%	Higher 50%~75%	High >75%	$n$	Low <25%	Lower 25%~50%	Higher 50%~75%	High >75%
Low	448	<b>0.946</b>	0.054	0.000	0.000	449	<b>0.931</b>	0.069	0.000	0.000
Lower	447	0.034	<b>0.897</b>	0.069	0.000	444	0.025	<b>0.876</b>	0.099	0.000
Higher	449	0.000	0.029	<b>0.889</b>	0.082	433	0.000	0.028	<b>0.935</b>	0.037
High	421	0.000	0.000	0.002	<b>0.998</b>	439	0.000	0.000	0.011	<b>0.989</b>

TABLE 3: Spatial Markov transition probability matrix of secondary residential prices in Hefei city in 2020–2021 (conditional on spatial lag).

Spatial lag	Local status $t/t+1$	2020.10–2021.3 (stage 1)					2021.4–2021.9 (stage 2)				
		$n$	Low <25%	Lower 25%~50%	Higher 50%~75%	High >75%	$n$	Low <25%	Lower 25%~50%	Higher 50%~75%	High >75%
Low	Low	313	<b>0.968</b>	0.032	0.000	0.000	318	<b>0.950</b>	0.050	0.000	0.000
	Lower	75	0.080	<b>0.920</b>	0.000	0.000	59	0.102	<b>0.831</b>	0.068	0.000
	Higher	25	0.000	0.000	<b>1.000</b>	0.000	16	0.000	0.063	<b>0.938</b>	0.000
	High	0	0.000	0.000	0.000	0.000	0	0.000	0.000	0.000	<b>0.000</b>
Lower	Low	94	<b>0.904</b>	0.096	0.000	0.000	91	<b>0.890</b>	0.110	0.000	0.000
	Lower	97	0.052	<b>0.887</b>	0.062	0.000	166	0.012	<b>0.898</b>	0.090	0.000
	Higher	98	0.000	0.020	<b>0.939</b>	0.041	89	0.000	0.067	<b>0.921</b>	0.011
	High	5	0.000	0.000	0.000	<b>1.000</b>	8	0.000	0.000	0.000	<b>1.000</b>
Higher	Low	37	<b>0.865</b>	0.135	0.000	0.000	40	<b>0.875</b>	0.125	0.000	0.000
	Lower	230	0.017	<b>0.909</b>	0.074	0.000	194	0.015	<b>0.856</b>	0.129	0.000
	Higher	249	0.000	0.040	<b>0.896</b>	0.064	253	0.000	0.016	<b>0.941</b>	0.043
	High	89	0.000	0.000	0.011	<b>0.989</b>	103	0.000	0.0	0.019	<b>0.981</b>
High	Low	4	<b>1.000</b>	0.000	0.000	0.000	0.000	<b>0.000</b>	0.000	0.000	0.000
	Lower	45	0.000	<b>0.822</b>	0.178	0.000	25	0.000	<b>1.000</b>	0.000	0.000
	Higher	77	0.000	0.013	<b>0.766</b>	0.221	75	0.000	0.013	<b>0.933</b>	0.053
	High	327	0.000	0.000	0.000	<b>1.000</b>	328	0.000	0.000	0.009	<b>0.991</b>

than that of downward transfer, indicating that the housing price presents an upward development trend in general.

- (4) Most residential prices maintain their original state, and the lowest probability of maintaining their original level is 87.6%, which is higher than the probability of the price type shifting to other types. This indicates that there is a “club convergence” of residential price types, and most residential prices tend to converge to higher or high-economic types.

**4.3.2. Spatial Markov Transfer Matrix.** The traditional Markov chain approach is based on the assumption that regions are independent of each other, thus ignoring the positive and negative influence of the neighborhood environment in the dynamic evolution of the region. Residential neighborhoods are relatively small regional units, but they do not exist independently, and they are interconnected with the surrounding areas. Therefore, based on the traditional Markov transfer probability matrix, the spatial Markov chain transfer probability matrix is constructed by introducing the condition of spatial lag through Matlab software (see Table 3).

- (1) Residential neighborhood background plays an important role in the development of residential economy. The probability of economic type transfer is different for a house in different neighborhood. If a house is adjacent to a house with a low price level, it will be negatively affected by the neighborhood, resulting in a negative spatial spillover effect and difficult to move up. When it is adjacent to the house with a higher price level, the positive spillover effect will be generated, which inhibits its downward transfer and promotes its upward transfer. As a result, the house price gradually tends to the same level in space, which provides a spatial explanation for the phenomenon of “club convergence.”
- (2) Different neighborhood environments play different roles in the process of residential price shift. For example, the probability of upward shift is 0.041 and downward shift is 0.020 when a higher type residential neighborhood is adjacent to a lower type residential neighborhood in Stage 2, and 0.178 and 0.000 when it is adjacent to a high type residential neighborhood. 13.7 percentage points, while its probability of downward shift decreases by 20 percentage points. This indicates that the probability of upward shift increases when a residential

neighborhood is adjacent to more developed residential houses; on the contrary, the probability of upward shift is suppressed when it is adjacent to less developed residential houses.

- (3) Policy adjustments play different roles for different types of residences. As a result of the policy, the upward shift of low and low-middle level types of residences is promoted, while the shift of middle and high level types of residences is suppressed. For example, the probability of upward shift for low and medium level types of dwellings in the same medium and high level neighborhood environment is 0.074 for Stage 1 and 0.129 for Stage 2, representing a 5.5% increase in the probability of upward shift. The probability of upward shift of stage 1 to 0.064 and upward shift of stage 2 to 0.043 for the higher level type of housing in the high level neighborhood environment decreased by 2.1%.

## 5. Research Conclusion

This paper analyzes the spatiotemporal evolution characteristics of residential prices in Hefei city by using kriging interpolation, spatial autocorrelation, and spatial Markov chain, based on the study of 353 residential community price seats in Hefei city from 2020 to 2021. The following conclusions are drawn.

- (1) From the perspective of spatial pattern, the residential houses in Hefei city show the situation of “two points on one side,” the high level of residential prices is mainly concentrated in the south and southwest of Hefei city, while Yaohai district and its surroundings have been at low level for a long time, and the residential development is divided between north and south. Hefei city residential prices have obvious clustering phenomenon, HH clustering of residential drive obvious role.
- (2) From the perspective of time evolution, under the influence of different neighborhood environments, the neighborhood with higher price level will increase the probability of upward transfer and inhibit the possibility of downward transfer, and the spatial convergence of clubs is presented overall. At the same time, under the short-term influence of the policy, the housing prices of low level and middle and low level are promoted in the same neighborhood environment, while the housing prices of high level and middle and high level are negatively affected.
- (3) Generally speaking, the policy has a positive regulation effect on the housing price of high level and high level type and promotes the transfer of low level and low level price type. Limited by the “old broken small, dirty and messy” living environment and the scarcity of development resources, the housing price in the old city of Hefei has a slow growth, but

compared with Shushan District, Binhu New Area and new station area, the housing price in the old city is still in the overall low value level for a long time, the growth power is insufficient. The government district and Binhu New Area, as new areas with policy-oriented resource input, always make clear the spatial layout of community living circle, gradually improve the future-oriented growth public service system, and provide public service guarantee for all ages, with strong community public service and green livable ability. As a result, housing prices have remained high for a long time.

## Data Availability

Sample data on second-hand home prices used to support the results of this study are included in the supplementary information document. These datasets were derived from the following public domain resources: <https://hf.anjuke.com/sale/?from=navigation> <https://map.baidu.com/@13057397.574469628,3719598.6899990723,12.3z>

## Conflicts of Interest

The authors have no conflicts of interest to declare.

## Acknowledgments

This study was supported by Humanities and Social Science Key Project of Education Department of Anhui Province (SK2019A0644).

## Supplementary Materials

Analyze data as attachment: Data.xlsx.(Supplementary Materials). (*Supplementary Materials*)

## References

- [1] P. Liu, *Research on the Spatial Evolution of Housingprices in Xi'an during the “13th Five-Year Plan” Period*, Northwestern university, Kirkland , Washington U.S, 2021.
- [2] *Bulletin of the Seventh National Population Census of Hefei City*, <http://tjj.hefei.gov.cn/public/14891/106487829.html>.
- [3] P. Wang and Z. B. Huang, “Study on the housing security of migrant populationb based on the new urbanization,” *Future and development*, vol. 39, no. 12, pp. 7–11+6, 2015.
- [4] Z. Yin, H. Zhuang, and J. Zhang, “Research on spatial differentiation and influencing factors of residential land price based on SDE model and kriging method—a case study of haidian district, beijing,” *Land and natural resources research*, vol. 190, no. 1, pp. 11–17, 2021.
- [5] S. Xu, Z. Zhang, and S. Zhang, “Spatial differentiation and influencing factors of second-hand housing prices: a case study of Binhu new district, hefei city, Anhui province, China,” *Journal of Mathematics*, vol. 2021, no. 6, pp. 1–8, 2021.
- [6] P. Liu, *A Study on the Spatial Evolution of Housing price in Xi'an during the 13th Five-Year Plan*, Northwest University, Kirkland , Washington, 2021.

- [7] J. Y. Gong, *Research on Spatial Differentiation Characteristics of Urban Residential Space Based on GIS*, Wuhan University, Hubei, China, 2018.
- [8] W. E. N. Li and H. Han, "Characteristics of commercial residential price spatial differentiation in Lanzhou city," *Science Surveying and Mapping*, vol. 43, no. 2, 2018.
- [9] J. Zhou and B. Jin, "Analysis for influencing factors of real estate price in Hefei based on spatial network auto-regressive transformation model," *Journal of university of Chinese Academy of Sciences*, vol. 37, no. 3, pp. 398–404, 2020.
- [10] Z. Zhang, X. Wang, and H. Li, "Spatial distribution of residential price in small and medium-sized cities and its influencing factors—a case study of Ganzhou City," *Science Surveying and Mapping*, vol. 45, no. 6, pp. 172–179, 2020.
- [11] Li. Yao, G. Gu, and J. Wang, "The spatial effect of building new housing in zhengzhou city—based on the spatial econometrics model," *Economic Geography*, vol. 34, no. 01, pp. 69–74+88, 2014.
- [12] J. Luo and Y. D. Wei, "A geostatistical modeling of urban land values in milwaukee, Wisconsin," *Annals of GIS*, vol. 10, no. 1, pp. 49–57, 2004.
- [13] L. Liu, J. Liu, and H. Qiao, "Real estate prices in China: an empirical study," *Of south fujian normal university (philosophy and social sciences edition)*, vol. 32, no. 4, pp. 67–74, 2018.
- [14] S. Holly, M. Hashem Pesaran, and T. Yamagata, "The spatial and temporal diffusion of house prices in the UK," *Journal of Urban Economics*, vol. 69, no. 1, 2010.
- [15] L. Xue and Z. Y. Ren, "The spatial-temporal dynamics of Guanzhong eco-security based on spatial Malcov Chains," *Journal of ecological environment*, vol. 20, no. 1, pp. 114–118, 2011.
- [16] L. Zhou and S. Xie, "Analysis of rural economic development level based on spatial Markov chain-Taking Sichuan province for example," *China Agricultural Resources and Zoning*, vol. 37, no. 12, pp. 186–191+208, 2016.
- [17] T. Yan, X. P. Zhang, H. Chen, and R. K. Li, "Evolution of regional differences in urban economic development in China from 2001 to 2016," *Economic Geography*, vol. 33, no. 12, pp. 11–20, 2019.
- [18] X. Liang, "Spatial Markov chain-based study on the spatial and temporal evolution of economic level in Guanzhong region," *Productivity Research*, vol. 29, no. 1, pp. 82–85, 2014.

## Research Article

# A Model for Surface Defect Detection of Industrial Products Based on Attention Augmentation

Gang Li <sup>1</sup>, Rui Shao <sup>1</sup>, Honglin Wan <sup>2</sup>, Mingle Zhou <sup>1</sup>, and Min Li <sup>1</sup>

<sup>1</sup>Shandong Computer Science Center, Qilu University of Technology (Shandong Academy of Sciences), Jinan 250014, China

<sup>2</sup>School of Physical and Electronic Sciences, Shandong Normal University, Jinan 250014, China

Correspondence should be addressed to Min Li; [lim@sdas.org](mailto:lim@sdas.org)

Received 30 March 2022; Accepted 26 April 2022; Published 14 May 2022

Academic Editor: Nian Zhang

Copyright © 2022 Gang Li et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Detecting product surface defects is an important issue in industrial scenarios. In the actual scene, the shooting angle and the distance between the industrial camera and the shooting object often vary, which results in a large variation in the scale and angle. In addition, high-speed cameras are prone to motion blur, which further deteriorates the defect detection results. In order to solve the above problems, this study proposes a surface defect detection model for industrial products based on attention enhancement. The network takes advantage of the lower-level and higher-resolution feature map from the backbone to improve Path Aggregation Network (PANet) in object detection. This study makes full use of multihead self-attention (MHSA), an independent attention block for enhancing the backbone network, which has made considerable progress for practical application in industry and further improvement of the surface defect detection. Moreover, some tricks have been adopted that can improve the detection performance, such as data augmentation, grayscale filling, and channel conversion of input images. Experiments in this study on internal datasets and four public datasets demonstrate that our model has achieved good performance in industrial scenarios. On the internal dataset, the mAP@.5 result of our model is 98.52%. In the RSDDs dataset, the model in this study achieves 86.74%. In the BSData dataset, the model reaches 82.00%. Meanwhile, it achieves 81.09% and 74.67% on the NRSD-MN and NEU-DET datasets, respectively. This study has demonstrated the effectiveness and certain generalization ability of the model from internal datasets and public datasets.

## 1. Introduction

The detection of industrial products has always received a great deal of attention in computer vision. Most traditional methods rely on manual parameter setting, which is not conducive to higher detection accuracy and speed. Object detection based on convolutional neural networks has significantly progressed in recent years. Some well-known benchmark datasets, including MS COCO [1] and PASCAL VOC [2], have promoted the development of object detection applications. However, most of the object detection tasks are designed for images of natural scenes. There are three problems caused by these models for object detection in industrial scenes. First, with large variations of orientation and position during the shooting process with the industrial camera, the shape of the object significantly changes. Second,

the images are captured at high speed, thus producing blurred objects. Third, images captured by industrial cameras often contain cluttered background because the targeted area is larger than the object area. For instance, a mirror image of the object produced by the object's background has interfered with the detection of the object, as shown in Figure 1. The above detection in industrial scenes is a challenging task. This study presents a targeted detection network in industrial settings. In this study, we come up with an attention-augmented object detection network, which improves the surface defect detection of existing industrial products. The network contains three key components, including the backbone, neck, and head. Referring to the backbone, it is composed of five convolutional groups and an independent attention block. Additionally, more feature information can be attained with the help of augmentation with independent

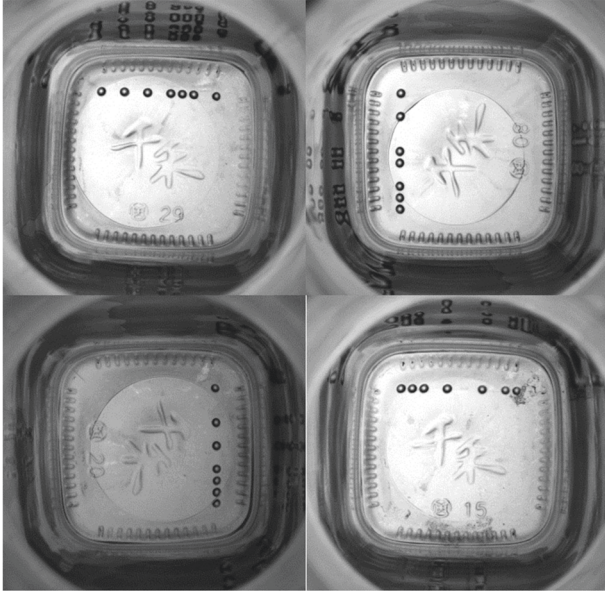


FIGURE 1: The mold points at the bottom of the glass bottle.

attention blocks, which serves as an innovation for the backbone. In the neck, the SPP [3] structure and the PANet [4] structure are used to make the network more suitable for the detection of industrial defects. An FDS Block adopted in this study (fast downsampling block) is designed to quickly downsample the high-resolution feature in the backbone to concatenate well with the feature information in the neck.

The contributions of this study are listed as follows:

- (1) This study aggregates image feature information of lower level and high resolution in the backbone network into PANet, making the model collect richer feature information and improve feature extraction capabilities.
- (2) This study integrates the multihead self-attention (MHSA) mechanism into the detection network's backbone, which can advance the feature extraction ability and pay more attention to information.
- (3) This study provides a fast downsampling block (FDS Block) for the high-resolution images of the backbone network that quickly reduces the resolution and simultaneously increases the number of feature map channels. Based on that, high-resolution images in the backbone can be used to connect information with the highest-level, lowest-resolution feature maps in PANet as soon as possible.
- (4) We validate the effectiveness of our module through extensive ablation studies.
- (5) We propose a surface defect detection method for industrial products based on attention augmentation, which can perform well in industrial scenarios.

The structure of this study is organized as follows. Related works and the proposed method are described in detail, respectively, in Sections 2 and 3. Sections 4 and 5 tell the experimental results and conclusions of the study.

## 2. Related Works

**2.1. Convolutional Neural Network.** Since AlexNet [5] won the ImageNet competition in 2012, more and more convolutional neural networks have been proposed. The VGG [6] network won second place in the 2014 ILSVRC competition. On the basis of AlexNet, it has been greatly improved, that is to say, multiple smaller convolution kernels can replace the receptive field of large convolution kernels. ResNet [7] has shown that a network with residual blocks can expand the network depth to 101 layers. ResNeXt [8] proposed a separable convolution between the depths of the common convolution kernels, which makes a balance between the two strategies by controlling the number of groups. DenseNet [9] widened the network structure. DarkNet53 [10] and CSPDarkNet53 [11] are also proposed as popular methods.

**2.2. Generic Object Detector.** Modern detectors usually consist of two stages, a backbone pretrained on ImageNet and a head for predicting object classes and bounding boxes. The most representative two-stage object detector is the R-CNN [12] family, including Fast R-CNN [13], Faster R-CNN [14], and R-FCN [15]. The most representative one-stage object detectors are the YOLO family, involving YOLOv1 [16], YOLOv2 [17], YOLOv3 [10], and YOLOv4 [11]. At the same time, SSD [18] and RetinaNet [19] are one-stage object detectors. In recent years, anchor-free object detectors have developed. Such detectors include CenterNet [20], RepPoints [21], FCOS [22], and so on. Recently, object detectors have often constructed some layers (necks) between the head and the backbone, and these layers usually collect feature maps at different stages. The neck can generally cover multiple bottom-up paths and multiple top-down paths. Networks with this mechanism include the feature pyramid network (FPN) [23] and Path Aggregation Network (PANet).

**2.3. Object Detection Effective Strategies.** Data augmentation expands the dataset and makes the model more robust among datasets with different environments. Well-known data augmentation methods include MixUp [24], CutMix [25], and Mosaic [11]. MixUp randomly selects two samples from the training image for random weighted summation, which is in line with the labels of the samples. Unlike occlusion, which typically uses a zero-pixel "black cloth" to occlude an image, CutMix resorts to another image area to cover the occlusion. Mosaic is an improved version of CutMix, stitching the four images to greatly enrich the background of the detected object. Other data augmentation methods include DropBlock [26], class label smoothing [11], Cross mini-Batch Normalization (CmBN) [11], CIoU loss [27], DIoU loss [27], and mesh sensitivity elimination [11]. With multiple anchors, mesh sensitivity elimination will obtain a single ground truth, cosine annealing schedule, optimal hyperparameters, and random training shape. BOS (Bag of Specials), such as Mish activation function [11], Cross-Stage Partial Connection (CSP [28]), SPP, and PANet,

can significantly heighten the accuracy of object detection with only a small increase in inference cost.

**2.4. Attention Mechanism.** In response to the defects of Seq2Seq [29], Bahdanau [30] proposed an attention mechanism to achieve a soft distinction and provide some visual effects of attention. Luong et al. [31] have put forward two improved versions of attention, such as global attention and local attention. Ahmed et al. [32] came up with a novel network structure transformer, which contains an attention mechanism called self-attention. Liu et al. [33] presented the gated multilingual attention (GMLATT) framework to solve such problems as data sparseness and monolingual ambiguity, using multilingual information combined with the attention mechanism to complete the task. Since the attention mechanism extracted key features by the weighted calculation of all local features and ignored the strong correlation between local features, there was robust information redundancy between features. In order to solve this problem, researchers from Meitu Cloud Vision and the Institute of Automation, Chinese Academy of Sciences, drew on the idea of PCA (principal component analysis) and proposed a self-attention model [34] that introduced the interactive perception of local features and combined them with the model embedded in a CNN network. The algorithm performs extremely well on behavior classification of multiple academic datasets and Meitu's internal industrial video dataset. Google released BoTNet [35], replacing the bottleneck of the fourth block in ResNet with the MHSA (multihead self-attention) module and forming a new module named Bottleneck Transformer (BoT). The final network structure, including blocks like BoT, is called BoTNet.

**2.5. Industrial Defects.** Zhang and Song [36] proposed a segmentation network to improve NRSD segmentation, which applied an attention mechanism to optimize the extracted information and performed well for both artificial and natural NRSDs. Niu and Song [37] proposed an unsupervised stereo saliency detection method based on a binocular line scanning system, which can simultaneously obtain high-precision image and contour information. He et al. [38] proposed a novel defect detection system based on deep learning and focused on steel plate defect detection, which uses a multilevel feature fusion network (MFN) to focus on multilevel features. Wu et al. [39] developed a more flexible deep learning method for industrial defect detection, and the author proposed a unified framework for detecting industrial products or flat surface defects. Xu et al. [40] established a defect detection network (D4Net) to detect deformed defects in a given image and its corresponding reference image.

### 3. Method

**3.1. Network Overview.** The proposed model includes three parts: backbone, neck, and head. The backbone covers CSPDarknet53 and an attention-enhancing structure

(purple block in the backbone in Figure 2). The neck part applies the PANet as the main part. In addition, this study decreases the resolution of the information on the  $104 \times 104$  feature map (blue block in Figure 2 backbone) and fuses it with the feature map in PANet to form the red block in Figure 2 neck, which is input to the third detection head. This study adopts a series of strategies, i.e., data augmentation, grayscale filling, and automatic conversion of images to three-channel RGB, making the model more robust.

**3.2. Structure.** In Figure 3, this study gives the most detailed explanation of the model, which is divided into two parts, A and B. Part A explains the backbone of the model, and Part B interprets the neck, SPP Block (spatial pyramid pooling block), and the FDS Block (fast downsampling block) that are connected with the neck by the trunk part.

**3.2.1. Attention-Enhanced Backbone.** The backbone used in this study is CSPDarkNet53, as shown in Figure 3, and it is divided into five parts. Block1, Block2, Block3, Block4, and Block5 (details are shown in Figure 4). To meet the actual needs of industrial product defect detection and strengthen the feature extraction ability of the model to focus on more information, this study resorts to an attention enhancement strategy for the backbone.

This study uses MHSA to enhance CSPDarknet53. Transformers based on the self-attention mechanism are first applied in the NLP domain. In order to make the CNN backbone network with such characteristics, an effective method is to replace the spatial convolution layer in CNN with the MHSA proposed in transformer. As for our method, instead of replacing the convolution of the last residual layer with an MHSA layer like BoTNet, we choose the MHSA structure as the entire attention block before Block5 of the backbone network. In this study, the input resolution and output resolution of the attention block, the number of input channels, and the number of output channels are not changed. The attention block is shown in Figure 5.

In addition, this study adopts some detection enhancement strategies for the backbone, including the use of the Mish activation function and the Mosaic data enhancement method.

**3.2.2. Neck.** As shown in Figure 3, the high-resolution feature map, SPP structure, and PANet structure in the backbone include the neck, making the network more suitable for detecting industrial product defects. The neck takes three feature maps of different output sizes from the backbone as input. The  $13 \times 13$  size feature map passes into the SPP structure after three convolutions. Then, the SPP structure is formed by the max-pooling layer from three different kernel sizes of 5, 9, and 13. In addition, the SPP structure also includes one shortcut layer. After the SPP structure, the data go to the PANet structure. The  $26 \times 26$  and  $52 \times 52$  feature maps are directly fed into the PANet



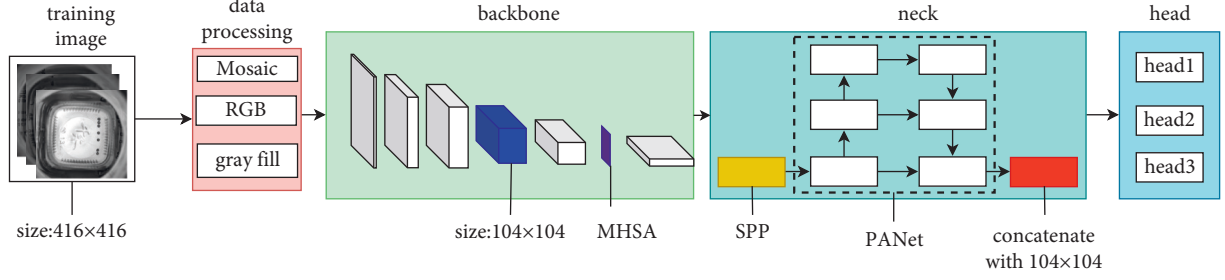


FIGURE 2: An overview of the detection process. Among them, the purple feature blocks in the backbone are attention enhancement structures.

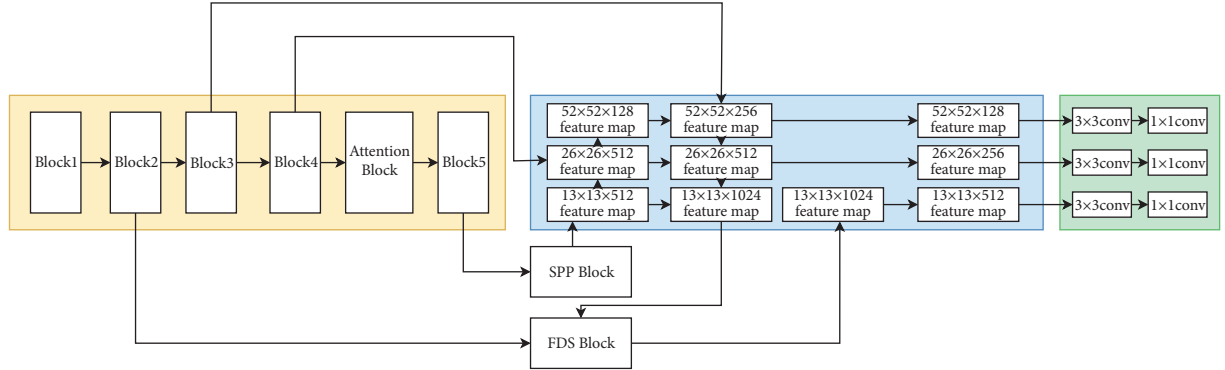


FIGURE 3: In the model structure diagram of this study, the yellow background is the backbone, the blue background is the neck, and the green background is the head.

network after three convolutions. The SPP structure is shown in Figure 6.

The PANet structure combines high-level, low-resolution feature maps with low-level, high-resolution feature maps bottom-up. Then, it connects the low-level, high-resolution feature maps with high-level, low-resolution feature maps from top to bottom. In this study, according to the experiments, the  $104 \times 104$  feature map has the function of delivering rich information to the highest and lowest resolution layers. Therefore, the  $104 \times 104 \times 128$  feature map and the  $13 \times 13 \times 1024$  feature map in the PANet will be input to the FDS Block, and then, the  $13 \times 13 \times 1024$  feature map will be accordingly output. In doing so, we can not only collect rich feature information involved in the backbone, but also the output does not change the feature map size of PANet. The structure of the FDS Block structure is shown in Figure 7.

**3.3. NMS of This Study.** In this study, DIOU-NMS is used to obtain the most suitable detection box for each object, thereby improving the discriminative ability of the model in this study in the detection of surface defects of industrial products. The mathematical formula of DIOU-NMS is defined as follows:

$$S_i = \begin{cases} S_i, & \text{IOU} - R(M, b_i) < N, \\ 0, & \text{IOU} - R(M, b_i) \geq N, \end{cases} \quad (1)$$

where  $b_i$  is removed by considering both the IoU and the distance between the center points of the two boxes,  $s_i$  is the

classification score, and  $N$  is the NMS threshold, where the mathematical formula for  $R$  is defined as follows:

$$R = \frac{\rho^2(b, b^{gt})}{c^2}, \quad (2)$$

where  $\rho$  is the distance,  $b$  and  $b^{gt}$  represent the two boxes, and  $c$  is the diagonal length of the smallest box containing the two boxes.

**3.4. Loss Function.** In this study, the expression of the loss function is as follows:

$$\begin{aligned} L = & - \sum_{i=0}^{S^2} 1_{ij}^{obj} \sum_{c \in N_{classes}} [\hat{p}_i(c) \log(p_i(c)) + (1 - \hat{p}_i(c)) \log(1 - \hat{p}_i(c))] \\ & + \lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^B 1_{ij}^{obj} [(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2] + \\ & \lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^B 1_{ij}^{obj} \left[ (\sqrt{w_i} - \sqrt{\hat{w}_i})^2 + (\sqrt{h_i} - \sqrt{\hat{h}_i})^2 \right] \\ & - \sum_{i=0}^{S^2} \sum_{j=0}^B 1_{ij}^{obj} [\hat{C}_i \log(C_i) + (1 - \hat{C}_i) \log(1 - C_i)] \\ & - \lambda_{noobj} \sum_{i=0}^{S^2} \sum_{j=0}^B 1_{ij}^{obj} [\hat{C}_i \log(C_i) + (1 - \hat{C}_i) \log(1 - C_i)]. \end{aligned} \quad (3)$$

Overall, the first line of formulas is classification loss, the second and third lines of formulas are box regression loss, and the fourth and fifth lines of formulas are confidence loss.

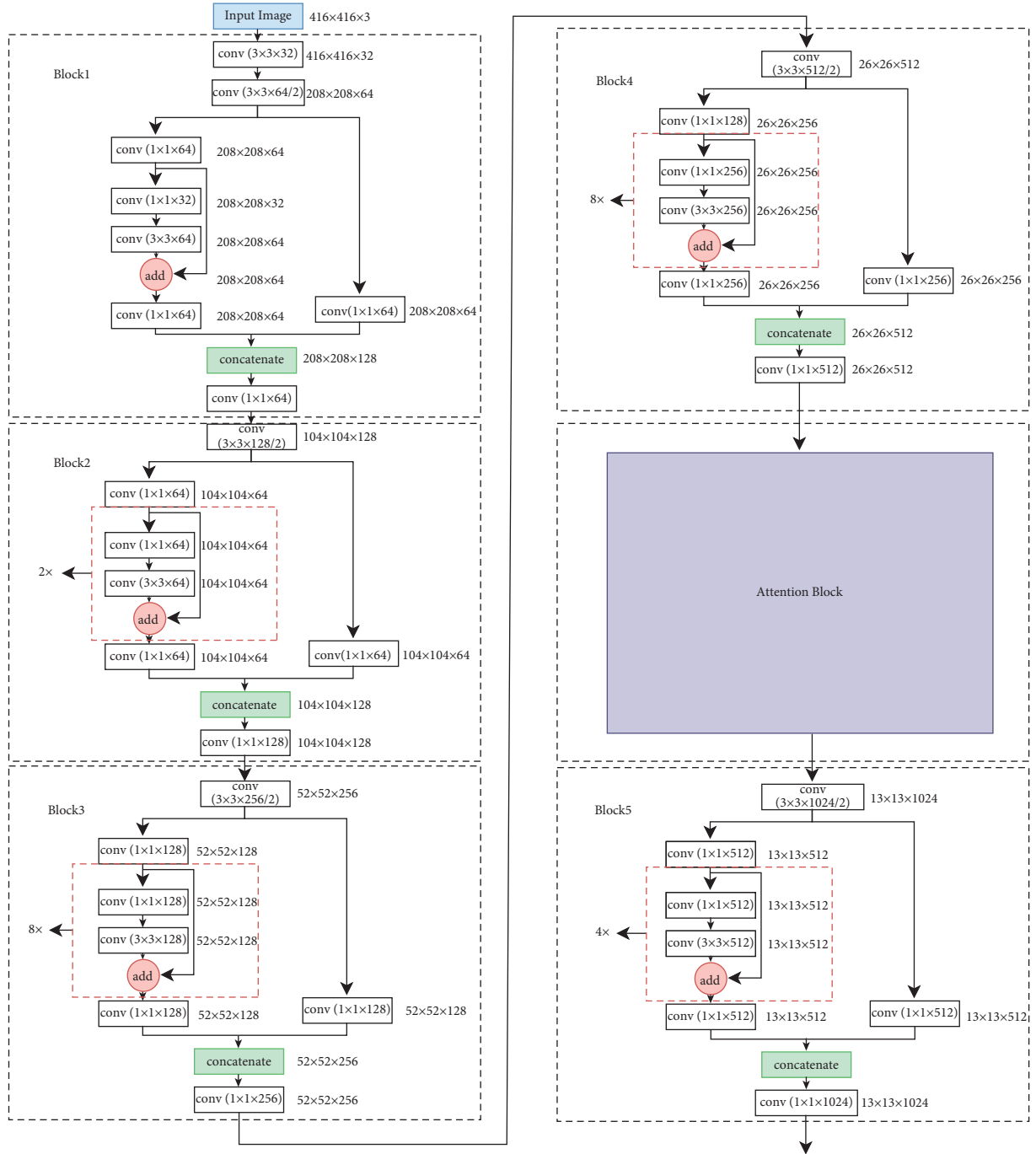


FIGURE 4: Details of blocks 1 to 5.

## 4. Experimental Results

### 4.1. Experiment Description

**4.1.1. Dataset.** Five datasets are exerted in this study, that is, four public datasets (RSDDs [41], BSDData [42], NRSD-MN [36], and NEU-DET [38]) and one internal dataset for comparison with existing methods. The internal dataset is the image of the mold point at the bottom of the glass bottle, which is collected and saved from the actual production line with a CCD camera. The image resolution is fixed at

$800 \times 780$ . In addition, each dataset is divided into training, validation, and testing, with an amount ratio of 5:2:3. All object detection images refer to different colors to represent the corresponding types. RSDDs (Railway Surface Defects Dataset) contains two types of datasets. The first one is the type I captured from the fast lane, covering 67 challenging images. The second is a class II captured from normal/heavy-transport tracks, containing 128 challenging images. Each image from both datasets includes at least one defect with complex and noisy backgrounds. Referring to the experiments in this study, object detection is performed on the

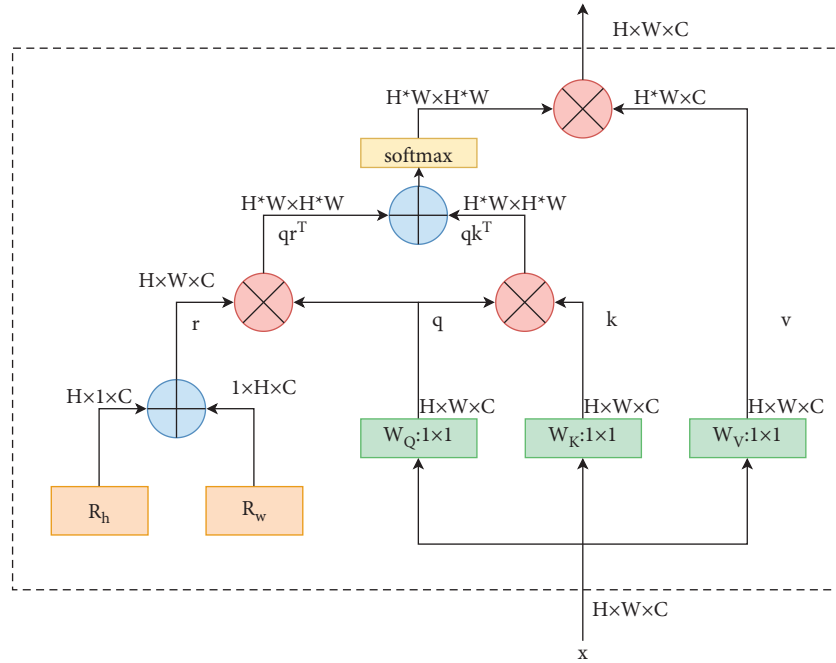


FIGURE 5: Attention block, where  $q$ ,  $k$ ,  $v$  and  $r$  refer to query, key, value, and position encoding, and  $R_h$  and  $R_w$  refer to height and height relative position encoding width.

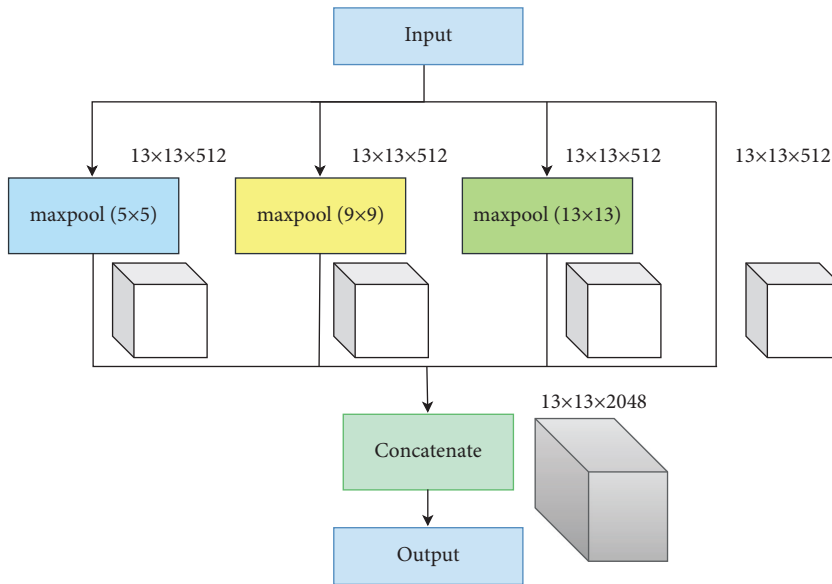


FIGURE 6: SPP structure.

type I dataset with a resolution of  $160 \times 1000$ . Thirty-two images are selected in this study as the training set, 14 as the validation set during training, and 21 as the test set. Figure 8 shows three examples of images and defects. The image size of BSData is  $1130 \times 460$ . This study chooses a subset of 394 defect images, of which 192 for training, 83 for training validation, and 119 for testing. Furthermore, Figure 9 shows some example pictures. The internal dataset of this study contains 481 training images, 207 validation images during training, and 295 testing images. Example images are shown in Figure 1. NEU-DET dataset is a defect classification data

set. There are six types of defects from hot-rolled steel plates, including crazing, inclusion, patches, pitted surface, rolled-in scales, and scratches. The dataset has a total of 1800 images. This study selects 1260 images as the training set and 540 as the test set. Example images are shown in Figure 10. The NRSD-MN dataset contains 4101 images, including 3936 man-made NRSD images and 165 natural NRSD images. In this study, 4101 images are selected as our training and test sets, compared with the state-of-the-art algorithm. Among them, we take 2971 images as training set and 1130 images as test set. Example images are shown in Figure 11.

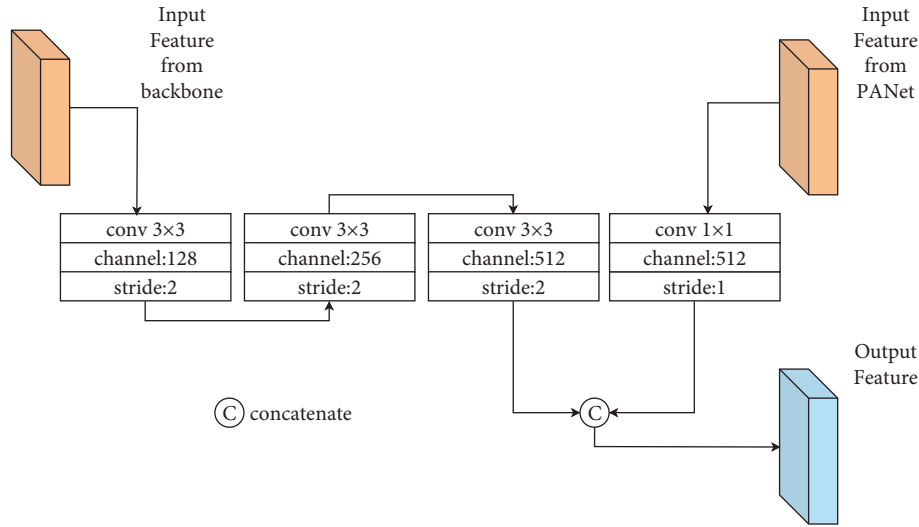


FIGURE 7: Structure of FDS block.

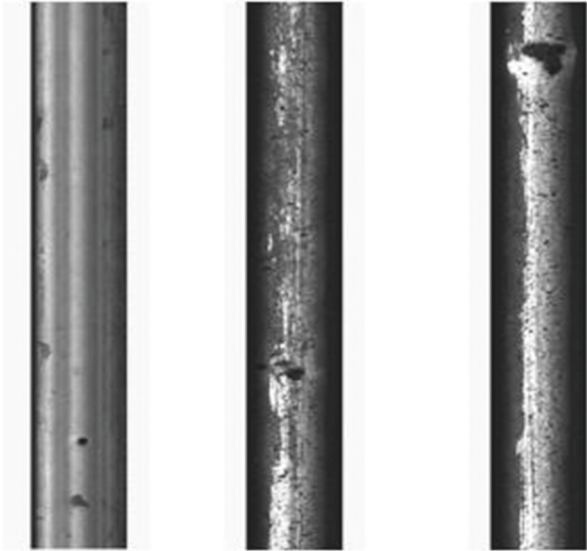


FIGURE 8: Example image of the RSDDs dataset.

**4.1.2. Implementation.** This study implements the model described in this study on PyTorch 1.9.0. The computing performance of industrial computers is strong or weak. In order to make the model have good detection ability on computers with different performances, this study trains and tests the model on GPUs with different performances. Ablation experiments are performed on NVIDIA RTX A6000 for the internal and RSDDs dataset, ablation experiments on NEU-DET and NRSD-MN datasets are performed on NVIDIA RTX 3080, and ablation experiments for the BSDData dataset are performed on NVIDIA Tesla k80. In the model in this study, after training and testing images entered into the network, the resolution is unified to  $416 \times 416$ . The batch size is 8 on NVIDIA RTX A6000, 8 on NVIDIA RTX 3080, 4 on NVIDIA Tesla k80, and 2 on NVIDIA RTX 3060. This study has two parts during training, which are divided into the freeze training part and the unfreeze training part. Freeze training sees that the

backbone is frozen and sees the unchanged feature extraction network. The initial learning rate becomes 0.001 during freeze training and 0.0001 after freeze training, and the learning rate decay is annealed cosine.

**4.1.3. Performance Metrics for Object Detection.** The method proposed in this study provides defect localization and defect type classification for defect objects.

The IoU (intersection over union) measures the degree of overlap between two regions, which is the ratio of the overlapping area of the two regions to the total area (the overlapping part is only calculated once). As for the object detection task, the model output is considered to be correct till the IoU value of the rectangular box output by the model and the rectangular box manually marked is greater than a certain threshold.

Precision and recall can be expressed as follows:

$$\text{Precision} = \frac{TP}{TP + FP}, \quad (4)$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

As the name indicates, AP means the average precision. Simply put, it is the average of the precision values on the PR curve. For the PR curve, this study employs the integral to calculate it.

$$AP = \int_0^1 p(r) dr. \quad (5)$$

The mAP is a general model evaluation criterion in the field of object detection. The object detection in this study is used for detecting multiple objects. Therefore, this study can calculate the mAP.

The mAP can be expressed as follows:

$$mAP = \frac{\sum_{n=1}^N \int_0^1 p(r) dr}{N}. \quad (6)$$

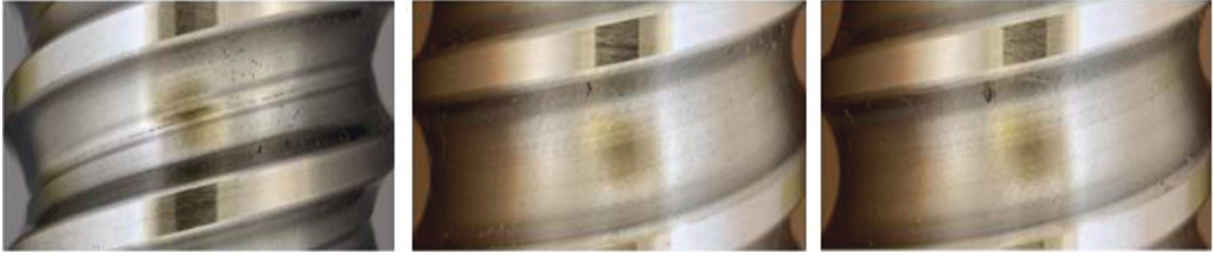


FIGURE 9: Example image of the BSData dataset.

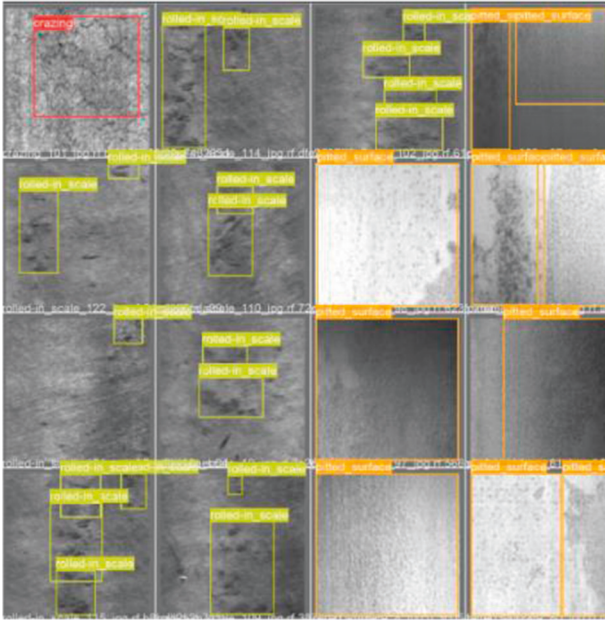


FIGURE 10: Example image of the NEU-DET dataset.

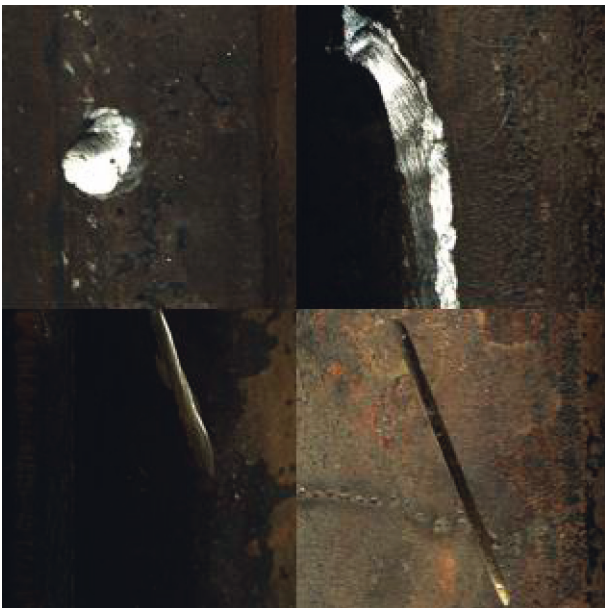


FIGURE 11: Example image of the NRSD-MN dataset.

**4.2. Results.** This study uses the test set of the internal dataset, the test set of the RSDDs dataset, the test set of the BSData dataset, the test of the NEU-DET dataset, and the test of the NRSD-MN dataset to evaluate our model and report mAP (using the metric mAP@.5 of the PASCAL VOC dataset).

In the end, this study achieves a good score of 98.52 on the internal dataset, which is higher than the state-of-the-art model yolov5x highest score of 98.45 on the internal dataset. Moreover, a good score of 86.74 on the RSDDs test dataset is performed, higher than the yolov5x highest score of 85.45 on the RSDDs dataset. It also gets a good score of 82.00 on the BSData test dataset, which is higher than the highest score of yolov5x on the BSData dataset of 81.79. Subsequently, in the NEU-DET dataset and the NRSD-MN dataset, our model achieved 74.67 and 81.09, respectively. As listed in Table 1, the scores of our model can compare their scores among YOLOV3, YOLOV4, YOLOv5l, and YOLOv5x algorithms. Figure 12 shows the comparison of mAP in the five datasets.

The comparative experiments on the internal dataset in this study are performed on NVIDIA RTX 3060 with batch size set to 2. This study conducts freeze training for 100 epochs followed by 100 epochs after unfreezing. The comparative experiments in Table 1 demonstrate the detection ability of the model in the internal dataset.

The comparative experiments on the RSDDs dataset are conducted on NVIDIA RTX 3060 with batch size set to 2. This study makes freeze training for 100 epochs and continues training for 100 epochs after unfreezing. The comparative experiments in Table 1 show the detection ability of the model in the RSDDs dataset.

The comparative experiments on the BSData dataset are made on NVIDIA RTX 3060 with batch size set to 2. In order to verify the ability of the model and fit the data faster and better, this study only carries out freeze training for 50 epochs. Only 50 epochs will be continuously trained after unfreezing. The comparative experiments in Table 1 demonstrate the detection ability of the model in this study in the BSData dataset.

The comparative experiments on the NEU-DET dataset and the NRSD-MN dataset are made on NVIDIA RTX 3080 with batch size 8. In order to verify the ability of the model and fit the data faster and better, this study only carries out freeze training for 200 epochs. The comparative experiments in Table 1 demonstrate the detection ability of the model in this study in the NEU-DET dataset and the NRSD-MN dataset.

TABLE 1: Comparison with state-of-the-art methods in three different datasets.

Methods	(Internal) mAP	(RSDDs) mAP	(BSData) mAP	(NEU-DET) mAP	(NRSD-MN) mAP
YOLOv3	91	76.60	77.63	54.7	63.09
YOLOv4	95.97	79.27	81.38	70.35	75.4
YOLOv5l	95.53	80.43	80.39	72.7	78.71
YOLOv5x	98.45	85.45	81.79	74.07	80.15
Our method	98.52	86.74	82.00	74.67	81.09

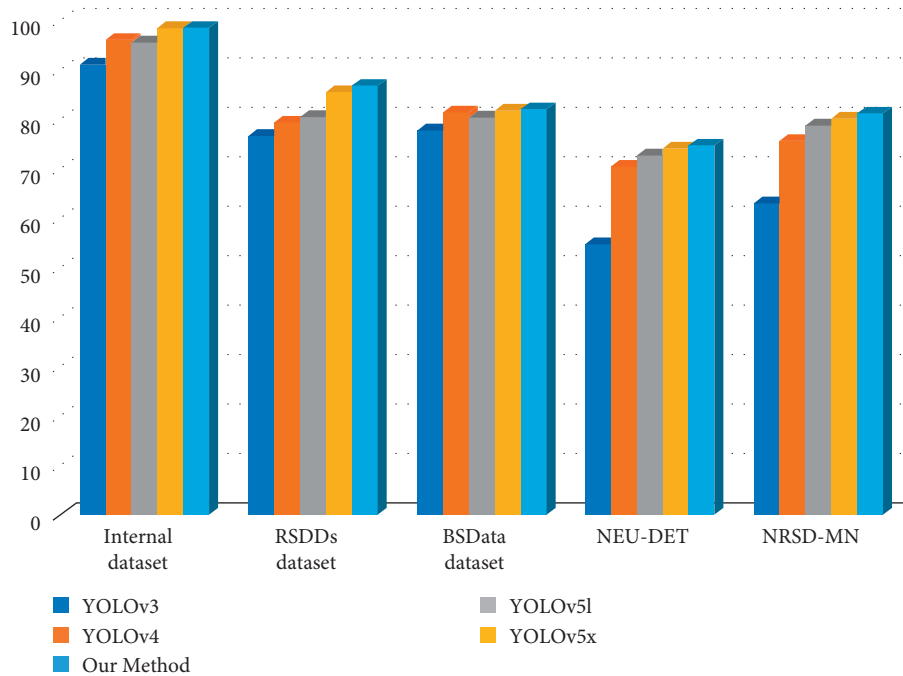


FIGURE 12: Comparison of mAP in five datasets.

### 4.3. Ablation

**4.3.1. On the Internal Dataset.** This study utilizes a test set of internal datasets (including images of glass bottle bottom mold points collected in actual industrial production) to evaluate our model and report mAP (including the metric mAP@.5 of the PASCAL VOC dataset). Table 2 recites the scores of the models.

The ablation experiments in Table 2 exhibit the detection ability of each improved module of the proposed model in the internal dataset. With attempts to advance the use of higher-resolution images ( $832 \times 832$ ), the fitting ability does not meet expectations, followed by an increased amount of computation by two times. Although this study tries to use  $104 \times 104$  feature information to add to P3 and pass the result to the detection head, the effect is far from expectations. Therefore, this study no longer makes an attempt on  $832 \times 832$  images in the subsequent experiments nor does it add  $104 \times 104$  feature information to P3 in the PANnet structure.

**Effects of Attention Blocks.** The addition of an attention block increases the amount of computation, but mAP improves very well. It can be seen from Table 2 that the CSPDarkNet53+ attention block performs well in object

TABLE 2: Ablation studies on the internal dataset.

Number	Methods	mAP
A	CSPDarkNet53 + SPP + PANet	95.97
B	A + $832 \times 832$ resolution	95.33 ( $\downarrow 0.64$ )
C	A + $104 \times 104$ feature to P3	95.75 ( $\downarrow 0.22$ )
D	A + FDS block	96.97 ( $\uparrow 1.00$ )
E	A + attention block	97.42 ( $\uparrow 1.45$ )
F	Our method (D + E)	98.52 ( $\uparrow 2.55$ )

detection, with an increase of 1.45%. The introduction of attention block is worthwhile.

**Effects of FDS Block.** Using lower-layer, higher-resolution feature maps for fusion improves mAP and does not have a large impact on computation and inference speed. It also plays a certain role in detecting dense and large objects. As shown in Table 2, by adding  $104 \times 104$  feature information to the P5 layer on PANet, mAP increases by 1%.

**Aggregation Effect.** This study lists the mAP of all the results of the ablation experiment. It is found that adding all innovation points to the model at the same time benefits the most, and mAP achieves the best results.

**4.3.2. On the RSDDs Dataset.** This study uses the test set of the RSDDs dataset to evaluate our model and report mAP (using the metric mAP@.5 of the PASCAL VOC dataset). Table 3 lists the scores of the models.

The ablation experiments in Table 3 demonstrate the detection ability of each improved module in the RSDDs dataset.

**Impact of Attention Block.** It can be seen from Table 3 that the CSPDarkNet53+ attention block performs well in object detection, with an increase of 1.33%. The introduction of attention block is worthwhile.

**The Effect of FDS Block.** As shown in Table 3, by adding  $104 \times 104$  feature information to the P5 layer in PANet, mAP increases by 2.11%.

**Aggregation Effect.** This study enumerates the mAP of all the results of the ablation experiment. It is found that adding all innovation points to the model simultaneously benefits the most, and mAP gets the best results. As shown in Table 3, mAP increases by 7.47%.

**4.3.3. On the BSData Dataset.** This study uses the test set of the BSData dataset to evaluate our model and report mAP (using the metric mAP@.5 of the PASCAL VOC dataset). Table 4 lists the scores of the models.

The ablation experiments in Table 4 demonstrate the detection ability of each improved module of the model on the BSData dataset.

**Effects of Attention Blocks.** As Table 4 shows, the CSPDarkNet53+ attention block performs well in object detection, with an increase of 0.21%. So it is well worth making an introduction to the attention block.

**Effects of FDS Block.** As shown in Table 4, by adding  $104 \times 104$  feature information to the P5 layer on PANet, mAP increases by 0.41%.

**Aggregation Effect.** This study shows the mAP of all the results of the ablation experiment. We find that adding all innovation points to the model at the same time behaves well, and mAP achieves the best results. As shown in Table 4, mAP has increased by 0.62%.

In order to test the fast fitting ability of the model, this study only performs freeze training for 50 epochs. In addition, 50 epochs will be carried out after thawing, which improves the detection effect. This study believes that the model will have better detection results in light of more sufficient computing conditions.

**4.3.4. On the NEU-DET Dataset.** This study uses the test set of the NEU-DET dataset to evaluate our model and report mAP (using the metric mAP@.5 of the PASCAL VOC dataset). Table 5 lists the scores of the models.

The ablation experiments in Table 5 demonstrate the detection ability of each improved module of the model on the NEU-DET dataset.

**Effects of Attention Blocks.** As Table 5 shows, the CSPDarkNet53+ attention block performs well in object detection, with an increase of 1.73%. So it is well worth making an introduction to the attention block.

TABLE 3: Ablation studies on the RSDDs dataset.

Number	Methods	mAP
A	CSPDarkNet53 + SPP + PANet	79.27
B	A + attention block	80.6 (↑1.33)
C	A + FDS block	81.38 (↑2.11)
D	Our method (B + C)	86.74 (↑7.47)

TABLE 4: Ablation experiments on the BSData dataset.

Number	Methods	mAP
A	CSPDarkNet53 + SPP + PANet	81.38
B	A + attention block	81.59 (↑0.21)
C	A + FDS block	81.79 (↑0.41)
D	Our method (B + C)	82.00 (↑0.62)

TABLE 5: Ablation experiments on the NEU-DET dataset.

Number	Methods	mAP
A	CSPDarkNet53 + SPP + PANet	70.35
B	A + attention block	72.08 (↑1.73)
C	A + FDS block	71.22 (↑0.87)
D	Our method (B + C)	74.67 (↑4.32)

TABLE 6: Ablation experiments on the NRSD-MN dataset.

Number	Methods	mAP
A	CSPDarkNet53 + SPP + PANet	75.4
B	A + attention block	77.91 (↑2.51)
C	A + FDS block	76.45 (↑1.05)
D	Our method (B + C)	81.09 (↑5.69)

**Effects of FDS Block.** As shown in Table 5, by adding  $104 \times 104$  feature information to the P5 layer on PANet, mAP increases by 0.87%.

**Aggregation Effect.** This study shows the mAP of all the results of the ablation experiment. We find that adding all innovation points to the model at the same time behaves well, and mAP achieves the best results. As shown in Table 5, mAP has increased by 4.32%.

**4.3.5. On the NRSD-MN Dataset.** This study uses the test set of the NRSD-MN dataset to evaluate our model and report mAP (using the metric mAP@.5 of the PASCAL VOC dataset). Table 6 lists the scores of the models.

The ablation experiments in Table 6 demonstrate the detection ability of each improved module of the model on the NRSD-MN dataset.

**Effects of Attention Blocks.** As Table 6 shows, the CSPDarkNet53+ attention block performs well in object detection, with an increase of 2.51%. So it is well worth making an introduction to the attention block.

**Effects of FDS Block.** As shown in Table 6, by adding  $104 \times 104$  feature information to the P5 layer on PANet, mAP increases by 1.05%.

**Aggregation Effect.** This study shows the mAP of all the results of the ablation experiment. We find that adding all

innovation points to the model at the same time behaves well, and mAP achieves the best results. As shown in Table 6, mAP has increased by 5.69%.

## 5. Conclusions

This study proposes a surface defect object detector for industrial products, which is especially good at detecting product surface defects in industrial scenarios. Experiments on an internal dataset and four public datasets (RSDDs, BSDData, NRSD-MN, and NEU-DET) have been carried out. Experiments show that the model in this study has good performance. This study argues that with rich computing resources, a longer training time can be used to allow the model to get better detection results. On the basis of previous research work, this study plans to further improve the object detection network structure to achieve better industrial detection performance in the future, including higher accuracy, faster test speed, and better prediction stability. At the same time, in the face of more difficult and deeper defect detection, on the one hand, this study intends to use camouflaged object detection to conduct experiments and research. On the other hand, it is committed to helping developers and researchers analyze and process scenes captured in industrial machine vision.

## Data Availability

The dataset can be accessed upon request to the corresponding author.

## Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

## Acknowledgments

This study was funded by the Key R&D Plan of Shandong Province (Soft Science Project) (2021RZA01024), and the Plan of Youth Innovation Team Development of Colleges and Universities in Shandong Province (SD2019-161).

## References

- [1] T.-Y. Lin, M. Maire, S. Belongie et al., "Microsoft COCO: Common Objects in Context," in *Proceedings of the 13th European Conference, ECCV (2014)*, Zurich, Switzerland, September 2014.
- [2] M. Everingham and L. Van Gool, "The pascal visual object classes (VOC) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2009.
- [3] K. He and X. S. J. Zhang, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.
- [4] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, June-2018.
- [5] Ts. Technicolor, A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Advances in neural information processing systems*, vol. 25, 2012.
- [6] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Computer Science*, vol. 6, 2014.
- [7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, San Juan, PR, USA, June 2016.
- [8] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated Residual Transformations for Deep Neural Networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, IEEE, Honolulu, July 2016.
- [9] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, July 2016.
- [10] J. Redmon and A. Farhadi, "YOLOv3: an incremental improvement," 2018, <https://arxiv.org/abs/1804.02767>.
- [11] A. Bochkovskiy, C. Y. Wang, and H. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection," 2020, <https://arxiv.org/abs/2004.10934>.
- [12] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Portland, OR, USA, June 2013.
- [13] R. Girshick, "Fast R-CNN," in *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1440–1448, Cambridge, MA, USA, June 2015.
- [14] K. He and K. R. Girshick, "Faster R-CNN: towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.
- [15] J. Dai, Y. Li, K. He, and J. Sun, "Object Detection via Region-Based Fully Convolutional Networks," 2016, <https://arxiv.org/abs/1605.06409>.
- [16] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look once: Unified Real-Time Object Detection," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 779–788, Las Vegas, NV, USA, June 2016.
- [17] J. Redmon and F. Ali, "YOLO9000: better, faster, stronger," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6517–6525, Honolulu, HI, USA, July 2017.
- [18] W. Liu, D. Anguelov, D. Erhan et al., "SSD: Single Shot MultiBox Detector," in *Proceedings of the 14th European Conference ECCV (2016)*, Amsterdam, Netherlands, October 2016.
- [19] T.-Y. Lin and P. Goyal, "Focal loss for dense object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 318–327, 2020.
- [20] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "CenterNet: Keypoint Triplets for Object Detection," in *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV) (2019)*, pp. 6568–6577, Seoul, Korea, June 2019.
- [21] Z. Yang, S. Liu, H. Hu, L. Wang, and S. Lin, "RepPoints: Point Set Representation for Object Detection," in *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9656–9665, Seoul, Korea, June 2019.



- [22] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully Convolutional One-Stage Object Detection," in *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9626–9635, Seoul, Korea, July 2019.
- [23] T.-Yi Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 936–944, Honolulu, HI, USA, July 2017.
- [24] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "Mixup: Beyond Empirical Risk Minimization," 2018, <https://arxiv.org/abs/1710.09412>.
- [25] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features," in *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6022–6031, Seoul, Korea, July 2019.
- [26] G. Ghiasi, T. Y. Lin, and Q. V. Le, "DropBlock: A regularization method for convolutional networks," *Advances in neural information processing systems*, vol. 31, 2018.
- [27] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, "Distance-IoU Loss: Faster and Better Learning for Bounding-Box Regression," in *Proceedings of the AAAI Conference on Artificial Intelligence*, AAAI, New York, NY, USA, February 2020.
- [28] C.-Y. Wang, H. Y. M. Liao, Y. H. Wu, P. Y. Chen, J. W. Hsieh, and I. H. Yeh, "CSPNet: a new backbone that can enhance learning capability of CNN," in *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1571–1580, Seattle, WA, USA, June 2020.
- [29] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to Sequence Learning with Neural Networks," NIPS, 2014, <https://arxiv.org/abs/1409.3215>.
- [30] D. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," 2015, <https://arxiv.org/abs/1409.0473>.
- [31] T. Luong, H. Pham, and C. D. Manning, "Effective Approaches to Attention-Based Neural Machine Translation," 2015, <https://arxiv.org/abs/1508.04025>.
- [32] K. Ahmed, N. S. Keskar, and R. Socher, "Weighted Transformer Network for Machine Translation," 2017, <https://arxiv.org/abs/1711.02132>.
- [33] J. Liu, Y. Chen, K. Liu, and J. Zhao, "Event Detection via Gated Multilingual Attention Mechanism," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, AAAI, New Orleans, LA, USA, February 2018.
- [34] Y. Du, C. Yuan, B. Li, L. Zhao, Y. Li, and W. Hu, "Interaction-aware Spatio-Temporal Pyramid Attention Networks for Action Classification," in *Proceedings of the European Conferences of Computer Vision ECCV*, Munich, Germany, September 2018.
- [35] A. Srinivas, T. Y. Lin, N. Parmar, J. Shlens, P. Abbeel, and A. Vaswani, "Bottleneck Transformers for Visual Recognition," in *Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16514–16524, Las Vegas, NV, USA, June 2021.
- [36] D. Zhang and K. Song, "MCnet: multiple context information segmentation network of No-service rail surface defects," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–9, 2021.
- [37] M. Niu and K. Song, "Unsupervised saliency detection of rail surface defects using stereoscopic images," *IEEE Transactions on Industrial Informatics*, vol. 99, p. 1, 2020.
- [38] Y. He, K. Song, Q. Meng, and Y. Yan, "An end-to-end steel surface defect detection approach via fusing multiple hierarchical features," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, p. 1, 2019.
- [39] Y. Wu, D. Guo, H. Liu, and Y. Huang, "An end-to-end learning method for industrial defect detection," *Assembly Automation ahead-of-print*, vol. 40, no. 1, 2019.
- [40] X. Xu, J. Chen, and H. Zhang, "D4Net: De-deformation Defect Detection Network for Non-rigid Products with Large Patterns," *Information Sciences*, vol. 547, no. 2, 2020.
- [41] J. Gan and Q. J. H. Li, "A hierarchical extractor-based visual rail surface inspection system," *IEEE Sensors Journal*, vol. 17, no. 23, pp. 7935–7944, 2017.
- [42] T. Schlagenhaut, M. Landwehr, and J. Fleischer, "Industrial Machine Tool Component Surface Defect Dataset," *Data in Brief*, vol. 39, Article ID 107643, 2021.