Research Article

Vehicle Reidentification Based on MAPANet and k-Reciprocal Encoding

Shuai Tian ^{1,2} and Xuedong Tian ^{1,2}

¹School of Cyber Security and Computer, Hebei University, Baoding 071002, China

²Institute of Intelligent Image and Document Information Processing, Hebei University, Baoding 071002, China

Correspondence should be addressed to Xuedong Tian; xuedong_tian@126.com

Received 26 May 2021; Accepted 12 August 2021; Published 30 August 2021

Academic Editor: Jiangbo Qian

Copyright © 2021 . 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.

Vehicle reidentification has important applications in intelligent monitoring systems. However, due to many factors, such as inaccurate vehicle image detection and viewing angle changes, distinguishing features cannot be effectively obtained when the vehicle is reidentified. To improve the recognition ability and robustness of vehicle reidentification, this study proposes a new multiattention part alignment network (MAPANet). The network uses different channels in the feature map to perceive different characteristics of the image clustering of the channels and achieves fine-grained attention to the vehicle. It can automatically locate the distinguishing subregions in the vehicle image and avoid the need for a large number of additional manual pretreatment steps. Moreover, an unsupervised reranking method based on multiple metrics is proposed. The k-reciprocal encoding algorithm can optimize the performance of the sorted list in the reordering problem, recalculate the interclass and intraclass distances of vehicle pictures, and improve sorting results. The experiments in this paper are carried out on the VeRi-776 and VehicleID datasets, and the mean average precision (mAP) results on the two datasets are 72.83% and 75.25%, respectively.

1. Introduction

The problem of vehicle reidentification refers to a traffic monitoring scene within a specific range, judging whether vehicle images taken at different times in nonoverlapping areas belong to the same vehicle. Vehicle reidentification [1] plays an important role in traffic flow monitoring, city monitoring and tracking suspicious vehicles, estimating vehicle trajectory and speed [2], improving the safety performance of automated vehicles, and reducing transportation energy consumption [3, 4]. Vehicle reidentification is more fine-grained than material object identification, and different vehicle instances must be distinguished. Therefore, it is particularly challenging to obtain distinctive features in the task of vehicle reidentification.

The vehicle reidentification method mainly relies on the global appearance features of the image, but due to the following problems, excessive reliance on the global features extracted by the deep neural network cannot achieve a good recognition effect. First, the similarity within the vehicle

category is high. Different vehicles, especially vehicles of the same make and model, are very similar in appearance and color. Second, subtle local cues in the vehicle image are very important for reidentification, such as interior decoration, tire texture, and other details. However, due to the large changes in the viewing angle of the vehicle image, it is impossible to accurately locate these detailed features. Third, the license plate is invalid. Vehicle license plates are the unique identification code of vehicle identity and have been widely used in real traffic flow analysis systems [5]. However, license plate recognition is too dependent on the external environment, and persons planning illegal activities may deliberately remove, obscure, or disguise the license plate, which makes license plate recognition technology unreliable. In summary, the method of extracting appearance attributes between different vehicle images based on a deep neural network has great limitations, and it must pay attention to smaller discriminative local areas to distinguish vehicles.

The contributions of this study can be summarized as follows:

- (1) Introduce local discriminant features into the vehicle recognition task, use different channel response information to obtain local features, and reduce the impact of local part misalignment and a large number of marked key points.
- (2) Propose a discriminative and fine-grained network composed of multiple branches—the multiattention part alignment network (MAPANet), which extracts and integrates global and local features in a multi-branch network structure. By combining complementary global and local features, the subtle differences between different vehicle logos with similar appearances can be accurately distinguished.
- (3) To further improve the recognition performance, based on the representative features of multiattention points and global images, the fusion measurement strategy and k-receptive algorithm are used to design the resorting stage.

2. Related Work

2.1. Vehicle Reidentification. The problem of vehicle reidentification has received extensive attention in recent years. Similar to pedestrian reidentification [6–9], it relies on the global appearance characteristics of the image, but the feature description of pedestrian reidentification can only be applied to pedestrians. Migrating the pedestrian reidentification method to vehicle reidentification did not achieve the expected results. The appearance attributes of vehicles are highly monotonous, so a targeted network structure needs to be well applied to vehicle reidentification.

With the development of deep learning models, Liu et al. [10] merged vehicle color and texture data with high-level semantic information extracted by neural networks for vehicle reidentification and constructed a vehicle reidentification dataset VeRi-776 [10]. The results showed that with deep features and manual “Compared with” features, vehicles are clearly distinguishable. Later, they [11] constructed a large-scale vehicle dataset VehicleID [11] and proposed a progressive vehicle reidentification (PROVID) method. The image was filtered by appearance information, while the license plate information was used for more refined verification, and finally, the images were reordered. In recent years, generative adversarial networks (GANs) have also been widely used for vehicle recognition. Zhou and Shao [12] proposed a viewpoint-aware attention multiview inference model (VAMI) based on visual information, which extracted single-view features of vehicle images and transformed them into global multiview features, and adopted the viewpoint-aware attention model to achieve an effective multiview inference model through GAN viewpoint feature reasoning. In addition, Chen et al. [13] designed a segmentation and reunion network structure (PRN) and proposed a 3D segmentation strategy. By using a double-branch structure, the network is divided into “height-channel” branches and “width-channel” branches to avoid certain spatial characteristics. Zheng et al. [14] proposed a multiscale attention framework (MSA) that considers multiscale

mechanisms and attention technologies. The multiscale mechanism is used in the feature map to obtain a comprehensive representation of the fusion of global and local cues, and it is used on each subnetwork attention mechanism to mine characteristic information. Chen et al. [15] used the cross-correlation method to propose a local calibration network (PAN) to address the vehicle deviation problem and obtained a more robust local feature representation.

In summary, to obtain effective regional information, most methods segment vehicle images or mark key points. As shown in Figures 1(b) and 1(c), although these methods have greatly promoted vehicle reidentification research, due to change in the vehicle’s perspective, the segmentation of parts can easily cause vehicle misalignment. The use of key points increases the data preprocessing work. To solve this problem, this study proposes a multimetric reordering method based on multiattention part alignment network (MAPANet) and k-reciprocal. As shown in Figure 1(d), by clustering the feature map channels, the local area to be considered is obtained, and the distinguishing local feature vector is obtained. Then, different distance formulas are used for reranking calculations to identify vehicles.

2.2. Multiattention Models in Person and Vehicle Re-ID. The attention mechanism [14] has certain advantages in dealing with local dislocation challenges and has a wide range of applications in computer vision tasks, including target detection [16], saliency detection [17, 18], aesthetic clipping [19], and cross-mode search [20, 21]. Hourglass [22] and cascaded pyramid networks [23] use the attention mechanism to estimate human body poses very well. The deconvolution head network proposed by Xiao et al. [24] uses key point detection models to locate human joints. There are relatively few studies on vehicle reidentification of local key attention. The multiscale attention framework (MSA) proposed by Zheng et al. [14], which considers multiscale mechanisms and attention technologies, uses attention blocks on subnetworks of each scale to mine complementary and distinguishing information. Tumrani et al. [25] proposed local attention and multiple attributes based on appearance features for vehicle reidentification, improved the number of output channels of the deconvolution head network proposed by Xiao et al. [24], and realized the use of vehicle key points to obtain local features.

2.3. Reranking in Person and Vehicle Re-ID. The results obtained by reordering representative information are consistently more representative than those obtained through other mechanisms. Existing metric learning and ranking algorithms have been successfully applied to various reidentification problems. The metric learning method is mainly based on Markov metric learning (KISSME) and cross-view quadratic discriminant analysis (XQDA) [26]. For reidentification of human individuals, Zhong et al. [27] proposed the k-reciprocal encoding algorithm, introduced the Jaccard distance, and merged it with the initial distance to improve the reidentification results. The reordering

FIGURE 1: Comparison of different division strategies for obtaining local features of vehicles.

technology of vehicle reidentification has also attracted attention. Wang et al. [26] used the k-reciprocal encoding algorithm to propose a discriminative fine-grained vehicle reidentification network based on a two-stage reordering framework. In the first stage, the k-reciprocal encoding feature is obtained from the fusion feature. In the second stage, the average feature of the sample is formed by extracting the average center of k-reciprocal encoding nearest neighbors. The final distance is weighted by the original and Jaccard distances.

3. Overview of Vehicle Reidentification and Reranking System Based on MAPANet

Subtle local clues are the key to the fine-grained image classification algorithm [28], which contains sufficient distinguishing feature information. Figures 2(a)–2(f) show different vehicles with similar appearances obtained from the two datasets. As shown in Figure 2(a), the vehicles are distinguished by the vehicle logo, and in Figures 2(c) and 2(e), the annual inspection labels on the hood and windshield are used, respectively. In Figure 2(b), it is determined whether the wheel hubs are the same, whether there is a receiving antenna on the roof of Figure 2(f), and the appearance of the same vehicle in Figure 2(d) is different due to different tasks. The reranking method is very important in reidentification. Figure 2(g) shows the top ten images in the reidentification. The red box indicates the incorrect sample, and the green box corresponds to the correct sample. The figure shows that some fake samples achieve higher rankings, while certain true samples achieve lower rankings. Fine-grained vehicle identification is the key to vehicle reidentification. Fine-grained networks can detect and extract more detailed and distinguishing information.

Aiming at the difficulty of extracting local subtle clues due to vehicle misalignment and angle changes, this study proposes a multiattention part alignment network (MAPANet), which probably realizes the extraction of distinctive local features of the vehicle. The overall structure of MAPANet network and reidentification is shown in Figure 3, which is mainly composed of two parts: feature extraction and reordering. The feature extraction part is used to extract multiattention local and global features. The reidentification part reorders the initial list to reduce the top ranking of false samples. In Figure 3, part A shows the picture of the MAPANet network. Due to the convincing performance of IBN-Net and the well-designed network architecture of ResNet50, the MAPANet model uses the

ResNet50 network with the IBN-Net structure as the backbone network of MAPANet. Following the effective modification strategy in the person-RE-ID model, the conv of the third to fourth stages of ResNet50-b is copied and divided into five branches. The five branches are named as Global Feature and Part Feature. Global Feature, which contains a branch, is used to extract the deep-level global features of the vehicle and perform similarity mapping from the image to Euclidean space through recognition and verification loss. Part Feature, which contains four branches, extract local images obtained by channel grouping and pay more attention to local features, so that subtle differences between vehicles can be recognized. The vehicle image with the size of $256 \times 256 \times 3$ is entered into the backbone network, and five feature maps are generated through five branches, including one global feature map and four local feature maps. When generating a local feature map, due to the channel grouping operation, the feature vector dimension is different, and 1×1 convolution is used on each vector to align with global feature graphs. After the global feature performs the GAP operation, the fully connected layer outputs a 512-dimensional vector. In order to complete the reordering stage, the local feature maps are connected in series, and the dimension of the vector is unified to 512 using 1×1 convolution. The two subnetworks of Global Feature and Part Feature share weights at the same time during training and combine recognition and verification losses. Part B is the proposed reordering algorithm. Based on the k-reciprocal encoding algorithm, the Mahalanobis distance of the global feature is calculated as the initial list and the Jaccard distance of the multiattention local feature. Finally, the two-part distance of the deep feature vector is merged for sorting.

4. Vehicle Reidentification Based on MAPANet and k-Reciprocal

4.1. Vehicle Global Feature Extraction. To extract global features with strong resolution, the IBN-Net structure is added to ResNet50 to extract global features, which helps to improve the accuracy of the model without increasing the amount of calculation. This network has good feature extraction capabilities and has been widely used in computer vision tasks.

The recognition model regards the task of vehicle reidentification as a multiclassification problem. The cross-entropy loss function is used to train the deep model directly on the vehicle reidentification dataset. Recognition loss pays

FIGURE 2: Examples of vehicle nuances.

FIGURE 3: MAPANet framework and reidentification.

more attention to learning local semantic features, which can retain enough identifiable information for reidentification. In this way, the classification loss will be more inclined to semantic local features rather than global features during training.

Definition 1 (neural network model symbol). $f_{\text{MAPANet}} = \text{Conv}(\text{VE}_{pi}, \theta_{id})$ is used to represent the extracted vehicle features. $\text{Conv}(\cdot)$ is the MAPANet network model, VE_{pi} is the image of the vehicle to be detected, f_{MAPANet} is the extraction of vehicle feature vectors, and θ is all the parameters that the model needs to learn.

The typical softmax loss function is used to train the model:

$$\text{Loss}(f_{\text{MAPANet}}, l, \theta_{tp}) = \frac{1}{\text{mVE}} \left[\sum_{l=1}^{\text{mVE}} \left(- \sum_{i=1}^N p_i^l \log(\bar{p}_i) \right) \right], \quad (1)$$

where mVE refers to the number of samples in the training set, θ_{tp} refers to the training parameters, N refers to the number of classifications, l refers to the classification target, and $(p_1^l, p_2^l, \dots, p_N^l)$ refers to the prior probability of the l st target vehicle. If $i = l$, then $p_i^l = 1$; otherwise, $p_i^l = 0$. $\bar{p}_i$ refers to the predicted probability of the softmax layer on class i .

The input of the verification model is a pair of vehicle images, which determines whether the input has the same identity. In the feature space, images of vehicles with the

same identity are closer than images of vehicles with different identities.

Definition 2 (verification loss).

$$\text{VE}_{pi}(f_{\text{MAPANet}}^i, f_{\text{MAPANet}}^j) = \begin{cases} \frac{1}{2} \|f_{\text{MAPANet}}^i - f_{\text{MAPANet}}^j\|_2^2, & \text{if } f_{\text{MAPANet}}^{ij} = 1, \\ \frac{1}{2} \max(0, m - \|f^i - f^j\|_2)^2, & \text{if } f_{\text{MAPANet}}^{ij} \neq 1, \end{cases} \quad (2)$$

where f_{MAPANet}^i and f_{MAPANet}^j represent the neural network feature vectors obtained from two vehicle images and minimize the L_2 distance between these two features. If $y_{\text{MAPANet}}^{ij} = 1$, then f_{MAPANet}^i and f_{MAPANet}^j have the same ID. Otherwise, they will have different vehicle IDs.

4.2. Vehicles Pay More Attention to Local Feature Extraction.

Figure 4 is the attention part alignment network framework; this clustering is realized by taking advantage of the different visual information corresponding to different channels in the feature map and the different characteristics of their peak response areas, clustering similar channels in the response area to obtain the local attention area. The original image of the vehicle in Figure 4(c) passes through the first stage of ResNet50-b to obtain 256 feature channels, and the white area on each feature channel is the peak response of different visual information.

In the experiment, in order to prevent too few clusters from making local features close to global features, excessive segmentation reduces the weight of global features. Therefore, in Figure 4(d), the channels are clustered into 4 sub-regions; after testing, 4 types are relatively moderate specific numbers. To make the channel clustering more optimized, the channel clustering operation is completed by using the full connection layer. Multiattention local feature extraction [29] is designed to adapt to vehicle features, as shown in Algorithm 1.

Definition 3. $(\theta_{i,d} * \text{VE}_{pi})_{(w \times h \times c)}$ is the feature map obtained from image block VE_{pi} , where $\theta_{i,d}$ is all the training parameters of the network, $*$ is the 9-layer convolution, pooling, activation, and other operations of the first block of ResNet50-b, w is the width of 56, h is the height of 56, and c is the number of channels of 256.

Definition

4. $\{(x_1, y_1), (x_2, y_2), \dots, (x_i, y_i), \dots, (x_{c_{\text{channel}}}, y_{c_{\text{channel}}})\}$ is the maximum response coordinate on the training image.

This coordinate serves as the representative feature of the channel and participates in clustering, where (x_i, y_i) is the coordinate of the maximum response position and c_{channel} is the number of channels.

Definition 5. C_{clu} refers to the number of channel clusters, and $F(\cdot) = (f_{1-c}(\cdot), \dots, f_{4-c}(\cdot))$ is the fully connected layer corresponding to the channel clusters, where the input of each $f_{i-c}(\cdot)$ is the feature map of the convolutional layer. $v_{i-c, \text{channel}}(\text{VE}_{pi})$ is the weight vector obtained by the fully connected layer.

Definition 6. $S_{i-c}(\text{VE}_{pi})$ refers to the channel grouping obtained by clustering, $P_{i-c}(\text{VE}_{pi})$ refers to the multi-attention local feature map, and $i-c \in C_{\text{clu}}$. $f_{\text{MAPANet}}(r_1, r_2 \dots r_p)^T$ refers to the attention local feature vector.

To adapt to the characteristics of the vehicle image, in the multiattention network, we set the channel grouping loss function [29], as shown in the following formula:

$$\text{Loss}(\text{VE}_{pi}) = \sum_{i=1}^4 [L_{\text{part}}(Y^i, Y^*)] + L_{\text{cng}}(S_{1c}, \dots, S_{4c}) + \frac{S^i}{\sum_{i=1}^4 S^i}, \quad (3)$$

where Y^i is the label of each subarea and S^i is the local area. In the last item representing the proportion of the entire area, the local area obtained in the experiment should be as large as possible and contain more information to prevent loss of details. L_{part} refers to the classification loss of the local region, which is used for comparison and classification of channel grouping with original features. L_{cng} loss function is used for channel group loss, and the formulas are shown as follows:

$$L_{\text{cng}}(S_{i-c}(\text{VE}_{pi})) = \text{Dis}(S_{i-c}(\text{VE}_{pi})) + \alpha \text{Div}(S_{i-c}(\text{VE}_{pi})), \quad (4)$$

$$\text{Dis}(S_{i-c}(\text{VE}_{pi})) = \sum_{(x,y) \in S_{i-c}} s_{i-c}(x, y) [\|x - x_i\|^2 + \|y - y_i\|^2], \quad (5)$$

$$\text{Div}(S_{i-c}) = \sum_{(x,y) \in S_{i-c}} s_{i-c}(x, y) [\max_k s_k(x, y) - \text{mrg}]_{k \neq i}, \quad (6)$$

FIGURE 4: Attention part alignment network. (a) Vehicle image. (b) Conv layers. (c) Feature channels. (d) Channel grouping. (e) Part representations. (f) Classification.

where Dis makes the parts more concentrated and Div() makes the parts more dispersed so that only a certain proportion of channels can exist in a certain part. The two loss functions strengthen each other and iterate continuously until their value no longer changes.

The different local features obtained in the branch network are connected in series, and the global and local feature vectors are finally obtained. The sorting result is obtained by comparing the distance between the query image VE_{pi} and the library set mVE_{pj} . Among them, this distance is obtained by using the complementarity of global or local features.

4.3. Vehicle Reordering Based on k -Reciprocal Encoding. Vehicle recognition is a subproblem of image retrieval. In the retrieval process, resorting can optimize the accuracy of the results. However, most studies determine the similarity by calculating the distance between the features, and the recognition results obtained are biased. The k -reciprocal encoding [27] algorithm has achieved good results in pedestrian recognition. This study proposes a fusion multi-metric vehicle reidentification and reranking method to determine the characteristics and differences between vehicles. If an image in mVE_{pj} and k of VE_{pi} are close to each other, they are likely to have the same ID; therefore, the performance of the vehicle recognition dataset is improved. The distance is defined as two parts. The original distance list *Old_list* is obtained by the Mahalanobis distance between the global feature $f_{MAPANet}(o_1, o_2 \dots o_p)^T$ vectors, and the Jaccard distance is obtained by multiple attention complementary local area feature $f_{MAPANet}(r_1, r_2 \dots r_p)^T$ vectors. The final *Recorder_list* list is obtained by the aggregation of two distances. The Mahalanobis distance and Jaccard distance are shown in the following formulas, respectively:

$$d(o, y) = \frac{1}{p} \sum_{i=1}^p \sqrt{(o_i - y_i)^T S^{-1} (o_i - y_i)}, \quad (7)$$

where S is the covariance matrix.

$$d_J(VE_{pi}, mVE_{pj}) = 1 - \left(\frac{|\mathfrak{R} * (VE_{pi}, k) \cap \mathfrak{R} * (mVE_{pj}, k)|}{|\mathfrak{R} * (VE_{pi}, k) \cup \mathfrak{R} * (mVE_{pj}, k)|} \right), \quad (8)$$

where mVE_{pi} is the picture in mVE and k is the number of neighboring images.

Definition 7. d^* is the final distance used to sort the list, weighted by the original distance and Jaccard distance.

$$d^*(p, mVE_{pj}) = d(o, y) + d_J(VE_{pi}, mVE_{pj}), \quad (9)$$

where $d_J(VE_{pi}, mVE_{pj})$ is the Jaccard distance and $d(o, y)$ is the Mahalanobis distance.

5. Results and Discussion

5.1. Datasets. This study uses VeRi-776 [10] and VehicleID [11] datasets to evaluate the performance of the network structure. A summary of the two datasets is shown in Table 1.

VeRi-776 is a widely used vehicle reidentification dataset. Twenty cameras are arranged in different areas of 1 km² to cover different traffic conditions. The images include lanes, intersections, and different lighting and background conditions while ensuring data quality. After manual processing and annotation, 49360 images of 776 cars were finally generated. Each vehicle is captured by at least two cameras to ensure that the data are suitable for reidentification tasks. The VeRi-776 dataset provides color, model, and make information. The vehicle colors include black, blue, brown, gray, green, gold, red, white, yellow, and purple; vehicle types include bus, sedan, hatchback, MPV, pickup, sedan, SUV, and truck; vehicle make information includes more than 30 makes.

VehicleID is a large-scale vehicle reidentification dataset presented in [11] that includes approximately 200,000 images of 26,267 cars. It mainly includes the front and rear perspectives of the vehicle. The dataset contains information such as car models and colors. Among them, there are 250 models, including Audi, Mercedes Benz, and Toyota; vehicle colors include black, blue, gray, red, silver, white, and yellow.

5.2. Network Training. The MAPANet model was trained on the dataset for 24 hours on a single NVIDIA GeForce GTX 1660Ti processor in the PyTorch environment. Multibranch neural network models usually do not allow training. Direct integration of multiparameter original network branch training requires a long convergence time. The network structure used in this study includes 5 branches, including a multiattention branch network, global feature network, and reordering network. Therefore, the network is divided into “Baseline,” representing a single-branch ResNet50 model for extracting global features; “MAPA + local,” referring to the

```

Input:  $(\theta_i)_d * VE_{pi})_{(w \times h \times c)}$ 
Output: multiattention part feature vectors  $(f_{\text{MAPANet}}(r_1, r_2 \dots r_p))^T$ 
(1) for  $i\_c$  in  $C_{\text{clu}}$  //channel grouping
(2)   for  $ic\_channel$  in  $c\_channel$ 
(3)      $v_{ic\_channel}(VE)_{pi} = f_{i\_c}((\theta_i)_d * VE_{pi})_{(w \times h \times c)}$  //weight vector
(4)   end for
(5)    $V_{ic\_channel}(VE_{pi}) = [v_1, \dots, v_{c\_channel}]$  //weight vector collection
(6)   if  $(v_{ic\_channel}(VE_{pi}) == 1)$ 
(7)      $v_{ic\_channel}(VE_{pi}) \in S_{i\_c}(VE_{pi})$ 
(8)   else
(9)      $v_{ic\_channel}(VE_{pi}) \in S_{i\_c}(VE_{pi})$ 
(10)  end if
(11) end for
(12) for  $j\_c$  in  $C_{\text{clu}}$  //generate local attention feature map
(13)    $S_{j\_c}(VE_{pi}) = \text{sigmoi } d(\sum_{j=1}^{c\_channel} ([\theta * VE_{pi}]_j v_j))$  //accumulation of similar channels, using sigmoid to generate probability
(14)    $L_{\text{eng}}(S_{j\_c}(VE_{pi}))$  //loss function training
(15)    $P_{j\_c}(VE_{pi}) = \sum_{j=1}^{c\_channel} ([\theta * VE_{pi}]_j \cdot S_{j\_c}(VE_{pi}))$  //features of the jth channel, multiplied element by element  $S_{j\_c}(VE_{pi})$ 
(16)   for  $k\_c$  in  $C_{\text{clu}}$ 
(17)      $\text{Conv1} \times 1(P_{k\_c}(VE_{pi}))$  //using  $1 \times 1$  convolution to change the number of channels
(18)    $\text{Stage}_{3-4}(\text{Conv1} \times 1(P_{k\_c}(VE_{pi})))$  //enter the remaining block to obtain the local feature map vector
(19) end for
(20)  $\oplus(\text{Stage}_{3-4}(\text{Conv1} \times 1(P_{2-4}(VE_{p2-4})))) \rightarrow f_{\text{MAPANet}}(r_1, r_2 \dots r_p)^T$ 
(21) END

```

ALGORITHM 1: Multiattention part feature extraction algorithm.

TABLE 1: Summary of VeRi-776 and VehicleID datasets.

	VeRi-776	VehicleID
Image_train	37,778 images of 576 vehicles	113,346 images of 13,164 vehicles
Image_query	1678 images of 200 vehicles	—
Image_test (full)	11,579 images of 200 vehicles	108,221 images of 13,164 vehicles
Image_test (800)	—	6493 images of 800 vehicles
Image_test (1600)	—	13,377 images of 1600 vehicles
Image_test (2400)	—	19,777 images of 2400 vehicles
Image_test (3200)	—	26,353 images of 3200 vehicles
Image_test (6000)	—	48,922 images of 6000 vehicles
Ref-average	65 images/vehicle	9 images/vehicle
Ref-large	289 images/vehicle	142 images/vehicle
Ref-small	11 images/vehicle	2 images/vehicle

proposed multiattention area; “MAPANet,” referring to the proposed network; and “MAPANet + FR,” corresponding to the entire network using the reordering method. ResNet50 pretrained on ImageNet was used as the backbone network to accelerate the training process, and pretraining was carried out through cross-entropy loss. All vehicle images are adjusted to 256×256 in order to maintain sufficient information on each segmentation part. To make the proposed model more robust, random horizontal flips and random erasures with a probability of 0.5 are used to increase the data. In order to output the 2048-dimensional feature vector of the initial FC layer, 512-dimensional feature vector and L2 distance are obtained after compression. The weight decay factor of L2 regularization is set to $5e-4$, the initial learning rate lr is set to $3e-4$, and the decay is $3e-6$ after 50 epochs. The batch size is set to 64, and the Adam optimizer based on AMSGrad is used for optimization, and the maximum epoch is 60.

5.3. Evaluation Criteria. Vehicle reidentification is a sub-problem of image retrieval. According to widely used protocols, the mean average precision (mAP) and Rank-N table evaluate the overall performance of vehicle reidentification on VeRi-776 and VehicleID vehicle data.

5.3.1. Mean Average Precision. The mean average precision represents the average of the precision of all search results.

It is an important index to evaluate the overall performance of the reidentification method, which considers both recall rate and accuracy. For each query image, calculate the average accuracy (AP) as shown in the following equation:

$$AP = \frac{\sum_{k=1}^n P(k) \times gt(k)}{N}, \quad (10)$$

where N refers to the total number of images of the target vehicle, k refers to the image sequence number of the picture

set, n refers to the total number of galleries and images, $p(k)$ refers to the accuracy before the k th position in the retrieval sequence, and $gt(k)$ refers to whether the k th image is the target vehicle. Finally, the mean average precision (mAP) of all query sets retrieved is calculated as shown in the following equation:

$$\text{mAP} = \frac{\sum_{q=1}^Q \text{AP}(q)}{Q}, \quad (11)$$

where Q is the total number of query images.

5.3.2. Rank- N Form. When comparing the performance of different methods, if the performance difference between the methods is low, the cumulative matching performance curve will mostly overlap, so it is impossible to accurately judge the performance. To compare the performance of different methods more concisely, Rank-1 and Rank-5 are selected to represent the probability of correct image matching.

5.4. Performance on the VeRi-776 Dataset. In this section, the method in this paper is compared with some traditional methods and current existing methods using the VeRi-776 dataset. Both LOMO [30] and BOW-CN [31] are traditional methods. LOMO used the manual feature local highest occurrence rate expression method that is robust to illumination changes in pedestrian reidentification for vehicle reidentification. BOW-CN used the BOW model and color name to obtain the color and color characteristics of the vehicle. Based on a deep learning network, GoogLeNet [32] was a neural network architecture that learns vehicle features and is pretrained on ImageNet. Different from conventional deep networks, two-branch convolutional networks were applied in PROVID [33], FACT + Plate-SNN + STR [34], Siamese-CNN [35], and DFN [26]. FACT + Plate-SNN + STR [34] was based on FACT, adding license plate information and spatiotemporal relationships to achieve distinguishing characteristics. DFN [26] used fine-grained network to obtain global and local features for vehicle identification. STR + ST [36] and PAMA [25] separately process the key points of the vehicle. The former extracts 20 well-aligned combination key points of local area features in different directions. PAMA [25] used multiple vehicle attributes and key points to extract vehicle features. VRSDNet [37] used the short and dense connection mechanism of the Siamese network to learn vehicle characteristics. VANet [38] used the learning method of viewpoint perception metrics to learn two metrics of similar viewpoints and different viewpoints in two feature spaces; thereby, viewpoint perception was generated. PNVR [39] relies on local regularization to discriminate feature retention, which enhances the ability to perceive subtle differences and merges local and global features through detection branches. Table 2 shows more details of the comparison results between this method and other methods.

Ablation Study. The proposed method includes three main parts, and the performance is continuously improved by

TABLE 2: The mAP, Rank-1, and Rank-5 comparison on the VeRi-776 dataset.

VeRi-776 methods	mAP (%)	R1 (%)	R5 (%)
GoogLeNet [32]	17.89	52.32	72.17
VAMI [13]	50.13	77.03	90.82
STR + ST [36]	51.42	68.3	89.7
VRSDNet [37]	53.45	83.49	92.55
Siamese-CNN [35]	54.21	79.32	88.92
EALN [40]	57.44	84.39	88.97
MSA [14]	62.89	92.07	96.19
DFN [26]	61.85	88.14	96.83
VANet [38]	66.35	89.79	95.92
PNVR [39]	72.39	92.36	96.77
Ours	72.83	93.35	97.09

gradually increasing the functions of each part. The final reordering of MAPANet has the best performance and is superior to the baseline in mAP/Rank_1/Rank_5: 13.83/5.75/2.79 (%), MAPA + local: 7.6/3.58/2.01 (%) and the proposed MAPANet: 2.96/2.04/0.63 (%), as shown in Figure 5(a).

5.5. Performance on the VehicleID Dataset. The method in this study is compared with the existing methods on the VehicleID dataset. Generally, all methods perform better on small-scale test sets because large test sets introduce more challenging and complex scenarios. EALN [40], VAMI [13], and XVGAN [41] all focused on using multiview information to obtain global features through the GAN network, so as to improve the results of vehicle reidentification. Among them, VAMI [13] viewpoint perception and attention multiview inference method used the vehicle's multiview information to extract distinguishing features from single-view features. EALN [40] designed an end-to-end embedding adversarial learning network to generate samples positioned in the embedding space, which greatly improves the network's ability to distinguish similar vehicles. DLCNN [42] used the Siamese network to combine verification and recognition losses, and the overall performance has been greatly improved. GS-TRE [43] used vehicle ID tags and uses intraclass variance to merge samples and individuals into an intermediate group. The VGG network is used as the backbone network for clustering, and the triple loss is used for training. In comparison, MAPANet achieves better performance and is much better than Baseline. Table 3 shows more details of the comparison results between this method and other methods.

Ablation Study. For the VehicleID data, when the multi-attention part is added to the global feature extraction method and the reranking algorithm is used, the vehicle recognition performance is constantly revised. MAPANet + FR finally achieves the best performance, which is most obvious in small datasets. Among them, mAP/Rank_1/Rank_5 are better than Baseline: 3.76/9.05/4.12 (%), MAPA + local: 2.57/4.94/3.39 (%), and MAPANet without reranking: 1.14/2.29/1.19 (%). Figures 5(b)–5(d) show more

FIGURE 5: Results of ablation assessment on VeRi-776 and VehicleID data. (a) VeRi-776. (b) VehicleID (800). (c) VehicleID (1600). (d) VehicleID (2400).

TABLE 3: The Rank-1 and Rank-5 comparison on the VehicleID dataset.

Method	Small (800)		Medium (1600)		Large (2400)	
	R1 (%)	R5 (%)	R1 (%)	R5 (%)	R1 (%)	R5 (%)
GoogLeNet [32]	47.88	67.18	43.40	63.86	38.27	59.39
FACT [10]	49.53	68.07	44.59	64.57	39.92	60.32
XVGAN [41]	52.79	80.69	49.47	71.42	44.92	66.71
VRSDNet [37]	56.98	86.90	50.57	80.05	42.92	73.44
VAMI [13]	63.12	83.25	52.87	75.12	47.34	70.29
DLCNN [42]	73.01	82.70	66.50	77.06	61.00	73.17
EALN [40]	75.11	—	78.69	—	69.30	—
GS-TRE [43]	75.90	84.20	74.80	83.60	72.30	82.70
DFN [26]	77.02	85.04	71.81	80.81	66.29	78.42
Ours	78.15	89.92	74.21	84.86	72.31	84.15

details. To provide more visualization to MAPANet, a more intuitive result is provided in Figure 6. The left column contains different query images, and the other columns are

the results obtained by the MAPANet method. The correct results sorted by similarity are displayed in the green box, and the incorrect results are displayed in the red box.

FIGURE 6: Example of vehicle reidentification scene.

6. Conclusions

This study proposes an unsupervised reranking method for vehicle reidentification and fusion of multiple metrics based on the MAPANet model. The MAPANet model is mainly composed of a multiattention local alignment network, local feature extraction, and global feature extraction. The design loss function is designed to include as much detail as possible by taking advantage of the different regions to which the different channels respond. The reordering method is used to further improve the reidentification performance of the vehicle. The method in this study focuses on more details of different attention areas and does not require complex strategies to maintain local matching. The proposed framework is evaluated on two public vehicle reidentification datasets (VeRi-776 and VehicleID) and compared with other current algorithms. The experimental results clearly show the competitive performance of this method.

In the future, research will be conducted in the following areas relative to vehicle reidentification:

- (1) Lighten the network structure and improve the problem of long training time for multibranch networks
- (2) Deal with special vehicles in special scenarios, such as ambulances, police cars, school buses, fire trucks, and muck trucks
- (3) Improve the local feature extraction method and increase the weight for some important local objects

Data Availability

The data used can be found at <https://pan.baidu.com/s/1UFc6VJioavlO3Po2FbJU-g>, extraction code: nnrw.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this study.

Acknowledgments

This work was supported by the Natural Science Foundation of Hebei Province, China (Grant No. F2019201329).

Portions of the research in this paper used the VehicleID dataset funded by the National Natural Science Foundation of China.

References

- [1] Q. Wang, W. Min, Q. Han et al., "Viewpoint adaptation learning with cross-view distance metric for robust vehicle reidentification," *Information Sciences*, vol. 564, pp. 71–84, 2021.
- [2] X. Wang, J. Zhao, Y. Wang, J. Lv, and W. Yan, "Vehicle feature point trajectory clustering and vehicle behavior analysis in complex traffic scenes," *Springer International Publishing*, vol. 316, pp. 182–205, 2020.
- [3] F. Zhang, X. Hu, R. Langari, and D. Cao, "Energy management strategies of connected HEVs and PHEVs: recent progress and outlook," *Progress in Energy and Combustion Science*, vol. 73, pp. 235–256, 2019.
- [4] D. Q. Tran and S.-H. Bae, "Improved responsibility-sensitive safety algorithm through a partially observable Markov decision process framework for automated driving behavior at non-signalized intersection," *International Journal of Automotive Technology*, vol. 22, no. 2, pp. 301–314, 2021.
- [5] P. Ingawale and L. Desai, "Deep learning based vehicle license plate recognition system for complex conditions," *Advances in Intelligent Systems and Computing*, vol. 1, pp. 499–507, 2020.
- [6] X. R. Song, S. Gao, and C. B. Chen, "A multispectral feature fusion network for robust pedestrian detection," *Alexandria Engineering Journal*, vol. 60, no. 1, pp. 73–85, 2020.
- [7] Y. Jin, Y. Zhang, Y. Cen, Y. Li, V. Mladenovic, and V. Voronin, "Pedestrian detection with super-resolution reconstruction for low-quality image," *Pattern Recognition*, vol. 115, no. 3, Article ID 107846, 2021.
- [8] Y. Wang, C. Han, G. Yao, and W. Zhou, "MAPD: an improved multi-attribute pedestrian detection in a crowd," *Neurocomputing*, vol. 432, no. 8, pp. 101–110, 2021.
- [9] X. Hu, D. Wei, Z. Wang, J. Shen, and H. Ren, "Hypergraph video pedestrian re-identification based on posture structure relationship and action constraints," *Pattern Recognition*, vol. 111, Article ID 107688, 2021.
- [10] X. C. Liu, W. Liu, H. D. Ma, and H. Y. Fu, "Large-scale Vvehicle re-identification in urban surveillance videos," in *Proceedings of the the 2016 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, IEEE Press, Seattle, WA, USA, July 2016.
- [11] 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 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2167–2175, Las Vegas, NV, USA, June 2016.
- [12] Y. Zhou and L. Shao, “Viewpoint-aware attentive multi-view inference for vehicle re-identification,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6489–6498, Salt Lake City, UT, USA, June 2018.
  - [13] H. Chen, B. Lagadec, and F. Bremond, “Partition and reunion: a two-branch neural network for vehicle re-identification,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 184–192, Long Beach, CA, USA, June 2019.
  - [14] A. H. Zheng, X. M. Lin, J. C. Dong, W. Z. Wang, J. Tang, and L. Bin, “Multi-scale attention vehicle re-identification,” *Neural Computing & Applications*, vol. 32, pp. 17489–17503, 2021.
  - [15] Y. Chen, B. Ma, and H. Chang, “Part alignment network for vehicle re-identification,” *Neurocomputing*, vol. 418, pp. 114–125, 2020.
  - [16] D. Fan, W. Wang, M. Cheng, and J. Shen, “Shifting more attention to video salient object detection,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8554–8564, Long Beach, CA, USA, June 2019.
  - [17] J. Zhao, J. Liu, D. Fan et al., “EGNet: edge guidance network for salient object detection,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8779–8788, Seoul, Korea, November 2019.
  - [18] K. Fu, Q. Zhao, and I. Yang, J. Yang, “Deepside: a general deep framework for salient object detection,” *Neurocomputing*, vol. 356, pp. 69–82, 2019.
  - [19] D. Fan, Z. Lin, J. Zhao et al., “Rethinking rgb-d salient object detection: models, datasets, and large-scale benchmarks,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, pp. 2075–2089, 2019.
  - [20] W. Wang, J. Shen, and H. Ling, “A deep network solution for attention and aesthetics aware photo cropping,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 7, pp. 1531–1544, 2019.
  - [21] X. Xu, F. Shen, Y. Yang, H. T. Shen, and X. Li, “Learning discriminative binary codes for large-scale cross-modal retrieval,” *IEEE Transactions on Image Processing*, vol. 26, no. 5, pp. 2494–2507, 2017.
  - [22] A. Newell, K. Yang, and J. Deng, “Stacked hourglass networks for human pose estimation,” in *Proceedings of the European Conference on Computer Vision*, pp. 483–499, Springer, Amsterdam, Netherlands, October 2016.
  - [23] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Wu, and J. Sun, “Cascaded pyramid network for multi-person pose estimation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7103–7112, Salt Lake City, UT, USA, June 2018.
  - [24] B. Xiao, H. Wu, and Y. Wei, “Simple baselines for human pose estimation and tracking,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 466–481, Munich, Germany, September 2018.
  - [25] S. Tumrani, Z. Deng, H. Lin, and J. Shao, “Partial attention and multi-attribute learning for vehicle re-identification,” *Pattern Recognition Letters*, vol. 138, pp. 290–297, 2020.
  - [26] Q. Wang, W. Min, D. He et al., “Discriminative fine-grained network for vehicle re-identification using two-stage re-ranking,” *Science China. Information Sciences*, vol. 63, no. 11, pp. 1–12, 2020.
  - [27] Z. Zhong, L. Zheng, D. Cao, and S. Li, “Re-ranking person re-identification with k-reciprocal encoding,” in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, July 2017.
  - [28] Z. Ye, F. Hu, Z. Xia, F. Lyu, P. Liu, and Y. Lin, “Associating multi-scale receptive fields for fine-grained recognition,” in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, pp. 1851–1855, Abu Dhabi, United Arab Emirates, November 2020.
  - [29] H. Zheng, J. Fu, T. Mei, and J. Luo, “Learning multi-attention convolutional neural network for fine-grained image recognition,” in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 5209–5217, Venice, Italy, October 2017.
  - [30] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, “Person re-identification by local maximal occurrence representation and metric learning,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2197–2206, IEEE, Boston, MA, USA, June 2015.
  - [31] 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 (ICCV)*, pp. 1116–1124, Santiago, Chile, December 2015.
  - [32] L. Yang, P. Luo, X. Tang, and C. C. Loy, “A large-scale car dataset for fine-grained categorization and verification,” in *Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition*, pp. 3973–3981, Boston, MA, USA, June 2015.
  - [33] 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.
  - [34] X. Liu, W. Liu, T. Mei, and H. Ma, “A deep learning-based approach to progressive vehicle re-identification for urban surveillance,” in *Proceedings of the European Conference on Computer Vision*, pp. 869–884, Amsterdam, Netherlands, October 2016.
  - [35] 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 IEEE International Conference on Computer Vision (ICCV)*, pp. 1900–1909, Venice, Italy, October 2017.
  - [36] Z. Wang, L. Tang, X. Liu et al., “Orientation invariant feature and spatial temporal regularization for vehicle re-identification,” in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 379–387, Venice, Italy, October 2017.
  - [37] J. Zhu, Y. Du, Y. Hu, L. Zheng, and C. Cai, “VRSDNet: vehicle re-identification with a shortly and densely connected convolutional neural network,” *Multimedia Tools and Applications*, vol. 78, no. 20, pp. 29043–29057, 2019.
  - [38] R. Chu, Y. Sun, Y. Li, Z. Liu, C. Zhang, and Y. Wei, “Vehicle re-identification with viewpoint-aware metric learning,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8282–8291, Seoul, Korea, October 2019.
  - [39] B. He, J. Li, Y. Zhao, and Y. Tian, “Part-regularized near-duplicate vehicle re-identification,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3997–4005, Long Beach, CA, USA, June 2019.
  - [40] Y. Lou, Y. Bai, J. Liu, S. Wang, and L.-Y. Duan, “Embedding adversarial learning for vehicle re-identification,” *IEEE Transactions on Image Processing*, vol. 28, no. 8, pp. 3794–3807, 2019.

- [41] Y. Zhou and L. Shao, "Cross-view GAN Based Vehicle Generation for Re-identification," in *Proceedings of the British Machine Vision Conference*, pp. 1–12, Imperial College, London, UK, September 2017.
- [42] H. Guo, C. Zhao, Z. Liu, J. Wang, and H. Lu, "Learning coarse-to-fine structured feature embedding for vehicle re-identification," *AAAI*, vol. 32, no. 1, pp. 6853–6860, 2018.
- [43] Y. Bai, Y. Lou, F. Gao, S. Wang, Y. Wu, and L.-Y. Duan, "Group-Sensitive triplet embedding for vehicle reidentification," *IEEE Transactions on Multimedia*, vol. 20, no. 9, pp. 2385–2399, 2018.