A Dual-Branch Network for Few-Shot Vehicle Re-Identification With Enhanced Global and Local Features

Wei Sun, Fan Xu, Xiaorui Zhang, Yahua Hu, Guangzhao Dai, and Xiaozheng He

***Abstract*— Traditional vehicle re-identification (Re-ID) methods mainly rely on large-size training samples to achieve better results. However, obtaining abundant training samples is challenging for applications of these methods to few-shot vehicle Re-ID in real-world traffic scenes. To address this challenge, we propose a dual-branch network (DB-Net) for few-shot vehicle Re-ID with enhanced global and local features. Our proposed method reduces the dependence of vehicle Re-ID network on a large number of training samples by improving its feature expression ability and feature quality, rather than synthesizing vehicle images in traditional methods. The proposed DB-Net, composed of global and local branches, extracts global appearance features and local detail features of the vehicle. A global feature optimization module in the DB-Net mines and maintains more global semantic information using convolution and concat operations. Also, a feature screening module selects the most relevant local features based on mutual information and Shell sorting to enhance local features. In addition, a local attention module (L-ATT) assigns adaptive weights to enhance salient local regions. Our approach is evaluated on a new dataset (Veri-FS) with small sample sizes and poor illumination conditions. It outperforms state-of-the-art methods on VehicleID, VeRi-776, and Veri-FS datasets, demonstrating its effectiveness in few-shot vehicle Re-ID.**

***Index Terms*— Attention mechanism, feature screening, few-shot learning, vehicle dataset, vehicle re-identification (Re-ID).**

I. Introduction

VEHICLE re-identification (Re-ID) aims to retrieve all images of a given query vehicle from the image data captured by surveillance cameras, which is of great significance for urban monitoring, vehicle detection, and intelligent transportation. In recent years, the development of deep convolutional neural networks (CNNs) [4] and large-scale datasets [5], [6] expedites the invention of vehicle Re-ID methods [8], [9], [10], [11], [12] to achieve remarkable success.

Existing vehicle Re-ID methods can be classified into global feature learning and local feature learning-based methods. Early research [9], [10] focused on learning global information of vehicles and incorporating information, such as car type, color, and perspective into the global features. However, the appearance of the vehicle changes greatly from different perspectives, and the acquired global information lack complete semantic capability. Therefore, the methods based on global information cannot identify distinguishable details from the local regions of the vehicle. To resolve the issue, recent studies [12], [13] use the bounding boxes to locate more salient local regions of the vehicle, such as car logos, lights, windows, and rear-view mirrors.

Although existing methods have achieved good progress, most of them rely on large-size samples in the training process. However, in real-world traffic scenes, it is difficult to obtain a large number of high-quality vehicle images. The reasons are given as follows: 1) the monitoring area is relatively large and the camera coverage is sparse, resulting in few sample images; 2) the illumination conditions are insufficient and the images taken are unclear; and 3) for some special scenes, the vehicles intentionally avoid the surveillance cameras, resulting in fewer vehicle images, as shown in Fig. 1(a). Therefore, it is difficult for traditional methods to effectively reidentify query vehicles in the abovementioned scenes.

To alleviate the problem of few-shot vehicle images, researchers propose to use GAN [8], [14], [15] to synthesize samples to provide more training samples for vehicle Re-ID. However, these methods are difficult to generate fine-grained vehicle texture information, especially when the vehicle types are similar, as shown in Fig. 1(b) [5]. The effectiveness of the generated samples for vehicle Re-ID will be greatly reduced, which in turn leads to a reduced accuracy in vehicle Re-ID.

To address the above challenge, this study proposes a dual-branch network (DB-Net) for few-shot vehicle Re-ID with enhanced global and local features, which reduces the dependence on large-size vehicle samples for network training. The method adopts two module branches, a global branch and a local branch, to extract the global and local features of the vehicle. In addition, the extracted global features are

Manuscript received 13 February 2023; revised 1 May 2023; accepted 23 May 2023. Date of publication 23 June 2023; date of current version 29 June 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 62272236 and Grant 61502240; in part by the Natural Science Foundation of Jiangsu Province under Grant BK20191401 and Grant BK20201136; in part by the Innovation and Entrepreneurship Training Project of College Students under Grant 202010300290, Grant 202010300211, and Grant 202010300116E; and in part by the Postgraduate Research and Practice Innovation Program of Jiangsu Province under Grant SJCX21_0363. The Associate Editor coordinating the review process was Dr. Manyun Huang. *(Corresponding author: Wei Sun.)*

Wei Sun is with the School of Automation and the Jiangsu Collaborative Innovation Center on Atmospheric Environment and Equipment Technology, Nanjing University of Information Science and Technology, Nanjing 210044, China (e-mail: sunw0125@163.com).

Fan Xu, Yahua Hu, and Guangzhao Dai are with the School of Automation, Nanjing University of Information Science and Technology, Nanjing 210044, China.

Xiaorui Zhang is with the School of Computer Science and the Wuxi Research Institute, Nanjing University of Information Science and Technology, Nanjing 210044, China.

Xiaozheng He is with the Department of Civil and Environmental Engineering, Rensselaer Polytechnic Institute, Troy, NY 12180 USA.

Digital Object Identifier 10.1109/TIM.2023.3285978

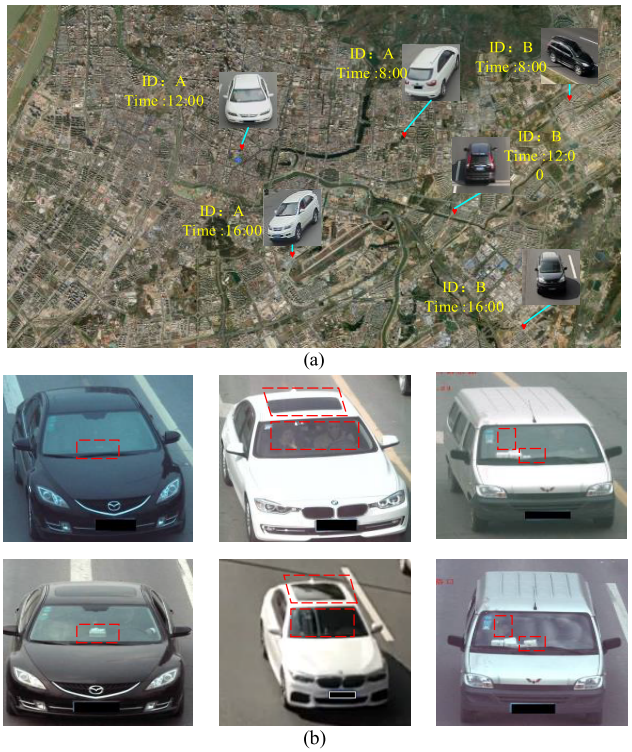


Fig. 1. Challenges of few-shot vehicle Re-ID. (a) Vehicles with the same identity have a large spatiotemporal span. (b) Red dashed boxes mark subtle differences between similar-looking vehicles.

optimized, and local features are enhanced to improve the extracted feature quality, which fundamentally boosts the accuracy of few-shot vehicle Re-ID. Because none of the existing datasets are dedicated to few-short vehicle Re-ID, this study constructs a few-shot vehicle Re-ID dataset (Veri-FS) to facilitate the performance investigation of the proposed method. The main contributions and innovation of this article are on the following three aspects.

1) This article develops a global feature optimization module, which solves the problem of weak semantic power in shallow global features. Different from the normalization and activation functions directly after the global features acquired by parsing-based view-aware embedding Network (PVEN) and three-branch embedding network (TBE-Net), the strong semantic information of global features is mined through the parallel convolution layers of the main branch of this module. In addition, in view of the problem that the deep convolution operation may reduce the resolution of global features, a shortcut branch is used to fuse the original global features with the convolved global features to obtain optimized global features. This branch ensures that the network can extract enhanced vehicle features and improve the ability to perceive unique details.
2) This article develops a meta-attention module (MAM), which is composed of a meta-learning-based feature screening module and a local attention module (L-ATT). Different from the direct input of local features into the attention mechanism, we filter the local features of vehicles by mutual information and Shell ranking algorithm. The features with high relevance are selected to improve the feature quality, thereby potentially resolving the network's dependence on large-size training samples. Attention weights are generated from these features through the built-in meta-weight generator of L-ATT and the weights are adaptively assigned to key regions, which further enhances the targeting ability of learned vehicle features.
3) We construct a Veri-FS with a large spatiotemporal span and few samples. Different from traditional datasets, this Veri-FS dataset can represent the monitoring scenes in the real world, such as few samples, large spatiotemporal span, and complex background. The Re-ID networks of few-shot vehicles can be effectively trained and verified upon the constructed Veri-FS dataset.

II. Related Work

This section first introduces the role and application of few-shot learning. Then, the existing methods and problems of vehicle Re-ID are introduced. Finally, we illustrate the reasons why existing vehicle Re-ID datasets are not suitable and the Veri-FS dataset proposed in this article.

A. Few-Shot Learning

The successful applications of deep learning to visual recognition tasks rely heavily on the availability of a large amount of high-quality labeled data, which is usually expensive to obtain. Recently, the few-shot learning [17], [18], [19], [20], [21], [22] problems have attracted increasing attention to tackling the issues associated with the small-size training samples. Most of the existing methods for few-shot learning in the field of image classification are based on meta-learning.

Meta-learning is also known as learning to learn, which aims to learn a paradigm that can be adapted to recognize novel classes upon a few training examples. The meta-learning-based methods can be further classified into two categories: metric- and optimization-based methods.

Metric-based methods focus on learning a good metric to measure the distance or similarity among the images. For example, Sung et al. [20] proposed a transferable depth metric method for comparing the relationship between images. Snell et al. [22] used a prototype network to calculate the distance of the prototype representation of different classes between the query samples and the support samples. The reviewed methods are suitable for scenarios with large differences between different categories. However, for vehicles, the interinstance discrepancy between different vehicles sometimes is subtle, especially when they share the same type and color. A large number of training samples are required to increase the recognition ability of the networks.

Optimization-based methods aim to design an optimization algorithm to reduce the network's dependence on samples [21]. For example, Ravi and Larochelle [18] trained a long short-term memory network (LSTM)-based meta-learning algorithm as an optimizer for feature learning. Jamal et al. [17] proposed a meta-learning algorithm based on task unbiased thinking, which improved network generalization ability to a certain extent. Because optimization-based

methods have achieved huge success in image classification, we introduce them to the few-shot vehicle Re-ID and use a meta-learning-based feature screening method to enhance the feature expression ability.

B. Vehicle Re-Identification

Currently, vehicle Re-ID methods can be classified into three categories.

1) *Based on Global Appearance Features:* Early research mainly focused on learning global features of the vehicle by training the networks, such as appearance [23], color, texture, and orientation of the vehicles. Liu et al. [6] proposed the fusion attributes and color features (FACT), which uses networks to fuse attributes and semantic information. Based on the FACT method, Liu et al. [24] further improved the Re-ID performance by a coarse-to-fine network framework to integrate multimodal attributes. Although the accuracy of these methods has been greatly improved, it suffers when the appearance of the vehicle changes drastically due to low resolution, lighting changes, and occlusion. To solve the problems of large differences in vehicle images from different camera perspectives [25] and aligning global appearance features, Chu et al. [26] proposed a new vehicle feature matching algorithm, which divided the vehicle image into two feature spaces by calculating the viewpoints and used different loss functions to improve the model performance.
2) *Based on Local Region Features:* To further improve the representation ability of the vehicle Re-ID networks, methods based on local region features have been proposed recently. These methods can be classified based on vehicle key points and local components. For example, Wang et al. [27] obtained local information of vehicles through 20 premarked key points and generated direction-invariant features according to the marked key points. Inspired by key points, He et al. [13] used YOLO [28] to detect salient regions, such as windows, license plates, and introduced more region features into the vehicle Re-ID framework to improve the local feature capability in the learning process. Li et al. [12] introduced a novel framework, which successfully encodes geometric local features to distinguish vehicle instances, optimized only by the supervision from original ID labels. However, the positions of these key feature points or local regions are predetermined, and their positions will change with the change in the shooting angle.
3) *Based on Attention Mechanism:* Recently, the visual attention mechanism has been widely used in computer vision tasks. For vehicle Re-ID, the modules can automatically highlight the important regions of the input vehicle images and ignore the useless regions by this mechanism, facilitating the effective extraction of recognizable global and local features. Li et al. [12] introduced an interpretable attention module, with the core of local maxima aggregation instead of fully automatic learning. Khorramshahi et al. [29] proposed a dual-path adaptive attention module, which can adaptively obtain important features and vehicle orientations to extract locally identifiable features in the vehicle Re-ID. When the camera's viewing angle changes greatly, the above two methods cannot adapt to the changes of key features and local regions, resulting in differences in the features learning each time. In addition, obtaining vehicle orientation information relies on deep learning algorithms and large-scale data for training, which consumes significant computing resources. The method we propose combines a meta-learning-based method with an attention mechanism to better retain the detailed information of local components of the vehicle through feature screening. The built-in meta-weight generator generates attention weights to highlight the locally salient regions of the query vehicle. The local regions can improve the feature expression ability of the network and reduce the network's dependence on sample size.

C. Datasets

Several datasets have been constructed for the vehicle Re-ID. VehicleID [5] is a large-scale vehicle Re-ID dataset that consists of a training set with 110 178 images of 13 134 vehicles and a testing set with 111 585 images of 13 133 vehicles. However, samples in the VehicleID dataset are captured under relatively restricted conditions, where vehicle images are captured in the daytime by multiple cameras. Thus, the perspective, illumination conditions, and background of the dataset are specific, causing issues when method transferability is a concern. Therefore, the dataset might not be suitable for the few-shot vehicle Re-ID with changing perspective, illumination, and background.

VeRi-776 [6] is the most widely used dataset for vehicle Re-ID tasks. It contains about 50 000 images of 776 vehicle identities from 20 cameras on a ring road. This dataset collects vehicles' images in an area of 1 km^2 in a 1-h period. The spatiotemporal span is small.

Cityflow [7] is one of the largest datasets, including 229 680 images from 666 vehicles. However, this dataset is based only on a 3.25-h video. In addition, the maximum distance between two cameras is 2.5 km. The captured vehicle images with the same ID have many repeated samples, and the spatial span is also small.

Although the abovementioned datasets have played an important role in the community of vehicle Re-ID, the limited spatiotemporal span of these datasets is not suitable for training and verification of few-shot vehicle Re-ID methods. Therefore, we construct a new dataset Veri-FS suitable for few-shot vehicle Re-ID. Compared to the existing vehicle Re-ID datasets, our dataset presents the characteristics of complex background information, illumination conditions, and different camera perspectives. The Veri-FS dataset fully embodies the characteristics of the large spatiotemporal span of vehicles, complex backgrounds, changeable perspectives, and a small scale of vehicle samples with the same ID.

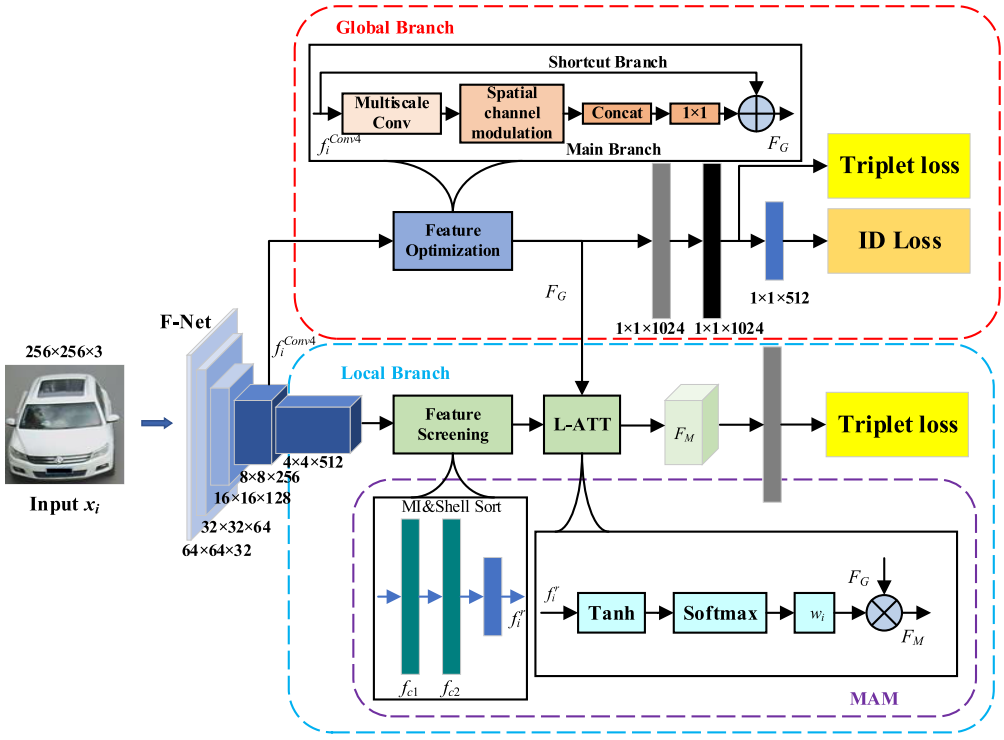


Fig. 2. Network structure of DB-Net. The image is fed into the feature extractor, and the fourth convolutional layer as the global feature is input into the feature optimization module to construct ID loss and triple loss. At the same time, the extracted local features are input into the local branches, and the mutual information and Shell sort algorithm are used to analyze and screen the local features. In addition, the attention weights of different regions are learned by the built-in meta-weight generator of L-ATT, and the key regions are adaptively assigned with larger weights.

III. Method

Fig. 2 shows the structure of our proposed DB-Net. The network consists of two branches, the global branch and the local branch. In the training phase, we perform feature extraction on the input image through F-Net and use the fourth convolutional layer that retains more spatial dimensions and information as our global feature. The global features are input into the feature optimization module (red dotted box in the figure) to mine stronger semantic information and improve the expressiveness of features through three parallel convolutions. At the same time, the extracted local features are input into the local branch (blue dotted box in the figure), which uses the mutual information and the Shell sort algorithm to analyze and screen local features to improve feature quality. The enhanced local features are beneficial to reduce the network's dependence on samples. Besides, the attention weights of different regions are learned through the built-in meta-weight generator of L-ATT to adaptively assign larger weights to key regions. The attention weights enhance the preservation of discriminative information about the extracted vehicle features. Finally, this study adopts global average pooling to obtain the salient information of feature maps. The combined ID loss and triplet loss are used to train DB-Net to obtain more abundant vehicle features. In the testing phase, we use the trained network to extract global features for vehicle Re-ID.

A. Feature Learning

This research applies the CNN pretrained on ImageNet [31] as a feature extraction network (F-Net [8]) to learn the intrinsic characteristics of the vehicle, including vehicle brand, model, color, and class. Zhou and Shao [8] showed that, when the backbone is deployed with five convolutional layers (Conv) and two fully connected (FC) layers, not only the number of layers and the amount of computation are small, but also the effect is the best. The first two convolutional layers are configured with 5 × 5 convolution kernels, and the number of channels is set to 32 and 64. The last three convolution layers are configured with a 3 × 3 convolution kernel, and the number of channels is set to 128, 256, and 512. The stride of the first convolutional layer is set to four, two for the remaining convolutional layers. The Leaky-ReLU [44] is set with a leak of 0.2 after each layer. The two 1024-D FC layers f_{c1} and f_{c2} are fed into the feature screening module. Taking the fourth convolutional layer that retains more spatial information and larger spatial size as the extracted global feature f_i^{Conv4}, after training, the proposed network can learn vehicle features from all training data.

B. Global Feature Optimization

Unlike these existing modules, we designed a global feature optimization module containing two branches, which locates between the global feature map and global averaging pooling, to capture global features effectively, as shown in Fig. 3. The main branch consists of three feature extraction branches, a spatial channel modulation module, a concat module, and a 1 × 1 convolution module. The three feature extraction branches are composed of three parallel different-size convolution kernels, which are 1 × 1, 3 × 3, and 5 × 5. Because a large-size

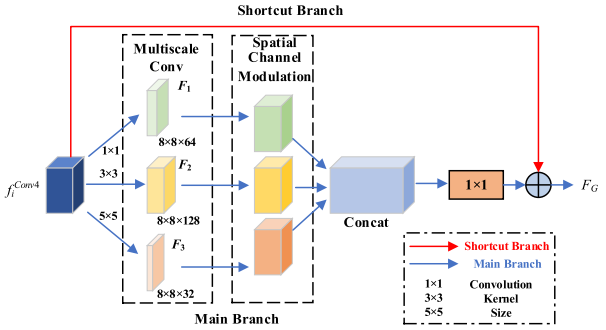


Fig. 3. Structure of feature optimization module. It consists of a main branch and a shortcut branch. The main branch further mines the semantic information of global features of vehicle by a series of convolution operations. The shortcut branch directly transmits the original global features to enhance the semantic capabilities of vehicle features by combining with the global features convoluted in the main branch.

convolution kernel involves a big receptive field [46], it can extract the whole vehicle appearance information and correlation information between components, and a small-size convolution kernel involves a small receptive field, it can focus on the texture details in the whole vehicle image. Therefore, the proposed parallel branch structure with different convolution kernel sizes can extract multiple-attribute features from the whole vehicle image, including the correlation information between image contours and between components, as well as the vehicle's texture details, which fundamentally improves the Re-ID performance.

The global feature f_i^{Conv4} extracted by CNN is input into the feature optimization module containing the main branch and shortcut branch for optimization. In the main branch, global feature f_i^{Conv4} is processed to obtain features at different scales through three different convolution kernels: 1×1, 3×3, and 5×5. The dimensions of the obtained feature maps of F_1, F_2, and F_3 are $8 \times 8 \times 64$, $8 \times 8 \times 128$, and $8 \times 8 \times 32$, respectively. However, since the reduction of the number of channels may lead to the loss of features, we add spatial channel modulation to extend the number of channels in the three feature maps of F_1, F_2, and F_3 to be consistent with that in the global feature map to avoid the loss of feature information. After the spatial channel modulation, the feature maps of F_1, F_2, and F_3 have the same dimension in the spatial dimension. Finally, they are fused by the concat operation, and then, the dimensions of the fused feature maps are reduced by 1×1 convolution to ensure that the fused feature maps are equal to the global feature maps of f_i^{Conv4} in the dimension. In addition, the feature optimization module, which combines the original global feature f_i^{Conv4} with the feature after the concat operation, makes up the damage to vehicle appearance due to deep convolution operation for global feature extraction. The optimization of global features can be expressed as

$$F_1 = \text{Conv}\left(f_i^{\text{Conv4}}, k_{1\times1}\right) \tag{1}$$

$$F_2 = \text{Conv}\left(f_i^{\text{Conv4}}, k_{3\times3}\right) \tag{2}$$

$$F_3 = \text{Conv}\left(f_i^{\text{Conv4}}, k_{5\times5}\right) \tag{3}$$

$$F_G = \text{Conv}\left(\text{Concat}(F_1, F_2, F_3), k_{1\times1}\right) \tag{4}$$

where F_1, F_2, and F_3 are features obtained by the convolution kernel with size of 1×1, 3×3, and 5×5, respectively, and F_G is the global feature after optimization.

Note that the activation function ReLU is not activated immediately after the last convolutional layer (the last convolutional layer is different from others, which is activated immediately after generating the output feature matrix). It is activated after fusing with f_i^{Conv4} passed by the shortcut branch.

We add BatchNorm after each layer of convolution to accelerate the training.

The optimized global feature F_G is fed into the average-pooling layer, which consists of a 1×1 convolution reduction block, a batch normalization (BN) layer, and an activation function ReLU to reduce the feature dimension. Next, the global feature vector of $1 \times 1 \times 512$ is input to the softmax layer for vehicle identity prediction under the constraint of the ID loss. The formula for the ID loss function is given as follows:

$$L_{\text{ID}} = \sum_{i=1}^{N} y_i \log\left(\hat{y}_i\right) \tag{5}$$

where N is the number of vehicle identities in the datasets, y_i is the ground-truth label for sample i, and $\hat{y}_i$ is the prediction probability. In (5), $\hat{y}_i = 1$ if the predicted identity of the sample is equal to the true identity; otherwise, $\hat{y}_i = 0$. In addition, the global features are also optimized by a triplet loss-based [41] function

$$L_{\text{Tri}}^{G} = \text{Max}(D_{a,p} - D_{a,n} + m, 0) \tag{6}$$

where $D_{a,p}$ represents the Euclidean distance between vehicle samples with the same identity, $D_{a,n}$ is opposite, and m is a margin parameter.

C. Local Feature Enhancement

Since vehicles may look similar from the same perspective, it is difficult to identify with global features alone. The vehicle logos, lights, windows, rearview mirrors, and other local regions are highly recognizable, which can provide reliable feature information for vehicle matching [6]. Considering that the methods of determining local areas in advance will consume more computing resources and few vehicle samples will lead to low matching accuracy, this study proposes the MAM based on meta-learning and attention mechanism to deal with the problem of few samples of vehicles. First, a meta-learning-based feature screening module is used to screen vehicle features so that the trained module can retain the local component information of the vehicle and reduce the dependence on samples. Then, based on the screened features, the built-in meta-weight generator of L-ATT is used to adaptively assign the corresponding weights to the local regions of the vehicle to enhance the salient local features.

Feature Screening: This selects M features from the existing N features, aiming to identify the most relevant features from the original features. Therefore, through the feature screening module based on meta-learning, we can not only better retain local features but also reduce the dimension of local feature

vectors. Besides, the most significant thing is to reduce the network's dependence on samples [18].

The feature screening module based on meta-learning proposed in this article is mainly composed of mutual information and Shell sort algorithm. First, we calculate the relevance of each feature on the last two FC layers f_{c1} and f_{c2} of the F-Net with dimension 1024 by using mutual information. The calculation of the relevance can be expressed by (4)

$$I\left(\alpha_l^1, \alpha_j^2\right) = D_{\mathrm{KL}}\left(p\left(\alpha_l^1, \alpha_j^2\right) \| p\left(\alpha_l^1\right) \otimes p\left(\alpha_j^2\right)\right) \tag{7}$$

where α_l^1 and α_j^2 represent the lth and jth features of the two feature vectors of the FC layers f_{c1} and f_{c2}, respectively, $p(\alpha_l^1, \alpha_j^2)$ is the joint distribution of α_l^1 and α_j^2, and $p(\alpha_l^1) \otimes p(\alpha_j^2)$ represents the product of the two feature distributions. $\mathrm{D_{KL}}$ represents the KL divergence. Note that $I(\alpha_l^1, \alpha_j^2) \in [0, 1]$, and when α_l^1 and α_j^2 are more relevant, the score of $I(\alpha_l^1, \alpha_j^2)$ is larger, indicating that there is a relevance between the two features, which is higher; otherwise, they are not related. From this, we can get the relevance score between features, indicated as (5)

$$I = [I_1, I_2, \ldots, I_{1024}]. \tag{8}$$

To facilitate the generation of the attention maps, the feature vector of the input L-ATT is required to keep consistent with the channel number of the optimized global feature F_G. Therefore, we select the features corresponding to the top 256 scores from the relevance score I as input. First, we rank the relevance score I, and considering the efficiency of the module, we use the Shell sort algorithm to rank the feature relevance scores, which does not require a lot of auxiliary space and performs well on moderately sized data. Then, we screen the 256 relevance features as local features $f_i^r \in \mathbb{R}^{1\times1\times256}$. After screening, f_i^r retains the local information of the vehicle to the greatest extent.

Local Attention Learning: The L-ATT module relies on the adaptive attention score learned by the attention mechanism and characters the importance of different local regions by assigning different weights to different local regions. Based on the generated weight vector ω_i and optimized global feature F_G, the weighted attention map F_M can be obtained, which can be expressed as

$$F_M = \omega_i \otimes F_G. \tag{9}$$

The specific implementation process of the L-ATT is summarized as follows.

1) First, the L-ATT has a built-in meta-weight generator G_ω, which uses the above-screened local feature vector $f_i^r \in \mathbb{R}^{1\times1\times256}$ as the input of it. Through the G_ω, all input local feature vectors f_i^r are weighted and averaged by (10).
2) The importance of different local regions is learned through softmax to obtain attention scores, with larger values emphasizing important salient local regions and smaller values representing less important local regions

$$H_i = \tanh\left(W_i f_i^r\right) \tag{10}$$

$$\omega_i = \mathrm{soft\,max}(W_h H_i) \tag{11}$$

where H_i represents the weighted intermediate parameter and ω_i is the generated weight.

Algorithm 1 Feature Screening and Attention Learning

Require: Output full connection layer f_{c1} and f_{c2} by F-Net.
1.Calculate the relevance of features in f_{c1} and f_{c2} by (7).
2.Get relevance score$I = [I_1, I_2, \ldots\ldots, I_{1024}]$.
3.The relevance score I is sorted by the Shell Sort algorithm, and the 256 features with the highest relevance score are screened as the local region feature vector f_i^r.
4.Train MAM to obtain attention weight ω_i.
5.Get attention map F_M.
6.Return F-Net.

To improve the accuracy of vehicle Re-ID when vehicles have a similar appearance, we use the triple loss to optimize the distance between features in the embedded feature space for local branch. The triple loss can try to reduce the distance between vehicle samples with the same identity while increasing the distance between samples from different vehicles, and the margin parameter m is set to 0.3. The loss function of the local branch is given as follows:

$$L_{\mathrm{Tri}}^L = \mathrm{Max}(D_{a,p} - D_{a,n} + m, 0). \tag{12}$$

Algorithm 1 shows the MAM learning process. The total loss function of DB-NET network is the sum of local branch loss and global branch loss

$$L = L_{\mathrm{ID}} + L_{\mathrm{Tri}}^L + L_{\mathrm{Tri}}^G. \tag{13}$$

IV. Veri-FS Dataset

The limited spatiotemporal span of existing datasets is not suitable for training and verification of few-shot vehicle Re-ID methods. To address this issue, a new small-scale dataset called Veri-FS is constructed. In real traffic scenes, the vehicle samples have a large spatiotemporal span, which leads to significant changes in the perspective information, illumination conditions, and background of the vehicles. The Veri-FS dataset accounts for the special needs for few-shot vehicle Re-ID in terms of the spatiotemporal span, perspective, illumination condition, and background. It adds 23 183 pictures of 964 vehicles taken in real-world scenes, in addition to the directly selected vehicle samples in the existing datasets. Among them, the training set contains 17 215 images, the test set contains 5179 images, and the validation set contains 789 images. Compared with the existing datasets, vehicles with the same ID have fewer images under similar perspectives. In addition, the Veri-FS dataset reflects the complexity of real-world monitoring scenes, and the samples contain different types and colors, which makes the dataset more diverse.

A. Data Collection and Annotation

Our Veri-FS dataset contains vehicle images captured by multiple cameras distributed in Nanjing, China, during daytime and night. We mainly choose different areas and use multiple cameras to capture images of vehicles traveling in that area. First, we perform vehicle detection on the captured video to obtain all bounding boxes that contain the

vehicle. The bounding box is a rectangle that encloses the entire body of the vehicle. Then, the captured vehicles are classified according to their license plates and are manually preprocessed to ensure the reliability of the selected pictures. In addition, the license plates of the vehicles are blocked to protect privacy. Finally, the vehicle samples are annotated carefully. To ensure the diversity and complexity of the dataset, we also select some vehicle images from the existing vehicle Re-ID datasets, including VehicleID, VeRi-776, and CityFlow.

B. Dataset Analysis

The main characteristics of the Veri-FS dataset are summarized as follows.

1) *Few Samples of Vehicles With the Same ID:* Most of the existing datasets contain more training samples so that vehicles with the same ID have 30–80 images with repeated perspectives. However, vehicles may deliberately avoid the monitoring areas, resulting in the problems of fewer vehicle types, a single perspective, and a small amount of data in the collected images. The existing datasets cannot character such real-world scenes. We have reduced the number of images of vehicles with the same ID, which only contains 20 images with different perspectives.
2) *Diverse Illumination Conditions:* Most of the existing datasets only obtain images during the daytime. For example, VeRi-776 selects the vehicles collected from 4:00 P.M. to 5:00 P.M. Considering the large temporal span of vehicles, our dataset includes vehicles collected not only in the daytime but also at night. In addition, we add vehicle images collected in bad weather to enrich the diversity of the dataset.
3) *Diverse Shooting Scenes:* Our dataset covers a variety of scenes, such as urban roads, crossroads, signalized areas, and suburban areas. Considering that vehicles may appear in various places, causing a relatively large spatial span, we capture images of the same vehicle when it appears in different scenes. This fully shows that the collected data cover a wide range and diverse scenes.
4) *Rich Attribute Annotations:* In the Veri-FS dataset, abundant attribute information is provided for the vehicles, as shown in Fig. 4. Including ten vehicle colors (white, black, silver, and so on) and ten vehicle types (cars, SUVs, vans, MPVs, passenger cars, trucks, and so on), these attributes can be used as auxiliary information to enhance the feature representation.

V. Experiments

Comparisons with the state-of-the-art vehicle Re-ID methods and ablation studies are performed on the VehicleID, VeRi-776, and proposed Veri-FS datasets. To evaluate the performance of the DB-Net, this study uses two indicators as evaluation indicators: mean average precision (mAP) and cumulative matching characteristics (CMC).

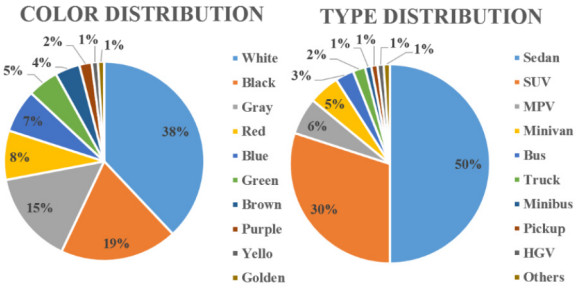


Fig. 4. Distribution of vehicle colors and types. The left one indicates the proportion of the vehicles with ten colors in the Veri-FS dataset, and the right one shows the proportion of the vehicles with ten types.

A. Datasets and Experimental Settings

In the experiments, the proposed method is first compared with several state-of-the-art vehicle Re-ID methods on two widely used datasets, VehicleID and VeRi-776. Through these two datasets, we can verify the performance of our method under a large amount of vehicle samples. Then, to further verify the performance of the proposed network with few samples, comparative experiments with traditional methods are conducted on the newly constructed Veri-FS dataset. Finally, ablation experiments on Veri-FS and VeRi-776 are performed to prove the effectiveness of individual modules in DB-Net.

Before the experiment, all vehicle images are resized to 256×256 pixels. Training of the F-Net and MAM is stopped after 40 and 50 epochs, respectively, when the losses converge to a stable value. During the network training, in order to maintain the stability and convergence speed of stochastic gradient descent (SGD) convergence, a momentum of 0.9 is used in the SGD optimizer. For the weight attenuation strategy, we set the basic learning rate to $2e^{-3}$ and gradually decrease to $2e^{-4}$ and $2e^{-5}$ in the 120th epoch and 180th epoch for faster convergence, respectively. To ensure the fairness of the experimental results, we set the training batch size to the common 64 and the total epochs are 220 times. We also enhanced the data by randomizing the horizontal flips and the erasure [38].

B. Evaluation Metric

To evaluate the performance of the DB-Net, this study uses two indicators as evaluation metrics: mAP and CMC. The average precision for each query q can be calculated as

$$\mathrm{AP}(q) = \frac{\sum_{k=1}^{n} P(k) \times \mathrm{rel}(k)}{N_{gt}} \tag{14}$$

where $P(k)$ denotes the accuracy of the first k; $\mathrm{rel}(k)$ is an indicator function, if the item at rank k is a matched vehicle image, it equals 1; otherwise, it is 0; n is the number for retrieval; and N_{gt} denotes the number of ground truths. The following equation calculates the mAP of all queried images:

$$\mathrm{mAP} = \frac{\sum_{q=1}^{Q} \mathrm{AP}(q)}{Q} \tag{15}$$

where Q is the number of all query images in the query set.

The metric CMC represents the probability of finding the correct result among the first k retrieval results, which can

be expressed as (16); rel(k) is equal to 1 when the correct matching target of image q is ranked among the first k retrieval results

$$\text{CMC@K} = \frac{\sum_{q=1}^{Q} \text{rel}(k)}{Q}. \tag{16}$$

C. Comparison With the State-of-the-Art Methods

1) Experiments on VehicleID Dataset: We compare the proposed method with the state-of-the-art methods on the three test subsets (small, medium, and large) of the VehicleID dataset in terms of CMC@1 and CMC@5. Table VII shows the comparison results with different sizes. As shown in Table VII, the proposed method achieves 85.6% and 97.13% on the small test subset. The performance outperforms other state-of-the-art methods. Among these methods, Self-supervised Attention for Vehicle Re-identification (SAVER) [33] learned important global features by exploiting preprocessing of all vehicle images to remove background noise. Unlike SAVER, deep relative distance learning (DRDL) [5] utilized a deep CNN to learn visual global appearance features. However, these two methods based on global features may ignore local features. Therefore, it is hard to perceive the importance of local information. What is more, VAMI [8] applied the viewpoint information of the vehicle, but it cannot solve the problem of the subtle interinstance discrepancy of different vehicles when they share the same type and color.

Part-regularized near-duplicate (PRN) [13] used the local regions of the vehicle, but not all regions learned by PRN can provide important identification information. In addition, Self-supervised Geometric Features Discovery (SGFD) [12] used the attention mechanism to extract distinguishable key information. However, it ignores that discerning cues can appear randomly anywhere in the entire vehicle image under the cross-camera. Disentangled feature network (DFNet) [37] applied disentangled feature learning to adaptively learn the orientation-specific and orientation-common features of the vehicle, but it consumed a lot of computing resources when dealing with matching pairs in different orientations. In contrast, we use feature optimization to improve the expressive power of global features and enhance them by adaptively assigning different weights according to the importance of different local regions through MAM, thereby obtaining more salient key local regions. All of these help improve the accuracy of vehicle Re-ID.

2) Experiments on the VeRi-776 Dataset: We also evaluate our method on the VeRi-776 dataset. Three metrics are used, such as mAP, CMC@1, and CMC@5. Table I presents the comparison results of DB-Net and the state-of-the-art methods. Our proposed DB-Net achieves good results with 81.72% mAP, surpassing the other comparative methods. In these methods, FACT [6] used the license plate and spatiotemporal information to realize vehicle Re-ID. However, FACT requires datasets to contain a large number of additional annotations, thereby consuming a lot of time and computing resources. Besides, part-attention and mentored network (PMNet) [45] distills vehicle part-specific features from part-attention network (PANet) and performs multiscale global-part feature extraction. Structured graph attention (SGAT) [41] relied on

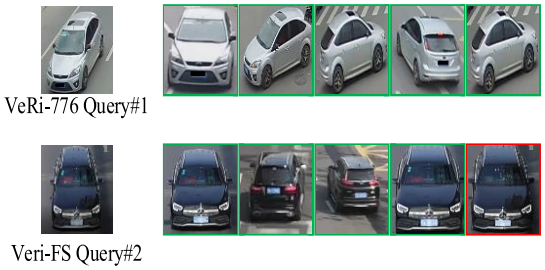


Fig. 5. Visualized detection results. The query images are shown in the first column. The top-5 images are shown in the right columns, in which the correctly matched images are marked in green boxes.

structural relationships between vehicle key points and images for Re-ID. However, when the appearance of the vehicle changes drastically under different viewpoints, this method is hard to obtain the structural relationships.

In addition, PRN [35] used the preset bounding boxes to obtain recognizable local regions to select local features, such as lights, signs, and grids. However, different local regions lead to inaccurate detection due to viewpoint changes. DFNet [37] incorporated orientation features and achieved matching of the same vehicles through an adaptive matching scheme. In contrast, our method improves the quality of features through a feature screening module when extracting vehicle local features, which can retain features with high relevance. In addition, the importance of local regions is learned through the L-ATT module and fused with the optimized global features to obtain key local regions.

3) Experiments on Veri-FS Dataset: In order to verify the performance of our DB-Net in the case of large spatiotemporal span and few vehicle samples with the same ID, we conduct experiments on the proposed Veri-FS dataset. We use mAP, CMC@1, and CMC@5 as metrics. Table II shows the experimental results on the Veri-FS dataset. Based on the Veri-FS dataset, our proposed method can reach 71.86% in terms of mAP. We also perform the open-source codes provided by recent methods [8], [21], [27], [34], [40], [44] to compare with the proposed method on the Veri-FS dataset.

To be fair, we conduct experiments under the same hardware environment. According to the experimental results provided in Table II, the above methods have low accuracy. Further analyses find that they all rely on a large amount of data to ensure the accuracy of the networks. After expanding the spatiotemporal span and reducing the repeated samples of vehicles with the same ID, their accuracy is reduced on the Veri-FS dataset. Our method outperforms methods that only obtain complete vehicle features, such as TransReID and TBE-Net, with 3.51% and 2.83% improvement, respectively, and surpasses PVEN based on the attention mechanism by 4.53%. In contrast, our method separately improves the expressiveness and quality of features through different branches, thereby effectively reducing the network's dependence on samples. Combining with an attention mechanism helps to improve the accuracy of vehicle Re-ID.

D. Ablation Experiment

To verify the proposed DB-Net, we conduct ablation experiments to explore the performance of the feature optimization

TABLE I

COMPARISON OF ACCURACY ON THREE TEST SUBSETS OF VEHICLEID DATASET IN TERMS OF CMC@1 AND CMC@5

Settings	Small		Medium		Large	
Method	CMC@1	CMC@5	CMC@1	CMC@5	CMC@1	CMC@5
VAMI[8]	63.12	83.25	52.87	75.12	47.34	70.29
DRDL[6]	49.0	73.5	42.8	66.8	38.2	61.6
FDA-Net[38]	65.91	86.15	59.84	77.09	55.53	74.65
TAMR[39]	79.71	-	76.80	-	73.87	-
EALN[24]	75.11	88.09	71.78	83.94	69.30	81.42
PRN[35]	78.92	94.81	74.94	92.02	71.58	88.46
AAVER[29]	74.7	93.8	68.6	90.0	63.5	85.6
SAVER[33]	79.90	95.20	77.60	91.10	75.30	88.30
Strong Baseline[40]	79.53	95.27	76.52	91.11	73.31	87.34
DFNet[37]	84.76	96.22	80.61	94.10	79.15	92.86
SGFD[12]	80.8	93.7	-	-	-	-
PMNet[45]	85.2	97.5	-	-	-	-
DB-Net (ours)	**86.32**	**97.61**	**82.47**	**95.98**	**81.24**	**93.85**

TABLE II

COMPARISON OF ACCURACY ON THE VERI-776 DATASET IN TERMS OF MAP, CMC@1, AND CMC@5

Method	mAP	CMC@1	CMC@5
FACT[6]	18.5	51.0	73.5
VAMI+STR[34]	61.32	85.92	91.84
AAVER[29]	61.20	89.00	94.70
VANet[26]	66.34	89.78	95.99
SGAT[11]	65.5	89.6	-
PRN[35]	74.30	94.34	98.91
PVEN[27]	79.50	95.62	98.43
SAVER[33]	79.60	96.40	98.61
SPAN[42]	68.9	94.0	97.6
V2I-CARLA[43]	78.03	93.84	-
DFNet[37]	80.97	97.08	99.01
SGFD[12]	81.0	96.7	98.6
PMNet[45]	81.5	96.8	98.6
DB-Net (ours)	**81.72**	**97.94**	**99.15**

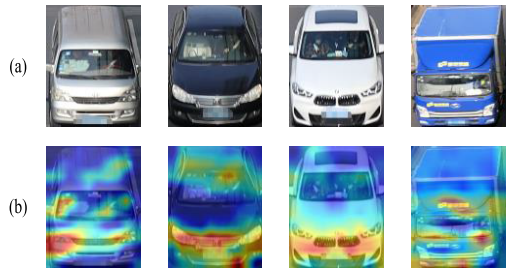


Fig. 6. Original image and heatmaps of query vehicles. (a) Original images. (b) Corresponding heatmaps. After feature enhancement, the Re-ID model assigns greater weights to local areas with important discriminative cues, such as vehicle lights, windows, and brand logos.

module and MAM on VeRi-776 and Veir-FS datasets. Fig. 5 shows the retrieval results of our DB-Net. Fig. 6 shows the heatmaps of query vehicles. The identifiable local regions are highlighted, such as decorations and drivers.

1) Effectiveness of Feature Optimization Module: We conduct ablation experiments on VeRi-776 to verify the effectiveness of the feature optimization module in the global feature branch. The experiment removes the feature optimization module of the network and directly uses the features extracted from the fourth layer of the feature extraction network f_i^{Conv4} as the global features. The experimental results are shown in Table III. f_i^{Conv4} represents the global feature

TABLE III

COMPARISON OF ACCURACY ON THE VERI-FS DATASET IN TERMS OF MAP, CMC@1, AND CMC@5

Method	mAP	CMC@1	CMC@5
VAMI[8]	50.80	84.63	94.84
Fast-reid[44]	69.15	93.55	96.78
PVEN[27]	67.33	88.68	95.34
Strong Baseline[40]	67.56	88.79	95.62
TransReID[34]	68.74	90.74	96.53
TBE-Net[44]	69.42	92.41	97.72
DB-Net (ours)	**72.25**	**94.14**	**98.71**

TABLE IV

ACCURACY COMPARISON FOR GLOBAL FEATURE OPTIMIZATION MODULE ON THE VERI-776 DATASET IN TERMS OF MAP, CMC@1, AND CMC@5

Settings	mAP	CMC@1	CMC@5
Baseline	75.12	91.27	93.45
f_i^{Conv4}	77.84	96.54	98.12
$\boldsymbol{F_G}$	**81.72**	**97.94**	**99.15**

before optimization, and F_G represents the global feature after feature optimization. It can be found that by adding the feature-optimized network, the overall performance is improved by 3.88%. This means that the feature optimization network helps to improve the expressive power of global features and it can also improve the recognition accuracy of the network.

2) Effectiveness of Feature Screening Module: To verify the effectiveness of the feature screening module in the local branch, we conduct the following ablation experiments on the VeRi-776 dataset. In the experiment, the feature screening module is removed first, and the feature dimension reduction measure [8] is taken directly so that the feature vector $f_i \in \mathbb{R}^{1\times1\times256}$ is used as the input of the L-ATT. f_i represents the feature vector after dimensionality reduction and fir represents the feature after feature screening. The results in Table IV show that the mAP is improved by 3.93% by adding the network after feature screening. This proves that feature screening is better than simple dimensionality reduction operations. Through screening, the network accuracy can be improved, while the feature quality is improved.

TABLE V

ACCURACY COMPARISON FOR FEATURE SCREENING MODULE ON THE VERI-776 DATASET IN TERMS OF MAP, CMC@1, AND CMC@5

Settings	mAP	CMC@1	CMC@5
Baseline	75.12	91.27	93.45
f_i [8]	77.79	96.54	98.12
$\boldsymbol{f_i^r}$	**81.72**	**97.94**	**99.15**

TABLE VI

ACCURACY COMPARISON FOR FEATURE SCREENING MODULE ON THE VERI-FS DATASET IN TERMS OF MAP, CMC@1, AND CMC@5

Settings	mAP	CMC@1	CMC@5
Baseline	66.39	90.27	92.64
f_i [4]	67.85	92.54	95.12
$\boldsymbol{f_i^r}$	**72.25**	**94.14**	**98.71**

TABLE VII

ACCURACY COMPARISON FOR L-ATT MODULE ON THE VERI-FS DATASET IN TERMS OF MAP, CMC@1, AND CMC@5

Settings	mAP	CMC@1	CMC@5
Baseline	66.39	90.27	92.64
f_i^r	65.84	89.45	91.21
$f_i^r + F_M$	67.63	90.82	95.35
$f_i^r + F_G$	69.57	92.47	97.61
$\boldsymbol{f_i^r + F_G + F_M}$	**72.25**	**94.14**	**98.71**

In order to further verify the role of feature screening when there are few vehicle samples, we also conducted the corresponding ablation experiments on the Veir-FS dataset. The experimental results are shown in Table V. After the feature screening module is added to the network, the mAP of the network is 4.01% higher than that without feature screening module. Therefore, the above ablation experiments show that the meta-learning-based feature screening module can effectively improve the feature quality, thereby reducing the network's dependence on samples and significantly improving the performance of few-shot vehicle Re-ID.

3) Effectiveness of L-ATT: Besides, we also investigate the effect of the L-ATT module on vehicle Re-ID during local feature learning. We conduct ablation experiments on it and summarize the results in Table VI. We first remove the attention module from the network and compare the local feature screening with or without the influence of L-ATT. By comparing the methods based on fir and $f_i^r + F_M$, where F_M indicates that the network has an L-ATT module, the experimental results show that removing the L-ATT will degrade the accuracy of vehicle re-recognition, with a maximum reduction of 4.40% in mAP. This shows that the L-ATT module can effectively improve the accuracy of vehicle Re-ID.

To further verify the effectiveness of the attention module, we add the optimized global feature FG based on the previous step. By exploring the two methods based $f_i^r + F_G$ and $f_i^r + F_G + F_M$, the method with L-ATT module in mAP is improved by 2.68%, compared to the method without it. The results of ablation experiments show that the L-ATT module can effectively enhance local regions.

VI. CONCLUSION

This article proposes a DB-Net with enhanced global and local features to address the challenge of few-shot vehicle samples in realistic traffic scenes. The global branch mines the deep semantic information of global features through the feature optimization module and preserves the original spatial size and spatial information. At the same time, the MAM adopted by the local branch improves the quality of features through screening to reduce the module's dependence on samples. Meanwhile, the local branch improves feature quality through feature screening, and the L-ATT module assigns weights to enhance extracted features, maintaining good Re-ID even with scarce samples. We also introduce the Vert-FS dataset, characterizing real traffic scenes with few samples and large spatiotemporal span of vehicles. Extensive experiments demonstrate the effectiveness of the proposed DB-Net.

For future research, local information maximization [16] can be used to generate better attention weights, instead of screening features in this article, for enhanced discriminative cues in few-shot vehicle Re-ID. In addition, lightweight research on the proposed network can enable its application on mobile devices, such as unmanned aerial vehicles, for high-efficiency real-time vehicle Re-ID.

REFERENCES

[1] S. Yang, L. Li, S. Wang, W. Zhang, Q. Huang, and Q. Tian, "SkeletonNet: A hybrid network with a skeleton-embedding process for multi-view image representation learning," *IEEE Trans. Multimedia*, vol. 21, no. 11, pp. 2916–2929, Nov. 2019.

[2] H. Guo, C. Zhao, Z. Liu, J. Wang, and H. Lu, "Learning coarse-to-fine structured feature embedding for vehicle re-identification," in *Proc. AAAI Conf. Artif. Intell.*, Apr. 2018, vol. 32, no. 1, pp. 6853–6860.

[3] P. Angelo, B. Luca, and C. Simone, "Robust re-identification by multiple views knowledge distillation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 93–110.

[4] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2017, pp. 23–30.

[5] H. Liu, Y. Tian, Y. Wang, L. Pang, and T. Huang, "Deep relative distance learning: Tell the difference between similar vehicles," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2167–2175.

[6] X. Liu, W. Liu, T. Mei, and H. Ma, "A deep learning-based approach to progressive vehicle re-identification for urban surveillance," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 869–884.

[7] Z. Tang et al., "CityFlow: A city-scale benchmark for multi-target multi-camera vehicle tracking and re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8797–8806.

[8] Y. Zhouy and L. Shao, "Viewpoint-aware attentive multi-view inference for vehicle re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6489–6498.

[9] C. Wang, Q. Zhang, C. Huang, W. Liu, and X. Wang, "Mancs: A multi-task attentional network with curriculum sampling for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 4226–4235.

[10] H. Li et al., "Attributes guided feature learning for vehicle re-identification," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 6, no. 5, pp. 1211–1221, Oct. 2022.

[11] D. Meng et al., "Parsing-based view-aware embedding network for vehicle re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7101–7110.

[12] M. Li, X. Huang, and Z. Zhang, "Self-supervised geometric features discovery via interpretable attention for vehicle re-identification and beyond," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 194–204.

[13] B. He, J. Li, Y. Zhao, and Y. Tian, "Part-regularized near-duplicate vehicle re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3992–4000.

[14] W.-T. Chen et al., "RVSL: Robust vehicle similarity learning in real hazy scenes based on semi-supervised learning," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 23–27.

[15] Y. Lou, Y. Bai, J. Liu, S. Wang, and L. Duan, "Embedding adversarial learning for vehicle re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 8, pp. 3794–3807, Aug. 2019.

[16] T. Bluche, "Joint line segmentation and transcription for end-to-end handwritten paragraph recognition," in *Proc. Adv. Neural Inf. Proc. Syst.*, 2016, pp. 838–846.

[17] M. A. Jamal and G. Qi, "Task agnostic meta-learning for few-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11719–11727.

[18] S. Ravi and H. Larochelle, "Optimization as a model for few-shot learning," in *Proc. Int. Conf. Learn. Represent.*, 2017, pp. 1–11.

[19] W. Chen, Y. Liu, Z. Kira, Y. Wang, and J. Huang, "A closer look at few-shot classification," in *Proc. IEEE Int. Conf. Learn. Represent.*, May 2019, pp. 1–16.

[20] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. S. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1199–1208.

[21] L. He, X. Liao, W. Liu, P. Cheng, and T. Mei, "FastReID: A PyTorch toolbox for general instance re-identification," *CoRR*, vol. abs/2006.02631, p. 8, Jun. 2020.

[22] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 4080–4090.

[23] R. Kuma, E. Weill, F. Aghdasi, and P. Sriram, "Vehicle re-identification: An efficient baseline using triplet embedding," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2019, pp. 1–9.

[24] X. Liu, W. Liu, T. Mei, and H. Ma, "PROVID: Progressive and multimodal vehicle reidentification for large-scale urban surveillance," *IEEE Trans. Multimedia*, vol. 20, no. 3, pp. 645–658, Mar. 2018.

[25] Y. Bai, Y. Lou, F. Gao, S. Wang, Y. Wu, and L. Duan, "Group-sensitive triplet embedding for vehicle reidentification," *IEEE Trans. Multimedia*, vol. 20, no. 9, pp. 2385–2399, Sep. 2018.

[26] R. Chu, Y. Sun, Y. Li, Z. Liu, C. Zhang, and Y. Wei, "Vehicle re-identification with viewpoint-aware metric learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8281–8290.

[27] Z. Wang et al., "Orientation invariant feature embedding and spatial temporal regularization for vehicle re-identification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 379–387.

[28] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[29] P. Khorramshahi, A. Kumar, N. Peri, S. S. Rambhatla, J. Chen, and R. Chellappa, "A dual-path model with adaptive attention for vehicle re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6131–6140.

[30] W. Sun, L. Dai, X. Zhang, P. Chang, and X. He, "RSOD: Real-time small object detection algorithm in UAV-based traffic monitoring," *Appl. Intell.*, vol. 52, no. 8, pp. 8448–8463, Jun. 2022.

[31] O. Russakovsky et al., "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 2385–2399, 2015.

[32] M. Yang and P. Yang, "A novel condensing tree structure for rough set feature selection," *Neurocomputing*, vol. 71, nos. 4–6, pp. 1092–1100, Jan. 2008.

[33] P. Khorramshahi, N. Peri, J.-C. Chen, and R. Chellappa, "The devil is in the details: Self-supervised attention for vehicle re-identification," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 369–386.

[34] S. He, H. Luo, P. Wang, F. Wang, H. Li, and W. Jiang, "TransReID: Transformer-based object re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 15013–15022.

[35] K. Yan, Y. Tian, Y. Wang, W. Zeng, and T. Huang, "Exploiting multi-grain ranking constraints for precisely searching visually-similar vehicles," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 562–570.

[36] Z. Sun, X. Nie, X. Xi, and Y. Yin, "CFVMNet: A multi-branch network for vehicle re-identification based on common field of view," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 3523–3531.

[37] Y. Bai, J. Liu, Y. Lou, C. Wang, and L. Duan, "Disentangled feature learning network and a comprehensive benchmark for vehicle re-identification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 6854–6871, Oct. 2022.

[38] Y. Lou, Y. Bai, J. Liu, S. Wang, and L. Duan, "VERI-wild: A large dataset and a new method for vehicle re-identification in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3230–3238.

[39] J. Qian, W. Jiang, H. Luo, and H. Yu, "Stripe-based and attribute-aware network: A two-branch deep model for vehicle re-identification," 2019, *arXiv:1910.05549*.

[40] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, "Bag of tricks and a strong baseline for deep person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 1487–1495.

[41] Y. Zhu, Z.-J. Zha, T. Zhang, J. Liu, and J. Luo, "A structured graph attention network for vehicle re-identification," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 646–654.

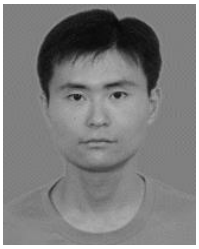
[42] T.-S. Chen, C.-T. Liu, C.-W. Wu, and S.-Y. Chien, "Orientation-aware vehicle re-identification with semantics-guided part attention network," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 330–346.

[43] H. Wang, X. Yuan, Y. Cai, L. Chen, and Y. Li, "V2I-CARLA: A novel dataset and a method for vehicle reidentification-based V2I environment," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–9, 2022.

[44] W. Sun, G. Dai, X. Zhang, X. He, and X. Chen, "TBE-Net: A three-branch embedding network with part-aware ability and feature complementary learning for vehicle re-identification," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 9, pp. 14557–14569, Sep. 2022.

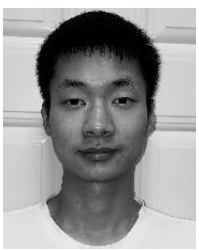
[45] L. Tang, Y. Wang, and L.-P. Chau, "Weakly-supervised part-attention and mentored networks for vehicle re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 12, pp. 8887–8898, Dec. 2022.

[46] X. Ding, X. Zhang, J. Han, and G. Ding, "Scaling up your kernels to 31×31: Revisiting large kernel design in CNNs," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11963–11975.



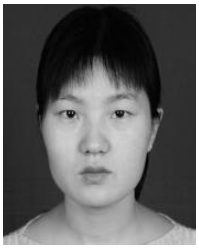
Wei Sun received the B.S. and M.S. degrees in mechanical manufacture and automation from the Henan University of Science and Technology, Luoyang, China, in 2004 and 2006, respectively, and the Ph.D. degree in instrument science and technology from Southeast University, Nanjing, China, in 2010.

From 2014 to 2015, he was a Post-Doctoral Researcher with the Next Generation Transportation Systems (NEXTRANS) Center, Purdue University, West Lafayette, IN, USA. He is currently a Professor of automation with the Nanjing University of Information Science and Technology, Nanjing. His research interests include vehicle re-identification, computer vision, deep learning, and environment perception for intelligent vehicles.



Fan Xu received the B.S. degree in electrical engineering and automation from the Nanjing University of Information Science and Technology, Nanjing, China, in 2020, where he is currently pursuing the M.S. degree in control engineering.

His research interests include large-scale vehicle retrieval and vehicle re-identification.



Xiaorui Zhang received the B.S. and M.S. degrees in mechanical manufacture and automation from the Henan University of Science and Technology, Luoyang, China, in 2004 and 2006, respectively, and the Ph.D. degree in instrument science and technology from Southeast University, Nanjing, China, in 2010.

From 2013 to 2014, she was a Post-Doctoral Researcher with the ViDi Center, University of Pennsylvania, Philadelphia, PA, USA. She is currently a Professor of computer science and technology with the Nanjing University of Information Science and Technology, Nanjing. Her research interests include virtual reality and human–computer interaction, haptic perception, and pattern recognition.

Yahua Hu received the B.S. degree in electrical engineering and automation from the Nanjing University of Information Science and Technology, Nanjing, China, in 2020, where he is currently pursuing the M.S. degree in control engineering.

His research interests include large-scale vehicle retrieval and vehicle re-identification.

Guangzhao Dai received the B.S. degree in automation from Wuxi Taihu University, Wuxi, China, in 2019. He is currently pursuing the M.S. degree in control engineering with the Nanjing University of Information Science and Technology, Nanjing, China.

His research interests include large-scale vehicle retrieval and fine-grained image recognition.

Xiaozheng He received the Ph.D. degree from the University of Minnesota, Twin Cities, Minneapolis, MN, USA, in 2010.

He is currently an Assistant Professor with the Department of Civil and Environmental Engineering, Rensselaer Polytechnic Institute, Troy, NY, USA. His research areas cover transportation system modeling and simulation, interdependent infrastructure resilience, and intelligent transportation systems. The research results have been published in over 90 technical articles in prestigious venues.

Dr. He was a recipient of the NSF CAREER Award. He serves as an Editorial Board Member for *Transportation Research Part B: Methodological*, an Associate Editor for *Frontiers in Future Transportation*, and a Special Issue Guest Editor for *Transportation Research Part D: Transport and Environment*, *Journal of Advanced Transportation*, and *Sustainability*.