TBE-Net: A Three-Branch Embedding Network With Part-Aware Ability and Feature Complementary Learning for Vehicle Re-Identification

Wei Sun, Guangzhao Dai, Xiaorui Zhang, Xiaozheng He, and Xuan Chen

***Abstract*— Vehicle re-identification (Re-ID) is one of the promising applications in the field of computer vision. Existing vehicle Re-ID methods mainly focus on global appearance features or pre-defined local region features, which have difficulties in handling inter-class similarities and intra-class differences among vehicles in various traffic scenarios. This paper proposes a novel end-to-end three-branch embedding network (TBE-Net) with feature complementary learning and part-aware ability. The proposed TBE-Net integrates complementary features, global appearance, and local region features into a unified framework for subtle feature learning, thereby obtaining more integral and diverse vehicle features to re-identify the vehicle from similar ones. The local region feature branch in the proposed TBE-Net contains an attention module that highlights the major differences among local regions by adaptively assigning large weights to the critical local regions and small weights to insignificant local regions, thereby enhancing the perception sensitivity of the network to subtle discrepancies. The complementary branch in the proposed TBE-Net exploits different pooling operations to obtain more comprehensive structural features and multi-granularity features as a supplement to the global appearance and local region features. The abundant features help accommodate the ever-changing critical local regions in vehicles' images due to the sensors' settings, such as the position and shooting angle of surveillance cameras. The extensive experiments on VehicleID and VeRi-776 datasets show that the proposed TBE-Net outperforms the state-of-the-art methods.**

***Index Terms*— Vehicle re-identification, attention mechanism, multi-granularity features learning, embedding.**

I. Introduction

VEHICLE re-identification (Re-ID) aims to retrieve all images of a query vehicle from a large image database captured in non-overlapping cameras in the context of traffic video monitoring. This task is challenging, specifically when license plates are obscured, blurred, or damaged. With the introduction of large datasets [1]–[5] and the advancement of Deep Convolutional Neural Networks (DCNNs) [6], the recently proposed models have achieved remarkable results, allowing vehicle Re-ID to be applied in intelligent transportation systems and smart city [7]–[13].

However, many vehicles share almost identical appearances, while the same vehicle can show its appearance from different perspectives in different surveillance systems, which introduces considerable challenges to vehicle Re-ID. Fig. 1 (a) illustrates that different vehicles may share the same color and even be from the same manufacturer. The similarity causes difficulty in distinguishing vehicles using the vehicle's overall appearance only (e.g., color, model, texture).

Local regions of vehicles, such as windows, lights, brand logo, personal decorations, carry abundant details that usually provide discriminative information [14] to distinguish similar vehicles. Therefore, recent vehicle Re-ID studies have applied the discriminative local regions to improve their accuracy. However, existing methods cannot obtain the desired accuracy because the same vehicle may present different appearances in the images captured by cameras at different locations. This situation requires the Re-ID model to have the ability to locate and identify local regions with critical discriminative clues. This situation also requires the Re-ID model to pay more attention to the critical local regions and less attention to the insignificant local regions [18], which will improve the accuracy of vehicle Re-ID while reducing the computational burden.

Vehicle Re-ID methods based on local regions [14]–[19] apply a prior knowledge to define the size and range of local regions. However, such a priori knowledge may not work

Manuscript received 23 January 2021; revised 9 June 2021, 18 August 2021, and 6 October 2021; accepted 17 November 2021. Date of publication 3 December 2021; date of current version 12 September 2022. This work was supported in part by the Natural Science Foundation of Jiangsu Province under Grant BK20191401 and Grant BK20201136; in part by the Postgraduate Research and Practice Innovation Program of Jiangsu Province under Grant SJCX21_0363; in part by the National Natural Science Foundation of China under Grant 61304205 and Grant 61502240; and in part by the Innovation and Entrepreneurship Training Project of College Students under Grant 202010300290, Grant 202010300211, and Grant 202010300116E. The Associate Editor for this article was J. Li. *(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).

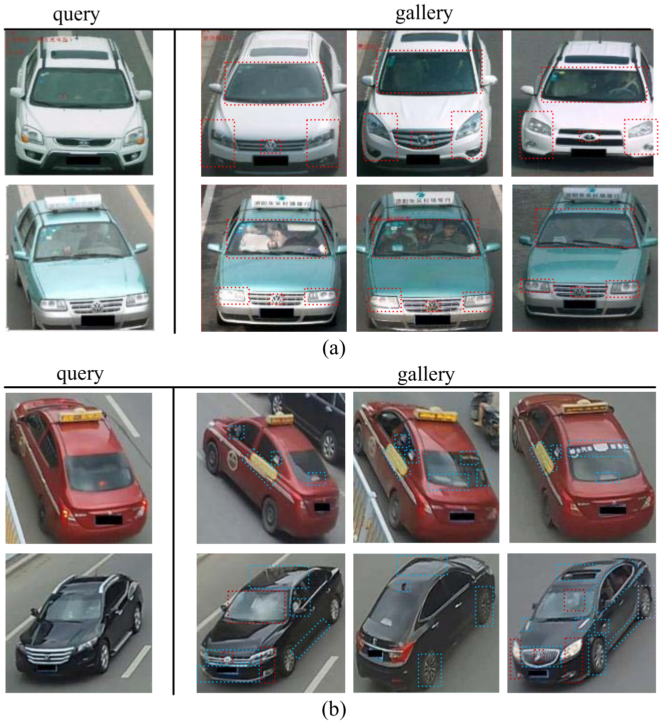
Guangzhao Dai is with the School of Automation, Nanjing University of Information Science and Technology, Nanjing 210044, China.

Xiaorui Zhang is with the School of Computer and Software and the Engineering Research Center of Digital Forensics, Ministry of Education, Nanjing University of Information Science and Technology, Nanjing 210044, China, and also with the Wuxi Research Institute, Nanjing University of Information Science and Technology, Wuxi 214100, China.

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

Xuan Chen is with the School of Computer and Software, Nanjing University of Information Science and Technology, Nanjing 210044, China.

Digital Object Identifier 10.1109/TITS.2021.3130403



(a)

(b)

Fig. 1. Challenges of vehicle Re-ID. It is difficult to differentiate vehicles simply by the global appearance in each row. (a) The vehicle images with different identities have a similar appearance and the same color, even from the same manufacturer. The red boxes in the vehicle images indicate subtle differences between these vehicles. (b) The essential clues to distinguish the four vehicle samples maybe randomly occur in any location of the vehicle image.

well in practice. When vehicles travel under non-overlapping cameras, the images captured by cameras can present their appearances and shapes differently due to viewpoint variations and occlusion, as illustrated in Fig. 1(b). These situations lead to an intractable fact, i.e., the essential clues to determine vehicle identity may appear at a random location in a vehicle image. Traditional methods based on fixed local regions using pre-defined a priori knowledge fail to meet these requirements.

To address the challenges in practice, this paper proposes a three-branch embedding network with part-aware ability and feature complementary learning, which integrates the features of three branches into a unified framework to realize end-to-end training. The global branch extracts the global appearance features of the whole vehicle. The local branch can identify and locate the local regions with discriminative clues and assign large weights to the critical local regions based on an attention mechanism. In order to adapt to the ever-changing vehicle appearance under non-overlapping cameras, this study proposes the third branch, named complementary branch, where we first divide the feature map of the backbone network into four sub-regions, and then encourage every sub-region to leverage the information of other sub-regions in addition to its own effective information. Moreover, we furthermore utilize the pooling operations with different sizes to obtain multi-granularity features. Integrating the complementary features with the global appearance and local region features can obtain a more integral and diverse representation of vehicle features and adapt to the ever-changing critical local regions in vehicle images. Our main contributions are threefold, summarized as follows:

1) The study proposes a three-branch embedding network (TBE-Net) with part-aware ability and feature complementary learning to enhance vehicle feature representation. In TBE-Net, the global branch learns the global appearance features of a vehicle, and the local branch learns the local features with subtle differences. More importantly, the complementary branch learns more abundant vehicle features by different-size pooling operations. Extensive experiments on two large-scale standard datasets, i.e., VeRi-776 [5] and VehicleID [2] demonstrate that TBE-Net outperforms state-of-the-art methods.
2) A compact and flexible local attention (L-ATT) module is proposed to determine the importance of every local region, which can accurately leverage the critical local regions to improve vehicle Re-ID performance. Through the L-ATT module, larger weights are adaptively assigned to highly important local regions to obtain more attention. In contrast, smaller weights are assigned to the less important local regions. This adaptive weight assignment helps to exploit the network's subtle perception ability to capture distinguishable clues.
3) This paper proposes a complementary branch to extract complete structural features and multi-granularity features. A new Sub-region Interaction Module (SIM) is introduced to make full use of the relations between the divided sub-regions. The divided narrow sub-regions in the feature map correlate to the corresponding whole region of the original image instead of small fixed part regions, shown in Fig. 2. Because the SIM accounts for the relations between vehicle sub-regions and every sub-region contains the effective information of other sub-regions besides its own information, the complementary branches can obtain more comprehensive structural information (e.g., symmetric relation of two lights or two rearview mirrors). In addition to the structural features, we further utilize different-size pooling operations to obtain multi-granularity features. As a supplement to global appearance and local region features, these features involve richer feature representations, which can adapt to the ever-changing local regions containing discriminative clues in vehicle images, thereby having more vital feature expression ability.

II. Related Work

A. Global Appearance Feature-Based Methods

Global appearance feature-based methods mainly focus on learning the global appearance features through training networks, such as vehicle appearance, color, texture, and orientation. Liu *et al.* [9] proposed a method based on Fusion Attributes and Color Features (FACT), which uses a convolutional neural network to integrate low-level color information and high-level semantic information for vehicle Re-ID.

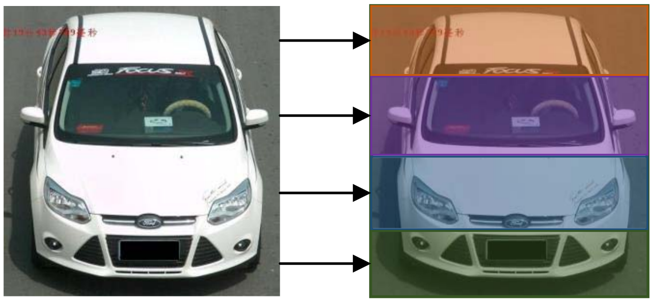


Fig. 2. Sub-region division and the mapping relationship with the original image. The left is an original vehicle image, and the right indicates four divided sub-regions.

Based on the FACT method, Liu *et al.* [1], [5] introduced a from-coarse-to-fine network framework to improve recognition performance by integrating multi-modal attributes such as vehicle color, model, texture, and spatial-temporal information. Although the recognition accuracy of these methods has been greatly improved, there is a drastic decline when the vehicle appearance changes dramatically due to low resolution, illumination variation and occlusion. Thanks to the great success of deep metric learning in image retrieval [20], [21], face verification [22]–[25], and person Re-ID [26], [27], researchers proposed deep metric learning methods [2], [3], [28], [29] to enhance identification ability of global appearance feature-based models. Liu *et al.* [2] transformed the entire vehicle image into Euclidean space to calculate the distance between vehicles and trained the neural network for vehicle Re-ID by vehicle ID and model information. Furthermore, Bai *et al.* [30] proposed the Group Sensitive Triplet embedding (GS-TRE) model, which alleviates large intra-class differences and small inter-class similarity between vehicle samples by using an intra-class deviation loss function. To solve the problems that vehicle images from different camera viewpoints vary greatly and global appearance features are difficult to align, Chu *et al.* [29] exploited a pre-trained viewpoint classifier to calculate the vehicle viewpoint, where the vehicle image is divided into s-view and d-view feature spaces and different loss functions are adopted, which improves vehicle Re-ID accuracy. Although the reviewed methods are helpful for grasping the global appearance feature of the vehicle, these models do not have the ability to perceive distinguishable details from local regions of vehicle.

B. Local Region Feature-Based Methods

Other than global appearance feature-based methods, methods based on local region features have also been introduced to enlarge the slight differences between vehicle images of the same type. Wang *et al.* [15] defined 20 key feature points of the vehicle body and clustered them according to the vehicle orientation into four directions: front, back, left, and right. The pre-defined key feature points help extract orientation-invariant features for vehicle Re-ID. Khorramshahi *et al.* [31] used a two-branch adaptive attention model, built upon [15], to extract features with locally identifiable information based on key feature points and vehicle orientations. Inspired by the region of interest, He *et al.* [14] proposed a Part Regularized Near-duplicate (PRN) network and extracted the bounding boxes of headlights, windows, and brand logo through the object detection algorithm to learn information from critical local regions. Zhang *et al.* [18] used a similar object detection algorithm to extract some critical local regions from vehicle images, such as license plate, vehicle lights, annual inspection signs, and personality decoration, as candidate searching regions during the network learning process for vehicle Re-ID. Because these methods learn pre-defined key feature points and local region locations, which helps it more accurately see the discrepancy features, they can improve the accuracy of vehicle Re-ID to some extent.

However, when vehicles travel under non-overlapping cameras, their appearance may vary obviously due to cameras' installation position and shooting angle. Therefore, the location and importance of these key feature points or local regions may constantly change under cross-camera video surveillance, limiting the application of these methods based on pre-defined fixed key feature points and local regions.

C. Attention-Based Methods

Recently, the visual attention mechanism has been applied to improve person Re-ID [32]–[34]. By the mechanism, the Re-ID models can automatically focus on the important regions of the input image and ignore the useless regions, and effectively extract discernable global and local features. To improve the neural network's ability to perceive subtle differences, recent studies have applied the attention mechanism to vehicle Re-ID to highlight critical local regions of vehicles. Khorramshahi *et al.* [31] proposed a dual-path model with adaptive attention network, which can adaptively select key feature points and vehicle orientations for extracting local identifiable features in vehicle Re-ID. However, obtaining vehicle orientation information requires deep learning algorithms, which consume significant computing resources.

In addition, Zhang *et al.* [18] proposed a Part-Guided Attention Network (PGAN). PGAN can extract the key parts from the vehicle and learn the importance of different local regions by an attention module, as well as assign appropriate weights to the corresponding local regions, thereby achieving a good vehicle Re-ID result. However, these methods [31], [18] based on pre-determined key feature points and local regions cannot adapt to the dynamic change of these key feature points and local regions under cross-camera video surveillance.

Further studies find that essential clues to determine vehicle identity may present at a random location in a vehicle image, as well as their size and shape in images constantly varies due to different camera viewpoints, changing illumination, and low resolution, which hinders traditional methods from capturing these essential clues. The work in literature [35] shows that different granularity features tend to have different information expressing abilities. Setting different-size pooling factors can obtain not only global information with large receptive fields and coarse granularities but also local detail information with small receptive fields and subtle granularities.

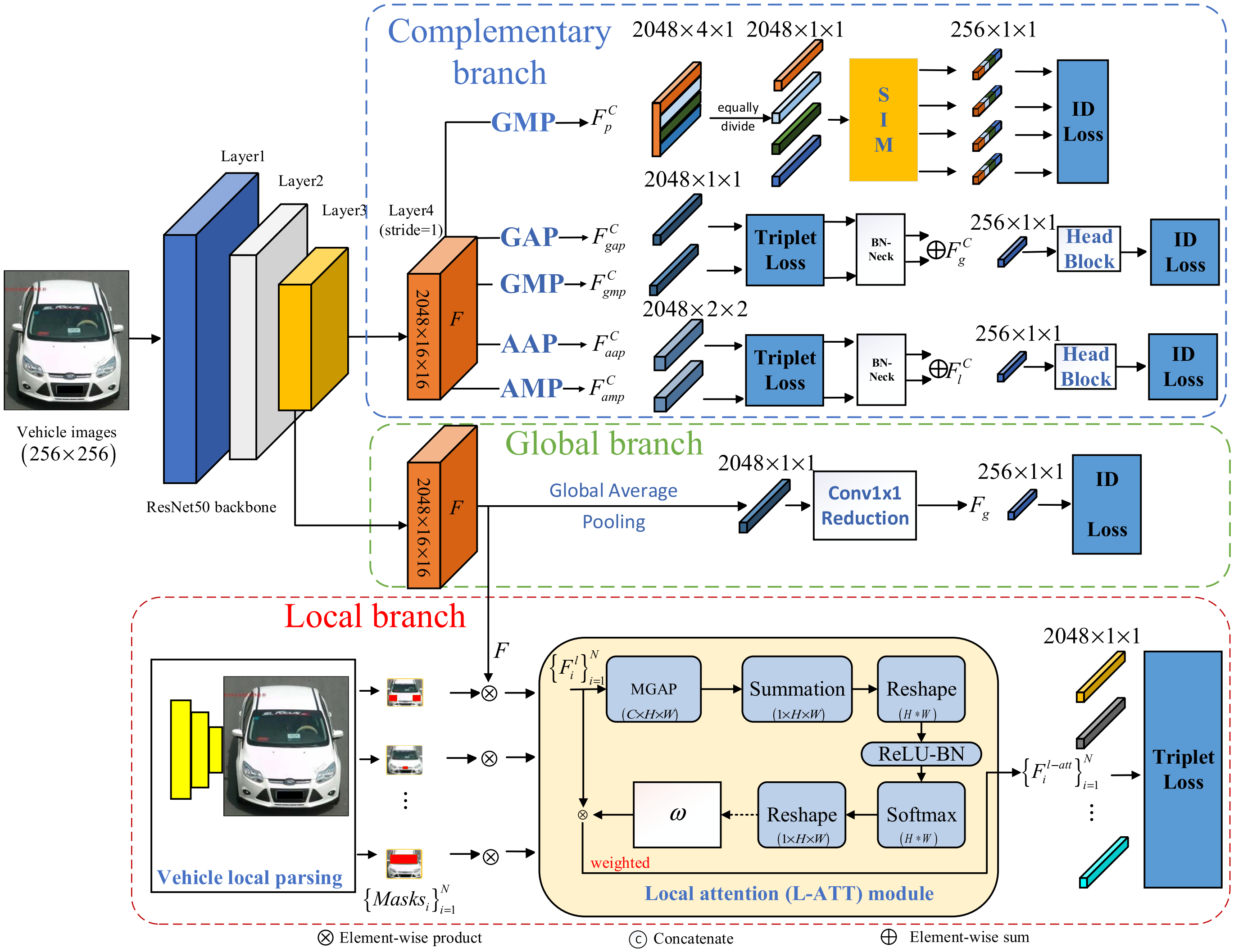


Fig. 3. Structure of the TBE-Net. Our network framework consists of three branches: global branch, local branch, and complementary branch.

If these multi-granularity features are jointly used in a Re-ID model, then the model can accommodate the ever-changing critical local regions in vehicle image, which will further enhance the robustness and accuracy of the Re-ID model. Moreover, the global average pooling operation can integrate the spatial relationship between feature maps, and the global maximum pooling can gather the salience information of the feature maps [36]. The experiments in reference [37] also demonstrate that fusing two pooling features of global pooling and global maximum pooling has a better effect than using only one pooling feature. Inspired by this observation, this paper proposes a complementary branch for extracting multi-granularity features. It adopts more pooling operations and fuses them to capture more integral and diverse vehicle features.

III. Method

Fig. 3 shows the network structure of the proposed TBE-Net. The framework consists of three branches, global branch, local branch, and complementary branch, respectively. During training stage, firstly, this study exploits a vehicle part parser to get masks of vehicle parts, by which local region features of vehicle parts can be extracted. Then, the extracted local region features are input into the L-ATT module to learn the attention scores of different parts, by which critical local regions with higher scores can receive more attention. Similarly, the global appearance features can be obtained through the global branch.

Note that, in the complementary branch, we copy the feature map of the backbone network and independently use it for feature learning. Subsequently, the study applies global max pooling to extract features, and then the pooled features are further divided into four sub-regions, by which vehicle Re-ID model can identify separately vehicle identity under the identification loss (ID loss) constraint. To obtain more discriminative and complete structural features, a new Sub-region Interaction Module (SIM) is introduced to learn the relations between the divided sub-regions and further encourage every sub-region to leverage the information of other sub-regions besides its own effective information. In addition, the study also uses different pooling operations by setting multiply pooling factors to extract multi-granularity features. By fusing these multi-granularity features according to their feature size, the TBE-Net are trained by combining triplet loss and ID loss to obtain abundant vehicle features. During the testing stage, we use the global feature extracted by the trained model for vehicle matching.

A. Global Branch Feature Learning

This paper adopts ResNet50 [6], which is pre-trained on ImageNet [38], as the feature extraction network. To obtain the feature map with larger size and more detail information, the stride of the last convolutional layer is set as 1. In the global branch, the global feature vector F_g of the vehicle is generated through the global average pooling operation. As shown in Fig. 3, the input vehicle image generates a 2048-dimensional feature map F with 16×16 shape through the feature extraction network. The feature map F is then applied to three branches for feature learning, namely global branch, local branch, and complementary branch. In the global feature learning, F is fed into the average pooling layer, and it is followed a Conv1 × 1 reduction block consisting of a 1×1 convolution layer, a batch normalization (BN) layer, and a rectified linear unit (ReLU) to reduce feature dimension. Next, the reduced feature $F_g \in \mathrm{R}^{256\times 1\times 1}$ is input to the softmax layer for vehicle identity prediction under the ID loss constraint. To improve recognition accuracy, an ID loss function in the global branch is set as softmax cross-entropy loss to punish the wrong prediction. The ID loss function is formulated as follows:

$$Loss_{cse}^{Global} = \sum_{i=1}^{N} y_i \log(\hat{p}_i) \tag{1}$$

where N is the number of vehicle identities in the dataset, y_i is the ground-truth label of sample i, and $\hat{p}_i$ is the prediction probability. In Eq. (1), if the predicted identity of sample i equals to ground truth, $y_i = 1$; otherwise $y_i = 0$.

B. Local Branch Feature Learning

Many studies in the field of vehicle Re-ID only consider global information. However, when vehicles are almost identical in appearance, it is difficult to distinguish them only by global appearance attributes. Some discernible local regions in vehicles, such as windows, lights, markings, interior accessories, can provide reliable essential clues to distinguish vehicles. To obtain local features of vehicles, we first establish a vehicle part parser to obtain the masks of vehicle parts and then perform feature extraction based on these local regions under the masks. In order to learn the discriminative local region features, the L-ATT module is used to obtain the attention map of different local regions, and different degrees of attention are paid according to the importance of these local regions to highlight these critical local regions. The local region feature learning contains two steps: 1) vehicle part parsing and 2) local attention learning.

1) Vehicle Part Parsing: As a classic object detection algorithm, YOLO [39] is adopted to obtain local regions of vehicle parts. The YOLO algorithm contains two steps. First, it selects local regions of vehicle parts, such as lights, brand logos, windows, and labels them. Next, it extracts these local regions by the bounding box detection network. Note that, if some local regions of vehicle parts cannot be detected well due to occlusion or low resolution, the undetected local region is estimated with the help of the local regions detected correctly by the YOLO algorithm, which can be formulated as Eq. (2), shown at the bottom of the next page, where $x_{\min}$ and $x_{\max}$ represent the left and right boundary of the Bounding Box, respectively; and $y_{\min}$, $y_{\max}$ are the bottom and up boundaries, respectively. Notation $B_j^0\,(x_{\min}, y_{\min}, x_{\max}, y_{\max})$ represents the estimated local region j, $B_j^i\,(x_{\min}, y_{\min}, x_{\max}, y_{\max})$ is the detected local region j of the i-th sample image with the same identity as the estimated image, and N is the number of the detected sample images.

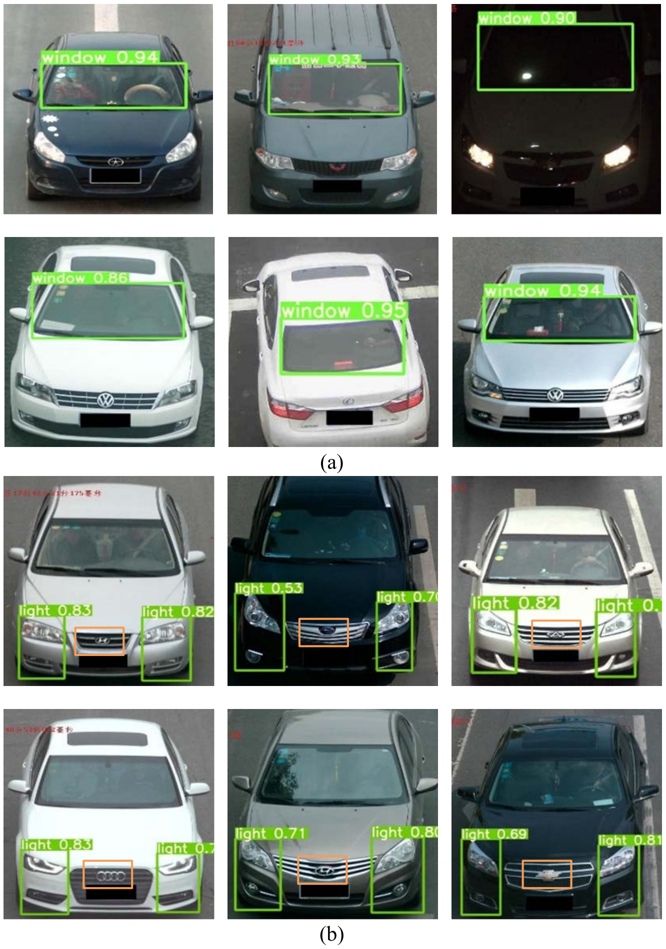


Fig. 4. Visualization of local region detection. Through YOLO algorithm, local vehicle parsing extracts the local regions of vehicle windows and headlights and infers the location and size of the brand logo from the detected headlights (left light, right light). (a) Detection of windshield from different viewpoints; (b) Detection of vehicle headlights and brand logo.

The vehicle part parsing results are visually shown in Fig. 4. For the i-th local region, the local vehicle parsing computes a series of mask-guided local region feature F_i^l as:

$$F_i^l = \{Masks_i\}_{i=1}^{N} \odot F \tag{3}$$

where $\odot$ represents element-wise product and $\{Masks_i\}$ denotes the i-th vehicle mask. In the study, the used masks include the light mask, brand logo mask, and window mask.

2) Local Attention Learning: The L-ATT module relies on a spatial attention mechanism to adaptively learns the attention scores ω_i of different local regions, such as lights, brand logos, and windows, which can indicate the importance of different local regions. Based on the obtained attention scores ω_i, the weighted local region feature F_i^{l-att} for the i-th local region can be expressed as: weighted local region feature F_i^{l-att} for the i-th local region can be expressed as:

$$F_i^{l-att} = \left\{\omega_i F_i^l\right\}_{i=1}^{N} \tag{4}$$

The specific learning process, i.e., the L-ATT module, is summarized as follows: 1) first apply mask global average pooling (MGAP) operation on the local region feature F_i^l, calculated by Eq. (6), where MGAP can precisely restrict the average pooling areas in feature map via different local region masks, since only the elements in the local region i can be activated while the values of the other position elements are all zero. 2) The local region features are summed along the feature channel, and the importance of different regions is learned through the softmax layer mapping function to obtain attention scores ω_i, which can emphasize the most important local regions with high values and represent the unimportant local regions with low values. For the i-th local region, the attention scores ω_i can be predicted by Eq. (5).

$$\omega_i = \frac{\exp(\varphi(MGAP(F_i^l), \theta_\varphi))}{\sum_{I=1}^{N} \exp(\varphi(MGAP(F_i^l), \theta_\varphi))} \tag{5}$$

$$MGAP\left(F_i^l\right) = \frac{\sum_{j,k=1}^{16} Masks_i(j,k) \odot F(j,k)}{\sum_{j,k=1}^{16} Masks_i(j,k)} \tag{6}$$

where $\varphi(\cdot)$ represents a learnable function to highlight the most informative local region feature and θ_φ is a learnable parameter. $Masks_i$ is the i-th element of $\{Masks_i\}_{i=1}^{N}$.

For the local region features, triplet loss [40] is used to optimize the distance between features in the embedding feature space. The optimization procedure attempts to reduce the distance between vehicle samples with identical identity while increasing the distance between samples from different vehicles. Therefore, it improves the accuracy of vehicle Re-ID. The loss function of the local branch is formulated as follows:

$$Loss_{tri}^{Local} = Max(D_{a,p} - D_{a,n} + \alpha, 0) \tag{7}$$

Algorithm 1 Local Feature Learning

```
Require: Train local vehicle parser
Require: Output feature map $F$ by feature extraction network.
1. Parse vehicle part components and obtain $N$ vehicle part masks$\{Masks_i\}_{i=1}^{N}$.
2. Obtain local region features $F_i^l$ by Eq. (2).
3. Restrict the average pooling areas indicated by the local region mask via MGAP operation.
4. Train the L-ATT module and calculate attention scores $\omega_i$.
5. Calculate the weighted local region features via Eq. (5).
6. Return $F_i^{l-att}$.
```

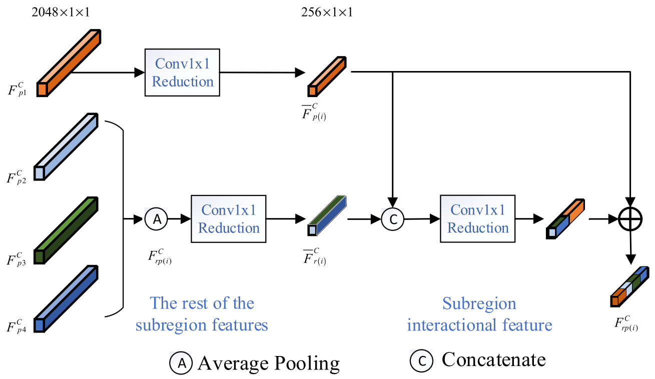


Fig. 5. Structure of Sub-region Interaction Module (SIM).

where D_{ap} represents the Euclidean distance between vehicle samples with identical identity, D_{an} is the Euclidean distance between samples of different vehicles, and α is a margin parameter set to 0.3. Algorithm 1 summarizes the process of local branch learning.

C. Complementary Branch Feature Learning

Inspired by the work of Zheng *et al.* [41], this study proposes a complementary branch to cope with the ever-changing critical local regions containing discriminative clues in the whole vehicle images. As shown in Fig. 3, the complementary branch generates structural features and multi-granularity features for a more integral and diverse representation of vehicle appearance.

We firstly copy the feature F generated by the backbone network and use the copied feature F for independent complementary feature learning. The main learning process can

$$\begin{aligned} B_j^0(x_{\min}, y_{\min}, x_{\max}, y_{\max}) &= \frac{\sum_{i=1}^{N} B_j^i(x_{\min}, y_{\min}, x_{\max}, y_{\max})}{N} \\ &= \left[\frac{\sum_{i=1}^{N} B_j^i(x_{\min})}{N}, \frac{\sum_{i=1}^{N} B_j^i(y_{\min})}{N}, \frac{\sum_{i=1}^{N} B_j^i(x_{\max})}{N}, \frac{\sum_{i=1}^{N} B_j^i(y_{\max})}{N}\right] \end{aligned} \tag{2}$$

be divided into two steps. The first step is to apply global maximum pooling (GMP) to extract features, and then divide the pooled feature map into four sub-regions to obtain structural features F_p^C. Subsequently, these four sub-regions are fed into a Conv1×1 reduction block to reduce 2048-dimensional features to 256-dimensional features of sub-regions. Inspired by the literature [35] and [42], based on the obtained sub-regions features, a new SIM is introduced to fully use the relations between the divided sub-regions, shown in Fig. 5. Here, we assume that the feature $F_{rp(i)}^C$ contains information of the original sub-regional feature $\overline{F}_{p(i)}^C$ itself and other sub-regional features $\overline{F}_{r(i)}^C$ of vehicles. For the *i*-th sub-region, we use a skip-connection [9] to transfer the interactional information of $\overline{F}_{p(i)}^C$ and $\overline{F}_{r(i)}^C$ to $F_{rp(i)}^C$ as follows:

$$F_{rp(i)}^C = \overline{F}_{p(i)}^C + CONV1 \times 1\left(Cat\left(\overline{F}_{p(i)}^C, \overline{F}_{r(i)}^C\right)\right) \quad (8)$$

where $CONV1 \times 1(\cdot)$ is a reduction block consisting of a 1×1 convolution layer, a batch normalization layer, and a ReLU to reduce feature dimension, and $Cat(\cdot)$ represents the concatenation operation. After obtaining sub-region interactional features, ID loss is adopted to optimize the network independently by supervised learning, which is denoted as Eq. (9). Because supervised learning can force the model to explore all useful details information in the sub-regions, the model can obtain discriminative and complete features.

$$Loss_{cse\_p}^{Comp} = \sum_{i=1}^{N} y_i \log(\hat{p}_i) \quad (9)$$

The second step is to apply different pooling operations with different pooling factors to obtain multi-granularity features. In this step, global average pooling (GAP) and global maximum pooling (GMP) operations are used in this step to obtain coarse-grained features, represented as F_{gap}^C and F_{gmp}^C, with 1×1 spatial size; meanwhile, adaptive average pooling (AAP) and adaptive maximum pooling (AMP) are used to obtain the fine-grained features, represented as F_{aap}^C and F_{amp}^C, with 2×2 spatial size. All these features are optimized by triplet loss-based function defined by Eq. (10).

$$Loss_{tri}^{Comp} = Max(D_{a,p} - D_{a,n} + \alpha, 0) \quad (10)$$

Moreover, the obtained multi-granularity features are fed into batch normalization layer for normalization and then fuse them according to the corresponding feature size (e.g., F_{gap}^C and F_{gmp}^C are fused according to the size of 1×1 and F_{aap}^C and F_{amp}^C are fused according to the size of 2×2). Subsequently, these fused multi-granularity features are fed into a 'Head Block' [41] to predict vehicle identity, where ID loss punishes the results with wrong vehicle predictions.

Note that, when ID loss and triplet loss are used simultaneously to optimize the same feature, synchronous convergence may fail because the gradient directions of ID loss and triplet loss may be inconsistent during the iterative process. To address this issue, this study inputs the multi-granularity features into a batch normalization layer before combining the ID loss to optimize the model by the 'BNNeck' strategy [43], instead of optimizing all features by directly connecting ID loss behind the optimization based on triplet loss. The integration of the BN operation improves the sparsity of each dimension of features in the batch and enables these features to be distributed near the hypersphere, which aids ID loss function to optimize these features. The ID loss function adopts the formation of softmax cross-entropy loss, as shown in Eq. (11).

$$Loss_{cse}^{Comp} = \sum_{i=1}^{N} y_i \log(\hat{p}_i) \quad (11)$$

Finally, the total loss function in TBE-Net is the sum of all loss functions, such as:

$$\begin{aligned} Loss = Loss_{cse}^{Global} + Loss_{tri}^{Local} + Loss_{tri}^{Comp} \\ + Loss_{cse\_p}^{Comp} + Loss_{cse}^{Comp} \end{aligned} \quad (12)$$

IV. Experiments

A. Datasets and Evaluation Metrics

1) Datasets: The performance of the proposed network is evaluated on two datasets: VehicleID [2] and VeRi-776 [5] VehicleID is a large-scale vehicle Re-ID dataset [2], which includes 221,763 images of 26,267 vehicles captured from front and rear views, with a total of 250 vehicle models. Among them, the training set contains 110,178 pictures of 13,134 vehicles, and the test set contains 111,585 pictures of 13,133 vehicles. Each image in VehicleID is tagged with vehicle ID, camera information, and vehicle model. For a complete performance evaluation of vehicle Re-ID, the test set in VehicleID is divided into three subsets with different sizes of large, medium, and small. In the test phase, our network was evaluated in the three subsets.

VeRi-776 dataset [5] is one of the most widely used datasets in the vehicle Re-ID community. It contains about 50,000 images of 776 vehicles, captured and collected from 20 cameras with different viewpoints, illumination, and occlusion. The train set contains 37,778 images of 576 vehicles, and the test set covers 11,579 images of the other 200 vehicles. All samples in the datasets contain abundant annotation information such as location boundary, license plate, time stamp, geographical location, vehicle model, color, and manufacturer, which facilitates training and test of vehicle Re-ID network based on vehicle local regions.

2) Evaluation Metrics: To evaluate the performance of the TBE-Net, this study uses two indicators as evaluation metrics: mean Average Precision (mAP) and Cumulated Matching Characteristics (CMC).

a) mAP: The metric mAP measures the overall performance of the vehicle Re-ID model and represents the average accuracy of the recognition results. Equation (13) formulates mAP,

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

where Q represents the total number of query images in query set and $AP(q)$ represents the average precision retrieval results of each query image q that is formulated as Eq. (14):

$$AP = \frac{\sum_{k=1}^{n} P(k) \times gt(k)}{N} \tag{14}$$

where n represents the total number of images in the dataset, N represents the total number of images of the target vehicle, $P(k)$ represents the accuracy of the first k retrieval result sequences, and $gt(k)$ represents whether the target vehicle ranks among the first k results.

b) CMC: The metric CMC represents the probability of finding the correct result in the first k retrieval results, which can be expressed as Eq. (15). When the correct matching target of image q ranks among the first k retrieval results, $gt(k)$ is equal to 1.

$$CMC@K = \frac{\sum_{q=1}^{Q} gt(q,k)}{Q} \tag{15}$$

B. Implementation Details

We deploy ResNet50 as backbone network to extract feature, which is pre-trained on ImageNet [38]. Before experiments, all input images are resized to 256 × 256. During the network training, SGD optimizer is used with a momentum of 0.9. As for the weight decay strategy, we set the base learning rate to 2e-4, and drops to 2e-5, 2e-6 in the 150th epoch, 220th epoch for faster convergence, respectively. The batch size of network training is set as 64, and the total number of epochs is 240. We also augment the data with random horizontal flipping and random erasing [44]. In the training and testing phases, we both use Euclidean distance to evaluate the feature similarity between the query images and the gallery images. In the Eq. (7) and Eq. (10), α is set to 0.3.

Before the Re-ID model is trained, 800 sample images are randomly selected from the VehicleID dataset and labeled in these local regions, such as windows, lights, and brand logos with bounding boxes, which are used to train the YOLO detection network. Next, the trained network will be applied to the VehicleID and VeRi-776 datasets to extract local regions of the vehicle image. For a vehicle image, we detect three critical local regions, i.e., the windows, lights, and brand logos, as shown in Fig. 4.

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

The proposed method is compared with state-of-the-art vehicle Re-ID methods. The detailed results are as follows:

1) Experiments On VehicleID: The proposed method is compared with several state-of-the-art methods using the metrics CMC@1 and CMC@5 on the three test subsets (small, medium, large) of VehicleID dataset. Table I and Table II show the comparison results of CMC@1 and CMC@5, respectively. Tables I and II show that the proposed method achieves 0.860 and 0.984 in CMC@1 and CMC@5, respectively. The performance outperforms all other methods by a large margin.

TABLE I
RESULTS OF CMC@1 ON VEHICLEID DATASET

Method	Year	Small	Medium	Large
DRDL[2]	2016	0.490	0.428	0.382
OIFE[15]	2017	-	-	0.670
VAMI[45]	2018	0.631	0.529	0.473
C2F[12]	2018	0.611	0.562	0.514
EALN[49]	2019	0.751	0.718	0.693
AAVER[31]	2019	0.747	0.686	0.635
PRN[14]	2019	0.784	0.750	0.742
SAVER[46]	2020	0.799	0.776	0.753
SGAT[48]	2020	0.781	0.739	0.718
CFVMNet[47]	2020	0.814	0.773	0.747
TBE-Net	2021	0.860	0.823	0.807

TABLE II
RESULTS OF CMC@5 ON VEHICLEID DATASET

Method	Year	Small	Medium	Large
DRDL[2]	2016	0.735	0.668	0.616
OIFE[15]	2017	-	-	0.829
VAMI[45]	2018	0.833	0.751	0.703
C2F[12]	2018	0.817	0.762	0.722
EALN[49]	2019	0.881	0.839	0.814
AAVER[31]	2019	0.938	0.900	0.856
PRN[14]	2019	0.923	0.883	0.864
DFLNet[50]	2020	0.950	-	0.905
SAVER[46]	2020	0.952	0.911	0.883
CFVMNet[47]	2020	0.941	0.904	0.887
TBE-Net	2021	0.984	0.966	0.949

Among these state-of-the-art methods, both OIFE [15] and VAMI [45] apply the viewpoint information of vehicles, but they cannot resolve the difficult-to-recognize problem that different vehicles from the same camera viewpoints are visually similar. To address the problem, both methods need more integral and diverse features besides the vehicle viewpoint information. DRDL [2] uses global appearance features for vehicle Re-ID based on metric learning, but the method ignores the role of local region features. PRN [14] uses the local regions of the vehicle for Re-ID; however, not all local regions in PRN [14] can provide equally critical distinguishing information, leading to lower Re-ID accuracy.

Different from these two methods of OIFE and VAMI, the proposed method learns the attention map of different local regions through the local attention module, which can adaptively assign different weights to different local regions according to their significance. Therefore, the proposed method pays more attention to critical local regions, and reduces the weight on irrelevant local regions. AAVER [31] applies an attention mechanism to extract discriminative key features. However, it ignores the fact that discriminative clues for the cross-camera vehicle Re-ID task could appear randomly anywhere in the whole vehicle image. Therefore, AAVER fails to obtain comprehensive vehicle features. Due to lacking training based on vehicle local regions, SAVER [46] only learns important global information in the vehicle through self-supervision learning, thereby hard to perceptive the significant local regions. Similarly, although CFVMNet [47] can determine whether vehicles belong to the same view, it cannot accurately find all salient local information. In contrast, our proposed method can

TABLE III
RESULTS OF MAP, CMC@1AND CMC@5 ON VERI-776 DATASET

Method	Year	mAP	CMC@1	CMC@5
BOW-CN[51]	2015	0.122	0.339	0.537
GoogLeNet[52]	2015	0.170	0.498	0.712
FACT[9]	2016	0.185	0.510	0.735
FACT+PLATE-REC[1]	2016	0.187	0.511	0.736
FACT+PLATE-SNN[1]	2016	0.259	0.611	0.774
FACT+PLATE+STR[1]	2016	0.278	0.614	0.788
OIFE[15]	2017	0.480	0.894	-
GSTE[30]	2018	59.40	-	-
VAMI[45]	2018	0.501	-	-
EALN[49]	2019	0.574	0.844	0.941
AAVER[31]	2019	0.612	0.890	0.947
PRN[14]	2019	0.743	0.943	0.989
SPAN[17]	2020	0.689	0.940	0.976
SGAT[48]	2020	0.656	0.896	-
DFLNet[50]	2020	0.732	0.932	0.975
TBE-Net	2021	0.795	0.960	0.985

detect important local regions through the attention mechanism and extract complete structural features and multi-granularity features through the complementary branch, thereby obtaining more abundant vehicle features, which helps to improve the accuracy of vehicle Re-ID.

2) Experiments on VeRi-776: To verify the general performance of the proposed method, we also evaluated our method on the VeRi-776 dataset using the criteria mAP, CMC@1 and CMC@5. Table III shows the results of our method compared to the state-of-the-art methods. Among these methods, FACT+PLATE-REC [1], FACT +PLATE-SNN [1] and FACT+PLATE+STR [1] all use the license plate information in the VeRi-776 dataset; OIFE [15] and FACT+PLATE+STR [1] also use the spatial-temporal information. Note that any Re-ID method using the spatial-temporal information is strict to the dataset, because it needs to cost lots of time for sample annotation; especially, it needs to make a large number of additional annotations for the pre-defined interested local regions before network training. Attributed to the use of vehicle local regions, AAVER [31] and PRN [14] show good performance in mAP.

In addition, SGAT [48] also used the structural relations of vehicle key points and the extrinsic relations between images for Re-ID. However, the accuracy is low, since these two relations are hard to obtain when vehicle appearance sharply changes under non-overlapping cameras. Chen *et al.* [17] proposed a Semantics-Guided Part Attention Network (SPAN), but it is difficult to find fine-grained local information within the view due to the lack of an attention mechanism to learn the importance of local features in the view. In contrast, in addition to extracting the local vehicle feature, our method learns the importance of different local regions through the L-ATT module and pays more attention to the high-importance local regions by assigning larger weights to them. Moreover, the complementary branch integrated into the proposed method extracts more comprehensive structural features and multi-granularity features, which can effectively cope with the constant changes of these local regions in vehicle image when the vehicle moves across non-overlapping cameras.

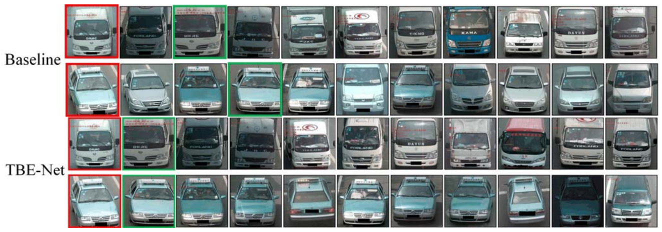


Fig. 6. Visualized detection results. The images marked in the red box in the first column is the query image. The remaining are the top-10 images, in which the correctly matched images are marked in the green boxes.

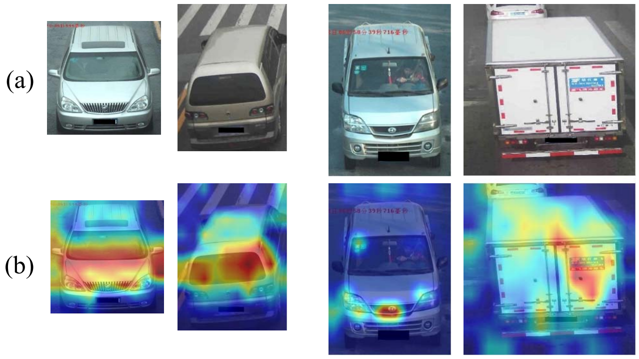


Fig. 7. Heatmaps of query vehicles. (a) Original images; (b) Corresponding heatmaps. Through the learning of TBE-Net, the Re-ID model assigns large weights to not only the local regions with high-importance discriminative clues, such as lights, windows and brand logo, but also the discriminative markings such as tags that appear at a random location of the vehicle image.

TABLE IV
ABLATION EXPERIMENTS BASED ON LOCAL REGION FEATURES

Method	CMC@1	CMC@5	CMC@10
F_g (Baseline)	0.749	0.887	0.922
$F_g + F_{light}$	0.750	0.874	0.910
$F_g + F_{window}$	0.764	0.899	0.934
$F_{light} + F_{brand}$	0.677	0.871	0.909
$F_{light} + F_{window}$	0.731	0.884	0.928
$F_g + F_{light} + F_{brand}$	0.768	0.896	0.934
$F_g + F_{light} + F_{window}$	0.770	0.905	0.942
$F_{light} + F_{window} + F_{brand}$	0.758	0.916	0.946
$F_g + F_{light} + F_{window} + F_{brand}$	0.777	0.920	0.948

TABLE V
ABLATION EXPERIMENTS BASED OF L-ATT IN VEHICLE RE-ID

Method	CMC@1	CMC@5	CMC@10
$F_{light} + F_{window} + F_{brand}$	0.758	0.916	0.946
$F_{light} + F_{window}+F_{brand}$ +*L-ATT*	0.791	0.925	0.963
$F_g + F_{light} + F_{window} + F_{brand}$	0.777	0.920	0.948
$F_g + F_{light} + F_{window} + F_{brand} +$ *L-ATT*	0.801	0.934	0.966
$F_g + F_{light} + F_{window} + F_{brand} + F_{comp}$	0.846	0.966	0.983
$F_g + F_{light} + F_{window} + F_{brand} + F_{comp}$ + *L-ATT*	0.860	0.984	0.993

D. Ablation Study

To validate the TBE-Net model, we carried out ablation experiments to explore the performance of local region features, L-ATT module, and complementary features on

TABLE VI

ABLATION EXPERIMENTS BASED ON VEHICLE COMPLEMENTARY FEATURES

Method	CMC@1	CMC@5	CMC@10
F_g (Baseline)	0.749	0.887	0.922
$F_g + F_{comp}$	0.839	0.953	0.976
$F_g + F_{light} + F_{window} + F_{brand}$	0.777	0.920	0.948
$F_g + F_{light} + F_{window} + F_{brand} + F_{comp}$	0.846	0.966	0.983
$F_g + F_{light} + F_{window} + F_{brand} +$ *L-ATT*	0.801	0.934	0.966
$F_g + F_{light} + F_{window} + F_{brand} + F_{comp} +$ *L-ATT*	0.860	0.984	0.993

TABLE VII

ABLATION EXPERIMENTS BASED ON USING DIFFERENT POOLING OPERATIONS

Method	CMC@1	CMC@5	CMC@10
F_g (Baseline)	0.749	0.887	0.922
GMP_GAP_GAP_AAP_AAP	0.855	0.971	0.990
GMP_GMP_GMP_AMP_AMP	0.852	0.969	0.989
GMP_GAP_GMP_AAP_AMP	0.860	0.984	0.993
GMP _GMP_GAP_AMP_AMP	0.849	0.967	0.988
GMP _GMP_GAP_AMP_AMP	0.847	0.965	0.986
GMP _GMP_GMP_AMP_AAP	0.839	0.951	0.977
GMP _GMP_GMP_AAP_AAP	0.846	0.962	0.984
GMP _GAP_GAP_AMP_AMP	0.856	0.972	0.990
GMP _GAP_GAP_AMP_AAP	0.845	0.962	0.981

TABLE VIII

ABLATION EXPERIMENTS ON VERI-776 DATASET

Settings	mAP	CMC@1	CMC@5
F_g (Baseline)	0.758	0.946	0.976
TBE-Net w/o Local	0.773	0.954	0.980
TBE-Net w/o F_{comp}	0.764	0.951	0.978
TBE-Net w/o L-ATT	0.788	0.958	0.983
TBE-Net	0.795	0.960	0.985

VehicleID dataset, respectively. The performance variations of ablation experiments in terms of CMC@1, CMC@5, and CMC@10 are summarized into Table IV, Table V, and Table VI, respectively. Fig. 6 gives the detection results of our TBE-Net and the baseline model. Fig. 6 shows that the top-10 results retrieved from TBE-Net are superior to those of the baseline model. Fig. 7 shows the heatmaps of query vehicles. The critical local regions containing discriminative clues are highlighted, demonstrating that our model can capture the local regions with high-importance discriminative clues, such as vehicle lights, brand logo, windows, and randomly appearing markings. Our model also pays more attention to the critical local regions. For instance, the blue tag on the wagon in the last column of Fig. 7 may appear at a random location in the vehicle image; however, this tag can determine whether two vehicles are the same. Moreover, we performed ablation experiments on the VeRi-776 dataset to further verify the effectiveness of the proposed method, shown in Table VIII.

The convergence speed and accuracy of TBE-Net are investigated on the VeRi-776 dataset. Fig. 8 compares the top 1 error change of TBE-Net to that of the global appearance features-based method during the network training. The figure shows that the two curves present a steep trend at first and then a gentle trend. This is because the learning rate of 60-130 epoch is ten times smaller than that of 60 epoch. Fig. 8 demonstrates that the TBE-Net model has a more robust convergence speed and a higher accuracy.

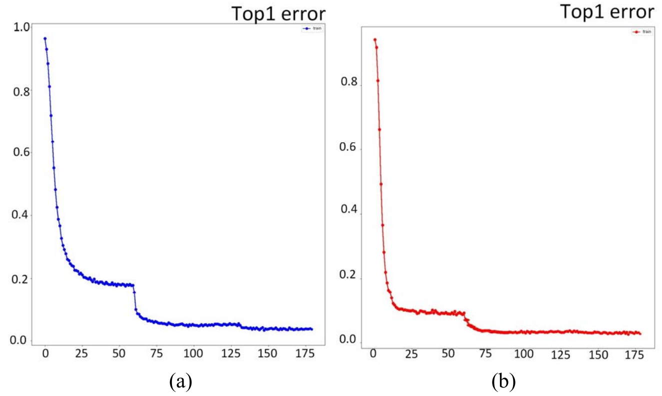


Fig. 8. Top1 error curves. (a) Global appearance feature-based model; (b) TBE-Net.

1) Vehicle Local Information: To investigate the significance of different local region features to vehicle Re-ID, we designed an ablation experiment to verify the significance of the vehicle lights, windows, and brand logo. In the experiment, the method using the global appearance feature only, F_g, is studied as a baseline model. Table IV shows that the accuracy of the baseline model is 0.749, 0.887, and 0.922 in terms of CMC@1, CMC@5, and CMC@10, respectively. Moreover, we conducted ablation experiments based on local regions such as vehicle lights, windows, and brand logo. Compared to the baseline model, the accuracy of the method based on $F_g + F_{light}$, i.e., the method integrating global appearance feature and vehicle light feature, increases by 0.1% due to the utilization of headlight information. The accuracy of the method based on $F_g + F_{window}$, i.e., the method integrating global appearance feature and vehicle window feature, increases by 1.5% attributed to the utilization of window information, and its accuracy is also higher than that of the method based on $F_g + F_{light}$ by 1.4%. These results show that the method surpasses the baseline method by exploiting the local region and global appearance features. Meanwhile, different local regions contribute different power to vehicle Re-ID. Further investigations illustrate that the window region contributes more than the light region. The main reason is that the window region of vehicles may contain more discriminative information such as annual service signs, customize paintings, and personal decorations, by which vehicles can be distinguished easily. To investigate the role of the combination of different local region features in vehicle Re-ID, we perform the ablation experiment based on $F_g + F_{light} + F_{brand}$ and $F_g + F_{light} + F_{window}$ by fixing the global appearance feature F_g and light feature F_{light}. Experimental results in Table IV show that the method based on $F_g + F_{light} + F_{brand}$ surpasses the method based on the global appearance feature and light feature $F_g + F_{light}$ by 1.8% and 2.2% in

CMC@1 and CMC@5, respectively. The method based on $F_g + F_{light} + F_{brand}$ also exceeds the baseline method in CMC@1 and CMC@5 by about 1.9% and 0.9%, respectively. When using the combined information of global image, lights, windows, and brand logo, the proposed method significantly improved by 2.8% compared to the baseline method. These experimental results suggest that local features can provide more refined information for distinguishing vehicle and proves the significance of local region features to vehicle Re-ID.

Moreover, we also conduct a series of experiments to study the effect of combining global features and local features on vehicle Re-ID. By comparing the methods based on $F_{light} + F_{window}$ and $F_g + F_{light} + F_{window}$, the experimental results show that removing global features significantly affects the performance of Re-ID, with a maximum reduction of 3.9% in CMC@1. This phenomenon is also found in the ablation experiments of $F_{light}+F_{brand}$ and $F_g+F_{light}+F_{brand}$. In addition, the experimental results based on $F_{light} + F_{window} + F_{brand}$ and $F_g + F_{light} + F_{window} + F_{brand}$ show that the method combining the global and local region features has a higher accuracy. These experimental results suggest that global features can provide more macro information as supplements for local features, and combining global and local region features can promote vehicle feature representation.

2) Local Attention (L-ATT) Module: To verify the effectiveness of L-ATT in feature learning, we conducted the following ablation experiments and summarized the results in Table V. We first remove the L-ATT module from the proposed network framework. By comparing the methods based on $F_{light} + F_{window} + F_{brand}$ and $F_{light} + F_{window} + F_{brand} + L - ATT$, experimental results show that removing the L-ATT module negatively affects the accuracy of vehicle Re-ID, with a maximum reduction of 3.3% in CMC@1. This demonstrates that the L-ATT module can effectively improve the accuracy of vehicle Re-ID. In addition, by exploring the two methods based on $F_g + F_{light} + F_{window} + F_{brand}$ and $F_g + F_{light} + F_{window} + F_{brand} + L - ATT$, the method with L-ATT module in CMC@1 was improved by 2.4%, compared to the method without the L-ATT module. These results from ablation experiments demonstrate that L-ATT has a more significant influence on the local branch. It also indicates that L-ATT can weigh adaptively to critical local regions according to their significance during feature learning.

3) Vehicle Complementary Features: We also investigate the effect of vehicle complementary features using ablation experiments, as summarized in Table VI. We compare the baseline method F_g with the method based on global appearance and complementary features, i.e., $F_g + F_{comp}$. Experimental results show that using the complementary features and the global appearance features can significantly improve the proposed network by 9.0%, 6.6%, and 5.4% in terms of CMC@1, CMC@5, and CMC@10, respectively. The results demonstrate that exploiting complementary features has a stronger vehicle Re-ID power, for which the proposed network has more complete and richer granularities than the methods using only global appearance features.

Additionally, to verify the effect of complementary featureson local region features, we performed the experiments based on $F_g + F_{light} + F_{window} + F_{brand} + F_{comp}$ and $F_g + F_{light} + F_{window} + F_{brand}$. Experimental results show that the method with the complementary features outperforms the method without the complementary features in recognition accuracy by a large margin, especially, with a 6.9% improvement in the CMC@1. The improved result demonstrates that the multi-granularity features extracted by the complementary branch can adapt to the ever-changing critical local regions in the whole vehicle image. Further experiments based on $F_g + F_{light} + F_{window} + F_{brand} + L - ATT$ and $F_g + F_{light} + F_{window} + F_{brand} + F_{comp} + L - ATT$ also gain a similar result. Namely, the network is improved when the complementary branch is used, with a maximum improvement of 5.9% in CMC@1. To sum up, the above ablation experiments illustrate the significance and effectiveness of the complementary branch in improving vehicle Re-ID.

4) Importance of Applying Different Pooling Operations on Vehicle Re-ID: In this part, we conduct ablation experiments to verify the performance of different pooling operations in the complementary branch for vehicle Re-ID. Table VII shows the results of the ablation experiment on the VehicleID dataset in terms of CMC@1, CMC@5, and CMC@10 after using different pooling operations. Note that we only change the second to the fifth pooling layers since these features will be fused to obtain multi-granularity features, and we fixed the first pooling layer since the features extracted by the first pooling layer are not fused. Seen from Table VII, the top three methods in performance are *GMP_GAP_GMP_AAP_AMP, GMP_GAP_GAP_AMP_AMP*, and *GMP_GAP_GAP_AAP_AAP* in turn, among which *GMP_GAP_GMP_AAP_AMP* has the best performance with 0.860 CMC@1, 0.984 CMC@5 and 0.993 CMC@10. Further analysis shows that the method fusing multiply different pooling operations in complementary branches performs better than those based on a single pooling operation.

Finally, we implement the ablation experiments on the VeRi-776 dataset to further verify the effectiveness of the proposed method. We select the mAP, CMC@1, CMC@5 as evaluation metrics, the results as shown in Table VIII. *TBE-Net w/o Local* indicates the TBE-Net model without the local branch in the feature learning. Similarly, *TBE-Net* w/oF_{comp} indicates that the proposed model adds the local regions and local attention module, but does not uses the complementary branch in the feature learning. *TBE-Net w/o L-ATT* expresses that the model adds the local regions in local branch, but does not use local attention module to learn the importance of different local regions. As seen from Table VIII, our TBE-Net achieves better performance than others by a large margin. This is because TBE-Net can perceive the most salient local regions of the vehicle and can also adapt to the ever-changing local regions containing discriminative clues in vehicle images. In addition, the model obtains limited feature representation using *TBE-Net* w/oF_{comp} in feature learning, leading to lower Re-ID accuracy. Finally, the experiment of *TBE-Net w/o L-ATT* is conducted to verify the effectiveness of L-ATT. Experimental results show that applying the local regional features directly without learning the importance between them or directly removing the local branch will degrade the performance of the model.

V. Conclusion

This paper proposes a three-branch embedding network with feature complementary learning and part-aware ability for vehicle Re-ID to solve the challenge that different vehicles from the same camera viewpoints are visually similar in the vehicle Re-ID task. In feature learning, the L-ATT module assigns large weights to the critical local regions with discriminative clues and small weights to the obscure or insignificant regions. The L-ATT module enables the feature learning procedure to capture more distinguishing features. Also, the proposed model contains a complementary branch to obtain abundant vehicle features, including structural features and multi-granularity features to accommodate the ever-changing critical local regions in vehicle images captured by different non-overlapping cameras. These abundant vehicle features can complement the global appearance and local region features, thereby enhancing the Re-ID accuracy of the proposed TBE-Net. Extensive experiments have demonstrated the effectiveness of the proposed TBE-Net.

Future studies can leverage information from multiple different perspectives of cameras to further improve the performance of vehicle Re-ID, instead of being constrained by one perspective of cameras such as vehicle front image. Vehicle images from multiple perspectives contain more discriminative clues for improving the accuracy and robustness of vehicle Re-ID. In addition, training vehicle Re-ID models require a large volume of samples from various perspectives of cameras, yet these samples are difficult to collect and annotate in the real world. To obtain more samples from various perspectives, GAN [53] can be used to generate more synthetic samples, including samples from different perspectives and those from the identical perspective yet having subtle distinctions, thus enabling deep neural network training.

References

[1] X. Liu, W. Liu, T. Mei, and H. Ma, “A deep learning-based approach to progressive vehicle re-identifification for urban surveillance,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, vol. 1, Oct. 2016, pp. 869–884.

[2] H. Liu, Y. Tian, Y. Yang, 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.

[3] 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.

[4] 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.

[5] 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.

[6] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[7] W. Sun, X. Zhang, X. He, Y. Jin, and X. Zhang, “A two-stage vehicle type recognition method combining the most effective Gabor features,” *Comput., Mater. Continua*, vol. 65, no. 3, pp. 2489–2510, 2020.

[8] W. Sun, X. Zhang, and X. He, “Lightweight image classifier using dilated and depthwise separable convolutions,” *J. Cloud Comput.*, vol. 9, no. 1, p. 55, Dec. 2020, doi: 10.1186/s13677-020-00203-9.

[9] X. Liu, W. Liu, H. Ma, and H. Fu, “Large-scale vehicle re-identification in urban surveillance videos,” in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2016, pp. 1–6.

[10] X. Liu, L. Li, S. Wang, Z.-J. Zha, D. Meng, and Q. Huang, “Adaptive reconstruction network for weakly supervised referring expression grounding,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2611–2620.

[11] 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.

[12] 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.*, vol. 32, no. 1, pp. 6853–6860, Apr. 2018.

[13] P. Angelo, B. Luca, and C. Simone, “Robust re-identification by multiple views knowledge distillation,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2020, pp. 93–110.

[14] 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.

[15] 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.

[16] 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.

[17] T. Chen, C. Liu, C. Wu, and S. Chien, “Orientation-aware vehicle re-identification with semantics-guided part attention network,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2020, pp. 330–346.

[18] X. Zhang, R. Zhang, J. Cao, D. Gong, M. You, and C. Shen, “Part-guided attention learning for vehicle instance retrieval,” 2019, *arXiv:1909.06023*.

[19] X. Liu, W. Liu, J. Zheng, C. Yan, and T. Mei, “Beyond the parts: Learning multi-view cross-part correlation for vehicle re-identification,” in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 907–915.

[20] M. Opitz, G. Waltner, H. Possegger, and H. Bischof, “BIER—Boosting independent embeddings robustly,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5199–5208.

[21] Y. Yuan, K. Yang, and C. Zhang, “Hard-aware deeply cascaded embedding,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 814–823.

[22] B. Bhattarai, G. Sharma, and F. Jurie, “CP-mtML: Coupled projection multi-task metric learning for large scale face retrieval,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4226–4235.

[23] Y. Wen, K. Zhang, and Z. Li, “A discriminative feature learning approach for deep face recognition,” in *Proc. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2016, pp. 499–515.

[24] F. Schroff, D. Kalenichenko, and J. Philbin, “FaceNet: A unified embedding for face recognition and clustering,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 815–823.

[25] O. Tadmor, Y. Wexler, T. Rosenwein, S. Shalev-Shwartz, and A. Shashua, “Learning a metric embedding for face recognition using the multibatch method,” 2016, *arXiv:1605.07270*.

[26] H. Shi *et al.*, “Embedding deep metric for person re-identication a study against large variations,” 2016, *arXiv:1611.00137*.

[27] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, “Person re-identification by multi-channel parts-based CNN with improved triplet loss function,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1335–1344.

[28] 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)*, Budapest, Hungary, Jul. 2019, pp. 1–9.

[29] 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.

[30] 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.

[31] P. Khorramshahi, A. Kumar, N. Peri, S. S. Rambhatla, J.-C. 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.

[32] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 2285–2294.

[33] F. Yang, K. Yan, S. Lu, H. Jia, X. Xie, and W. Gao, "Attention driven person re-identification," 2018, *arXiv:1810.05866*.

[34] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan, "End-to-end comparative attention networks for person re-identification," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3492–3506, Jul. 2017.

[35] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou, "Learning discriminative features with multiple granularities for person re-identification," 2018, *arXiv:1804.01438*.

[36] V. Christlein, L. Spranger, M. Seuret, A. Nicolaou, P. Kral, and A. Maier, "Deep generalized max pooling," in *Proc. Int. Conf. Document Anal. Recognit. (ICDAR)*, Sep. 2019, pp. 1090–1096.

[37] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 3–19.

[38] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," 2014, *arXiv:1409.0575*.

[39] 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.

[40] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," 2017, *arXiv:1703.07737*.

[41] Z. Zheng *et al.*, "Going beyond real data: A robust visual representation for vehicle re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 2550–2558.

[42] P. Hyunjong and H. Bumsub, "Relation network for person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, Feb. 2020, pp. 11839–11847.

[43] 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.

[44] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *Proc. AAAI Conf. Artif. Intell.*, Feb. 2020, pp. 13001–13008.

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

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

[47] 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.

[48] 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.

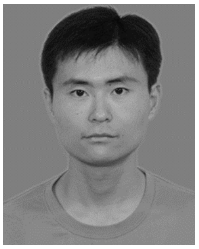
[49] Y. Lou, Y. Bai, J. Liu, S. Wang, and L.-Y. Duan, "Embedding adversarial learning for vehicle re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 8, pp. 3794–3807, Aug. 2019.

[50] Y. Bai, Y. Lou, Y. Dai, J. Liu, Z. Chen, and L.-Y. Duan, "Disentangled feature learning network for vehicle re-identification," in *Proc. 29th Int. Joint Conf. Artif. Intell. (IJCAI)*, Jul. 2020, pp. 474–480.

[51] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1116–1124.

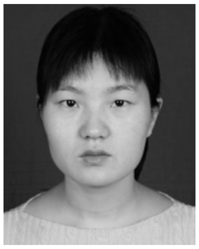
[52] L. Yang, P. Luo, C. C. Loy, and X. Tang, "A large-scale car dataset for fine-grained categorization and verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3973–3981.

[53] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.



Wei Sun received the B.S. and M.S. degrees in mechanical manufacture and automation from the Henan University of Science and Technology, China, in 2004 and 2006, respectively, and the Ph.D. degree in instrument science and technology from Southeast University, China, in 2010. From 2014 to 2015, he was a Post-Doctoral Researcher with the NEXTRANS Center, Purdue University, USA. He is currently an Associate Professor of automation with the Nanjing University of Information Science and Technology. His research interests include vehicle re-identification, computer vision, deep learning, and environment perception for intelligent vehicles.

Guangzhao Dai received the B.S. degree in automation from Wuxi Taihu University, China, in 2019. He is currently pursuing the M.S. degree in control engineering with the Nanjing University of Information Science and Technology. His research interests include large-scale vehicle retrieval and fine-grained image recognition.



Xiaorui Zhang received the B.S. and M.S. degrees in mechanical manufacture and automation from the Henan University of Science and Technology, China, in 2004 and 2006, respectively, and the Ph.D. degree in instrument science and technology from Southeast University, 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. Her research interests include virtual reality and human–computer interaction, haptic perception, and pattern recognition.

Xiaozheng He received the Ph.D. degree from the University of Minnesota, Twin Cities. He is currently an Assistant Professor with the Department of Civil and Environmental Engineering, Rensselaer Polytechnic Institute. 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. He was a recipient of the NSF CAREER Award. He serves as an Editorial Board Member for *Transportation Research Part B*; an Associate Editor for *Frontiers in Future Transportation*; and a Special Issue Guest Editor for *Transportation Research Part D*, *Journal of Advanced Transportation*, and *Sustainability*.

Xuan Chen received the B.S. degree in Internet of Things from the Nanjing University of Information Science and Technology, China, in 2019, where she is currently pursuing the M.S. degree in computer science and technology. Her research interests include large-scale vehicle retrieval and computer vision.