IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received January 18, 2022, accepted February 7, 2022, date of publication February 9, 2022, date of current version February 17, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3150411

Vehicle Re-Identification Based on Global Relational Attention and Multi-Granularity Feature Learning

XIN TIAN, XIYU PANG, GANGWU JIANG, QINGLAN MENG, AND YANLI ZHENG
School of Information Science and Electrical Engineering, Shandong Jiaotong University, Jinan 250357, China

Corresponding author: Xiyu Pang (xiyupang@126.com)

This work was supported in part by the Focus on Research and Development Plan in Shandong Province under Grant 2019GGX101055, and in part by the Shandong Education Science 13th Five Year Plan Special Project "Research on Examinee Online Identity Verification System Based on Multimodal Biometrics" under Grant BYZK201904.

ABSTRACT Vehicle Re-identification (Re-ID) refers to finding the same vehicle shot by other cameras from a given vehicle image library, which can also be regarded as a sub-problem of image retrieval. It plays an important role in intelligent transportation and smart cities. The key of vehicle Re-ID is to extract discriminative vehicle features. To better extract such features from the vehicle image to improve the recognition accuracy, we propose a three-branch adaptive attention network—Global Relational Attention and Multi-granularity Feature Learning (GRMF) to improve feature representation and discrimination. First, we divide the network into three branches, extracting different and useful features from three perspectives: spatial location, channel information, and local information. Second, we propose two effective global relational attention modules, which capture the global structural information for better attention learning. Specifically, to determine the importance level of a node, we use the global relationship between the node and all other nodes to infer the attention weight of the node directly. Third, according to the characteristics of the vehicle re-identification task, we introduce a suitable local partition strategy. It not only can simply capture subtle local information but also solve the problem of misalignment and within-part consistency disruption to a great extent. Extensive experiments demonstrate the effectiveness of our approach, and we achieve state-of-the-art results on two challenging datasets, including VeRi776 and VehicleID.

INDEX TERMS Vehicle re-identification, image retrieval, global relational attention, multi-granularity.

I. INTRODUCTION

Vehicle Re-identification (Re-ID) refers to the recognition of target vehicles in different cameras, which plays an important role in urban intelligent traffic systems. It has many applications in real life. For example, in the real traffic monitoring system, vehicle Re-ID can play the role of positioning, supervision, and criminal investigation of the target vehicle. In recent years, with the rise of deep neural networks and the proposal of large datasets, improving the accuracy of vehicle Re-ID has become an important task in the field of image recognition. However, due to the different perspectives under multiple cameras and the influence of illumination, occlusion, and other aspects, the feature distance intra-class becomes larger, while the distance inter-class becomes smaller, which further increases the difficulty of recognition.

In essence, person Re-ID and vehicle Re-ID are both image retrieval tasks. Recently, CNN-based methods have achieved great progress on person Re-ID [1]–[8]. Therefore, the CNN-based model applied to person Re-ID will also have a good performance in vehicle Re-ID. Most state-of-the-art CNN-based person Re-ID methods adopt pretrained CNN models (e.g., ResNet [9] and VGG [10]) on ImageNet and fine-tune them on Re-ID datasets under the supervision of different losses (e.g., Cross-Entropy loss and triplet loss [11]). In this paper, the backbone of our network is based on ResNet50 [9].

CNN-based vehicle and person Re-IDs usually focus on extracting the global features [12]–[15] of a person or a vehicle image. In this way, complete feature information can be obtained from the whole, but the global feature

The associate editor coordinating the review of this manuscript and approving it for publication was Michele Magno.

cannot well depict the intra-class variation caused by factors such as perspective. To address this problem, Part-based Convolutional Baseline (PCB) [1] and Multiple Granularity Network (MGN) [16] network models with local branches are designed. These networks divided feature maps into several stripes to extract local features from corresponding regions of the image. Moreover, the latter combined global information for multi-granularity learning, which further improves the model's performance. For vehicle Re-ID, the vehicles belonging to the same model are similar in overall appearance. However, in some small areas, such as checking marks, decorations, and use traces, they may have great differences. Therefore, it is very important to obtain and use local information to identify the same vehicle.

However, these part-based models [1], [16], which divide feature maps into three or even six parts, demand relatively well-aligned parts for the same pedestrian to acquire salient part information. Although both vehicle Re-ID and person Re-ID are essentially image retrieval problems, the division of parts of vehicles is not as clear as that of pedestrians, and the parts of the same vehicle observed from different angles vary greatly. In addition, strict uniform partitioning of the feature map breaks within-part consistency. This will make it difficult for the deep neural network to obtain useful features from the part, thus reducing the performance. Therefore, it is not feasible to simply apply the local partition method in the person Re-ID task to the vehicle. We need to adopt an effective division strategy according to the characteristics of the vehicle Re-ID task.

Work [1] shows that the destruction degree of within-part consistency is generally in direct proportion to the number of partitions, that is, the more the partitions are, the easier it is to destroy within-part consistency. In [17], Guo *et al.* Found that these different local details can be summarized into two parts. One is the windscreen part, which involves annual inspection signs and decorations. Another is the car-head section, which may contain some subtle using traces, etc. Inspired by the above findings, they proposed hard part-level attention to localize the salient vehicle parts by utilizing Spatial Transformer Network (STN) [18]. However, to locate the vehicle body parts, [17] needs an auxiliary model for rigid space segmentation, which increases the complexities and uncertainties of algorithms. In addition, by observing the actual vehicles and the vehicle images in the data set, we found that the height of the two parts was basically the same. So, on the local branch, we just divide feature maps evenly into two parts. In this way, it can not only simply extract the subtle information of the two parts but also solve the problem of misalignment and within-part disruption to a great extent.

The attention mechanism plays an important role in the human perceptual system, helping people to focus on identifying useful features and eliminating some noise and background distractions. For network models, an attention mechanism can make the model focus on the target subject rather than the background, and it is widely used in Re-ID tasks. So, a lot of networks with attention modules have been proposed (e.g., Further Non-local and Channel attention (FNC) [19] and Second-order Non-local Attention (SONA) [20]). However, they mainly reconstruct a node using pairwise relationships between nodes, without explicitly and directly modeling its global importance. We believe that to determine the importance of a node, it will be helpful if we know its global relationship with all other nodes, because people can intuitively determine the relative importance of a thing by comparing it with all other things. In this paper, two novel and effective global attention modules for vehicle re-identification are proposed. Specifically, we model the average pairwise relationship between nodes and all other nodes, obtain the global relationship of nodes, and then deduce the global importance of nodes. On the one hand, it reduces the difficulty of attention learning and computational complexity. On the other hand, global relations can comprehensively and robustly evaluate the importance of nodes, making attention learning more accurate. In the channel global attention module (CGAM), we utilize the global relationship importance of a channel relative to other channels to get the weight of that channel in all channels. Similarly, in the spatial global attention module (SGAM), we use the global relationship importance of a location relative to other locations to get the weight of that location in all locations. They will be described in Sections 3.2 and 3.3.

On the whole, we combine global and local information and further explore spatial and channel attention, so as to perform vehicle Re-ID tasks in a highly robust and efficient manner. The contribution of this paper mainly includes the following four aspects.

1) An adaptive attention network with three branches is constructed, which aims to extract a large number of useful information.
2) Based on the global relationship of nodes, two attention modules, namely, CGAM and SGAM, are built. The average pairwise relationship between nodes is used to model the global relationship of nodes and infer the importance of nodes, which reduces the computational complexity and improves the reliability of the extracted salient features.
3) On the local branch, we only horizontally divide the feature map obtained from the vehicle image into two parts, which can solve the problem of misalignment and within-part consistency disruption to a great extent.
4) Experiments on two vehicle Re-ID datasets verify the effectiveness of GRMF. It achieves better performance than SOTA methods.

II. RELATED WORKS

A. VEHICLE RE-IDENTIFICATION

With the development of CNN and the introduction of large-scale vehicle datasets, the performance of vehicle Re-ID has improved significantly. At present, vehicle Re-ID mainly focuses on how to extract more informative features. Guo *et al.* [21] and Zhu *et al.* [22] sought a better feature encoding method. He *et al.* [23] were based on the local to extract highly distinguished features. Zhou *et al.* [24] adopted the adversarial training architecture and the aware attention module to realize meaningful feature inference. Zakria *et al.* [25] used some appearance attributes and license plate information to identify the target vehicle. Liu *et al.* [19] proposed a two-branch FNC network to capture multiple useful information. Shen *et al.* [26] designed a network framework that can extract the temporal and spatial information of vehicle images and integrate them to improve vehicle Re-ID performance. Some works [27]–[29] automatically located vehicle key points through existing models and then used them to learn local feature representations.

B. ATTENTION MECHANISM

Attention mechanism plays an essential role in the human brain. When the human brain receives external information, such as visual and auditory information, the brain often only focuses on some important information rather than all. This helps to filter unimportant information and improves the efficiency of information processing.

For network models, an attention mechanism can make the model focus on the target subject rather than the background, and it is widely used in Re-ID tasks. In Two-level Attention network supervised by a Multi-grain Ranking loss (TAMR) [17], Guo *et al.* proposed a two-level attention network, which can adaptively extract distinctive features from vehicle images. Fang *et al.* [30] proposed an Attention in Attention (AiA) mechanism to make full use of the extracted high-order statistical information. In [31], Wang *et al.* designed a Fully Attentional Block (FAB), which considers both spatial and channel information to make the learned features more robust. Recently, some work shows that the relationship between variables plays an important role in constructing attention mechanism and extracting detailed information. Xia *et al.* [20] proposed a SONA Module to directly model long-distance relationships through a large amount of feature statistical information. Liu *et al.* [19] utilized the relationship between spatial positions to construct spatial attention block (SAB). Cheng *et al.* [32] used the relationship between three-dimensional data to extract deep information, which greatly reduced the dimensional reduction and computational algorithm complexity. In this paper, we use the pairwise relationship between nodes to model the global relationship of nodes and then infer the attention weight of nodes.

C. LOCAL FEATURE EXTRACTION

In the process of global feature learning, some insignificant information is easily ignored, which makes the network unable to capture discriminative information. To address this problem, Wang *et al.* [16] designed a multiple granularity network (MGN), which consists of two local branches and one global branch. To extract local information, the images are divided into several parts. Other studies have suggested similar ideas. TAMR [17] utilized the STN [18] network to locate the head and the windshield on two branches, respectively. In [33], Shervan Fekri-Ershad proposed an improved version of local ternary patterns (ILTP) as texture descriptor which extracts high-discriminative features and is more resistant to rotation and noise. In [15], Dai *et al.* designed a "feature dropping branch," in which some information of the feature map will be deleted. In general, the purpose of these methods is to let the network extract subtle local features from the image, thereby improving the accuracy of recognition.

III. THE PROPOSED METHOD

In this section, we describe our proposed GRMF Network (GRMF-Net). The network consists of (1) a backbone architecture similar to what was used in MGN [16]; (2) Two global branches with an attention module and a local branch that divides an image into two parts; and (3) Improved spatial and channel attention module.

A. NETWORK ARCHITECTURE

FIGURE. 1 shows the overall network architecture, including a backbone network, a local branch, and two global branches. For the backbone network, we choose ResNet-50 [9] as the baseline for feature extraction. Similar to previous work [2], [16], [34], [35], we further modify the original ResNet-50 by adjusting the stages and removing the original fully connected layers for multi-loss training. The ResNet50 backbone is split into three branches after the *res_conv4_1* residual block. To enhance the resolution, we changed the stride size of the down-sampling in the *res_conv5_1* block of *Global-C* Branch and *Local Branch* from 2 to 1. Then, we add spatial and channel attention after the *res_conv5* block on the two global branches, respectively, which will enhance discrimination feature information. We use global average pooling (GAP) to cover the whole body parts of the vehicle image. For the *Local Branch*, we only divide feature maps into two parts horizontally, which can solve the problem of misalignment and within-part consistency disruption to a great extent. Subsequent experiments show that this branch can improve the performance of the model. The Reduction module includes a 1*1 convolution, a BN layer, and a rectified linear unit (ReLU) function. We apply it on each branch to reduce the dimension to 256, which enables the network to obtain a more compact feature representation. In our work, each branch is trained with Triplet Loss and Cross-Entropy Loss. During testing phases, we concatenate the features before the fully-connected (FC) layer of the three branches as the final feature.

B. CHANNEL GLOBAL ATTENTION MODULE

The overview of the CGAM is displayed in FIGURE. 2. Let tensor $x \in \mathbb{R}^{C\times H\times W}$ be the input to the CGAM, where C is the number of the channels, H and W are the spatial height and width of the tensor, respectively. We get the tensors x^e and x^f from the functions e and f, and reshape x^e to $C \times HW$, reshape x^f to $HW \times C$. FIGURE. 3 shows the $e(x)$

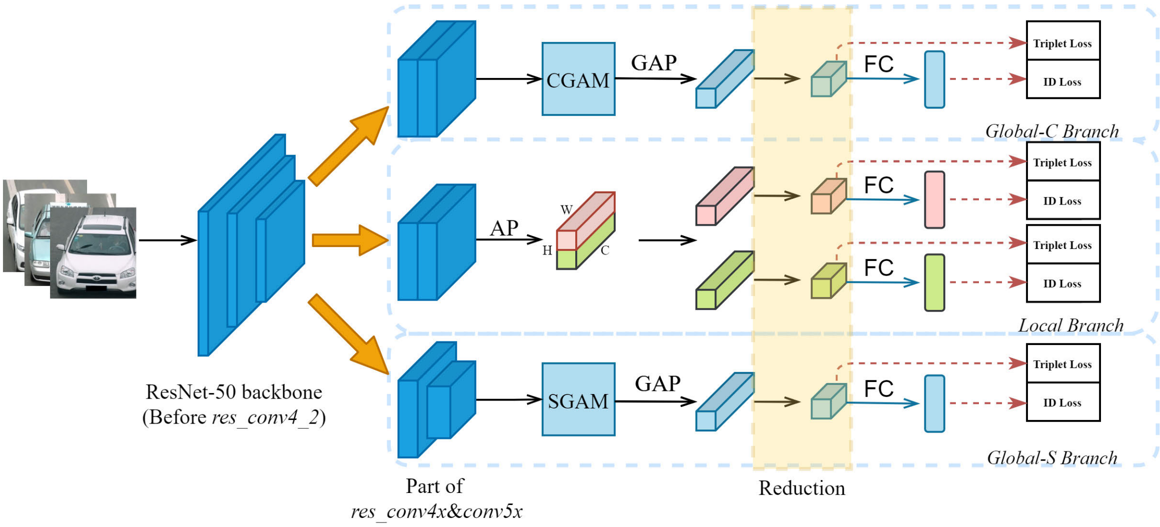


FIGURE 1. **The overall architecture of the proposed GRMF-Net for the vehicle Re-ID task. The ResNet50 backbone is split into three branches after *res_conv4_1* residual block: *Global-C Branch*, *Local Branch*, and *Global-S Branch*. GAP refers to Global Average pooling. CGAM and SGAM refer to channel and spatial global attention modules, respectively. AP refers to Average pooling. Reduction is a dimension reduction module composed of a 1*1 convolution layer, a BatchNormal, and a ReLU function.**

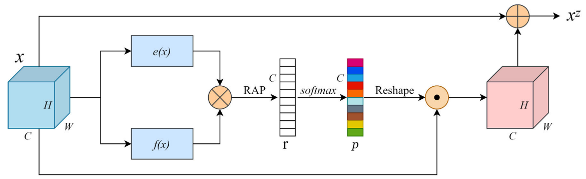


FIGURE 2. **The architecture of the channel global attention module (CGAM). *e*(*x*) contains four convolutions, two BN layers, and two ReLU functions to increase the receptive field and reduce the number of parameters. *f*(*x*) is the same as *e*(*x*). RAP refers to Relation Average pooling. In the CGAM, we apply RAP to determine the weight of each channel among all channels. "⊗" indicates matrix multiplication, "⊙" indicates Hadamard product and "⊕" indicates element-wise sum.**

architecture. In detail, we utilize two 3*3 Group Convolution to increase the receptive field and reduce the number of parameters. Subsequently, we apply matrix multiplication to x^e and x^f to obtain the matrix $R_c \in \mathbb{R}^{C\times C}$, which represents the pairwise relations for all the channels. The R_c can be written as:

$$R_c = x^e \otimes x^f. \tag{1}$$

where "⊗" represents matrix multiplication.

In addition, each row element of the matrix R_c represents the pairwise relationships between each channel and all other channels. We model the average pairwise relationship to obtain the global relationship of a channel. Then, we utilize the global relationship importance of a channel relative to other channels to get the weight of that channel in all channels. Specifically, we apply the relation average pooling (RAP) to matrix R_c to get a vector $\mathbf{r} = (r_1, r_2, r_3, \ldots, r_C) \in \mathbb{R}^C$, where C is the number of the channels. At this time, each element of vector $\mathbf{r}$ represents the global relationship between each channel and all channels. The n-th element of the vector $\mathbf{r}$ is defined as $\{\mathbf{r}_n | n \in \{1, 2, 3, \ldots, C\}\}$. The $\mathbf{r}_n$ is calculated by

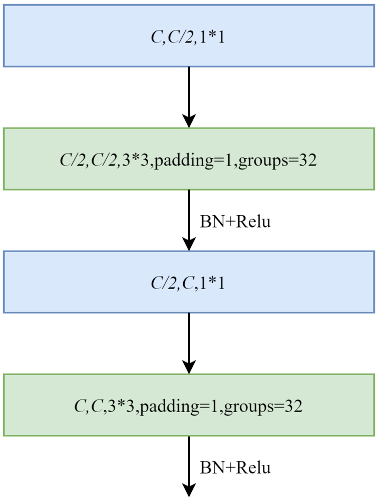


FIGURE 3. **The overall architecture of the *e*(*x*). The blue boxes represent convolutions of size 1*1, and the green boxes represent group convolutions of size 3*3. *C* refers to the number of channels.**

$$\mathbf{r}_n = \frac{\sum_{m=1}^{C} R_c(n, m)}{C}. \tag{2}$$

Then, we adopt the softmax function to change all global relationships into the weight of each channel. The global relationship reflects the intimacy between the node and the whole. By using the global relationship to directly model the learning of the node's attention, the network can extract more reliable and significant information.

$$p = softmax(\mathbf{r}). \tag{3}$$

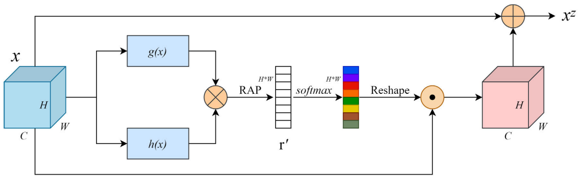


FIGURE 4. The architecture of the spatial global attention module (SGAM). $g(x)$ contains a 1*1 convolution, a BN layer, and a ReLU function to reduce the number of channels C to C/r. $h(x)$ is the same as $g(x)$. In the SGAM, we apply RAP to determine the weight of each location among all spatial locations.

In order to acquire an attention map x^m, we first reshape the vector p to $C\times1\times1$, and then reshape it to $C\times H\times W$. Finally, we apply element-wise multiplication and element-wise sum to the origin feature map to obtain the final feature x^z. The x^z can be expressed as:

$$x^z = x + x \odot x^m. \tag{4}$$

C. SPATIAL GLOBAL ATTENTION MODULE

FIGURE. 4 shows the SGAM architecture. Spatial attention and channel attention utilize the relationships between locations and between channels, respectively, to determine the importance of each location and channel, and they work in a similar way. But compared with CGAM, SGAM has three differences. First, let tensor $x \in \mathbb{R}^{C\times H\times W}$ be the input to the SGAM. $g(x)$ contains a 1*1 convolution, a BN layer, and a ReLU function to reduce the number of channels C to C/r. r refers to the reduction factor, which is set to 2 in our experiments, $h(x)$ plays the same role as $g(x)$. We get the tensors x^g and x^h from the functions g and h, and reshape x^g to $HW \times C/r$, reshape x^h to $C/r \times HW$. Then, we adopt matrix multiplication to determine the Non-local relation and obtain matrix $R_s \in \mathbb{R}^{HW\times HW}$.

$$R_s = x^g \otimes x^h. \tag{5}$$

Second, to determine the importance level of a location, we apply RAP to matrix R_s to get a vector $\mathbf{r}' \in \mathbb{R}^{HW}$. The n-th member of the vector $\mathbf{r}'$ can be expressed as:

$$\mathbf{r}'_n = \frac{\sum_{m=1}^{H*W} R_s(n,m)}{H*W}, \quad n \in \texttt{\{1,2,3,\textbackslash ldots,H*W\}}. \tag{6}$$

Third, we first reshape the vector generated by the softmax function to $1 \times H \times W$, and then reshape it to $C \times H \times W$.

It should be noted that in CGAM and SGAM, we add both the attention feature map and the original feature map to obtain the final output feature. There are two reasons for using the addition operation here. First, the normalized function used here is Softmax, the function of Softmax is to map the weight value to the range from 0 to 1, and the sum of all weight values is 1. Due to the existence of a large number of weights, the feature map element value of the output of the attention module may be small, which will break the characteristics of the original network and bring great difficulties to training. Second, when the attention map x^m is equal to 0, the output of the attention module is equal to the original feature map x, so the effect of the attention is impossible to be worse than the original x, which also refers to the idea of identity mapping in ResNet. At the same time, this addition operation also highlights the useful features in the feature map x. The experiment also shows the model has very good performance through this residual structure. In comparison with the model without addition operation, the GRMF can achieve a 1.2%/1.5% advance on mAP and Top-1.

D. LOSS FUNCTIONS

We use the most common cross-entropy loss and triplet loss.

1) CROSS-ENTROPY LOSS FUNCTION

Cross-entropy can measure the degree of difference between two different probability distributions in the same random variable. The cross-entropy loss can be expressed as:

$$L_{\text{ID}} = \sum_{i=1}^{N_1} -q_i \log(p_i) \begin{cases} q_i = 0, & y \neq i \\ q_i = 1, & y = i. \end{cases} \tag{7}$$

where N_1 represents the number of classes, y is the truth label of ID, and p_i is the ID prediction logits of class i.

2) TRIPLET LOSS

The goal of the triplet loss is to make the samples with the same label as close as possible in the embedding space, while the samples with different labels keep the distance as far as possible. In this paper, we use the batch-hard triplet loss [11], an improved version based on the original semi-hard triplet loss [36]. We randomly select P identities and K images for each mini-batch from the training set to cooperate with the requirement of triplet loss. It enhances the robustness in metric learning and further improves the model's performance. The batch-hard triplet loss can be expressed as:

$$L_{\text{Triplet}} = \sum_{i=1}^{P}\sum_{a=1}^{K}\Big[\alpha + \max_{p=1,\ldots,K} \|a_i - p_i\|_2 - \min_{n=1,\ldots,K,j=1,\ldots,P,j\neq i} \|a_i - n_j\|_2\Big]_+ . \tag{8}$$

where a_i, p_i, and n_j are the features extracted from anchor, positive and negative samples receptively. We set the value of α to 1.2 as in [16].

3) TOTAL LOSS

The total loss is the sum of the cross-entropy loss and the batch-hard triplet loss, which is defined as:

$$L_{\text{total}} = \lambda_1 \sum_{i=1}^{N} L_{ID}^{i} + \lambda_2 \sum_{j=1}^{N} L_{Triplet}^{j}. \tag{9}$$

where λ_1 and λ_2 are hyperparameters to balance the two loss terms and are both set to 1 in our experiments. N is set to 4.

IV. EXPERIMENTS

A. DATASETS AND EVALUATION PROTOCOL

1) DATASETS

We evaluate our model on two popular vehicle Re-ID datasets, including VeRi776 [37] and VehicleID [38].

VeRi776: It consists of about 50,000 images of 776 cars taken from 20 cameras in different locations at different viewing angles. The training set contains 576 vehicles, and the test set includes the remaining 200 vehicles.

VehicleID: It is a large-scale dataset composed of 221,763 images. These images are obtained by many cameras shooting about 26,267 vehicles from different perspectives. Each image in the dataset has the corresponding ID attribute, but only a part has the model attribute. Three test sets were extracted based on size, namely small, medium, and large. In the test stage, each image in the gallery set belongs to different vehicles. Therefore, only one image in the test results is correct at most.

2) EVALUATION PROTOCOL

In order to make a fair comparison with the existing methods, we adopt the mean average precision (mAP) and the cumulative matching characteristics (CMC) to evaluate our proposed model.

B. IMPLEMENTATION DETAILS

ResNet50 [9] is chosen as the backbone to generate the feature. We adopt the same training strategy on two datasets. We adjust the image size to 256*256 before inputting the image to the network. Each mini-batch is sampled with selected P identities and selected K instances for each identity from the training set to meet the demand of batch-hard triplet loss. For hyperparameters P and K, we set them to 16 and 4. In all our experiments, we set the margin parameter of triple loss to 1.2. We choose Adam as the optimizer. For the learning rate strategy, we set the initial learning rate to 2e-4, which decays to 2e-5 after the 120th epoch and further drops to 2e-6 and 2e-7 in the 220th and 320th epochs for faster convergence. The total training process lasts for 450 epochs. We adopt cross-entropy loss and batch-hard triplet loss [11] together to train all the branches.

During testing phases, the Veri776 dataset is tested in the form of image-to-track. By calculating the distance between the query image and all images in the gallery track, we regard the minimum distance of image-to-image as the distance of image-to-track. For the VehicleID dataset, we test its three test sets, respectively. We concatenate four reduced 256-dim features of the three branches to form a 1024-dim feature vector as the final feature representation. All experiments are implemented with PyTorch 1.9 on 2 Nvidia GTX 2080Ti GPUs.

C. EXPERIMENTAL RESULTS

As shown in FIGURE. 5, we plot the variation in the performance of GRMF on the VeRi776 dataset during training. Note that the rank metric in the figure is the same as the CMC. For example, rank1 and CMC@1 are equal. Then, we compare the results of the proposed model with other state-of-the-art models on two datasets. Local Maximal Occlusion Representation (LOMO) [45] was designed to solve the problem of different visual and light variations. In order to obtain better results on the CompCars [39] dataset, the Googlenet [40] model was fine-tuned, and the fine-tuned model was called GoogLeNet [41]. FACT [42] discriminated vehicles in joint domains by utilizing SIFT, Color Name, and GoogLeNet features. Embedding Adversarial Learning Net (EALN) [43] designed a new network that can adaptively generate a large number of samples. Region-aware deep model (RAM) [44] first divided the image horizontally into three parts and then embedded detailed visual cues in these local areas. To improve the ability to perceive subtle differences, Part-regularized Near-duplication (PRN) [23] introduced the local normalization (PR) constraint into the vehicle Re-ID task. Parsing-based view-aware embedding network (PVEN) [46] can avoid the mismatch of local features under the different perspectives. Generative Adversarial Networks (GAN) [47] used Generative and Discriminative models to learn from each other to produce good output. With the help of GAN [47], viewpoint-aware attentive multi-view inference (VAMI) [48] generated features of different views. TAMR [17] proposed the two-level attention network to gradually focus on the subtle but distinct local details in the visual appearance of vehicles, and proposed a multi-grain Ranking loss to learn a structured deep feature embedding. Multi-attention-based soft partition (MUSP) [50] divided vehicle areas and extracted features without metadata using attention. Part-Mentored Attention Network (PMANet) [51] designed a two-stage attention structure and performed a coarse-to-fine search among vehicles.

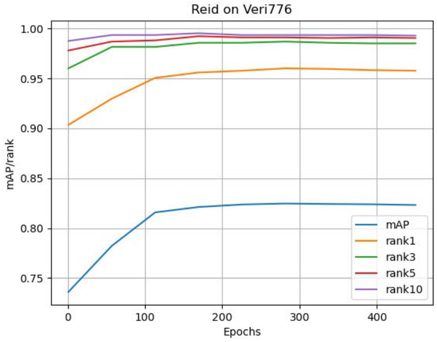


FIGURE 5. The performance of GRMF on test set of VeRi776.

The experimental results on VeRi776 and VehicleID are detailed in TABLE 1 and TABLE 2, respectively. From TABLE 1, we find that, first, the GRMF obtains an improvement of 2.7% on mAP and 0.1% CMC@1 over the PVEN. Second, the CMC@5 of our method has exceeded the 99.1%, which is a promising performance for the actual vehicle Re-ID scenario. TABLE 2 shows the comparison results on

TABLE 1. Performance (mAP, CMC@1 and CMC@5) comparisons with the state-of-the-arts on VeRi776.

Method	mAP	CMC@1	CMC@5
LOMO [45]	0.096	0.253	0.465
GoogLeNet [41]	0.170	0.498	0.712
FACT [42]	0.185	0.510	0.735
EALN [43]	0.574	0.844	0.941
RAM [44]	0.615	0.886	0.940
PRN [23]	0.743	0.943	0.989
PVEN [46]	0.795	0.956	0.984
SPAN [49]	0.689	0.940	0.976
MUSP [50]	0.78	0.956	0.979
PMANet [51]	0.818	**0.964**	0.986
GRMF (ours)	**0.822**	0.957	**0.991**

TABLE 2. Performance (CMC@1, CMC@5) comparisons with the state-of-the-arts on VehicleID.

Method	small		medium		large	
	@1	@5	@1	@5	@1	@5
LOMO [45]	0.197	0.320	0.188	0.291	0.153	0.252
GoogLeNet [41]	0.478	0.671	0.434	0.638	0.382	0.593
VAMI [48]	0.631	0.832	0.528	0.751	0.473	0.702
TAMR [17]	0.660	0.797	0.629	0.768	0.596	0.738
FACT [42]	0.495	0.680	0.445	0.645	0.399	0.603
EALN [43]	0.751	0.881	0.718	0.839	0.693	0.814
PRN [23]	0.784	0.923	0.750	0.883	**0.742**	0.864
RAM [44]	0.752	0.915	0.723	0.870	0.677	0.845
GRMF (ours)	**0.815**	**0.963**	**0.768**	**0.936**	0.728	**0.910**

three test datasets with different sizes. We observe that our GRMF achieve the improvement at the CMC@5 by 4.0%+ over the SOTA PRN on different test data. Moreover, note that some advanced network models need to utilize other auxiliary models, which increases the algorithm's complexity. For example, PVEN [46] used the U-Net [52] to parse a vehicle into four different views. PRN [23] took YOLO [53] as the detection network of part positioning. TAMR adopted STN to localize on vehicle head and windshield parts automatically. However, our model still has better performance without any auxiliary model.

Compared with other models, our GRMF model achieves the best performance. We report mAP of 82.24%, CMC@1 of 95.77% and CMC@5 of 99.11% on the test set of VeRi776 and CMC@1 of 81.51%, 95.54%, 72.81% and CMC@5 of 96.38%, 93.69%,91.01% on three test sets of VehicleID. All the results are achieved under the single-query mode without re-ranking.

D. ABLATION STUDY

To validate the effectiveness of critical components in GRMF and find the optimal structure, we conduct extensive comparison experiments on VeRi776 and VehicleID datasets. Detailed experimental results are shown in Tables 3 and 4.

1) THE EFFECTIVENESS OF CGAM AND SGAM

Our model consists of three branches. On the two global branches, channel attention and spatial attention are used to pay attention to discriminative features. We separately validate the effects of SGAM and CGAM on the model's advance and show the results in TABLE 3. We have the following observations. "Baseline+SGAM" brings an improvement of 0.6% in mAP and 0.6% in CMC@1 on the test set of VeRi776 compared to Baseline. This demonstrates the effectiveness of leveraging global relations for learning attention. "Baseline + SGAM + CGAM" consists of two branches with SGAM and CGAM. The model yields 5.0%/1.6% mAP and CMC@1 accuracy advance compared with the Baseline. When removing SGAM or CGAM from GRMF, the performance of the model decreases. This indicates that SGAM and CGAM can capture important discriminative information from spatial and channel dimensions. Note that compared with "Baseline + SGAM" and "Baseline (ResNet50)", "Baseline + SGAM + CGAM" has a significant performance improvement. This is because SGAM and CGAM can provide complementary information so that the two modules can extract more useful information under cooperation.

Furthermore, we make a qualitative analysis of the attention module to see its effectiveness more intuitively. FIGURE. 6 shows the qualitative visualization results of our

TABLE 3. "CGAM" and "SGAM" are the channel attention module and spatial attention module, respectively. "w/o" refers to without. The results are on the test set of VeRi776.

Method	mAP	CMC@1
Baseline (ResNet50)	73.8%	92.0%
Baseline + SGAM	74.4%	92.6%
Baseline + SGAM + CGAM	78.8%	93.6%
GRMF w/o CGAM	81.7%	94.5%
GRMF w/o SGAM	81.6%	95.3%
GRMF (ours)	**82.2%**	**95.7%**

TABLE 4. Experiments to verify the ***Local Branch*** on VeRi776. "w/o" refers to without. "local" refers to the local branch of GRMF. "Part-3" and "Part-4" refer to dividing the feature map into three or four parts, respectively.

Method	mAP	CMC@1
GRMF w/o local	78.8%	93.6%
GRMF (Part-3)	81.7%	95.1%
GRMF (Part-4)	81.6%	94.6%
GRMF (ours)	**82.2%**	**95.7%**

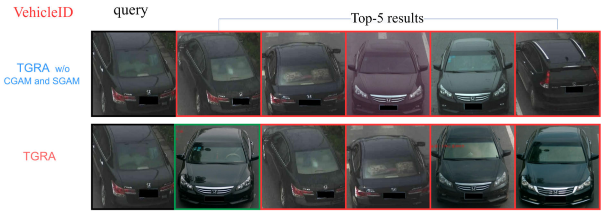


FIGURE 6. Visualization of the ranking list on vehicle Re-ID task. The images in the first column are the query images. The rest images are retrieved top-5 ranking results. In the retrieval results, the images with a green border are correct, and the rest images with a red border are wrong.

GRMF on the VehicleID. As can be seen from the figure, the network with the attention module can accurately find the same vehicle image. Although it is difficult to identify the same vehicle when the query and target images are in different views, our model can also identify the same vehicle very well. So, our global relational attention module has a good performance in enhancing discrimination pixels and suppressing noise pixels.

2) THE VALIDATION OF THE LOCAL BRANCH

We investigated the effectiveness of the *Local Branch* by comparing GRMF and ''GRMF w/o local''. We cut off the *Local Branch* of the GRMF model to obtain the model ''GRMF w/o local''. ''GRMF w/o local'' has only two global branches. Therefore, ''GRMF w/o local'' can extract global information from vehicle images, but cannot obtain detailed local information. To fully verify the effectiveness of our proposed *Local Branch*, we further conduct two experiments. One is to divide the image into three parts, and the other is to divide it into four parts. As in TABLE 4, we observe that, first, among the four models, the performance of GRMF without the *Local Branch* is the worst, indicating that local detail information is crucial in the vehicle re-identification task. Second, compared to ''GRMF (Part-3)'', ''GRMF (ours)'' brings an improvement of 0.5% in mAP and 0.6% in CMC@1 on the test set of VeRi776. In addition, the more parts, the worse performance. This is caused by misalignment and within-part disruption. However, our proposed local branch can be largely solved these questions. The ablation experiment strongly proves its effectiveness.

V. CONCLUSION

In this paper, we propose a three-branch adaptive attention network for vehicle Re-ID. The model can extract the distinctive features of vehicles from multiple angles. In addition, to solve the problem of misalignment and within-part consistency disruption to a great extent, we only divide feature maps into two parts evenly. The last, through the improved attention, the network can focus on the most important part and learn more discriminant and robust features in the vehicle re-identification task. In the inference stage, we concatenate the features of the three branches for better performance. Experimental results demonstrate that our model is significantly superior to the current best on the VeRi776 and VehicleID datasets.

In the future, we will explore more lightweight and effective attention modules to overcome the large amount of computation brought by computing all pairwise relationships.

REFERENCES

[1] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, ''Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline),'' in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 480–496.

[2] F. Zheng, C. Deng, X. Sun, X. Jiang, X. Guo, Z. Yu, F. Huang, and R. Ji, ''Pyramidal person re-identification via multi-loss dynamic training,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8514–8522.

[3] M. M. Kalayeh, E. Basaran, M. Gokmen, M. E. Kamasak, and M. Shah, ''Human semantic parsing for person re-identification,'' in *Proc. CVPR*, 2018, pp. 1062–1071.

[4] 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.

[5] H. Zhao, M. Tian, S. Sun, S. Jing, and X. Tang, ''Spindle net: Person re-identification with human body region guided feature decomposition and fusion,'' in *Proc. CVPR*, 2017, pp. 1077–1085.

[6] W. Li, X. Zhu, and S. Gong, ''Harmonious attention network for person re-identification,'' in *Proc. CVPR*, 2018, pp. 2285–2294.

[7] W. Zhang, S. Hu, K. Liu, and Z. Zha, ''Learning compact appearance representation for video-based person re-identification,'' 2017, *arXiv:1702.06294*.

[8] J. Liu, Z. Zha, D. Chen, R. Hong, and M. Wang, ''Adaptive transfer network for cross-domain person re-identification,'' in *Proc. CVPR*, 2019, pp. 7202–7211.

[9] K. He, X. Zhang, S. Ren, and J. Sun, ''Deep residual learning for image recognition,'' in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[10] K. Simonyan and A. Zisserman, ''Very deep convolutional networks for large-scale image recognition,'' in *Proc. Int. Conf. Learn. Represent.*, May 2015, pp. 1–14.

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

[12] Y. Sun, L. Zheng, W. Deng, and S. Wang, ''SVDNet for pedestrian retrieval,'' in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3800–3808.

[13] Z. Zheng, L. Zheng, and Y. Yang, ''Pedestrian alignment network for large-scale person re-identification,'' *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 10, pp. 3037–3045, Oct. 2019.

[14] D. Chen, D. Xu, H. Li, N. Sebe, and X. Wang, ''Group consistent similarity learning via deep CRF for person re-identification,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8649–8658.

[15] Z. Dai, M. Chen, X. Gu, S. Zhu, and P. Tan, ''Batch dropblock network for person re-identification and beyond,'' 2018, *arXiv:1811.07130*.

[16] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou, ''Learning discriminative features with multiple granularities for person re-identification,'' in *Proc. 26th ACM Int. Conf. Multimedia*, Oct. 2018, pp. 274–282.

[17] H. Guo, K. Zhu, M. Tang, and J. Wang, ''Two-level attention network with multi-grain ranking loss for vehicle re-identification,'' *IEEE Trans. Image Process.*, vol. 28, no. 9, pp. 4328–4338, Sep. 2019.

[18] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, ''Spatial transformer networks,'' in *Proc. NIPS*, Jun. 2015, pp. 2017–2025.

[19] K. Liu, Z. Xu, Z. Hou, Z. Zhao, and F. Su, ''Further non-local and channel attention networks for vehicle re-identification,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 2494–2500.

[20] N. Xia, Y. Gong, Y. Zhang, and C. Poellabauer, ''Second-order non-local attention networks for person re-identification,'' in *Proc. ICCV*, 2019, pp. 3759–3768, 2019.

[21] H. Guo, C. Zhao, Z. Liu, J. Wang, and H. Lu, ''Learning coarse-to-fine structured feature embedding for vehicle re-identification,'' in *Proc. Conf. Artif. Intell.*, 2018, pp. 6853–6890.

[22] J. Zhu, H. Zeng, J. Huang, S. Liao, Z. Lei, C. Cai, and L. Zheng, ''Vehicle re-identification using quadruple directional deep learning features,'' *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 1, pp. 410–420, Jan. 2019.

[23] 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. 3997–4005.

[24] Y. Zhou and L. Shao, ''Aware attentive multi-view inference for vehicle reidentification,'' in *Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6489–6498.

[25] C. Zakria, ''Efficient and deep vehicle re-identification using multi-level feature extraction,'' *Appl. Sci.*, vol. 9, no. 7, p. 1291, 2019.

[26] Y. Shen, T. Xiao, H. Li, S. Yi, and X. Wang, ''Learning deep neural networks for vehicle re-ID with visual-spatio-temporal path proposals,'' in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1900–1909.

[27] Z. Wang, L. Tang, X. Liu, Z. Yao, S. Yi, J. Shao, J. Yan, S. Wang, H. Li, and X. Wang, ''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] P. Khorramshahi, N. Peri, A. Kumar, A. Shah, and R. Chellappa, "Attention driven vehicle re-identification and unsupervised anomaly detection for traffic understanding," in *Proc. CVPR Workshops*, 2019, pp. 239–246.

[29] P. Khorramshahi, A. Kumar, N. Peri, S. S. Rambhatla, J. Chen, and R. Chellappa, "Adual-path model with adaptive attention for vehicle re-identification," in *Proc. Int. Conf. Comput. Vis.*, 2019, pp. 6132–6141.

[30] P. Fang, J. Zhou, S. K. Roy, L. Petersson, and M. Harandi, "Bilinear attention networks for person retrieval," in *Proc. ICCV*, 2019, pp. 8029–8038.

[31] 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. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 365–381.

[32] K. Cheng, M. Khokhar, M. Ayoub, and Z. Jamali, "Nonlinear dimensionality reduction in robot vision for industrial monitoring process via deep three dimensional Spearman correlation analysis (D3D-SCA)," *Multimedia Tools Appl.*, vol. 80, no. 4, pp. 5997–6017, 2021.

[33] S. Fekri-Ershad, "Bark texture classification using improved local ternary patterns and multilayer neural network," *Expert Syst. Appl.*, vol. 158, Nov. 2020, Art. no. 113509.

[34] Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao, and J. R. Smith, "Learning locally-adaptive decision functions for person verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3610–3617.

[35] Z. Dai, M. Chen, X. Gu, S. Zhu, and P. Tan, "Batch dropblock network for person re-identification and beyond," 2018, *arXiv:1811.07130*.

[36] 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.

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

[38] H. Liu, Y. Tian, Y. Wang, L. Pang, and T. Huang, "Deep relative distance learning: Tell the difference between similar vehicles," in *Proc. CVPR*, 2016, pp. 2167–2175.

[39] P. Luo, C. C. Loy, and X. Tang, "A large-scale car dataset for fine-grained categorization and verification," in *Proc. CVPR*, 2015, pp. 3973–3981.

[40] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. CVPR*, 2015, pp. 1–9.

[41] L. Yang, P. Luo, C. C. Loy, and X. Tang, "A large-scale car dataset for fine-grained categorization and verification," in *Proc. CVPR*, 2015, pp. 3973–3981.

[42] X. Liu, W. Liu, H. Ma, and H. Fu, "Large-scale vehicle re-identification in urban surveillance videos," in *Proc. ICME*, 2016, pp. 1–6.

[43] 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.

[44] X. Liu, S. Zhang, Q. Huang, and W. Gao, "RAM: A region-aware deep model for vehicle re-identification," in *Proc. ICME*, 2018, pp. 1–6.

[45] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proc. CVPR*, Jun. 2015, pp. 2197–2206.

[46] D. Meng, L. Li, X. Liu, Y. Li, S. Yang, Z. Zha, X. Gao, S. Wang, and Q. Huang, "Parsing-based view-aware embedding network for vehicle re-identification," in *Proc. CVPR*, 2020, pp. 7101–7110.

[47] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[48] Y. Zhou and L. Shao, "Viewpoint-aware attentive multi-view inference for vehicle re-identification," in *Proc. CVPR*, 2018, pp. 6489–6498.

[49] 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. ECCV*, 2020, pp. 330–346.

[50] S. Lee, T. Woo, and S. Lee, "Multi-attention-based soft partition network for vehicle re-identification," *CoRR*, vol. abs/2104.10401, pp. 1–10, Apr. 2021.

[51] L. Tang, Y. Wang, and L. Chau, "Looking twice for partial clues: Weakly-supervised part-mentored attention network for vehicle re-identification," *CoRR*, vol. abs/2107.08228, pp. 1–13, Jan. 2021.

[52] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf.*, Munich, Germany, Oct. 2015, pp. 234–241.

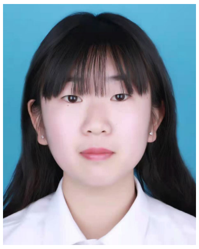
[53] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

XIN TIAN received the B.E. degree in computer science and technology from Shandong Jiaotong University, Jinan, China, in 2020, where he is currently pursuing the M.E. degree. His research interests include deep learning, neural networks, and image recognition.

XIYU PANG received the M.S. degree in computer science from the College of Computer Science, Shandong University, Jinan, China, in 2007. He is currently an Associate Professor with the School of Information Science and Electrical Engineering, Shandong Jiaotong University. His research interests include machine learning and data mining and image processing.

GANGWU JIANG received the B.E. degree in computer science and technology from Shandong Jiaotong University, Jinan, China, in 2020, where he is currently pursuing the M.E. degree. His research interests include deep learning, machine learning, and image recognition.

QINGLAN MENG received the B.E. degree in computer science and technology from Shandong Jiaotong University, Jinan, China, in 2019, where she is currently pursuing the M.E. degree. Her research interests include deep learning, image segmentation, and neural networks.



YANLI ZHENG received the B.E. degree in computer science and technology from Shandong Jiaotong University, Jinan, China, in 2020, where she is currently pursuing the M.E. degree. Her research interests include deep learning, computer vision, and image recognition.

• • •