Vehicle Re-identification in Multi-branches Network

Leilei Rong
Shandong University of Science & Technology
Yan Xu (x1y5@163.com)
Shandong University of Science & Technology
Xiaolei Zhou
Shandong University of Science & Technology
Lisu Han
Shandong University of Science & Technology
Linghui Li
Shandong University of Science & Technology
Xuguang Pan
Shandong University of Science & Technology

Research Article

Keywords: Vehicle re-identification, multi-branches network, Global-Local feature fusion, Channel Attention Mechanism, Weighted Local Feature

Posted Date: June 9th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-598208/v1

Vehicle Re-identification in Multi-branches Network

Leilei Rong[1*], Yan Xu[1*✉], Xiaolei Zhou[1], Lisu Han[1], Linghui Li[1], Xuguang Pan[1]

1 College of Electronic and Information Engineering, Shandong University of Science & Technology, Qingdao 266590, China

E-mail addresses: rong15305385721@163.com (L. Rong), 1462873363@qq.com (X. Zhou), 1164953726@qq.com (L. Han), 357206527@qq.com (L. Li), 975624756@qq.com (X. Pan)

✉ Corresponding author. E-mail: x1y5@163.com;

* Equal contribution.

Abstract

Vehicle re-identification (Re-ID) aims to solve the problem of matching and identifying the same vehicles under the scene of cross multiple surveillance cameras. Finding the target vehicle quickly and accurately in the massive vehicle database is extremely important for public security, traffic surveillance and applications on smart city. However, it is very challenging due to the orientation variations, illumination changes, occlusion, low resolution, rapid vehicle movement, and amounts of similar vehicle models. In order to overcome these problems and improve the accuracy of vehicle re-identification, a multi-branches network is proposed, which is integrated by global-local feature fusion, channel attention mechanism, and weighted local feature. First, the fusion of global and local features is to obtain more complete information of the vehicle and enhance the learning ability of the model; second, the purpose of embedding the channel attention module in the feature extraction branch is to extract the personalized feature of the vehicle; finally, the influence of sky area and noise information on feature extraction is weakened by weighted local feature. The comprehensive experiments implemented on the mainstream evaluation datasets including VeRi-776, VRIC, and VehicleID indicate that our method can effectively improve the accuracy of vehicle re-identification and is superior to the state-of-the-art methods.

Keywords: Vehicle re-identification, multi-branches network, Global-Local feature fusion, Channel Attention Mechanism, Weighted Local Feature

1 Introduction

Vehicle re-identification (Re-ID), an intelligent surveillance camera analysis technology, is part and parcel to smart and safer cities. Vehicle Re-ID is similar to person re-identification [1-4] both of them belong to the problem of object re-identification, and are closely related to object recognition and fine-grained classification. The task of vehicle Re-ID is to retrieve a given vehicle among all gallery vehicle images captured across multiple surveillance cameras. However, it's a hard work due to the various viewpoint, occlusion, motion blur, illumination, and the low resolution, as shown in the left five columns(***a***-***e***) of Fig. 1. Furthermore, vehicle Re-ID has its particular challenge that different identities may have similar or even the same appearance especially for the vehicles with the same model, as shown in the last column(***f***) of Fig. 1.

In the field of re-identification, the mainstream method is mainly feature learning method. Learning and extracting more discriminative and robust vehicle features is the principal task of feature learning. For example,

[5] proposed a Shortly and Densely convolutional neural Network (VRSDNet), which utilized a list of short and dense units (SDUs), necessary pooling, and spatial normalization layers to enhance the feature learning ability. [6] encouraged the deep model to place emphasis on more details in local regions and results in more discriminative features. [7] introduced MSDeep to learn multi-scale and multi-level representations for vehicle Re-ID. The proposed MSF and MLF blocks not only extract features of different scales and different levels, but also learn to emphasize the distinct features for different vehicles dynamically, thus comprehensive and discriminative descriptors can be aggregated. The prosperity of feature learning method has introduced more powerful representations with better discrimination and robustness for vehicle images, pushing the performance of Re-ID to a new level.

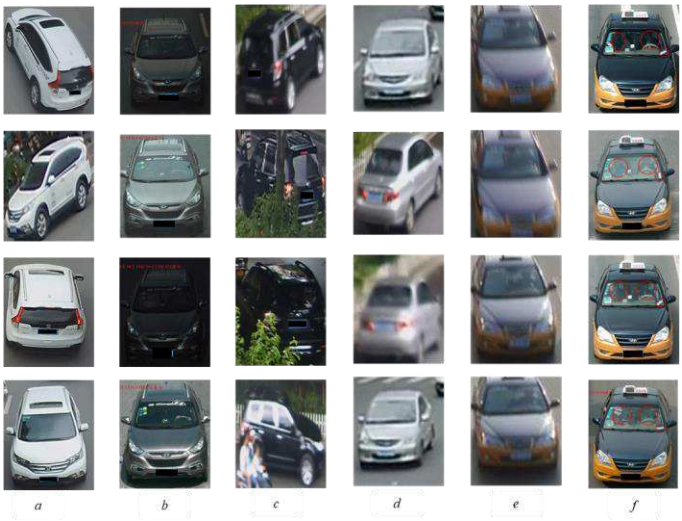

Fig.1 Illustration of challenges in vehicle Re-ID. Some examples of vehicle images selected from the VehicleID, VeRi-776, and VRIC datasets. The vehicle images(***a-e***) in each column are collected with the same vehicle, but their appearances are quite different due to various challenging factors, e.g., viewpoint, illumination, occlusion, low resolution and motion blur. The last column(***f***) illustrates the challenge of different vehicle identities with extremely similar appearance, where the red circles indicate the difference in local features.

Based on the above discussion, aiming at extracting more discriminative and robust feature for vehicle images, we propose a vehicle re-identification method based on global-local feature fusion, channel attention mechanism, and weighted local feature. We first choose ResNet-50 as the backbone network and lead to three characteristic branches (Global Branch, Local Branch1, and Local Branch2) after res_conv5 layer. By fusing global and local features to obtain more complete information of the vehicle and enhance the learning ability of the model. In the second place, we embed the channel attention module in the Local Branch1 and the Local Branch2 so that the network can extract the personalized features of the vehicle. In the last place, the influence of sky area and noise information on feature extraction is weakened by weighted local feature. During the training, each branch does not share the weight, and trains separately. But when testing, all branch information will be assembled into a comprehensive feature to enhance network performance. Finally, extensive

experimental results on three vehicle datasets verify the promising performance of the proposed method compared to state-of-the-art methods.

The rest of this paper is organized as follows. Section 2 briefly introduces the recent related works. Section 3 describes the proposed multi-branches network framework. Section 4 demonstrates the experimental results on public vehicle datasets with comprehensive evaluations on the proposed method, while Sect. 5 concludes this paper.

2 Related Work

In this section, we shall briefly review the recent vehicle re-identification works including vehicle Re-ID, local feature, and attention vehicle Re-ID tasks.

2.1 Vehicle Re-ID datasets

In recent years, Person Re-ID has attracted wide attention, which promotes the research of vehicle re-identification. Several vehicle Re-ID datasets have been proposed. Liu et al. [8] released the first vehicle Re-ID dataset VeRi-776 which contains 37,778 images of 576 vehicles as training set, 11,579 images of 200 vehicles as gallery set and 1,678 images of 200 vehicles as query set. In addition, it provides attributes (color and type) information and a part of license plate information. More recently, Kanaci et al. [9] introduced a more realistic and challenging vehicle Re-ID benchmark, called VRIC. It contains 54,808 images of 2,811 vehicles as training set, 2,811 images of 2,811 vehicles as probe set and 2,811 images of 2,811 vehicles as gallery set. Liu et al. [10] proposed a larger dataset VehicleID with 221,763 images of 26,267 vehicles from multiple real-world surveillance cameras, including the training set with 110,178 images of 13,134 vehicles and testing set with 111,585 images of 13,133 vehicles. The details of these three datasets are shown in Table 1.

Table 1: Characteristics of three benchmark vehicle Re-ID datasets.

Dataset	VeRi-776 [6]	VRIC [7]	VehicleID [8]
Images	51,035	60,430	221,763
Identities	776	5,622	26,267
Cameras	20	120	12
Capture Time	18h	24h	N/A
Views	6	Unconstrained	2
Resolutions Width×Height (Mean)	376.1×345.4	65.9×103.0	345.4×376.1
Motion Blur	No	Unconstrained	No
Illumination	Limited	Unconstrained	Limited
Occlusion	No	Unconstrained	No
Spatio-temporal Relation Annotation	√	×	×
Timestamp	×	×	×
Camera ID	×	×	×
Morning	√	√	√
Afternoon	√	√	√
Night	×	√	×

2.2 Vehicle Re-ID methods

With the prosperity of deep learning, feature learning [11-12] by deep networks has become a mainstream practice in vehicle Re-ID tasks. A lot of effective network architectures have been designed to achieve better matching performance for vehicle Re-ID. Khorramshahi et al. [13] presented a dual path model AAVER, which is a robust end-to-end framework, combining macroscopic global features with localized discriminative features to efficiently identify a probe image in a gallery of varying sizes. Zheng et al. [14] proposed a multi-scale attention framework (MSA) to fusing the discriminative local cues and effective global information. Wang et al. [15] designed AGNet with attribute-guided attention module which could learn global representation with abundant attribute features in an end-to-end manner. He et al. [16] used a simple and efficient part-regularized discriminative feature preserving method, which improves the recognition ability of subtle information. Huang et al. [17] introduced a Position-Dependent Deep Metric unit, which is capable of learning a similarity metric adaptive to local feature structure. Cui et al. [18] designed a network that combined attention mechanisms and long short-term memory network (LSTM) for the recognition of spatial relations.

2.3 Local Feature

In the past, vehicle re-identification methods often just used global feature. Due to the limited scale and weak diversity of vehicle Re-ID training datasets, some non-salient or infrequent detailed information can be easily ignored and even make no contribution for better discrimination during global feature learning procedure, making global features hard to adapt similar inter-class common properties or large intra-class differences. To relieve this dilemma, locating significant vehicle parts from images to represent local information of identities has been confirmed to be an effective approach for better Re-ID accuracy in many previous works [19-21]. Liu et al. [7] explored a Region-Aware deep Model (RAM) to extract regional features from three overlapped local regions and pay more attention to the details in local regions. Zhang et al. [22] proposed a novel Part-Guided Attention Network (PGAN) for vehicle instance retrieval (IR) to extract part regions of each vehicle image from an object detection model. Suprem et al. [23] presented GLAMOR, a small and fast model, which extracts additional global features and performs self-guided local feature extraction using global and local attention, respectively.

2.4 Attention mechanism in vehicle Re-ID

Attention mechanism [24-25] is widely implemented on various fields of deep learning and it has been first employed in [26] in vehicle re-identification field. Teng et al. [26] proposed a spatial and channel attention network to mine the discriminative features in vehicle Re-ID task. Channel attention mechanism as a kind of soft attention, its final function is based on the degree of information contained in each location of the feature map to produce a weight value, and then multiply the feature map and the weight value, so that the area containing differential information to obtain a higher weight. To this end we introduce channel attention mechanism that can aggregate semantic similarity channels and attain more discriminative feature representation for vehicle Re-ID.

3 Our Algorithm

The algorithm model framework of this paper is shown in Figure 2. Firstly, the proposed multi-branches network is used to extract vehicle features of training set. Then the similarity between Query and Gallery vehicle features is calculated. Finally, the similarity scores are sorted to obtain the retrieval results of all the images of Query in the Gallery. Based on the ResNet-50 network, the model in this paper introduces global and local feature fusion, channel attention mechanism and weighted local features to enhance feature extraction and expression capabilities and effectively improve the recognition rate.

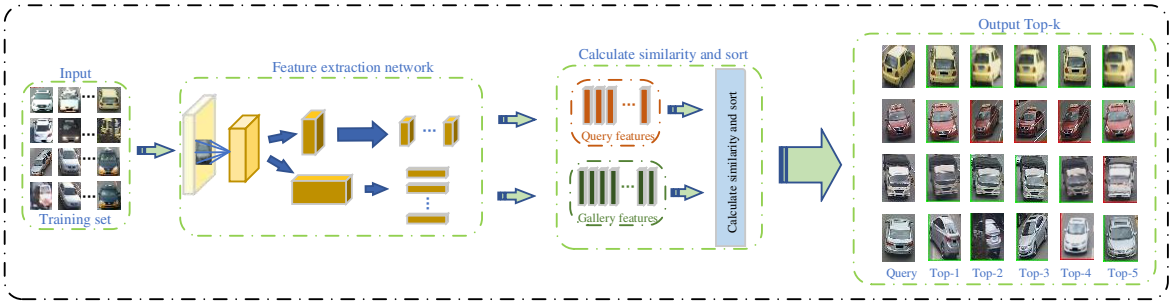

Fig.2 The overall framework of the algorithm model.

3.1 Multi-branches network architecture

The architecture of multi-branches network is shown in Figure 3. The backbone of our network is ResNet-50 which helps to acquire competitive performances in some Re-ID systems. The most obvious modification different from the original version is that we divide the subsequent part after res_conv5 block into three independent branches, sharing the similar architecture with the original ResNet-50.

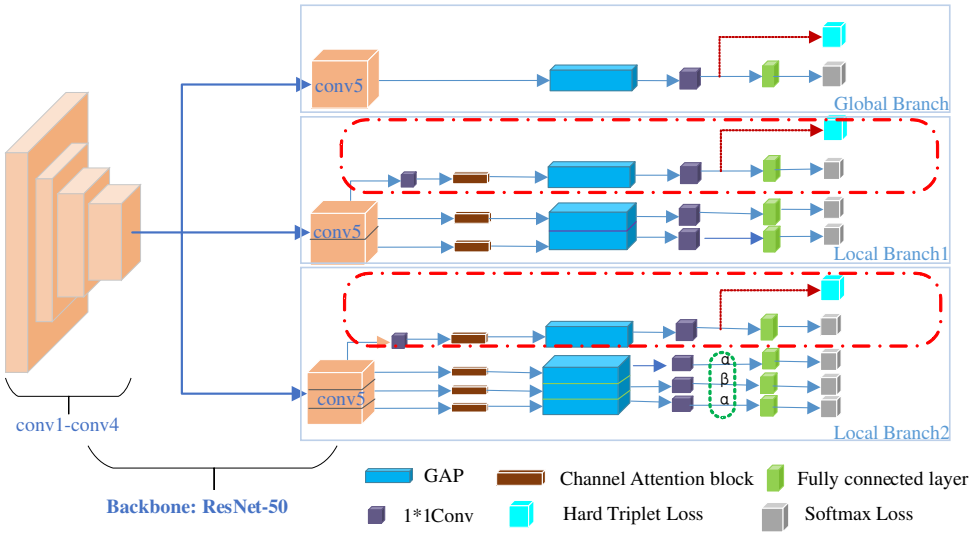

Fig.3 Multi-branches network architecture. The ResNet-50 backbone is split into three branches: Global Branch, Local Branch1, and Local Branch2. GAP and 1*1 Conv refer to Global Average Pooling and 1*1

convolutional layer, respectively. During the training, each branch does not share the weight and trains separately. But when testing, all branch in-formation will be assembled into a comprehensive feature to increase network performance.

The first is a Global Branch, which learns the global feature representations without any partition information. The second and third are Local Branch1 and Local Branch2, respectively. They both share the similar network architecture, the difference is that the Local Branch1 divides the height of the feature map into two pieces, while the Local branch2 divides the height of the feature map into three parts. In particular, Local Branch1 and Local Branch2 all contain a global branch which aims to solve the problem of low robustness of learning local features by focusing on specific semantic regions.

In Local Branch1 and Local Branch2, we use the channel attention mechanism on local features to give higher weight for important feature information. Global average pooling (GAP) [27] is used to average each feature map and output a value. GAP replaces the fully connected layer and greatly reduces the number of parameters. It is worth mentioning that we also used a 1*1 convolution before the GAP block of the global branches of Local Branch1 and Local Branch2. This can not only reduce the number of channels, but also decrease the number of calculations later. After the GAP block, 1*1 convolution block is used to increase the dimension, which can extract high dimensional features, and enhance the effect of feature extraction.

During the training, each branch does not share the weight and trains separately. But when testing, all branch information will be assembled into a comprehensive feature to increase network performance.

3.2 Weighted Local Feature and Channel Attention Mechanism

In the process of analyzing and processing the vehicle dataset, the vertical segmentation of the vehicle image will destroy the integrity of some key features of the vehicle and affect the re-identification effect. The horizontal segmentation will not only retain the characteristics of the most part of the vehicle, but also conducive to the subsequent weighting of local features. Local Branch2 performs horizontal three-division processing on the input vehicle images, and then performs weighting processing on the three pieces of parts when calculating the loss, as shown in Figure 4.

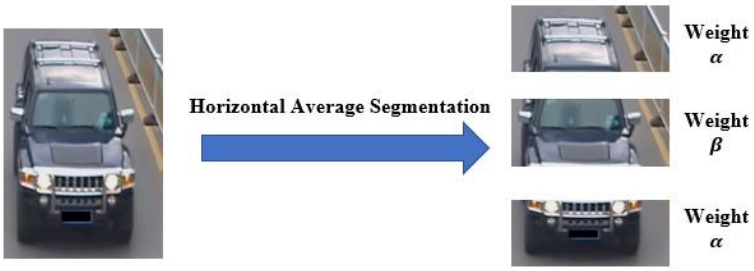

Fig.4 Segmentation and weighted vehicle image

The vehicle in the input image is usually located in the middle of the image, the upper and lower parts of the image usually contain a lot of background information. Therefore, we assign the weight $\boldsymbol{\alpha}$ to the upper and lower parts of the image, and the weight of the middle part to $\boldsymbol{\beta}$ ($\boldsymbol{\alpha} < \boldsymbol{\beta}$).

In addition to weighted local feature, we also design an attention module. The purpose is to enable the network to better find the detailed features of the vehicle, such as windshield stickers, vehicle scratches. Figure 5 shows our channel attention module.

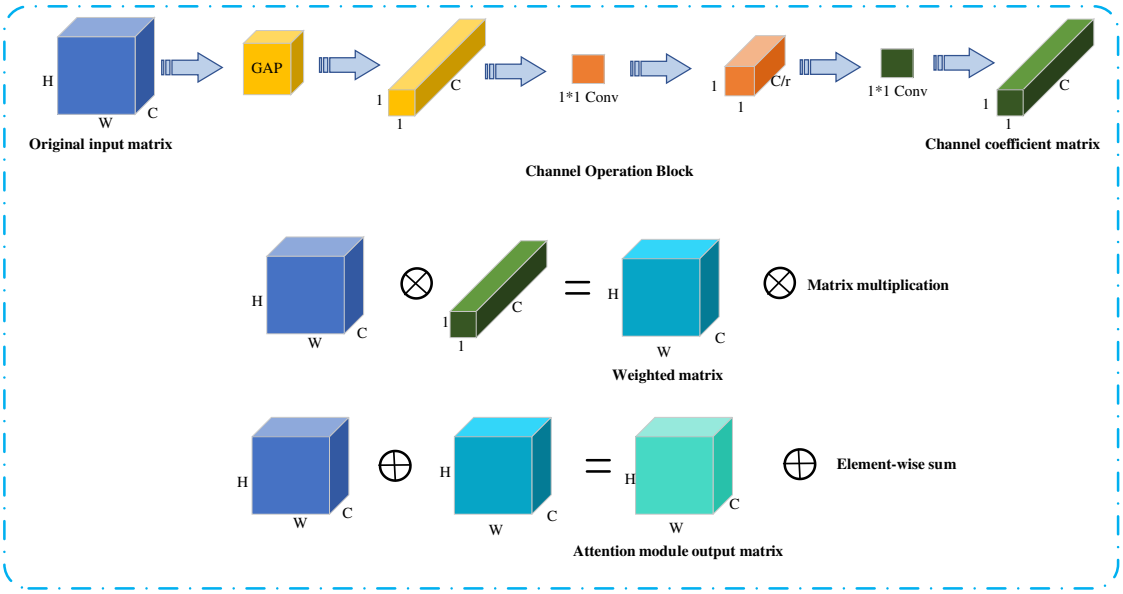

Fig.5 Channel Attention Module (CAM). H, W, C represent the height, width, and channel number of the feature map respectively. r is the scaling factor.

As shown in Figure 5, we divide the channel attention mechanism into three stages: channel operation stage, channel weighting stage, and channel superposition stage. During the channel operation stage, the global average pooling is carried out on the original input matrix, so that the original input matrix with the dimension of H*W*C is changed into a channel descriptor of 1*1*C, which can reduce the computational cost and accelerate the network training speed. Then two 1*1 convolution modules are used to first reduce the dimension of channel descriptor and then increase the dimension. There is a dimensionality reduction factor r between the two 1*1 convolution modules, and the dimension change is controlled by r. Finally, through the rise and fall of dimensions, the characteristic information of different channels is fused and the correlation between channels is captured to obtain a 1*1*C channel weight matrix. The channel weighting matrix is composed of C weight values, each weight value corresponds to a channel, and the weight value is the importance degree of the features represented by the corresponding channel. If the feature in the channel contains rich discriminant information, the corresponding weight value of the channel is large; on the contrary, the weight value of the channel is small. Then the original input matrix is multiplied by the channel weight matrix to get the weighted matrix, this process is called channel weighting stage. Finally, the output matrix of the attention module is obtained by adding the weighted matrix to the original input matrix in the channel superposition stage.

3.3 Loss Function

Softmax Cross-entropy loss. Softmax Cross-entropy loss is the most commonly used loss function in image classification tasks. In our experiment, the loss can be described as:

$$L_{Softmax} = -\sum_{i=1}^{N_i} \log\left(\frac{\exp(x_y)}{\sum_{j=1}^{N_{id}} \exp(x_j)}\right) \quad (1)$$

where N_i and N_{id} denotes the number of vehicle images per batch and vehicle identities in the whole training dataset, respectively. x_j and y respectively represent the output of fully connected layer for jth identity and the ground truth identity of input vehicle image.

Triplet loss. During the training process, each batch of processed vehicle pictures contains Q identities and each identity has K images. Among them, there are $K-1$ images (*Positives*) that are the same as each vehicle identity (*Anchor*), and the remaining $(Q-1)*K$ vehicles have different image identities (*Negatives*). The purpose of triplet loss is to pull the positive pair (A, P) together while pushing the negative pair (A, N) away by a margin, as shown in Fig. 6.

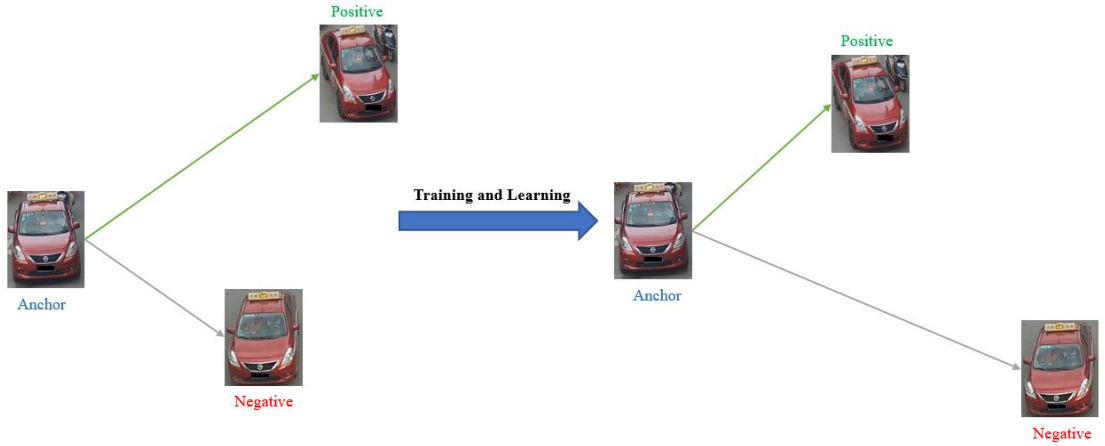

Fig.6 Triplet loss

Hard mining triplet loss, introduced in [28] as a variant of traditional triplet loss, aims at paying more attention to the closest negative and the farthest positive pairs. The hard mining triplet loss can be defined as:

$$L_{hard\ mining\ triplet} = \sum_{i=1}^{Q}\sum_{A=1}^{K}[\overbrace{\max_{P=1,\dots,K}\|A_i - P_i\|_2}^{hardest\ positive} - \overbrace{\min_{\substack{N=1,\dots,K \\ j=1,\dots,Q \\ i\neq j}}\|A_i - N_j\|_2}^{hardest\ negative} + \delta]_+ \quad (2)$$

where A_i, P_i and N_i respectively represent the feature vectors of Anchor, Positive and Negative samples, and δ is a minimal margin (Euclidean distance) and controls the difference between positive and negative pair distances.

Total loss. The total loss combining Softmax cross entropy losses with hard mining triplet losses is used to our training experiment. The total loss can be described as:

$$L_{total} = L_{Softmax} + \lambda * L_{hard\ mining\ triplet} \quad (3)$$

where parameter λ balances the contribution of Softmax cross entropy losses and hard mining triplet losses.

4 Experiment

To evaluate our model, we conducted experiments on three large-scale vehicle Re-ID datasets: VeRi-776, VRIC, and VehicleID. Firstly, we report a set of ablation studies (mainly on VeRi-776) to validate the

effectiveness of each component. Secondly, we compare the performance of our model against existing state-of-the-art methods on all three datasets. Finally, we provide more visualizations and analysis to illustrate how our model has achieved its effectiveness.

4.1 Datasets

VeRi-776 [8],[29],[30], a large-scale urban surveillance vehicle dataset for Re-ID, is collected by 20 non-overlapping traffic surveillance cameras in unconstrained traffic scenarios, which contains 51,035 images of 776 vehicles with identity annotations, camera geo-locations, image timestamps, vehicle types and color information. In addition, each vehicle is captured by 2 to 18 cameras. Following the evaluation protocol of [24], VeRi-776 is separated into a training subset and a testing subset (the testing subset includes gallery and query subset). In this paper, the number of training, gallery, and query sets are collected as 37,778 images of 576 vehicles, 11,579 images of 200 vehicles and 1,678 images of 200 vehicles, respectively.

VRIC [9] is a newer dataset, which consists of 60,430 images of 5,656 vehicle IDs collected from 60 different cameras in traffic scenes. VRIC differs significantly from existing datasets in that unconstrained vehicle appearances were captured with variations in imaging resolution, motion blur, weather condition, and occlusion. The training set has 54,808 images of 2,811 vehicles, while the rest 5,622 images of 2,811 identities are employed to testing.

VehicleID [10] is a larger-scale vehicle Re-ID dataset from multiple non-overlapping surveillance cameras in a small city in China. It contains 221,763 images of 26,267 vehicles in total (8.44 images/vehicle in average), and the vehicle in each picture is either captured from the front or the back. Following the evaluation protocol of [10], VehicleID is separated into training set with 110,178 images of 13,134 vehicles and testing set with 111,585 images of 13,133 vehicles. We further extract three subsets (i.e. small, medium and large) ordered by their size from the original testing dataset for our vehicle re-identification evaluation tasks. Specifically, the small subset includes 800 gallery images and 6,532 query images of 800 vehicles, the medium subset contains 1,600 gallery images and 11,395 query images of 1,600 vehicles, the large subset consists of 2,400 gallery images and 17,638 query images of 2,400 vehicles.

4.2 Implementation Details and Evaluation Metric

In our experiments, the software tools are *PyTorch*, *CUDA11.1*, and *CUDNN V8.0.4.30*. The hardware device is a workstation equipped with *AMD Ryzen 5 3600X CPU 32G*, *NVIDIA GeForce RTX 3080* and 256GB+2TB memory. During training, the input images are re-sized to 384*128 and then augmented by random horizontal flip, normalization, and random erasing. We set the training batch size to 32, the initial learning rate is $3*10^{-4}$, and the learning rate decreased to 0.1 times at 20 and 40 epoch. At the same time, we chose the AMSGrad optimizer to train the network. The testing images are re-sized to 384*128 and augmented only by normalization. The weight of Local Branch 2 is 0.3 for $\boldsymbol{\alpha}$ and 0.4 for $\boldsymbol{\beta}$. After many experiments and tests, the attenuation factor $\boldsymbol{r}$ of the channel attention module is set to 4. The margin δ in triplet loss is set to 1.2 in all experiments and the parameter λ in total loss is set to 0.1.

Following the evaluation protocol of re-identification work [30-31], we utilize the mean average precision (mAP) and Rank-n (the expected correct matching pair in the top n matches) as the evaluation metrics.

In a typical re-identification evaluation setting, we have a query set and a gallery set. For each vehicle in the query set, the goal is to retrieve similar identities from the test set (that is, the gallery set). $AP(q)$ for a query image q is defined as

$$AP(q) = \frac{\sum_{k=1}^{n} P(k) \times gt(k)}{N} \tag{4}$$

where k, n, and N represent the image sequence number of the gallery collection, the total number of images in the gallery collection, and the total number of the target vehicle images, respectively. $P(k)$ represents precision at rank k. $gt(k)$ represents whether the k-th image is the target vehicle and it's equal to 1 when the matching of query image q to a test image is correct at $rank \leq k$. mAP is then computed as average over all query images

$$mAP = \frac{\sum_{q=1}^{Q} AP(q)}{Q} \tag{5}$$

where Q is the total number of query images.

4.3 Ablation Experiment

In order to verify the effectiveness of multi-branches network architecture, we conducted ablation experiments on the VeRi-776 dataset. We choose ResNet-50 with the global feature branch as the baseline. Seven variants are then constructed on the top of the baseline (***Best view in color.***):

a) baseline + Local Branch1(no red dotted area) + CAM;

b) baseline + Local Branch2(no red dotted area) + CAM;

c) Local Branch1(no red dotted area) + Local Branch2(no red dotted area and green dotted area);

d) Local Branch1(no red dotted area) + Local Branch2(no red dotted area) + Green dotted area;

e) Local Branch1(no red dotted area) + Local Branch2(no red dotted area) + Green dotted area +CAM;

f) baseline + Local Branch1(no red dotted area) + Local Branch2(no red dotted area) + Green dotted area + CAM;

g) baseline + Local Branch1 + Local Branch2 + Green dotted area + CAM.

The detailed results of the ablation studies in VeRi-776 dataset are illustrated in Table 2.

Table 2: The mAP and Rank-1 in Ablation Experiment (in %)

Methods	**VeRi-776**	
	mAP	**Rank-1**
baseline	69.18	93.21
a) baseline + Local Branch1(no red dotted area) + CAM	72.13	94.04
b) baseline + Local Branch2(no red dotted area) + CAM	73.05	94.16
c) Local Branch1(no red dotted area) + Local Branch2(no red dotted area and green dotted area)	63.90	90.82
d) Local Branch1(no red dotted area) + Local Branch2(no red dotted area) + Green dotted area	69.57	91.66
e) Local Branch1(no red dotted area) + Local Branch2(no red dotted area) + Green dotted area +CAM	71.17	93.86
f) baseline + Local Branch1(no red dotted area) + Local Branch2(no red dotted area) + Green dotted area + CAM	76.32	95.83
g) ***baseline + Local Branch1 + Local Branch2 + Green dotted area + CAM (ours)***	**77.12**	**96.30**

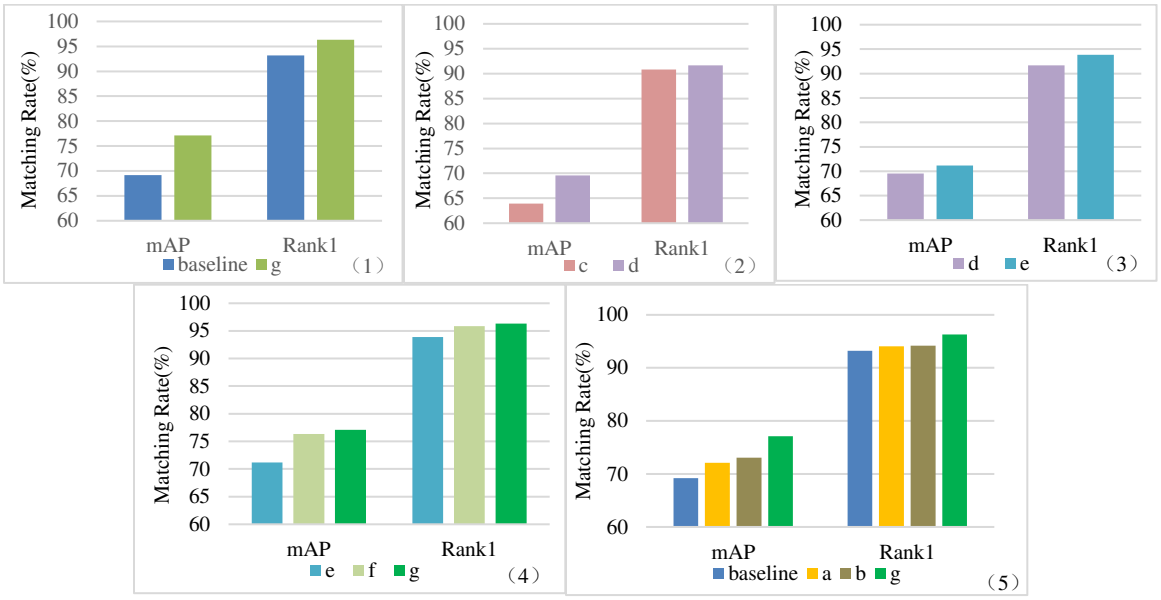

Fig.7 Ablation Experiment of the proposed framework on VeRi-776 dataset (in %).

It can be observed from Table 2 and Fig.7(1) that compared with the baseline network (only the branch of global features), our improved network has increased by 7.94% and 3.09% on mAP and Rank-1 respectively. It proves that our network has strong robustness.

As shown in Fig.7(5), comparing the baseline network, network ***a***, network ***b***, with our improved network, we can draw two conclusions: first, adding local features can greatly advance the recognition accuracy; second, Local Branch1 is to divide the image into two pieces (Horizontal segmentation), the purpose is to increase the proportion of the vehicle area in the respective images, thereby improving the accuracy of vehicle feature recognition; the method of Branch1 and Branch2 is similar, the difference is that the image is divided into three pieces and further reduces the proportion of the sky area, so that the network pays more attention to the feature extraction of the vehicle area. Dividing the image into two and three at the same time can increase the learning effect of the network on vehicle features.

Compared with network ***c***, network ***d*** performs weighting processing on local features, and mAP improves by 5.67%, also proving the effectiveness of weighting processing, as shown in Fig.7(2). In Fig.7(3), compared with network ***d***, mAP and Rank-1 of network ***e*** are improved by 1.60% and 2.20% respectively after adding channel attention block. Fig.7(4) shows that by compared with the experimental results of networks ***e***,***f***, ***g***, the importance of global features can be proved.

4.4 Performance Comparison with State-of-the-Art Methods

We compare our proposed multi-branches network framwork with multiple state-of-the-art vehicle re-identification approaches on three mainstream datasets, i.e., VeRi-776, VRIC, and VehicleID with corresponding evaluation metrics (mAP and Rank-n).

A. Results on VeRi-776 Dataset

For VeRi-776 dataset, the test is conducted following the standard evaluation of [30]. Table 3 presents the result comparison between previous state-of-the-art and our model on VeRi-776 dataset. Our proposed method

achieves 96.30% on Rank-1 accuracy, 98.11% on Rank-5 accuracy and 77.12% on mAP without re-ranking. These results surpass previous state-of-the-art on almost the three metrics, especially on mAP. In this paper, our method only relies on the supervised information of ID, while VGG+C+T [32], GS-TRE [33], VAMI+ST [34], and AGNet-ASL+STR [15] exploit spatial-temporal information, and other methods also utilize extra annotations such as attributes (RAM [6]), but our accuracy still exceeds them. In a word, re-ranking can further enhance the performance of our model. A good mAP score demonstrates that our model has a stronger potential to retrieve all the corresponding images of the same identity in the gallery set, regardless of different camera properties and viewpoint changes.

Table 3: The mAP, Rank-1 and Rank-5 on VeRi-776 dataset (in %)

Methods	mAP	Rank-1	Rank-5	References
VRSDNet [5]	53.45	83.49	92.55	Multimed Tools Appl 2019
VGG+C+T [32]	58.78	86.41	92.91	ICME 2017
GS-TRE [33]	59.47	96.24	**98.97**	IEEE TMM 2018
AAVER [13]	61.18	88.97	94.70	ICCV 2019
VAMI+ST [34]	61.32	85.92	91.84	CVPR 2018
RAM [6]	61.50	88.60	94.00	ICME 2018
QD-DLF [35]	61.83	88.50	94.46	IEEE TITS 2019
MSA [14]	62.89	92.07	96.19	Neural Computing and Applications 2020
AGNet-ASL+STR [15]	71.59	95.61	96.56	arXiv 2020
Ours	**77.12**	**96.30**	98.11	**Proposed**

B. Results on VRIC Dataset

VRIC is a relatively newly released dataset, hence, the results of previous work reports are few. For VRIC dataset, the test is conducted following the standard evaluation of [9], we compared the results of our proposed method with other models and frameworks on the VRIC dataset. As shown in Table 4, by comparison, we can find that our model outperforms the latest methods by 1.97% in Rank-1 and 0.89% in Rank-5, respectively, and significantly improves the recognition effect of vehicle re-identification on both Rank-1 and Rank-5 accuracy.

Table 4: The mAP, Rank-1 and Rank-5 on VRIC dataset (in %)

Methods	mAP	Rank-1	Rank-5	References
MSVF [9]	47.50	46.61	65.58	arXiv 2018
GLAMOR [23]	76.48	78.58	93.63	arXiv 2020
PGAN [22]	**84.80**	78.00	93.20	arXiv 2020
Ours	82.75	**79.97**	**94.09**	**Proposed**

C. Results on VehicleID Dataset

For VehicleID dataset, all the tests are conducted following the standard evaluation of [8]. Our model outperforms the latest methods by 5.32%-7.39% in mAP, 0.07%-4.41% in Rank-1 and 1.53%-3.91% in Rank-5 on three test subsets. Generally speaking, the larger testing sets (1,600 and 2,400 test size) introduce more challenging and complex scenarios in real life, therefore, the methods perform better on the small size (800) testing set. It can be found from Table 5, our model outperforms all other methods in all testing sets (800, 1,600, and 2,400 test size), and improves about 4.0% in mAP, Rank-1, and Rank-5 on all three testing sets, compared with the second best methods achieved by MSA [14] and AAVER [13], respectively. The results demonstrate the robustness and superiority of our proposed method.

Table 5: The mAP, Rank-1, and Rank-5 on VehiceID dataset (in %)

Methods	Test800			Test1,600			Test2,400			References
	mAP	Rank1	Rank5	mAP	Rank1	Rank5	mAP	Rank1	Rank5	
VRSDNet [5]	63.52	56.98	86.90	57.07	50.57	80.05	49.68	42.92	73.44	Multimed Tools Appl 2019
VAMI [34]	N/A	63.12	83.25	N/A	52.87	75.12	N/A	47.34	70.29	CVPR 2018
VGG+C+T+S [32]	N/A	69.90	87.30	N/A	66.20	82.30	N/A	63.20	79.40	ICME 2017
AGNet-ASL [15]	74.05	71.15	83.78	72.08	69.23	81.41	69.66	65.74	78.28	arXiv 2020
GS-TRE [33]	75.40	75.90	84.20	74.30	74.80	83.60	72.40	74.00	82.70	IEEE TMM 2018
QD-DLF [35]	76.54	72.32	92.48	74.63	70.66	88.90	68.41	64.14	83.37	IEEE TITS 2019
AAVER [13]	N/A	74.69	93.82	N/A	68.62	89.95	N/A	63.54	85.64	ICCV 2019
RAM [6]	N/A	75.20	91.50	N/A	72.30	87.00	N/A	67.70	84.50	ICME 2018
MSA [14]	80.31	77.55	90.50	77.11	74.41	86.26	75.55	72.91	84.35	Neural Computing and Applications 2020
Ours	**87.70**	**81.96**	**95.35**	**84.26**	**77.85**	**92.44**	**80.87**	**74.07**	**89.55**	**Proposed**

5 Conclusion and future work

In this work, we propose a multi-branches network for vehicle re-identification. First of all, a channel attention mechanism strategy integrates discriminative information with global and local features. At the same time, feature extraction is optimized through attention modules and weighted local feature, so that more discriminative detail features are extracted. Finally, we conduct ablation experiments and discuss in terms of vehicle Re-ID performance so as to investigate the architecture design of multi-branches. Extensive comparative evaluations have indicated that our method not only exceeds state-of-the-art results on three challenging vehicle Re-ID datasets: VeRi-776, VRIC, and VehicleID, but also pushes the performance to an exceptional level comparing to existing methods.

Although our method outperforms the state-of-the-art approaches without employing other auxiliary information of the vehicle, it still faces limitations because of the challenging scenarios in real-world surveillance. In addition to the similar appearance, there are many other useful features about the vehicle Re-ID, such as spatio-temporal relation and road network information. On the basis of our method, we will try to use spatio-temporal information as an auxiliary means to narrow the scope of vehicle retrieval and further improve the accuracy of vehicle re-identification in the future.

Author Contributions

Conceptualization, Y.X.; Methodology, L.R.; Software, L.L. and X.P.; Validation, X.Z.; Formal analysis, L.H.; Writing—original draft preparation, L.R.; Writing—review and editing, Y.X. All authors have read and agreed to the published version of the manuscript.

Acknowledgments

This work was supported by the Natural Science Foundation of China (11547037, 11604181), Shandong Province Postgraduate Education Quality Curriculum Project (SDYKC19083), Shandong Province Postgraduate Education Joint Training Base Project (SDYJD18027), Hisense Group research and development Center Project, and the Scholarship Fund of SDUST.

Reference

1 Xiong M, Chen D, Lu X. Mobile person re-identification with a lightweight trident CNN. Sci China Inf Sci, 2020, 63: 1-3

2 Hu B, Xu J, Wang X. Learning generalizable deep feature using triplet-batch-center loss for person re-identification. Sci China Inf Sci, 2021, 64: 1-2

3 Zhang S, Wei C. Deep learning network for uav person re-identification based on residual block. Sci China Inf Sci, 2020, 63: 1-3

4 Ye M, Shen J, Lin G, et al. Deep learning for person re-identification: a survey and outlook. IEEE Trans Pattern Anal Mach Intell, 2021

5 Zhu J, Du Y, Hu Y, et al. Vrsdnet: vehicle re-identification with a shortly and densely connected convolutional neural network. Multimed Tools Appl, 2019, 78: 29043-29057

6 Liu X, Zhang S, Huang Q, et al. Ram: a region-aware deep model for vehicle re-identification. In: Proceedings of IEEE International Conference on Multimedia and Expo, Seattle, 2018: 1-6

7 Cheng Y, Zhang C, Gu K, et al. Multi-scale deep feature fusion for vehicle re-identification. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, Virtual Barcelona, 2020: 1928-1932

8 Liu X, Liu W, Mei T, et al. A deep learning-based approach to progressive vehicle re-identification for urban surveillance. In: Proceedings of European conference on computer vision, Amsterdam, 2016: 869-884

9 Kanacı A, Zhu X, Gong S. Vehicle re-identification in context. In: Proceedings of German Conference on Pattern Recognition, Stuttgart, 2018: 377-390

10 Liu H, Tian Y, Yang Y, et al. Deep relative distance learning: tell the difference between similar vehicles. In: Proceedings of IEEE conference on computer vision and pattern recognition, Las Vegas, 2016: 2167-2175

11 Sun Y, Zheng L, Yang Y, et al. Beyond part models: person retrieval with refined part pooling (and a strong convolutional baseline). In: Proceedings of the European conference on computer vision, Munich, 2018: 480-496

12 Sun Y, Xu Q, Li Y, et al. Perceive where to focus: learning visibility-aware part-level features for partial person re-identification. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, 2019: 393-402

13 Khorramshahi P, Kumar A, Peri N, et al. A dual-path model with adaptive attention for vehicle re-identification. In: Proceedings of IEEE/CVF International Conference on Computer Vision, Seoul, 2019: 6132-6141

14 Zheng A, Lin X, Dong J, et al. Multi-scale attention vehicle re-identification. Neural Comput Appl, 2020: 1-15

15 Wang H, Peng J, Chen D, et al. Attribute-guided feature learning network for vehicle reidentification.

IEEE Multimedia, 2020, 27: 112-121

16 He B, Li J, Zhao Y, et al. Part-regularized near-duplicate vehicle re-identification. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, 2019: 3997-4005

17 Huang C, Loy C C, Tang X. Local similarity-aware deep feature embedding. arXiv preprint arXiv:1610.08904, 2016

18 Cui W, Wang F, He X, et al. Multi-scale semantic segmentation and spatial relationship recognition of remote sensing images based on an attention model. Remote Sens, 2019, 11: 1044

19 Li Y, Wang S, Dong S, et al. Person reidentification model based on multiattention modules and multiscale residuals. Complexity, 2021, 2021

20 Chen T, Ding S, Xie J, et al. Abd-net: attentive but diverse person re-identification. In: Proceedings of IEEE/CVF International Conference on Computer Vision, Seoul, 2019: 8351-8361

21 Wang G, Yuan Y, Chen X, et al. Learning discriminative features with multiple granularities for person re-identification. In: Proceedings of 26th ACM international conference on Multimedia, Seoul, 2018: 274-282

22 Zhang X, Zhang R, Cao J, et al. Part-guided attention learning for vehicle re-identification. arXiv preprint arXiv:1909.06023, 2019

23 Suprem A, Pu C. Looking glamorous: vehicle re-id in heterogeneous cameras networks with global and local attention. arXiv preprint arXiv:2002.02256, 2020

24 Chen X, Zheng L, Zhao C, et al. Rrgccan: re-ranking via graph convolution channel attention network for person re-identification. IEEE Access, 2020, 8: 131352-131360

25 Li W, Zhu X, Gong S. Harmonious attention network for person re-identification. In: Proceedings of IEEE conference on computer vision and pattern recognition, Salt Lake, 2018: 2285-2294

26 Teng S, Liu X, Zhang S, et al. Scan: spatial and channel attention network for vehicle re-identification. In: Proceedings of Pacific Rim conference on multimedia, Hefei, 2018: 350-361

27 Almazan J, Gajic B, Murray N, et al. Re-id done right: towards good practices for person re-identification. arXiv preprint arXiv:1801.05339, 2018

28 Hermans A, Beyer L, Leibe B. In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737, 2017

29 Liu X, Liu W, Mei T, et al. Provid: progressive and multimodal vehicle reidentification for large-scale urban surveillance. IEEE Trans Multimedia, 2017, 20: 645-658

30 Liu X, Liu W, Mei T, et al. A deep learning-based approach to progressive vehicle re-identification for urban surveillance. In: Proceedings of European conference on computer vision, Amsterdam, 2016: 869-884

31 Shen Y, Xiao T, Li H, et al. Learning deep neural networks for vehicle re-id with visual-spatio-temporal path proposals. In: Proceedings of IEEE International Conference on Computer Vision, Shenzhen, 2017: 1900-1909

32 Zhang Y, Liu D, Zha Z J. Improving triplet-wise training of convolutional neural network for vehicle re-identification. In: Proceedings of IEEE International Conference on Multimedia and Expo, Hongkong, 2017: 1386-1391

33 Bai Y, Lou Y, Gao F, et al. Group-sensitive triplet embedding for vehicle reidentification. IEEE Trans Multimedia, 2018, 20: 2385-2399

34 Zhou Y, Shao L. Aware attentive multi-view inference for vehicle re-identification. In: Proceedings of IEEE conference on computer vision and pattern recognition, Salt Lake, 2018: 6489-6498

35 Zhu J, Zeng H, Huang J, et al. Vehicle re-identification using quadruple directional deep learning features. IEEE Trans Intell Transp Syst, 2020, 21: 410-420