



Joint metric learning of local and global features for vehicle re-identification

Junge Shen¹ · Jian Sun¹ · Xin Wang^{1,2} · Zhaoyong Mao¹

Received: 6 October 2021 / Accepted: 3 February 2022
© The Author(s) 2022, corrected publication 2022

Abstract

Vehicle re-identification (ReID) means to identify the target vehicle in large-scale surveillance videos captured by multiple cameras, where robust and distinctive visual features of vehicles are critical to the performance. Recently, the researchers have approached the problem with attention based models. However, most of these models use strongly-supervised methods, which rely on expensive extra labels, e.g., keypoints(vehicle wheels, logo and lamps) and attributes(e.g., color and type). Therefore, we propose a joint metric learning approach to solve the problem. We present an end-to-end Partition and Fusion Multi-branch Network (PFMN), a novel approach to effectively learn discriminative features without any annotations or additional attributes. For hard samples, which means different vehicles with similar appearance or the same vehicle with different appearances, a novel variant of hard sampling triplet loss is proposed. Based on extensive experiments, we have proved the effectiveness of our proposed method. On the challenging public data sets VeRi-776 and VehicleID, our model outperforms most state-of-the-art algorithms on mAP and rank-1. Especially on mINP, which measures the cost of model retrieval hard samples, we can achieve a significant improvement.

Keywords Vehicle re-identification · Feature partition · Global and local feature fusion · Metric learning

Introduction

Object re-identification (ReID), the task to match images of the same object in a gallery, has made notable progress and achieved high performance in recent years. Vehicle ReID is an important application in the field of object re-identification, which aims to identify a particular vehicle across different camera views. With the continuously increasing quantity of urban vehicles, vehicle ReID has diverse applications in real-world video surveillance and intelligent transportation [1–3]. One challenging task of vehicle ReID is cross-camera retrieval, which means to retrieve the same vehicle in the gallery images captured across different non-overlapping city security cameras. Because of low resolution, occlusions and viewpoint changes, it is difficult for the cam-

era to get license plate information clearly. At this time, the use of vehicle ReID technology can achieve both rapid and accurate retrieval or track of the target vehicle within a certain area.

The retrieval of hard samples is the most significant impact on the accuracy of vehicle ReID. Vehicles are mass-produced objects, which brings high similarity to the same vehicle. Moreover, the appearance variance of a vehicle is significant across different viewpoints or different illuminating levels. Therefore, models need to pay more attention on discriminative local information to deal with the problem of large intra-class variance and small inter-class variance.

Manual features are used by traditional vehicle ReID [4,5]. Unfortunately, they have the problem of poor generalization. With the emergence of deep learning [6–9], especially the breakthroughs of the convolutional neural network (CNN), global deep representation are extracted for vehicle ReID automatically with strong generalization. However, several challenges exist when extracting features with deep learning methods, as shown in Fig. 1: (a) represents the occlusion problem. Global deep features can easily add irrelevant semantic information to the learned features, which affects the accuracy of retrieval; (b) and (c) indicate that different

✉ Junge Shen
shenjunge@nwpu.edu.cn

¹ Unmanned System Research Institute, Northwestern Polytechnical University, Xi'an 710000 Shaanxi, China

² Shaanxi Transportation Holding Group CO.,LTD., Northwestern Polytechnical University, Xi'an 710000 Shaanxi, China

views from the same ID vehicle have a huge change in the overall appearance, which may be classified as different vehicles. (d) and (e) refer the difficulty to distinguish vehicles with tiny differences.

To address above issues, researchers have been studying vehicle ReID by extracting local features based on attention mechanisms. Zhang et al. [10] have learnt vehicle feature representations by perceiving attention from multi-perspectives of the vehicle. Many vehicle ReID models use additional annotations to obtain more robust features, and image segmentation [13,14] is one effective approach. For example, Meng et al. [11] have introduced a parsing network to parse a vehicle into four different views. In this paper, the network is trained using a data set labeled with parsing information. Wang et al. [12] have proposed a posture-invariant vehicle ReID model by detecting 20 key points of the vehicle to extract local features from different perspectives.

Although existing methods achieve effective retrieval of vehicles with similar appearances, they rely on expensive key point labels, component annotations, and attributes, such as the manufacturer, model, and color, which increases the complication and the training time of the model. Various additional supervision information brings a huge workload, which is unfavorable to the deployment of the model. As a result, two local branches which uniformly partitioning and re-fusion features in the horizontal and vertical directions, respectively, are introduced. These two branches can obtain fine-grained local features of key regions without additional auxiliary information.

In addition, traditional triplet loss function [24] or the hard sampling triple loss function [15] is widely applied in almost every vehicle ReID model. The triplet loss function can alleviate the problems of high inter-class similarity and large intra-class differences. Since the vehicle ReID data set has an obvious long tail effect, most of the vehicles with the same ID only have a few samples. By constructing triples, it is possible to generate a combination of triples far more than the number of images, which can effectively ease the over-fitting problem. However, the traditional triple loss still has some shortcomings. For example, it does not take the absolute distance of positive sample pairs between different triples into consideration. To this end, a constraint on the absolute distance of positive sample pairs is introduced to the metric learning loss to distinguish hard samples. The main contributions of this paper are summarized as follows:

- We have introduced a novel method for uniformly partitioning feature maps in horizontal and vertical directions to further utilize the local information of the image. With no additional supervision information, the model focus on the local regions with distinctive features of the vehicle.

- A new end-to-end partition and fusion multi-branch network (PFMN), which combines global features and local features, has been proposed to handle similar samples. Based on global features, two local branches are introduced, and the local features of different directions are fused to obtain higher performance.
- We have proposed a new variant of the hard sampling triplet loss which employs the absolute distance of positive samples in a batch to pull the positive samples closer in the learned feature space. Extensive comparative experiments demonstrate that our proposed method outperforms many state-of-the-art approaches. The significant improvement of mINP shows our model has greatly improved the retrieval ability of hard samples.

The rest of the paper is organized as follows. “Related Work” briefly reviews recent works about representation learning and metric learning in vehicle ReID. The detailed architecture of the proposed approach is discussed in “Methodology”. Through extensive experiments in “Experiments”, we validate our design choices and show the effectiveness of the proposed network structure and loss function on two challenging data sets. Finally, “Conclusion” concludes the paper.

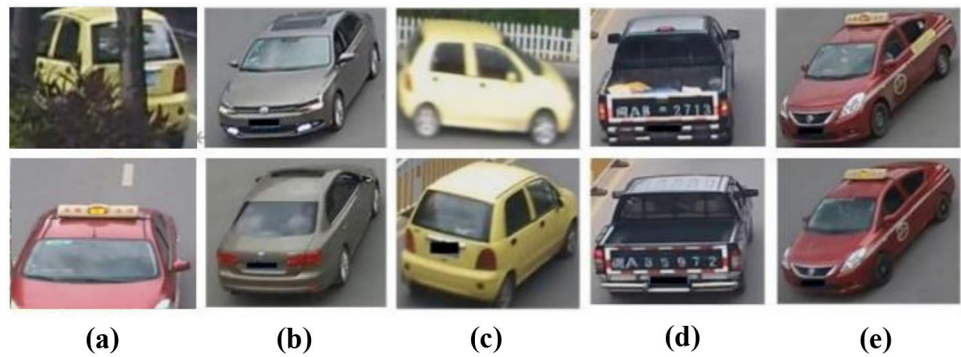
Related work

Since person ReID and vehicle ReID are similar tasks, methods also can be shared. First, we will give a brief review to some famous person ReID methods. Then, we will focus on reviewing the existing research of vehicle ReID using representation learning and metric learning.

Person ReID

The vigorous development of person ReID provides a wide range of ideas for vehicle ReID. Luo et al. [30] have proposed a strong baseline method called “Bag of Tricks”, which summarized and improved the tricks used in the previous person ReID model. The performance of the “Bag of Tricks” baseline has greatly surpassed the state-of-the-art person ReID methods at the time. Then the vehicle re-identification model regards Luo’s model as the baseline. Sun et al. [31] have used uniformly segmented feature maps to extract local features of different parts of the person, and have obtained better results than that of global features for retrieval. Wang et al. [32] have designed a Multiple Granularity Network, and have proposed a feature learning strategy that combines global and local information with different granularities, making full use of the coarse-to-fine mechanism in person ReID.

Fig. 1 Challenges in vehicle ReID: **a** occlusion of irrelevant objects, **b, c** different camera views, **d, e** vehicles with similar appearance



Representation learning based Vehicle ReID

With the development of deep learning and the emergence of large-scale vehicle re-identification data sets, such as VeRi-776 [4], VehicleID [5], CityFlow [18], CompCar [19], breakthrough has been made in vehicle ReID tasks.

Liu et al. [20] have proposed progressive vehicle ReID, which uses license plate and time-geographic information to reorder the results. They employ CNN to extract the appearance attributes to obtain coarse results and Siamese neural network to verify license plate to obtain fine search results. Zheng et al. [21] have proposed a novel end-to-end deep learning network architecture. In addition to the vehicle's identity information, the model uses three auxiliary attribute training models: camera view, vehicle color, and vehicle information to learn global feature with rich attribute features. [22] have proposed Generalized Pairwise Ranking and Multi-Grain based List Ranking to alleviate the precise vehicle search problem.

Apart from global feature representation learning, the network that uses annotation information to learn local features of vehicles has become a popular approach, which can enhance the ability to represent features. Zhang et al. [10] have proposed a partial attention network for vehicle re-identification. They have visualized the feature map to show important information in 16 component modules, such as lights, pendants, and logos in an interpretable manner. He et al. [23] have used areas such as windows, logos, and light bounding boxes to learn discriminative features, and have proposed a simple and effective method for regularized discriminative feature-preserving to enhance the model's ability to perceive subtle differences.

Metric learning based Vehicle ReID

A triplet loss function [24] can effectively reduce the distance between positive samples and expand the distance between negative ones in the feature space. Liu et al. [25] have origi-

nally introduced metric learning to the task of vehicle ReID, and have proposed coupled clusters loss, which solves the local optimization problem of traditional triple loss by optimizing the movement direction of positive samples. Guo et al. [26] have proposed a coarse-to-fine structured feature embedding method, which enables the network to gradually complete the tasks from rough to specific. The model gradually has a strong sensitivity to the position of the image in the feature space through the constraint of the loss function. Bai et al. [27] have designed a group-sensitive triplet embedding method, where K-means method is applied. In this way, variance for the intra-class is reduced and that for inter-class is increased. Hao et al. [47] have proposed a two-branch Partition and Reunion Network (PRN) to extract more local features.

In metric learning, the idea of multi-task learning is also widely used. To avoid the model's restriction on the distance between the selected triples too simple, a classification task is added to the network. Combining classification loss and triplet loss to optimize the network [28,29] has become a common method for subsequent vehicle ReID based on metric learning.

We propose a vehicle ReID model combining global and local features. In contrast to prior methods, we partition the global feature horizontally and vertically to obtain discriminative local features. The ability to retrieve hard samples is enhanced with our fine-grained fusion of features and improvements to the metric loss function.

Methodology

In this section, we first introduce our proposed network structure, which is an end-to-end Partition and Fusion Multi-branch Network (PFMN) utilized for global and local features extraction, and then we explain the feature extraction of the two local branches in detail. Finally, we focus on our improved triplet loss function.

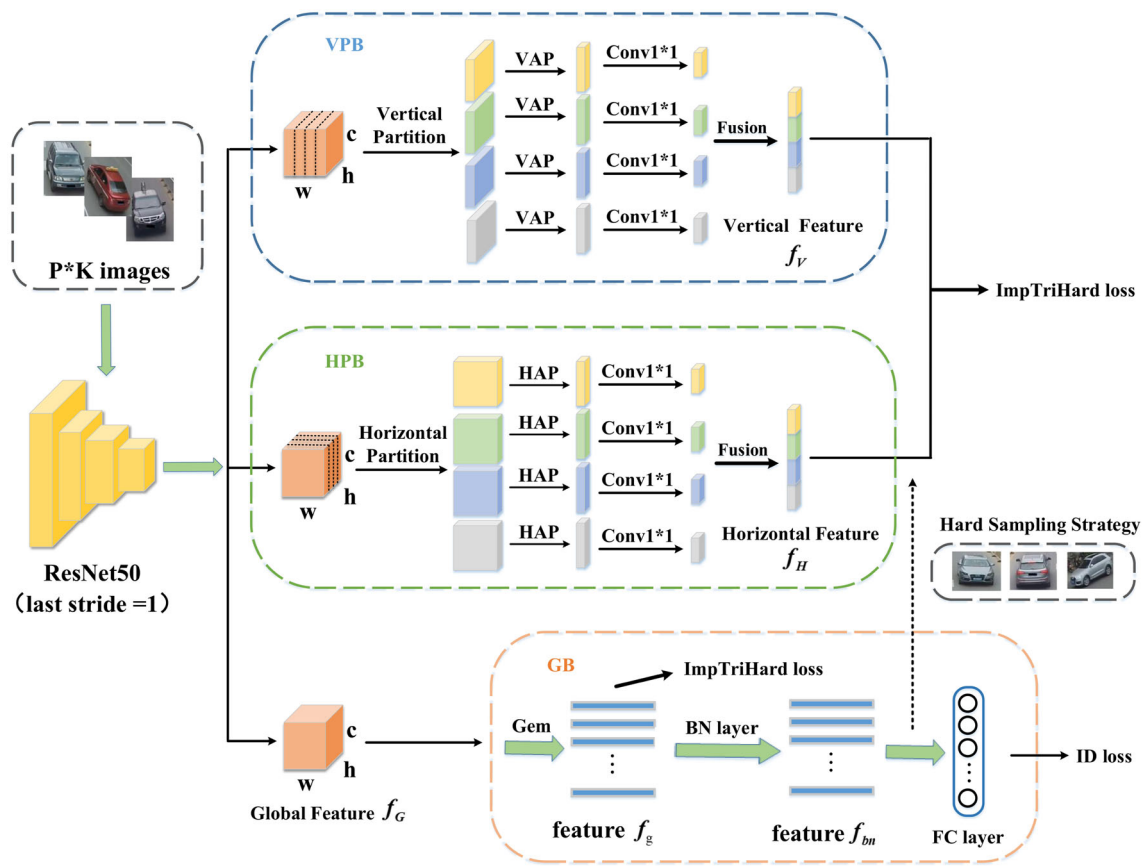


Fig. 2 Framework of joint local and global features for vehicle re-identification on metric learning

Table 1 Details of the branches of our PFMN model

Branch	Direction	Inputsize(h×w)	Outputsize(h×w)	Dim
GB	None	256×256	16×16	2048
VPB	Horizontal	256×256	1×4	256×4
HPB	Vertical	256×256	4×1	256×4

Overview of network structure

The framework of our proposed method is shown in Fig. 2. Following the strategy of selecting samples by the triplet loss function, we set the input of the model to $P * K$ images (P is the number of vehicle IDs, K is the number of pictures selected for each ID). When an image is fed into the model, the Global Features(f_G) is generated from the backbone network.

And then, a multi-branch network is designed, which consists of three modules: Global Branch (GB), Horizontal Partition Branch (HPB) and Vertical Partition Branch (VPB). These three branches use the f_G in different ways. The Global Branch use the original f_G without any partition. In the other two branches, first, we partition the global feature maps into four parts in the horizontal and vertical directions, respectively. After that, the horizontal average pooling and vertical

average pooling operations are applied on the feature map to extract discriminative local features. Then, in the feature fusion part, the eight local features extracted from the two local branches are grouped according to the directions and fused in the channel dimension. The re-fused local features are called Vertical Feature(f_V) and Horizontal Feature(f_H).

Finally, in the loss function part, the triplet loss function with the absolute distance constraint of positive sample pairs is employed for training. To make the model easier to converge, the global branch and the local branch use the same hard sample sampling strategy.

Deep feature extraction

In this section, we elaborate the details of our deep feature extraction in different branches.

Table 2 Structure of ResNet50 that we apply

Layer name	Output size	Layers
conv1	128×128	7×7, 64, stride 2
maxpool	64×64	3×3, stride 2
conv2_x	64×64	$\begin{Bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{Bmatrix} \times 3$
conv3_x	32×32	$\begin{Bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{Bmatrix} \times 4$
conv4_x	16×16	$\begin{Bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{Bmatrix} \times 6$
conv5_x	16×16	1×1, 512 3×3, 512 1×1, 2048 1×1, 2048, stride 1, then, $\begin{Bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 2048 \end{Bmatrix} \times 2$

Global and local joint learning methods are employed to enable the model to learn distinguishing features and achieve more effective distinctions between vehicles with similar appearances. The details of the branches for deep feature extraction are shown in Table 1, where *Direction* represents the direction in which the feature map is partitioned, and *none* represents no partition in this branch. According to prior knowledge, the information of local parts such as wheels and logos of a vehicle has better distinguish ability. In addition, the vehicle is a rigid body, it is easy to get these areas with distinctive features by dividing the feature map. Hence, we combined the horizontal and vertical partition of the feature map, and designed two local branches.

Inputsize represents the size of the feature map input to each branch, *Outputsize* is the size of the output feature map of each branch. *h* and *w* denote height and width, respectively. Finally *Dim* means the depth of the feature map, which is the size of the channel dimension.

In the GB branch, because we resize the input image size to 256×256, after the resnet50 network with the last pooling layer stride of 1, we get a global feature map with a size of 16×16 and a channel number of 2048. By this method, we obtained a global feature map four times as large as using the original resnet50 network, so we were able to utilize more valid information from the input images.

In the VPB branch, after vertical average pooling and 1×1 convolutional dimensional reduction operation, a feature map of size 1×4 (height×weight) is obtained. Similar to the VPB branch, but the HPB branch uses horizontal averaging pooling to obtain a 4×1 feature map.

Finally, each local branch fuses the four parts of the vertically or horizontally partitioned feature maps to a final local feature map with 256×4 channels.

Backbone module

Table 2 provides the detailed architectures of the ResNet50 network that we use. Down sampling is performed by conv3_1 and conv4_1 with a stride of 2. The pre-trained ResNet50 [33] network on ImageNet is utilized as our backbone to extract vehicle global features. Different branches of our PFMN model share the same ResNet50 network. The last fully connected layer of the network is removed to obtain more useful information from the feature. The stride size of the last convolutional layer of ResNet50 is changed from 2 to 1 [34]. “output size” indicates the height × width of the output feature map for each layer with an original input image shape of 256×256.

Global branch module

The global feature is extracted by ResNet50 without any partition operations on the features in the global branch. It is first pooled through the Generalized-Mean Pooling (Gem) layer [34], which can improve retrieval performance by focusing on different fine-grained areas of the vehicle. Then the pooled feature is input to the *BN* layer for normalization, which accelerates the convergence of the network. Finally, we set a fully connected layer with a hidden size *N*, where *N* denotes the number of all vehicle IDs in the training set.

Local branches module

Since the vehicle can be roughly regarded as a cube and rigid body whose shape is stable, the prior knowledge can be used to uniformly partition the vehicle global feature map.

As in Fig. 2, the local part of our PFMN consists of two main modules. They are designed for feature partition and re-fusion, termed VPB and HPB. They share with similar

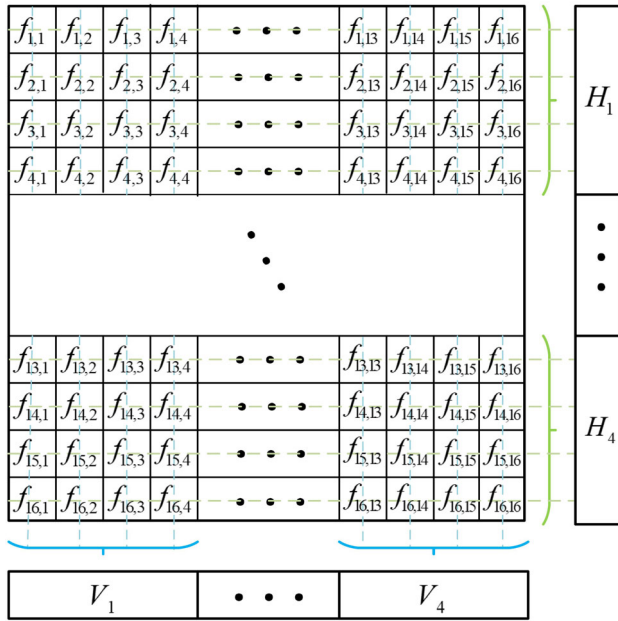


Fig. 3 Schematic diagram of vertical average pooling and horizontal average pooling

structures, yet apply either a vertical or horizontal average pooling layer for pooling.

Vertical average pooling The vertical average pooling layer partitions f_G into m parts in the vertical direction to extract the local features of each row, representing m vertical parts of the vehicle.

Horizontal average pooling The horizontal average pooling layer is similar to vertical average pooling. In this layer, the global feature map f_G is divided into n parts in the horizontal direction, corresponding horizontal part of the vehicle.

Different from ordinary convolutional pooling operations, our horizontal and vertical average pooling are performed in a complete sub-region of the feature map. This enables us to obtain local features in broader receptive field under prior knowledge.

In this paper, $f_G = 16 \times 16$ and we set $m = 4, n = 4$. $f_{i,j}^G$ represents the feature value in the i th row and j th column of the global feature f_G . As shown in Fig. 3, the global feature map is uniformly divided into four parts in the horizontal and vertical directions, respectively. H_1, H_2, H_3, H_4 , respectively, represent the horizontal average pooling results and V_1, V_2, V_3, V_4 , respectively, refer to those in vertical. H_1 and V_1 are calculated as follows. Other H_i s and V_i s are obtained in the same way:

$$H_1 = (f_{1,1}^G + f_{1,2}^G + \dots + f_{1,16}^G + f_{2,1}^G + f_{2,2}^G + \dots + f_{2,16}^G + \dots + f_{4,1}^G + f_{4,2}^G + \dots + f_{4,16}^G) \times \frac{1}{64} \quad (1)$$

$$V_1 = (f_{1,1}^G + f_{1,2}^G + \dots + f_{1,4}^G + f_{2,1}^G + f_{2,2}^G + \dots + f_{2,4}^G + \dots + f_{16,1}^G + f_{16,2}^G + \dots + f_{16,4}^G) \times \frac{1}{64} \quad (2)$$

After that, a 1×1 convolutional layer (Conv1*1) is used to reduce the dimension of the feature map, which can reduce the amount of calculation and improve the accuracy. After that, We fuse the features according to channel dimension. The fused features are used as the single representation of the vehicle. The same hard sample sampling strategy and an improved triplet loss function are shared within each local branch.

Improved triplet loss of hard samples

No bells and whistles, we make further improvements on the basis of the hard sampling triplet loss. By introducing the absolute distance of positive sample pairs based on hard sampling strategy to the loss function, the intra-class distance is further reduced.

Traditional triplet loss

A triplet loss contains three samples, namely, a randomly selected anchor sample, a positive sample and another negative sample. The anchor and positive samples share the same ID. The features extracted from the selected samples are recorded as $\{f_a, f_p, f_n\}$. After training with the triplet loss as a constraint, the distance between the positive samples gradually decreases, and the distance between the negative samples gradually increases, thus to achieve the target of gathering similar features in the feature space.

The traditional triplet loss function is calculated as

$$d_{a,p} = \|f_a - f_p\|_2 \quad (3)$$

$$d_{a,n} = \|f_a - f_n\|_2 \quad (4)$$

$$L_{Trp} = \frac{1}{N} \sum (d_{a,p} - d_{a,n} + \alpha)_+ \quad (5)$$

where $d_{a,p}$ means the Euclidean distance between features of anchor and positive samples, whereas $d_{a,n}$ means that of anchor and negative samples. $(z)_+$, a , p , and n represent $\max(z, 0)$ [35], anchor sample, positive sample, and negative sample, respectively. N denotes the number of batch size. α is a margin value used to distinguish between simple triples and hard triples. Only when $d_{a,p} > d_{a,n} + \alpha$, triples can be considered as valid and provide gradient changes. Otherwise the gradient is 0. In this case, these triples are not used for training.

Hard sampling triplet loss

The triplet loss function based on hard sample sampling uses hard samples that are not easy to distinguish for training, i.e., selecting the closest negative sample and the farthest positive sample in the mini-batch for training. Hard sampling triplet loss is formulated as following:

$$L_{TrpHard} = \frac{1}{N} \sum \left(\max_{p \subseteq A} d_{a,p} - \min_{n \subseteq B} d_{a,n} + \alpha \right)_+ \quad (6)$$

where A and B represent the positive and negative sample set, respectively. $\max_{p \subseteq A} d_{a,p}$ means selecting the hard sample with the largest relative distance from the anchor in the A set. $\min_{n \subseteq B} d_{a,n}$ means selecting the smallest relative distance from the anchor in the B set.

Our improved triplet loss of hard sampling

The original loss function for hard sampling triples only calculates the relative distance between the sample pairs. No difference exists when dealing the positive pairs with large absolute distance and small absolute distance. Therefore, they provide the same gradient. However, the sample pairs with large absolute distances are harder, thus should have a greater contribution to the change of the network gradient.

To tackle this problem, we add the absolute distance of the positive sample pairs to the original hard sampling triplet loss. Since if the positive and negative sample pairs in the triplet loss function are simple and easy to distinguish, it is not conducive to the training of the network. We select the hard positive sample pairs in a batch to calculate the loss. Because only hard samples are used for training, the model can obtain better gradients during training, leading to a more robust model.

Furthermore, We also designed different degrees of absolute distance, namely, the semi-hard sample absolute distance constraint and the easy-sample absolute distance constraint. The semi-hard sample constraint means adding the average of the absolute distances of all positive sample pairs in an input batch. In the same way, simple sample constraint is the minimum absolute distance of all positive sample pairs in a batch.

Improved triplet loss of hard sample constraint is calculated as following:

$$L_{hard-ImpTrpHard} = L_{TrpHard} + \lambda \times \frac{1}{N} \sum_{p \subseteq A} \max d_{a,p} \quad (7)$$

Table 3 Comparison of different branches on VeRi-776

Algorithm	mAP	Rank-1	Rank-5
Baseline	74.3	94.8	97.5
Baseline + HPB	75.8	94.3	97.5
Baseline + VPB	74.3	95.3	97.7
PFMN	77.2	95.4	98.0

The best performances are marked bold

The improved triplet loss function with the addition of the semi-hard sample constraint and the easy sample constraint are, respectively, calculated as follows:

$$L_{semi-ImpTrpHard} = L_{TrpHard} + \lambda \times \frac{1}{N} \sum_{p \subseteq A} \max d_{a,p} \quad (8)$$

$$L_{easy-ImpTrpHard} = L_{TrpHard} + \lambda \times \frac{1}{N} \sum_{p \subseteq A} \min d_{a,p} \quad (9)$$

As shown in Fig. 4, a denotes the selected anchor sample, p denotes to the positive sample, and n denotes to the negative sample. The distance representation of triples is composed of anchor, positive and negative samples in the feature space. If the Euclidean distance between the positive and negative sample pairs is less than the margin in (a), these sample pairs will be used to optimize the network. After training, as shown in (b), the Euclidean distance of the positive sample pair is reduced, which means the intra-class distance is reduced and the inter-class distance is increased. As a result, the vehicles with the same ID are closer in the feature space.

End-to-end training

As a preliminary of our work, a single-branch ResNet50 is used as the backbone to extract global feature. Then we jointly train the two local branches and the global branch. In this way, the well-learned representation by the global branch can benefit the learning of the local branches. Meanwhile, the ability to distinguish local details learned by the local branches can also enhance the global appearance representation.

In addition, after the modelling step, we designed different losses for the different branches. We adopt the same ImpTriHard loss on both of the local branches:

$$L_{VPB} = L_{ImpTriHard} \quad (10)$$

and

$$L_{HPB} = L_{ImpTriHard} \quad (11)$$

Fig. 4 Representation of the distance relationship of triples in feature space

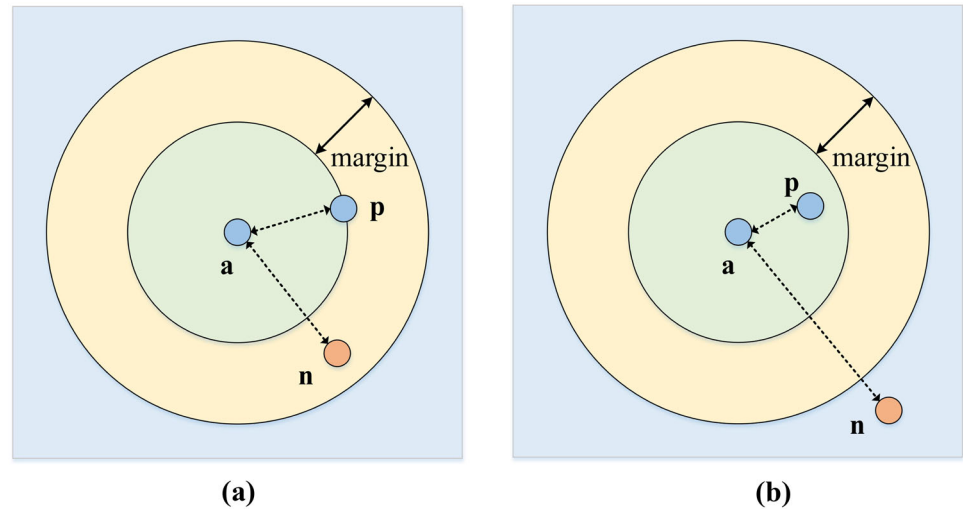


Table 4 Comparison of different branches on VehicleID

Algorithm	Test 800		Test 1600		Test 2400	
	Rank-1	Rank-5	Rank-1	Rank-5	Rank-1	Rank-5
Baseline	80.5	93.7	75.3	89.8	74.6	87.0
Baseline + HPB	82.5	95.4	76.5	91.3	74.8	89.0
Baseline + VPB	81.1	95.1	76.8	92.4	74.7	88.9
PFMN	83.1	95.0	77.4	92.4	74.7	89.2

The best performances are marked bold

Table 5 mAP, Rank-1 and Rank-5 on VeRi-776

Algorithm	mAP	Rank-1	Rank-5	mINP
PFMN	77.2	95.4	98.0	36.7
PFMN + $d_{ap,min}$	74.5	93.8	97.4	33.9
PFMN + $d_{ap,mean}$	78.4	96.5	98.0	39.2
PFMN + $d_{ap,max}$	78.9	95.9	98.4	40.6
PFMN + $d_{ap,max}$ + re-rank	80.8	96.8	97.4	44.6

The best performances are marked bold

For the global branch, we use the feature before normalization to optimize the triplet, and apply the normalized feature to calculate ID loss [49], which is mostly used in image classification tasks. We develop ImpTriHard loss and ID loss to learn a more discriminative feature for vehicle ReID. The loss function of the global branch can be described as follows:

$$L_G = L_{ID} + L_{ImpTriHard} \quad (12)$$

where

$$L_{ID} = - \sum_{i=1}^N \log \left(\frac{\exp(W_g^T x_g)}{\sum_{j=1}^T \exp(W_j^T x_j)} \right) \quad (13)$$

In Eq. 13, N denotes the number of images in an input batch. T is the number of categories in the whole training

data set. In addition, W_j^T is the weight vector with class j of the fully connected layer. g represents the ground truth identity of input image batch. x_j is the probability of the input image belonging to the j th identity.

In summary, the total loss of our PFMN network is calculated as follows:

$$L_{total} = L_G + L_{VPB} + L_{HPB} \quad (14)$$

Experiments

Data sets and protocols

We evaluate our model on two large-scale vehicle ReID data sets, namely, VeRi-776 [17] and VehicleID [5]. We introduce these data sets and their evaluation protocols before we show our results.

VeRi-776 is a classic vehicle re-identification data set. It consists of more than 50,000 images of 776 vehicles, which are collected by 20 cameras in the city block under different camera views. The training set contains 37,778 images of 576 vehicles, whereas the test set contains 200 vehicles, of which 11,579 pictures are used to construct the gallery set, and 1,678 pictures are used as the query set. The data set is marked with vehicle color and model information. All images of each vehicle are captured by 2–18 surveillance

Table 6 Comparison of different hard constraints on VehicleID

Algorithm	Test 800			Test 1600			Test 2400		
	Rank-1	Rank-5	mINP	Rank-1	Rank-5	mINP	Rank-1	Rank-5	mINP
PFMN	83.1	95.0	88.3	77.4	92.4	83.8	74.7	89.2	81.2
PFMN + $d_{ap.min}$	73.6	87.9	79.6	72.6	83.4	77.6	71.1	80.6	75.9
PFMN + $d_{ap.max}$	78.5	92.7	84.5	75.9	88.9	81.5	74.1	86.6	79.7
PFMN + $d_{ap.mean}$	84.1	95.8	89.1	81.1	94.0	86.7	78.8	91.6	84.4
PFMN + $d_{ap.mean}$ + re-rank	85.6	96.8	90.3	81.4	94.1	87.0	80.0	92.0	85.1

The best performances are marked bold

Table 7 Comparison of different parameter settings

λ	VeRi-776				VehicleID					
					Test 800		Test 1600		Test 2400	
	mAP	Rank-1	Rank-5	mINP	Rank-1	Rank-5	Rank-1	Rank-5	Rank-1	Rank-5
0.1	74.3	95.2	97.7	40.1	79.7	93.6	76.3	90.2	73.7	86.1
0.3	78.8	95.7	97.8	40.9	75.4	90.7	75.6	88.2	73.6	84.9
0.5	79.3	95.4	97.7	41.8	83.7	96.4	78.3	93.5	77.7	91.3
0.7	78.6	95.9	98.5	41.2	84.6	96.9	80.7	93.2	77.0	90.6
0.9	79.2	96.3	98.0	41.7	78.9	93.9	76.1	89.6	74.6	86.3
1	78.9	95.8	98.1	40.6	84.1	95.8	81.1	94.0	78.8	91.6

The best performances are marked bold

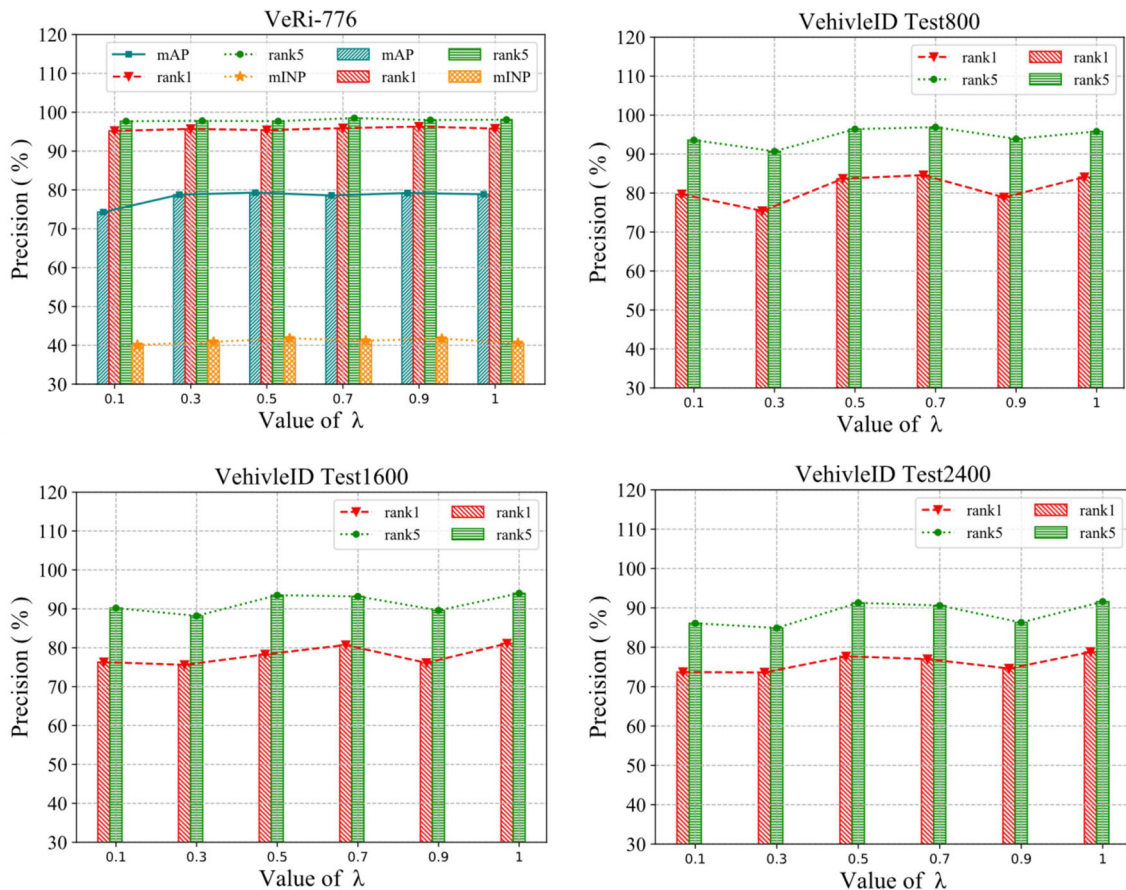
**Fig. 5** Results of ablation experiments with different λ values on VeRi-776 and VehicleID

Fig. 6 Visualization of vehicle feature maps by Grad-CAM [52]. **a** Input images. These two pictures are selected from the VeRi-776 data set. **b, d** Visualization of feature maps which trained without ImpTriHard loss. **c, e** Visualization of feature maps which trained with ImpTriHard loss

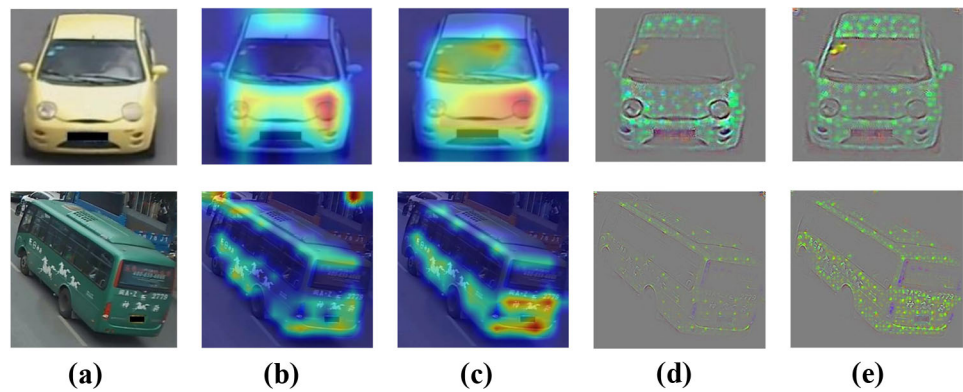


Fig. 7 Rank-10 results of vehicle re-identification on the VeRi-776 data set



cameras under different angles, illumination, resolution, and occlusion.

VehicleID is a large-scale vehicle ReID data set that contains images captured by multiple real surveillance cameras distributed in different hours in a small city in China. It contains 221,763 pictures, about 26,267 cars, each with an ID tag corresponding to the identity in the real world. Among them, the *Test800* test data set contains 800 query pictures and 6532 gallery pictures, the *Test1600* test data set contains 1600 query pictures and 11,395 gallery pictures, and the *Test2400* test data set contains 2400 query pictures and 17,638 gallery pictures. The images of this data set are captured in the front or back perspective.

Protocols To evaluate the accuracy of the model, we apply Rank-1, Rank-5, *mAP* (mean average precision) and a new

evaluation standard *mINP* (mean Inverse Negative Penalty) [36].

Rank-k indicates the accuracy of the highest *k* similarity that belong to the same ID with the query one.

mAP refers to calculating the average precision (*AP*) of all query images, which contemplating the recall rate and accuracy rate to evaluate the global performance of the model.

For a query image, *k* denotes the number of pictures with the same ID in the gallery set. *AP* is calculated as follows:

$$AP = \frac{1}{k} \sum_{i=1}^k \frac{1}{p_i} \quad (15)$$

where p_i is the minimum number of images that need to be queried to obtain the top *i* correct results in the image

Fig. 8 Rank-10 results of vehicle re-identification on the VehicleID data set

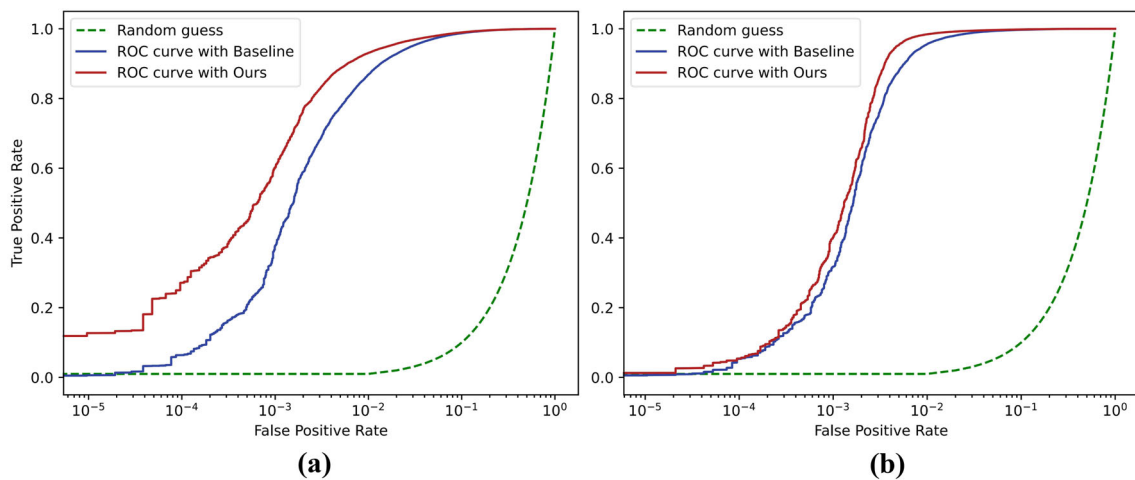
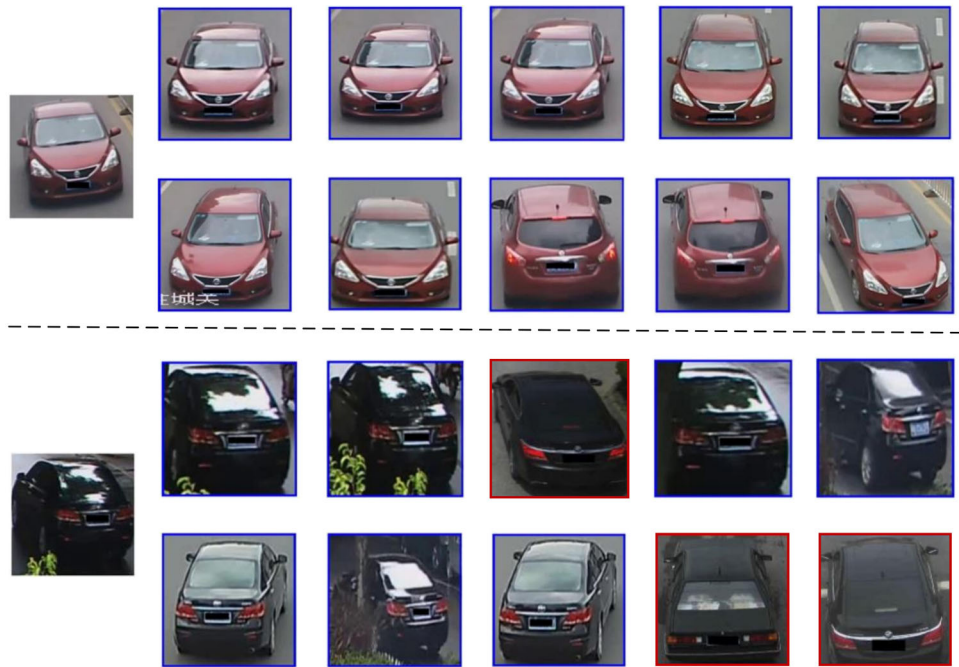


Fig. 9 ROC curve generated using our proposed model (PFMN+ImpTriHard loss) and the Baseline model using only global features. **a** shows the results on the VeRi-776 and **b** shows the results on VehicleID

sequence sorted according to the similarity. Then mAP can be obtained by averaging AP . Because it reflects the degree to which we rank all correct samples within the result queue. mAP can measure the performance of the vehicle ReID model more comprehensively.

$mINP$ is used to evaluate the cost of finding the most hard matching samples. R_i^{hard} is the position of the most hard sample in the ranking list, and i is the position of the last positive sample in the ranking list. Ranking list is made up by the retrieved results in the gallery set. G_i is the number of all positive samples in the ranking list. The negative penalty (NP) is formulated as follows:

$$NP_i = \frac{R_i^{hard} - G_i}{R_i^{hard}} \quad (16)$$

where NP_i is the total number of query set samples. The $mINP$ value of all query samples is formulated as following:

$$mINP = \frac{1}{N_q} \sum_{i=1}^{N_q} (1 - NP_i) = \frac{1}{N_q} \sum_{i=1}^{N_q} \frac{G_i}{R_i^{hard}} \quad (17)$$

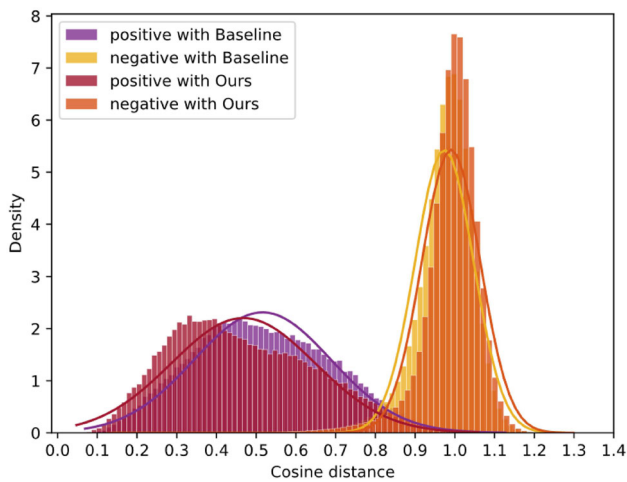


Fig. 10 Positive and negative pairs distribution on VeRi-776

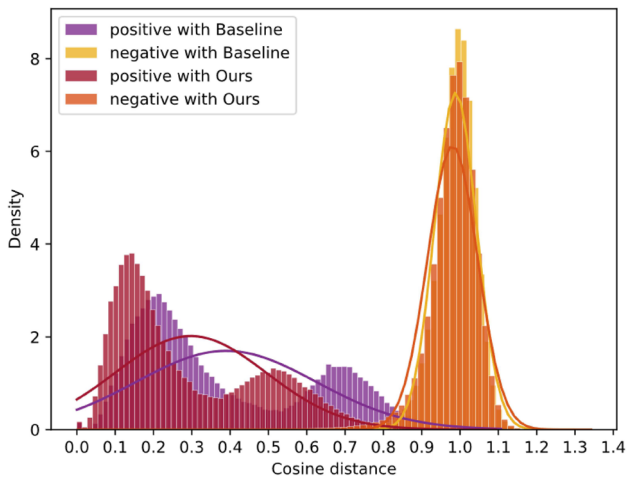


Fig. 11 Positive and negative pairs distribution on VehicleID

Implementation details

Here, we discuss the implementation of both the global and local branches. The size of the input image is scaled to 256×256 . Data augmentation techniques such as horizontal flipping and random erasure are used to improve the generalization. We use *ResNet50* as backbone for global feature extraction, remove its last fully connected layer and set the stride of the last pooling layer to 1. To complete the sampling part of the triplet, we select $P \times K$ pictures randomly among sixteen vehicles with different IDs in each mini batch, i.e., $P = 16$, and the number of pictures sampled by each ID is four, i.e., $K = 4$. The training batch size is 128 and the training epoch is 60. We choose Adam as the optimizer and the learning rate of the warm up strategy is adopted. The learning rate has a gradual increase from 3.5×10^{-5} to 3.5×10^{-4} in the first 10 epochs. The learning rate is set to 3.5×10^{-4} in 10 to 30 epochs, 3.5×10^{-5} in 30 to 50 epochs, and 3.5×10^{-6}

Table 8 mAP, Rank-1 and Rank-5 on VeRi-776

Algorithm	mAP	Rank-1	Rank-5
LOMO [37]	9.6	25.3	46.5
BOW-CN [38]	12.2	33.9	53.7
GoogLeNet [39]	17.9	52.3	72.2
FACT [6]	18.8	52.2	72.9
DenseNet121 [40]	45.1	80.3	91.1
NuFACT [20]	48.5	76.8	91.4
VAMI [42]	50.1	77.0	90.8
PROVID [20]	53.4	81.6	95.1
Siamese-CNN+Path-LSTM [41]	58.3	83.5	90.0
AAVER [46]	66.3	90.1	94.3
VCAM [50]	68.6	94.4	96.9
VehicleX [48]	73.2	94.9	97.9
Ours	79.3	95.4	97.7
Ours + re-rank	81.2	96.8	97.6

The best performances are marked bold

in 50 to 60 epochs. The local and global branch use the same batch of hard samples. The loss function used to optimize the model is the softmax cross-entropy loss and our ImpTriHard loss. All of our experiments on the two different data sets follows the above settings.

Ablation studies

In this subsection, extensive comparative experiments on different network branches and parameter settings of our ImpTriHard loss are presented to illustrate the effectiveness of our proposed method. All the results are based on VeRi-776 and VehicleID.

Effectiveness of local branches

We adopt the same network structure as BOT (“Bag of Tricks”) [30] as the baseline for comparison. We have conducted experiments on the baseline with different local branches that partition the feature map in horizontal and vertical directions, as shown in Tables 3 and 4. Since the additional branch pays more attention to local features, the performance of the model can be improved in variety degrees.

These results suggest Horizontal Partition Branch (HPB) performs better than Vertical Partition Branch (VPB) on both data sets. The model reaches its highest performance when the two branches are applied together. On the VeRi-776 data set, compared with the baseline, mAP is increased by 2.9%, and Rank-1 and Rank-5 are also increased by about 0.5%. Meanwhile, on the VehicleID data set, Rank-1 is improved by 2.6% and 2.1%, respectively, on the Test 800 and Test 1600 data sets.

Table 9 Rank-1 and Rank-5 on VehicleID

Algorithm	Test 800		Test 1600		Test 2400	
	Rank-1	Rank-5	Rank-1	Rank-5	Rank-1	Rank-5
BOW-CN [38]	13.1	22.7	12.9	21.1	10.2	17.9
LOMO [37]	19.7	32.1	19.0	29.5	15.3	25.6
GoogLeNet [39]	47.9	67.4	43.3	63.5	38.2	59.5
NuFACT [20]	48.9	69.5	43.6	65.3	38.6	60.7
FACT [6]	49.5	68.0	44.6	64.2	39.9	60.5
DRDL [17]	49.0	73.5	42.8	66.8	38.2	61.6
VAMI [42]	63.1	83.2	52.9	75.1	47.3	70.3
TAMR [44]	66.0	79.7	62.9	76.8	59.7	73.9
QD-DLF [45]	72.3	92.5	70.7	88.9	64.1	83.4
PRN [47]	78.9	94.8	74.9	92.0	71.5	88.4
SAVER [51]	79.9	95.2	77.6	91.1	75.3	88.3
Ours	84.1	95.8	81.1	94.0	78.8	91.6
Ours + re-rank	85.6	96.8	81.4	94.1	80.0	92.0

The best performances are marked bold

Improved performance of ImpTriHard loss

As in Tables 5 and 6, we have added three positive sample absolute distances (i.e., $d_{ap.min}$, $d_{ap.mean}$, $d_{ap.max}$) to the loss function of the proposed PFMN model. Through comparative experiments, we will illustrate the influence of different degrees of absolute distance constraints on the experimental results. On the VeRi-776 data set, adding the $d_{ap.max}$ can achieve the best performance. The mAP rate is increased by 1.7%, and the mINP rate is increased by 3.9%. However, on the VehicleID data set, comparing with adding other kinds of positive sample absolute distance, using $d_{ap.mean}$ has obvious improvement, which is caused by the data set itself.

There are many hard samples in the VeRi-776 data set. Images have lower resolution than that in VehicleID and the camera views changes greatly. In contrast, there are only two different camera views in the VehicleID data set, which are front and back. When the re-rank strategy is not used, the best mINP rate of our method on the VehicleID data set is 89.1%. However, the best mINP rate on the VeRi-776 data set is only 40.6%, which illustrates that the retrieval of hard samples on the VehicleID data set is much easier. Therefore, a softer $d_{ap.mean}$ instead of $d_{ap.max}$ is preferred on VehicleID.

Ablation Study of λ

Table 7 illustrates the influence of the trade-off coefficient λ . We set different λ values (0.1, 0.3, 0.5, 0.7, 0.9, 1.0) to conduct comparative experiments on the two data sets to achieve the best results. We visualize the results in Fig. 5. On the VeRi-776 data set, when $\lambda = 0.5$, the model achieves 79.3% mAP and 41.8% mINP. As the λ value continues to increase, the performance of the model on VeRi-776 is degraded. Obviously, when $\lambda=1$, the model reaches its peak on the VehicleID

data set. On the test sets of Test1600 and Test2400, Rank-1 is 81.1% and 78.8%, respectively, and Rank-5 is 94.0% and 91.6%, respectively. Moreover, We have achieved consistent experimental results on test sets of different sizes on VehicleID. The experimental results above show the effectiveness of our proposed ImpTriHard loss.

From Fig. 6, we can see that the ImpTriHard loss can effectively enhance the model's attention to local discriminative features, e.g., vehicle wheels and lamps, which makes it easier to retrieve hard samples. Meanwhile, the model's ability to learn global context information is also improved.

Evaluation on VeRi-776 and VehicleID

Ranking results in gallery set As shown in Figs. 7 and 8, we visualize the retrieval results of the model on the VeRi-776 and VehicleID gallery set. The leftmost of these two figures shows a query image and the rest shows the Rank-10 images using our proposed model. The vehicles with blue border are correct retrieved results, while the vehicles with red border are incorrect results. Most of the search results are correct, but there are still some wrong ones. This is because vehicles with different IDs have very similar appearances or vehicles with the same ID have large angle changes under different cameras, resulting in unsuccessful retrieval. It's easy to see that our model has high accuracy of retrieving the target vehicle in challenging vehicle ReID scenes.

Receiver Operating Characteristic It can be seen intuitively from Fig. 9 that the curves generated using our proposed model on VeRi-776 and VehicleID are closer to the upper left corner, and the area under the curve is larger, which shows that the accuracy of retrieval using our model is higher than the baseline.

Positive and negative pairs distribution As can be seen in Figs. 10 and 11, there is a different distribution of the cosine distance between positive and negative sample pairs, where the same color indicates samples of the same subject, either positive or negative sample pairs. The distribution of the negative pairs in the cosine distance is much larger than that of the positive sample pairs for both baseline and our proposed method. Obviously, compared with baseline, our model significantly reduces the distance between positive sample pairs, and simultaneously the distance between positive sample pairs is also expanded, indicating more robust of our model. When faced with the retrieval problem of hard samples, our model can identify positive and negative samples more accurately.

Comparison with state-of-the-art methods

In this subsection, our proposed PFMN network and ImpTriHard loss are compared with the current state-of-the-art methods. We compare our proposed method with both traditional feature extraction methods (i.e., LOMO [37] and BOW-CN [38]) and deep learning based models (i.e., GoogLeNet [39], FACT [6], DenseNet121 [40], NuFACT [20], Siamese-CNN+Path-LSTM [41], PROVID [20], VAMI [42], DRDL [17], TAMR [43], and QD-DLF [45]).

Results on VeRi-776 are shown in Table 8. The result based on deep learning methods exceeds that of using traditional feature extraction by a large degree. Models of using a single global branch for re-identification (i.e., GoogLeNet [39], FACT [6], DenseNet121 [40] and VAMI [42]) is often unable to achieve relatively high rate of mAP, Rank-1 and Rank-5. Because many important details of vehicles are ignored when only global features are considered. Multiple branch networks (i.e., PROVID [20], Siamese-CNN+Path-LSTM [41], AAVER [46], VehicleX [48]) can achieve higher accuracy than the single global branch models. When local features of the vehicle are used, our proposed PFMN can achieve better results than the baseline. Without using the re-rank method, our proposed ImpTriHard loss has further increased mAP to 79.3%, Rank-1 to 95.4% on VeRi-776.

Results on VehicleID are listed in Table 9. Similar to the results on the VeRi-776 data set, our method on the VehicleID data set outperforms most of state-of-the-art methods by a large degree. To prove the effectiveness of our partition and fusion strategy with ImpTriHard loss, we also compare with the partition method proposed in PRN [47]. Our model can achieve a relatively large improvement when compared with the PRN network. Specifically, On Test800 of VehicleID, our method can, respectively, increase by 4.2% in Rank-1, and 1% in Rank-5. Especially on Test2400, the biggest test set of VehicleID, our method outperforms PRN on both Rank-1 and Rank-5 accuracy by 7.3% and

3.2%, respectively. In the case of using the re-rank strategy, our approach can make further improvement of 1% ~ 1.5%. In conclusion, the above results clearly indicate the effectiveness of our PFMN network and ImpTriHard loss.

Conclusions

In this paper, we have proposed an end-to-end Partition and Fusion Multi-branch Network, where both global and local features are utilized for vehicle re-identification. Furthermore, a new variant of metric learning loss named ImpTriHard loss is proposed to deal with the retrieval problem of hard samples. In more detail, local features are extracted using the method of uniformly partition the feature map in the horizontal and vertical directions without any additional annotation information. The results of ablation studies have demonstrated that the improvement of the ability to retrieval hard samples produced by ImpTriHard loss is more obvious on the VeRi-776 data set, which has more camera views, lower resolution and more complicated environment. Governed by the re-rank post-processing method, the accuracy of the model will be further promoted. Extensive experiments have indicated that our method shows better performance than most of other methods. Designing more efficient feature alignment method plays a significant role in improving the accuracy of the model, which uses local features. In the future research, we will consider introducing new feature alignment mechanisms.

Funding This work is supported by National Natural Science Foundation of China (Grant No. 61603233).

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Zhang J, Wang FY, Wang K, Lin WH, Xu X, Chen C (2011) Data-driven intelligent transportation systems: a survey. *IEEE Trans Intell Transp Syst* 12(4):1624–1639
2. Wang Z, Li X, Zhu X et al (2021) Big data-driven public transportation network: a simulation approach. *Complex Intell Syst*, 1–13

3. Alsufyani A, Alotaibi Y, Almagrabi A.O et al.(2021) Optimized intelligent data management framework for a cyber-physical system for computational applications. *Complex Intell Syst*, 1-13
4. FERENC Z A, LEARNED-MILLER E G, MALIK J (2005) Building a classification cascade for visual identification from one example. In: *The IEEE International Conference on Computer Vision*, pp 286-293
5. LIU X, LIU W, MEI T, et al (2016) A deep learning-based approach to progressive vehicle re-identification for urban surveillance. In: *The European Conference on Computer Vision*. Heidelberg: Springer, pp 869-884
6. Liu X, Liu W, Ma H, et al. (2016) Large-scale vehicle re-identification in urban surveillance videos[C]/2016 IEEE International Conference on Multimedia and Expo (ICME). IEEE, pp 1-6
7. Zhang N, Ju Z, Yang C et al (2021) Special issue on interpretation of deep learning: prediction, representation, quantification and visualization. *Complex Intell Syst*, 1-3
8. Priya S, Uthra RA (2021) Deep learning framework for handling concept drift and class imbalanced complex decision-making on streaming data. *Complex Intell Syst*, 1-17
9. Xia Y, Zhang J, Jiang T et al.(2021) HatchEnsemble: an efficient and practical uncertainty quantification method for deep neural networks. *Complex Intell Syst*, 1-15
10. Anuse A, Vyas V (2016) A novel training algorithm for convolutional neural network. *Complex Intell Syst* 2:221–234
11. Meng D , Li L , Liu X , et al (2020) Parsing-Based View-Aware Embedding Network for Vehicle Re-Identification. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp 7101-7110
12. Wang Z , Tang L , Liu X , et al (2017) Orientation Invariant Feature Embedding and Spatial Temporal Regularization for Vehicle Re-identification. In: *IEEE International Conference on Computer Vision (ICCV)*. IEEE, pp 379-387
13. Pasupa K, Kittiworapanya P, Hongngern N et al (2021) Evaluation of deep learning algorithms for semantic segmentation of car parts. *Complex Intell Syst*, 1-13
14. Saleem S, Amin J, Sharif M et al (2021) A deep network designed for segmentation and classification of leukemia using fusion of the transfer learning models. *Complex Intell Syst*, 1-16
15. Hermans A, Lucas B, Bastian L (2017) In defense of the triplet loss for person re-identification. *arXiv preprint [arXiv:1703.07737](https://arxiv.org/abs/1703.07737)*
16. Zhun Z, Liang Z, Donglin C, Shaozi L (2017) Re-ranking person re-identification with k-reciprocal encoding. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp 3652-3661
17. Liu H, Tian Y, Yang Y, Pang L and Huang T (2016) Deep relative distance learning: Tell the difference between similar vehicles. In: *Proceedings of the Conference on Computer Vision and Pattern Recognition (Piscataway, NJ: IEEE)*, pp 2167-2175
18. Tang Z, Naphade M, Liu M, Yang X, Birchfield S, Wang S, Kumar R, Anastasiu D.C, Hwang J (2019) Cityflow: A city-scale benchmark for multi-target multi-camera vehicle tracking and re-identification. In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pp 8797-8806
19. Yang L, Luo P, Loy C.C, Tang X (2015) A large-scale car dataset for fine-grained categorization and verification. In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pp 3973-3981
20. Liu X, Liu W, Mei T (2018) PROVID: progressive and multimodal vehicle reidentification for large-scale urban surveillance. In: *IEEE Transactions on Multimedia*, pp 645-658
21. Aihua Z, Xianmin L, Chenglong L, Ran H, Jin T (2019) Attributes guided feature learning for vehicle re-identification
22. Yan K, Tian Y, Wang Y (2017) Exploiting multi-grain ranking constraints for precisely searching visually-similar vehicles. In: *IEEE International Conference on Computer Vision*, pp 562-570
23. Bing H, Jia L, Yifan Z, Yonghong T (2019) Partregularized near-duplicate vehicle re-identification. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp 3997–4005
24. Liu H, Feng J, Qi M, Jiang J, Yan S(2017) End-to-end comparative attention networks for person re-identification.' *IEEE Transactions on Image Processing*, , pp 3492-3506
25. Liu H, Tian Y, Yang Y (2016) Deep relative distance learning: tell the difference between similar vehicles. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp 2167–2175
26. GUO H, ZHAO C, LIU Z (2018) Learning coarse-to-fine structured feature embedding for vehicle re-identification. In: *The 3rd AAAI Conference on Artificial Intelligence*, pp 6853–6860
27. Bai Y, Lou Y, Gao F (2018) Group-sensitive triplet embedding for vehicle re-identification. In: *IEEE Transactions on Multimedia*, pp 2385–2399
28. Zheng Z, Ruan T, Wei Y, Yang Y, Mei T(2020) VehicleNet: Learning Robust Visual Representation for Vehicle Re-identification. *IEEE Transactions on Multimedia*
29. Wang G, Yuan Y, Chen X, et al (2018) Learning discriminative features with multiple granularities for person re-identification. In: *The 2018 ACM Multimedia Conference on Multimedia Conference*, pp 274–282
30. Luo H, Gu Y, Liao X, Lai S, Jiang W (2019) Bag of tricks and a strong baseline for deep person re-identification. In: *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp 1487–1495
31. Yifan S, Liang Z, Yi Y, Qi T, Shengjin W (2018) Beyond part models: Person retrieval with refined part pooling (and A strong convolutional baseline). In *ECCV*, pp 501–518
32. Wang G, Yuan Y, Li J, Ge S, Zhou X(2020) Receptive Multi-Granularity Representation for Person Re-Identification. In: *IEEE Transactions on Image Processing*, pp 6096–6109
33. Kaiming H, Zhang X, Ren S, Sun J (2016) Deep Residual Learning for Image Recognition. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp 770–778
34. Maxim B et al (2019) Multigrain: a unified image embedding for classes and instances. *arXiv preprint [arXiv:1902.05509](https://arxiv.org/abs/1902.05509)*
35. Rosasco L, De Vito E, Caponnetto A, Piana M, Verri A.(2004) Are loss functions all the same? *Neural Computation*,pp 1063–1076
36. Ye M, Shen J, Lin G, Xiang T, Shao L, Hoi SCH (2021) Deep Learning for Person Re-identification: A Survey and Outlook. *IEEE Trans Pattern Anal Mach Intell*
37. - Liao S, Hu Y, Zhu X, Li S Z (2015) Person re-identification by local maximal occurrence representation and metric learning. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp 2197–2206
38. Zheng L, Shen L, Tian L, Wang S, Wang J, Tian Q (2015) Scalable person re-identification: A benchmark. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp 1116–1124
39. Yang L, Luo P, Change Loy C, Tang X (2015) A large-scale car dataset for fine-grained categorization and verification. In: *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pp 3973–3981
40. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp 4700–4708
41. Shen Y, Xiao T, Li H, Yi S, Wang X (2017) Learning deep neural networks for vehicle re-ID with visual-spatio-temporal path proposals. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp 1900-1909

42. Zhou Y, Shao L (2018) Aware attentive multi-view inference for vehicle re-identification. In: *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pp 6489–6498
43. Guo H, Zhu K, Tang M, Wang J (2019) Two-level attention network with multi-grain ranking loss for vehicle re-identification. *IEEE Trans. Image Process*, pp 4328–4338
44. He B, Li J, Zhao Y, Tian Y (2019) Part-regularized near-duplicate vehicle re-identification. In: *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pp 3997–4005
45. Zhu J, Zeng H, Huang J, Liao S, Lei Z, Cai C, Zheng L (2019) Vehicle re-identification using quadruple directional deep learning features. *IEEE Trans Intell Transp Syst*, pp 2410–2420
46. Khorramshahi P, Kumar A, Peri N, Rambhatla SS, Chen JC, Chellappa R (2019) A dual-path model with adaptive attention for vehicle re-identification. In: *The IEEE International Conference on Computer Vision (ICCV)*, pp 6131–6140
47. Chen H, Lagadec B, Bremond F (2019) A Two-Branch Neural Network for Vehicle Re-identification. In: *CVPR Workshops, Partition and Reunion*, pp 184–192
48. Yao Y, Zheng L, Yang X, Naphade M, Gedeon T (2020) Simulating content consistent vehicle datasets with attribute descent. In: *Computer Vision-ECCV 2020*. Springer International Publishing, Part VI 16:775–791
49. Zhedong Z, Zheng L, Yang Y (2017) A discriminatively learned cnn embedding for person reidentification. In: *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 14.1, pp 1–20
50. Chen TS, Lee MY, Liu CT, Chien SY (2020) Aware Channel-Wise Attentive Network for Vehicle Re-Identification. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp 574–575
51. Khorramshahi P, Peri N, Chen JC, Chellappa R. (2020) The devil is in the details: Self-supervised attention for vehicle re-identification. In: *European Conference on Computer Vision*, Springer, Cham, pp 369–386
52. Ramprasaath RS, Michael C, Abhishek D, Ramakrishna V, Devi P, Dhruv B (2017) Grad-cam: Visual explanations from deep networks via gradient-based localization. In: *ICCV*, pp 618–626

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.