Neural Computing and Applications
https://doi.org/10.1007/s00521-020-05108-x

REVIEW

Multi-scale attention vehicle re-identification

Aihua Zheng[1] · Xianmin Lin[1] · Jiacheng Dong[1] · Wenzhong Wang[1] · Jin Tang[1] · Bin Luo[1]

Received: 18 June 2019 / Accepted: 5 June 2020

Abstract

Vehicle re-identification (Re-ID) aims to match the vehicle images with the same identity captured by the non-overlapping surveillance cameras. Most existing vehicle Re-ID methods focus on effective deep network architectures to extract discriminative features from single-scale images. However, these methods ignored the complementary information from different scales, which is a crucial factor in computer vision tasks. Attention mechanism, a commonly used technique in recognition and detection tasks, can selectively focus on discriminative local cues of the image. In this work, we propose a multi-scale attention framework which jointly considers multi-scale mechanism and attention technique for vehicle Re-ID. Specifically, we exploit multi-scale mechanism in feature maps, which can acquire more comprehensive representations for fusing global and local cues. Meanwhile, we exploit attention blocks on each scale subnetwork, which aims to mine complementary and discriminative information. We conduct extensive experiments on three vehicle datasets, VeRi-776, VehicleID and PKU-VD. The promising results demonstrate the effectiveness of the proposed method and yield to a new state of the art for vehicle Re-ID.

Keywords Vehicle re-identification · Multi-scale · Attention

1 Introduction

Vehicle re-identification (Re-ID) is to verify whether vehicle shot in one camera appears in other non-overlapping cameras. It is of increasing importance in computer vision task due to the wide range of potential applications such as cross-camera tracking, intelligent monitoring and urban surveillance. Although license plates are unique identities for vehicles, their applications in uncontrolled urban surveillance are limited since the current LPR (licence plate recognition) techniques are struggling in such complex environments where low-quality images, arbitrary viewpoints, motion blur, poor lighting conditions are pervasive. Therefore, vehicle Re-ID approaches mainly devote to exploring the vehicle appearance information. Similar to the person Re-ID, vehicle Re-ID suffers from many challenges due to the viewpoint and illumination changes, occlusion, which bring large appearance variations for the same identity across different cameras, as shown on the top three rows in Fig. 1. Furthermore, vehicle Re-ID has its particular challenge: different identities may have similar or even the same appearance especially for the vehicles with the same model from the same manufacturer, as shown at the bottom row in Fig. 1.

Recently, deep learning has been applied in numerous computer vision problems such as object detection [5], object recognition [3, 43], data representation [10]. A lot of vehicle Re-ID methods based on CNN networks [19, 44, 47] have been developed recently. They mainly focus on either designing new network architectures to learn more discriminative features or introducing extra information to boost the performance of the vehicle Re-ID

Xianmin Lin and Jiacheng Dong have been contributed equally to this paper.

✉ Wenzhong Wang
wenzhong@ahu.edu.cn

Aihua Zheng
ahzheng214@ahu.edu.cn

Xianmin Lin
xmlin1995@gmail.com

Jiacheng Dong
jiachengdong@foxmail.com

Jin Tang
tj@ahu.edu.cn

Bin Luo
luobin@ahu.edu.cn

[1] School of Computer Science and Technology, Anhui University, Hefei 230601, China

Published online: 23 June 2020

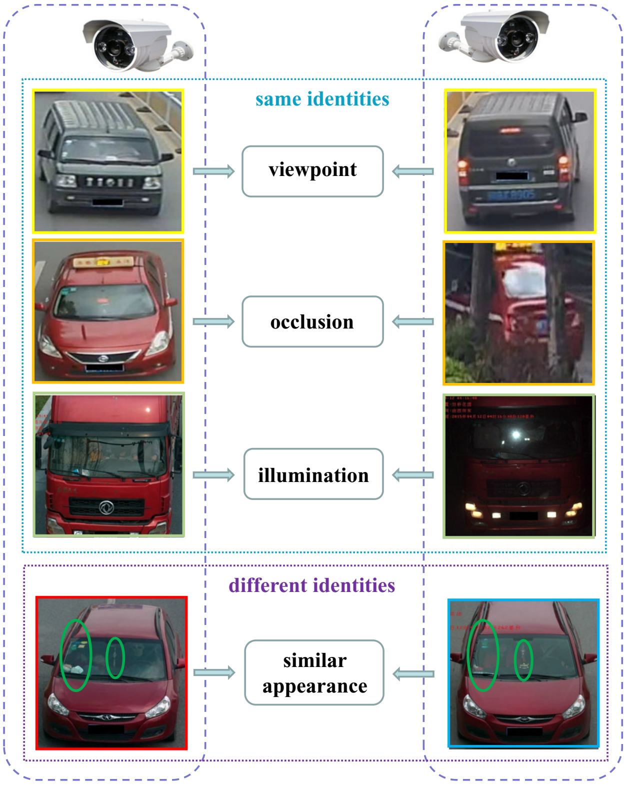


Fig. 1 Illustration of challenges in vehicle Re-ID. The vehicle parked in the first three rows demonstrate that the same vehicle identities appear differently in vision due to the changes of the viewpoint, illumination, and the occlusion. The last row illustrates the challenge of different vehicle identities with extremely similar appearance, where the green circles indicate the local difference in tags and hangings

models. Most of them focused on the information learnt from the single scale of the original vehicle images while ignoring the information from different scales. Local information is crucial cues in vehicle Re-ID. As shown in Fig. 1, the different identities with similar global appearance can be better distinguished according to there local appearance cues. Multi-scale mechanism has been successfully applied in many computer vision tasks such as video content-based advertising [45], object detection [9, 21], face recognition [30] and segmentation [13], considering that small scales contain more global information and large scales contain more local cues. Therefore, it is important to learn rich hierarchical features from multiple scales in order to resolve the challenging appearance ambiguity in vehicle re-identification.

Meanwhile, inspired by the fact that human visual system (HVS) always concentrates on a certain part of visual data, visual attention mechanism has been exploited in many tasks such as person Re-ID [16], pedestrian counting [38], detection [39] and image search [1]. Visual attention model can automatically produce the regions of interest from the image. To mine the discriminative local parts, we propose to introduce the visual attention mechanism into each scale in a unified end-to-end network.

Pooling layers are essential parts of CNN network which are used to expand the fields of perception. However, they shrink the size of feature maps to decrease the resolution. Therefore, the fine-grained information in extracted feature maps is always lost in conventional deep learning models, which is crucial in Re-ID task. Considering the over-small size problem of the original feature maps caused by pooling operations, we apply multi-scale mechanism on feature map level to supply the missing information caused by pooling operations [46].

Based on the above discussion, we propose a novel multi-scale attention framework (MSA) for vehicle Re-ID, as shown in Fig. 2. Firstly, we feed the image into a backbone network, and rescale the produced feature maps into multiple scales using bilinear interpolation. The feature maps at each scale are fed into a scale-specific subnetwork, each followed by a spatial-channel attention block. After progressively training these subnetworks, we finally use two convolutional layers to fuse the multi-scale feature maps and fine-tune the whole network.

Compared with conventional approaches for vehicle reidentification, this paper makes the following three main contributions:

- We propose to explore the multi-scale mechanism into the vehicle Re-ID task. By fusing the global and local information from different scales, our method can acquire more discriminative features which contain global and local cues.
- We propose a novel multi-scale attention framework (MSA) by integrating the attention model into the multi-scale framework to further mine discriminative cues, which provides a general framework to exploit multi-scale attention for Re-ID.
- We jointly consider the multi-scale and spatial-channel attention mechanism in a unified framework for vehicle Re-ID. Comprehensive experimental results on three benchmark 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 reviews some recent related works. Section 3 elaborates the proposed multi-scale attention (MSA) framework. Section 4 demonstrates the experimental results on public benchmark datasets with comprehensive evaluations on the proposed method, while Sect. 5 concludes our paper.

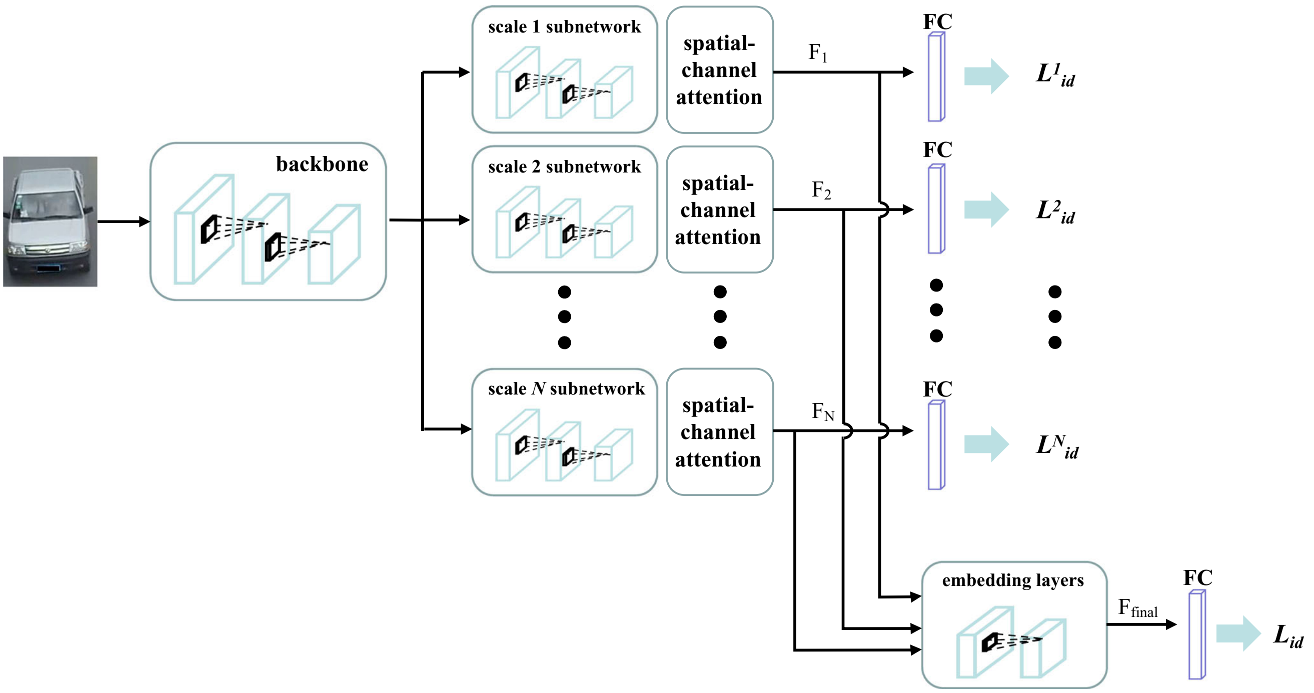


Fig. 2 Overall architecture of the proposed framework. We feed the input image into the backbone network, and then we exploit bilinear interpolation to generate *N* scale feature maps. Each scale feature map is then fed into corresponding subnetworks, followed by the spatial-channel attention block. After training these subnetworks, we use the embedding layers to fuse the multi-scale attentional feature maps and fine-tune the whole network

2 Related work

In this section, we shall briefly review the recent related works including vehicle Re-ID, multi-scale and attention person Re-ID tasks.

2.1 Vehicle Re-ID datasets

Person Re-ID [16, 17, 50, 53] has drawn much attention recently, which boosts the research of vehicle re-identification. Several vehicle Re-ID datasets have been proposed. Liu et al. [26] 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 1678 images of 200 vehicles as query set. Furthermore, it provides attributes (color and type) information and a part of license plate information. Liu et al. [23] released a larger dataset VehicleID, which contains 221,763 images of 26,267 vehicles, including the training set with 110,178 images of 13,134 vehicles and testing set with 111,585 images of 13,133 vehicles. Before above two dataset, Yang et al. [41] proposed CompCars dataset with 136,726 images of 1716 models from two scenarios (web-nature and surveillance-nature). It artificially divided the vehicle images into 100,000 pairs; therefore, it can also been used for vehicle Re-ID task. More recently, Liu et al. [24] proposed the Vehicle-1M dataset with nearly 1 million vehicle images, which is the largest vehicle Re-ID dataset at present. NVIDIA released a large CityFlow dataset [36] in AI City Challenge in CVPR 2019, which contains 36,935 vehicle images from 40 cameras together with the spatio-temporal path information. Yan et al. [12] proposed a large-scale vehicle search dataset PKU-VD, which is collected by high-resolution traffic cameras. There are 1,097,649 images of 141,756 vehicles in total. It contains training set with 422,326 images of 70,591 vehicles and testing set with 424,032 images of 71,165 vehicles.

2.2 Vehicle Re-ID methods

The development of deep learning model accelerate the research of vehicle re-identification, a lot of effective network architectures have been designed to achieve the better matching performance for vehicle Re-ID. Zapletal et al. [44] detected 3D bounding boxes of vehicles and extracted robust vehicle representation based on color histograms and histograms of oriented gradients. Zhang et al. [47] improved the triplet-wise training based on classification-oriented loss and a novel triplet sampling method for vehicle Re-ID. Kanacı et al. [11] exploited cross-level vehicle recognition method to avoid expensive and time-consuming label collection. Zhu et al. [52] proposed novel short and dense units, which can combine the advantages of VGGNet and DenseNet. Furthermore, auxiliary information has been integrated in vehicle Re-ID task, such as the spatiotemporal information [33], or attributes information such as color and type [27]. Li et al. [19] proposed a unified vehicle Re-ID framework which exploited vehicle attributes and the relationship among samples. Zhou et al. [51] first considered exploiting conditional generative network to generate different viewpoint vehicle images for vehicle Re-ID task. However, they mainly focus on the global appearance of the vehicles while neglected the local information in different scales and the various contribution from different spatial and channel aspects. Herein, we propose jointly consider the multi-scale and spatial-channel attention mechanism in a unified framework for vehicle Re-ID.

2.3 Multi-scale person Re-ID

Multi-scale mechanism has been successfully exploited in person Re-ID task. One of the pioneer work is Li et al. [18], which proposed a novel multi-scale discriminant distance metric learning method to align the cross-scale image domain. Liu et al. proposed a multi-scale deep person Re-ID model [25] which down-sampled different scales of the input image and fed them into different sub-networks to extract multi-scale features, followed by fusing network. To automatically fuse the contributions from different scales of the person images, Fu et al. [4] designed a novel multi-scale deep learning model which learnt deep robust feature representations at different scales and automatically selected the most discriminative scales for metric learning. Chen et al. [2] proposed an deep pyramidal feature learning CNN architecture to solve the challenge of the person images with different scales (resolutions) in person Re-ID. To better capture both the global and local information, this paper propose to learn more discriminative descriptor by multi-scale mechanism for vehicle Re-ID.

2.4 Attention models in person and vehicle Re-ID

Attention mechanism has been first explored in person Re-ID task due to its benefit of handling the matching misalignment challenge [15, 35, 48]. For example, to learn discriminative representations from global and local parts of the person image, Su et al. [35] proposed a pose-driven deep convolutional network which integrated the human part cues into part-based Re-ID model. Li et al. [15] designed a multi-scale context-aware network which could locate the latent discriminative regions of image. To mine the important human body as spatial constraint, Zhao et al. [48] proposed a part-aligned human representation method based on the spatial transformer network [8]. Liu et al. [29] exploited a soft attention model to learn the discriminative regions at the pixel level in person Re-ID task. Li et al. [16] jointly learnt soft pixel attention and hard region attention to boost the person Re-ID performance. Recently, Teng et al. [37] proposed a spatial and channel attention network to mine the discriminative features in vehicle Re-ID task. However, they only simply employed single attention block, while this paper introduces multiple attention blocks with multi-scale mechanism, to achieve more comprehensive feature representation for vehicle Re-ID.

3 Our algorithm

In this paper, we propose a multi-scale attention (MSA) framework for vehicle re-identification. In this section, we first make a brief summarization for the overall architecture of MSA, followed by the elaboration of the multi-scale mechanism and spatial-channel attention mechanism. Finally, we summarize the implementation details of our framework.

3.1 Overall architecture of MSA

The proposed multi-scale attention framework is shown in Fig. 2. Our proposed framework consists of one backbone network, multi-branch subnetworks and the embedding layers.

Due to the compelling performance of residual learning, we use the first three residual blocks of ResNet-50 [6] as our backbone. Then we resize the feature map generated by backbone into different scales and feed them into the following subnetworks correspondingly.

As shown in Fig. 2, each scale subnetwork is composed of the Block-4 of ResNet-50 and a spatial-channel attention block. We use these subnetworks to extract N-scale high-level features $(F_1, F_2, \ldots, F_N)$. Each subnetwork is learned using the cross-entropy loss $L_{\text{id}}^n, n = 1, 2, \ldots, N$.

The embedding layers consist of two convolution layers and one FC layer. After resizing $F_2, \ldots, F_N$ to the F_1 size, the convolution layer is used to embed the comprehensive features $F_n (n = 1, 2, \ldots, N)$ into the higher level joint feature F_{final}. Then, the feature F_{final} is fed into the FC layer with the C-way softmax to predict the labels. We use cross-entropy loss L_{id} to learn the parameters of the model,

$$L_{\text{id}} = -\sum_{c=1}^{C} \log(p(c))q(c), \tag{1}$$

where C is the number of vehicle identity in the training set. $q(c)$ is the one-hot vector of the ground-truth. $p(c)$ is the predicted probability by softmax. L_{id}^n for $n \in 1, \ldots, N$ is defined in the same manner as L_{id}, We shall elaborate the multi-scale and attention mechanisms in the following sections.

3.2 Bilinear interpolation multi-scale mechanism

We introduce multi-scale in vehicle Re-ID task since different scales contains complimentary information. Jointly learn from multiple scales would hopefully exploit this complementary nature. We expand the scale of the original feature maps, which can also deal with the over-small size problem of the original feature maps caused by pooling operations. Herein, we generate feature maps at different scales to consider both coarse and fine information in vehicle Re-ID. The low-resolution feature map and high-resolution feature map will generate different representations, respectively.

3.3 Spatial-channel attention mechanism

Inspired by the theory of human vision, deep attention learning methods have made great progress in computer vision tasks such as person Re-ID [22, 34], image classification [31, 32], image question answering [42]. Attention mechanism plays an important role in vehicle Re-ID especially for the vehicles with highly similar global appearance. By enforcing the attention on different scales, a more discriminative representation is expected to be obtained. In this paper, we exploit the spatial-channel attention model [16] as the attention block, which can select discriminative regions of image and important channel of feature maps for vehicle Re-ID. There are two branches in the spatial-channel attention block, (1) the Spatial Attention Branch (SAB) to select discriminative pixels, and (2) the Channel Attention Branch (CAB) to select important channels. Figure 3 shows the structure of the spatial-channel attention model.

The input of spatial-channel attention block is an original feature map $f \in R^{h\times w\times c}$, where h, w, and c represent the size of height, width, and channel of the feature map, respectively. After the processing of spatial attention branch and channel attention branch, we will obtain spatial attention maps $s \in R^{h\times w\times 1}$ and channel attention maps $c \in R^{1\times 1\times c}$, respectively. To better integrate the generated attention maps from SAB and CAB, a convolution layer is followed after tensor product. We exploit the sigmoid function to normalize the attention map, which has the same size with the input feature map. Finally, we fuse the weights of the final attention maps with the original feature maps by element-wise multiplication, which generates the final attentional feature maps.

3.3.1 Spatial attention branch (SAB)

We exploit SAB to automatically select discriminative pixels of vehicle images. As shown in Fig. 3. We feed a set of tensors $f \in R^{h\times w\times c}$ into the SAB branch which contains four layers. We use a global channel-wise average-pooling layer to compress the input feature map, which is defined as follows [7]:

$$s_{i,j} = \frac{1}{c}\sum_{k=1}^{c} f_{i,j,k}, \qquad (2)$$

where $f_{i,j,k}$ represents the value of the k-th channel at the location (i, j). After global channel-wise average-pooling layer, a convolutional layer (with 3 × 3 filter) and a resize layer is followed. Finally, we use a scaling convolutional layer (with 1 × 1 filter) to learn an adaptive feature scale for the fusing processing with channel attention map. Here we employ ReLU as active function.

3.3.2 Channel attention branch (CAB)

We use CAB to automatically concentrate the discriminative channels of feature maps. CAB contains three layers as shown is Fig. 3. Firstly, we exploit an average pooling layer to aggregate spatial feature cues into channel signature (c_k), which is defined as:

$$c_k = \frac{1}{h\times w}\sum_{i=1}^{h}\sum_{j=1}^{w} f_{i,j,k}. \qquad (3)$$

Then we add two convolution layers to learn an adaptive feature scale for the fusing processing with spatial attention map.

Figure 4 visualizes the spatial attention in vehicle Re-ID. To better illustrate the benefit of the attention, we demonstrate the corresponding attention maps for the two vehicles with the same model and color but different IDs, which is one of the key challenge in vehicle Re-ID. From Fig. 4 we can observe that the spatial attention maps can highlight the discriminative spatial regions, such as the headlights, the vehicle tags or decorations, which are crucial to distinguish the local difference for similar vehicle images.

3.4 Implementation details

We progressively learn the three subnetworks and fine-tune the whole multi-scale attention framework, as described in Algorithm 1, which can significantly reduce the computational complexity. We initialize the backbone with the parameters learned on ImageNet [14], and the rest of Re-ID model from scratch. We train our model using mini-batch gradient descent, and perform Adam optimizer at

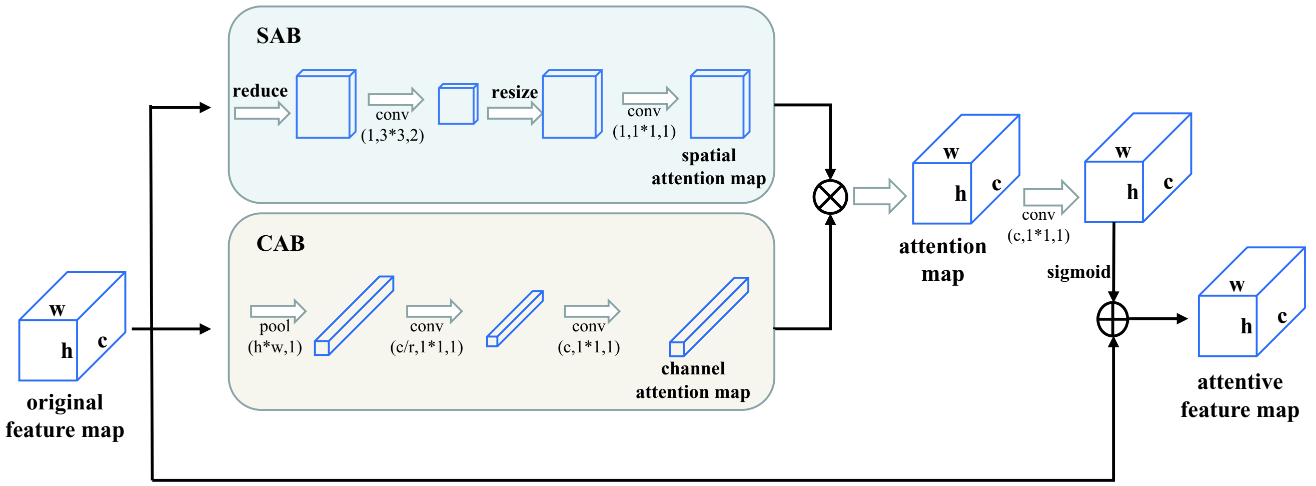


Fig. 3 The architecture of spatial-channel attention mechanism. The original feature maps are firstly fed into 2 branches, the spatial attention branch (SAB) to select discriminative pixels and the channel attention branch (CAB) to select important channels. For better combining generated feature maps, a convolution layer is followed after tensor product. After normalizing the feature maps by sigmoid function, we fuse the weights of final attention maps with the original feature maps to achieve the final attentive feature maps

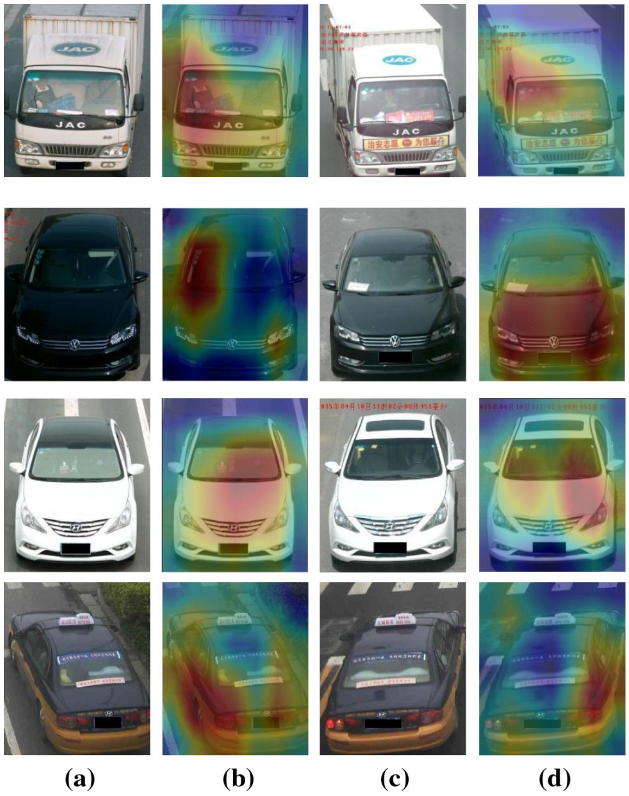


Fig. 4 Visualization of our spatial attention in vehicle Re-ID. From left to right, **a** Query images, **b** Attention maps for (**a**), **c** Vehicle images in gallery with the same type and color but with different ID from the query image, **d** Attention maps for (**c**)

recommended parameters with an initial learning rate of 0.0001 and a decay of 0.96 every epoch. With more passes over the training data, the model improves until it converges. In this paper, we set the number of scales $N = 3$ and experimentally set the size of each scale as $(7 \times 7, 14 \times 14, 28 \times 28)$. The reason why we only evaluated on three scales is that more scales will introduce higher computational complexity without significant improvement in accuracy. Three scales is an appropriate trade-off between accuracy and efficiency.

Algorithm 1 The training process of MSA

```
Input: Vehicle Re-ID training data I, Identity labels Y
Output: $F_{final}$
$n = 1$
while $n <= N$ do
    minimize $L_{id}^{n}$
    update $\theta_n$
    $n = n + 1$
end while
minimize $L_{id}$
update $\theta_0$ (parameters of backbone), $\theta_1, \theta_2, \cdots, \theta_N$
Return $F_{final}$
```

4 Experiments

We evaluate the effectiveness of the proposed multi-scale attention framework on three vehicle Re-ID datasets VeRi-776 [26], VehicleID [23] and PKU-VD [12], compared with twelve state-of-the-art methods.

4.1 Evaluation settings

4.1.1 Datasets

VeRi-776 [26] is collected by 20 non-overlapping traffic surveillance cameras, which contains 51,035 images with corresponding type and color labels of 776 vehicles. Specifically, the number of training, gallery and query sets are collected as 37,778 images of 576 vehicles, 11,579 images of 200 vehicles and 1678 images of 200 vehicles, respectively.

VehicleID [23] is a larger vehicle Re-ID dataset from real-world traffic surveillance environment, which contains 221,763 images of 26,267 vehicles in total. It contains training set with 110,178 images of 13,134 vehicles and testing set with 111,585 images of 13,133 vehicles. Following the protocols in [23], we utilize three different size testing sets for evaluation, which contains 800, 1600 and 2,400 vehicles, respectively.

PKU-VD [12] is a large-scale vehicle Re-ID dataset collected from high-resolution traffic cameras. It contains two subsets VD1 and VD2, which were captured from high-resolution traffic cameras and low-resolution surveillance videos, respectively. The VD1 and VD2 subsets initially contain 1,097,649 and 807,260 images, respectively. After removing the vehicle images only containing one viewpoint, the dataset has been split into 422,326 images of 70,591 vehicles in VD1 for training, while 424,032 images of 71,165 vehicles in VD1 for testing. In the same manner, there are 342,608 images of 39,619 vehicles in VD2 for training while 347,910 images of 40,144 vehicles for testing. Both subsets consist of three testing sets, e.g., small, medium and large, with 106,887, 604,032 and 1,097,649 samples, respectively, for VD1 while 105,550, 457,910 and 807,260 testing samples, respectively, for VD2.

4.1.2 Evaluation metric

Following the evaluation protocol of re-identification work [27, 33], we utilize mAP and Rank-n as the evaluation metrics, where mAP represents the mean average precision, Rank-n indicates the expected correct matching pair in the top n matches, which are the two commonly used metrics in Re-ID task.

4.1.3 Compared state-of-the-art methods

We briefly describe the twelve compared state-of-the-art methods as follows:

(1) *LOMO* [20] Local Maximal Occurrence Representation (LOMO) is a novel local feature extractor, which is proposed for solving the problem of viewpoint changes.

(2) *BOW-CN* [49] Bag-of-Word with Color Names (BOW-CN) descriptor is a hand-crafted feature for vehicle Re-ID. It considers the local cues and can speed up the global feature matching during Re-ID.

(3) *GoogLeNet* [40] GoogLeNet is an deep neural network architecture to learn the vehicle features. It is pre-trained on ImageNet [14] and then fine-tuned on the CompCars [41] dataset.

(4) *FACT* [26] Fusion of Attributes and Color feaTures (FACT) is a discriminative feature learning network which fuses color, texture and semantic information.

(5) *FACT+Plate-SNN+STR* [27] FACT [26] based on Plate Siamese Neural Network and SpatioTemporal Relations (FACT+Plate-SNN+STR) is a vehicle Re-ID scheme, which adds plate information and spatio-temporal relations to achieve discriminative features.

(6) *NuFACT* [28] Null space base Fusion of Attribute and Color feaTures (NuFACT), which fuses the appearance features and the attribute features to extract effective and robust representations.

(7) *Siamese-Visual* [33] Siamese-CNN with pairwise visual branch is an novel deep learning architecture to generate the similarity of query vehicle pairs.

(8) *Siamese+Path-LSTM* [33] Siamese-CNN together with Path LSTM, which exploits additional spatio-temporal path information in Siamese-CNN.

(9) *VAMI* [51] Viewpoint-aware Attentive Multi-view Inference (VAMI) method exploits multi-viewpoint information of vehicle from single-viewpoint feature to extract discriminative features.

(10) *C2F-Rank* [24] Coarse-to-Fine Ranking (C2F-Rank) Loss is deigned to explore structured feature embedding which can improve the performance of vehicle Re-ID.

(11) *CLVR* [11] Cross-Level Vehicle Recognition (CLVR) is an novel method, which combines the structured information of vehicle identity and vehicle model classification.

(12) *VRSDNet* [52] Shortly and Densely convolutional neural Network (VRSDNet) exploits the short and dense connection mechanism in a siamese network to learn the vehicle features.

4.2 Evaluation on benchmarks

4.2.1 Evaluation on the VeRi-776 dataset

The results of the proposed method on VeRi-776 dataset [26] comparing with the state-of-the-art methods are reported in Table 1. Our method outperforms all the other methods, by improving mAP and Rank-1 about 5% and 9%, respectively, compared with the second best method Siamese+Path-LSTM [33]. Note that four of the compared methods in Table 1 have used the auxiliary information of the vehicles except for the appearance. e.g., FACT+Plate-SNN+STR [27], NuFACT [28], Siamese+Path-LSTM

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

Method	mAP	Rank-1	Rank-5	Reference
(1) LOMO [20]	9.64	25.33	46.48	CVPR2015
(2) BOW-CN [49]	12.20	33.91	53.69	ICCV2015
(3) GoogLeNet [40]	17.89	52.32	72.17	CVPR2015
(4) FACT [26]	18.49	50.95	73.48	ICME2016
(5) FACT+Plate-SNN+STR [27]	27.70	61.44	78.78	ECCV2016
(6) NuFACT [28]	48.47	76.76	***91.42***	TMM2018
(7) Siamese-Visual [33]	29.48	41.12	60.31	ICCV2017
(8) Siamese+Path-LSTM [33]	*58.27*	*83.49*	90.04	ICCV2017
(9) VAMI [51]	50.13	77.03	90.82	CVPR2018
(12) VRSDNet [52]	***53.45***	*83.49*	*92.55*	ICPR2018
MSA	**62.89**	**92.07**	**96.19**	Ours

The top three results are highlighted in bold, italic and bold italic, respectively

[33] and VAMI [51] used the auxiliary plate information, attribute information (color and type), spatio-temporal path information and viewpoint information, respectively. Our method is developed only on the vehicle appearance and still beats these methods with auxiliary information, which verifies the promising effectiveness of our method.

Figure 5 shows five examples of matching results on the VeRi-776 [26], where the left column indicates the query images, and the following ten columns are the corresponding top-10 hits obtained by our multi-scale attention framework. From which we can see that our MSA framework can hit most correct matchings in the top-10 rankings. The false matching, such as the Rank-8 of the first query, has almost the same appearance to the query, which is challenging case for all existing vehicle Re-ID methods. The Rank-7 for the second query in Fig. 5 appears with different views to the query, but contains similar local information such as the green strips along the bottom of the trucks highlighted in yellow. Although it is a false matching, it is still meaningful with the high probability of being a right hit from human perception.

4.2.2 Evaluation on the VehicleID dataset

The results of the proposed method on VehicleID dataset [23] comparing with the state-of-the-art methods are reported in Table 2. Generally speaking, all the methods perform better on the small size testing set since larger

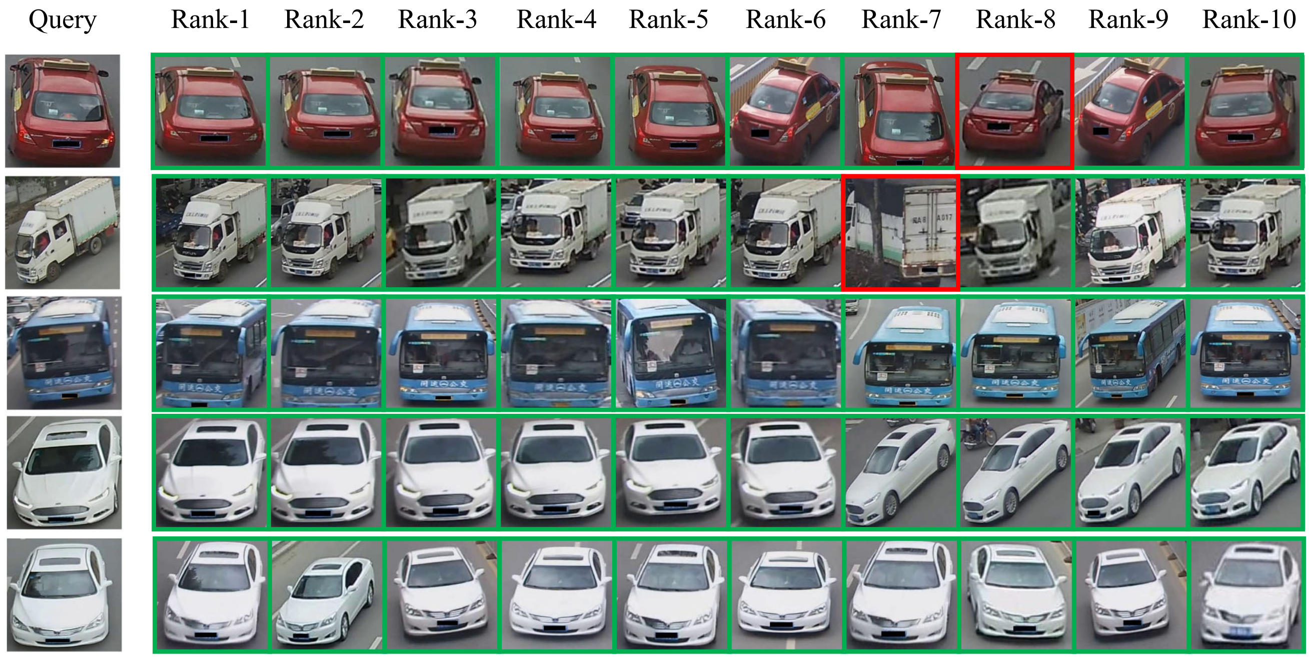


Fig. 5 Example of our method (MSA) on VeRi-776 dataset. The green and red boxes indicate the right hits and the wrong hits, respectively

Table 2 The mAP, Rank-1 and Rank-5 comparison on VehicleID dataset (in %)

Test size	800			1600			2400			Reference
Method	mAP	Rank-1	Rank-5	mAP	Rank-1	Rank-5	mAP	Rank-1	Rank-5	
(1) LOMO [20]	–	19.76	32.01	–	18.85	29.18	–	15.32	25.29	CVPR2015
(2) BOW-CN [49]	–	13.14	22.69	–	12.94	21.09	–	10.20	17.89	ICCV2015
(3) GoogLeNet [40]	46.20	47.88	67.18	44.00	43.40	63.86	38.10	38.27	59.39	CVPR2015
(4) FACT [26]	–	49.53	68.07	–	44.59	64.57	–	39.92	60.32	ICME2016
(6) NuFACT [28]	–	48.90	69.51	–	43.64	65.34	–	38.63	60.72	TMM2018
(9) VAMI [51]	–	*63.12*	***83.25***	–	52.87	75.12	–	47.34	70.29	CVPR2018
(10) C2F-Rank [24]	***63.50***	61.10	81.70	*60.00*	*56.20*	***76.20***	*53.00*	*51.40*	***72.20***	AAAI2018
(11) CLVR [11]	–	***62.00***	76.00	–	***56.10***	71.80	–	***50.60***	68.00	BMVC2017
(12) VRSDNet [52]	*63.52*	56.98	86.90	***57.07***	50.57	*80.05*	***49.68***	42.92	*73.44*	ICPR2018
MSA	**80.31**	**77.55**	*90.50*	**77.11**	**74.41**	**86.26**	**75.55**	**72.91**	**84.35**	Ours

The top three results are highlighted in bold, italic and bold italic, respectively

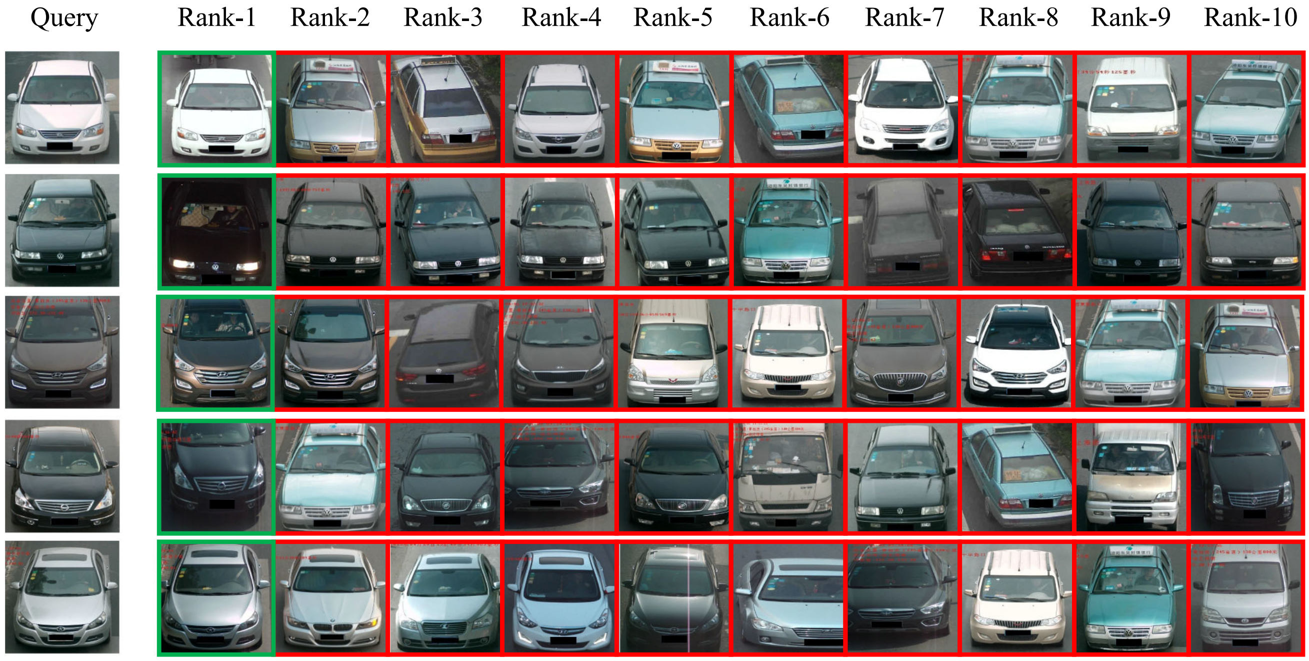


Fig. 6 Example of our method (MSA) on VehicleID dataset (800 Test size). The green and red boxes indicate the right hits and the wrong hits, respectively

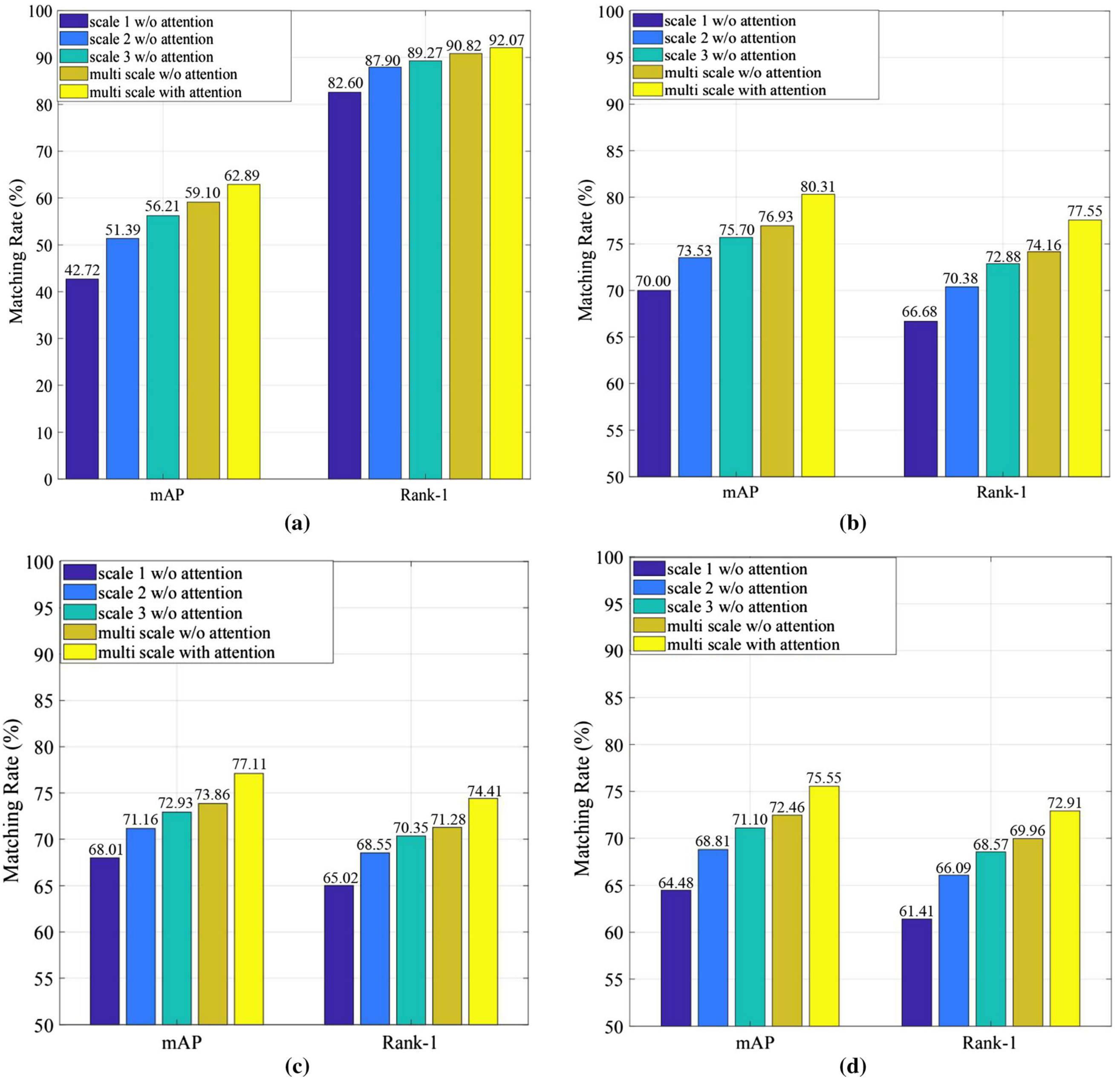
testing sets introduce more challenging and complex scenarios. Our MSA outperforms all other methods in all testing sets by a large margin. It improves about 20% in both mAP and Rank-1 assurances on all three testing sets, comparing with the second best methods achieved by VAMI [51], VRSDNet [52], CLVR [11].

There are some examples of matching result on the VehicleID dataset (800 Test size) [23] shown in Fig. 6, where the left column is query images, and others are the corresponding top-10 result obtained by our multi-scale attention framework. We can see that our MSA can hit the correct matching at the early ranks (all Rank-1 in Fig. 6).

Table 3 The mAP (in %) of different variants of our method (MSA) on PKU-VD dataset

Dataset	Component	Small	Medium	Large
VD1	Scale 1 (7 × 7)	82.74	71.50	68.90
	Scale 2 (14 × 14)	84.91	74.62	72.45
	Scale 3 (28 × 28)	85.55	75.96	74.65
	MSA	**86.46**	**79.60**	**76.59**
VD2	Scale 1 (7 × 7)	78.99	62.91	57.32
	Scale 2 (14 × 14)	81.29	67.55	61.04
	Scale 3 (28 × 28)	82.63	69.53	63.25
	MSA	**83.75**	**71.84**	**64.95**

The best results are highlighted in bold



(a)

(b)

(c)

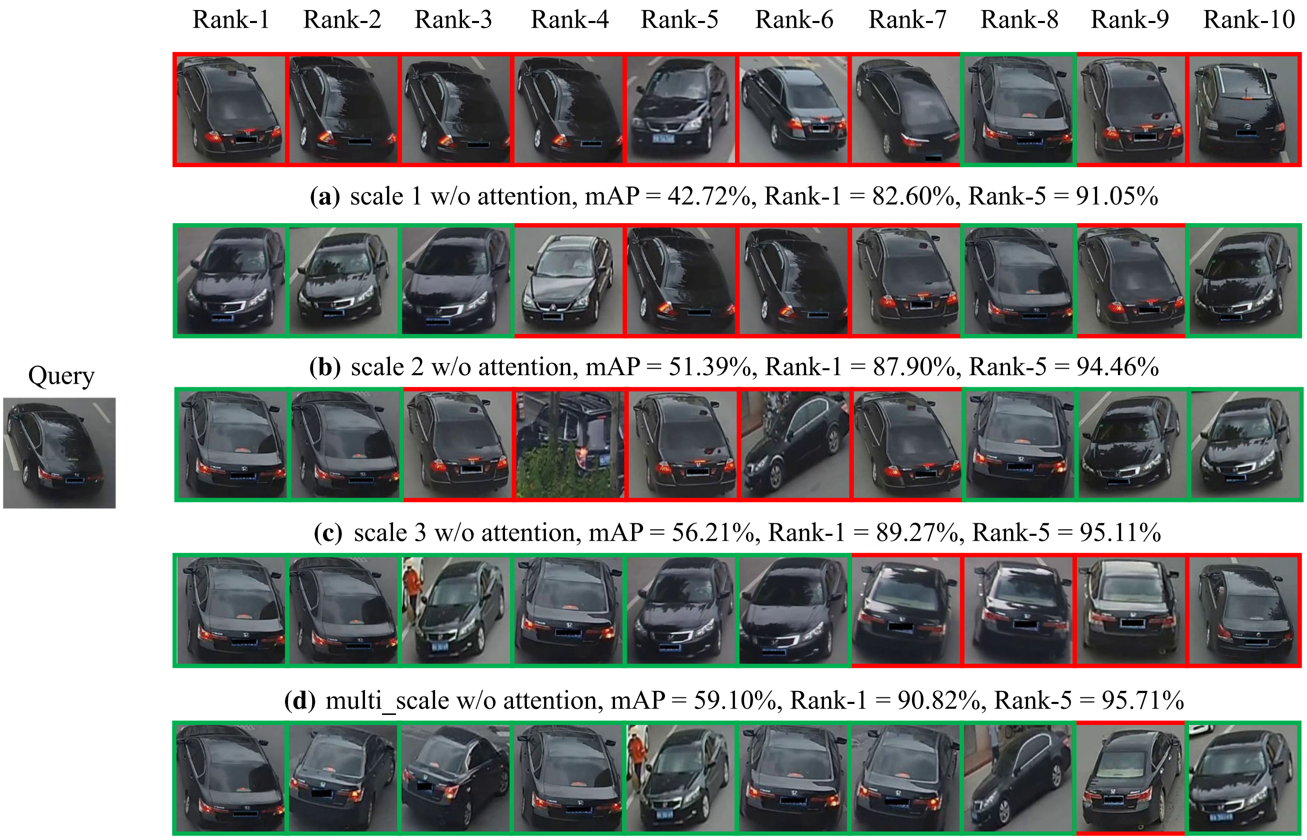
(d)

Fig. 7 Evaluation of The Proposed Framework on the benchmark dataset (in %). **a** is the result in VeRi-776 dataset, **b–d** is the result in VehicleID dataset test size 800, 1600, 2400, respectively

Note that there is only one ground truth (correct matching) vehicle image in gallery in the VehicleID dataset. Some of the other false hits are with high similarity to the query, such as Rank-4 and Rank-7 for the first query, Rank-2 to Rank-5 for the second query. It is worth mentioning that there is only one ground truth image for each query in the gallery. And the reason that Rank-7 for the first query is quite similar to query than the Rank-2 may be, our method has paid higher attention on the local difference, such as the color of the auto logos, or the position distribution of the logos and the vents of the vehicles between Rank-7 to the query, which may lead to higher distance to the query.

4.2.3 Evaluation on the PKU-VD dataset

We evaluate three variants in different scales comparing with our method on all the three test settings on both VD1 and VD2 subsets to verify the effectiveness of our framework. Table 3 reports the evaluation results. Generally speaking, the larger scale tends to outperform the smaller



(a) scale 1 w/o attention, mAP = 42.72%, Rank-1 = 82.60%, Rank-5 = 91.05%

(b) scale 2 w/o attention, mAP = 51.39%, Rank-1 = 87.90%, Rank-5 = 94.46%

(c) scale 3 w/o attention, mAP = 56.21%, Rank-1 = 89.27%, Rank-5 = 95.11%

(d) multi_scale w/o attention, mAP = 59.10%, Rank-1 = 90.82%, Rank-5 = 95.71%

(e) multi_scale with attention, mAP = 62.89%, Rank-1 = 92.07%, Rank-5 = 96.19%

Fig. 8 Example of different variants of our method (MSA) on VeRi-776 dataset. The green and red boxes indicate the right hits and the wrong hits, respectively

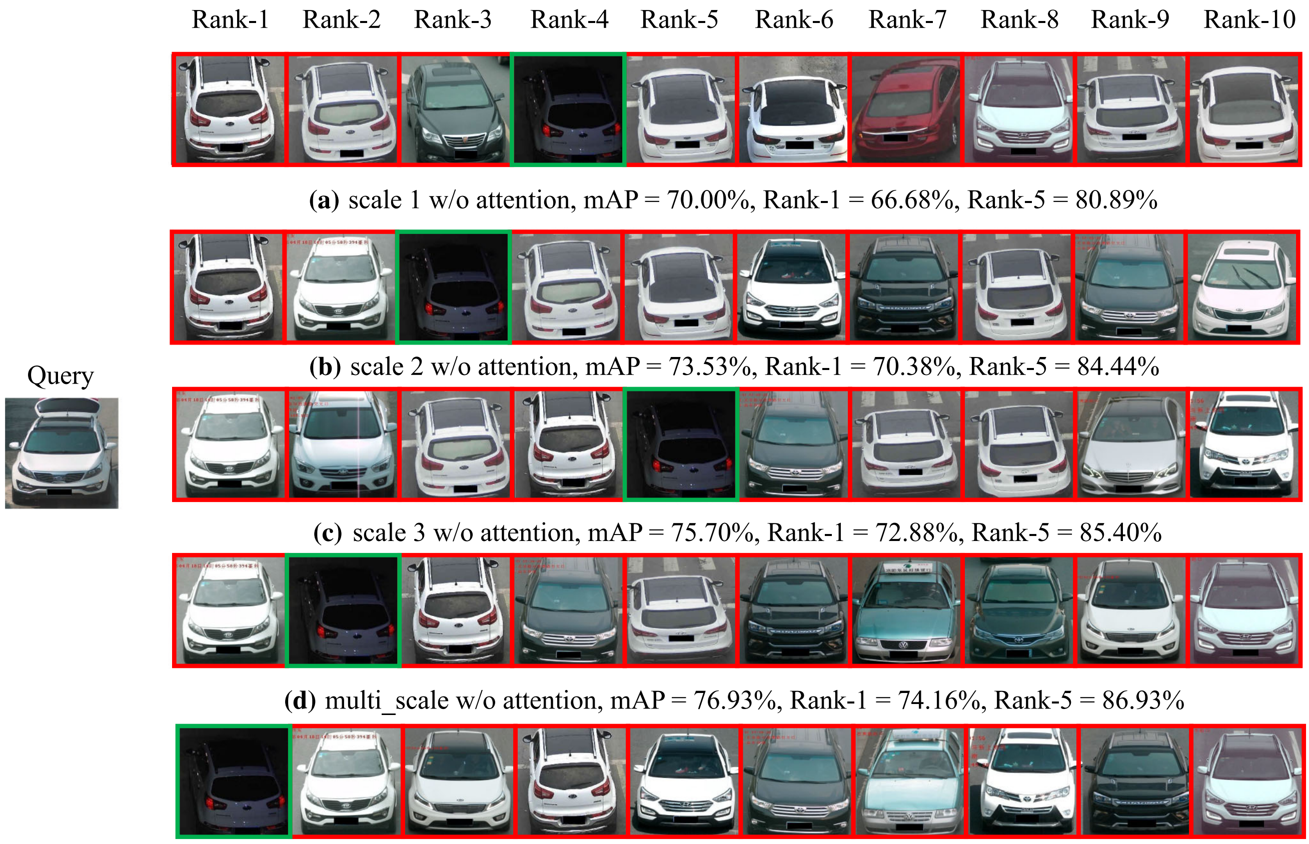
scales, this may be because the larger scale contains both global information and local information. The multi-scale integrating mechanism surpasses all the three single-scale case, which promises the effectiveness of our MSA. Note that, PKU-VD [12] is much larger and more challenging compared to VeRi-776 [26] and VehicleID [23]. Our method still achieves promising performance, which implies the robustness of our method on large-scale real-life applications.

4.3 Ablation study

In order to evaluate the components of the proposed multi-scale attention framework, we first study several variants of our method by ablating the attention mechanism or changing the number of scales in our framework. Then we further study on each spatial and channel attention branches in the spatial-channel attention mechanism.

Study on multi-scale and spatial-channel attention mechanisms Figure 7 reports the results of this ablation study on the proposed multi-scale and spatial-channel attention mechanisms, where Fig. 7a demonstrates the evaluation result on VeRi-776 [26] dataset and Fig. 7b–d corresponds to the evaluation result on VehicleID [23] dataset with 800, 1600, 2400 testing size, respectively. From Fig. 7 we can see that (1) the larger scale tends to outperform the smaller scales, this may be because larger scale contain both global information and local information. (2) The multi-scale performs superior by integrating both the global and local cues from different scales, which verifies the effectiveness of the multi-scale integration of the proposed method. (3) By introducing the attention mechanism into each variant, it can further boost the performance of corresponding tasks, which verifies the effectiveness of the attention technique in our method.

Figures 8 and 9 demonstrate an example of the matching results of different variants of our framework on VeRi-776 [26] and VehicleID [23] (800 Test size), respectively, where Fig. 8a–c demonstrates the ranking results on VeRi-776 [26] without attention technique on scale 1, scale 2 and scale 3, respectively. Figure 8d indicates the ranking results of by exploring the multi-scale mechanism. Figure 8e shows the ranking results of our MSA which further introduces the attention technique into the multi-scale



(a) scale 1 w/o attention, mAP = 70.00%, Rank-1 = 66.68%, Rank-5 = 80.89%

(b) scale 2 w/o attention, mAP = 73.53%, Rank-1 = 70.38%, Rank-5 = 84.44%

(c) scale 3 w/o attention, mAP = 75.70%, Rank-1 = 72.88%, Rank-5 = 85.40%

(d) multi_scale w/o attention, mAP = 76.93%, Rank-1 = 74.16%, Rank-5 = 86.93%

(e) multi_scale with attention, mAP = 80.31%, Rank-1 = 77.55%, Rank-5 = 90.50%

Fig. 9 Example of different variants of our method (MSA) on VehicleID dataset (800 Test size). The green and red boxes indicate the right hits and the wrong hits, respectively

Table 4 Evaluation on Channel Attention Branch (CAB) and Spatial Attention Branch (SAB) in Spatial-Channel Attention (SCA) model in scale of (7 × 7) on VeRi-776 dataset (in %)

Method	mAP	Rank-1
Baseline (w/o SAB or CAB)	42.72	82.60
+ CAB	48.15	84.45
+ SAB	51.22	86.53
+ CAB + SCA (Ours)	**54.54**	**88.68**

The best results are highlighted in bold

framework. Figure 9 is organized in the same manner as Fig. 8 but on VehicleID [23] dataset.

From Figs. 8 and 9 we can observe that (1) when the appearance of gallery image is very similar to the query image, it is hard to hit the correct matching only with one single scale, as shown in (a)–(c) in both in Fig. 8 and in Fig. 9. (2) By fusing local cues and global cues from multiple scales, MSA can hit the correct matchings even with local appearance difference on the images, such as the third correct matching in Fig. 8d. (3) After enforcing the attention mechanism, MSA can hit more correct matchings in the top-10 rankings as shown in Fig. 8e or shift forward the ground truth as shown in Fig. 9e.

Study on spatial-channel attention We further evaluated the effect of each attention branch in spatial-channel attention model (SCA): spatial attention branch (SAB) and channel attention branch (CAB). Table 4 reports the results of this experiment, from which we can see, (1) by introducing spatial attention branch or channel attention branch into the baseline, both achieve satisfactory improvement, which verifies the contribution of both SAB and CAB. (2) SAB plays more important role than CAB by improve more performance on both mAP and Rank-1 accuracies, which in turn means it is more important to select discriminative regions of the vehicle image than important channel of feature maps. (3) Cooperating both SAB and CAB further boosts the Re-ID performance, which verifies the contribution of both spatial and channel attention branches.

4.4 Parameter analysis

Analysis on the number of scale N As one of the key parameters in our method, we first analyze the effect of parameter N in our framework. As indicated in Table 5, we evaluate our MSA on five different scales, together with

Table 5 Parameter analysis of the number of scales (N) of MSA without spatial-channel attention on VeRi-776 dataset (in %)

Method	mAP	Rank-1
Scale 1 (7×7)	42.72	82.60
Scale 2 (14×14)	51.39	87.90
Scale 3 (28×28)	56.21	89.27
Scale 4 (56×56)	57.67	89.21
Scales 1 to 3	59.10	90.82
Scales 1 to 4	**59.68**	**90.84**

The best results are highlighted in bold

Table 6 Evaluation of various size combinations of three scales on MSA without spatial-channel attention on VeRi-776 dataset (in %)

Method	mAP	Rank-1
Scale 1 & 2 & 3	59.10	**90.82**
Scale 1 & 3 & 4	**59.23**	90.41

The best results are highlighted in bold

three different combinations of multi-scale cases. We observe that (1) multi-scale cases consistently outperform the single-scale cases, which verifies the contribution of the multi-scale mechanism. (2) Introducing more scales does not achieve significant improvement, however will obviously bring higher computational complexity. Therefore, we fix $N = 3$ to keep the balance between accuracy and efficiency.

Analysis on the size of each scale To analyze the impact of the size of each scale in our three-scale MSA model, we further evaluate our method with various combinations of sizes. As shown in Table 6, our method is insensitive to the size of each scale. We set the sizes of each scale as $(7 \times 7, 14 \times 14, 28 \times 28, 56 \times 56)$ in this paper.

4.5 Limitations and future improvements

Although our model achieves a new state of the art for vehicle Re-ID, it still faces limitations due to the challenging scenarios in real-world surveillance. First, some important parameters, including the number of scales and the size of each scale, are set empirically. More intelligent scheme, such as network architecture searching technique, can be developed for possible improvement in the future. Second, despite the benefits from scales and spatial or channel attention, semantic feature and view information also play important roles in vehicle Re-ID, which can further boost the robustness of the performance for future application.

5 Conclusion

In this work, we propose a multi-scale attention framework (MSA) to fusing the discriminative local cues and effective global information. Specifically, we exploit bilinear interpolation technique on the backbone of network, to generate different scale feature maps containing local cues and global information. For further mining discriminative information of vehicles, we add multiple attention block on each subnetwork. Finally, we fuse complementary feature maps to acquire more discriminative vehicle features. For all we know, we are the first to introduce multi-scale mechanism into vehicle Re-ID task and the first to combine multi-scale and attention block in one convolutional network. Experimental results on benchmark datasets VeRi-776, VehicleID and PKU-VD demonstrate the framework we proposed is so far the most effective way to solve the vehicle Re-ID problem.

Acknowledgements This work was supported by the National Natural Science Foundation of China (61976002, 61860206004), the Natural Science Foundation of Anhui Higher Education Institutions of China (KJ2019A0033), and the Open Project Program of the National Laboratory of Pattern Recognition (NLPR) (201900046).

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Chen K, Bui T, Fang C, Wang Z, Nevatia R (2017) Amc: Attention guided multi-modal correlation learning for image search. arXiv preprint arXiv:1704.00763
2. Chen Y, Zhu X, Gong S (2017) Person re-identification by deep learning multi-scale representations. In: IEEE international conference on computer vision, pp 2590–2600
3. Fei G, Teng H, Sun J, Wang J, Hussain A, Yang E (2018) A new algorithm of sar image target recognition based on improved deep convolutional neural network. Cognit Comput 11(6):809–824
4. Fu XQY, Jiang YG, Xue TXX (2017) Multi-scale deep learning architectures for person re-identification. In: IEEE international conference on computer vision, pp 1–2
5. Gao F, Ma F, Wang J, Sun J, Zhou H (2017) Visual saliency modeling for river detection in high-resolution SAR imagery. IEEE Access 6:1000–1014
6. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: IEEE conference on computer vision and pattern recognition, pp 770–778
7. Hu J, Shen L, Sun G (2018) Squeeze-and-excitation networks. In: IEEE conference on computer vision and pattern recognition, pp 7132–7141
8. Jaderberg M, Simonyan K, Zisserman A et al (2015) Spatial transformer networks. In: Advances in neural information processing systems, pp 2017–2025
9. Ji Y, Zhang H, Wu QJ (2018) Salient object detection via multi-scale attention CNN. Neurocomputing 322:130–140

10. Jiang B, Zhang Z, Lin D, Tang J, Luo B (2019) Semi-supervised learning with graph learning-convolutional networks. In: IEEE conference on computer vision and pattern recognition, pp. 11313–11320
11. Kanacı A, Zhu X, Gong S (2017) Vehicle reidentification by fine-grained cross-level deep learning. In: British machine vision conference workshop, pp 1–6
12. Ke Y, Tian Y, Wang Y, Wei Z, Huang T (2017) Exploiting multi-grain ranking constraints for precisely searching visually-similar vehicles. In: IEEE international conference on computer vision, pp 1–1
13. Kim TH, Eom IK, Kim YS (2009) Multiscale bayesian texture segmentation using neural networks and Markov random fields. Neural Comput Appl 18(2):141–155
14. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, pp 1097–1105
15. Li D, Chen X, Zhang Z, Huang K (2017) Learning deep context-aware features over body and latent parts for person re-identification. In: IEEE conference on computer vision and pattern recognition, pp 384–393
16. Li W, Zhu X, Gong S (2018) Harmonious attention network for person re-identification. In: IEEE conference on computer vision and pattern recognition, p 2
17. Li X, Wu A, Zheng WS (2018) Adversarial open-world person re-identification. arXiv preprint arXiv:1807.10482
18. Li X, Zheng WS, Wang X, Xiang T, Gong S (2015) Multi-scale learning for low-resolution person re-identification. In: ieee international conference on computer vision, pp 3765–3773
19. Li Y, Li Y, Yan H, Liu J (2017) Deep joint discriminative learning for vehicle re-identification and retrieval. In: IEEE international conference on image processing, pp 395–399
20. Liao S, Hu Y, Zhu X, Li SZ (2015) Person re-identification by local maximal occurrence representation and metric learning. In: IEEE conference on computer vision and pattern recognition, pp 2197–2206
21. Lin B, Wang F, Zhao F, Sun Y (2018) Scale invariant point feature (sipf) for 3d point clouds and 3d multi-scale object detection. Neural Comput Appl 29(5):1209–1224
22. Liu H, Feng J, Qi M, Jiang J, Yan S (2017) End-to-end comparative attention networks for person re-identification. IEEE Trans Image Process 26(7):3492–3506
23. Liu H, Tian Y, Yang Y, Pang L, Huang T (2016) Deep relative distance learning: tell the difference between similar vehicles. In: IEEE conference on computer vision and pattern recognition, pp 2167–2175
24. Liu HGCZZ, Lu JWH (2018) Learning coarse-to-fine structured feature embedding for vehicle re-identification. In: Association for the advancement of artificial intelligence, pp 1–8
25. Liu J, Zha ZJ, Tian Q, Liu D, Yao T, Ling Q, Mei T (2016) Multi-scale triplet CNN for person re-identification. In: the ACM on multimedia conference, pp 192–196
26. Liu X, Liu W, Ma H, Fu H (2016) Large-scale vehicle re-identification in urban surveillance videos. In: IEEE international conference on multimedia and expo, pp 1–6
27. Liu X, Liu W, Mei T, Ma H (2016) A deep learning-based approach to progressive vehicle re-identification for urban surveillance. In: European conference on computer vision, pp 869–884
28. Liu X, Liu W, Mei T, Ma H (2018) Provid: progressive and multimodal vehicle reidentification for large-scale urban surveillance. IEEE Trans Multimed 20(3):645–658
29. Liu X, Zhao H, Tian M, Sheng L, Shao J, Yi S, Yan J, Wang X (2017) Hydraplus-net: Attentive deep features for pedestrian analysis. arXiv preprint arXiv:1709.09930
30. Liu Z, Song X, Tang Z (2015) Fusing hierarchical multi-scale local binary patterns and virtual mirror samples to perform face recognition. Neural Comput Appl 26(8):2013–2026
31. Mnih V, Heess N, Graves A et al (2014) Recurrent models of visual attention. In: Advances in neural information processing systems, pp 2204–2212
32. Sermanet P, Frome A, Real E (2014) Attention for fine-grained categorization. arXiv preprint arXiv:1412.7054
33. 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: IEEE international conference on computer vision, pp 1918–1927
34. Song C, Huang Y, Ouyang W, Wang L (2018) Mask-guided contrastive attention model for person re-identification. In: IEEE conference on computer vision and pattern recognition, pp 1179–1188
35. Su C, Li J, Zhang S, Xing J, Gao W, Tian Q (2017) Pose-driven deep convolutional model for person re-identification. In: IEEE international conference on computer vision, pp 3980–3989
36. Tang Z, Naphade M, Liu MY, Yang X, Birchfield S, Wang S, Kumar R, Anastasiu D, Hwang JN (2019) Cityflow: a city-scale benchmark for multi-target multi-camera vehicle tracking and re-identification. arXiv preprint arXiv:1903.09254
37. Teng S, Liu X, Zhang S, Huang Q (2018) Scan: spatial and channel attention network for vehicle re-identification. In: Pacific Rim conference on multimedia, pp 350–361
38. Wang B, Cao G, Shang Y, Zhou L, Zhang Y, Li X (2020) Single-column cnn for crowd counting with pixel-wise attention mechanism. Neural Comput Appl 32:2897–2908
39. Wang Z, Du L, Wang F, Su H, Zhou Y (2015) Multi-scale target detection in SAR image based on visual attention model. In: Synthetic Aperture Radar, pp 704–709
40. Yang L, Luo P, Change Loy C, Tang X (2015) A large-scale car dataset for fine-grained categorization and verification. In: IEEE conference on computer vision and pattern recognition, pp 3973–3981
41. Yang L, Luo P, Chen CL, Tang X (2015) A large-scale car dataset for fine-grained categorization and verification. In: IEEE conference on computer vision and pattern recognition, pp 3973–3981
42. Yang Z, He X, Gao J, Deng L, Smola A Stacked attention networks for image question answering. In: IEEE conference on computer vision and pattern recognition, pp 1–2
43. Yue Z, Gao F, Xiong Q, Wang J, Huang T, Yang E, Zhou H (2019) A novel semi-supervised convolutional neural network method for synthetic aperture radar image recognition. Cognit Comput. https://doi.org/10.1007/s12559-019-09639-x
44. Zapletal D, Herout A (2016) Vehicle re-identification for automatic video traffic surveillance. In: IEEE conference on computer vision and pattern recognition workshops, pp 25–31
45. Zhang H, Ji Y, Huang W, Liu L (2019) Sitcom-star-based clothing retrieval for video advertising: a deep learning framework. Neural Comput Appl 31(11):7361–7380
46. Zhang J, Du J, Dai L (2018) Multi-scale attention with dense encoder for handwritten mathematical expression recognition. In: International conference on pattern recognition (ICPR), pp 2245–2250
47. Zhang Y, Liu D, Zha ZJ (2017) Improving triplet-wise training of convolutional neural network for vehicle re-identification. In: ieee international conference on multimedia and expo, pp 1386–1391
48. Zhao L, Li X, Zhuang Y, Wang J (2017) Deeply-learned part-aligned representations for person re-identification. In: IEEE international conference on computer vision, pp 3239–3248

49. Zheng L, Shen L, Tian L, Wang S, Wang J, Tian Q (2015) Scalable person re-identification: a benchmark. In: IEEE international conference on computer vision, pp 1116–1124
50. Zheng L, Yang Y, Hauptmann AG (2016) Person re-identification: past, present and future. arXiv preprint arXiv:1610.02984
51. Zhou Y, Shao L (2018) Viewpoint-aware attentive multi-view inference for vehicle re-identification. In: IEEE conference on computer vision and pattern recognition, pp 6489–6498
52. Zhu J, Du Y, Hu Y, Zheng L, Cai C (2018) Vrsdnet: vehicle re-identification with a shortly and densely connected convolutional neural network. Multimed Tools Appl 78(20):29043–29057
53. Zhu X, Jing XY, Ma F, Cheng L, Ren Y (2018) Simultaneous visual-appearance-level and spatial-temporal-level dictionary learning for video-based person re-identification. Neural Comput Appl 31(11):7303–7315

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