Eliminating cross-camera bias for vehicle re-identification

Jinjia Peng¹ · Guangqi Jiang¹ · Dongyan Chen¹ · Tongtong Zhao¹ · Huibing Wang¹ · Xianping Fu¹,²

Received: 20 December 2019 / Revised: 14 July 2020 / Accepted: 24 September 2020 / Published online: 27 October 2020 © Springer Science+Business Media, LLC, part of Springer Nature 2020

Abstract
Vehicle re-identification (reID) often requires to recognize a target vehicle in large datasets captured from multi-cameras. It plays an important role in the automatic analysis of the increasing urban surveillance videos, which has become a hot topic in recent years. However, the appearance of vehicle images is easily affected by the environment that various illuminations, different backgrounds and viewpoints, which leads to the large bias between different cameras. To address this problem, this paper proposes a cross-camera adaptation framework (CCA), which smooths the bias by exploiting the common space between cameras for all samples. CCA first transfers images from multi-cameras into one camera to reduce the impact of the illumination and resolution, which generates the samples with the similar distribution. Then, to eliminate the influence of background and focus on the valuable parts, we propose an attention alignment network (AANet) to learn powerful features for vehicle reID. Specially, in AANet, the spatial transfer network with attention module is introduced to locate a series of the most discriminative regions with high-attention weights and suppress the background. Moreover, comprehensive experimental results have demonstrated that our proposed CCA can achieve excellent performances on benchmark datasets VehicleID and VeRi-776.

Keywords Cross-camera · Attention alignment · Vehicle re-identification

1 Introduction
The related research to vehicles has attracted wide attention and made some progress in the field of computer vision, such as vehicle detection [4, 10], tracking [5, 26] and classification [22, 29]. Different from the tasks above, the purpose of vehicle reID is to accurately
match the target vehicle captured from multiple non-overlapping cameras, which is of great significance to intelligent transportation. Meanwhile, the large amount of videos or images could be processed automatically carried out by vehicle reID to exploit the meaningful information, which plays an important role in modern smart surveillance systems.

With the development of deep learning, lots of excellent deep learning-based methods [6, 8, 11, 46] are proposed for the vehicle reID task. However, there still exist many limitations for the application in the real-world. Different with the person reID [12, 34, 36], fine-grained classification [31, 35, 38] and other methods [30, 33, 37] that could extract rich features from the images with various poses and colors, the vehicles are generally rigid structure with solid colors and appearance is easily affected by various illuminations, viewpoints. Most existing works only focus on learning the discriminative features while neglecting the influence of different cameras. Actually, images captured from different cameras often have obviously different styles. Usually, cameras differ from each other regarding resolution, illumination, background, etc. As Fig. 1 shows, for each row, the images with the same identity have different appearances in different camera views. This could lead to serving cross-camera bias and affect the vehicle reID task. Some vehicle reID researches also noticed the challenges, thus preferred to make use of spatial-temporal information and plate license to achieve promising results. However, the spatial-temporal information is usually not annotated in some datasets. Besides that, the high-resolution images of in front or rear viewpoints are required for license plate recognition, which is not impractical in the real-world scenes.

In order to solve these problems, some methods consider learning the global features from multi-view images, such as VAMI [45] and DHMV [43]. VAMI adopts cross-view generative adversarial network to transform the features into a global multi-view feature representation. DHMV aims to learn transformations across different viewpoints for inferring the multi-view representation from one input vehicle image. There are also some methods...
that exploit the constraint among cross-cameras by proposing the cross-view losses. For instance, MVR loss [15] introduced several latent groups to represent multiple views and ranked them by calculating the intra and inter loss. However, these methods only consider the influence of various viewpoints to solve the problem between different cameras and neglect the background and other factors.

The above issues prompt us to focus on the changes of images caused by different cameras. To solve aforementioned problems, this paper proposes a cross-camera adaptation network (CCA) to smooth the bias between different cameras and learn powerful features. In our paper, the single camera is regarded as an independent domain. CCA aims at transforming the multi-domains into one common domain that has the similar background, illumination and resolution. Different with existing methods, CCA firstly generates vehicle images by StarGAN, which transfers the same vehicle images from other cameras into one camera and doesn’t augment the quantity of original datasets. Besides, it could be observed in the Fig. 1, the images captured from different always have different backgrounds, which may interfere with the training of vehicle reID model. Hence, to eliminate the impact of background, the attention alignment network (AANet) is proposed to locate discriminative features. Specially, the STN with attention module is employed to select a series of regions from vehicle images for training a powerful reID model. The main contributions of our work can be summarized as follows,

- A cross-camera adaptation framework is proposed for better smoothing the bias between different cameras, which reduces the influence of illumination, background and resolution for vehicle reID task by transferring images into a common space and learning a powerful discriminative feature.
- The attention alignment network is proposed to obtain a series of local regions for vehicle reID, which focuses on locating the meaningful parts while suppressing background. Moreover, Extensive experiments demonstrate that our proposed method achieves competitive performance on challenging benchmark datasets.

The rest of this paper is organized as follows. In Section 2, we review and discuss the related works. Section 3 illustrates the proposed method in detail. Experimental results on two vehicle reID datasets are discussed in Section 4, followed by conclusions in Section 5.

2 Related works

In this section, existing vehicle reID works are reviewed. With the prosperity of deep learning, vehicle reID has achieved some progress in recent years.

A series of methods attempt to identify vehicles based on the visual appearance. In [39], 3D bounding boxes of vehicles were detected and then were processed by the color histograms and histograms of oriented gradients for vehicle reID. In [41], a ROIs-based vehicle reID method was proposed to detect the ROIs’ as discriminative identifiers. And then encode the structure information of a vehicle for reID task. DHMVI [43] utilized the LSTM bi-directional loop to learn transformations across different viewpoints of vehicles, which could infer all viewpoints’ information from the only one input view. RAM [18] was proposed for vehicle reID task with several branches including region and attribute branches to extract distinct features from several overlapped local regions. MRM [24] introduced a multi-region model to extract features from a series of local regions for learning powerful features for vehicle reID task. EALN [21] was introduced to improve the capability of the reID model by automatically generating hard negative samples in the specified
embedding space to train the reID model. VAMI [45] tried to better optimize the reID model by transforming single-view feature into a global multi-view feature representation through generative adversarial network. CV-GAN [44] was conducted to generate the various viewpoints vehicle images by generative adversarial network for training an adaptive reID model. AGNet [28] exploited the discriminative features by the generated mask from attribute branches, which learns features with attribute cues for vehicle reID task.

Apart from the visual appearance, a series of metric losses for deep feature embedding to achieve higher performance. In [17], coupled cluster loss was proposed to push those negative ones far away and pull the positive images closer, which could minimize maximize inter-class distance and intra-class distance to train the vehicle reID model. GST loss [1] was proposed to deal with intra-class variance in learning representation. Besides that, it introduced the mean-valued triplet loss to alleviate the negative impact of improper triplet sampling during training stage. MGR [7] was presented to further enhanced the discriminative ability of reID model by enhancing the discrimination that not only between different vehicles but also different vehicle models.

Besides, spatio-temporal information is an important cue for vehicle reID task. Hence, some approaches exploit spatial and temporal information for vehicle images to improve vehicle reID performance. PROVID [20] employed visual features, spatial-temporal relations and the information of license plates with a progressive strategy to learn similarity scores between vehicle images. OIFE [32] refined the retrieval results of vehicles by utilizing the log-normal distribution to model the spatio-temporal constrains in camera networks. Siamese-Cnn+Path-LSTM [25] model was proposed to incorporate complex spatio-temporal information for regularizing the reID results.

### 3 Cross-camera adaptation framework

The overall structure of the proposed framework is depicted in Fig. 2. Cross-camera Adaptation Framework (CCA) is composed of the camera transfer adversarial network and the attention alignment network. Firstly, the samples from different cameras are transferred into one domain by the camera transfer adversarial network. Then the images with similar distribution could be obtained, which are fed into the proposed attention alignment network.

![Fig. 2 Cross-camera Adaptation Framework. The proposed framework is composed of the camera transfer adversarial network and attention alignment network. In the camera transfer adversarial network, the samples in different cameras are transferred into the common space, which means the generated samples have the similar distribution. Subsequently, the transferred samples are employed to train the reID model with the attention alignment network.](image-url)
feature learning network for training the reID task. Specially, the attention alignment network focuses on different meaningful parts of vehicle images for improving the discriminate ability of reID model.

In this section, we introduce our method from two aspects: 1) a camera transfer adversarial network is introduced in Section 3.2, which learns transfer mappings for different cameras; 2) an attention alignment feature learning network is illustrated in section 3.3, which optimizes the reID model utilizing the generated images from the camera transfer adversarial network. The Table 1 shows the notations that are frequently utilized in our paper.

3.1 Camera transfer adversarial network

The same vehicles always have different appearances in different camera views and the bias is shown in Fig. 1. In this paper, to smooth the bias between different cameras, we require to transfer images in different cameras into one camera, which means that all images have the similar distribution. To achieve this, StarGAN [3] is utilized as the camera transfer adversarial network. StarGAN [3] utilizes generator $G$ and discriminator $D$ to implement the conversion between multiple cameras, which learns the mapping relations among multiple cameras using only a single model, as shown in Fig. 3.

In StarGAN, in order to generate a more realistic fake sample, an adversarial loss function is employed to obtain the high-quality image, which could be written as:

$$L_{adv} = E_x \left[ \log D_{src}(x) \right] + E_{x,c} \left[ \log(1 - D_{src}(G(x,c))) \right]$$

(1)

where $G$ generates an image $G(x,c)$ to fake $D$. $D$ tries to distinguish the real image from the generated image. The target of StarGAN is to translate $x$ to an output images $y$ that is classified as the target domain $c$. For this goal, a domain classifier of real images is added on the $D$, which could be defined as:

$$L_{dom}^r(x,c^\star) = E_{x,c^\star} \left[ - \log D_{dom}(c^\star|x) \right]$$

(2)

where $D_{dom}(c^\star|x)$ is the probability distribution over the camera labels of a given real image $x$, and $c^\star$ represents the source camera labels.

Similar with the (2), the domain classifier of fake images is defined as :

$$L_{dom}^f(x,c) = E_{x,c} \left[ - \log D_{dom}(c|x) \right]$$

(3)

| Notation | Representation |
|----------|----------------|
| reID     | re-identification |
| CCA      | Cross-camera Adaptation |
| AANet    | Attention Alignment Network |
| $G$      | Generator in StarGAN |
| $D$      | Discriminator in StarGAN |
| STN      | Spatial Transform Network |
| GAP      | Global Average Pooling |
| CMC      | Cumulative Match Characteristic |
| mAP      | mean Average Precision |
To guarantee that generated images could preserve the identity information of original images, StarGAN employs the cycle consistent loss [47], which is defined as:

$$L_{rec}(x, c, c^*) = E_{x,c,c^*} \left[ \| x - G(G(x, c), c^*) \|_1 \right]$$

(4)

Through StarGAN, one image could be transferred into any other cameras. Hence, there are $N$ times images than original dataset. $N$ is the number of cameras. However, in our paper, we aim to transfer images into one common domain. So we just select images from one camera for training vehicle reID model. As illustrated in Fig. 1, the images with irrelevant background or less discriminative parts of objects of interest may confuse the reID model, which would degenerate the model’s performance. To solve this problem, in this paper, the attention alignment network is proposed to utilize the style-translated images as training set to guide the reID model to focus on the discriminative parts, to be detailed in Section 3.2.

3.2 Attention alignment network

Redundancy of background information is another important factor that obstructs vehicle reID performance. We propose the attention alignment network (AANet) to reduce the discrepancy of attention maps across non-overlapping cameras. The AANet is designed to focus on the meaningful parts of vehicle images and neglect the background when training the reID model, which is illustrated in Fig. 4.

The AANet is designed as a multi-branch structure, which is composed of one global branch and two local region branches. As shown in Fig. 4, given an input vehicle image, a set of features are generated by AANet. There are two categories of features generated in our network. For the global branch, it is utilized to learn the context features with the attention module. Besides that, the differences among similar vehicles may lie on some local
regions. So for the local regions, in order to obtain key information from the local region, we divided the output feature map generated by several convolutional layers with the size of 112 × 112 × 64 into two non-overlapping local regions, which could be named as “Upper-Local” and “Lower-Local”, respectively. Then, the feature maps are fed into two branches to generate different features. Specially, the alignment module and attention module are utilized in each sub-network to generate more discriminative features. Hence given an input vehicle image, the local region branch and global branch could generate a series of features for vehicle reID.

To address the problems of excessive background and extract the remarkable features, in each branch, an STN-based alignment module is employed. The alignment module includes three components, a localization network to learn the transformation parameters, a grid generator to calculate the coordinate of the input feature maps by applying the transformation parameters and bilinear sampler to make up the missing pixels. Meanwhile, as shown in Fig. 4, to focus on the meaningful parts of vehicle images and neglect the background when training the feature learning model, an attention module is introduced to generate discriminative features. In the attention module, after a global average pooling layer, we employ the Softmax layer to re-weight the feature maps and generate the mask, which could be computed as:

\[ M = \text{Softmax}(\text{Conv}(\text{GAP}(f_r))) \]  

where the Conv operator is 1 × 1 convolution. The M is the weight matrix. After obtaining M, the attended feature map could be calculated by \( f_m = f_a \otimes M \). The operator \( \otimes \) is performed in an element-wise product. Then the attended feature map \( f_m \) is fed into the subsequent structure.

The structure of global branch is also a two-branch network that is introduced in [42], as shown in the Fig. 5. In our paper, the ResNet50 [9] is adopted as the base model for vehicle classification, which consists of residual units that preserve the identity and maintain a deeper structure. “Feature maps conv1” represents the output features from conv1. After convolutional layers from conv1 through conv5, the feature vector \( f \) could be obtained.
Similar with the local region branch, the features $f$ is fed into the attention module to obtain distinct features. Then the output feature is utilized to train the identification task with the cross-entropy (CE) loss.

At last, for all branches, the obtained features are named as $f_g$, $f_u$ and $f_l$, respectively. During the training phase of each branch in local region features learning network, Fully Connected (FC) layers are added to identify vehicles only with a part of feature maps as input. This procedure enforces the network to extract discriminative details in each part. At last, the prediction identity classification is given by the FC layer with the CE loss that could be described as:

$$L(\theta) = \ell_{id}^g + \lambda_1 \ell_{id}^u + \lambda_2 \ell_{id}^l$$  \hspace{1cm} (6)$$

where $\theta$ denotes the parameters in the deep model. $\ell_{id}^g$, $\ell_{id}^u$ and $\ell_{id}^l$ represent the identification loss in global features extraction module, respectively. $\lambda_1$ and $\lambda_2$ are the weights for corresponding loss. The CE loss is calculated based on softmax.

Specially, during the test phase, the final features from AANet could be described as follows:

$$f = [f_g \times \alpha, f_u \times (1 - \alpha), f_l \times (1 - \alpha)]$$ \hspace{1cm} (7)$$

where $\alpha$ is the weight for features. $f$ is the feature for the phrase of testing. The size of features from different branches is $1 \times 1 \times 4096$ in our paper.

### 4 Experiments

#### 4.1 Datasets and evaluation metrics

- **VeRi-776** [20]. The dataset is a large-scale urban surveillance vehicle dataset for reID, which contains over 50,000 images of 776 vehicles across 20 cameras. Each vehicle
is from 2-18 cameras with various viewpoints, illuminations and occlusions. In this dataset, 37,781 images of 576 vehicles are split as a train set and 11,579 images of 200 vehicles are employed as a test set. A subset of 1,678 images in the test set generates the query set.

– VehicleID [17]. It is a widely-used vehicle reID dataset, which contains 26267 vehicles and 221763 images in total. The training set contains 110,178 images of 13,134 vehicles. For the testing data, three subsets which contain 800, 1600, and 2400 vehicles are extracted for vehicle search in different scales. Two other test sets are processed in the same way. Some samples are listed in the Fig. 6

– Evaluation Metrics. To measure the performance for vehicle reID task, the CMC and mAP [16] are utilized as evaluation criterions. For each query, its average precision (AP) is computed from its precision-recall curve. And mAP is the mean value of average precisions across all queries.

4.2 Implementation details

For the translation module, the model is trained by Pytorch [23] on NVIDIA GeForce GTX Tian GPU. We utilize the Adam optimizer [13] with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. The initial learning rate is 0.0001 for the first 100 epochs and linearly decays to the learning to 0 over the next 100 epochs. The batch size is 16. For VeRi-776, cameras annotations are provided, whose samples from 20 cameras are first transferred by StarGAN into one camera. The index of the selected camera in our experiment is 1. Because the camera labels are not annotated in VehicleID. So all images in VehicleID are transferred into VeRi-776 which is regarded as the selected camera. And then generated images are utilized for training reID model. Some samples of transferred images by StarGAN in VeRi-776 dataset are shown in Fig. 7.

For the feature learning network, we implement the proposed vehicle reID model in the Matconvnet [27] framework. SGD [2] is employed to update the parameters of the network with a momentum of $\mu = 0.0005$ during the training procedure on both VehicleID and VeRi-776. The batch size is set to 16. In addition, the learning rate of the first 40 epochs is set to 0.1 while the last 25 is 0.01.
4.3 Comparison with the state-of-the-art methods

4.3.1 Comparison on VeRi-776

The results of the proposed method are compared with state-of-the-art methods on VeRi-776 dataset in Tables 2 and 3, which includes: (1) LOMO [14]; (2) FACT+Plate-SNN+STR [19]; (3) VAMI [45]; (4) OIFE+ST [32]; (5) PROVID [20]; (6) VGG+C+T+S [40]; (7) Siamese-CNN+Path-LSTM [25]; (8) VAMI+ST [45]; (9) RAM [18]; (10) AAVER [11]. From the Tables 2 and 3, it should be noted that the proposed method achieves the best performance among the compared methods with rank-1 = 91.71%, mAP = 68.05% on VeRi-776, which acquires the highest mAP and rank-1 among all methods under comparisons. More details are analyzed as follows.

Compared with those methods that learn multi-view features, the proposed also show satisfactory performance. For instance, compared with VAMI [45], our method has a gain of 17.92 in terms of mAP and 14.68 in terms of rank-1 accuracy. This is because that our method eliminates background interference information. It strongly proves that the bias among cameras has a serve influence on the vehicle reID task. Among these methods, RAM [18] is similar with our methods that divides the feature maps into several parts to be local

Table 2  Experimental results on VeRi-776

| Method                      | mAP  | Rank1 | Rank5 |
|-----------------------------|------|-------|-------|
| LOMO [14]                   | 9.64 | 25.33 | 46.48 |
| FACT+Plate-SNN+STR [19]     | 27.77| 61.64 | 78.78 |
| VAMI [45]                   | 50.13| 77.03 | 90.82 |
| OIFE+ST [32]                | 51.42| 68.30 | 89.70 |
| PROVID [20]                 | 53.42| 81.56 | 95.11 |
| VGG+C+T+S [40]              | 57.4 | 86.5  | 92.8  |
| Siamese-CNN+Path-LSTM [25]  | 58.27| 83.49 | 90.04 |
| VAMI+ST [45]                | 61.32| 85.92 | 91.84 |
| RAM [18]                    | 61.5 | 88.6  | 94.0  |
| AAVER [11]                  | 66.35| 90.17 | 94.34 |
| CCA                         | **68.05** | **91.71** | **96.90** |

The mAP (%) and cumulative matching scores (%) at rank 1, 5 are listed.
| Method          | Test size = 800 |                     | Test size = 1600 |                     | Test size = 2400 |                     |
|-----------------|-----------------|---------------------|------------------|---------------------|------------------|---------------------|
|                 | mAP  | Rank1 | Rank5 | mAP  | Rank1 | Rank5 | mAP  | Rank1 | Rank5 |
| LOMO [14]       | –    | 19.76 | 32.14 | –    | 18.95 | 29.46 | –    | 15.26 | 25.63 |
| NuFACT [20]     | –    | 48.90 | 69.51 | –    | 43.64 | 65.34 | –    | 38.63 | 60.72 |
| DRDL [17]       | –    | 49.0  | 73.5  | –    | 42.8  | 66.8  | –    | 38.2  | 61.6  |
| OIFE [32]       | –    | –     | –     | –    | –     | –     | –    | 67.0  | 82.9  |
| C2F-Rank [6]    | 63.5 | 61.1  | 81.7  | 60.0 | 56.2  | 76.2  | 53.0 | 51.4  | 72.2  |
| VAMI [45]       | –    | 63.12 | 83.25 | –    | 52.87 | 75.12 | –    | 47.34 | 70.29 |
| TAMR [7]        | 67.64| 66.02 | 79.71 | 63.69| 62.90 | 76.80 | 60.97| 59.69 | 73.87 |
| VGG+C+T+S [40]  | –    | 69.9  | 87.3  | –    | 66.2  | 82.3  | –    | 63.2  | 79.4  |
| AAVER [11]      | –    | 74.69 | **93.82** | –    | 68.62 | **89.95** | –    | 63.54 | **85.64** |
| CCA             | **78.89** | **75.51** | 91.14 | **76.53** | **73.60** | 86.46 | **73.11** | **70.08** | 83.20 |

The mAP (%) and cumulative matching scores (%) at Rank 1, 5 are listed.
features. Compared with RAM, our method also has a large improvements. This verifies that the step of camera transfer is effectiveness and smooths the bias between cameras as shown in Fig. 7. Moreover, although our proposed method only utilizes visual information, it also has significant improvements when compared with methods with spatio-temporal information. such as FACT+Plate-SNN+STR [19], PROVID [20], Siamese-CNN+Path-LSTM [25], OIFE+ST [32] and VAMI+ST [45], the proposed method has higher mAP, rank-1 and rank-5 than them, which demonstrates that our CCA could extract more discriminative features without other information besides the vehicle images.

4.3.2 Comparison on VehicleID

There are 9 methods are compared with our proposed method, which are (1) LOMO [14]; (2) NuFACT [20]; (3) DRDL [17]; (4) OIFE [32]; (5) C2F-Rank [6]; (6) VAMI [45]; (7) TAMR [7]; (8) VGG+C+T+S [40]; (9) AAVER [11]. Table 3 illustrates the rank-1, rank-5 and mAP of our method and other comparison methods on VehicleID. Firstly, the proposed method outperforms all deep learning based methods under comparison on the test sets with different sizes on VehicleID, which obtains 75.51%, 73.60%, 70.08% in rank-1, respectively. Secondly, different VeRi-776, there is no spatio-temporal labels in VehicleID. Hence, there are no methods that consider the spatio-temporal information. All compared methods utilize the appearance information only from vehicle images. At last, it can be observed that our method obviously outperforms traditional methods, such as LOMO [14]. Compared with LOMO [14], CCA has 55.75%, 54.65% and 54.82% gains in rank-1 on different test sets, respectively. This also shows that our proposed CCA could smooth the camera bias and generate more distinct features for vehicle reID task.

4.4 Evaluation of proposed method

To validate the necessity of the proposed method, some ablation experiments are conducted. The comparison results on VeRi-776 and VehicleID are presented in Tables 4 and 5. “O” means the the training set is original samples while “T” is the generated samples. “Rigid” represents the training network doesn’t employ the STN module and attention module, which is divided into two parts from resnet50 directly. “Part-n” is the descriptor of i-th branch. “Global” means the descriptor is only composed of the features from global branch.

| Table 4 | Performance of features fusion on VeRi-776 |
|-----------------|----------------------|--------|--------|
| Descriptor      | mAP      | Rank1 | Rank5  |
| Global          | 56.71    | 86.55 | 92.14  |
| O-Rigid-Part1   | 48.15    | 81.58 | 90.04  |
| O-Rigid-Part2   | 47.51    | 81.10 | 90.88  |
| O-Rigid-All     | 61.19    | 87.24 | 93.32  |
| O-AANet-Part1   | 50.77    | 83.07 | 92.01  |
| O-AANet-Part2   | 50.41    | 82.24 | 92.19  |
| O-AANet-All     | 65.45    | 89.92 | 94.39  |
| T-AANet-Part1   | 54.04    | 86.94 | 93.74  |
| T-AANet-Part2   | 53.78    | 86.59 | 93.98  |
| T-AANet-All     | 68.05    | 91.71 | 96.90  |

The mAP (%) and cumulative matching scores (%) at Rank 1, 5 are listed.
Table 5 Performance of features fusion on VehicleID

| Descriptor         | Test size = 800 |                    |                    | Test size = 1600 |                    |                    |
|--------------------|----------------|--------------------|--------------------|----------------|--------------------|--------------------|
|                    | mAP            | Rank1              | Rank5              | mAP            | Rank1              | Rank5              |
| Global             | 69.78          | 66.51              | 79.25              | 67.71          | 64.79              | 78.86              |
| O-Rigid-Part1      | 67.82          | 65.69              | 74.28              | 65.37          | 63.38              | 71.22              |
| O-Rigid-Part2      | 67.43          | 65.25              | 73.98              | 64.64          | 62.60              | 72.57              |
| O-Rigid-All        | 74.93          | 71.45              | 88.32              | 72.36          | 69.34              | 82.41              |
| O-AANet-Part1      | 69.22          | 66.69              | 77.00              | 67.44          | 64.81              | 75.37              |
| O-AANet-Part2      | 70.64          | 67.97              | 79.15              | 68.57          | 64.11              | 74.19              |
| O-AANet-All        | 77.27          | 73.79              | 89.98              | 74.47          | 71.07              | 84.90              |
| T-AANet-Part1      | 71.78          | 69.47              | 79.43              | 70.17          | 68.06              | 76.60              |
| T-AANet-Part2      | 71.24          | 68.87              | 78.93              | 70.10          | 67.91              | 76.67              |
| T-AANet-All        | 78.89          | 75.51              | 91.14              | 76.53          | 73.60              | 86.46              |
| Descriptor         | Test size = 2400 |                    |                    | Test size = 3200 |                    |                    |
|                    | mAP            | Rank1              | Rank5              | mAP            | Rank1              | Rank5              |
| Global             | 64.43          | 60.68              | 74.37              | 62.88          | 59.79              | 71.53              |
| O-Rigid-Part1      | 63.42          | 61.50              | 68.85              | 61.96          | 60.21              | 66.78              |
| O-Rigid-Part2      | 63.48          | 61.66              | 68.54              | 61.69          | 59.94              | 66.61              |
| O-Rigid-All        | 67.58          | 63.60              | 81.01              | 67.91          | 65.21              | 75.71              |
| O-AANet-Part1      | 65.30          | 62.75              | 73.07              | 65.12          | 63.13              | 77.35              |
| O-AANet-Part2      | 64.77          | 62.45              | 71.62              | 63.17          | 60.96              | 69.50              |
| O-AANet-All        | 71.19          | 68.01              | 81.54              | 69.11          | 66.18              | 78.17              |
| T-AANet-Part1      | 68.04          | 66.06              | 73.85              | 66.12          | 64.23              | 71.40              |
| T-AANet-Part2      | 67.65          | 65.62              | 73.64              | 66.16          | 64.28              | 71.36              |
| T-AANet-All        | 73.11          | 70.08              | 83.20              | 70.75          | 67.98              | 79.35              |

The mAP (%) and cumulative matching scores (%) at Rank 1, 5 are listed.

Firstly, the difference of “O-AANet-All” and “T-AANet-All” is only the source images of training sets. Hence, compared with “O-AANet-All”, the “T-AANet-All” has gains of 2.6%, 1.79% in mAP and rank-1 on VeRi-776, which demonstrates that through the cross-camera transfer network, the bias of different cameras has dropped. This is because that through the StarGAN, vehicle samples from different cameras are transferred into one common space, which leads to generate images with the similar distribution and smooths the bias between different cameras.

Besides that, because our descriptor is learned by multiple branches in the proposed network, we design an ablation experiment analyzing the effectiveness of global, part and fusion feature. “T-AANet-All” is our proposed method that combines all features for reID task. “T-AANet-Part1” and “T-AANet-Part2” denote the features are extracted by the upper branch and lower branch, respectively. As reported in Tables 4 and 5, it is worth noting that, for each group, the match rates of all independent features are lower than the combination features, such as the “Global”, “T-AANet-Part1” and “T-AANet-Part2”. However, the match rate further increases slightly when adding the part features and global features. For instance, on VeRi-776, compared with “global”, “T-AANet-All” improves 11.34% in
mAP. It shows that combining with global and part feature can provide more useful information. This is because that the global branch learns the context features, and the part branch learn some distinct features that are different the global features. Hence, the combination of global features and part features could form strong features for reID task. This also indicates the advantage of the combination features.

To verify the effectiveness of localization model, we remove the STN module and attention module in the AANet and divide the vehicle image into two regions directly as rigid parts. On VeRi-776, compared with “o-Rigid-All”, “O-AANet-all” has gains of 4.26%, 2.68% in mAP and rank-1, respectively. For VehicleID, we also observe improvements of 2.34%, 2.11%, 3.61%, 1.2% in mAP on test set with the size of 800, 1600, 2400 and 3200. All of these shows that the proposed AANet could learn more discriminative features for vehicle reID.

4.5 Visualization of results

Furthermore, to illustrate the validate of the proposed CCA, some experiment results on VehicleID are visualized. Examples are shown in Fig. 8. In Fig. 8, There are two group results on VehicleID. For each group, the left column shows query images, while images on the right-hand side are the top-5 results obtained by the proposed CCA. Vehicle images with green border are right results while other images are wrong results. For all results, the number on the left-top means Vehicle ID. The same Vehicle ID represents the same vehicle. The Camera ID is the camera number that images are captured. From Fig. 8, it is significant that our proposed CCA has high accuracy and good robustness to different viewpoints and illumination.

5 Conclusion

In this paper, we propose cross-camera adaptation framework for better smoothing the bias between different cameras, which reduces the influence of illumination, background and resolution for vehicle reID task by transferring images into a common space and learning a
powerful discriminative feature. Besides that, AANet is designed to obtain a series of local regions for vehicle reID, which focuses on locating the meaningful parts while suppressing background. Form the detailed experimental analysis, it is obvious that our proposed method achieves competitive results with other methods, which verifies the effectiveness of the proposed CCA.

**Acknowledgments** This work was supported in part by the National Natural Science Foundation of China Grant 61370142, Grant 61272368 and Grant 62002041, by the Fundamental Research Funds for the Central Universities Grant 3132016352, by the Fundamental Research of Ministry of Transport of P. R. China Grant 2015329225300, by the Dalian Science and Technology Innovation Fund 2018J12GX037 and Dalian Leading talent Grant, by the Foundation of Liaoning Key Research and Development Program, by the China Postdoctoral Science Foundation 3620080307, by the Liaoning Revitalization Talents Program XLYC1908007 and by the Dalian Science and Technology Innovation Fund 2019J11CY001.

**References**

1. Bai Y, Lou Y, Gao F, Wang S, Wu Y, Duan L-Y (2018) Group-sensitive triplet embedding for vehicle reidentification. IEEE Transactions on Multimedia 20(9):2385–2399
2. Bottou L (2010) Large-scale machine learning with stochastic gradient descent. In: Proceedings of COMPSTAT'2010, pp 177–186, Springer
3. Choi Y, Choi M, Kim M, Ha J-W, Kim S, Choo J (2018) Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 8789–8797
4. Chu W, Liu Y, Shen C, Cai D, Hua X-S (2018) Multi-task vehicle detection with region-of-interest voting. IEEE Trans Image Process 27(1):432–441
5. Fang Y, Wang C, Yao W, Zhao X, Zhao H, Zha H (2019) On-road vehicle tracking using part-based particle filter. IEEE Trans Intell Transp Syst 20(12):4538–4552
6. Guo H, Zhao C, Liu Z, Wang J, Lu H (2018) Learning coarse-to-fine structured feature embedding for vehicle re-identification. In: Thirty-second AAAI conference on artificial intelligence
7. 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 28(9):4328–4338
8. He B, Li J, Zhao Y, Tian Y (2019) Part-regularized near-duplicate vehicle re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3997–4005
9. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
10. Hu X, Xu X, Xiao Y, Chen H, He S, Qin J, Heng P-A (2018) Sinet: A scale-insensitive convolutional neural network for fast vehicle detection. IEEE Trans Intell Transp Syst 20(3):1010–1019
11. Khorramshahi P, Kumar A, Peri N, Rambhatla SS, Chen J-C, Chellappa R (2019) A dual path model with adaptive attention for vehicle re-identification. arXiv preprint arXiv:1905.03397
12. Li D, Chen X, Zhang Z, Huang K (2017) Learning deep context-aware features over body and latent parts for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 384–393
13. Li W, Zhao R, Xiao T, Wang X (2014) Deepreid: Deep filter pairing neural network for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 152–159
14. Liao S, Hu Y, Zhu X, Li SZ (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
15. Lin W, Li Y, Yang X, Peng P, Xing J (2019) Multi-view learning for vehicle re-identification. In: 2019 IEEE international conference on multimedia and expo (ICME), pp 832–837, IEEE
16. Lin Y, Zheng L, Zheng Z, Wu Y, Hu Z, Yan C, Yang Y (2019) Improving person re-identification by attribute and identity learning. Pattern Recogn 95:151–161
17. Liu H, Tian Y, Yang Y, Pang L, Huang T (2016) Deep relative distance learning: Tell the difference between similar vehicles. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 2167–2175
18. Liu X, Zhang S, Huang Q, Gao W (2018) Ram: A region-aware deep model for vehicle re-identification. In: 2018 IEEE international conference on multimedia and expo (ICME), pp 1–6, IEEE
19. 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, Springer
20. Liu X, Liu W, Mei T, Ma H (2017) Provid: Progressive and multimodal vehicle reidentification for large-scale urban surveillance. IEEE Transactions on Multimedia 20(3):645–658
21. Lou Y, Bai Y, Liu J, Wang S, Duan L-Y (2019) Embedding adversarial learning for vehicle re-identification. IEEE Trans Image Process 28(8):3794–3807
22. Ma Z, Chang D, Xie J, Ding Y, Wen S, Li X, Si Z, Guo J (2019) Fine-grained vehicle classification with channel max pooling modified cnns. IEEE Trans Veh Technol 68(4):3224–3233
23. Paszke A, Gross S, Chintala S, Chanan G, Yang E, DeVito Z, Lin Z, Desmaison A, Antiga L, Lerer A (2017) Automatic differentiation in pytorch
24. Peng J, Wang H, Zhao T, Fu X (2019) Learning multi-region features for vehicle re-identification with context-based ranking method. Neurocomputing 359:427–437
25. 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
26. Tang Z, Naphade M, Liu M-Y, Yang X, Birchfield S, Wang S, Kumar R, Anastasius D, Hwang J-N (2019) Cityflow: A city-scale benchmark for multi-target multi-camera vehicle tracking and re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 8797–8806
27. Van Der Maaten L (2014) Accelerating t-sne using tree-based algorithms. The Journal of Machine Learning Research 15(1):3221–3245
28. Wang H, Peng J, Chen D, Jiang G, Zhao T, Fu X (2020) Attribute-guided feature learning network for vehicle re-identification. IEEE Multimedia
29. Wang H, Wang Y, Zhang Z, Fu X, Xu M, Wang M (2020) Kernelized multiview subspace analysis by self-weighted learning. IEEE Transactions on Multimedia
30. Wang Y (2020) Survey on deep multi-modal data analytics: Collaboration, rivalry and fusion. ACM Trans. on Multimedia Computing Communications and Applications
31. Wang Y, Lin X, Wu L, Zhang W (2017) Effective multi-query expansions: Collaborative deep networks for robust landmark retrieval. IEEE Trans Image Process 26(3):1393–1404
32. Wang Z, Tang L, Liu X, Yao Z, Yi S, Shao J, Yan J, Wang S, Li H, Wang X (2017) Orientation invariant feature embedding and spatial temporal regularization for vehicle re-identification. In: Proceedings of the IEEE International Conference on Computer Vision, pp 379–387
33. Wei W, Guo C (2019) A text semantic topic discovery method based on the conditional co-occurrence degree. Neurocomputing 368:11–24
34. Wu L, Wang Y, Gao J, Li X (2018) Where-and-when to look: Deep siamese attention networks for video-based person re-identification. IEEE Transactions on Multimedia 21(6):1412–1424
35. Wu L, Wang Y, Li X, Gao J (2018) Deep attention-based spatially recursive networks for fine-grained visual recognition. IEEE transactions on cybernetics 49(5):1791–1802
36. Wu L, Wang Y, Li X, Gao J (2018) What-and-where to match: deep spatially multiplicative integration networks for person re-identification. Pattern Recogn 76:727–738
37. Wu L, Wang Y, Shao L, Wang M (2019) 3-d personvlad: Learning deep global representations for video-based person reidentification. IEEE transactions on neural networks and learning systems 30(11):3347–3359
38. Yang Z, Luo T, Wang D, Hu Z, Gao J, Wang L (2018) Learning to navigate for fine-grained classification. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 420–435
39. Zapletal D, Herout A (2016) Vehicle re-identification for automatic video traffic surveillance. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp 25–31
40. Zhang Y, Liu D, Zha Z (2017) Improving triplet-wise training of convolutional neural network for vehicle re-identification, pp 1386–1391
41. Zhao Y, Shen C, Wang H, Chen S (2019) Structural analysis of attributes for vehicle re-identification and retrieval. IEEE Trans Intell Transp Syst 21(2):723–734
42. Zheng Z, Zheng L, Yang Y (2018) A discriminatively learned cnn embedding for person reidentification. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 14(1):13
43. Zhou Y, Liu L, Shao L (2018) Vehicle re-identification by deep hidden multi-view inference. IEEE Trans Image Process 27(7):3275–3287
44. Zhou Y, Shao L (2017) Cross-view gan based vehicle generation for re-identification. In: BMVC, 1, pp 1–12
45. Zhou Y, Shao L (2018) Aware attentive multi-view inference for vehicle re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 6489–6498
46. 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 21(1):410–420
47. Zhu J-Y, Park T, Isola P, Efros AA (2017) Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision, pp 2223–2232

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

Affiliations

Jinjia Peng1 · Guangqi Jiang1 · Dongyan Chen1 · Tongtong Zhao1 · Huibing Wang1 · Xianping Fu1,2

Jinjia Peng
jinjiapeng@dlmu.edu.cn

Guangqi Jiang
guangqi-j@dlmu.edu.cn

Dongyan Chen
chendongyan@dlmu.edu.cn

Tongtong Zhao
zhaotongtong94@163.com

1 College of Information and Science Technology, Dalian Maritime University, Dalian, Liaoning, 116021, China
2 Pengcheng Laboratory, Shenzhen, Guangdong, 518055, China