Information Sciences 564 (2021) 71–84

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Viewpoint adaptation learning with cross-view distance metric for robust vehicle re-identification

Qi Wang [a], Weidong Min [b,c,*], Qing Han [a], Ziyuan Yang [a], Xin Xiong [a], Meng Zhu [a], Haoyu Zhao [a]

[a] School of Information Engineering, Nanchang University, Nanchang 330031 China
[b] School of Software, Nanchang University, Nanchang330047 China
[c] Jiangxi Key Laboratory of Smart City, Nanchang 330047 China

ARTICLE INFO

Article history:
Received 12 July 2020
Received in revised form 4 December 2020
Accepted 12 February 2021
Available online 26 February 2021

2010 MSC:
00-01
99-00

Keywords:
Vehicle Re-identification
VANet
CVLSR
Cross-view Distance Metric

ABSTRACT

Many vehicle re-identification (Re-ID) problems require the robust recognition of vehicle instances across multiple viewpoints. Existing approaches for dealing with the vehicle re-ID problem are insufficiently robust because they cannot distinguish among vehicles of the same type nor recognize high-level representations in deep networks for identical vehicles with various views. To address these issues, this paper proposes a viewpoint adaptation network (VANet) with a cross-view distance metric for robust vehicle Re-ID. This method consists of two modules. The first module is the VANet with cross-view label smoothing regularization (CVLSR), which abstracts different levels of a vehicle's visual patterns and subsequently integrates multi-level features. In particular, CVLSR based on color domains assigns a virtual label to the generated data to smooth image-image translation noise. Accordingly, this module supplies the viewing angle information of the training data and provides strong robust capability for vehicles across different viewpoints. The second module is the cross-view distance metric, which designs a cascaded cross-view matching approach to combine the original features with the generated ones, and thus, obtain additional supplementary viewpoint information for the multi-view matching of vehicles. Results of extensive experiments on two large scale vehicle Re-ID datasets, namely, VeRi-776 and VehicleID demonstrate that the performance of the proposed method is robust and superior to other state-of-the-art Re-ID methods across multiple viewpoints.

1. Introduction

Vehicle re-identification (Re-ID) aims to search for a specific target vehicle in a large-scale database across multiple camera scenes. This task is widely used in urban surveillance, cross-camera vehicle tracking, and intelligent security fields. Similar to person Re-ID problems, vehicle Re-ID problems are inevitably affected by domain variations [1–3].

Viewpoint variations pose several challenges to the robustness of vehicle Re-ID. Existing deep learning approaches utilize the output of the final layer of a convolutional neural network (CNN) to obtain high-level semantic features, achieving good performance in target classification recognition as global feature representation [4–8]. However, training models based on

* Corresponding author at: School of Software, Nanchang University, Nanchang 330047 China.
E-mail addresses: qiwang@email.ncu.edu.cn (Q. Wang), minweidong@ncu.edu.cn (W. Min), hanqing@ncu.edu.cn (Q. Han), cziyuanyang@gmail.com (Z. Yang), 15070017693@163.com (X. Xiong), mengzhu@email.ncu.edu.cn (M. Zhu), zhaohaoyu@email.ncu.edu.cn (H. Zhao).

https://doi.org/10.1016/j.ins.2021.02.013
0020-0255/

these methods are not robust to many scenarios, particularly in terms of viewpoint awareness for vehicle retrieval tasks across different viewpoints due to the abstraction of high-level semantics. Moreover, unbalanced training data make the discriminative capability of the network biased toward viewing angle samples. Thus, a vehicle Re-ID training model requires the annotation of large-scale datasets to enhance robustness. In consideration of the aforementioned issues, generative adversarial networks (GANs) [9] are recently used to generate images via adversarial training; the generated images will be used to supplement training data. However, these networks introduce noise and pollute the generated samples. Thus, examining the robustness of vehicle Re-ID based on data denoising is highly significant.

Another complicated problem originates from multi-view information [10], particularly in multi-view matching [11,12]. Single viewpoint information may be unreliable in situations where in a vehicle exhibits varying appearance in different viewpoints because of viewpoint variation. For example, identifying a vehicle with a rear viewpoint feature based on its front viewpoint during the ranking stage of vehicle Re-ID is difficult.

The objective of this paper is to enhance the viewpoint-aware robustness of training models by denoising the generated samples and fusing multi-level information, enabling the accurate identification of vehicles across views and achieving multi-view matching from a single view. This paper proposes a viewpoint adaptation network (VANet) with a cross-view distance metric for robust vehicle Re-ID. Its major contributions are summarized as follows.

1. A VANet is proposed for extracting feature maps from different layers to integrate multi-level information. View-transferred GANs are adopted to generate images by learning from the real training data across various viewpoints. In addition, cross-view label smoothing regularization (CVLSR) is performed to assign the corresponding labels to generate cross-view images on the basis of color domains. The proposed network is combined with CVLSR to provide a robust viewpoint-aware capability and mitigate the effect of noise data caused by GANs.
2. A cross-view distance metric is designed. This metric adaptively combines the original and generated viewpoint information to address the multi-view matching issue.
3. Extensive experiments are performed using the proposed network on two vehicle Re-ID datasets, and the proposed network achieves better performance than other state-of-the-art methods. Moreover, ablation studies are conducted, and the results verify that the comprehensive view features of the proposed network can outperform state-of-the-art vehicle Re-ID methods.

The remainder of this paper is organized as follows. Section 2 reviews the related work. Section 3 describes the overall process of the proposed architecture. Sections 4 and 5 introduce the viewpoint adaptation learning and cross-view distance metric modules for vehicle Re-ID, respectively. Section 6 presents the experimental results that validate the superiority of the proposed method. Section 7 provides the conclusion of the study.

2. Related work

Research on vehicle Re-ID is relevant to person re-ID. With the increasing importance of public security, many researchers have proposed different image-based person Re-ID algorithms, and CNNs have been widely used to address Re-ID problems [1]. Wu et al. [2] presented a novel method for gradually improving the features learned from CNN. This method involves selecting a few samples with the most reliable pseudo labels and updating the CNN model. Deng et al. [3] trained the Re-ID model with translated images by using supervised images. To resolve the problem caused by image style variations, Zhong et al. [13] proposed a camera style adaptation, called CamStyle, which can smoothen camera style disparities. In addition, many researchers have attempted to reduce the difference between the source and the target [14,15]. For example, Yu et al. [16] proposed a deep discriminative representation network. Various methods have been proposed to improve the classification model and obtain an effective scheme for the Re-ID task [17,18]. Lin et al. [19] designed a bottom-up clustering approach for jointly optimizing an unsupervised CNN. Researchers have also proposed several methods to address the problems of video-based Re-ID [20,21].

Previous works [22,23] have presented automatic license plate recognition as a globally unique identifier. However, license plate occlusion is common in actual scenes. Popular deep learning feature representations for vehicle Re-ID have been utilized to learn the discriminative feature of vehicle images effectively [24,25]. Alfasly et al. [26] designed a twofold framework, called Mob.VFL*, which contains variational feature learning and long short-term memory networks to achieve an efficient representation. Bashir et al. [27] adopted a progressive two-step cascaded framework, i.e., VR-PROUD, to conduct feature extraction and unsupervised learning. Lou et al. [28] designed a feature distance adversarial network, named FDA-Net, to generate hard negative samples in feature space and obtain a highly discriminative vehicle Re-ID model. To solve multi-view vehicle Re-ID issues, Zhou et al. [11] proposed a viewpoint-aware attentive multi-view inference model that can obtain multi-view features from single-view ones. Peng et al. [29] developed two frameworks to achieve domain adaptation and extract many distinctive cues. Guo et al. [30] presented a coarse-to-fine ranking method with three types of loss to reduce distance among similar images. Wang et al. [31] established two modules. In one module, a model utilizes 20 key point locations to extract and align discriminative part-level features; in the other module, spatial–temporal regularization is introduced to improve retrieval results. Several studies have explored car parts to obtain local features, and most of these studies have focused on the feature extraction problem [32,33]. He et al. [34] proposed a part-regularized discriminative

feature method for enhancing perceptive capability for subtle discrepancies and then combined it with global constraints. Zhu et al. [35] designed quadruple directional deep learning features with the same basic architectures but different pooling layers, including horizontal, vertical, diagonal, and anti-diagonal average pooling. Gou et al. [36] presented a two-level attention network supervised by a multi-grain ranking loss (TAMR) to extract discriminative features from the appearance of a vehicle. This network also learns a distance metric to pull similar images close and separate dissimilar ones [37,38].

GANs have recently elicited the attention of many researchers in different fields, such as image generation and translation [39,9]. Zhu et al. [9] proposed the CycleGAN model with two mapping functions to achieve image translation between two different domains. Moreover, some studies have used GAN to generate multi-view or multi-style images for enhancing data. Lou et al. [40] developed an embedding adversarial learning network to improve the capability for similar vehicle discrimination by learning the embedding feature of hard negative and cross-view images. Several studies have focused on image transfer among different datasets to achieve data enhancement. Zheng et al. [41] presented label smoothing regularization for outliers to combine unlabeled GAN samples with real ones for semi-supervised learning.

3. Overview of the proposed framework

The challenge to existing vehicle Re-ID approaches is the difficulty in distinguishing subtle discrepancies between vehicles of the same type through high-level representations in deep networks and the bottleneck of single-view feature matching, as shown in Fig. 1.

In this section, a novel framework for vehicle Re-ID is introduced to solve the aforementioned issues. The framework consists of two modules: viewpoint adaptation learning and cross-view distance metric modules (Fig. 2). The viewpoint adaptation learning module focuses on subtle discrepancies among vehicles across multiple views to develop a robust viewpoint-aware model. Meanwhile, the purpose of the cross-view distance metric module is to design a novel multi-view feature matching method that can supply additional viewpoint information to improve the performance of a single-view matching task. First, CycleGAN is trained using real images from different viewpoints, and the viewpoint transfer model is adopted to generate cross-view images as newly added training data. Then, a VANet is utilized to train the real and generated samples. CVLSR is implemented on the generated samples and involved in several color domains (details are discussed in Section 4.2). Consequently, the cross-view features adaptively adjust weights by using a penalty weight matrix that aggregates multilevel features. Lastly, the original and cross-view feature vectors are utilized to measure the distance for multi-view vehicle Re-ID.

4. Viewpoint adaptation learning

The proposed VANet is described in Section 4.1, and the training strategy with CVLSR is introduced in Section 4.2.

4.1. Viewpoint adaptation network

In a practical surveillance scenario, recognizing a vehicle from one visible view is difficult. The current study aims to utilize and fuse information from different levels to adapt viewpoint variations. Traditional deep architectures consider the out-

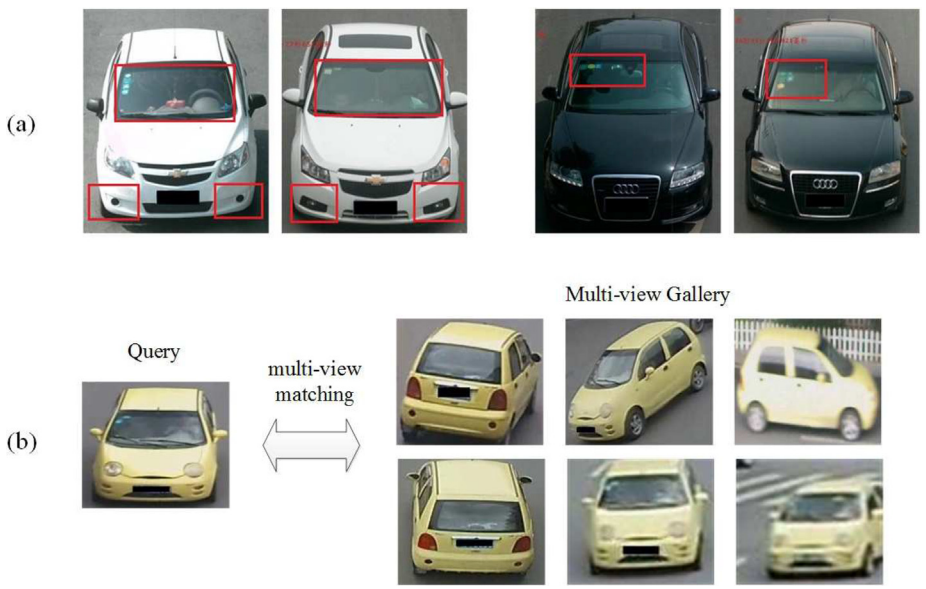

Fig. 1. Motivations for studying robust vehicle Re-ID.

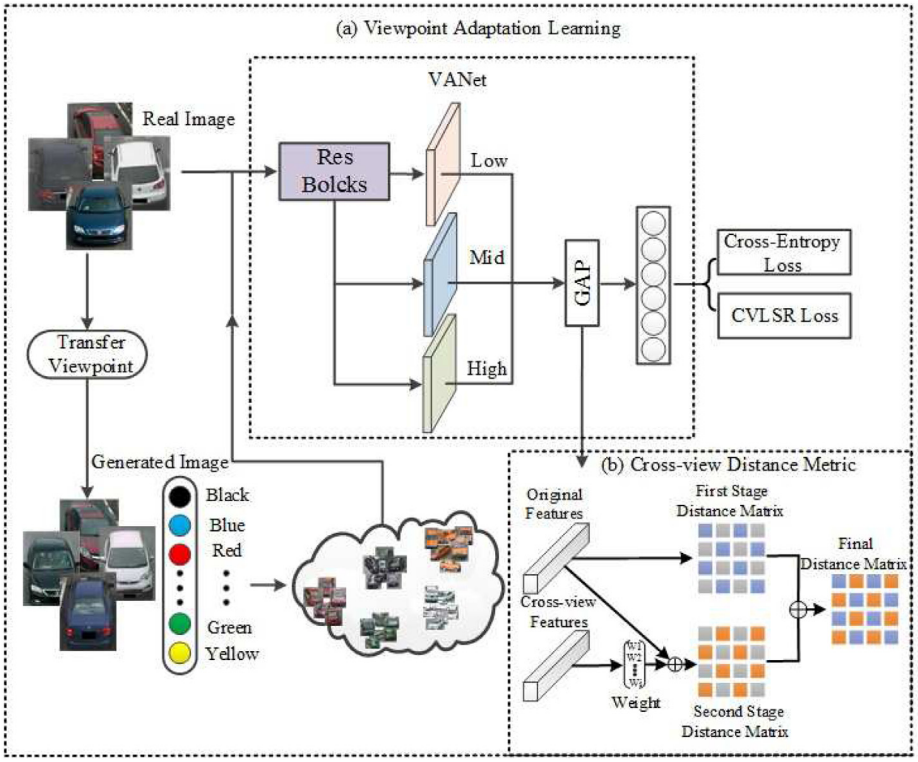

Fig. 2. Pipeline of the proposed method for vehicle Re-ID.

put of a last layer of the network as the final feature representation. However, the high-level information captured by the last convolution layer is abstract and insensitive to cross-domain variation. To obtain robust representations across multiple viewpoints, the proposed VANet contains information from different CNN levels. The pipeline of the proposed network is shown in Fig. 2. The viewpoint adaptation model is learned from the real and generated images. We utilize three information levels to extract feature maps from each image, denoted as $f_i, i \in \{low, mid, high\}$, where i represents the different levels of the semantic feature maps of VANet. Their outputs are then extracted through global average pooling (GAP) and concatenated as the final image representation.

We visualize the response maps of different layers of ResNet-50 on the basis of vehicle Re-ID data, as shown in Fig. 3(a). Each row corresponds to the information activated by different layers. As indicated in the first row of Fig. 3(a), the responses of low-level features are highly focused on identifying the contour of a vehicle. The low-level semantic information of VANet captures texture and shape information and accurately locates vehicles. Fig. 3(a), particularly from the second row to the fourth row, shows that the responses of different semantic levels focus on several local regions, such as the annual inspection mark, headlight, and taillight. Early layers are highly sensitive to subtle differences, but the last layers (Conv5_x) merely extract global features, which are most sensitive to category-level abstract information. Consequently, the extracted high-level representations cannot be expressed as an effective local region; such representations are particularly effective for vehicles of the same type. The response maps of specific layers are visualized in VANet, as shown in Fig. 3(b). VANet can capture localized information such as the annual inspection mark and headlight. VANet uses subtle regions to identify vehicles of the same model.

4.2. Cross-view label smoothing regularization

The unbalanced allocation of viewpoints during the training phase makes the network biased toward front samples. The ratio of front viewpoints to the overall training samples is higher than that of the rear viewpoint on the VehicleID dataset. To learn the complete viewpoint information of each vehicle, we use the image generation method based on CycleGAN to effectively alleviate the unbalanced viewpoint distribution of vehicles in the training images. The CycleGAN model generates the new training sample set that corresponds to the view domain, including the front view of vehicle $X = \{x_1, x_2, \cdots, x_j\}$ and the rear view of vehicle $Y = \{y_1, y_2, \cdots, y_i\}$.

Although CycleGAN can produce many samples for training, it transfers the viewpoint to the domain set level, causing noise or distortion in the generated images. As shown in Fig.4, the generated images of several classes are blurred and distorted due to lack of viewpoint pairs.

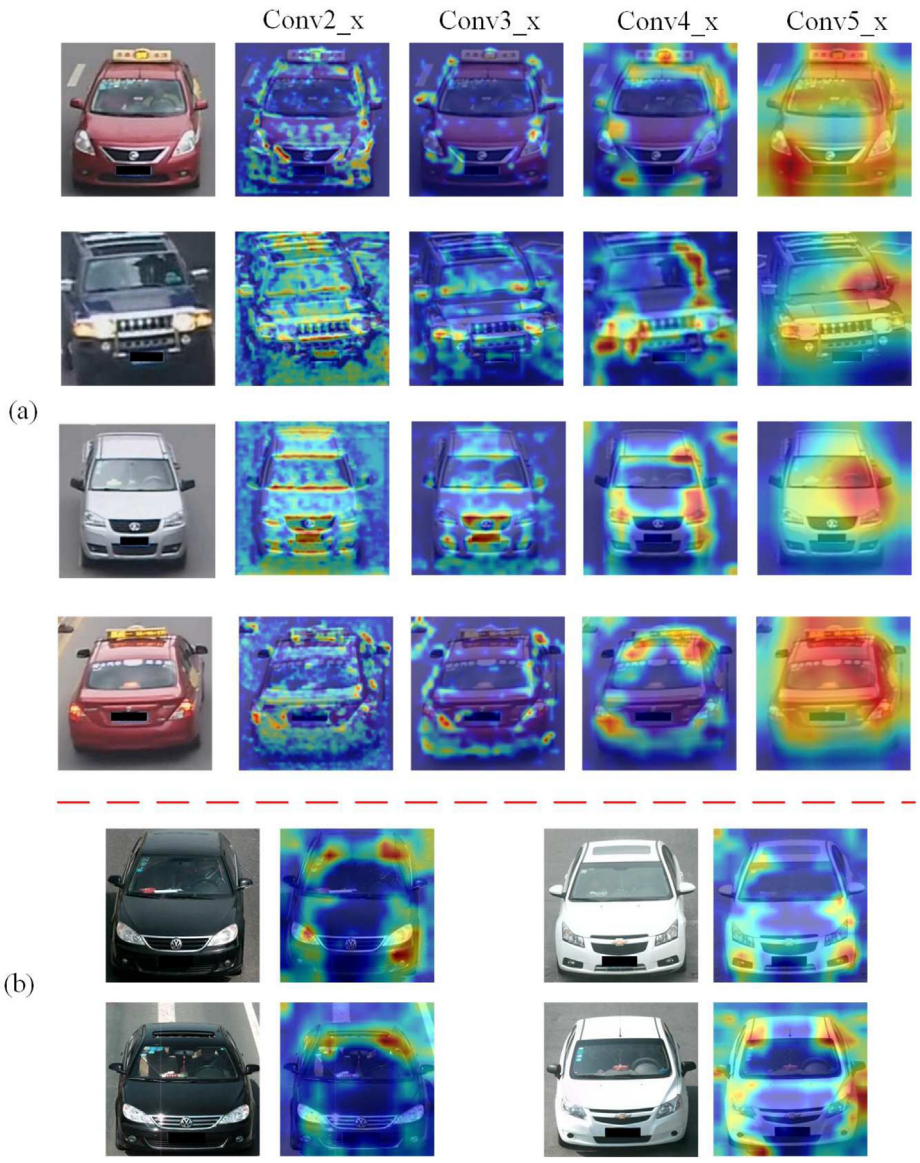

Fig. 3. Visualizations of the feature maps of VANet with different level information.

Fig. 4. Examples of cross-view generation on VeRi-776.

Existing GAN-based semi-supervised approaches are used to assign virtual labels to produce many data for training, and label smoothing regularization (LSR) is adopted to consider non-ground truth distribution. LSR assigns small confidence and weight values to the ground truth label and other classes, respectively, and provides equal possibilities to the classes of each generated image to avoid overtting. The label distribution of each generated image is expressed as

$$q_{lsr}(k) = \begin{cases} \frac{\varepsilon}{K} & k \neq y \\ 1 - \varepsilon + \frac{\varepsilon}{K} & k = y \end{cases} \tag{1}$$

where $\varepsilon \in [0, 1]$, K is the number of classes in original training samples, $k \in [1, \cdots, K]$ represents the k-th predefined training class. For a generated image, y denotes the ground truth class label, and the loss of LSR is defined as

$$L_{lsr} = -(1 - \varepsilon)\log(p(y)) - \frac{\varepsilon}{K}\sum_{k=1}^{K}\log(p(k)) \tag{2}$$

where $p(k) \in [0, 1]$ is the predicted probability of the input belonging to class k and $\varepsilon \in [0, 1]$ is a hyperparameter.

The loss function suggests that LSR depends on ground truth class and also assigns weights to other classes. However, excessive smoothing may occur when numerous training classes are present. Our goal is to eliminate the weights of some classes that are dissimilar to the generated image.

The generated images are shown in Fig. 5. Each row corresponds to different generative viewpoints of a vehicle. We can observe that the subtle information of the generated images with the same ID has changed, particularly for the red arrows, such as logos, lights, and mirrors. Therefore, the generated images that introduce noise cannot assign virtual labels accurately. Color information, which functions as prior knowledge in the generated blurred image, is retained in many cases. Inspired by LSR, we design CVLSR to assign virtual labels to the generated data that are set on the predefined training classes of the same color to filter irrelevant classes. CVLSR assigns generated images to virtual labels in accordance with their dominant color, eliminating the interference of other classes (i.e. different color classes) and reducing their weight. Compared with LSR, CVLSR enables a training model to avoid the over-smoothness issue with a large amount of classes.

To reduce other non-ground truth classes of interference, we utilize the color property of the generated images to converge groups with similar classes. Fig.6 illustrates the label distribution of CVLSR. Given a generated image I, where $C = \{1, 2, \cdots, c\}$ is the number of color labels, the proposed label distribution $q_{cvlsr}(k)$ is defined as

Fig. 5. Examples of generated noise images from different viewpoints on VehicleID.

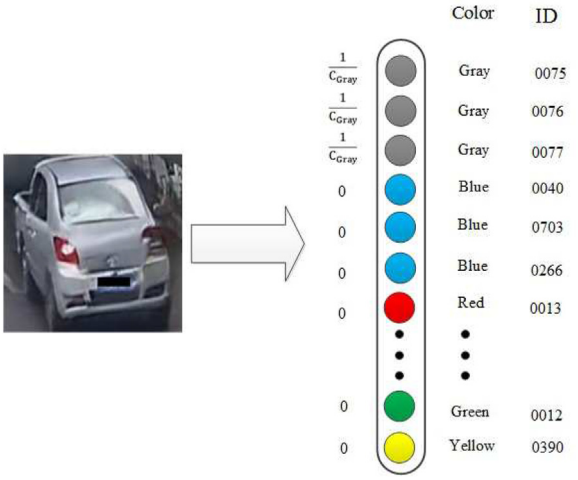

Fig. 6. Label distributions of a training image in the proposed method.

$$q_{cvlsr}(k) = \begin{cases} 0 & k \neq y_c \\ \frac{1}{K_c} & k = y_c \end{cases} \tag{3}$$

where y_c represents the vehicle color label and Kp_c is the class number of the c-th color. The loss function of cross-entropy loss for the training data is written as

$$L_{cvlsr} = -(1-\varepsilon)\log(p(y)) - \frac{\varepsilon}{K_c}\sum_{c=1}^{C}\log(p(k)) \tag{4}$$

The proposed VANet has two types of losses. For real and generated images, we set ε to 0 and 1, respectively.

In the proposed label distribution strategy, the similarity of the generated images is maintained using color properties. Color constraints enable noisy images to reduce label contribution on uncorrelated classes. Compared with LSR, CVLSR effectively resolves the over-smoothness problem by reducing the weights of several non-ground classes.

5. Cross-view distance metric

This component consists of two sub-modules: (1) cross-view feature extraction and aggregation and (2) cascaded cross-view matching modules.

5.1. Cross-view features extraction and aggregation

The feature extraction step involves original and cross-view features. ResNet-50 is adopted as the baseline network. All original images are inputted into the VANet model to obtain a multi-level feature vector, and the images generated through the trained CycleGAN model are used to extract the cross-view model. The multi-level feature vector comprises three level branches of the feature maps from different layers of ResNet-50. The res4 features are selected as the low-level semantic information and denoted as f_{res4}. In accordance with [42], the feature maps of res5a and res5b are adopted as the middle-level semantic information, which are respectively denoted as f_{res5a} and f_{res5b}. These feature maps and the feature map from the last convolution layer f_{pool5} are fused into the final feature representation. Multiple layers output feature vectors by adding a GAP layer. Three vectors are concatenated as the final image representation f and subsequently fed into the full connection layer. The concatenation result is as expressed as

$$f = \sum_{i=1}^{3} f_i = concat(f_{res4}, f_{res5a}, f_{res5b}, f_{pool5}) \tag{5}$$

Viewpoint variations are inherent in vehicle Re-ID in real scenarios. Vehicles with different IDs exhibit similar appearances from the same viewpoint, whereas those with the same IDs appear different in various viewpoints. The majority of existing approaches infer multi-view information from a single-view input to obtain viewpoint invariant features. However, if noise is presented in the generated multi-view information, then misalignment is likely to occur. To overcome this issue, an adaptive multi-view features fusion aggregation approach is proposed to reassign weights to the multi-view feature vector on the basis of the original distance between the original and cross-view images.

During feature fusion, the entire feature vector f is encoded using an adaptive penalty weight matrix ω. The corresponding images for a given probe vehicle p and the gallery set $G = \{g_i | i = 1, 2, \cdots, N\}$ with N images are represented as p' and $G' = \{g_i' | i = 1, 2, \cdots, N\}$, respectively. The view pair distance is recalculated to compare the subtle deviation between the original and cross-view images. The weight of the view pair distance in the entire set is small, hence, the deviation between the two images should also be small.

We adaptively assign the weights to the generated cross-view feature vector in accordance with the pairwise distance; that is, large and small weights are assigned to close and far distances, respectively. The adaptive penalty weight matrix $\omega = \sum_i \omega_i, (i = 1, 2, \cdots, N)$ is defined using the Gaussian kernel, in which each element ω_i in ω is defined as

$$\omega_i = \frac{\exp\left(-d_{x_i^a, x_i^b}\right)}{\sum_{i=1}^{N} \exp\left(-d_{x_i^a, x_i^b}\right)} \tag{6}$$

where x_i^a and x_i^b represent the multilevel feature vectors of the i-th original and cross-view images, respectively, and $d_{x_i^a, x_i^b}$ is the Euclidean distance between x_i^a and x_i^b.

The generated cross-view feature vector is assigned with soft weighting by ω, and the original view feature vector is concatenated to achieve the fused feature vector. The final concatenation result f_{final} is determined as

$$f_{final} = [f_a; \omega_i \cdot f_b] \tag{7}$$

where f_a and f_b are a pair of vehicle images that belongs to the same class ID in both views.

5.2. Cascaded cross-view matching

Measuring identical vehicles with various views in Re-ID poses a significant challenge. Positive images with different viewpoints may be excluded from the ranking list L due to viewpoint variations. We use the image generation technology of CycleGAN to generate the cross-view of vehicles to deal with the aforementioned problem. The cross-view information can optimize the retrieval result because positive pairs from different viewpoints are frequently included. Conversely, positive pairs from the same viewpoint may hinder retrieval performance due to the noise generated by the cross-view information. To address this trade-off issue, cascaded cross-view matching (CCM) is designed to obtain a robust ranking list.

Fig. 7 illustrates the CCM process, in which a cross-view image is generated for each input image in the probe and gallery sets. First, the original view features from VANet are fed to calculate the pairwise distance. To include many positive samples from a similar viewpoint, the method proposed in [43] is adopted to obtain $\mathcal{L}$ as the ranking list of the first stage. Second, the cross-view feature vector for each generated image is weighted using adaptive penalty weights. Then, the original feature vector is concatenated to achieve the aggregated feature vector, which is applied to measure distance for ranking in the second stage. Lastly, the pairwise distance between p and G is recalculated by comparing their multi-view feature vectors to obtain a more reliable viewpoint-aware ranking list based the initial ranking list.

To obtain additional positive samples with different viewpoints and retain the original ones with similar viewpoints, we revise the final ranking results by using multi-view confidence weights on the basis of the initial list of such weights. The final distance set $D(p, g_i)$ based on probe p is defined by encoding the original and multi-view distances as the confidence weights in Formula (8)

$$D(p, g_i) = \omega^1 \cdot \omega^2 \tag{8}$$

where ω^1 is the confidence weight between p and g_i in the first phase and ω^2 corresponds to the multi-view confidence, which is redefined using the Gaussian kernel of the multi-view pairwise distance in the second phase.

6. Experiments

6.1. Dataset and evaluation protocols

In the experiment, the proposed VANet is applied to the VeRi-776 and VehicleID datasets. VeRi-776 is a vehicle Re-ID dataset derived from real city surveillance scenarios. It includes labels for the different attributes of a vehicle, such as color, model, and space time information. The entire dataset is captured using 218 cameras at different viewpoints, illuminations, and occlusions. This dataset contains more than 50,000 images of 776 vehicles; among which 37,778 images of 576 vehicles are used for training and 11,579 images of 200 vehicles are used for testing. A subset containing 1678 images taken from the test set is used as query images in the evaluation phase.

VehicleID is a large-scale vehicle Re-ID dataset derived from the daytime data of multiple real-world surveillance cameras in a small city in China. The entire dataset contains 221,567 images of 26,267 vehicles, each of which includes a corresponding front and rear viewpoints. The images are also labeled with model and color information. VehicleID is divided into a

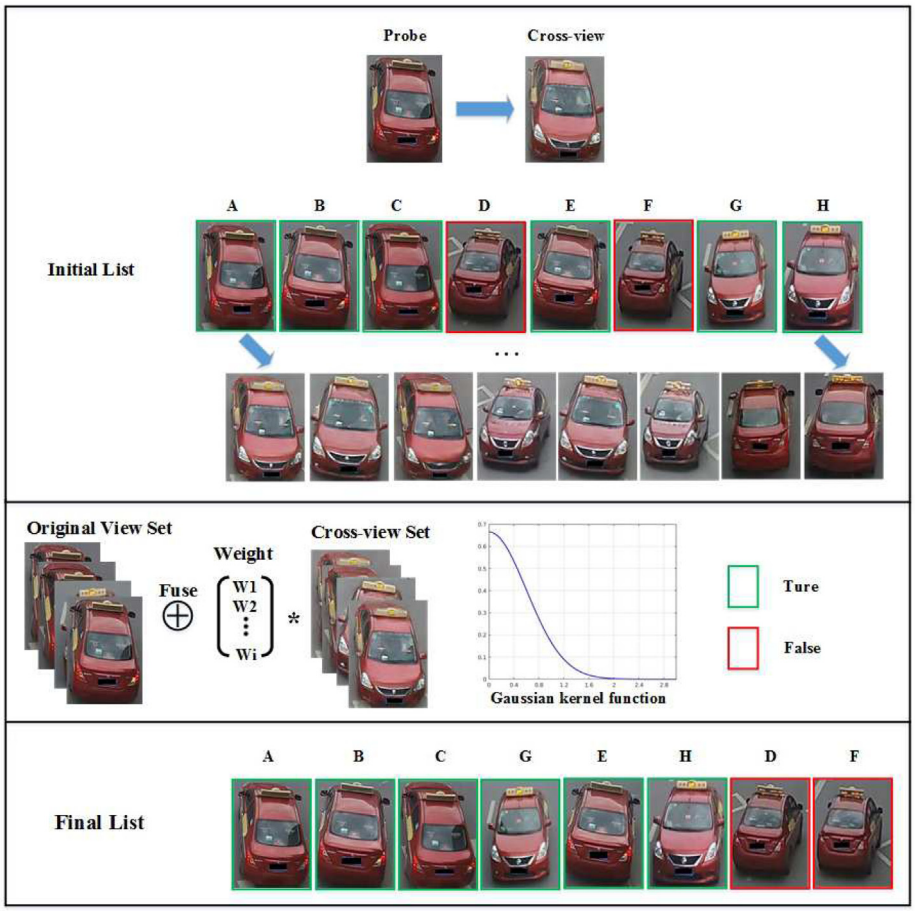

Fig. 7. Examples of the vehicle Re-ID results and illustration of the process of cascaded cross-view feature metric on VeRi-776.

training set with 110,178 images of 13134 vehicles and a test set with 111,585 images of 13,133 vehicles. The latter is further divided into three different scale subsets: small (800 vehicles), medium (1,600 vehicles) and large (2,400 vehicles).

For the VeRi-776 dataset, the mean average precision (mAP) and cumulative matching cure (CMC) are adopted for evaluation. For the VehicleID dataset, standard CMC metric is adopted with random gallery selection.

6.2. Experiment settings

All the experiments are performed on PyTorch platform. The learning rate of the generator and the discriminator during image generation is 0.0002. During training, all input images are resized to 224×224. ResNet-50 is used as the baseline network, which is initialized using the pretrained weights of ImageNet. The learning rate decay factor and the initial learning rate are set as 5e-4 and 0.05, respectively. The batch size is set as 64, and the VANet is trained for 64 epochs.

6.3. Ablation Study

We experimentally verify the effect of parameter ε (Formula 4) on vehicle Re-ID performance. This parameter is used to assign different losses to the real and generated data. That is, the effectiveness of CVLSR in VANet is first evaluated in Tables 1 and 2. "CE" denotes cross entropy and "CVLSR" corresponds to the proposed cross-view label smoothing regularization. As shown in Table 1 and 2, when $\varepsilon = 0$ for the real and generated data, cross entropy is used on the real and generated data. Replacing cross entropy with CVLSR on the generated data can achieve superior performance compared with those of the others when we set ε as 0 and 1 on the real and generated data, respectively. This result indicates that CVLSR performs better in smoothing noise data.

To demonstrate the effectiveness of VANet, several alternative selections obtained by fusing different layers are compared to achieve an appropriate strategy. We explore the activation of four residual blocks from ResNet-50(Baseline), which correspond to Res2, Res3, Res4 and Res5(a,b) + Res5c. The experimental results are listed in Tables 3 and 4. The results indicate

Table 1
Performance evaluation on VeRi-776 by using different loss functions.

Method	Rank-1	mAP
Real(CE)	87.02	60.11
Real(CE)+ Generated(CE)	87.91	61.15
Real(CE)+ Generated(CVLSR)	**89.99**	**64.99**

Table 2
Performance evaluation on VehicleID by using different loss functions.

Method	Small		Medium		Large	
	Rank-1	Rank-5	Rank-1	Rank-5	Rank-1	Rank-5
Real(CE)	82.16	88.75	80.90	86.90	77.92	83.05
Real(CE)+ Generated(CE)	83.88	90.25	81.31	88.05	80.04	86.04
Real(CE)+ Generated(CVLSR)	**86.38**	**93.01**	**84.31**	**90.75**	**81.12**	**89.05**

Table 3
Comparison results the fusion of different layers on VeRi-776

Method	Rank-1	mAP
Baseline	89.50	61.86
Res2 + Res5(a,b) +Res5c	88.92	64.81
Res3 + Res5(a,b) +Res5c	89.51	64.50
Res4 + Res5(a,b) +Res5c(Ours)	**89.99**	**64.99**

Table 4
Comparison results the fusion of different layers on VehicleID

Method	Small		Medium		Large	
	Rank-1	Rank-5	Rank-1	Rank-5	Rank-1	Rank-5
Baseline	82.01	88.39	80.69	86.56	76.08	83.89
Res2 + Res5(a,b) +Res5c	82.39	88.51	81.69	87.00	77.12	84.89
Res3 + Res5(a,b) +Res5c	82.51	88.61	81.31	86.75	77.04	84.22
Res4 + Res5(a,b) +Res5c(Ours)	**86.38**	**93.01**	**84.31**	**90.75**	**81.12**	**89.05**

that specifically adopting VANet with CVLSR to extract robust features can improve the performance of the baseline network. In particular, Fig. 3 show that using our selected fusion layers can obtain a more robust training model that is more informative and better at capturing regions of interest.

As mentioned in Section 5.1, the proposed method is an adaptive cross-view feature aggregation. Several alternative ratios of weight strategies are set and compared with the proposed fusion strategy to analyze its effectiveness comprehensively. The cross-view feature vector ratios of the weights are varied to investigate their effect on vehicle Re-ID. The rank-1 and rank-5 accuracy results are presented in Fig. 8, indicating that the proposed feature aggregation achieves the best performance among the compared methods.

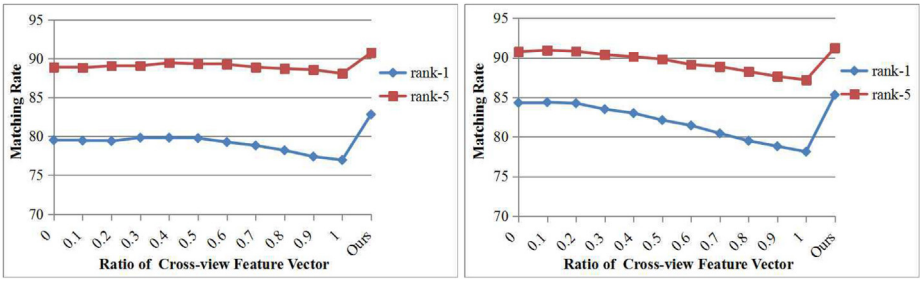

Fig. 8. Evaluation with different ratio of cross-view feature vector (1: N) on VehicleID. (a) CMC results of VANet (LSR) method on the Medium test set of VehicleID. (b) CMC results of VANet (CVLSR) method on the Medium test set of VehicleID.

6.4. Comparison with different multi-view learning methods

To validate the proposed distance metric, several recent multi-view learning methods are presented in Tables 5 and 6. Evidently, our cascaded cross-view matching approach is remarkably superior to the others. This experiment justifies that using CCM can improve robustness to multi-view matching.

6.5. Comparison with different metric learning methods

The performance of the proposed CCM method and other metrics is summarized in Tables 7 and 8, where "VANet" and "CCM" correspond to the proposed network and cascaded cross-view matching approach, respectively. In the VeRi-776 dataset, the proposed method with CVLSR exceeds rank-1 and mAP of LSR by 2.98% and 5.40%, respectively. When CVLSR and CCM are applied, the proposed method outperforms the baseline with only CVLSR by 1.49% and 3.18% in the rank-1 rate and mAP, respectively. Similarly to the results on VeRi-776, the proposed method also achieves higher improvement than other methods in Rank-1, Rank-5 on VehicleID. Furthermore, the proposed method outperforms VANet that combines re-

Table 5
Comparison with multi-view methods on VeRi-776

Method	Rank-1	mAP
VAMI [11]	77.03	50.13
XVGAN [12]	60.20	24.65
OIFE [31]	65.92	48.00
EALN [40]	84.39	57.44
VANet(CVLSR)+ CCM	**91.48**	**68.17**

Table 6
Comparison results with multi-view methods on VehicleID

Method	Small		Medium		Large	
	Rank-1	Rank-5	Rank-1	Rank-5	Rank-1	Rank-5
VAMI [11]	63.08	83.12	52.69	75.08	47.28	70.06
XVGAN [12]	52.79	80.69	49.47	71.42	44.92	66.72
EALN [40]	75.11	88.09	71.78	83.94	69.30	81.42
VANet(CVLSR)+ CCM	**87.50**	**93.88**	**85.31**	**91.25**	**82.00**	**89.88**

Table 7
Comparison results with metric learning methods on VeRi-776

Method	Rank-1	mAP
VANet(LSR)	87.01	59.59
VANet(LSR) + Re-ranking [43]	88.02	65.01
VANet(LSR) + CCM	89.81	65.12
VANet(CVLSR)	89.99	64.99
VANet(CVLSR)+ Re-ranking [43]	90.41	68.13
VANet(CVLSR)+ CCM	**91.48**	**68.17**

Table 8
Comparison results with metric learning methods on VehicleID

Method	Small		Medium		Large	
	Rank-1	Rank-5	Rank-1	Rank-5	Rank-1	Rank-5
VANet(LSR)	82.00	92.13	79.50	88.87	73.96	86.08
VANet(LSR)+ Re-ranking [43]	82.35	92.78	81.06	90.06	75.71	87.13
VANet(LSR)+ CCM	85.38	92.88	82.81	90.75	78.13	88.29
VANet(CVLSR)	86.38	93.01	84.31	90.75	81.12	89.05
VANet(CVLSR)+ Re-ranking [43]	87.25	93.65	85.12	91.05	81.75	89.46
VANet(CVLSR)+ CCM	**87.50**	**93.88**	**85.31**	**91.25**	**82.00**	**89.88**

Table 9
The performance comparison of state-of-the-art methods on VeRi-776

Method	Rank-1	mAP
Siamese-Visual [45]	41.12	29.48
BOW-CN [46]	33.82	9.63
FACT [44]	51.89	18.69
OIFE [31]	65.92	48.00
NuFACT [48]	76.76	48.47
VRSDNet [47]	83.49	53.45
VR-PROUD [27]	83.19	40.50
FDA-Net [28]	84.27	55.49
QD-DLF [35]	88.50	61.83
RAM [49]	88.60	61.50
VANet	87.02	60.11
VANet(CVLSR)	89.99	64.99
VANet(CVLSR)+CCM	**91.48**	**68.17**

Table 10
The performance comparison of state-of-the-art methods on VehicleID

Method	Small		Medium		Large	
	Rank-1	Rank-5	Rank-1	Rank-5	Rank-1	Rank-5
VGG + CCL [37]	43.92	65.01	38.84	61.91	34.58	55.72
MixedDiff + CCL [37]	48.52	74.55	43.94	67.96	40.85	62.79
NuFACT [48]	48.90	69.51	43.64	65.34	38.63	60.72
FACT [44]	49.93	68.37	45.01	64.75	40.12	60.59
VRSDNet [47]	56.98	86.90	50.57	80.05	42.92	73.44
C2F-Rank [30]	61.10	81.70	56.20	76.20	51.40	72.20
VAMI [11]	63.08	83.12	52.69	75.08	47.28	70.06
TAMR [36]	66.02	79.71	62.90	76.80	59.69	73.87
Mob.VFL* [26]	73.37	85.52	69.52	81.00	67.41	78.48
QD-DLF [35]	72.32	92.48	70.66	88.90	64.14	83.37
RAM [49]	75.20	91.50	72.30	87.00	67.70	84.50
VANet	82.16	88.75	80.90	86.90	77.92	83.05
VANet(CVLSR)	86.38	93.01	84.31	90.75	81.12	89.05
VANet(CVLSR)+CCM	**87.50**	**93.88**	**85.31**	**91.25**	**82.00**	**89.88**

ranking with either LSR or CVLSR. Thus, the results confirm that the presented CCM method integrates multi-view information better than the other metric methods.

6.6. Comparison with state-of-the-art methods

The proposed framework is compared with several state-of-the-art vehicle Re-ID methods. The comparison results of the proposed approach and other state-of-art methods, including FACT [44], Siamese-Visual[45], BOW-CN [46], VGG + CCL [37], MixedDiff + CCL [37], VAMI [11], XVGAN [12], OIFE [31], VRSDNet [47], VR-PROUD [27], FDA-Net [28], C2F-Rank [30], TAMR [36], Mob.VFL* [26], NuFACT [48], QD-DLF[35], and RAM[49], are summarized in Tables 9 and 10. The proposed method achieves 91.48% rank-1 accuracy and 68.17% mAP in the VeRi-776 dataset, which are higher than those of the other methods. The proposed method in the three subsets of the VehicleID dataset reaches 87.50%, 85.31%, and 82.00% rank-1 accuracy and 93.88%, 91.25% and 89.88% rank-5 rate, signifying that it achieves the best performance among the compared methods. In conclusion, VANet demonstrates better multi-level feature extraction performance than the other methods.

7. Conclusion

A VANet with a cross-view distance metric for robust vehicle Re-ID is proposed in this paper. This scheme consists of two major components: viewpoint adaptation learning and cross-view distance metric. It combines multi-level features to obtain a robust training model. In particular, CVLSR is applied to the generated data to smooth image-image translation noise. Moreover, a CCM method is designed to merge the features of the input and generated images to preserve the integrity of vehicle viewpoint during the multi-view matching process. The results show that the proposed method is robust to viewpoint variations, and it outperforms other state-of-the-art approaches for vehicle Re-ID.

CRediT authorship contribution statement

Qi Wang: Conceptualization, Methodology, Formal analysis, Data curation, Software, Writing - original draft. **Weidong Min:** Validation, Visualization, Supervision, Funding acquisition. **Qing Han:** Validation, Writing - review & editing. **Ziyuan Yang:** Validation, Writing - review & editing. **Xin Xiong:** Writing - review & editing. **Meng Zhu:** Writing - review & editing. **Haoyu Zhao:** Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (Grant No. 61762061 and 62076117), the Natural Science Foundation of Jiangxi Province, China (Grant No. 20161ACB20004) and Jiangxi Key Laboratory of Smart City (Grant No. 20192BCD40002).

References

[1] H. Ling, Z. Wang, P. Li, Y. Shi, J. Chen, F. Zou, Improving person re-identification by multi-task learning, Neurocomputing 347 (2019) 109–118, https://doi.org/10.1016/j.neucom.2019.01.027.
[2] Y. Wu, Y. Lin, X. Dong, Y. Yan, W. Ouyang, Y. Yang, Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5177–5186. doi:10.1109/CVPR.2018.00543..
[3] W. Deng, L. Zheng, Q. Ye, G. Kang, Y. Yang, J. Jiao, Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 994–1003. doi:10.1109/CVPR.2018.00110..
[4] X. Xiong, W. Min, W.-S. Zheng, P. Liao, H. Yang, S. Wang, S3d-cnn: skeleton-based 3d consecutive-low-pooling neural network for fall detection, Appl. Intell. (2020), https://doi.org/10.1007/s10489-020-01751-y.
[5] H. Yang, L. Liu, W. Min, X. Yang, X. Xiong, Driver yawning detection based on subtle facial action recognition, IEEE Trans. Multimedia (2020), https://doi.org/10.1109/TMM.2020.2985536.
[6] Q. Wang, W. Min, D. He, S. Zou, T. Huang, Y. Zhang, R. Liu, Discriminative fine-grained network for vehicle re-identification using two-stage re-ranking, Sci. China Inform. Sci. (2020), https://doi.org/10.1007/s11432-019-2811-8.
[7] L. Zhou, W. Min, D. Lin, Q. Han, R. Liu, Detecting motion blurred vehicle logo in iov using filter-deblurgan and VL-YOLO, IEEE Trans. Veh. Technol. 69 (4) (2020) 3604–3614, https://doi.org/10.1109/TVT.2020.2969427.
[8] W. Min, M. Fan, X. Guo, Q. Han, A new approach to track multiple vehicles with the combination of robust detection and two classifiers, IEEE Trans. Intell. Transp. Syst. 19 (1) (2018) 174–186, https://doi.org/10.1109/TITS.2017.2756989.
[9] J. Zhu, T. Park, P. Isola, A. A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2242–2251. doi:10.1109/ICCV.2017.244..
[10] X. Peng, Z. Huang, J. Lv, H. Zhu, J. T. Zhou, COMIC: multi-view clustering without parameter selection, in: International Conference on Machine Learning, (ICML), Vol. 97, 2019, pp. 5092–5101..
[11] Y. Zhou, L. Shao, Viewpoint-aware attentive multi-view inference for vehicle re-identification, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6489–6498. doi:10.1109/CVPR.2018.00679..
[12] Y. Zhou, L. Shao, Cross-view GAN based vehicle generation for re-identification, in: British Machine Vision Conference (BMVC), 2017..
[13] Z. Zhong, L. Zheng, Z. Zheng, S. Li, Y. Yang, Camstyle: A novel data augmentation method for person re-identification, IEEE Trans. Image Process. 28 (3) (2019) 1176–1190, https://doi.org/10.1109/TIP.2018.2874313.
[14] J. Wang, X. Zhu, S. Gong, W. Li, Transferable joint attribute-identity deep learning for unsupervised person re-identification, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2275–2284. doi:10.1109/CVPR.2018.00242..
[15] X. Zhang, J. Cao, C. Shen, M. You, Self-training with progressive augmentation for unsupervised cross-domain person re-identification, in: IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 8221–8230. doi:10.1109/ICCV.2019.00831..
[16] J. Yu, D. Ko, H. Moon, M. Jeon, Deep discriminative representation learning for face verification and person re-identification on unconstrained condition, in: IEEE International Conference on Image Processing (ICIP), 2018, pp. 1658–1662. doi:10.1109/ICIP.2018.8451494..
[17] F. Yang, K. Yan, S. Lu, H. Jia, X. Xie, W. Gao, Attention driven person re-identification, Pattern Recognit. 86 (2019) 143–155, https://doi.org/10.1016/j.patcog.2018.08.015.
[18] L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang, Q. Tian, Person re-identification in the wild, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3346–3355. doi:10.1109/CVPR.2017.357..
[19] Y. Lin, X. Dong, L. Zheng, Y. Yan, Y. Yang, A bottom-up clustering approach to unsupervised person re-identification, in: Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence, AAAI, 2019, pp. 8738–8745, https://doi.org/10.1609/aaai.v33i01.33018738.
[20] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, X. Chen, VRSTC: occlusion-free video person re-identification, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7183–7192. doi:10.1109/CVPR.2019.00735..
[21] F. Ma, X. Jing, X. Zhu, Z. Tang, Z. Peng, True-color and grayscale video person re-identification, IEEE Trans. Inform. Forensics Security 15 (2020) 115–129, https://doi.org/10.1109/TIFS.2019.2917160.
[22] Y. He, C. Dong, Y. Wei, Combination of appearance and license plate features for vehicle re-identification, in: IEEE International Conference on Image Processing (ICIP), 2019, pp. 3108–3112. doi:10.1109/ICIP.2019.8803323..
[23] S. M. Silva, C. R. Jung, License plate detection and recognition in unconstrained scenarios, in: Europeon Conference on Computer Vision (ECCV), Vol. 11216, 2018, pp. 593–609. doi:10.1007/978-3-030-01258-8_36..
[24] Y. Li, Y. Li, H. Yan, J. Liu, Deep joint discriminative learning for vehicle re-identification and retrieval, in: IEEE International Conference on Image Processing (ICIP), 2017, pp. 395–399. doi:10.1109/ICIP.2017.8296310..
[25] M. Wu, Y. Zhang, T. Zhang, W. Zhang, Background segmentation for vehicle re-identification, in: MultiMedia Modeling (MMM), Vol. 11962, 2020, pp. 88–99. doi:10.1007/978-3-030-37734-2_8..
[26] S. A. S. Alfasly, Y. Hu, T. Liang, X. Jin, Q. Zhao, B. Liu, Variational representation learning for vehicle re-identification, in: IEEE International Conference on Image Processing (ICIP), 2019, pp. 3118–3122. doi:10.1109/ICIP.2019.8803366..
[27] R.M.S. Bashir, M. Shahzad, M.M. Fraz, VR-PROUD: vehicle re-identification using progressive unsupervised deep architecture, Pattern Recognit. 90 (2019) 52–65, https://doi.org/10.1016/j.patcog.2019.01.008.

[28] Y. Lou, Y. Bai, J. Liu, S. Wang, L. Duan, Veri-wild: A large dataset and a new method for vehicle re-identification in the wild, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3235–3243. doi:10.1109/CVPR.2019.00335..
[29] J. Peng, H. Wang, T. Zhao, X. Fu, Cross domain knowledge transfer for unsupervised vehicle re-identification, in: IEEE International Conference on Multimedia & Expo Workshops, (ICME), 2019, pp. 453–458. doi:10.1109/ICMEW.2019.00084..
[30] H. Guo, C. Zhao, Z. Liu, J. Wang, H. Lu, Learning coarse-to-fine structured feature embedding for vehicle re-identification, in, in: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI), 2018, pp. 6853–6860.
[31] Z. Wang, L. Tang, X. Liu, Z. Yao, S. Yi, J. Shao, J. Yan, S. Wang, H. Li, X. Wang, Orientation invariant feature embedding and spatial temporal regularization for vehicle re-identification, in: IEEE International Conference on Computer Vision (ICCV), 2017, pp. 379–387. doi:10.1109/ICCV.2017.49..
[32] H. Chen, B. Lagadec, F. Brémond, Partition and reunion: A two-branch neural network for vehicle re-identification, in: IEEE Conference on Computer Vision and Pattern Recognition Workshops, (CVPR), 2019, pp. 184–192..
[33] Z. Zheng, L. Zheng, Y. Yang, Pedestrian alignment network for large-scale person re-identification, IEEE Trans. Circuits Syst. Video Techn. 29 (10) (2019) 3037–3045, https://doi.org/10.1109/TCSVT.2018.2873599.
[34] B. He, J. Li, Y. Zhao, Y. Tian, Part-regularized near-duplicate vehicle re-identification, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3997–4005. doi:10.1109/CVPR.2019.00412..
[35] J. Zhu, H. Zeng, J. Huang, S. Liao, Z. Lei, C. Cai, L. Zheng, Vehicle re-identification using quadruple directional deep learning features, IEEE Trans. Intell. Transp. Syst. 21 (1) (2020) 410–420, https://doi.org/10.1109/TITS.2019.2901312.
[36] H. Guo, K. Zhu, M. Tang, J. Wang, Two-level attention network with multi-grain ranking loss for vehicle re-identification, IEEE Trans. Image Process. 28 (9) (2019) 4328–4338, https://doi.org/10.1109/TIP.2019.2910408.
[37] H. Liu, Y. Tian, Y. Wang, L. Pang, T. Huang, Deep relative distance learning: Tell the difference between similar vehicles, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2167–2175. doi:10.1109/CVPR.2016.238..
[38] R. Kumar, E. Weill, F. Aghdasi, P. Sriram, Vehicle re-identification: an efficient baseline using triplet embedding, in: IEEE International Joint Conference on Neural Networks (IJCNN), 2019, pp. 1–9. doi:10.1109/IJCNN.2019.8852059..
[39] A. Brock, J. Donahue, K. Simonyan, Large scale GAN training for high fidelity natural image synthesis, in: International Conference on Learning Representations (ICLR), 2019..
[40] Y. Lou, Y. Bai, J. Liu, S. Wang, L. Duan, Embedding adversarial learning for vehicle re-identification, IEEE Trans. Image Process. 28 (8) (2019) 3794–3807, https://doi.org/10.1109/TIP.2019.2902112.
[41] Z. Zheng, L. Zheng, Y. Yang, Unlabeled samples generated by GAN improve the person re-identification baseline in vitro, in: IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3774–3782. doi:10.1109/ICCV.2017.405..
[42] Q. Yu, X. Chang, Y. Song, T. Xiang, T. M. Hospedales, The devil is in the middle: Exploiting mid-level representations for cross-domain instance matching, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017. arXiv:1711.08106..
[43] Z. Zhong, L. Zheng, D. Cao, S. Li, Re-ranking person re-identification with k-reciprocal encoding, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3652–3661. doi:10.1109/CVPR.2017.389..
[44] X. Liu, W. Liu, H. Ma, H. Fu, Large-scale vehicle re-identification in urban surveillance videos, in: IEEE International Conference on Multimedia and Expo (ICME), 2016, pp. 1–6. doi:10.1109/ICME.2016.7553002..
[45] Y. Shen, T. Xiao, H. Li, S. Yi, X. Wang, Learning deep neural networks for vehicle re-id with visual-spatio-temporal path proposals, in: IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1918–1927. doi:10.1109/ICCV.2017.210..
[46] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, Q. Tian, Scalable person re-identification: A benchmark, in: IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1116–1124. doi:10.1109/ICCV.2015.133..
[47] J. Zhu, Y. Du, Y. Hu, L. Zheng, C. Cai, Vrsdnet: vehicle re-identification with a shortly and densely connected convolutional neural network, Multim. Tools Appl. 78 (20) (2019) 29043–29057, https://doi.org/10.1007/s11042-018-6270-4.
[48] X. Liu, W. Liu, T. Mei, H. Ma, PROVID: progressive and multimodal vehicle reidentification for large-scale urban surveillance, IEEE Trans. Multimedia 20 (3) (2018) 645–658, https://doi.org/10.1109/TMM.2017.2751966.
[49] X. Liu, S. Zhang, Q. Huang, W. Gao, RAM: A region-aware deep model for vehicle re-identification, in: IEEE International Conference on Multimedia and Expo ICME, IEEE Computer Society, 2018, pp. 1–6. doi:10.1109/ICME.2018.8486589..