Partial Attention-Based Direction-Aware Vehicle Re-identification

Yujie Zhou, Caihong Yuan(✉), Chenshuang Su, Mingdong Zou, Xiaoke Zhu, and Wenjuan Liang

School of Computer and Information Engineering, Henan University Henan Engineering Research Center of Intelligent Technology and Application, Henan Province Spatial Information Processing Engineering Technology Research Center, Kaifeng, China
yuanch@henu.edu.cn, whuzxk@whu.edu.cn, 10120085@vip.henu.edu.cn

Abstract. With the rapid development of urban transportation, vehicle re-identification has become a focal point in traffic management and vehicle tracking problems. In order to address the problem of small inter-class similarity among vehicles, previous studies utilize vehicle parsing models to extract local features. Therefore, we introduce the Squeeze-and-Excitation attention mechanism to extract important discriminative information from these local features. Furthermore, we propose a local co-occurrence attention mechanism to represent the proportion of common parts feature matching. To address the issue of large intra-class differences caused by vehicle direction change, we propose a lightweight and effective direction weighted fusion strategy. Experiments on two large datasets show that the proposed algorithm performs competitively.

Keywords: Attention mechanism · Discriminative local feature · Feature alignment

1 Introduction

Vehicle Re-identification (ReID) is an image retrieval problem that aims to find the most similar images in the gallery captured in another camera views. This task plays an important role in intelligent transportation and city surveillance systems [3,23]. For example, the police may quickly lock the suspect's motion trajectory through vehicle re-identification. However, due to the similar color and type, or the same brand, different vehicles may have similar appearance. And affected by different illumination, different vehicle direction, etc. , the same vehicle may look very different in different camera views.

Some previous works learn global feature [1,2,6,13,22,25], but it is difficult to distinguish vehicles solely based on global features. Different vehicles with similar appearances need to be distinguished by subtle differences in local features. Therefore, some works learn both global and local feature. For example, Zhao et al. [20] extracts local features by designing a gradually expanding circular ROI

W. Hong and G. Kanaparan (Eds.): ICCSE 2023, CCIS 2023, pp. 162–172, 2024.
https://doi.org/10.1007/978-981-97-0730-0_15

projection. Liu et al. [11] obtains local features by horizontal splitting the feature map. But simple division of vehicles can result in misalignment issues. Therefore, vehicle ReID must learn vehicle component to capture subtle differences. He et al. [4] uses vehicle predefined regions to learn more discriminative regions. However, this method neglects that discriminative differences between vehicles may appear in any part of the vehicle. Zhang et al. [19] extracts vehicle components using object detection model. This method could achieve more accurate locate information. However, these local features are just simply separated from the global, which may reduce to a suboptimal performance. Therefore, we first use vehicle parsing model to parse the vehicle into four different views (front, back, top, and side), and then introduce the Squeeze-and-Excitation attention mechanism(SE) to learn the important information in the local features. Furthermore, we propose a local co-occurrence attention mechanism(LCA) to discover the proportion of co-occurring parts of two instances, for automatically the importance of each part features, and helping to overcome the challenge of the small differences of similar vehicles in different viewpoints.

Vehicle direction change is a very large challenge, which could make the same vehicle look very different. But vehicle ReID with direction changing is not well studies. For example, Zhu et al. [24] replaces the vehicle ID of a network model with a direction ID to identify the direction. Tang et al. [16] determines the direction through a pose estimation model. Teng et al. [17] designs a multi-view branch network that uses CNN as a viewpoint classifier, each branch learning features specific to a particular viewpoint. Although this method has achieved some success, its disadvantage is that multiple models are complex, training is difficult, and the important information of local features is not fully mined. Some works do not use directional classification model to determine the direction of vehicles. Zhu et al. [22] uses different directional pooling layers to compress the feature maps into horizontal, vertical, diagonal, and anti-diagonal directional feature maps respectively. Finally, these feature maps are spatially normalized and concatenated into four directional deep learning features. Chu et al. [2] divides feature space into similar perspectives and different perspectives, and learns two constraints to improve recognition accuracy. The disadvantages of these methods is neglect the learning of local features. However our method combines global features with local features, fully exploring the important information of local features. We propose a lightweight and effective direction weighted fusion strategy (DF). This strategy determines the similarity of vehicle directions through four components, and automatically adjusts the distance of image features based on their similarity, so as to improve recognition accuracy.

We evaluated our approach on two widely used large vehicle datasets, VehicleID and VeRi776. The experimental results show that our method is competitive. The main contributions of this work can be summarized as follows:

1) We propose a local co-occurrence attention mechanism, which aims to improve the alignment of local features by focusing on the proportion of common parts that appear in two images.
2) We introduce the SE attention mechanism to learn subtle differences in features with greater robustness.
3) We design a direction weighted fusion that determines the vehicle's direction based on its components. This helps to mitigate the re-identification deviation caused by changes in viewpoint.

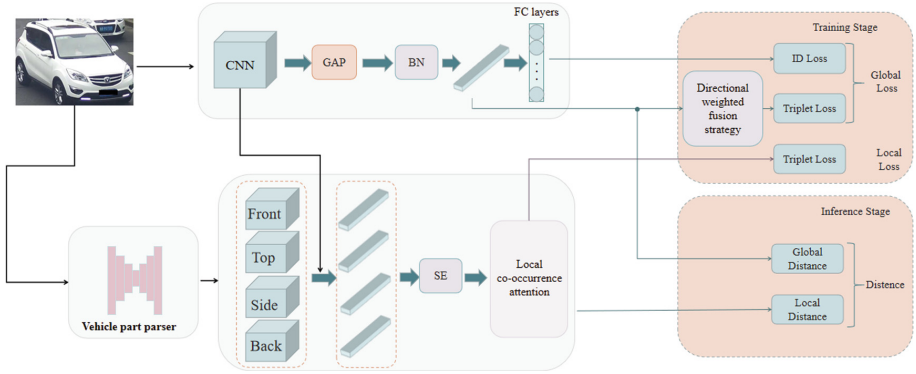

Fig. 1. An overview of the proposed system. It consists of two blocks for global and local feature learning, respectively. "GAP" denotes global average pooling, "BN" denotes batch normalization, and "FC layers" denotes fully connected layers.

2 Methodology

In Sect. 2.1, we show the basic architecture of this paper. In Sect. 2.2, SE attention mechanism is introduced. In Sect. 2.3, local co-occurrence attention mechanism is introduced. Finally, we will introduce the direction weighted fusion strategy in Sect. 2.4.

2.1 Network Architecture

Our network architecture is shown in Fig. 1.

Backbone. We use ResNet50 [5], pre-trained on the ImageNet [15] dataset, as our feature extractor. As shown in Fig. 1, the feature extractor network has two output branches. The first branch is the global branch, which is used to obtain a feature map of the overall appearance of the vehicle. Another branch is the local branch, which obtains local features through vehicle part parser.

2.2 Squeeze-and-Excitation Attention Mechanism

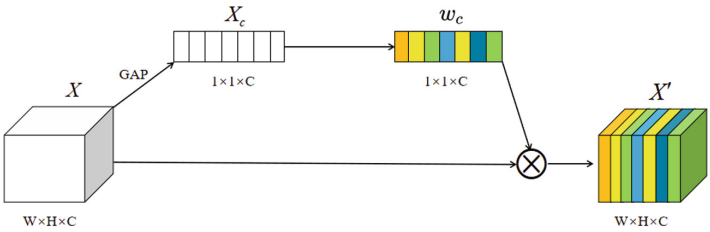

Fig. 2. The structure of SE attention mechanism.

SE attention mechanism is introduced into local branch to enhance the features with a high discriminative degree and suppress the interference of features with a low discriminative degree. This enhancement aims to improve the network's capacity to learn subtle differences. The structure of SE attention module is illustrated in Fig. 2.

The input feature map X is first compressed along the spatial dimension using a global average pooling layer. Then, two FC layers are used to calculate the weight for each feature channel. Finally, the learned features weights w_c of each channel are multiplied by X to obtain a new feature map $X^{'}$. The SE attention mechanism is formulated as follows.

$$X_c = GAP(X) = \frac{1}{H \times W} \sum_{i}^{H} \sum_{j}^{W} X(i,j) \tag{1}$$

$$w_c = \sigma_2(w_2 \sigma_1(w_1 X_c)) \tag{2}$$

$$X^{'} = w_c \bullet X \tag{3}$$

GAP is the global average pooling, w_1 and w_1 are the weight parameters of the two fully connected layers, and σ_1 and σ_2 are the ReLU and Sigmoid activation functions, respectively.

2.3 Local Co-occurrence Attention Mechanism

The local branch is designed with local co-occurrence attention to achieve local features alignment, avoiding the features mismatch problem.

Firstly, We use ResNet50 to obtain a $16 \times 16 \times 2048$ feature map F. Secondly, we pool vehicle parsing mask to 16×16 by max pooling, which is defined as $\{M_i | i \in \{1,2,3,4\}\}$. Thirdly, We multiply F with the mask to obtain local features $\{f_i | i \in \{1,2,3,4\}\}$. They represent the front, back, side, and top views of the car.

Given two images p, q and their masks M_i^p and M_i^q, we calculate the visible scores v_i^p and v_i^q for each part, which represent the size of each region of the part. The visible score v_i is defined as

$$v_i = \sum_{j,k=1}^{16} M_i(j,k) \tag{4}$$

We compute the matching score $C_i^{p,q}$ as follows.

$$C_i^{p,q} = \frac{\frac{v_i^p v_i^q}{\left|v_i^p - v_i^q\right|}}{\sum_{i=1}^{N} \frac{v_i^p v_i^q}{\left|v_i^p - v_i^q\right|}} \tag{5}$$

where $C_i^{p,q}$ measures the matching scores of each corresponding component of the two images. N is the number of local features. Then, the distance of local features $\hat{D}_l^{p,q}$ between the two vehicles is calculated as

$$\hat{D}_l^{p,q} = \sum_{i=1}^{N} C_i^{p,q} D\left(f_i^p, f_i^q\right) \tag{6}$$

where D denotes the Euclidean distance. If there are missing parts of the vehicle, the corresponding area visible score will be relatively small, resulting in a low matching score. A higher matching score indicates a larger proportion of the matched area. In this paper, we optimize the network by constructing ID loss and triplet loss for global feature, as well as triplet loss for local features. The triplet loss of local features is calculated as.

$$L_{triplet}^l = max(\hat{D}_l^{ap} - \hat{D}_l^{an} + \gamma, 0) \tag{7}$$

2.4 Direction Weighted Fusion Strategy

To correct the bias caused by the viewpoint and to better expand the role of local features, we propose a direction weighted fusion strategy.

$w^{p,q}$ denotes the direction similarity between two vehicle images and is defined as follows.

$$w^{p,q} = \frac{\sum_{i=1}^{N} f(\frac{v_i^p}{v_i^q})}{\sigma} \tag{8}$$

$$f(x) = \begin{cases} \frac{1}{x} & \text{if } x > 1 \\ x & \text{otherwise} \end{cases} \tag{9}$$

The larger the direction similarity w^{pq} is, the closer the two vehicles are. For Eq. (8), we experimentally conclude that it works best when $\sigma = 6$. Then, the global feature distance between the two vehicles is calculated D as following,

$$\hat{D}_g^{p,q} = w^{p,q} D\left(f^p, f^q\right) \tag{10}$$

The triplet loss of the global feature is computed from the distance of the above global feature as:

$$L_{triplet}^{g} = max(\hat{D}_{g}^{ap} - \hat{D}_{g}^{an} + \gamma, 0) \tag{11}$$

Finally, the total loss in this paper contains the following loss functions:

$$L = L_{id}^{g} + L_{triplet}^{g} + L_{triplet}^{l} \tag{12}$$

3 Experiments

3.1 Datasets

We evaluate our model on two popular vehicle datasets, including VeRi776 and VehicleID.

VeRi776 [12] is the benchmark dataset for the vehicle task. It consists of about 50,000 images of 776 vehicles captured by 20 cameras with different viewpoints. The training set contains 576 vehicles and the test set contains another 200 vehicles.

VehicleID [9] is a large-scale vehicle re-identification benchmark dataset. It contains a total of 221,763 images of about 26,267 vehicles. The images in the dataset are taken in either front or back view. Three test sets, small, medium, and large, are extracted based on size of the test set. In the inference phase, one image is randomly selected for each car as a gallery set, and other images are used as query images.

3.2 Implementation Details

We train models for 120 epochs with warm-up strategy. The initial learning rate is 3.5e-5 and increases to 3.5e-4 after the 10th calendar element. We first fill the image boundary with 10 pixels and then randomly crop it to 256×256. We also augment the data by random erasure using Adam as the optimizer.

To evaluate our method, we first compute the Euclidean distance $\hat{D}_{gloabal}$ between global features. Then, we compute the distance $\hat{D}_{local}$ between the local features defined in Eq. 6. The final distance between the query set and the gallery set is computed as $\lambda_1 \hat{D}_{gloabal} + \lambda_2 \hat{D}_{local}$. Here, we set $\lambda_1 = 1$ and $\lambda_2 = 0.5$.

3.3 Experiments on VeRi776 Dataset

We evaluate our method on the VeRi776 dataset. Table 1 shows the performance comparison between our proposed method and other methods. In the Baseline method, LCA, SE attention, and DF are removed. From the results, it can be seen that recent mainstream vehicle ReId methods combine the learning of global and local features, which greatly improves their effectiveness on the VeRi776 dataset. Compared with the baseline, our method improves by 4.1% on mAP, and 1.8 % on CMC@1. Other than that, our method improves both mAP and CMC@1 over the other methods, and CMC@5 improves over the majority of other methods.

Table 1. The mAP, CMC@1 and CMC@5 on VeRi776

Method	mAp	CMC@1	CMC@5
FACT [12]	0.185	0.510	0.735
OIFE [18]	0.480	0.894	-
VAMI [21]	0.501	0.770	0.908
PROVID [13]	0.534	0.816	0.951
EALN [14]	0.574	0.844	0.941
AAVER [7]	0.612	0.890	0.947
RAM [10]	0.615	0.886	0.940
VANET [2]	0.663	0.897	0.959
PAMTRI [16]	0.718	0.929	0.970
PRN [4]	0.743	0.943	**0.989**
PGAN [19]	0.793	0.965	0.983
SAVER [8]	0.796	0.964	0.986
Baseline	0.759	0.948	0.978
Ours	**0.800**	**0.966**	0.982

3.4 Experiments on VehicleID Dataset

We compare the scores of CMC@1 and CMC@5 on this dataset because there is only one ground truth for each query vehicle. Table 2 gives the comparison results for three different sizes of test datasets. In the Baseline method, LCA, SE attention, and DF are removed. We observed that, when the scale is small, our method improves the CMC@1 by 1.3% compared to the baseline. At a medium scale, our approach achieves 6% and 4% improvement over baseline at CMC@1 and CMC@5, respectively. At a large scale, our approach improved 7.7% and 6.1% over baseline at CMC@1 and CMC@5, respectively. Compared to other methods, our method is superior to the majority of other methods. The above comparison result proves that our method is effective in not only improving retrieval accuracy but also enhancing the capacity to identify more challenging samples.

4 Ablation Study

4.1 The Effects of Key Components

In this section, we perform ablation experiments to evaluate the contribution of each part. LCA is the local co-occurrence attention mechanism, SE is the SE attention mechanism added to local branches, and DF is the direction weighted fusion strategy. The baseline is to remove LCA, SE attention mechanism, and DF. The effectiveness of each part is shown in Table 3.

Table 2. The CMC@1 and CMC@5 on VehicleID

Method	small		medium		large	
	@1	@5	@1	@5	@1	@5
OIFE [18]	–	–	–	–	0.670	0.829
VAMI [21]	0.631	0.833	0.529	0.751	0.473	0.703
AAVER [7]	0.747	0.938	0.686	0.900	0.635	0.856
EALN [14]	0.751	0.881	0.718	0.839	0.693	0.814
RAM [10]	0.752	0.915	0.723	0.870	0.677	0.845
PRN [4]	0.784	0.923	0.750	0.883	0.742	0.864
SAVER [8]	0.799	0.952	0.776	0.911	0.753	0.883
PGAN [19]	–	–	–	–	0.778	0.921
VANET [2]	**0.881**	**0.972**	0.831	0.951	0.803	0.929
Baseline	0.821	0.962	0.779	0.927	0.758	0.904
Ouers	0.834	0.964	**0.839**	**0.967**	**0.835**	**0.965**

Table 3. Ablation study about each part on VeRi776

settings	mAP	CMC@1	CMC@5
Baseline	0.759	0.948	0.978
+LCA	0.794	0.956	0.979
+SE	0.794	0.964	0.979
+LCA+SE	0.796	0.958	**0.985**
+LCA+SE+DF	**0.800**	**0.966**	0.982

Compared to the baseline, the separate local co-occurrence attention mechanism and the addition of the SE attention mechanism both increased mAP by 3.5%. When learning features together with local co-occurrence attention and SE attention, the mAP accuracy increased by 3.7% and CMC@1 by 1%. On the basis of the above results, adding DF, the map increased by 0.4% and CMC@1 increased by 0.8%. The result indicates that the designed parts can effectively identify discriminative features and improve the accuracy of recognition.

5 Qualitative Analysis

Figure 3 shows the qualitative results of our method on the VeRi776 dataset, where the top 5 predictions are contained in the corresponding query image. Our method is better able to retrieve the correct image when the query image is in a different viewpoint from the target image. Also our method recognizes the correct image when the image appears to be occluded. This shows that our proposed method can better match local features and reduce the effect of viewpoint on recognition compared to the baseline.

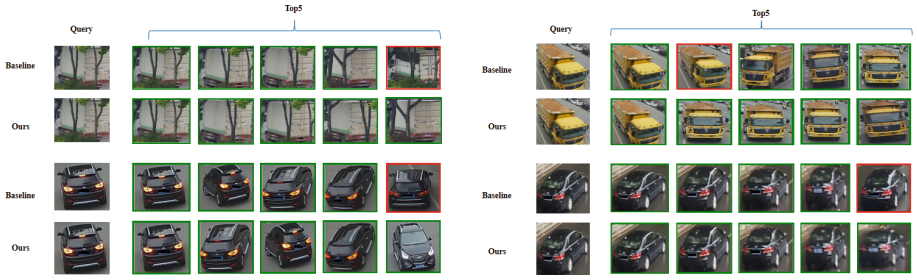

Fig. 3. Visualization of the ranking list on VeRi776. The images in the first column are query images. The remaining images are retrieved from the top 5 ranking results. The correctly retrieved images are indicated by a green border, while false instances are indicated by a red border. (Color figure online)

6 Conclusion

In this article, we propose a new features learning framework. This framework combines global features and local features based on vehicle parsers for joint learning. The Squeeze-and-Excitation attention mechanism is introduced to extract distinctive local features. For the matching of local features, local co-occurrence attention mechanism is designed to better measure the matching of vehicle parts. To reduce the impact of orientation on recognition accuracy, we propose a direction weighted fusion strategy. We evaluate our method on two large-scale vehicle ReID datasets. Experimental results demonstrate the effectiveness of our method.

Acknowledgment. This work was supported by the Young Scientists Fund of the National Natural Science Foundation of China (No. 62006070), and partly supported by Key Scientific and Technological Project of Henan Province of China (Nos. 222102210197,222102210204,232102211013 and 222102210238).

References

1. Chen, X., Sui, H., Fang, J., Feng, W., Zhou, M.: Vehicle re-identification using distance-based global and partial multi-regional feature learning. IEEE Trans. Intell. Transp. Syst. **22**(2), 1276–1286 (2021). https://doi.org/10.1109/TITS.2020.2968517
2. Chu, R., Sun, Y., Li, Y., Liu, Z., Zhang, C., Wei, Y.: Vehicle re-identification with viewpoint-aware metric learning. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 8281–8290 (2019). https://doi.org/10.1109/ICCV.2019.00837
3. Guo, H., Zhao, C., Liu, Z., Wang, J., Lu, H.: Learning coarse-to-fine structured feature embedding for vehicle re-identification. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence. AAAI'18/IAAI'18/EAAI'18, AAAI Press (2018)

4. He, B., Li, J., Zhao, Y., Tian, Y.: Part-regularized near-duplicate vehicle re-identification. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3992–4000 (2019). https://doi.org/10.1109/CVPR.2019.00412
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90
6. Huynh, S.V., Nguyen, N.H., Nguyen, N.T., Nguyen, Q.V., Huynh, C., Nguyen, C.: A strong baseline for vehicle re-identification. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 4142–4149 (2021). https://doi.org/10.1109/CVPRW53098.2021.00468
7. Khorramshahi, P., Kumar, A., Peri, N., Rambhatla, S.S., Chen, J.C., Chellappa, R.: A dual-path model with adaptive attention for vehicle re-identification. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 6131–6140 (2019). https://doi.org/10.1109/ICCV.2019.00623
8. Khorramshahi, P., Peri, N., Chen, J.c., Chellappa, R.: The devil is in the details: self-supervised attention for vehicle re-identification. In: Computer Vision - ECCV 2020, pp. 369–386 (2020)
9. Liu, H., Tian, Y., Wang, Y., Pang, L., Huang, T.: Deep relative distance learning: tell the difference between similar vehicles. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2167–2175 (2016). https://doi.org/10.1109/CVPR.2016.238
10. Liu, X., Zhang, S., Huang, Q., Gao, W.: Ram: a region-aware deep model for vehicle re-identification. In: 2018 IEEE International Conference on Multimedia and Expo (ICME), pp. 1–6 (2018). https://doi.org/10.1109/ICME.2018.8486589
11. Liu, X., Zhang, S., Wang, X., Hong, R., Tian, Q.: Group-group loss-based global-regional feature learning for vehicle re-identification. IEEE Trans. Image Process. **29**, 2638–2652 (2020). https://doi.org/10.1109/TIP.2019.2950796
12. Liu, X., Liu, W., Ma, H., Fu, H.: Large-scale vehicle re-identification in urban surveillance videos. In: 2016 IEEE International Conference on Multimedia and Expo (ICME), pp. 1–6 (2016). https://doi.org/10.1109/ICME.2016.7553002
13. Liu, X., Liu, W., Mei, T., Ma, H.: PROVID: progressive and multimodal vehicle reidentification for large-scale urban surveillance. IEEE Trans. Multimedia **20**(3), 645–658 (2018). https://doi.org/10.1109/TMM.2017.2751966
14. Lou, Y., Bai, Y., Liu, J., Wang, S., Duan, L.Y.: Embedding adversarial learning for vehicle re-identification. IEEE Trans. Image Process. **28**(8), 3794–3807 (2019). https://doi.org/10.1109/TIP.2019.2902112
15. Russakovsky, O., et al.: ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. (IJCV) **115**(3), 211–252 (2015). https://doi.org/10.1007/s11263-015-0816-y
16. Tang, Z., et al.: PAMTRI: pose-aware multi-task learning for vehicle re-identification using highly randomized synthetic data. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 211–220 (2019). https://doi.org/10.1109/ICCV.2019.00030
17. Teng, S., Zhang, S., Huang, Q., Sebe, N.: Multi-view spatial attention embedding for vehicle re-identification. IEEE Trans. Circuits Syst. Video Technol. **31**(2), 816–827 (2021). https://doi.org/10.1109/TCSVT.2020.2980283
18. Wang, Z., et al.: Orientation invariant feature embedding and spatial temporal regularization for vehicle re-identification. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 379–387 (2017). https://doi.org/10.1109/ICCV.2017.49

19. Zhang, X., Zhang, R., Cao, J., Gong, D., You, M., Shen, C.: Part-guided attention learning for vehicle re-identification. ArXiv abs/1909.06023 (2019)
20. Zhao, J., Zhao, Y., Li, J., Yan, K., Tian, Y.: Heterogeneous relational complement for vehicle re-identification. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 205–214 (2021). https://doi.org/10.1109/ICCV48922.2021.00027
21. Zhouy, Y., Shao, L.: Viewpoint-aware attentive multi-view inference for vehicle re-identification. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6489–6498 (2018). https://doi.org/10.1109/CVPR.2018.00679
22. Zhu, J., Zeng, H., Huang, J., Liao, S., Lei, Z., Cai, C., Zheng, L.: Vehicle re-identification using quadruple directional deep learning features. IEEE Trans. Intell. Transp. Syst. **21**(1), 410–420 (2020). https://doi.org/10.1109/TITS.2019.2901312
23. Zhu, W., Hu, R., Wang, Z., Li, D., Gao, X.: Tell the truth from the front: anti-disguise vehicle re-identification. In: 2020 IEEE International Conference on Multimedia and Expo (ICME), pp. 1–6 (2020). https://doi.org/10.1109/ICME46284.2020.9102939
24. Zhu, X., Luo, Z., Fu, P., Ji, X.: VOC-ReLD: vehicle re-identification based on vehicle-orientation-camera. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 2566–2573 (2020). https://doi.org/10.1109/CVPRW50498.2020.00309
25. Zhuge, C., Peng, Y., Li, Y., Ai, J., Chen, J.: Attribute-guided feature extraction and augmentation robust learning for vehicle re-identification. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 2632–2637 (2020). https://doi.org/10.1109/CVPRW50498.2020.00317