Occlusion Based Discriminative Feature Mining for Vehicle Re-identification

Xianmin Lin[1], Shengwang Peng[1], Zhiqi Ma[1], Xiaoyi Zhou[2],
and Aihua Zheng[1](✉)

[1] Anhui Provincial Key Laboratory of Multimodal Cognitive Computation,
School of Computer Science and Technology, Anhui University, Hefei, China
ahzheng214@foxmail.com

[2] School of Computer Science and Cyberspace Security, Hainan University,
Haikou, China

Abstract. Existing methods of vehicle re-identification (ReID) focus on training robust models on the fixed data while ignore the diversity in the training data, which limits generalization ability of the models. In this paper, it proposes an occlusion based discriminative feature mining (ODFM) method for vehicle re-identification, which increases the diversity of the training set by synthesizing occlusion samples, to simulate the occlusion problem in the real scene. To better train the ReID model on the data with large occlusions, an attention mechanism was introduced in the mainstream network to learn the discriminative features for vehicle images. Experimental results on two public ReID datasets, VeRi-776 and VehicleID verify the effectiveness of the proposed method comparing to the state-of-the-art methods.

Keywords: Vehicle re-identification · Occlusion · Attention

1 Introduction

With the hot research of person re-identification, vehicle re-identification has attracted more and more attention from researchers. Vehicle re-identification (ReID) is the task of retrieving particular vehicles across different cameras. It is very important task for many other fields such as intelligent video surveillance and smart transportation and social security. Despite of the great progress in recent years, it still faces challenges such as occlusion, light, and similar appearance of different vehicles.

Existing methods of vehicle ReID mainly reply on deep learning technology to learn robust features. They roughly fall into two categories, either constructing complex network models to extract discriminative features, or proposing various loss functions for metric learning. However, due to the relatively clean background with rare occlusion, the models are generally easy to fit on the training data while hard during testing especially with background clutters. Eg., the loss function easily converges during training with however unsatisfactory performance during testing.

On the one hand, due to the view differences across the non-overlapping cameras, one can consider that the vehicle occludes itself. For instance, as shown in Fig. 1(a), the front view vehicle images will occlude the rear parts. On the other hand, the

P. Qin et al. (Eds.): ICPCSEE 2020, CCIS 1258, pp. 246–257, 2020.
https://doi.org/10.1007/978-981-15-7984-4_19

background occlusion is ubiquitous, such as the trees, other vehicles or pedestrians, as shown in Fig. 1(b).

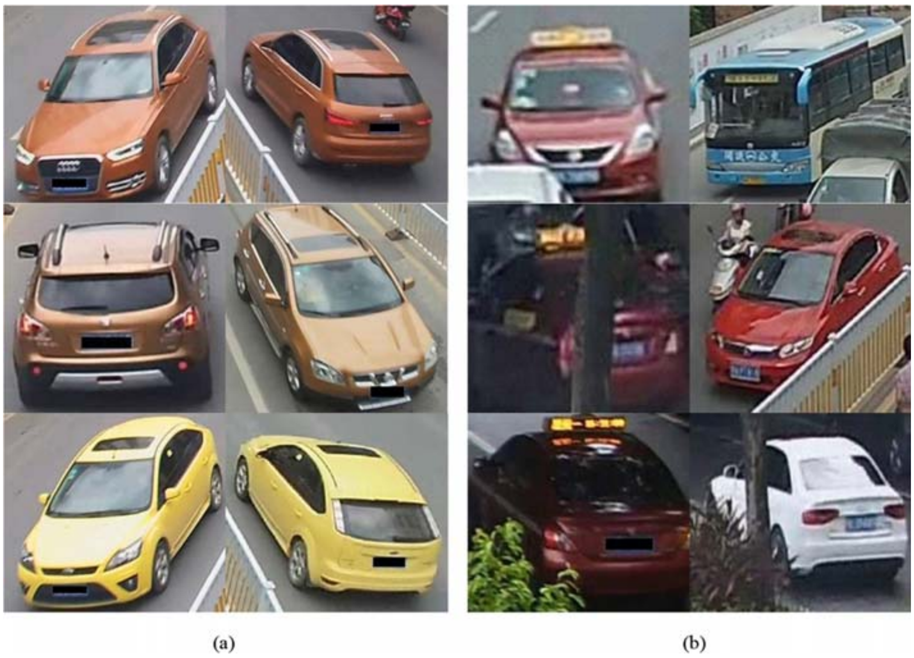

Fig. 1. Occluded samples from VeRi-776. (a) Image pairs of vehicles with different viewpoints. (b) Some samples occluded by vehicles, trees, pedestrians and other objects.

Unfortunately, existing methods lack of consideration of the occlusion issue during vehicle ReID. Furthermore, training set in the existing datasets are generally with rare occlusion which results in the limited generalization ability of the deep networks while handling the complex scenarios in the real scenes.

Related work [6] evidences that increasing the diversity of the training set can effectively improve the generalization ability of the model. Therefore, we propose to increase the diversity of the training by synthesizing occlusion samples and propose an attention mechanism to mine the discriminative feature for vehicle ReID.

After obtaining the more diverse training data with synthesized occlusion samples, the key issue is how to learn a robust discriminative feature representation while relieve the influence of the occlusions. It is well known that attention plays a very important role in the human visual system [2, 5], which does not process the entire scene information when facing a scene. It instead selectively pay more attention on the prominent parts to better capture the useful information [7]. Inspired by this, we propose to utilize an attention model based feature mining framework for vehicle ReID in this paper. Specifically, we embed spatial attention module and channel attention module into ResNet-50 network and train the ReID model.

As summary, we propose a novel occlusion based discriminative feature mining method for vehicle ReID in this paper. The main contributions of this paper can be summarized as:

- We introduce a data augmentation scheme via synthesizing the occlusion samples to the training data, which can significantly increase the generality of the ReID model especially in the challenging scenarios with self or environmental occlusions.
- We utilize the spatial and channel attention module to emphasize the discriminative region for vehicle ReID by exploring the relationship between both the channel and spatial level on the features.
- Comprehensive experiments on the benchmark datasets evidence the promising performance of the proposed method comparing with the state-of-the-art methods on vehicle ReID.

2 Related Work

We briefly introduce the related vehicle ReID methods in two folds: i) the appearance based vehicle ReID, ii) the attributes and temporal information enhanced vehicle ReID.

2.1 Appearance Based Vehicle ReID

As the main information in computer vision, most of existing works rely on the appearance information for vehicle ReID. Zapletal et al. [18] and Sochor et al. [14] propose to using 3D-boxes to align different vehicle surfaces and use three visible surfaces feature extraction. Zhang et al. [20] design an improved triplet-wise training by classification-oriented loss for vehicle ReID. Li et al. [8] integrate the identification, attribute recognition, verification and triplet tasks into a unified CNN framework. Liu et al. [3] propose a coarse-to-fine ranking method for vehicle ReID, consisting of a vehicle model classification loss, a coarse-grained ranking loss, a fine-grained ranking loss and a pairwise loss. Zhouy et al. [22] propose a Viewpoint-aware Attentive Multi-view Inference (VAMI) model, by first exploiting conditional generative network to generate vehicle images in different views to solve the multi-view problem. Peng et al. [12] propose a cross-camera adaptation framework (CCA), which smooths the bias by exploiting the common space between cameras for all samples. He et al. [4] propose a two-module framework that combines appearance and corresponding license plate features for vehicle Re-ID. In the appearance module, they designed a Two-Branch Network to extract comprehensive global features.

2.2 Attribute and Spatio-Temporal Information Enhanced Vehicle ReID

To overcome the appearance limitation across the cameras, auxiliary information, including attributes and spatio-temporal information have been integrated in vehicle ReID task. Liu et al. [11] design a progressive searching scheme which employed the appearance attributes of the vehicle for coarse filtering. Li et al. [8] introduced the attribute recognition into the vehicle ReID framework together with the verification loss and triplet loss into a unified vehicle ReID framework. Shen et al. [13] combine the visual spatio-temporal path information for regularization. Wang et al. [16] introduce the spatial-temporal regularization into the proposed orientation invariant feature embedding module to boost the vehicle ReID. Liu et al. [10] propose a deep region-aware model (RAM) to extract features from a series of local regions and encourage the deep model to learn discriminative features in both global and local levels. Zhong et al. [21] propose to predict the spatio-temporal motion of the vehicle via the Gaussian distribution probability model, followed by the driving direction estimation via CNN embedded pose classifier, to boost the ReID task. Wang et al. [15] propose a novel Attribute-Guided Network (AGNet) to learn global representation with the abundant attribute features in an end-to-end manner.

3 Methods

This paper mainly propose an occlusion based discriminative feature mining (ODFM) method to consider the occlusion issue in vehicle ReID. The pipeline of the proposed method is shown in Fig. 2, which falls into three phases: First, we train the ReID model until convergence. Second, we find the discriminative regions according to some strategy (which we will describe in Sect. 3.2) to synthesize the occluded samples. Finally, we retrain the ReID model on both original and synthesized occlusion samples (selected according to a certain strategy introduced below) until convergence. It is worth noting that the labels of the synthesized occluded samples are consistent with their original ones.

3.1 Re-identification Network

During the training phase, each vehicle is treated as a class like the common re-identification model. The probability of the vehicle image belonging to each class can be obtained after passing through the FC layer as the classifier. During the test, we remove the classifier layer and utilize the remaining layers for feature extraction.

Given a training set T containing N images of M vehicles (i.e.: M classes), each training image is recorded as $(I_i, m_i), i \in 1, 2, \ldots, N$, where m_i represents the label of the image I_i. The ReID network can be regarded as a function that maps the image I_i to the classification score vector $S_i = g(I_i)$, which is further normalized into a probability distribution by a softmax function (y_{ij} is the predicted value):

$$p(y_{ij}|I_i) = \frac{\exp(s_{ij})}{\sum_{m=1}^{M} \exp(s_{im})}, j = 1, 2..M \tag{1}$$

We use the cross-entropy loss function to train the model:

$$l_i(\theta) = -\log\left(p(y_{im_i}|I_i)\right) \tag{2}$$

where θ represents the parameters of the network. Therefore, the loss function on the entire training set T can be calculated as:

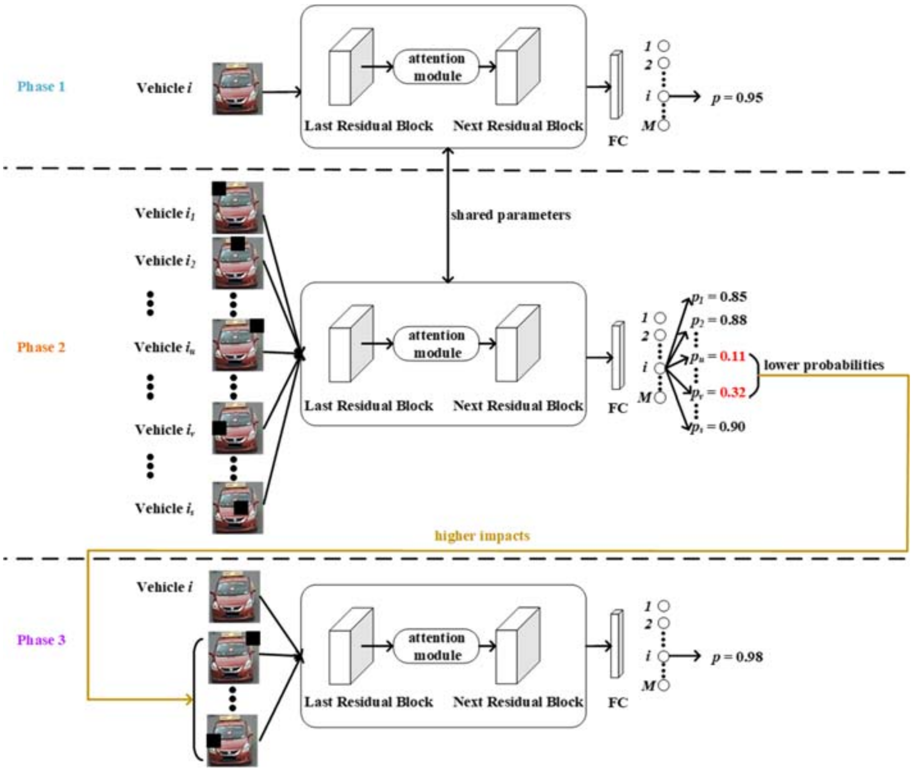

Fig. 2. The pipeline of the proposed occlusion based discriminative feature mining (ODFM) method. Phase 1 trains a ReID model to convergence with the original data. Phase 2 synthesizes the candidate occluded samples and simultaneously evaluates the impact values of each candidate via the pretrained ReID network. Phase 3 inputs both the original data and the selected occlusion samples with high impact values to retrain the ReID model. Note that, we use the attention mechanism all through the procedure.

$$l(\theta) = \frac{1}{N}\sum\nolimits_{i=1}^{N} l_i(\theta) \tag{3}$$

Minimizing the loss function is equivalent to maximizing the posterior probability of the true value. We use stochastic gradient descent and mini-batch samples to optimize the loss function.

3.2 Occlusion Sample Synthesizing Strategy

To increase the diversity of the training set with more severe scenarios, we propose to synthesize the occlusion samples for data augmentation. Specifically, we use a sliding occlusion mask to mine the discriminative regions. For an image of size $H * W$ and an occlusion mask of size $d * d$, we move from left to right and top to bottom with steps

S_w and S_h, respectively. Therefore, a total of $S = \left(\left\lfloor \frac{W-d}{S_w} \right\rfloor + 1\right) * \left(\left\lfloor \frac{H-d}{S_h} + 1 \right\rfloor\right)$ new samples are synthesized to the candidate sample pool.

Give a training image and its predicted true value probability p, the impact value of the l-th candidate can be defined as:

$$\widetilde{p_l} = \begin{cases} p - p_l, & p > p_l \\ 0, & p \leq p_l \end{cases} \tag{4}$$

where $p_l, l(l \in 1, 2, \ldots, S)$ indicates the probability value of the l-th candidate. To fairly consider the impact caused by different occlusion positions, we normalize the impact values of the S candidates of each training image into a distribution:

$$\overline{p_l} = \frac{\widetilde{p_l}}{\sum_{j=1}^{S} \widetilde{p_j}}, l = 1, 2, \ldots, S \tag{5}$$

according to which we sample the occluded images, which tend to have higher impact to the ReID task.

3.3 Spatial and Channel Attention Mechanism

To better explore the most discriminative regions in the vehicle images, we propose to introduce the channel and spatial attention mechanism. Given any feature map $X \in R^{C*H*W}$, one-dimensional channel attention feature $M_c \in R^{c*1*1}$ and two-dimensional spatial attention feature $M_s \in R^{1*H*W}$ can be obtained respectively through the channel attention module and the spatial attention module. The entire attention process is as follows:

$$X_1 = M_c \odot X, X_2 = M_s \odot X_1 \tag{6}$$

where $\odot$ represents the element-wise production. During the multiplication process, the attention values are sequentially propagated: the channel attention values are propagated along the spatial dimension, and vice versa. X_2 is the feature produced by the attention module. We shall elaborate each attention modules in the following.

Channel Attention Module. We use feature-to-channel relationship to produce channel attention maps. As mentioned in [19], each channel of a feature map can be considered as a feature detector, and channel attention will focus on "what information" of a given image is meaningful. In order to effectively obtain the channel attention map, we compress the input 3D features along the spatial dimension.

First, we use average pooling and maximum pooling to aggregate spatial information to respectively produce two different spatial context descriptors: F_{avg} and F_{max}, followed by the shared network to produce a channel attention map $M_c \in R^{C*1*1}$. The shared network consists of a multilayer perceptron (MLP) and a hidden layer. To reduce the parameters, we set the output size of the hidden layer to $R^{\frac{C}{r}*1*1}$, where r

represents the reduction ratio. Then the two features from the shared network are added together. In short, the channel attention feature is obtained as:

$$\begin{aligned} M_c =& sigmoid(MLP(AvgPool(X)) + MLP(MaxPool(X))) \\ =& sigmoid\left(W_1\left(W_0\left(F_{avg}\right)\right) + W_1(W_0(F_{max}))\right) \end{aligned} \tag{7}$$

where $sigmoid()$ is the activation function, $W_0 \in R^{\frac{C}{r}*C}$, $W_1 \in R^{C*\frac{C}{r}}$ are the shared weights of the multilayer perceptron. A ReLU activation function is added after W_0 (Fig. 3).

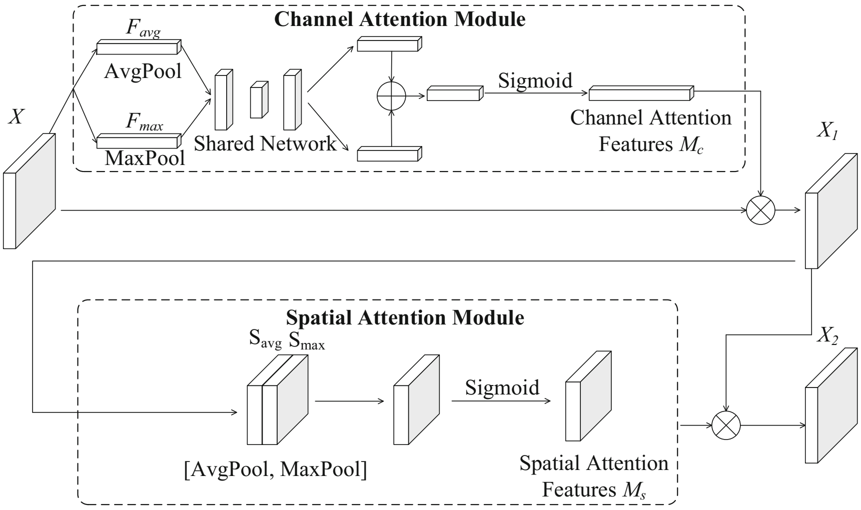

Fig. 3. Attention module diagram. The upper and lower parts demonstrate the channel and spatial attention modules respectively.

Spatial Attention Module. We use the spatial relationship of features to produce spatial attention maps. Unlike channel attention, spatial attention focuses meaningful regions. Note that applying the pooling operation along the channel dimension can effectively highlight the informative regions [17]. Herein, to obtain the spatial attention map, we first apply the average pooling and maximum pooling operations along the channel dimension, to aggregate the channel information of the input features as $S_{avg} \in R^{1*H*W}$ and $S_{max} \in R^{1*H*W}$. Then we concatenate them to synthesize an effective spatial feature descriptor. The concatenated these features are fed into a standard convolutional layer to obtain a spatial attention map $M_s \in R^{H*W}$. The process of obtaining the spatial attention map can be summarized by the following formula:

$$\begin{aligned} M_s =& sigmoid\left(f^{7*7}([AvgPool(X); MaxPool(X)])\right) \\ =& sigmoid\left(f^{7*7}\left(\left[S_{avg}; S_{max}\right]\right)\right) \end{aligned} \tag{8}$$

where f^{7*7} represents a convolution operation with a kernel size of 7 * 7.

3.4 Implement Details

We use ResNet-50 as our backbone network and employ Euclidean distance to measure similarity during the testing phase. We employ Gaussian weights with the bias as zero to initialize the last layer of ResNet-50, while preserving ImageNet pre-trained weights in all other layers. We use Stochastic Gradient Descent (SGD) to optimize the model with momentum as 0.9 and weight decay as 5e-4. We set the learning rate of the last layer to 0.02, while 0.01 for the other layers, followed by multiplication by 0.1 for every 25 iterations. During the training, we introduce a dropout layer before the classifier to regularize the network. We perform our experiment on Pytorch under Ubuntu16.04 system with a single TITAN xp. The input image is randomly flipped with a probability of 0.5. The batch size is set to 32.

4 Experiments

We evaluate our ODFM on two benchmark vehicle ReID datasets VeRi-776 [11] and VehicleID [9] comparing with the state-of-the-art methods. We use the commonly used metrics, mAP and rank-*n* to evaluate our experimental results. mAP indicates the mean average precision, while rank-*n* indicates the right hits in the first *n* rankings to the query image.

4.1 Experimental Results and Analysis

Results on VeRi-776. Table 1 reports the comparison results with other methods on the VeRi-776 dataset. It can be observed that our method significantly beats the prevalent methods especially on mAP, by 5.83% improvement than the second best method VFL [1]. It is worth noting that our method hasn't used any auxiliary information, like the license plate or spatio-temporal information in Siamese + Path-LSTM [13] and FACT + Plate-SNN + STR [11]. However, it still surpasses them on all metrics, which verifies the competitive performance of the proposed method.

Figure 4 demonstrates five examples of matching results on VeRi-776 dataset [11]. Our model generally achieves satisfactory results. Although some query vehicles are severely occluded by the environment, such as query (a) and (e), our method can still hit most of the right matchings. Furthermore, as shown in the results of query (b) to (d) in Fig. 4, it can correctly hit the gallery vehicles with occlusions or viewpoint changes even the query images are without occlusion or in totally different viewpoints.

Results on VehicleID. Table 2 demonstrates the comparison results with prevalent methods on VehicleID dataset. Generally speaking, our ODFM outperforms the state-of-the-art methods in a large margin. Our method consistently beats the prevalent methods on all the three subsets with promising performance, which yields to a new state-of-the-art for vehicle ReID.

Figure 5 shows several examples of the matching results on VehicleID dataset. Note that most of the vehicle images in VehicleID dataset is either front or back view with only one correct matching image in the gallery. From Fig. 5 we can see, our

Table 1. Comparison results on VeRi-776 [11]. The best three results are highlighted in red, blue and green respectively.

Method	mAP	rank1	rank5	reference
LOMO	9.64	25.33	46.48	CVPR2015
BOW-CN	12.20	33.91	53.69	ICCV2015
GoogLeNet	17.89	52.32	72.17	CVPR2015
FACT	18.49	50.95	73.48	ICME2016
FACT+Plate-SNN+STR	27.70	61.44	78.78	ECCV2016
Siamese-Visual	29.48	41.12	60.31	ICCV2017
Siamese+Path-LSTM	58.27	83.49	90.04	ICCV2017
NuFACT	48.47	76.76	91.42	TMM2018
VAMI	50.13	77.03	90.82	ICPR2018
VRSDNet	53.45	83.49	92.55	CVPR2018
AAVER	58.52	88.68	94.10	ICCV2019
VFL	59.18	88.08	94.63	ICIP2019
ODFM	65.01	91.00	97.85	**Ours**

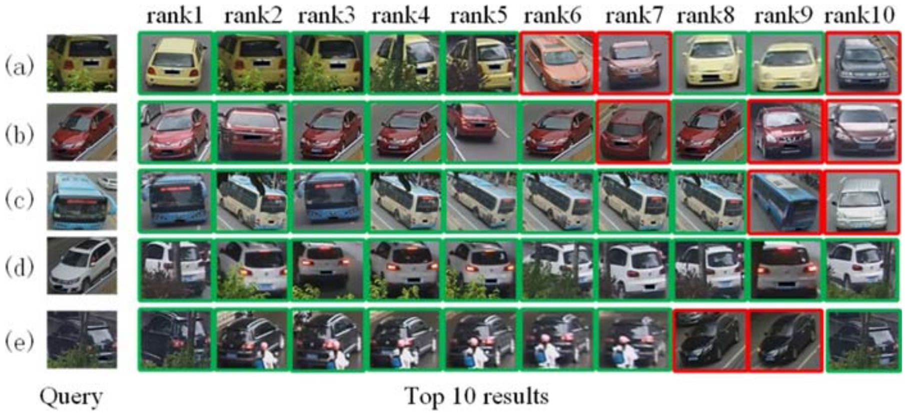

Fig. 4. Examples of matching result on VeRi-776 dataset [11]. The first column represents the query image, followed by the top 10 ranking results highlighted in green bounding boxes (indicating the right hits) and red bounding boxes (indicating the wrong hits), respectively. (Color figure online)

method can hit the unique right matching with different viewpoint (as shown in query (c)) and occlusion (as shown in query (d)). Meanwhile, it can handle the scenarios with severe illumination or scale changes across the cameras, such as query (a), (b) and (e).

4.2 Ablation Study

To verify the contribution of our method, we further evaluate the two key components in our model: occluded sample synthesizing strategy and attention module on VeRi-776 and VehicleID dataset on 800 test size. As shown in Table 3, both occluded sample synthesizing strategy and the spatial channel attention module play important roles in the ReID model, which verifies the contribution of the two components.

Table 2. Comparisons with state-of-the-art methods on VehicleID [9] (in %). The best three results are highlighted in red, blue and green respectively.

Methods	800		1600		2400		reference
	mAP	rank1	mAP	rank1	mAP	rank1	
LOMO	-	19.76	-	18.85	-	15.32	CVPR2015
BOW-CN	-	13.14	-	12.94	-	10.20	ICCV2015
GoogLeNet	46.20	47.88	44.00	43.40	38.10	38.27	CVPR2015
FACT	-	49.53	-	44.59	-	39.92	ICME2016
NuFACT	-	48.90	-	43.64	-	38.63	TMM2018
VAMI	-	63.12	-	52.87	-	47.34	CVPR2018
VRSDNet	63.52	56.98	57.07	50.57	49.68	42.92	ICPR2018
AAVER	-	72.47	-	66.85	-	63.54	ICCV2019
VFL	-	73.37	-	69.52	-	67.41	ICIP2019
ODFM	83.16	76.57	77.87	71.55	74.75	68.23	**Ours**

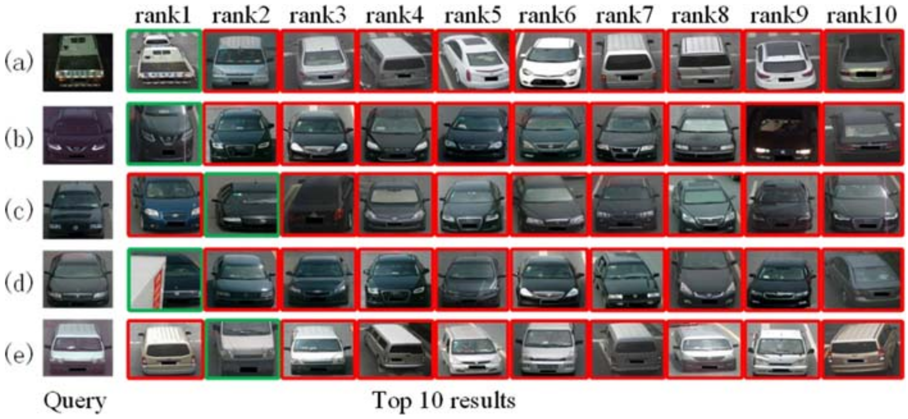

Fig. 5. Examples of matching result on VehicleID [9]. The first column represents the query image, followed by the top 10 ranking results highlighted in green bounding boxes (indicating the right hits) and red bounding boxes (indicating the wrong hits), respectively. (Color figure online)

Table 3. Ablation study on the occluded sample synthesizing strategy and the spatial channel attention module.

Methods	VeRi-776		VehicleID (800)	
	mAP	rank1	mAP	rank1
baseline	60.55	88.20	80.71	74.20
+attention	62.03	88.92	81.01	74.69
+sample synthesizing	64.62	90.41	82.29	76.29
+sample synthesizing + attention	65.01	91.00	83.16	76.57

5 Conclusion

In this paper, we propose an occlusion based discriminant feature mining (ODFM) method for vehicle re-identification. It first simulated the occlusion problem in real scenes by synthesizing occlusion samples, thereby increasing the diversity of the training set. In addition, the attention model was utilized in the mainstream network to learn the discriminative features of vehicle images. The experimental results on two public datasets VeRi-776 and VehicleID verify the superiority of the proposed method, which yields a new state-of-the-art for vehicle ReID.

Acknowledgement. This research is supported in part by the National Natural Science Foundation of China (Nos. 61976002), Hainan Provincial Natural Science Foundation (Grant No. 117063), the Natural Science Foundation of Anhui Higher Education Institutions of China (KJ2019A0033), and the National Laboratory of Pattern Recognition (NLPR) (201900046).

References

1. Alfasly, S.A.S., et al.: Variational representation learning for vehicle re-identification, pp. 3118–3122 (2019)
2. Corbetta, M., Shulman, G.L.: Control of goal-directed and stimulus-driven attention in the brain. Nat. Rev. Neurosci. **3**(3), 201–215 (2002)
3. Guo, H., et al.: Learning coarse-to-fine structured feature embedding for vehicle re-identification, pp. 6853–6860 (2018)
4. He, Y., Dong, C., Wei, Y.: Combination of appearance and license plate features for vehicle re-identification, pp. 3108–3112 (2019)
5. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Trans. Pattern Anal. Mach. Intell. **20**(11), 1254–1259 (1998)
6. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks, pp. 1097–1105 (2012)
7. Larochelle, H., Hinton, G.E.: Learning to combine foveal glimpses with a third-order Boltzmann machine, pp. 1243–1251 (2010)
8. Li, Y., et al.: Deep joint discriminative learning for vehicle re-identification and retrieval, pp. 395–399 (2017)
9. Liu, H., et al.: Deep relative distance learning: tell the difference between similar vehicles, pp. 2167–2175 (2016)

10. Liu, X., et al.: RAM: a region-aware deep model for vehicle re-identification, pp. 1–6 (2018)
11. Liu, X., et al.: A deep learning-based approach to progressive vehicle re-identification for urban surveillance. In: European Conference on Computer Vision (2016)
12. Peng, J., et al.: Eliminating cross-camera bias for vehicle re-identification. In: arXiv: Computer Vision and Pattern Recognition (2019)
13. Shen, Y., et al.: Learning deep neural networks for vehicle re-ID with visual-spatio-temporal path proposals, pp. 1918–1927 (2017)
14. Sochor, J., Herout, A., Havel, J.: BoxCars: 3D boxes as CNN input for improved fine-grained vehicle recognition, pp. 3006–3015 (2016)
15. Wang, H., et al.: Attribute-guided feature learning network for vehicle re-identification. In: arXiv: Computer Vision and Pattern Recognition (2020)
16. 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) 2017
17. Zagoruyko, S., Komodakis, N.: Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer (2017)
18. Zapletal, D., Herout, A.: Vehicle re-identification for automatic video traffic surveillance, pp. 1568–1574 (2016)
19. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks, pp. 818–833 (2014)
20. Zhang, Y., Liu, D., Zha, Z.: Improving triplet-wise training of convolutional neural network for vehicle re-identification, pp. 1386–1391 (2017)
21. Zhong, X., et al.: Poses guide spatiotemporal model for vehicle re-identification, pp. 426–439 (2019)
22. Zhouy, Y., Shao, L.: Viewpoint-aware attentive multi-view inference for vehicle re-identification, pp. 6489–6498 (2018)