SCAN: Spatial and Channel Attention Network for Vehicle Re-Identification

Shangzhi Teng[1(✉)], Xiaobin Liu[2], Shiliang Zhang[2], and Qingming Huang[1]

[1] University of Chinese Academy of Sciences, Beijing, China
shangzhi.teng@vipl.ict.ac.cn, qmhuang@ucas.ac.cn
[2] Peking University, Beijing, China
{xbliu.vmc,slzhang.jdl}@pku.edu.cn

Abstract. Most existing methods on vehicle Re-Identification (ReID) extract global features on vehicles. However, as some vehicles have the same model and color, it is hard to distinguish them only depend on global appearance. Compared with global appearance, some local regions could be more discriminative. Moreover, it is not reasonable to use feature maps with equal channels weights for methods based on Deep Convolutional Neural Network (DCNN), as different channels have different discrimination ability. To automatically discover discriminative regions on vehicles and discriminative channels in networks, we propose a Spatial and Channel Attention Network (SCAN) based on DCNN. Specifically, the attention model contains two branches, *i.e.*, spatial attention branch and channel attention branch, which are embedded after convolutional layers to refine the feature maps. Spatial and channel attention branches adjust the weights of outputs in different positions and different channels to highlight the outputs in discriminative regions and channels, respectively. Then feature maps are refined by our attention model and more discriminative features can be extracted automatically. We jointly train the attention branches and convolutional layers by triplet loss and cross-entropy loss. We evaluate our methods on two large-scale vehicle ReID datasets, *i.e.*, *VehicleID* and *VeRi-776*. Extensive evaluations on two datasets show that our methods achieve promising results and outperform the state-of-the-art approaches on *VeRi-776*.

Keywords: Vehicle Re-Identification
Deep Convolutional Neural Network · Attention

1 Introduction

Vehicle Re-Identification (ReID) aims to match a query vehicle image against a large-scale vehicle image gallery set [3,11–13,27]. The ability to quickly find and track suspect vehicles makes vehicle ReID important for traffic surveillance and applications on smart city. Vehicle ReID is related with several extensively studied tasks on vehicle identification, such as vehicle attribute prediction [29] and fine-grained vehicle classification [8,11,29]. Different from these tasks that

© Springer Nature Switzerland AG 2018
R. Hong et al. (Eds.): PCM 2018, LNCS 11166, pp. 350–361, 2018.
https://doi.org/10.1007/978-3-030-00764-5_32

Fig. 1. Illustration of challenging issues on vehicle ReID. The first row shows four images that capturing the same vehicle. The second row shows different vehicles with the same model and color.

Fig. 2. Examples of attention regions on different vehicles. The first row shows input vehicle images. The second row shows salient image regions which are learned by our attention model.

mainly focus on identifying the fine-grained categories of vehicles, vehicle ReID focuses on the instance-level identification. As different instances of the same maker and model may be similar with each other, vehicle ReID is more challenging and far from being solved.

Vehicle ReID task requires highly discriminative features to precisely identify different vehicles. However, in surveillance scenario, the quality of vehicle images could be easily affected by many factors, such as illumination, view point, and occlusion. This makes hand-crafted features unstable and prevents them from working. Recently, Deep Convolutional Neural Networks (DCNNs) have made breakthrough in many tasks including person ReID [9,10,19,21,24–26] and fine-grained categorization [6,22,30]. Existing works have designed many DCNN based model for vehicle feature learning and deep model have dominated the methods of vehicle ReID. More details of related works will be summarized in Sect. 2.1.

Although previous works have achieved significant success, there still remain several open issues that make vehicle ReID a challenging task. Firstly, different views of the same vehicle may capture little common region, as shown in the first row of Fig. 1. This results in large intra-class distance and false negative samples in ReID. Secondly, lots of vehicles with the same model (maker, product year and type) and color are quite similar. For example, Fig. 1 shows some different vehicles with similar appearance in the second row. It can be observed that only from some small regions, *e.g.*, the annual inspection marks on the front window, can we tell the difference. Lastly, the misalignment problem is serious in vehicle ReID. Vehicle image is difficult to align each part with a fixed order. As most existing DCNN based approaches [3,11,13] extract features from the whole vehicle image, they fail to conquer aforementioned issues very well.

Inspired to conquer these issues, we propose a Spatial and Channel Attention Network (SCAN) based on DCNN. SCAN contains two branches, *i.e.*, spatial attention branch and channel attention branch, to explore discriminative regions on vehicles and discriminative channels in networks, respectively. SCAN

produces saliency weight maps to highlight discriminative areas and channels. Some examples of generated spatial saliency maps are shown in Fig. 2. Compared with forcing the model to extract regional features from rigid local regions and some certain channels, SCAN uses a automatical soft attention strategy and thus can adaptively explores discriminative regions and channels for different input vehicles. The convolutional layers and attention branches are jointly trained in an end-to-end manner to simultaneously learn feature maps and spatial-channel attention, respectively.

Our methods are evaluated on two large-scale vehicle ReID datasets. Experimental results show that our methods achieve promising performance compared with recent works. The contributions of our method is two-fold. (1) A deep learning attention network is proposed for vehicle ReID. The attention module is designed to refine the feature maps in CNN. This reinforces the useful details and weakens the useless information. Thus, our model can make the feature representation more discriminating. As far as we know, this is the first attempt of learning attention network for solving the vehicle ReID problem without any extra annotation. (2) We test the effectiveness of our attention network on two vehicle ReID datasets. Experiments show that our attention model outperforms the state-of-the-art methods on *VeRi*.

2 Related Work

2.1 Vehicle ReID

With the development of smart city and public security, vehicle ReID has gained more and more attentions. Liu *et al.* [13,14] propose a vehicle ReID dataset *VeRi-776* which contains over 50,000 images of 776 vehicles captured by 20 cameras covering an elliptical area. Number plate, vehicle appearances and spatio-temporal relation are separately used in [14] to learn the similarity scores between pairs of images. Shen *et al.* [18] also use spatio-temporal information for improving the ReID results. A chain MRF model is used to generate a visual-spatio-temporal path which gives a similarity score by LSTM. Wang *et al.* [23] pre-train a region proposal module in order to produce the response maps of 20 vehicle key points. They then extract regional features based on key points prediction. And the spatial-temporal constraints is also adopted to refine the retrieval results. This work needs extra notation of key points on a large-scale dataset to pre-train the model, and the model is complex. Moreover, in most cases we do not have spatial and temporal information. So the above methods may not work out. Liu *et al.* [11] propose a Coupled Clusters Loss (CCL) which modifies traditional triplet loss. And a vehicle ReID dataset is proposed. Yan *et al.* [28] use multi-grain relations to improve vehicle search performance. These two methods only extract features from global appearance, resulting in features lack discrimination power.

2.2 Fine-Grained Vehicle Classification

Fine-grained classification is relevant to ReID task. They all focus on learning discriminative feature representations. Wang *et al.* [22] propose a patch-based framework. It acquires triplets of patches with geometric constraints and automatically mines discriminative geometrically-constrained triplets for classification. However, this method is not an end-to-end framework and it can not handle large view changes. Huang *et al.* [6] propose another part based models which pre-trains a region proposal network to capture the discriminative object regions. Different from these part-based methods, Zhang *et al.* [31] and Qian *et al.* [17] use distance metric learning to reduce the intra-class distance and increase inter-class distance. Metric learning is also widely used in ReID task. Yang *et al.* [29] propose a fine-grained vehicle model classification dataset (CompCars) which is the largest vehicle model dataset.

2.3 Attention Modelling

Recently, deep attention learning methods have been proposed in many tasks to handle the matching misalignment challenge. Many tasks have demonstrate that the attention model is valid. Such as semantic segmentation [2], visual question answering [15], tracking [33] and person ReID [9,10,19]. Most of these attention models are designed to select several parts from the whole image. Local and global characteristics are concatenated together to get a more comprehensive representation. In our work, SCAN produces saliency weight maps to explore discriminative regions on vehicles and discriminative channels in networks.

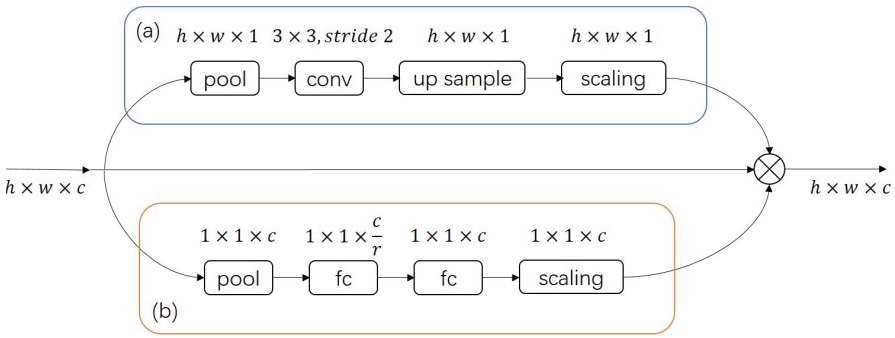

Fig. 3. The structure of SCAN consists of (a) spatial attention branch and (b) channel attention branch. The ReLU and Reshape operation are not shown for brevity.

3 Spatial and Channel Attention Network

This section presents the end-to-end trainable framework of Spatial and Channel Attention Network (SCAN). SCAN consists of two branches: one for spatial attention learning and one for channel attention learning, respectively. The

structure of proposed attention module is shown in Fig. 3. These two attention branches learn weights separately and then fuse weights with the original feature maps. The SCAN module can be integrated into any modern DCNN architectures.

3.1 Spatial Attention Branch

We employ a Spatial Attention Branch (SAB) to automatically discover discriminative regions on vehicles. The structure of SAB is illustrated in Fig. 3(a). The input of SAB is a set of feature maps denoted as $f \in R^{h \times w \times c}$, where h,w, and c denote the size of height, width and channels of feature maps, respectively. SAB first uses a channel-wise global average pooling layer without involving more parameter. The channel-wise global average pooling at the spatial location (i, j) is defined as follows:

$$s_{i,j} = \frac{1}{c} \sum_{k=1}^{c} f_{i,j,k}, \tag{1}$$

where $f_{i,j,k}$ is the value of feature maps at the location (i, j) of k^{th} channel. SAB then uses a convolutional layer of 3×3 filter with stride 2, and an up sampling layer is then added after the convolutional layer. A convolutional layer of 1×1 filter is added to automatically learn an adaptive attention scale. The ReLU function is applied as active function to each convolutional layer. A deconvolutional layer is finally used to generated spatial attention feature maps. We use sigmoid function to normalise the value of each output of the generated attention feature maps into the range between 0.5 and 1.

Instead of learning a rigid spatial decomposition of input images as in [4], SAB automatically identifies salient regions in each vehicle image. The salient regions could be an irregular shape, which helps to find more useful information than regular shape area. Some salient regions on different vehicles generated SAB are shown in Fig. 2. It can be observed that some discriminative regions are highlighted by SAB, such as annual inspection marks and logos of makers. Thus, features extracted from this regions are potential to convey more details and have stronger discrimination ability compared with ones directly extracted from entire images.

3.2 Channel Attention Branch

The purpose of designing Channel Attention Branch (CAB) is to improve the discrimination power of the network by explicitly modelling the interdependencies between the channels of its convolutional features [5]. The structure of CAB is illustrated in Fig. 3(b). The input to CAB is the same as SAB. CAB first uses a global average pooling layer to integrate spatial information in each feature map. The global average pooling is defined as:

$$c_k = \frac{1}{h \times w} \sum_{i=1}^{h} \sum_{j=1}^{w} f_{i,j,k} \tag{2}$$

Then two fully connected layers with $\frac{c}{r}$ and c outputs are used. Parameter r is designed for reducing the model parameter number from c^2 to ($\frac{c^2}{r} + \frac{c^2}{r}$), e.g. only $\frac{c}{8}$ parameters are involved when r = 16. A 1×1 convolutional layer is used to scale the output, and a deconvolutional layer is finally used to produce the attention weight map of the same size as the input feature map.

3.3 Vehicle ReID by SCAN

We add proposed two attention branch, *i.e.*, SAB and CAB, after the conv5 layer in VGG_CNN_M_1024 and the conv5_3 layer in VGG16. After the SCAN we add a global average pooling layer, a 512-D fully connected layer and two loss layers as illustrated in Fig. 4. Given a trained SCAN model, we use the 512-D fully connected layer as the vehicle feature. For vehicle ReID , we calculate L_2 distance between query images and each gallery image using this 512-D deep feature. We then rank all gallery images in ascendant order by their L_2 distances to the probe image. Based on ranking results, we could find and track vehicles in surveillance video analysis.

4 Experiments

4.1 Datasets and Base Model

We use two existing vehicle datasets to validate SCAN. *VeRi-776* [12] is a benchmark dateset for vehicle ReID that is collected from real-world surveillance scenarios, with over 50,000 images of 776 vehicles in total. Each vehicle is captured by 2 to 18 cameras in an urban area of 1 km^2 during a 24-hour time period. *VehicleID* [1] is a surveillance dataset, which consists 26,267 vehicles and 22,1763 images in total. The numbers of identities and images for training and testing are listed in Table 1.

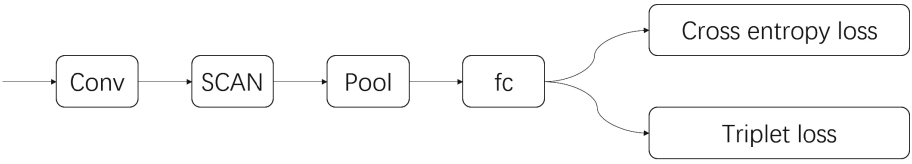

Fig. 4. We use cross entropy loss and triplet loss to train our spatial channel attention network.

Experiments are conducted based on the widely-used deep learning framework Caffe [7]. We use VGG_CNN_M_1024 [1] network and VGG16 [20] network as the basic network. Different from original VGG_CNN_M_1024 and VGG16 network, our baseline network replace all fully connected layers of original network with a 512-D layer. This change greatly reduces the number of parameters

in the network. For VGG_CNN_M_1024, our model size is 12 times smaller than the original one. For VGG16, our model size is 16 times smaller than the original one. Model size are illustrated in Table 4. And then we call our baseline network light_vgg_m and light_vgg16. The mean Average Precision (mAP), Top-1 accuracy and Top-10 accuracy are chosen as the evaluation metric.

Table 1. Statistics of the two datasets used in our experiment.

Dataset	Train ID/image	Probe ID/image	Gallery ID/image
VeRi-776	576/37778	200/1678	200/11579
VehicleID	13164/100182	2400/17638	2400/2400

The batch size used to train the network in our experiments is set to 64. The initial learning rate is set to 0.001. The learning rate decay factor is 0.8 for every 2,0000 iterations. The weight decay factor is set to 0.0004. The momentum is set to 0.9. The loss weight for cross entropy loss is set to 1 and the loss weight for triplet loss is also set to 1.0.

4.2 Results on VeRi-776

The experimental results on *VeRi* are summarised in Table 2. It can be observed that our proposed attention approach achieves the best performance on *VeRi-776* Dataset. light_vgg_m+SCAN denotes the light_vgg_m network with proposed SCAN. And light_vgg16+SCAN denotes the light_vgg16 network with proposed SCAN. Note that the Top-1 accuracy of our baseline model light_vgg_m is higher than most existing methods by large margins. But the mAP of our baseline model is lower than OIFE [23] and VAMI [32]. Significant performance gain can be observed from light_vgg_m+SCAN compared with light_vgg_m, which means our spatial-channel-attention network (SCAN) is effective. Our attention approach has a performance gain of 6% in Top-1 accuracy and 11% in mAP compared with our baseline light_vgg_m network. Compared with OIFE [23], which takes pairwise visual and spatio-temporal information into account, our approach light_vgg_m+SCAN has a performance gain of 15% in terms of Top-1 accuracy and 1% in terms of Top-5 accuracy. VAMI [32] also uses additional perspective information to train their network and the Top-1 accuracy is lower than our attention model. The same improvements can also be observed in performance by light_vgg16 and light_vgg16+SCAN. The performance of our model light_vgg_m+SCAN is a little higher than light_vgg16+SCAN. One possible reason is that the image resolution in *VeRi-776* dataset is small and the small size network can learn its parameters well with low resolution images.

Table 2. Experiment results of the proposed method and other compared methods on *VeRi-776* dataset.

VeRi-776	Top-1 (%)	Top-5 (%)	mAP (%)
KEPLER [16]	48.2	64.3	33.53
FACT [14]	50.95	73.48	18.49
FACT+Plate-SNN+STR [14]	61.44	78.78	27.77
OIFE [23]	68.3	89.7	51.42
VAMI [32]	77.03	90.82	50.13
light_vgg_m	76.02	86.05	38.94
light_vgg_m+SCAN	**82.24**	**90.76**	49.87
light_vgg16	76.82	86.71	39.91
light_vgg16+SCAN	79.92	88.32	50.15

4.3 Results on VehicleID

The experimental results on *VehicleID* are summarised in Table 3. On this dataset we also trained two baseline networks: light_vgg_m and light_vgg16. Because our baseline model has fewer parameters than normal VGG_M and VGG16 networks, the performance of our baseline on *VehicleID* dataset is not very good. However, after adding our attention module the performance has been significant improved. Our attention approach light_vgg_m+SCAN has a improvements gain of 11% in terms of Top-1 accuracy and 6% in terms of Top-5 accuracy compared with light_vgg_m network. light_vgg16+SCAN has a gain of 3% in terms of Top-1 accuracy and 5% in terms of Top-5 accuracy compared

Table 3. Experiment results of the proposed method and other compared methods on *VehicleID* dataset. light_vgg16+SCAN* indicates we use the SCAN feature vector and light_vgg16 feature vector together.

VehicleID	Top-1 (%)	Top-5 (%)
KEPLER [16]	45.4	68.9
VGG + Triplet Loss [11]	31.9	50.3
VGG + CCL [11]	32.9	53.3
Mixed Diff + CCL [11]	38.2	61.6
OIFE [23]	67.0	82.9
VAMI [32]	47.34	70.29
light_vgg_m	44.14	65.21
light_vgg_m+SCAN	55.73	71.73
light_vgg16	60.63	72.67
light_vgg16+SCAN	63.52	77.53
light_vgg16+SCAN*	**65.44**	**78.47**

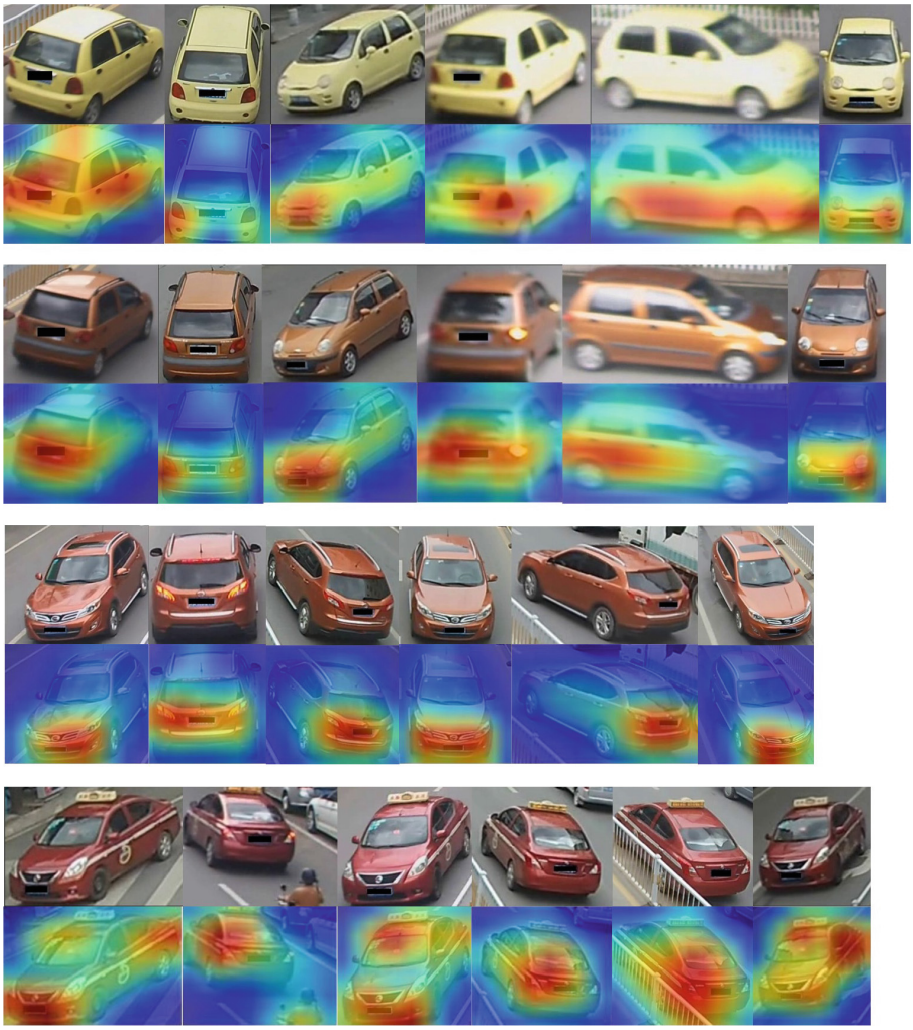

Fig. 5. Visualisation of our spatial attention in vehicle ReID. The odd rows show the original vehicle. The even rows show the attention weight maps learned from our spatial attention module. Best viewed in color.

with light_vgg16 network. If we concatenate features extracted by light_vgg16 and light_vgg16+SCAN, the performance may have further improved. The final Top-1 accuracy of light_vgg16+SCAN* is 65.44%, which is a little lower than OIFE [23]. However, [23] needs extra annotation information to pre-train a key point regressor network. And four vehicle datasets are used in [23]. We only use *VehicleID* without extra annotation to train our model. So our method has been proved to be effective.

Table 4. Comparisons of model size. NP denotes the number of parameters in each model.

Model	NP (million)
VGG_M_1024	86.2
VGG16	127.2
light_vgg_m	6.8
light_vgg16	7.9
SCAN	0.033

4.4 Visualisation of Spatial Attention and Model Size

We visualise our learned spatial attention of SCAN in Fig. 5. It can be observed that spatial attention branch locates some spatial regions of vehicles, which approximately corresponds to headlights, taillights, vehicle signs, and vehicle marks. This compellingly shows the effectiveness of our spatial attention learning. We compare model size of original models and our light models in Table 4. It is clear that our proposed light models use less parameters than original models, while achieve better performance.

5 Conclusions

In this work, we focus on the problem of vehicle re-identification, which aims at finding out the images belonging to exactly the same vehicle with the query image. To address this problem, we proposed an end-to-end trainable framework, namely Spatial Channel Attention Network (SCAN), for joint learning attention weights and feature representation. SCAN consists of two branches, *i.e.*, spatial attention branch and channel attention branch, to adjust the weight of outputs in different positions and channels. With our SCAN model we could explore discriminative regions and channels for powerful feature extraction. The proposed SCAN does not need bounding box or part annotations for training. We evaluated our proposed approach on two vehicle ReID datasets and a series of experiments show the validity of our model. Our two baseline network are all lightweight CNN architectures. So it's easy to embed our model in mobile devices.

Acknowledgments. This work was supported in part by National Natural Science Foundation of China under Grant: No. 61572050, 91538111, 61429201, 61620106009, 61332016, U1636214, 61650202, and the National 1000 Youth Talents Plan, in part by National Basic Research Program of China (973 Program): 2015CB351800, in part by Key Research Program of Frontier Sciences, CAS: QYZDJ-SSW-SYS013.

References

1. Chatfield, K., Simonyan, K., Vedaldi, A., Zisserman, A.: Return of the devil in the details: delving deep into convolutional nets. arXiv preprint arXiv:1405.3531 (2014)
2. Chen, L.C., Yang, Y., Wang, J., Xu, W., Yuille, A.L.: Attention to scale: scale-aware semantic image segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3640–3649 (2016)
3. Feris, R.S., et al.: Large-scale vehicle detection, indexing, and search in urban surveillance videos. IEEE Trans. Multimed. **14**(1), 28–42 (2012)
4. Fu, J., Zheng, H., Mei, T.: Look closer to see better: recurrent attention convolutional neural network for fine-grained image recognition. In: CVPR, vol. 2, p. 3 (2017)
5. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks, vol. 7. arXiv preprint arXiv:1709.01507 (2017)
6. Huang, S., Xu, Z., Tao, D., Zhang, Y.: Part-stacked CNN for fine-grained visual categorization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1173–1182 (2016)
7. Jia, Y., et al.: Caffe: convolutional architecture for fast feature embedding. In: Proceedings of the 22nd ACM International Conference on Multimedia, pp. 675–678. ACM (2014)
8. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3D object representations for fine-grained categorization. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 554–561 (2013)
9. Li, S., Bak, S., Carr, P., Wang, X.: Diversity regularized spatiotemporal attention for video-based person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 369–378 (2018)
10. Li, W., Zhu, X., Gong, S.: Harmonious attention network for person re-identification. In: CVPR, vol. 1, p. 2 (2018)
11. Liu, H., Tian, Y., Yang, Y., Pang, L., Huang, T.: Deep relative distance learning: tell the difference between similar vehicles. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2167–2175 (2016)
12. Liu, X., Zhang, S., Huang, Q., Gao, W.: Ram: a region-aware deep model for vehicle re-identification. In: ICME (2018)
13. 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. IEEE (2016)
14. Liu, X., Liu, W., Mei, T., Ma, H.: A deep learning-based approach to progressive vehicle re-identification for urban surveillance. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9906, pp. 869–884. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46475-6_53
15. Lu, J., Yang, J., Batra, D., Parikh, D.: Hierarchical question-image co-attention for visual question answering. In: Advances In Neural Information Processing Systems, pp. 289–297 (2016)
16. Martinel, N., Micheloni, C., Foresti, G.L.: Kernelized saliency-based person re-identification through multiple metric learning. IEEE Trans. Image Process. **24**(12), 5645–5658 (2015)
17. Qian, Q., Jin, R., Zhu, S., Lin, Y.: Fine-grained visual categorization via multi-stage metric learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3716–3724 (2015)

18. Shen, Y., Xiao, T., Li, H., Yi, S., Wang, X.: Learning deep neural networks for vehicle re-ID with visual-spatio-temporal path proposals. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 1918–1927. IEEE (2017)
19. Si, J., et al.: Dual attention matching network for context-aware feature sequence based person re-identification. arXiv preprint arXiv:1803.09937 (2018)
20. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
21. Su, C., Li, J., Zhang, S., Xing, J., Gao, W., Tian, Q.: Pose-driven deep convolutional model for person re-identification. In: ICCV (2017)
22. Wang, Y., Choi, J., Morariu, V., Davis, L.S.: Mining discriminative triplets of patches for fine-grained classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1163–1172 (2016)
23. Wang, Z., et al.: Orientation invariant feature embedding and spatial temporal regularization for vehicle re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 379–387 (2017)
24. Wei, L., Liu, X., Li, J., Zhang, S.: VP-ReID: vehicle and person re-identification system. In: ICMR (2018)
25. Wei, L., Zhang, S., Gao, W., Tian, Q.: Person transfer GAN to bridge domain gap for person re-identification. In: CVPR (2018)
26. Wei, L., Zhang, S., Yao, H., Gao, W., Tian, Q.: Glad: global-local-alignment descriptor for pedestrian retrieval. In: ACM MM (2017)
27. Xu, Q., Yan, K., Tian, Y.: Learning a repression network for precise vehicle search. arXiv preprint arXiv:1708.02386 (2017)
28. Yan, K., Tian, Y., Wang, Y., Zeng, W., Huang, T.: Exploiting multi-grain ranking constraints for precisely searching visually-similar vehicles. In: The IEEE International Conference on Computer Vision (ICCV) (2017)
29. Yang, L., Luo, P., Change Loy, C., Tang, X.: A large-scale car dataset for fine-grained categorization and verification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3973–3981 (2015)
30. Yao, H., Zhang, S., Zhang, Y., Li, J., Tian, Q.: Coarse-to-fine description for fine-grained visual categorization. IEEE Trans. Image Process. **25**(10), 4858–4872 (2016)
31. Zhang, X., Zhou, F., Lin, Y., Zhang, S.: Embedding label structures for fine-grained feature representation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1114–1123 (2016)
32. Zhou, Y., Shao, L.: Aware attentive multi-view inference for vehicle re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6489–6498 (2018)
33. Zhu, Z., Wu, W., Zou, W., Yan, J.: End-to-end flow correlation tracking with spatial-temporal attention. arXiv preprint arXiv:1711.01124 (2017)