Deep Ensemble Effectively and Efficiently for Vehicle Instance Retrieval

Zhengyan Ding[1], Xiaoteng Zhang[1], Shaoxi Xu[1], Lei Song[1,2(✉)], and Na Duan[1]

[1] The Third Research Institute of the Ministry of Public Security, Shanghai 201204, China
[2] Shenzhen Key Laboratory of Media Security, Shenzhen University, Shenzhen 518060, China
dzy_wlw@163.com, zxt_wlw@126.com, gbzfx_sunny@163.com, songlei9312@126.com, naduan323@163.com

Abstract. This paper aims to highlight instance retrieval tasks centered around 'vehicle', due to its wide range of applications in surveillance scenario. Recently, image representations based on the convolutional neural network (CNN) have achieved significant success for visual recognition, including instance retrieval. However, many previous retrieval methods have not exploit the ensemble abilities of different models, which achieve limited accuracy since a certain kind of visual representation is not comprehensive. So we propose a Deep Ensemble Efficiently and Effectively (DEEE) framework, to preserve the impressive performance of deep representations and combine various deep architectures in a complementary way. It is demonstrated that a large improvement can be acquired with slight increase on computation. Finally, we evaluate the performance on two public vehicle datasets, VehicleID and VeRi, both outperforming state-of-the-art methods by a large margin.

Keywords: Instance retrieval · Vehicle · Deep ensemble

1 Introduction

Visual instance search is one of the core tasks in the field of computer vision and has been evolving rapidly in recent years. Given a query image example, the basic goal of instance-level retrieval is to search for images that contain the same instance, also viewed as a re-ID task.

With the ground-breaking success of deep learning based methods, image descriptors produced by CNN are significantly improving state-of-the-art performance for various problems including image classification [1–4], object detection [5, 6], etc. It comes as no surprise that pre-trained CNN models for a source task (e.g., classification) are capable of being applied to another target domain (e.g., retrieval), due to the generalization power of transfer learning [7].

Motivated by these advances, instance search approaches based on deep features have attracted sustained attention [7–9] both from academic research and from industrial applications. In this paper, we focus on the problem of instance-level retrieval centered around 'vehicle'. As shown in Fig. 1, an image is considered to match the query if it contains the same vehicle across different surveillance camera views.

A. M. Tjoa et al. (Eds.): CONFENIS 2017, LNBIP 310, pp. 192–201, 2018.
https://doi.org/10.1007/978-3-319-94845-4_17

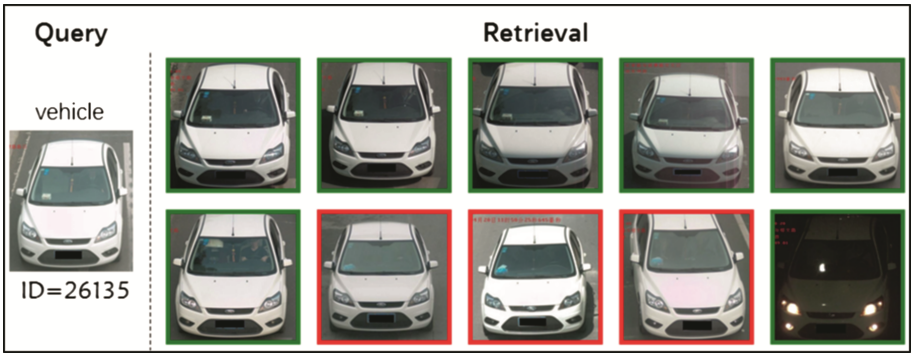

Fig. 1. Example retrieval results on VehicleID [15] dataset. The images in the first column are the query images. The retrieval images are sorted according to the similarity scores from left to right, which are color-coded as (green): correct, (red): incorrect. (Color figure online)

Traditional CNN-based retrieval methods [10–13] have not emphasized the importance of fusing various deep architectures into an ensemble model with slight increase on computation. In contrast, our proposed method has payed more attention to the complementarity of different models and implemented the deep ensemble framework via an effective and efficient way. To this end, this paper addresses two main challenges: (1) Model selection: How to ensemble various deep architectures to obtain an evident performance boost with marginal extra cost; (2) Feature selection: How to exploit multi-level features to generate a more comprehensive and compact fusion feature.

Firstly, we improve the retrieval performance by fusing various deep architectures into a single model. After comparing the advantages of different CNN models, residual-like network [4] is selected, as it can avoid the vanishing gradient problem significantly and some recent works [14] indicate that ResNet behaves like ensembles of relatively shallow networks, with the fusion strengths inherently. In terms of training loss functions, we finetune the network with both the verification and identification losses, inspired by 'Mixed Difference Network' [15]. Secondly, we utilize the Feature Pyramid Network (FPN) [6] method to ensemble multi-level features by a top-down architecture with lateral connections. After FPN fusion, we not only improve the retrieval performance significantly, but also obtain a more compact and efficient feature representation.

The rest of this paper is organized as follows: In Sect. 2, we review some related works and in Sect. 3, we introduce the proposed approaches in details. The experimental results are presented and analyzed in Sect. 4. Finally, we draw a conclusion in Sect. 5.

2 Related Work

In this section, we will describe previous works relevant to the approach discussed in this paper. Most of the works utilize CNN models for feature extraction and shed light on 'vehicle' retrieval tasks, especially in real-word surveillance scene.

2.1 Deep Representation for Instance Retrieval

Many works in the literature have proposed CNN-based representations for image retrieval. Razavian et al. [7] first investigate the use of CNN features for various computer vision tasks, including image retrieval. A typical CNN consists of several convolutional layers, followed by fully connected layers and ends with a softmax layer producing a distribution over the training classes. However, different from classification task, the pooled convolutional features often perform better than the fully connected layers [10]. Local convolutional features are similar to traditional hand-crafted features, e.g., SIFT, and various aggregation methods are proposed to improve the retrieval performance. To our best knowledge, most of the previous retrieval works focus on aggregating convolutional features from a single layer, but overlook the complementary properties of features from multiple layers. In fact, a series of excellent approaches have improved detection and segmentation performance by fusing different layers in a CNN model. For example, the FCN [16] algorithm sums partial scores for each category over multiple scales to compute semantic segmentations. Hypercolumns [17] uses a similar method for object instance segmentation. FPN [6] develops a top-down architecture with lateral connections for building high-level semantic feature maps at all scales. In this paper, our proposed method firstly introduces an extension of the FPN architecture to instance retrieval task and improves the results significantly.

2.2 Vehicle Instance Retrieval

Vehicle-related research includes detection, tracking, joint detection and 3D parsing. The growing explosion in the use of surveillance cameras highlights the importance of vehicle search from a large-scale image or video database. Liu et al. [15] released a carefully-organized largescale image database 'VehicleID', which includes multiple images of the same vehicle captured by different real-world cameras in a city. To facilitate progressive vehicle Re-Id research, Liu et al. [18] collect vehicle instance dataset named VeRi-776 from large-scale urban surveillance videos, which contains not only massive vehicles with diverse attributes and high recurrence rate, but also sufficient license plates and spatiotemporal labels. In this paper, comprehensive evaluations on the above two datasets have shown that ensemble of various deep architectures (e.g., verification and identification loss) and multi-level deep features will make a good contribution to boost the retrieval performance, compared with state-of-the-art results in previous works.

3 Proposed Method

3.1 Baseline Framework

Our baseline framework is illustrated in Fig. 2. It can be divided into offline stage processing and online stage query. In the online stage, a query is provided by a user

based on his intension. Then, the image is transform to the corresponding feature representation through a deep CNN model. Finally, the retrieval results from image dataset are generated, ranking by similarities with the query image feature.

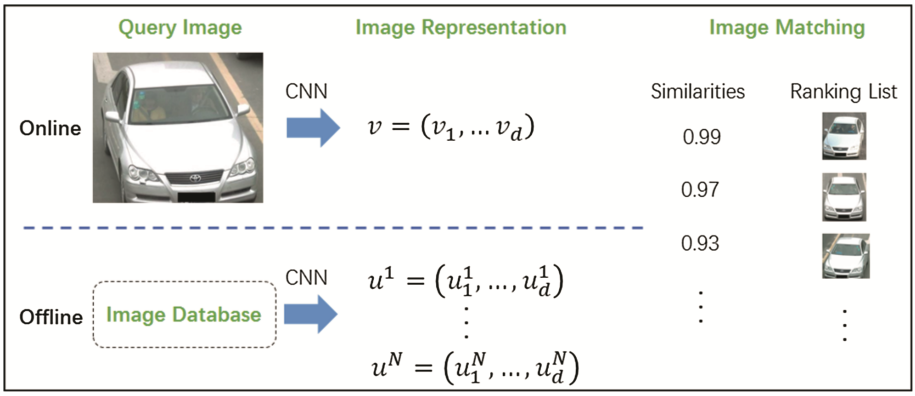

Fig. 2. Our baseline framework with online stage query and offline stage processing

As shown in Fig. 2, image representation is the core part of our framework, in which we adopt a compact representation pooled from activations of convolutional layers. This kind of global feature is very effective for instance-level retrieval, as formulated in Eq. 1:

Query Feature:

$$v = (v_1, \ldots, v_d) \tag{1}$$

where d denotes the feature dimension. In this paper, all fully connected layers are discarded in the inference part, and the image representation, called Average Activations of Convolutions (AAC), is simply constructed by average pooling over all dimensions per feature map. The dimension of our pooled feature is equal to the number of feature map channels, e.g. 1024 or 2048. To yield a shorter feature vector and improve retrieval efficiency, previous methods [10] use PCA of a post-processing tool for dimensionality reduction, by analyzing the covariance matrix of all descriptors. After acquiring the final image representation of query image (indicated as v) and a certain database image (indicated as u^i), the corresponding similarity can be calculated as inner product of the two feature vectors.

3.2 Effective and Efficient Ensemble Methods

In this part, we will further introduce our effective and efficient ensemble methods for image representation, which can be divided into two main components: model selection and feature selection as shown in Fig. 3.

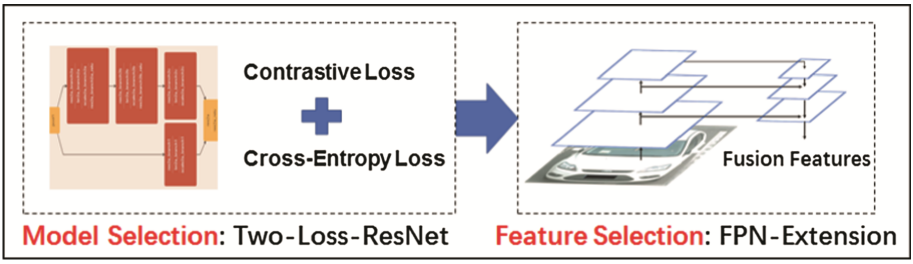

Fig. 3. Ensemble methods for image representation: model selection (Two-Loss-ResNet) and feature selection (FPN Extension)

Two-Loss-ResNet: Ensemble of Shallower Networks and Multiple Tasks
With the rapid development of CNN models on the image classification, more and more excellent CNN models emerged, like VGG-like [2], Inception-like [3] and residual-like [4] models. Compared with other previous architectures, residual networks avoid the vanishing gradient problem by introducing short paths which can carry gradient throughout the extent of very deep networks. Veit et al. [14] even proposes a novel interpretation of residual networks showing that a single ResNet model behaves as an ensemble of shallower networks, which results from ResNet's additive operations. This property is very important for improving the feature representation since that averaging a set of deep networks is an effective solution to improve the accuracy performance, widely adopted in various computer vision tasks. Considering the above observations, we choose Residual-like architectures as our basic deep pretrained model.

In consideration of the common issues for training deep CNN models, such as infeasible computational cost and model size, we decide to select ResNet-50 as our baseline model, which is a tiny-version residual network and keeps a good balance between effectiveness and efficiency. Compared with traditional VGG-based image retrieval methods [15], ResNet-50 demonstrates a better speed with 3.8 billion FLOPs, which is only 25% of standard VGG-16 model (15.3 billion FLOPs). More significantly, ResNet-50 also yields a much better accuracy for instance retrieval task, as shown in the following experiments (See Tables 1 and 2).

Training a deep model with different loss functions is always a good way to ensemble the representative abilities for multiple tasks [5, 15]. To this end, two commonly-used loss functions are selected, cross-entropy loss and contrastive loss. Similar to conventional multiclass recognition approaches, we use the cross-entropy loss for identity prediction. Meanwhile, Siamese architecture and contrastive loss are adopted, in which we train a two-branch network and each branch is a clone of the other, meaning that they share the same parameters. Then we sum the above two losses together, to measure multiple outputs simultaneously.

Our ensemble method of deep architectures can be denoted as **Two-Loss-Res50**, fusing two loss functions based on the output of ResNet-50 model,

FPN Extension: Ensemble of Multi-level Deep Features

As a basic component in visual recognition systems, feature image pyramids are heavily used to generate scale-invariant representations. In order to fuse multi-level features in the training phase efficiently, we introduce an extension of Feature Pyramid Network (FPN) [6], which has been a popular method used in the object detection model recently. The top-down architecture with lateral connections is developed to exploit the inherent multi-level, pyramidal hierarchy of deep CNN models with marginal extra cost. Our experimental results have demonstrated that such multi-level ensemble representation is very compact, suitable for instance-level retrieval tasks. (See Table 1)

Through the above fusion strategies in training phase, the obtained CNN model is capable of yielding three kinds of features. As shown in Fig. 4, Feature_1 is a linear transformation of output from the 5th stage, which consist of only high-level semantics. On the contrary, Feature_3 is an ensemble of output from three stages (3&4&5), which represents more local and detailed information.

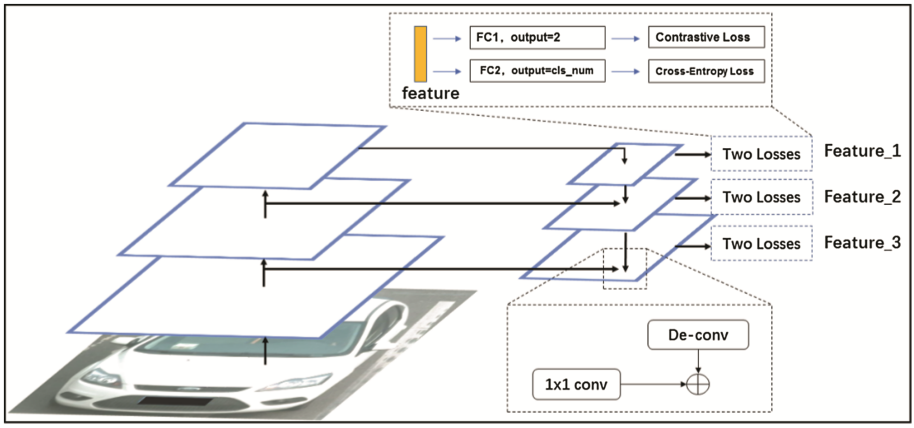

Fig. 4. Extension of FPN [6] to ensemble multi-level deep features

In this paper, we adopt Feature_3 as our final multi-level image representation for instance retrieval. The corresponding ensemble method can be denoted as **Two-Loss-Res50+FPN**.

4 Experiments

The proposed DEEE framework is evaluated by a standard performance metric, i.e., MAP, which is the mean of average precision scores for all query images over all the returned images. In addition, this paper makes most efforts on the retrieval tasks of 'vehicle', and two related datasets are used in our experiments. For fair comparison with existing methods, we follow the standard protocol of train/test split. All the results are obtained by **single-query**.

4.1 Datasets

VehicleID [15]: VehicleID dataset is a large-scale vehicle dataset that contains 221,763 images of 26,267 vehicles, where the training set contains 110,178 images of 13,134 vehicles and the testing set contains 111,585 images of 13,133 vehicles. Following the settings in [15], we use 3 test splits of different sizes constructed from the testing set: small, medium and large.

VeRi [18]: VeRi dataset contains over 50,000 images of 776 vehicles captured by 20 cameras covering an 1.0 km^2 area in 24 h, which makes the dataset scalable enough for vehicle Re-Id and other related research.

4.2 Experiment Setup

According to the above analysis, we choose ResNet-50 as our base model for retrieval tasks. The model pretrained on ImageNet classification dataset is used to initialize the weight parameters. All the experiments are implemented on the CAFFE platform. We use the mini-batch SGD algorithm to learn the network parameters, where the batch size is set to 256, and momentum set to 0.9. We only experiment multi-level feature ensemble method with FPN on VehicleID dataset, because that the image samples in VehicleID have higher resolution and the detailed information is abundant, compared with VeRi dataset.

4.3 Comparison with State of the Art

We compare our method with state-of-the-art methods on the two datasets.

For the VehicleID dataset: (1) DRDL [15] method exploits a two-branch deep convolutional network to project raw vehicle images into an Euclidean space. The triplet loss used for deep metric learning is replaced by a novel loss function: coupled clusters loss (CCL). (2) HDC [19] method ensembles a set of models with different complexities in cascaded manner and mine hard examples adaptively. (3) GS-TRS [20] method forms the triplet samples across different categories as well as different groups, through partitioning training images within each category into a few groups. The comparison of statistic results can be found in Table 1.

For the VeRi dataset, we compared with PROVID method proposed in [18], which is a novel deep learning-based approach to PROgressive Vehicle re-ID. Table 2 lists the detailed results.

Table 1. Comparison results for VehicleID dataset (MAP)

Method name	Small	Medium	Large
DRDL+CCL [15]	0.546	0.481	0.455
HDC+Contrastive [19]	0.655	0.631	0.575
GS-TRS [20]	0.746	0.734	0.715
Two-Loss-Res50 (2048d)	0.823	0.775	0.737
Two-Loss-Res50+FPN (256d)	**0.843**	**0.794**	**0.760**

Table 2. Comparison results for VeRi dataset (MAP)

Method name	Standard train-test split [18]
PROVID [18]	0.278
Two-Loss-Res50 (2048d)	**0.672**

4.4 Ablation Studies

Improving Retrieval Effectiveness by Fusing Multi-scale Features
In this section, we will analyze the typical example that our ensemble framework improves the retrieval result effectively. As shown in Fig. 5, more discriminative features are taken into account after FPN fusion, such as the color of the car roof.

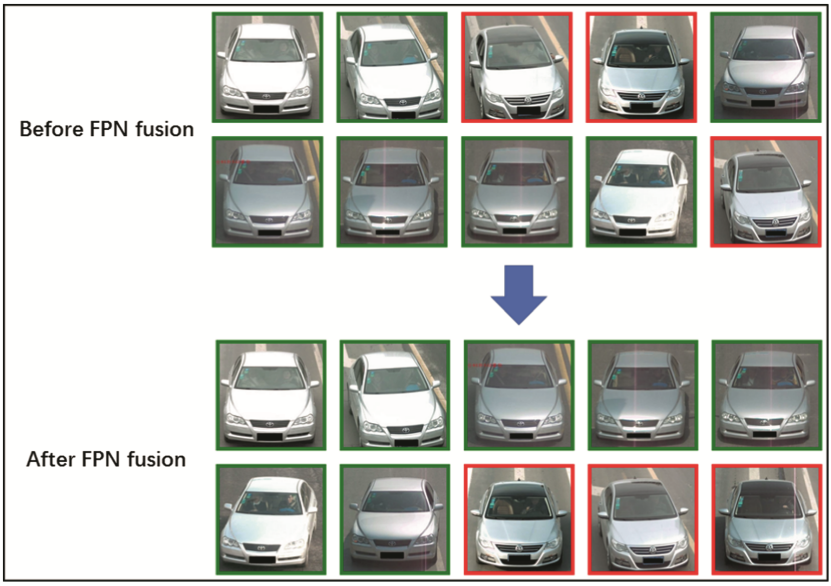

Fig. 5. Results comparison before and after FPN fusion. All the correct images (coded as green) rank before the incorrect images (coded as red) with FPN fusion. (Color figure online)

Improving Retrieval Efficiency by Feature Dimension Reduction
As shown in Table 1, the ensemble image representation using **Two-Loss-Res50+FPN** is 256d, which is only 1/8 of the original image representation using **Two-Loss-Res50**. This property will accelerate the retrieval procedure substantially, especially for large-scale image database.

5 Conclusion

In this paper, we propose a deep ensemble framework that simultaneously considers the effectiveness and efficiency, focusing on the instance-level retrieval task centered around 'vehicle'. A set of problems are investigated comprehensively, including the selection of base CNN models, loss functions and multi-level features for training a discriminative ensemble model. The final experimental results indicate that the proposed DEEE framework is very effective and achieve the state-of-the-art results with a considerably short image representation. Our study also suggests that an efficient fusion method is capable of generating strong representation for instance retrieval tasks, providing a practical solution for balancing the speed and accuracy issues in the future research.

Acknowledgements. The authors of this paper are members of Shanghai Engineering Research Center of Intelligent Video Surveillance. Our research was sponsored by following projects: the National Natural Science Foundation of China (61403084, 61402116); Program of Science and Technology Commission of Shanghai Municipality (No. 15530701300, 15XD1520200); 2012 IoT Program of Ministry of Industry and Information Technology of China; Key Project of the Ministry of Public Security (No. 2014JSYJA007); the Project of the Key Laboratory of Embedded System and Service Computing, Ministry of Education, Tongji University (ESSCKF 2015-03); Shanghai Rising-Star Program (17QB1401000); The Special Fund for Basic R&D Expenses of Central Level Public Welfare Scientific Research Institutions (C17384).

References

1. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
2. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR (2015)
3. Szegedy, C., Liu, W., Jia, Y., et al.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9 (2015)
4. He, K., Zhang, X., Ren, S., et al.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
5. Ren, S., He, K., Girshick, R., et al.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, pp. 91–99 (2015)
6. Lin, T.Y., Dollár, P., Girshick, R., et al.: Feature pyramid networks for object detection. arXiv preprint arXiv:1612.03144 (2016)
7. Razavian, A.S., Azizpour, H., Sullivan, J., et al.: CNN features off-the-shelf: an astounding baseline for recognition. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 512–519. IEEE (2014)
8. Razavian, A.S., Sullivan, J., Carlsson, S., et al.: Visual instance retrieval with deep convolutional networks. ITE Trans. Media Technol. Appl. **4**(3), 251–258 (2016)
9. Tolias, G., Sicre, R., Jégou, H.: Particular object retrieval with integral max-pooling of CNN activations. In: ICLR (2016)

10. Yue-Hei Ng, J., Yang, F., Davis, L.S.: Exploiting local features from deep networks for image retrieval. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 53–61 (2015)
11. Kalantidis, Y., Mellina, C., Osindero, S.: Cross-Dimensional weighting for aggregated deep convolutional features. In: Hua, G., Jégou, H. (eds.) ECCV 2016. LNCS, vol. 9913, pp. 685–701. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46604-0_48
12. Babenko, A., Lempitsky, V.: Aggregating local deep features for image retrieval. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1269–1277 (2015)
13. Hoang, T., Do, T.T., Tan, D.K.L., et al.: Selective Deep Convolutional Features for Image Retrieval. arXiv preprint arXiv:1707.00809 (2017)
14. Veit, A., Wilber, M.J., Belongie, S.: Residual networks behave like ensembles of relatively shallow networks. In: Advances in Neural Information Processing Systems, pp. 550–558 (2016)
15. Liu, H., Tian, Y., Yang, Y., et al.: 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)
16. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440 (2015)
17. Hariharan, B., Arbeláez, P., Girshick, R., et al.: Hypercolumns for object segmentation and fine-grained localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 447–456 (2015)
18. 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
19. Yuan, Y., Yang, K., Zhang, C.: Hard-aware deeply cascaded embedding. In: Proceedings of the IEEE International Conference on Computer Vision (2017)
20. Bai, Y., Gao, F., Lou, Y., et al.: Incorporating Intra-Class Variance to Fine-Grained Visual Recognition. arXiv preprint arXiv:1703.00196 (2017)