IMPROVING TRIPLET-WISE TRAINING OF CONVOLUTIONAL NEURAL NETWORK FOR VEHICLE RE-IDENTIFICATION

Yiheng Zhang, Dong Liu, Zheng-Jun Zha*

CAS Key Laboratory of Technology in Geo-Spatial Information Processing and Application System,
University of Science and Technology of China, Hefei 230027, China
zhangyih@mail.ustc.edu.cn, {dongeliu,zhazj}@ustc.edu.cn

ABSTRACT

Vehicle re-identification (re-id) plays an important role in the automatic analysis of the drastically increasing urban surveillance videos. Similar to the other image retrieval problems, vehicle re-id suffers from the difficulties caused by various poses of vehicles, diversified illuminations, and complicated environments. Triplet-wise training of convolutional neural network (CNN) has been studied to address these challenges, where the CNN is adopted to automate the feature extraction from images, and the training adopts triplets of (query, positive example, negative example) to capture the relative similarity between them to learn representative features. The traditional triplet-wise training is weakly constrained and thus fails to achieve satisfactory results. We propose to improve the triplet-wise training at two aspects: first, a stronger constraint namely classification-oriented loss is augmented with the original triplet loss; second, a new triplet sampling method based on pairwise images is designed. Our experimental results demonstrate the effectiveness of the proposed methods that achieve superior performance than the state-of-the-arts on two vehicle re-id datasets, which are derived from real-world urban surveillance videos.

***Index Terms*—** Convolutional neural network (CNN), Triplet-wise training, Vehicle re-identification.

1. INTRODUCTION

With the explosive growth of the video data captured by surveillance cameras, the demand for surveillance video analysis capabilities increases rapidly. A large number of vehicle related tasks, such as vehicle detection, tracking, classification and verification, play important roles in surveillance video analysis [1]. Among these tasks, vehicle re-identification (re-id) is to find out the images captured by other cameras that contain the same vehicle with the query image. With the help of vehicle re-id, the target vehicle can be automatically discovered, located and tracked across multiple cameras, saving the manual labor and cost.

Traditionally, the problem of vehicle re-id is solved by the combination of multiple clues and/or sensor data, such as the passage time [2] and the wireless magnetic sensors [3]. However, these methods either need extra cost of hardware, or are sensitive to the fickle environment. In addition, the license plate that contains the unique ID of a vehicle is an important clue, thus license plate related technologies have been researched in depth. Nevertheless, the license plate may be intentionally occluded, removed, or even forged, especially in criminal circumstances. Consequently, a vehicle re-id method based purely on appearance has both practical value and research significance.

Appearance based re-id has been studied especially for person re-id for several years. Well designed low-level features, such as color histogram [4], local binary pattern (LBP) [5], scale-invariant feature transform (SIFT) [6], have been adopted to represent the target person so as to be immune to the variations across cameras. Zhao et al. [7] proposed to learn mid-level filters to resolve the difficulty of person re-id caused by view variance. Recently, deep convolutional neural network (CNN), which has been shown efficient in several computer vision tasks, was also introduced into the person re-id research, and obtained remarkable results [8, 9].

Similar to person re-id as well as the other image retrieval problems, appearance based vehicle re-id also suffers from the difficulties caused by various poses of vehicles, diversified illuminations, and complicated environments. Besides, one special difficulty of vehicle re-id is due to the fact that the vehicles belonging to the same model are extremely similar to each other. It is challenging even for human eyes to distinguish the vehicles that have different IDs but belong to the same model. Therefore, vehicle re-id needs to be robust to tolerate the large intra-ID variance and at the same time discriminative to identify the small inter-ID difference.

In recent years, appearance based vehicle re-id has attracted the attention of researchers [10, 11, 12]. Specifically, in [10], an appearance based method known as "FACT"

* Corresponding author. This work was supported by the National Key Research and Development Plan under Grant 2016YFC0801001, by the Natural Science Foundation of China (NSFC) under Grants 61390512, 61331017, 61472392, 61622211, and by the Fundamental Research Funds for the Central Universities under Grant WK3490000001.

978-1-5090-6067-2/17/$31.00 ©2017 IEEE

is proposed, which combines the color, texture, and high-level semantic information extracted by deep neural network; a large-scale benchmark dataset for vehicle re-id, named "VeRi," is also presented. Then, in [11], the information of license plates and spatiotemporal relations (STR) of vehicles are augmented with FACT to enhance the performance of vehicle re-id, where the information of license plates is utilized by training a siamese network. In [12], another deep learning approach termed deep relative distance learning (DRDL) is proposed for vehicle re-id, which features a coupled cluster loss function and a mixed difference network structure, and trains the network with both vehicle ID and vehicle model; also, a large-scale dataset for vehicle re-id known as "VehicleID" is presented.

Inspired by the great successes of deep network based methods in computer vision, we investigate the vehicle re-id problem via an appearance-based deep learning approach in this paper. Essentially, vehicle re-id can be solved by identifying the relative similarity between vehicle images, where the similarity can be calculated by a pre-trained deep network, i.e. distance learning. Triplet-wise training, which adopts triplets of (query, positive example, negative example) to train the network, is an intuitive manner for distance learning and has been studied in e.g. face recognition [13] and person re-id [14]. However, the traditional triplet-wise training suffers from inefficiency due to its too loose constraint, that is, it only requires the distance between query and positive example to be less than the distance between query and negative example. The *accurate* distance between images, especially between query and negative example, is not well formulated. Therefore, we propose to improve the triplet-wise training at two aspects:

- Since the original triplet loss is too weak to constrain the distance learning, we propose to augment it with a stronger constraint, namely classification-oriented loss. This loss requires the images of the same class to be as similar as possible, thus regularizes the distance learning to some extent.

- We also propose a new triplet sampling method based on pairwise images during network training. Instead of random sampling of triplets, our proposed triplet sampling based on pairs can ensure the simultaneous constraint of both intra-ID similarity and inter-ID dissimilarity on sampled images.

We have performed experiments on both VeRi and VehicleID datasets, and experimental results demonstrate that the proposed classification-oriented loss and triplet sampling method help to improve the triplet-wise training and lead to better re-id results. Overall, our scheme achieves much better results than the previously reported in [10, 11, 12], without increasing the complexity of the trained network.

2. THE PROPOSED APPROACH

In this section, the proposed architecture and method are presented. The architecture is illustrated in Fig. 1. It consists of three parts: a shared deep CNN which is expected to learn a mapping from raw images to Euclidean space where the distance can reflect the relevance between the images, a triplet stream for calculating the distances and providing the constraint of triplet loss, and a classification stream for ID-level supervision provided by the classification-oriented loss. Moreover, the image triplets are generated by our proposed triplet sampling method.

The original triplet loss and its shortcomings will be discussed in Section 2.1. The two contributions proposed to improve the triplet-wise training will be discussed in Section 2.2 and Section 2.3, respectively.

2.1. Triplet Loss

In order to learn a mapping from raw images to their representations in a Euclidean space, where the distance in the Euclidean space reflects the relevance between images, training a deep CNN with triplet loss is intuitive and has been shown efficient [13, 14].

Let $\mathcal{X} = \{x_i | i = 1, 2, \ldots, N\}$ denotes the training set, where x_i represents the i-th image in the dataset and N is the amount of images in the training set. The embedding of the image can be described as a function $f(x_i) \in \mathbb{R}^d$, where d is the dimensionality of the Euclidean space. Additionally, in order to regularize the representations, the embedded image features are constrained on a d-dimensional hypersphere, i.e. $\|f(x_i)\|_2^2 = 1$. This constraint can be imposed by l_2-normalization [13].

For an image triplet $\{x_i, x_i^+, x_i^-\}$, $\{x_i,\, x_i^+\}$ is termed a positive pair of images that captured the same vehicle, and $\{x_i,\, x_i^-\}$ is termed a negative pair of images that belong to different vehicles. The objective of triplet loss is making the mapped features of triplets to satisfy the constraint:

$$\begin{aligned} &\|f(x_i) - f(x_i^-)\|_2^2 - \|f(x_i) - f(x_i^+)\|_2^2 \geq \alpha \\ &\forall (x_i, x_i^+, x_i^-) \in \mathcal{T} \end{aligned} \tag{1}$$

where α is the parameter of expected gap between the distances of positive pair and negative pair. $\mathcal{T}$ is the set of triplets, which can contain all possible combinations of the images in $\mathcal{X}$. Hence, the triplet loss to be minimized during the training process can be represented as,

$$\begin{aligned} L_T = &\sum_{(x_i, x_i^+, x_i^-) \in \mathcal{T}} \max(0, \\ &\|f(x_i) - f(x_i^+)\|_2^2 - \|f(x_i) - f(x_i^-)\|_2^2 + \alpha) \end{aligned} \tag{2}$$

During the minimization of the triplet loss, the point of $f(x_i^+)$ is pulled towards the anchor point $f(x_i)$ and the point of

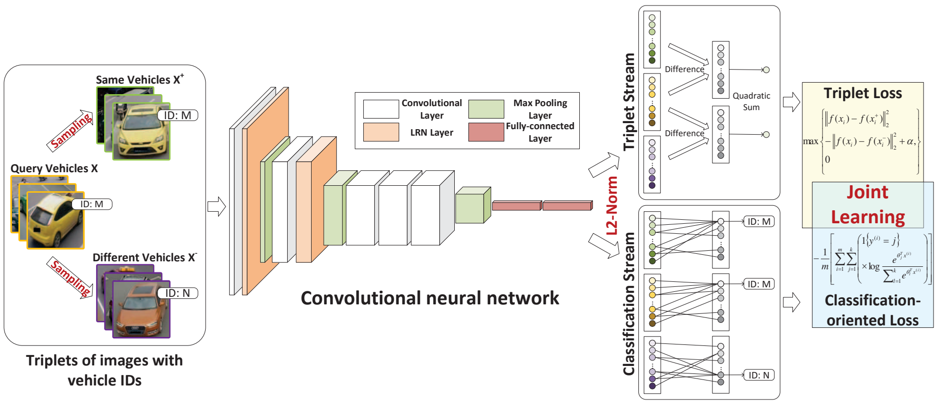

Fig. 1. Our proposed triplet-wise training architecture for vehicle re-id. The input to the network is in the form of image triplets, which are generated by our proposed triplet sampling method. The shared deep CNN followed by two streams is exploited to learn the representations of images. The triplet stream provides the constraint of triplet loss. The classification stream provides the classification-oriented loss, which strengthens the constraint on image representations.

$f(x_i^-)$ is pushed in the opposite direction with respect to the anchor point. After training, the features of the same vehicle are closer to each other than the features of different vehicles, then the image with less distance to the query image has higher probability to be the same vehicle ID.

Although the triplet loss can help to train a deep CNN with some discriminative ability, it also has some limitations. A triplet only focuses on three images belonging to two vehicles and minimizing triplet loss can easily lead to misleading of the negative point. As shown in Fig. 2, the negative point N is pushed away from the points that belong to the same vehicle, when triplet loss is applied on the selected triplet. As a result, the speed of convergence is slowed down by the misleading of the negative point. What is worse, the misleading may not be corrected, since the misleading negative point may not appear in the other triplets selected for training (as the amount of all possible triplets is huge, a sampling is typically used in training). Another weakness of the traditional triplet-wise training is that, for a large number of randomly selected triplets, the triplet loss can be easily satisfied as it is quite loose; such already satisfied triplets cannot help in training, which lead to a waste of time.

2.2. Classification-Oriented Loss

这也能发论文?????

The problems associated with the original triplet loss are essentially caused by the weak constraint of distance comparison. Indeed, how to pose a constraint in learning similarity/distance for re-id remains a difficult problem, since the training data do not provide accurate labels of *distances*. Intuitively, the distances between same-vehicle images should be small and the distances between different-vehicle images should be large, but how small/large is not quantitatively available from the training data.

We propose a classification-oriented loss in addition to

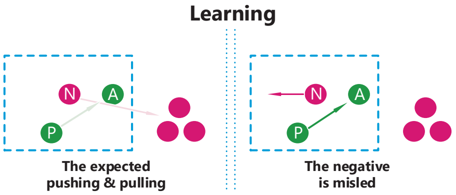

Fig. 2. This figure depicts the possible inappropriate moving of the negative point when minimizing the original triplet loss. For an image triplet (A, P, N), where A is anchor, P is positive, and N is negative, the negative point N is pushed away from A to minimize the original triplet loss, but it is also pushed away from the points that belong to the same vehicle (red-colored points in the figure).

triplet loss for vehicle re-id. Specifically, a classification stream is added at the end of the CNN based feature extractor, as shown in Fig. 1. We simply consider each vehicle ID as a "class," i.e. the number of classes in training is equal to the amount of vehicle IDs in the training set. The classification-oriented loss is formulated as that in usual softmax based classifiers,

$$L_C = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K}\mathbf{1}\left(y_i = j\right)\log\frac{e^{\theta_j^T f(x_i)}}{\sum_{l=1}^{K} e^{\theta_l^T f(x_i)}} \tag{3}$$

where $\mathbf{1}(\cdot)$ is the indicator function, $y_i \in \{1, 2, ..., K\}$ is the class (vehicle ID) of the image x_i, K is the number of classes, and θ's are the parameters of the full-connection layer of the classification stream.

The classification-oriented loss implicitly imposes a constraint that the embedded features of the images of the same vehicle should be similar. Indeed, by using the classification-oriented loss, the features learnt by the CNN tend to form clusters, which is a stronger constraint than the original triplet loss. Especially, the misleading problem shown in Fig. 2 can be eliminated by the classification-oriented loss.

It is worth noting that, the vehicle IDs in test are probably not those in the training set, thus re-id problem is not equivalent to classification. However, using classification-oriented loss can benefit the training of the CNN as a feature extractor, and the CNN applies to the images of the other vehicles as well. Moreover, in [12], an additional branch for classification is also trained, but that branch is separate and trained for vehicle models; on the contrary, in our scheme we combine the network for triplet-wise training and for classification, and we use vehicle IDs as classes without requiring the additional information of vehicle models.

2.3. Improved Triplet Sampling

As the amount of possible triplets constructed from a given training set can be very huge, a proper sampling is usually adopted in triplet-wise training. As mentioned before, randomly sampled triplets may not be capable in correcting the misleading shown in Fig. 2.

We propose an improved triplet sampling method to correct the misleading problem. Since the misleading only affects negative examples, we ensure a negative example in one triplet to be an anchor or a positive example in another triplet, so as to provide counteraction.

As the training adopts stochastic gradient descent with mini-batches, we present our triplet sampling method at mini-batch level, as shown in Fig. 3. To generate a mini-batch of triplets, we first sample several positive pairs, i.e. pairs of images of the same vehicle. Next, each pair (A, B) results in two triplets: in one triplet, A is anchor and B is positive, in another B is anchor and A is positive, and negative examples are randomly chosen from the other pairs.

During implementation, the sampling method is applied on the features rather than original images, in order to avoid an image to pass through the CNN multiple times. In this manner, the time cost of training is also decreased because of the reuse of image features.

3. EXPERIMENTS

3.1. Datasets

We perform experiments using VeRi [10] and VehicleID [12], both of which are released recently for vehicle re-id and the images are extracted from real-world surveillance videos. VeRi contains over 40,000 bounding boxes of 619 vehicles captured by 20 surveillance cameras in unconstrained traffic scene. Each vehicle is captured by 2–18 cameras at different viewpoints, different resolutions, and under different illuminations; consecutive frames of the same vehicle in the same scene are available. VehicleID contains 221,763 images of 26,267 vehicles, the vehicle in each image being captured from either the front or the back. The license plates of the vehicles are intentionally removed from the images for privacy concern.

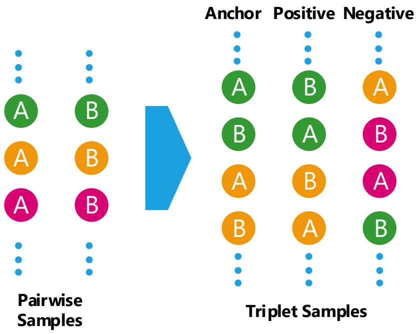

Fig. 3. The proposed triplet sampling method applies on positive image pairs. One positive pair results in two triplets, in which the anchor and positive are exchanged, and the negative is selected from one image of the other pairs.

3.2. Settings

Our improved triplet-wise training is implemented based on the deep learning framework Caffe [15]. The deep CNN for feature extraction is fine tuned from VGG_CNN_M (for VeRi) or VGG_CNN_M_1024 (for VehicleID) [16], to make fair comparisons with [11] and [12], respectively. Both models were pre-trained on the ILSVRC-2012 dataset [17]. We use stochastic gradient descent with a momentum of $\mu = 0.9$ and weight decay $\lambda = 0.0005$ during the training process. The base learning rate is set to 0.001 at the beginning and is gradually decreased by multiplying 0.1 after every k iterations. k is set to 4000 for VeRi dataset, where the gap parameter α in (1) is also gradually increased. For VehicleID dataset, k is set to 20000 and the gap parameter is fixed. When using both triplet loss and classification-oriented loss, both losses are directly summed up. The mini-batch size of the proposed triplet sampling method is set to 24 pairs ($24 \times 2 = 48$ images), and that of the randomly sampling method is set to 48 triplets ($48 \times 3 = 144$ images).

3.3. Results on VeRi Dataset

We follow the test process (cross-camera retrieval task) described in [10]. The cumulative match curve (CMC) [18], the mean average precision (mAP), HIT1, HIT5 are adopted as metrics to evaluate the overall performance. Table 1 presents the HIT1, HIT5 and mAP of different methods, and Fig. 4 illustrates the CMC curves.

It can be observed that the deep CNN based methods achieve significant improvement than the methods in [10, 11], which did not utilize CNN for extracting features from images but rather adopted handcrafted features known as FACT.

Table 1. Performance comparison on the VeRi dataset

Method		HIT1	HIT5	mAP
FACT [10]		0.5095	0.7348	0.1849
FACT+Plate-SNN+STR [11]		0.6144	0.7878	0.2770
Ours[1]	VGG+T	0.7294	0.8683	0.4730
	VGG+T+S	0.7604	0.8754	0.4850
	VGG+C	0.8480	0.9243	0.5364
	VGG+C+T	0.8641	**0.9291**	**0.5878**
	VGG+C+T+S	**0.8659**	0.9285	0.5740

[1] In our methods, C, T, S represent using classification-oriented loss, using triplet loss, and using the proposed triplet sampling method, respectively. The same hereafter.

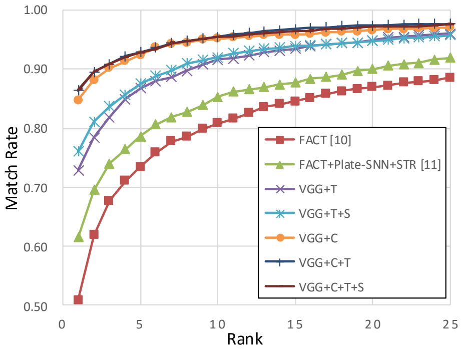

Fig. 4. The CMC curves of different methods on the VeRi dataset.

Compared with the method in [11], which additionally utilizes license plate information (Plate-SNN) and spatiotemporal relation (STR), our methods based solely on appearance achieve much better results, improving 41% in HIT1, 18% in HIT5 and 112% in mAP.

Our proposed classification-oriented loss helps to achieve improvement than the original triplet loss. Using only classification-oriented loss (VGG+C) is already better than using only triplet loss (VGG+T), and the combination (VGG+C+T) is even better. Such results demonstrate the effectiveness of using classification-oriented loss in vehicle re-id. Our proposed triplet sampling method also helps to achieve improvement than the random sampling method. The improvement of VGG+T+S (using the improved sampling) than VGG+T (using the random sampling) is 4.2% in HIT1, 0.8% in HIT5, and 2.5% in mAP. When the proposed triplet sampling method is used together with the classification-oriented loss and the triplet loss, the improvement in HIT1 is still visible.

Observing the CMC curves in Fig. 4, the proposed methods achieve better results especially at the small rank end, i.e. the first several results are more reliable. Nonetheless, using classification-oriented loss still provides improvement at the large rank end.

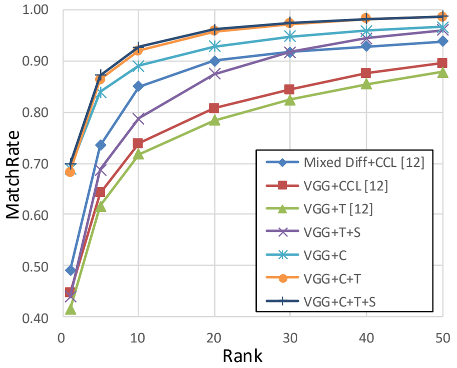

Fig. 5. The CMC curves of different methods on VehicleID dataset (gallery size is 800).

3.4. Results on VehicleID Dataset

We follow the process of training/testing described in [12], the Top-1, Top-5 and CMC curves are adopted to evaluate the performance. Table 2 and Fig. 5 illustrate the performances of the proposed methods and the methods in [12], please note that the result of using only triplet loss is already presented in [12] and thus not included in our results.

We can observe that the proposed method with both classification-oriented loss and improved triplet sampling achieves the best results in Top-5 metric. It leads to significant gain than the best method in [12] (Mixed Diff+CCL), as high as 42–65% in Top-1 and 19–29% in Top-5 on the three test sets with different gallery sizes (denoted by Small, Medium, and Large in Table 2). Similar to the results on VeRi dataset, our proposed classification-oriented loss (VGG+C) helps to achieve improvement than the original triplet loss (VGG+T). The combination of both (VGG+C+T) is also better than using either one except for the Top-1 metric, where using VGG+C is better. Also, the proposed triplet sampling method helps to achieve improvement, as can be observed by comparing VGG+T+S with VGG+T, or comparing VGG+C+T+S with VGG+C+T.

Last but not the least, benefiting from the improved triplet sampling method, the number of situations that a sample is misled by other samples of different classes in the training set of VehicleID is decreased from 359 million (VGG+T) to 271 million (VGG+T+S). And for VeRi which has less training samples and classes, this number is reduced from 74847 (VGG+T) to 47260 (VGG+T+S). Additionally, the time consumption of the training phase can be reduced from 25.3 hours to 18.2 hours in our experiments, using an NVIDIA Titan X Pascal GPU.

4. CONCLUSION

In this paper, we have proposed an improved triplet-wise training method for vehicle re-id task. Specifically, a shared

Table 2. Performance comparison on the VehicleID dataset

Method		Top-1			Top-5		
		Small	Medium	Large	Small	Medium	Large
VGG+T [12]		0.404	0.354	0.319	0.617	0.546	0.503
VGG+CCL [12]		0.436	0.370	0.329	0.642	0.571	0.533
Mixed Diff+CCL [12]		0.490	0.428	0.382	0.735	0.668	0.616
Ours	VGG+T+S	0.439	0.386	0.335	0.688	0.622	0.562
	VGG+C	0.690	**0.666**	**0.639**	0.840	0.801	0.775
	VGG+C+T	0.683	0.647	0.610	0.864	0.812	0.775
	VGG+C+T+S	**0.699**	0.662	0.632	**0.873**	**0.823**	**0.794**

deep CNN followed by triplet and classification streams is exploited to learn the representations of images from the sampled triplets. The triplet stream aims to ensure the learnt features to have the ability to represent the relative similarities between images in a triplet, while the classification stream supervised by the classification-oriented loss is to strengthen the constraint directly from the information of the vehicle IDs. At the same time, a well-designed triplet sampling method is presented to eliminate the misleading problem of the triplet loss and thus reduces the time consumption of the training process. Extensive experiments have been performed on the real-world large-scale datasets for vehicle re-id, and the results have demonstrated the effectiveness of the proposed methods.

As appearance based vehicle re-id has not been researched in depth, future explorations are still promising to further improve the performance. For example, we may use video sequences or consecutive frames rather than single images for vehicle re-id, then how to exploit the correlation between multiple frames is worth study. Moreover, training a camera adaptive model seems attractive to take into account the characteristics of different cameras, but also quite challenging due to the scarcity of training data.

5. REFERENCES

[1] Linjie Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang, “A large-scale car dataset for fine-grained categorization and verification,” in *CVPR*, 2015, pp. 3973–3981.

[2] Wei-Hua Lin and Daoqin Tong, “Vehicle re-identification with dynamic time windows for vehicle passage time estimation,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, no. 4, pp. 1057–1063, 2011.

[3] Karric Kwong, Robert Kavaler, Ram Rajagopal, and Pravin Varaiya, “Arterial travel time estimation based on vehicle re-identification using wireless magnetic sensors,” *Transportation Research Part C: Emerging Technologies*, vol. 17, no. 6, pp. 586–606, 2009.

[4] Michela Farenzena, Loris Bazzani, Alessandro Perina, Vittorio Murino, and Marco Cristani, “Person re-identification by symmetry-driven accumulation of local features,” in *CVPR*, 2010, pp. 2360–2367.

[5] Wei Li and Xiaogang Wang, “Locally aligned feature transforms across views,” in *CVPR*, 2013, pp. 3594–3601.

[6] Rui Zhao, Wanli Ouyang, and Xiaogang Wang, “Unsupervised salience learning for person re-identification,” in *CVPR*, 2013, pp. 3586–3593.

[7] Rui Zhao, Wanli Ouyang, and Xiaogang Wang, “Learning mid-level filters for person re-identification,” in *CVPR*, 2014, pp. 144–151.

[8] Ejaz Ahmed, Michael Jones, and Tim K Marks, “An improved deep learning architecture for person re-identification,” in *CVPR*, 2015, pp. 3908–3916.

[9] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang, “Deepreid: Deep filter pairing neural network for person re-identification,” in *CVPR*, 2014, pp. 152–159.

[10] Xinchen Liu, Wu Liu, Huadong Ma, and Huiyuan Fu, “Large-scale vehicle re-identification in urban surveillance videos,” in *ICME*, 2016, pp. 1–6.

[11] Xinchen Liu, Wu Liu, Tao Mei, and Huadong Ma, “A deep learning-based approach to progressive vehicle re-identification for urban surveillance,” in *ECCV*, 2016, pp. 869–884.

[12] Hongye Liu, Yonghong Tian, Yaowei Yang, Lu Pang, and Tiejun Huang, “Deep relative distance learning: Tell the difference between similar vehicles,” in *CVPR*, 2016, pp. 2167–2175.

[13] Florian Schroff, Dmitry Kalenichenko, and James Philbin, “Facenet: A unified embedding for face recognition and clustering,” in *CVPR*, 2015, pp. 815–823.

[14] Jiawei Liu, Zheng-Jun Zha, Qi Tian, Dong Liu, Ting Yao, Qiang Ling, and Tao Mei, “Multi-scale triplet cnn for person re-identification,” in *ACM MM*, 2016, pp. 192–196.

[15] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell, “Caffe: Convolutional architecture for fast feature embedding,” in *ACM MM*, 2014, pp. 675–678.

[16] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, “Return of the devil in the details: Delving deep into convolutional nets,” in *BMVC*, 2014.

[17] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al., “Imagenet large scale visual recognition challenge,” *IJCV*, vol. 115, no. 3, pp. 211–252, 2015.

[18] Douglas Gray, Shane Brennan, and Hai Tao, “Evaluating appearance models for recognition, reacquisition, and tracking,” in *Proc. IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS)*, 2007.