

Vehicle Re-identification Using Quadruple Directional Deep Learning Features

Jianqing Zhu, Huanqiang Zeng, Jingchang Huang, Shengcai Liao, Zhen Lei, Canhui Cai and LiXin Zheng

Abstract—In order to resist the adverse effect of viewpoint variations for improving vehicle re-identification performance, we design quadruple directional deep learning networks to extract quadruple directional deep learning features (QD-DLF) of vehicle images. The quadruple directional deep learning networks are with similar overall architecture, including the same basic deep learning architecture but different directional feature pooling layers. Specifically, the same basic deep learning architecture is a shortly and densely connected convolutional neural network to extract basic feature maps of an input square vehicle image in the first stage. Then, the quadruple directional deep learning networks utilize different directional pooling layers, i.e., horizontal average pooling (HAP) layer, vertical average pooling (VAP) layer, diagonal average pooling (DAP) layer and anti-diagonal average pooling (AAP) layer, to compress the basic feature maps into horizontal, vertical, diagonal and anti-diagonal directional feature maps, respectively. Finally, these directional feature maps are spatially normalized and concatenated together as a quadruple directional deep learning feature for vehicle re-identification. Extensive experiments on both VeRi and VehicleID databases show that the proposed QD-DLF approach outperforms multiple state-of-the-art vehicle re-identification methods.

I. INTRODUCTION

Vehicle re-identification with the goal of matching the same vehicle image captured by different cameras plays an important role in video surveillance for public security, since vehicle has been an indispensable part of human daily life [1]. In practical scenarios, vehicle re-identification is a very challenging computer vision problem, due to the fact that vehicle images usually contain a lot of adverse factors, such as

This work was supported in part by the National Natural Science Foundation of China under the Grants 61602191, 61401167, 61605048, 61473291 and 61372107, in part by the Natural Science Foundation of Fujian Province under the Grants 2018J01090 and 2016J01308, in part by High-level Talent Innovation Program of Quanzhou City under the Grant 2017G027, in part by the Promotion Program for Young and Middle aged Teacher in Science and Technology Research of Huaqiao University under the Grants ZQN-PY418 and ZQN-YX403, in part by the Scientific and Technology Funds of Xiamen under the Grant 3502Z20173045, and in part by the Scientific Research Funds of Huaqiao University under the Grant 16BS108, 14BS201 and 14BS204. (Corresponding author: Huanqiang Zeng).

Jianqing Zhu, Canhui Cai and LiXin Zheng are with the College of Engineering, Huaqiao University, Quanzhou, 362021, China and Fujian Provincial Academic Engineering Research Centre in Industrial Intellectual Techniques and Systems (e-mail: {jqzhu, chcai, zlx}@hqu.edu.cn).

Huanqiang Zeng is with the College of Information Science and Engineering, Huaqiao University, Xiamen, 361021, China (e-mail: zeng0043@hqu.edu.cn).

Jingchang Huang is with IBM-Research China, Building 10/6F, 399 Ke-Yuan Road, Zhangjiang Hi-Tech Park, Pudong New District, Shanghai 201203, China (e-mail: hjingc@cn.ibm.com).

Shengcai Liao and Zhen Lei are with the Center for Biometrics and Security Research and National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: {seliao, zlei}@cbsr.ia.ac.cn).

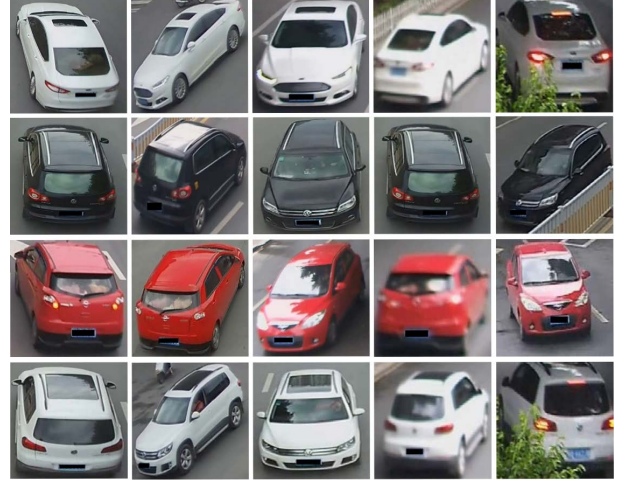


Fig. 1. Classical vehicle samples from the VeRi [1] database. Each row denotes the same vehicle captured by cameras from different viewpoints.

viewpoint variation, illumination change, blur, occlusion and low resolution, as shown in Fig. 1. Therefore, how to design an effective vehicle re-identification method has attracted more and more attentions.

To address the problem of vehicle re-identification, two large benchmark databases, VeRi [1, 2] and VehicleID [3], are released by the Institute of Digital Media, Peking University. Based on these two databases, multiple vehicle re-identification methods were developed and will be highlighted in the following Section II. Note that for vehicle re-identification, viewpoint variation is most crucial challenge and often encountered factor among those above-mentioned adverse ones, because vehicle images are usually captured under different camera viewpoints. Hence, this paper focuses on designing a method to resist adverse viewpoint variations so as to improve the vehicle re-identification performance. For that, we propose quadruple directional deep learning features (QD-DLF) to comprehensively describe vehicle images for improving the vehicle re-identification performance. The novelty and contribution of the proposed QD-DLF is that we make the first attempt to fuse quadruple directional deep features learned by using quadruple directional average pooling layers to improve the robustness of viewpoint variations. Experimental results have shown that the proposed approach is able to significantly improve the vehicle re-identification accuracy.

The rest of this paper is organized as follows. Section II introduces the related work. Section III describes the proposed quadruple directional deep learning features for vehicle re-identification. Section IV presents the experimental results to

validate the superiority of the proposed method. Section V concludes this paper.

II. RELATED WORKS

In this section, we briefly review the existing vehicle re-identification works. Since feature representation and similarity metric are two key roles in vehicle re-identification, the existing vehicle re-identification works mainly focus on two aspects: (1) feature representation for vehicle re-identification, and (2) similarity metric for vehicle re-identification.

A. Feature Representation for Vehicle Re-identification

Feature representation methods for vehicle re-identification can be mainly divided into two classes: hand crafted feature representations and deep learning features representations. For hand crafted feature representations, LOMO [4] and BOW-CN [5] features originally used in person re-identification are directly applied to vehicle re-identification. For deep learning feature representations, some well-known deep feature learning networks, such as AlexNet [6], VGGNet [7] and GoogLeNet [8], are used as feature extractors for vehicle re-identification. For example, FACT [2] uses AlexNet [6] to extract features of vehicles. NuFACT [1] takes GoogLeNet [8] as a feature extractor. DRDL [3] utilizes VGGNet [7] to extract features of vehicles. It can be further found that the deep learning features obviously outperforms hand crafted features on VeRi and VehicleID databases as reported in [1–3].

To effectively learn discriminative feature representations of vehicle images, multiple types of loss functions are applied to train deep learning based vehicle re-identification models. For example, the deep joint discriminative learning (DJDL) [9] method trains a convolutional neural network to extract discriminative feature representations of vehicle images using identification, verification and triplet loss functions simultaneously. The improved triplet convolutional neural network [10] using the classification-oriented loss function and the original triplet loss function is proposed for learning deep feature representations of vehicle images. Most of the above-mentioned deep learning features [6–10] are holistical, which are learned by a deep learning network ended up with several fully connection layers. Although these deep learning methods make great progresses of vehicle re-identification, they do not design a specific solution for processing viewpoint variations, which is a crucial challenge for vehicle re-identification.

For better handling viewpoint variations, the adversarial bi-directional long short-term memory (LSTM) network (ABLN) is proposed in [11]. ABLN uses LSTM to model transformations across continuous view variations of a vehicle and adopts the adversarial architecture to enhance training. Thus, a global vehicle representation containing all views' information can be inferred from only one visible view, and then used for learning to measure the distance between two vehicles with arbitrary views. Similar to ABLN, the spatially concatenated convolutional network (SCCN) and CNN-LSTM bi-directional loop (CLBL) are further proposed in [12] to address the challenging of viewpoint variations. However, all of ABLN [11], SCCN [12] and CNN-LSTM [12] require a

vehicle dataset, where each vehicle has images in densely sampled camera viewpoints. However, this is hard to acquire in practical video surveillance systems. Therefore, there is still ample room for vehicle re-identification by thoroughly considering viewpoint variations.

B. Similarity Metric for Vehicle Re-identification

For similarity metrics for vehicle re-identification, FACT [2] applies the Euclidean or Cosine distance between a vehicle pair described with deep learning features to measure the similarity, as done in many face recognition algorithms [13, 14]. Moreover, NuFACT [1] calculates the similarity of the query and gallery vehicle images measured by the Euclidean distance in the discriminative null space [15]. In addition, DRDL [3] proposes a deep relative distance learning method. This method exploits a two-branch convolutional neural network to transform raw vehicle images into an Euclidean space, where the distance can be directly used to measure the similarity of arbitrary two vehicles.

Besides, multi-modal vehicle re-identification methods are also proposed to improve vehicle similarity metrics. For example, the progressive and multi-modal vehicle re-identification (PROVID) [1] is presented to obtain a more accurate vehicle searching. The PROVID method first applies the NuFACT method to make a coarse searching. Then, it makes a fine searching based on a vehicle license plate verification model so that the re-identification accuracy is improved. In addition, the two-stage framework (i.e., siamese convolutional neural network (Siamese-CNN) and path long short-term memory (Path-LSTM) network) [16] incorporates complex spatial-temporal information for effectively regularizing the vehicle re-identification results. However, it is obvious that those multi-modal vehicle re-identification methods require the extra vehicle information (e.g., license plate, spatial-temporal information) and thus the additional computational load.

Based on the above analysis, vehicle re-identification method could improve the performance by resisting the viewpoint variation. Furthermore, from the viewpoint of practical applications, the vehicle re-identification method could not require extra vehicle information and introduce extra computational load. Motivated by these, the main objective of our proposed approach lies in: how to develop a more effective deep learning feature representation method for vehicle re-identification, without using extra vehicle information. For that, we propose quadruple directional deep learning features (QD-DLF) for vehicle re-identification in this paper. The most significant contribution of this paper is to fuse quadruple directional deep features learned by using quadruple directional average pooling layers, which constructs viewpoint robust features to significantly boost the vehicle re-identification performance. The details of the proposed QD-DLF will be described in the following Section.

III. VEHICLE RE-IDENTIFICATION USING QUADRUPLE DIRECTIONAL DEEP LEARNING FEATURES

A. Quadruple Deep Feature Learning Networks

As shown in Fig. 2, the proposed feature learning approach consists of quadruple directional (i.e., horizontal, vertical,

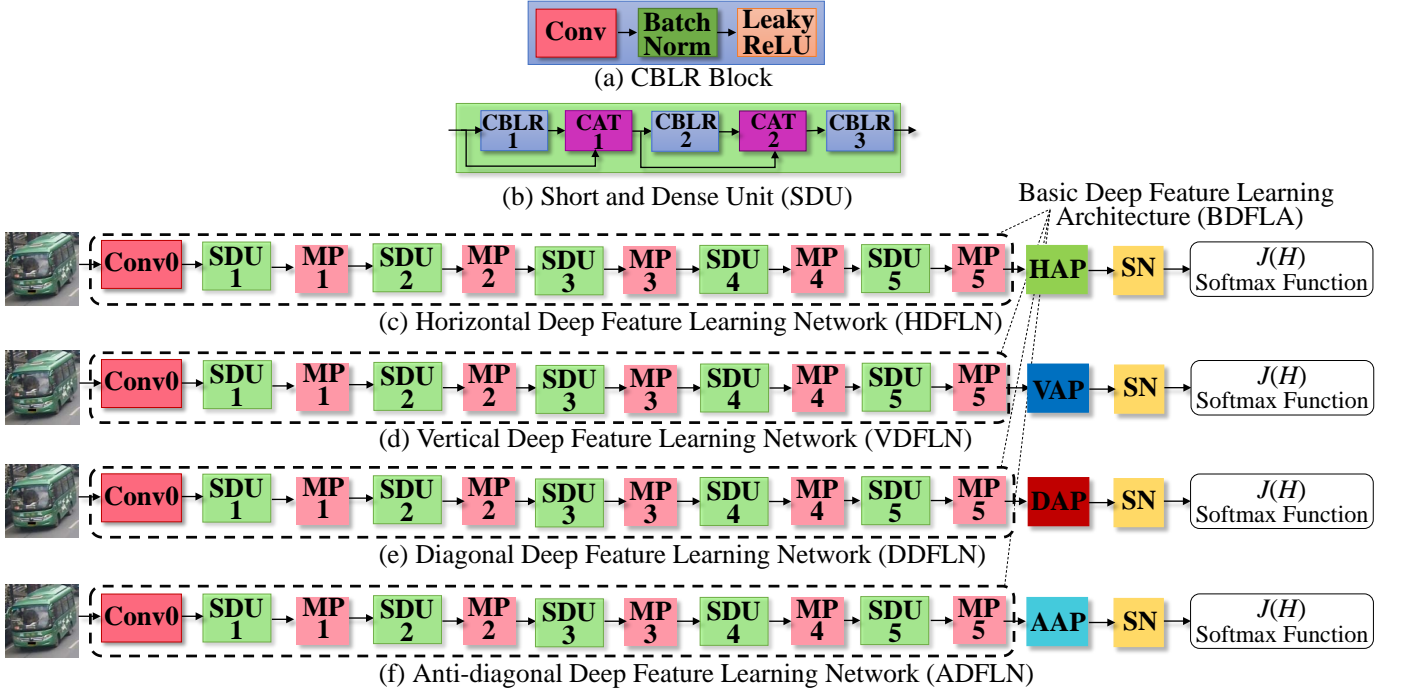


Fig. 2. The diagrams of the proposed quadruple directional deep feature learning networks. Here, MP, HAP, VAP, DAP, AAP and SN represents max pooling, horizontal average pooling, vertical average pooling, diagonal average pooling, anti-diagonal average pooling and spatial normalization layers, respectively.

diagonal and anti-diagonal) deep feature learning networks (i.e., HDFLN, VDFLN, DDFLN, ADFLN). More specifically, each directional deep feature learning network is composed of a common basic deep feature learning architecture (BDFLA), the corresponding directional average pooling layer and a spatial normalization (SN) layer [17].

1) **Basic Deep Feature Learning Architecture:** It can be seen from Fig. 2 that the basic deep feature learning architecture (BDFLA) is realized by a shortly and densely connected convolutional neural network, which is constructed by a list of short and dense units (SDUs) and max pooling layers. For convenient description, convolutional, batch normalization [18] and Leaky ReLU [19] layers are sequently packaged to construct a CBLR block, as shown in Fig. 2 (a). Then, three CBLR blocks are densely connected with two concatenation layers (i.e., CAT1 and CAT2) to build a short and dense unit (SDU), as shown in Fig. 2 (b). Each concatenation layer concatenates the input images according to the channel dimension. Finally, one convolutional layer (i.e., Conv0), a batch normalization, five SDUs (i.e., SDU1-SDU5), and five max pooling layers (i.e., MP1-MP5) are packaged in turn to obtain the basic deep feature learning architecture.

2) **Quadruple Directional Average Pooling Layers:** To comprehensively describe vehicle images from different directions, quadruple directional (i.e., horizontal, vertical, diagonal and anti-diagonal) average pooling layers are designed. Assume that the basic feature maps produced by the BDFLA is $X \in \mathbb{R}^{d \times d \times c}$, where d and c represent the height/width, and channel sizes, respectively. The developed quadruple directional average pooling layers can be described as follows.

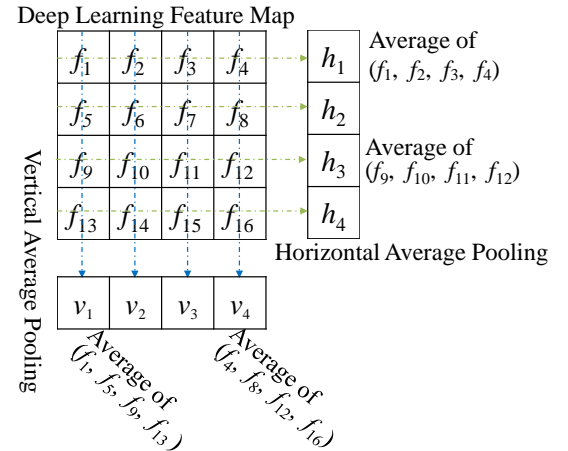


Fig. 3. The schematic diagram of horizontal and vertical average pooling operations.

- i) **Horizontal Average Pooling (HAP) Layer:** The HAP layer averages each row of X into a single point to obtain the horizontal average pooling feature map $P \in \mathbb{R}^{d \times 1 \times c}$, as shown in Fig. 3. For example, h_1 is equal to the average of f_1, f_2, f_3 and f_4 , that is, $h_1 = \frac{1}{4}(f_1 + f_2 + f_3 + f_4)$.
- ii) **Vertical Average Pooling (VAP) Layer:** The VAP layer averages each column of $X \in \mathbb{R}^{d \times d \times c}$ into a single point to obtain the vertical average pooling feature $Q \in \mathbb{R}^{1 \times d \times c}$, as shown in Fig. 3. Note that Q is transposed into $Q_t \in \mathbb{R}^{d \times 1 \times c}$ in a practical testing process to make the dimension of Q_t be compatible with that of P (i.e., the output of the HAP layer). For instance, v_4 is equal to the average of f_4, f_8, f_{12} and f_{16} , that is, $v_1 = \frac{1}{4}(f_4 + f_8 + f_{12} + f_{16})$, as show in Fig. 3.

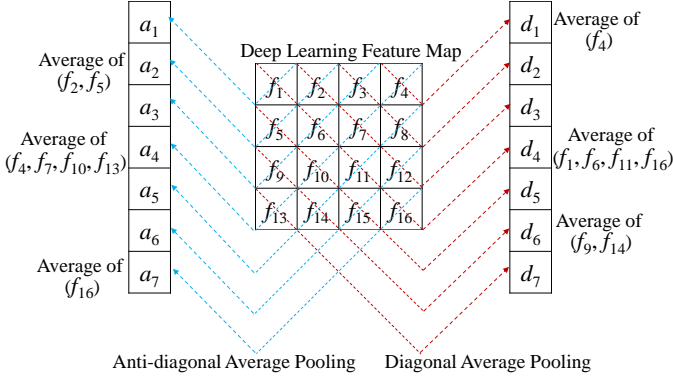


Fig. 4. The schematic diagram of diagonal and anti-diagonal average pooling operations.

- iii) **Diagonal Average Pooling (DAP) Layer:** The DAP layer averages multiple elements of the feature map $X \in \mathbb{R}^{d \times d \times c}$ according to the diagonal direction, as shown in Fig. 4. For example, d_6 is equal to the average of f_9 and f_{14} , that is, $d_6 = \frac{1}{2}(f_9 + f_{14})$.
- iv) **Anti-diagonal Average Pooling (AAP) Layer:** The AAP layer averages multiple elements of the feature map $X \in \mathbb{R}^{d \times d \times c}$ according to the anti-diagonal direction, as shown in Fig. 4. For instance, a_4 is equal to the average of f_4, f_7, f_{10} and f_{13} , that is, $a_4 = \frac{1}{4}(f_4 + f_7 + f_{10} + f_{13})$.

Since all the above-mentioned HAP, VAP, DAP and AAP layers use a average pooling operation, the forward and backward propagations are briefly introduced as follows. Assume that the input feature map of an average pooling layer is $X = [X_1, X_2, \dots, X_d] \in \mathbb{R}^{d \times d}$, the average pooling window size is $d \times 1$, and the output feature map is $Y \in \mathbb{R}^{1 \times d} = [y_1, y_2, \dots, y_d]$. Then, the forward propagation of this average pooling layer is calculated as follows:

$$y_i = \frac{1}{d} \sum_{j=1}^d X_{ij}, \quad (1)$$

where y_i is the i -th element of Y ; X_{ij} is j -th element of i -th column vector of X . According to the chain rule, the backward propagation of this average pooling layer can be calculated as follows:

$$\begin{aligned} \frac{\partial J}{\partial X_{i1}} &= \frac{\partial J}{\partial y_i} \frac{\partial y_i}{\partial X_{i1}} = \frac{1}{d} \frac{\partial J}{\partial y_i}, \\ \frac{\partial J}{\partial X_{i2}} &= \frac{\partial J}{\partial y_i} \frac{\partial y_i}{\partial X_{i2}} = \frac{1}{d} \frac{\partial J}{\partial y_i}, \\ &\dots \\ \frac{\partial J}{\partial X_{id}} &= \frac{\partial J}{\partial y_i} \frac{\partial y_i}{\partial X_{id}} = \frac{1}{d} \frac{\partial J}{\partial y_i}, \end{aligned} \quad (2)$$

and the matrix form can be formulated as follows:

$$\frac{\partial J}{\partial X_i} = \frac{1}{d} \left[\frac{\partial J}{\partial y_i}, \frac{\partial J}{\partial y_i}, \dots, \frac{\partial J}{\partial y_i} \right]^T, \quad (3)$$

where $X_i = [X_{i1}, X_{i2}, \dots, X_{id}]^T$ is i -th column vector of X ; J is the objective function (i.e., Eq. (6)) of the overall learning framework and will be discussed in the following subsection.

3) **Spatial Normalization Layer:** As shown in Fig. 2, a spatial normalization (SN) layer [17] is exploited to follow each directional average pooling layer. It is to make each dimension of the directional average pooling feature maps unified distributing in $[0, 1)$, which is beneficial to prevent a specific dimension whose value is too predominate. Assume that the input of a SN layer is $P \in \mathbb{R}^{d \times c}$. Then, the corresponding output Z of the SN layer can be calculated as follows:

$$Z_j^k = \frac{P_j^k}{\sqrt{1 + \sum_{l \in N_j} (P_l^k)^2}}, \quad (4)$$

where Z_j^k is j -th element of the k -th feature map of Z , P_j^k represents the j -th element of the k -th feature map of P , and N_j is the neighborhood size.

The backward propagation of the SN layer can be formulated as follows:

$$\frac{\partial J}{\partial P_j^k} = \frac{\frac{\partial J}{\partial Z_j^k} - Z_j^k \sum_{l \in N_j} \frac{\partial J}{\partial Z_l^k} Z_l^k}{\sqrt{1 + \sum_{l \in N_j} (P_l^k)^2}}. \quad (5)$$

Based on the above-mentioned basic deep feature learning network, quadruple directional average pooling layers, and SN layer, quadruple deep feature learning networks (QDFLNs) are constructed, consisting of horizontal deep feature learning network (HDFLN), vertical deep feature learning network (VDFLN), diagonal deep feature learning network (DDFLN), and anti-diagonal deep feature learning network (ADFLN), as shown in Fig. 2.

The recommended parameter configuration of the proposed QD-DLF method is listed in Table I. The channel numbers of Conv0, SDU1, SDU2, SDU3, SDU4 and SDU5 are 64, 64, 128, 192, 256 and 320, respectively. The scope of Leaky ReLU layer of SDU5 is 0, and that of the others is 0.15. The sub-window for Conv0 and SDU represents a filter size. Specifically, for pooling layers (i.e. MP1-MP5), it means a pooling window size, while for the spatial normalization (i.e. SN) layers, it denotes a normalization window size. Conv0 and five SDUs (i.e., SDU1-SDU5) apply 3×3 sized filters. Five max pooling layers use 3×3 sized pooling windows. Among the quadruple directional average pooling layers (i.e., HAP, VAP, DAP and AAP), the pooling window sizes for HAP and VAP layers are 1×4 and 4×1 , respectively, while for DAP and AAP the maximum number of pooled elements is 4, as show in Fig. 4. The normalization neighborhood size (i.e., N_j in Eq. (4)) in all spatial normalization layers is 4. Moreover, only those strides working on five MP layers are set as 2 pixels, and those on the others are set as 1 pixel.

B. Objective Function

Similar to [13, 20], the softmax function is utilized to build the objective function of the proposed method, as follows:

$$J(W) = \frac{1}{K} \left[\sum_{k=1}^K \sum_{c=1}^C \ell(y^{(k)} = c) \log \frac{e^{W_c^T X^{(k)}}}{\sum_{p=1}^C e^{W_p^T X^{(k)}}} \right] + \frac{1}{2} \alpha \|W\|_2^2, \quad (6)$$

where $W = [W_1, W_2, \dots, W_C] \in \mathbb{R}^{d \times C}$ is the projection matrix used to predicate a vehicle's class label, $X^{(k)}$ is the deep

TABLE I
THE PARAMETER CONFIGURATION OF THE PROPOSED QDFLNS.

Name	Channels	Scope of Leaky ReLU	Sub-window ($h \times w$)	Stride	Output Size
Conv0	64	0.15	3×3	1	$128 \times 128 \times 64$
SDU1	64	0.15	3×3	1	$128 \times 128 \times 64$
MP1	64	-	3×3	2	$64 \times 64 \times 64$
SDU2	128	0.15	3×3	1	$64 \times 64 \times 128$
MP2	128	-	3×3	2	$32 \times 32 \times 128$
SDU3	192	0.15	3×3	1	$32 \times 32 \times 192$
MP3	192	-	3×3	2	$16 \times 16 \times 192$
SDU4	256	0.15	3×3	1	$16 \times 16 \times 256$
MP4	256	-	3×3	2	$8 \times 8 \times 256$
SDU5	320	0	3×3	1	$8 \times 8 \times 320$
MP5	320	-	3×3	2	$4 \times 4 \times 320$
HAP	320	-	1×4	1	$4 \times 1 \times 320$
VAP	320	-	4×1	1	$1 \times 4 \times 320$
DAP	320	-	4	1	$7 \times 1 \times 320$
AAP	320	-	4	1	$7 \times 1 \times 320$

learning feature of k -th training sample, $y^{(k)} \in \{1, 2, 3, \dots, C\}$ is the corresponding class label, α is a constant used to control the contribution of the L_2 regularization item, K and C represent the numbers of the training samples and classes, respectively, and $\ell(\cdot)$ is an indicator function.

IV. EXPERIMENT AND ANALYSIS

To validate the superiority of the proposed quadruple directional deep learning feature (QD-DLF) approach, the performance comparison with multiple state-of-the-art methods is conducted on two challenging databases, namely, VeRi [1] and VehicleID [3]. In our experiments, the Euclidean distance is employed to measure the similarity of a vehicle pair described with quadruple deep learning features. And two commonly used criteria in the re-identification field, i.e., cumulative match curve (CMC) [21, 22] and mean average precision (MAP) [5, 23], are used to evaluate the performance. The CMC shows the identification accuracy rates that a query identity appears in different sized candidate lists. The MAP is used to evaluate the overall performance. For each query, the area under the precision-recall curve is calculated, which is known as average precision (AP). Then, the mean value of APs of all queries is calculated as MAP, which considers both precision and recall of a re-identification method, and thus provides a more comprehensive performance evaluation.

A. Training Configuration

In our experiments, the software tools are Matconvnet [17], CUDA 8.0, CUDNN V5.1, MATLAB 2014 and Visual Studio 2012. The hardware device is workstation configured with a Intel Xeon E3-1505 M v5 CPU @2.80 GHz, a NVIDIA Titan X GPU and 128 GB DDR3 Memory. Moreover, the training settings similar to [22, 24] are adopted and summarized as

follows. All images in these two databases are scaled to 128×128 pixels, and each image is further augmented by the horizontal mirror and randomly rotating operations. The randomly rotating operation is applied to randomly rotate an image in ranges $[-3^\circ, 0^\circ]$ and $[0^\circ, 3^\circ]$. The weights in each layer are initialized based on a normal distribution $N(0, 0.01)$, and the biases are initialized to 0. The L_2 regularization weights α in Eq. (6) is set as 0.005 on the VeRi [1] database, while for the larger VehicleID [3] database, it is set as 0.001. The size of mini-batch is 128 including 64 positive and 64 negative image pairs, and both positive and negative pairs are randomly selected from the whole database. The momentums are set to 0.9. The learning rates start with 0.01 and are gradually decreased along the training progress. That is, if the objective function is convergent at a stage, the learning rates are reduced to 1/10 of the current values, and the minimum learning rates are 0.001.

B. Databases

VeRi [1] is captured by 20 cameras in unconstrained traffic scenarios and each vehicle is captured by 2-18 cameras under different viewpoints, illuminations, occlusions and resolutions. The VeRi dataset is divided into a training subset containing 37,781 images of 576 subjects and a testing subset with 13,257 images of 200 subjects. For the evaluation, one image of each vehicle captured from each camera is applied as a query, then a query set containing 1,678 images of 200 subjects and a gallery including 11,579 image of 200 subjects are finally obtained. Furthermore, only the cross-camera vehicle re-identification is evaluated, which means that if a probe image and a gallery image are captured under the same camera viewpoint, the corresponding matching result will be excluded in the final performance evaluation.

VehicleID [3] is captured in daytime by multiple real-world surveillance cameras distributed in a small city of China. There are 221,763 images of 26,267 subjects in the entire database. Each vehicle is captured from either a front viewpoint or a back viewpoint. The training subset consists of 110,178 images of 13,134 subjects. In addition, VehicleID provides three testing subsets, Test800, Test1600 and Test2400, for evaluating the performance in different data scales. Specifically, Test800 includes 800 gallery images and 6,532 probe images of 800 subjects. Test1600 contains 1600 gallery images and 11,395 probe images of 1,600 subjects. Test2400 consists 2400 gallery images and 17,638 probe images of 2,400 subjects.

C. Performance Evaluation

1) *Comparison on VeRi*: The performance comparison of the proposed QD-DLF and multiple state-of-art methods on the VeRi database is shown in Table II. It can be found that the proposed QD-DLF acquires the highest MAP (i.e., 61.83%) and rank-1 identification rate (i.e., 88.50%) among all methods under comparisons. More details are analyzed as follows.

Firstly, compared with four multi-modal deep learning based vehicle re-identification methods (i.e., NuFACT + Plate-SNN [1], NuFACT + Plate-REC [1], PROVID [1] and Siamese-CNN+Path-LSTM [16]), the proposed QD-DLF method consistently defeats NuFACT + Plate-SNN [1], NuFACT +

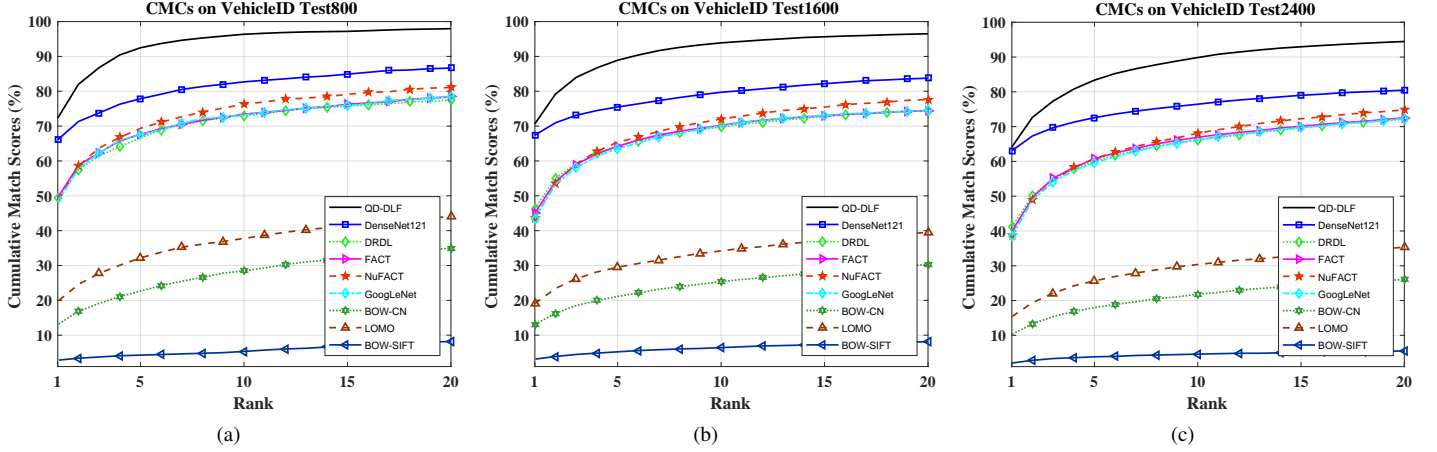


Fig. 5. The CMC curve comparisons of the proposed QD-DLF method and multiple state-of-the-art methods on (a) Test800, (b) Test1600, and (c) Test2400 of VehicleID, respectively.

TABLE II
THE PERFORMANCE (%) COMPARISON OF THE PROPOSED QD-DLF AND MULTIPLE STATE-OF-THE-ART METHODS ON VERI.

Methods	MAP	Rank=1	Rank=5
Proposed QD-DLF	61.83	88.50	94.46
Siamese-CNN+Path-LSTM [16]	58.27	83.49	90.04
PROVID [1]	53.42	81.56	95.11
NuFACT + Plate-SNN [1]	50.87	81.11	92.79
NuFACT + Plate-REC [1]	48.55	76.88	91.42
NuFACT [1]	48.47	76.76	91.42
DenseNet121 [25]	45.06	80.27	91.12
SCCN-Ft+CLBL-8-Ft [12]	25.12	60.83	78.55
ABLN-Ft-16 [11]	24.92	60.49	77.33
FACT [2]	18.75	52.21	72.88
GoogLeNet [26]	17.89	52.32	72.17
VGG-CNN-M-1024 [3]	12.76	44.10	62.63
BOW-CN [5]	12.20	33.91	53.69
LOMO [4]	9.64	25.33	46.48
BOW-SIFT [27]	1.51	1.91	4.53

Plate-REC [1] and Siamese-CNN+Path-LSTM [16] by higher MAPs, rank-1 identification rates and rank-5 identification rates. Although the rank-5 identification rate of the proposed QD-DLF method is a bit lower than that of PROVID [1], the proposed QD-DLF method achieves much higher MAP and rank-1 identification rate, and thus is still superior to PROVID [1].

Secondly, compared with those single modal deep learning based vehicle re-identification methods (i.e., NuFACT [1], DenseNet121 [25], SCCN-Ft+CLBL-8-Ft [12], ABLN-Ft-16 [11], FACT [2], GoogLeNet [26] and VGG-CNN-M-1024 [3]), the proposed QD-DLF methods shows a larger accuracy improvement. Specifically, the best single modal deep learning based vehicle re-identification method, i.e., NuFACT [1], only obtains a 48.47% MAP, a 76.76% rank-1 identification rate and a 91.42% rank-5 identification rate, which are much lower than those of the proposed QD-DLF method. Moreover, it can

be seen that SCCN-Ft+CLBL-8-Ft [12] and ABLN-Ft-16 [11] do not obviously show the superiority on the VeRi database, although they specially consider the viewpoint variation. This is because each vehicle in the VeRi database is not densely captured by different camera viewpoints, which does not fully meet the requirement of the training data for CCN-Ft+CLBL-8-Ft [12] and ABLN-Ft-16 [11], limiting the performance improvement.

Thirdly, the proposed QD-DLF also obtains much better performance than those hand-crafted feature representation methods, i.e., BOW-CN [5], LOMO [4] and BOW-SIFT [27].

2) *Comparison on VehicleID*: The performance comparisons of the proposed QD-DLF and multiple state-of-art methods on VehicleID database are shown in Fig. 5 and Table III. Firstly, it can be observed that deep learning based methods (i.e., DJDL [9], DenseNet121 [25], Improved Triplet CNN [10], DRDL [3], FACT [2], NuFACT [1] and GoogLeNet [26]) obviously defeat traditional methods (i.e., LOMO [4], BOW-CN [5] and BOW-SIFT [27]) on this large scale database. Secondly, the proposed QD-DLF method outperforms all deep learning based methods under comparison, including DJDL [9], DenseNet121 [25], Improved Triplet CNN [10], DRDL [3], FACT [2], NuFACT [1] and GoogLeNet [26], on the Test800, Test1600 and Test2400 subsets of the VehicleID database.

3) *Comparison on Each Directional Deep Learning Feature*: In addition, we further comprehensively analyze the performance resulted from each directional deep learning feature. The corresponding results on VeRi and VehicleID databases are shown in Fig. 6, and Table IV, where horizontal, vertical, diagonal and anti-diagonal deep learning features are denoted as H-DLF, V-DLF, D-DLF and A-DLF, respectively. The analyses are described in detailed as follows.

i) Which directional deep learning feature is better for vehicle re-identification?

From Fig. 6, it can be found that the D-DLF/A-DLF defeats H-DLF/V-DLF. It means that the diagonal/anti-diagonal directional deep learning feature is more appropriate than the horizontal/vertical directional deep learning feature for vehicle re-identification. This may be

TABLE III
THE PERFORMANCE (%) COMPARISON OF THE PROPOSED QD-DLF AND MULTIPLE STATE-OF-THE-ART METHODS ON VEHICLEID.

Method	Test800			Test1600			Test2400			Average		
	MAP	Rank=1	Rank=5	MAP	Rank=1	Rank=5	MAP	Rank=1	Rank=5	MAP	Rank=1	Rank=5
Proposed QD-DLF	76.54	72.32	92.48	74.63	70.66	88.90	68.41	64.14	83.37	73.19	69.04	88.25
DJDL [9]	N/A	72.3	85.7	N/A	70.8	81.8	N/A	68.0	78.9	N/A	70.4	82.1
DenseNet121 [25]	68.85	66.10	77.87	69.45	67.39	75.49	65.37	63.07	72.57	67.89	65.52	75.31
Improved Triplet CNN [10]	N/A	69.9	87.3	N/A	66.2	82.3	N/A	63.2	79.4	N/A	66.4	83.0
DRDL [3]	N/A	48.91	66.71	N/A	46.36	64.38	N/A	40.97	60.02	N/A	45.41	63.70
FACT [2]	N/A	49.53	67.96	N/A	44.63	64.19	N/A	39.91	60.49	N/A	44.69	64.21
NuFACT [1]	N/A	48.90	69.51	N/A	43.64	65.34	N/A	38.63	60.72	N/A	43.72	65.19
GoogLeNet [26]	N/A	47.90	67.43	N/A	43.45	63.53	N/A	38.24	59.51	N/A	43.20	60.04
LOMO [4]	N/A	19.74	32.14	N/A	18.95	29.46	N/A	15.26	25.63	N/A	17.98	3.76
BOW-CN [5]	N/A	13.14	22.69	N/A	12.94	21.09	N/A	10.20	17.89	N/A	12.09	20.56
BOW-SIFT [27]	N/A	2.81	4.23	N/A	3.11	5.22	N/A	2.11	3.76	N/A	2.68	3.76

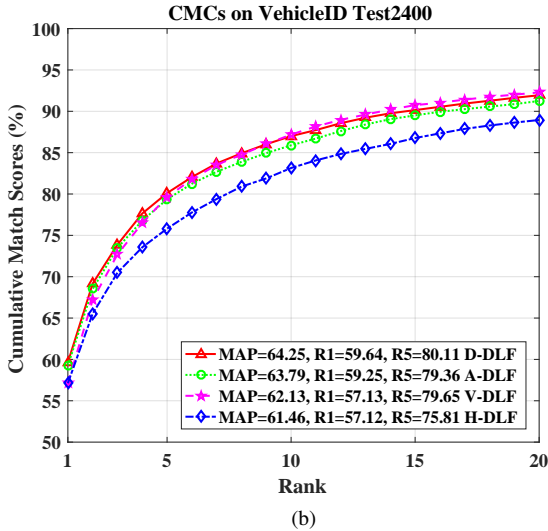
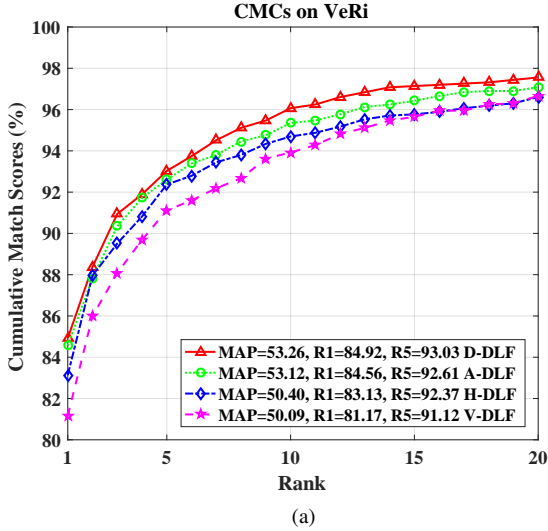


Fig. 6. The performance (%) comparison of diagonal deep learning feature (D-DLF), anti-diagonal deep learning feature (A-DLF), horizontal deep learning feature (H-DLF) and vertical deep learning feature (V-DLF) on (a) VeRi, (b) Test2400 of VehicleID, respectively.

TABLE IV
THE PERFORMANCE (%) COMPARISON OF QD-DLF, DAH-DLF, DA-DLF AND D-DLF ON VeRi AND TEST2400 OF VEHICLEID.

Methods	VeRi			Test2400 of VehicleID		
	MAP	Rank=1	Rank=5	MAP	Rank=1	Rank=5
QD-DLF	61.83	88.50	94.46	68.41	64.14	83.37
DAH-DLF	60.31	88.62	94.34	68.24	63.88	83.72
DA-DLF	58.16	87.19	94.46	66.90	62.53	82.07
D-DLF	53.26	84.92	93.03	64.25	59.64	80.11

attributed to two aspects. Firstly, although vehicle images are captured from different camera viewpoints, most vehicle images still have a prominent symmetry. Secondly, for the vehicle images with symmetry, the average pooling corresponding to the diagonal or anti-diagonal direction is able to involve more diverse local feature regions to construct more effective features so as to obtain a better performance of vehicle re-identification.

ii) Is there contribution from the fusion of various directional deep learning features?

By starting from the D-DLF as an example, we gradually fuse more directional deep learning features to see their performance. Here, DA-DLF means the fusion of D-DLF and A-DLF, DAH-DLF means the fusion of D-DLF, A-DLF and H-DLF. Note that the proposed QD-DLF means the fusion of all four directional deep learning features. From Table IV, it can be observed that QD-DLF fusing quadruple directional deep learning features obtains the best performance, DAH-DLF defeats DA-DLF and D-DLF, and DA-DLF outperforms D-DLF. It can be found that the more directional deep learning features are, the better the performance is. This study shows that the fusion of various directional deep learning features can indeed contribute to the improvement of the performance.

4) *Comparison on Directional and Deep Holistic Learning Features:* In this work, we also compare the performance resulted from the proposed method with directional deep learn-

TABLE V
THE PERFORMANCE (%) COMPARISON OF QD-DLF, D-DLF, A-DLF, H-DLF, V-DLF, AND F-DLFS ON VERI.

Methods	MAP	Rank=1	Rank=5
QD-DLF	61.83	88.50	94.46
D-DLF	53.26	84.92	93.03
A-DLF	53.12	84.56	92.61
H-DLF	50.40	83.13	92.37
V-DLF	50.09	81.17	91.12
F-DLF-256	40.99	80.15	90.64
F-DLF-512	39.68	80.21	89.87
F-DLF-128	39.39	79.68	91.12
F-DLF-1024	39.10	79.86	89.33

ing feature and a deep holistic feature learning configuration. This deep holistic feature learning configuration is built by keeping the basic deep feature learning architecture (BDLFA) unchanged and using a **Full** connection layer instead of one directional pooling layer. The corresponding learned feature is denoted as F-DLF. Since the F-DLF is able to produce holistical features with different dimensions and we do not know the suitable feature dimension in advance, the performances resulted from the holistical features with different dimensions produced by the full connection layer are all obtained. The F-DLF-128, F-DLF-256, F-DLF-512 and F-DLF-1024 represents the corresponding learning configuration that produces 128, 256, 512 and 1024 dimensional holistical features, respectively.

From the results shown in Table V, one can find that among F-DLF-128, F-DLF-256, F-DLF-512 and F-DLF-1024, F-DLF-256 obtains the highest MAP, but is inferior to our weakest directional deep learning feature, V-DLF. Specifically, the MAP of V-DLF is 9.10% higher than that of F-DLF-256 and the rank-1 identification rate of V-DLF is 1.02% higher than that of F-DLF-256, respectively. Moreover, the proposed QD-DLF is significant better than F-DLF-256 (i.e., 20.84% higher MAP and 8.35% higher rank-1 identification rate). This study indicates that the proposed directional deep learning features are more suitable and effective than deep holistic learning features to represent vehicle images.

5) *Running Time Analysis*: In addition to the accuracy, the efficiency is also very important for vehicle re-identification methods. For that, the running time comparison is conducted in terms of the feature extraction time (FET) per image as suggested in the similar person re-identification task [22]. The running time comparison of the proposed QD-DLF method and several state-of-the-art vehicle re-identification methods is shown in Table VI, where all the methods are implemented in the GPU mode.

Firstly, one can see that the running time of each proposed directional deep learning feature (i.e., D-DLF, A-DLF, H-DLF, V-DLF) is similar to that of each other. And the FETs of each directional deep learning features are a bit slower than that of VGG-CNN-M-1024 [3] and are comparable to that of GoogLeNet [26]. Moreover, compared to the ultra-deep model DenseNet121 [25], the FET of each directional deep

TABLE VI
THE RUNNING TIME COMPARISON OF THE PROPOSED QD-DLF METHOD AND MULTIPLE STATE-OF-THE-ART VEHICLE RE-IDENTIFICATION METHODS. FET REPRESENTS THE FEATURE EXTRACTION TIME.

Methods	FET (msec/image)
D-DLF	2.321
A-DLF	2.304
H-DLF	2.296
V-DLF	2.312
QD-DLF	11.199
DenseNet121 [25]	13.647
GoogLeNet [26]	2.345
VGG-CNN-M-1024 [3]	1.872

learning feature is about 17% of that of DenseNet121 [25], which demonstrates all of D-DLF, A-DLF, H-DLF, V-DLF are much faster than DenseNet121 [25].

Secondly, the running time of the proposed QD-DLF is analyzed. Intuitively, the FET of QD-DLF would be 4 times of that of each proposed single directional deep learning feature. However, from Table VI, one can find that the FET of QD-DLF is about 5 times of that of each proposed directional deep learning feature. This is due to that the efficiency of simultaneously making the communalization between CPU and GPU for quadruple directional deep learning model is lower than that for each proposed directional deep learning model, since quadruple directional deep learning model is naturally larger than each single directional deep learning model. In addition, it should be pointed out that the FET of proposed QD-DLF is 2.448ms faster than that of DenseNet121 [25].

V. CONCLUSION

In this paper, quadruple directional deep learning networks are designed for vehicle re-identification. The quadruple directional deep learning networks are with the same basic deep learning architecture but different directional feature pooling layers. The same basic deep learning architecture is a shortly and densely connected convolutional neural network to extract basic feature maps of an input square vehicle image. After that, a horizontal average pooling (HAP) layer, a vertical average pooling (VAP) layer, a diagonal average pooling (DAP) layer and an anti-diagonal average pooling (AAP) layer are repetitively applied in the proposed quadruple directional deep learning network to compress the basic feature maps into horizontal, vertical, diagonal and anti-diagonal directional feature maps. Finally, these obtained directional feature maps are spatially normalized and concatenated together as a quadruple directional deep learning feature for vehicle re-identification. Through quadruple directional deep learning features learned by the proposed quadruple directional deep learning network, the adverse effect of viewpoint variations is effectively resisted and the performance of vehicle re-identification is thus significantly improved. Extensive experiments on both VeRi and VehicleID databases show that the proposed method is obviously superior to multiple state-of-the-art vehicle re-identification methods.

REFERENCES

- [1] X. Liu, W. Liu, T. Mei, and H. Ma, "Provid: Progressive and multi-modal vehicle re-identification for large-scale urban surveillance," *IEEE Transactions on Multimedia*, vol. 20, no. 3, pp. 645–658, 2018.
- [2] X. Liu, W. Liu, H. Ma, and H. Fu, "Large-scale vehicle re-identification in urban surveillance videos," in *IEEE International Conference on Multimedia and Expo*, 2016, pp. 1–6.
- [3] H. Liu, Y. Tian, Y. Wang, L. Pang, and T. Huang, "Deep relative distance learning: Tell the difference between similar vehicles," in *IEEE International Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2167–2175.
- [4] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *IEEE International Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2197–2206.
- [5] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *IEEE International Conference on Computer Vision*, 2015, pp. 1116–1124.
- [6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Annual Conference on Neural Information Processing Systems*, 2012, pp. 1097–1105.
- [7] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
- [8] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *IEEE International Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.
- [9] Y. Li, Y. Li, H. Yan, and J. Liu, "Deep joint discriminative learning for vehicle re-identification and retrieval," in *IEEE International Conference on Image Processing*, 2017, pp. 395–399.
- [10] Y. Zhang, D. Liu, and Z.-J. Zha, "Improving triplet-wise training of convolutional neural network for vehicle re-identification," in *IEEE International Conference on Multimedia and Expo*, 2017, pp. 1386–1391.
- [11] Y. Zhou and L. Shao, "Vehicle re-identification by adversarial bi-directional lstm network," in *2018 IEEE Winter Conference on Applications of Computer Vision*. IEEE, 2018, pp. 653–662.
- [12] Y. Zhou, L. Liu, and L. Shao, "Vehicle re-identification by deep hidden multi-view inference," *IEEE Transactions on Image Processing*, vol. 27, no. 7, pp. 3275–3287, 2018.
- [13] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *IEEE International Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1891–1898.
- [14] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," in *Annual Conference on Neural Information Processing Systems*, 2014, pp. 1988–1996.
- [15] L. Zhang, T. Xiang, and S. Gong, "Learning a discriminative null space for person re-identification," in *IEEE International Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1239 – 1248.
- [16] Y. Shen, T. Xiao, H. Li, S. Yi, and X. Wang, "Learning deep neural networks for vehicle re-id with visual-spatio-temporal path proposals," in *IEEE International Conference on Computer Vision*, 2017, pp. 1918–1927.
- [17] A. Vedaldi and K. Lenc, "Matconvnet: Convolutional neural networks for matlab," in *ACM International Conference on Multimedia*, 2015, pp. 689–692.
- [18] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning*, 2015, pp. 448–456.
- [19] B. Xu, N. Wang, T. Chen, and M. Li, "Empirical evaluation of rectified activations in convolutional network," *arXiv preprint arXiv:1505.00853*, 2015.
- [20] L. Cai, J. Zhu, H. Zeng, J. Chen, C. Cai, and K.-K. Ma, "Hog-assisted deep feature learning for pedestrian gender recognition," *Journal of the Franklin Institute*, vol. 355, no. 4, pp. 1991–2008, 2018.
- [21] D. Gray, S. Brennan, and H. Tao, "Evaluating appearance models for recognition, reacquisition, and tracking," in *IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*, 2007.
- [22] J. Zhu, H. Zeng, S. Liao, Z. Lei, C. Cai, and L. Zheng, "Deep hybrid similarity learning for person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. PP, no. 99, pp. 1–1, 2017.
- [23] Z. Zheng, L. Zheng, and Y. Yang, "A discriminatively learned CNN embedding for person reidentification," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 14, no. 1, p. 13, 2017.
- [24] J. Zhu, H. Zeng, Y. Du, Z. Lei, L. Zheng, and C. Cai, "Person re-identification based on novel triplet convolutional neural network," *Journal of Electronics & Information Technology*, vol. 40, no. 4, pp. 1012–1016, 2018.
- [25] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *IEEE International Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2261–2269.
- [26] L. Yang, P. Luo, C. L. Chen, and X. Tang, "A large-scale car dataset for fine-grained categorization and verification," in *IEEE International Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3973–3981.
- [27] L. Zheng, S. Wang, W. Zhou, and Q. Tian, "Bayes merging of multiple vocabularies for scalable image retrieval," in *IEEE International Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1963–1970.



Jianqing Zhu received the B.S. degree in communication engineering and the M.S. degree in communication and information system from the School of Information Science and Engineering, Huaqiao University, Xiamen, China, in 2009 and 2012, respectively. He received the Ph.D. degree in Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2015. He is currently an Assistant Professor at the College of Engineering, Huaqiao University, Quanzhou, China. His current research interests include computer vision and pattern

recognition, with a focus on image and video analysis, particularly person re-identification, object detection and video surveillance. He was awarded the Best Biometrics Student Paper award at the International Conference on Biometrics in 2015.



Huanqiang Zeng (S'10-M'13-SM'18) received the B.S. and M.S. degrees from Huaqiao University, Xiamen, China and the Ph.D. degree from Nanyang Technological University, Singapore, all in electrical engineering. He was a Postdoctoral Fellow at the Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong from 2012 to 2013, and a Research Associate at the Temasek Laboratories, Nanyang Technological University, Singapore in 2008. He is now a Professor at the School of Information Science and Engineering, Huaqiao

University, Xiamen, China. His research interests are in the areas of image processing and video coding, machine learning and pattern recognition, and computer vision. He has published more than 80 papers in well-known international journals and conferences. He has been actively serving as the Associate Editor for IET Electronics Letters and International Journal of Image and Graphics, Guest Editor for multiple international journals, including Journal of Visual Communication and Image Representation, Multimedia Tools and Applications, Journal of Ambient Intelligence and Humanized Computing, the General Co-Chair for IEEE International Symposium on Intelligent Signal Processing and Communication Systems 2017 (ISPACS2017), the Technical Program Co-Chair for Asia-Pacific Signal and Information Processing Association Annual Summit and Conference 2017 (APSIPA ASC2017), the Area Chair for IEEE International Conference on Visual Communications and Image Processing (VCIP2015), the Technical Program Committee Member for multiple flagship international conferences. He received the Best Paper Award from Chinese Conference on Signal Processing 2017 (CCSP2017). He is an IEEE Senior member, and a Member of International Steering Committee of International Symposium on Intelligent Signal Processing and Communication Systems.



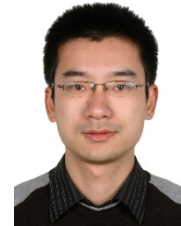
Jingchang Huang received the Ph.D. degree in Communication and Information Systems from University of Chinese Academic and Sciences, China in 2015. He joined IBM Research - China in 2015 and currently he is a research staff member in Cognitive-IoT team. He mainly focuses on the cognitive data analytics, including but not limited to acoustic, image and time sequences, all of which comes from Internet-of-Thing related applications. His current research interests also include patterns recognition, big data and wireless sensor networks.

Dr. Huang has designed and implemented many efficient target detection and classification projects, during which contributes more than 10+ scientific paper in IEEE journals & Con. and 20+ patent inventions.



Shengcai Liao received the B.S. degree in mathematics and applied mathematics from the Sun Yatsen University, Guangzhou, China, in 2005 and the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2010. He was a Post Doctoral Fellow in the Department of Computer Science and Engineering, Michigan State University during 2010-2012. He is currently an Associate Professor in the Institute of Automation, Chinese Academy of Sciences. His research interests include computer vision, pattern recognition, with

a focus on image and video analysis, particularly face recognition, object detection, person re-identification, metric learning and video surveillance. He was awarded the Motorola Best Student Paper award and the 1st Place Best Biometrics Paper award at the International Conference on Biometrics in 2006 and 2007, respectively, for his work on face recognition. He was also awarded the best reviewer award in IJCB 2014.



Zhen Lei received the B.S. degree in automation from the University of Science and Technology of China, in 2005, and the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, in 2010, where he is currently an Associate Professor. He has published over 100 papers in international journals and conferences. His research interests are in computer vision, pattern recognition, image processing, and face recognition in particular. He served as an area chair of the International Joint Conference on biometrics in 2014, the IAPR/IEEE

International Conference on Biometric in 2015, 2016 and 2018, and the IEEE International Conference on Automatic Face and Gesture Recognition in 2015.



Canhui Cai received the B.S. degree from Xidian University, Xian, China in 1982, M.S. degree from Shanghai University, Shanghai, China in 1985, and Ph.D. degree from Tianjin University, Tianjin, China in 2003, all in electronic engineering. Since 1984, he has been with the faculty of Huaqiao University, Quanzhou, China, and is currently a Professor at the College of Engineering. He was a Visiting Professor at Delft University of Technology, Delft, the Netherlands from 1991 to 1992, and a Visiting Professor at University of California at Santa Barbara, Santa

Barbara, CA from 1999 to 2000. His research areas include video communications, image and video signal processing, and computer vision. He is the author or co-author of 4 books, and has published over 150 papers in journals and conference proceedings. Dr. Cai was a General Co-Chair of Intelligent Signal Processing and Communication Systems in 2007..



Lixin Zheng received the B.S. degree in electronic technology application and the M.S. degree in machinery manufacturing from the Huaqiao University, in 1987 and 1990, respectively. He received the Ph.D. degree in measurement technology and instrument from Tianjin University in 2002. He is a Professor and the Dean of the College of Engineering, Huaqiao University, Quanzhou, China. His current research interests include motion control, power supply technology, computer vision and pattern recognition.