SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

November 2020, Vol. 63 212102:1–212102:12
https://doi.org/10.1007/s11432-019-2811-8

Discriminative fine-grained network for vehicle re-identification using two-stage re-ranking

Qi WANG[1], Weidong MIN[2,3*], Daojing HE[4], Song ZOU[1], Tiemei HUANG[1], Yu ZHANG[1] & Ruikang LIU[1]

[1]*School of Information Engineering, Nanchang University, Nanchang* 330031*, China;*
[2]*School of Software, Nanchang University, Nanchang* 330047*, China;*
[3]*Jiangxi Key Laboratory of Smart City, Nanchang University, Nanchang* 330047*, China;*
[4]*School of Computer Science and Software Engineering, East China Normal University, Shanghai* 200062*, China*

Received 26 August 2019/Revised 15 November 2019/Accepted 31 January 2020/Published online 13 October 2020

Abstract Research on the application of vehicle re-identification to video surveillance has attracted increasingly growing attention. Existing methods are associated with the difficulties of distinguishing different instances of the same car model owing to the incapability of recognizing subtle differences among these instances and the possibility that a subtle difference may lead to incorrect results of ranking. In this paper, a discriminative fine-grained network for vehicle re-identification based on a two-stage re-ranking framework is proposed to address these issues. This discriminative fine-grained network (DFN) is composed of fine-grained and Siamese networks. The proposed hybrid network can extract discriminative features of the vehicle instances with subtle differences. The Siamese network is rather suitable for general object re-identification using two streams of the network, while the fine-grained network is capable of detecting subtle differences. The proposed two-stage re-ranking method allows obtaining a more reliable ranking list by using the Jaccard metric and merging the first and second re-ranking lists, where the latter list is formed using the sample mean feature. Experimental results on the VeRi-776 and VehicleID datasets show that the proposed method achieves the superior performance compared to the state-of-the-art methods used in vehicle re-identification.

Keywords vehicle re-identification, DFN, two-stage re-ranking, fine-grained, Jaccard metric

Citation Wang Q, Min W D, He D J, et al. Discriminative fine-grained network for vehicle re-identification using two-stage re-ranking. Sci China Inf Sci, 2020, 63(11): 212102, https://doi.org/10.1007/s11432-019-2811-8

1 Introduction

Re-identification is aimed to identify the same target within the different shooting scenes and periods, which is an important field of computer vision, and vehicle re-identification is one of major topics. A straightforward application is to distinguish whether a vehicle corresponds to the same car model by identifying the license plate [1–3]. Vehicle re-identification can be executed successfully if license plate characters can be accurately registered. However, the analysis of surveillance videos is still associated with the issues owing to license plate recognition loss, various viewpoints, and blurred image resolution. Illumination may vary owing to changes in the angle of view and a nature of the camera. Surveillance videos from different cameras also make the task of vehicle re-identification challenging.

Vehicle re-identification is deemed more difficult than person re-identification [4–7] as the vehicles belonging to the same model can only be distinguished by subtle differences. Several previous approaches [8,9] have focused on the appearance attributes of vehicles, such as the color, shape, and model.

* Corresponding author (email: minweidong@ncu.edu.cn)

However, different vehicle identifications (IDs) may correspond to the same model in particular cases, and only subtle distinctions may exist among different vehicles registered by the same camera. The subtle inter-instance differences between different vehicle images and large intra-instance differences between the same vehicle images hinder the improvement of vehicle re-identification performance. Distinguishing such vehicles on the basis of simple appearance attributes is difficult, resulting in larger intra-instance differences compared with the inter-instance ones.

The spatio-temporal relationship has been frequently considered in object association [10]. Several approaches [11, 12] combine the space-time and location data to estimate the relationship between every pair of vehicle images to improve the re-identification results. However, the lack of relevant datasets representing the spatio-temporal information is a key problem, which also incurs additional computational costs. Fine-grained classification [13] is also closely related to the problem of re-identification. However, it should be noted that minor visual differences can affect the accuracy of a ranking list.

In this study, we aim to distinguish different vehicle IDs having the similar appearance accurately and to obtain the improved quality ranks. Therefore, hybrid architecture is proposed to address the considered vehicle re-identification issues. The contributions of this paper are summarized as follows.

(1) A discriminative fine-grained network (DFN) composed of the two main parts is proposed. The Siamese network is introduced to enhance the expression of features in the first part. Then, the fine-grained network is utilized to obtain a strong discriminative feature in the second part. The proposed network can accurately distinguish subtle differences in different vehicle IDs with similar appearances.

(2) Two-stage re-ranking is proposed to obtain a reliable ranking list by applying the fusion metric strategy, which combines the two stages of formulating the ranking lists considering the Jaccard distance.

(3) A comprehensive experiment is conducted on the two representative vehicle datasets. The results confirm that the proposed approach has superior performance compared with the several state-of-the-art methods.

The rest of this paper is organized as follows. Section 2 provides an overview on the related research work dedicated to re-identification. Section 3 describes the overall proposed architecture. Sections 4 and 5 introduce the details about the proposed DFN and two-stage re-ranking, respectively. The experimental results are presented in Section 6. The conclusion of the paper is presented in Section 7.

2 Related work

Re-identification is a widely-discussed application in the field of computer vision, and most of existing methods focus on person re-identification that is aimed to search for the same people corresponding to a probe person in the target gallery set. Several hand-crafted features [14–17], including texture, color, and local maximal occurrence (LOMO) [18], have been applied to person re-identification. Moreover, at present, deep neural networks are widely applied to several computer vision tasks [19–24]. Zhao et al. [25] have proposed a cross-view training strategy to learn the filters with invariant and discriminative perspectives to distinguish pedestrians. Zheng et al. [26] have implemented a Siamese network using identification and verification losses to learn more discriminative pedestrian features. Cheng et al. [27] have designed a multi-channel convolution neural network (CNN) that employs the improved triplet loss function to achieve the final result. A novel joint spatio-temporal attention pooling network (ASTPN) [28] has been utilized to solve the problem of pedestrian video recognition. It can be used to improve re-identification performance. Zhao et al. [29] have proposed the multilevel dropout method and the improved Monte Carlo strategy to solve the overfitting problem and to reduce the impact of uncertain pedestrian representations, respectively. Many researches have considered person re-identification as a deep metric learning problem [30, 31].

The rapid advance of person re-identification applied to surveillance video has facilitated the research on vehicle re-identification. Many previous methods have identified vehicles by obtaining attributes manually [32]. Tang et al. [33] have combined local binary patterns and BOW-CN features to enable the robust discriminating ability. Zapletal et al. [34] have extracted the 3D bounding box, color histogram,

and gradient histogram of images, and then have applied linear regression to distinguish whether a vehicle corresponds to the same car model. Recent researches on vehicle re-identification have mainly focused on deep learning approaches, and several state-of-the-art methods have utilized CNN features to achieve excellent results. Yang et al. [35] have employed CNN to extract the overall and local features of vehicles aiming to perform fine-grained classification and attribute prediction. Zhang et al. [36] have designed the classification-oriented loss based on the original triplet loss enhanced with the distance learning and have proposed a new sampling method to solve the misleading problem. Zhou et al. [37] have proposed an adversarial bi-directional long short-term memory network (ABLN) that is capable of synthesizing the unknown view information about vehicles based on a partially visible view. Zhou et al. [38] have proposed a viewpoint-aware attentive multi-view inference (VAMI) model, which focuses the attention of the network model on the intersection of the input viewpoint and the target viewpoint of an image. Then, it converts single-view features into multi-view ones using the developed adversarial training architecture. Zhou et al. [39] have proposed a novel cross-view generative adversarial network (XVGAN), which combines the features of the original input and generated images to compute the distances aiming to improve the vehicle identification performance. Zhu et al. [40] designed the joint horizontal and vertical deep learning feature to describe the horizontal and vertical directions of vehicles in a more comprehensive manner so as to enhance the robustness of vehicle viewpoint variations. Shen et al. [11] have developed two structurally similar sub-networks of the Siamese neural networks (Siamese-Visual) to learn the similarity of an image pair. Liu et al. [32] have proposed a new loss function called coupled clusters loss (CCL) corresponding to the vehicle search problem. This function has been improved on the basis of the triplet loss and has replaced the original triplet input with positive and negative input sets. A mixed network structure base on CCL, namely, MixedDiff + CCL has been also proposed to learn a similarity measure accurately [32]. This structure can effectively extract the vehicle information and distinguish the differences in similar vehicles. The FACT model [9] has adopted the fusion strategy of color and texture, as well as the high-order semantic features, such as BOW-CN and BOW-SIFT, and deep semantic features obtained from GoogLeNet.

Most of existing metric learning and ranking algorithms have been successfully applied to the re-identification problem. The metric learning methods are mainly based on the Mahalanobis metric learning (KISSME) [41] and cross-view quadratic discriminant analysis (XQDA) [18]. Li et al. [42] have employed labels to measure learning, and then have combined semantic information to improve the image retrieval performance and accuracy. Zhong et al. [43] have proposed the k-reciprocal encoding method. Then, the Jaccard distance has been also introduced and fused with the initial distance to improve the results of person re-identification. Ding et al. [44] have proposed an improved triplet loss function to narrow the distance between positive sample pairs and enlarge the distance between the negative sample pairs for achieving accurate image retrieval. Li et al. [45] constructed the local adaptive decision function by using the combined model of locally adapted thresholds and metric distance. This method allows achieving outstanding results.

In turn, fine-grained identification of a vehicle is also relevant to vehicle re-identification. Fine-grained networks [46–48] can be used to detect and extract more detailed information, such as window labels and wheel bones. These specific regions of information are vital to identify the subtle differences between images in vehicle re-identification. Yu et al. [49] have proposed a deep learning model based on the vehicle detection model and vehicle fine-grained detection, and have concluded that a classification model can identify more details corresponding to vehicles. Hu et al. [50] have designed a multi-task that leverages multiple characteristics jointly to CNNs. It can be used to identify vehicle types on a fine-grained level. Zhang et al. [51] have proposed a fine-grained vehicle recognition method, which combines pre-training and hierarchical fine-tuning to provide better robustness with respect to visual changes.

3 Overview of the proposed method

The challenge to the existing re-identification methods lies in the subtle differences between vehicles of

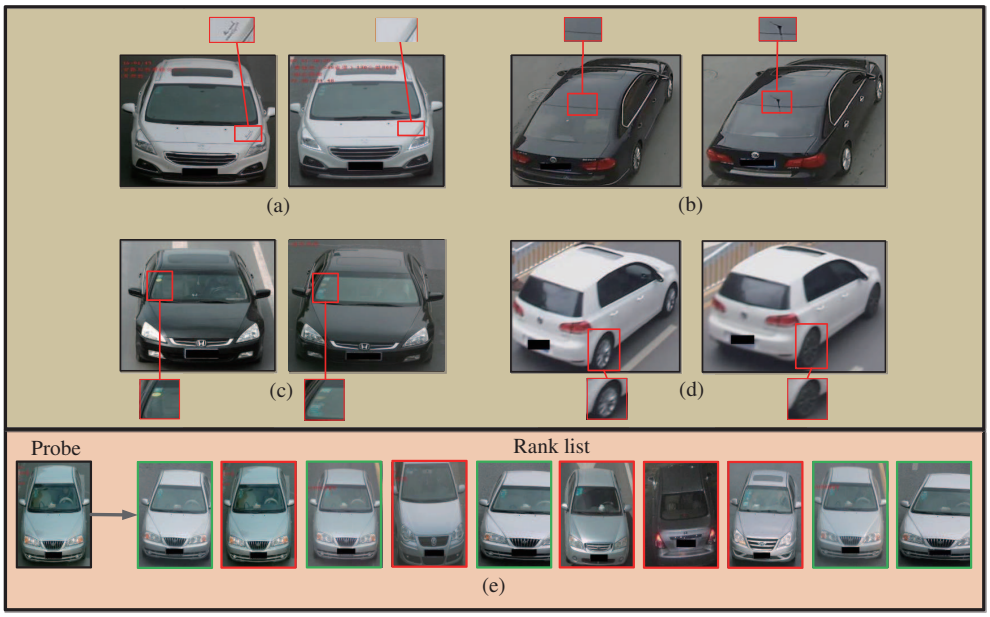

Figure 1 (Color online) The challenges associated with vehicle re-identification.

the same model and the lack of positive samples with high ranks. Figures 1(a)–(d) represent several examples of subtle differences between different vehicle IDs having the similar appearance, which are obtained from the two benchmark datasets, namely, VeRi-776 [52] and VehicleID [32]. The appearance of the vehicles presented in Figures 1(a) and (c) is rather similar and therefore, they can be distinguished by particular marks, such as stickers on the car hood and windshield. Vehicles corresponding to the same model can be distinguished only by subtle differences in their parts. For example, it is possible to determine whether there are receiving antenna devices on both vehicle roofs presented in Figure 1(b), or whether the styles of a wheel hub shown in Figure 1(d) are the same. The re-ranking method is also vital in re-identification. Figure 1(e) represents the top ten ranked images of the probe. The red box denotes an incorrect sample, and the green box corresponds to a positive sample. It should be noted that several false samples have received higher ranks, while particular positive samples have obtained lower ranks.

The proposed architecture mainly consists of the two components: DFN and the two-stage re-ranking. The proposed DFN is implemented as the first component. The pipeline of DFN is illustrated in the left part of Figure 2. The overall proposed network architecture adopts the multi-loss as supervision signals. It is composed of the two parts: the Siamese network and the fine-grained one. First, the Siamese network simultaneously learns the deep features of images and conducts similarity mapping from image pairs to the Euclidean space by identification and verification losses. Then, the fine-grained network is used to identify subtle differences between the vehicles by applying fine-grained classification loss. The two-stage re-ranking method is proposed, as shown in the right part of Figure 2. It implies fusing the two parts of the deep feature vectors to compute fusion features and is divided into two stages. In stage 1, we obtain k-reciprocal features from the fusion features. In stage 2, sample mean features are formed by extracting the mean center of the k-reciprocal nearest neighbor. The final distance is weighted by the original and Jaccard distances.

4 Discriminative fine-grained network

The Siamese network [26] is introduced according to the scheme presented in the left part of Figure 2. Similarly to the approach proposed by Zheng et al., this network mainly consists of the two sub-networks of the same structure, which simultaneously share the weights during the training period. The shared CNNs combine the identification and verification losses at once. Therefore, two deep networks can be jointly regulated by verification supervision while being separately managed by identification supervision.

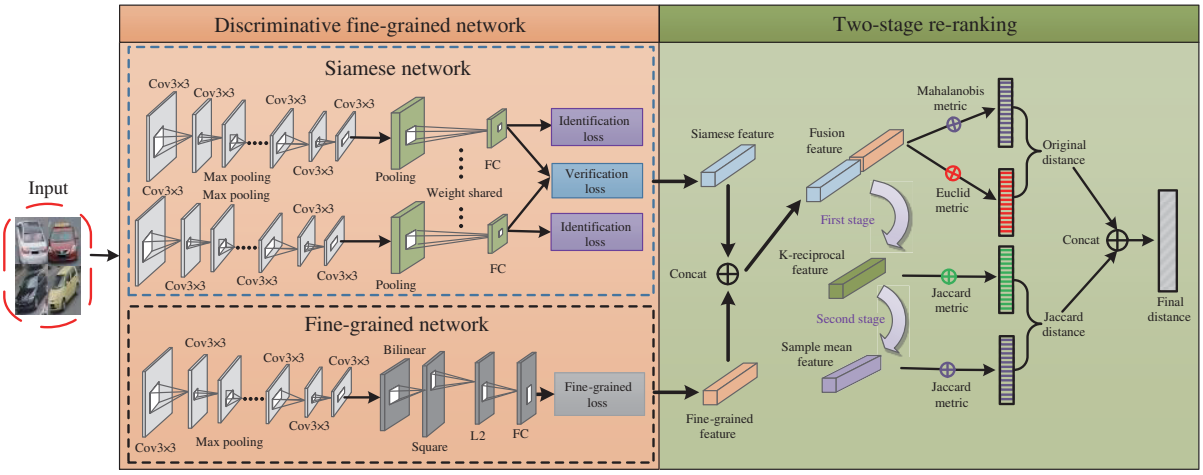

Figure 2 (Color online) The overview of the proposed architecture for vehicle re-identification. Firstly, the dataset is input into the network. Then the discriminative fine-grained network part including the Siamese network in the upper part and the fine-grained network in the lower part is applied. Finally, the two-stage re-ranking part is executed to merge the feature vectors of the two sub-networks to obtain the final distance by the two-stage calculation.

The identification model considers re-identification as a multi-classification task. This model obtains deep network features by supervised learning on the basis of the strong label information. By using cross entropy loss, the identification function is defined similarly to the traditional softmax loss function and can be written as follows:

$$\mathrm{Loss}_i = \frac{1}{K}\left[\sum_{k=1}^{K}\left(\tilde{y}_k \cdot \log\left(\frac{\mathrm{e}^{f\cdot\omega_t}}{\sum_{m=1}^{c}\mathrm{e}^{f\cdot\omega_m}}\right)\right)\right], \tag{1}$$

where K represents the number of samples in the training set, C indicates the total number of classes, $\tilde{y}_k$ is the correct probability of the target vehicle, $\omega = [\omega_1, \omega_2, \ldots, \omega_C]$ is the predictive vehicle probability matrix, ω_t represents the matrix of the correct vehicle labels, and f denotes the corresponding extracted features.

In turn, the verification model addresses the problem as a two-class similarity regression task. The image pairs input in the network are analyzed to predict whether they correspond to the same class. The Siamese deep network can be used to learn similarity metrics in the Euclidean space by identification supervision. We note the presence of the risk that the contrast loss may lead to overfitting at the data size. Therefore, the cross-entropy loss is still applied to train the validation model, which can be formulated as follows:

$$\mathrm{Loss}_v = \frac{1}{G}\left[\sum_{g=1}^{G}\left(\tilde{y}_{1,2} \cdot \log\left(\frac{\mathrm{e}^{(f_1-f_2)^2\cdot\omega_s}}{\sum_{n=1}^{2}\mathrm{e}^{(f_1-f_2)^2\cdot\omega_{sn}}}\right)\right)\right], \tag{2}$$

where G is the number of image pairs, and $\tilde{y}_{1,2}$ indicates that the pair of images corresponds to the same target. When the detected pair of images matches the same target, $\tilde{y}_{1,2} = 1$; otherwise, $\tilde{y}_{1,2} = 0$. In the present study, a square layer is added to fuse the features extracted from the Siamese network. The two input eigenvectors are calculated in accordance with $(f_1 - f_2)^2$, and then, the fused feature output is obtained. ω_s represents the matrix of the correctly detected vehicle labels.

Although the identification and verification losses can facilitate extracting features with the considerable discriminative ability, they still have several deficiencies. First, the network does not learn fine features effectively owing to the subtle differences in many positive input sample pairs. In the problem of vehicle re-identification, the vehicles with similar appearance may belong to different IDs. However, the vehicles belonging to the same class ID can be mismatched owing to various factors, such as the angle and illumination conditions. This phenomenon is illustrated in Figure 3. Owing to the angle and similar appearance factors, people often misjudge the left and middle vehicles represented in this

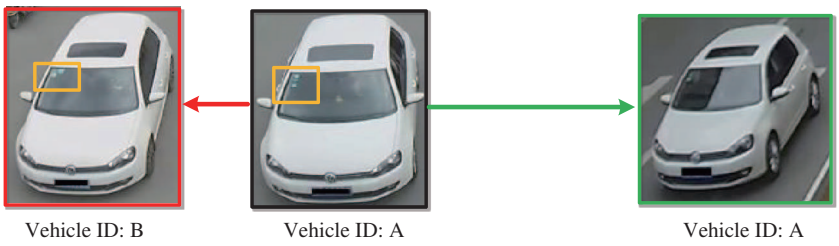

Figure 3 (Color online) The influence of the subtle feature information on vehicle re-identification.

figure as the same target. In fact, the right and middle vehicles correspond to the same vehicle. To mitigate this problem, a deep network can be applied to extract precise and discriminative features. At present, fine-grained networks are reported to achieve the excellent performance in image classification. However, the extracted features are high-dimensional and may reach millions of levels, which makes the computational load too large and difficult to process. Gao et al. [13] have designed the compact bilinear pooling method with the same distinguishing power as the bilinear representation, which benefits from the compact bilinear representation of thousands of dimensions.

Unlike the study of Zheng et al. [26], we apply the compact bilinear pooling method to obtain fine-grained features based on the previous vehicle re-identification effort. The sub-network considered in this section is the VGG-16 model. According to the study described in [13], this model replaces the compact bilinear layer with the original pooling layer and adds the element-wise signed square-root layer ($y \leftarrow \mathrm{sign}(x)\sqrt{|x|}$) and L2 regularization ones to perform a normalization step. As a result, a global image descriptor is extracted by using the compact bilinear layer according to

$$C(X) = \sum_{s \in S} x_s x_s^{\mathrm{T}}, \tag{3}$$

where S denotes a series of spatial locations, and $X = \left(x_1, \ldots, x_{|S|}\right)$ represents a set of local descriptors. X_S are the local descriptors output from HOG, SIFT or by a forward pass based on a CNN.

The main challenge of using the fine-grained classification network is to efficiently detect and extract the important local area information from an image. These specific regions of information are vital to identify subtle differences between vehicle images. From the above, we conclude that the compact bilinear layer can be successfully used to combine the features corresponding to the different local positions of an image to obtain a discriminative global representation vector representing fine-grained features. Moreover, a normalization step is performed in which the compact bilinear vector passes through the element-wise signed square-root and the L2 regularization layers. More detailed description about these concepts can be found in [13].

This specifically added network structure is represented in the lower left part of Figure 2. To extract the fine-grained features of vehicles, we utilize this element of the sub-network, which is supervised by fine-grained losses during training. We also employ the softmax loss function as a fine-grained one, which can be written as follows:

$$\mathrm{Loss}_f = \frac{1}{K}\left[\sum_{k=1}^{K}\left(\tilde{y}_k \cdot \log\left(\frac{\mathrm{e}^{f \cdot \omega_t}}{\sum_{m=1}^{C} \mathrm{e}^{f \cdot \omega_m}}\right)\right)\right] + \lambda \sum_{m=1}^{C} \omega_m, \tag{4}$$

where K represents the number of samples in the training set, C indicates the total number of classes, $\tilde{y}_k$ is the correct probability of the target vehicle, $\omega = [\omega_1, \omega_2, \ldots, \omega_C]$ is the predictive vehicle probability matrix, ω_t represents the matrix of the correct vehicle labels, f represents the corresponding extracted fine-grained features, and λ is an L_2 regularization parameter.

We extract the features in the two sub-networks of DFN, and then merge them. The fusion method is based on simple dimension superposition and can be calculated as follows:

$$f_{\mathrm{all}} = [f_1, f_2, \ldots, f_N], \tag{5}$$

where N represents the number of sub-networks, and f_n is the N-th extracted feature vector.

The calculation method of formula (5) is simple; however, it can effectively retain the feature representation of strong discriminative power. The right part of Figure 2 represents the fusion feature vector of 12288 dimensions, including the 4096- and 8192-dimensional features of the outputs obtained from the Siamese network and fine-grained network parts.

5 Two-stage re-ranking

Zhong et al. [43] have proposed the k-reciprocal encoding method to improve the results of person re-identification. On its basis, in the present paper, we propose a two-stage re-ranking method for vehicle re-identification to determine the characteristics and differences among vehicles.

In the first stage, the k-reciprocal encoding method is introduced to obtain a k-reciprocal feature. We assume that the gallery set of N images is defined as $G = \{g_i | i = 1, 2, \ldots, N\}$. $H(p,k) = \{g_1, g_2, \ldots, g_k\}$ is the top-k similar sample set of the probe vehicle p and can be defined according to (6). $N(g_i, k)$ represents the top-k similar sample set of $\boldsymbol{g}_i$. Assuming the similarity of the two sets, their intersection is considered as the most similar candidate target as follows:

$$H(p,k) = \{g_i | (g_i \in N(p,k)) \cap (p \in N(g_i,k))\}. \tag{6}$$

In [43], the 1/2 k-reciprocal nearest neighbors of each candidate are added into a more robust set. It is proposed to concentrate the positive samples in the forefront of the ranking list:

$$H^*(p,k) = H(p,k) \cup H\left(q, \frac{1}{2}k\right). \tag{7}$$

Confidence values corresponding to the highly ranked samples are often influenced by neighboring samples. In the second stage, we define p of the confidence item $\overline{p}$ to enhance the confidence of the positive samples to retrieve a rank-list after calculating $H(p,k)$ using the fusion feature. The selection strategy is to extract the mean of the top-k list samples of the probe vehicle p, which can be obtained according to (8) as follows:

$$\overline{p} = \text{avg}(H(p,k)). \tag{8}$$

The robust set $H^*(\overline{p},k)$ is calculated from the confidence sample $\overline{p}$ corresponding to the probe vehicle p. Figure 4 represents the process of deriving $H^*(\overline{p},k)$. First, we set image Q as the probe vehicle, and image C is the sample mean candidate for $H(\text{Q},20)$ in the first row of Figure 4. Then, we obtain $H(\text{C},20)$ in the second row. Finally, we incrementally add the 1/2 k-reciprocal nearest neighbors of the candidate in $H(\text{C},20)$ into $H^*(\text{C},20)$ in the third and fourth rows of Figure 4. $H^*(\text{C},20)$ has more positive samples than $H(\text{Q},20)$. Accordingly, the k-reciprocal nearest neighbors of the mean sample candidate $\overline{p}$ can map appropriately to the positive samples corresponding to the ranking-list of the probe vehicle p that are difficult to distinguish.

The Jaccard distance is applied to measure the difference between two sets. As described in [43], if two images are similar, their k-reciprocal nearest neighbor sets will have a larger number of duplicate samples. The Jaccard distance between p and $\boldsymbol{g}_i$ is calculated as follows:

$$d_J(p, g_i) = 1 - \frac{|H^*(p,k) \cap H^*(g_j,k)|}{|H^*(p,k) \cup H^*(g_j,k)|}. \tag{9}$$

The final distance d^* is weighted by the original distance (namely, Mahalanobis and Euclidean distances) and the Jaccard one, and is defined as

$$d^* = \sum_{x \in (M,E)} d_x(p, g_i) + \sum_{y \epsilon (p,\overline{p})} d_J(y, g_i), \tag{10}$$

where $d_M(p, g_i)$ denotes the Euclidean distance between p and $\boldsymbol{g}_i$, $d_E(p, g_i)$ is the Mahalanobis distance between p and $\boldsymbol{g}_i$, $d_J(y, g_i)$ is the Jaccard distance between y and $\boldsymbol{g}_i$, and y contains p and $\overline{p}$.

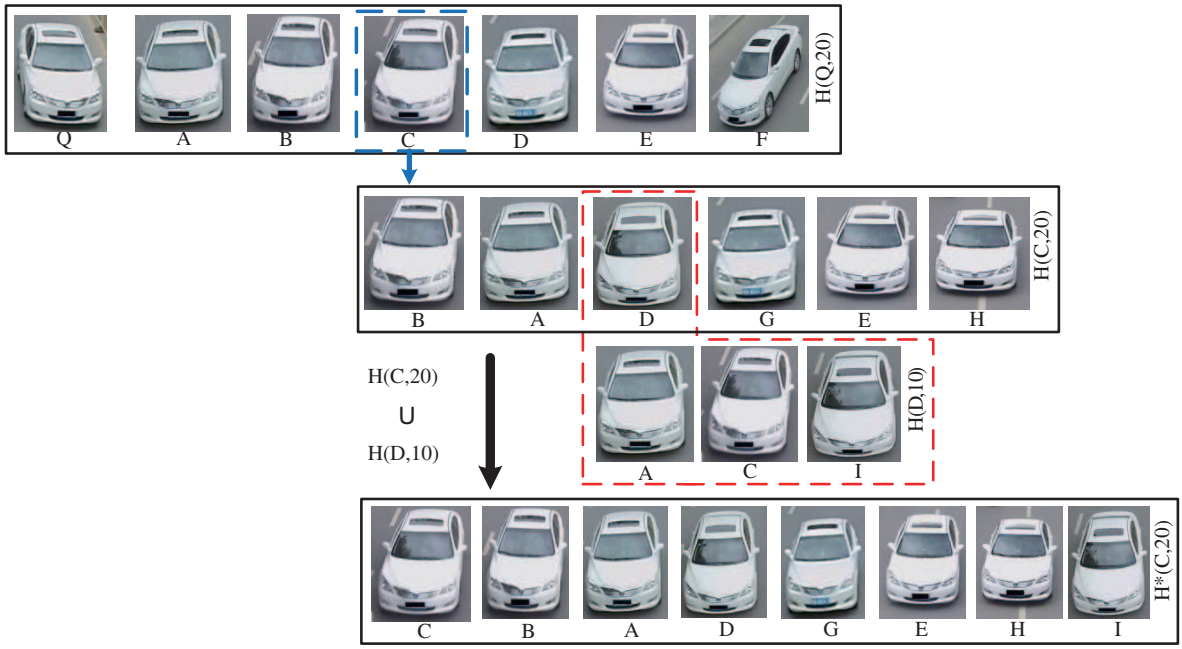

Figure 4 (Color online) Example of the selection of candidate $\overline{p}$ and the definition process of robust set $H^*(\mathrm{C}, 20)$ in the second step of re-ranking.

6 Experiments

6.1 Datasets and comparison methods

To evaluate the efficiency of the proposed DFN, we conduct the experiments on VeRi-776 and VehicleID, the vehicle re-identification datasets, extracted from large-scale surveillance videos.

VeRi-776 is an urban surveillance vehicle dataset with the multi-view and spatio-temporal location information. It contains about 50000 images of 776 vehicles from 20 cameras. Several labels of other attributes, such as car model and color, are also included in the dataset. The whole dataset is split into three subsets as follows: a training set consisting of 37778 images of 576 vehicles, a test set of 11579 images corresponding to 200 vehicles, and a subset of 1678 images from the test set is extracted as the query set.

VehicleID is a large-scale vehicle re-identification dataset that contains front and rear views of cars. It contains 221763 images of 26267 vehicles. Furthermore, 10319 vehicles are labeled using the model information. The whole dataset comprises the training and test sets in which the training part contains 110178 images of 13134 vehicles, and the test part contains 111585 images of 13133 vehicles, respectively. Considering that the test set is rather large, three subsets (small, medium and large) are extracted from the VehicleID dataset test data. We randomly select one image of each vehicle from the probe set, and consider that all other images belong to the gallery set.

The proposed framework is compared with the state-of-the-art vehicle re-identification methods, such as LOMO [18], BOW-CN [16], FACT [9], VGG+CCL [32], MixedDiff + CCL [32], VAMI [38], XVGAN [39], DLCNN [26], and Siamese-Visual [11].

6.2 Evaluation metric

Liu et al. [52] have proposed a standard evaluation protocol for the VeRi-776 dataset. On this basis, we employ the cumulative matching characteristic (CMC) curve for evaluation. In turn, mean average precision (mAP) is a common evaluation index used in the multi-label image classification task and is an important criterion for measuring the quality of a similar task model. Therefore, mAP is also adopted to evaluate the vehicle re-identification performance. For the VehicleID dataset, the probe image set

Table 1 Comparison of the state-of-the-art methods on the VeRi-776 dataset

Method	rank1 (%)	mAP (%)
LOMO [18]	24.59	9.68
FACT [9]	51.89	18.69
Siamese-Visual [11]	41.12	29.48
BOW-CN [16]	33.82	9.63
VAMI [38]	77.03	50.13
XVGAN [39]	60.20	24.65
DLCNN [26]	82.42	49.88
Ours	**88.14**	**61.85**

Table 2 Comparison of the state-of-the-art methods on the VehicleID dataset

Method	Small		Medium		Large	
	rank1 (%)	rank5 (%)	rank1 (%)	rank5 (%)	rank1 (%)	rank5 (%)
LOMO [18]	19.92	32.83	19.52	29.91	15.72	25.56
FACT [9]	49.93	68.37	45.01	64.75	40.12	60.59
VGG+CCL [32]	43.92	65.01	38.84	61.91	34.58	55.72
MixedDiff+CCL [32]	48.52	74.55	43.94	67.96	40.85	62.79
VAMI [38]	63.08	83.12	52.69	75.08	47.28	70.06
XVGAN [39]	52.79	80.69	49.47	71.42	44.92	66.71
DLCNN [26]	73.01	82.70	66.50	77.06	61.00	73.17
Ours	**77.02**	**85.04**	**71.81**	**80.81**	**66.29**	**78.42**

comprises randomly selected images corresponding to the one identity in the gallery set. Following the method described in [52], we apply the CMC curve to evaluate the re-identification performance.

6.3 Experimental results

The experimental results are represented in Tables 1 and 2, and Figure 5. Bold data represent the best experimental results. In the results corresponding to VeRi-776, "Ours" indicates the overall pipeline of DFN. It can be seen that the proposed method achieves the best results among all considered vehicle re-identification methods [9, 11, 16, 18, 26, 32, 38, 39]. LOMO and BOW-CN show poor performance in terms of the hand-crafted feature. FACT and Siamese-Visual, which adopt the deep network to learn semantic features, achieve acceptable performance. Both XVGAN and VAMI improve the result of vehicle re-identification results by generating multi-view representation. They focus on exploiting multi-view information to obtain a global feature, rather than extracting the fine-grained features. However, they cannot distinguish the subtle differences of vehicles accurately because the appearance of the same model of the vehicles captured in same viewpoint is still similar. DLCNN outperforms the above mentioned methods owing to the combination of verification and identification losses in the Siamese network. Compared with DLCNN, the proposed method achieves a gain of 11.97% in mAP and an increase of 5.72% in the rank-1 accuracy. Similarly to the results on VeRi-776, VGG + CCL and MixedDiff + CCL also demonstrate the effectiveness of utilizing CNN with the improved loss function to extract semantic features in VehicleID. However, they cannot distinguish the subtle differences of vehicles accurately. The proposed method achieves a 4.01% improvement in terms of the rank-1 accuracy and a 2.34% improvement in the rank-5 accuracy compared with the second best method (DLCNN) in the case of the small scale test dataset. In the case of the medium scale and large-scale test datasets, the proposed method also achieves 5.31% and 5.29% higher accuracy in terms of the rank-1 rate, and 3.75% and 5.25% higher accuracy in terms of the rank-5 rate, respectively. Evaluating on both VeRi-776 and VehicleID, we can observe that significant performance improvements can be achieved by extracting fine-grained features. This demonstrates that the proposed method can better distinguish similar vehicles than other considered state-of-the-art approaches.

Applying re-ranking methods can also improve the results of vehicle re-identification. Therefore, we

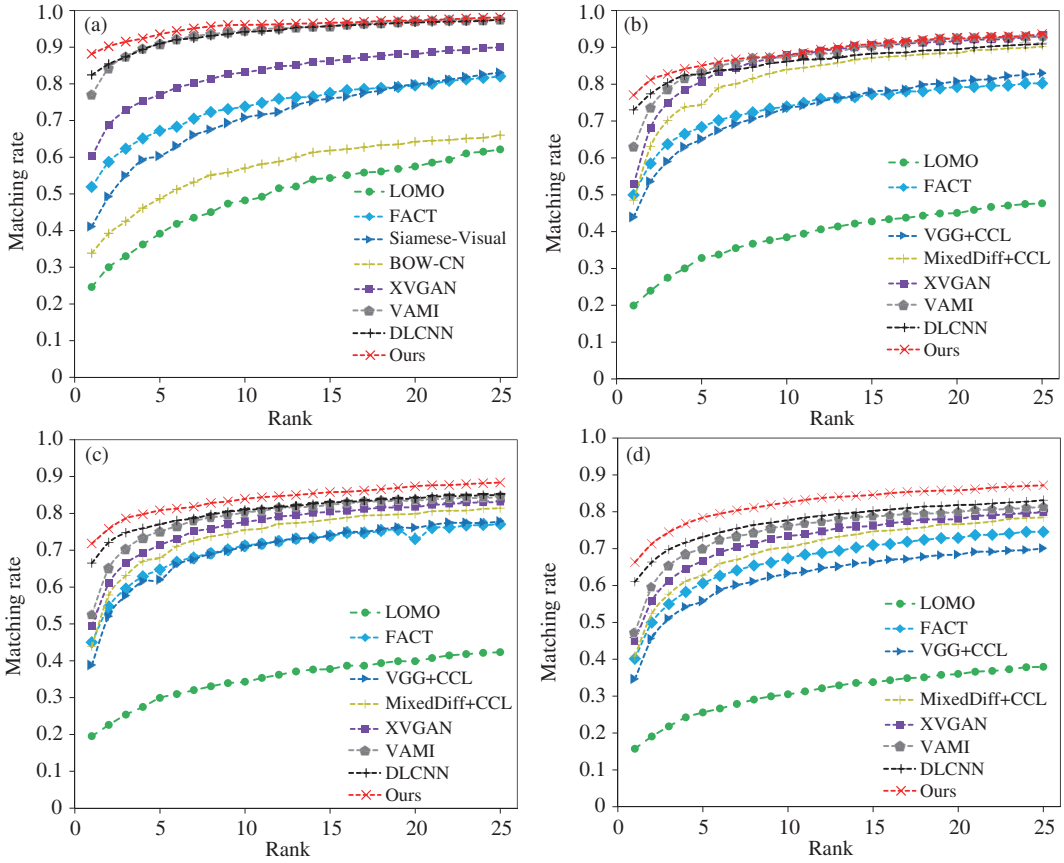

Figure 5 (Color online) CMC curves of different methods. (a) VeRi-776; (b) the small test set of VehicleID; (c) the medium test set of VehicleID; (d) the large test set of VehicleID.

Table 3 Comparison of the results obtained using the methods with and without re-ranking on VeRi-776

Method	rank1 (%)	mAP (%)
Base	88.14	61.85
Base+Zhong [43]	89.03	65.19
Base+TR	**90.11**	**66.10**

Table 4 Comparison of the results obtained using the methods with and without re-ranking on VehicleID

Method	Small		Medium		Large	
	rank1 (%)	rank5 (%)	rank1 (%)	rank5 (%)	rank1 (%)	rank5 (%)
Base	77.02	85.04	71.81	80.81	66.29	78.42
Base+Zhong [43]	77.89	85.28	72.38	81.06	67.92	79.17
Base+TR	**79.00**	**86.01**	**74.06**	**82.19**	**69.50**	**79.79**

utilize two re-ranking methods to evaluate the aforementioned datasets. Tables 3 and 4 represent the results obtained using the methods with and without re-ranking on the VeRi-776 and VehicleID datasets. "Base" denotes the proposed network (DFN) and "Base + TR" corresponds to the overall network using the two-stage re-ranking method. The method described in [43] allows gaining additional improvements based on the DFN. However, the TR method yields better performance compared with the method presented in [43]. Therefore, the proposed method allows improving the re-ranking results more efficiently compared with the other two methods.

7 Conclusion

In this paper, the DFN for vehicle re-identification using the two-stage re-ranking framework is proposed. It can be used to extract more subtle features and improve the re-ranking method. First, the Siamese and fine-grained networks are combined to extract fusion features. Owing to the combined effect of verification, identification, and fine-grained losses, the extracted features have strong discriminative capability. The two-stage re-ranking is applied to obtain the sample mean feature, which is then added to the final distance metric. As a result, the number of the positive samples in the top-k list increases. The conducted experiments demonstrate that the proposed method outperforms other considered state-of-the-art approaches on the VehicleID and VeRi-776 datasets.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant Nos. 61762061, 62076117), National Key R&D Program of China (Grant Nos. 2017YFB0801701, 2017YFB0802805), Natural Science Foundation of Jiangxi Province (Grant No. 20161ACB20004), and Jiangxi Key Laboratory of Smart City (Grant No. 20192BCD40-
002).

References

1 Gou C, Wang K, Yao Y, et al. Vehicle license plate recognition based on extremal regions and restricted Boltzmann machines. IEEE Trans Intell Transp Syst, 2016, 17: 1096–1107
2 Min W, Li X, Wang Q, et al. New approach to vehicle license plate location based on new model YOLO-L and plate pre-identification. IET Image Process, 2019, 13: 1041–1049
3 Wang Y, Zhao C, Liu X, et al. Fast cartoon-texture decomposition filtering based license plate detection method. Math Problems Eng, 2018, 2018: 1–9
4 Wang T Q, Gong S G, Zhu X T, et al. Person re-identification by discriminative selection in video ranking. IEEE Trans Pattern Anal Mach Intell, 2016, 38: 2501–2514
5 Zhao R, Oyang W L, Wang X G. Person re-identification by saliency learning. IEEE Trans Pattern Anal Mach Intell, 2017, 39: 356–370
6 Cho Y J, Yoon K J. Improving person re-identification via pose-aware multi-shot matching. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2016
7 Zhao H Y, Tian M Q, Sun S Y, et al. Spindle net person re-identification with human body region guided. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2017
8 Feris R S, Siddiquie B, Petterson J, et al. Large-scale vehicle detection, indexing, and search in urban surveillance videos. IEEE Trans Multimedia, 2012, 14: 28–42
9 Liu X C, Liu W, Ma H D, et al. Large-scale vehicle re-identification in urban surveillance videos. In: Proceedings of IEEE International Conference on Multimedia and Expo, 2016
10 Loy C C, Xiang T, Gong S G. Multi-camera activity correlation analysis. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2009
11 Shen Y T, Xiao T, Li H S, et al. Learning deep neural networks for vehicle re-id with visual-spatio-temporal path proposals. In: Proceedings of IEEE International Conference on Computer Vision, 2017
12 Wang Z D, Tang L M, Liu X H, et al. Orientation invariant feature embedding and spatial temporal regularization for vehicle re-identification. In: Proceedings of IEEE International Conference on Computer Vision, 2017
13 Gao Y, Beijbom O, Zhang N, et al. Compact bilinear pooling. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2016
14 Matsukawa T, Okabe T, Suzuki E, et al. Hierarchical Gaussian descriptors with application to person re-identification. IEEE Trans Pattern Anal Mach Intell, 2019. doi: 10.1109/TPAMI.2019.2914686
15 Chen D P, Yuan Z J, Chen B D, et al. Similarity learning with spatial constraints for person re-identification. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2016
16 Zheng L, Shen L Y, Tian L, et al. Scalable person re-identification: a benchmark. In: Proceedings of IEEE International Conference on Computer Vision, 2015
17 Varior R R, Wang G, Lu J W, et al. Learning invariant color features for person reidentification. IEEE Trans Image Process, 2016, 25: 3395–3410
18 Liao S C, Hu Y, Zhu X Y, et al. Person re-identification by local maximal occurrence representation and metric learning. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2015
19 Min W, Cui H, Rao H, et al. Detection of human falls on furniture using scene analysis based on deep learning and activity characteristics. IEEE Access, 2018, 6: 9324–9335
20 Liao Y, Xiong P, Min W, et al. Dynamic sign language recognition based on video sequence with BLSTM-3D residual networks. IEEE Access, 2019, 7: 38044–38054
21 Min W, Fan M, Li J, et al. Real-time face recognition based on pre-identification and multi-scale classification. IET Comput Vision, 2019, 36: 165–171

22 Zhang K, Liu N, Yuan X F, et al. Fine-grained age estimation in the wild with attention LSTM networks. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2018
23 Ji Z, Xiong K, Pang Y, et al. Video summarization with attention-based encoder-decoder networks. IEEE Trans Circ Syst Video Technol, 2020, 30: 1709–1717
24 Ji Z, Sun Y, Yu Y, et al. Attribute-guided network for cross-modal zero-shot hashing. IEEE Trans Neural Netw Learn Syst, 2020, 31: 321–330
25 Zhao R, Ouyang W L, Wang X G. Learning mid-level filters for person re-identification. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2014
26 Zheng Z D, Zheng L, Yang Y. A discriminatively learned CNN embedding for person reidentification. ACM Trans Multimedia Comput Commun Appl, 2018, 14: 1–20
27 Cheng D, Gong Y H, Zhou S P, et al. Person re-identification by multi-channel parts-based CNN with improved triplet loss function. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2016
28 Xu S J, Cheng Y, Gu K, et al. Jointly attentive spatial-temporal pooling networks for video-based person re-identification. In: Proceedings of IEEE International Conference on Computer Vision, 2017
29 Zhao C R, Chen K, Zang D, et al. Uncertainty-optimized deep learning model for small-scale person re-identification. Sci China Inf Sci, 2019, 62: 220102
30 Paisitkriangkrai S, Shen C H, Hengel A V D. Learning to rank in person re-identification with metric ensembles. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2015
31 Yi D, Zhen L, Liao S C, et al. Deep metric learning for person re-identification. In: Proceedings of the 1st International Conference on Pattern Recognition Applications and Methods, 2014
32 Liu H Y, Tian Y H, Wang Y W, et al. Deep relative distance learning: tell the difference between similar vehicles. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2016
33 Tang Y, Wu D, Jin Z, et al. Multi-modal metric learning for vehicle re-identification in traffic surveillance environment. In: Proceedings of IEEE International Conference on Image Processing, 2017
34 Zapletal D, Herout A. Vehicle re-identification for automatic video traffic surveillance. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2016
35 Yang L J, Lou P, Loy C C, et al. A large-scale car dataset for fine-grained categorization and verification. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2015
36 Zhang Y H, Liu D, Zha Z J. Improving triplet-wise training of convolutional neural network for vehicle re-identification. In: Proceedings of IEEE International Conference on Mutimedia and Expo, 2017
37 Zhou Y, Shao L. Vehicle re-identification by adversarial bi-directional LSTM network. In: Proceedings of IEEE Winter Conference on Applications of Computer Vision, 2018
38 Zhou Y, Shao L. Viewpoint-aware attentive multi-view inference for vehicle re-identification. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2018
39 Zhou Y, Shao L. Cross-view GAN based vehicle generation for re-identification. In: Proceedings of British Machine Vision Conference, 2017
40 Zhu J Q, Zeng H Q, Jin X, et al. Joint horizontal and vertical deep learning feature for vehicle re-identification. Sci China Inf Sci, 2019, 62: 199101
41 Martin K, Hirze M, Wohihart P, et al. Large scale metric learning from equivalence constraints. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2012
42 Li Z C, Tang J H. Weakly supervised deep metric learning for community-contributed image retrieval. IEEE Trans Multimedia, 2015, 17: 1989–1999
43 Zhong Z, Zheng L, Cao D L, et al. Re-ranking person re-identification with k-reciprocal encoding. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2017
44 Ding S Y, Lin L, Wang G R, et al. Deep feature learning with relative distance comparison for person re-identification. Pattern Recogn, 2015, 48: 2993–3003
45 Li Z, Chang S Y, Liang F, et al. Learning locally-adaptive decision functions for person verification. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2013
46 Valev K, Schumann A, Sommer L, et al. A systematic evaluation of recent deep learning architectures for fine-grained vehicle classification. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2018
47 Ma Z, Chang D, Li X. Channel max pooling layer for fine-grained vehicle classification. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2019
48 Wang Q, Ding Y D. A novel fine-grained method for vehicle type recognition based on the locally enhanced PCANet neural network. J Comput Sci Technol, 2018, 33: 335–350
49 Yu S, Wu Y, Li W, et al. A model for fine-grained vehicle classification based on deep learning. Neurocomputing, 2017, 257: 97–103
50 Hu B, Lai J H, Guo C C. Location-aware fine-grained vehicle type recognition using multi-task deep networks. Neurocomputing, 2017, 243: 60–68
51 Zhang Q, Zhuo L, Hu X, et al. Fine-grained vehicle recognition using hierarchical fine-tuning strategy for urban surveillance videos. In: Proceedings of International Conference on Progress in Informatics and Computing, 2017
52 Liu X C, Liu W, Mei T, et al. A deep learning-based approach to progressive vehicle re-identification for urban surveillance. In: Proceedings of Europeon Conference on Computer Vision, 2016