Attributes Guided Feature Learning for Vehicle Re-identification

Hongchao Li, Xianmin Lin, Aihua Zheng, Chenglong Li, Bin Luo*, Ran He, and Amir Hussain

***Abstract*—Vehicle Re-ID has recently attracted enthusiastic attention due to its potential applications in smart city and urban surveillance. However, it suffers from large intra-class variation caused by view variations and illumination changes, and inter-class similarity especially for different identities with a similar appearance. To handle these issues, in this paper, we propose a novel deep network architecture, which guided by meaningful attributes including camera views, vehicle types and colors for vehicle Re-ID. In particular, our network is end-to-end trained and contains three subnetworks of deep features embedded by the corresponding attributes. For network training, we annotate the view labels on the VeRi-776 dataset. Note that one can directly adopt the pre-trained view (as well as type and color) subnetwork on the other datasets with only ID information, which demonstrates the generalization of our model. Extensive experiments on the benchmark datasets VeRi-776 and VehicleID suggest that the proposed approach achieves the promising performance and yields to a new state-of-the-art for vehicle Re-ID.**

***Index Terms*—Vehicle Re-identification, Deep Features, Attributes.**

I. Introduction

VEHICLE re-identification (Re-ID) is a frontier and important research problem in computer vision, which has many potential applications, such as intelligent transportation, urban surveillance and security since vehicle is the most important object in urban surveillance. The aim of vehicle Re-ID is to identify the same vehicle across non-overlapping cameras. Although license plate can uniquely identify the vehicle, it is scarcely recognizable due to the challenging factors of motion blur, challenging camera view and low resolution, to name a few. Some researchers have explored spatio-temporal information [1], [2], [3], [4] to boost the performance of appearance based vehicle Re-ID. However, it is difficult to obtain the complete spatio-temporal information since the vehicles may only appear in a few of the cameras in the large scale camera networks. Therefore, the prevalent vehicle Re-ID methods still focus on appearance based models.

Extensive works dedicate on person Re-ID in the past decade [5], [6], [7], [8], [9], [10], [11], which focus on two mainstreams: (1) Appearance modelling [6], [7], [11], which develops robust feature descriptors to encode the various changes and occlusions among different camera views. (2) Learning-based methods [5], [12], [13], [8], [9], [10], which learns metric distance to mitigate the appearance gaps between the low-level features and high-level semantics. Recently, deep neural networks have made a marvelous progress on both feature learning [5], [9], [10] and metric learning [12], [13], [8] for person Re-ID. However, directly employing person Re-ID models for vehicle Re-ID could not guarantee the satisfactory performance, since the appearance of pedestrians and vehicles varies in the different manner from different viewpoints.

Although much progress has been made on vehicle Re-ID [14], [15], [2], [3], [16], [1], [17], [18], [19], [20], it still encounters many challenges as in addition to the common challenges in person Re-ID such as occlusion and illumination, etc. The first crucial challenge of vehicle Re-ID is the large intra-class variation caused by the viewpoint variation across different cameras, which has been widely explored in person Re-ID [21], [22], [23], [24]. This issue is even more challenging in vehicle Re-ID since most of the vehicle images under a certain camera are almost in the same viewpoint due to the rigid motion of the vehicles. Unfortunately, it might not achieve the satisfactory performance when directly employing the methods from person Re-ID to vehicle Re-ID since the appearance distributes totally different between persons and vehicles. Some vehicle Re-ID methods [25], [26] use adversarial learning schemes to generate multi-view images or features from a single image, and can thus address the challenge of view variation to some extent. But they might be difficult to distinguish different vehicles with very similar appearance. Furthermore, they neglect the attributes information, such as type and color, which would be critical cues for boosting the performance of vehicle Re-ID.

The second challenge is the high inter-class similarity especially for different identities with a similar appearance. Incorporating the attributes information suffices to generate better discriminative representation for person Re-ID [9], [28], [10], [29]. Therefore, it is essential to learn the deep features with the supervision of attributes in vehicle Re-ID, enforcing the same identity with the consistent attributes. Li et al. [15] introduce the attribute recognition into the vehicle Re-ID framework, and use extra semantic information to assist vehicle identification especially for different identities with a

H. Li, X. Lin, and B. Luo are with School of Computer Science and Technology, Anhui University, Hefei 230601, China.

A. Zheng, and C. Li are with Information Materials and Intelligent Sensing Laboratory of Anhui Province, Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, School of Artificial Intelligence, Anhui University, Hefei, 230601, China.

R. He is with Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China.

Amir Hussain is with Edinburgh Napier University, School of Computing, Edinburgh EH10 5DT, Scotland, U.K.

This research is supported in part by the National Key Research and Development Program of China (No. 2018AAA0100400) and the National Natural Science Foundation of China (No. 61976002, 61976003, 62076003 and 61860206004).

* Corresponding author.

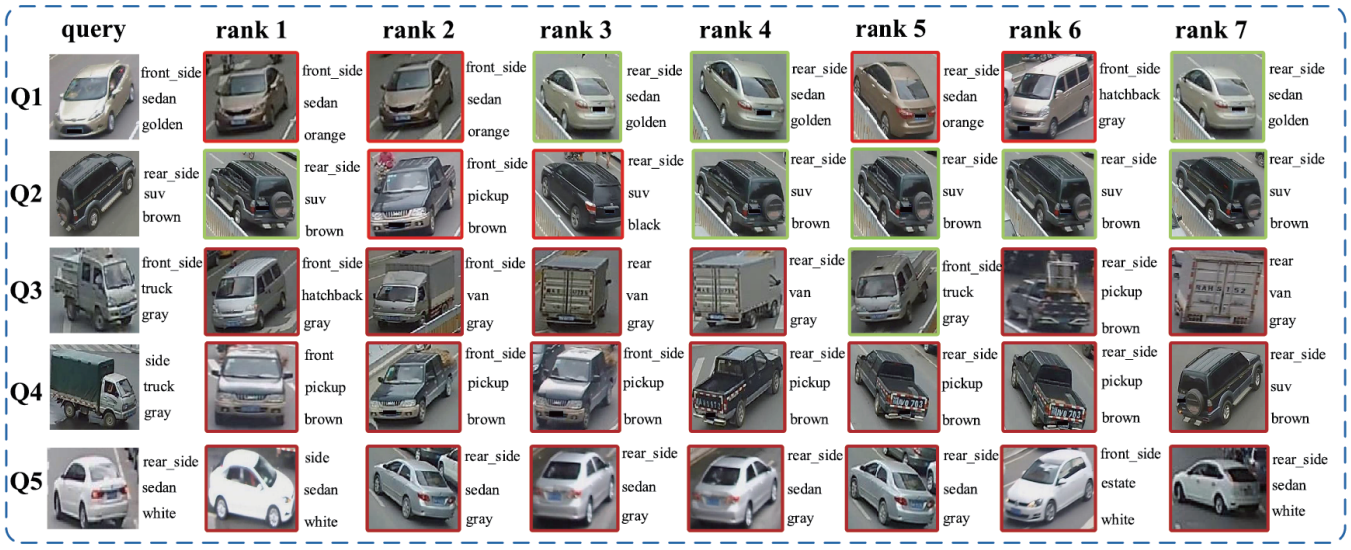

Fig. 1. Benefits of camera views, types and colors on vehicle Re-ID. The blue dash box demonstrates several ranking results of conventional vehicle Re-ID based on ResNet-50 [27], where the red and green solid boxes of the first 7 ranks indicate the wrong and right matching respectively. The results show that extra semantics or attributes play critical role in handling the challenges of vehicle Re-ID.

similar appearance. Qian et al. [30] propose a two-branch stripe-based and attribute-aware deep convolutional neural network (SAN), in which attribute information and part-level features are combined to enhance the discriminative capability for vehicle Re-ID. However, none of the methods handles both of two challenges (intra-class difference and inter-class similarity) simultaneously.

Motivated by the human visual system that recognizes the vehicle by progressively identifying the color, type with various viewpoint, we propose a unified deep convolutional framework to learn Deep Feature representations jointly guided by the meaningful attributes, including Camera Views, vehicle Types and Colors (DF-CVTC) for vehicle Re-ID. Attribute information has been successfully investigated as the mid-level semantics to boost person Re-ID. It can also help vehicle Re-ID in challenging scenarios. First of all, the camera view is one of the key attributes and challenges in Re-ID. As shown in Fig. 1, the query vehicle image may have completely different views from their counterparts under other cameras, such as query Q1 and Q2 and their right ranks marked as green solid boxes. Second, vehicle types and colors, as the representative attributes for vehicles, also play an important role in vehicle Re-ID especially for the different vehicles with a similar appearance. As shown in Fig. 1, the wrong hits of query Q3 and Q4, which present with similar appearance, could be effectively evaded by the vehicle type. Furthermore, the vehicles with different colors may present with similar shapes (such as the wrong hits rank 1 and rank 2 of query Q1), similar overall appearance (such as the wrong hits of query Q4), or even the similar color (such as the query Q5 with white color while the wrong hits of rank 2-4 with gray color) due to illumination changes. Integrating the color attribute may relieve this inter-class similarity. These challenges motivate us to utilize the above attributes to help network distinguish different vehicles with very similar appearance and also identify the same vehicles with different viewpoints.

It is worth noting that we jointly learn deep features, camera views, vehicle types and colors in an end-to-end framework. At last, we annotate the view labels in the benchmark datasets for network training, which can be directly used for other datasets with only ID labels. Comprehensive evaluations on two benchmark datasets, i.e., VeRi-776 and VehicleID, demonstrate the promising performance of the proposed method which yields to a new state-of-the-art for vehicle Re-ID.

In summary, this paper makes the following contributions to vehicle Re-ID and related applications:

- It proposes a unified attributes guided deep learning framework that jointly learns Deep Feature representations, Camera Views, vehicle Types and Colors (DF-CVTC) for vehicle Re-ID. These components are collaborative to each other, and thus boost the performance of vehicle Re-ID significantly.
- We annotate the view labels for the benchmark dataset VeRi-776 for view predictor training, which could be easily employed for the situation with only ID information available in vehicle Re-ID. We will release the annotation information of view labels to the public for free academic usage [1].

II. Related Work

A. Vehicle Re-ID

With great progress in person Re-ID [31], [32], [33], vehicle Re-ID has gradually gained a lot of attention recently since vehicles are the most important object in urban surveillance. Liu et al. [14], [2], [3] released a benchmark dataset VeRi-776 and considered the vehicle Re-ID task as a progressive recognition process by using visual features, license plates and

[1] http://www.escience.cn/people/AihuaZheng/Code-Dataset.html

spatial-temporal information. Liu et al. [2] released another big surveillance-nature dataset (VehicleID) and designed coupled clusters loss to measure the distance of two arbitrary similar vehicles. Tang et al. [17] introduced a dataset CityFlow, which is currently the largest-scale dataset in terms of spatial coverage and the number of cameras/videos in an urban environment. Lou et al. [18] collected a new dataset called VERI-Wild for vehicle Re-ID community in the wild, and designed a novel feature distance adversary scheme to online generate hard negative samples in feature space to facilitate Re-ID model training. He et al. [19] developed a new end-to-end framework to integrate part constrains with the global Re-ID modules by introducing an detection branch. Zhang et al. [16] designed an improved triplet-wise training by classification-oriented loss. Li et al. [15] integrated the identification, attribute recognition, verification and triplet tasks into a unified CNN framework. Liu et al. [34] proposed a coarse-to-fine ranking method consisting of a vehicle model classification loss, a coarse-grained ranking loss, a fine-grained ranking loss and a pairwise loss.

In addition to appearance information, Shen et al. [1] combined the visual spatio-temporal path information for regularization. Hsu et al. [4] exploited the temporal attention model to extract the most discriminant feature of each trajectory. Chen et al. [35] proposed a dedicated Semantics-guided Part Attention Network (SPAN) to robustly predict part attention masks for different views of vehicles given only image-level semantic labels during training. However, the large intra-class variation and inter-class similarity in different viewpoints have not been well studied in existing works. In this paper, we propose to embed the viewpoint as well as the attributes information into the appearance information for better discriminative feature learning.

B. View-aware Re-ID

Viewpoint changes introduce a large variation of the intra-class variation in person Re-ID. Zhao et al. [22] proposed a novel method based on human body region guided for person Re-ID which can boost the performance well. Wu et al. [36] proposed an approach called pose prior to make identification more robust to the viewpoint. Zheng [37] introduced the PoseBox structure which is generated through pose estimation followed by affine transformations. Qian et al. [38] use GAN to generate eight pre-defined pose for each image which augment the data and address the viewpoint variation to some extent. Liu et al. [6] transferred various person pose instances from one dataset to another to improve the generalization ability of the model. Zhou et al. [25] designed a conditional generative network to obtain cross-view images from input view pairs to address the vehicle Re-ID task. Later on, Zhou et al. [26] proposed a Viewpoint-aware Attentive Multi-view Inference (VAMI) model to infer multi-view features from single-view image inputs.

This issue is even crucial in vehicle Re-ID, since the viewpoint of the images are almost the same due to the rigid motion of the vehicles. Prokaj et al. [39] presented a method based on pose estimation to deal with multiple viewpoints. However, different vehicle identities might present a similar appearance while th auxiliary attributes information could help to distinguish them. Therefore, we propose to integrate several attributes information into a joint deep feature learning framework in this paper.

C. Attribute Embedded Re-ID

Attributes have been extensively investigated as the mid-level semantic information to boost the person Re-ID. Su et al. [9] introduced a low rank attribute embedding into the multi-task learning framework for person Re-ID. Khamis et al. [10] jointly optimized the attributes classification loss and triplet loss for person Re-ID. Lin et al. [40] integrated the identification loss and the attributes prediction into a simple ResNet framework and annotated the pedestrian attributes in two benchmark person Re-ID datasets Market-1501 and DukeMTMC-reID. Su et al. [41] proposed a weakly supervised multi-type attribute learning framework based on the triplet loss by pre-training the attributes predictor on independent data. Despite the previous works focusing on image-based query, Li et al [42] and Yin et al. [43] investigated attribute-based query for person retrieval and Re-ID task.

In vehicle Re-ID, Li et al. [15] introduced the attribute recognition into the vehicle Re-ID framework together with the verification loss and triplet loss. Qian et al. [30] proposed a novel two-branch stripe-based and attribute-aware deep convolutional neural network (SAN) to learn the efficient feature embedding for vehicle Re-ID task. Different from these methods, we take the view-aware identification and attributes recognition into a unified vehicle Re-ID framework.

III. DF-CVTC Network Architecture

In this paper, we propose a novel Deep Feature learning method, which embeds attributes information, including Camera views, Vehicle Types and Colors (DF-CVTC), for vehicle Re-ID. We shall elaborate the proposed method in this section.

A. Architecture Overview

The overall architecture is demonstrated in Fig. 2. To explore and make full use of the auxiliary information of the vehicle, we firstly employ a backbone to learn the shared features, followed by three subnetworks to extract the corresponding attribute weighted feature specifically. Finally, we use an embedding layer to generate the attribute embedded features for vehicle Re-ID. We shall elaborate our model in the following of this section.

B. Backbone

Due to the compelling performance with deeper layers by residual learning, ResNet-50 has been widely used in many research [44], [45]. Therefore, we adopt the first three residual blocks of ResNet-50 as the baseline network for our backbone as shown in Fig. 2. One could also configure other networks such as Inception-v4 [46], VGG16 [47] and MobileNet [48] architectures without limitation. The task of each attribute recognition shares the low-level properties which could be efficiently learned by the shared backbone.

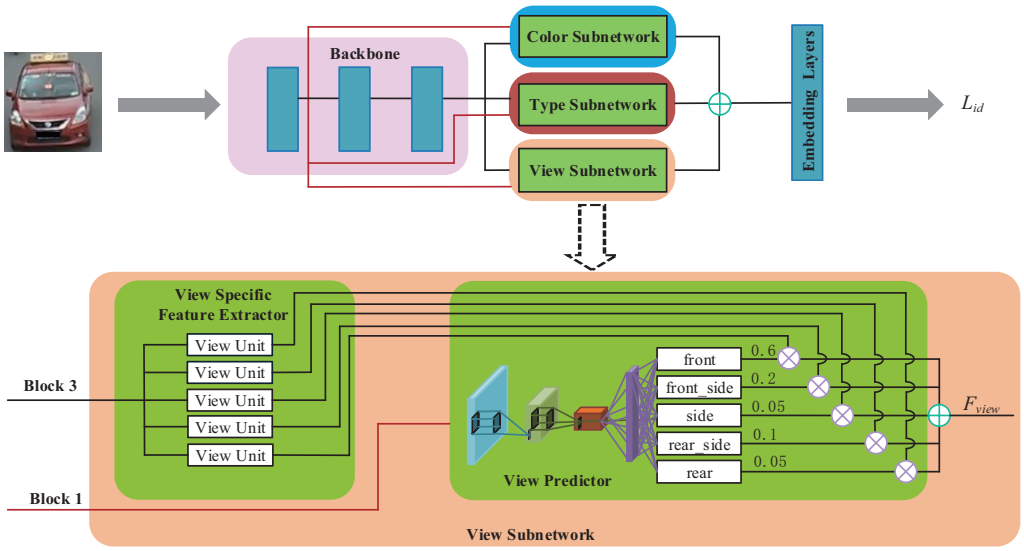

Fig. 2. Overview of our DF-CVTC, which consists of one backbone (the first three blocks of ResNet-50), three subnetworks and one embedding network.

C. Subnetworks

As shown in Fig. 2, each subnetwork consists of a predictor part and a feature extraction part. To share the low-level information and reduce the complexity of our network, we use the feature of Block-1 as the input of the following predictor part.

The predictor is composed of three convolutional (Conv) layers and one fully-connected (FC) layer which outputs a probability distribution over the corresponding (view, type or color) values. The kernel sizes in the three Conv layers are 5×5, 3×3, 5×5, respectively. The strides for these kernels are 3, 2 and 1, respectively. We use ReLU activation in all three layers and add a *batch normalization* layer after each Conv layer. The resulting feature vector is fed into the following FC layer to predict the attribute scores via the K-way softmax.

The feature extractor is composed of K units, each of which is a Conv-net responsible for extracting high-level features corresponding to one of the K view or attribute classes. We use the Block-4 of ResNet-50 as feature extractor.

The features from each specific feature extractor E_Φ can be formulated as,

$$f_{\Phi_k} = E_\Phi(x; \alpha_\Phi) \tag{1}$$

where $\Phi \in \{view, type, color\}$, $k = 1, 2, ..., K_\Phi$. K_Φ is the number of corresponding units, which also indicates the possible classes of each view or attribute. x is an image, α_Φ denotes the parameters of E_Φ.

The probability distribution w_Φ over corresponding view or attribute values from the predictor network P_Φ is,

$$w_\Phi = P_\Phi(x; \beta_\Phi) \tag{2}$$

where β_Φ denotes the parameters of P_Φ, which is learnt using the cross-entropy loss $\mathcal{L}_\Phi$,

$$\mathcal{L}_\Phi = -\sum_{k=1}^{K_\Phi} log(w_\Phi(k)) q_\Phi(k) \tag{3}$$

where q_Φ is a one-hot vector of the ground truth of corresponding view or attributes values.

After progressively learning of the three subnetworks, we achieve the specific feature maps via:

$$F_\Phi = (f_{\Phi_1} \odot w_{\Phi_1}) \oplus ... \oplus (f_{\Phi_K} \odot w_{\Phi_K}) \tag{4}$$

where $\oplus$ denotes the element-wise sum, and $\odot$ denotes the element-wise multiply. F_Φ is the augmented features by element-wise sum operation in each channel dimension, which can be beneficial for classification without extra computation. The joint deep features with camera view, type and color are achieved as the fusion of feature maps of three subnetworks,

$$F = F_{view} \oplus F_{type} \oplus F_{color} \tag{5}$$

F is the fused deep features containing the complementary view and attributes information. The feature dimensions of F_{view}, F_{type}, and F_{color} are all 2048. Next, we describe the details of each subnetwork as follows.

1) View subnetwork: Viewpoint changes bring a crucial challenge for the Re-ID task. We use the view subnetwork to incorporate the view information into the Re-ID model. The view predictor predicts K_{view}-way softmax scores which are used to weight the output of each corresponding view unit. Followed by [26], we horizontally flip each vehicle image, which means we do not distinguish left and right viewpoints. Therefore the vehicle images fall into five viewpoints in this paper: $front$, $front\_side$, $side$, $rear\_side$, and $rear$.

For instance, for the training sample in the rear orientation to the camera, the corresponding view unit will be assigned a strong weight and updated strongly during the back propagation.

2) Type subnetwork: Type is useful to distinguish the vehicles with similar appearance, which can relieve the inter-class similarity. In the same manner as the view subnetwork, we use the type subnetwork to learn the attribute specific deep features. The K_{type} scores predicted by type predictor are used to weight the output of each corresponding type unit in the type specific feature extractor. Following the protocol of type attribute information in VeRi-776 dataset [14], in this paper, we set K_{type} = 9, indicating 9 types of the vehicles: $sedan$, suv, van, $hatchback$, mpv, $pickup$, bus, $truck$ and $estate$. Similar to previous view specific feature extractor, each type unit will learn a feature map specialized for one of the K_{type} types.

3) Color subnetwork: Color is another discriminative attribute in vehicles. Therefore, we analogously use the color subnetwork to learn the color-specific features. The color predictor predicts the color scores of the vehicle then weight to each color unit. In the same manner as the color attribute is given in VeRi-776 dataset [14], in our implementation, we set K_{color} =10 denoting 10 colors of the vehicles: $yellow$, $orange$, $green$, $gray$, red, $blue$, $white$, $golden$, $brown$ and $black$. The color-specific feature extractor is designed in the same manner as in view and type subnetworks.

D. Embedding Layers

The embedding layers consist of two FC layers. It embeds the fused feature F in Eq. (5) into the higher level joint deep feature F_{joint}, which is used for the final Re-ID task.

In order to train the Re-ID model, we add a softmax layer into the embedding network for ID classification. We use the cross-entropy loss of $\mathcal{L}_{id}$ for model training,

$$\mathcal{L}_{id} = -\sum_{n=1}^{N} log(p_{id}(n))q_{id}(n) \tag{6}$$

where N is the number of the vehicle IDs in the training set. q_{id} is the one-hot ground-truth of the ID label of the vehicle. $p_{id}(n)$ is the predicted probability indicating the ID of the input vehicle image,

$$p_{id} = softmax(F_{joint}). \tag{7}$$

Fig. 3 demonstrates the effectiveness of the jointly learnt deep features of the proposed DF-CVTC. We can observe that, the vehicle images of the same identity fall into the same cluster regardless of the different visible appearance caused by different camera views (as shown in Fig. 3(a) and (b)) or illumination changes (as shown in Fig. 3 (c)).

E. Difference from Previous Work

Our method is significantly different from [25], [26], [21] from the following aspects. First, [25], [26] infer the multi-view images or features using adversarial learning. However, they render vehicle Re-ID as a verification task while our method employs a classification CNN to learn the deep features. Furthermore, our learnt features embeds attributes information (type and color) in addition to the view information. Second, [21] incorporate both fine and coarse pose/view information to learn a feature representations and propose a novel re-ranking method for person Re-ID. While our DF-CVTC further integrates the attributes information and jointly learns the deep features embedded by camera views, vehicle types and colors into an end-to-end framework.

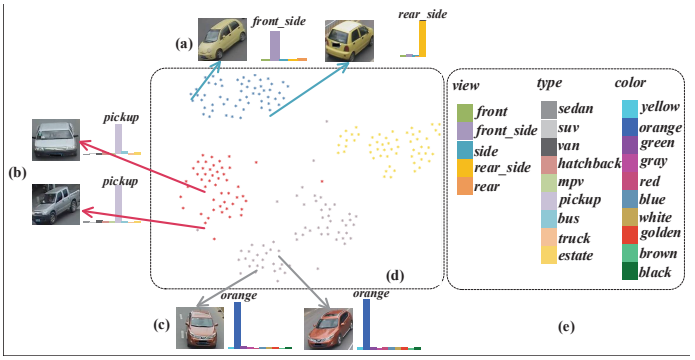

Fig. 3. Demonstration of the DF-CVTC features. (a) (b) and (c) denote three vehicle pairs sampled from VeRi-776 dataset under two distinct camera views and their corresponding learned probabilities of varying camera views, types and colors, where visible appearances are distinct due to the different camera views or illumination changes. (d) illustrates the 2D feature projections of the vehicle images learnt by the proposed DF-CVTC. (e) represents the corresponding annotation categories of camera views, types, and colors.

IV. Training Details

A. Progressive Learning

We progressively learn the three subnetworks and fine-tune the DF-CVTC model, which achieves comparative performance as the multi-task learning (minimizing the combination of the four losses). Furthermore, it can significantly reduce the computational complexity.

1) View subnetwork training: We fine-tune the backbone network pre-trained on ImageNet classification [49] and the rest of Re-ID model are initialized from scratch. First, we minimize $\mathcal{L}_{view}$ to obtain α_{view}, then we minimize $\mathcal{L}_{id}$ to obtain $\{\beta_{view}, \theta_2\}$ while fixing all the other parameters in Re-ID model.

2) Type subnetwork training: We first minimize $\mathcal{L}_{type}$ to obtain α_{type}, and then minimize $\mathcal{L}_{id}$ using $F_{view} \oplus F_{type}$ to obtain $\{\beta_{type}, \theta_2\}$ while fixing all the other parameters in Re-ID model.

3) Color subnetwork training: In the same manner, we first minimize $\mathcal{L}_{color}$ to obtain α_{color}, then minimize $\mathcal{L}_{id}$ using $F_{view} \oplus F_{type} \oplus F_{color}$ to obtain $\{\beta_{color}, \theta_2\}$ while fixing all the other parameters in Re-ID model.

4) Joint learning: After training the three subnetworks, we fine-tune $\{\alpha_\Phi, \beta_\Phi, \theta_1, \theta_2\}$, $\Phi \in \{view, type, color\}$ of the whole Re-ID model by minimizing $\mathcal{L}_{id}$ until convergence.

B. Implementation Details

In practice, we use a stochastic approximation of the objective since the training set is quite huge. The training set

TABLE I
COMPARISONS WITH STATE-OF-THE-ART RE-ID METHODS ON VERI-776 (IN %). THE TOP THREE RESULTS ARE HIGHLIGHTED IN RED, GREEN AND BLUE, RESPECTIVELY.

Method	mAP	rank 1	rank 5
(1) LOMO [13]	9.64	25.33	46.48
(2) BOW-CN [51]	12.20	33.91	53.69
(3) GoogLeNet [52]	17.89	52.32	72.17
(4) FACT [14]	18.49	50.95	73.48
(5) FACT+Plate-SNN+STR [2]	27.70	61.44	78.78
(6) Siamese-Visual [1]	29.48	41.12	60.31
(7) Siamese-CNN+Path-LSTM [1]	58.27	83.49	90.04
(8) NuFACT [3]	48.47	76.76	91.42
(9) VAMI [26]	50.13	77.03	90.82
(11) EALN [53]	57.44	84.39	94.05
(12) AAVER [54]	**58.52**	**88.68**	**94.10**
(13) VFL [55]	**59.18**	**88.08**	**94.63**
DF-CVTC	**61.06**	**91.36**	**95.77**

is stochastically divided into mini-batches with 16 samples. The network performs forward propagation on the current mini-batch, followed by the backpropagation to compute the gradients with simple cross-entropy loss for network parameters updating. We perform Adam optimizer at recommended parameters with an initial learning rate of 0.0001 and a decay of 0.96 every epoch. With more passes over the training data, the model improves until it converges. To reduce overfitting, we artificially augment the data by performing random 2D translation as the same protocol in [50]. In our implementation, all the input images are resized to $W \times H = 256 \times 256$.

V. EXPERIMENTS

We carry out a comprehensive evaluation of the proposed DF-CVTC comparing to the state-of-the-art methods on two public vehicle Re-ID datasets, VeRi-776 [14] and VehicleID [2]. We use the Cumulative Matching Characteristics (CMC) curves and mAP to evaluate our results [2]. The type and color labels are available in VeRi-776, therefore, we annotate the view labels for network training. In VehicleID, we directly employ the view, type and color subnetworks pre-trained on VeRi-776 and only ID labels are used.

A. Experiments on VeRi-776 Dataset

1) Setting: The VeRi-776 dataset contains 776 identities collected with 20 cameras in a real-world traffic surveillance environments. The whole dataset is split into 576 identities with 37,778 images for training and 200 identities with 11,579 images for testing. An additional set of 1,678 images selected from the test identities are used as query images. In order to evaluate the view subnetwork, we annotate all the vehicle images in VeRi-776 dataset into five viewpoints as $front$, $front\_side$, $side$, $rear\_side$ and $rear$. We follow the evaluation protocol in [2]. We use mean average precision (mAP) metric for evaluation. We first calculate the average precision for each query. Then, the mAP can be obtained by calculating the mean of each average precision. The cumulative match curve (CMC) metric is also used for evaluation. First, we sort the Euclidean distance between each query and gallery images in ascending order. Then, the CMC curve can be obtained by the average of sorted value. Noted that, only the vehicles in non-overlap cameras are counted during evaluation.

2) Qualitative examples: Fig. 4 demonstrates the qualitative examples of six ranking results of our DF-CVTC on VeRi-776 dataset. From which we can observe that our method successfully hits the vehicles with large view variations to the query such as rank 2 and ranks 10-11 in Fig. 4 (a), rank 4 and ranks 8-9 in Fig. 4 (b), rank 8 in Fig. 4 (c), rank 1 and rank 5 in Fig. 4 (d). The wrong hits generally result from the high inter-class similarity with homologous visual appearance, such as ranks 4-5, rank 9 and rank 12 in Fig. 4 (a), rank 10 in Fig. 4 (c), ranks 7-9 in Fig. 4 (e), ranks 1-3 and rank 5 in Fig. 4 (f). Fig. 6 demonstrates the qualitative examples of ranking result of our DF-CVTC on VeRi-776 dataset. Fig. 6 (b) shows the view/attributes probability which is predicted by each subnetwork. Fig. 6 (c) show the ranking result. From the Fig. 6, we can find that the ranking result is improving by introduction each subnetwork progressively.

From the above observation, we can conclude that our method can reduce the influence of viewpoint changes to some extends.

3) Quantitative results: Table I reports the performance of our approach comparing with the published state-of-the-arts on VeRi-776 dataset. From which we can see, our DF-CVTC significantly surpasses the state-of-the-art. Compared with the second best method VFL [55], our method achieves 1.88% and 3.28% improvements in terms of mAP and rank 1 respectively. Note that we haven't utilized any license plates or spatial temporal information as in Siamese-CNN+Path-LSTM [1] and FACT+Plate-SNN+STR [2]. Even though, our method still achieves the superior mAP and ranking accuracies by a large margin. Such a huge improvement verifies the effectiveness of our proposed method for the vehicle Re-ID.

B. Experiments on VehicleID Dataset

1) Setting: The VehicleID dataset [2] consists of the training set with 110,178 images of 13,134 vehicles and the test set with 111,585 images of 13,133 vehicles. Followed by the protocol in [2], we test VehicleID dataset in three distinct settings with different number of testing samples: 800, 1600 and 2400. Specifically, since some of the type and color information is missing and no view labels in this dataset, we adopt the view, type and color subnetworks pre-trained on VeRi-776 dataset and fine-tuned during the Re-ID training. Which in turn means one can easily apply our model on the dataset with only ID information. The mean average precision (mAP), cumulative match curve (CMC) are used as the evaluation metric in the same manner as in VeRi-776. The only difference is we randomly select a image from test dataset as gallery, while consider the remaining images in test dataset as query. The experimental results are based on the average of 10 random trials.

2) Qualitative examples: Fig. 5 demonstrates six ranking results of our DF-CVTC on VehicleID. From which we can observe that, our method can successfully hit the right matching with large inter-class difference caused by the illumination/color changes, such as Fig. 5 (a), (c) and (f), as well

Fig. 4. Examples of ranking results on VeRi-776 dataset. The green and red boxes indicate the right matchings and the wrong matchings respectively.

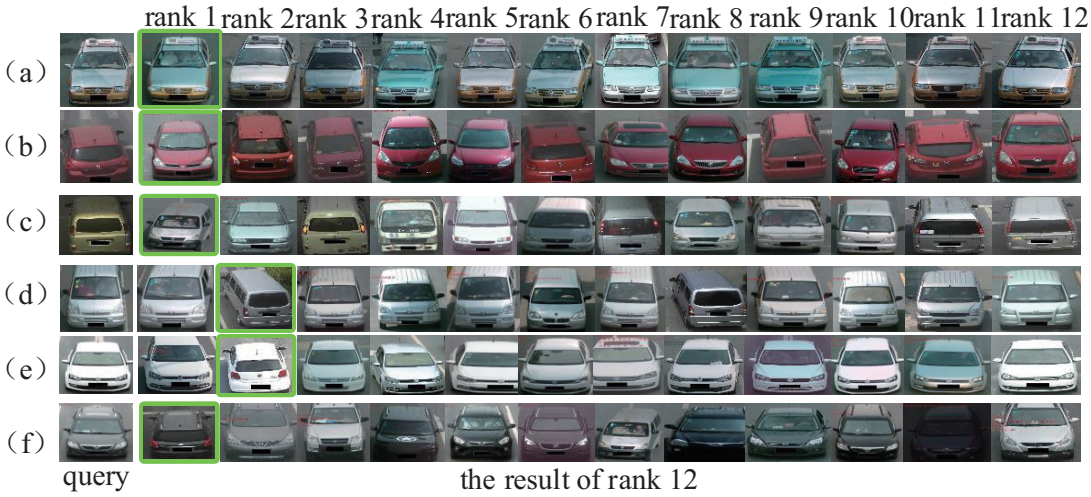

Fig. 5. Examples of ranking results on VehicleID dataset. The green boxes indicate the right matchings. Note that there is only one ground truth vehicle image in gallery set in the VehicleID dataset.

as the viewpoint changes, such as Fig. 5 (b), (c), (d) and (f). The wrong hits of rank 1 on Fig. 5 (d) and (e) result from the inter-class similarity between vehicles, despite of which, our method still hit the right matchings in the early ranks. Note that there is only one ground truth vehicle image in gallery set in the VehicleID dataset.

We can easily draw the conclusion that our method tends to hit the vehicles with the same type and color despite the change of viewpoints.

3) Quantitative results: Table II reports the performance of our method against the state-of-the-arts on VehicleID dataset. Clearly, our method significantly beats the existing state-of-the-arts in mAP, rank 1 and rank 5. Based on above results, we can conclude that our method achieves significant improvement for vehicle Re-ID.

C. Ablation Study

1) Analysis on subnetworks: Table III reports the effective of a different component on VehicleID and VeRi-776 dataset. Obviously, by introducing the view, type and color subnetworks progressively, the performance of our method is consistently improved. Compare to the base model of ResNet-50, our DF-CVTC make great progress in all metric. In specific, we increase the rank1 by 7.48%, 6.36% and 7.01% in

TABLE II
COMPARISONS WITH STATE-OF-THE-ART RE-ID METHODS ON VEHICLEID (IN %). THE TOP THREE RESULTS ARE HIGHLIGHTED IN RED, GREEN AND BLUE, RESPECTIVELY.

Method	*Test Size = 800*			*Test Size = 1600*			*Test Size = 2400*		
	mAP	rank 1	rank 5	mAP	rank 1	rank 5	mAP	rank 1	rank 5
(1) LOMO [13]	-	19.76	32.01	-	18.85	29.18	-	15.32	25.29
(2) BOW-CN [51]	-	13.14	22.69	-	12.94	21.09	-	10.20	17.89
(3) GoogLeNet [52]	46.20	47.88	67.18	44.00	43.40	63.86	38.10	38.27	59.39
(4) FACT [14]	-	49.53	68.07	-	44.59	64.57	-	39.92	60.32
(8) NuFACT [3]	-	48.90	69.51	-	43.64	65.34	-	38.63	60.72
(9) VAMI [26]	-	63.12	83.25	-	52.87	75.12	-	47.34	70.29
(10) C2F-Rank [34]	**63.50**	61.10	81.70	**60.00**	56.20	76.20	**53.00**	51.40	72.20
(11) EALN [53]	**77.50**	**75.11**	**88.09**	**74.20**	**71.78**	**83.94**	**71.00**	**69.30**	**81.42**
(13) VFL [55]	-	**73.37**	**85.52**	-	**69.52**	**81.00**	-	**67.41**	**78.48**
DF-CVTC	**78.03**	**75.23**	**88.11**	**74.87**	**72.15**	**84.37**	**73.15**	**70.46**	**82.13**

TABLE III
INFLUENCE OF DIFFERENT COMPONENT ON VEHICLEID AND VERI-776 DATASET (IN %). THE TOP THREE RESULTS ARE HIGHLIGHTED IN RED, GREEN AND BLUE, RESPECTIVELY.

Method	**VehicleID**									**VeRi-776**		
	Test Size = 800			*Test Size = 1600*			*Test Size = 2400*					
	mAP	rank 1	rank 5	mAP	rank 1	rank 5	mAP	rank 1	rank 5	mAP	rank 1	rank 5
ResNet-50 (baseline)	70.50	67.75	79.13	68.48	65.79	76.64	66.19	63.45	74.70	51.58	86.71	92.43
+view	75.44	72.63	84.82	72.41	69.62	81.36	70.71	68.02	79.19	54.52	89.69	94.40
+view+type	76.06	73.14	86.25	73.39	70.77	81.75	71.75	69.10	80.40	60.47	91.66	95.59
+view+type+color (DF-CVTC)	78.03	75.23	88.11	74.87	72.15	84.37	73.15	70.46	82.13	61.06	91.36	95.77

TABLE IV
ABLATION STUDY ON DIFFERENT BACKBONES WITH VARYING COMPONENTS ON VERI-776 DATASET (IN %). THE TOP THREE RESULTS ARE HIGHLIGHTED IN RED, GREEN AND BLUE, RESPECTIVELY.

Component	(a)			(b)			(c)			(d)		
view	×			✓			✓			✓		
type	×			×			✓			✓		
color	×			×			×			✓		
Backbone	mAP	rank 1	rank 5	mAP	rank 1	rank 5	mAP	rank 1	rank 5	mAP	rank 1	rank 5
VGG16 [47]	42.35	77.77	88.14	44.17	80.63	89.57	45.43	81.17	90.35	45.62	81.76	91.12
MobileNet [48]	52.55	86.23	94.10	54.48	87.60	93.92	58.49	89.15	94.64	59.23	89.45	94.87
Inception-v4 [46]	49.78	84.62	91.90	52.74	87.66	93.68	59.49	89.27	94.76	60.50	89.51	95.47

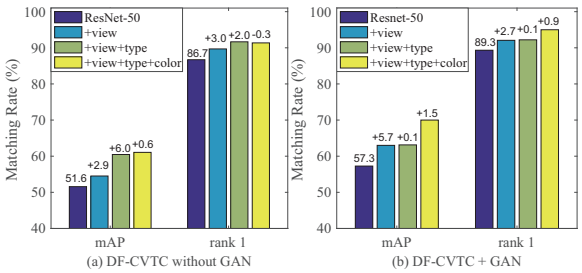

Fig. 7. Performance of GAN on VeRi-776 dataset. (a) and (b) demonstrate the mAP and rank 1 scores of the proposed DF-CVTC and its variants without and with GAN respectively. The digits on the top of the last three bars on each metric indicate the degree of improvement by progressively introducing view, type and color, comparing to the first blue bar of the baseline ResNet-50.

three difference scaler test sets respectively in VehicleID. And increase the rank 1 by 4.65%, mAP by 9.46% respectively in VeRi-776.

Fig. 6 demonstrates an example of ranking results of the proposed DF-CVTC for a query from VeRi-776 dataset by progressively introducing the view, type and color subnetworks into the ResNet-50 backbone. We observe that: 1) By introducing the view subnetwork, it can eliminate the wrong ranks with quite similar visible appearance especially with similar views to the query especially, such as rank 1 and rank 2 in Fig. 6 (c1). 2) By further introducing the type subnetwork, it can eliminate the wrong ranks with obviously distinct types, such as rank 2 and rank 9 in Fig. 6 (c2). 3) Our full model DF-CVTC (Fig. 6 (c4)) hits the rightest ranks by progressively introduce the three subnetworks.

Fig. 6 further visualizes the feature maps via CAM (Class Activation Mapping) [56] to demonstrate the attended area by progressively introducing the attribute subnetworks. From which we can see, by introducing the view subnetwork, our method tends to emphasize the common area with different views from the query, such as the right hit rank 1 in as shown in Fig. 6 (c2). Furthermore, progressively introducing the type subnetwork leads to higher attention on the areas reflecting the vehicle type information, comparing Fig. 6 (c3) to Fig. 6 (c2). Similarly, introducing the color subnetwork leads to higher

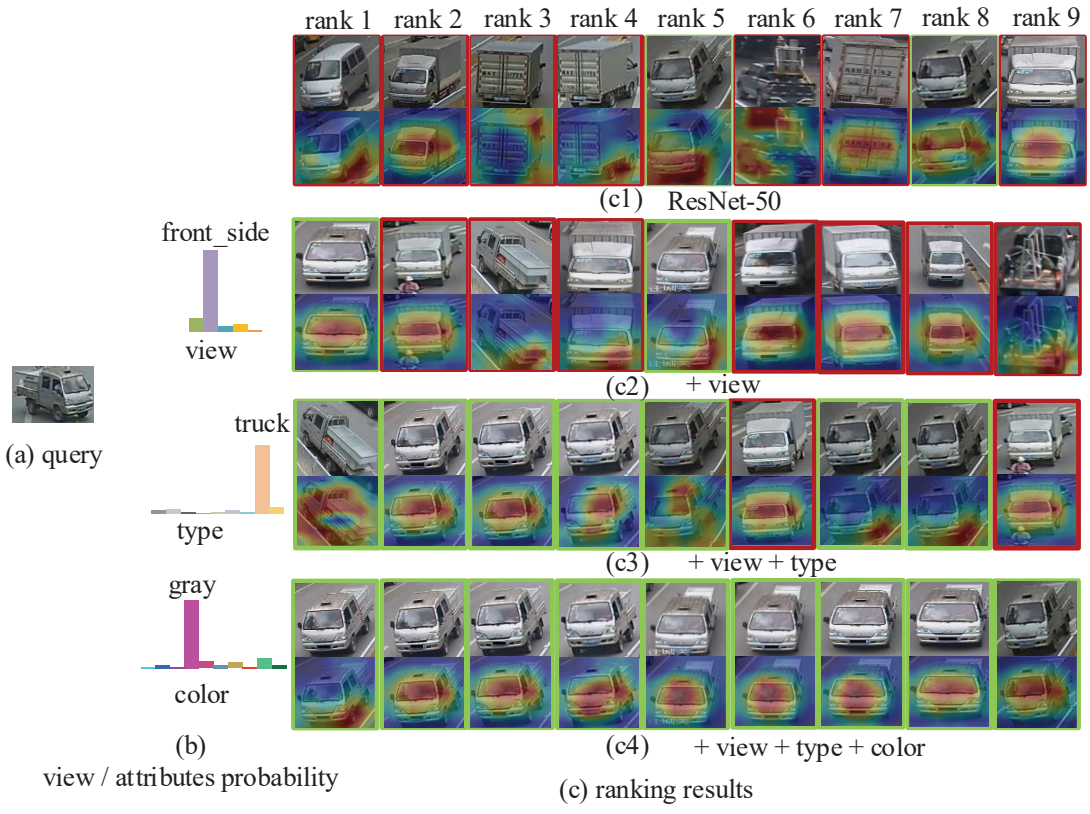

Fig. 6. An example of ranking results together with corresponding CAM visualizations of DF-CVTC on ResNet-50 backbone by progressively introducing the view, type and color subnetworks on VeRi-776 dataset. The green and red boxes indicate the right matchings and the wrong matchings respectively. The histograms denote the probability distributions learnt from the view, type and color subnetworks respectively.

attention on discriminative color regions, as shown in Fig. 6 (c4).

2) Analysis on backbones: As we mentioned in Section III-B, any other CNN architecture could be used in our framework instead of ResNet-50 without any limitation. We further evaluate three prevalent CNN architectures, Inception-v4 [46], VGG16 [47] and MobileNet [48] as the backbone respectively while remaining the other part of the proposed model unchanged. The results on VeRi-776 dataset are reported in Table IV. From which we can see, all the three CNN counterparts achieve satisfactory performance. Specifically, Inception-v4 and MobileNet achieve competitive performance on all the metrics. VGG16 works slight worse than the other two architecture, but it is still competitive to the state-of-the-art methods, which demonstrates that the high performance of the proposed model is not totally due to the superiority of the ResNet-50. Furthermore, by progressively introducing the view, type and color subnetworks, the performance of the corresponding variants based on all the backbones consistently improves, which verifies the contribution of the proposed jointly learning model.

3) Analysis on Attribute Predictors: Table V reports the result of our method while fixing the weighs of three attributes. Specifically, we fix the weights of the view, type and color subnetworks as 1/5, 1/9 and 1/10, respectively. After progressively training the model, we can find that with the introduction of the three subnetworks, there is no significant improvement. The proposed DF-CVTC is even inferior than the baseline, which implies the importance of the adaptive weights learning during the proposed method.

TABLE V
EVALUATION ON OUR METHOD WHILE FIXING THE WEIGHTS OF THREE ATTRIBUTES.

	VeRi-776		
Method	mAP	rank 1	rank 5
ResNet-50 (baseline)	51.58	86.71	92.43
+view	51.13	87.66	94.82
+view+type	52.02	86.47	93.50
+view+type+color (DF-CVTC)	50.09	84.33	92.31

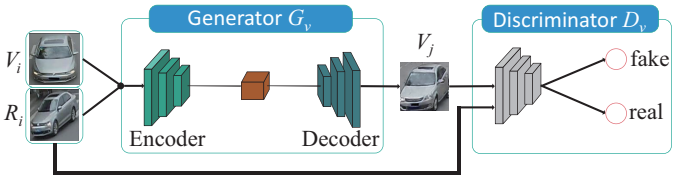

Fig. 8. The architecture of GAN based on the architecture of pix2pix [57]. For the input single view vehicle image V_i ($front$ view as shown), it aims to synthesize a vehicle image V_j with the same view as the target vehicle image R_i ($front\_side$ view as shown).

D. Data Augmentation

As we observed, most of the vehicle images under a certain camera are almost in the same viewpoint due to the rigid

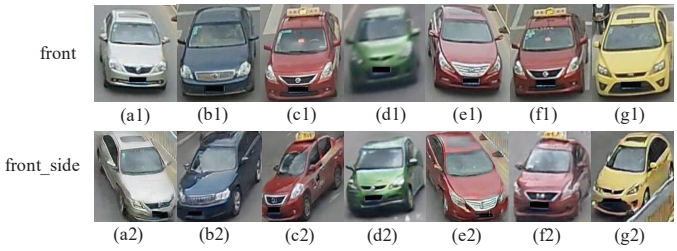

Fig. 9. Examples of synthesizing the $front$ view vehicle images to $front\_side$ view on VeRi-776 dataset via GAN. The first and the second rows indicate the vehicle images with the original $front$ view and the synthesized $front\_side$ view respectively.

motion of vehicles, and thus the number of vehicle images with different views is very limited which brings a big challenge to train deep networks. To handle this issue, we design a generative adversarial network (GAN) to generate the multi-view vehicle images. In this paper, we simply employ pix2pix [45] for its generality. The generation architecture is illustrated as Fig. 8. In a specific, given an input vehicle image V_i and a target vehicle image R_i with a slightly different view the same ID, our GAN aims to generate a new vehicle image V_j with the same view as R_i. GAN constitutes a Generator G_v learning a map conditional on the given target, and a Discriminator D_v discriminating real data samples from the generated samples, such that the distribution of image V_j is indistinguishable from the distribution image V_i. The loss function can be expressed as,

$$\mathcal{L}(G_v, D_v) = \mathbb{E}_{V_i,R_i}[logD_v(V_i, R_i)] + \mathbb{E}_{V_i,R_i}[log(1- D_v(V_i, G_v(V_i, R_i))] + \lambda\mathbb{E}_{V_i,R_i}[\|R_i - G_v(V_i, R_i)\|_1] \quad (8)$$

where G_v tries to minimize this objective against an adversarial D_v that tries to maximize it, ℓ_1 distance is used to encourage less blurring. λ is the weighting coefficient. Fig. 9 demonstrates several examples of synthesizing the $front$ view vehicle images to $front\_side$ view on VeRi-776 dataset via GAN. One more thing we would like to mention is the pair of the input images of our pix2pix GAN is from the same ID but slightly different viewpoint.

Due to the computational complexity, we have simply transferred 1400 front view vehicles into the front side view images for training data augmentation on the VeRi-776 dataset as shown in Fig. 9. Fig. 7 demonstrates the performance of GAN. From which we can see, by augmenting even only 1400 synthetic multi-view images into a total of 37729 training samples, it can benefit the Re-ID model with various components. Moreover, it can further boost the contribution of the view subnetwork by improving 5.7% and 2.7% in mAP and rank 1 respectively, comparing to 2.9% and 3.0% improvements of DF-CVTC without GAN. We believe that more generated images with more viewpoints will further boost the performance.

VI. Conclusion

In this paper, we have proposed a novel end-to-end deep convolutional network to jointly learn deep features, camera views, types and colors for vehicle Re-ID. We expand the backbone of ResNet-50 with three consolidated subnetworks incorporating the view, type and color cues respectively. These three tasks benefit each other and learn an informative discriminative representation for vehicle Re-ID. Furthermore, we have increased the diversity of the views for vehicle images via a generative adversarial network. By jointly learning the deep features, camera views, vehicle types and vehicle colors in a single unified framework, our method can achieve superior performance comparing to the state-of-the-art methods. Comprehensive evaluation on two benchmark datasets demonstrates the clear contribution of each subnetwork and the capability of informative representation for vehicle Re-ID.

References

[1] 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 (ICCV)*, 2017, pp. 1918–1927.

[2] X. Liu, W. Liu, T. Mei, and H. Ma, "A deep learning-based approach to progressive vehicle re-identification for urban surveillance," in *European Conference on Computer Vision (ECCV)*, 2016, pp. 869–884.

[3] X. Liu, W. Liu, T. Mei, and H.-D. Ma, "Provid: Progressive and multimodal vehicle reidentification for large-scale urban surveillance," *IEEE Transactions on Multimedia (TMM)*, pp. 645–658, 2018.

[4] H.-M. Hsu, T.-W. Huang, G. Wang, J. Cai, Z. Lei, and J.-N. Hwang, "Multi-camera tracking of vehicles based on deep features re-id and trajectory-based camera link models," in *AI City Challenge Workshop, IEEE/CVF Computer Vision and Pattern Recognition (CVPR) Conference, Long Beach, California*, 2019.

[5] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by gan improve the person re-identification baseline in vitro," in *IEEE International Conference on Computer Vision (CVPR)*, 2017, pp. 3754–3762.

[6] J. Liu, B. Ni, Y. Yan, P. Zhou, S. Cheng, and J. Hu, "Pose transferrable person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 4099–4108.

[7] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer gan to bridge domain gap for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 79–88.

[8] P. Chen, X. Xu, and C. Deng, "Deep view-aware metric learning for person re-identification." in *the International Joint Conference on Artificial Intelligence (IJCAI)*, 2018, pp. 620–626.

[9] C. Su, F. Yang, S. Zhang, Q. Tian, L. S. Davis, and W. Gao, "Multi-task learning with low rank attribute embedding for person re-identification," in *IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 3739–3747.

[10] S. Khamis, C.-H. Kuo, V. K. Singh, V. D. Shet, and L. S. Davis, "Joint learning for attribute-consistent person re-identification," in *European Conference on Computer Vision (ECCV)*, 2014, pp. 134–146.

[11] H. Fan, L. Zheng, C. Yan, and Y. Yang, "Unsupervised person re-identification: Clustering and fine-tuning," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 14, no. 4, pp. 1–18, 2018.

[12] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3908–3916.

[13] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 2197–2206.

[14] 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 (ICME)*, 2016, pp. 1–6.

[15] 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 (ICIP)*, 2017, pp. 395–399.

[16] 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 (ICME)*, 2017, pp. 1386–1391.

[17] Z. Tang, M. Naphade, M.-Y. Liu, X. Yang, S. Birchfield, S. Wang, R. Kumar, D. C. Anastasiu, and J.-N. Hwang, "Cityflow: A city-scale benchmark for multi-target multi-camera vehicle tracking and re-identification," in *CVPR 2019: IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[18] Y. Lou, Y. Bai, J. Liu, S. Wang, and L. Duan, "Veri-wild: A large dataset and a new method for vehicle re-identification in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3235–3243.

[19] B. He, J. Li, Y. Zhao, and Y. Tian, "Part-regularized near-duplicate vehicle re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3997–4005.

[20] Z. Tang, M. Naphade, S. Birchfield, J. Tremblay, W. Hodge, R. Kumar, S. Wang, and X. Yang, "Pamtri: Pose-aware multi-task learning for vehicle re-identification using highly randomized synthetic data," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 211–220.

[21] M. Saquib Sarfraz, A. Schumann, A. Eberle, and R. Stiefelhagen, "A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 420–429.

[22] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, and X. Tang, "Spindle net: Person re-identification with human body region guided feature decomposition and fusion," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1077–1085.

[23] L. Zhao, X. Li, Y. Zhuang, and J. Wang, "Deeply-learned part-aligned representations for person re-identification." in *IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 3239–3248.

[24] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Pose-driven deep convolutional model for person re-identification," in *IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 3980–3989.

[25] Y. Zhouy and L. Shao, "Cross-view gan based vehicle generation for re-identification," in *British Machine Vision Conference (BMVC)*, 2017, pp. 1–12.

[26] Y. Zhou and L. Shao, "Aware attentive multi-view inference for vehicle re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 6489–6498.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[28] C. Su, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Deep attributes driven multi-camera person re-identification," in *European conference on computer vision (ECCV)*, 2016, pp. 475–491.

[29] R. Layne, T. M. Hospedales, S. Gong, and Q. Mary, "Person re-identification by attributes." in *The British Machine Vision Conference (BMVC)*, 2012, p. 8.

[30] J. Qian, W. Jiang, H. Luo, and H. Yu, "Stripe-based and attribute-aware network: A two-branch deep model for vehicle re-identification," *Measurement Science and Technology*, vol. 31, no. 9, p. 095401, 2020.

[31] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang, "Camera style adaptation for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 5157–5166.

[32] X. Zhu, B. Wu, D. Huang, and W.-S. Zheng, "Fast open-world person re-identification," *IEEE Transactions on Image Processing (TIP)*, pp. 2286–2300, 2018.

[33] X. Wang, S. Zheng, R. Yang, B. Luo, and J. Tang, "Pedestrian attribute recognition: A survey," *arXiv preprint arXiv:1901.07474*, 2019.

[34] G. Haiyun, Z. Chaoyang, L. Zhiwei, W. Jinqiao, L. Hanqing *et al.*, "Learning coarse-to-fine structured feature embedding for vehicle re-identification," in *Association for the Advancement of Artificial Intelligence (AAAI)*, 2018, pp. 1–8.

[35] T.-S. Chen, C.-T. Liu, C.-W. Wu, and S.-Y. Chien, "Orientation-aware vehicle re-identification with semantics-guided part attention network," in *European Conference on Computer Vision*, 2020, pp. 330–346.

[36] Z. Wu, Y. Li, and R. J. Radke, "Viewpoint invariant human re-identification in camera networks using pose priors and subject-discriminative features," *IEEE transactions on pattern analysis and machine intelligence (TPAMI)*, pp. 1095–1108, 2015.

[37] L. Zheng, Y. Huang, H. Lu, and Y. Yang, "Pose invariant embedding for deep person re-identification," *IEEE Transactions on Image Processing (TIP)*, 2019.

[38] X. Qian, Y. Fu, T. Xiang, W. Wang, J. Qiu, Y. Wu, Y.-G. Jiang, and X. Xue, "Pose-normalized image generation for person re-identification," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 650–667.

[39] J. Prokaj and G. Medioni, "3-d model based vehicle recognition," in *Applications of Computer Vision (WACV), 2009 Workshop on*, 2009, pp. 1–7.

[40] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, Z. Hu, C. Yan, and Y. Yang, "Improving person re-identification by attribute and identity learning," *Pattern Recognition*, vol. 95, pp. 151–161, 2019.

[41] C. Su, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Multi-type attributes driven multi-camera person re-identification," *Pattern Recognition*, pp. 77–89, 2018.

[42] S. Li, T. Xiao, H. Li, B. Zhou, D. Yue, and X. Wang, "Person search with natural language description," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1970–1979.

[43] Z. Yin, W.-S. Zheng, A. Wu, H.-X. Yu, H. Wan, X. Guo, F. Huang, and J. Lai, "Adversarial attribute-image person re-identification." in *International Joint Conference on Artificial Intelligence (IJCAI)*, 2018, pp. 1100–1106.

[44] X. Liu and Z. Deng, "Segmentation of drivable road using deep fully convolutional residual network with pyramid pooling," *Cognitive Computation*, vol. 10, no. 2, pp. 272–281, 2018.

[45] G. Ding, M. Chen, S. Zhao, H. Chen, J. Han, and Q. Liu, "Neural image caption generation with weighted training and reference," *Cognitive Computation*, vol. 11, no. 6, pp. 763–777, 2019.

[46] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[47] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015.

[48] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[49] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.

[50] W. Li, R. Zhao, T. Xiao, and X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 152–159.

[51] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1116–1124.

[52] L. Yang, P. Luo, C. Change Loy, and X. Tang, "A large-scale car dataset for fine-grained categorization and verification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3973–3981.

[53] Y. Lou, Y. Bai, J. Liu, S. Wang, and L.-Y. Duan, "Embedding adversarial learning for vehicle re-identification," *IEEE Transactions on Image Processing*, vol. 28, no. 8, pp. 3794–3807, 2019.

[54] P. Khorramshahi, A. Kumar, N. Peri, S. S. Rambhatla, J.-C. Chen, and R. Chellappa, "A dual-path model with adaptive attention for vehicle re-identification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 6132–6141.

[55] S. A. S. Alfasly, Y. Hu, T. Liang, X. Jin, Q. Zhao, and B. Liu, "Variational representation learning for vehicle re-identificati," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 3118–3122.

[56] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.

[57] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1125–1134.