Multi-Query Vehicle Re-Identification: Viewpoint-Conditioned Network, Unified Dataset and New Metric

Aihua Zheng, Chaobin Zhang, Chenglong Li, Jin Tang, and Chang Tan

***Abstract*— Existing vehicle re-identification methods mainly rely on the single query, which has limited information for vehicle representation and thus significantly hinders the performance of vehicle Re-ID in complicated surveillance networks. In this paper, we propose a more realistic and easily accessible task, called multi-query vehicle Re-ID, which leverages multiple queries to overcome viewpoint limitation of single one. Based on this task, we make three major contributions. First, we design *a novel viewpoint-conditioned network* (VCNet), which adaptively combines the complementary information from different vehicle viewpoints, for multi-query vehicle Re-ID. Moreover, to deal with the problem of missing vehicle viewpoints, we propose a cross-view feature recovery module which recovers the features of the missing viewpoints by learnt the correlation between the features of available and missing viewpoints. Second, we create *a unified benchmark dataset*, taken by 6142 cameras from a real-life transportation surveillance system, with comprehensive viewpoints and large number of crossed scenes of each vehicle for multi-query vehicle Re-ID evaluation. Finally, we design *a new evaluation metric*, called mean cross-scene precision (mCSP), which measures the ability of cross-scene recognition by suppressing the positive samples with similar viewpoints from the same camera. Comprehensive experiments validate the superiority of the proposed method against other methods, as well as the effectiveness of the designed metric in the evaluation of multi-query vehicle Re-ID. The codes and dataset are available at: https://github.com/zhangchaobin001/VCNet**

***Index Terms*— Vehicle Re-ID, multiple queries, benchmark dataset, mean cross-scene precision, viewpoint-conditioned learning.**

Manuscript received 14 September 2022; revised 10 July 2023 and 26 August 2023; accepted 7 October 2023. Date of publication 27 October 2023; date of current version 2 November 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 62372003 and Grant 61976002, in part by the Natural Science Foundation of Anhui Province under Grant 2308085Y40 and Grant 2208085J18, in part by the University Synergy Innovation Program of Anhui Province under Grant GXXT-2022-036, and in part by the Natural Science Foundation of Anhui Higher Education Institution under Grant 2022AH040014. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Zhu Li. *(Corresponding author: Chenglong Li.)*

Aihua Zheng and Chenglong Li are with the 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 (e-mail: ahzheng214@foxmail.com; lcl1314@foxmail.com).

Chaobin Zhang and Jin Tang are with the Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, School of Computer Science and Technology, Anhui University, Hefei 230601, China (e-mail: chaobinzhang@foxmail.com; tangjin@ahu.edu.cn).

Chang Tan is with iFLYTEK Company Ltd., Hefei 230088, China (e-mail: changtan2@iflytek.com).

Digital Object Identifier 10.1109/TIP.2023.3326691

I. Introduction

VEHICLE re-identification (Re-ID) aims to correlate the images of the same vehicle captured by non-overlapping cameras. This task has been widely applied in urban security monitoring and intelligent transportation systems, and has received more and more attention in recent years [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14].

Most existing image-based vehicle Re-ID methods [16], [17], [18], [19], [20], [21] rely on the single query. However, the dramatic appearance changes caused by different viewpoints lead to the huge intra-class discrepancy.

To solve the viewpoint diversity, some works use vehicle keypoint information [22], [23] or vehicle local area features [24], [25], [26] to perform local feature alignment. Moreover, meta-information (e.g. vehicle attributes, spatial-temporal information) has also been explored to alleviate the difference from different viewpoints. Li et al. [27] introduce attribute fusion and Liu et al. [28] utilize attributes and spatial-temporal information to learn global vehicle representations. However, the key issue in this single-shot fashion is that one vehicle image only has a specific viewpoint as the single query, thus it is very challenging to match the gallery images with different viewpoints. To explore the comprehensive information in multi-shot images, Jin et al. [29] propose a multi-shot teacher branch that mines the information of multi-viewpoint images to guide the single-image student branch during training. However, they cannot guarantee that the teacher network contains information from multiple viewpoints, and they only use one query image in the inference phase, which still can not fully utilize the information from multiple query images. Zheng et al. [30] evaluate person Re-ID in the so-called multi-query fashion by average or max operations on the multiple person images from the same camera to obtain a new query feature in the inference process. Zhou et al. [31] propose to utilize the fused multiple features for person re-identification. However, most of the existing "multi-query" person Re-ID mainly refers to using additional queries from the same tracking list of a pedestrian to form multiple queries. Therefore the variability between query images in the same camera is small, and the average or max operations cannot fully utilize the diversity and complementarity among the multiple queries.

In fact, we can easily access multiple images of a certain vehicle from diverse viewpoints or scenes as the query in

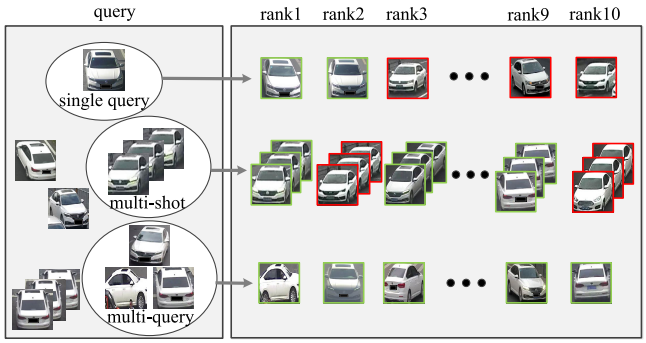


Fig. 1. Examples of ranking results comparison among the conventional single query, multi-shot and the proposed multi-query ReID on our collected MuRI dataset via ResNet-50 [15], where multi-shot (or multi-query) ReID is achieved by averaging the consecutive (or the different viewpoints) vehicle image features. The true and false matchings are bounded in green and red boxes, respectively.

real-life surveillance. On the one hand, in a certain scene, we can easily obtain multi-view images of the same vehicle via the crowded dome and box cameras or robust tracking algorithms. On the other hand, we can obtain the cross-scene multi-view images of the same vehicle by correlating the corresponding cross-scene tracklets. In addition to the above intelligent acquisition, we can also manually construct the multi-view query images in the surveillance system.

By contrast, we rethink vehicle Re-ID in the more realistic multi-query inference setting in this paper, which is expected to significantly overcome the dramatic appearance changes caused by different viewpoints. However, how to effectively integrate the complementary information among the multi-query images scenario (or with random viewpoint missing), thus to learn more comprehensive appearance feature representation of a certain vehicle presents great research potential.

As shown in Fig. 1, due to the limited view information, single query and multi-shot Re-ID tend to easily matching the vehicle images with similar viewpoints. By contrast, multi-query Re-ID can hit the more challenging right matchings with diverse viewpoints since it can integrate the complementary information among the diverse viewpoint queries. Giving the easily accessible multiple query images captured from a single or several non-overlapping scenes/cameras, how to take the advantage of multiple queries with diverse viewpoint and illumination changes to achieve more accurate vehicle Re-ID? **In this paper, we propose a novel viewpoint-conditioned network (VCNet), which effectively combines the complementary information from different vehicle viewpoints, for multi-query vehicle Re-ID.** First, in the training process, to make full use of diverse viewpoint information of the vehicle, we propose a viewpoint conditional coding (VCC) module, which uses the learned vehicle viewpoint features as viewpoint conditional coding information and integrates it into the feature learning process of vehicle appearance. Second, we propose the viewpoint-based adaptive fusion (VAF) module to adaptively fuse the viewpoint coded appearance features of the vehicle in the multi-query inference process. In particular, it adaptively assigns weights to the appearance features of the multiple queries according to their viewpoint similarity to gallery. The higher viewpoint similarity between the query image to the current gallery image, the larger weight to the appearance feature of the corresponding query image. **Finally, to tolerate the missing viewpoints in the query set in the inference, we propose a cross-view feature recovery module (CVFR) to recover the features of the missing viewpoints.** The cross-view feature recovery module learns the consistency of multiple viewpoints by maximizing the common information of different viewpoint features, and achieves the recoverability of cross-view features by minimizing the conditional entropy between different viewpoint features.

In addition, although conventional vehicle Re-ID metrics (namely CMC, mAP and mINP) avoid the easy matching from the same camera between query and gallery, they mainly focus on the global relation between the query and gallery while ignoring the local relations within the gallery. Therefore, they tend to result in virtual high scores when retrieving easy positive samples with similar viewpoints from one single camera. Although Zhao et al. [32] propose the evaluation metric Cross-camera Generalization Measure (CGM) to improve the evaluations by introducing position-sensitivity and cross-camera generalization penalties. It still suffers from the influence of similar viewpoint samples under the same camera. Since they only easily divide target images captured from the same cameras into individual groups, and fail to consider the positive samples with similar viewpoints from the individual group. In this paper, we argue that the realistic Re-ID cares more about the cross-scene retrieval ability of the model, which is more crucial to the intelligent transportation society to trace the trajectory of the certain vehicle among the identity of the vast Skynet in the smart city. **Therefore, we propose a new metric, the mean Cross-scene Precision (mCSP), which focuses on the cross-scene retrieval ability by suppressing the positive samples with similar viewpoints from the same camera.**

At last, although existing vehicle Re-ID datasets, including VehicleID [33], VeRI-776 [34], and VERI-Wild [35], provide important benchmarks to evaluate the state-of-the-art methods, the crucial issue is the number of cameras is limited (12 in Vechile ID, 20 in VeRI-776 and 174 in VERI-Wild). Therefore, each vehicle only appears with limited cameras. Furthermore, although they contain vehicle images from multiple viewpoints, the number of viewpoints for each vehicle ID is still limited. Herein, **we propose a new vehicle image dataset captured by a large number of cameras (*i.e.*, 6142 cameras) from a real-life transportation surveillance system, named Multi-query Re-Identification dataset (MuRI).** MuRI contains diverse viewpoints, including $front$ ($side\ front$), $side$ and $rear$ ($side\ rear$), and a large number of crossed scenes/cameras for each vehicle (*i.e.*, 34.6 in average), which provides more realistic and challenging scenarios for multi-query vehicle Re-ID.

To the best of our knowledge, we are the first to launch the multi-query setting in vehicle Re-ID, which jointly uses images from multiple scenes/viewpoints of a vehicle as a query. The contributions of this paper are mainly in the following four aspects.

- We introduce a new task called multi-query vehicle re-identification, which devotes to inferring the cross-scene re-identification by exploring the complementary information among the multiple query images with different viewpoints. The task is challenging, but easily accessible and very practical in realistic transportation systems.
- We propose a viewpoint-conditioned network (VCNet) for multi-query vehicle Re-ID, which learns special viewpoint information through the viewpoint conditional coding (VCC) module during the training and integrate the complementary viewpoints information through the viewpoint-based adaptive fusion (VAF) module in the testing. We propose the cross-view feature recovery (CVFR) module to deal with the missing viewpoint problem during inference, which maximizes the mutual information of different viewpoints by contrast learning to learn informative and consistent representations and recover the missing viewpoints.
- To measure the cross-scene retrieval ability of Re-ID, we further design a new metric, namely mean cross-scene precision (mCSP), by suppressing the positive samples with similar viewpoints from the same camera, which provides a more crucial measure in real-life Re-ID applications.
- To evaluate the effectiveness of the proposed VCNet for multi-query inference in vehicle Re-ID, we collect a multi-query vehicle Re-ID dataset with a large number of crossed scenes from real city traffic.

II. Related Work

A. Vehicle Re-ID Methods

Most of the vehicle Re-ID methods rely on single query image. In order to learn the detailed features of vehicles and expand the subtle differences between the same models, some works introduce the idea of the region of interest prediction or attention models to mine the salient regions of vehicles. He et al. [36] propose a simple and effective partial regularization method, which detects the regions of interest using pre-trained detectors and introduces multi-dimensional constraints at the part level (windows, lights, and make alike) into the vehicle Re-ID framework. It improves the model's ability to learn local information while enhancing the subtle difference perception. An et al. [37] propose an attention network using local region guidance, which mines the most important local regions by learning the weights of candidate search regions to increase the weights of discriminative features in vehicle images while reducing the effect of irrelevant background noise. Khorramshshi et al. [38] learn to capture localized discriminative features by focusing attention on the most informative keypoints based on different orientations.

To handle the similar appearance of the different vehicles, some works propose to use the additional annotation information of the dataset to learn more accurate local features of the vehicle. Liu et al. [28] exploit multi-modal data from large-scale video surveillance, such as visual features, license plates, camera locations, and contextual information, to perform a coarse-to-fine search in the feature domain and near-to-far search in the physical space. Wang et al. [22] extract local area features in different directions based on 20 keypoint locations, which are aligned and combined by embedding into directional features. The spatio-temporal constraints are modeled by spatio-temporal regularization using log-normal distribution to refine the retrieval results. Li et al. [27] introduce a deep network to fuse the camera views, vehicle types and color into the vehicle features.

Metric learning based approaches focus on solving the problem of intra-class variation and inter-class similarity caused by view variation. Bai et al. [39] propose a deep metric learning method that divides samples within each vehicle ID into groups using an online grouping method, and create multi-granularity triple samples across different vehicle IDs as well as different groups within the same vehicle ID to learn fine-grained features. Jin et al. [40] propose a multi-center metric learning framework for multi-view vehicle Re-ID that models potential views directly from the visual appearance of vehicles, and constrains the vehicle view centers by intra-class ranking loss and cross-class ranking loss to increase the discriminative information of different vehicles. In addition, zero-shot learning [41] can be combined with re-identification to achieve more accurate and fine-grained recognition and classification. To explore more information in the query, Jin et al. [29] explore the comprehensive information of multi-shot images of an object in a teacher-student manner. Although they use the multi-shot teacher branch to guide the single-image branch during training, it still contained only single-image information during the inference phase.

B. Vehicle Re-ID Metrics

Vehicle Re-ID is an image retrieval subproblem, and to evaluate the performance of Re-ID methods. Cumulative Matching Characteristics (CMC) [33] and mean Average Precision (mAP) [30] are two widely used measures. CMC-k (also known as k-level matching accuracy) [33] indicates the probability of a correct match among the top k ranked retrieval results. When comparing the performance of different methods, if there is little difference in performance between methods, the cumulative matching performance curves will overlap for the most part, making it impossible to accurately determine good or bad performance. In order to compare the performance differences between methods more concisely, the cumulative matching accuracy at some key matching positions is generally selected for comparison, where rank1 and rank5 are more common, indicating the probability of correctly matching the first 1 and the first 5 images in the result sequence, respectively.

Another metric, the mean accuracy (mAP) [30], is used to evaluate the overall performance of the Re-ID methods and represents the average of the accuracy of all retrieval results. It is originally widely used in image retrieval. For Re-ID evaluation, it can address the issue of two systems performing equally well in searching the first ground truth, but has different retrieval abilities for other hard matches. However, these two widely used measures cannot assess the ability of the model to retrieve difficult samples. To address

this issue, Ye et al. [42] propose a computationally efficient metric, namely a negative penalty (NP), which measures the penalty to find the hardest correct match. To measure the results derived from individual cameras [32] propose a cross-scene generalization measure (CGM). It first divides the vehicle images captured by the same camera into individual groups, then calculate the average ranking values for each camera.

However, all of the above metrics ignore the similar positive samples in the same camera, which leads to the virtual high metric scores.

C. Vehicle Re-ID Datasets

Recent vehicle Re-ID methods are mainly evaluated on three public datasets, including VeRI-776 [34], VehicleID [33] and VERI-Wild [35]. VeRI-776 [34] dataset contains 49,360 images of 776 vehicles, of which the samples are obtained by 20 cameras on a circular road in a 1.0 Square Kilometers area for a short period of time (4:00 PM to 5:00 PM during the day), with each vehicle being captured by at least 2 and at most 18 cameras. VehicleID [33] includes 221,763 images about 26,267 vehicles, mainly containing both front and rear views. For comprehensive evaluation of the vehicle Re-ID methods, VehicleID [33] divides the test set into 3 subsets, large, medium and small, according to the size of the vehicle images. VehicleID [33] contains limited views (only two views, i.e., front view and rear view). In addition, the images in this dataset mainly contain less complex backgrounds, occlusions and illumination changes. VERI-Wild [35] is collected in a 200 Square Kilometers suburban areas and contains 416,314 images of 40,671 vehicles taken by 174 traffic cameras. The training set consists of 277,797 images (30,671 cars) and the testing set consists of 138,517 images (10,000 cars). Similarly, the testing set of VERI-Wild [35] is divided into three subsets according to image size: large, medium, and small. The vehicle images in VERI-Wild [35] mainly have little variability in views, mostly in front and rear views.

Although impressive results have been achieved on these datasets, the vehicle Re-ID problem is still far from being addressed in the real-world scenarios. First, these datasets contain only a limited number of scenarios and cameras. The samples in VeRI-776 [34], VehicleID [33] and VERI-Wild [35] are captured by 20, 12 and 174 cameras, respectively. This is inconsistent with the real-life surveillance system in the smart city which contains tens of thousands cameras. Second, the distribution of vehicle views is uneven, with most vehicles containing images of only the front and rear views and lacking side images. Moreover, the number of cameras that each vehicle crosses is limited in the existing datasets, and thus it is difficult to evaluate the cross-scene retrieval capability of the models.

III. VCNet: Viewpoint-Conditioned Network

In this work, to effectively combine the complementary information from different vehicle viewpoints, we propose a novel viewpoint-conditioned network (VCNet) for multi-query vehicle Re-ID.

A. Network Architecture

Our VCNet includes two stages: multi-query inference as shown in Fig. 2. First, we propose a viewpoint conditional coding (VCC) module to learn specific viewpoint information. By encoding the vehicle's viewpoint features and embedding them into the learning process of vehicle detail features, it enforces the model to focus on the detail information under a specific viewpoint of the vehicle. As shown in Fig. 3, we use the vehicle's viewpoint features as conditional encoding information, to fuse with the vehicle detail features obtained at each layer of the network. It thus enables the model to focus on the vehicle viewpoint information while learning the discriminative features at that viewpoint.

To integrate the complementary information among different viewpoints of the vehicle, we propose a viewpoint-based adaptive fusion (VAF) module for multi-query inference. As shown in Fig. 2, we first assign the appearance feature weights of the query according to the similarity between the multi-query and gallery viewpoint features, then adaptively fuse the features of the multi-query according to the obtained weights, so as to take into account the complementarity and specificity between the different viewpoint features of the vehicle.

To handle the scenario of multi-query images with missing viewpoints, we further propose a cross-view feature recovery (CVFR) module to recover the missing appearance features. CVFR module maximizes the common information between different viewpoints through comparative learning, and completes the reconstruction between different viewpoint features based on the common information.

B. Viewpoint Conditional Coding (VCC) Module

The large intra-class variability due to different viewpoints is a huge challenge for vehicle Re-ID. Therefore, for the vehicle with different viewpoints, the network should focus on different detailed regions. To make full use of the information of vehicle viewpoint, we propose a viewpoint conditional coding (VCC) module, as shown in Fig. 3. We introduce a two-stream network structure in the VCC module, the upper and lower branch is used to learn the appearance and viewpoint features of the vehicle respectively, and both branches use ResNet-50 [15] as feature extractor.

First, different from Wang et al. [22] which mark 8 viewpoints ($front$, $rear$, $left$, $left\ front$, $left\ rear$, $right$, $right\ front$ and $right\ rear$) on VeRI-776 dataset, we re-divide the 8 viewpoints into 3 viewpoint labels ($front$, $rear$ and $side$) to maximize the variation between different viewpoints for the training of viewpoint prediction. To obtain a more robust and refined viewpoint features, on the basis of the training on the VeRI-776 dataset, we re-train the viewpoint prediction network on our MuRI dataset. To regress viewpoints, we use the cross-entropy loss as the supervision of the training of viewpoints as follows,

$$\mathcal{L}_{view} = -\frac{1}{N}\sum_{i=1}^{N} log(p(v_i|x_i)), \tag{1}$$

where N represents the number of images in a training batch, x_i denotes the input image, and v_i denotes the viewpoint label.

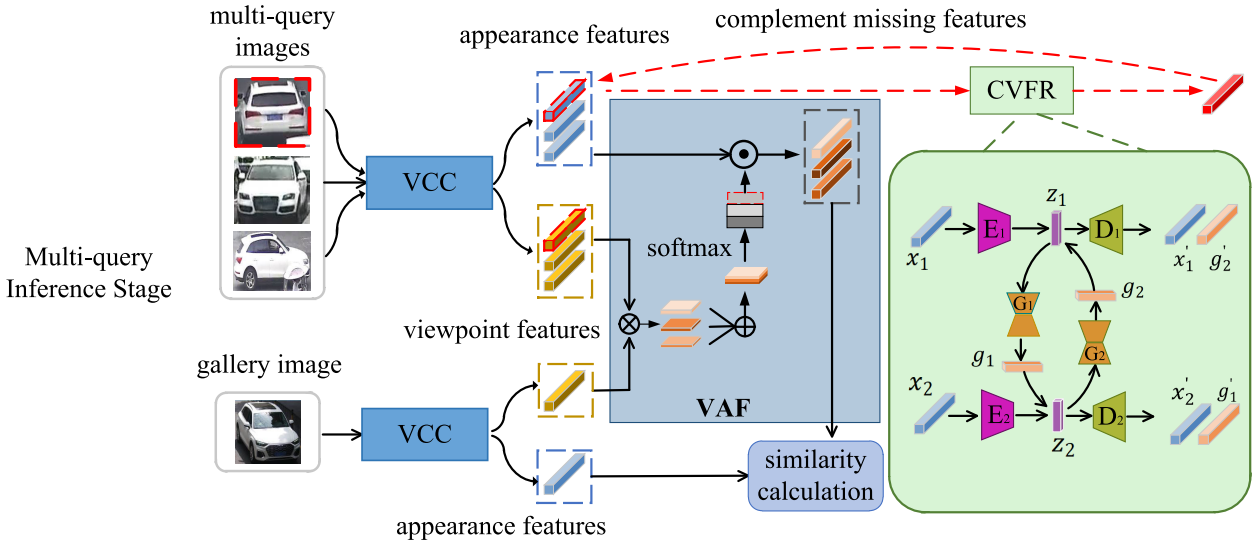
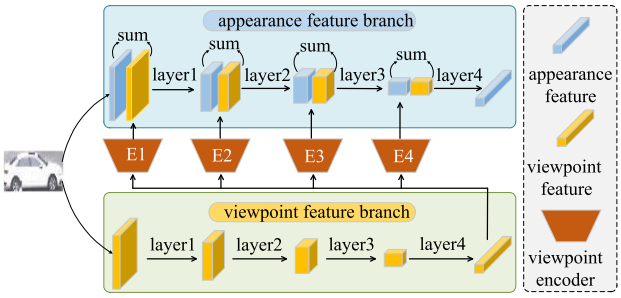


Fig. 2. The pipeline of our inference framework. In the multi-query inference stage, viewpoint weights will be calculated between query and gallery viewpoint features, to integrate the complementary information among different viewpoints of the vehicle, we adaptively fuse the generated viewpoint weights with appearance features by viewpoint-based adaptive fusion (VAF) module. When we miss a query image from a random viewpoint, the appearance feature will be recovered by the cross-view feature recovery (CVFR) module. "⊕", "⊗" and "⊙" denote concatenation, cosine similarity calculation, and multiply respectively.



Viewpoint Conditional Coding Module

Fig. 3. The framework of the proposed VCC module. First, we learn the vehicle viewpoint features by the yellow branch below, then pass them through different deconvolution encoders (E1, E2, E3, and E4) to obtain viewpoint encoding features in different scales. These viewpoint encoding features are added to the vehicle appearance feature learning branch to learn vehicle detail features based on specific viewpoints.

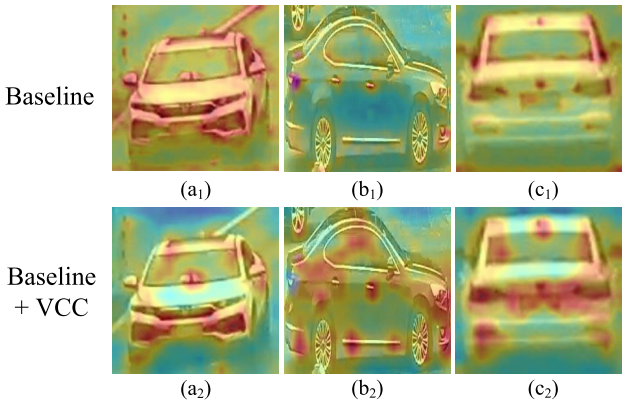


Fig. 4. Visualization of the feature maps of our VCC module comparing with the baseline.

Then, to enforce the network focus on more discriminative regions based on the viewpoint information, the learned viewpoint features are encoded to the vehicle appearance feature learning branch. Here, we use different deconvolution functions as the viewpoint encoders to map the viewpoint features to the embedded features whose dimensions are same with the corresponding layers. Next, we sum the appearance features and the embedded features and then send to the next layer of the network as viewpoint encoding information. Finally, we can obtain the vehicle features which contain specific viewpoint information. The cross-entropy loss and triplet loss are used for the training of appearance features as follows,

$$\mathcal{L}_{appearance} = -\frac{1}{N}\sum_{i=1}^{N} log(p(y_i|x_i)) + \frac{1}{N}\sum_{i=1}^{N}(m + d(f_a^i, f_p^i) - d(f_a^i, f_n^i)), \quad (2)$$

where y_i denotes the appearance label, m denotes the margin, $d(\cdot)$ indicates the Euclidean distance, f_a, f_p and f_n denotes anchor, positive and negative appearance features respectively.

At last, the training loss of VCC module can be formulated as,

$$\mathcal{L}_{vcc} = \mathcal{L}_{view} + \mathcal{L}_{appearance}. \quad (3)$$

To demonstrate the effectiveness of VCC module, we visualize the features of the last layer, as shown in Fig. 4. VCC module can reduce the interference of background and enforce the model to better focus on the vehicles, comparing Fig. 4 (a_2) with Fig. 4 (a_1). In addition, VCC module encourages the model to focus on the main clues for classification and explores more discriminative regions, comparing Fig. 4 (b_2) and (c_2) with Fig. 4 (b_1) and (c_1).

C. Viewpoint-Based Adaptive Fusion (VAF)

Although we have got a robust features that contains viewpoint information in VCC module, the limited information in

a single query image significantly hinders the performance of vehicle Re-ID in the inference stage. To integrate multi-viewpoint information of the vehicle in the inference stage and solve diverse viewpoint gaps between query and gallery, we propose a viewpoint-based adaptive fusion (VAF) module in the inference process, which adaptively fuses the generated viewpoint weights with appearance features. The inference process is shown in Fig. 2.

First, we jointly use 3 vehicle images with different viewpoints in the query set and send them into the pre-trained VCC module to extract the appearance features and viewpoint features respectively. To obtain the viewpoint similarity between 3 query images with gallery images, we calculate the features cosine distance between 3 query viewpoint features with gallery viewpoint features as follows,

$$s_i = \frac{< f_v^{q_i}, f_v^g >}{||f_v^{q_i}|| \times ||f_v^g||}, \tag{4}$$

where $i = 1, 2, 3$, $f_v^{q_i}$ and f_v^g denotes query and gallery viewpoint features, $< x, y >$ indicates the inner product of x and y.

Then, we can obtain the similarity weight set $\mathbf{W} = \{w_i | i = 1, 2, 3\}$ by computing the similarity of query and gallery viewpoint features using the concatenation and softmax function. To adaptively fuse the viewpoint information in the appearance features, we multiply the multi-query appearance features $\mathbf{F} = \{ f_a^{q_i} | i = 1, 2, 3 \}$ with the similarity weight set $\mathbf{W}$ to obtain the weighted appearance features F' as follows,

$$\bar{F} = \left\{ f_a^{\bar{q}_1} \times w_1, f_a^{\bar{q}_1} \times w_2, f_a^{\bar{q}_1} \times w_3 \right\}, \tag{5}$$

To this end, the query appearance features with the similar viewpoint as the gallery image will be assigned a large weight. If a query image from a random viewpoint is missing, the appearance feature will be recovered by the cross-view feature recovery (CVFR) module. For the final recognition task, we perform a similarity calculation between the fused appearance features with the gallery appearance features and obtain the corresponding scores, which are summed to obtain the final recognition scores.

D. Cross-View Feature Recovery (CVFR) Module

In some scenarios, the query data might not contain some viewpoints of vehicle images. While our network accepts three query vehicle images as inputs, and thus can not handle such data with missing viewpoints. To solve this problem, referring to Lin et al. [43] in multi-view, we propose cross-view feature recovery (CVFR) module to recover the missing appearance features. To learn information-rich consistent representations, CVFR module maximizes the mutual information between different viewpoints by contrast learning. To recover the missing appearance features, CVFR module minimizes the conditional entropy of different viewpoints by dual prediction. For the sake of convenience, we assume that one viewpoint from $front$, $rear$, and $side$ is randomly missing, and the recovery process of the missing appearance features is as follows. First, we send two known images from different viewpoints to the pre-trained VCC module and obtain the appearance features x_1 and x_2, respectively. Then the latent representations Z_1, Z_2 are obtained after the respective encoders E_1, E_2, and the reconstructed features x_1', x_2' are obtained after the decoders D_1, D_2. The reconstructed differences will be minimized by the mean squared error loss function as follows,

$$\mathcal{L}_{mse} = \sum_{v=1}^{2} \sum_{t=1}^{m} ||x_v^t - x_v'^t||^2, \tag{6}$$

where x_v^t denotes the t-th sample of x_v. To facilitate data recovery ability, contrastive learning is used to learn the common information between different viewpoints and to maximize the common information. The contrastive loss mathematical formula is as follows:

$$\mathcal{L}_{cl} = -\sum_{t=1}^{m} (I(Z_1^t, Z_2^t) + \alpha(H(Z_1^t) + H(Z_2^t))), \tag{7}$$

where I denotes the mutual information, H is the information entropy, and the parameter α is set as 9 to regularize the entropy in our experiments. From information theory, information entropy is the average amount of information conveyed by an event. Hence a larger entropy $H(Z^i)$ denotes a more informative representation Z^i. The viewpoint predictors G_1, G_2 are used to generate latent representations of the corresponding viewpoints, and the differences are generated by minimizing the loss function,

$$\mathcal{L}_{pre} = \sum_{v=1}^{2} \sum_{t=1}^{m} ||g_v^t - Z_{3-v}^t||^2, \tag{8}$$

where v represents the number of available viewpoints, and we can obtain the missing appearance feature with random viewpoints from available ones.

To further narrow the differences between the generated and the original viewpoint features, we feed the latent representations generated by the viewpoint predictor into the corresponding decoders separately. Then we obtain the generation of reconfiguration features g_1', g_2', which are constrained by mean squared error loss,

$$\mathcal{L}_{mse} = \sum_{v=1}^{2} \sum_{t=1}^{m} ||x_v^t - D_v(g_k^t)||^2, \tag{9}$$

where D_v denotes Decoders. Therefore, even with the conventional single query setting, our framework can still learn more comprehensive appearance feature representation of a certain vehicle by integrating the complementary information among the given and recovered features, which is expected to significantly overcome the dramatic appearance changes caused by different viewpoints.

IV. MuRI Dataset

To evaluate the proposed VCNet on multi-query vehicle Re-ID, we propose a multi-views and unconstrained vehicle Re-ID dataset, MuRI, to integrate the complementary information among different viewpoints during inference.

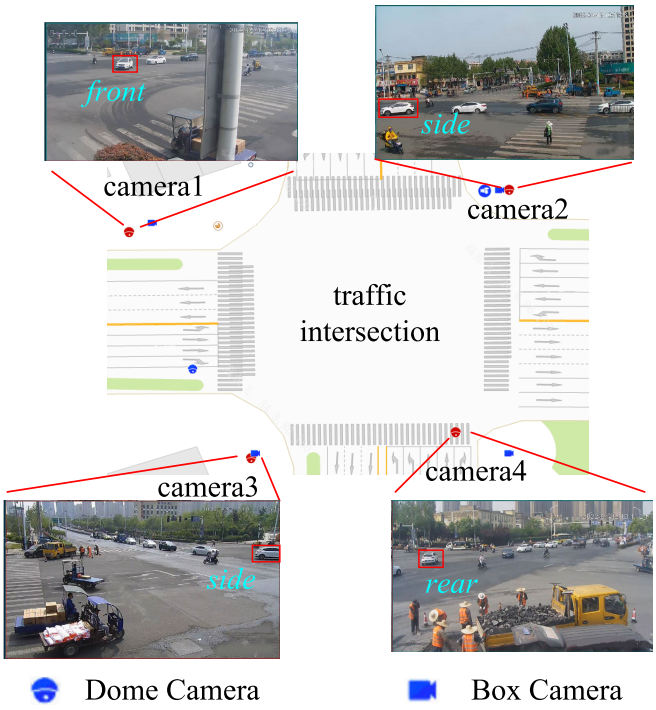


Fig. 5. Illustration of data acquisition environment of the vehicle images obtained in the MuRI dataset. Vehicle images with diverse viewpoints are captured by the dome cameras at traffic intersection.

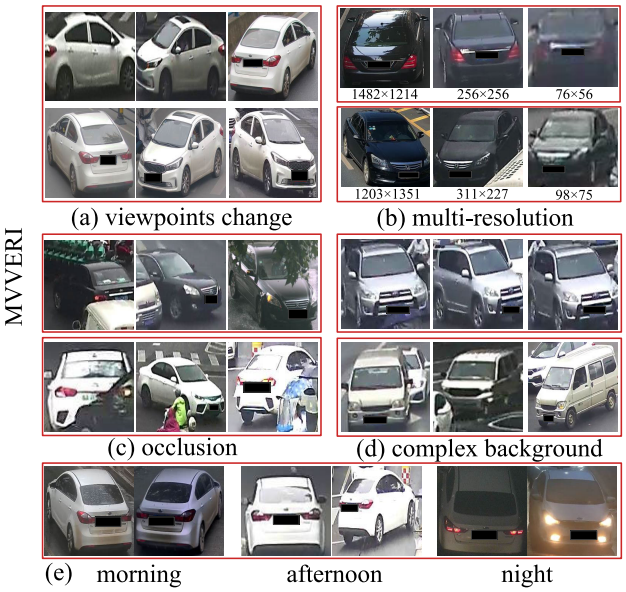


Fig. 6. Challenges of MuRI dataset. The images in the common red box indicate the same vehicle ID.

A. Data Acquisition

The MuRI dataset is collected in a large city with more than 1000 Squares Kilometers. to obtain the vehicle images from more diverse cameras, we first search the corresponding vehicle images by license plate in the Public Security City Service Platform, which monitors tens of thousands of cameras in the city. To ensure that each vehicle has rich viewpoint information, we obtain the vehicle images of different viewpoints at a traffic intersection, which has three or four surveillance cameras from different directions. As shown in Fig. 5, to ensure the diverse viewpoint information, we choose the rotatable dome cameras from the Public Security City Service Platform as the shooting cameras. For the videos captured from the platform, we generate the surrounding boxes by the tracking detection algorithm [44]. For effective evaluation, we automatically select the vehicle images in every 3 adjacent frames, followed by manual checking to avoid data redundancy. The time span of vehicle appearance in the data set is about half the year, and the vehicle resolution is variable due to the varying distances between cameras and vehicles of interest.

B. Dataset Description

The MuRI dataset contains 200 identities in five viewpoints ($front$, $side\ front$, $side$, $rear$, and $side\ rear$) with diverse resolution and illumination conditions. Due to the similar appearance between $front$ and $side\ front$, as well as $rear$, and $side\ rear$, we merge the five viewpoints into three in this paper, i.e., $front$, $side$ and $rear$. For effective evaluation, we automatically select the vehicle images from the traffic intersection in every 3 adjacent images, followed by manual checking to avoid data redundancy. Together with the images collected by the Public Security City Service Platform, our MuRI forms 23637 vehicle images from 6142 cameras in total, which provides more near-reality smart city scenarios with mass cameras for multi-query vehicle Re-ID. We select 150 identities for training, and 50 identities for testing/inference. In the inference stage, we use the entire testing set as gallery set, while randomly selecting 3 records from different viewpoints as multi query images.

C. Dataset Challenges

Our MuRI mainly contains five different challenges as shown in Fig. 6. First, our dataset provides comprehensive five viewpoints, including $front$ ($side\ front$), $side$ and $rear$ ($side\ rear$) for each vehicle, which produce huge intra-class for Vehicle Re-ID. Then, due to the varying distances between cameras and vehicles of interest, the vehicle images present different resolutions as shown in Fig. 6 (b). The poor detailed information in the low resolution as well as the appearance gap between different resolutions further bring huge challenge for veichle Re-ID. Moreover, MuRI dataset is collected in a large city surveillance system spanned more than 1000 km^2, and the urban environment is very complex. To this end, MuRI contains many vehicle images with occlusion and complex background as shown in Fig. 6 (c, d), which brings severe challenges for vehicle Re-ID. Finally, the vehicle images in MuRI dataset are collected in a long time span with more than half a year, which provides large number of cross-time vehicle data with different illuminations, such as morning, afternoon, and evening as shown in Fig. 6 (e). The large illumination changes results in huge difference in vehicle appearance. Furthermore, the strong lighting from the headlights and the taillights during the night brings additional challenge for vehicle Re-ID.

D. Dataset Characteristics

Compared with existing prevalent Re-ID datasets as shown in Table I, in calculating the average number of viewpoints for

TABLE I
COMPARISONS AMONG VEHICLEID, VERI-776, VERI-WILD, AND THE CREATED MURI DATASETS FOR VEHICLE REID

Dataset	VehicleID	VeRI-776	VERI-Wild	MuRI
Images	221,763	49,360	416,314	23,637
Identities	26,267	776	40,671	200
Cameras	12	20	174	6142
Viewpoints/id	2.0	4.2	3.4	5.0
Cross-resolution	×	×	×	√
Occlusion	×	×	√	√
Complex Background	×	×	√	√
Morning	√	×	√	√
Afternoon	√	√	√	√
Night	×	×	√	√

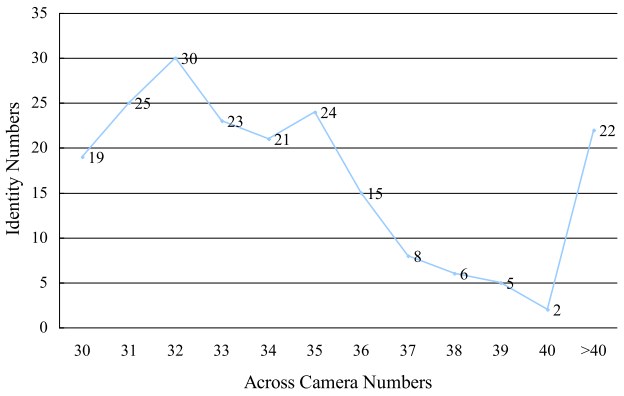


Fig. 7. Distribution of the number of identities across the number of cameras.

each id in the dataset, we divided the vehicle viewpoints into five categories according to *front*, *side front*, *side*, *side rear*, and *rear*. MuRI has the following major advantages.

1) **Numerous cameras with wide area**. MuRI contains 200 vehicle IDs captured by 6142 cameras from a real-life transportation surveillance system covering over 1000 km^2 urban area.
2) **Comprehensive viewpoints of each ID**. MuRI provides comprehensive five viewpoints, including *front* (*side front*), *side* and *rear* (*side rear*) for each vehicle, which provides a more realistic and challenging scenario for vehicle Re-ID.
3) **Large number of cameras crossed by each ID**. Each vehicle in MuRI crosses 34.6 cameras in average, varying from 30 to 50 cameras, as shown in Fig. 7.

V. MCSP METRIC

One problem with existing metrics is the lack of consideration of cross-scene scenarios. To this end, a new metric, named mean Cross-scene Precision (mCSP) is proposed in this paper to ensure the cross-scene retrieval capability of the network. The main idea of mCSP can be summarized as follows: if there exist positive samples with similar viewpoints from the same camera, we consider them as the same scene and remove them from the ranked list. Given a ranked list, we use TP to denote the number of positive samples retrieved, f_c^i and f_c^j denotes the viewpoint feature of the two positive sample image retrieved under the same camera c. When their Euclidean distance $d(f_c^i, f_c^j)$ is smaller than a threshold hyperparameter

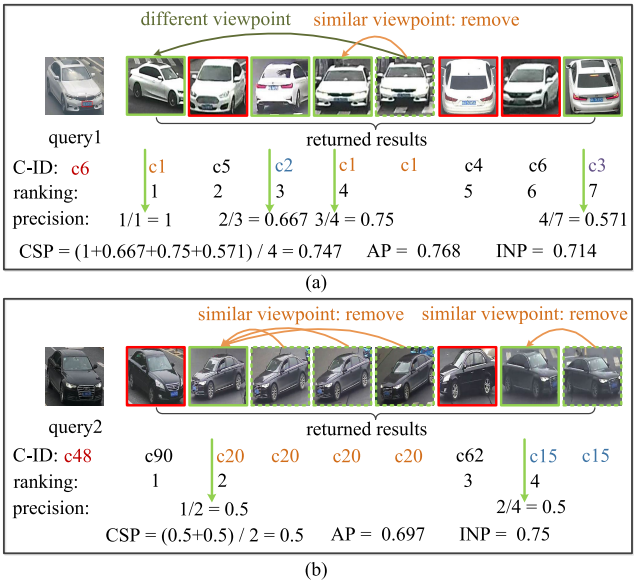


Fig. 8. Illustration of the calculation process of CSP metric. C-ID denotes the camera id, true matching and false matching are bounded in green and red boxes, respectively. For positive samples retrieved from the same camera, the CSP metric removes them during the calculation when their viewpoints are similar.

ε, we consider that the viewpoints of f_c^j and f_c^i is similar. We use SC to denote the number of samples with the similar viewpoint under the same camera ID in TP, and FP denotes the number of positive samples with prediction errors, mCSP can be expressed in the following form,

$$mCSP = \frac{\sum_{i=0}^{N_{cs}} \frac{TP-SC}{TP-SC+FP}}{N_{cs}}, \tag{10}$$

where N_{cs} denotes the captured target images from positive samples with different cameras. We visualize the calculation process of CSP, as shown in Fig. 8. As shown in Fig. 8 (a), even though the conventional Re-ID metrics [6], [45] such as AP and INP remove the vehicle images with the same camera as the query from the gallery set, positive samples with similar viewpoints under another same camera (different from the query camera) still tend to be more easily identified, which results in virtually high scores in the existing metrics. By contrast, for the positive samples from the different cameras with the query, the proposed CSP metric further removes the ones with similar viewpoints from another same camera. Therefore it can better reflect the ability of cross-camera retrieval by further suppressing the positive samples with similar viewpoints as shown in Fig. 8 (b).

VI. EXPERIMENTS

A. Experiments Setup

1) Train: We use ResNet-50 [15] pre-trained on ImageNet [46] as our backbone. The model is trained for 80 epochs with the SGD optimizer. We warm up the learning rate to 5e-2 in the first 5 epochs and the backbone is frozen in the warm-up step. The learning rate of 5e-2 is kept until the 60th, drops to 5e-3 in the 60th epoch, and drops to

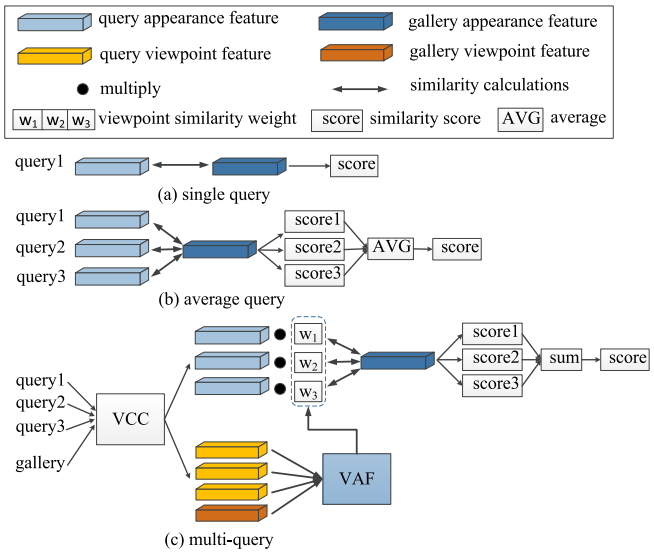


Fig. 9. Diagrams of different inference settings.

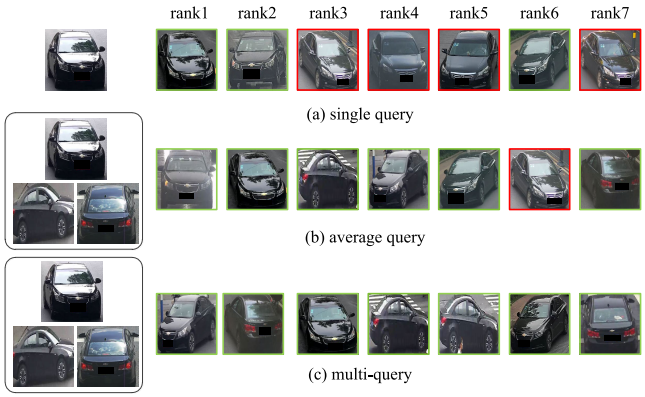


Fig. 10. Examples of ranking results with different inference settings.

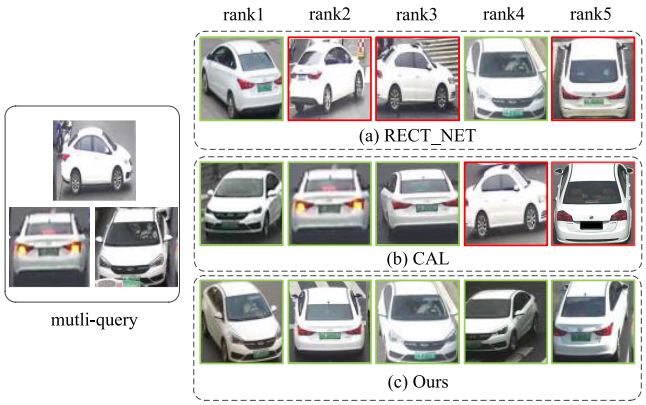


Fig. 11. The top five results of different methods for multi-query inference.

5e-4 in the 75th epoch for faster convergence. We first pad 10 pixels on the image border, and then randomly crop it to 256×256. We also augment the data with random erasing. Further, we add a Batch Normalization layer after the global feature. A fully connected layer is added to map the global feature to the ID classification score. The batch size is 36 in the MuRI dataset.

2) Inference: In our inference process, we evaluate the methods in three inference ways, including single query, average query and multi-query. As demonstrated in Fig. 9 (a), single inference directly calculate the cosine distance between each query and the gallery set, which ignores the multi-view information during the inference. When facing the multiple queries, the intuitive inference way is the average inference, which computes the average value of multiple query features, as shown in Fig. 9 (b). However, simply averaging the query features can not effectively use the different viewpoint information of the vehicle. To adaptively utilize the complementary information in the multiple queries from different cameras with diverse viewpoints combinations, we propose the multi-inference for the proposed VCNet, as shown in Fig. 9 (c). Specifically, it generates viewpoint weights by the similarity between multiple queries and gallery view features, then fuses the generated viewpoint weights with appearance features. To demonstrate the significant advantage of the proposed multi-query setting over previous single or average query settings, we illustrate the ranking results with different settings as shown in Fig. 10. Compared with the single query and the average query, multi-query can hit earlier positive samples, since it can break the information barrier by combining the vehicle information from different viewpoints to obtain the overall features of the vehicle.

B. Experimental Results

1) Comparison With the State-of-the-Arts: To verify the effectiveness of the proposed VCNet with the multi-query setting, we compare four state-of-the-art vehicle Re-ID methods on the collected MuRI dataset. Specifically, we construct the multi-query setting with the number of query $N_Q = 3$, which including three different viewpoints $front$, $rear$ and $side$ respectively. We evaluate the state-of-the-art methods in both single and average inferences for comparison. As shown in Table II, all the state-of-the-art methods achieve significant improvement in the average inference compared to the single inference, which evidences that using multiple queries can better incorporate the complementary information among the images. By progressively embedding viewpoint features to appearance feature learning via the viewpoint conditional coding (VCC), and integrating the complementary information among different viewpoints via the viewpoint-based adaptive fusion (VAF), our VCNet with multi-inference achieves superior performance compared to the state-of-the-art methods. This validates the effectiveness of the proposed VCNet while handling the multi-query inference for vehicle Re-ID. Fig. 11 shows the corresponding ranking results of multiple queries from different viewpoints, which further evidences the promising performance of our method while handling the challenging cross-scene retrieval problem compared to other methods.

2) Ablation Study of VCNet: To verify the effective contribution of the components in our model, we implement the ablation study on the viewpoint conditional coding (VCC) module, viewpoint-based adaptive fusion (VAF) module, and cross-view feature recovery (CVFR) module on our MuRI

TABLE II
PERFORMANCE COMPARISONS ON MURI BENCHMARK

Method	Venue	Inference way	Rank1	Rank5	Rank10	mAP	mINP	mCGM	mCSP
DMML [16]	ICCV 2019	single query	0.644	0.767	0.814	0.362	0.060	0.175	0.198
		average query	0.766	0.884	0.927	0.524	0.088	0.281	0.272
RECT_Net [17]	CVPR 2020	single query	0.729	0.811	0.859	0.415	0.074	0.212	0.225
		average query	0.806	0.924	0.956	0.602	0.108	0.316	0.330
GRF [18]	TIP 2020	single query	0.683	0.806	0.841	0.398	0.070	0.188	0.214
		average query	0.786	0.906	0.942	0.565	0.101	0.302	0.314
CAL [19]	ICCV 2021	single query	0.754	0.842	0.890	0.441	0.082	0.228	0.266
		average query	0.816	0.960	0.980	0.645	0.138	0.360	0.359
VCNet	Ours	multi-query	**0.843**	**0.962**	**0.980**	**0.677**	**0.145**	**0.400**	**0.393**

TABLE III
ABLATION STUDY OF VCNET WHEN THE NUMBER OF QUERY N_Q = 3 WITH ONE VIEWPOINT RANDOM MISSING

Settings	Rank1	mAP	mINP	mCGM	mCSP
Baseline	0.774	0.506	0.104	0.310	0.307
+VCC	0.801	0.535	0.122	0.336	0.342
+VCC+VAF	0.820	0.554	0.127	0.344	0.360
+VCC+VAF +CVFR	0.832	0.565	0.130	0.358	0.371

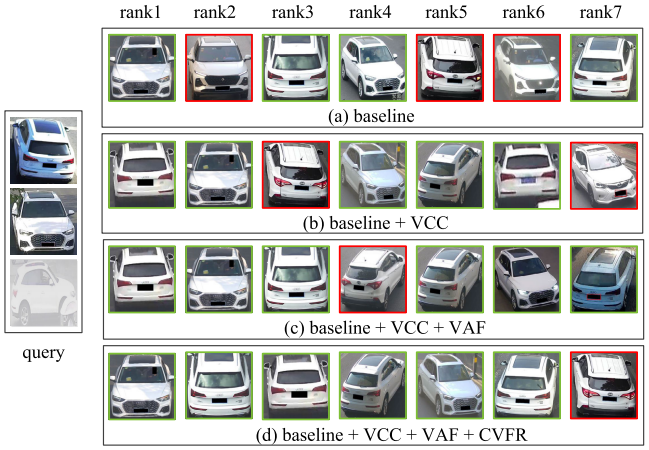


Fig. 12. Examples of ranking results by progressively introducing the proposed three components in VCNet when the number of query N_Q = 3 with one viewpoint random missing (shown as gray in the query).

dataset, as shown in Table III. We employ ResNet-50 [15] as our baseline to extract vehicle appearance features in an average query inference. Note that introducing VCC significantly boosts the baseline, which evidences the effectiveness of the proposed VCC module which can integrate the complementary information among different viewpoints of the vehicle. VAF consistently brings a significant improvement on all the metrics by adaptively fusing the generated viewpoint weights with appearance features. At last, CVFR further enhances the performance by recovering features of the missing viewpoint. To further evidences the effectiveness of the proposed VCC, VAF and CVFR modules, we demonstrate the ranking results of VCNet by integrating the three key components when the number of query $N_Q = 3$ with one viewpoint random missing, as shown in Fig. 12 demonstrates. By progressively

TABLE IV
EVALUATION OF VAF WHEN THE NUMBER OF QUERIES N_Q=3

Methods	Rank1	mAP	mINP	mCGM	mCSP
RECT_NET [17]	0.806	0.602	0.108	0.316	0.330
RECT_NET + VAF	0.820	0.620	0.114	0.349	0.338
CAL [19]	0.816	0.645	0.138	0.360	0.359
CAL + VAF	0.838	0.667	0.142	0.372	0.380
VCC (Ours)	0.826	0.659	0.141	0.374	0.370
VCC + VAF	**0.843**	**0.677**	**0.145**	**0.400**	**0.393**

TABLE V
EVALUATION ON CVFR WHEN THE NUMBER OF QUERY $N_Q = 3$ WITH ONE RANDOM MISSING

Inference way	Rank1	mAP	mINP	mCGM	mCSP
(a) single	0.715	0.426	0.087	0.256	0.272
(b) average	0.801	0.535	0.122	0.336	0.342
(c) average + CVFR	0.814	0.544	0.126	0.345	0.351
(d) multi (2 views)	0.820	0.554	0.127	0.344	0.360
(e) multi + CVFR	0.832	0.565	0.130	0.358	0.371

introducing the three key components, it can hit more correct vehicle images at earlier rankings, which verifies their effectiveness for multi-query vehicle Re-ID.

3) Evaluation on VAF: To further demonstrate the effectiveness and applicability of viewpoint-based adaptive fusion (VAF) module, we plugin VAF into two state-of-the-art methods with the number of query $N_Q = 3$ with three different viewpoints ($front$, $rear$ and $side$) in the query set. To obtain the viewpoint features for the VAF module, we use a pre-trained viewpoint prediction network for all the other methods. In vehicle re-identification, the main challenge is the intra-class variability and inter-class similarity problem due to the difference in vehicle viewpoints. We can use the images from different views of the vehicle during the inference through multi-query. VAF can assign weights adaptively according to the similarity between the viewpoints of vehicles in query and gallery. More similarity between the query and gallery viewpoints, the greater the weight when retrieving, which reduce the difficulty of identifying positive samples. As shown in Table IV, after integrating the proposed VAF into RECT_NET [17] and CAL [19], it brings a large margin improvement over the original methods by fusing the generated viewpoint weights with their appearance features. This verifies that VAF can better integrate the complementary information among different viewpoints.

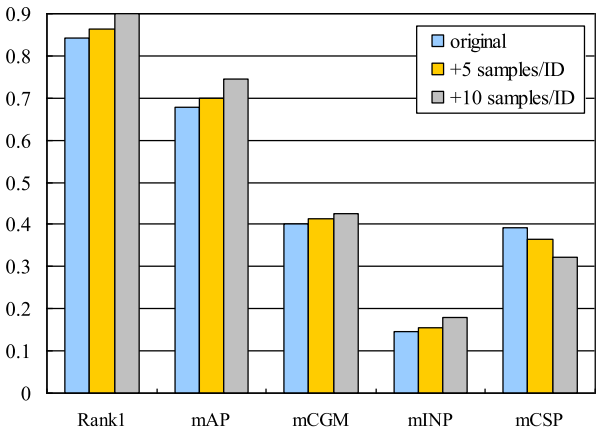


Fig. 13. The changes of each metric on the modified gallery sets in MuRI.

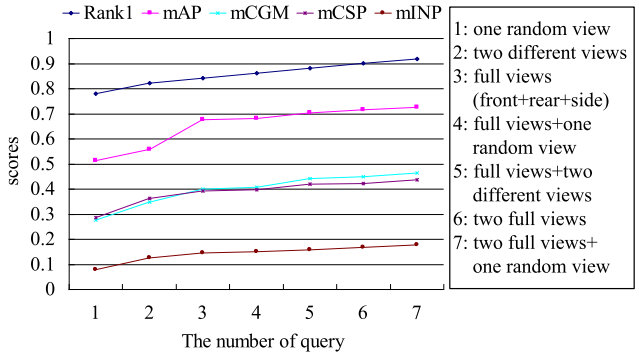


Fig. 14. Evaluation of the number of queries on MuRI.

4) Evaluation on CVFR: To handle the scenario with viewpoint missing, we propose the cross-view feature recovery (CVFR) module to recover the appearance features of the missing viewpoints. To validate the compatibility of our proposed multi-query inference method with viewpoint missing, we evaluate our method with different viewpoint missing cases as shown in Table V. The average query as shown in Table V (b) makes a significant improvement by combining two viewpoints, compared to the single inference in Table V (a). By recovering the missing appearance features via the proposed CVFR module, Table V (c) brings further improvement. Our proposed multi-query inference can further boost the performance with only two existing viewpoints, as shown in Table V (d), which strongly verifies the effectiveness of the proposed multi-query inference. Finally, the multi-query inference together with recovering the missing appearance features via the proposed CVFR module achieves the best performance, as shown in Table V (e), which indicates the necessity of the supplementary information in multi-query for vehicle Re-ID.

5) Evaluation on mCSP: To evaluate the capability of cross-scene retrieval of the Re-ID methods, we reconstruct by adding 5 and 10 images of the same viewpoint under the same camera for each vehicle in the original gallery set. To make a fair comparison, we further delete 5 and 10 images of that vehicle under different cameras to ensure the total number of images of each vehicle remain unchanged. Fig. 13 shows the comparison of the proposed mCSP comparing with existing metrics on both the original and modified gallery sets. By adding the images with the same viewpoint under the same camera, all the Rank1, mAP, and mINP increase in the modified galleries due to more easy matching samples. This is irrational in realistic Re-ID where the capability of matching more positive vehicle images across more diverse scenes is even crucial. By contrast, the proposed mCSP declines with the images with the same viewpoint under the same camera increase, since the positive samples recognized under different cameras become less. The mCSP only focuses on retrieving positive sample images from different cameras in the gallery set and images from different views in the same camera, which can more realistically reflect the retrieval accuracy across cameras in the realistic Re-ID.

6) Evaluation on the Number of Query: Fig. 14 evaluates the multi-query inference with the different number of queries. We can see that as the number of queries increases, the performance of the multi-query consistently improves, benefiting from the richer information in the multiple diverse images about the vehicle. Note that the significant improvement in each metric is achieved by increasing from 1 random view to 2 different views, and then to 3 full views ($front + back + side$). When the number of queries continues to increase from 3 to 7, the performance of each metric consistently increases, but with a slightly slower increase. This indicates that the larger diversity in viewpoints between queries, the better improvement in multi-query inference, since more complementary information between different viewpoints of the vehicle.

C. Limitation

This is the first work to launch the multi-query vehicle Re-ID by jointly utilizing multi-scene/multi-viewpoint images as the query for more realistic and robust vehicle re-identification. Despite the effectiveness of the proposed viewpoint-conditioned network (VCNet) for multi-query vehicle Re-ID, there are still some limitations remaining improved. First, VCNet extracts the vehicle viewpoint information to assist the appearance feature learning of vehicle images. However, this requires the annotations to pre-train the vehicle viewpoint predictor. Second, the vehicle multi-viewpoint information is only utilized in the multi-query testing process, and the complementarity between different viewpoint features of the vehicle has not fully explored in the training process. In the future, one can consider the unsupervised approach to learning vehicle viewpoint information. In addition, we can further explore the viewpoint complementarity in both training and testing, and infer the overall information of vehicles through a single vehicle image.

VII. Conclusion

In this paper, we first launch the multi-query vehicle Re-ID task which leverages multiple queries to overcome the viewpoint limitation of a single one, and propose a viewpoint-conditioned network (VCNet) for multi-query vehicle Re-ID.

First, we propose a viewpoint conditional coding (VCC) module in the training process to learn specific viewpoint information. Then, we propose a viewpoint-based adaptive fusion (VAF) module to integrate the complementary information among different viewpoints in the inference process. To handle the scenario when query images from random viewpoints, we propose the cross-view feature recovery (CVFR) module to recover the missing appearance feature. Finally, a new metric (mCSP) and a new dataset (MuRI) are proposed to measure the ability of cross-scene recognition and conduct multi-query evaluation experiments respectively. Comprehensive experiments demonstrate the necessity of the multi-query inference and the effectiveness of the proposed VCNet. This work provides new research direction for vehicle Re-ID and related areas.

REFERENCES

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

[2] H.-M. Hsu, J. Cai, Y. Wang, J.-N. Hwang, and K.-J. Kim, "Multi-target multi-camera tracking of vehicles using metadata-aided Re-ID and trajectory-based camera link model," *IEEE Trans. Image Process.*, vol. 30, pp. 5198–5210, 2021.

[3] N. Dilshad and J. Song, "Dual-stream Siamese network for vehicle re-identification via dilated convolutional layers," in *Proc. IEEE Int. Conf. Smart Internet Things (SmartIoT)*, Aug. 2021, pp. 350–352.

[4] H. Guo, K. Zhu, M. Tang, and J. Wang, "Two-level attention network with multi-grain ranking loss for vehicle re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 9, pp. 4328–4338, Sep. 2019.

[5] X. Liu, L. Li, S. Wang, Z.-J. Zha, D. Meng, and Q. Huang, "Adaptive reconstruction network for weakly supervised referring expression grounding," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2611–2620.

[6] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, "Bag of tricks and a strong baseline for deep person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 1487–1495.

[7] H. Guo, C. Zhao, Z. Liu, J. Wang, and H. Lu, "Learning coarse-to-fine structured feature embedding for vehicle re-identification," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 6853–6860.

[8] L. Mai, X.-Z. Chen, C.-W. Yu, and Y.-L. Chen, "Multi-view vehicle re-identification method based on Siamese convolutional neural network structure," in *Proc. IEEE Int. Conf. Consum. Electron.*, Sep. 2020, pp. 1–2.

[9] X. Liu, W. Liu, J. Zheng, C. Yan, and T. Mei, "Beyond the parts: Learning multi-view cross-part correlation for vehicle re-identification," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 907–915.

[10] S. Alfasly et al., "Multi-label-based similarity learning for vehicle re-identification," *IEEE Access*, vol. 7, pp. 162605–162616, 2019.

[11] F. Shen, J. Zhu, X. Zhu, Y. Xie, and J. Huang, "Exploring spatial significance via hybrid pyramidal graph network for vehicle re-identification," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 7, pp. 8793–8804, Jul. 2022.

[12] Z. Lu, R. Lin, X. Lou, L. Zheng, and H. Hu, "Identity-unrelated information decoupling model for vehicle re-identification," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 10, pp. 19001–19015, Oct. 2022.

[13] Y. Bai, J. Liu, Y. Lou, C. Wang, and L.-Y. Duan, "Disentangled feature learning network and a comprehensive benchmark for vehicle re-identification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 6854–6871, Oct. 2022.

[14] M. Li, J. Liu, C. Zheng, X. Huang, and Z. Zhang, "Exploiting multi-view part-wise correlation via an efficient transformer for vehicle re-identification," *IEEE Trans. Multimedia*, vol. 25, pp. 919–929, 2023.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[16] G. Chen, T. Zhang, J. Lu, and J. Zhou, "Deep meta metric learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9546–9555.

[17] X. Zhu, Z. Luo, P. Fu, and X. Ji, "VOC-RelD: Vehicle re-identification based on vehicle-orientation-camera," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 2566–2573.

[18] X. Liu, S. Zhang, X. Wang, R. Hong, and Q. Tian, "Group-group loss-based global-regional feature learning for vehicle re-identification," *IEEE Trans. Image Process.*, vol. 29, pp. 2638–2652, 2020.

[19] Y. Rao, G. Chen, J. Lu, and J. Zhou, "Counterfactual attention learning for fine-grained visual categorization and re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 1005–1014.

[20] X. Chen, H. Sui, J. Fang, W. Feng, and M. Zhou, "Vehicle re-identification using distance-based global and partial multi-regional feature learning," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 2, pp. 1276–1286, Feb. 2021.

[21] S. He, H. Luo, P. Wang, F. Wang, H. Li, and W. Jiang, "TransReID: Transformer-based object re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 14993–15002.

[22] Z. Wang et al., "Orientation invariant feature embedding and spatial temporal regularization for vehicle re-identification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 379–387.

[23] O. Moskvyak, F. Maire, F. Dayoub, and M. Baktashmotlagh, "Keypoint-aligned embeddings for image retrieval and re-identification," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 676–685.

[24] X. Liu, S. Zhang, Q. Huang, and W. Gao, "RAM: A region-aware deep model for vehicle re-identification," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2018, pp. 1–6.

[25] D. Meng et al., "Parsing-based view-aware embedding network for vehicle re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7101–7110.

[26] W. Sun, G. Dai, X. Zhang, X. He, and X. Chen, "TBE-Net: A three-branch embedding network with part-aware ability and feature complementary learning for vehicle re-identification," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 9, pp. 14557–14569, Sep. 2022.

[27] H. Li et al., "Attributes guided feature learning for vehicle re-identification," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 6, no. 5, pp. 1211–1221, Oct. 2022.

[28] X. Liu, W. Liu, T. Mei, and H. Ma, "PROVID: Progressive and multimodal vehicle re-identification for large-scale urban surveillance," in *Proc. IEEE Trans. Multimedia (TMM)*, 2017, pp. 645–658.

[29] X. Jin, C. Lan, W. Zeng, and Z. Chen, "Uncertainty-aware multi-shot knowledge distillation for image-based object re-identification," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 11165–11172.

[30] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1116–1124.

[31] R. Zhou, X. Chang, L. Shi, Y.-D. Shen, Y. Yang, and F. Nie, "Person reidentification via multi-feature fusion with adaptive graph learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 5, pp. 1592–1601, May 2020.

[32] J. Zhao, Y. Zhao, J. Li, K. Yan, and Y. Tian, "Heterogeneous relational complement for vehicle re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 205–214.

[33] H. Liu, Y. Tian, Y. Wang, L. Pang, and T. Huang, "Deep relative distance learning: Tell the difference between similar vehicles," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2167–2175.

[34] X. Liu, W. Liu, H. Ma, and H. Fu, "Large-scale vehicle re-identification in urban surveillance videos," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2016, pp. 1–6.

[35] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3230–3238.

[36] B. He, J. Li, Y. Zhao, and Y. Tian, "Part-regularized near-duplicate vehicle re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3992–4000.

[37] H. An, H. Fan, K. Deng, and H.-M. Hu, "Part-guided network for pedestrian attribute recognition," in *Proc. IEEE Vis. Commun. Image Process. (VCIP)*, Dec. 2019, pp. 1–4.

[38] 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 *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6131–6140.

[39] Y. Bai, Y. Lou, F. Gao, S. Wang, Y. Wu, and L.-Y. Duan, "Group-sensitive triplet embedding for vehicle reidentification," *IEEE Trans. Multimedia*, vol. 20, no. 9, pp. 2385–2399, Sep. 2018.

[40] Y. Jin, C. Li, Y. Li, P. Peng, and G. A. Giannopoulos, "Model latent views with multi-center metric learning for vehicle re-identification," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 3, pp. 1919–1931, Mar. 2021.

[41] C. Yan et al., "ZeroNAS: Differentiable generative adversarial networks search for zero-shot learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 9733–9740, Dec. 2022.

[42] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. H. Hoi, "Deep learning for person re-identification: A survey and outlook," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 6, pp. 2872–2893, Jun. 2022.

[43] Y. Lin, Y. Gou, Z. Liu, B. Li, J. Lv, and X. Peng, "COMPLETER: Incomplete multi-view clustering via contrastive prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 11169–11178.

[44] A. Bochkovskiy, C.-Y. Wang, and H.-Y. Mark Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.

[45] L. He, X. Liao, W. Liu, X. Liu, P. Cheng, and T. Mei, "FastReID: A PyTorch toolbox for general instance re-identification," 2020, *arXiv:2006.02631*.

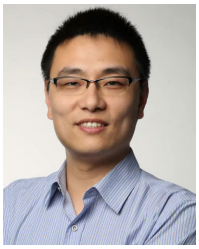
[46] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

Aihua Zheng received the B.Eng. and the integrated master's and Ph.D. degrees in computer science and technology from Anhui University, China, in 2006 and 2008, respectively, and the Ph.D. degree in computer science from the University of Greenwich, U.K., in 2012. She is currently an Associate Professor of artificial intelligence with Anhui University. Her current research interests include computer vision and artificial intelligence, especially on person/vehicle re-identification, audio-visual learning, and multi-modal and cross-modal learning.

Chaobin Zhang received the B.Eng. degree in computer science and technology from Nanchang Hangkong University, Nanchang, China, in 2020. He is currently pursuing the master's degree in computer science and technology with Anhui University. His current research interests include vehicle re-identification and multimodal fusion.

Chenglong Li received the M.S. and Ph.D. degrees from the School of Computer Science and Technology, Anhui University, Hefei, China, in 2013 and 2016, respectively. From 2014 to 2015, he was a Visiting Student with the School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China. He was a Postdoctoral Research Fellow with the Center for Research on Intelligent Perception and Computing (CRIPAC), National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing, China. He is currently an Associate Professor with the School of Artificial Intelligence, Anhui University. His current research interests include computer vision and deep learning. He was a recipient of the ACM Hefei Doctoral Dissertation Award in 2016.

Jin Tang received the B.Eng. degree from the School of Automation, Anhui University, Hefei, China, in 1999, and the Ph.D. degree from the School of Computer Science, Anhui University, in 2007. He is currently a Professor with the School of Computer Science and Technology, Anhui University. His current research interests include computer vision, pattern recognition, machine learning, and deep learning.



Chang Tan received the Ph.D. degree in computer science from the University of Science and Technology of China. He is currently the Vice President of the Smart City Business Group, iFLYTEK; the President of the Big Data Research Institute, iFLYTEK; and a system architect (senior). He is currently in charge of the research and development and application promotion of big data core technologies in smart cities, smart transportation, computational advertising, and personalized recommendations with IFLYTEK. He is a member of the Standing Committee of the Big Data Expert Committee of the China Computer Society and a member of the editorial board of the academic journal *Big Data*.