DAIR-V2XReid: A New Real-World Vehicle-Infrastructure Cooperative Re-ID Dataset and Cross-Shot Feature Aggregation Network Perception Method

Hai Wang, *Senior Member, IEEE*, Yaqing Niu, Long Chen, Yicheng Li, Miguel Angel Sotelo, *Fellow, IEEE*, Zhixiong Li, *Senior Member, IEEE*, and Yingfeng Cai, *Senior Member, IEEE*

***Abstract*— As an emerging research field, vehicle re-identification (Re-ID) can realize identity search between the vehicles, which plays an important role in the over-the-horizon perception of Vehicle-Infrastructure Cooperative Autonomous Driving (VICAD). At present, due to the lack of data sets, the relevant research on Vehicle-Infrastructure Cooperative (VIC) Re-ID can only be evaluated in the cross-view monitoring test set which leads to the lack of persuasion of the research. Therefore, based on the DAID-V2X dataset of Tsinghua University, this paper constructs a VIC Re-ID dataset "DAIR-V2XReid" from real vehicle scenarios through vehicle-road end target tag association, thereby making it better applicable to the research of VIC Re-ID. Owing to different task scenarios, existing algorithms trained on monitoring test sets are unable to effectively complete the Re-ID task in this new dataset. Therefore, Cross-shot Feature Aggregation Network (CFA-Net) is also proposed in this paper, to tackle the case where a vehicle becomes unrecognizable due to a large change in its visual appearance across different cameras. Firstly, we put forward a camera embedding module and add it to the Backbone, to group different cameras and solve the problem of cross-shot perspective mutation. Secondly, in order to address the situation where background and vehicle division are not distinguishable, we propose a cross-stage feature fusion module, which integrates low-order semantics with high-order semantics. Finally, we use multi-directional attention network to achieve the final feature extraction. The experimental results show that our proposed CFA-Net method achieves new state-of-the-art in DAIR-V2XReid, with mAP of 58.47%.**

***Index Terms*— Vehicle-infrastructure cooperative, Re-ID, datasets, automatic driving.**

Manuscript received 1 August 2022; revised 30 January 2023, 25 November 2023, and 24 January 2024; accepted 16 February 2024. Date of publication 6 March 2024; date of current version 1 August 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 52225212, Grant U20A20333, and Grant 52072160; and in part by the Key Research and Development Program of Jiangsu Province under Grant BE2020083-3. The Associate Editor for this article was X. Li. *(Corresponding author: Yingfeng Cai.)*

Hai Wang and Yaqing Niu are with the School of Automotive and Traffic Engineering, Jiangsu University, Zhenjiang 212013, China (e-mail: wanghai1019@163.com; 774716392@qq.com).

Long Chen, Yicheng Li, and Yingfeng Cai are with the Automotive Engineering Research Institute, Jiangsu University, Zhenjiang 212013, China (e-mail: chenlong@ujs.edu.cn; liyucheng070@163.com; caicaixiao0304@126.com).

Miguel Angel Sotelo is with the Department of Computer Engineering, University of Alcalá, Alcalá de Henares, 28801 Madrid, Spain (e-mail: miguel.sotelo@uah.es).

Zhixiong Li is with Yonsei Frontier Laboratory, Yonsei University, Seoul 03722, Republic of Korea, and also with the Faculty of Mechanical Engineering, Opole University of Technology, 45758 Opole, Poland (e-mail: zhixiong.li@yonsei.ac.kr).

Digital Object Identifier 10.1109/TITS.2024.3367723

I. Introduction

VEHICLE Re-ID aims to find the images of same vehicle identity taken by different cameras. Considering the continuous development of autonomous driving technology, vehicle Re-ID has broad application prospects in intelligent transportation systems [1], [2], and has become an essential technology to realize the automatic driving. For vehicle Re-ID tasks, comprehensive, reliable, and fair dataset is helpful to objectively evaluate the performance of a vehicle Re-ID algorithm, which is one of the key tasks in realizing the vehicle Re-ID. With the development of computer vision, a large number of vehicle Re-ID datasets have emerged [3], [4], [5], [6], and which have in turn brought many excellent algorithms for vehicle Re-ID tasks [7], [8], [9]. However, it is difficult for these datasets to effectively adapt to the field of autonomous driving. The root cause is that the existing vehicle Re-ID datasets have been originally developed for the purpose of security, and hence, they commonly adopted roadside cameras for data acquisition. Nevertheless, in automatic driving scenarios, the images taken by roadside cameras have the following two shortcomings: (1) Due to uncertain practical factors such as variation in the number of cameras and randomness of the path of target vehicles, there exist many data occlusion, single perspective and other problems in the dataset. (2) Due to the problems of shadow reflection, color temperature and brightness interference in the real scene, the background has great interference to the foreground in some complex situations. However, the roadside cameras with single view angle are difficult to solve such problems effectively.

At the same time, for the current automatic driving technology [2], the popularity of on-board camera is extremely high in this field, but it usually has problems such as occlusion in

the process of data acquisition, which makes it impossible to realize the environment perception without any dead corners. In this regard, Bishop et al. [10] pointed out that the single vehicle intelligence cannot well-solve the problem of automatic driving. Akki et al. [11] indicated that in addition to vehicle's own perception ability, there should also be roadside equipment to assist in jointly completing the automatic driving tasks, thus highlighting the high application value of Vehicle-Infrastructure Cooperative Autonomous Driving (VICAD). VICAD refers to the cooperation between vehicles and infrastructure, providing the vehicles with a global perspective far beyond the current field of vision along with the information covering the low vision areas, to effectively complete the target detection, recognition, tracking and other tasks to ensure correct and safe subsequent control decisions.

Considering the above discussion, we herein construct a new vehicle Re-ID dataset "DAID-V2XReid". This dataset is constructed based on the Vehicle-Infrastructure Cooperative (VIC) DAID-V2X dataset [12] for real vehicle scenes proposed by Tsinghua University. Furthermore, the data collected by vehicle cameras and roadside cameras are used to complete the vehicle Re-ID tasks. DAID-V2XReid dataset offers the following advantages: (1) Using vehicle camera as the mobile end and road end camera as the fixed end, the vehicle camera can make up for the shortcomings of a fixed camera, reduce vehicle occlusion, thus making the collected perspective of the same vehicle more comprehensive. (2) Due to highly inconsistent characteristics of the two cameras, the same vehicle collected by two cameras can be slightly different even under the same perspective, thereby adding diversity to the dataset. (3) The dataset is obtained from real scenes by using the two devices together, which can get different backgrounds for the same vehicle under different devices and obtain variable background and increase the background diversity.

At present, the existing vehicle Re-ID methods can be roughly divided into two categories. The first category of methods we call the detail capture approach [7], [8], [13], which attempts to separately handle the tasks of vehicle feature extraction and re-identification. This method first uses network training to obtain the easily distinguishable vehicle features (such as color, perspective, license plate, etc.), and then performs the subsequent Re-ID training. Although this method can achieve satisfactory results, it is time-consuming and resource-consuming due to the separate treatment of the two tasks. In contrast, the second category of methods [9], [14], [15], [16] provides feature fusion methods, which are further divided into distance metrics [9], [15], part segmentation [14], [16], etc. This method performs only the Re-ID task in an end-to-end manner, and most of them innovatively process high-order semantics. This type of method has a simple structure and convenient calculation, but often ignores the importance of details, and inaccurate high-order semantics may lead to serious errors in feature fusion.

In addition, the existing vehicle Re-ID algorithm on the DAIR-V2XReid dataset also has the following shortcomings: (1) It is difficult to distinguish similar vehicles from the same perspective, as shown in Fig 1(d). (2) The same vehicle has significant differences in different perspectives, as shown in Fig 1 (b) (c). The differences in device, mounting positions, field of view, and other aspects between on-board cameras and roadside cameras result in differences in brightness and color in the final image, as shown in Fig 1 (a) (c).

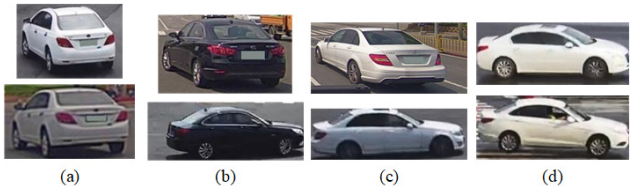


Fig. 1. Comparison between the images of DAIR-V2XReid dataset.

To solve the above problems, we propose a new network labelled as Cross-shot Feature Aggregation Network (CFA-Net), which consists of three modules. First of all, to deal with the different angles of view of different cameras, we use the camera label in the dataset to propose a camera embedding module, to embed the camera information into the backbone network and achieve a simple grouping of different cameras. Then, in order to avoid the disclosure of local information of high-order semantics, we propose a cross-stage feature fusion module, which integrates high-order semantics with low-order semantics, and distinguishes between the background and the vehicle. Finally, a multi-directional attention module is proposed to obtain the spatial features of the target and further refine them for the final Re-ID task.

The main contributions of this paper include the following three aspects:

1) To overcome the lack of datasets, we established a VIC Re-ID dataset "DAIR-V2XReid" in real vehicle scenarios.

2) For vehicle Re-ID tasks, we designed CFA-Net to solve the problem related to the greatly varying vehicle perspective in cross-shot scenes. Through the cooperation of camera embedding module, cross-stage feature fusion module, and multi-direction attention module, a better vehicle feature matching effect has been demonstrated.

3) To validate the performance of proposed model, we used the DAIR-V2XReID dataset with our proposed model, and the performance reached the highest ever reported. Moreover, to verify the generalization ability of the model, we also carried out experiments on VeRi776 dataset and obtained a good accuracy.

II. Related Work

A. Vehicle Re-ID

The existing vehicle Re-ID technology can be divided into four categories. (1) Method based on local area learning: This method usually utilizes local areas to provide the identification clues of vehicles and obtain the local features. Likewise, He et al. [8] first used YOLOv1 [17] to detect the three parts of the vehicle that are easy to identify from the image (window, lamp and vehicle brand), and proposed a simple and effective ROI projection method based on the region of interest method of object detection, that is, to combine the detection branch and the Re-ID task to complete the vehicle Re-ID task. Meng et al. [7] proposed to use a segmentation algorithm [18]

to train four view masks (front, back, top and side), and then aligned local view features through the masks using an average pooling layer based on a global feature map. (2) Methods based on attentional learning: The attentional mechanism usually makes the network focus on the most critical information of the current task, and reduces network's attention to other information. Rao et al. [19] studied the attention mechanism and proposed a counterfactual attention learning module to analyze the impact of learned visual attention on the network prediction, and maximized the learning of useful information by the network for vehicle Re-ID. Subsequently, Zhu et al. [20] proposed a dual cross-attention learning algorithm to enhance the interaction between the image pairs. (3) Method based on metric learning: Such method can learn a feature space and convert all the data into feature vectors to distinguish among the data. Cheng et al. [21] improved the triplet loss function to shorten the distance between the same ID and widen the distance between the different IDs to improve the accuracy. Yan et al. [22] proposed multi-grain ranking loss to distinguish between the vehicles with similar appearance. Chu et al. [9] proposed a viewpoint-aware metric learning approach to learn the extreme perspective variation problem. (4) Methods based on Generative Adversarial Network (GAN) [6], [23], [24]. Nowadays, with rapid development of neural networks, more and more people are using new methods, such as GAN, for Re-ID. Among them, Zhou et al. [23] dealt with the viewpoint problem by using a GAN that generates the opposite features. Zhou et al. [25] used Long and Short-Term Memory (LSTM) to simulate of continuous view transformation of vehicles, and used the adversarial architecture network to enhance the training. Progressively, Lou et al. [24] proposed an end-to-end embedding adversarial learning network to generate the local samples in the embedding space, to improve the recognition ability and robustness of the algorithm.

Different from the existing work, we design a Cross-shot Feature Aggregation Network (CFA-Net) for vehicle Re-ID tasks without any additional labeling. CFA-Net is not only simple in structure, but also achieves a high precision.

B. Vehicle Re-ID Dataset

In recent years, many research studies have been reported for vehicle Re-ID, and similarly, more and more vehicle Re-ID datasets have been surfaced. Currently, the vehicle Re-ID was dominated by three datasets: VehicleID [4], VeRi776 [5] and Veri-wild [6]. VehicleID [4] and its extended [22] datasets were the real-world data obtained using cameras in different scenes in a short time. However, this dataset only covered the two perspectives i.e., front and rear of the vehicle, which is not suitable for real-world application. Moreover, the dataset did not mark the camera ID, and thus, the scene transformation could not be distinguished. The dataset VeRi776 [5] was photographed in 24 hours covering an area of 1.0 square km. Veri-wild [6] conducted long-term continuous shooting with the camera, taking into account the weather and lighting issues in the real vehicle scene. The above two datasets (VeRi776 and Veri-wild) contained a large number of vehicle models with sufficient license plates and time-space labels. However, since these were shot and selected only by the lens at the road end, they represented only the data of a single device, which is not applicable for the Vehicle-Infrastructure Cooperative (VIC) scenarios. In order to make up for these deficiencies, we construct the first cross-shot VIC real scenes dataset: DAIR-V2XReid.

C. VIC Perception and Dataset

VIC technology aims to realize global perception of road target information through information interaction between vehicles and infrastructure, thus alleviating the shortcomings of limited scope and frequent occlusion in one-car perception. However, the research on the VIC perception [26], [27], [28] had just started, and there was a lot of research space that needs to be explored. Kim et al. [26] proposed a multi-mode collaborative perception system for the first time, realizing collaborative driving such as front-collision warning, automatic hiding and obstacle avoidance. More recently, Li et al. [27] proposed a novel distilled collaboration graph to realize a trainable adaptive collaboration, in an attempt to better improve the performance and bandwidth of multi-device perception.

Additionally, there are two main methods for the collection of datasets, one of which [27], [29] was to use simulation environments (such as Sim4cv [30], Carla [31] and other simulators) to generate well-annotated large-scale datasets. This method is simple and convenient, and can quickly generate all kinds of required data for free. Another approach is to take the real scene data and the make datasets. In this regard, Maalej et al. [32], [33] combined their own data with KITTI dataset ··· [34] to obtain a dataset of V2V real scenes. Yu et al. [12] obtained the DAIR-V2X dataset in the real scene to compensate for the discrepancies between real and virtual scenes. This dataset is detailed with category information and can complete tasks such as target detection.

On the basis of DAIR-V2X dataset, we assign IDs to vehicles and produce DAIR-V2XReid dataset, to obtain the first-ever VIC Re-ID dataset in real vehicle scenes, which is then used to complete the cross-shot cooperative perception Re-ID task.

III. DAIR-V2XREID

A. Data Acquisition

1) Sensors: The dataset was collected at 28 intersections in the Beijing Advanced autonomous driving Demonstration zone, and four pairs of high-resolution cameras were deployed at each intersection as the road end devices. At the same time, the vehicle was equipped with a forward-looking high-quality camera, serving as the vehicle end device, to jointly complete the acquisition process.

2) Data Processing: Since the two devices jointly collect the data for Re-ID, the data collected by the two devices should be time matched. If the time difference between the data of two devices was less than 10ms, it was recorded as the synchronization time.

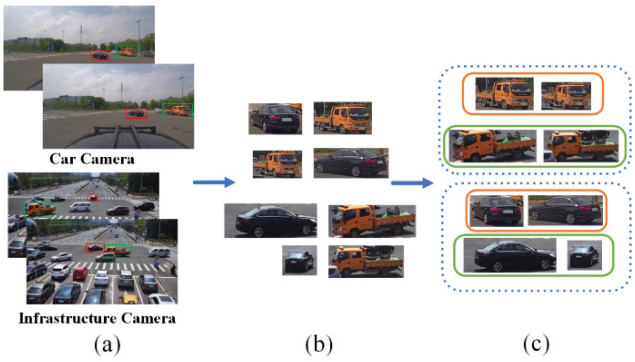


Fig. 2. DAIR-V2XReid dataset construction process. (a) Data matching between the two devices, and selection of the vehicles that meet the matching requirements. (b) Vehicle image capture. (c) Assigning values to the vehicle ID and camera ID, respectively. The orange box represents the vehicle camera, the green box represents the roadside camera.

3) Data Labeling: We manually selected the pictures of same vehicles in the data of the two devices, and assigned them with a same vehicle label value. At the same time, we also marked the camera ID on the data, setting the vehicle camera ID as 0 and the roadside camera ID as 1. The specific steps were shown in Fig 2. In the end, we obtained 205 matched vehicles with at least 2 photos in each group, resulting in a total of 2556 photos. We followed the sample distribution convention of the existing vehicle Re-ID datasets and divided the vehicle sample into two sub-datasets i.e., train and gallery, with a ratio of 2:1. We then randomly selected one image under each camera ID in the gallery dataset to generate the query dataset.

Among them, the train dataset was used for training, and the gallery and query datasets were used for testing.

B. Dataset Contribution

The proposed DARI-V2XReid dataset has the following outstanding contributions:

1) First Dual Device Real Scene Sampling: To overcome occlusion, a data set extracted using both vehicle and roadside cameras is proposed for vehicle Re-ID for the first time. Compared with the data in virtual scene, this dataset could essentially provide a more realistic and comprehensive perspective of the vehicle, as close to the real driving conditions as possible, thus narrowing the gap between theory and practice.

2) Complex Capture Conditions: Each vehicle ID in the dataset appeared at least once under the two lenses. Moreover, the vehicle camera and the roadside camera acquired images from different angles, which would not only allow the data to have variable background, resolution and angle of view, but also bring some local random occlusion to the data, thereby improving the model robustness and enabling effective execution of the following tasks such as vehicle cross-lens tracking [35] and vehicle behavior modeling [36].

3) Better Privacy: Before the entire dataset was published, all the information that could be suspected of violating the privacy was blocked, including license plates, faces, road signs and so on, to protect the public privacy to the greatest extent.

Dataset download address: https://github.com/Niuyaqing/DAIR-V2XReid.git. Our website has information about the dataset's annotation in detail.

IV. Methodology

A. Review

To solve the problems caused by cross-lens Re-ID of vehicles, including large differences in the appearance and color of the same vehicle and small differences in the appearance of similar vehicles, and meet the needs of VIC Re-ID, we proposed a Cross-shot Feature Aggregation Network (CFA-Net), and the proposed network framework is shown in Fig 3. First of all, in Section IV-B, we put forward a camera embedding module, which is used to solve the problems related to large variation in the appearance of same vehicle captured by the two cameras at vehicle end and road end. Subsequently, to better separate the background and target and extract more obvious identifiable features, in Section IV-C, we designed a cross-stage feature fusion module to fuse low-order features with high-order features. Finally, in Section IV-D, we used the features extracted in Section IV-C, to propose a multi-directional attention module for feature enhancement.

B. Camera Embedded Module

In Re-ID dataset, there would be huge differences in the ID of the same vehicle obtained by the two devices, due to variation in camera angle and other reasons. Because the two-dimensional pictures we took had few angles, it was impossible to completely describe the perspective of vehicles and accurately obtain the identity characteristics of objects in a three-dimensional space. This made it challenging for the model to identify the same vehicle under different cameras, after the training was completed. To solve this issue, we proposed a camera embedding module, which used a camera ID to group and embed the camera information into features for aggregation, so as to learn the characteristics of 3D object. The embedding position is illustrated in Fig 3.

Specifically, there were N cameras in the dataset, denoted as $ID_r, r \in [1, N]$. We used a randomly generated sequence to initialize the module. After initialization, the camera embedding was realized as $E_c \in R^{N_C \times A}$, where $A = H \times W$, and H and W respectively represent the height and width of the corresponding picture under the current V_0 channel. Therefore, for a photo img_i taken by a certain camera ID_r, the corresponding camera embedded feature could be represented as Ec_i^r, as shown in Fig 4. Finally, the camera embedding feature E_c was input into the Backbone, expressed as:

$$V_0' = V_0 + \lambda E_C[r], \tag{1}$$

where V_0 is a low-order feature in the Backbone and λ is a hyperparameter of the balanced camera embedding module. Through the feature embedding of camera embedding module, each input image could have its own camera position code, thus bringing in more discriminative features for the subsequent feature extraction.

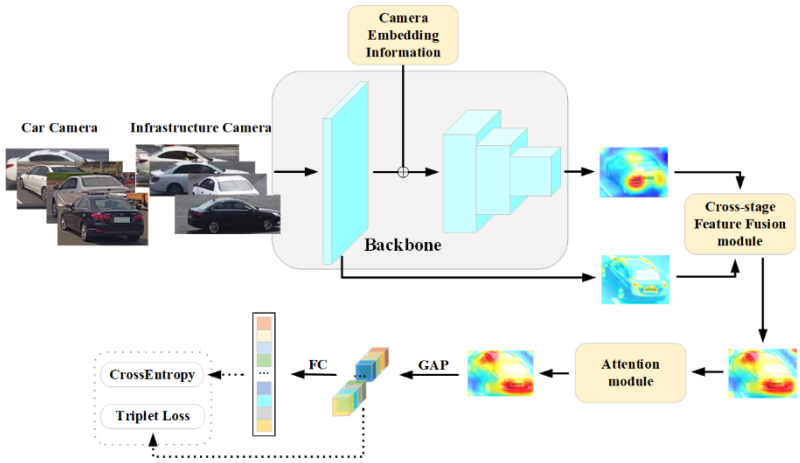


Fig. 3. The structure of Cross-shot Feature Aggregation Network. Firstly, the camera embedded module is added to the Backbone for feature extraction. Secondly, we have preserved a low-level feature of Backbone, fused it with the global characteristics, and jointly completed the cross-phase feature fusion. Finally, multi-directional attention feature acquisition is carried out for the obtained features.

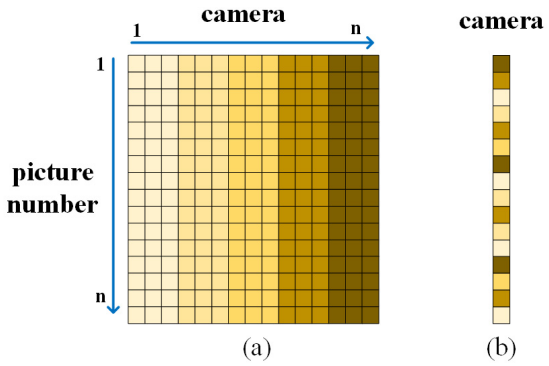


Fig. 4. Camera Embedded Module. (a) represents the generated random sequence, where the horizontal direction represents the number of cameras and the vertical direction represents the number of input images. (b) represents the camera embedding features generated from the camera IDs and the sequence in (a).

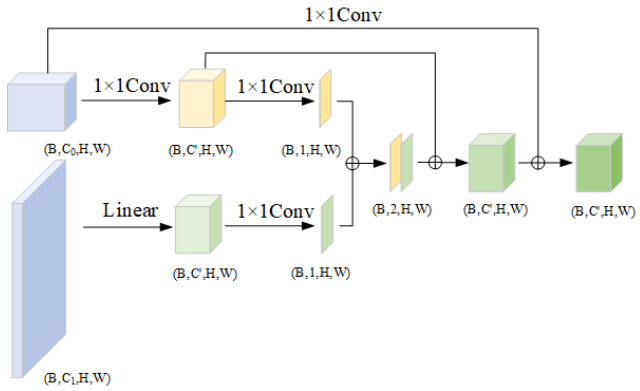


Fig. 5. Cross-stage feature fusion module.

C. Cross-Stage Feature Fusion Module

Theoretically, the background features are significantly different from the foreground features, and hence, the model should have a better fitting ability in case of sparse samples. However, in actual use, although high-level semantics had the advantage of clear features, they led to the mixing of environmental information and the loss of vehicle information, resulting in the final target feature localization is not accurate. On the other hand, although the low-level semantic features were not clear enough, they contained a lot of location information and background features, as opposed to high-level semantics. Therefore, this paper proposed a cross-stage feature fusion module to realize the feature fusion of low-level semantics and high-level semantics, and clearly distinguished the background from the target vehicle. In this way, we mitigated the possibility of high-level semantics falling into non-critical areas and improved the anti-interference ability of the model. The specific process of this module is shown in Fig 5.

To practically realize the above idea, we first obtained the low-level semantic V_0 and high-level semantic V_1 features in the model. However due to their different dimensions, we needed to fuse them to achieve the final feature extraction. To solve this problem, we first mapped the two features into a same space and then fused them. For low-level semantics, we first multiplied the two dimensions of height and width, and then changed the multiplied dimension through a linear transformation, denoted as V_{01}. Meanwhile, high-order semantics used 1×1 Conv for the reduction of channel dimension, to obtain the same number of channels as low-order semantics, while reducing the amount of computation, and the resultant is denoted as V_{11}. Next, concatenate the two vectors into a whole, expressed as:

$$V'_{C1} = \mathbb{C}\left(V_{01}, V_{11}\right), \tag{2}$$

where $\mathbb{C}$ represents the concatenation operation. We observed that in high-level semantics, although the perception ability for details was not enough, the obtained features were less noisy and more specific, so we chose to focus on high-level features during the fusion, to fully utilize the advantages of high-level semantics. After the splicing was completed, we channel-transformed V'_{C1} again through 1×1 Conv to make it have the same number of channels as that of V_{11}, and obtained V'_{C2}. Next, we added V'_{C2} to higher-order semantics V_{11} to obtain V_2, which fully reflects the advantages of higher-order semantics.

$$V_2 = \mathbb{C}\left(conv\left(V'_{C1}\right), V_{11}\right), \tag{3}$$

This concluded the operation of cross-stage feature fusion module, where the current features not only integrated the

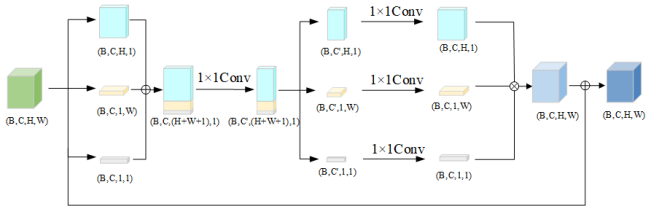


Fig. 6. Flow chart of Multi-directional attention module.

background information of low-level semantics, but also did not lose the clear features of high-level semantics.

It is worthwhile to note that such a simple integration would inevitably lead to confusion of features. Therefore, we then passed the features into a multi-direction attention module for further refinement.

D. Multi-Directional Attention Module

Following the cross-stage feature fusion, the background of the feature was integrated with the scene object information, but we observed that there was still some information confusion in the feature, which hindered the realization of perfectly clear and definite features. Therefore, we designed a multi-direction attention module to further refine the features and complete the final feature extraction. This module consisted of two parts: spatial feature coding and attention acquisition, and the relevant process is shown in Fig 6.

For final feature extraction, the common approach is to use a global pooling layer for global encoding and capture the global information. However, this will ignore the location information of features. For the vehicle Re-ID task, the recognized object is a three-dimensional model, where spatial information plays a key role in obtaining such three-dimensional structure. Thus, we decided to add the necessary spatial information to obtain more accurate vehicle appearance features. First, we decomposed the global pooling layer, and calculated the horizontal direction, vertical direction and information deviation in the feature map. Then, we obtained the position encoding along the horizontal and vertical directions, and the global encoding of the overall feature. Subsequently, the spatial information was encoded in the horizontal, vertical and offset directions, and the expressions are:

$$V_c^H(H) = \frac{1}{W}\sum_{i=1}^{W} V_2(h,i), \tag{4}$$

$$V_c^W(W) = \frac{1}{H}\sum_{j=1}^{H} V_2(j,w), \tag{5}$$

$$V_c^I(I) = \frac{1}{W}\frac{1}{H}\sum_{i=1}^{H}\sum_{j=1}^{W} V_2(j,i), \tag{6}$$

Among them, V_2 represents the output characteristics of the Cross-stage Feature Fusion Module. The above three changes were respectively aggregated along the three directions i.e., horizontal, vertical and information deviation, and three perceptual feature maps were obtained. In this way, the features could be encoded with the accurate spatial information along their respective attention directions, which was helpful for locating more important features and realizing the acquisition of position information points of the 3D model.

After encoding the spatial information, the model could use the attention module to achieve the aggregation of object features in the channel dimension. First, to reduce the model complexity, the features were dimensionally reduced to obtain more discriminative channels. Specifically, the spatially encoded features were first concatenated to obtain a multi-feature fusion representation vector, and then a set of 1×1 Conv was used for dimensionality reduction of the representation vector. The relevant calculation process is given as follows:

$$f = conv\left(\mathbb{C}\left(V_c^H, V_c^W, V_c^I\right)\right), \tag{7}$$

where, $\mathbb{C}$ represents the concatenation operation along the spatial dimensions of H, W and I, $f \in \mathbb{R}^{\frac{C}{r}\times(H+W+1)}$ is the intermediate feature map, and r is the control channel dimension compression ratio. To reduce the computation cost, we used a smaller compression ratio r to reduce the number of channels, with value $r = 32$.

Then, we split f into three separate tensors $f_h \in \mathbb{R}^{C/r\times H}$, $f_w \in \mathbb{R}^{C/r\times W}$, and $f_I \in \mathbb{R}^{C/r\times 1}$ along the resulting spatial dimension, which were then transformed into tensors with the same number of channels as that of input V_2 using three 1×1 Conv. In this way, correlations were generated between the channels of the three features. The relevant formulation is as follows:

$$\begin{cases} g_h = \sigma\left(conv\left(f_h\right)\right), \\ g_w = \sigma\left(conv\left(f_w\right)\right), \\ g_I = \sigma\left(conv\left(f_I\right)\right), \end{cases} \tag{8}$$

where σ represents the sigmoid function. Finally, to make full use of the spatial information encoded in the features, enable the final features to have a more correct localization, and obtain the attention feature, we fused the three tensors, and the corresponding output V can be expressed as:

$$V = (g_w \times g_I) \times g_h \times V_2. \tag{9}$$

Furthermore, we introduced the feature V into the loss function through a linear transformation to calculate the loss. Finally, the obtained loss was continuously optimized by back propagation.

V. Experiments

A. Experimental Settings

Dataset: We have carried out a large number of experiments on the DAIR-V2Xreid dataset and the mainstream VeRi776 vehicle reidentification dataset [5]. The VeRi776 dataset is based on images taken by 20 roadside cameras, covering an area of 1.0 km^2, over 24 hours, making it suitable for vehicle Re-ID studies. However, it does not include the vehicle cameras. Table I shows the image distribution of the two datasets.

1) Backbone: We selected the Resnet-50 [37] pre-trained on ImageNet [38] as our backbone network in order to achieve the feature extraction of vehicle look. We remove the number of layers after pooling5 in Resnet-50 and add an ibn-a block after it [39].

TABLE I
DISTRIBUTION OF THE NUMBER OF FOLDERS FOR THE TWO DATASETS

Dataset	Classify	Total	Train	Gallery	Query
DAIR-V2XReid	IDs	205	139	66	66
	Images	2556	1669	754	133
VeRi776	IDs	776	576	200	200
	Images	51035	37778	11579	1678

2) Implementation Details: We randomly cut the input image to a size of 256 × 256, and used random erasing, horizontal flip and other methods to enhance the data. In addition, we set the batch size to 48, resulting in 4 images per ID, and trained the network for 80 epochs using the cross-entropy loss function [40] and the triplet loss function [21]. Progressively, we used the SGD optimizer [40].

3) Evaluation Index: After training, to verify the effectiveness of the proposed network in solving the cross-camera vehicle matching problems, we used two evaluation metrics, mean Average Precision (mAP) [41] and Cumulative Matching Characteristics (CMC) [41], for the testing.

4) Training: All of our experiments were built on Pytorch 1.8 deep learning framework and performed on NVIDIA RTX 2080Ti GPU.

B. Ablation Experiments

1) Performance Analysis of Each Module: To validate the performance of each module, we conducted ablation experiments on the DAIR-V2XReid and VeRi776 datasets, as illustrated in Table II.

a) Effectiveness of Camera Embedded Module (CEM): To verify its effectiveness, we first put our camera embedding module into the Baseline for experiment. This is shown in the line 4 of Table II. Comparing our CFF with the Baseline, for VeRi776 and DAIR-V2XReid, the Rank-1 improved by 1.61% and 11.28%, respectively, and the mAP improved by 1.86% and 7.03%.

b) Effectiveness of Cross-stage Feature Fusion Module (CFF): As shown in the line 3 of Table II, compared with the Baseline, our CFF improved the Rank-1 by 2.50% and 6.77%, and mAP by 3.29% and 3.80% on VeRi776 and DAIR-V2XReid datasets, respectively.

c) Effectiveness of Multi-directional Attention Module (MAM): Evident from the line 2 of Table II, compared with the Baseline, our MAM improved the Rank-1 by 1.61% and 7.52%, and mAP by 2.48% and 3.89% on VeRi776 and DAIR-V2XReid, respectively.

d) Effects of Different Installation Stages of CEM on Accuracy: We also investigated the impact of CEM on overall accuracy under different stages of Resnet-50, as given in Table III. We tested on two datasets, VeRi776 and DAIR-V2XReid, and as can be seen from Table III, both datasets achieved the best results at stage 2.

e) Effects of Different Installation Stages of CFF.: In table IV, we studied the impact of CFF on overall accuracy under different stages of Resnet-50, using two datasets, VeRi776 and DAIR-V2XReid. We could see that the highest accuracy and the best effect were obtained in stage1. However, with the backward shift of stage, the lower-order semantics obtained by the module was closer to the higher-order semantics. When they were fused with higher-order semantics, the combination of background and objects was often ignored, resulting in a worse accuracy.

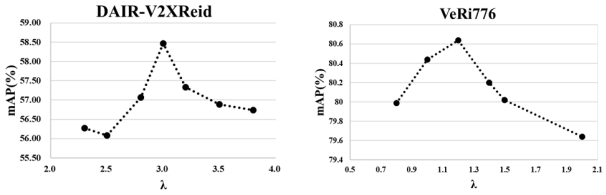


Fig. 7. Broken line comparison diagram of mAP under different λ in two datasets.

f) Effects of the Number of CFFs on Accuracy: In Table V, we examine the effect of different amounts of CFF on the final performance. It was found that with the increase of the number of the CFF module, the calculation speed slowed down and the accuracy also reduced.

g) CEM Parametric Analysis: We analyzed the influence of the coefficient λ of the CEM module on the overall performance. Under different coefficient the model accuracy was varied differently, however the change in overall performance was not significant. From Figure 7, we can see that the DAIR-V2XReid dataset works best when λ =3.0, while the VeRi776 dataset works best when λ =1.2.

2) Visualization Analysis:

a) Visualization of Activation Maps: To intuitively understand the learning method of our model, we visualized it on two datasets, VeRi776 and DAIR-V2XReid, as shown in Fig 8, where we show the original image, the baseline heatmap and the heatmap of our network. Certainly, the results showed that our method could better focus on the vehicle, avoided the interference of background information, and encouraged the module to focus on the more discriminative information of the vehicle (such as lights, windows, etc.), without limiting the exploration of overall information.

b) Visualization of Retrieval Results: We visualized the retrieval results of samples by different methods and verified the effectiveness of the proposed module (Fig. 9). Evidently, similar-looking vehicles are difficult to be distinguished using the Baseline model. Contrarily, since our CFA-Net allowed a better focus on local information, it could better extract the discriminative information, and could also easily judge the vehicles with similar appearance.

c) Vehicle Road Perception Effect: We put the correct Re-ID results of the DAIR-V2XReid dataset into the original image for vehicle road perception effect comparison. We show the result with the same object ID of matching, as shown in Fig 10. The left side of Fig 10 shows the correct result of vehicle Re-ID. Fig 10 (a) shows the data collected from the perspective of a single vehicle. In this perspective, the vehicle could only perceive two vehicles at present. However, with the addition of road-end devices (Fig 10 (b)), the perceived receptive field increased, and the number of photographed vehicles also increased significantly. Through the matching of

TABLE II
COMPARISON OF DIFFERENT GROUPINGS ON VERI776 AND DAIR-V2XREID DATASETS

Method			VeRi776		DAIR-V2XReid	
CEM	CFF	MAM	mAP	Rank-1	mAP	Rank-1
×	×	×	77.06	94.40	51.44	48.12
×	×	√	78.55	95.47	52.64	52.63
×	√	×	80.35	96.90	55.24	54.89
√	×	×	79.54	96.01	55.33	55.64
√	√	×	78.92	96.36	56.33	57.14
√	×	√	78.34	96.54	52.83	48.12
×	√	√	78.01	95.71	53.21	48.12
√	√	√	**80.67**	**96.90**	**58.47**	**59.40**

Among them, the first row is the result of Baseline training, and the last row is the result of our method, CFA-Net training.

TABLE III
COMPARISON OF CEM AT DIFFERENT STAGES

Stage	VeRi776		DAIR-V2XReid	
	mAP	Rank-1	mAP	Rank-1
No	78.01	95.71	51.44	48.12
stage1	79.91	96.60	56.82	55.88
stage2	**80.67**	**96.90**	**58.47**	**59.40**
stage3	78.70	96.62	55.32	52.63
stage4	79.04	95.89	53.99	54.89

TABLE IV
COMPARISON OF CFF AT DIFFERENT STAGES

Stage	VeRi776		DAIR-V2XReid	
	mAP	Rank-1	mAP	Rank-1
No	78.34	96.54	52.83	48.12
stage1	**80.67**	**96.90**	**58.47**	**59.40**
stage2	77.99	96.42	57.46	54.14
stage3	78.65	96.78	54.19	51.88
stage4	78.16	96.25	52.33	47.37

TABLE V
COMPARISON OF CFFS AT DIFFERENT QUANTITIES

Nums			VeRi776		DAIR-V2XReid	
Stage2	Stage3	Stage4	mAP	Rank-1	mAP	Rank-1
√			**80.67**	**96.90**	**58.47**	**59.40**
√	√		78.08	96.13	58.00	56.39
√	√	√	77.56	96.07	54.53	51.88

two devices, problems such as limited sensing range under the single-vehicle perspective can be avoided, and control and decision-making tasks such as self-vehicle positioning, judgment of road conditions ahead, subsequent vehicle tracking, and vehicle path planning can be completed effectively.

C. Comparison With The State-of-the-Art

In this section, we compare the proposed method, CFA-Net, with other state-of-the-art methods on two vehicle Re-ID benchmarks: DAIR-V2XReid and VeRi776.

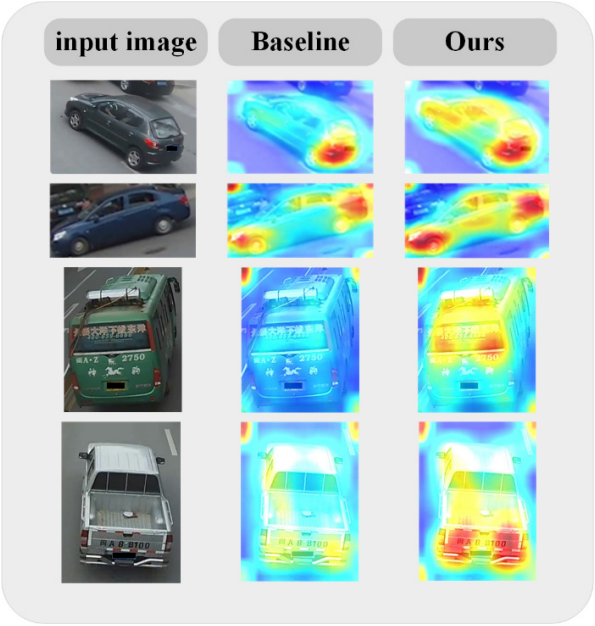


Fig. 8. Visualization of thermogram comparison between CFA-Net and Baseline model. In Baseline, the model ultimately focuses only on the local information of the vehicle. CFA-Net pays more attention to the vehicle target, avoids the interference of background information, and can better distinguish the background from the vehicle.

1) Compared Methods: Among the compared methods, BOW_CN [42] uses human-designed operators for the feature extraction, whereas FACT [13] fuses the color information with the semantic information. The VAMI [23] generates multi-view representations through GAN training on the input single-view, and VANet [9] learns two metrics for similar viewpoints and different viewpoints in two feature spaces. Moreover, PRN [8] and PVEN [7] utilize the additional annotations to train key feature cues and local cues to lock onto the targets. CAL [19] and DCAL [20] enable the model to learn more useful features by designing an attention mechanism. Lastly, HRCN [43] learns the data features by designing a fusion module.

2) Results Analysis: Our method outperformed most of the compared state-of-the-art methods in both datasets, as highlighted in Table VI. First, we use CFA-Net with the produced

TABLE VI
COMPARISON OF PROPOSED METHOD WITH STATE-OF-THE-ART METHODS ON VERI776 AND DAIR-V2XREID

Method	DAIR-V2XReid			VeRi776		
	mAP	Rank-1	Rank-5	mAP	Rank-1	Rank-5
BOW_CN[42]	-	-	-	12.20	33.1	53.69
FACT[13]	-	-	-	18.49	50.95	73.48
VAMI[23]	-	-	-	50.13	77.03	90.83
VANet[9]	-	-	-	66.34	89.28	-
PRN*[8]	-	-	-	74.3	94.3	98.9
PVEN*[7]	-	-	-	79.50	95.60	98.40
CAL[19]	44.9	42.9	62.4	74.3	95.4	-
HRCN[43]	**47.05**	**40.91**	**68.94**	**83.1**	**97.3**	98.9
DCAL[20]	-	-	-	80.2	96.9	-
Baseline	51.44	48.12	51.44	77.06	94.40	-
Ours	**58.47**	**59.40**	**76.69**	**80.67**	**96.90**	98.21

The best results published before and the best results of our method are marked in bold, - is no result. The star * in the superscript indicating that the model requires additional annotation, which is not provided by DAIR-V2XReid.

Fig. 9. A sorted list of the top 5 Ranks from the gallery corresponding to the query image on the VeRi776 dataset. The first two rows are the final results of Baseline, and the last two rows are the final results of our model. Green represents the correct results while red represents incorrect the results.

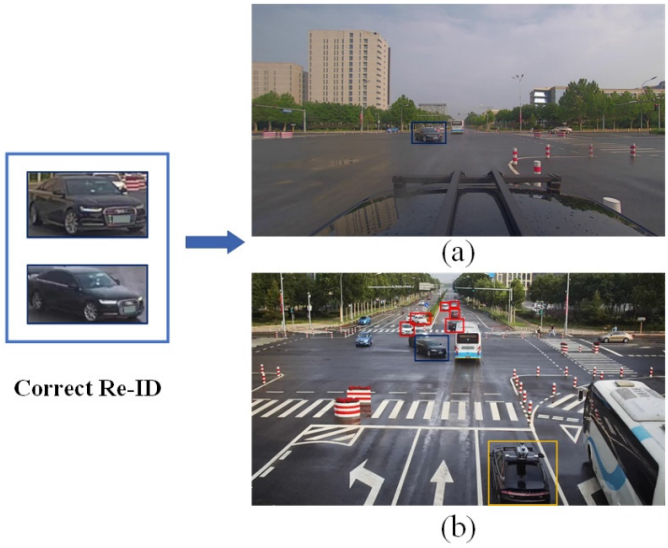


Fig. 10. Schematic diagram of matching. (a) A shot by vehicle-end device, (b) A shot by road-end device.

dataset DAIR-V2XReid for experiments, with three evaluation metrics, namely Rank-1, Rank-5 and mAP. In order to compare the effectiveness of this model, we also utilize this same dataset for the previously proposed methods. As can be seen from the second column of Table VI, our method greatly outperformed all the previously proposed methods in terms of mAP.

At the same time, to analyze the generalization ability of the model, we conducted experiments on VeRi776 dataset, using the same evaluation metrics as mentioned above. We then compared our method with 9 existing mainstream methods. Compared with the existing methods, our method does not require extra annotations, which not only alleviates the influence of annotation quality and number of annotations, but also makes the model run faster and more accurately. From column 3 of Table VI, it can be seen that the proposed model achieves a higher accuracy than most of the published methods, where mAP of our method is 0.47% higher than the mAP of recently proposed DCAL [20].

VI. CONCLUSION

In this work, we investigated the current problems in vehicle Re-ID, aiming to achieve over-the-horizon perception for autonomous driving. In order to solve problems including perspective occlusion and achieve comprehensive perception, we contributed with a new Vehicle-Infrastructure Cooperative (VIC) Re-ID dataset DAIR-V2XReid for real vehicle scenes, which contains rich variations of background, viewpoint, and light. At the same time, we also proposed a new network for vehicle Re-ID tasks, Cross-shot Feature Aggregation Network (CFA-Net), to overcome the issues related to the large variation of cross-shot perspective in VIC perception. The proposed network focuses on more specific and discriminative features of the vehicle, leading to improved training accuracy of the model. Extensive experimental results showed that the proposed method also has a good generalization ability and

achieves a good accuracy in the general dataset Veri776. Essentially, this work contributes towards the development of VIC Re-ID technology by providing the relevant dataset as well as an improved perception model, thus better enabling the utility of Re-ID in critical autonomous driving tasks such as control and decision-making.

REFERENCES

[1] Y. Li, F. Feng, Y. Cai, Z. Li, and M. A. Sotelo, "Localization for intelligent vehicles in underground car parks based on semantic information," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 2, pp. 1317–1332, Feb. 2024.

[2] H. Bagheri et al., "5G NR-V2X: Toward connected and cooperative autonomous driving," *IEEE Commun. Standards Mag.*, vol. 5, no. 1, pp. 48–54, Mar. 2021.

[3] L. Yang, P. Luo, C. C. Loy, and X. Tang, "A large-scale car dataset for fine-grained categorization and verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3973–3981.

[4] 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.

[5] X. Liu, W. Liu, T. Mei, and H. Ma, "A deep learning-based approach to progressive vehicle re-identification for urban surveillance," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 869–884.

[6] 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)*, Long Beach, CA, USA, Jun. 2019, pp. 3230–3238, doi: 10.1109/CVPR.2019.00335.

[7] 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.

[8] 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.

[9] R. Chu, Y. Sun, Y. Li, Z. Liu, C. Zhang, and Y. Wei, "Vehicle re-identification with viewpoint-aware metric learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8281–8290.

[10] R. Bishop, "A survey of intelligent vehicle applications worldwide," in *Proc. IEEE Intell. Vehicles Symp.*, Apr. 2000, pp. 25–30.

[11] A. S. Akki and F. Haber, "A statistical model of mobile-to-mobile land communication channel," *IEEE Trans. Veh. Technol.*, vol. VT-35, no. 1, pp. 2–7, Feb. 1986.

[12] H. Yu et al., "DAIR-V2X: A large-scale dataset for vehicle-infrastructure cooperative 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 21329–21338.

[13] 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.

[14] H. Park and B. Ham, "Relation network for person re-identification," 2019, *arXiv:1911.09318*.

[15] J. Zhu et al., "Vehicle re-identification using quadruple directional deep learning features," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 1, pp. 410–420, Jan. 2020.

[16] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 480–496.

[17] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[18] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.

[19] Y. Rao, G. Chen, J. Lu, and J. Zhou, "Counterfactual attention learning for fine-grained visual categorization and re-identification," 2021, *arXiv:2108.08728*.

[20] H. Zhu, W. Ke, D. Li, J. Liu, L. Tian, and Y. Shan, "Dual cross-attention learning for fine-grained visual categorization and object re-identification," 2022, *arXiv:2205.02151*.

[21] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, "Person re-identification by multi-channel parts-based CNN with improved triplet loss function," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1335–1344.

[22] K. Yan, Y. Tian, Y. Wang, W. Zeng, and T. Huang, "Exploiting multi-grain ranking constraints for precisely searching visually-similar vehicles," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 562–570.

[23] Y. Zhouy and L. Shao, "Viewpoint-aware attentive multi-view inference for vehicle re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6489–6498.

[24] 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.

[25] Y. Zhou and L. Shao, "Vehicle re-identification by adversarial bi-directional LSTM network," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 653–662.

[26] S.-W. Kim et al., "Multivehicle cooperative driving using cooperative perception: Design and experimental validation," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 2, pp. 663–680, Apr. 2015.

[27] Y. Li, S. Ren, P. Wu, S. Chen, C. Feng, and W. Zhang, "Learning distilled collaboration graph for multi-agent perception," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 29541–29552.

[28] T.-H. Wang, S. Manivasagam, M. Liang, B. Yang, W. Zeng, and R. Urtasun, "V2VNet: Vehicle-to-vehicle communication for joint perception and prediction," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 605–621.

[29] X. Weng et al., "All-in-one drive: A large-scale comprehensive perception dataset with high-density long-range point clouds," 2021.

[30] M. Müller, V. Casser, J. Lahoud, N. Smith, and B. Ghanem, "Sim4CV: A photo-realistic simulator for computer vision applications," *Int. J. Comput. Vis.*, vol. 126, no. 9, pp. 902–919, Sep. 2018.

[31] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "CARLA: An open urban driving simulator," in *Proc. Conf. Robot Learn.*, 2017, pp. 1–16.

[32] Y. Maalej, S. Sorour, A. Abdel-Rahim, and M. Guizani, "VANETs meet autonomous vehicles: A multimodal 3D environment learning approach," in *Proc. IEEE Global Commun. Conf.*, Dec. 2017, pp. 1–6.

[33] Q. Chen, X. Ma, S. Tang, J. Guo, Q. Yang, and S. Fu, "F-cooper: Feature based cooperative perception for autonomous vehicle edge computing system using 3D point clouds," in *Proc. 4th ACM/IEEE Symp. Edge Comput.*, Nov. 2019, pp. 88–100.

[34] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3354–3361.

[35] Z. Lu, V. Rathod, R. Votel, and J. Huang, "RetinaTrack: Online single stage joint detection and tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 14656–14666.

[36] K. Messaoud, N. Deo, M. M. Trivedi, and F. Nashashibi, "Trajectory prediction for autonomous driving based on multi-head attention with joint agent-map representation," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jul. 2021, pp. 165–170.

[37] 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.

[38] 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.

[39] X. Pan, P. Luo, J. Shi, and X. Tang, "Two at once: Enhancing learning and generalization capacities via ibn-net," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 464–479.

[40] J. Cherry, "SGD: Saccharomyces genome database," *Nucleic Acids Res.*, vol. 26, no. 1, pp. 73–79, Jan. 1998.

[41] D. Gray, S. Brennan, and H. Tao, "Evaluating appearance models for recognition, reacquisition, and tracking," in *Proc. IEEE Int. Workshop Perform. Eval. Track. Surveill.*, Feb. 2007, pp. 1–7.

[42] 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.

[43] 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)*, Montreal, QC, Canada, Oct. 2021, pp. 205–214, doi: 10.1109/ICCV48922.2021.00027.

Hai Wang (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees from the School of Instrument Science and Engineering, Southeast University, Nanjing, China.

In 2012, he joined the School of Automotive and Traffic Engineering, Jiangsu University, where he is currently a Professor. He has published more than 50 articles in the field of machine vision-based environment sensing for intelligent vehicles. His research interests include computer vision, intelligent transportation systems, and intelligent vehicles.

Yaqing Niu received the B.S. degree from Jiangsu University, Zhenjiang, China, where she is currently pursuing the Ph.D. degree.

Her research interests include computer vision, deep learning, and intelligent vehicles.

Long Chen received the Ph.D. degree in vehicle engineering from Jiangsu University, Zhenjiang, China, in 2002.

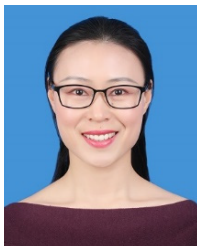
His research interests include intelligent automobiles and vehicle control systems.

Yicheng Li received the Ph.D. degree in vehicle engineering from Wuhan University of Technology, Wuhan, China, in 2018.

He is currently an Assistant Professor with the Automotive Engineering Research Institute, Jiangsu University. His research interests include intelligent vehicle localization, intelligent transportation systems, computer vision, and 3-D data processing.

Miguel Angel Sotelo (Fellow, IEEE) received the Ph.D. degree in electrical engineering from the University of Alcalá (UAH), Madrid, Spain, in 2001. He is currently a Full Professor with the Department of Computer Engineering, UAH. His research interests include autonomous vehicles and prediction of intentions. He has served as a project evaluator, a rapporteur, and a reviewer for the European Commission in the field of ICT for intelligent vehicles and cooperative systems in FP6 and FP7. He is a member of the IEEE ITSS Board of Governors and Executive Committee. He is the President of the IEEE Intelligent Transportation Systems Society. He served as the Editor-in-Chief for *IEEE Intelligent Transportation Systems Magazine* and *ITSS Newsletter* and an Associate Editor for IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS.

Zhixiong Li (Senior Member, IEEE) received the Ph.D. degree in transportation engineering from Wuhan University of Technology in 2013. He is with Yonsei Frontier Laboratory, Yonsei University, Seoul, Republic of Korea; and also with the Faculty of Mechanical Engineering, Opole University of Technology, Poland. He is the Director of the International Joint Research Centre on Renewable Energy and Sustainable Marine Vehicles. He is the author/coauthor of two books and over 100 papers. His research interests include dynamic system modeling, renewable energy, and machine learning applications. He is an Associate Editor of IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS and a Column Editor of *IEEE Intelligent Transportation Systems Magazine*.



Yingfeng Cai (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees from the School of Instrument Science and Engineering, Southeast University, Nanjing, China. In 2013, she joined the Automotive Engineering Research Institute, Jiangsu University, where she is currently a Professor. She has published more than 100 articles in high-level journals, including IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION VEHICLES, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, and IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY in the field of sensing and control for intelligent vehicles. She received the National Fund for Distinguished Young Scholars of China. Her research interests include computer vision, intelligent transportation systems, and intelligent automobiles.