Multimodality Adaptive Transformer and Mutual Learning for Unsupervised Domain Adaptation Vehicle Re-identification

Xin Zhang, Yunan Ling, Kaige Li, Weimin Shi, Zhong Zhou (✉)

***Abstract*—Unsupervised Domain Adaptation Vehicle Re-Identification (UDA vehicle re-ID) aims to enable the model trained in the source domain dataset to adapt to the target domain data and obtain accurate re-identification results, which has received widespread attention due to its practicality in the field of intelligent transportation systems. Most current UDA vehicle re-ID research ignores the mining and utilization of attribute information. Meanwhile, the Convolutional Neural Networks-based (CNN-based) network will cause the loss of fine-grained information, reducing the expression and generalization ability of vehicle features. To alleviate such issues, we are motivated by the Transformer, which can exploit distinguishable attribute information and fuse multimodal features effectively. Therefore, this paper proposes a Multimodality Adaptive Transformer Network (MATNet) to intensify the ability to learn vehicle fine-grained features related to attributes. Moreover, the noise contained in pseudo-labels assigned by cluster algorithms interferes with the performance of the UDA vehicle re-ID method. We also design the Dual Mutual Dynamic Update Pseudo-Label generation strategy (DMDU) to improve the accuracy of pseudo-labels and alleviate error accumulation. The strategy is based on mutual learning, which can effectively utilize the congruous and particular knowledge of the two models to generate pseudo-labels. Extensive experiments on two large-scale public datasets, including VeRi-776 and VehicleID, illustrate that our method outperforms the state-of-the-art methods.**

***Index Terms*—Domain Adaption, Fine-Grained Attribute, Transformer-Based, Unsupervised Vehicle Re-identification**

I. Introduction

THE objective of vehicle re-identification (Re-ID) is to recognize and identify vehicle images captured from multiple non-overlapping camera views[1]. Due to its practicality in intelligent transportation system research, it has evolved into a study focus in academia and industry[2, 3]. Although deep learning-based models that train with supervised data have achieved satisfactory performance in experiments. However, the well-trained models have performed poorly when directly transferred to a new dataset (i.e., target domain) due to the domain gap and inconsistency of hidden knowledge in different vehicle datasets. Consequently, the Unsupervised Domain Adaption (UDA) vehicle Re-ID research is proposed to address these issues that arise in the practical process [4, 5].

To tackle the UDA Re-ID problem, researchers have proposed various frameworks to implement the unsupervised domain adaptation task. In current, UDA based on progressive learning has achieved excellent performance. However, there remain two challenges that need to be addressed. First, due to the characteristics of vehicle data, there will be such a phenomenon in different shooting angles that different classes of vehicle images have analogous appearances at different shooting angles, while the same class of vehicle images have a significant difference in appearance, resulting in the weak expressive ability of vehicle appearance features, as shown in Fig. 1 (a). Furthermore, as a result of the drawback of clustering algorithms, the noise included in pseudo-labels and error accumulation is inescapable, as shown in Fig. 1 (b). This leads to that the low-quality pseudo-labels cannot provide accurate supervision information to guide the network to adapt to the new domain data. This makes the model unable to accurately identify the new domain vehicles, which is not

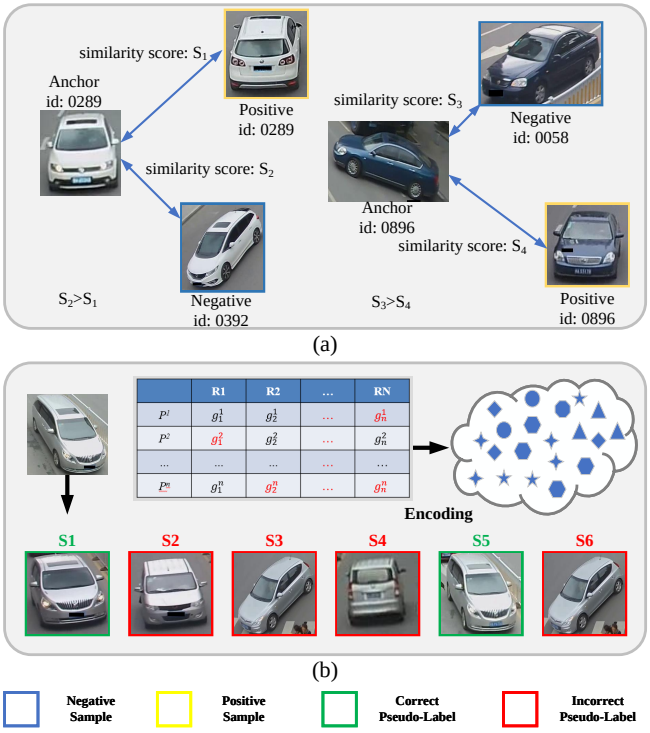


Fig. 1. (a) Example of similarity between appearance features of vehicles from different shooting angles. The change in shooting angle will cause the appearance features of the different vehicles to be more similar than the identical vehicles. (b) Example of pseudo-label generation in UDA vehicle re-ID with the clustering algorithm. The noise of pseudo-labels impedes the process of optimizing the network.

Xin Zhang is with the School of Artificial Intelligence, Hebei University of Technology, Tianjin, 300130, China.(e-mail: 2023927@hebut.edu.cn;) Yunan Ling, Kaige Li, Weimin Shi, and Zhong Zhou are with the State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing 100191, China (e-mail: lingyunan; lkg; shiwm; zz@buaa.edu.cn)

conducive to its practical performance. Consequently, we have analyzed the deep reason behind such challenges and propose a novel UDA vehicle Re-ID method.

First, in the actual application scene of vehicle Re-ID, various factors inevitably affect vehicle images, which will hurt the distinguishability of the vehicle appearance feature. This phenomenon is more severe in UDA Re-ID, which shows that it is not enough to describe the vehicle from its appearance only to identify vehicles in the new dataset. Accordingly, to more accurately identify target domain vehicle images, extracting more robust and expressive vehicle features is first necessary. The discrimination is associated mainly with the amount of information contained in vehicle features. Researchers improve the expressive ability of vehicle features by guiding the network to learn different vehicle attributes information, which improved the effect of vehicle Re-ID[6, 7]. How to effectively use the vehicle attribute information is the point that should be focused on. Attributed to the self-attention mechanism, Transformer [8] has been employed in computer vision and achieved outstanding performance. The Transformer can help the model to achieve discrimination feature filtering and multi-model information fusion, thereby extracting the appearance and attribute feature simultaneously to improve the distinguishability of the features. Meanwhile, the Transformer-based object ReID framework (TransReID)[9] and Vision Transformer (ViT)[10] have shown that the Transformer multi-model information fusion capabilities can help improve recognition accuracy. Inspired by this, we propose the Multimodality Adaptive Transformer Network (MATNet) as the UDA vehicle Re-ID model to use the Transformer's excellent multi-model information fusion capabilities to extract more distinguishable vehicle features.

After dissecting the pseudo-label generation process in progressive learning methods, we believe that previous studies have ignored two pivotal ingredients. First, researchers typically select the pseudo-labels according to the cluster results and a strict threshold, making random noise and error accumulation inevitable[11, 12]. Second, the network cannot learn high-quality target domain features due to the domain gap. For target domain images with high pseudo-label credibility, their style and distribution are closer to the source domain images. As a result, correct and effectively utilizing the target domain data (i.e., hard samples) with a large gap from the source domain data during domain adaptation training is difficult. To alleviate such issues, we design a novel Dual Mutual Dynamic Update Pseudo-Lable generation strategy (DMDU), which consists of two steps to alleviate unreliable pseudo-label. The main idea is that, based on mutual learning[13], two independent models are used to extract the target domain feature, and each model can leverage external knowledge to correct its own errors. In this way, it can improve the pseudo-label accuracy of simple target domain samples. For outlier samples and hard samples, they are dynamically updated through the category centers of simple sample pseudo-labels, restraining the interference of noise on the improvement of model domain adaptability.

The contributions of this paper could be summarized as follows:

- We propose the Multimodality Adaptive Transformer Network for UDA vehicle re-identification. This work is the first attempt to utilize the Transformer in conjunction with descriptive sentences to introduce vehicle attribute information, thereby improving feature expression capabilities in UDA vehicle Re-ID.
- We design a novel Dual Mutual Dynamic Update Pseudo-Label generation strategy. It integrates the double consistency and online correction ideas to reduce pseudo-label noises in a coarse-to-finely manner, achieving more accurate supervision information.
- Our proposed UDA vehicle Re-ID method was extensively evaluated on two public datasets, VeRi-776[14] and VehicleID[1]. It outperformed several state-of-the-art methods.

The remainder of this paper is arranged as follows. In section II, we introduce the related work about the UDA vehicle ReID. In section III, we explain our proposed method. In section IV, we illustrate systematic experimental results to verify the effectiveness of our method. In section V, we summarize this work and propose future research plans.

II. Related Works

A. Vehicle Feature Extraction

The quality of vehicle features can directly affect the effect of vehicle Re-ID. Most current research is devoted to extracting the distinguishability information of vehicles to generate discriminative features. The first type of method mainly mines the discriminative information contained in the vehicle appearance[15–17] to improve the accuracy of vehicle Re-ID. In [18], Wang et al. manually marked 20 key points for the vehicle to obtain the detailed feature of the vehicle according to the position of the key points. In [19], the Region of Interest (ROI) is introduced so that the network can focus on the vehicle area in the image. In addition, the ability of the network to extract vehicle appearance features can be improved by designing loss functions[15, 17, 20].

The second category includes approaches focusing on fusing vehicle global features and local features to exploit the discriminative information [21–23]. In [24], Cheng et al. extracted and fused different scales and levels of information contained in vehicle images. They distinguish different vehicles through the generated fusion feature descriptor. In [25], Liu et al. designed the parsing-guided cross-part reasoning network (PCRNet), which models the relationship between vehicle components and estimates the occluded local features of vehicles with graph convolutional networks [26].

The third category considers the attribute information of vehicles as essential cues to guide the network to generate discriminative vehicle descriptors [6, 27, 28]. Multi-task learning networks are often used to introduce vehicle attribute information directly. In [29], researchers introduced multi-dimensional attributes of vehicle images to enhance features, which is based on the multi-task cooperation method. In [30], Zhu et al. proposed the vehicle-orientation-camera joint re-identification method. They introduce the information on vehicle orientation and shooting angle into the vehicle feature

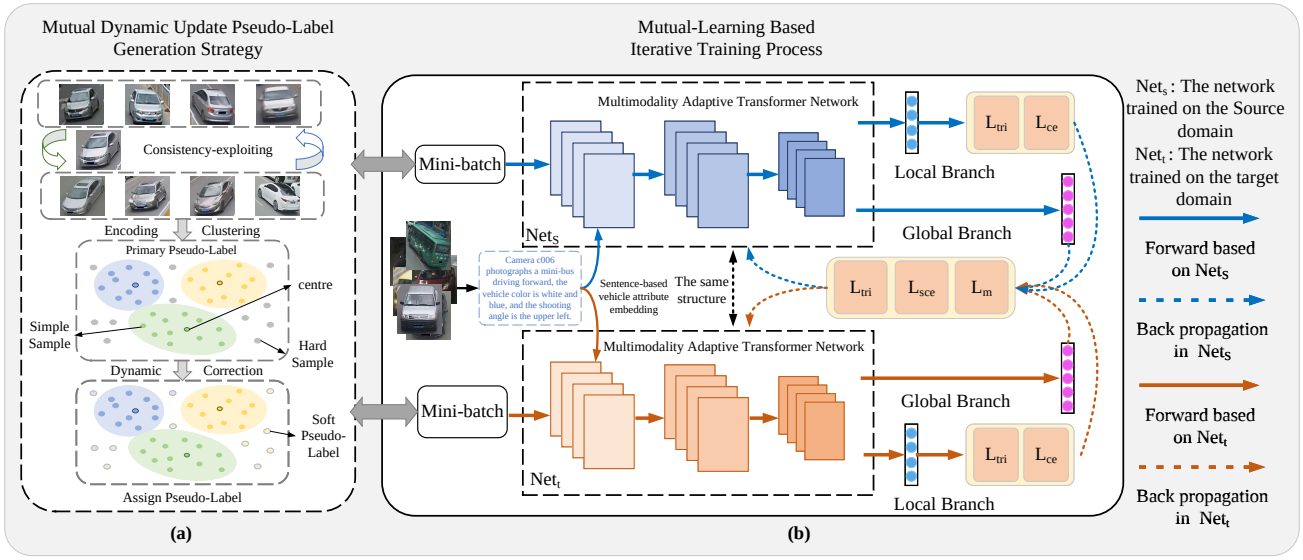


Fig. 2. The overall framework of the proposed Multimodality Adaptive Transformer and Mutual Learning Unsupervised Domain Adaptation Vehicle Re-Identification method with two collaborative models jointly optimized. The MATNet is used to learn and fuse multimodal attribute information of vehicles to extract vehicle features. The DMDU strategy is used to generate the accurate pseudo-label for training. Meanwhile, the iterative training is carried out in the mutual learning framework.

as auxiliary information to alleviate the issues of similar appearance features of different vehicles under the same shooting angle. Therefore, after rethinking the mentioned methods, we propose MATNet to learn the attribute information of vehicles. Different from the above methods, we utilize the powerful feature extraction ability and multi-modal information fusion ability of the Transformer[8, 10] to extract more discriminative vehicle features for the UDA vehicle Re-ID task.

B. Unsupervised Domain Adoption

The UDA has become hotspot research due to the ability to improve the practicality of deep learning models and reduce manual annotation workload. It focuses on learning highly generalizable features and alleviating the domain gap between the source and target domains[4, 31]. Exiting UDA methods are mainly centered on cross-domain transfer learning[32, 33] and progressive learning[31, 34, 35].

The first type of method mainly utilizes the generative adversarial network (GAN) to generate source domain images with the target domain style to train the network, thereby adapting the network to the target domain [5, 36]. In [37], Peng et al. introduced the identity information of the source domain and the style information of the target domain through the GAN network to achieve domain adaptation. In [5], researchers designed the GAN-Siamese network, which can alleviate the gap between image features in the two domains. However, the identification accuracy of these methods in the new dataset is limited by the quality of generated images.

The second category focuses on directly utilizing the supervision information contained in the target domain, which usually generates the pseudo-label for network fine-tuning[35, 38]. In [39] Zheng et al. introduced viewpoint attribute information to decrease the noise in generated vehicle pseudo-labels. Then, researchers were inspired by the idea of Mean Teacher [40] and used the mutual learning framework to conduct unsupervised domain adoption tasks [41–44]. In [45], Chen et al. employed three different networks to guide each other to avoid the interference of noise in pseudo-labels. Zhou et al.[46] used two asymmetric networks to complement each other and improve the supervision information in pseudo-labels. Progressive learning methods have been widely used in UDA research and have achieved promising performance. However, limited by the expression generalization ability of vehicle features and the defects of the clustering algorithm itself, noise inevitably appears in the generated pseudo-labels, which prompted us to propose the DMDU pseudo-lable generation strategy. The strategy generates pseudo-labels by employing mutual learning and online correction to alleviate the issue of unreliable pseudo-labels.

III. Method

This section will elaborate on the proposed multimodality adaptive Transformer and mutual learning unsupervised domain adoption vehicle Re-identification method. The proposed method improves the domain adaptability of the model by improving the distinguishability of vehicle features and generating accurate pseudo-labels. The proposed method comprises the MATNet and the DMDU pseudo-label generation strategy, as shown in Fig. 2.

A. Overview of Framework

Formally, we denote the vehicle data in the source domain as $\mathcal{D}_s = \{(v_i^s, l_i^s)\}$. Where v_i^s represents the source domain vehicle sample used for training, l_i^s denotes the identity label of the sample, $i \in \{1, 2, 3, ..., n\}$ indicates that there are n classes of vehicle samples with different identities in the source domain. In addition, we set $D_s = \{v_i^t\}$ to represent

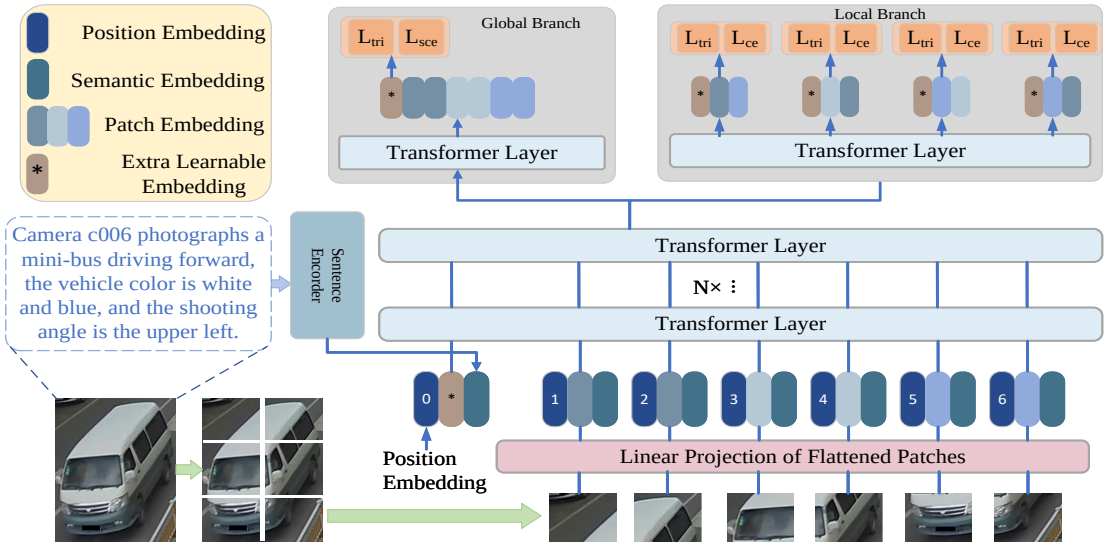


Fig. 3. The framework of proposed MATNet. The vehicle image is split into different patches. The attribute information of the vehicle is introduced by generating short sentences describing the vehicle image. Then, the semantic embeddings are input into the network along with patch and position embedding.

the vehicle sample in the target domain, it does not have the corresponding identity label. The proposed UDA framework needs to use the source domain well-trained model to extract the discriminative information in the target domain samples, which guides the model to adapt the target domain dataset.

The proposed UDA vehicle Re-ID method is based on mutual learning. The first is to pre-train the baseline vehicle Re-ID model, utilizing the designed MATNet as the baseline. The Net_s and Net_t represent source and target domain-trained initial models, respectively. Net_s is pre-trained with the cross-entropy loss L_{ce}^s and triplet loss function L_{tri}^s. Afterward, the model Net_t is pre-trained under the supervised information from initial pseudo-labels, which is cluster-generated based on extracting target domain vehicle features by Net_s.

The proposed framework is pre-training two models with two different datasets for two reasons. The first is to prevent the two models from overfitting the source domain data, which will interfere with improving the adaptability of the model to the target domain data. The second is that using two different datasets for training can enable the two models to obtain different learning capabilities to guide each other better.

After obtaining the initial pseudo-label and pre-trained model, iterative training is carried out under the mutual learning framework. The two models can guide each other to supervise each other until the UAD vehicle Re-ID method reaches the maximum number of iterations. In this way, each model can leverage the unique knowledge contained in the corresponding model to improve its adaptability to the target domain. In addition, consistent with the iterative training process in [47], Net_s and Net_t selected vehicle samples usually have a smaller loss value in a mini-batch training sample and fed them into its peer network for training.

B. Multimodality Adaptive Transformer Network

Extracting highly discriminative and robust features is essential for the UDA vehicle Re-ID task. This is because vehicle images will be disturbed by factors such as image resolutions, shooting angles, lighting change, and occlusion, reducing the discriminative ability of extracted vehicle features. Therefore, we proposed the MATNet to extract vehicle features, and the network structure is shown in Fig. 3.

Observing vehicle data, we found that vehicle attribute information such as vehicle color and vehicle model is difficult to affect by interference factors. Even if affected by interference factors, most of the attribute information will only change slightly and are still highly discriminative. Inspired by this, we use the powerful multimodality information integration capability of the Transformer network to introduce attribute information into the vehicle feature through knowledge refinement. The anti-jamming capability of these attributes information are utilized to make vehicle feature representation more robust and discriminative. In addition, as mentioned above, the shooting angle will also cause serious interference with vehicle Re-ID. Specifically, when the same vehicle is captured by cameras from different angles, its appearance will inevitably change or even take on a completely different appearance. Correspondingly, under the influence of the shooting angle, different vehicles with the same model and color will show the same appearance, making it difficult to distinguish. In this process, external factors such as illumination, occlusion, and image quality will aggravate the phenomenon. Therefore, to alleviate such issues, we also introduce the viewpoint information of the vehicle image into the features. Finally, we selected vehicle model, viewpoint, and color, which are valuable and discriminating attribute information to generate the descriptors of the vehicle.

For how to enable the network to efficiently learn the attribute information mentioned above. Inspired by the Contrastive Language-Image Pre-Training (CLIP) network[48] and Common Objects in Context (COCO) dataset[49], we describe this information in a short sentence. There are two reasons for choosing short sentences rather than words for descriptions. First, from the perspective of the Transformer network,

the early Transformer model was used in Natural Language Processing (NLP) research for adaptive learning of the long and short associations that exist in sentences. Therefore, more discriminative and robust attribute knowledge can be exploited from the described sentences. Secondly, from the perspective of human behavior patterns, we also search for vehicles through short sentences, not just a few words. In summary, we can easily generate a descriptive sentence containing the above attributes by a fixed sentence pattern during the training process. These concise descriptive sentences not only conform to semantic expression but also have discriminative. Moreover, we employ SentenceTransformer[50] provided the pre-trained model to convert the descriptive sentence into the regularized vector of length 768 we needed.

During the training, we split the vehicle image v into N fixed-sized image patches denoted as $\left\{v_p^i | i = 1, 2, ..., N\right\}$. Then concatenate with patch embedding, position embedding, and semantic embedding, which is similar to ViT[10] and TransReID[9]. Finally, each patch input to MATNet encoder consists of three parts as follows:

$$V_p = \mathcal{F}\left(x_p^i\right) + \varphi_i + \mu\mathcal{ST}\left(s\right) \tag{1}$$

Where $\mathcal{F}$ is the linear projection, which is used to flatten the image patch block into D dimensions. φ_i is position embedding of image patch. s represents describe sentences corresponding to the vehicle image. $\mathcal{ST}$ is the pre-trained sentence Transformer model, and μ is an adjustable hyper-parameter. It should be noted that since the description sentence is determined by the whole vehicle image, $\mathcal{ST}\left(s\right)$ is the same for each patch. In addition, a learnable embedding token x_{cls} needs to be input in the encoder of MATNet to output the learned global feature f_g after the last Transformer layer[9]. Therefore, the overall input of MATNet is:

$$V_{in} = \left[x_{cls}; \mathcal{F}\left(x_p^i\right); \ldots; \mathcal{F}\left(x_p^N\right)\right] + \varphi + \mu\mathcal{ST}\left(s\right) \tag{2}$$

Where V_{in} is the input sequence embedding, and the number of the Transformer layer is set as t. We define the hidden features input to the last Transformer layer as $H_{l-1} = \{h_{l-1}^0, h_{l-1}^1, ..., h_{l-1}^n\}$. In the global branch, we concatenate all patch embeddings into a standard Transformer output as the global feature f_g of the vehicle. In the local branch, we adopt the same operation as in TransReID[9] to learn the local features. Specifically, we disrupt and reorganize the learned patch embedding through the shift operation and the patch shuffle operation to generate different segment groups as local feature f_l of the vehicle. This enables the network to learn information from different local areas of the vehicle and fuse it to improve the distinguishability of features. It should be noted that we only use global feature f_g as the final vehicle feature representation to identify different vehicles, which is beneficial to improve the model's adaptability and robustness to target domain data.

C. Dual Mutual Dynamic Update Pseudo-Label Generation

In UDA vehicle Re-ID, we design the DMDU pseudo-label generation strategy to alleviate the noise contained in pseudo-labels. We first enforce constraints with the dual mutual consistency strategy, which can obtain more accurate supervision information (pseudo-labels). However, this strategy works better on simple samples, and the generated accurate pseudo-labels are not enough to train the model in the target domain. The large amount of supervised information from hard samples is not fully utilized. Therefore, in the second, we employ the online correction method to generate more reliable pseudo-labels for hard samples so that the network can get more supervision information for training.

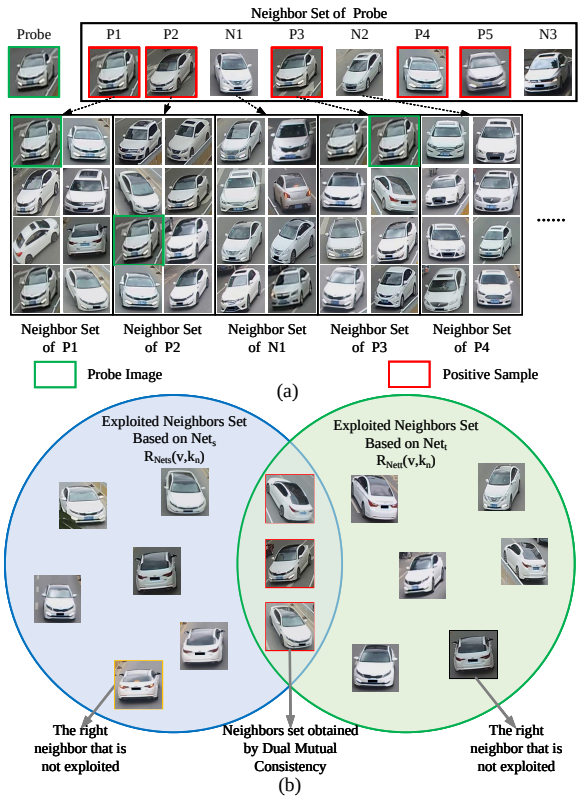


Fig. 4. The illustration of the dual mutual consistency strategy. (a): The sample neighbor sets mutual consistency. (b): The model neighbor sets mutual consistency.

The process of generating pseudo-labels for simple samples using the dual mutual consistency strategy is shown in Fig. 4. For probe image v, we define its k-reciprocal nearest neighbors as $R\left(v, k_n\right)$, where $k_n = 30$. There are samples P of the right neighbors and N of the wrong neighbors in these 30 nearest neighbors, as shown in Fig. 4 (a). This is because limited by the quality of extracted features, the neighbor samples selected with the k-reciprocal nearest neighbors strategy will inevitably contain noise. In order to obtain accurate pseudo-labels, the key is that when sample a is in the top-K neighbor set of sample b, sample b also needs to be in the top-K neighbor set of sample a. Only in this way are the two samples considered correct neighbors and belong to the same class. We call it sample neighbor set mutual consistency, which is the first mutual consistency.

Nonetheless, in the training process, we found that because of the finite learning ability of a single model, it is impossible to accurately and entirely extract features, resulting in noisy neighbors still existing. Especially in vehicle Re-ID, this issue

is exacerbated by reason of the slight inter-class discrepancy and considerable intra-class discrepancy of vehicle images. For this phenomenon, the knowledge in pseudo-labels can be refined and utilized with the help of vehicle features extracted from different models. Accordingly, we utilize the pre-trained Net_s and Net_t to generate $R_{nets}(v, k_n)$ and $R_{nett}(v, k_n)$, which indicate two k-reciprocal nearest neighbors, as shown in Fig. 4 (b). After that, the model mutual consistency constraint is performed to obtain the common k-reciprocal nearest $R_{co}(v, k_n)$, which is the second mutual consistency as follows:

$$R_{co}(v, k_n) = R_{nets}(v, k_n) \cap R_{nett}(v, k_n) \tag{3}$$

Intuitively, $R_{co}(v, k_n)$ is a more reliable neighbor set, generating more accurate pseudo-labels for simple samples. This operation inevitably loses some positive samples (i.e., hard and outlier samples), which makes the network lack supervision information during training. Intending to effectively adopt these samples to train the network, inspired by the traditional self-supervised study, we try to assign a soft pseudo-label for these samples. The classification probability of these remaining samples can usually be calculated by the loss function[11], and the category with the highest probability is used as its soft pseudo-label as follows:

$$L_{ce} = -\frac{1}{N}\sum_{i}\sum_{c=1}^{N_t} \hat{v}_t^{N_k} \log\left(p_t^{N_k}\right) \tag{4}$$

$$\hat{v}_t^{(i,N_k)} = \begin{cases} 1, if\ N_k = arg\ max_{N_k'} p_t^{\left(i,(N_k')\right)} \\ 0, \quad otherwise \end{cases} \tag{5}$$

Where N_t is the number of vehicle categories in the target domain generated by the pseudo-labels of simple samples. $p_t^{(i,N_k)}$ represents the softmax probability of target domain sample $v_t^{(i,N_k)}$ belong to N_ith class of target domain. Meanwhile, we denote the process of generating soft pseudo-labels as $v_t^{(i,N_K)} = \varepsilon\left(p_t^{(i,N_k)}\right)$. However, the hard sample pseudo-labels generated in this way will contain noise information, which will cause the network to overfit the noisy labels and fail to learn effective discriminative knowledge. On the other hand, concurrently updating the network weights and pseudo-labels will interfere with the training of the network on the target domain data.

Hence, the proposed strategy introduces a dynamic update pseudo-label denoising operation. By fixing the soft pseudo-labels and gradually weighting them by the distance from the centroid of simple samples pseudo-labels, with the update following the freshly learned knowledge:

$$v_t^{(i,N_k)} = \varepsilon\left[\left(\omega_t^{(i,N_k)} + \omega_s^{(i,N_k)}\right) p_t^{(i,N_k)}\right] \tag{6}$$

Where $\omega_t^{(i,N_k)}$ and $\omega_s^{(i,N_k)}$ denote the weights calculated by Net_s and Net_t to adjust the soft pseudo-labels, respectively, and change as the iterative training. We utilize the pre-trained model Net_s to initialize the $p_t^{(i,N_k)}$ and remain fixed. For the vehicle sample feature f_v, if it is far away from the centroid of class N_i, it means that the quality of the extracted feature is poor and more likely to be an outlier. Hence, we need to reduce the probability of it being classified in the N_ith class. We defined $\omega^{(i,N_k)}$ in the equation 6 as:

$$\omega^{(i,N_k)} = \frac{\exp\left(-\left\|f_{v_t^i} - \eta_{N_k}\right\| / \tau\right)}{\sum_{N_t}\exp\left(-\left\|f_{v_t^i} - \eta_{N_t}\right\| / \tau\right)} \tag{7}$$

Where $f_{v_t^i}$ is the feature of sample v, η_{N_k} is the centroid of kth class which is calculated with pseudo-label of simple samples, these are generated by the corresponding pre-trained model. The temperature coefficient τ is set to 1 by default.

D. Training Loss

We employ ordinary cross-entropy loss and triplet loss for optimization for the local feature branch. Moreover, we employ the symmetric cross-entropy loss[51] and triplet loss with soft margins to optimize the global feature branch. This is because the symmetric cross-entropy loss is an improved and robust variant of cross-entropy loss for noisy datasets, which is more suitable for iterative training of UDA vehicle Re-ID, is formulated as:

$$L_{sce} = \alpha L_{ce}\left(y_t^i, p_t^i\right) + \beta L_{ce}\left(p_t^i, y_t^i\right) \tag{8}$$

Where α and β represent the balancing coefficients, and we set them to 0.1 and 0.9, respectively. y_t^i is the label of the sample, and p_t^i is the predicted probability. In addition, to prevent the numerical issue of $\log(0)$, we generate the one-hot label y_t^i in $[1e-4, 1]$. Furthermore, we also introduce the mutual learning loss[47] to guide the global feature branch during the iterative training, which is:

$$\begin{aligned} L_s^{ml} &= \frac{1}{|\bar{D}_t|}\sum_{v_t^i \in \bar{D}_t}\left(C_t\left(v_t^i\right) \cdot \log C_s\left(v_t^i\right)\right) \\ L_t^{ml} &= \frac{1}{|\bar{D}_s|}\sum_{v_s^i \in \bar{D}_s}\left(C_s\left(v_s^i\right) \cdot \log C_t\left(v_s^i\right)\right) \end{aligned} \tag{9}$$

Where C_s and C_t is the classifier of Net_s and Net_t, respectively. The overall loss of the global feature branch is:

$$\begin{aligned} L_s &= L_{sce} + \lambda_{tri} L_{tri} + \lambda_{ml} L_s^{ml} \\ L_t &= L_{sce} + \lambda_{tri} L_{tri} + \lambda_{ml} L_t^{ml} \end{aligned} \tag{10}$$

Where L_s and L_t are the loss of the Net_s and Net_t, respectively. λ_{tri} and λ_{ml} are the weight of triple and mutual learning loss, set $\lambda_{tri} = 0.6$ and $\lambda_{ml} = 0.4$, respectively. We adopt feature distance to determine positive and negative samples for the triplet loss. Firstly, we employ Net_s to extract vehicle features, followed by calculating the feature distance to measure similarity. Subsequently, we set the threshold T_{tri} to select samples above the threshold as positive samples and the rest as negative samples. We set the initial T_{tri} to 0.65 and dynamically decreased during the training process.

IV. Experiments

In this section, we will conduct multiple experiments to evaluate the performance and effectiveness of the proposed method on two benchmark vehicle Re-ID datasets.

A. Datasets and Settings

We conduct experiments on two publicly available Vehicle Re-ID datasets: VeRi-776[14] and VehicleID[1], to validate the effectiveness of our approach. To assess the efficacy of different methods with precision and comprehensiveness, we employed Cumulative Matching Characteristics (CMC)[52] and mean Average Precision (mAP)[53] as the evaluative metrics in our experiment.

VeRi-776: The dataset includes 776 vehicles captured by 20 cameras and contains 50,000 images. Its training set consists of 576 vehicles, including 37781 images. The test set contains 11579 images of 200 vehicles.

VehicleID: It includes 26267 vehicles captured by real-world scenarios and contains 221763 images. Four subsets containing 800, 1600, 2400, and 3200 vehicles are extracted for testing in different scales.

Implementation details. First, we employ the Pytorch framework for training and finetuning the proposed model. During the processing, we set the input image size of the network to $256 * 128$ and adjusted the padding with 0 values. In addition, to better train the network, we employ data augmentation operations, including random erasing, flipping, and random cropping. We employ ImageNet pre-trained ViT-Base as the initial model. We use the "paraphrase-MiniLM-L6-v2 pre-trained model", which is the SentenceTransformer network to generate semantic embedding. A training batch contains 8 classes of vehicle samples that are based on actual or pseudo-labels, with 8 vehicle images for each class, for a total of 64 images. In the iteration process, the pseudo-labels of the target domain images are updated every 5 epochs. In this way, the training batch images of the target domain must also be reselected every 5 epochs based on the updated pseudo-labels. We generate the pseudo-labels of the target domain with the DBSCAN clustering algorithm[54], so each clustering result denotes a class of the target domain vehicle. The Stochastic Gradient Descent (SGD) optimizer is configured to $1e-4$ during the training procedure. All the training and experiments are performed with GeForce RTX 2080Ti GPU.

B. Ablation study

1) Analysis of Vehicle Re-identification Method: The experiments are conducted under different query settings of VehicleID. The comparison methods mainly select the representative excellent vehicle Re-ID methods. We presented the results in Table I. In the VeRi-776 dataset, our vehicle re-ID network achieves competitive performance. Our method achieves 97.72% in Rank-1, 98.65% in Rank-5, and 90.56% in mAP. The following conclusion can be got: 1) For appearance-based vehicle Re-ID methods (such as SN[20] and DGPM[55]) and feature fusion vehicle Re-ID methods (such as LABNet[23], SGFD[56], MED[57] and URRNet[58]), our vehicle Re-ID network shows superior performance, significantly outperforming them. 2) For introducing attribute information methods (such as EMPC[59], TransReID[9], VAT[60], GiT[61] and SOFCT[62]), our method also exceptionally outshines these approaches in an extensive margin. Especially compared to the current best method GiT[61], our proposed model outperforms 0.86% in Rank-1 and 0.22% in mAP. Combined with the characteristics of vehicle orientation and shooting angle are various in the VeRi-776 dataset, which shows that the MATNet can gain improve generalization expression ability with the help of vehicle attributes when facing such issues.

TABLE I
THE PERFORMANCE(%) COMPARISON ON VERI-776. THE BEST RESULT FOR EACH INDICATOR WILL BE BOLDED.

Method	Reference	Rank-1	Rank-5	mAP
SN[20]	TNNLS'21	95.10	98.10	75.70
DGPM[55]	TITS'21	96.19	98.09	79.39
LABNet[23]	Neurocput'21	95.70	-	79.50
SGFD[56]	ICCV'21	96.70	98.60	91.00
MED[57]	ITS'23	97.20	98.60	83.40
URRNet[58]	IVT'23	93.10	97.10	72.20
EMPC[59]	TMM'21	96.20	98.40	80.90
TransReID[9]	ICCV'21	97.10	-	82.00
VAT[60]	IPM'22	97.50	98.70	80.40
GiT[61]	TIP'23	96.86	-	90.34
SOFCT[62]	ITS'23	96.60	**98.80**	80.70
MATNet(Ours)	-	**97.72**	98.65	**90.56**

TABLE II
THE PERFORMANCE(%) COMPARISON ON VEHICLEID. THE BEST RESULT FOR EACH INDICATOR WILL BE BOLDED.

Method	Reference	Small		Mdeium		Large	
		Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
UMTS[63]	AAAI'20	84.50	72.70	79.30	66.10	72.80	54.20
VARID[64]	TITS'22	75.30	75.40	68.80	70.80	63.20	64.20
HPDG[65]	TITS'21	91.37	80.42	88.21	75.17	82.58	65.04
DFNet[66]	TPAMI'21	95.79	83.09	93.22	77.27	89.38	69.85
MED[57]	ITS'23	87.80	-	83.10	-	81.40	-
EMPC[59]	TMM'21	95.30	85.00	93.30	79.80	90.20	**73.00**
GiT[61]	TIP'23	92.65	81.76	89.92	75.64	85.41	67.50
MATNet(Ours)	-	**95.83**	**85.29**	**93.89**	**80.32**	**91.08**	72.65

Moreover, the comparative experiment will be carried out on the VehilceID to indicate the efficacy of the MATNet additionally. The VehicleID dataset has many vehicles of the same vehicle model and various shooting scenes. Table II shows the experiment result. We obtain the best results in Rank-1 compared with current representative methods on three testing sets and achieve competitive results in mAP. Compared with the current best method EMPC[59], we surpass them by 0.53%, 0.69%, and 0.88% in Rank-1 on small, medium, and large test modes, respectively. Consequently, it's from the results that the designed MATNet is effective for vehicle Re-ID and can attain excellent results under different settings and scales datasets.

2) Analysis of Vehicle Attributes Introducing Methods: As discussed in Sec.III-B, how to introduce vehicle attribute information has a meaningful impact on improving the distinguishability of vehicle features. Therefore, we investigate the effect of different methods of learning vehicle attribute information in vehicle Re-ID and perform comparative experiments on the VeRi-776. In the experiment, we chose the CNN-based and Transformer-based multi-task learning method, the Transformer attribute words learning (the Transformer introducing

TABLE III
THE DIFFERENT ATTRIBUTE INTRODUCING METHODS PERFORMANCE(%) COMPARISON ON VERI-776.

Method	Rank-1	Rank-5	mAP
Multi-task Learning based on CNN	91.64	95.30	72.53
Multi-task Learning based on Transformer	92.32	97.22	79.81
Transformer Attribute Words Learning	94.15	96.87	80.30
CLIP[48] (Direct Transfer)	95.31	97.16	81.67
Sentence Description(Ours)	**96.98**	**97.34**	**82.35**

TABLE IV
THE DIFFERENT PSEUDO-LABEL GENERATION STRATEGY PERFORMANCE(%) COMPARISON ON VEHICLEID-TO-VERI-776 TASK.

Method	Rank-1	Rank-5	mAP
Baseline	58.27	67.32	20.67
DBSCAN[54]	61.31	70.60	28.83
Mutual Learning	63.87	71.58	29.32
Dynamic Update	65.50	73.28	33.29
Dual Mutual Consistency	68.76	75.31	39.64
DMDU(Ours)	**76.32**	**83.25**	**45.20**

TABLE V
THE DIFFERENT PSEUDO-LABEL GENERATION STRATEGY PERFORMANCE(%) COMPARISON ON VERI-776-TO-VEHICLEID TASK.

Method	Small		Medium		Large	
	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
Baseline	36.84	41.85	29.68	32.95	27.80	29.05
DBSCAN[54]	39.26	42.63	34.76	36.08	31.25	33.52
Mutual Learning	42.92	44.19	38.52	40.41	35.86	38.83
Dynamic Update	46.35	49.69	40.02	43.27	37.18	40.77
Dual Mutual Consistency	48.52	51.05	41.25	46.50	39.21	43.41
DMDU(Ours)	**50.36**	**53.98**	**44.50**	**48.17**	**40.38**	**45.09**

TABLE VI
THE DIFFERENT PSEUDO-LABEL GENERATION STRATEGY PERFORMANCE(%) COMPARISON ON UDA VEHICLE RE-ID TASK.

Pseudo-label Generation Strategy	Method	VehicleID-to-Veri-776		Veri-776-to-VehicleID	
		Rank-1	mAP	Rank-1	mAP
original	PAL	68.17	42.04	44.25	48.05
	ML	77.83	36.92	40.30	48.72
	PLM	77.59	47.37	45.40	49.41
DMDU(Ours)	PAL	70.98	44.38	45.39	49.47
	ML	78.93	38.25	42.74	50.03
	PLM	78.60	48.39	47.62	53.81

vehicle attribute information based on descriptive words), and CLIP[48]. We show the experiment results in Table III. Introducing attribute information by short sentence has achieved 96.98% in Rank-1, 97.34% in Rank-5, and 82.35% in mAP. Compared with the Transformer method based on descriptive words, the proposed method improves 2.83% in Rank-1, 0.47% in Rank-5, and 2.05% in mAP. It is evident that the Transformer network structure has a more vital ability to learn hidden features from sentences than words. In summary, the multi-modal feature fusion capability of the Transformer can effectively guide the network to extract vehicle attitude information better, which is beneficial to generate a more generalizable and expressive vehicle feature.

3) Analysis of Pseudo-Label Generation Strategy: We perform experiments on the UDA vehicle re-ID task to confirm the efficacy of the DMDU pseudo-label generation strategy. This is achieved by comparing the identification accuracy of different generation strategies in the target domain. Other operations during the experiment remained the same. For the feature extraction model, we choose to employ ResNet-50 in the experiment, and the baseline is the direct transfer model. In the experiment, we chose the DBSCAN clustering algorithm, the mutual learning method, the dual mutual consistency strategy, and the dynamic update strategy for comparison.

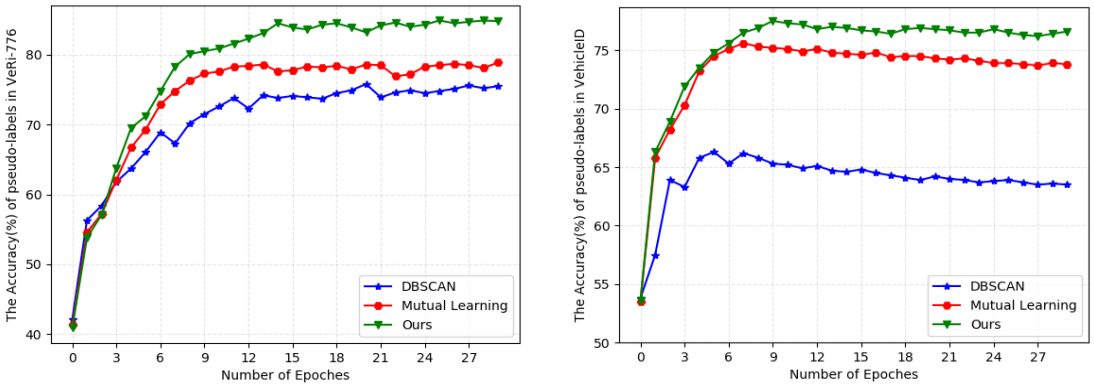
We report the comparisons of VehicleID-to-VeRi-776 in Table IV. The source dataset is VehicleID and tested on VeRi-776. Form the compare results, the proposed DMDU strategy achieves 76.32% Rank-1, 83.25% Rank-5, and 45.20% mAP for the VehicleID-to-VeRi-776 experiment. Our pseudo-label generation strategy has shown significant re-identification accuracy in the UDA vehicle Re-ID task. Compared to Dynamic Update and Dual Mutual Consistency, our method can achieve 7.56% Rank-1, 7.94% Rank-5, and 5.56% mAP improvement on the VehicleID-to-VeRi-776 experiment.

The results of VeRi-776-to-VehicleID are shown in Table V. The source dataset is VeRi-776 and tested on VehicleID. Our proposed method accomplishes the best-identified result for the VeRi-776-to-VehicleID experiment. Compared to Dynamic Update and Dual Mutual Consistency, our method achieves significant improvement on the VeRi-776-to-VehicleID experiment. These display that the DMDU pseudo-label generation strategy can reduce the noise in the pseudo-label by mutual consistency constraints and dynamic updates.

Meanwhile, to further validate the effectiveness and portability of the DMDU pseudo-label generation strategy, we replace the pseudo-label generation strategies in common progressive learning-based UDA vehicle Re-ID methods and observe the changes in their performance, including PAL[67], ML[68], and PLM[69]. The experimental results are shown in TableVI. It can be seen that after replacing the original pseudo-label generation strategy in these methods with the DMDU strategy, the accuracy of UDA vehicle Re-ID has been improved, which proves the effectiveness and portability of the DMDU strategy.

On this basis, in order to make a more intuitive comparison of the quality of different strategies generated pseudo-labels. We show the accurateness comparison of the generated pseudo-labels on VehicleID-to-VeRi-776 and VeRi-776-to-VehicleID UDA vehicle Re-ID task, as shown in Fig. 5. It can be seen that with the help of the proposed DMDU strategy, the quality of the generated pseudo-labels has been significantly improved. This demonstrates that the dual mutual consistency can help reduce the noise introduced in the clustering process and generate more reliable simple sample pseudo-labels in the target domain. Meanwhile, dynamic updating can further reduce the noise of hard sample pseudo-labels. Thus, more supervision information can be provided to assist in improving the accuracy of UDA vehicle Re-ID.

4) Analysis of Hyper-Parameters: First, we conduct some experiments to analyze the impact of vehicle attributes embedding weight μ on the vehicle feature learning effect, and the



(a) The Accuracy of Pseudo-Label in VeRi-776 (b) The Accuracy of Pseudo-Label in VehicleID

Fig. 5. The illustration of comparison of pseudo-label accuracy generated by different pseudo-label generation methods in VehicleID-to-VeRi-776 and VeRi-776-to-VehicleID UDA vehicle Re-ID tasks. (a) is on VeRi-776. (b) is on VehicleID.

TABLE VII
THE DIFFERENT VEHICLE ATTRIBUTE EMBEDDING WEIGHT PERFORMANCE(%) COMPARISON ON VEHICLE RE-ID TASK.

Dataset		$\mu = 0$	$\mu = 1.0$	$\mu = 2.0$	$\mu = 3.0$	$\mu = 4.0$
VeRi-776	Rank-1	48.26	64.85	89.17	**97.57**	95.30
	mAP	36.47	53.11	75.69	**82.56**	80.75
VehicleID	Rank-1	42.89	59.73	81.37	**92.47**	88.32
	mAP	29.73	49.27	69.84	**80.56**	75.21

TABLE VIII
THE DIFFERENT LOSS WEIGHT PERFORMANCE(%) COMPARISON ON VEHICLEID-TO-VERI-776 TASK.

λ_{tri} ($\lambda_{ml} = 0.4$)	Rank-1	mAP	λ_{ml} ($\lambda_{tri} = 0.6$)	Rank-1	mAP
0.3	71.29	42.18	0.3	75.89	46.72
0.4	76.03	45.04	0.4	**79.05**	**48.07**
0.5	78.51	46.93	0.5	78.02	47.16
0.6	**79.05**	**48.07**	0.6	74.57	44.85
0.7	73.72	48.82	0.7	71.34	42.58

result is shown in Table VII. When $\mu = 0$, we achieve 48.26% Rank-1 and 36.47% mAP on VeRi-776, 42.89% Rank-1 and 29.73% mAP on VehicleID, respectively. When μ increases the effect of vehicle Re-ID also improves. When $\mu = 3.0$, the best performance is achieved with 97.57% Rank-1 and 82.56% mAP on VeRi-776, 92.47% Rank-1 and 80.56% mAP on VehicleID, respectively. The experimental results demonstrate that vehicle semantic attribute embedding is beneficial for improving the discriminative of vehicle features.

Then, to determine λ_{tri} and λ_{ml} in the loss function, we conduct some experiments on the VehicleID-to-VeRi-776 task, as shown in Table VIII. It is clearly shown that $\lambda_{tri} = 0.6$ and $\lambda_{ml} = 0.4$ is the best choice, which obtains the best performance of 79.05%, 48.07% on Rank-1 and mAP respectively. This is because two different loss functions can achieve balance and improve the training effect of the network.

C. Comparison with the state-of-the-art methods and Discuss

To evaluate the superiority of our proposed UDA vehicle Re-ID method, we compare our method with several state-of-the-art UDA techniques. In this experiment, we conducted the comparisons on the two domain adaption tasks, VehicleID-to-VeRi-776 and VeRi-776-to-VehicleID. Table IX and Table X report the comparisons. Our proposed UDA vehicle Re-ID method significantly outperforms all existing UDA vehicle Re-ID methods. In the experiment, the direct transfer means directly utilizing the well-trained ResNet-50-based vehicle Re-ID model on the source domain to the target domain. The baseline method is direct employs the MATNet model trained in the source domain data to the target domain. It should be noted that, same as [69], since there are few pieces of methods focused on UDA vehicle Re-ID methods at present, the UDA pedestrian Re-ID method PUL is introduced to the experiment for comparison. Moreover, we did not use re-ranking or multi-query fusion in experiments. The selected methods for comparison can be divided into cross-domain transfer learning-based methods and progressive learning-based methods.

TABLE IX
COMPARISON OF THE PROPOSED METHOD WITH THE DIFFERENT UDA VEHICLE RE-ID METHODS ON VEHICLEID-TO-VERI-776 TASK. THE BEST RESULT FOR EACH INDICATOR WILL BE BOLDED.

Type	Method	Reference	Rank-1	Rank-5	mAP
	Direct Transfer	-	57.41	67.02	18.93
	Baseline	-	65.58	79.42	38.49
Cross-domain Transfer Learning	VR-PROUD[34]	PR'19	55.78	70.02	22.75
	ECN[70]	CVPR'19	57.41	70.53	20.06
	UDAR[33]	PR'20	76.90	85.80	35.80
	VDAF[71]	MTA'23	46.32	55.17	24.86
Progressive Learning	PAL[67]	IJCAI'20	68.17	79.91	42.04
	ML[68]	ICME'21	77.80	85.50	36.90
	PLM[69]	Sci.China'22	77.59	87.00	47.37
	CSP+FCD[72]	Eletronics'23	74.30	83.70	45.60
	MATNet+DMDU(Ours)	-	**79.13**	**88.97**	**49.25**

From the experiment result in Table IX, it can be seen that in the VehicleID-to-VeRi-776 task, the proposed methods achieve 79.13% Rank-1, 88.97% Rank-5, and 49.25% mAP. Compared to the cross-domain transfer learning-based methods, the progressive learning-based methods can obtain more competitive UDA vehicle Re-ID performance because they can

TABLE X
COMPARISON OF THE PROPOSED METHOD WITH THE DIFFERENT UDA VEHICLE RE-ID METHODS ON VERI-776-TO-VEHICLEID TASK. THE BEST RESULT FOR EACH INDICATOR WILL BE BOLDED.

Type	Method	Reference	Small			Medium			Large		
			Rank-1	Rank-5	mAP	Rank-1	Rank-5	mAP	Rank-1	Rank-5	mAP
	Direct Transfer	-	36.58	56.92	41.83	27.67	49.34	32.95	27.08	45.17	31.60
	Baseline	-	45.11	59.42	49.69	39.85	55.19	45.37	38.10	53.14	43.28
Cross-domain Transfer Learning	CycleGAN[73]	ICCV'17	37.29	58.56	43.32	30.00	49.96	34.92	27.15	46.52	31.89
	UDAR[33]	PR'20	54.00	66.10	59.60	48.10	64.10	55.30	45.20	**62.60**	52.90
	VDAF[71]	MTA'23	-	-	-	47.03	64.86	-	43.69	61.76	-
Progressive Learning	PAL[67]	IJCAI'20	50.25	64.91	53.50	44.25	60.95	48.05	41.08	59.12	45.14
	ML[68]	ICME'21	54.80	69.20	61.60	40.30	57.70	48.70	36.50	54.10	45.00
	PLM[69]	Sci.China'22	51.23	67.11	54.85	45.40	63.37	49.41	41.73	60.94	46.00
	CSP+FCD[72]	Eletronics'23	54.40	67.40	51.90	52.70	**65.60**	46.50	45.90	60.30	42.70
	MATNet+DMDU(Ours)	-	**55.61**	**68.25**	**61.83**	**53.28**	63.56	**56.73**	**47.59**	61.85	**53.97**

more comprehensively introduce the information in the target domain. In the experiment, the proposed method achieves the optimum. Especially compared with the best cross-domain transfer learning-based method UDAR[33], our method has achieved 2.23% in Rank-1, 3.17% in Rank-5, and 13.45% in mAP improvement. Meanwhile, compared with PLM[69], which is also based on progressive learning, the proposed method has also significantly improvement, improving 1.54% on Rank-1, improving 1.97% on Rank-5, and improving 1.88% on mAP. This is attributed to the fact that the MATNet is able to extract more discriminative vehicle features while the DMDU strategy alleviates the noise contained in the generated pseudo-labels.

Further, the experiment results in the VehicleID, which has a more complex vehicle Re-ID scene, are shown in Table X. We can learn that the proposed method achieves the most satisfactory performance among all approaches, with Rank-1 = 55.61%, 53.28%, 47.59%, Rank-5 = 68.25%, 63.56%, 61.85% and mAP = 61.83%, 56.73%, 53.97% on VehicleID with the test set of small, medium, large, respectively. Similarly, compared with the progressive learning-based method CSP+FCD[72], our method has 1.21%, 0.58%, and 1.69% gains in Rank-1, 0.85% and 1.55% improvement in Rank-5, 9.93%, 10.23%, and 11.27% gains in mAP on VehicleID with different test sets. It shows that the proposed method can perform better UDA vehicle Re-ID in more complex scenarios.

In addition, the UDA vehicle Re-ID method performs better in the VehicleID-to-VeRi-776 task than in the VeRi-776-to-VehicleID task. The reason for that is the feature extraction ability of the pre-trained model is stronger when the source domain data is complex and challenging. In this way, it is more adaptable to the training of the target domain and can obtain better re-identification results. In addition, when the pre-trained model is directly transferred to the new domain, the Re-ID performance will be affected. However, the proposed Re-ID model achieves the best result in both datasets while also achieving competitive Re-ID performance when direct transfer. These two experiment results reflect the effectiveness of our designed MATNet.

V. CONCLUSION

In this paper, we present a novel UDA vehicle Re-ID method that optimizes from the two perspectives of improving vehicle feature discrimination and alleviating pseudo-label noise, including MATNet and DMDU strategy. MATNet aims to utilize the multi-modal feature fusion capability of the Transformer network to introduce vehicle attribute information, which is beneficial to lighten typical challenges in vehicle Re-ID. Then, the DMDU strategy can rely on dual mutual consistency to generate more credible pseudo-labels for simple samples. The dynamic update operation can effectively alleviate the noise in hard sample pseudo-labels. Thus providing more accurate supervisory information for the training of UDA vehicle Re-ID. Moreover, we demonstrate the effectiveness and extendibility of our proposed MATNet and DMDU through extensive experiments on two challenging vehicle re-identification datasets. Observing the experimental results, we found that the performance of cross-domain is closely related to attribute labeling. In future work, we will conduct more in-depth research from two aspects. First, we will augment the classification of attribute multi-labels or directly generate more accurate textual descriptions with the Sentence Transformer model. Secondly, we try to employ semantic segmentation methods to introduce precise and detailed semantic attribute information as guidance, so that the network can extract more discriminative features. These two measures will enhance the performance of the UDA vehicle Re-ID.

ACKNOWLEDGMENTS

This work was supported in part by the National Key Research and Development Program of China under Grant 2022YFB3303800, the National Natural Science Foundation of China under Grant U21A20482, No.62073117, No.62272018.

REFERENCES

[1] H. Liu, Y. Tian, *et al.*, "Deep relative distance learning: Tell the difference between similar vehicles," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2167–2175, 2016.
[2] J. Zhao, Y. Zhao, J. Li, *et al.*, "Heterogeneous relational complement for vehicle re-identification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 205–214, 2021.

[3] Z. Lu, R. Lin, and H. Hu, "Mart: Mask-aware reasoning transformer for vehicle re-identification," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 2, pp. 1994–2009, 2022.

[4] J. Peng, H. Wang, F. Xu, *et al.*, "Cross domain knowledge learning with dual-branch adversarial network for vehicle re-identification," *Neurocomputing*, vol. 401, pp. 133–144, 2020.

[5] Z. Zhou, Y. Li, J. Li, *et al.*, "Gan-siamese network for cross-domain vehicle re-identification in intelligent transport systems," *IEEE Transactions on Network Science and Engineering*, pp. 1–12, 2022.

[6] R. Quispe, C. Lan, W. Zeng, *et al.*, "Attributenet: Attribute enhanced vehicle re-identification," *Neurocomputing*, vol. 465, pp. 84–92, 2021.

[7] R. Zhang, X. Zhong, X. Wang, *et al.*, "Graph-based structural attributes for vehicle re-identification," in *2022 IEEE International Conference on Multimedia and Expo*, pp. 1–6, IEEE, 2022.

[8] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, pp. 6000–6010, 2017.

[9] S. He, H. Luo, P. Wang, *et al.*, "Transreid: Transformer-based object re-identification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15013–15022, 2021.

[10] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[11] Y. Zou, Z. Yu, *et al.*, "Unsupervised domain adaptation for semantic segmentation via class-balanced self-training," in *Proceedings of the European Conference on Computer Vision*, pp. 289–305, 2018.

[12] Y. Zou, Z. Yu, X. Liu, *et al.*, "Confidence regularized self-training," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5982–5991, 2019.

[13] Y. Zhang, T. Xiang, T. M. Hospedales, *et al.*, "Deep mutual learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4320–4328, 2018.

[14] X. Liu, W. Liu, H. Ma, *et al.*, "Large-scale vehicle re-identification in urban surveillance videos," in *2016 IEEE International Conference on Multimedia and Expo*, pp. 1–6, IEEE, 2016.

[15] J. Hou, H. Zeng, J. Zhu, *et al.*, "Deep quadruplet appearance learning for vehicle re-identification," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 9, pp. 8512–8522, 2019.

[16] L. Wei, X. Liu, J. Li, *et al.*, "Vp-reid: Vehicle and person re-identification system," in *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, pp. 501–504, 2018.

[17] Y. Zhang, D. Liu, *et al.*, "Improving triplet-wise training of convolutional neural network for vehicle re-identification," in *2017 IEEE International Conference on Multimedia and Expo*, pp. 1386–1391, IEEE, 2017.

[18] Z. Wang, L. Tang, X. Liu, *et al.*, "Orientation invariant feature embedding and spatial temporal regularization for vehicle re-identification," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 379–387, 2017.

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

[20] K. Li, Z. Ding, K. Li, *et al.*, "Vehicle and person re-identification with support neighbor loss," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 2, pp. 826–838, 2020.

[21] Y. Zhou, L. Liu, and L. Shao, "Vehicle re-identification by deep hidden multi-view inference," *IEEE Transactions on Image Processing*, vol. 27, no. 7, pp. 3275–3287, 2018.

[22] Y. Chen, B. Ma, and H. Chang, "Part alignment network for vehicle re-identification," *Neurocomputing*, vol. 418, pp. 114–125, 2020.

[23] A. M. N. Taufique and A. Savakis, "Labnet: Local graph aggregation network with class balanced loss for vehicle re-identification," *Neurocomputing*, vol. 463, pp. 122–132, 2021.

[24] Y. Cheng, C. Zhang, K. Gu, *et al.*, "Multi-scale deep feature fusion for vehicle re-identification," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1928–1932, IEEE, 2020.

[25] X. Liu, W. Liu, J. Zheng, *et al.*, "Beyond the parts: Learning multi-view cross-part correlation for vehicle re-identification," in *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 907–915, 2020.

[26] S. Zhang, H. Tong, *et al.*, "Graph convolutional networks: a comprehensive review," *Computational Social Networks*, vol. 6, pp. 1–23, 2019.

[27] H. Li, X. Lin, A. Zheng, *et al.*, "Attributes guided feature learning for vehicle re-identification," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 6, no. 5, pp. 1211–1221, 2021.

[28] H. Li, C. Li, A. Zheng, *et al.*, "Attribute and state guided structural embedding network for vehicle re-identification," *IEEE Transactions on Image Processing*, vol. 31, pp. 5949–5962, 2022.

[29] Y. Tang, D. Wu, *et al.*, "Multi-modal metric learning for vehicle re-identification in traffic surveillance environment," in *2017 IEEE International Conference on Image Processing*, pp. 2254–2258, IEEE, 2017.

[30] X. Zhu, Z. Luo, P. Fu, *et al.*, "Vocreld: Vehicle re-identification based on vehicle-orientationcamera. 2020 ieee," in *CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 2566–2573, 2020.

[31] L. Zhang, Q. Diao, N. Jiang, *et al.*, "Mutual purification for unsupervised domain adaptation in person re-identification," *Neural Computing and Applications*, vol. 34, no. 19, pp. 16929–16944, 2022.

[32] L. Wei, S. Zhang, W. Gao, *et al.*, "Person transfer gan to bridge domain gap for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 79–88, 2018.

[33] L. Song, C. Wang, *et al.*, "Unsupervised domain adaptive re-identification: Theory and practice," *Pattern Recognition*, vol. 102, pp. 107–173, 2020.

[34] R. M. S. Bashir, M. Shahzad, and M. Fraz, "Vr-proud: Vehicle re-identification using progressive unsupervised deep architecture," *Pattern Recognition*, vol. 90, pp. 52–65, 2019.

[35] H. Fan, L. Zheng, *et al.*, "Unsupervised person re-identification: Clustering and fine-tuning," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 14, no. 4, pp. 1–18, 2018.

[36] J. Liu, Z. Zha, *et al.*, "Adaptive transfer network for cross-domain person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7202–7211, 2019.

[37] J. Peng, H. Wang, *et al.*, "Cross domain knowledge transfer for unsupervised vehicle re-identification," in *2019 IEEE International Conference on Multimedia and Expo Workshops*, pp. 453–458, IEEE, 2019.

[38] Y. Fu, Y. Wei, G. Wang, *et al.*, "Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification," in *proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6112–6121, 2019.

[39] A. Zheng, X. Sun, C. Li, *et al.*, "Aware progressive clustering for unsupervised vehicle re-identification," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 8, pp. 11422–11435, 2021.

[40] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *Advances in Neural Information Processing Systems*, vol. 30, pp. 1195–1204, 2017.

[41] Z. Chen, Z. Cui, C. Zhang, J. Zhou, and Y. Liu, "Dual clustering co-teaching with consistent sample mining for unsupervised person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 5908–5920, 2023.

[42] L. Qi, J. Liu, *et al.*, "Unsupervised generalizable multi-source person re-identification: A domain-specific adaptive framework," *Pattern Recognition*, vol. 140, p. 109546, 2023.

[43] X. Yun, Q. Wang, *et al.*, "Discrepant mutual learning fusion network for unsupervised domain adaptation on person re-identification," *Applied Intelligence*, vol. 53, pp. 2951–2966, 2023.

[44] M. Zhang, K. Li, J. Ma, and X. Wang, "Asymmetric double networks mutual teaching for unsupervised person re-identification," *Neural Networks*, vol. 169, pp. 744–755, 2024.

[45] S. Chen, L. Qiu, Z. Tian, Y. Yan, D.-H. Wang, and S. Zhu, "Mtnet: Mutual tri-training network for unsupervised domain adaptation on person re-identification," *Journal of Visual Communication and Image Representation*, vol. 90, p. 103749, 2023.

[46] H. Zhou, J. Kong, M. Jiang, and T. Liu, "Heterogeneous dual network with feature consistency for domain adaptation person re-identification," *International Journal of Machine Learning and Cybernetics*, vol. 14, no. 5, pp. 1951–1965, 2023.

[47] B. Han, Q. Yao, X. Yu, *et al.*, "Co-teaching: Robust training of deep neural networks with extremely noisy labels," *Advances in Neural Information Processing Systems*, vol. 31, pp. 8536–8546, 2018.

[48] A. Radford, J. W. Kim, C. Hallacy, *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*, pp. 8748–8763, PMLR, 2021.

[49] T.-Y. Lin, M. Maire, S. Belongie, *et al.*, "Microsoft coco: Common objects in context," in *Proceedings of the European Conference on Computer Vision*, pp. 740–755, 2014.

[50] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *arXiv preprint arXiv:1908.10084*, 2019.

[51] Y. Wang, X. Ma, Z. Chen, *et al.*, "Symmetric cross entropy for robust learning with noisy labels," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 322–330, 2019.

[52] S. Paisitkriangkrai, C. Shen, *et al.*, "Learning to rank in person re-identification with metric ensembles," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1846–1855, 2015.

[53] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1116–1124, 2015.

[54] M. Ester, H.-P. Kriegel, J. Sander, *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *International Conference on Knowledge Discovery and Data Mining*, vol. 96, pp. 226–231, 1996.

[55] X. Chen, H. Sui, *et al.*, "Vehicle re-identification using distance-based global and partial multi-regional feature learning," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, pp. 1276–1286, 2020.

[56] M. Li, X. Huang, and Z. Zhang, "Self-supervised geometric features discovery via interpretable attention for vehicle re-identification and beyond," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 194–204, 2021.

[57] J. Lian, D.-H. Wang, Y. Wu, and S. Zhu, "Multi-branch enhanced discriminative network for vehicle re-identification," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1263–1274, 2023.

[58] J. Qian, M. Pan, W. Tong, R. Law, and E. Q. Wu, "Urrnet: A unified relational reasoning network for vehicle re-identification," *IEEE Transactions on Vehicular Technology*, pp. 11156–11168, 2023.

[59] M. Li, J. Liu, C. Zheng, *et al.*, "Exploiting multi-view part-wise correlation via an efficient transformer for vehicle re-identification," *IEEE Transactions on Multimedia*, vol. 25, pp. 919–929, 2021.

[60] Z. Yu, J. Pei, M. Zhu, *et al.*, "Multi-attribute adaptive aggregation transformer for vehicle re-identification," *Information Processing & Management*, vol. 59, no. 2, p. 102868, 2022.

[61] F. Shen, Y. Xie, J. Zhu, *et al.*, "Git: Graph interactive transformer for vehicle re-identification," *IEEE Transactions on Image Processing*, vol. 32, pp. 1039–1051, 2023.

[62] Z. Yu, Z. Huang, *et al.*, "Semantic-oriented feature coupling transformer for vehicle re-identification in intelligent transportation system," *IEEE Transactions on Intelligent Transportation Systems*, pp. 2803–2813, 2023.

[63] X. Jin, C. Lan, W. Zeng, *et al.*, "Uncertainty-aware multi-shot knowledge distillation for image-based object re-identification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 11165–11172, 2020.

[64] Y. Li, K. Liu, Y. Jin, *et al.*, "Varid: Viewpoint-aware re-identification of vehicle based on triplet loss," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 2, pp. 1381–1390, 2020.

[65] F. Shen, J. Zhu, X. Zhu, *et al.*, "Exploring spatial significance via hybrid pyramidal graph network for vehicle re-identification," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 7, pp. 8793–8804, 2021.

[66] Y. Bai, J. Liu, Y. Lou, *et al.*, "Disentangled feature learning network and a comprehensive benchmark for vehicle re-identification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 6854–6871, 2021.

[67] J. Peng, Y. Wang, *et al.*, "Unsupervised vehicle re-identification with progressive adaptation," *arXiv preprint arXiv:2006.11486*, 2020.

[68] H. Wang, J. Peng, *et al.*, "Learning multiple semantic knowledge for cross-domain unsupervised vehicle re-identification," in *2021 IEEE International Conference on Multimedia and Expo*, pp. 1–6, IEEE, 2021.

[69] Y. Wang, J. Peng, H. Wang, *et al.*, "Progressive learning with multi-scale attention network for cross-domain vehicle re-identification," *Science China Information Sciences*, vol. 65, no. 6, p. 160103, 2022.

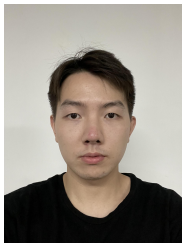
[70] Z. Zhong, L. Zheng, *et al.*, "Invariance matters: Exemplar memory for domain adaptive person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 598–607, 2019.

[71] F. Zhang, L. Zhang, H. Zhang, and Y. Ma, "Image-to-image domain adaptation for vehicle re-identification," *Multimedia Tools and Applications*, vol. 82, no. 26, pp. 40559–40584, 2023.

[72] G. Zhan, Q. Wang, *et al.*, "Unsupervised vehicle re-identification based on cross-style semi-supervised pre-training and feature cross-division," *Electronics*, vol. 12, no. 13, p. 2931, 2023.

[73] A. Almahairi, S. Rajeshwar, *et al.*, "Augmented cyclegan: Learning many-to-many mappings from unpaired data," in *International Conference on Machine Learning*, pp. 195–204, 2018.

Xin Zhang is a postdoctoral researcher at the School of Artificial Intelligence, Hebei University of Technology, Tianjin, China. He received his B.S. and M.S. degrees from the North University of China, Taiyuan, China, in 2015 and 2018, respectively. He got his Ph.D. degree from Beihang University in 2023. His research interests include Vehicle Re-Identification, Person Re-Identification, Multi-Object Tracking, and Computer Vision.



Yunan Ling received the B.E. degree in software engineering from Jilin University, Changchun, China, in 2021. He is currently pursuing the M.S. degree with the State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, China. His research interests include Person Re-Identification and Deep Learning.

Kaige Li received M.S. degree from the Ocean University of China, Qingdao, China, in 2020. He is currently pursuing the Ph.D. degree with the State Key Laboratory of Virtual Reality Technology and Systems, Computer Science and Technology, Beihang University, Beijing, China. His current research interests include deep learning, image semantic segmentation, computer vision, and smart city.

Weimin Shi received the MS. degree from Beijing University of Chemical Technology in 2021. He is currently pursuing the Ph.D. degree in the State Key Laboratory of Virtual Reality Technology and Systems, computer science and technology from Beihang University, Beijing, China. His current research interests include deep learning, multi-modal learning, computer vision, and smart city.

Zhong Zhou Professor, Ph.D. adviser, State Key Lab of Virtual Reality Technology and Systems, Beihang University, Beijing, China. He got his B.S. degree from Nanjing University and Ph.D. degree from Beihang University in 1999 and 2005 respectively. His main research interests include Virtual Reality/Augmented Reality/Mixed Reality, Computer Vision and Artificial Intelligence. He is the member of IEEE, ACM, and CCF.