Viewpoint-Aware Progressive Clustering for Unsupervised Vehicle Re-Identification

Aihua Zheng, Xia Sun, Chenglong Li, and Jin Tang

***Abstract*— Vehicle re-identification (Re-ID) is an active task due to its importance in large-scale intelligent monitoring in smart cities. Despite the rapid progress in recent years, most existing methods handle vehicle Re-ID task in a supervised manner, which is both time and labor-consuming and limits their application to real-life scenarios. Recently, unsupervised person Re-ID methods achieve impressive performance by exploring domain adaption or clustering-based techniques. However, one cannot directly generalize these methods to vehicle Re-ID since vehicle images present huge appearance variations in different viewpoints. To handle this problem, we propose a novel viewpoint-aware clustering algorithm for unsupervised vehicle Re-ID. In particular, we first divide the entire feature space into different subspaces according to the predicted viewpoints and then perform a progressive clustering to mine the accurate relationship among samples. Comprehensive experiments against the state-of-the-art methods on two multi-viewpoint benchmark datasets VeRi-776 and VeRi-Wild validate the promising performance of the proposed method in both with and without domain adaption scenarios while handling unsupervised vehicle Re-ID.**

***Index Terms*— Viewpoint-aware, progressive clustering, vehicle Re-ID, unsupervised learning.**

I. Introduction

VEHICLE re-identification aims to identify a specific vehicle in non-overlapping camera networks. It is a crucial task in modern society with potential applications in artificial transportation, smart city and public security, to name a few. Similar to the person Re-ID task, vehicle Re-ID faces common challenges such as illumination and viewpoint changes across cameras, background clutters, and occlusions. Besides, vehicle Re-ID dramatically suffers from the challenges of large intra-class discrepancy and inter-class similarity. This is because different vehicles might present exactly similar appearance while the same vehicle might present totally different features, as shown in Fig. 1. Therefore, one cannot directly deploy person Re-ID models to achieve satisfactory performance in vehicle Re-ID.

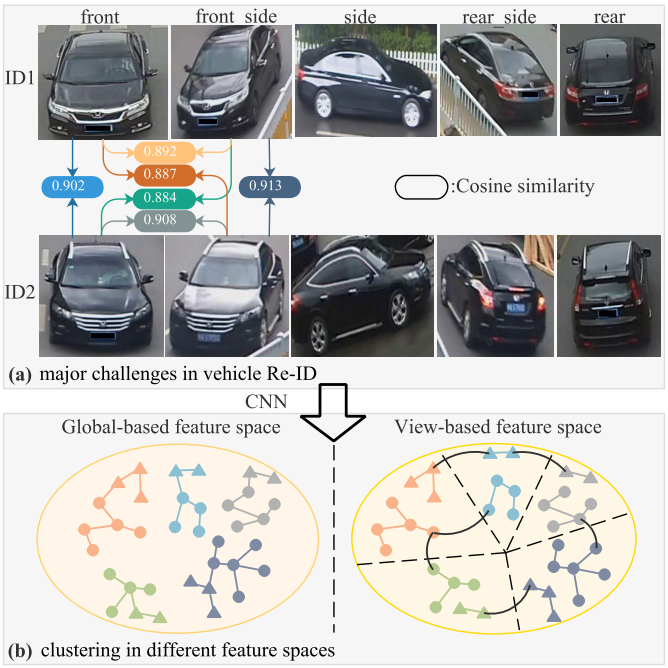


Fig. 1. Major challenges and clustering in different feature spaces in vehicle Re-ID. As shown in (a), different vehicles with the same viewpoint have higher visual similarity than those same vehicles with different viewpoints. These examples demonstrate that vehicle Re-ID greatly suffers from the challenges of large intra-class discrepancy and inter-class similarity. In (b), the same color represents the same viewpoint, and the same shape represents the same identity. The global-based method tends to prioritize connections between the same viewpoint (as shown with the same color) even for the different identities (as shown with the different shape) instead of the same identity samples from different viewpoints. Our method can correctly connect the samples from the same identity with different viewpoints (solid black line) through progressive clustering of view-based division (dashed line).

With the blossom of deep learning techniques and its powerful learning ability on large labeled data, various supervised learning architectures [1]–[9] have been proposed and achieved remarkable performance for vehicle Re-ID. Despite great progress, supervised learning-based methods require numerous annotations to train the deep models, which are time and labor-consuming and significantly limit real-life applications of vehicle Re-ID.

Domain adaptation, which transfers the learned information from the source domain (labeled data) to the target domain (unlabeled data), has been widely explored in the past decade as one of unsupervised learning manners in both person Re-ID [10]–[13] and vehicle Re-ID [14], [15]. However,

Manuscript received November 17, 2020; revised May 21, 2021; accepted August 1, 2021. This work was supported in part by the University Synergy Innovation Program of Anhui Province under Grant GXXT-2019-007 and Grant GXXT-2020-051; in part by the National Natural Science Foundation of China under Grant 61976002, Grant 61976003, and Grant 62076003; and in part by the Natural Science Foundation of Anhui Higher Education Institutions of China under Grant KJ2019A0033. The Associate Editor for this article was W. Lin. *(Corresponding author: Chenglong Li.)*

Aihua Zheng and Chenglong Li are with the Information Materials and Intelligent Sensing Laboratory of Anhui Province, Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, School of Artificial Intelligence, Anhui University, Hefei 230601, China (e-mail: ahzheng214@foxmail.com; lcl1314@foxmail.com).

Xia Sun and Jin Tang are with the Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, School of Computer Science and Technology, Anhui University, Hefei 230601, China (e-mail: sunxia233@foxmail.com; tangjin@ahu.edu.cn).

Digital Object Identifier 10.1109/TITS.2021.3103961

they still require large annotations in the source domain. In addition, when the style gap between the two domains is too large, these transfer learning methods are also limited.

Different from domain adaptation-based methods, we study the problem of vehicle Re-ID in the target-only unsupervised learning framework, which does not rely on any labeled data in the source domain. As one of the target-only unsupervised methods, clustering-based methods have been widely explored in the related computer vision tasks [15]–[20]. Recent efforts on clustering-based methods in person Re-ID are to assign pseudo labels for samples by clustering algorithms and then use these labeled samples to train Re-ID models [15], [17], [18], [21].

However, one cannot directly apply these techniques to vehicle Re-ID. One of the key reasons is large viewpoint variations of vehicles, which bring big challenges to clustering algorithms. As shown in Fig. 1 (a), by directly calculating the cosine similarity between vehicle images, we can see the similarity between the same vehicle images with different viewpoints is even lower than that between different vehicles in the same viewpoint, which is referred to as the similarity dilemma of vehicles in this paper. Due to the inter-instance similarity and intra-instance discrepancy caused by large viewpoint variations of vehicles, the accuracy of clustering algorithms is significantly affected. As shown in Fig. 1 (b), in the global-based feature space, the same viewpoints with different identities will be clustered preferentially than different viewpoints with the same identity, resulting in the extremely degraded performance of vehicle Re-ID.

To handle this problem, we propose a novel viewpoint-aware progressive clustering framework (VAPC) for robust unsupervised vehicle Re-ID. In Fig. 1 (a), we observe that vehicle images from different viewpoints of the same ID are more similar than vehicle images from different viewpoints of different IDs, e.g., image pairs {ID1 (*front*), ID1 (*front_side*)} are more similar than {ID1 (*front*), ID2 (*front_side*)}. Therefore we can divide the vehicles into different view-based feature spaces. After clustering within the same viewpoint, the same ID from different viewpoints can be correctly classified according to the degree of similarity, as shown in Fig. 1 (b). In addition, the vehicles in each viewpoint space exclude the effects of large viewpoint variations. When only performing the clustering between samples of the same viewpoint, the comparison of different viewpoints with the same ID is excluded, which further reduces the intra-class differences and simplifies the clustering task. Therefore, we propose a viewpoint-aware progressive clustering framework, which can be regarded as three parts. First, considering the extreme viewpoint changes of the vehicle, we design a viewpoint-aware network, which can be pre-trained using viewpoint annotations [22], to predict viewpoints of vehicle images as the prior information. Second, feature extraction is crucial to the performance of clustering. To extract the discriminative feature of each sample, it is necessary to train an initial model with strong feature extraction capabilities. In this paper, we use a self-supervised manner to learn the discriminative feature of each sample. Without the ground truth labels in the target-only unsupervised learning, we treat each sample as a category and force the network to learn the discriminative feature of each sample via the repelled loss [17], [23], which we call the recognition stage. Third, we design a viewpoint-aware progressive clustering algorithm to handle the problem of the similarity dilemma discussed above. Specifically, we first perform clustering in each vehicle image set with the same viewpoint and then cluster them by comparing the similarity of clusters across different viewpoints. In this way, we can distinguish small gaps between different identities in the same viewpoint, and mine the same identity samples with large gaps between different viewpoints.

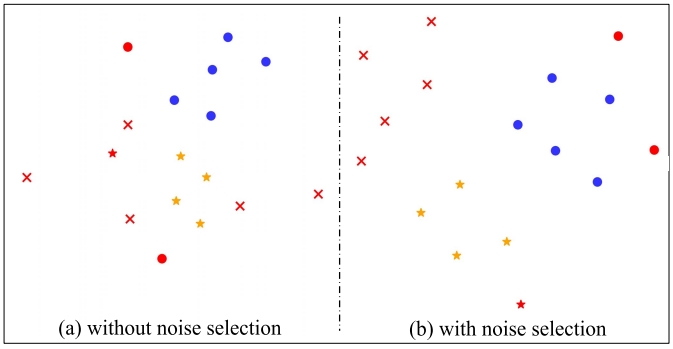


Fig. 2. t-SNE [24] distance distribution with and without noise selection on VeRi-776 [1]. The same shape and color represent the same identities belong to the same cluster, where red ones indicates the noise samples that are not clustered (with pseudo label as -1).

We use the clustered results to train the Re-ID network in a supervised way after progressive viewpoint-aware clustering. However, the clustering performance of different viewpoints significantly relies on the clustering results from the same viewpoint. Therefore, we introduce the k-reciprocal encoding [15], [20], [25] as the distance metric to feature comparison of the same viewpoint due to its powerful ability in mining similar samples.

In addition, recent methods [15], [19], [20] achieve remarkable performance on target-only unsupervised person Re-ID. They directly employ the prevalent Density Based Spatial Clustering of Applications with Noise (DBSCAN) [26] to obtain pseudo labels. The key idea of DBSCAN [26] is to cluster the tightly distributed samples while regarding the others as noise. However, they all discard noisy samples (the hard positive and hard negative samples with pseudo labels assigned as -1). We argue that it is hard to deal with various challenges (illumination, occlusion, and camera offset) in real scenes by merely learning close simple samples. As shown in Fig. 2 (a), only learning closely distributed samples leaves noise samples to form feature embeddings with unclear boundaries. It is more important to learn the discriminative embeddings by mining hard positive samples, which has been proven in a large number of machine learning tasks [27]–[32]. To this end, we propose a noise selection method to mine hard positive samples from noise samples. Specifically, we classify each noise sample into a suitable cluster by the similarity between the noise sample and other clusters. After noise selection, Fig. 2 (b) pulls the noise samples closer to the samples with the same identity, which improves the generalization ability of the model to deal with more challenging scenarios.

Based on the above discussion, VAPC focuses on addressing unsupervised vehicle Re-ID through a viewpoint-aware progressive clustering framework. We alleviate the impact of vehicle similarity dilemmas on clustering by transforming global comparisons into progressive clustering based on viewpoint. To improve the clustering quality of the same viewpoint cluster, we introduced k-reciprocal encoding [15], [20], [25] as a distance metric for DBSCAN [26] clustering. In order to deal with outlier noise samples, we propose a noise selection method to improve the generalization ability of the model further. The major contributions of this work are summarized as follows.

- We propose a novel progressive clustering method to handle the similarity dilemma of vehicles in unsupervised vehicle Re-ID. To our best knowledge, this is the first time to employ the viewpoint-aware progressive clustering algorithm to achieve unsupervised vehicle Re-ID.
- We designed a noise selection scheme to mine the hard positive samples with the same identity while considering their relationship to the hard negative samples, which significantly improves the discriminative ability of our network.
- Comprehensive experimental results on two benchmark datasets, including VeRi-776 [1] and VeRi-Wild [33] demonstrate the promising performance of our method and yield to a new state-of-the-art for unsupervised vehicle Re-ID.

II. Related Works

Since most vehicle Re-ID methods are in a supervised fashion, we briefly review the progress in supervised vehicle Re-ID and recent advances in unsupervised person/vehicle Re-ID.

A. Vehicle Re-ID

Most existing deep vehicle re-identification methods follow a supervised setting. Pioneer vehicle Re-ID methods [3], [4], [34] focus on discriminative feature learning. Lou *et al.* [34] by mining similar negative samples, the features learned by the model are more robust. He *et al.* [4] proposed an efficient feature preserving method, which can enhance the perception ability of subtle differences. Some works introduce [5]–[7], [35]–[37] additional attribute information, such as color or type, to improve the discrimination of the deep feature for vehicle Re-ID. Temporal path information is also auxiliary information and has been widely employed [38], [39], to improve the robustness of vehicle Re-ID, especially for the vehicles with a similar appearance from the same manufacture. To handle the viewpoint variation issue, in person Re-ID, Lin *et al.* [40] propose to use more fine-grained information to describe individual persons and learn the matching probability of all patch between a camera pair [41] to guide fine matching to eliminate the influence of viewpoint variation. In vehicle Re-ID, Zhou *et al.* [5], [6], and Liu *et al.* [7] employ GAN to infer multi-view information from a single-view of the input image in either image or feature level to boost the performance by integrating the input and generated images or features. Chu *et al.* [42] separate the Re-ID into similar and different viewpoint modes and learn the respective deep metric for each case. In the case of a known 3D bounding box for the vehicle image, Sochor *et al.* [43] calculated orientation information through 3D coordinates and added it to the feature map to improve performance. Despite the significant progress on vehicle Re-ID, these supervised deep learning-based methods require extensive training data, which is expensive in both time and labor-consuming.

B. Unsupervised Person/Vehicle Re-ID

Along with the great achievement on person Re-ID, unsupervised person Re-ID offers more challenges, which has attracted more and more attention recently. Recent advances of unsupervised person Re-ID methods generally fall into two categories. 1) The domain adaption based methods [10]–[13], [44], [45], which aims to transfer the knowledge in the labeled source domain to the unlabeled target domain. Although the domain adaption based methods make impressive achievement in unsupervised Re-ID by exploring domain-invariant features, they still require a large amount of label annotation in the source domain. Furthermore, the huge diversity in different domains limits their transferring capabilities. 2) The target-only based methods [17], [18], [46], which fulfill the unsupervised task by dividing the unlabeled samples into different categories based on specific similarity. Lin *et al.* [17] treat each image as a single category and then gradually reduces the number of categories in subsequent clusters. Lin *et al.* [46] propose a framework that mines the similarity as a soft constraint and introduces camera information to encourage similar samples under different cameras to approach. In addition, the work related to video surveillance [47], [48] infers pseudo-labels based on an idea that samples with the same identity shoule be close to each other.

To the best of our knowledge, there are few works on unsupervised vehicle Re-ID. Peng *et al.* [14] propose to use a style GAN to generate vehicle pictures in the source domain more like the target domain. They assume that the source domain contains more viewpoints than the target domain for a better generation. Song *et al.* [15] introduce the theoretical guarantees of unsupervised domain adaptive Re-ID based on and use a self-training scheme to iteratively optimize the unsupervised domain adaptation model. However, it only focuses on unsupervised domain adaptation, not target-only unsupervised learning. Bashir *et al.* [49] employ clustering and reliable result selection with embedded color information to iteratively fine-tune the cascade network. However, despite the annotation on color information, this method requires a specific number of identities, which is hard to be known in real-life scenarios.

III. Proposed Approach

The pipeline of the proposed framework is shown in Fig. 3, which includes three parts: 1) viewpoint prediction, which identifies the viewpoint information through a viewpoint prediction network on input data, 2) recognition stage, which learns the discriminative feature for each sample using the

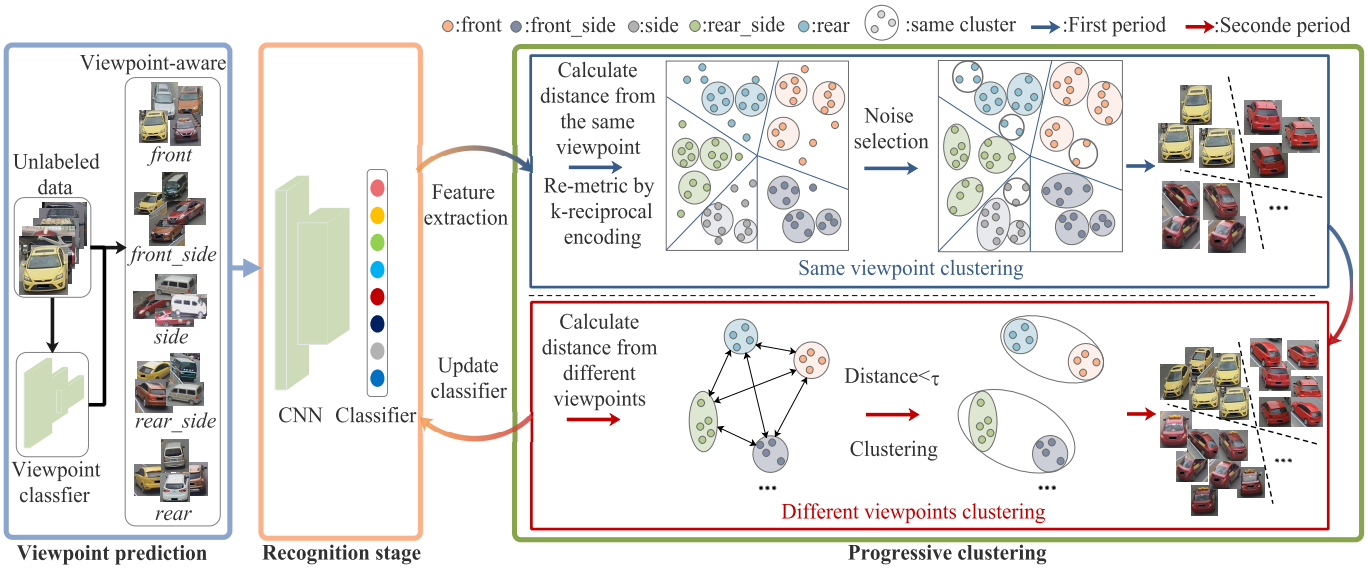


Fig. 3. The overview of our method framework. We first predict each viewpoint, and then the viewpoint-aware unlabeled training set is input to the CNN model for feature extraction, which can be divided into different directional feature clusters. Then we will go through a recognition stage to make each sample feature extracted by the network more identifying. We design a clustering method that divides direction and period. In the first period, we use DBSCAN [26] to generate initial clusters (colored background) within the same viewpoint. For the noise samples found in the clustering process, we design a noise selection method to select. After noise selection, the noise will be merged with the initial cluster or generating a new cluster (white background). In the second period, comparing the distances of all different viewpoints, clusters smaller than the distance threshold τ will be merged. The network is iteratively trained based on the final clustering results.

TABLE I
IMPORTANT NOTATIONS USED IN THIS PAPER

Notation	Meaning
y_{ind}	Index label
y_p	Pseudo label
N	Number of all images
N_v	Number of images in viewpoint v
s_i	The i-th noise sample index
x_i^v	The i-th image in viewpoint v
f_i^v	Feature of the i-th image in viewpoint v
d_{ij}	Distance between the i-th and j-th images
F_v	The set of all features in viewpoint v
$D(F_i, F_j)$	Distance between the feature set of viewpoint i and viewpoint j

repelled loss, and 3) progressive clustering, which uses the two-period algorithm to handle the problem of the similarity dilemma in clustering. For better understanding, we list the important notations used in this paper as in TABLE I. The detailed optimization process is shown in Algorithm 1.

A. Viewpoint Prediction

Due to the extreme viewpoint changes in vehicles, there are relatively small inter-class differences between different vehicles. We argue that global comparison in previous unsupervised clustering methods [15], [17], [18] tends to group the different vehicles with the same viewpoint into the same cluster. Therefore, this global comparison scheme cannot guarantee the promising performance for target-only unsupervised vehicle Re-ID without any label supervision in network training. To handle this problem, we propose to introduce a viewpoint prediction model to identify the vehicle's viewpoint information during the forthcoming clustering.

In specific, we use a viewpoint prediction network to predict the viewpoint of each unlabeled vehicle image x_i in training set $\{X \mid x_1, x_2, \ldots, x_N\}$. We train our viewpoint prediction model on VeRi-776 [1], which contains all the visible viewpoints of the vehicle. Following the viewpoints annotation in previous work [22], we divide vehicle images into five viewpoints, e.g., $front, front\_side, side, rear\_side, rear$. Furthermore, we have additionally labeled 3000 samples in VeRi-Wild [33] data to fine-tune the model to improve the robustness of the viewpoint prediction. We use the commonly used cross-entropy loss L_η to optimize the viewpoint classifier $W(x_i \mid \theta)$,

$$L_\eta = -\Sigma_i^N y_v log\left(W\left(x_i \mid \theta\right)\right), \tag{1}$$

where y_v is a one-hot vector of the ground truth of corresponding viewpoint labels.

B. Recognition Stage

After the viewpoint prediction, we can obtain viewpoint-aware unlabeled training set $X^v = \{x_1^v, x_2^v, \ldots, x_N^v\}$, and the current training set can be regarded as the clusters divided according to the viewpoint. For example, VeRi-776 [1] falls into five different viewpoint clusters. For each image in X^v, we assign a unique index-label $y_{ind} = \{1, 2, 3, \ldots, N\}$ to indicate the category of each sample. In order to learn the discriminative feature, one can achieve this objective by directly using triplet loss [27], [50] or cross-entropy loss via classification. However, the learning driven by these losses, which mainly calculate the similarity among each batch, will become inefficient and difficult to converge with the dataset's scale growth. Herein, we employ the more efficient repelled

loss [17], [23], [51], which calculates the feature similarity between the current sample and all the training samples at once.

It is equipped with a key-value structure to store the features of all training samples, and the index-label y_{ind} is stored in the key memory. The y_{ind} will not change during the entire training process. We calculate the feature similarity between the i-th image in the v-th viewpoint f_i^v and all the samples,

$$p(y_p|x_i^v) = \frac{\exp(\left(M[i]^T f_i^v/\beta\right)}{\sum_{j=1}^{N}\exp(\left(M[j]^T f_i^v/\beta\right)}, \tag{2}$$

where $M[i]$ denotes the i-th slot of the value memory M. β is a hyper-parameter to control the softness of the probability distribution over classes, which is set to 0.1 followed by [17]. N indicates the number of clusters. y_p is the pseudo label, and we initialize $y_p = y_{ind}$. We maximize the distance between samples by assigning each sample to its own slot,

$$L_\alpha = -\log(p(y_p|x_i^v)). \tag{3}$$

During the back propagation, the feature memory is updated by the formula $M[y_i] \leftarrow \frac{1}{2}(M[y_i] + f_i^v)$. At the recognition stage, $M[y_i]$ stores the features of each training sample. At the subsequent progressive clustering stage, the pseudo label y_p of each sample will be redistributed according to the clustering results, while each slot stores the features of each cluster.

C. Progressive Clustering

Without any identification information, we propose a progressive clustering algorithm for unsupervised vehicle Re-ID. It mainly contains three aspects, two-period algorithm to avoid the similarity dilemma caused by the extreme viewpoint changes of vehicles, the k-reciprocal encoding to re-metric the distance for more robust clustering, and clustering with noisy sample selection to deal with outliers that are difficult to be clustered in real scenes.

The First Period: Through the recognition stage, the model learned more recognizable identity features of each image. The features obtained from the training set $F^* = \{F_1, F_2, F_v, \ldots, F_V\}$,

$$F_v = \{f_1^v, f_2^v, \ldots, f_{N_v}^v\}, \tag{4}$$

where F_v and N_v represent the feature set and the number of samples in the v-th viewpoint. We compare the similarity of all features F_v belonging to the same viewpoint cluster to obtain the distance matrix $D(F_v, F_v)$, $v = 1, 2, \ldots, V$. D represents the scoring matrix of Euclidean distance $d_{ij} = \|f_i - f_j\|^2$. There is no doubt that the same vehicle with the same viewpoint has the highest similarity and thus tends to be clustered together (assigned to the same pseudo label) with the highest priority. For the distance matrix under each viewpoint, we obtain pseudo labels by the prevalent cluster algorithm DBSCAN [26], which can effectively deal with noise points and achieve spatial clusters of arbitrary shapes without information of the number of clusters compared to the conventional k-means [52] clustering.

The Second Period: In the second period, we compare the distance between different viewpoint clusters. We take the shortest distance between features in two clusters as a measure of the distance between clusters. Considering that we have no idea whether the current sample has positive samples (with the same identity) in other viewpoints, we comprehensively compare the distance between all different viewpoint clusters,

$$D^* = \{D(F_1, F_2), \ldots, D(F_m, F_n)\}, \quad m \neq n. \tag{5}$$

We argue that the higher similarity, the more likely the same identity. Thus adopt a progressive strategy to merge the clusters between different viewpoints gradually. Therefore, We first calculate a rank list R,

$$R = argsort(D^*), \tag{6}$$

where R finds the most similar clusters among all different viewpoints. We set a strict distance threshold τ, and merge clusters from different viewpoints only when the distance of the candidates in R^* is less than τ, i.e,

$$R^* = R[1 : C(d = \tau)], \tag{7}$$

where $C = \{c_i, c_j\}$ is the last sample pair in different clusters with distance less than τ. Intuitively, due to the style diversity of different datasets, we expect the setting of τ to be irrelevant to datasets. In our method, after the recognition stage, we ascending sort the calculated D^*, and set the distance value between the ti-th lowest sample pair as the threshold τ. The distance threshold is only calculated after the recognition stage and then fixed in the whole training process. We alternately execute the above two periods during each iteration. The model learns the features of vehicles from the same viewpoint while continuously mining the features of vehicles with the same identity from different viewpoints.

Distance Metric by k-Reciprocal Encoding: Clearly, more positive samples in the same-viewpoint cluster in the first period, higher clustering quality at different viewpoints in the second period, which in turn will benefit the performance in the next iteration. Note that the clustering method significantly relies on the distance metric. We propose introducing the widely used k-reciprocal encoding [15], [20], [25] as the distance metric for feature comparison. For the sample x_i^v in X^v, we record its k nearest neighbors with index-labels $K_k(x_i^v)$, for all indexes $ind \in K_k(x_i^v)$, if $\left|K_k(x_i^v) \cap K_{\frac{k}{2}}(x_{ind}^v)\right| \geqslant \frac{2}{3}\left|K_{\frac{k}{2}}(x_{ind}^v)\right|$, x_i^v's mutual k nearest neighbors set $G_i \leftarrow \left|K_k(x_i^v) \cup K_{\frac{k}{2}}(x_{ind}^v)\right|$. In this case, all reliable samples similar to x_i^v are recorded in G_i. Then distance d_{ij} of the sample pair in the same viewpoint distance matrix, $D(F_m, F_n)$, $m = n$ reassigns weight by,

$$\tilde{d}_{ij} = \begin{cases} e^{-d_{ij}} & if \ j \in G_i, \\ 0 & else. \end{cases} \tag{8}$$

For each image pairs (x_i^v, x_j^v) at the same viewpoint, we get a new distance matrix $D_J(F_m, F_n)$, $m = n$ for clustering, it can be calculated by,

$$d_J(x_i^v, x_j^v) = 1 - \frac{\sum_{l=1}^{N_v} min(\tilde{d}_{il}, \tilde{d}_{jl}))}{\sum_{l=1}^{N_v} max(\tilde{d}_{il}, \tilde{d}_{jl}))}, \tag{9}$$

where N_v is the total number of samples in viewpoint v.

Algorithm 1 The Viewpoint-Aware Progressive Clustering Method (VAPC)

Require: Unlabeled training set $X = \{x_1, x_2, x_3, \ldots, x_N\}$; Recognition stage epoch E_r; Set the distance of the most similar ti-th sample pair as the distance threshold; CNN model $\tilde{M}$; index-label $y_{ind} = 1, 2, 3, \ldots, N$.

Viewpoint prediction: $X \rightarrow X^v$, $V = 5$.
Recognition stage:
for $i < E_r$ **do**
  Train CNN model $\tilde{M}$ with X and y_{ind} according to Eq. (3).
end for
Calculate threshold τ.

Ensure: Best CNN model $\tilde{M}$

Progressive clustering stage:
First period:
for $i < V$ **do**
  Calculate distance matrix: $D(F_i, F_i)$.
  Re-metric distance by Eq. (9) to obtain $D_J(F_i, F_i)$.
  Use DBSCAN to obtain clustering results.
end for
Mine noise samples according to Eq. (11).
Second period:
Compare feature sets at the different viewpoint to obtain distance matrix $D(F_m, F_n), m \neq n$.
Select the clusters need merged from different viewpoints through Eq. (6) and Eq. (7).
Retrain CNN model $\tilde{M}$ with X and y_p according to Eq. (3).
Evaluate on the test set $\rightarrow$ performance P.
if $P > P^*$ **then**
  $P^* = P$.
  Save the best model $\tilde{M}$.
end if

Clustering With Noisy Sample Selection: Our viewpoint-aware clustering strategy avoids comparing different viewpoints of vehicles during the first period of clustering, which alleviates the intra-class gap and reduces the difficulty of clustering to a great extent. However, due to the complexity of the real scene, some hard samples are still difficult to cluster and then regarded as noises. The reason is, although DBSCAN [26] can generate clusters for data of any spatial shape, it uses two parameters eps and $minPts$ to define the density conditions that need to be meet when forming clusters in the training set, which tends to cluster the samples with small intra-class gaps and treat the samples with larger intra-class gaps as noises, as shown in Fig. 4. We observe that these noises usually derive from two situations which are shown as P_1 and P_2 in Fig. 4. In P_1, due to occlusion, misalignment of the bounding box, or deviation of the viewpoint prediction, samples with the same identity but far from the already formed clusters (the blue cluster as shown in Fig. 4) will be regarded as noise. In P_2, some samples deriving from the same identity fail to form into the same cluster since they cannot meet the density condition due to large intra-class differences.

The basic idea is to mine positive samples from noise samples. To achieve this objective, we first use set S_n to collect

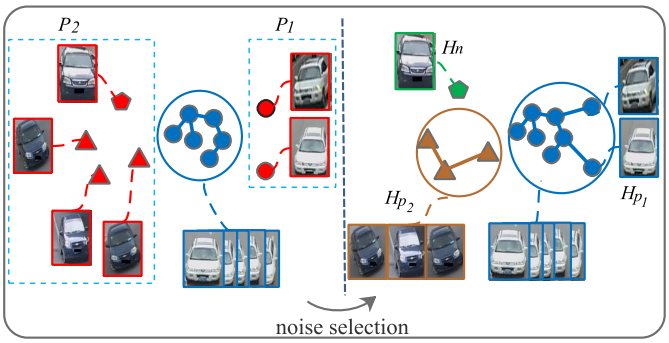


Fig. 4. Illustration of noise selection. Samples in the same color belong to the same cluster except the red color is for noisy samples (pseudo label is -1). P_1, P_2 represent two different noise situations. After noise selection, we reconstruct the cluster for the noise samples by comparing each noise and other clusters. H_{p1} and H_{p2} represent different sets of hard positive samples, and H_n represent set of hard negative samples.

all the indexes of noise samples. For each member s_i in S_n, we find $\tilde{k}$ nearest neighbors of s_i in same viewpoint feature set F_v and define the indexes of them as $top\_\tilde{k}(s_i)$. The nearest one in $top\_\tilde{k}(s_i)$ is $top\_1(s_i)$. In this way, we can establish a correlation for each noise sample and recorded it in the correlation set T,

$$T = \{(s_1, top\_1(s_1)), (s_2, top\_1(s_2)), \ldots, (s_n, top\_1(s_n))\}, \tag{10}$$

where the similarities ranked in descending order. Then we judge which situation the noise belongs to based on $top\_1(s_i)$. If $top\_1(s_i)$ belongs to a cluster, it corresponds to the first situation P_1, recorded in set H_{p_1}. Otherwise, $top\_1(s_i)$ is a noise sample, it belongs to the second situation P_2, recorded in set H_{p_2}. For P_1 caused by outlier noise samples (best viewed in Fig. 4), we expect noise samples to be classified into clusters with the same identity. For P_2 caused by a large intra-class gap, we expect that noise samples with the same identity to be clustered together to form a new cluster.

Directly merge noise samples is not reliably caused by hard negative samples. We take a more reliable approach as follows. Inspired by k-reciprocal encoding, if $(s_i, top\_1(s_i))$ belong to the same identity, their neighbor image sets should be similar, which also means that they should be located in each other's $\tilde{k}$-nearest neighbors. Therefore, we calculate the $\tilde{k}$-nearest neighbor image sets of the $top\_1(s_i)$. If s_i appears in $top\_\tilde{k}(top\_1(s_i))$, s_i is regarded as a reliable hard positive sample and will be merged with $top\_1(s_i)$. Otherwise, the noisy sample s_i will be treated as a hard negative sample and recorded in set H_n, which is divided into new clusters to learn its discriminative feature further. Formally, we construct:

$$\begin{aligned} H_{p_1} &= \left\{(s_i, j) \mid j = top\_1[s_i], j \notin S_n, s_i \in top\_\tilde{k}[j]\right\}, \\ H_{p_2} &= \left\{(s_i, j) \mid j = top\_1[s_i], j \in S_n, s_i \in top\_\tilde{k}[j]\right\}, \\ H_n &= \left\{(s_i) \mid j = top\_1[s_i], s_i \notin top\_\tilde{k}[j]\right\}. \end{aligned} \tag{11}$$

For H_{p_1}, we assign s_i the same pseudo-label as the j. For H_{p_2}, we first create a new cluster C_{si} containing

s_i and j, which also means that s_i and j are assigned the same new unique pseudo-label. Then, we search for noise samples belonging to C_{si}, for index $ind \in C_{si}$, $C_{si} \leftarrow \left| top\_\tilde{k}[ind] \cap S_n \right|$. For H_n, each item is treated as a separate cluster by assigning a new unique pseudo-label. Note that we process noises in order of similarity in the correlation set T, and when one noise is merged, it will not be merged with other clusters.

IV. Experiments

We evaluate our proposed method VAPC on two benchmark datasets VeRi-776 [1] and VeRi-Wild [33], which contain 5 and 4 view-points respectively. We compare our method with the prevalent domain adaption based unsupervised, and target-only methods without domain adaption for evaluation.

A. Datasets and Evaluation Protocol

VeRi-776 [1] is a comprehensive vehicle re-identification dataset providing rich attributes information such as color, type and temporal path. It contains 776 different vehicles captured in 20 cameras, yielding more than 49,357 images and 9,000 tracks. The training and testing sets contain 37,728 images of 576 vehicles and 11579 images of 200 vehicles. Both training and testing sets contain 5 common visible viewpoint situations, including $front, front\_side, side, rear\_side, rear$. Following the protocol in [1], we only return the matchings from the different cameras for the query vehicles as the results. We use the mean average precision (mAP) and cumulative matching characteristic (CMC) at Rank-1, Rank-5 and Rank-20 as the measurement metrics.

VeRi-Wild [33] is a large-scale vehicle Re-ID dataset, containing more than 400 thousand images of 40 thousand vehicle IDs captured by 174 cameras in the surveillance system. It contains complex backgrounds, various viewpoints, and illumination variations in real-world scenes. The training set contains 277,797 images of 30,671 vehicles. After the viewpoint prediction of the training set, VeRi-Wild contains 4 viewpoints, $front, front\_side, rear\_side, rear$, respectively containing 110204, 52716, 64968, 49909 images. Due to hardware limitations, we use all the training data in the recognition stage, and each viewpoint in the clustering stage takes 10,000 images, respectively. While the testing set consists of three subsets, test-3000, test-5000 and test-10000, with different testing sizes. Following the protocol in [33], the match rate protocol on VeRi-Wild is that all the references of the given query are in the gallery. We use mAP, Rank-1 and Rank-5 as the evaluation metrics.

VehicleID [53] is another large-scale vehicle Re-ID dataset, it includes 110,178 real scene images of 13134 types of vehicles as a training set. 111,585 images of 13,113 vehicles were used as a test set. In this article, to compare the results of other existing unsupervised domain adaptation methods, we also use the VehicleID dataset as the source domain for supervised training.

B. Implementation Details

We use ResNet50 [55] as the backbone by eliminating the last classification layer. All experiments are implemented on two NVIDIA TITAN Xp GPUs. We initialize our model with pre-trained weights on ImageNet [56]. For the viewpoint prediction network, we set the batch size as 32 and the learning rate as 0.001, with a maximum of 20 epochs. If not specified, we use stochastic gradient descent with a momentum of 0.9 and the dropout rate as 0.5 to optimize the model. For the Re-ID feature extraction network, we resize the input images of VeRi-776 [1] and VeRi-Wild [33] as (384,384). The batch size is set to 16. The learning rates at the recognition stage is set to 0.1 and divided by 10 after every 15 epochs, and set to 0.001 in the clustering stage. We only use a random horizontal flip as a data augmentation strategy. Following the protocol in [25] we set k to 20.

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

We compare our method with the state-of-the-art unsupervised Re-ID methods on VeRi-776 [1] and VeRi-Wild [33] in both target-only and domain adaption scenarios, as shown in TABLE II.

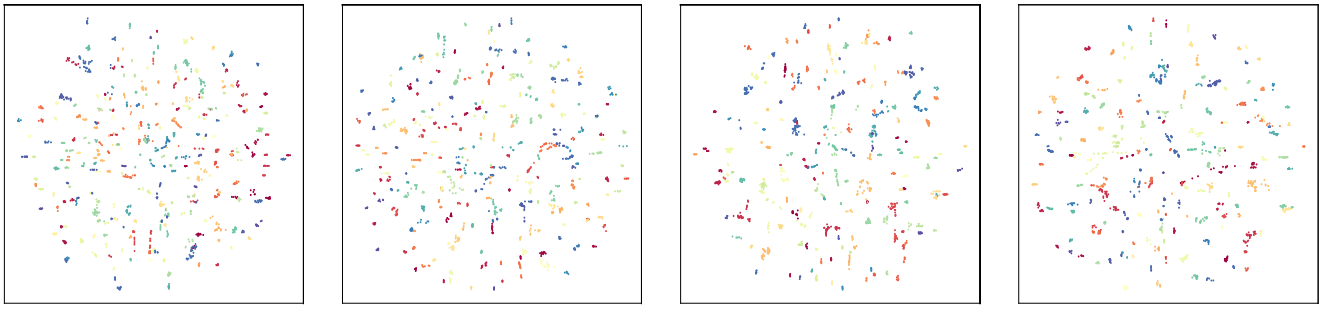
Compared With the Target-Only Method: We first compare our method with three state-of-the-art target-only unsupervised methods OIM [23], Bottom [17] and AE [54]. Generally speaking, our method (VAPC_TO) outperforms the three state-of-the-art target-only methods by a large margin by exploring the intra-class relationship. OIM [23] devotes to extracting discriminative features efficiently, which ignores the intra-class relationship, thus results in stumbling performance. Bottom [17] designs a bottom-up clustering strategy by merging the fixed clusters during each step. However, each clustering may produce the wrong classification, and more clustering steps, more clustering errors. Especially on the VeRi-776 [1], almost all visible viewpoints are included, which brings greater clustering challenges. Each clustering step only focuses on the same viewpoint and cannot bring more samples of different viewpoints together. Our method effectively alleviates this problem and brings greater improvement. AE [54] clusters the samples via a similarity threshold and constrains the cluster size by embedded a balance term into the loss. However, due to the similarity dilemma of vehicles, where the same viewpoints of different identities may have higher similarities, it is difficult to set an optimal similarity threshold for clustering. In addition, more and more samples meeting the similarity threshold are treated as the same identity during the training, especially on larger scaled dataset VeRi-Wild [33], it will cause more severe data imbalance in each cluster and damage the feature representation. Therefore the performance of AE [54] on VeRi-Wild [33] declines comparing with Bottom [17].

We further use t-SNE [24] to visualize the feature space distribution of our method compared to the three state-of-the-art target-only methods, as shown in Fig. 5. Compared with ours, the distribution between the points is sparser in OIM [23] and Bottom [17], while more points of different colors gathering in AE [54]. our method presents a better feature distribution, which demonstrates that VAPC_TO can successfully cluster more images of vehicles with the same identity and effectively improve the feature representation for unsupervised vehicle Re-ID.

TABLE II

COMPARISON WITH THE STATE-OF-THE-ART OF TARGET-ONLY RE-ID AND DOMAIN ADAPTIVE RE-ID METHODS ON VERI-776 [1] AND VERI-WILD [33]. "src" DENOTES THE SOURCE DOMAIN/DATASET, WHERE "N/A" AND "VEHICLEID" INDICATE THE TARGET-ONLY AND DOMAIN ADAPTIVE METHODS ON VEHICLEID DATASET [53] RESPECTIVELY. "VAPC_TO", "VAPC_DT" AND "VAPC_DA" INDICATE OUR VAPC IN TARGET-ONLY, DIRECT TRANSFER AND DOMAIN ADAPTION RESPECTIVELY

method	VeRi-776					VeRi-Wild								
	src	R1	R5	R20	mAP	test-3000			test-5000			test-10000		
						R1	R5	mAP	R1	R5	mAP	R1	R5	mAP
OIM [23]	N/A	45.1	62.2	78.1	12.2	48.7	66.6	14.4	45.0	60.9	12.6	38.8	54.4	10.0
Bottom [17]	N/A	63.7	73.4	83.4	23.5	70.5	86.0	30.7	64.2	82.2	27.1	55.2	75.1	21.6
AE [54]	N/A	73.4	**82.5**	**89.7**	26.2	68.5	87.0	29.9	61.8	81.5	26.2	53.1	73.7	20.9
VAPC_TO (ours)	N/A	**76.2**	81.2	85.3	**30.4**	**72.1**	**87.7**	**33.0**	**64.3**	**83.0**	**28.1**	**55.9**	**75.9**	**22.6**
SPGAN [44]	VehicleID	57.4	70.0	-	16.4	59.1	76.2	24.1	55.0	74.5	21.6	47.4	66.1	17.5
ECN [51]	VehicleID	60.8	70.9	85.4	27.7	73.4	88.8	34.7	68.6	84.6	30.6	61.0	78.2	24.7
UDAP [15]	VehicleID	76.9	**85.8**	-	35.8	68.4	85.3	30.0	62.5	81.8	26.2	53.7	73.9	20.8
VAPC_DT (ours)	VehicleID	69.1	79.0	88.2	35.5	74.0	88.6	37.7	68.1	84.8	33.1	60.2	78.7	26.3
VAPC_DA (ours)	VehicleID	**77.4**	84.6	**91.6**	**40.3**	**75.3**	**89.0**	**39.7**	**69.0**	**85.5**	**34.5**	**61.0**	**79.7**	**27.4**

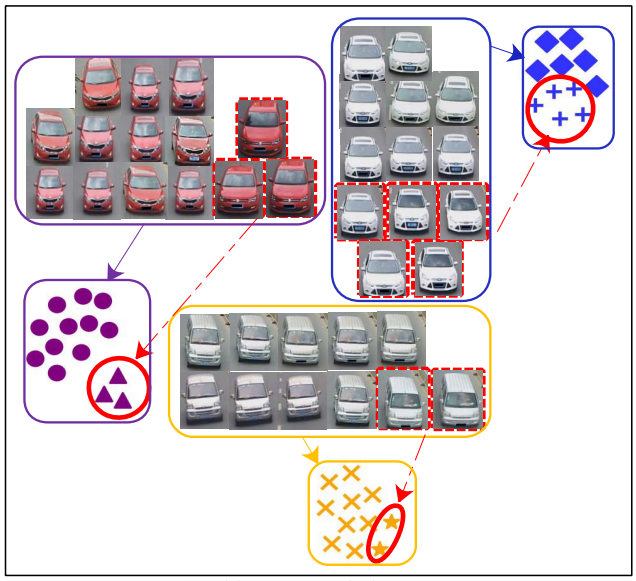


(a) OIM (b) Bottom (c) AE (d) VAPC_TO (ours)

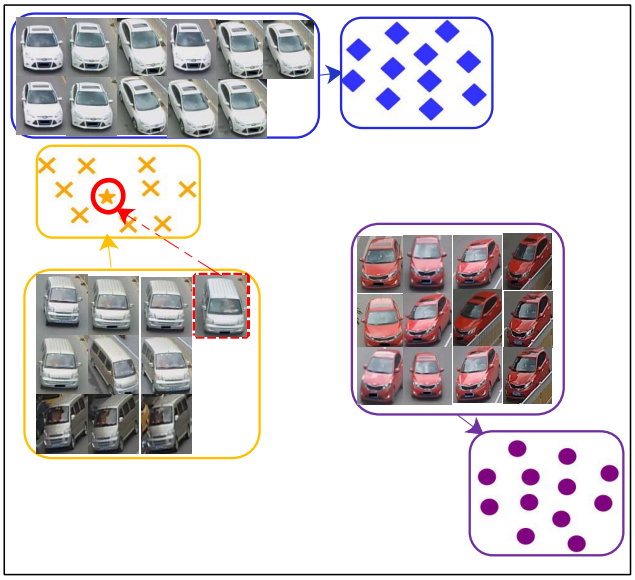
Fig. 5. Visualization for features extracted by target-only method, OIM [23], Bottom [17], AE [54] and our method. 37 identities with 2000 images in the gallery of VeRi-776 [1] are used. Each point represents an image, and each color represents a vehicle identity.

Compared With Unsupervised Domain Adaptation: To evidence the effectiveness of our method on unsupervised vehicle Re-ID, we further evaluate our method in the domain adaption fashion. Following the protocol in [15], we use VehicleID [53] as the source domain and employ repelled loss [17] for supervised training, replacing the recognition stage in III-B. We compare our method in the domain adaption fashion (VAPC_DA) with three state-of-the-art unsupervised domain adaptation methods, including SPGAN [44], ECN [51] and UDAP [15], as shown in the lower half part in TABLE II.

SPGAN [44] considers the style change among different datasets and trains a style conversion model to bridge the style discrepancy between the source domain data and the target domain. However, due to the huge gap between the vehicle datasets in the real scene, e.g., the diverse viewpoints, resolution and illumination, it is challenging to obtain the desired translated image, which is crucial in SPGAN [44], and thus results in poor performance for vehicle Re-ID. ECN [51] joins the source domain for model constraints while using the k-nearest neighbor algorithm to mine the same identity in the target domain. The setting of the k value not only has a greater impact on the experimental results, but the most similar top k samples are always at the same viewpoint. UDAP [15] uses source domain data to initialize the model and theoretically analyzes the rules that the model needs to follow when adapting to the target domain from the source domain. It achieves satisfactory results on vehicle Re-ID due to the strengthening of the constraints on the target domain training. The target domain feature extractor has stronger learnability while obtaining the source domain knowledge. However, it relies on global comparison, which may cause more clustering errors, especially on VeRi-Wild [33] dataset presents much smaller inter-class differences than VeRi-776 [1]. In addition, we evaluate our method in the "Direct Transfer" fashion by training on the source domain and directly testing on the target domain indicated as (VAPC_DT) in TABLE II. First of all, by leveraging the information in the training data, VAPC_DT generally outperforms VAPC_TO, which verifies the knowledge of the source domain during the training improves the vehicle retrieval ability of the model. The only exception is the rank-1 score on VeRi-776 [1]. The main reason is the huge gap between VehicleID [53] and VeRi-776 [1] datasets, e.g., VeRi-776 has lower resolution and more viewpoints, which results in poor generalization performance. Even though VAPC_DT significantly boosts the mAP score on both VeRi-776 [1] comparing to the target-only fashion (VAPC_TO). Second, VAPC_DT is even significantly superior to the domain adaption methods SPGAN [44] and ECN [51], and comparable to UDAP [15] on mAP, which proves the robustness of our method for unsupervised vehicle Re-ID.



(a) without two-period strategy



(b) with two-period strategy

Fig. 6. t-SNE [24] distance distribution with and without a two-period strategy on VeRi-776 [1]. The different color boxes indicate different clusters while the samples with the same shape representing the same identity. The red circle marks the false clustered samples.

Note that our method in target-only fashion (VAPC_TO) even surpasses most unsupervised domain adaptation methods such as SPGAN [44] and ECN [51], and works comparably to UDAP [15]. This further verifies the promising performance of our method while handling unsupervised vehicle Re-ID especially without prior annotation information or source data.

D. Ablation Study

In this section, we will thoroughly analyze the effectiveness of three critical components in the VAPC framework, including the two-period (tP) clustering strategy based on viewpoint prediction, *k*-reciprocal encoding (kR) and noise selection (NS), as reported in TABLE III.

Quantitative Study: One of the key contributions of our progressive clustering is the two-period clustering on both the same and different viewpoints for vehicle Re-ID. As shown in TABLE III (b), without dividing the viewpoints and removing the two-period (tP) strategy, we cluster all training samples directly after the recognition stage, both mAP and rank scores significantly drop, -7.5% in Rank-1 and -5.4% in mAP on VeRi-776 [1], while -3.6% and -2.7% on VeRi-Wild [33] test-3000. Which verifies the effectiveness of the progressive clustering for unsupervised vehicle Re-ID. Similar phenomenons happen to the *k*-reciprocal encoding (kR) and the noise selection (NS), as shown in TABLE III (c) and TABLE III (d). By removing the corresponding components, both mAP and rank scores significantly decline, which evidences the role of each component. Without any of the three components, the baseline (as shown in TABLE III (e)) results in stumble performance on both datasets due to the inability to cope with the various challenges brought about by the extreme viewpoint changes of vehicles. By integrating all the three components, our method, as shown in TABLE III (a) achieves promising results for unsupervised Re-ID.

TABLE III

RESULTS EVALUATED ON THE VeRi-776 [1] AND TEST-3000 SET OF VeRi-WILD [33]. tP REPRESENTS OUR FIRST PERIOD AND SECOND PERIOD CLUSTERING STRATEGY, kR MEANS DISTANCE METRIC BY *k*-RECIPROCAL ENCODING, NS MEANS NOISE SELECTION

method	VeRi-776			VeRi-Wild		
	R1	R5	mAP	R1	R5	mAP
(a) Ours	76.2	81.2	30.4	72.1	87.7	33.0
(b) w/o tP	68.7	73.2	25.0	68.5	85.0	30.3
(c) w/o kR	71.0	78.9	24.1	69.2	86.0	29.7
(d) w/o NS	71.3	78.8	27.8	70.1	87.0	32.5
(e) w/o tP + kR + NS	61.4	72.5	18.2	48.7	66.6	14.4

Qualitative Study: To further understand the contribution of the three components, we visualize the results of different variants as discussed in Table III in terms of sample distribution or ranking list, as shown in Fig. 6 to Fig. 8. From Fig. 6 (a), we can see that more hard negative samples (different identities with highly similar appearance) with the same viewpoints tend to cluster without a two-period clustering strategy. Our method successfully gathers vehicle images with diverse viewpoints, even with large appearance differences due to the viewpoint and illumination changes. This further evidence the effectiveness of the proposed two-period clustering strategy, which can distinguish small gaps between different identities in the same viewpoint and mine the same identity samples with large gaps between different viewpoints. The role of *k*-reciprocal encoding is to mine samples sharing the most similar features despite appearance differences. As shown in Fig. 8 (a), the result without *k*-reciprocal encoding tends to split the same identity with difference appearance caused by viewpoint and illumination changes into individual clusters, while it can merge them into one single cluster after introducing the *k*-reciprocal encoding, as shown in Fig. 8 (b). Fig. 7 demonstrates the qualitative comparison of ranking results of three queries with or without noise selection. Clearly, after introducing the noise selection scheme, our method can

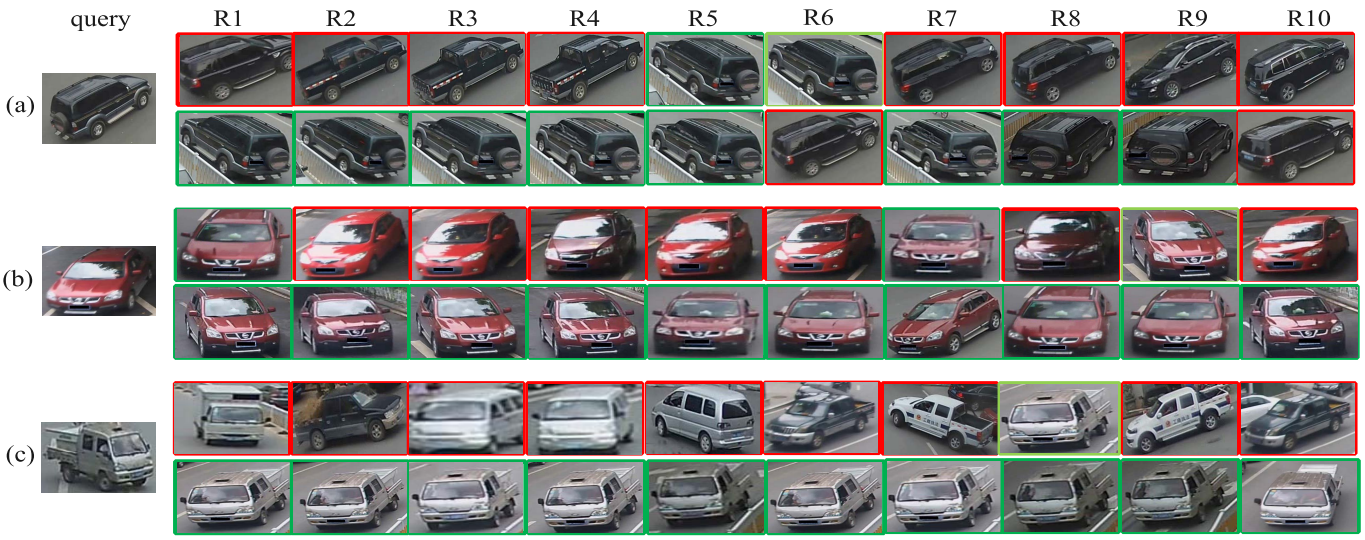


Fig. 7. Examples of ranking results with and without noise selection on VeRi-776 [1]. For each query, the top and the bottom rows show the ranking result without and with noise selection, respectively. The green and red boxes indicate the right and the wrong matchings, respectively.

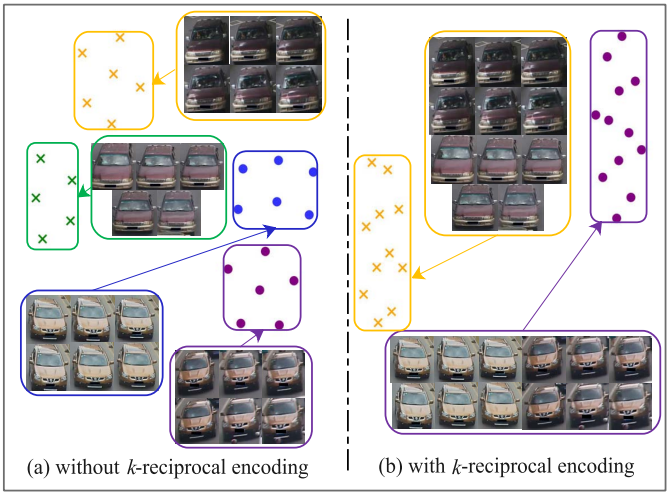


Fig. 8. t-SNE [24] distance distribution with and without distance metric by k-reciprocal encoding on VeRi-776 [1]. The different color boxes indicate different clusters while the samples with the same shape representing the same identity.

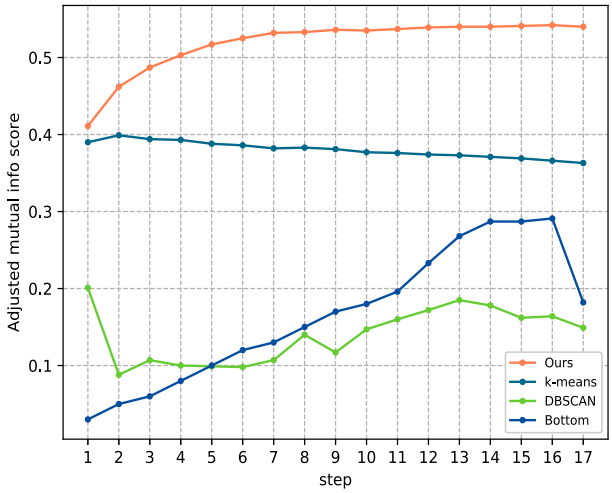


Fig. 9. The performance of clustering quality (AMI) on VeRi-776 [1]. Each step represents an iteration of progressive clustering and retraining the model.

hit more correct matchings in the earlier rankings and can remove the false matchings with a similar appearance as the queries.

E. Analysis of Clustering Quality

Clustering quality is a crucial factor in clustering-based methods for vehicle Re-ID. Therefore, we measure the clustering quality via Adjusted Mutual Information (AMI) [57] on our method compared to the state-of-the-art methods. AMI measures the distribution of ground truth and pseudo labels generated by clustering through mutual information. A larger AMI means the closer distribution of the ground truth and pseudo labels, which in turn means better clustering quality. We compare our method with Bottom [17], k-means [52] and DBSCAN [26] which also allocate pseudo labels during clustering.

As illustrated in Fig. 9, the classic clustering algorithm k-means [52] and DBSCAN [26] work stumblingly in the global comparison fashion. Furthermore, k-means [52] specifies the number of clusters, which makes the change of samples in the cluster relatively stable. However, due to global comparison, a large number of samples with the same viewpoint and different identities appear in the same cluster, which makes model training continue to decline. DBSCAN [26] is sensitive to noise; therefore, a large number of noise samples under various challenges in real scenes deteriorates the clustering quality. Bottom [17] causes the final collapse due to the accumulation of the number of clustering errors each step. Since clustering based on viewpoint division greatly simplifies the clustering task, and the strategy progressively merges different viewpoints and gradually gathers vehicles of the same identity from different viewpoints, our method continues to improve with training.

F. Investigation of Viewpoint Prediction

Viewpoint prediction is a prerequisite component in our framework, as discussed in III-A. To investigate the influence

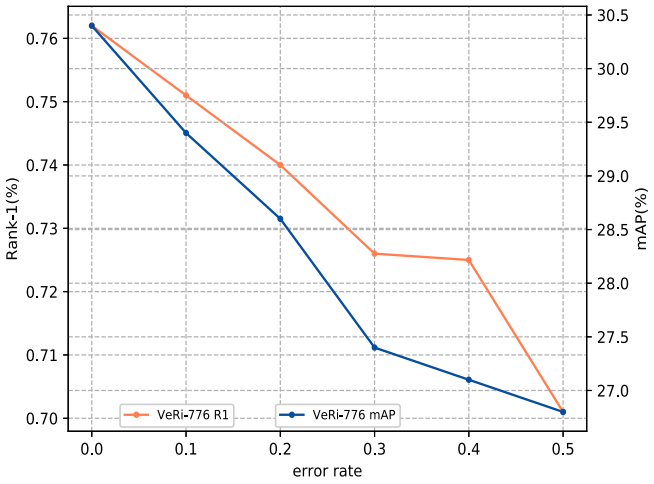


Fig. 10. The performance along with different error rate viewpoint predictors on VeRi-776 [1].

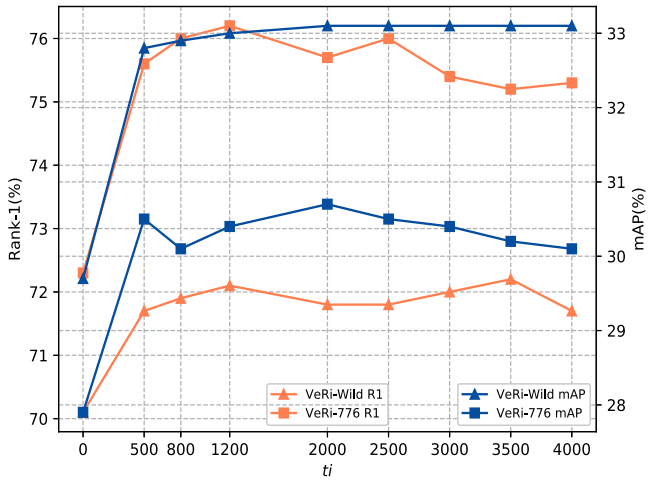
of viewpoint prediction in our method, we have trained a series of classifiers with different accuracy rates for viewpoint prediction. The experimental results are shown in Fig. 10. As expected, the Re-ID accuracy of VAPC_TO decreases as the accuracy of the viewpoint classification classifier decreases. When the accuracy of the viewpoint classifier drops to 0.5, the accuracy of Rank-1 drops from 76.2% to 70% (-6.2%) on VeRi-776 [1]. Even though our method with only 0.5 viewpoint classifier accuracy still outperforms the most unsupervised algorithms, as shown in TABLE II. We can see that a robust viewpoint classifier can significantly improve the performance of our algorithm. And due to our more reasonable clustering strategy and effective noise processing, we can also perform well on a poor viewpoint classifier.

G. Parameter Analysis

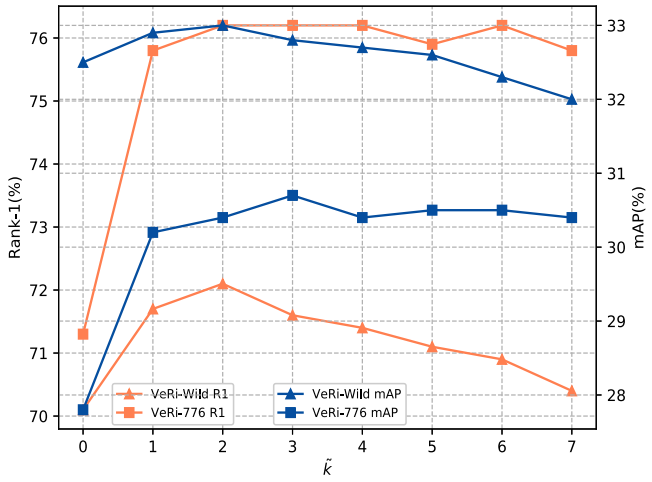
There are two essential parameters in our methods, ti denoting the distance of the ti-th sample pair as the threshold for combining different viewpoint clusters as explained in III-C Eq. (7), and $\tilde{k}$ in Eq. (11), indicating the judgment condition when selecting noise as explained in III-C. We shall evaluate the impact of these two parameters in this section.

The Impact of the Number ti: As shown in Fig. 11 (a), we vary ti from 0 to 4000 to calculate the distance threshold τ and test the model performance. $ti = 0$ means only the same viewpoint clustering. bigger ti, larger threshold τ. A large ti will harm the model performance. For example, when $ti > 3500$, a substantial performance drop can be observed. This is because the over large ti may cause too many clusters of different viewpoints to be merged at one time, which resulting in a large number of incorrect classifications. However, over small ti selects a few correct clusters, which also leads to poor performance. For the comprehensive performance of ti on VeRi-776 [1] and VeRi-Wild [33], we set ti to 1200.

The Impact of the Number $\tilde{k}$: Fig. 11 (b) reports the analysis on $\tilde{k}$ during the noise selection. As discussed in III-C, $\tilde{k}$ plays the role of limiting noise combined with clusters or other noises. The larger $\tilde{k}$, the weaker limitation. The larger $\tilde{k}$ declines the performance on VeRi-Wild [33], while remaining



(a) The parameter ti



(b) The parameter $\tilde{k}$

Fig. 11. Parameter and method analysis. (a) The impact of ti in progressive clustering. (b) The impact of the number of $\tilde{k}$ in noise selection.

stable on VeRi-776 [1]. The reason is VeRi-Wild [33] has a smaller inter-class difference compared to VeRi-776 [1]. When $\tilde{k}$ increases, the constraint of judging whether the two clusters are merged is weakened, which increases the error rate. Based on the results on Fig. 11 (b), we set $\tilde{k} = 2$ for the best balance.

H. Complexity Study

Complexity Analysis: The computation complexity of baseline is $O(nlogn)$, therefore the computation complexity of the first period is $O(V\bar{N}_v log(\bar{N}_v))$, where V, $\bar{N}_v$ represent the total number of viewpoints and average number of images per viewpoint, respectively. For k-reciprocal encoding, most of the computation costs focus on pairwise distance computing and ranking process is $O(V{\bar{N}_v}^2 + V{\bar{N}_v}^2 log(\bar{N}_v))$. For noisy sample selection, the computation complexity is $O(N)$. For the second period, the computational complexity of sorting all samples and traversal is $O(Nlog(N) + N^2)$.

Running Time Study: In our framework, each step mainly includes feature extraction, progressive clustering and 2 epochs of model training. We evaluate the relationship between

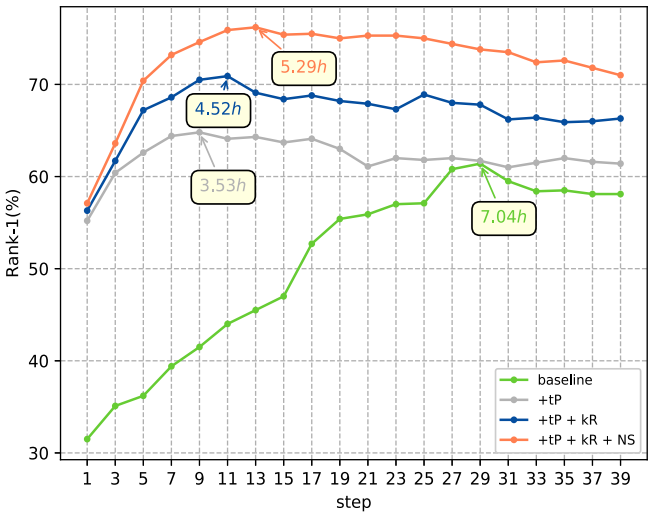


Fig. 12. Evaluation of accuracy and training steps on VeRi-776 [1]. The color box indicates the time required to obtain the best model. Each step represents an iteration of progressive clustering and retraining the model.

TABLE IV

AVERAGE RUNNING TIME OF EACH PART IN ONE STEP ON VeRi-776 [1]. tP REPRESENTS OUR FIRST PERIOD AND SECOND PERIOD CLUSTERING STRATEGY, kR MEANS DISTANCE METRIC BY k-RECIPROCAL ENCODING, AND NS MEANS NOISE SELECTION

feature extraction	progressive clustering			model training
	tP	kR	NS	
115s	445s	172s	40s	791s

accuracy change and running time in steps, as shown in Fig. 12. Although our method introduces more computation on progressive clustering comparing to the baseline, our method takes fewer steps (about 5.29 hours to step 13) than the baseline (about 7.04 hours to step 29) to achieve the best results. The main reason is that the two-period (tP) clustering strategy and k-reciprocal encoding (kR) can mine more samples with the same identity from different viewpoints in the early stage of model training. And, the noise selection (NS) further can enhance the generalization ability of the model. In addition, we further evaluate the average running time of each part in one step as shown in the TABLE IV, which is tested on an 8-core Intel(R) Xeon(R)@2.10GHz CPU platform. From which we can see, although introducing progressive clustering brings more cost on the running time in a single step. It can help the model reach the best results at early steps (as shown in Fig. 12) with even less total running time. This further evidences the benefit of the proposed progressive clustering while handling the inter-instance similarity and intra-instance discrepancy caused by large viewpoint variations among vehicles.

V. CONCLUSION

In this paper, we propose a viewpoint-aware progressive clustering method to solve the unsupervised Re-ID problem of vehicles. We analyzed the similarity dilemma of vehicle comparison, and it is first time to explored the progressive clustering by dividing the training set into different subsets according to the viewpoint. In addition, we propose a noise selection strategy to solve the noise problem generated in the clustering process. Extensive experimental results demonstrate the effectiveness of the proposed methods in unsupervised Vehicle Re-ID.

Our method is based on the observation that images of vehicles from adjacent views normally share a large degree of common appearance thus the adjacent views can be merged based on similarity. Therefore, it is more suitable for the real-life multi-view scence. In the other words, it is still challenging to deal with the scenario with only two large discrepancy viewpoints, such as *front* and *rear*. In addition, due to the diverse similarity among the viewpoints, it is a bit crude to merge all the different viewpoints of the same ID by a fixed empirical similarity threshold. In the future, we will further explore a more effective method to deal with these more challenging situations.

REFERENCES

[1] 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.*, 2016, pp. 869–884.
[2] H. Guo, C. Zhao, Z. Liu, J. Wang, and H. Lu, "Learning coarse-to-fine structured feature embedding for vehicle re-identification," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 6853–6860.
[3] Z. Zheng, T. Ruan, Y. Wei, and Y. Yang, "VehicleNet: Learning robust feature representation for vehicle re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jan. 2019, pp. 1–4.
[4] B. He, J. Li, Y. Zhao, and Y. Tian, "Part-regularized near-duplicate vehicle re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 3997–4005.
[5] Y. Zhou and L. Shao, "Cross-view GAN based vehicle generation for re-identification," in *Proc. Brit. Mach. Vis. Conf.*, 2017, pp. 1–12.
[6] 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.
[7] X. Liu, S. Zhang, Q. Huang, and W. Gao, "RAM: A region-aware deep model for vehicle re-identification," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2018, pp. 1–6.
[8] C. Li, X. Liang, Y. Lu, N. Zhao, and J. Tang, "RGB-T object tracking: Benchmark and baseline," *Pattern Recognit.*, vol. 96, Dec. 2019, Art. no. 106977.
[9] A. Lu, C. Li, Y. Yan, J. Tang, and B. Luo, "RGBT tracking via multi-adapter network with hierarchical divergence loss," *IEEE Trans. Image Process.*, vol. 30, pp. 5613–5625, 2021.
[10] Z. Zhong, L. Zheng, S. Li, and Y. Yang, "Generalizing a person retrieval model hetero- and homogeneously," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 172–188.
[11] H. Fan, L. Zheng, C. Yan, and Y. Yang, "Unsupervised person re-identification: Clustering and fine-tuning," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 14, no. 4, pp. 1–18, 2018.
[12] P. Peng *et al.*, "Unsupervised cross-dataset transfer learning for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1306–1315.
[13] J. Wang, X. Zhu, S. Gong, and W. Li, "Transferable joint attribute-identity deep learning for unsupervised person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2275–2284.
[14] J. Peng, H. Wang, T. Zhao, and X. Fu, "Cross domain knowledge transfer for unsupervised vehicle re-identification," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops*, Jul. 2019, pp. 453–458.
[15] L. Song *et al.*, "Unsupervised domain adaptive re-identification: Theory and practice," *Pattern Recognit.*, vol. 102, Jun. 2020, Art. no. 107173.
[16] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, "Deep clustering for unsupervised learning of visual features," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 132–149.
[17] Y. Lin, X. Dong, L. Zheng, Y. Yan, and Y. Yang, "A bottom-up clustering approach to unsupervised person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 8738–8745.
[18] G. Ding, S. H. Khan, and Z. Tang, "Dispersion based clustering for unsupervised person re-identification," in *Proc. Brit. Mach. Vis. Conf.*, 2019, p. 264.

[19] X. Zhang, J. Cao, C. Shen, and M. You, "Self-training with progressive augmentation for unsupervised cross-domain person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2019, pp. 8222–8231.

[20] Y. Fu *et al.*, "Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6112–6121.

[21] F. Zhao, S. Liao, G.-S. Xie, J. Zhao, K. Zhang, and L. Shao, "Unsupervised domain adaptation with noise resistible mutual-training for person re-identification," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 1–18.

[22] A. Zheng, X. Lin, C. Li, R. He, and J. Tang, "Attributes guided feature learning for vehicle re-identification," 2019, *arXiv:1905.08997*. [Online]. Available: http://arxiv.org/abs/1905.08997

[23] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang, "Joint detection and identification feature learning for person search," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3415–3424.

[24] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

[25] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with k-reciprocal encoding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1318–1327.

[26] M. Ester, H. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. 2nd Int. Conf. Knowl. Discov. Data Mining*, 1996, pp. 226–231.

[27] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," 2017, *arXiv:1703.07737*. [Online]. Available: http://arxiv.org/abs/1703.07737

[28] D. Wang and S. Zhang, "Unsupervised person re-identification via multi-label classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10981–10990.

[29] H.-X. Yu, W.-S. Zheng, A. Wu, X. Guo, S. Gong, and J.-H. Lai, "Unsupervised person re-identification by soft multilabel learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 2148–2157.

[30] K. Sohn, "Improved deep metric learning with multi-class n-pair loss objective," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 1857–1865.

[31] H. Shi *et al.*, "Embedding deep metric for person re-identification: A study against large variations," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 732–748.

[32] H. O. Song, Y. Xiang, S. Jegelka, and S. Savarese, "Deep metric learning via lifted structured feature embedding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4004–4012.

[33] 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 Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 3235–3243.

[34] 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.

[35] M. Cormier, L. Sommer, and M. Teutsch, "Low resolution vehicle re-identification based on appearance features for wide area motion imagery," in *Proc. IEEE Winter Appl. Comput. Vis. Workshops*, Mar. 2016, pp. 1–7.

[36] 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.

[37] X. Liu, W. Liu, T. Mei, and H. Ma, "PROVID: Progressive and multimodal vehicle reidentification for large-scale urban surveillance," *IEEE Trans. Multimedia*, vol. 20, no. 3, pp. 645–658, Mar. 2018.

[38] Y. Shen, T. Xiao, H. Li, S. Yi, and X. Wang, "Learning deep neural networks for vehicle re-ID with visual-spatio-temporal path proposals," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1900–1909.

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

[40] W. Lin *et al.*, "Group reidentification with multigrained matching and integration," *IEEE Trans. Cybern.*, vol. 51, no. 3, pp. 1478–1492, Mar. 2021.

[41] W. Lin *et al.*, "Learning correspondence structures for person re-identification," *IEEE Trans. Image Process.*, vol. 26, no. 5, pp. 2438–2453, May 2017.

[42] 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. 8282–8291.

[43] J. Sochor, A. Herout, and J. Havel, "BoxCars: 3D boxes as CNN input for improved fine-grained vehicle recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 3006–3015.

[44] W. Deng, L. Zheng, Q. Ye, G. Kang, Y. Yang, and J. Jiao, "Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 994–1003.

[45] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer GAN to bridge domain gap for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 79–88.

[46] Y. Lin, L. Xie, Y. Wu, C. Yan, and Q. Tian, "Unsupervised person re-identification via softened similarity learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3390–3399.

[47] M. Ye, X. Lan, and P. C. Yuen, "Robust anchor embedding for unsupervised video person re-identification in the wild," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 170–186.

[48] X. Lan, W. Zhang, S. Zhang, D. K. Jain, and H. Zhou, "Robust multi-modality anchor graph-based label prediction for RGB-infrared tracking," *IEEE Trans. Ind. Informat.*, early access, Oct. 14, 2019, doi: 10.1109/TII.2019.2947293.

[49] R. M. S. Bashir, M. Shahzad, and M. M. Fraz, "VR-PROUD: Vehicle re-identification using progressive unsupervised deep architecture," *Pattern Recognit.*, vol. 90, pp. 52–65, Jun. 2019.

[50] 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.*, Jun. 2016, pp. 1335–1344.

[51] Z. Zhong, L. Zheng, Z. Luo, S. Li, and Y. Yang, "Invariance matters: Exemplar memory for domain adaptive person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 598–607.

[52] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, "An efficient K-means clustering algorithm: Analysis and implementation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 881–892, Jul. 2002.

[53] 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.

[54] Y. Ding, H. Fan, M. Xu, and Y. Yang, "Adaptive exploration for unsupervised person re-identification," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 16, no. 1, pp. 1–19, Apr. 2020.

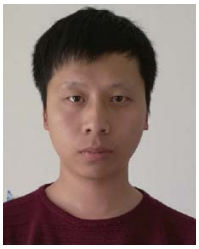
[55] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.

[56] 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.

[57] N. X. Vinh, J. Epps, and J. Bailey, "Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance," *J. Mach. Learn. Res.*, vol. 11, pp. 2837–2854, Jan. 2010.

Aihua Zheng received the B.Eng. degree in computer science and technology from Anhui University, China, in 2006. She is currently an Associate Professor and the Ph.D. Supervisor with the School of Artificial Intelligence, Anhui University. She visited the University of Stirling and Texas State University from June 2013 to September 2013 and from September 2019 to August 2020, respectively. As the first author or corresponding author, she has published more than 40 academic papers, including top conferences papers in AAAI and IJCAI, and authoritative journals in IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS: SYSTEMS (TSMCS), *PR*, *PRL*, *NeuCom*, *CogCom*, and *NCA*. Her main research interests include vision-based artificial intelligence and pattern recognition, especially on person/vehicle re-identification, audio visual computing, and multi-modal intelligence. She is a member of China Computer Federation (CCF) and China Society of Image and Graphics (CSIG). She has been awarded the Best Paper in SERA 2017 and the Best Student Paper in the Workshop in ICME 2019. She is serving as a Reviewer for representative conferences and journals, including AAAI, IJCAI, IEEE TRANSACTIONS ON IMAGE PROCESSING (TIP), IEEE TRANSACTIONS ON MULTIMEDIA (TMM), IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS (TITS), and *PR*.

Xia Sun received the B.Eng. degree in computer science and technology from Hefei University, Hefei, China, in 2018. He is currently pursuing the M.Eng. degree in computer science and technology from Anhui University, Hefei. His research interests include computer vision, vehicle re-identification, and machine learning.



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

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