Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2024.058586

ARTICLE

Pseudo Label Purification with Dual Contrastive Learning for Unsupervised Vehicle Re-Identification

Jiyang Xu[1], Qi Wang[1,*], Xin Xiong[2], Weidong Min[1,3], Jiang Luo[4], Di Gai[1] and Qing Han[1,3]

[1]School of Mathematics and Computer Sciences, Nanchang University, Nanchang, 330031, China
[2]The First Affiliated Hospital, Jiangxi Medical College, Nanchang University, Nanchang, 330006, China
[3]Institute of Metaverse, Nanchang University, Nanchang, 330031, China
[4]Jiangxi Fangxing Technology Company Limited, Nanchang, 330025, China

* Corresponding Author: Qi Wang. Email: wangqi@ncu.edu.cn

Received: 15 September 2024; Accepted: 13 December 2024; Published: 06 March 2025

ABSTRACT: The unsupervised vehicle re-identification task aims at identifying specific vehicles in surveillance videos without utilizing annotation information. Due to the higher similarity in appearance between vehicles compared to pedestrians, pseudo-labels generated through clustering are ineffective in mitigating the impact of noise, and the feature distance between inter-class and intra-class has not been adequately improved. To address the aforementioned issues, we design a dual contrastive learning method based on knowledge distillation. During each iteration, we utilize a teacher model to randomly partition the entire dataset into two sub-domains based on clustering pseudo-label categories. By conducting contrastive learning between the two student models, we extract more discernible vehicle identity cues to improve the problem of imbalanced data distribution. Subsequently, we propose a context-aware pseudo label refinement strategy that leverages contextual features by progressively associating granularity information from different bottleneck blocks. To produce more trustworthy pseudo-labels and lessen noise interference during the clustering process, the context-aware scores are obtained by calculating the similarity between global features and contextual ones, which are subsequently added to the pseudo-label encoding process. The proposed method has achieved excellent performance in overcoming label noise and optimizing data distribution through extensive experimental results on publicly available datasets.

KEYWORDS: Unsupervised vehicle re-identification; dual contrastive learning; pseudo label refinement; knowledge distillation

1 Introduction

The purpose of vehicle re-identification (Re-ID) is to retrieve vehicles with specific identities under cross-camera surveillance systems [1,2]. Unsupervised vehicle Re-ID refers to the accurate retrieval of a given vehicle image from gallery datasets without any data annotation [3–5]. Due to the high similarity in appearance of vehicles, the goal of unsupervised vehicle Re-ID is to effectively distinguish the feature distribution within the data domain. Recently, with the development of neural networks [6,7] in the field of computer vision, unsupervised vehicle Re-ID methods have achieved significant performance on public datasets using clustering labeled pseudo labels [8–11]. However, existing works only consider the use of global features as clustering inputs to generate pseudo-labels, which not only generates a large amount of pseudo-label noise but also affects the optimization of unsupervised data distribution.

Due to the high similarity among vehicles, this poses challenges to the judgment in some unsupervised vehicle Re-ID works [8,9,12] that rely on visual representation. Recent studies have employed contrastive learning techniques to address the distance relationships between samples [11,13–15]. By constructing positive and negative sample pairs and optimizing the inter-cluster distance, these methods effectively segregate hard negative samples and train the network. For instance, Dai et al. [13] established contrastive learning within the centroid of clustering to optimize the distance between samples and centroids. Lan et al. [15] segmented the image into three parts and utilized three centroid bank for contrastive learning. However, this approach may lead to excessive memory consumption. Although these contrastive learning methods have demonstrated remarkable performance in unsupervised vehicle Re-ID tasks, they overlook the distributional characteristics in the data domain. Consequently, the development of methodologies that optimize data distribution and construct superior visual representations remains an issue deserving of further investigation.

To tackle the issue of pseudo-label noise, previous works utilize the pseudo-label refinement strategies [9,10,15,16]. These methods use clustering filtering or knowledge distillation-based feature optimization to generate accurate label information. For example, Chen et al. [10] used contrastive learning by calculating the similarity score between the original image and the enhanced image as a pseudo label of the image. Wang et al. [9] adopted a joint clustering filtering method with a teacher network to filter out labels with low similarity scores and realize the assignment of false labels to images. However, due to the loss of feature information extracted by convolutional neural network (CNN), these methods rely on global features as the basis for image clustering or pseudo-label assignment, while neglecting the information between image contexts. The above-mentioned methods may make it more difficult for the Re-ID model to distinguish between hard samples.

Our motivation is to explore the granularity information of bottleneck blocks, reduce the noise interference by clustering, and optimize the distribution of vehicle features. Specifically, our goal is to increase the distance between inter-class features while reducing the distance between intra-class features. To overcome the limitations of aforementioned methods, this paper proposes a novel fully unsupervised vehicle Re-ID framework consisting of two components: the distillation-based dual contrastive learning method (DCL) and context-aware pseudo label refinement (CPLR). The proposed method gradually correlates the granularity of information at different levels of the network, effectively reducing noise interference in the process of generating pseudo labels for global features. Additionally, we constructed a contrastive learning method between the student network and the teacher network to deeply explore the feature distribution within the data domain.

Our contributions can be summarized as follows:

- A dual contrastive learning framework based on knowledge distillation is designed to to improve the distribution of unsupervised sample features. In the clustering stage, the teacher model is used to divide the domain data after clustering, provide the student model joint contrastive learning, and discover the sample information with more discriminative ability.
- We propose a context-aware pseudo label refinement strategy to improve the awareness of image context. The contextual features of images are calculated using the differences in granularity information between different levels of the network, and the context-aware score calculated with the global features is used to provide reliable pseudo-labels.
- Extensive experimental results demonstrate the effectiveness of our method, significantly outperforming existing state-of-the-art methods on several mainstream vehicle Re-ID tasks.

The remaining structure of this paper as follows. Section 2 reviews the related work. Section 3 provides a detailed introduction to the proposed methods. Section 4 presents experimental data to validate the superiority of the proposed methods. Section 5 concludes this paper.

2 Related Work

2.1 Unsupervised Vehicle Re-ID

Existing unsupervised vehicle Re-ID methods mainly focus on how to smoothly assign the one-hot label weights to other categories after clustering. The feature space-based label smoothing methodology primarily entails establishing potential correlation between global and localized features [15–18], or augmenting the global feature representation through the incorporation of supplementary modal information [1,19]. Cho et al. [16] proposed the partially guided pseudo-label refinement (PPLR) method, which exploits the complementary relationship between global and local features to reduce label noise. He et al. [19] proposed a graph-based progressive fusion network to fuse the RGB features and multi-infrared features of vehicles. Furthermore, inspired by transfer learning methodologies that leverage intra-domain category relationships, several endeavors [4,20–22] have employed style transfer techniques to generate samples characterized by distinct domain styles and then mine the intra-domain and inter-domain category relationships to smooth label weights. Wang et al. [20] proposed dual constrained label smoothing to monitor unlabeled source domain data from few-sample source domain data to mine the information of source domain data, and guide the style transfer of different domain data through domain difference penalty. Ding et al. [21] proposed adaptive exploration to deal with the uneven distribution of image domains after clustering. These methods have demonstrated excellent performance in addressing noise and data domains. However, label smoothing in these methods only utilizes global and local features, and irrelevant local features may introduce excessive redundancy, thereby increasing computational load and potentially reducing feature discriminative ability. This work explores how to extract contextual features from the model blocks to improve the quality of pseudo labels.

2.2 Contrastive Learning

The contrastive learning method in unsupervised vehicle Re-ID tasks is mainly based on Momentum Contrast (MoCo) [23] to optimize features distribution by constructing positive and negative samples. The first paradigm relies on clustering outcomes, wherein datasets are divided into positive and negative sample pairs, followed by comparative learning conducted on these instances [10,11,24,25]. Ge et al. [11] proposed a self-paced contrastive learning framework to provide hybrid supervision through multiple different forms of category prototypes to fully exploit the distribution of data within clusters. Hu et al. [25] proposed a hard sample-guided hybrid contrastive learning method to improve feature representation by contrastive learning clustering centers and instance samples. The second paradigm conducts samples to perform contrastive learning on the centroids of clusters, and update the features of centroids with momentum in each learning process [13,15,26]. Dai et al. [13] proposed the centroid contrastive learning method to better partition the feature distribution of different instances. Yang et al. [26] proposed a contour-guided mask autoencoder method to extract the edge information of the vehicle contour to improve the quality of the label. These methods have contributed to exploring and optimizing the feature distribution of data, but they have generally neglected the imbalance of sample distribution in unsupervised processes. In contrast, our motivation is to explore the imbalance of the initial pseudo-label assignment and further optimize the feature distribution of the samples.

2.3 Knowledge Distillation

Knowledge distillation is an approach for teaching the knowledge of a complicated model to a simple model [27], which tries to lead the training of a student model on a downstream task by using the prior knowledge of a teacher model on an upstream job. Several recent research studies [9,15,28,29] have used knowledge distillation into unsupervised vehicles Re-ID task. Typically, these methods use the teacher model

for clustering to guide the training of the student model, and the student network model updates the teacher network using exponential moving average (EMA). Wang et al. [9] proposed an uncertainty-aware clustering method that assigns pseudo-labels through collaborative filtering of teacher and student networks. Lan et al. [15] employed an off-line distillation approach by training a teacher model from noisy pseudo-labels, which is then used to guide the learning of a student model. Ge et al. [28] utilized the joint training method of multi-teacher networks to perform joint label smoothing operations on the labels of images through the joint learning of two student networks. The aforementioned methods illustrates the superior efficacy of distillation techniques in the Re-ID field. To some extent, these methods rely on the quality of the pseudo-labels extracted by teacher model but ignore the noise existing in the initial teacher model. Distinct from those methods, we endeavor not only to use the teacher model for the extraction of high-quality pseudo-labels but also to derive more reliable information from the student model.

3 Proposed Method

3.1 The Overall Framework

As shown in Fig. 1, We extract contextual features of the extracted model from the training set. In the clustering stage, the teacher model is used to calculate the similarity distance matrix of the global features and the contextual features respectively, and the two pair-wise distances are clustered after similarity fusion to reduce the noise influence of global features. At each iteration, the training set is randomly divided into two subdomains according to the pseudo label category to simulate different data distributions for the learning of the dual contrastive model. Then, each student model predicts the two extracted features and performs label smoothing between the prediction vector and the one-hot label to obtain more reliable pseudo-labels for the learning of the loss function. The student model performs parameter updates under the joint supervision of softmax-triplet contrastive loss, context-aware identity discrimination loss, and centroid contrastive loss. After each round of learning, the dual contrastive model updated the parameters of the teacher model by collaboration EMA (Co-EMA) to obtain a more stable distillation effect.

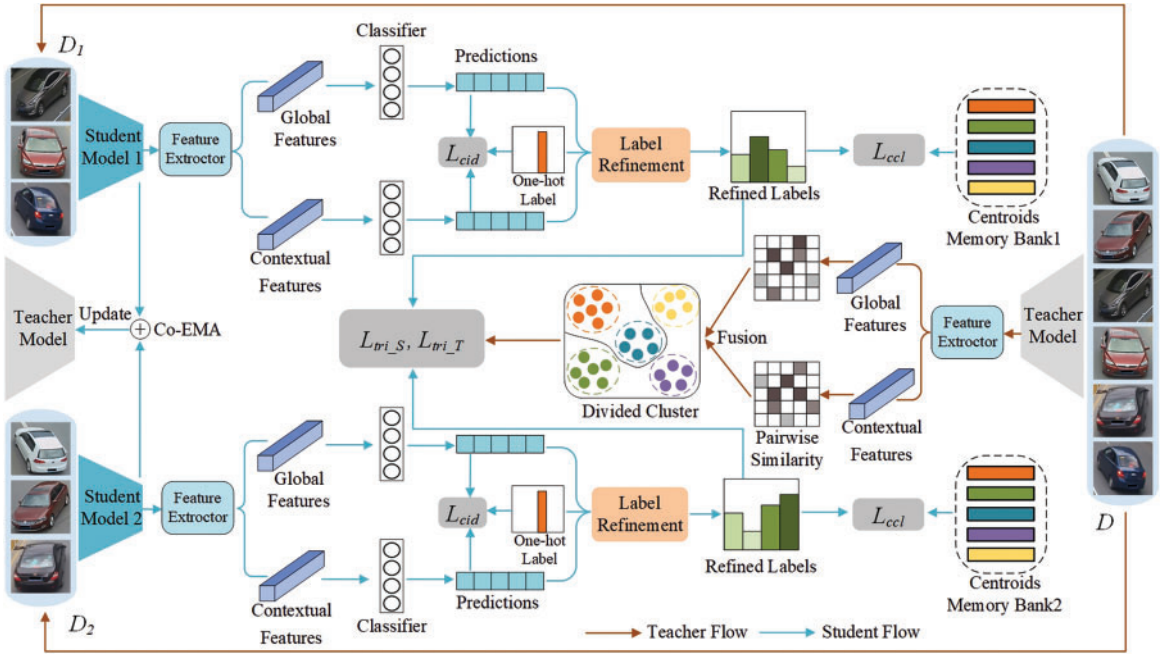

Figure 1: The overall flow of the proposed method

3.2 Distillation-Based Dual Contrastive Learning

The classification of unsupervised samples is limited by the quality of feature extraction, particularly during the early training phase when the pseudo-label distribution is uneven, which makes it challenging for the model to learn to differentiate hard negative data. Inspired by Shi et al. [30], the study employed contrastive learning to optimize the distance of data features in the dataset. We consider the variances in data distribution within clusters and introduce a dual contrastive learning framework incorporating knowledge distillation. This framework consists of two student models with identical network topology and a teacher model.

Given $\mathcal{D} = \{x_1, x_2, \ldots, x_n\}$ denote the unlabeled training dataset, the global feature $F_g = \left\{f_1^g, f_2^g, \ldots, f_n^g\right\}$ is obtained by the network $f_\theta(x_i)$ extraction. In the clustering phase, we first use the teacher network to cluster the data domain $\mathcal{D}$. Subsequently, the samples of each cluster are randomly divided into two subdomains $\mathcal{D}_1$ and $\mathcal{D}_2$ based on the $\beta \in (0,1)$ ratio. We allocate centroid memory banks for each subdomain to facilitate online learning for students. The centroid is obtained through Eq. (1):

$$\varphi_j = \frac{1}{|C_j|} \sum_{i \in C_j} f_i^g, \tag{1}$$

where $\boldsymbol{C_j}$ is the number of samples in category $\boldsymbol{j}$, The value of $\boldsymbol{C_j}$ changes with each round clustering results. During each iteration of training, the centroid will be updated. The update for: $\boldsymbol{\varphi_j^t \leftarrow m\varphi_j^{t-1} + (1-m)\, f_i^g}$, where m is momentum update factor.

The pioneering work Wei et al. [3] and Lan et al. [15] employed the cluster-level noise-contrastive estimation (ClusterNCE) loss for contrastive learning, optimizing feature distribution by generating pair-wise positive and negative samples. However, information from hard samples may be erroneously grouped into the same cluster. We utilize soft label information (Section 3.3) to guide ClusterNCE, aiming to balance the convergence rates of different types of samples through soft labels. Shown in Fig. 2, we employ a strongly supervised prototype as supervisory information, considering pairs within the same cluster as positive and those between different clusters as negative. The centroids contrastive loss $\mathcal{L}_{ccl}$ can be defined as Eq. (2):

$$\mathcal{L}_{ccl} = -\sum_{i=0}^{D} y_i \cdot log \frac{exp\left(\left\langle f_i^g \cdot \varphi_i^T \right\rangle / \tau\right)}{\sum_{j=0}^{C} exp\left(\left\langle f_i^g \cdot \varphi_j^T \right\rangle / \tau\right)}, \tag{2}$$

where $\langle \cdot \rangle$ indicates cosine similarity, $\boldsymbol{\tau}$ is a temperature hyper-parameter, $\boldsymbol{y_i}$ is the pseudo label of instance $\boldsymbol{i}$. The pseudo labels are allocated from each subdomain. Therefore, the total loss for one student training is as Eq. (3):

$$\mathcal{L}_{net_i} = \mathcal{L}_{ccl} + \mathcal{L}_{cid}, \tag{3}$$

where $\mathcal{L}_{cid}$ represents the context-aware identity discrimination loss (Section 3.3). After each iteration, the teacher model will be jointly updated by the dual contrastive model in the Eq. (4) in a Co-EMA manner:

$$f_{\theta_T}^t = \alpha f_{\theta_T}^{t-1} + (1-\alpha)\left(\beta f_{\theta_1} + (1-\beta) f_{\theta_2}\right), \tag{4}$$

where $\boldsymbol{f_{\theta_T}^{t-1}}$ denotes $\boldsymbol{t-1}$ hour iterative teachers model parameters, $\boldsymbol{f_{\theta_1}}$ and $\boldsymbol{f_{\theta_2}}$ respectively two students model parameters. The parameter $\boldsymbol{\alpha}$ is the hyper-parameter for updating the momentum of the model, and $\boldsymbol{\beta}$ is the scaling factor used to partition the data domain $\mathcal{D}$.

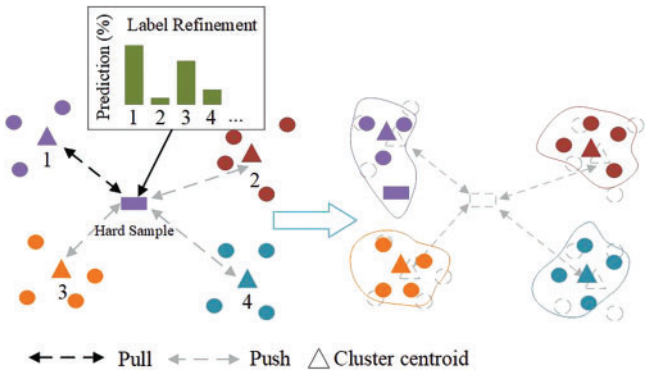

Figure 2: Loss $\mathcal{L}_{\mathbf{ccl}}$ description, monitored by label smoothing method, simple samples are closer to the centroid, while hard samples have a slower speed of approaching the centroid

To handle the global features $\boldsymbol{f_i^g}$ and their pseudo-labels $\boldsymbol{y_i}$ that are extracted by each student network, we also utilize softmax-triplet loss [31]. In addition, we merge the features extracted from the sub-domains, and jointly calculate the softmax-triplet contrast loss of the dual contrastive network and the teacher network, to better eliminate mistake amplification. The expression is as Eq. (5):

$$\begin{aligned} \mathcal{L}_{tri\_S} &= -\sum log\left(\frac{e^{\|f_i^g - f_{i,n}^g\|}}{e^{\|f_i^g - f_{i,p}^g\|} - e^{\|f_i^g - f_{i,n}^g\|}}\right) \\ \mathcal{L}_{tri\_T} &= -\sum log\left(\frac{e^{\|\widetilde{f_i^g} - \widetilde{f_{i,n}^g}\|}}{e^{\|\widetilde{f_i^g} - \widetilde{f_{i,p}^g}\|} - e^{\|\widetilde{f_i^g} - \widetilde{f_{i,n}^g}\|}}\right) \end{aligned}, \tag{5}$$

where $\boldsymbol{f_i^g}, \widetilde{\boldsymbol{f_i^g}}$ respectively from the student network $\boldsymbol{f_\theta}$ and teacher network $\boldsymbol{f_{\theta_T}}$ extract query instance $\boldsymbol{i}$ global features. $(\boldsymbol{i}, \boldsymbol{n}), (\boldsymbol{i}, \boldsymbol{p})$ denote the positive and negative samples of query instance $\boldsymbol{i}$, respectively.

3.3 Context-Aware Pseudo Label Refinement

Contextual feature extraction. Granularity information typically refers to the degree to which data or information is subdivided at different levels or scales. As illustrated in Fig. 3, there are significant differences in the scale of image features presented by different layers of network blocks. It is noteworthy that in this work, our granularity information is cross-hierarchical. We hierarchically extract granularity information from four network blocks and achieve progressive correlation through a self-attention (SA) mechanism, effectively establishing long-term dependencies between these features.

Specifically, we applied $\mathbf{1} \times \mathbf{1}$ convolution, global average pooling (GAP), ReLU and batch norm (BN) to feature maps from each stage at different scales, aligning their feature dimensions to 2048. Subsequently, we utilize the feature maps from the previous layer as Q, and those from the subsequent layer as K and V for feature enhancement. The contextual features of each layer can be defined as Eq. (6):

$$SA\left(f_i, f_j\right) = softmax\left(\frac{f_j \cdot f_i^T}{\sqrt{d_k}}\right) f_i, \tag{6}$$

where f_i from the current layer as K & V, f_j from the prior layer as Q. In particular, we employ the multi-head self-attention (MSA) mechanism to establish the relationship of features, which is an extended form of multiple independent SA modules. It is denoted as $MSA = [SA_0, SA_1, \ldots, SA_m]\, W$, where W is the projection matrix. We empirically set m to 4.

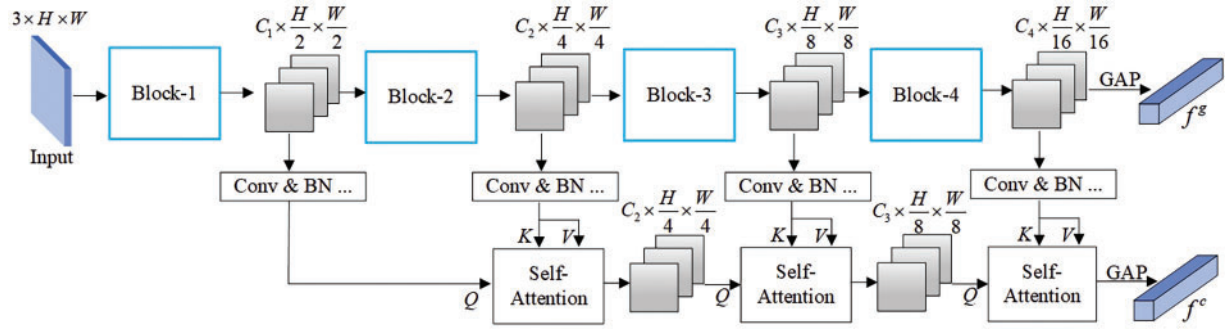

Figure 3: Illustration of the contextual feature extraction. Association of feature maps extracted from different levels in the network through a progressive approach

Based on Eq. (6), we compute the contextual feature information between the current layer and the previous layer in a progressive manner, thereby deriving the contextual features of the image. Let us review whether SA based contextual feature extraction is necessary? Although the implementation of SA undoubtedly escalates the complexity of the model, it is noteworthy that, in comparison to traditional methods of direct feature concatenation, SA can effectively capture the dependency relationships among the various components of the input data. This capability facilitates a more robust contextual understanding during the feature extraction process, especially considering the recognition interference that may arise from the high similarity between vehicle models. We deem that, in addition to relying on the network for extracting global features of the image, it is also necessary to focus on the contextual information of the image (such as fine-grained vehicle information: license plates, headlights, logos, etc.). In subsequent experiments, we conducted further analysis on the impact of contextual feature extraction on the overall performance of the model.

Optimization of cluster noise. Owing to the inherent informational biases in isolated global feature clustering, we employ a teacher network to extract both contextual features $\boldsymbol{f_i^c}$ and global features $\boldsymbol{f_i^g}$. Subsequently, during the clustering phase, we compute the Jaccard distance matrix for each feature across the entire sample denoted $\boldsymbol{D_g}, \boldsymbol{D_c}$. The weighted pair-wise distance is implemented as follows:

$$D = (1 - \lambda_D)\, D_g + \lambda_D D_c, \tag{7}$$

where $\boldsymbol{\lambda_D}$ is the pair-wise matrix weighting factor. In line with cluster-based methods like Chen et al. [10] and PPLR [16], we generate one-hot labels using the DBSCAN [32] clustering algorithm. This allows us to establish hard pseudo-labels $\boldsymbol{y} = \{\boldsymbol{y_1}, \boldsymbol{y_2}, \ldots, \boldsymbol{y_{n'}}\}$ for the training dataset. Because of outliers, the number of clustered samples $\boldsymbol{n'}$, is smaller than the training set sample $\boldsymbol{n}$.

Pseudo-label refinement. Although contextual features are employed during the clustering phase to mitigate biases introduced by global features, the resulting labels remain fundamentally hard labels. Feature extraction and clustering algorithms can impact the quality of label assignment, thereby complicating the attainment of effective generalization. Owing to variations in visual attention regions, contextual features and global features convey complementary information. By leveraging the global features $\boldsymbol{f_i^g}$ and contextual features $\boldsymbol{f_i^c}$ of a query image, we employ cosine similarity to compute a similarity-aware score $\boldsymbol{S_i}$, as Eq. (8):

$$S_i = \frac{f_i^g \cdot f_i^c}{\| f_i^g \|_2 * \| f_i^c \|_2}, \tag{8}$$

A high similarity perception score means that there is a significant correlation between global features and semantic context features, and the two information can complement each other to provide a more comprehensive feature representation. Conversely, a low similarity perception score means that the intersection of the two provides unreliable information. $\boldsymbol{q} = \{\boldsymbol{q_1}, \boldsymbol{q_2}, \ldots, \boldsymbol{q_C}\}$ is obtained by $\boldsymbol{f_\theta}(\boldsymbol{x_i})$ prediction get $\boldsymbol{C}$ category labels. The label $\boldsymbol{y_i}$ is smoothed as Eq. (9):

$$y_i = (1 - \lambda_h)\, y_i^h + \lambda_h \left[(1 - S_i)\, q_i^g + S_i\, q_i^c\right], \tag{9}$$

where $\boldsymbol{\lambda_h}$ is a constant and is set to 0.7 in the experiments, $\boldsymbol{y_i^h}$ represents one hot label, derived from clustering algorithm. $\boldsymbol{q_i^g}, \boldsymbol{q_i^c}$ extracted from global features and contextual features respectively. During the training phase, the student model's loss function (Section 3.2) is computed using the trustworthy pseudo labels that we acquired from label smoothing. The loss associated with context-aware identity discrimination $\mathcal{L}_{\boldsymbol{cid}}$ can be defined as Eq. (10):

$$\mathcal{L}_{cid} = -y_i \cdot \log\left((1 - S_i)\, \frac{exp\left(q_i^g\right)}{\sum_{j=1}^{C} exp\left(q_j^g\right)} + S_i \frac{exp\left(q_i^c\right)}{\sum_{j=1}^{C} exp\left(q_j^c\right)}\right), \tag{10}$$

3.4 Training Objective and Real-Life Application

Overall, the training loss arises from two student networks and a teacher network. The calculation of the overall framework's training loss is as Eq. (11):

$$\mathcal{L}_{total} = \beta \mathcal{L}_{net_1} + (1 - \beta)\, \mathcal{L}_{net_2} + \gamma \mathcal{L}_{tri\_S} + (1 - \gamma)\, \mathcal{L}_{tri\_T}, \tag{11}$$

where $\boldsymbol{\gamma}$ is the comparison weight parameter of the triplet loss of the teacher network and the dual contrastive network, and $\boldsymbol{\beta}$ is the training dataset partition factor, which is used to balance the loss weight of the two student networks.

Clearly, the two student networks are updated the parameters by $\mathcal{L}_{\mathbf{total}}$, and the teacher network is jointly updated with the two student networks in a Co-EMA manner (details in Eq. (4)). The overall training process is shown in Algorithm 1.

In summary, the real-life application scenarios of the model include:

(1) Vehicle tracking: achieve continuous tracking of target vehicles across cameras. For example, in cross camera traffic scenarios, the vehicle re identification system can integrate data from various security surveillance cameras in the city, accurately identify and associate the same vehicle in different video frames, and construct a complete driving trajectory of suspected vehicles including fake license plates and obscured license plates, providing key clues for case investigation.

(2) Cross city model deployment: based on unsupervised learning methods, the system can explore the potential patterns and structures of the data itself, learn directly from a large amount of unlabeled monitoring data, and do not rely on labeled data for specific traffic scenarios. Therefore, it has better cross city retrieval capabilities, can adapt to the traffic environment of different cities, and achieve generalized deployment of the model.

Algorithm 1: The training procedure of proposed method

Inputs: Initialize the student models $\{f_{\theta_1}, f_{\theta_2}\}$, The teacher model f_{θ_T}, DBSCAN, Unlabelled training dataset $\mathcal{D}$, Dataset partitioning factor β.

for i = 0 in $[1, epochs]$ **do**
  $f_{\theta_T} \rightarrow \{f_i^g, f_i^c\}$;
  Calculate Jaccard distance by $\{f_i^g, f_i^c\}$, and cluster into C clusters through DBSCAN;
  Divide D in proportion β to category $\{\mathcal{D}_1, \mathcal{D}_2\}$ and generate pseudo-label;
  Initialize cluster memory dictionaries $\{\varphi_{\mathcal{D}_1}, \varphi_{\mathcal{D}_2}\}$ with Eq. (1);
  for j = 0 in $[1, iterations]$ **do**
    P × K sample query images $x_{\mathcal{D}_1}, x_{\mathcal{D}_2}$ from $\mathcal{D}_1, \mathcal{D}_2$ respectively;
    Extract features $\{f_i^g, f_i^c\}$ by $\{f_{\theta_1}(\mathcal{D}_1), f_{\theta_2}(\mathcal{D}_2)\}$;
    $S_i \leftarrow \{f_i^g, f_i^c\}$ with Eq. (9), assign soft label with Eq. (10);
    Calculate the total loss with Eq. (11);
    Update centroid memory banks and teacher model parameter with Eq. (4).
  end
end

4 Experiments

In this section, we will experimentally analyze the performance of the proposed method. The following four issues need to be considered: **RQ1:** Is the effect of label smoothing better than other methods. **RQ2:** How DCL affects the performance of models. **RQ3:** How to assess the impact of label smoothing beyond accuracy. **RQ4:** How to evaluate the contribution of contextual feature to global feature representation.

4.1 Dataset and Evaluation Protocols

VeRi-776 [33] is a basic dataset widely used in vehicle re-identification research. It consists of over 50,000 images captured by 20 cameras covering 776 different vehicles. The training set contains 37,781 images of 576 vehicles, the query set contains 1678 images of 200 vehicles, and the gallery set contains 11,579 images of the same 200 vehicles.

VERI-Wild [34] is a large-scale vehicle re-identification dataset consisting of 416,314 images of 40,671 vehicles captured by 174 cameras. Different from the VeRi-776 [33] datasets, the VERI-Wild dataset has differences in illumination, weather and night changes caused by time span. The training set contains 277,797 images of 30,671 vehicles, and the test set contains 128,517 images of 10,000 vehicles and is further subdivided into three subsets of different sizes: Test3000, Test5000, and Test10000.

Following the general evaluation metrics in the vehicle Re-ID task, we use cumulative matching curve (CMC) and the mean average precision (mAP) proposed by Zheng et al. [35] to evaluate the performance of the proposed method.

Rank-k. Rank-k in the CMC curve is used to evaluate the matching degree of the model at different rankings. Rank-k calculation is as follows:

$$Rank - K = \frac{\sum_{i=1}^{N} gt(i,k)}{N}, \tag{12}$$

where $\boldsymbol{N}$ represents the total number of vehicle images in the query set. When there are accurately matched images in the K-th retrieved images $\boldsymbol{gt}(\boldsymbol{i},\boldsymbol{k}) = \mathbf{1}$, otherwise $\boldsymbol{gt}(\boldsymbol{i},\boldsymbol{k}) = \mathbf{0}$.

mAP. The average precision (AP) for each image in the query set is calculated as follows:

$$AP = \frac{\sum_{k=1}^{M} P(k) \times gt(k)}{N_{gt}}, \tag{13}$$

where M is the length of the entire gallery set, $\boldsymbol{N_{gt}}$ denotes the number of images in the gallery set with the same ID as the query image, and $\boldsymbol{P(k)}$ denotes the accuracy of the top k query result. If the ID of the k-th image is the same as the query image $\boldsymbol{gt(k)} = \mathbf{1}$, otherwise $\boldsymbol{gt(k)} = \mathbf{0}$. MAP is the average AP value of the entire query set N, which can be defined as:

$$mAP = \frac{\sum_{i=1}^{N} AP_i}{N}. \tag{14}$$

mAP comprehensively reflects the accuracy of the model across all retrieval results. A higher mAP value indicates superior performance of the model in accurately matching vehicles. Conversely, Rank-k metric signifies the probability that at least one of the top k retrieved results is a positive sample, with Rank-1 and Rank-5 being common evaluation criteria. These Rank-k metrics can more directly indicate the retrieval performance of the model in comparison to mAP. Nevertheless, neither Rank-k nor mAP alone can accurately describe the performance of the re identification system. Therefore, both indicators must be considered simultaneously to comprehensively represent the retrieval performance of the model.

4.2 Implementation Details

We adopt ResNet50 [36] as ours backbone, we remove all sub-module layers after the fourth layer and add the GAP operation as the representation of global features. Initialize these two student network parameters using ImageNet [37]. All experiments were performed on 2× NVIDIA Tesla V100 GPU. Our training process divided into two stages. The whole training process is divided into 50 epochs, in the image clustering stage, we use DBSCAN [32] as the clustering algorithm to assign pseudo-labels to images. On the VeRi-776 dataset, the maximum distance d is set to 0.7, while on the VERI-Wild datasets, the maximum distance d is set to 0.6. Due to device limitations, all our VERI-Wild experiments use a uniform set of 40,000 images as the training set. During training phase, we chose an initial learning rate of 3×10^{-4}, decreasing by a factor of 10 every 20 epochs. Use the Adam optimizer to optimize the weights of the network and set the weight decay to 5×10^{-4}. To augment the data, we used random horizontal flips and random occlusion [38], both with probability set to 0.5. During training phase, we set the weighting factor λ_D of the pair-wise matrix to 0.7. In addition, the weight coefficient γ of $\mathcal{L}_{tri\_S}$ and $\mathcal{L}_{tri\_T}$ is set to 0.8.

4.3 Ablation Studies

Influence of the Different Modules: To verify the effectiveness of the proposed framework in the unsupervised vehicle Re-ID task, we conduct experiments to analyze the combination effect of the CPLR and DCL method. The explanation for each ablation module is as follows:

- "Baseline" means using the traditional clustering-based unsupervised pipeline without any ablation modules.
- "**w/** CPLR" indicates that only context-aware pseudo label refinement strategy based on "Baseline".
- "**w/** DCL" indicates employing only dual contrastive learning methods based on "Baseline".
- "**Ours**" indicates the use of two proposed ablation modules.

In the Table 1, we compare the results of different combinations of modules. The results show that DCL exhibits significant performance in both datasets, owing to its ability to differentiate the distribution of samples within the two simulated subdomains. Additionally, employing CPLR improved R-1 and mAP by 4.9% and 4.7% respectively over the "Baseline" in VeRi-776, demonstrating its effectiveness in label purification. When integrated, the combination further enhanced accuracy, clearly illustrating the complementarity of the two modules, which offers more robust information in the realm of unsupervised vehicle enrichment.

Table 1: Ablation studies on the impacts of individual components in VeRi-776 and VERI-Wild. "**w/**" denotes only using individual ablation modules. "R-1" and "R-5" represent the accuracy of Rank-1 and Rank-5, respectively. In subsequent experiments, we will keep the definitions of these indicators unchanged

Methods	VeRi-776			VERI-Wild (Test3000)		
	R-1	R-5	mAP	R-1	R-5	mAP
Baseline	79.6	85.6	35.1	51.8	75.9	27.3
w/ DCL	85.2	90.5	40.6	60.7	79.7	31.4
w/ CPLR	84.5	90.0	39.8	58.9	80.0	30.8
Ours	**87.8**	**92.1**	**43.2**	**62.8**	**82.8**	**32.8**

Influence of the Partitioning Factor: To explore the effect of random partitioning factor $\boldsymbol{\beta}$ in different dataset domains, we conduct experiments on VeRi-776 and VERI-Wild datasets. In each experiment, we keep the other hyper-parameters fixed and only adjust the partition factor $\boldsymbol{\beta}$. The experimental results are shown in Table 2. Experimental results show that a small value of $\boldsymbol{\beta}$ will lead to unreliable identity information and confidence, while a high value of $\boldsymbol{\beta}$ will reduce the accuracy of the model. Due to the partition of the data domain, the centroids update frequency in the subdomain and the loss function of the student network change, which leads to the network influence on the learning of the information in the sample. Experimental results prove that the process of pseudo-label assignment will lead to the difference of data distribution in the domain. Based on the above experimental results, we choose the value of $\boldsymbol{\beta}$ 0.7 and 0.8 as the basic parameters of VERI-Wild and VeRi-776 datasets in the subsequent experiments.

Table 2: Factor of partitioning factor β value results on the VERI-Wild and VeRi-776 datasets

Parameters $\boldsymbol{\beta}$	VeRi-776			VERI-Wild (Test3000)		
	R-1	R-5	mAP	R-1	R-5	mAP
0.1	80.1	87.8	34.9	51.7	74.9	24.3
0.2	81.5	88.3	35.4	53.7	77.8	25.4
0.3	81.8	88.9	35.5	56.6	78.4	28.8
0.4	83.6	90.5	37.6	59.1	80.0	30.3
0.5	84.7	90.5	39.6	61.5	81.1	31.4
0.6	85.6	91.7	40.8	61.8	81.1	32.0
0.7	87.2	91.5	42.1	**62.8**	**82.8**	**32.8**
0.8	**87.8**	**92.1**	**43.2**	62.4	**82.8**	32.6
0.9	86.0	90.4	41.2	61.3	80.9	31.2

Analysis of Loss Function: We explore the effect of different loss functions on model performance on VeRi-776 and VERI-Wild, and Table 3 displays the results of our loss function ablation. The first row illustrates the results of utilizing only $\mathcal{L}_{\text{tri}}$ loss, which reduces mAP and R-1 by 8.1% and 12%, respectively. This indicates that the softmax-triplet loss cannot effectively optimize the model performance. Rows 2 and 3 show the ablation results of removing $\mathcal{L}_{\text{ccl}}$ and $\mathcal{L}_{\text{cid}}$, respectively, with mAP dropped by 4.4% and 2.7% in VeRi-776, respectively. This demonstrates that by including contextual feature, feature quality can be improved and features with distinct identities may be effectively distinguished in $\mathcal{L}_{cid}$ loss. Additionally, $\mathcal{L}_{\text{ccl}}$ push inter-class data farther and intra-class data closer, which will enhance the feature distribution.

Table 3: Ablation studies on the effects of different loss function in VeRi-776 and VERI-Wild

$\mathcal{L}_{tri}(\mathcal{L}_{tri\_S} + \mathcal{L}_{tri\_T})$	$\mathcal{L}_{ccl}$	$\mathcal{L}_{cid}$	VeRi-776			VERI-Wild (Test3000)		
			R-1	R-5	mAP	R-1	R-5	mAP
√			80.1	87.6	35.2	51.6	73.3	25.1
√		√	83.6	90.0	38.8	59.6	80.6	29.6
√	√		85.1	90.6	40.5	60.8	81.4	31.0
√	√	√	**87.8**	**92.1**	**43.2**	**62.8**	**82.8**	**32.8**

Effect of the label refinement strategy: To answer **RQ1**. We investigated multiple pseudo-labeling strategies, ensuring fair comparisons by conducting all experiments within the "Baseline" model, as detailed in Table 4. The "One-hot" label is derived from clustering results, with the correct cluster assigned a value of 1 and all others set to 0. The "LSR" strategy, introduced by Szegedy et al. [39], assigns a weight of 0.9 to the correct label, distributing the remaining weights evenly at 0.1 each. The "OLS" strategy, proposed by Zhang et al. [40], is an online label smoothing technique that leverages correct classifications from past epochs to refine label smoothing in the current epoch. The aforementioned method demonstrates that the hard labels produced by clustering are unreliable. However, these approaches all depend on the accuracy of global feature extraction. In contrast, "CPLR" exhibits superior effectiveness by synergistically smoothing labels using both contextual and global features, without relying on the quality of single feature extraction, thereby more effectively mitigating the impact of label noise.

Table 4: Ablation studies on the effects of the label refinement strategy in VeRi-776 and VERI-Wild

Methods	VeRi-776			VERI-Wild (Test3000)		
	R-1	R-5	mAP	R-1	R-5	mAP
One-hot	79.6	85.6	35.1	51.8	75.9	27.3
LSR [39]	80.6	88.5	37.1	53.2	77.2	27.5
OLS [40]	81.7	89.2	38.8	55.6	78.8	28.5
CPLR	**84.5**	**90.0**	**39.8**	**58.9**	**80.0**	**30.8**

Analysis of DCL: To answer **RQ2**, we conducted an in-depth analysis of DCL's impact on model performance. As shown in Table 5, removing the Memory Bank (row 2) led to a significant drop in performance. This is because DCL depends on contrastive learning to extract intra-domain feature distributions. When momentum updates were removed (row 3), the model became heavily reliant on the quality of

feature extraction at the start of each iteration, failing to utilize historical data feature distributions. We further examined the reliance on label information during contrastive learning (row 4). By removing label information from $\mathcal{L}_{ccl}$ (Eq. (2)) and using ClusterNCE for the contrastive loss, the mAP of both datasets decreased by 1% and 0.8%, respectively, suggesting that label refinement somewhat lessens the reliance on label information. Additionally, we analyzed the impact of the teacher model on performance (row 5) by substituting the teacher model task in framework Fig. 1 with a student model, encompassing pseudo-label clustering assignments, data domain partitioning, and loss training. The results showed that if the dual contrast network only learns intra-domain information, it can easily lead to model overfitting. The teacher model enhances model robustness by jointly updating parameters with feature information from their respective sub-domain distributions, facilitated by Co-EMA methods of the student networks.

Table 5: Analysis of DCL on model performance in VeRi-776 and VERI-Wild

Methods	VeRi-776			VERI-Wild (Test3000)		
	R-1	R-5	mAP	R-1	R-5	mAP
Ours	**87.8**	**92.1**	**43.2**	**62.8**	**82.8**	**32.8**
w/o Memory bank	83.6	90.0	38.8	59.6	80.6	29.6
w/o Momentum updating	83.7	90.8	39.9	60.3	81.6	30.6
w/o Label refinement	86.9	91.5	42.3	62.1	82.2	31.6
w/o Knowledge distillation	84.9	91.2	40.1	57.5	80.0	30.7

Additionally, we analyzed the impact of various momentum update rates m on the model, as illustrated in Fig. 4. The momentum coefficient m closer to 1 indicates a slower update rate. A higher m value increases the model training process's dependency on the quality of centroid feature extraction. Conversely, as m approaches 0, centroid feature updates tend towards the current sample, potentially causing frequent updates and a consequent loss of information from other features within the same cluster. The optimal model accuracy is achieved when m approaches 0.1.

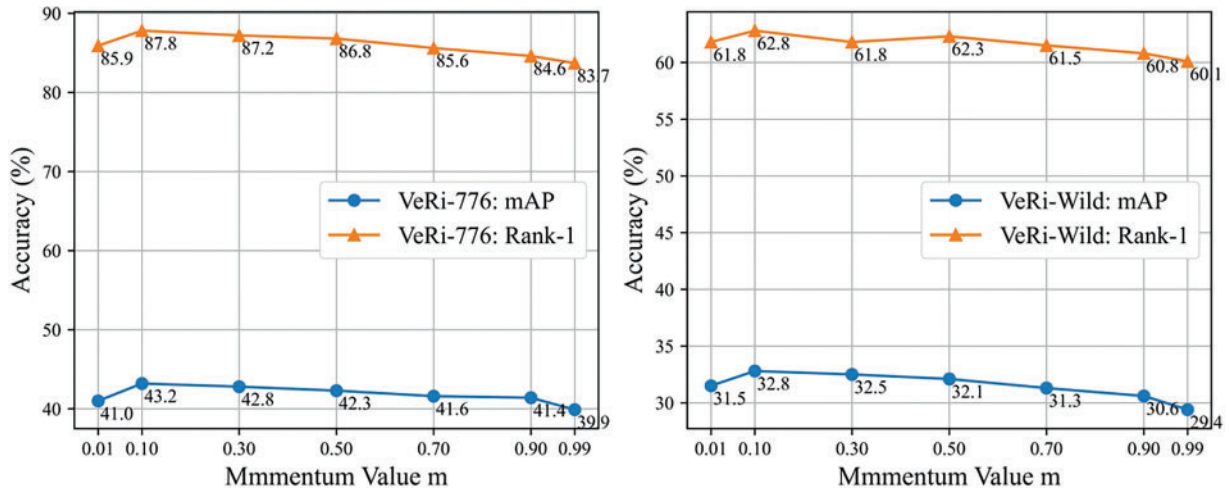

Figure 4: Ablation study of the momentum value m on model performance in VeRi-776 and VERI-Wild

4.4 Comparison with State-of-the-Arts

We evaluate our method with other state-of-the-art unsupervised vehicle Re-ID techniques, including UDA and USL methods. As shown in Table 6, the results for the two widely used vehicle datasets. We also use some open-source code in the field of object or person Re-ID in our research, as pure unsupervised vehicle re-identification approaches are currently limited. Such: MMT [28], ICE [10], HHCL [25], RLCC [41], UCF [9], Lan et al. [15]. All results are from experiments conducted in their open-source code or on their published vehicle Re-ID dataset. For fair comparison, we have kept the basic parameters of the model consistent in the open-source code. Despite the simplicity of the approach, we demonstrated strong competitive performance across both datasets.

UDA models (e.g., MMT [28], UCF [9], CTFRN [29], and TDSR [3]) first undergo a fully supervised learning phase in the source domain, followed by an unsupervised training phase in the target domain. These methods primarily address the challenges of data adaptability and domain discrepancies. For instance, MMT learns representations from the source domain data and uses a dual-teacher network to perform joint smoothing operations on the images, thereby facilitating domain adaptation. Our method is fully unsupervised and does not require the use of fully supervised source domain data for training. Our method has been implemented on ResNet50 [36] and IBN-ResNet50 [42] backbone networks. Specifically, the performance is achieved at mAP = 43.2%, 32.8%, R-1 = 87.8%, 62.8%, and R-5 = 92.1%, 82.8% on the ResNet50 backbone for VeRi-776 and VERI-Wild (Test3000), respectively.

Table 6: Comparision with the state-of-the-art methods on VeRi-776 and VERI-Wild. "*" represents unsupervised domain adaptation (UDA) method

Methods	References	VeRi-776			VERI-Wild								
					Test3000			Test5000			Test10000		
		R-1	R-5	mAP	R-1	R-5	mAP	R-1	R-5	mAP	R-1	R-5	mAP
MMT [28]*	ICLR'20	60.9	69.0	25.4	45.6	67.1	21.8	39.2	61.2	18.3	30.4	50.8	14.1
SPCL [11]	NIPS'20	79.9	86.8	36.9	52.8	77.6	27.6	48.5	72.8	26.4	38.1	61.9	20.3
HHCL [25]	IC-NIDC'21	69.6	75.6	31.0	56.3	79.7	30.2	49.2	73.3	26.1	38.3	62.3	20.5
RLCC [41]	CVPR'21	83.4	88.8	39.6	55.2	79.4	29.1	47.3	72.8	24.9	37.0	62.2	19.5
ICE [10]	ICCV'21	82.1	87.1	37.9	54.7	78.6	28.7	47.0	71.3	24.8	37.0	61.1	19.3
CACL [43]	TIP'22	62.4	73.5	27.3	57.0	80.3	30.3	48.8	74.0	26.1	38.2	63.6	20.5
CTFRN [29]*	PR'22	76.7	81.5	37.1	61.3	82.4	32.3	51.8	74.9	27.5	42.1	65.7	22.1
Cluster-Contrast [13]	ACCV'22	86.2	90.5	40.8	56.2	78.7	29.5	48.6	72.9	26.0	37.8	62.2	19.5
PPLR [16]	CVPR'22	85.6	88.7	41.6	59.6	82.1	31.4	51.5	74.5	26.9	42.1	65.3	22.2
Lan et al. [15]	TIP'23	78.5	84.8	35.1	56.3	80.4	30.2	47.3	71.0	25.5	36.5	60.8	19.9
TDSR [3]*	TITS'23	86.8	92.1	40.0	–	–	–	–	–	–	–	–	–
UCF [9]*	TMM'23	85.2	89.2	40.5	62.3	82.9	32.3	51.0	74.3	26.1	41.8	65.2	21.8
Ours (ResNet50)	This paper	87.8	**92.1**	**43.2**	62.8	82.8	32.8	53.0	75.6	27.5	43.5	67.3	22.6
Ours (IBN-ResNet50)	**This paper**	**88.4**	91.8	42.6	**63.4**	**83.3**	**33.2**	**53.2**	**76.2**	**28.1**	**44.2**	**68.5**	**22.8**

The methods SPCL [11], Cluster-Contrast [13], HHCL [25] and CACL [43] are based on contrastive learning techniques. These methods focus on how to optimize features within the data domain. Such SPCL proposes an automatic contrastive learning framework that makes use of a range of distinct category prototypes to provide hybrid supervision. HHCL proposes instance contrastive learning to mine the information between instances. Although these methods have explored optimizing feature space distribution, they have higher requirements for pseudo labels in data allocation. It is notable that the primary parameters of contrastive learning methods encompass the quantity of input data and update size for each iteration. Under

the same momentum update parameters and batch size, our mAP on VeRi-776 and VERI-Wild (Test3000) improves by 6.3% and 5.2%, respectively, in comparison to SPCL.

Furthermore, our method outperforms several label refinement techniques, including UCF [9], Lan et al. [15], ICE [10] and PPLR [16]. These methods employ teacher models or local features for label smoothing to mitigate pseudo-label noise. Although PPLR uses a label smoothing method, like most cluster-based algorithms, it only uses global features for cluster assignment pseudo-labels and relies on the reliability of the clusters. To lessen the inherent bias of global features, our method considers the contextual information of the images during the clustering process. These methods are based on the clustering label smoothing paradigm, and we uphold the consistency of the clustering algorithm, encompassing same clustering parameters, backbone architecture, training epochs and batch size. Compared to PPLR, our mAP in VeRi-776 and VERI-Wild (Test3000) improved by 1.6% and 1.4%, respectively.

4.5 Further Analysis

Effective of different backbones. Considering that unsupervised vehicle Re-ID relies on the quality of visual feature extraction, we studied various mainstream CNN visual extraction backbone architectures (PCB, OSNet, DenseNet121, ResNet50) in Table 7. It can be observed that the proposed method achieves optimal Re-ID performance when using ResNet50 as the backbone network. Compared to DenseNet, ResNet establishes residual connections between network blocks without additional parameters. This allows ResNet to maintain efficient performance while having fewer parameters and shorter back propagation time. In addition, although PCB and OCN are widely used as universal backbone networks for Re-ID, their scalability is limited due to their main design for pedestrian features, and they cannot effectively extract a wide range of contextual features for the granularity information of vehicles. Therefore, the experimental results and analysis have verified the rationality and superior performance of choosing ResNet50 as the backbone network in this paper.

Table 7: Comparison of different backbones on VeRi-776 and VERI-Wild

Methods	VeRi-776			VERI-Wild (Test3000)		
	R-1	R-5	mAP	R-1	R-5	mAP
PCB [44]	78.3	81.8	36.5	52.3	76.5	28.3
OSNet [45]	84.9	89.3	39.6	59.2	80.6	30.7
DenseNet121 [46]	83.4	88.0	39.8	56.7	78.6	29.6
ResNet50 [36]	**87.8**	**92.1**	**43.2**	**62.8**	**82.8**	**32.8**

Performance analysis of the model. We evaluated the performance distinctions between "Ours" and other unsupervised methodologies in terms of models. Specifically, we conducted a thorough evaluation of the model across two dimensions: spatial complexity and time complexity. To maintain parity in the comparison, the time complexity indicator only uses the time consumption of a single epoch during the model training phase. As delineated in Table 8, Lan et al. [15] introduced a teacher-guide student model optimization framework that markedly escalates the time complexity when contrasted with single contrastive learning techniques such as cluster concentrate [13] and HHCL [25] by segmenting the image into three parts for contrastive learning. Furthermore, MMT [28] and CTFRN [29] implement a dual teacher-student model, which collaboratively smooths the labels of their respective region images, thereby incurring supplementary time consumption during the backpropagation process. "Ours" leverages self-attention to associate network

bottleneck blocks to extract contextual features, demanding additional memory resources and failing to exhibit superiority over a solitary ResNet50 architecture. In summation, our approach is at an intermediary level, but within an acceptable performance overhead margin, it demonstrates superior model performance compared to other methods.

Table 8: Compare the performance of the model with other methods. "Params" represents the size of model parameters and are used to evaluate spatial complexity. "Time (VeRi-776)" and "time (VERI-Wild)" represent the running time in each epoch of their respective datasets, which are employed to evaluate time complexity

Methods	Params (M)	Time (VeRi-776)	Time (VERI-Wild)
MMT [28]	90	16.2 m	19.3 m
SPCL [11]	23.5	6.7 m	9.5 m
HHCL [25]	23.5	8.5 m	10.2 m
ICE [10]	23.5	9.1 m	12.7 m
CACL [43]	44.9	14.6 m	19.2 m
CTFRN [29]	90	15.2 m	20.6 m
Cluster-Contrast [13]	23.5	7.7 m	11.4 m
PPLR [16]	23.5	7.3 m	10.5 m
Lan et al. [15]	44.9	12.3 m	15.6 m
UCF [8]	44.9	10.2 m	14.1 m
Ours (ResNet50)	33.4	9.5 m	13.8 m

4.6 Visual Quality

T-SNE visualization: To intuitively demonstrate the clustering effectiveness of our proposed method, we employed T-SNE [47] to analyze the feature extraction results of different components. We randomly selected 15 samples from the VeRi-776 dataset, with each category represented by one sample. As illustrated in Fig. 5, considering the red ID sample, the features extracted by the "baseline" have a relatively scattered distribution in the feature space, making it difficult to effectively distinguish individual samples. The DCL method demonstrates a more focused feature distribution for simple data, yet it is not sufficiently distinct for distinguishing challenging samples. Following the integration of the CPLR, the discriminability of the feature distribution has improved; however, the features are still not sufficiently concentrated within the feature space, leading to confusion with similar data. By combining both modules, the extracted features demonstrate a more tightly clustered distribution within the feature space, facilitating clear differentiation between categories.

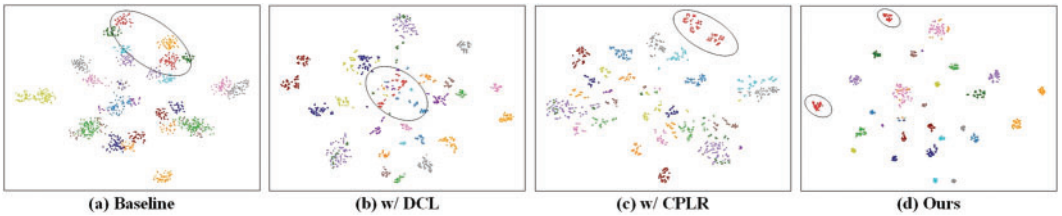

Figure 5: We selected 15 categories from VeRi-776 for T-SNE of different ablation modules, with different color points representing different categories. (a) Baseline; (b) w/ DCL; (c) w/ CPLR; (d) Ours

Cluster pseudo-label quality: To answer **RQ3**, we followed the pair-wise precision proposed by Wang et al. [48]. To evaluate the quality information of pseudo labels generated by our method and baseline during clustering. Due to the limitations of the DBSCAN on clustering results, we uniformly set the maximum distance d to 0.7. Firstly, we construct $\omega_{n\times n}$, $\Omega_{n\times n}$ matrixs for the entire sample. $\omega_{i,j}$ indicates whether samples i and j share the same pseudo label, where $\omega_{i,j} = 1$ represents a same cluster, $\omega_{i,j} = 0$ represents different clusters. Similar usage of $\Omega_{i,j}$ represents whether samples i and j have the same true label, $\Omega_{i,j} = 1$ represents a same cluster, $\Omega_{i,j} = 0$ represents different clusters. We use TP to represent positive sample pairs and FP to represent negative sample pairs. However, when calculating TP and FP, we consider the existence of outliers. The accuracy P of clustering is calculated as: $P = TP/(TP + FP)$, where TP denotes $\omega_{i,j} = 1 \,\&\, \Omega_{i,j} = 1$, FP denotes $\omega_{i,j} = 1 \,\&\, \Omega_{i,j} = 0$. As shown in Fig. 6, our method significantly improves the accuracy of clustering, with higher quality clustering and fewer outliers, providing reliable pseudo labels for model training.

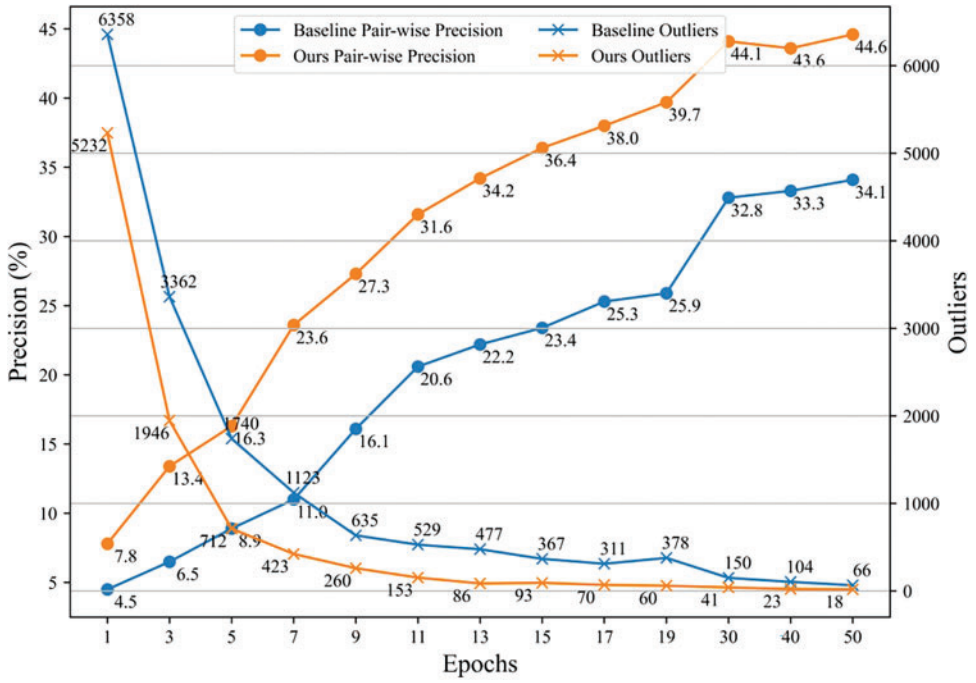

Figure 6: Quantification of pseudo label quality (%) for the proposed method and baseline on the VeRi-776 dataset. Pair-wise precision represents the accuracy of clustering

Rank-list visualization: To verify the qualitative outcomes of our proposed method, we compared the rank-list visualization with the "Baseline" and chose three different viewing angles for the retrieval. As shown in Fig. 7, the experimental results reveal that the "Baseline" tends to include images with similar backgrounds and resolutions in the query results, leading to incorrect matches that resemble the query samples in appearance and viewpoint. In contrast, "Ours" mitigates the disturbances from variations in viewpoint, background, and lighting intensity. This indicates that our approach can better distinguish, capturing their contextual information, and effectively distinguish negative samples.

Attention map visualization: To answer **RQ4**. We utilized Grad-CAM [49] for the ablation analysis of model feature predictions. We randomly selected three groups of images with different views, as shown in Fig. 8. The focus distribution of the "Baseline" is relatively scattered, making it susceptible to viewpoint

changes. However, with the integration of the CPLR module, additional contextual information is incorporated. This suggests that semantic contextual features can focus on fine-grained information within the image, resulting in more precise prediction outcomes. When DCL is combined with CPLR, the attention information in the image is further amplified.

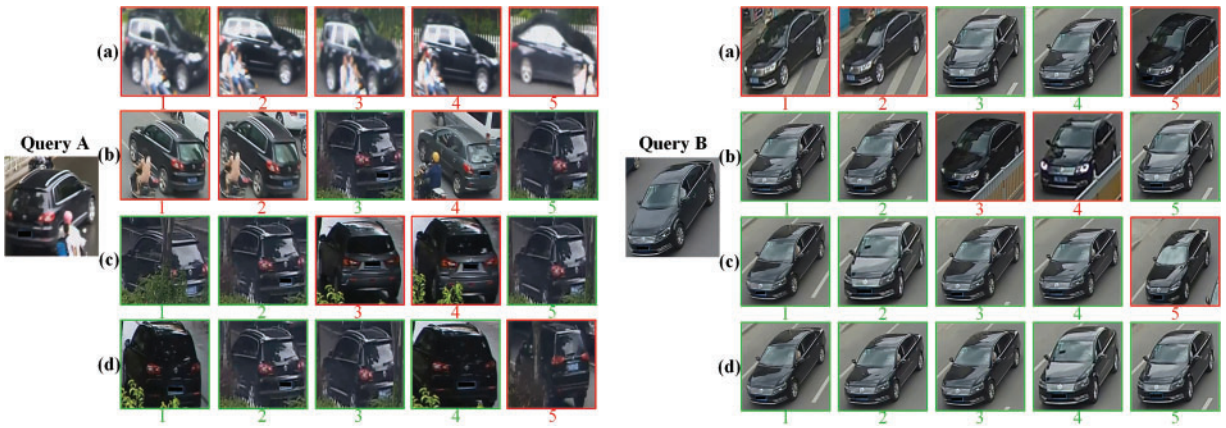

Figure 7: Retrieve Rank-5 visualization. Matched and unmatched images are marked in green and red, respectively. (a) Baseline; (b) w/ DCL; (c) w/ CPLR; (d) Ours

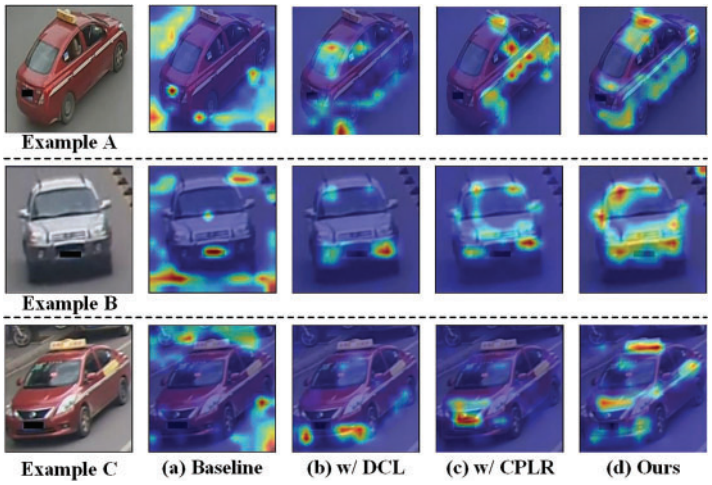

Figure 8: Visual analysis of Grad CAM ablation models for three sample vehicles. (a) Baseline; (b) w/ DCL; (c) w/ CPLR; (d) Ours

5 Conclusion and Future Works

In this work, we propose a novel unsupervised vehicle Re-ID framework to mitigate the label noise and data domain distribution problem. First, we design a DCL training method to optimize the distance distribution of data features, which effectively improves the imbalance of data domain distribution through online joint training of teachers and dual contrastive networks. Furthermore, we introduce a CPLR strategy that progressively integrates granular information from various layers of the network to extract contextual features, thereby generating more reliable pseudo-labels in conjunction with global features. Extensive experiments have confirmed the effectiveness of our approach. In future work, we will extend our research

to vehicle Re-ID in video streams, with a particular emphasis on enhancing the model's comprehension of spatial-temporal features. Additionally, we will address the challenges posed by occlusion and variations in image quality, while further investigating the performance of vehicle Re-ID in real-world scenarios.

Acknowledgement: The authors thank all editors and anonymous reviewers for suggestions, as well as to all members who have supported and contributed to this work.

Funding Statement: This work was supported by the National Natural Science Foundation of China under Grant Nos. 62461037, 62076117 and 62166026, the Jiangxi Provincial Natural Science Foundation under Grant Nos. 20224BAB212011, 20232BAB202051, 20232BAB212008 and 20242BAB25078 and the Jiangxi Provincial Key Laboratory of Virtual Reality under Grant No. 2024SSY03151.

Author Contributions: Study conception and design: Jiyang Xu, Qi Wang, Xin Xiong, Weidong Min; data collection: Jiang Luo, Di Gai, Qing Han; analysis and interpretation of results: Jiyang Xu, Qi Wang, Weidong Min, Di Gai; draft manuscript preparation: Jiyang Xu, Qi Wang, Xin Xiong. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: Data will be made available on request.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Li H, Chen J, Zheng A, Wu Y, Luo Y. Day-night cross-domain vehicle re-identification. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2024 Jun. p. 12626–35. doi:10.1109/cvpr52733.2024.01200.
2. Zhang X, Chen X, Sun W, He X. Vehicle re-identification model based on optimized DenseNet121 with joint loss. Comput Mater Contin. 2021 Jan;67(3):3933–48. doi:10.32604/cmc.2021.016560.
3. Wei R, Gu J, He S, Jiang W. Transformer-based domain-specific representation for unsupervised domain adaptive vehicle re-identification. IEEE Trans Intell Transp Syst. 2022 Dec;24(3):2935–46. doi:10.1109/TITS.2022.3225025.
4. Wang Q, Min W, Han Q, Liu Q, Zha C, Zhao H, et al. Inter-domain adaptation label for data augmentation in vehicle re-identification. IEEE Trans Multimed. 2021 Aug;24:1031–41. doi:10.1109/TMM.2021.3104141.
5. Sun W, Chen X, Zhang X, Dai G, Chang P, He X. A multi-feature learning model with enhanced local attention for vehicle re-identification. Comput Mater Contin. 2021 Jan;69(3):3549–61. doi:10.32604/cmc.2021.021627.
6. Li J, Pang M, Dong Y, Jia J, Wang B. Graph neural network explanations are fragile. Proc 41st Int Conf Mach Learn. 2024;235:28551–67.
7. Pang M, Wang B, Ye M, Cheung Y-M, Zhou Y, Huang W, et al. Heterogeneous prototype learning from contaminated faces across domains via disentangling latent factors. IEEE Trans Neural Netw Learn Syst. 2024 Jan;1–15. doi:10.1109/TNNLS.2024.3393072.
8. Zheng A, Sun X, Li C, Tang J. Viewpoint-aware progressive clustering for unsupervised vehicle re-identification. IEEE Trans Intell Transp Syst. 2021 Aug;23(8):11422–35. doi:10.1109/TITS.2021.3103961.
9. Wang P, Ding C, Tan W, Gong M, Jia K, Tao D. Uncertainty-aware clustering for unsupervised domain adaptive object re-identification. IEEE Trans Multimed. 2022 Feb;25:2624–35. doi:10.1109/TMM.2022.3149629.
10. Chen H, Lagadec B, Bremond F. ICE: inter-instance contrastive encoding for unsuper-vised person re-identification. In: Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV); 2021 Oct. doi:10.1109/iccv48922.2021.01469.
11. Ge Y, Zhu F, Chen D, Zhao R, Li H. Self-paced contrastive learning with hybrid memory for domain adaptive object Re-ID. Neural Inform Process Syst. 2020 Jun;33:11309–21.
12. Lu Z, Lin R, He Q, Hu H. Mask-aware pseudo label denoising for unsupervised vehicle re-identification. IEEE Trans Intell Transp Syst. 2023 Jan;24(4):4333–47. doi:10.1109/TITS.2022.3233565.

13. Dai Z, Wang G, Yuan W, Zhu S, Tan P. Cluster contrast for unsupervised person re-identification. In: Proceedings of the Asian Conference on Computer Vision; 2022. p. 1142–60.
14. Pang Z, Wang C, Wang J, Zhao L. Reliability modeling and contrastive learning for unsu- pervised person re-identification. Knowl-Based Syst. 2023 Jan;263(6):110263. doi:10.1016/j.knosys.2023.110263.
15. Lan L, Teng X, Zhang J, Zhang X, Tao D. Learning to purification for unsupervised person re-identification. IEEE Trans Image Process. 2023 Jan;32(34):3338–53. doi:10.1109/TIP.2023.3278860.
16. Cho Y, Kim WJ, Hong S, Yoon S-E. Part-based pseudo label refinement for unsupervised person re-identification. In: Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun. doi:10.1109/cvpr52688.2022.00716.
17. Ni H, Li Y, Gao L, Shen HT, Song J. Part-aware transformer for generalizable person re-identification. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV); 2023. doi:10.1109/iccv51070.2023.01036.
18. Zhang Z, Zhang H, Liu S, Xie Y, Durrani TS. Part-guided graph convolution networks for person re-identification. Pattern Recognit. 2021 Jun;120(3):108155. doi:10.1016/j.patcog.2021.108155.
19. He Q, Lu Z, Wang Z, Hu H. Graph-based progressive fusion network for multi-modality vehicle re-identification. IEEE Trans Intell Transp Syst. 2023 Jun;24(11):12431–47. doi:10.1109/TITS.2023.3285758.
20. Wang Q, Zhong Y, Min W, Zhao H, Gai D, Han Q. Dual similarity pre-training and domain difference encouragement learning for vehicle re-identification in the wild. Pattern Recognit. 2023 Jul;139(2):109513. doi:10.1016/j.patcog.2023.109513.
21. Ding Y, Fan H, Xu M, Yang Y. Adaptive exploration for unsupervised person re-identification. ACM Trans Multimed Comput Commun Appl. 2020 Feb;16(1):1–19. doi:10.1145/3369393.
22. Zhong Z, Zheng L, Luo Z, Li S, Yang Y. Invariance matters: Exemplar memory for domain adaptive person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2019. p. 598–607.
23. He K, Fan H, Wu Y, Xie S, Girshick R. Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2020. p. 9729–38.
24. Qiu M, Lu Y, Li X, Lu Q. Camera-aware differentiated clustering with focal contrastive learning for unsupervised vehicle re-identification. IEEE Trans Circuits Syst Video Technol. 2024 Jan;34(10):10121–34. doi:10.1109/TCSVT.2024.3402109.
25. Hu Z, Zhu C, He G. Hard-sample guided hybrid contrast learning for unsupervised person re-identification. In: 2021 7th IEEE International Conference on Network Intelligence and Digital Content (IC-NIDC); 2021. p. 91–5. doi:10.1109/ic-nidc54101.2021.9660560.
26. Yang J, Fang J, Xu H. ConMAE: Contour guided MAE for unsupervised vehicle re-identification. In: 2023 35th Chinese Control and Decision Conference (CCDC); 2023 May; Yichang, China. p. 4616–22. doi:10.1109/CCDC58219.2023.10327202.
27. Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. arXiv:1503.02531. 2015.
28. Ge Y, Chen D, Li H. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. In: Proceedings of the International Conference on Learning Representations; 2020.
29. Zheng D, Xiao J, Chen K, Huang X, Chen L, Zhao Y. Soft pseudo-label shrinkage for unsupervised domain adaptive person re-identification. Pattern Recognit. 2022 Feb;127(2):108615. doi:10.1016/j.patcog.2022.108615.
30. Shi J, Zheng S, Yin X, Lu Y, Xie Y, Qu Y. CLIP-guided federated learning on heterogeneity and long-tailed data. Proc AAAI Conf Artif Intell. 2024 Mar;38(13):14955–63. doi:10.1609/aaai.v38i13.29416.
31. Hermans A, Beyer L, Leibe B. In defense of the triplet loss for person re-identification. arXiv:1703.07737. 2017.
32. Ester M, Kriegel H -P, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining; 1996 Jan. p. 226–31.
33. Liu H, Tian Y, Wang Y, Pang L, Huang T. Deep relative distance learning: tell the difference between similar vehicles. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2016 Jun. p. 2167–75. doi:10.1109/cvpr.2016.238.

34. Lou Y, Bai Y, Liu J, Wang S, Duan L. VERI-Wild: A large dataset and a new method for vehicle re-identification in the wild. Comput Vis Pattern Recognit. 2019 Jun;3230–8. doi:10.1109/cvpr.2019.00335.
35. Zheng L, Shen L, Tian L, Wang S, Wang J, Tian Q. Scalable person re-identification: a benchmark. In: IEEE/CVF Conference on Computer Vision;. 2015 Dec. p. 1116–24. doi:10.1109/iccv.2015.133.
36. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2016 Jun. doi:10.1109/cvpr.2016.90.
37. Deng J, Dong W, Socher R, Li L -J, Li NK, Fei-Fei L. ImageNet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition; 2009 Jun. doi:10.1109/cvpr.2009.5206848.
38. Zhong Z, Zheng L, Kang G, Li S, Yang Y. Random erasing data augmentation. Proc AAAI Conf Artif Intell. 2020 Apr;34(7):13001–8. doi:10.1609/aaai.v34i07.7000.
39. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2016. p. 2818–26.
40. Zhang C-B, Jiang P-T, Hou Q, Wei Y, Han Q, Li Z, et al. Delving deep into label smoothing. IEEE Trans Image Process. 2021 Jan;30:5984–96. doi:10.1109/TIP.2021.3089942.
41. Zhang X, Ge Y, Qiao Y, Li H. Refining pseudo labels with clustering consensus over generations for unsupervised object re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2021 Jun. p. 3435–44. doi:10.1109/cvpr46437.2021.00344.
42. Pan X, Luo P, Shi J, Tang X. Two at once: enhancing learning and generalization capacities via IBN-Net. In: Proceedings of the European Conference on Computer Vision (ECCV); 2018. p. 484–500. doi:10.1007/978-3-030-01225-0_29.
43. Li M, Li C-G, Guo J. Cluster-guided asymmetric contrastive learning for unsupervised person re-identification. IEEE Trans Image Process. 2022 Jan;31:3606–17. doi:10.1109/TIP.2022.3173163.
44. Sun Y, Zheng L, Yang Y, Tian Q, Wang S. Beyond part models: person retrieval with refined part pooling (and a strong convolutional baseline). In: Proceedings of the European Conference on Computer Vision (ECCV); 2018. p. 480–96.
45. Zhou K, Yang Y, Cavallaro A, Xiang T. Omni-scale feature learning for person re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2019 Oct. p. 3701–11. doi:10.1109/iccv.2019.00380.
46. Huang G, Liu Z, van der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2017. p. 4700–8.
47. van der Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res. 2008 Jan;9(86):2579–605.
48. Wang H, Yang M, Liu J, Zheng W-S. Pseudo-label noise prevention, suppression and softening for unsupervised person re-identification. IEEE Trans Inf Forensics Secur. 2023 Jan;18:3222–37. doi:10.1109/TIFS.2023.3277694.
49. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. Int J Comput Vis. Oct. 2019;128(2):336–59. doi:10.1007/s11263-019-01228-7.