SCIENCE CHINA
Information Sciences

June 2022, Vol. 65 160103:1–160103:15
https://doi.org/10.1007/s11432-021-3383-y

• RESEARCH PAPER •
Special Focus on Deep Learning for Computer Vision

Progressive learning with multi-scale attention network for cross-domain vehicle re-identification

Yang WANG[1,2], Jinjia PENG[3,4], Huibing WANG[3] & Meng WANG[1,2*]

[1]*School of Computer Science and Information Engineering, Hefei University of Technology, Hefei* 230009*, China;*
[2]*Intelligent Interconnected Systems Laboratory of Anhui Province, Hefei University of Technology, Hefei* 230009*, China;*
[3]*Information Science and Technology College, Dalian Maritime University, Dalian* 116026*, China;*
[4]*School of Cyber Security and Computer, Hebei University, Baoding* 071002*, China*

Received 29 January 2021/Revised 3 July 2021/Accepted 19 November 2021/Published online 14 April 2022

Abstract Vehicle re-identification (reID) aims to identify vehicles across different cameras that have non-overlapping views. Most existing vehicle reID approaches train the reID model with well-labeled datasets via a supervised manner, which inevitably causes a severe drop in performance when tested in an unknown domain. Moreover, these supervised approaches require full annotations, which is limiting owing to the amount of unlabeled data. Therefore, with the aim of addressing the aforementioned problems, unsupervised vehicle reID models have attracted considerable attention. It always adopts domain adaptation to transfer discriminative information from supervised domains to unsupervised ones. In this paper, a novel progressive learning method with a multi-scale fusion network is proposed, named PLM, for vehicle reID in the unknown domain, which directly exploits inference from the available abundant data without any annotations. For PLM, a domain adaptation module is employed to smooth the domain bias, which generates images with similar data distribution to unlabeled target domain as "pseudo target samples". Furthermore, to better exploit the distinct features of vehicle images in the unknown domain, a multi-scale attention network is proposed to train the reID model with the "pseudo target samples" and unlabeled samples; this network embeds low-layer texture features with high-level semantic features to train the reID model. Moreover, a weighted label smoothing (WLS) loss is proposed, which considers the distance between samples and different clusters to balance the confidence of pseudo labels in the feature learning module. Extensive experiments are carried out to verify that our proposed PLM achieves excellent performance on several benchmark datasets.

Keywords data adaptation module, weighted label smoothing loss, multi-scale attention network, vehicle re-identification

Citation Wang Y, Peng J J, Wang H B, et al. Progressive learning with multi-scale attention network for cross-domain vehicle re-identification. Sci China Inf Sci, 2022, 65(6): 160103, https://doi.org/10.1007/s11432-021-3383-y

1 Introduction

Vehicle re-identification (reID) aims to search for and locate the same vehicle images from a variety of images captured by multiple non-overlapping cameras; this is an important issue for modern, smart surveillance systems. Moreover, through vehicle reID, the task of a search can automatically be conducted with less manual labor and less time, which is vitally significant for intelligent transport, particularly for deep learning models that rely on supervised approaches [1–4] for an ideal performance.

However, supervised reID models suffer from certain limitations; in different scenes, various illumination conditions, resolutions, backgrounds and viewpoints could cause domain bias. These limitations could result in the vehicle reID models, though well-trained under these supervised methods, performing poorly when directly deployed to the real-world large-scale camera networks. Moreover, full annotations are necessary for these supervised methods, for example, identity labels, which are labor intensive and impractical when annotating a large number of unlabeled data in real-world scenes when there are many cameras. In particular, for vehicle reID, the same vehicle is required to be annotated under all cameras. Hence, the incremental optimization of vehicle reID algorithms utilizing a combination of the abundant unlabeled data and existing well-labeled data is a practical but challenging problem.

* Corresponding author (email: eric.mengwang@gmail.com)

To tackle the aforementioned problem, various strategies [5, 6] have been proposed to finish the task of domain adaptation. These can be divided into two categories: cross-domain unsupervised transfer learning and progressive learning. Cross-domain unsupervised transfer learning employs a generative adversarial network to transfer labeled images from the source domain to the unlabeled domain [7–9], together with the transfer of discriminative information to train the unsupervised reID model. Though some improvements have been made, the learned style remains different from the authentic one, which limits its performance. Conversely, progressive learning trains the reID model while estimating the pseudo labels for unlabeled images until it converges after several iterations [10, 11]. Assigning reliable labels for unlabeled images is of vital importance, but is a challenging process. Although some approaches, such as VR-PROUD [12], adopt feature clustering to calculate the pseudo labels, incorrect annotations are inevitable, which causes severe adverse effects on unsupervised reID models.

In this paper, we propose a novel unsupervised domain adaptation method together with a weighted label smoothing (WLS) based loss to better exploit the unlabeled data. This method progressively adapts the unknown domain for vehicle reID. Considering that there are always some existing well-labeled samples, it is better to make full use of these data rather than abandoning them. Hence, unlike the existing unsupervised vehicle reID methods, an adaptation module is employed to generate "pseudo target images", which learns the style of the unlabeled domain while preserving the identity cues of the labeled domain. Owing to the uncertainty of the classes in the unknown domain, DBSCAN is employed to cluster the unlabeled samples and assign the pseudo labels. Moreover, after clustering and selecting reliable pseudo-labeled data from large, unlabeled samples, fusion data that combine the "pseudo target images" and reliable pseudo-labeled data are employed to train the vehicle reID model in the subsequent processes.

In most approaches, the feature maps generated by the last convolutional layer carry high-level semantic information, which is employed for visual tasks. However, feature maps from the intermediate layers also have abundant texture cues that contain important details for vehicle reID. To this end, we propose a multi-scale attention network for learning latent features with different cues from various scales. In particular, to better adapt the unknown domain, WLS loss is presented to construct the pseudo label distribution as a weighted distribution over all clusters, which effectively regularizes the network to the target training data distribution. The major framework is illustrated in Figure 1.

In our previous work [13], a progressive learning method was proposed for unsupervised domain adaptation vehicle reID. As an extension to [13], our new method exploits the potential similarity of various samples in unknown domains in a progressive manner. In contrast to [13], we propose a multi-scale attention network to learn rich features for clustering in the feature learning step. Owing to a lack of labels in the unknown domain, the extracted rich semantic features are beneficial in distinguishing various vehicles. As the depth of the network increases, some important cues gradually disappear. Therefore, the attention-based multi-scale network is proposed, which constructs the attention structure in different layers in order to achieve features with discriminative information, combining the progressive learning framework in [13] with our proposed feature multi-scale fusion network, named PLM, to exploit generalized cues to adapt to unknown domains. More visualized results than [13] are offered over benchmark datasets.

Beyond [13], our contributions are summarized as follows.

• A novel progressive learning method is proposed for vehicle reID to better adapt to an unknown domain; the method iteratively updates the model using a WLS-based multi-scale attention learning network while adopting the clustering approach to assign labels with various weights for selected reliable unlabeled data.

• Unlike [13], by learning feature extracting model with ResNet only, a multi-scale attention network is developed to integrate features from multiple layers for training the reID model, including low-layer texture features and high-level semantic features. Moreover, considering the distances between the samples and different clusters, we propose the WLS loss to balance the confidence of pseudo labels for improved performance.

• To make full use of the labeled data, PLM employs a data adaptation module based on the generative adversarial network to generate images, labeled as the "pseudo target samples". These are then combined with the selected samples from the unlabeled domain for training the model.

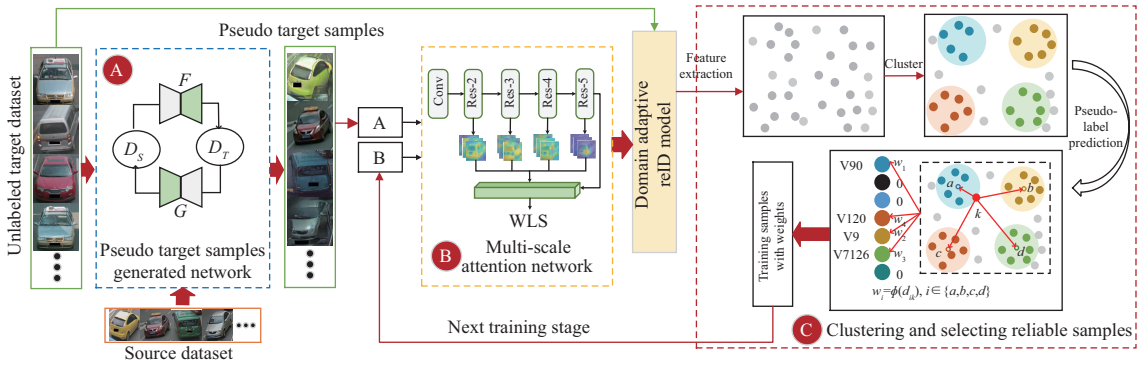

Figure 1 (Color online) Illustration of our proposed PLM framework. The images are transferred from the source domain to the target domain to generate pseudo target samples. Iterations then commence. During each iteration, we (1) train the reID model using the proposed WLS-based multi-scale attention network, which utilizes the fusion data and combines the pseudo target samples with selected samples, and (2) assign pseudo labels with various weights for unlabeled images and select reliable samples according to a dynamic sampling strategy.

2 Related work

This paper considers the efficiency and accuracy of unsupervised vehicle reID in an unknown domain. In this section, we briefly review the methods of supervised reID methods and unsupervised reID methods.

2.1 Supervised vehicle reID methods

There are many supervised vehicle reID methods; they can be divided into four categories. The first is structure-based reID methods, which exploit various deep learning networks to tap potential discriminative cues, such as VAMI [14] or DHMVI [15]. According to the characteristics of vehicles, these methods learn multi-view features by solely using sigle-view features to enhance the representation of vehicles. In addition to the viewpoints cues, attributes such as models and colors enable more distinctive features to be learned in MGR [16]. There are also some methods that try to locate details using detection or segmentation approaches in the vehicle images as the local features, such as PGANet [17]. The second category includes approaches that focus on designing metric losses to better optimize the training stage, such as CCL loss [18] and GST loss [19]. Taking the group into consideration, GST loss aims to minimize the intra-class variance to achieve better performance. The third category considers that spatial and temporal information are important cues to improve vehicle reID performance. The most popular solution is to construct the spatial-temporal model to regularize the vehicle reID results. In PROVID [20] and OIFE [21], the spatial-temporal cue is employed as the auxiliary information to refine the ranking results obtained by other feature learning models. In contrast to these methods, the complex model based on spatial-temporal information is exploited in [22], which mainly employs the Markov model to construct the relationship between different samples. The final category includes the popular application of generative adversarial networks (GAN) [23,24] in vision tasks, which some methods propose for vehicle reID, such as CV-GAN [25] and EALN [2]. Various images with multiple viewpoints are generated in CV-GAN [25] to construct the complete representation for vehicle images. Through EALN [2], hard negative samples could be generated automatically to improve the discriminative ability of the vehicle reID model, especially for vehicles with similar appearances.

However, supervised reID models are not effective in observing real-world scenes with considerable unlabeled images. Therefore, we propose a novel and effective unsupervised vehicle reID model.

2.2 Domain adaptation methods for reID

Unsupervised vehicle reID means that there is no label information in the target dataset. Though some methods have been exploited in person reID, only a few of these methods explore unsupervised reID. VR-PROUD [12] presents a self-paced progressive unsupervised learning architecture that adopts the unsupervised K-means clustering to infer the vehicle IDs in a semi-unsupervised manner. PUL [11] proposes to train the reID model of the unlabeled domain by iterating the clustering and fine-tuning. Similar to PUL, in [10], a self-training scheme is presented to assign labels for unlabeled target samples with an encoder. These labels are utilized for training the encoder of the vehicle reID model.

It is obvious that domain adaptation has been widely utilized for vehicle reID. However, domain adaptation has multiple limitations, which presents the motivation of our proposed method. Hence, this paper provides a feasible model by leveraging domain adaptation into progressive learning to solve the unsupervised vehicle reID problem.

3 Progressive learning with a multi-scale attention network

3.1 Framework overview

The structure of PLM is shown in Figure 1. For unsupervised vehicle reID, domain adaptation is necessary to transfer discriminative information between domains. Therefore, a domain adaptation module based on GAN is trained to transfer the well-labeled images to the unlabeled target domain, as in part A of Figure 1, which could decrease the domain bias and make full use of existing source domain images. Then, as in part B of Figure 1, the generated images are employed as the "pseudo target samples" to be input for the proposed multi-scale network for feature learning, which adapts the target domain progressively. When the model is trained, WLS loss is proposed to balance the confidence of unlabeled samples and different clusters, which exploits pseudo labels with different weights according to the ability of the model trained by the last iteration. Next, the output features of the reID model are employed to select reliable samples using a dynamic sampling strategy, by utilizing the various clustering results and selection strategies, as in part C of Figure 1. Lastly, the "pseudo target samples" with accurate labels and the selected samples from the unlabeled domain with pseudo labels are combined to be the training sets for the next iteration. In this way, a more stable adaptive model can be learned progressively.

3.2 Domain adaptation module

Domain adaptation is a pivotal part of unsupervised vehicle reID. With the intervention of domain adaptation, an unsupervised reID model can obtain discriminative information transferred from the labeled source domain. As in part A of Figure 1, there are two types of vehicle images for the task of unsupervised vehicle reID: well-labeled images in the source domain and unlabeled images from the target domain. However, although there are some well-labeled samples, directly applying them to the target domain may result in a poor performance due to the domain bias. Moreover, for the target domain, the supervised learning approaches are limited by the unlabeled samples, which cannot be utilized to train the reID model. Hence, CycleGAN [26, 27] is employed as a data adaptation module to make full use of these well-labeled data. This generates "pseudo target samples", which decreases the domain bias by transferring the style between the source domain and target domain. Noteably, although "pseudo target samples" with labels are employed, we do not utilize any labeled data from the target domain.

CycleGAN introduces two generator-discriminator pairs, (G, D_T) and (F, D_S), which map a sample from a source (target) domain to a target (source) domain and generate a sample, which is indistinguishable from those in the target (source) domain [28]. For our method, besides the traditional adversarial losses and cycle-consistent loss in CycleGAN, a content loss [29] is utilized to maintain the label information from the source domain, which is formulated as follows:

$$L_{\mathrm{id}}(G, F, X, Y) = E_{y\sim p_{\mathrm{data}}(y)}||F(y) - y||_1 + E_{x\sim p_{\mathrm{data}}(x)}||G(x) - x||_1, \tag{1}$$

where X and Y represent the source domain and target domain, respectively, and $p_{\mathrm{data}}(x)$ and $p_{\mathrm{data}}(y)$ denote the sample distributions in the source and target domain, respectively. Through the generated network, we can make full use of the well-labeled data. There are two reasons. The first is that through the CycleGAN, the generated "pseudo target samples" have a similar distribution as the target domain, which reduces the bias between the source and target domains. The second reason is that the identity information of the source domain is also preserved by turning the content loss during the transferring phrase, which means that the well labeled annotations could be reused in the subsequence.

3.3 Multi-scale attention network

Training the reID model with a feature learning network plays a vital role in the PLM, which trains the model by combining the generated "pseudo target images" with the selected pseudo labeled samples. Owing to the lack of labels in the unknown domain, more distinctive and strong features from the deep

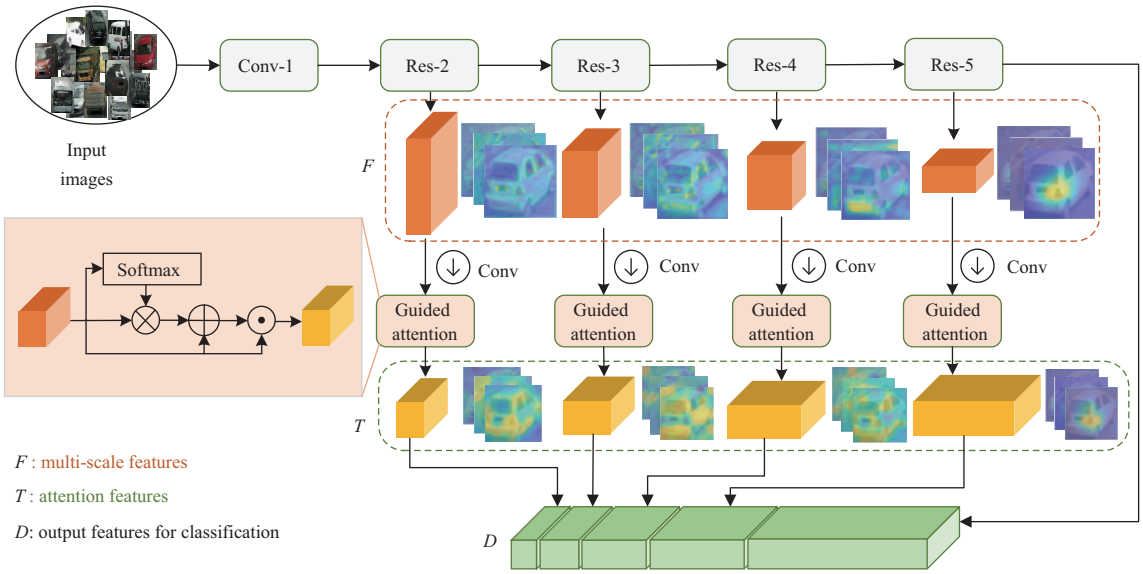

Figure 2 (Color online) Structure of the multi-scale attention network. Based on the original ResNet [30], several sub-branches are added to seek attentive features from high layers and low layers. $\otimes$ is performed in an element-wise product. $\odot$ represents the concatenate operation. $\oplus$ is performed in an element-wise sum.

learning network are beneficial to train the vehicle reID model. Thus, in contrast to [13], a multi-scale attention network is proposed in this subsection (part B of Figure 1).

Visual cues are a vital component in vehicle reID. In a typical network, visual cues optimize the convolutional neural networks for achieving deep features from the last layer. However, the texture of different scales or other discriminative cues from low layers are neglected. Moreover, with the increase of the network's depth, these features can easily and gradually disappear. Therefore, in our paper, the multi-scale fusion network is introduced to integrate features from multiple layers, which contain low-layer texture features and high-level semantic features with various scales. The features from intermediate layers are redundant, as they may cause some noise for the vehicle reID task. Therefore, a self-attention structure is added into our network to select the most relevant piece of information for visual analysis.

The structure of the proposed multi-scale attention network is illustrated in Figure 2. Based on ResNet [30], 4 branches are added to generate multi-scale feature maps, which follow Block2, Block3, Block4 and Block5, as shown in Figure 2. In our network, in each branch, to reduce the dimension of features and extract effective features, several additional layers are utilized, which contain ReLU layers and convolutional layers. From Figure 2, it can be seen that it is significant that the features with different scales from multiple layers concentrate on different regions, which are important cues for vehicle reID. For the unlabeled domain, pseudo labels are achieved by clustering, which is inaccurate. Hence, more distinct features could partly make up for the deficiency. Moreover, considering a large amount of redundant information from different layers, we can build a soft attention mechanism to extract attentive features while enhancing the discriminative information.

Specifically, for each branch, after several convolutional layers and ReLU layers, the set of the attention mask M corresponding to each branch can be computed as follows:

$$M_c^{i,j} = \mathrm{Softmax}(F_c^{i,j}), \tag{2}$$

where $M_c^{i,j}$ and $F_c^{i,j}$ are the values for the corresponding features maps at pixel location (i, j) of the cth channel. We use Softmax to generate the weights of pixels in $F_c^{i,j}$, which is similar to the ranking of importance for every pixel. Softmax is calculated by $\frac{\exp(F_c^{i,j})}{\sum \exp(F_c^{i,j})}$. M is used to re-weight the pixels to produce the attentive features as follows:

$$T_c^{i,j} = [F_c^{i,j}, F_c^{i,j} + F_c^{i,j} \otimes M_c^{i,j}], \tag{3}$$

where $T_c^{i,j}$ is the attentive feature of each sub-branch. Owing to the spatial misalignment of some of the images in the training set, the mask M could be imprecise, which could subsequently result in the

attentive feature being disturbed by noise. Hence, in our paper, a shortcut connection architecture is utilized to fuse the features from the low layer with the attended features.

After achieving attentive features from different branches, a concatenate layer is employed to fuse these features in order to exploit latent information from these features adaptively. This is described as follows:

$$D_c = [T_c^2, T_c^3, T_c^4, T_c^5, F_c^5], \tag{4}$$

where T_c^2, T_c^3, T_c^4 and T_c^5 represent the output features from each of the sub-branches, respectively. D_c is the final feature for optimizing the vehicle reID model with the loss function.

3.4 Weighted label smoothing loss

Through the multi-scale attention network, rich features could be obtained to train the reID model. To match up with these abundant features, reasonable labels should be considered in the unknown domain in order to better optimize the reID model. Therefore, a novel weighted label smoothing loss is proposed in this subsection.

For the "pseudo target images", it is trivial to obtain the label information. However, assigning reasonable labels for the pseudo labeled samples is a big challenge. In the unlabeled domain, if the clustering centroids are regarded as the pseudo labels directly, ambiguous prediction in training owing to inaccurate clustering results could occur. Moreover, it is improper to assign the same labels for all the samples regardless of the distance to the clustering centroids.

Hence, WLS loss is presented to set the pseudo label distribution as a weighted distribution over all clusters, which effectively regularizes the feature learning network to the target training data distribution. Each generated sample that is well-labeled in the training set is assigned with only one ground-truth label, which can be formulated as follows:

$$q_k(g) = \begin{cases} 0, & \text{if } k \neq y, \\ 1, & \text{if } k = y, \end{cases} \tag{5}$$

where y is the ground-truth class label of g. However, through the aforementioned analysis, it is not suitable to the unknown domain with pseudo labels. Hence, according to the result of clustering, we model the virtual label distribution as a weighted distribution over all clusters for unlabeled data, according to the distance between the features and each centroid of clusters.

Thus, the weights over all clusters are different in WLS loss. In this way, a dictionary α is constructed to record the weights. For an image g, the weights of the label can be calculated as

$$w_k^g = \frac{1}{K}\alpha_k^g, \quad k \in [1, K], \tag{6}$$

where α_k^g represents the weight of the image g over the kth cluster. To obtain α_k^g, unlabeled samples are clustered to obtain the centroids set $C = \{c_1, c_2, \ldots, c_k\}, k \in [1, K]$, which is introduced in Subsection 3.5. K is the number of clusters; the similarity between g and c_k can be calculated as $d_k^g = ||f_g - f_{c_k}||_2$, where f represents the feature of images or centroids. The set of distance of image g over K centroids can be described as $d^g = \{d_1^g, d_2^g, \ldots, d_k^g\}, k \in [1, K]$. Inspired by [31], all elements in d^g are sorted with descending order, and saved to ds^g. α_k is obtained by taking the corresponding index of ds_k^g in the set of ds^g:

$$\alpha_k^g = \left(1 - \frac{d_k^g}{\max(d^g)}\right) \cdot \psi_{ds^g}(d_k^g), \tag{7}$$

where $\psi_{ds^g}(\cdot)$ is the index of d_k^g in ds^g. Thus, the corresponding relationship between images and cluster centroids is constructed with different weights. Hence, the WLS loss of unlabeled data ℓ_{wls} can be formulated as follows:

$$\ell_{\text{wls}} = -\sum_{k=1}^{K} w_k \log(p(k)). \tag{8}$$

In addition to the real unlabeled samples, there are some generated images by CycleGAN that are combined to train the reID model. The training loss is defined as follows:

$$\ell = -(1-\sigma) \cdot \log(p(y)) - \sigma \cdot \lambda \cdot \sum_{k=1}^{K} w_k \log(p(k)). \tag{9}$$

For a generated image $\lambda = 0$, the loss is equivalent to the cross-entropy loss and y is the label of the generated image. When $\lambda = 1$, the image is from the unlabeled data and belongs to the cluster y. For the unlabeled data, σ is a smoothing factor between cross-entropy loss and WLS loss.

3.5 Clustering and selecting reliable samples

Appropriate candidates are crucial to exploit the unlabeled domain. When the model is weak, a small reliable measure is set, which is near the cluster centroids in the feature space. As the model becomes stronger in subsequent iterations, various instances should be adaptively selected as the training examples. Hence, a dynamic sampling strategy is proposed to ensure the reliability of the selected pseudo-labeled samples, as in part C of Figure 1. Images in the target domain are processed by the well-trained reID model to output features with high dimensions. Most methods select the K-means to generate clusters, which need to be initialized by the cluster centroids. However, it is uncertain how many categories are required in the target domain. Hence, DBSCAN is selected as the clustering method. Instead of employing the fixed clustering radius, this paper employs a dynamic cluster radius rad that is calculated by K-nearest neighbors (KNN). After implementing DBSCAN, in order to filter noise, some of the top reliable samples are selected to be assigned with soft labels, according to the distance between the features of the samples and cluster centroids. In our method, samples with $||f_g, c_{f_g}||_2 < \gamma$ are satisfied for the next iteration for training the model, where f_g is the feature of the gth image and c_{f_g} is the feature of the cluster centroid to which f_g belongs. Our method is summarized in Algorithm 1.

Algorithm 1 PLM for cross-domain vehicle reID

Require: Number of images on the target domain N, labeled source domain S, unlabeled target domain T, iteration number M, cluster number K, reliability threshold γ, new training set D;
Ensure: An encoder E for target domain;
Transfer style from S to T by GAN to generate pseudo target images ST;
Initialization $E^{(0)}$ with ST, D: $D = ST$;
for i=1 to M **do**
  Train $E^{(i)}$ with D utilizing WLS-based multi-scale attention network; compute $ft = E^{(i)}(D)$;
  Reduce dimension by manifold learning $f = \text{mad}(ft)$; calculate number and centroids of clusters: $(K, C) = \text{DBSCAN}(f)$;
  Select features of centroids: $\{c_k\}_{k=1}^{K} \rightarrow \{f_{c_k}\}_{k=1}^{K}$;
  D= ST;
  for k=1 to K **do**
    for $g = 1$ to N **do**
      if $||f_g, f_{c_k}||_2 > \gamma$ **then**
        $D = D \cup T_g$;
        Calculate weights w_k^g by (2)–(4);
      end if
    end for
  end for
end for

4 Experiments

4.1 Datasets and evaluation metrics

In this subsection, detailed analyses are conducted to verify the effectiveness of the proposed PLM. Additionally, the cumulative match characteristic (CMC) curve and the mean average precision (mAP) are utilized to evaluate all methods in our experiments. In this subsection, in addition to the comparison with state-of-the-art approaches, each part of PLM is analyzed in detail in the ablation studies. Our experiments are trained and tested on two benchmark datasets: VeRi-776 and VehicleID.

• VeRi-776 [20] is a vehicle dataset that contains over 50000 images of 776 vehicles. In addition to identity annotations, it also has information about the vehicle's color and type. The training set owns 37781 images of 576 vehicles and the test set has 11579 images of 200 vehicles. The query set is a subset of 1678 images from the test set.

• VehicleID [18] is a large vehicle dataset captured from multiple non-overlapping cameras. It contains 221763 images of 26267 vehicles. In our experiments, there are four subsets for test sets, which contain 800, 1600, 2400 and 3200 vehicles, respectively. During the testing stage, the gallery set is generated from test sets. The probe set is generated with the remaining images after selecting one image of one identity.

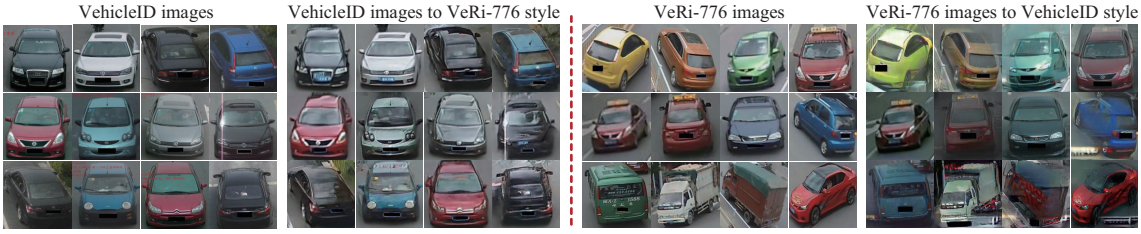

Figure 3 (Color online) Sample images of (a) VehicleID dataset, (b) VehicleID images which are translated to VeRi-776 style, (c) VeRi-776 dataset, and (d) VeRi-776 images translated to VehicleID style.

Moreover, the generated "pseudo target samples" from the above 2 datasets are also included in the training of PLM. The generated samples have a similar distribution as the target domain, which reduces the bias between the source and target domains, as shown in Figure 3.

4.2 Implementation details

In the data adaptation module, the tensorflow [32] is utilized for the training platform to train the translated model. During the training procedure, the learning rate is set to 0.0002, the min-batch size is 16, and epoch is 6. For feature learning, the proposed multi-scale network is employed as the backbone network. The vehicle reID model is trained in the Matconvnet [33]. We utilize SGD to optimize the training procedure with a momentum of $\mu = 0.0005$. The batch size and the iteration are set to 64 and 6, respectively.

For PLM, images are transferred by CycleGAN from the source domain to the target domain, which is used as initial samples for training the feature learning model. Considering the limit of the device, when training the reID model on VeRi-776, 10000 transferred images from VehicleID are utilized as the "pseudo target images". On VehicleID, the same implementations are conducted to train the reID model. When training the unsupervised model on VehicleID, only 35000 images from the VehicleID are selected for the training set.

4.3 Comparison with the state-of-the-art methods

To better validate the effectiveness of PLM, some existing methods are compared with our proposed method. Additionally, the results of the comparison between PLM and other state-of-the-art methods are reported in Tables 1 and 2 and Figure 4. The methods to be compared with PLM are (1) FACT [34]; (2) FACT+Plate-SNN+STR [34]; (3) Mixed Diff+CCL [18]; (4) VR-PROUD [12]; (5) CycleGAN [26], a method of style transfer, which is employed for the domain adaptation; (6) ECN [35]; (7) UDAR [10]; (8) direct transfer, which directly employs the well-trained reID model of [36] on the source domain to the target domain; (9) baseline system, which, compared with the framework of PLM, utilizes the original samples from the source domain instead of the generated data and only trains the reID model with cross-entropy (CE) loss; (10) PUL [11]; (11) PAL [13].

Methods (1)–(3) are supervised vehicle reID approaches, whereas the remaining methods are unsupervised methods. In particular, the PUL is an unsupervised adaptation method of person reID. As there are only a few studies focused on unsupervised vehicle reID, PUL is also compared with our proposed PLM. There are some other methods that are similar to PUL; however, most of them require special annotations, such as labels for segmenting or detecting keypoints, which are not annotated in the existing vehicle reID datasets. From Tables 1 and 2, we note that the proposed PLM achieves the best performance among all approaches, with Rank-1 = 77.59%, mAP = 47.37% on VeRi-776, Rank-1 = 51.23%, 45.40%, 41.73%, 39.25%, mAP = 54.85%, 49.41%, 46.00%, 43.46% on VehicleID with the test size of 800, 1600, 2400, 3200, respectively.

Compared with PAL, PLM also performs better for both VeRi-776 and VehicleID. In particular, in the VeRi-776 dataset, there are 6.35% and 9.42% improvements in mAP and Rank-1, respectively. The main reason for this is that the proposed attention-based multi-scale network learns more distinct features that integrate high-level semantic features with low-layer texture features, which are stronger for vehicle reID than the feature from ResNet. This also validates that the excellent feature learning network is conductive to exploit the unknown domain.

Table 1 Performance of different methods on VeRi-776. The best results are shown in bold face. PLM achieves the best performance.

Method	mAP (%)	Rank-1 (%)	Rank-5 (%)
FACT [34]	18.75	52.21	72.88
FACT+Plate-SNN+STR [34]	27.77	61.44	78.78
VR-PROUD [12]	22.75	55.78	70.02
PUL [11]	17.06	55.24	66.27
CycleGAN [26]	21.82	55.42	67.34
ECN [35]	20.06	57.41	70.53
UDAR [10]	35.8	76.9	85.8
Direct transfer	19.39	56.14	68.00
Baseline system	31.94	58.58	73.24
PAL [13]	42.04	68.17	79.91
PLM	**47.37**	**77.59**	**87.00**

Table 2 Performance of various methods over different reID methods on VehicleID. The best results are shown in bold face. PLM can achieve the best performance in most situations. Mixed Diff+CCL can also achieve a good performance.

Method	Test size = 800 (%)			Test size = 1600 (%)			Test size = 2400 (%)			Test size = 3200 (%)		
	mAP	Rank-1	Rank-5	mAP	Rank-1	Rank-5	mAP	Rank-1	Rank-5	mAP	Rank-1	Rank-5
FACT [34]	–	49.53	67.96	–	44.63	64.19	–	39.91	60.49	–	–	–
Mixed Diff+CCL [18]	–	49.00	**73.50**	–	42.80	**66.80**	–	38.20	**61.60**	–	–	–
PUL [11]	43.90	40.03	56.03	37.68	33.83	49.72	34.71	30.90	47.18	32.44	28.86	43.41
CycleGAN [26]	42.32	37.29	58.56	34.92	30.00	49.96	31.89	27.15	46.52	29.17	24.83	42.17
Direct transfer	40.58	35.48	57.26	33.59	28.86	48.34	30.50	26.08	44.02	27.90	23.85	39.76
Baseline system	42.96	39.11	55.24	38.03	34.04	50.91	34.04	30.10	48.41	31.98	28.24	43.77
PAL [13]	53.50	50.25	64.91	48.05	44.25	60.95	45.14	41.08	59.12	42.13	38.19	55.32
PLM	**54.85**	**51.23**	67.11	**49.41**	**45.40**	63.37	**46.00**	**41.73**	60.94	**43.46**	**39.25**	**57.99**

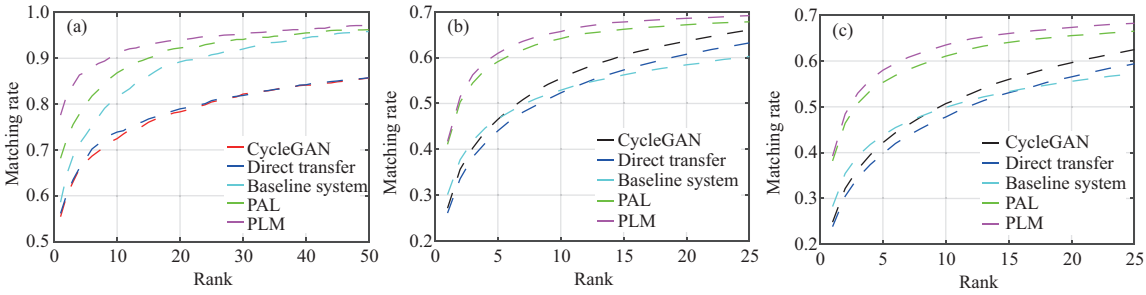

Figure 4 (Color online) CMC curves of several typical methods on VeRi-776 and VehicleID. The proposed PLM outperforms other compared methods, especially the "CycleGAN" and the "direct transfer' method. This demonstrates that progressive learning could increase the adaptive ability for the reID model in the unlabeled target domain. (a) VeRi-776; (b) VehicleID (test size = 2400); (c) VehicleID (test size = 3200).

Compared with PUL [11] and VR-PROUD [12], PLM has 30.31% and 24.62% gains in mAP on VeRi-776, respectively. Our model also outperforms the PUL and VR-PROUD in Rank-1, Rank-5 and mAP on VehicleID. For these methods, the K-means is employed to assign pseudo-labels for unlabeled samples. Owing to the uncertainty over the number of categories, it is not appropriate to utilize the K-means in reID. Compared with UDAR [10], a clustering-based domain adaptation reID method whose structure is similar to our method, our method again achieves better results on VeRi-776, with a 11.57% improvement in mAP. In addition, compared with the "direct transfer" method, our proposed PLM achieves 27.98% and 21.45% gains in mAP and Rank-1 on VeRi-776, which is significant. Our method also demonstrates similar improvements on VehicleID. Furthermore, compared with the supervised approaches, such as FACT [34], Mixed Diff+CCL [18] and FACT+Plate-SNN+STR [34], PLM performs better on VeRi-776 and VehicleID, validating that PLM is more adaptive to different domains.

Compared with CycleGAN [26], which adapts the domain bias by style transfer, our method significantly outperforms on both VeRi-776 and VehicleID. For VeRi-776, our proposed PLM achieves 25.55% and 22.17% improvements in mAP and Rank-1, respectively. Similarly, our method has 13.94%, 15.40%,

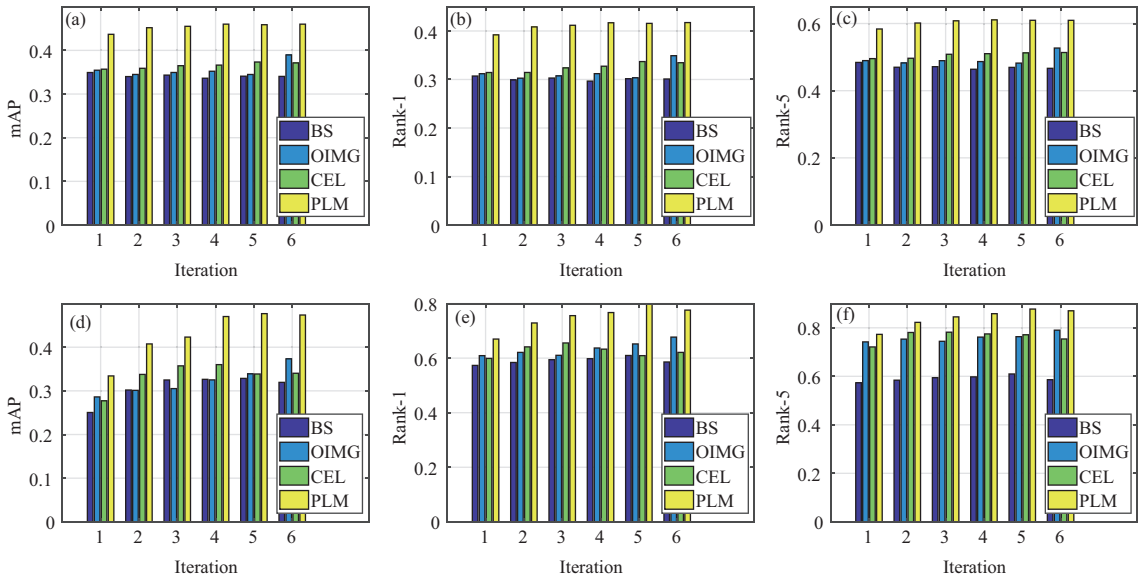

Figure 5 (Color online) Comparison results. (a)–(c) are the mAP, Rank-1 and Rank-5 of four comparison methods on VeRi-776; (d)–(f) are the results of four comparison methods in every iteration on VehicleID.

Table 3 Settings for different ablation models

Method	Generated images	Original images	WLS	CE
BS	×	√	×	√
CEL	√	×	×	√
OIMG	×	√	√	×
PLM	√	×	√	×

14.58% and 14.42% gains in Rank-1 on VehicleID with the test sets of 800, 1600, 2400 and 3200, respectively. These significant improvements are mainly due to the fact that PLM exploits the similarity among unlabeled samples through iterations for unsupervised vehicle reID. Though the generated images have the style of the target domain, they merely serve as pseudo samples. The real samples in the target domain could be more reliable in generating the discriminative features during the training stage. These results suggest that reliable samples in the target domain are an important component for unsupervised reID, which indicates that PLM could make full use of the unlabeled samples in the target domain. Compared with ECN [35], which employs exemplar memory to assign soft labels for unlabeled samples, our methods also have 27.31% and 20.18% improvements in mAP and Rank-1 on VeRi-776 dataset. Though this method solves the problem of assigning labels, it has a lot of noisy labels in the training set, which could cause errors in the training stage.

Compared with the baseline system, PLM has large improvements both on VeRi-776 and VehicleID. The PLM achieves 15.43% increase in mAP on VeRi-776, and 11.89%, 11.38%, 11.96%, 11.48% improvements in mAP on VehicleID with different test sets. These indicate that the "pseudo target images" and "weighted label smoothing" are two core components in PLM. This leads the reID model trained by our method to be more robust to different domains. More details will be discussed in Subsection 4.4.

4.4 Ablation studies

To better analyze PLM, we conducted ablation studies on two major components of PLM: the data adaptation module and WLS, which are both shown in Figure 5. The settings are depicted in Table 3. In addition to the backbone, all of them share a similar structure to PLM. For a fair comparison, the backbone of the feature learning network is ResNet50 [30] when training the reID model in these ablation studies. When the target domain is the VehicleIS dataset, 35000 images are randomly selected from the VehicleID dataset to be employed to train the reID model. "Generated images" indicates that the transferred images from the source domain and the image of the target domain are employed to train the models, while "Original images" indicates that the original images from the source domain and samples from the target domain are utilized for unsupervised vehicle reID. WLS and CE indicate that the WLS

Table 4 Performance comparison between CEL, OIMG and BS on VeRi-776

Iteration	CEL (%)			OIMG (%)			BS (%)		
	mAP	Rank-1	Rank-5	mAP	Rank-1	Rank-5	mAP	Rank-1	Rank-5
Iter1	27.71	59.89	72.10	28.61	60.90	74.19	25.04	57.33	71.33
Iter2	33.76	64.12	78.06	30.11	62.09	75.32	30.19	58.40	73.53
Iter3	35.73	65.55	78.18	30.52	61.02	74.43	32.49	59.41	73.06
Iter4	36.01	63.28	77.47	32.51	63.70	76.16	32.63	59.77	74.07
Iter5	33.86	60.90	77.11	33.90	65.19	76.34	32.86	60.96	74.91
Iter6	34.03	62.09	75.38	37.33	67.69	79.02	31.94	58.58	73.24

Table 5 Performance comparison between CEL, OIMG and BS on VehicleID (2400)

Iteration	CEL (%)			OIMG (%)			BS (%)		
	mAP	Rank-1	Rank-5	mAP	Rank-1	Rank-5	mAP	Rank-1	Rank-5
Iter1	35.69	31.43	49.54	35.45	31.19	48.95	34.93	30.71	48.41
Iter2	35.89	31.44	49.65	34.48	30.26	48.25	34.00	29.90	46.98
Iter3	36.49	32.39	50.84	34.94	30.74	48.93	34.33	30.29	47.14
Iter4	36.62	32.73	51.01	35.22	31.20	48.66	33.62	29.67	46.38
Iter5	37.33	33.69	51.25	34.48	30.35	48.21	34.08	30.15	46.94
Iter6	37.12	33.45	51.36	38.95	34.90	52.69	34.04	30.10	46.94

loss and cross-entropy loss are employed to train reID models, respectively. Figure 5 shows that PLM achieves the best performance on both datasets, demonstrating that the data adaptation module and WLS are effective in adapting to an unlabeled domain.

Effectiveness of the data adaptation module. To demonstrate the effectiveness of the generated samples, BS and CEL are compared, and the results are reported in Tables 4 and 5. For CEL, CycleGAN is employed to translate some of the labeled images from the source domain to the target domain, and these generated images are regarded as the "pseudo target samples". Then, the "pseudo target samples" are combined with the images in the target domain in order to train the reID model. Examples of translated images by CycleGAN are shown in Figure 3. It can be seen that the generated images are able to learn some styles from different domains for VehicleID, such as low illumination or high resolution. Both CEL and BS are trained by cross-entropy loss. From Tables 4 and 5, according to the last iteration, compared with BS, the mAP of CEL increases by 2.09% for VeRi-776. Additionally, it rises to 37.12% and 33.45% in mAP and Rank-1 on VehicleID, demonstrating that some important latent style information is learned by the generated images from the target domain, which could subsequently smooth the domain bias and achieve a better performance.

Effectiveness of WLS. A comparison between BS and OIMG is conducted to validate the effectiveness of the WLS. Tables 4 and 5 show the comparisons for VeRi-776 and VehicleID. Our proposed WLS achieves a better performance than cross-entropy loss. The last iteration for OIMG, compared with BS, has an increased accuracy of 5.39% and 9.11% on VeRi-776 for mAP and Rank-1, respectively. The similar conclusion holds for VehicleID, which achieves 4.91% and 4.8% increases for mAP and Rank-1, respectively. These results indicate that the WLS loss has a better ability to achieve discriminative representation during the training stage. Additionally, Tables 4 and 5 show that the accuracy of CEL changes slowly during iterating, which is significant. The reason for this is that the pseudo labels assigned by clustering are inaccurate, which could cause ambiguous predictions during the training phase. Therefore, it is not desirable to use the clustering results for the pseudo labels of unlabeled data. In our method, the WLS loss sets the pseudo label distribution as a weighted distribution over all clusters, which effectively regularizes the feature learning network to the target training data distribution. This demonstrates that WLS is more suitable for an unknown domain in an unsupervised setting.

Effectiveness of the multi-scale network. A comparison between PAL and PLM is conducted to validate the effectiveness of the multi-scale network. Figure 6 shows the comparisons for VeRi-776 and VehicleID. In particular, compared with PAL, PLM achieves significant improvements for every iteration. This is because the multi-scale network in PLM can generate abundant features that integrate the high semantic information into low texture cues, demonstrating that features from the multi-scale network are more distinctive and beneficial for vehicle reID in the unknown domain. Additionally, to verify the effectiveness of the soft attention mechanism in the multi-scale network, we removed the structure

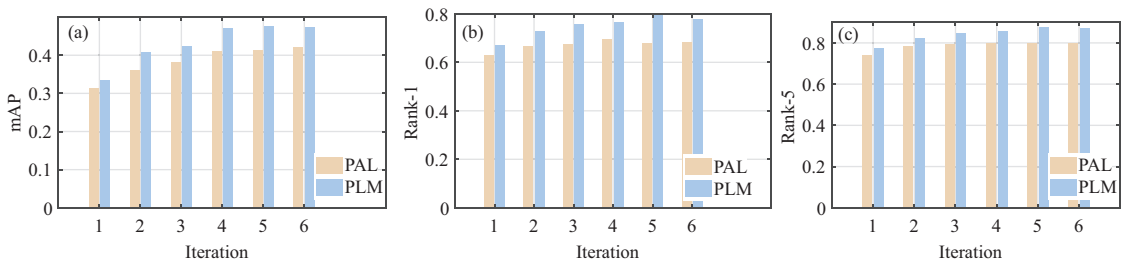

Figure 6 (Color online) Comparison results. (a)–(c) are the mAP, Rank-1 and Rank-5 of four comparison methods on VeRi-776 in different iterations, respectively.

Table 6 Results for different structures on VeRi-776

Method	mAP (%)	Rank-1 (%)	Rank-5 (%)	Rank-10 (%)
PLM-Sum	53.19	84.02	91.59	94.33
PLM-Concat-noAtt	53.44	83.61	91.41	93.98
PLM-Concat	54.16	83.84	91.89	94.33

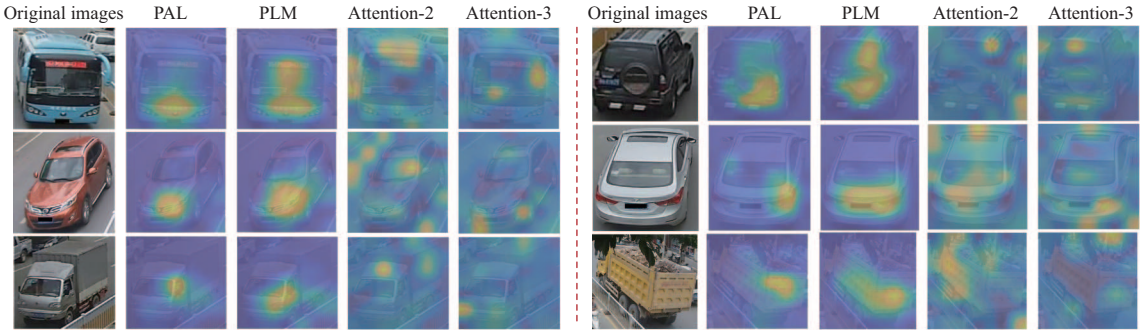

Figure 7 (Color online) Visualization of attention maps. The first column is original images, and the last four columns are the visualization of feature maps from ResNet, the multi-scale attention network and the output of the soft-attention structure in the second branch and the third branch, respectively.

and directly concatenated features from different layers, denoted as "PLM-Concat-noAtt". Table 6 shows the comparison results. It is obvious that, after adding the soft attention structure, the mAP has a 0.72% improvement for the VeRi-776 dataset. Moreover, the concatenation in (4) is replaced by the summarization, denoted as "PLM-Sum", to exploit the impact of different fusion methods on our framework. As shown in Table 6, the mAP of "PLM-Sum" drops to 53.19%, which demonstrates that concatenation is better than summarization in our network compared to "PLM-Concat". Moreover, we summarize our computational complexity in terms of memory and training time. For one single model, the saved parameter files of using ResNet50 and our multi-scale attention network are 186.8 and 336.9 Mb, respectively. This is because, based on ResNet50, there are some convolutional layers added into our network, which increases the amount of parameters. Additionally, for one epoch, the computing time for our network is approximately 3 min more than that of using ResNet50.

4.5 Qualitative analysis

Visualization of attention maps. To verify the effectiveness of the multi-scale attention network, features from ResNet and our proposed multi-scale attention network are visualized. This is outlined in Figure 7, where there are five columns in each group: original images, visualization of feature maps from ResNet, the multi-scale attention network, the output of the soft-attention structure in the second branch and the third branch. It is significant that the reID model trained by the multi-scale attention network can seek more distinctive parts.

Visualization of feature distributions. To better demonstrate the effectiveness of PLM for an unlabeled domain, the features of vehicle images in the gallery set are visualized in the different datasets, and the features are projected to 2-dimensional space via t-SNE [37] for dimension reduction and visualization. Figure 8 shows the visualization of feature distribution. Specifically, there are features of a total

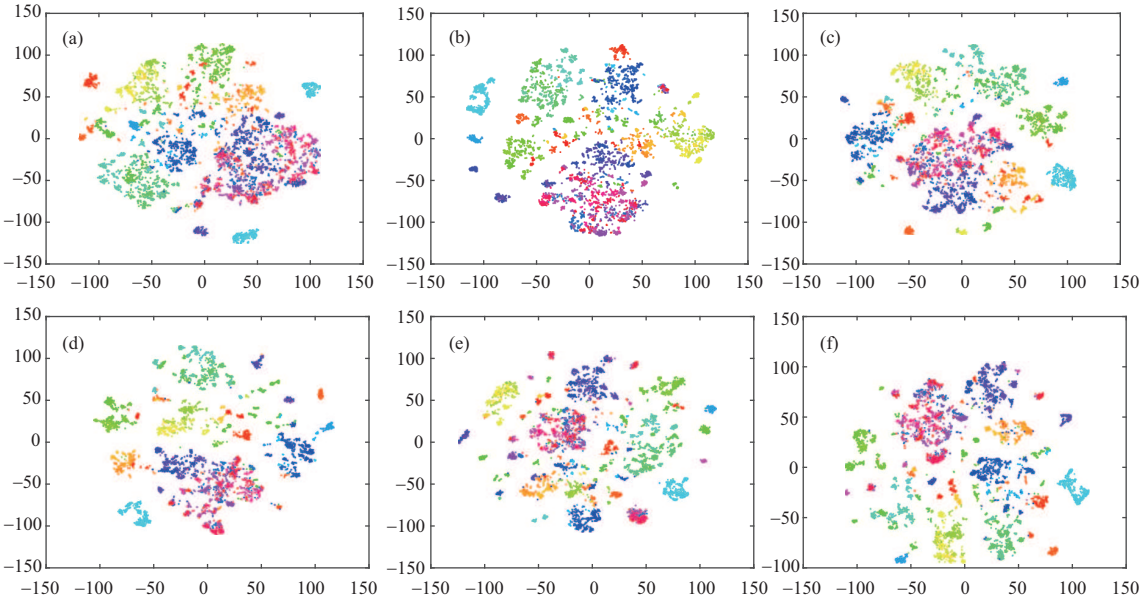

Figure 8 (Color online) Visualization of feature distribution by t-SNE [37]. Different colors represent different vehicle IDs. (a) iter = 1; (b) iter = 2; (c) iter = 3; (d) iter = 4; (e) iter = 5; (f) iter = 6.

Figure 9 (Color online) Retrieval results for VeRi-776. The image in the left-hand column is the query image, whereas the right-hand side shows the top-11 retrieval results obtained in different iterations.

of 200 vehicles that are extracted for different iterations on VeRi-776. The reID model is trained by our proposed PLM. It is observed that as the iteration grows, the result becomes better than the previous iteration.

Visualization of retrieval results. To further validate the effectiveness of PLM, some visualized results are provided in Figure 9 for VeRi-776 and VehicleID, where, for different datasets, the 6 separate rows represent the retrieval results from the first iteration to the sixth one. The number in the top-left represents the Vehicle ID/Camera ID. Each individual vehicle has its own Vehicle ID. The Camera ID is the number of the camera that captured the images. The image in the left-hand column is the query image, whereas the right-hand side shows the top-11 retrieval results obtained in different iterations. In each row, the correct results are indicated by a green border on the vehicle, whereas all other images are wrong. Figure 9 shows that a better performance is obtained as the iterations increase, for both VeRi-776 and VehicleID.

5 Conclusion

In this paper, we proposed a network for unsupervised vehicle reID, named PLM, which iteratively updates feature learning and estimates pseudo labels for unlabeled data to adapt the reID model in a target domain. The proposed domain adaptation module makes full use of the source domain, while the WLS loss treats the labels as a distribution over all pseudo labels, according to the distance between the samples and clustering centroids. PLM balances the confidence of different pseudo labels well. As an extension of this, the attention-based multi-scale network is proposed to learn more distinct features in unknown domains. The experimental results along with a detailed analysis have been carried out, demonstrating the advantages of PLM.

Acknowledgements This work was partially supported by National Natural Science Foundation of China (Grant Nos. 62172136, U21A20470, U1936217, 61725203, 62002041), Key Research and Technology Development Projects of Anhui Province (Grant No. 202004a5020043), Liaoning Doctoral Research Start-up Fund Project (Grant No. 2021-BS-075), and Dalian Science and Technology Innovation Fund (Grant No. 2021JJ12GX028).

References

1 Wang Y. Survey on deep multi-modal data analytics: collaboration, rivalry, and fusion. ACM Trans Multimedia Comput Commun Appl, 2021, 17: 1–25
2 Lou Y, Bai Y, Liu J, et al. Embedding adversarial learning for vehicle re-identification. IEEE Trans Image Process, 2019, 28: 3794–3807
3 Wu L, Wang Y, Gao J, et al. Deep coattention-based comparator for relative representation learning in person re-identification. IEEE Trans Neural Netw Learn Syst, 2021, 32: 722–735
4 Wang H, Peng J, Chen D, et al. Attribute-guided feature learning network for vehicle reidentification. IEEE Multimedia, 2020, 27: 112–121
5 Wu Y, Lin Y, Dong X, et al. Exploit the unknown gradually: one-shot video-based person re-identification by stepwise learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018. 5177–5186
6 Wu Y, Lin Y, Dong X, et al. Progressive learning for person re-identification with one example. IEEE Trans Image Process, 2019, 28: 2872–2881
7 Deng W, Zheng L, Ye Q, et al. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018. 994–1003
8 Wang J, Zhu X, Gong S, et al. Transferable joint attribute-identity deep learning for unsupervised person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018. 2275–2284
9 Wu L, Wang Y, Yin H, et al. Few-shot deep adversarial learning for video-based person re-identification. IEEE Trans Image Process, 2020, 29: 1233–1245
10 Song L, Wang C, Zhang L, et al. Unsupervised domain adaptive re-identification: theory and practice. Pattern Recogn, 2020, 102: 107173
11 Fan H, Zheng L, Yan C, et al. Unsupervised person re-identification. ACM Trans Multimedia Comput Commun Appl, 2018, 14: 1–18
12 Bashir R M S, Shahzad M, Fraz M M. VR-PROUD: vehicle re-identification using progressive unsupervised deep architecture. Pattern Recogn, 2019, 90: 52–65
13 Peng J, Wang Y, Wang H, et al. Unsupervised vehicle re-identification with progressive adaptation. In: Proceedings of International Joint Conference on Artificial Intelligence (IJCAI), 2020
14 Zhou Y, Shao L. Aware attentive multi-view inference for vehicle re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018. 6489–6498
15 Zhou Y, Liu L, Shao L. Vehicle re-identification by deep hidden multi-view inference. IEEE Trans Image Process, 2018, 27: 3275–3287
16 Guo H, Zhu K, Tang M, et al. Two-level attention network with multi-grain ranking loss for vehicle re-identification. IEEE Trans Image Process, 2019, 28: 4328–4338
17 Zhang X, Zhang R, Cao J, et al. Part-guided attention learning for vehicle re-identification. 2019. ArXiv:1909.06023
18 Liu H, Tian Y, Yang Y, et al. Deep relative distance learning: tell the difference between similar vehicles. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016. 2167–2175
19 Bai Y, Lou Y, Gao F, et al. Group-sensitive triplet embedding for vehicle reidentification. IEEE Trans Multimedia, 2018, 20: 2385–2399
20 Liu X, Liu W, Mei T, et al. PROVID: progressive and multimodal vehicle reidentification for large-scale urban surveillance. IEEE Trans Multimedia, 2018, 20: 645–658
21 Wang Z, Tang L, Liu X, et al. Orientation invariant feature embedding and spatial temporal regularization for vehicle re-identification. In: Proceedings of the IEEE International Conference on Computer Vision, 2017. 379–387
22 Shen Y, Xiao T, Li H, et al. Learning deep neural networks for vehicle Re-ID with visual-spatio-temporal path proposals. In: Proceedings of the IEEE International Conference on Computer Vision, 2017. 1900–1909
23 Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets. In: Proceedings of Advances in Neural Information Processing Systems, 2014. 2672–2680
24 Wu L, Hong R, Wang Y, et al. Cross-entropy adversarial view adaptation for person re-identification. IEEE Trans Circ Syst Video Technol, 2020, 30: 2081–2092
25 Zhou Y, Shao L. Cross-view GAN based vehicle generation for re-identification. In: Proceedings of the British Machine Vision Conference (BMVC), 2017. 1–12
26 Zhu J Y, Park T, Isola P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision, 2017. 2223–2232
27 Wu L, Wang Y, Shao L. Cycle-consistent deep generative hashing for cross-modal retrieval. IEEE Trans Image Process, 2019, 28: 1602–1612

28 Almahairi A, Rajeswar S, Sordoni A, et al. Augmented CycleGAN: learning many-to-many mappings from unpaired data. 2018. ArXiv:1802.10151
29 Taigman Y, Polyak A, Wolf L. Unsupervised cross-domain image generation. 2016. ArXiv:1611.02200
30 He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016. 770–778
31 Huang Y, Xu J, Wu Q, et al. Multi-pseudo regularized label for generated data in person re-identification. IEEE Trans Image Process, 2019, 28: 1391–1403
32 Abadi M, Barham P, Chen J, et al. TensorFlow: a system for large-scale machine learning. In: Proceedings of 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16), 2016. 265–283
33 Vedaldi A, Lenc K. MatConvNet: convolutional neural networks for MATLAB. In: Proceedings of the 23rd ACM International Conference on Multimedia, 2015. 689–692
34 Liu X, Liu W, Mei T, et al. A deep learning-based approach to progressive vehicle re-identification for urban surveillance. In: Proceedings of European Conference on Computer Vision. Berlin: Springer, 2016. 869–884
35 Zhong Z, Zheng L, Luo Z, et al. Invariance matters: exemplar memory for domain adaptive person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019. 598–607
36 Zheng Z, Zheng L, Yang Y. A discriminatively learned CNN embedding for person reidentification. ACM Trans Multimedia Comput Commun Appl, 2018, 14: 1–20
37 van Der Maaten L. Accelerating t-SNE using tree-based algorithms. J Machine Learn Res, 2014, 15: 3221–3245