Attribute-Aligned Domain-Invariant Feature Learning for Unsupervised Domain Adaptation Person Re-Identification

Huafeng Li, Yiwen Chen, Dapeng Tao, *Member, IEEE*, Zhengtao Yu, and Guanqiu Qi

***Abstract*— Domain invariance and discrimination of learned features as two crucial factors affect the performance of unsupervised domain adaptation (UDA) person re-identification (Re-ID). Person attributes (such as "backpack", "boots", "handbag", etc) remaining unchanged across multiple domains have been used as mid-level visual-semantic information in UDA person Re-ID. As two main challenges, both misalignment of attribute-related regions across multiple images and domain shift between source and target domains affect the learning of domain-invariant features (DIF). To address the above two challenges, this article proposes to take advantage of the stability of person attributes and the complementarity of person attributes and the corresponding low-level visual features to guide the learning of discriminative DIF. Specifically, the proposed solution contains the generation of latent attribute-correlated visual features (GLAVF), DIF learning under the guidance of person attributes, and the alignment of person attributes corresponding to the local regions of pedestrian images. Due to the gap between person attributes and visual features, person attributes are first converted into latent attribute-correlated visual features (LAVF) without any specific domain information in GLAVF, and then LAVF are used as the substitutions of person attributes to guide the learning of DIF. To enhance the discrimination of learned features, the proposed solution mainly explores the alignment between person attributes and corresponding local regions, and the alignment of the same person attributes across multiple pedestrian images. A fully connected layer is used to achieve the above two types of alignment in the proposed framework, which reduces the adverse impacts of inference information and ensures the semantic consistency between person attributes and corresponding local regions across multiple pedestrian images. The effectiveness of the proposed solution is confirmed on four existing datasets by comparative experiments.**

***Index Terms*— Person Re-ID, domain adaptation, person attributes, semantic alignment.**

Manuscript received April 22, 2020; revised August 31, 2020 and October 18, 2020; accepted October 18, 2020. Date of publication November 9, 2020; date of current version December 11, 2020. This work was supported in part by the National Natural Science Foundation of China under Grant 61966021, Grant 61772455, Grant U1713213, and Grant 61562053; in part by the National Key Research and Development Plan Project under Grant 2018YFC0830105 and Grant 2018YFC0830100; in part by the Major Science and Technology Project of Precious Metal Materials Genetic Engineering in Yunnan Province under Grant 2019ZE001-1 and Grant 202002AB080001; in part by the Yunnan Natural Science Funds under Grant 2018FY001(-013) and Grant 2019FA-045; in part by the Yunnan University Natural Science Funds under Grant 2018YDJQ004; and in part by the Project of Innovative Research Team of Yunnan Province under Grant 2018HC019. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Domingo Mery. *(Huafeng Li and Yiwen Chen contributed equally to this work.) (Corresponding authors: Dapeng Tao; Zhengtao Yu.)*

Huafeng Li, Yiwen Chen, and Zhengtao Yu are with the Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, China (e-mail: lhfchina99@kust.edu.cn; cyw9620@gmail.com; ztyu@hotmail.com).

Dapeng Tao is with the FIST Lab, School of Information Science and Engineering, Yunnan University, Kunming 650091, China, and also with United Vision Innovations Technology Company Ltd., Kunming 650091, China (e-mail: dapeng.tao@gmail.com).

Guanqiu Qi is with the Computer Information Systems Department, State University of New York at Buffalo State, Buffalo, NY 14222 USA (e-mail: qig@buffalostate.edu).

Digital Object Identifier 10.1109/TIFS.2020.3036800

I. Introduction

PERSON re-identification (Re-ID) aims to identify the same person across various scenes in multiple images captured by non-overlapping cameras. Compared with the traditional image retrieval, image querying of person Re-ID involving image matching across multiple disjoint camera views is more challenging [1]–[6]. On the basis of essential applications in intelligent surveillance, person Re-ID has attracted considerable attention in both academia and industry [7]–[16]. Although great progress has been made in person Re-ID, most of existing solutions focus on supervised deep learning. However, due to the difference between training data and target data, "domain shift" (also called "domain bias) causes the poor scalability and usability of supervised learning based solutions in practical applications. UDA person Re-ID methods can effectively alleviate the "domain shift" issues in supervised deep learning on labeled source datasets [17]–[22]. Although the effectiveness of these existing methods was confirmed on public datasets, a lot of challenges still exist in practical applications due to the ambiguity in pedestrian appearances across multiple disjoint camera views.

As a generic solution, the ambiguity of pedestrian appearances across multiple domains is alleviated by improving the discrimination of learned features. Person attributes such as "boots", "handbag", and "shoulder bag" (as shown in Fig. 1) can describe a person from visual-semantic aspects, and these attributes keep unchanged across multiple disjoint camera views, thus can be used as the complement to low-level visual features. In person Re-ID, person attributes are often utilized to improve visual feature learning. In most existing attribute-based person Re-ID methods [7], [11], [23]–[27], both mid-level attributes and low-level visual features extracted from pedestrian images work together on image matching. These methods usually design specific attribute classifiers first, and then the attribute classifiers are applied to global features to

Fig. 1. Pedestrian images and corresponding person attributes across multiple disjoint camera views. Person attributes are unchanged across multiple disjoint camera views, and have certain discrimination for distinguishing different people.

achieve the attribute prediction. Actually, most of mid-level attributes (such as "boots", "handbag", and "shoulder bag"), are only associated with specific local regions of pedestrian images, so person attributes directly from global features are not conducive to enhancing the performance of person Re-ID. Moreover, the extraction of DIF is difficult for these existing methods. One main reason is the spatial misalignment of local regions associated with the same attributes across multiple images (as shown in Fig. 1), which is mainly caused by the cluttered background, pose/viewpoint variation, and imperfect person detection. Another reason is that attributes cannot be effectively aligned with the local regions of pedestrian images as described.

In this article, the stability of person attributes is exploited to ensure the domain invariance of learned features, and the complementarity of person attributes and the corresponding low-level visual features is utilized to enhance the discrimination of learned features. As a basic principle, when the learned discriminative identity features are aligned with LAVF at domain level, the domain invariance of the learned features is guaranteed due to the domain invariance of LAVF. In addition, local regions of pedestrian images corresponding to person attributes are often discriminative, so they are helpful in distinguishing personal identities. If the local responses at the corresponding local regions of person attributes are strengthened, the discrimination of learned features can be enhanced. Furthermore, if the attribute-related regions are semantically aligned across multiple images, the discrimination of learned features can be further improved. According to the above analysis, a novel solution is proposed for UDA person Re-ID. As shown in Fig. 2, in addition to the "Baseline", the proposed method contains the generation of latent attribute-related visual features (i.e. GLAVF module), the learning of discriminative domain-invariant features (i.e. DIFL module), and the alignment between attributes and the corresponding local-region features (i.e. LAAF module).

Assuming that a certain correlation exists between attributes and the corresponding visual features, attributes are converted into LAVF in GLAVF module, and all the LAVF constitute a space, which is called LAVF space. After that, an adversarial learning strategy is proposed to achieve the domain alignment between the learned features and LAVF. In this way, the learned features are guaranteed to share the same domain information with LAVF. To enhance the discrimination of the learned DIF, a new fully connected layer is added to the proposed network to select and combine attribute-related features, in which the alignment between LAVF and the corresponding local features is achieved by a specific objective function. Under the supervision of attribute labels, feature maps associated with person attributes are purified, and the local regions corresponding to the same attribute across multiple images are aligned semantically in feature map channels. Therefore, the semantic alignment of attribute-related features is achieved across multiple images. Correspondingly, the discrimination of learned features can be improved. So far as we know, it is the first time to exploit attributes to obtain DIF and enhance the discrimination of DIF by semantically aligning local regions with the same attributes across multiple images.

In summary, this article has three main contributions as follows.

- According to the unchanged attributes across multiple domains, a domain-invariant feature extraction method is proposed. The proposed method converts attributes into LAVF first, and then obtains DIF by aligning the domain information of visual features extracted from pedestrian images with LAVF. So far as we know, it is the first time to apply attributes to the learning of DIF.
- To align the domain information of pedestrian visual features with LAVF, an adversarial learning mechanism is proposed to make both camera classifier and feature encoder pit against each other. During the learning process, camera classifier and feature encoder are optimized alternatively. Following the optimization, the camera classifier gradually becomes unable to distinguish whether the source of domain information is from extracted pedestrian visual features or LAVF. So, the domain information of pedestrian visual features is aligned with LAVF.
- To improve the discrimination of the learned domain-invariant features, a fully connected layer is added to the proposed network to extract features corresponding to attributes from global features obtained by Global Average Pooling (GAP), and then the discriminative information in the global features is strengthened by aligning LAVF with corresponding local-region features in pedestrian images. The semantic alignment of local regions corresponding to the same attributes across multiple images is achieved, which helps improve the representation quality of attribute-related regions.

The rest of this article is organized as follows: Section II discusses the related work; Section III specifies the proposed solution in detail; Section IV compares the proposed solution with existing solutions; and Section V concludes this article.

II. Related Work

A. Unsupervised Person Re-ID

Due to the lack of labeled data in practical application, unsupervised person Re-ID has attracted considerable

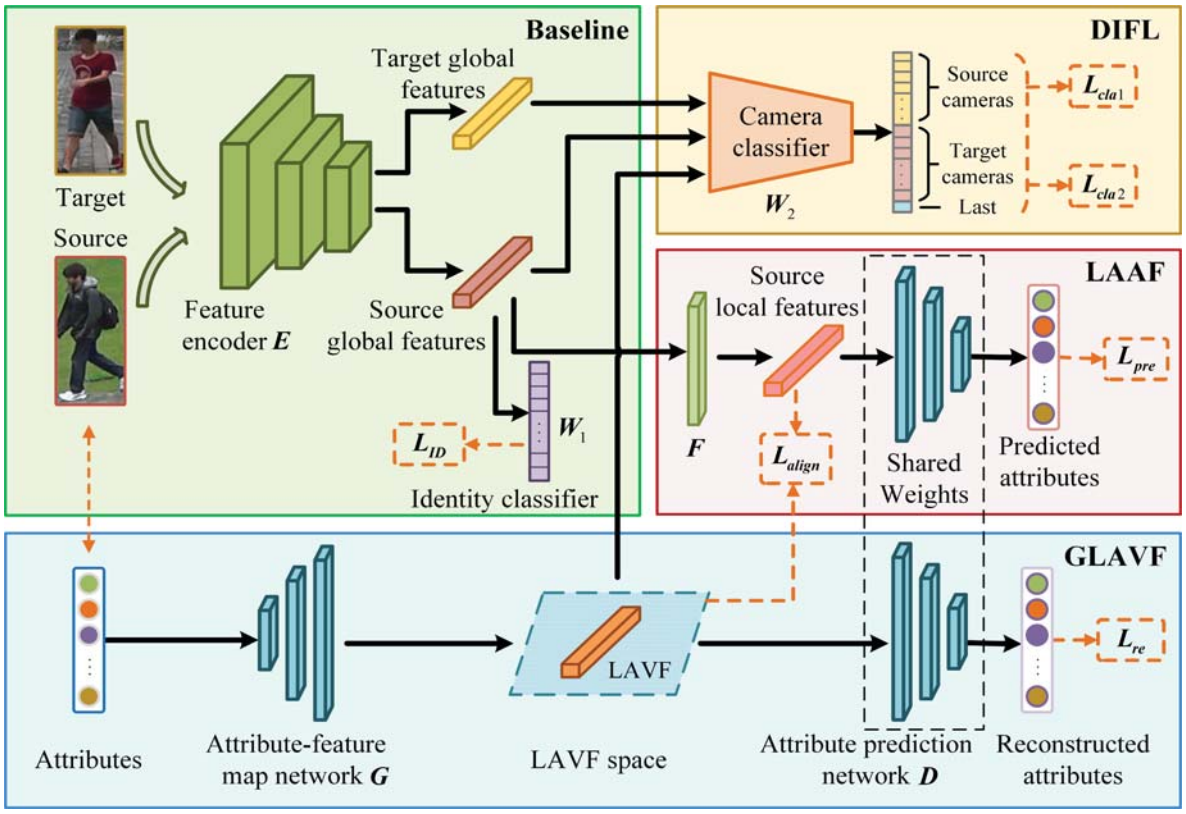

Fig. 2. Overall architecture of the proposed model. Attribute-feature map network $\boldsymbol{G}$ converts the attributes into LAVF so as to guide the learning of domain-invariant features. The attribute prediction network $\boldsymbol{D}$ is used to predict the original attributes from LAVF so as to encourage $\boldsymbol{G}$ to give an accurate prediction. The camera ID classifier $\boldsymbol{W}_2$ is used to identify the camera labels of global features and LAVF. $\boldsymbol{F}$ as a fully connected layer is used to achieve the alignment between attributes and their corresponding local regions. In the framework, $\boldsymbol{L}_{ID}$, $\boldsymbol{L}_{align}$, $\boldsymbol{L}_{re}$, $\boldsymbol{L}_{pre}$, $\boldsymbol{L}_{cla_1}$ and $\boldsymbol{L}_{cla_2}$ denote ID loss, alignment (align) loss, attribute reconstruction (re) loss, attribute prediction (pre) loss, and two camera classification (cla) losses respectively.

attention from researchers. Handcrafted features were first applied to unsupervised person Re-ID [28], [29]. However, it is difficult to obtain handcrafted features with high discrimination. In addition, this type of methods ignore the distribution of samples, resulting in the poor performance on large-scale datasets. Unsupervised feature learning such as [30] can avoid this problem. In person Re-ID, to obtain domain-invariant features across different views, some feature representation learning methods were proposed to learn the discriminative features by exploring invariant cross-view person information [31], [32]. Nevertheless, due to the absence of labeled pairwise samples in target domain, these methods have poor recognition performance. To solve this issue, some unsupervised self-training methods were developed to predict the labels for the unlabeled target data, and employ them to retrain the unsupervised person Re-ID models [33], [34]. In practical applications, target datasets often have really large size, but only contain a small number of paired samples. So, it is extremely difficult to select the paired samples from a target dataset. Due to the limited number of paired samples, even if all the paired samples are selected, it is still not enough to train a deep learning model with large-scale parameters to significantly enhance the performance of the propose solution on target dataset.

B. UDA in Person Re-ID

UDA person Re-ID has been widely concerned, because it can effectively alleviate the problem of poor recognition performance caused by the lack of labeled samples in target data. UDA person Re-ID methods usually train a Re-ID model to learn transferable features under the supervision of the labeled source dataset, and apply the learned Re-ID model to the unlabeled target dataset for person identity matching [18], [35]–[37]. Camera style translation, domain-invariant feature learning, and pseudo label prediction are commonly used in UDA person Re-ID methods. The camera style translation based methods often transfer images from labeled source domain to target domain, which makes the translated images share the same camera style with the samples in target domain. With the transferred images, the Re-ID model can be trained in a supervised manner. In particular, Liu *et al.* [37] developed an adaptive transfer network for UDA person Re-ID, where the image style transfer is achieved by multiple factor-wise CycleGANs and an ensemble CycleGAN. To alleviate the influence of complicated background on the discrimination of domain-invariant features, Huang *et al.* [19] proposed a generative adversarial network of background shift to generate images with suppressed backgrounds. This method guarantees the style consistency of the transferred image across different

domains, which narrows the domain gap. However, during the image style transfer, it is difficult to ensure that the visual cues associated with the corresponding ID information in an image are not changed [20]. Although this issue was noticed a long time ago, it has not been well solved.

Unlike image style transfer methods, which use the transferred images to train Re-ID models in a supervised manner, clustering and pseudo-label prediction based methods [14], [17], [18], [35], [38]–[40] solve lack of pairwise labels in target domains by assigning pseudo-labels to target samples. Particularly, to mitigate the negative effects of noisy labels introduced by clustering algorithms, Ge *et al.* [18] presented a mutual mean-teaching for UDA person Re-ID. To reduce the restrictions caused by lack of pairwise samples on the improvement of recognition performance, Yu *et al.* [35] developed a soft multi-label learning method to select paired samples from unlabeled target domain to train the corresponding Re-ID model. To improve the label estimation process, Ye *et al.* [40] designed a dynamic graph matching method for the task of video-based person Re-ID. Because of the excellent performance of dictionary learning in computer vision tasks [41]–[44], Li *et al.* [7] proposed a scalable pedestrian re-recognition based on dictionary learning. This method addresses the absence of labeled target samples by performing pseudo-label prediction on target data samples. These methods achieve excellent performance on public datasets, but lack of paired samples may cause such algorithms impractical in real world scenarios.

Compared with the clustering and pseudo-label prediction based methods, unsupervised domain-invariant feature learning is more practical in real-world applications [21], [45]–[48]. This type of methods usually train Re-ID models on the labeled source domain in a supervised way, so the domain-invariant features can be extracted across multiple domains. For example, to improve the discrimination of the learned features, Yang *et al.* [46] developed a patch-based unsupervised learning framework known as PatchNet for UDA person Re-ID. Song *et al.* [21] presented a domain-invariant mapping network to learn a generalizable person Re-ID model by exploring the mapping relationship between pedestrian images and an identity classifier. To learn the discriminative information from unlabeled target domains, Qi *et al.* [45] developed an unsupervised camera-aware domain adaptation approach to reduce the discrepancy among different domains.

C. Attributes for Person Re-ID

As mid-level semantic information, person attributes are intrinsically unchanged across multiple non-overlapping camera views, so they have been widely used in person Re-ID [7], [11], [23]–[27]. Wang *et al.* [23] developed a transferable deep learning framework to extract discriminative identity attributes by exploring both labeled attributes and identity information for UDA person Re-ID. To reduce the reliance on annotations, Wang *et al.* [11] used a plug-and-play method to extract valuable information for attribute prediction in an unsupervised way. During the testing process, the outputs of both feature and attribute layers are used as the discriminative information for personal identity matching. Tay *et al.* [25] presented an attribute attention network for person Re-ID, in which an identity classification framework is used to analyze both human body parts and key attributes. The above attribute-based methods mostly concatenate the predicted attributes with the extracted visual features for similarity measurement. Unlike the above methods, Li *et al.* [24] proposed to use person attributes to aid the detection and refinement of human parts for person Re-ID. Different from existing methods, the proposed solution in this article only utilizes attributes to guide the learning of domain-invariant features and enhance their discrimination by semantically aligning local regions with the same attribute across pedestrian images.

III. The Proposed Method

A. Overview of Our Framework

As shown in Fig. 2, the proposed model primarily consists of three functional modules: generation of LAVF (GLAVF), domain-invariant feature learning (DIFL), and local alignment of attributes and features (LAAF). The feature encoder $\boldsymbol{E}$ is trained on the labeled source domains by supervised learning, so that it has an basic ability to extract global discriminative features. In pre-training process, ResNet-50 [49] pre-trained on ImageNet [50] is used as the backbone, and followed by a fully connected layer, i.e., identity classifier $\boldsymbol{W}_1$ as shown in Fig.2. In GLAVF, we propose to learn a translator, i.e. attribute-feature map network $\boldsymbol{G}$, and an attribute prediction network $\boldsymbol{D}$. $\boldsymbol{G}$ is used to convert attributes into LAVF, and $\boldsymbol{D}$ is used to predict the attributes from LAVF.

In fact, $\boldsymbol{G}$ and $\boldsymbol{D}$ can be regarded as a pair of encoder and decoder. Attribute prediction network ensures that the relationship between the generated LAVF and the original attributes stay unchanged. The global features of both source and target domain samples are obtained by the pre-trained encoder $\boldsymbol{E}$. In DIFL, the encoder $\boldsymbol{E}$ is further trained to extract domain-invariant features. Specifically, a camera ID classifier $\boldsymbol{W}_2$ is introduced to conduct adversarial learning between $\boldsymbol{W}_2$ and $\boldsymbol{E}$, so that the visual features extracted by $\boldsymbol{E}$ and the LAVF share the same domain information. Local alignment is used to enhance the discrimination of the learned features by aligning attributes and the corresponding local visual features.

B. Generation of LAVF

Due to the gap between attributes and visual features, we can not directly employ attributes to guide the learning of domain-invariant features. To solve this issue, we propose to construct an attribute-feature map network $\boldsymbol{G}$ to convert attributes into LAVF. Given a labeled source domain $\boldsymbol{X}_s = \{\boldsymbol{x}_s^i\}_{i=1}^{N_s}$, and the real labels $\boldsymbol{Y} = \{y_s^i\}_{i=1}^{N_s}$, $y_s^i \in [1, 2, \cdots, n_s]$, where N_s is the number of images and n_s is the number of identities, both feature encoder $\boldsymbol{E}$ and identity classifier $\boldsymbol{W}_1$ can be optimized by minimizing the following cross-entropy loss.

$$\boldsymbol{L}_{ID}(\boldsymbol{E}, \boldsymbol{W}_1) = -\sum_{c=1}^{n_s} \boldsymbol{I}_{[c=y_s^i]} \log\left(p\left(\boldsymbol{W}_1\left(\boldsymbol{E}(\boldsymbol{x}_s^i)\right)\right)\right), \quad (1)$$

where $p(\cdot)$ indicates the logits of the i-th person image $\boldsymbol{x}_s^i$ belonging to c-th identity, and $\boldsymbol{I}_{[c=y_s^i]}$ denotes a common indicator function.

Since the feature extractor $\boldsymbol{E}$ is trained without the supervision of the labels of target samples, which may cause the over-fitting. To mitigate this issue, a label smooth operation defined in Eq.(2) [51] is introduced to replace the indicator function $\boldsymbol{I}_{[c=y_s^i]}$.

$$\hat{\boldsymbol{I}}_{[c=y_s^i]} = \begin{cases} 1 - \dfrac{N_s - 1}{N_s}\varepsilon, & \text{if } c = y_s^i \\ \dfrac{\varepsilon}{N_s}, & \text{otherwise} \end{cases}, \tag{2}$$

where ε is a small hyper-parameter. In this work, it is set to 0.1 according to [52]. So the improved Eq.(1) can be formulated as follows.

$$\boldsymbol{L}_{ID}(\boldsymbol{E}, \boldsymbol{W}_1) = -\sum_{c=1}^{n_s} \hat{\boldsymbol{I}}_{[c=y_s^i]} \log\left(p\left(\boldsymbol{W}_1\left(\boldsymbol{E}(\boldsymbol{x}_s^i)\right)\right)\right), \tag{3}$$

Under the supervision of attributes, both $\boldsymbol{G}$ and $\boldsymbol{D}$ are optimized by minimizing the following reconstruction loss.

$$\boldsymbol{L}_{re}(\boldsymbol{G}, \boldsymbol{D}) = -\frac{1}{n_b}\sum_{i=1}^{n_b} \|\boldsymbol{a}_s^i - \boldsymbol{D}(\boldsymbol{G}(\boldsymbol{a}_s^i))\|_1, \tag{4}$$

where $\|\cdot\|_1$ denotes the l_1 norm, n_b is the batch size and $\boldsymbol{a}_s^i$ is the attributes of the i-th person in source domain. With the learned $\boldsymbol{G}$, $\boldsymbol{D}$ and $\boldsymbol{E}$, the global visual features $\boldsymbol{f}_s^i = \boldsymbol{E}(\boldsymbol{x}_s^i)$, and the LAVF $\boldsymbol{f}_{s,a}^i = \boldsymbol{G}(\boldsymbol{a}_s^i)$ can be obtained. Due to the domain invariance of attributes, the converted LAVF from attributes are also domain-invariant, thus they can be used to replace the attributes to guide the learning of domain-invariant features.

C. DIF Learning

With the target dataset $\boldsymbol{X}_t = \{\boldsymbol{x}_t^i\}_{i=1}^{N_t}$, where N_t denotes the number of person images, the features $\boldsymbol{f}_t^i$ of input image $\boldsymbol{x}_t^i$ can be obtained by $\boldsymbol{f}_t^i = \boldsymbol{E}(\boldsymbol{x}_t^i)$ for identity matching. Nevertheless, the above method does not guarantee the learned features are domain-invariant. According to the natural stability of person attributes cross different camera views, we develop a novel adversarial learning mechanism to make camera ID classifier $\boldsymbol{W}_2$ and feature encoder $\boldsymbol{E}$ pit against each other. In this process, LAVF are used as the game director between $\boldsymbol{W}_2$ and $\boldsymbol{E}$ to promote the extracted features are aligned with LAVF at domain level by letting the features extracted by $\boldsymbol{E}$ share the same domain information with LAVF.

As shown in Fig. 3, a camera ID classifier $\boldsymbol{W}_2$ is designed to classify the features $\boldsymbol{f}_s^i$, $\boldsymbol{f}_t^i$ and LAVF $\boldsymbol{f}_{s,a}^i$ into the corresponding camera ID classes and a separate class respectively. During the training process, $\boldsymbol{W}_2$ is trained by supervised learning first. So $\boldsymbol{W}_2$ can classify the input images correctly. Once $\boldsymbol{W}_2$ is updated, we leave it unchanged and further optimize the encoder $\boldsymbol{E}$ to make $\boldsymbol{W}_2$ classify the features $\boldsymbol{f}_s^i$, $\boldsymbol{f}_t^i$ into the separate class. After that, we leave $\boldsymbol{E}$ unchanged and update $\boldsymbol{W}_2$ to make it can classify the input samples into corresponding camera classes. After this game, the domain of the extracted features by encoder $\boldsymbol{E}$ is aligned with the LAVF. Due to the domain invariance of LAVF, the domain invariance of the extracted features by $\boldsymbol{E}$ is guaranteed. Let $\boldsymbol{C}_s = \{c_c^i\}_{i=1}^{N_s}, c_c^i \in \{1, 2, \cdots, k_s\}$ and $\boldsymbol{C}_t = \{c_c^i\}_{i=1}^{N_t}, c_c^i \in \{1, 2, \cdots, k_t\}$ be the camera ID labels in both source and target domains. Thus, the dimension of the classification output of $\boldsymbol{W}_2$ is $n_c = k_s + k_t + 1$. With the updated $\boldsymbol{E}$, $\boldsymbol{W}_2$ can be updated by minimizing the following loss.

$$\begin{aligned} \boldsymbol{L}_{cla1}(\boldsymbol{W}_2) &= \boldsymbol{L}_{cla1,s}(\boldsymbol{W}_2) + \boldsymbol{L}_{cla1,t}(\boldsymbol{W}_2) + \boldsymbol{L}_{cla1,a}(\boldsymbol{W}_2) \\ &= -\sum_{c=1}^{k_s} \boldsymbol{I}_{[c=c_s^i]} \log\left(p_c\left(\boldsymbol{W}_2\left(\boldsymbol{E}(\boldsymbol{x}_s^i)\right)\right)\right) \\ &\quad -\sum_{c=1}^{k_t} \boldsymbol{I}_{[c=c_t^i]} \log\left(p_c\left(\boldsymbol{W}_2\left(\boldsymbol{E}(\boldsymbol{x}_t^i)\right)\right)\right) \\ &\quad -\sum_{c=1}^{n_c} \boldsymbol{I}_{[c=c_a]} \log\left(p_c\left(\boldsymbol{W}_2\left(\boldsymbol{G}(\boldsymbol{a}_s^i)\right)\right)\right), \end{aligned} \tag{5}$$

where p_c is the camera prediction logits of class c, and c_a is the camera label of LAVF.

After updating $\boldsymbol{W}_2$, the encoder $\boldsymbol{E}$ is updated by minimizing the following loss.

$$\begin{aligned} \boldsymbol{L}_{cla2}(\boldsymbol{E}) &= -\sum_{c=1}^{n_c} \boldsymbol{I}_{[c=c_a]} \log\left(p_c\left(\boldsymbol{W}_2\left(\boldsymbol{E}(\boldsymbol{x}_s^i)\right)\right)\right) \\ &\quad -\sum_{c=1}^{n_c} \boldsymbol{I}_{[c=c_a]} \log\left(p_c\left(\boldsymbol{W}_2\left(\boldsymbol{E}(\boldsymbol{x}_t^i)\right)\right)\right), \end{aligned} \tag{6}$$

In the above-mentioned adversarial learning, the recognition ability of classifier $\boldsymbol{W}_2$ is gradually improved by minimizing the loss functions in Eq.(5) and (6), which further promotes the encoder $\boldsymbol{E}$ accordingly to extract more domain-invariant features. Since equal probability classification can not guarantee the learned features share the same domain in theory, the proposed solution does not classify features into each camera identity with equal probability as [53] to achieve domain alignment.

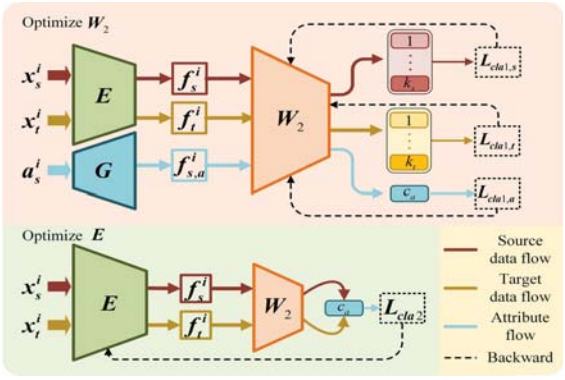

Fig. 3. Illustration of the DIFL module. This process includes two stages. The first one aims to optimize $\boldsymbol{W}_2$ by classifying features from different domains into corresponding camera labels. Here, $\boldsymbol{L}_{cla1,s}$, $\boldsymbol{L}_{cla1,t}$ and $\boldsymbol{L}_{cla1,a}$ are classified losses of $\boldsymbol{f}_s^i$, $\boldsymbol{f}_t^i$ and $\boldsymbol{f}_{s,a}^i$ respectively. The second one optimizes $\boldsymbol{E}$ with fixed $\boldsymbol{W}_2$ by classifying the features of both source and target domains into the separate class (last class). This process is achieved by minimizing the classified loss $\boldsymbol{L}_{cla2}$. "Backward" means the back propagation. Two stages are performed alternately.

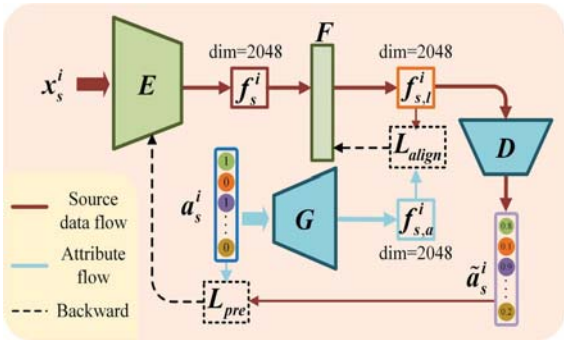

Fig. 4. Illustration of the LAAF module. $\boldsymbol{F}$ is used to extract the features $\boldsymbol{f}_{s,l}^i$ corresponding to attributes from global features $\boldsymbol{f}_s^i$ obtained by GAP, for achieving the alignment between attributes and their corresponding local features. "Backward" means the back propagation.

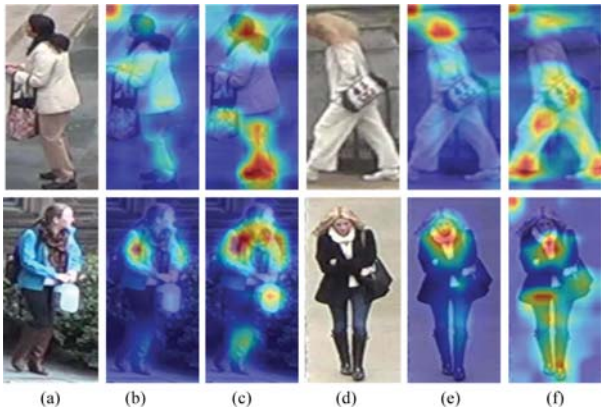

Fig. 5. Illustration of the effect of the fully connected layer $\boldsymbol{F}$ on global feature extraction under attribute supervision. Columns (a) and (d) show the original pedestrian images. Columns (b) and (e) show the responses (i.e. heatmaps) of the encoder $\boldsymbol{E}$ without the fully connected layer $\boldsymbol{F}$. Columns (c) and (f) show the responses of the encoder $\boldsymbol{E}$ followed by the fully connected layer $\boldsymbol{F}$. The strong activation regions are marked in red.

D. Local Alignment Between Attributes and Features

It is challenging to automatically align attributes with the corresponding local regions of pedestrian images due to the gap between attributes and visual features. As shown in recent work, the corresponding connections exist between feature channels of person Re-ID network and local regions of pedestrian images [54]–[56]. After assigning different weights to the features from different channels and learning these features under the supervision of attributes, the attribute-related feature channels can be selected to describe the corresponding attribute-related regions, so the alignment between attributes and the corresponding local visual features can be achieved.

So far, the encoder $\boldsymbol{E}$ trained without the supervision of attributes may ignore the discriminative local details of pedestrian images corresponding to attributes. As we know, person mid-level attributes such as "boots", "backpack", etc, are usually related to the specific local regions in pedestrian images. If person Re-ID network pays more attention to the attribute-related regions, the discrimination of the learned features can be improved. As one intuitive alignment method, an image is divided into different patches first, and then the attributes are aligned with these patches. However, this patch-based feature extraction method may damage the latent relationship between each patch, resulting in the reduction of feature discrimination. To solve the above issue, a fully connected layer $\boldsymbol{F}$ followed the encoder $\boldsymbol{E}$ is added into the Re-ID network to select relevant feature channels for describing each attribute-related region, as shown in Fig. 4. This process can be achieved by minimizing the following loss.

$$\boldsymbol{L}_{align}(\boldsymbol{F}) = \frac{1}{n_b}\sum_{i=1}^{n_b} \|\boldsymbol{G}(\boldsymbol{a}_s^i) - \boldsymbol{F}(\boldsymbol{E}(\boldsymbol{x}_s^i))\|_1, \tag{7}$$

where n_b denotes the batchsize.

Since $\boldsymbol{f}_s^i = \boldsymbol{E}(\boldsymbol{x}_s^i)$ contains a lot of redundant information, it is challenging to predict attributes directly from $\boldsymbol{f}_s^i$. To mitigate this issue, the local alignment between attributes and corresponding visual features is introduced in Eq.(7). In this process, if the global feature $\boldsymbol{E}(\boldsymbol{x}_s^i)$ contains attribute-related information, the local features $\boldsymbol{F}(\boldsymbol{E}(\boldsymbol{x}_s^i))$ aligned with attribute should be able to correctly predict the corresponding attributes $\boldsymbol{a}_s^i$ via $\boldsymbol{D}(\boldsymbol{F}(\boldsymbol{E}(\boldsymbol{x}_s^i)))$. In this article, the following Sigmoid Cross Entropy loss function is utilized to predict the attributes by considering all m attribute classes.

$$\boldsymbol{L}_{pre}(\boldsymbol{E}) = \sum_{i=1}^{n_b}\sum_{j=1}^{m}\left(a_s^{i,j}\log\left(p_a(\tilde{a}_s^{i,j})\right) + (1-a_s^{i,j})\log\left(1-p_a(\tilde{a}_s^{i,j})\right)\right), \tag{8}$$

where $p_a(\tilde{a}_s^{i,j})$ represents the predicted classification probability of the ground truth class $a_s^{i,j}$ of $\boldsymbol{x}_s^i$, and $\tilde{a}_s^{i,j}$ is the j-th element in attribute vector $\tilde{\boldsymbol{a}}_s^i = [\tilde{a}_s^{i,1}, \tilde{a}_s^{i,2}, \cdots, \tilde{a}_s^{i,m}]^T$. Since $\tilde{\boldsymbol{a}}_s^i$ are the attributes predicted by $\boldsymbol{D}$ from the local features $\boldsymbol{F}(\boldsymbol{E}(\boldsymbol{x}_s^i))$, $\tilde{\boldsymbol{a}}_s^i$ can be calculated by $\tilde{\boldsymbol{a}}_s^i = \boldsymbol{D}(\boldsymbol{F}(\boldsymbol{E}(\boldsymbol{x}_s^i)))$.

In the above process, the fully connected layer can not only align the attributes with the corresponding local regions of pedestrian images, but also align the local regions with the same attributes across multiple images in the feature channels of Re-ID network semantically. The alignment between attributes and corresponding local regions can promote the attribute-related regions to get more attention in Re-ID network training. Meanwhile, the channel alignment is conducive to the learning of features for semantic alignment. The above two alignments are unquestionably beneficial to the enhancement of the discrimination of global features. As shown in Fig. 5, the visualization results are generated by the method in [57] under the supervision of attributes. After introducing the fully connected layer $\boldsymbol{F}$, the encoder $\boldsymbol{E}$ can pay more attention to the attribute-related regions. At the same time, other discriminative regions irrelevant to attributes also receive sufficient attention from the encoder $\boldsymbol{E}$.

As shown in Fig. 6, one attribute-related channel is selected from the same channels of different images to further demonstrate the effect of the fully connected layer $\boldsymbol{F}$. As shown in columns (b) and (e) of Fig. 6, the encoder $\boldsymbol{E}$ can not effectively focus on the regions related to the attributes from

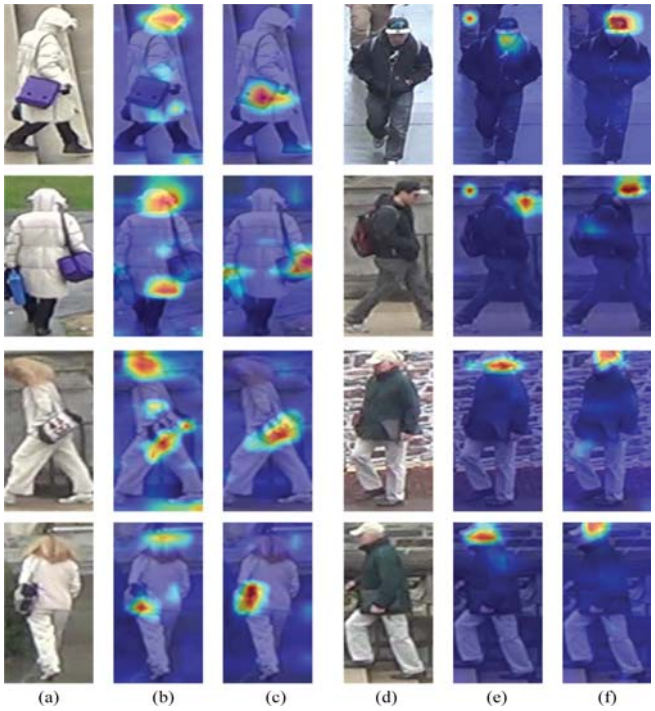

Fig. 6. The effect of the fully connected layer $\boldsymbol{F}$ on the semantic alignment of attribute-related regions. These heatmaps are the responses of our model to a specific attribute on the related single channel. Columns (a) and (d) show the source images. Columns (b) and (c) are from the same channel related to the attribute "shoulder bag", and columns (e) and (f) are from the same channel related to the attribute "hat" . Columns (b) and (e) show the results without the fully connected layer $\boldsymbol{F}$. Columns (c) and (f) show the results with the fully connected layer $\boldsymbol{F}$.

the same feature channels of Re-ID network without the fully connected layer $\boldsymbol{F}$. After introducing the fully connected layer $\boldsymbol{F}$, the regions activated by encoder $\boldsymbol{E}$ focus at the attribute-related regions, and the regions of the same attributes in different pedestrian images are activated at the same feature channels. It indicates that the encoder $\boldsymbol{E}$ can semantically align the attribute-related region features in global features across multiple images once the encoder $\boldsymbol{E}$ is followed by the fully connected layer $\boldsymbol{F}$. Thus the attribute-aligned global features can be obtained. Furthermore, the encoder $\boldsymbol{E}$ followed by the fully connected layer $\boldsymbol{F}$ can purify the attribute-related feature channels, which is helpful to enhance the discrimination of the learned features.

E. Final Loss of the Proposed Model

In sum, the final loss is formulated as the combination of the improved identity loss $\boldsymbol{L}_{ID}(\boldsymbol{E}, \boldsymbol{W}_1)$, reconstruction loss $\boldsymbol{L}_{re}(\boldsymbol{E}_a, \boldsymbol{D}_a)$, adversarial loss $\alpha_1 \boldsymbol{L}_{cla1}(\boldsymbol{W}_c) + \alpha_2 \boldsymbol{L}_{cla2}(\boldsymbol{E})$, alignment loss $\boldsymbol{L}_{align}(\boldsymbol{F})$ and prediction loss $\boldsymbol{L}_{pre}(\boldsymbol{E})$:

$$\begin{aligned} &\boldsymbol{L}(\boldsymbol{E}, \boldsymbol{G}, \boldsymbol{D}, \boldsymbol{F}, \boldsymbol{W}_1, \boldsymbol{W}_2) \\ &\quad = (\boldsymbol{L}_{ID}(\boldsymbol{E}, \boldsymbol{W}_1) + \boldsymbol{L}_{re}(\boldsymbol{G}, \boldsymbol{D})) + \alpha_1 \boldsymbol{L}_{cla1}(\boldsymbol{W}_2) \\ &\qquad + \alpha_2 \boldsymbol{L}_{cla2}(\boldsymbol{E}) + \alpha_3 \boldsymbol{L}_{align}(\boldsymbol{F}) + \alpha_4 \boldsymbol{L}_{pre}(\boldsymbol{E}) \end{aligned} \tag{9}$$

where α_1, α_2, α_3 and α_4 are hyper-parameters that adjust the weight of the corresponding loss respectively in the final loss. In the training process, identity loss $\boldsymbol{L}_{ID}(\boldsymbol{E}, \boldsymbol{W}_1)$ is used to train the "Baseline" on the labeled source domain in a supervised way, while reconstruction loss $\boldsymbol{L}_{re}(\boldsymbol{G}, \boldsymbol{D})$ is used to train the attribute-feature map network $\boldsymbol{G}$ and attribute prediction network $\boldsymbol{D}$ on source domain under the supervision of the manually labeled attributes. Then, adversarial loss $\alpha_1 \boldsymbol{L}_{cla1}(\boldsymbol{W}_2) + \alpha_2 \boldsymbol{L}_{cla2}(\boldsymbol{E})$ is used to train the encoder $\boldsymbol{E}$ on the labeled source domain and the unlabeled target domain for learning domain-invariant features. The last two losses $\boldsymbol{L}_{align}(\boldsymbol{F})$ and $\boldsymbol{L}_{pre}(\boldsymbol{E})$ are minimized to learn more discriminative features. The above processes are summarized in **Algorithm 1**.

Algorithm 1 Attribute-Aligned Domain-Invariant Feature Learning for Unsupervised Domain Adaptation Person Re-Identification

Input: Labeled source samples $\boldsymbol{X}_s = \{\boldsymbol{x}_s^i\}_{i=1}^{N_s}$, corresponding labels $\boldsymbol{Y} = \{y_s^i\}_{i=1}^{N_s}$, corresponding attribute annotations $\boldsymbol{A} = \{\boldsymbol{a}_s^i\}_{i=1}^{N_s}$, unlabeled target samples $\boldsymbol{X}_t = \{\boldsymbol{x}_t^i\}_{i=1}^{N_t}$.
Output: The trained encoder $\boldsymbol{E}$.
Step I: Generation of Latent Attribute-Correlated Visual Features (Sec.III.B)
1:Sample a batch of labeled source data to $\boldsymbol{E}$.
2:Initialize $\boldsymbol{E}$, $\boldsymbol{W}_1$, $\boldsymbol{G}$, $\boldsymbol{D}$
3:**for** *iter*=1, $\cdots$, $Iteration_1$ **do**
4: Update $\boldsymbol{E}$ and $\boldsymbol{W}_1$ by minimizing the loss in Eq.(3).
5: Update $\boldsymbol{G}$ and $\boldsymbol{D}$ by minimizing the loss in Eq.(4).
5:**end for**
Step II: Domain-Invariant Feature Learning (Sec.III.C)
6:Sample a batch of labelled source data.
7:Sample a batch of unlabeled source data.
8:Load the learned $\boldsymbol{E}$, $\boldsymbol{W}_1$, $\boldsymbol{G}$, $\boldsymbol{D}$.
9:Initialize camera classifier $\boldsymbol{W}_2$;
10: **for** *iter*=1, $\cdots$, $Iteration_2$ **do**
11: Update $\boldsymbol{W}_2$ by minimizing the loss in Eq.(5).
12: Update $\boldsymbol{E}$ by minimizing the loss in Eq.(3) and Eq.(6).
13: **end for**
Step III: Local Alignment Between Attributes and Features (Sec.III.D)
14:Sample a batch of labelled source data.
15:Sample a batch of unlabeled source data.
16:Load the learned $\boldsymbol{E}$, $\boldsymbol{W}_1$, $\boldsymbol{E}_a$, $\boldsymbol{D}$ and $\boldsymbol{W}_2$.
17:Initialize $\boldsymbol{F}$.
18: **for** *iter*=1, $\cdots$, $Iteration_3$ **do**
19: Update $\boldsymbol{W}_2$ by minimizing the loss in Eq.(5).
20: Update $\boldsymbol{F}$ by minimizing the loss in Eq.(7).
21: Update $\boldsymbol{E}$ by minimizing the loss in Eqs.(3), (6) and (8).
22: **end for**

IV. Experiments

A. Datasets and Evaluation Protocol

The proposed solution is applied to four large-scale person Re-ID datasets, Market1501 [67], DukeMTMC-reID [68], [69], CUHK03 [70] and MSMT17 [59]. The corresponding

results of the proposed solution are compared with state-of-the-art UDA person Re-ID methods.

Market1501 consists of 32,668 person images which contain to 1,501 identities and are captured by six cameras. This dataset is divided into training and testing sets. The training set has 12,936 images from 751 identities, and the testing set has 19,732 images from 750 identities. During the testing process, 3,368 query images are used to match the images in gallery. 27 attributes annotated in this dataset [26] are used as auxiliary information to guide the learning of domain-invariant features.

DukeMTMC-reID as a commonly used large-scale person Re-ID dataset contains 36,411 images from 1,404 identities captured by eight non-overlapping cameras. 702 identities and the remaining 702 identities are used as training and testing sets respectively. Similar to the split setting in [68], [69], 2,228 and 17,661 images in the testing set are used as query and gallery images respectively, and 23 attributes annotated are used as auxiliary information [26]. For simplicity,"Duke" is short for "DukeMTM-reID" in comparative experiments.

CUHK03 contains 14,096 images from 1,467 identities captured by five pairs of non-overlapping cameras. This dataset contains two image sets of people. One consists of the images captured by pedestrian detectors, and the other is composed by the images annotated by hand-drawn bounding boxes. In this article, we only perform experiments on the images captured by pedestrian detectors, because it is more challenging and closer to real scenes. The protocol in [70] is used, where the corresponding images from 100 identities are randomly selected as testing images and the remaining images are used for training. This procedure repeats 10 times, and the average score is reported.

MSMT17 as the largest pedestrian Re-ID dataset so far contains 126,441 images from 4,101 identities captured by 15 cameras within four days. According to [59], this dataset is randomly divided into both training and testing sets at a ratio of 1:3. The training set contains 32,621 bounding boxes marked on 1,041 people, and the testing set contains 93,820 bounding boxes marked on 3,060 people. Since the images in MSMT17 contain more complex background conditions, cover multiple time periods, and diverse illuminations, it is more challenging to process than other datasets.

Since Market1501 and Duke are annotated with various attributes, they are used as both source and target domains in comparative experiments. Due to lack of annotated attributes, CUHK03 and MSMT17 are only used as the target domain in comparative experiments.

Evaluation Protocol. Cumulative Match Characteristic (CMC) and mean average precision (mAP) are used to evaluate the performance of each method under a single query setting. CMC and mAP are used to measure the accuracy of identity matching at each rank and the accuracy of overall retrieval respectively.

B. Implementation Details

Network settings. Before applying an image to the proposed network, each image is resized to 256×128, and both random flip and crop are adopted for data augmentation. ResNet-50 is used as the backbone in the proposed method, which is followed by GAP to resize the features to 2,048-dimensional vectors. Subsequently, two parallel fully connected layers are add into the proposed framework, which are used to identify pedestrian images and extract local features respectively. Besides the backbone, an attribute-feature map network $\boldsymbol{G}$, an attribute prediction network $\boldsymbol{D}$, and a camera classifier $\boldsymbol{W}_2$ are designed. These networks mainly consist of fully convolutional layers and fully connected layers. The input of $\boldsymbol{G}$ is a low-dimensional attribute vector. The dimensions of an attribute vector of one person are 30 in Market1501 and 23 in Duke respectively. The output dimension of classifier $\boldsymbol{W}_2$ is one more than the total number of cameras in both source domain and target domain. During the testing process, Euclidean distance is used to measure the similarity of global features. All experiments are performed on one NVIDIA Tesla P100 GPU with 16GB memory.

Optimization. The complete training process has 130 epochs. Feature encoder $\boldsymbol{E}$ is only trained with the identity loss defined in Eq.(3) in the first 80 epochs. Simultaneously, attribute related networks $\boldsymbol{G}$ and $\boldsymbol{D}$ are optimized by $\boldsymbol{L}_{re}(\boldsymbol{E}, \boldsymbol{G})$. After that, classifier $\boldsymbol{W}_2$ and encoder $\boldsymbol{E}$ are trained by $\boldsymbol{L}_{cla1}(\boldsymbol{W}_2)$ and $\boldsymbol{L}_{cla2}(\boldsymbol{E})$, and this process lasts 20 epochs. In the last 30 epochs, the model are fine-tuned by $\boldsymbol{L}_{align}(\boldsymbol{F})$ and $\boldsymbol{L}_{pre}(\boldsymbol{E})$ on source domains to achieve the local alignment between attributes and visual features. Adam optimizer [71] with mini-batch of 16 is applied to all networks in the proposed method. When the training is performed on Market1501 and Duke, the initial learning rate of $\boldsymbol{E}$ is 1×10^{-4}, and the decay factor is 0.0005. According to [52], a warm-up strategy [72] is used to adjust the learning rate linearly. Specifically, the learning rate rises from 1×10^{-4} to 1.12×10^{-4} linearly for the first 30 epochs and declines to 1.12×10^{-5} at the 31th epoch. In 31∼55 epochs, the learning rate increases from 1.12×10^{-5} to 1.22×10^{-5} and then decays to 1.22×10^{-6} after the 55th epochs. Finally, the learning rate increases to 1.52×10^{-6} at the 130th epoch. In addition, the learning rate is set to 3.5×10^{-4} for $\boldsymbol{G}$ and $\boldsymbol{D}$ on all the datasets.

When Market1501 servers as the source domain, the initial learning rate is 2.5×10^{-5} for training $\boldsymbol{W}_2$. The initial learning rate of $\boldsymbol{F}$ is 1×10^{-3}, and then decays to one tenth of the initial learning rate after 20 epochs. When Duke is the source domain, the initial learning rates of $\boldsymbol{W}_c$ and $\boldsymbol{F}$ are set to 3×10^{-5} and 7.5×10^{-3} respectively. The differences in the above settings are only related to source domains, so the practical applications of the proposed method involving target domains are not affected. $\alpha_1 = 1$ and $\alpha_2 = 1$ are set empirically after the 81th epoch, and then $\alpha_1 = 0.1$ and $\alpha_2 = 0.01$ are set to fine-tune the proposed model at the 101th epoch. $\alpha_3 = 10$ and $\alpha_4 = 20$ are set throughout all the experiments. The settings of these hyper-parameters will be discussed in detail later.

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

In this section, the proposed method is compared with the state-of-the-art approaches on six experimental settings: Duke → Market1501, Duke → CUHK03, Duke → MSMT17,

TABLE I

COMPARISON OF THE PROPOSED METHOD WITH SOME STATE-OF-THE-ART METHODS ON DUKE→ MARKET1501 AND DUKE→ CUHK03. "MAP" INDICATES MEAN AVERAGE PRECISION, AND "–" INDICATES NO REPORTED DATA

Methods	Duke→ Market1501				Duke→ CUHK03			
	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP
HHL(ECCV'18) [58]	62.2	78.8	84.0	31.4	–	–	–	–
PTGAN(CVPR'18) [59]	38.6	–	66.1	–	24.8	–	–	–
TJ-AIDL(CVPR'18) [23]	58.2	–	–	26.5	–	–	–	–
ATNet(CVPR'19) [37]	55.7	73.2	79.4	25.6	–	–	–	–
CamStyle(TIP'19) [60]	58.8	78.2	84.3	27.4	–	–	–	–
SBSGAN(ICCV'19) [19]	58.5	–	–	27.3	33.7	–	–	27.3
UCDA(ICCV'19) [45]	64.3	–	–	34.5	–	–	–	–
CFSM(AAAI'19) [61]	61.2	–	–	28.3	–	–	–	–
CASC(ICCV'19) [22]	64.7	80.2	85.6	35.6	–	–	–	–
PAUL(CVPR'19) [46]	66.7	–	–	36.8	–	–	–	–
FMC(TFS'20) [62]	63.4	79.5	84.8	32.4	–	–	–	–
SSAE(PR'20) [63]	60.7	–	–	26.6	–	–	–	–
LVRP(TMM'20) [64]	63.9	81.1	86.4	33.9	–	–	–	–
CaNE(WACV'20) [65]	57.2	73	80	27.4	–	–	–	–
CSGLP(TIFS'20) [36]	61.2	77.5	83.2	31.5	–	–	–	–
DG-Net(ECCV'20) [66]	52.2	70.7	77.0	28.6	–	–	–	–
Ours	**71.8**	**85.9**	**90.1**	**39.6**	**35.1**	**52.9**	**62.8**	**29.7**

Market1501 → Duke, Market1501 → CUHK03 and Market1501 → MSMT17. In these experiments, A → B indicates the datasets A and B are used as the labeled source domain and the unlabeled target domain respectively. The proposed method neither performs the pseudo-label prediction on unlabeled target samples, nor selects the paired samples from target domains to fine-tune the proposed model. Therefore, the proposed method is not compared with pseudo-label prediction based methods.According to the review of unsupervised person Re-ID in [73], on Duke → Market1501 and Duke → CUHK03, the comparative methods include domain-invariant feature learning methods, i.e. TJ-AIDL [23], PAUL [46], CaNE [65], CASC [22], CFSM [61], FMC [62], UCDA [45], and SSAE [63], as well as style transfer learning methods, i.e. HHL [58], PTGAN [59], CamStyle [60], CSGLP [36], LVRP [64], SBSGAN [19], ATNet [37], and DG-Net [66] (without self-training). The comparative results of Duke → Market1501 and Duke → CUHK03 are shown in Table I.

As shown in Table I, the proposed method achieves the best matching accuracy in Rank-1, Rank-5, Rank-10, and mAP on Duke → Market1501 and Duke → CUHK03. Specifically, the recognition rate of the proposed method reaches 71.8%/39.6% and 35.1%/29.7% in Rank-1/mAP on Duke → Market1501 and Duke → CUHK03 respectively. As a domain-invariant feature learning method, the proposed solution outperforms the second best domain-invariant feature learning method 5.1%/2.8% in Rank-1/mAP on Duke → Market1501. Compared with the style transfer learning based methods such as LVRP, CSGLP, and PTGAN, the proposed method also shows better performance, and outperforms the second best style transfer learning method LVRP 7.9%/5.7% in Rank-1/mAP. On Duke → CUHK03, the proposed method outperforms SBSGAN 1.4%/2.4% in Rank-1/mAP. The above results confirm the effectiveness of the proposed method.

To further test both effectiveness and scalability of the proposed method on different datasets, the proposed method is applied to Market1501 → Duke and Market1501 → CUHK03, and the corresponding results obtained by the proposed method are compared with domain-invariant feature learning methods (including TJ-AIDL [23], PAUL [46], CFSM [61], FMC [62], CASC [22], UCDA [45], ENC [47], and SSAE [63]), and style transfer learning methods (including HHL [58], PTGAN [59], CamStyle [60], CSGLP [36], LVRP [64], SBSGAN [19], ATNet [37] and DG-Net [66] (without self-training)). All the obtained recognition rates on Market1501 → Duke and Market1501 → CUHK03 are listed in Table II.

As shown in Table II, the proposed method outperforms other comparative methods, when Maket1501 is used as the labeled source domain, and Duke and CUHK03 are used as the unlabeled target domains. As a domain-invariant feature learning method based on attribute guidance, the proposed solution is compared with existing domain-invariant feature learning methods to test the its effectiveness. As existing methods, ENC and SBSGAN are the second best methods are Duke and CUHK03 respectively. According to Table II,

TABLE II

EXPERIMENTAL RESULTS OF THE PROPOSED METHOD AND STATE-OF-THE-ART METHODS ON MARKET1501→ DUKE AND MARKET1501→ CUHK03. "MAP" INDICATES MEAN AVERAGE PRECISION. "–" INDICATES NO REPORTED DATA

Methods	Market1501→ Duke				Market1501→ CUHK03			
	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP
HHL(ECCV'18) [58]	46.9	61.0	66.7	27.2	–	–	–	–
PTGAN(CVPR'18) [59]	27.4	–	50.7	–	26.9	–	–	–
TJ-AIDL(CVPR'18) [23]	44.3	–	–	23.0	–	–	–	–
ATNet(CVPR'19) [37]	45.1	59.5	64.2	24.9	-	–	–	–
CamStyle(TIP'19) [60]	48.4	62.5	68.9	25.1	–	–	–	–
SBSGAN(ICCV'19) [19]	53.5	–	–	30.8	42.2	–	–	32.5
CFSM(AAAI'19) [61]	49.8	–	–	27.3	–	–	–	–
CASC(ICCV'19) [22]	51.5	66.7	71.7	30.5	–	–	–	–
UCDA(ICCV'19) [45]	55.4	–	–	36.7	–	–	–	–
PAUL(CVPR'19) [46]	56.1	–	–	35.7	–	–	–	–
ENC(CVPR'19) [47]	63.3	75.8	80.4	40.4	–	–	–	–
FMC(TFS'20) [62]	48.0	62.3	68.1	27.8	–	–	–	–
SSAE(PR'20) [63]	50.2	–	–	28.1	–	–	–	–
LVRP(TMM'20) [64]	36.3	54.0	61.6	17.9	–	–	–	–
CSGLP(TIFS'20) [36]	47.8	62.3	68.3	27.1	–	–	–	–
DG-Net(ECCV'20) [66]	53.2	68.7	73.8	36.3	–	–	–	–
Ours	**64.1**	**77.2**	**81.4**	**43.1**	**44.5**	**63.8**	**74.0**	**40.0**

TABLE III

PERFORMANCE(%) COMPARISON OF THE PROPOSED METHOD AND STATE-OF-THE-ART METHODS ON TWO TASKS OF DUKE→ MSMT17 AND MARKET1501→ MSMT17

Methods	Duke→ MSMT17			
	Rank-1	Rank-5	Rank-10	mAP
PTGAN(CVPR'18) [59]	11.8	-	27.4	3.3
ENC(CVPR'19) [47]	30.2	41.5	46.8	10.0
Ours	**38.6**	**50.8**	**56.1**	**14.0**
Methods	Market1501→ MSMT17			
	R-1	R-5	R-10	mAP
PTGAN(CVPR'18) [59]	10.2	-	24.4	2.9
ENC(CVPR'19) [47]	25.3	36.3	42.1	8.5
Ours	**30.5**	**42.6**	**48.8**	**11.4**

the accuracy obtained by the proposed solution is 0.8% and 2.7% higher than ENC in Rank-1 and mAP on Market1501 → Duke, and 2.3% and 7.5% higher than SBSGAN in Rank-1 and mAP on Market1501 → CUHK03. The above results further confirm both effectiveness and scalability of the proposed method.

As the largest dataset in comparative experiments, MSMT17 as target domain, and Duke and Market1501 with attribute annotations are used as source domains respectively to further test the performance of the proposed solution on a large-scale dataset. The proposed method is compared with two state-of-the-art unsupervised person Re-ID methods PTGAN [59] and ENC [47] on Duke→ MSMT17 and Market1501→MSMT17. As shown in Table III, the results confirm that the proposed method significantly outperforms PTGAN [59] and ENC [47]. Compared with the second best results, the proposed method improves the recognition accuracy from 30.2% and 10.0% to 38.6% and 14.0% in Rank-1 and mAP respectively, when Duke serves as the source domain. Similarly, when Market1501 is used as the source domain, the recognition rates obtained by the proposed solutions are 30.5% and 11.4% in Rank-1 and mAP respectively, which are increased by 5.2% and 2.9% respectively based on the second best results. According to the above results, the effectiveness of the proposed method is confirmed on a large-scale dataset.

D. Ablation Study

In this section, a series of experiments are conducted to evaluate the effectiveness of GLAVF, DIFL, and LAAF. In DIFL, the generated latent attribute-correlated visual features in GLAVF are utilized to guide the learning of domain-invariant features. To demonstrate the contributions of GLAVF, all attribute-related losses are removed, and a method based on the proposed approach is constructed to learn domain-invariant features and compared with "Baseline+DIFL". Since the learning of domain-invariant features is not guided by attributes, this method is called "Baseline+DIFL w/o GLAVF". In "Baseline+DIFL w/o GLAVF", $\boldsymbol{W}_2$ is optimized to classify the learned features into the corresponding camera

TABLE IV

ABLATION STUDY OF THE DEVELOPED METHOD. DUKE→ MARKET1501 AND MARKET1501→ DUKE ARE USED TO TEST THE EFFECTIVENESS OF EACH MODULE. THE "BASELINE (RESNET-50)" IS TRAINED ONLY WITHIDENTITY LOSS $\boldsymbol{L}_{ID}(\boldsymbol{E}, \boldsymbol{W}_1)$

Methods	Source	Target:Market1501				Source	Target:Duke			
		Rank-1	Rank-5	Rank-10	mAP		Rank-1	Rank-5	Rank-10	mAP
Baseline(ResNet-50)	Duke	63.6	78.0	83.2	33.2	Market1501	48.1	62.9	68.3	30.6
Baseline+DIFL\GLAVF		67.3	82.6	87.5	35.9		61.4	74.9	79.0	40.7
Baseline+DIFL		69.1	83.7	88.8	37.3		62.7	76.5	81.0	42.5
Baseline+DIFL+LAAF		**71.8**	**85.9**	**90.1**	**39.6**		**64.1**	**77.2**	**81.4**	**43.1**

ID classes, and $\boldsymbol{E}$ is also optimized to classify the learned features into an additional class that does not belong to any camera identity. All experiments are performed on Duke and Market1501, and the corresponding results are shown in Table IV. When one dataset is used as the source domain, the other is used as the target domain.

Effectiveness of DIFL. To alleviate the domain bias between both source and target datasets, attributes are used to guide the learning of domain-invariant features. As shown in Table IV, "Baseline+DIFL" achieves 69.1%/37.3% accuracy rate in Rank-1/mAP on Duke→Market1501, and 62.7%/42.5% accuracy rate in Rank-1/mAP on Market1501→Duke, which are significantly outperforms the recognition accuracy of "Baseline (ResNet-50)". The results demonstrate the effectiveness of the proposed learning of domain-invariant features.

Effectiveness of Attribute Guidance. Although the domain-invariant feature learning methods based on attribute guidance are effective, it does not show how the attribute guidance affects the domain-invariant feature learning. To explore the effects of attribute guidance, the attribute-related loss is removed from "Baseline+DIFL", and the method "Baseline+DIFL w/o GLAVF" is obtained and compared with "Baseline+DIFL". As shown in Table IV, when GLAVF is removed from "Baseline+DIFL", the recognition rate of Rank-1/mAP is reduced by 1.8%/1.4% (reduced from 69.1%/37.3% to 67.3%/35.9%) on Duke→Market1501, and 1.3%/1.8% (reduced from 62.7%/42.5% to 61.4%/40.7%) on Market1501→Duke. The results demonstrate that domain-invariant feature learning with attribute guidance is more effective than "Baseline+DIFL w/o GLAVF" for UDA person Re-ID.

Effectiveness of LAAF. To further improve the discrimination of the learned features, LAAF is added to "Baseline+DIFL" and the method "Baseline+DIFL+LAAF" is obtained. As shown in Table IV, "Baseline+DIFL+LAAF" outperforms "Baseline+DIFL", and improves the Rank-1/mAP accuracy from 69.1%/37.3% to 71.8%/39.6 on "Duke→ Market1501", and from 62.7%/42.5% to 64.1%/43.1 on "Market1501→Duke". This performance improvement benefits from the alignment between attributes and the corresponding local features. It indicates that the alignment between attributes and the corresponding local features can effectively improve the discrimination of features, which further confirms that the algorithm proposed in this article is reasonable and effective.

E. Parameter Selection and Analysis

In this section, a series of experiments are conducted to investigate the effects of the hyper-parameters α_1, α_2, α_3, and α_4 involved in the proposed model. The discussion of hyper-parameters analyzes the effects of α_1 and α_2 in DIFL and the effects of α_3 and α_4 in LAAF. The value of each parameter is changed within a certain range at a time to analyze the impacts of different parameter values. The analysis of these hyper-parameters is carried out on Duke and Market1501, but the selected values of these parameters are applied to each dataset. Only one parameter is changed during analyzing the effect of one parameter.

1) Effects of α_1 and α_2 in DIFL: To make the module play a better role in domain-invariant feature learning, α_1 and α_2 are set to different values in different training stages. In 81 to 100 epochs, the effects of α_1 and α_2 are shown in Figs. 7 (a) and (b). After 100th epoch, we change the values of α_1 and α_2 to adjust the contribution of $\boldsymbol{L}_{cla1}(\boldsymbol{W}_c)$ and $\boldsymbol{L}_{cla2}(\boldsymbol{E})$.

Effect of the parameter α_1. In 81∼100 epochs, the effect of α_1 with different values is shown in Fig. 7 (a). According to these results, the accuracies of both Rank-1 and mAP are improved, when $\alpha_1 \in [0.5, 1]$ is set for Duke→Market1501. On Market1501→Duke, the proposed method achieves higher accuracies in Rank-1 and mAP, when $\alpha_1 \in [0.5, 2]$. So α_1 is set to 1 for the proposed model in 80 ∼100 epochs. After the 100th epoch, the effect of the value of α_1 should be explored again on the proposed model. As shown in Fig. 7 (c), when $\alpha_1 = 1$ remains unchanged, the Rank-1 and mAP accuracies of the proposed method are not obviously optimal, and the optimal performance is achieved at $\alpha_1 = 0.1$ for all testing tasks.

Effect of the parameterα_2. The parameter α_2 is used to control the contributions of $\boldsymbol{L}_{cla2}(\boldsymbol{E})$. In Figs. 7 (b) and (d), the effects of α_2 are evaluated. First, in 81 ∼100 epochs, the value of α_2 varies from 0.01 to 10 to investigate its effect on the proposed model. As shown in Fig. 7 (b), $\alpha_2 = 1$ is a good choice on both Duke→Market1501 and Market1501→Duke. Therefore, in the initial stage of training, α_2 is set to a larger value (i.e. $\alpha_2 = 1$), so that $\boldsymbol{L}_{cla2}(\boldsymbol{E})$ can play a more important role in model training. Once the performance of our model becomes stable, α_2 is set to a smaller value to fine-tune the proposed model. According to Fig. 7 (d), when $\alpha_2 = 0.01$, the proposed method can achieve the best performance after the 100th epoch.

TABLE V

ANALYSIS OF EXPERIMENTAL RESULTS UNDER DIFFERENT SETTINGS. THE "BASELINE (RESNET-50)" IS TRAINED ONLY WITH IDENTITY LOSS $\boldsymbol{L}_{ID}(\boldsymbol{E}, \boldsymbol{W}_1)$

Methods	Source	Target:Market1501				Source	Target:Duke			
		Rank-1	Rank-5	Rank-10	mAP		Rank-1	Rank-5	Rank-10	mAP
Baseline(ResNet-50)	Duke	63.6	78.0	83.2	33.2	Market1501	48.1	62.9	68.3	30.6
Our model w/o $\boldsymbol{F}$		68.7	84.1	88.5	36.8		63.0	76.4	81.5	42.4
Our model w/o CamID		64.1	78.8	83.3	34.0		53.2	67.2	72.1	34.5
Our model + 2attri		68.0	82.5	87.4	36.8		61.8	76.1	80.5	41.8
Our model + Equ		68.8	83.8	88.2	36.8		63.6	77.2	81.5	43.0
Our model		**71.8**	**85.9**	**90.1**	**39.6**		**64.1**	**77.2**	**81.4**	**43.1**

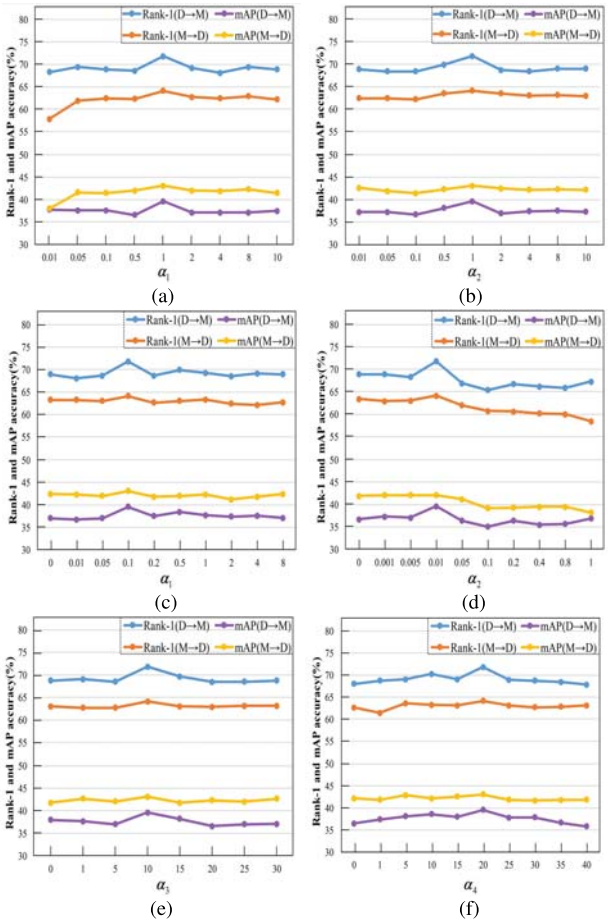

Fig. 7. Performance Analysis with Different Values of Hyper-parameters. (a)In 81∼100 epochs, Rank-1 and mAP accuracy when α_1 takes different values. (b)In 81∼100 epochs, Rank-1 and mAP accuracy when α_2 takes different values. (c)After the 100th epoch, Rank-1 and mAP accuracy when α_1 takes different values. (d)After the 100th epoch, Rank-1 and mAP accuracy when α_2 takes different values. (e)Rank-1 and mAP accuracy when α_3 takes different values. (f)Rank-1 and mAP accuracy when α_4 takes different values. D denotes the Duke dataset, and M denotes the Market1501 dataset.

2) Effects of α_3 and α_4 in LAAF: This sub-section evaluates how α_3 and α_4 impact the performance of the proposed model.

Effect of the parameter α_3. The effect analysis of parameter α_3 is illustrated in Fig. 7(e). When $\alpha_3 \geq 5$, the performance of the proposed model gradually improves on both Duke→Market1501 and Market1501→Duke. When $\alpha_3 \geq 10$, the accuracies of Rank-1 and mAP obtained by the proposed method reach the peak simultaneously, and then begin to decline. This indicates that $\alpha_3 = 10$ is the best value for the proposed method. In this article, α_3 is set to 10 for all the experiments.

Effect of the parameter α_4. In Eq.(9), α_4 is used to adjust the role of $\boldsymbol{L}_{pre}(\boldsymbol{E})$. In Fig. 7(f), the effect of α_4 is evaluated by varying its value from 0 to 40. As shown in Fig. 7(f), the proposed model consistently improves the accuracies of both Rank-1 and mAP within a wide range of parameter α_4. The best performance is achieved at $\alpha_4 = 20$. This indicates that $\alpha_4 = 20$ is a good value for the proposed method. The experiments performed on all datasets are under the fixed hyper-parameter settings.

F. Further Analysis and Discussion

Effect of fully connected layer $\boldsymbol{F}$. In LAAF module, a fully connected layer $\boldsymbol{F}$ is added into the encoder $\boldsymbol{E}$ to select attribute-related features from global features so as to achieve the alignment between attributes and the corresponding local regions. The proposed method realizes the above alignment by aligning LAVF with the corresponding local regions in pedestrian images. As a result, the discriminative information related to attributes in global features can be strengthened correspondingly. In addition, the alignment between attributes and the corresponding local regions is helpful to correctly predict the attributes from the corresponding local-region features in pedestrian images. The encoder $\boldsymbol{E}$ is promoted to pay more attention to attribute-related regions. To demonstrate the effect of $\boldsymbol{F}$, the fully connected layer $\boldsymbol{F}$ is removed from the proposed person Re-ID model and the corresponding model is named as "Our model w/o $\boldsymbol{F}$". "Our model w/o $\boldsymbol{F}$" is applied to "Duke→Market1501" and "Market1501→Duke" to test its performance. According to the corresponding experiment results listed in Table V, the Rank1/mAP of "Our model w/o $\boldsymbol{F}$" is 3.1%/2.8% and 1.1%/0.7% lower than the corresponding ones of "Our model" on Duke→Market1501 and Market1501→Duke respectively, which demonstrate the effectiveness of the fully connected layer $\boldsymbol{F}$ on improving the performance of the proposed person Re-ID model.

Effect of Camera ID. The proposed model introduces camera ID labels to train camera ID classifier $\boldsymbol{W}_2$ to distinguish the camera IDs of input features. To achieve the alignment of different domains, an adversarial mechanism is proposed to make $\boldsymbol{W}_2$ and encoder $\boldsymbol{E}$ pit against each other. To demonstrate the effect of the supervision of camera IDs,

the camera ID loss is removed from Eqs.(5) and (6), and the input features of $\boldsymbol{W}_2$ are classified into two classes. One class is composed of the sample features from both source and target domains, and the other class is composed of the LAVF. This method is named as "Our model w/o CamID". According to Table V, Rank1/mAP obtained by "Our model w/o CamID" only reaches 64.1%/34.0% and 53.2%/34.5% on Duke→Market1501 and Market1501→Duke respectively. Compared with the "Baseline" method, the performance of "Our model w/o CamID" is improved, but its performance is far less than the corresponding one obtained by using camera IDs in supervised training. The reason is that the learned classifier $\boldsymbol{W}_2$ under the supervision of camera IDs has stronger identification ability, and the ability of encoder $\boldsymbol{E}$ in extraction of domain-invariant features is also improved accordingly in adversarial training.

Effect of Person Attributes. The proposed method introduces person attributes to guide the learning of domain-invariant features. In practice, each person contains more than one attribute. To show the effect of different number of attributes, only two attributes are used in the training of the proposed person Re-ID model. This model is named as "Our model+2attri". As listed in Table V, the performance of "Our model + 2attri" is better than the corresponding one of "Baseline", when only two attributes are used. But its recognition accuracy on Rank1/mAP is 3.8%/2.8% and 2.3%/1.3% lower than the corresponding ones obtained by "Our model" on Duke→Market1501 and Market1501→Duke respectively. The experiment results confirm the comprehensive pedestrian descriptions are helpful to improve the recognition performance of the proposed Re-ID model.

Effect of Adversarial Mechanism. In DIFL module, the encoder $\boldsymbol{E}$ is updated based on the following expectation. Classifier $\boldsymbol{W}_2$ can classify all the extracted features into the separate classes of camera IDs. To achieve the domain alignment, traditional person Re-ID methods based on adversarial learning usually classify the extracted features into each camera ID with equal probability. For instance, the probability of image features from an image belonging to each camera is $\frac{1}{n_c}$, where n_c denotes the number of cameras. This method is named as "Our model + Equ". According to Table V, "Our model + Equ" achieves 68.8%/36.8% and 63.6%/43.0% Rank1/mAP accuracy on Duke→Market1501 and Market1501→Duke respectively. "Our model + Equ" has a lower performance than "Our model". The reason is that equal probability classification in "Our model+Equ" can not theoretically guarantee the domain alignment of the learned features.

Effect of LAVF in The Supervised Person Re-ID. The effectiveness of LAVF is furtherly evaluated in the supervised person Re-ID. Only the data from a single domain and the corresponding camera labels are used in the training of the supervised person Re-ID, so the output dimension n_c of camera classifier $\boldsymbol{W}_2$ equals $k_s + 1$. The supervised training process of the first 100 epochs uses the same settings of the unsupervised person Re-ID training. After adding the LAAF module to the 101st epoch, the training lasted 20 and 15 epochs in Duke and Market1501 respectively. For the experiments on Duke, the initial learning rate of camera classifier $\boldsymbol{W}_2$ is set to 2.25×10^{-5}, the hyper-parameters α_3 and α_4 are set to 1 respectively, and the remaining parameters are set to be same as the corresponding ones used in the unsupervised person Re-ID when Duke is used as the source domain. For the experiments on Market1501, the initial learning rate of camera classifier $\boldsymbol{W}_2$ and fully connected layer $\boldsymbol{F}$ are set to 1.8×10^{-5} and 0.01 respectively, the hyper-parameters α_3 and α_4 are set to 0.5 and 1 respectively, and the remaining parameters are set to be same as the corresponding ones used in the unsupervised person Re-ID when Market1501 is used as the source domain. The recognition performance of the proposed model in the supervised person Re-ID is named as "Supervised" in Table VI. The "Supervised w/o LAVF" method is obtained by removing LAVF learning from the "Supervised" method. These two methods are applied to the same dataset with the same experiment settings to demonstrate the effect of LAVF. As shown in Table VI, the recognition rate of Rank-1/mAP is reduced by 0.8%/1.2% (reduced from 85.5%/71.8% to 84.7%/70.6%) on Duke, and 0.7%/0.2% (reduced from 94.9%/83.8% to 94.2%/83.6%) on Market1501 after removing LAVF learning from the "Supervised" method. The results confirm that LAVF is conducive to improving the performance of supervised person Re-ID.

TABLE VI
ANALYSIS THE EFFECT OF LAVF IN THE SUPERVISED PERSON RE-ID

Methods	Duke	
	Rank-1	mAP
Supervised	**85.5**	**71.8**
Supervised w/o LAVF	84.7	70.6
Methods	Market1501	
	Rank-1	mAP
Supervised	**94.9**	**83.8**
Supervised w/o LAVF	94.2	83.6

V. CONCLUSION

In this article, a novel domain-invariant feature learning method is proposed for UDA person Re-ID. This approach makes full use of the domain-invariant attributes to improve the domain invariance and the discrimination of the learned features. The proposed method has three main sub-modules, GLAVF, DIFL, and LAAF. DIFL allows the model to extract domain-invariant features under the guidance of attributes, and LAAF improves the robustness of features by aligning attributes with their corresponding local features. The model structure and loss functions are discussed in detail. A series experiments are conducted on four convictive datasets, and the results confirm that the proposed method outperforms the state-of-the-art solutions. The ablation study demonstrates the effectiveness of attributes and each sub-module. Under the guidance of attributes, the scalability ability of the proposed model is enhanced, which is conducive to the practical applications in real-word scenarios. In future, the learning of person discriminative features will be further explored to improve the generalization of this model in practical applications.

REFERENCES

[1] M. Ye, X. Lan, Z. Wang, and P. C. Yuen, "Bi-directional center-constrained top-ranking for visible thermal person re-identification," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 407–419, 2020.

[2] C. Deng, E. Yang, T. Liu, J. Li, W. Liu, and D. Tao, "Unsupervised semantic-preserving adversarial hashing for image search," *IEEE Trans. Image Process.*, vol. 28, no. 8, pp. 4032–4044, Aug. 2019.

[3] K. L. Navaneet, R. K. Sarvadevabhatla, S. Shekhar, R. V. Babu, and A. Chakraborty, "Operator-in-the-loop deep sequential multi-camera feature fusion for person re-identification," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 2375–2385, 2020.

[4] C. Deng, E. Yang, T. Liu, and D. Tao, "Two-stream deep hashing with class-specific centers for supervised image search," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 6, pp. 2189–2201, Jun. 2020.

[5] E. Yang, C. Deng, C. Li, W. Liu, J. Li, and D. Tao, "Shared predictive cross-modal deep quantization," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 11, pp. 5292–5303, Nov. 2018.

[6] C. Yan, B. Gong, Y. Wei, and Y. Gao, "Deep multi-view enhancement hashing for image retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Feb. 24, 2020, doi: 10.1109/TPAMI.2020.2975798.

[7] H. Li, S. Yan, Z. Yu, and D. Tao, "Attribute-identity embedding and self-supervised learning for scalable person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 10, pp. 3472–3485, Oct. 2020.

[8] M. Ye and P. C. Yuen, "PurifyNet: A robust person re-identification model with noisy labels," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 2655–2666, 2020.

[9] H. Li, W. Zhou, Z. Yu, B. Yang, and H. Jin, "Person re-identification with dictionary learning regularized by stretching regularization and label consistency constraint," *Neurocomputing*, vol. 379, pp. 356–369, Feb. 2020.

[10] H. Li, J. Xu, J. Zhu, D. Tao, and Z. Yu, "Top distance regularized projection and dictionary learning for person re-identification," *Inf. Sci.*, vol. 502, pp. 472–491, Oct. 2019.

[11] Z. Wang, J. Jiang, Y. Wu, M. Ye, X. Bai, and S. Satoh, "Learning sparse and identity-preserved hidden attributes for person re-identification," *IEEE Trans. Image Process.*, vol. 29, pp. 2013–2025, 2020.

[12] A. R. Lejbolle, K. Nasrollahi, B. Krogh, and T. B. Moeslund, "Person re-identification using spatial and layer-wise attention," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 1216–1231, 2020.

[13] Q. Yang, A. Wu, and W.-S. Zheng, "Person re-identification by contour sketch under moderate clothing change," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Dec. 18, 2019, doi: 10.1109/TPAMI.2019.2960509.

[14] H.-X. Yu and W.-S. Zheng, "Weakly supervised discriminative feature learning with state information for person identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5528–5538.

[15] F. Ma, X.-Y. Jing, X. Zhu, Z. Tang, and Z. Peng, "True-color and grayscale video person re-identification," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 115–129, 2020.

[16] H. Li, J. Xu, Z. Yu, and J. Luo, "Jointly learning commonality and specificity dictionaries for person re-identification," *IEEE Trans. Image Process.*, vol. 29, pp. 7345–7358, 2020.

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

[18] Y. Ge, D. Chen, and H. Li, "Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2020, pp. 5157–5166.

[19] Y. Huang, Q. Wu, J. Xu, and Y. Zhong, "SBSGAN: Suppression of inter-domain background shift for person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9527–9536.

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

[21] J. Song, Y. Yang, Y.-Z. Song, T. Xiang, and T. M. Hospedales, "Generalizable person re-identification by domain-invariant mapping network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 719–728.

[22] A. Wu, W.-S. Zheng, and J.-H. Lai, "Unsupervised person re-identification by camera-aware similarity consistency learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6922–6931.

[23] J. Wang, X. Zhu, S. Gong, and W. Li, "Transferable joint attribute-identity deep learning for unsupervised person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2275–2284.

[24] S. Li, H. Yu, and R. Hu, "Attributes-aided part detection and refinement for person re-identification," *Pattern Recognit.*, vol. 97, Jan. 2020, Art. no. 107016.

[25] C.-P. Tay, S. Roy, and K.-H. Yap, "AANet: Attribute attention network for person re-identifications," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7134–7143.

[26] Y. Lin *et al.*, "Improving person re-identification by attribute and identity learning," *Pattern Recognit.*, vol. 95, pp. 151–161, Nov. 2019.

[27] Q. Zhou, B. Zhong, X. Lan, G. Sun, Y. Zhang, and M. Gou, "LRDNN: Local-refining based deep neural network for person re-identification with attribute discerning," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 1041–1047.

[28] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2197–2206.

[29] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato, "Hierarchical Gaussian descriptor for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1363–1372.

[30] M. Ye, J. Shen, X. Zhang, P. C. Yuen, and S.-F. Chang, "Augmentation invariant and instance spreading feature for softmax embedding," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Aug. 3, 2020, doi: 10.1109/TPAMI.2020.3013379.

[31] H.-X. Yu, A. Wu, and W.-S. Zheng, "Unsupervised person re-identification by deep asymmetric metric embedding," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 4, pp. 956–973, Apr. 2020.

[32] Z. Liu, D. Wang, and H. Lu, "Stepwise metric promotion for unsupervised video person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2429–2438.

[33] Y. Lin, X. Dong, L. Zheng, Y. Yan, and Y. Yang, "A bottom-up clustering approach to unsupervised person re-identification," in *Proc. 33rd AAAI Conf. Artif. Intell. (AAAI)*, 2019, pp. 8737–8745.

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

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

[36] C.-X. Ren, B. Liang, P. Ge, Y. Zhai, and Z. Lei, "Domain adaptive person re-identification via camera style generation and label propagation," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 1290–1302, 2020.

[37] J. Liu, Z.-J. Zha, D. Chen, R. Hong, and M. Wang, "Adaptive transfer network for cross-domain person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7202–7211.

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

[39] F. Yang *et al.*, "Asymmetric co-teaching for unsupervised cross-domain person re-identification," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2020, pp. 12597–12604.

[40] M. Ye, J. Li, A. J. Ma, L. Zheng, and P. C. Yuen, "Dynamic graph co-matching for unsupervised video-based person re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2976–2990, Jun. 2019.

[41] H. Li, X. He, Z. Yu, and J. Luo, "Noise-robust image fusion with low-rank sparse decomposition guided by external patch prior," *Inf. Sci.*, vol. 523, pp. 472–491, Mar. 2020.

[42] Z. Zhu, H. Yin, Y. Chai, Y. Li, and G. Qi, "A novel multi-modality image fusion method based on image decomposition and sparse representation," *Inf. Sci.*, vol. 432, pp. 516–529, Mar. 2018.

[43] H. Li, Y. Wang, Z. Yang, R. Wang, X. Li, and D. Tao, "Discriminative dictionary learning-based multiple component decomposition for detail-preserving noisy image fusion," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 4, pp. 1082–1102, Apr. 2020.

[44] H. Li, X. He, D. Tao, Y. Tang, and R. Wang, "Joint medical image fusion, denoising and enhancement via discriminative low-rank sparse dictionaries learning," *Pattern Recognit.*, vol. 79, pp. 130–146, Jul. 2018.

[45] L. Qi, L. Wang, J. Huo, L. Zhou, Y. Shi, and Y. Gao, "A novel unsupervised camera-aware domain adaptation framework for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8080–8089.

[46] Q. Yang, H.-X. Yu, A. Wu, and W.-S. Zheng, "Patch-based discriminative feature learning for unsupervised person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3633–3642.

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

[48] A. Wu, W.-S. Zheng, X. Guo, and J.-H. Lai, "Distilled person re-identification: Towards a more scalable system," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1187–1196.

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

[50] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and F. Li, "ImageNet: A larage-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 79–88.

[51] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.

[52] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, "Bag of tricks and a strong baseline for deep person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019.

[53] K. Kansal, A. V. Subramanyam, Z. Wang, and S. Satoh, "SDL: Spectrum-disentangled representation learning for visible-infrared person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 10, pp. 3422–3432, Oct. 2020.

[54] H. Yao, S. Zhang, R. Hong, Y. Zhang, C. Xu, and Q. Tian, "Deep representation learning with part loss for person re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2860–2871, Jun. 2019.

[55] S. Zhang, J. Yang, and B. Schiele, "Occluded pedestrian detection through guided attention in CNNs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6995–7003.

[56] C. Ding, K. Wang, P. Wang, and D. Tao, "Multi-task learning with coarse priors for robust part-aware person re-identification," 2020, *arXiv:2003.08069*. [Online]. Available: http://arxiv.org/abs/2003.08069

[57] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 336–359, Feb. 2020.

[58] Z. Zhong, L. Zheng, S. Li, and Y. Yang, "Generalizing a person retrieval model Hetero- and homogeneously," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 172–188.

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

[60] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang, "CamStyle: A novel data augmentation method for person re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 3, pp. 1176–1190, Mar. 2019.

[61] X. Chang, Y. Yang, T. Xiang, and T. M. Hospedales, "Disjoint label space transfer learning with common factorised space," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 33, 2019, pp. 3288–3295.

[62] Z. Zhang, M. Huang, S. Liu, B. Xiao, and T. Durrani, "Fuzzy multilayer clustering and fuzzy label regularization for unsupervised person re-identification," *IEEE Trans. Fuzzy Syst.*, vol. 28, no. 7, pp. 1356–1368, Jul. 2020.

[63] H. Li, Z. Kuang, Z. Yu, and J. Luo, "Structure alignment of attributes and visual features for cross-dataset person re-identification," *Pattern Recognit.*, vol. 106, Oct. 2020, Art. no. 107414.

[64] F. Yang, Z. Zhong, Z. Luo, S. Lian, and S. Li, "Leveraging virtual and real person for unsupervised person re-identification," *IEEE Trans. Multimedia*, vol. 22, no. 9, pp. 2444–2453, Sep. 2020.

[65] Y. Yuan *et al.*, "Calibrated domain-invariant learning for highly generalizable large scale re-identification," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 3589–3598.

[66] Y. Zou, X. Yang, Z. Yu, B. V. Kumar, and J. Kautz, "Joint disentangling and adaptation for cross-domain person re-identification," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 87–104.

[67] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1116–1124.

[68] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *Proc. Eur. Conf. Comput. Vis. Workshops (ECCVW)*, 2016, pp. 17–35.

[69] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by GAN improve the person re-identification baseline *in vitro*," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3754–3762.

[70] W. Li, R. Zhao, T. Xiao, and X. Wang, "DeepReID: Deep filter pairing neural network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 152–159.

[71] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980

[72] X. Fan, W. Jiang, H. Luo, and M. Fei, "SphereReID: Deep hypersphere manifold embedding for person re-identification," *J. Vis. Commun. Image Represent.*, vol. 60, pp. 51–58, Apr. 2019.

[73] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. H. Hoi, "Deep learning for person re-identification: A survey and outlook," 2020, *arXiv:2001.04193*. [Online]. Available: http://arxiv.org/abs/2001.04193

Huafeng Li received the M.S. degree in applied mathematics and the Ph.D. degree in control theory and control engineering from Chongqing University in 2009 and 2012, respectively. He is currently a Professor with the School of Information Engineering and Automation, Kunming University of Science and Technology, China. His research interests include image processing, computer vision, machine learning, and information fusion.

Yiwen Chen received the B.S. degree in automation from the Kunming University of Science and Technology, Yunnan, China, in 2014, where she is currently pursuing the master's degree in pattern recognition and intelligent system with the School of Information Engineering and Automation. Her research interests include machine learning and computer vision.

Dapeng Tao (Member, IEEE) received the B.E. degree from Northwestern Polytechnical University, Xian, China, in 1999, and the Ph.D. degree from the South China University of Technology, Guangzhou, China, in 2014. He is currently a Professor with the School of Information Science and Engineering, Yunnan University, Kunming, China. He has authored or coauthored more than 50 scientific articles. His research interests include machine learning, computer vision, and robotics. He has served for more than ten international journals, including IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON CYBERNETICS, IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, *Pattern Recognition*, and *Information Sciences*.

Zhengtao Yu received the Ph.D. degree in computer application technology from the Beijing Institute of Technology, Beijing, China, in 2005. He is currently a Professor with the School of Information Engineering and Automation, Kunming University of Science and Technology, China. His main research interests include natural language process, image processing, and machine learning.

Guanqiu Qi received the Ph.D. degree in computer science from Arizona State University, Tempe, AZ, USA, in 2014. He is currently an Assistant Professor with the Computer Information Systems Department, State University of New York at Buffalo State, Buffalo, NY, USA. His primary research interests include deep learning, machine learning, and image processing, and also span many aspects of software engineering, such as Software-as-a-Service (SaaS), Testing-as-a-Service (TaaS), big data testing, combinatorial testing, and service-oriented computing.