Body Part-Level Domain Alignment for Domain-Adaptive Person Re-Identification With Transformer Framework

Yiming Wang, Guanqiu Qi, Shuang Li, Yi Chai, and Huafeng Li

***Abstract*— Although existing domain-adaptive person re-identification (re-ID) methods have achieved competitive performance, most of them highly rely on the reliability of pseudo-label prediction, which seriously limits their applicability as noisy labels cannot be avoided. This paper designs a Transformer framework based on body part-level domain alignment to solve the above-mentioned issues in domain-adaptive person re-ID. Different parts of the human body (such as head, torso, and legs) have different structures and shapes. Therefore, they usually exhibit different characteristics. The proposed method makes full use of the dissimilarity between different human body parts. Specifically, the local features from the same body part are aggregated by the Transformer to obtain the corresponding class token, which is used as the global representation of this body part. Additionally, a Transformer layer-embedded adversarial learning strategy is designed. This strategy can simultaneously achieve domain alignment and classification of the class token for each human body part in both target and source domains by an integrated discriminator, thereby realizing domain alignment at human body part level. Compared with existing domain-level and identity-level alignment methods, the proposed method has a stronger fine-grained domain alignment capability. Therefore, the information loss or distortion that may occur in the feature alignment process can be effectively alleviated. The proposed method does not need to predict pseudo labels of any target sample, so the negative impact caused by unreliable pseudo labels on re-ID performance can be effectively avoided. Compared with state-of-the-art methods, the proposed method achieves better performance on the datasets that are in line with real-world scene settings.**

***Index Terms*— Person re-ID, domain adaptation, body part-level, domain alignment.**

Manuscript received 4 February 2022; revised 3 August 2022; accepted 7 September 2022. Date of publication 19 September 2022; date of current version 27 September 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 61966021, Grant 61562053, and Grant U2034209; and in part by the National Key Research and Development Plan Project under Grant 2018YFC0830105 and Grant 2018YFC0830100. The associate editor coordinating the review of this manuscript and approving it for publication was Mr. Remi Cogranne. *(Corresponding authors: Shuang Li; Yi Chai.)*

Yiming Wang and Yi Chai are with the School of Automation, Chongqing University, Chongqing 400044, China (e-mail: cquawang1ming@cqu.edu.cn; chaiyi@cqu.edu.cn).

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).

Shuang Li and Huafeng Li are with the Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, China (e-mail: shuangli936@gmail.com; lhfchina99@kust.edu.cn).

Digital Object Identifier 10.1109/TIFS.2022.3207893

I. Introduction

PERSON re-ID is used to associate a single pedestrian image captured by one camera with a/multiple pedestrian images captured by another/other cameras in a camera monitoring network, thereby realizing the positioning and tracking of the target pedestrian across time and space. Since person re-ID has high practical values, it has attracted considerable attention of researchers and a series of effective related methods have been proposed [1], [2]. These methods can achieve excellent recognition performance on the testing set consistent with the domain information of the corresponding training set. However, a large number of labeled samples are often required for supervised model training. When a domain shift occurs between testing set and training set, the corresponding re-ID performance drops sharply. In practical applications, the time and labor costs of manually labeling large-scale training samples on the target dataset are unacceptable [3]. According to the above-mentioned factors, it is difficult to apply supervised person re-ID methods to real-world surveillance scenes.

Unsupervised domain adaptation (UDA) is an effective method to solve the above-mentioned issues. Compared with both unsupervised person re-ID methods of domain generalization and unsupervised person re-ID methods without the participation of source-domain data (i.e. fully unsupervised methods), UDA method has better stability. Therefore, UDA person re-ID methods [4], [5] have attracted the attention of researchers. They usually apply both labeled source-domain data and unlabeled target-domain training data to model training. UDA method based on pseudo-label prediction and domain alignment are two main categories. UDA methods based on pseudo-label prediction usually perform pseudo-label prediction on training samples in the unlabeled target domain first [6]. Then, the predicted pseudo labels are used to further supervise model training. Therefore, the trained model is generally more adaptive to target data distribution.

Although UDA methods based on pseudo-label prediction show excellent re-ID performance on well-constructed training sets (each sample has at least a positive sample across multiple camera views), they have weak applicability in real-world surveillance scenes. As a main reason, these methods highly rely on pseudo-label prediction. When the accuracy of the predicted pseudo labels increases, the corresponding model performance on target data is improved. In real-world surveillance scenes, many pedestrians may only move in a

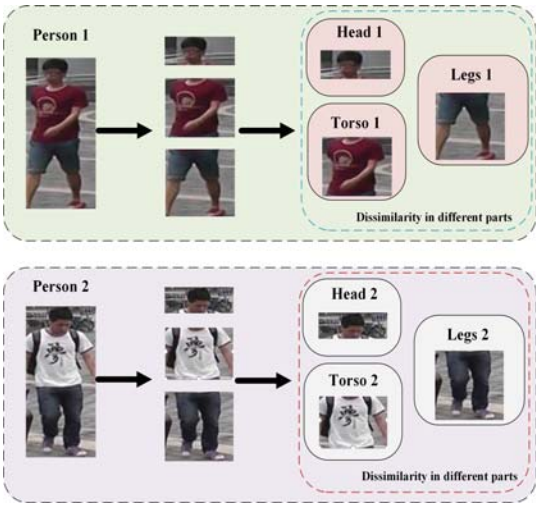


Fig. 1. Different parts of the human body.

small-range local area and pedestrian destination directions vary. Some pedestrians may be involved in the images captured by only one camera in a local camera network. Therefore, the corresponding pseudo-label prediction is seriously affected, resulting in a lot of noisy labels. In addition, when the distance between different cameras is relatively large (such as across distant scenes), it is highly possible that no pedestrian with the same identity appears in different cameras [7]. In this case, the predicted pseudo labels will all be noisy labels. As a main factor, the performance of pseudo-label prediction-based methods decreases considerably in such scenes.

Domain alignment-based feature learning solves domain shift by making both source-domain and target-domain sample features have the same distribution. In this process, the corresponding model performance can be improved without pseudo-label prediction. Therefore, the negative impact of noisy labels is avoided. Existing domain alignment-based methods often make the learned features achieve distribution consistency at domain level (between different datasets or between different cameras) or identity level (between different pedestrians). Although these methods finally achieve domain alignment, unexpected consequences may occur, such as the related information is only extracted from a local pedestrian image region or identity-related clues are sacrificed to satisfy the alignment of feature distribution.

To alleviate the above-mentioned issues, a novel domain alignment-based feature learning method is proposed. Compared with existing methods, the proposed method explores more fine-grained domain alignment. As shown in Fig. 1, different parts of the human body (such as head, torso, and legs) show different shapes and structures in a single pedestrian image. Different human body parts are essentially dissimilar (without considering the same color of both tops and pants). Therefore, this paper proposes a body part-level domain alignment method embedded in the Transformer layer to solve the issues of UDA person re-ID. Local features from the same body part are first aggregated through the Transformer to obtain class tokens for different body parts. Then, the aggregated features are used as a global representation of the corresponding body part. The class token for a body part in both source and target domains is aligned to achieve domain alignment at body-part level.

Specifically, this paper achieves body part-level domain alignment by the cooperation of two discriminators. Each discriminator is composed of four Transformer layers with the same structure (the related parameters are not shared) and four fully connected (FC) layers with different structures. This can ensure that the two discriminators can discriminate the features extracted by the backbone network from different views. In addition, the discriminators can simultaneously identify the class token for each pedestrian part belonging to the body part category and the corresponding original domain.

The integrated design avoids mode collapse caused by the independent design of identity and domain classifiers [8]. In the above-mentioned design, the input of the Transformer layer is the local features classified according to pedestrian body parts. As the main purpose, all the local features of the same human body part are used to further refine the class token of each body part. Additionally, the domain information is simultaneously extracted from the input features to prepare for the subsequent classification. Through the cooperative adversarial training between the backbone network and two discriminators, the main network is encouraged to extract the identity features consistent with source-domain distribution from target-domain samples, which ensures that the extracted features have strong capability to distinguish body parts. This is conducive to alleviating the loss and distortion of the identity information of different parts in the domain alignment process.

This paper has three main contributions as follows.

- A fine-grained domain alignment method embedded in the Transformer layer is proposed to solve the domain shift between the source and target domains. Features extracted from image patches of the same body part are input into the Transformer layer to obtain class tokens for different body parts and support subsequent body part-level domain alignment.
- The adversarial learning mechanism embedded in the Transformer layer is designed to align the class tokens of source-domain and target-domain samples from pedestrian bodies. Moreover, a priori knowledge based on the dissimilarity of different pedestrian parts is used to facilitate domain alignment at the body-part level.
- The proposed body part-level domain alignment enables the network to focus on discriminative features from different pedestrian body parts. Additionally, the potential loss of feature information or the distortion of identity-related clues can be effectively avoided in the domain alignment process.

The rest of this paper is organized as follows. Section II discusses related work; Section III elaborates the proposed method; Section IV analyzes the comparative experimental results; and Section V concludes this paper.

II. Related Work

A. Pseudo-Label Prediction-Based UDA Person Re-ID

UDA methods based on pseudo-label prediction perform model training in a supervised manner on unlabeled target data samples (training set), so the trained model has strong adaptability to the target dataset. As the core problem, this type of method needs to solve how to assign reliable pseudo labels to unlabeled target samples participating in model training. Although clustering is a commonly used method, it is easy to introduce noisy labels. In order to suppress the negative impact of noisy labels, Yang *et al.* [9] proposed an asymmetric collaborative teaching framework to suppress the generation of noisy labels through the cooperation of two networks. Ge *et al.* [10] proposed a mutual mean-teaching (MMT) method, which used both offline refining of hard pseudo labels and online refining of soft pseudo labels. Additionally, an alternate training method was applied to soft refining of pseudo labels in target domain. Zhai *et al.* [11] proposed a new augmented discriminative clustering method to achieve pseudo-label prediction. Both hierarchical clustering and hard-batch triplet loss (HCT) were integrated to improve pseudo-label prediction performance by Zeng *et al.* [12]. Zhao *et al.* [13] applied two networks to collaborative clustering and interactive instance selection to predict pseudo labels in the training process. Luo *et al.* [14] proposed to first cluster target data according to camera views and then predict and refine sample labels across camera views, so the reliability of confident labels was improved.

The above-mentioned methods achieve excellent performance on manually constructed datasets. Due to the limited range of pedestrian activities in real-world scenes, the different directions of entering and exiting a local camera network and a long distance between cameras can cause a large number of interference samples (the samples without cross-camera pairs of the same pedestrian) to be mixed in the training data. The existence of these samples inevitably causes the introduction of noisy labels, thereby reducing the corresponding re-ID performance of such methods.

B. Domain Alignment Based UDA Person Re-ID

UDA person re-ID methods based on domain alignment mainly solve the issues of domain shift by aligning the distribution of source and target domains. They do not use any pseudo label of the target dataset to supervise model training. The number and scale of cross-camera paired pedestrian samples do not have much impact on model performance. Therefore, the trained model has a strong generalization capability. In order to achieve domain alignment, PT-GAN [15], SPGAN [16], ATNet [17] and CR-GAN [18] first transfer the labeled source-domain samples to target domain, and then use the transferred samples to supervise model training.

Although these methods can alleviate domain shift, they ignore the intra-domain changes of samples, thereby limiting re-ID performance improvement. To alleviate this problem, Zhong *et al.* [19] conducted a comprehensive study on the intra-domain changes of target domain and proposed to assign three basic invariances (i.e. sample invariance, camera invariance, and neighborhood invariance) to re-ID models for achieving model performance improvement. Li *et al.* [20] integrated pedestrian pose information into an adversarial generation mechanism to obtain pose-invariant features after domain information alignment. Qi *et al.* [21] proposed a camera-aware domain-adaptive person re-ID framework. The data distribution discrepancy between source and target domains is addressed from different representation learning perspective. Aiming at the extraction of robust features for cross-domain person re-ID, Zou [22] proposed to improve model's domain adaptability by purifying the representation space to be adapted. Li *et al.* [23] made full use of the domain invariance of pedestrian features to guide the learning of domain-invariant features, which ensured the consistency of the distribution of both source-domain and target-domain features.

Most of the above-mentioned methods achieve domain alignment on the entire training sample space. Nevertheless, it is difficult to ensure the domain alignment between samples with the same identity. Although Li *et al.* [24] proposed an effective solution, the negative impact of domain alignment on feature quality was not considered. According to the dissimilarity of different pedestrian body parts, this paper proposes a body part-level fine-grained domain alignment framework to eliminate the domain difference between source domain and target domain. This method considers both domain alignment-related issues and the change of identity clues during the alignment process, and introduces the consistency classification of both body parts and domains to implementation.

C. Transformer in Person Re-ID

As a deep learning model, Transformer was designed for machine translation by Vaswani *et al.* [25]. Unlike convolutional neural networks, this method uses a self-attention mechanism to extract features from the entire input data. Inspired by the great success of the Transformer in natural language processing, researchers have applied the Transformer to image processing [26], object detection [27], semantic segmentation [28], object tracking [29], which achieves good performance. He *et al.* [30] first proposed a Transformer-based framework, which initiated the trend of Transformer-related applications in person re-ID. Lai *et al.* [31] proposed a Transformer-based framework for local fine-grained feature extraction,which can adaptively generate non-overlapping masks for robust part division. Zhu *et al.* [32] proposed an automatic alignment-based Transformer framework to realize the semantic alignment of features for person re-ID. Ma *et al.* [33] proposed a gesture-guided inter-part and intra-part relational Transformer to solve the issues of occluded person re-ID. The Transformer was mainly used to establish part-aware long-term correlations in this method. Li *et al.* [34] proposed an end-to-end part-aware Transformer to solve the negative impact of obstructions on pedestrian identity matching. Due to the lack of image-to-image attention, the Vision Transformer (ViT) and the vanilla Transformer with decoder are not able to achieve the matching of pedestrian images. Liao *et al.* [35] introduced query-gallery

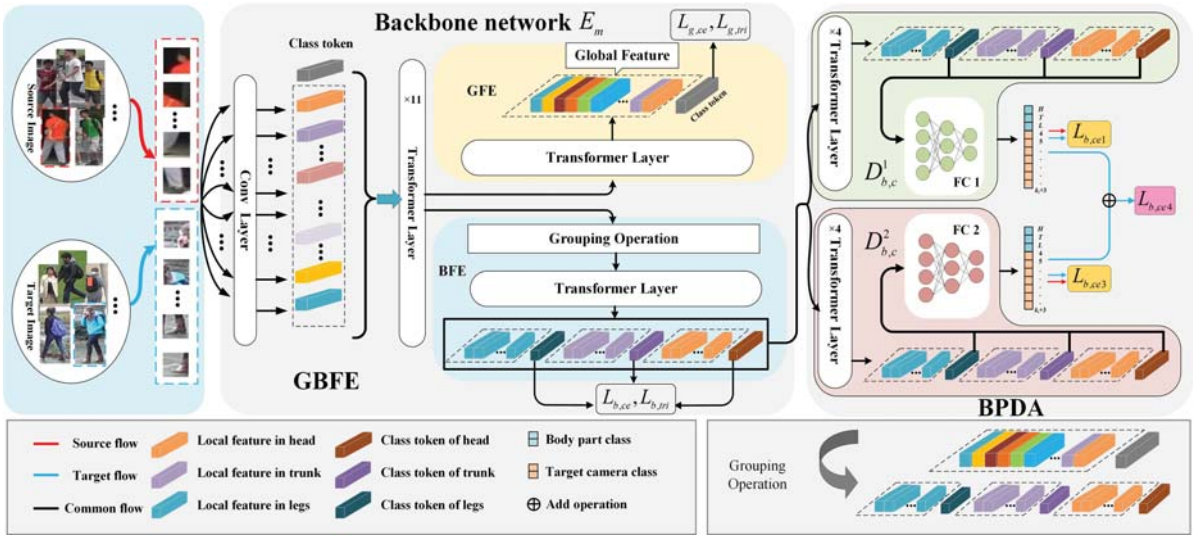


Fig. 2. The framework of the proposed method. The input image is first partitioned into different patches, and then the partitioned patches are convolved to obtain feature vectors in the convolution layer. The feature vectors of these local regions are input into the backbone network composed of Transformer layers to extract class tokens for global features. The last layer of the backbone network has two Transformer layers that share the same parameters. One is used for the global class token extraction of pedestrian bodies. The other is used for the class token extraction of pedestrian body parts. According to the position of the local feature output from the penultimate layer of the backbone network, each local feature is first classified into the category corresponding to the body part, and then input into the last Transformer layer $\boldsymbol{E}_m$ of the backbone network to extract the class token for each pedestrian body part. The class token for each body part is used to realize body part-level domain alignment of pedestrian samples.

concatenation and query-gallery cross-attention to VIT and vanilla Transformer respectively. Liang *et al.* [36] introduced Transformer to cross-modality person re-ID and achieved excellent performance in visible-infrared person re-ID.

Different from existing methods, this paper embeds transformer layers into both feature extraction backbone network and discriminators respectively. The backbone network is used to extract the class tokens describing each single pedestrian part. The discriminators are used to assist the backbone network in extracting the features of domain distribution alignment.

III. The Proposed Method

Preliminary. UDA person re-ID aims to obtain the model by training on the labeled source-domain dataset and the unlabeled target-domain dataset, thereby achieving good re-ID performance on the target dataset. Suppose the source domain dataset is $S = \{x_{s,i}, y_{s,i}, c_{s,i}\}|_{i=1}^{N_s}$, where $x_{s,i}$ represents the i-th pedestrian image in S, $y_{s,i} \in \{1, 2 \ldots, n_s\}$ and $c_{s,i} \in \{1, 2, \ldots, k_s\}$ represent both identity label and camera label of $x_{s,i}$ respectively, N_s is the total sample size, n_s is the number of pedestrians, and k_s is the total number of cameras in source domain. Additionally, $T = \{x_{t,i}, c_{t,i}\}|_{i=1}^{N_t}$ is set as the target dataset sample, N_t is the total number of sample images, $c_{t,i} \in \{1, 2, \ldots, k_t\}$ represents the camera label of $x_{t,i}$, and k_t represents the total number of cameras in target domain.

A. Overview

As shown in Fig.2, the proposed method consists of global and body-part feature extraction (GBFE) and body part-level domain alignment (BPDA). Specifically, the network of GBFE as the backbone network is composed of global feature extraction (GFE) and body-part feature extraction (BFE). GFE is mainly used to obtain the global features of the entire pedestrian body. BFE is mainly used to obtain the features of each pedestrian body part and build the foundation for the subsequent realization of part-level domain alignment based on the dissimilarity of body parts. BPDA is mainly composed of two discriminators. Each discriminator is embedded with four Transformer layers to extract the domain information contained in each body part. The dual adversarial learning is introduced between the backbone network and the two discriminators to realize the domain alignment of each class token for target-domain samples with the corresponding class token in source domain.

B. Global and Body-Part Feature Extraction

1) Global Feature Extraction: The backbone network $\boldsymbol{E}_m$ consists of a convolution layer and 12 transformer layers. $\boldsymbol{E}_m$ is used in GFE and BFE to extract global and local appearance features, respectively. The convolutional layer is used to convert the input image into tokens to be processed by Transformer layers. Each Transformer layer consists of a multi-head self-attention (MHSA) layer and a multilayer perception (MLP). There is a layer norm (LN) in front of each MLP and MHSA layer. Each sample $x_{s,i}$ in source domain and its corresponding identity label $y_{s,i}$ are known. In addition, similar to [7], [37], the identity labels of intra-camera samples can be assumed to be known, because they can be easily obtained various target tracking techniques. Therefore, the samples in T are grouped as $T^c = \{x_{t,i}^c, y_{t,i}^c\}_{i=1}^{n_t^c}$ according to the camera labels, where $x_{t,i}^c$ denotes the i-th image collected by the c-th camera, $y_{t,i}^c$ is the label identity of $x_{t,i}^c$, n_t^c is

the total sample size of the c-th camera. The source-domain samples and target-domain intra-camera samples are used to train $\boldsymbol{E}_m$ in a supervised manner, which ensures that the global features extracted by $\boldsymbol{E}_m$ are discriminative. Similar to [30], this paper uses cross-entropy loss and soft triple loss to optimize $\boldsymbol{E}_m$ as follows:

$$\boldsymbol{L}_{g,ce}(\boldsymbol{E}_m, \boldsymbol{W}_g, \boldsymbol{W}_c) = -\frac{1}{n_b}\Bigg(\sum_{i=1}^{n_b} \boldsymbol{q}_{s,i} \log(\boldsymbol{W}_g(\boldsymbol{E}_m(x_{s,i}))) + \frac{1}{k_t}\sum_{c=1}^{k_t}\sum_{i=1}^{n_b} \boldsymbol{q}_{t,i}^{c} \log(\boldsymbol{W}_c(\boldsymbol{E}_m(x_{t,i}^{c})))\Bigg), \tag{1}$$

$$\begin{aligned}\boldsymbol{L}_{g,tri}(\boldsymbol{E}_m) = \frac{1}{n_b}\Bigg(\sum_{i=1}^{n_b} \log[1 + \exp(\|\boldsymbol{E}_m(x_{s,i}) &- \boldsymbol{E}_m(x_{s,i}^{p})\|_2 - \|\boldsymbol{E}_m(x_{s,i}) \\ &- \boldsymbol{E}_m(x_{s,i}^{n})\|_2)] + \frac{1}{k_t}\sum_{c=1}^{k_t}\sum_{i=1}^{n_b} \log[1 \\ &+ \exp(\|\boldsymbol{E}_m(x_{t,i}^{c}) - \boldsymbol{E}_m(x_{t,i}^{c,p})\|_2 \\ &- \|\boldsymbol{E}_m(x_{t,i}^{c}) - \boldsymbol{E}_m(x_{t,i}^{c,n})\|_2)]\Bigg),\end{aligned} \tag{2}$$

where n_b is the batch size. $\boldsymbol{q}_{s,i} \in \mathbb{R}^{n_s \times 1}(\boldsymbol{q}_{t,i}^{c} \in \mathbb{R}^{n_s^c \times 1})$ is the one-hot vector, if and only if the $y_{s,i}(y_{t,i}^{c})$-th element are 1. $\boldsymbol{W}_g$ is the source domain pedestrian identity classifier and $\boldsymbol{W}_c$ is the pedestrian identity classifier of the samples captured by the c-th camera of target domain. $x_{s,i}^{p}(x_{t,i}^{c,p})$ and $x_{s,i}^{n}$ $(x_{t,i}^{c,n})$ are the hard positive samples and hard negative samples in a batch size corresponding to $x_{s,i}(x_{t,i}^{c})$.

2) Body-Part Feature Extraction: Since different pedestrian body parts (such as head, torso, legs) have different shapes and structures, they show strong dissimilarity. According to the above-mentioned a priori knowledge, body part-level domain alignment is proposed to avoid the loss or distortion of feature information in the domain alignment process. The proposed method needs to extract the features of different body parts. As shown in Fig. 2, a Transformer layer is copied from the last layer of $\boldsymbol{E}_m$ and used to extract body-part features. The input features of the last layer of $\boldsymbol{E}_m$ are listed as follows:

$$\boldsymbol{Z}_{l,11} = [z_{l,11}^{0}; z_{l,11}^{1}, z_{l,11}^{2}, \ldots, z_{l,11}^{N}], \tag{3}$$

which come from the input of the 11-th Transformer layer of $\boldsymbol{E}_m$ (the penultimate layer of the backbone network). In Eq.(3), N is the number of partitioned patches. $l \in \{s, t\}$ indicates that the input sample comes from either source domain or target domain. $z_{l,11}^{0}$ represents the class token for global features. $z_{l,11}^{i}, (i = 1, 2, \ldots, N)$ is the feature vector corresponding to the i-th local patch.

According to the feature dissimilarity of different parts of the human body, the entire pedestrian image is partitioned into three parts, head, torso, and legs. In this paper, local feature vectors are categorized into head, torso, and legs according to their positions in the source image. Then, they are spliced with the class token $z_{l,11}^{k}(k = hea, tor, leg)$ of the corresponding body parts as follows:

$$\begin{aligned}\boldsymbol{Z}_{l,11}^{hea} &= [z_{l,11}^{hea}; z_{l,11}^{1}, z_{l,11}^{2}, \ldots, z_{l,11}^{24}] \\ \boldsymbol{Z}_{l,11}^{tor} &= [z_{l,11}^{tor}; z_{l,11}^{25}, z_{l,11}^{26}, \ldots, z_{l,11}^{68}] \\ \boldsymbol{Z}_{l,11}^{leg} &= [z_{l,11}^{leg}; z_{l,11}^{69}, z_{l,11}^{70}, \ldots, z_{l,11}^{128}],\end{aligned} \tag{4}$$

Next, $\boldsymbol{Z}_{l,11}^{hea}$, $\boldsymbol{Z}_{l,11}^{tor}$, and $\boldsymbol{Z}_{l,11}^{leg}$ are input into the last Transformer layer of $\boldsymbol{E}_m$ to obtain $\boldsymbol{Z}_{l,12}^{hea}$, $\boldsymbol{Z}_{l,12}^{tor}$, and $\boldsymbol{Z}_{l,12}^{leg}$. Additionally, the corresponding class tokens $z_{l,12}^{hea}$, $z_{l,12}^{tor}$, and $z_{l,12}^{leg}$ are obtained. For the source domain image $x_{s,i}$ $(x_{t,i}^{c})$, the corresponding hard positive sample is denoted as $x_{s,i}^{p}$ $(x_{t,i}^{c,p})$ and the hard negative sample is denoted as $x_{s,i}^{n}$ $(x_{t,i}^{c,n})$. The class tokens of different body parts of $x_{s,i}$ obtained by $\boldsymbol{E}_m$ are expressed as $(z_{s,i}^{hea}, z_{s,i}^{tor}, z_{s,i}^{leg})$, $(z_{s,i}^{hea,p}, z_{s,i}^{tor,p}, z_{s,i}^{leg,p})$ and $(z_{s,i}^{hea,n}, z_{s,i}^{tor,n}, z_{s,i}^{leg,n})$. Similarly, the class tokens of different body parts of $x_{t,i}^{c}$ are expressed as $(z_{t,i}^{c,hea}, z_{t,i}^{c,tor}, z_{t,i}^{c,leg})$, $(z_{t,i}^{c,hea,p}, z_{t,i}^{c,tor,p}, z_{t,i}^{c,leg,p})$ and $(z_{t,i}^{c,hea,n}, z_{t,i}^{c,tor,n}, z_{t,i}^{c,leg,n})$. The parameters of $\boldsymbol{E}_m$ are optimized by the following loss to make the class tokens of various body parts discriminative, (5) and (6), as shown at the bottom of the next page, where $\boldsymbol{W}_{hea}(\boldsymbol{W}_{c,hea})$, $\boldsymbol{W}_{tor}(\boldsymbol{W}_{c,tor})$, and $\boldsymbol{W}_{leg}(\boldsymbol{W}_{c,leg})$ represent the local feature classifiers corresponding to different body parts in source-domina (c-th camera of target domain). To simplify the discussion, this paper only considers the cases of the pedestrian detection frame standard.

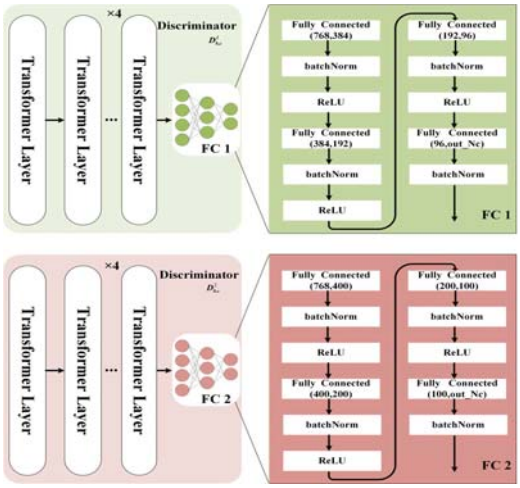


Fig. 3. Structure diagram of two discriminators.

C. Body Part-Level Domain Alignment

Since the pedestrian body parts are essentially dissimilar, the class tokens $z_{s,i}^{hea}$, $z_{s,i}^{tor}$, and $z_{s,i}^{leg}$ describing different pedestrian body parts should not be similar to each other. A self-dissimilarity domain alignment method embedded in the Transformer layer is proposed to obtain the features of domain alignment. This method is mainly realized by the cooperation between the backbone network and two discriminators.

1) Self-Dissimilarity Domain Alignment Embedded in the Transformer Layer: In domain alignment, two discriminators are mainly used to check whether the class tokens $\boldsymbol{z}_{t,i}^{hea}$, $\boldsymbol{z}_{t,i}^{tor}$, and $\boldsymbol{z}_{t,i}^{leg}$ contain target-domain information. As shown in Fig.3, each of the two discriminators is composed of four Transformer layers and four fully connected layers. The first discriminator denoted as $\boldsymbol{D}_{b,c}^{1}$ is obtained by integrating the source-domain body part classifier and the camera identity classifier of the target-domain sample into one classifier. In other words, $\boldsymbol{D}_{b,c}^{1}$ can simultaneously differentiate the body part category of the source-domain sample feature and the camera identity of the target-domain sample. Therefore, the output dimension of $\boldsymbol{D}_{b,c}^{1}$ is k_t+3, where k_t is the number of cameras of the target domain, and 3 represents the number of pedestrian body parts. This integrated design is conducive to avoiding mode collapse caused by the independent design of task and domain classifiers [8].

During the optimization of $\boldsymbol{D}_{b,c}^{1}$, the input source-domain features are expected to be correctly classified into the corresponding body part categories. $\boldsymbol{D}_{b,c}^{1}$ is expected to classify the input features of the target-domain sample into the camera category (identity) corresponding to the sample. Since different body parts contain the domain information that is inconsistent with source domain, the input features of the target-domain sample cannot be classified into the category of the corresponding body part. This process can be achieved by minimizing the loss function shown in Eq. (7):

$$\begin{aligned} L_{b,ce1}(\boldsymbol{D}_{b,c}^{1}) = -\frac{1}{3n_b}\sum_{i=1}^{n_b}[&\boldsymbol{q}_{s,i}^{hea}\log(\boldsymbol{D}_{b,c}^{1}(\boldsymbol{Z}_{s,i,12}^{hea})) \\ &+\boldsymbol{q}_{s,i}^{tor}\log(\boldsymbol{D}_{b,c}^{1}(\boldsymbol{Z}_{s,i,12}^{tor}))+\boldsymbol{q}_{s,i}^{leg}\log(\boldsymbol{D}_{b,c}^{1} \\ &\times(\boldsymbol{Z}_{s,i,12}^{leg}))]+[\boldsymbol{c}_{t,i}^{hea}\log(\boldsymbol{D}_{b,c}^{1}(\boldsymbol{Z}_{t,i,12}^{hea})) \\ &+\boldsymbol{c}_{t,i}^{tor}\log(\boldsymbol{D}_{b,c}^{1}(\boldsymbol{Z}_{t,i,12}^{tor})) \\ &+\boldsymbol{c}_{t,i}^{leg}\log(\boldsymbol{D}_{b,c}^{1}(\boldsymbol{Z}_{t,i,12}^{leg}))], \end{aligned} \tag{7}$$

where $\boldsymbol{q}_{s,i}^{hea} \in \mathbb{R}^{(k_t+3)\times 1}$, $\boldsymbol{q}_{s,i}^{tor} \in \mathbb{R}^{(k_t+3)\times 1}$, and $\boldsymbol{q}_{s,i}^{leg} \in \mathbb{R}^{(k_t+3)\times 1}$ are the label vectors of head, torso, and lower body categories, respectively. The values of $\boldsymbol{q}_{s,i}^{hea}$, $\boldsymbol{q}_{s,i}^{tor}$, and $\boldsymbol{q}_{s,i}^{leg}$ at the first, second, and third element positions are 1, and the values at other element positions are 0. $\boldsymbol{c}_{t,i}^{hea} \in \mathbb{R}^{(k_t+3)\times 1}$, $\boldsymbol{c}_{t,i}^{tor} \in \mathbb{R}^{(k_t+3)\times 1}$, and $\boldsymbol{c}_{t,i}^{leg} \in \mathbb{R}^{(k_t+3)\times 1}$ indicate the corresponding camera labels of different parts. If and only if the $(\boldsymbol{c}_{t,i}+3)$-th element is 1, other elements are zero, where $\boldsymbol{c}_{t,i}$ is the camera label of the input target-domain sample.

The target-domain sample is input into the backbone network. If the features consistent with the domain information of the source-domain sample can be extracted, $\boldsymbol{D}_{b,c}^{1}$ can classify the input target-domain features into the categories corresponding to body parts. To achieve this, the discriminator $\boldsymbol{D}_{b,c}^{1}$ is fixed to update the backbone network by minimizing L_{b1} as follows:

$$\begin{aligned} L_{b,ce2}(\boldsymbol{E}_m) = -\frac{1}{n_b}\sum_{i=1}^{n_b}[&\boldsymbol{q}_{t,i}^{hea}\log(\boldsymbol{D}_{b,c}^{1}(\boldsymbol{Z}_{t,i,12}^{hea})) \\ &+\boldsymbol{q}_{t,i}^{tor}\log(\boldsymbol{D}_{b,c}^{1}(\boldsymbol{Z}_{t,i,12}^{tor})) \\ &+\boldsymbol{q}_{t,i}^{leg}\log(\boldsymbol{D}_{b,c}^{1}(\boldsymbol{Z}_{t,i,12}^{leg}))], \end{aligned} \tag{8}$$

Same as $\boldsymbol{q}_{s,i}^{hea}$, $\boldsymbol{q}_{s,i}^{tor}$, and $\boldsymbol{q}_{s,i}^{leg}$, the values of $\boldsymbol{q}_{t,i}^{hea}$, $\boldsymbol{q}_{t,i}^{tor}$, and $\boldsymbol{q}_{t,i}^{leg}$ at the first, second, and third element positions are 1 respectively, and the elements at other positions are 0.

$$\begin{aligned} &L_{b,ce}(\boldsymbol{E}_m, \boldsymbol{W}_{hea}, \boldsymbol{W}_{tor}, \boldsymbol{W}_{leg}, \boldsymbol{W}_{c,hea}, \boldsymbol{W}_{c,tor}, \boldsymbol{W}_{c,leg}) \\ &\quad= -\frac{1}{3n_b}\Bigg(\sum_{i=1}^{n_b}\boldsymbol{q}_{s,i}[\log(\boldsymbol{W}_{hea}(\boldsymbol{z}_{s,i}^{hea})+\log(\boldsymbol{W}_{tor}(\boldsymbol{z}_{s,i}^{tor}) \\ &\qquad+\log(\boldsymbol{W}_{leg}(\boldsymbol{z}_{s,i}^{leg})]+\frac{1}{k_t}\sum_{c=1}^{k_t}\sum_{i=1}^{n_b}\boldsymbol{q}_{t,i}^{c}[\log(\boldsymbol{W}_{c,hea}(\boldsymbol{z}_{t,i}^{c,hea}) \\ &\qquad+\log(\boldsymbol{W}_{c,tor}(\boldsymbol{z}_{t,i}^{c,tor})+\log(\boldsymbol{W}_{c,leg}(\boldsymbol{z}_{t,i}^{c,leg})]\Bigg), \end{aligned} \tag{5}$$

$$\begin{aligned} L_{b,tri}(\boldsymbol{E}_m) = &\frac{1}{3n_b} \\ &\times\Bigg(\sum_{i=1}^{n_b}\log[1+\exp(\|\boldsymbol{z}_{s,i}^{hea}-\boldsymbol{z}_{s,i}^{hea,p}\|_2-\|\boldsymbol{z}_{s,i}^{hea}-\boldsymbol{z}_{s,i}^{hea,n}\|_2)] \\ &+\log[1+\exp(\|\boldsymbol{z}_{s,i}^{tor}-\boldsymbol{z}_{s,i}^{tor,p}\|_2-\|\boldsymbol{z}_{s,i}^{tor}-\boldsymbol{z}_{s,i}^{tor,n}\|_2)] \\ &+\log[1+\exp(\|\boldsymbol{z}_{s,i}^{leg}-\boldsymbol{z}_{s,i}^{leg,p}\|_2-\|\boldsymbol{z}_{s,i}^{leg}-\boldsymbol{z}_{s,i}^{leg,n}\|_2)] \\ &+\sum_{c=1}^{K_t}\sum_{i=1}^{n_b}\log[1+\exp(\|\boldsymbol{z}_{t,i}^{c,hea}-\boldsymbol{z}_{t,i}^{c,hea,p}\|_2-\|\boldsymbol{z}_{t,i}^{c,hea}-\boldsymbol{z}_{t,i}^{c,hea,n}\|_2)] \\ &+\log[1+\exp(\|\boldsymbol{z}_{t,i}^{c,tor}-\boldsymbol{z}_{t,i}^{c,tor,p}\|_2-\|\boldsymbol{z}_{t,i}^{c,tor}-\boldsymbol{z}_{t,i}^{c,tor,n}\|_2)] \\ &+\log[1+\exp(\|\boldsymbol{z}_{t,i}^{c,leg}-\boldsymbol{z}_{t,i}^{c,leg,p}\|_2-\|\boldsymbol{z}_{t,i}^{c,leg}-\boldsymbol{z}_{t,i}^{c,leg,n}\|_2)]\Bigg), \end{aligned} \tag{6}$$

$\boldsymbol{Z}^{hea}_{t,i,12}$, $\boldsymbol{Z}^{tor}_{t,i,12}$, $\boldsymbol{Z}^{leg}_{t,i,12}$ are the output of the last Transformer layer of $\boldsymbol{E}_m$.

2) Domain Alignment Based on Cooperation of Dual Discriminators: In the above process, the domain alignment ability of the backbone network depends on the discriminator $\boldsymbol{D}^1_{b,c}$. When the discriminability of $\boldsymbol{D}^1_{b,c}$ increases, the domain alignment capability of the backbone network improves. This paper introduces the second discriminator $\boldsymbol{D}^2_{b,c}$ that is different from the structure of the discriminator $\boldsymbol{D}^1_{b,c}$ to further enhance the domain alignment ability of the backbone network through collaborative training. Specifically, $\boldsymbol{D}^2_{b,c}$ is composed of four Transformer layers and four fully connected layers. Like the Transformer layer in $\boldsymbol{D}^1_{b,c}$, it is used to determine whether the information input into $\boldsymbol{D}^2_{b,c}$ contains target-domain information. In the optimization process, the parameters of the backbone network are fixed, and the function shown in Eq. (9) is used to optimize the parameters of $\boldsymbol{D}^2_{b,c}$ as follows:

$$\begin{aligned} L_{b,ce3}(\boldsymbol{D}^2_{b,c}) = -\frac{1}{3n_b}\sum_{i=1}^{n_b}[&\boldsymbol{q}^{hea}_{s,i}\log(\boldsymbol{D}^2_{b,c}(\boldsymbol{Z}^{hea}_{s,i,12})) \\ &+ \boldsymbol{q}^{tor}_{s,i}\log(\boldsymbol{D}^2_{b,c}(\boldsymbol{Z}^{tor}_{s,i,12})) + \boldsymbol{q}^{leg}_{s,i}\log(\boldsymbol{D}^2_{b,c} \\ &\times(\boldsymbol{Z}^{leg}_{s,i,12}))] + [\boldsymbol{c}^{hea}_{t,i}\log(\boldsymbol{D}^2_{b,c}(\boldsymbol{Z}^{hea}_{t,i,12})) \\ &+ \boldsymbol{c}^{tor}_{t,i}\log(\boldsymbol{D}^2_{b,c}(\boldsymbol{Z}^{tor}_{t,i,12})) \\ &+ \boldsymbol{c}^{leg}_{t,i}\log(\boldsymbol{D}^2_{b,c}(\boldsymbol{Z}^{leg}_{t,i,12}))], \end{aligned} \tag{9}$$

In the training process of the backbone network, this paper proposes the adversarial loss of the cooperation of two discriminators, which ensures that the backbone network extracts the discriminative features consistent with source domain. Therefore, the domain alignment between source domain and target domain is realized on feature representation. For this loss, the output results of $\boldsymbol{D}^1_{b,c}$ and $\boldsymbol{D}^2_{b,c}$ are added together as the final body part classification result. When updating the backbone network, the loss shown in Eq. (10) can be minimized, if and only if $\boldsymbol{D}^1_{b,c}$ and $\boldsymbol{D}^2_{b,c}$ output consistent discrimination results:

$$\begin{aligned} L_{b,ce4}(\boldsymbol{E}_m) = -\frac{1}{n_b}\sum_{i=1}^{n_b}[&\boldsymbol{q}^{hea}_{t,i}\log\frac{1}{2}(\boldsymbol{D}^1_{b,c}(\boldsymbol{Z}^{hea}_{t,i,12}) \\ &+ \boldsymbol{D}^2_{b,c}(\boldsymbol{Z}^{hea}_{t,i,12})) + \boldsymbol{q}^{tor}_{t,i}\log\frac{1}{2}(\boldsymbol{D}^1_{b,c}(\boldsymbol{Z}^{tor}_{t,i,12}) \\ &+ \boldsymbol{D}^2_{b,c}(\boldsymbol{Z}^{tor}_{t,i,12})) + \boldsymbol{q}^{leg}_{t,i}\log\frac{1}{2}(\boldsymbol{D}^1_{b,c}(\boldsymbol{Z}^{leg}_{t,i,12}) \\ &+ \boldsymbol{D}^2_{b,c}(\boldsymbol{Z}^{leg}_{t,i,12}))], \end{aligned} \tag{10}$$

In the proposed method, Eq. (10) is substituted for Eq. (8).

Due to the different structures of the two discriminators, the above process allows discriminators to discriminate the input features from two different views [38]. When the discrimination results of the two discriminators are consistent, the loss function can be minimized. This facilitates the backbone network to extract both features after domain alignment, and fine-grained features of different pedestrian parts, thereby enhancing the comprehensiveness and expression ability of the corresponding features. Additionally, the domain alignment of pedestrian body parts is realized in the same discriminator. Mode collapse caused by the independent design of task and domain classifiers can be avoided. The loss or distortion of features that may occur in the domain alignment process can also be alleviated.

Algorithm 1 Body Part-Level Domain Alignment (BPDA) for Domain-Adaptive Person Re-ID

Input: Labeled source samples $\boldsymbol{X}_s = \{x_{s,i}\}_{i=1}^{N}$, corresponding pedestrian labels $\boldsymbol{Y}_s = \{y_{s,i}\}_{i=1}^{N_s}$. Unlabeled target samples $\boldsymbol{X}_t = \{x_{t,i}\}_{i=1}^{N_t}$, corresponding intra-camera pedestrian labels $\boldsymbol{Y}_t = \{y^c_{t,i}\}_{i=1}^{n^c_t}$ and camera labels $\boldsymbol{C}_t = \{c_{t,i}\}_{i=1}^{K_t}$.
Output: The trained encoder $\boldsymbol{E}_m$.
Step I: Global and Body-part Feature Extraction (Sec.III.B)
1:Sample a batch of labeled source data.
2:**for** $c = 1, \cdots, K_t$ **do**
3: Sample a batch of c-th camera target data.
4:Initialize $\boldsymbol{E}_m$, $\boldsymbol{W}_g$, $\boldsymbol{W}_c$, $\boldsymbol{W}_{hea}$, $\boldsymbol{W}_{tor}$, $\boldsymbol{W}_{leg}$, $\boldsymbol{W}_{c,hea}$, $\boldsymbol{W}_{c,tru}$ and $\boldsymbol{W}_{c,leg}$.
5:**for** *iter*=1, $\cdots$, $Iteration_1$ **do**
6: Update $\boldsymbol{E}_m$, $\boldsymbol{W}_c$ and $\boldsymbol{W}_g$ by minimizing the loss in Eqs.(1) and (2).
7: Update $\boldsymbol{E}_m$, $\boldsymbol{W}_{hea}$, $\boldsymbol{W}_{tru}$, $\boldsymbol{W}_{leg}$, $\boldsymbol{W}_{c,hea}$, $\boldsymbol{W}_{c,tru}$ and $\boldsymbol{W}_{c,leg}$ by minimizing the loss in Eqs.(5) and (6).
8:**end for**
Step II: Body-part level Domain Alignment (Sec.III.C)
9:Sample a batch of labelled source data.
10:Sample a batch of unlabeled source data.
11:Load the learned $\boldsymbol{E}_m$, $\boldsymbol{W}_g$, $\boldsymbol{W}_c$, $\boldsymbol{W}_{hea}$, $\boldsymbol{W}_{tor}$, $\boldsymbol{W}_{leg}$, $\boldsymbol{W}_{c,hea}$, $\boldsymbol{W}_{c,tru}$ and $\boldsymbol{W}_{c,leg}$.
12:Initialize the classifier $\boldsymbol{D}^1_{b,c}$, $\boldsymbol{D}^2_{b,c}$;
13: **for** *iter*=1, $\cdots$, $Iteration_2$ **do**
14: Update $\boldsymbol{D}^1_{b,c}$ and $\boldsymbol{D}^2_{b,c}$ by minimizing the loss in Eqs.(7) and (9).
15: Update $\boldsymbol{E}_m$, $\boldsymbol{W}_g$, $\boldsymbol{W}_c$, $\boldsymbol{W}_{hea}$, $\boldsymbol{W}_{tor}$, $\boldsymbol{W}_{leg}$, $\boldsymbol{W}_{c,hea}$, $\boldsymbol{W}_{c,tru}$ and $\boldsymbol{W}_{c,leg}$ by minimizing the loss in Eqs.(1),(2),(5),(6) and (10).
16: **end for**

D. Entire Loss Function

The total loss function of network parameter optimization in this paper can be formulated as follows:

$$\begin{aligned} \boldsymbol{L} = &\boldsymbol{L}_{g,id} + \boldsymbol{L}_{g,ce} + \boldsymbol{L}_{b,ce} + \boldsymbol{L}_{b,tri} \\ &+ \lambda_1\boldsymbol{L}_{b,ce1} + \lambda_2\boldsymbol{L}_{b,ce3} + \lambda_3\boldsymbol{L}_{b,ce4}, \end{aligned} \tag{11}$$

where λ_1, λ_2 and λ_3 are hyperparameters. The above process is summarized in **Algo. 1**.

TABLE I
THE SETTINGS OF DIFFERENT PERSON RE-ID DATASETS. PN: NUMBER OF PEDESTRIANS, IN: NUMBER OF IMAGES, CN: NUMBER OF CAMERAS, POCN: NUMBER OF PEDESTRIANS CAPTURED BY ONLY ONE CAMERA

Datasets	CN	Training				Query(Testing)		Gallery(Testing)	
		POCN		Other PN	Other IN	PN	IN	PN	IN
		PN	IN						
Market1501-new	6	458	1732	159	1465	750	3368	750	15913
Duke-new	8	413	3184	140	2116	702	2228	1110	17661
MSMT17-new	15	1335	8376	455	6980	1041	2900	1041	29721
Market-SCT	6	751	3561	0	0	750	3368	750	15913
Duke-SCT	8	702	5993	0	0	702	2228	1110	17661

IV. EXPERIMENTS

A. Datasets and Evaluation Protocol

This paper uses five challenging datasets as the target dataset to verify the effectiveness and superiority of the proposed method, including Market1501-new [39], DukeMTMC-reID-new (Duke-new) [39], MSMT17-new [39], Market-SCT [7], and DukeMTMC-SCT (Duke-SCT) [7]. The training sets in Market1501-new, Duke-new and MSMT17-new are set according to the probability of pedestrians appearing under different cameras in a local camera network, Market-SCT and Duke-SCT are partitioned by Zhang et al [7]. These two datasets assume that the span between different cameras is relatively large, and each pedestrian in the training set appears in only one camera.

1) Market1501-New: The training set in Market1501-new is reset from the samples in the training set of Market1501 [40]. Specifically, the dataset assumes that pedestrians appear on one road. There are several intersections on this road. Since person re-ID focuses on pedestrian identity matching between cameras without overlapping, only one camera is installed at each intersection. According to the moving direction of each pedestrian, there is a 25% probability that pedestrians captured by each camera appear in other adjacent cameras. Therefore, the training set of Market1501-new contains 3,197 images of 617 pedestrians. Since many pedestrians change moving direction at one or multiple intersections, some of them are not captured by more than one camera. Therefore, 1,732 images of 458 pedestrians in the training set were captured by a single camera. This setting is close to the real-world scenes where pedestrians appear in each camera, so it is considerably challenging. In addition, the testing in Market1501-new is the same as that of Market1501.

2) Duke-New: The testing set in Duke-new is the same as the testing set in DukeMTMC-reID (Duke) [41]. As the main difference, the training set in Duke-new is reset by the samples in the training set of Duke according to the probability of pedestrians appearing under each camera in the local camera network. This dataset assumes that a camera is installed at each intersection and any pedestrian captured by each camera has a 25% probability of appearing in other adjacent cameras. As a result, the training set in Duke-new contains 5,300 images of 553 pedestrians, in which 3,184 images of 413 pedestrians were captured by a single camera.

3) MSMT17-New: The training set in MSMT17-new is re-partitioned according to the setting of Duke-new training set. As the main difference, the samples of MSMT17-new training set come from the samples of MSMT17 [42] testing set. Since the number of samples in the testing set of MSMT17 is more than that in the training set, the MSMT17 testing set samples are used to construct the MSMT17-new training set, thereby obtaining the MSMT17-new training set with a large sample size. Correspondingly, all the training samples of MSMT17 are used as the testing set samples of MSMT17-new. MSMT17-new contains 47,977 images of 2,831 pedestrians in total, of which the training set contains 15,356 images of 1,790 pedestrians, and the testing set contains 32,621 images of 1,041 pedestrians. In the training set, 8,376 images are composed of 1,335 pedestrians captured by only one camera.

4) Market-SCT: The training set of Market-SCT is also constructed by using the Market1501 training set samples under the new setting. Different from the above-mentioned training set construction method, each pedestrian in the Market-SCT training set is captured by only one camera. Since no pedestrian was captured by two or more cameras, all cross-camera samples in the Market-SCT training set come from different pedestrians. In this case, the Market-SCT training set contains 3,561 images of 751 pedestrians. According to the partition setting of the Market1501 testing set, the Market-SCT testing set consists of 19,281 images of 750 pedestrians.

5) Duke-SCT: The training set of Duke-SCT is composed of pedestrian samples in the Duke training set under the new partition setting. According to the Market-SCT partition setting, each pedestrian in the Duke-SCT training set was captured by only one camera. Therefore, each sample in the Duke-SCT training set does not have any positive sample across cameras. Therefore, the training set of Duket-SCT contains 5,993 images of 702 pedestrians in total. According to the original Duke partition setting, the testing set consists of 19,889 images of 702 pedestrians. The details of each dataset are shown in Tab.I.

6) Evaluation Protocol: Cumulative match characteristic (CMC) [43] and mean average precision (mAP) [40] are used to evaluate the performance of each method under a single query setting. They are also used to measure the accuracy of identity matching at each rank and the accuracy of overall retrieval, respectively.

B. Implementation Details

1) Network Settings: In the training process, the size of all images is uniformly adjusted to 256×128. Similar to [44], data augmentation is achieved through random cropping, random horizontal flipping, and image padding. In the experiments,

the batch size is set to 16 and each pedestrian has four images in a batch. All networks use SGD optimizer [45]. Momentum is set to 0.9. Weight decay is set to 1×10^{-4}. The learning rate of $\boldsymbol{E}_m$, $\boldsymbol{W}_c$, $\boldsymbol{W}_g$, $\boldsymbol{W}_{hea}$, $\boldsymbol{W}_{tor}$, $\boldsymbol{W}_{leg}$, $\boldsymbol{W}_{c,hea}$, $\boldsymbol{W}_{c,tor}$ and $\boldsymbol{W}_{c,leg}$ is set to 1.6×10^{-3}. The learning rate of the classifiers $\boldsymbol{D}_{b,c}^1$ and $\boldsymbol{D}_{b,c}^2$ is set to 1.2×10^{-3}. All hyperparameters λ_1, λ_2, and λ_3 are set to 1.0, 0.5, 2, respectively. The proposed method is implemented on the pytorch framework [46]. All experiments were done on a platform equipped with a single NVIDIA GeForce RTX 2080 Ti GPU.

2) Optimization: The entire model is totally trained for 180 epochs. In the 0∼10 epochs, the learning rate is linearly adjusted through the warm-up strategy [47]. In the 40-th epoch, the learning rate is decayed again at a 10% rate. In the training process, the first 50 epochs are used to update the parameters of $\boldsymbol{E}_m$ by minimizing the loss functions shown in Eqs. 1, 2, 5 and 6. In the remaining 130 epochs, the loss functions shown in Eqs. 7 and 9 are used to train the classifiers $\boldsymbol{D}_{b,c}^1$ and $\boldsymbol{D}_{b,c}^2$, and the loss functions shown in Eqs. 1, 2, 5, 6 and 10 are used to further update the parameters of $\boldsymbol{E}_m$. In the testing process, Euclidean distance is used to match pedestrian identity. The specific optimization algorithm is shown in Algo.1.

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

In this section, the proposed method is compared with state-of-the-art unsupervised person re-ID methods to verify its effectiveness and superiority. There are three main types of methods involved in performance comparison: fully unsupervised person re-ID (FUPR) methods, person re-ID methods based on domain generalization (DG), and unsupervised domain-adaptive (UDA) person re-ID methods. For FUPR methods, source-domain data is not generally available. Therefore, the related models are only obtained by unsupervised training on target dataset. For person re-ID methods based on DG, the target dataset is unknown, and only the labeled source-domain data can be applied to model training. For UDA person re-ID methods, both source-domain and target-domain data is known. But only the source-domain data is already labeled. The target-domain data is not labeled. Generally, UDA person re-ID methods can achieve stable performance because both source-domain and unlabeled target-domain data participates in model training. The performance comparison between this method and the above two types of method is mainly used to illustrate the advantages of UDA. The proposed method is compared with state-of-the-art UDA methods to verify its effectiveness and superiority over existing methods.

Market1501-new and Duke-new are set to verify model performance in the local monitoring networks. In the above two scenes, there are cross-camera paired training samples between different adjacent cameras. However, as the span between cameras increases, the number of pedestrians with the same identity appearing across cameras gradually decreases, which is highly consistent with the situation in real-world scenes. Tab.II shows the re-ID performance obtained by different methods on both Market1501-new and Duke-new datasets. The results listed in Tab.II were obtained by using the codes provided by the corresponding original authors and retraining the corresponding models under the original parameter settings. All comparative experiments only used a set of optimal preset parameters provided by their authors in their papers.

For FUPR methods, IICS obtained the best re-ID performance on Duke→Market1501-new (Market1501→Duke-new), and the corresponding re-ID accuracy of rank-1 and mAP reached 61.5% and 34.2% (54.5% and 33.7%), respectively. In contrast, the re-ID accuracy of the proposed method on rank-1 and mAP reached 78.1% and 51.7% (68.7% and 48.2%) respectively, which was significantly better than the corresponding re-ID performance of IICS. There are two main reasons. On one hand, UDA methods apply source-domain samples to model training, so they naturally have a certain expansion capability. On the other hand, the number of cameras that can capture the same pedestrian is different in Market1501-new and Duke-new, resulting in an imbalance in the number of cross-camera samples of different pedestrians in the training set. In addition, pedestrians only appearing in a single camera increase the risk of noisy label introduction. These factors limit the performance of the FUPR methods. Since DG-based methods assume that the target dataset is unknown, the above problems do not affect their performance. However, the methods based on DG require multiple source-domain datasets on model training. If only a single dataset is applied to model training, only 69.2% and 35.9%(55.2% and 33.1%) re-ID accuracy can be obtained on rank-1 and mAP.

The proposed method is a domain-adaptive method. In this paper, the performance of the proposed method is compared with the corresponding ones of state-of-the-art domain-adaptive methods. According to Tab.II, on Duke→ Market1501-new (Market1501→ Duke-new), the re-ID accuracy obtained by the proposed BPDA on Rank-1 and mAP reached 78.1% and 51.7% (68.7% and 48.2%) respectively. When the intra-camera sample labels are involved in model training, the performance of the proposed method (BPDA+) is further improved. However, the highest re-ID accuracy obtained by domain-adaptive methods based on pseudo-label prediction is only 59.7% and 33.7%(50.5% and 30.7%) on Rank-1 and mAP, respectively. The re-ID accuracy of the latest IDM method only reached 47.5% and 25.5% (32.6% and 20.3%) on Rank-1 and mAP, respectively. The proposed BPDA exceeds the re-ID accuracy of IDM by 30.6% and 26.2% (36.1% and 27.9%) on Rank-1 and mAP, respectively. The main reason is that a large number of pedestrians only appearing under a single camera and the imbalance of different pedestrian samples in both Market1501-new and Duke-new datasets. The above results confirm that the proposed method has stronger practical value than existing methods. To further verify the above statement, this paper tests the performance of different algorithms on Duke→ MSMT-new (Market1501→ MSMT-new). According to Tab. III, the same conclusion consistent with the above statement can be drawn, which further verifies the effectiveness of the proposed method and its superiority over existing methods.

To further verify the practicability of the proposed method, the following experiments use Market1501 and Duke as the

TABLE II

THE PERFORMANCE COMPARISON BETWEEN THE PROPOSED METHOD AND STATE-OF-THE-ART METHODS WHEN MARKET1501-NEW AND DUKE-NEW ARE USED AS TARGET DOMAIN DATA. BOTH CMC AND MAP RATE (%) OBTAINED BY EACH METHOD ARE LISTED. THE BEST RESULTS ARE MARKED IN BOLD FONT. BPDA INDICATES THAT THE PROPOSED MODEL WAS NOT TRAINED USING THE TARGET-DOMAIN INTRA-CAMERA SAMPLE LABELS, BPDA+ INDICATES THAT THE TARGET-DOMAIN INTRA-CAMERA SAMPLE LABELS ARE INVOLVED IN MODEL TRAINING

Methods	Target: Market1501-new					Target: Duke-new				
	Source	Rank-1	Rank-5	Rank-10	mAP	Source	Rank-1	Rank-5	Rank-10	mAP
FUPR										
IICS(CVPR'21) [48]	None	61.5	80.1	86.2	34.2	None	54.5	68.7	74.4	33.7
STS(arXiv'21) [49]	None	20.9	34.0	40.3	8.7	None	37.5	49.7	56.4	21.3
ICE(ICCV'21) [50]	None	35.2	50.3	56.4	16.0	None	29.3	40.7	46.1	17.5
DG										
MetaBIN(CVPR'21) [51]	Duke	69.2	83.1	87.8	35.9	Market1501	55.2	69.0	74.4	33.1
Mixstyle(ICLR'21) [52]	Duke	58.2	74.9	80.9	29.0	Market1501	48.2	62.7	68.4	27.2
UDA										
MMT-500(ICLR'20) [10]	Duke	59.7	75.1	80.9	33.7	Market1501	45.4	61.0	67.6	30.8
MMT-700(ICLR'20) [10]	Duke	57.3	73.5	80.2	32.0	Market1501	45.2	60.5	67.0	30.5
MMT-900(ICLR'20) [10]	Duke	58.4	74.3	80.3	32.5	Market1501	43.9	61.0	67.1	30.8
SPCL(NeurIPS'20) [53]	Duke	14.1	26.1	33.0	5.6	Market1501	13.2	21.5	25.3	5.5
Meb-Net(ECCV'20) [54]	Duke	57.3	73.0	79.1	33.4	Market1501	44.2	59.1	65.4	30.7
CAC(INS'21) [55]	Duke	58.9	75.0	80.2	28.2	Market1501	50.5	65.0	70.4	30.7
IDM (ICCV'21) [4]	Duke	47.5	62.8	68.9	25.5	Market1501	32.6	47.0	54.5	20.3
Dual-Refine(TIP'21) [5]	Duke	56.1	70.2	76.2	34.3	Market1501	41.9	55.9	63.3	29.8
BPDA(Proposed)	Duke	**78.1**	**89.2**	**92.6**	**51.7**	Market1501	**68.7**	**80.7**	**84.5**	**48.2**
BPDA+(Proposed)	Duke	**83.4**	**92.5**	**95.5**	**64.6**	Market1501	**82.5**	**90.4**	**93.0**	**67.7**

TABLE III

THE PERFORMANCE COMPARISON BETWEEN THE PROPOSED METHOD AND STATE-OF-THE-ART METHODS WHEN MSMT-NEW USED AS TARGET DOMAIN DATA. BOTH CMC AND MAP RATE (%) OBTAINED BY EACH METHOD ARE LISTED. THE BEST RESULTS ARE MARKED IN BOLD FONT. BPDA INDICATES THAT THE PROPOSED MODEL WAS NOT TRAINED USING THE TARGET-DOMAIN INTRA-CAMERA SAMPLE LABELS, BPDA+ INDICATES THAT THE TARGET-DOMAIN INTRA-CAMERA SAMPLE LABELS ARE INVOLVED IN MODEL TRAINING

Methods	MSMT17-new					MSMT17-new				
	Source	Rank-1	Rank-5	Rank-10	mAP	Source	Rank-1	Rank-5	Rank-10	mAP
FUPR										
IICS(CVPR'21) [48]	None	45.1	58.3	64.0	20.4	None	45.1	58.3	64.0	20.4
STS(arXiv'21) [49]	None	42.9	56.0	61.3	18.7	None	42.9	56.0	61.3	18.7
ICE(ICCV'21) [50]	None	37.9	49.2	55.9	17.6	None	37.9	49.2	55.9	17.6
DG										
MetaBIN(CVPR'21) [51]	Duke	44.7	58.2	63.5	17.2	Market1501	39.9	52.7	58.1	15.2
Mixstyle(ICLR'21) [52]	Duke	31.4	45.4	51.3	12.2	Market1501	25.3	39.2	44.7	9.8
UDA										
MMT-1000(ICLR'20) [10]	Duke	36.5	50.4	55.9	16.3	Market1501	30.7	43.9	50.9	13.7
MMT-2000(ICLR'20) [10]	Duke	37.8	51.7	57.5	17.2	Market1501	33.9	47.5	54.9	15.4
SPCL(NeurIPS'20) [53]	Duke	19.8	31.7	37.7	8.8	Market1501	18.8	30.4	36.6	9.0
Meb-Net(ECCV'20) [54]	Duke	33.9	46.5	52.2	15.9	Market1501	26.1	37.3	43.5	12
CAC(INS'21) [55]	Duke	33.5	44.1	50.1	12.4	Market1501	20.7	32.7	38.1	7.6
IDM(ICCV'21) [4]	Duke	32.5	43.3	50.0	14.7	Market1501	31.0	42.7	48.6	14.6
Dual-Refine(TIP'21) [5]	Duke	28.0	40.0	45.4	12.1	Market1501	27.2	39.2	46.1	12.3
BPDA(Proposed)	Duke	**50.9**	**63.9**	**69.4**	**25.2**	Market1501	**48.0**	**61.3**	**67.2**	**23.0**
BPDA+(Proposed)	Duke	**65.7**	**76.9**	**81.6**	**41.3**	Market1501	**67.6**	**77.9**	**82.1**	**42.1**

source domain and Duke-SCT and Market-SCT as the target domain. Both Duke-SCT and Market-SCT are set to simulate large-scale camera networks. As three assumptions, the distance span between different cameras is large, each pedestrian in the training set appears in a single camera, and there is no sample of the same pedestrian across cameras. In such training sets, the labels predicted by clustering-based pseudo-label prediction methods are all noisy labels. The following experiments use such datasets as the target datasets to evaluate the adaptability of different methods in different large-scale surveillance networks.

As shown in Tab. IV, except IICS, the overall performance of FUPR methods and UDA-based methods decreased. The methods based on clustering pseudo-label prediction are weak in adapting to this type of scene. In contrast, the proposed method has stronger stability. On Duke→ Market-SCT, the re-ID accuracy of rank-1 and mAP obtained by the proposed method can still reach 77.2% and 51.3%, respectively. Additionally, on Marke→ Duke-SCT, the re-ID accuracy of rank-1 and mAP obtained by the proposed method reached 68.6% and 47.8%, respectively. As a specially designed method MCNL for the problems involved in both Market-SCT and Duke-SCT, its performance is surpassed by the proposed method.

D. Ablation Study

The proposed method is composed of a global and body part feature extraction module (GBFE) and a body part level domain alignment module (BPDA). GBFE consists of GFE and BFE. In this paper, the GEF trained by minimizing $L_{g,ce}(\boldsymbol{E}_m, \boldsymbol{W}_g)$ and $L_{g,tri}(\boldsymbol{E}_m)$ is regarded as the baseline.

TABLE IV

THE PERFORMANCE COMPARISON BETWEEN THE PROPOSED METHOD AND STATE-OF-THE-ART METHODS WHEN MARKET1501-NEW AND DUKE-NEW ARE USED AS TARGET DOMAIN DATA. BOTH CMC AND MAP RATE (%) OBTAINED BY EACH METHOD ARE LISTED. THE BEST RESULTS ARE MARKED IN BOLD FONT. BPDA INDICATES THAT THE PROPOSED MODEL WAS NOT TRAINED USING THE TARGET-DOMAIN INTRA-CAMERA SAMPLE LABELS, BPDA+ INDICATES THAT THE TARGET-DOMAIN INTRA-CAMERA SAMPLE LABELS ARE INVOLVED IN MODEL TRAINING

Methods	Market-SCT					Duke-SCT				
	Source	Rank-1	Rank-5	Rank-10	mAP	Source	Rank-1	Rank-5	Rank-10	mAP
FUPR										
MCNL(AAAI'20) [7]	None	67.0	82.8	87.9	41.6	None	67.1	80.9	84.7	45.2
IICS(CVPR'21) [48]	None	64.3	81.2	86.8	36.4	None	47.7	62.1	68.1	26.5
STS(arXiv'21) [49]	None	21.1	34.3	41.3	8.5	None	33.0	45.8	50.9	18.4
ICE(ICCV'21) [50]	None	29.3	41.1	47.2	13.4	None	20.4	28.2	33.0	11.6
CCFP(ACMMM'21) [37]	None	82.4	92.6	95.4	63.9	None	80.3	89.0	91.9	64.5
DG										
MetaBIN(CVPR'21) [51]	Duke	69.2	83.1	87.8	35.9	Market1501	55.2	69.0	74.4	33.1
Mixstyle(ICLR'21) [52]	Duke	58.2	74.9	80.9	29.0	Market1501	48.2	62.7	68.4	27.2
UDA										
MMT-500(ICLR'20) [10]	Duke	50.0	68.0	75.9	27.8	Market1501	38.9	56.3	63.5	26.8
MMT-700(ICLR'20) [10]	Duke	49.1	66.9	74.3	27.7	Market1501	40.9	58.1	65.5	29.2
MMT-900(ICLR'20) [10]	Duke	51.0	70.0	76.9	28.5	Market1501	42.3	59.6	67.6	30.4
SPCL(NeurIPS'20) [53]	Duke	11.5	23.5	30.2	4.5	Market1501	12.3	19.7	24.2	5.6
Meb-Net(ECCV'20) [54]	Duke	54.4	71.1	78.1	30.7	Market1501	41.6	58.1	64.0	27.8
CAC(INS'21) [55]	Duke	62.1	76.6	81.8	30.6	Market1501	49.6	64.0	69.8	30.0
IDM (ICCV'21) [4]	Duke	32.3	48.3	56.1	14.3	Market1501	37.9	51.2	58.4	23.6
Dual-Refine(TIP'21) [5]	Duke	47.7	63.4	70.1	23.3	Market1501	39.8	53.4	60.2	28.1
BPDA(Proposed)	Duke	**77.2**	**89.0**	**92.2**	**51.3**	Market1501	**68.6**	**80.7**	**84.1**	**47.8**
BPDA+(Proposed)	Duke	**84.0**	**92.9**	**96.0**	**66.0**	Market1501	**81.6**	**89.9**	**92.8**	**67.8**

TABLE V

COMPARISON OF EXPERIMENTAL PERFORMANCE ON DUKE→MARKET1501-NEW AND MARKET1501→DUKE-NEW AFTER ADDING DIFFERENT MODULES. DUKE IS SHORT FOR DUKEMTMC

Methods	Duke→Market1501-new				Market1501→Duke-new			
	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP
Baseline	70.1	83.5	88.2	42.5	66.6	77.3	82.3	45.7
Baseline+BFE	75.3	87.5	91.3	49.4	67.9	79.8	83.2	47.6
Baseline+CDD	75.4	88.1	92.0	49.1	67.9	79.9	83.8	48.1
Baseline+BFE+CDD	78.1	89.2	92.6	51.7	68.7	80.7	84.5	48.2
Baseline+BFE+CDD+ICL	83.4	92.5	95.5	64.6	82.5	90.4	93.0	67.7

The model after adding discriminator $\boldsymbol{D}^1_{b,c}$ and BFE to the baseline is denoted as Baseline+BFE. The cooperation training of dual discriminators (CDD) is introduced into the Baseline, and the corresponding model is denoted as Baseline+CDD. The model obtained by introducing the CDD into Baseline+BFE is denoted as Baseline+BFE+CDD. The model Baseline+BFE+CDD with the intra-camera sample labels (ICL) of the target-domain dataset involved in training is denoted as Baseline+BFE+CDD+ICL. Tab.V shows the experimental results of different models on Duke→ Market1501-new and Market1501 →Duke-new under different settings.

1) Effectiveness of BFE: Baseline+BFE and Baseline are compared to verify the effectiveness of BFE. According to Tab.V, the fine-grained domain alignment at part level can effectively improve model performance. Specifically, on Duke→Market1501-new (Market1501→Duke-new), the Rank-1 recognition accuracy increases from 70.1% (66.6%) to 75.3% (67.9%), and the mAP recognition accuracy increases from 42.5% (45.7%) to 49.4% (47.6%). With the assistance of a single discriminator $\boldsymbol{D}^1_{b,c}$, BFE can improve model's adaptability in target domain to a certain extent, so that the model achieves better performance.

2) Effectiveness of CDD: To verify the validity of CDD, CDD is added to baseline and Baseline+BFE. According to Tab. V, on Duke→Market1501-new (Market1501→Duke-new), compared with Baseline+BFE, Baseline+BFE+CDD improves the recognition accuracy of Rank-1 from 75.3% (67.9%) to 78.1% (68.7%), and also improves the recognition accuracy of mAP from 49.4% (47.6%) to 51% (48.2%). The introduction of the CDD-based adversarial mechanism effectively improves the discriminator's discriminating ability, so the feature distribution alignment is achieved in both source and target domains, and the model's adaptability is improved in target domain.

3) Effectiveness of ICL: To verify the effectiveness of intra-camera sample label (ICL), Baseline+BFE+CDD+ICL is compared with Baseline+BFE+CDD. According to Tab. V, Baseline+BFE+CDD+ICL improves the Rank-1 recognition accuracy from 78.1% (68.7%) to 83.4% (82.5%), and also improves the mAP recognition accuracy from 51.7% (48.2%) to 64.6% (67.7%). It confirms that the intra-camera label participation in training can effectively improve model's feature representation ability. Fig.4 further shows the pedestrian retrieval results under different ablation settings. The contribution of each module is intuitively demonstrated.

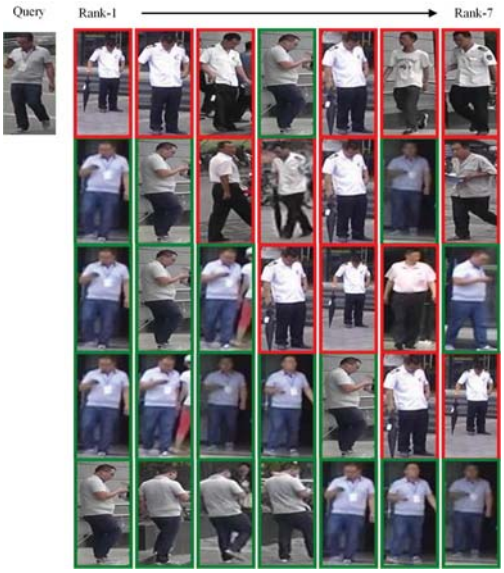


Fig. 4. From top to bottom, the first row to the fifth row represent the results retrieved by Baseline, Baseline+BFE, Baseline+CDD, Baseline+BFE+CDD, and Baseline+BFE+CDD+ICL, respectively.

TABLE VI

COMPARISON OF EXPERIMENTAL PERFORMANCE ON DUKE→MARKET1501-NEW AND MARKET1501→DUKE-NEW AFTER ADDING DIFFERENT MODULES. DUKE IS SHORT FOR DUKEMTMC

Methods	Duke→Market1501-new			Market1501→Duke-new		
	Rank-1	Rank-5	mAP	Rank-1	Rank-5	mAP
D6	76.5	88.8	51.4	67.2	79.0	46.9
D3	77.3	89.1	51.0	68.4	79.9	47.9
Proposed	78.1	89.2	51.7	68.7	80.7	48.2

E. Further Discussion

According to the dissimilarity of different pedestrian parts, the proposed method divides pedestrian features into three parts: head, torso, and legs to achieve part-level domain alignment. Compared with the traditional vertical average dividing features, two experiments "D6" and "D3" are designed in this section to verify the superiority of dividing pedestrian features by parts. "D6" and "D3" indicate that the pedestrian features are divided into 6 blocks and 3 blocks on average, respectively. As shown in Tab.VI, the more blocks are evenly divided in vertical direction, the lower performance is exhibited. Vertical average division ignores the dissimilarity of different parts, resulting in the loss of some key features of the same parts of different blocks during the model training process. This reduces the discriminability of features. Since the proposed method considers the dissimilarity of different body parts, it achieves better performance than the above two experiment settings. None of the experiments shown in Tab.VI involves target-domain intra-camera labels.

Additionally, the proposed body-part-level domain alignment facilitates the network to extract features from the entire pedestrian image. The first and second rows in Fig. 5 give the attention maps (heatmaps) of Baseline+CDD and Baseline+BFE+CDD, respectively, where red indicates that the model pays more attention to this region. According to these results, Baseline+CDD cannot always pay attention to the entire pedestrian body without the division of pedestrian body parts. In this case, the extracted features are not complete, which are not conducive to improving the discriminability of pedestrian features. When dividing the body parts of pedestrian images, in order to achieve correct classification of different parts, the network is encouraged to extract discriminative features from different pedestrian parts, which makes the pedestrian features extracted by Baseline+BFE+CDD more complete. Therefore, Baseline+BFE+CDD achieves higher recognition performance than Baseline+CDD.

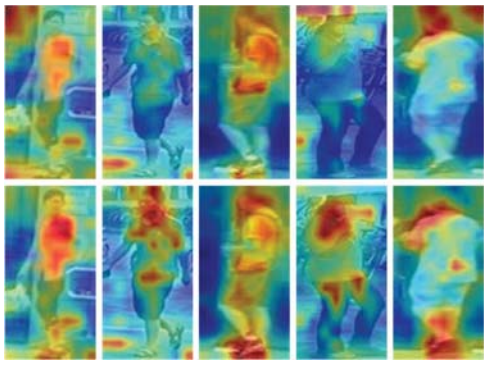


Fig. 5. The impact of pedestrian body part alignment on feature extraction. The first row shows the attention maps (heat maps) of Baseline+CDD, and the second row shows the attention maps (heat maps) of Baseline+BFE+CDD. The strong activation regions are marked in red.

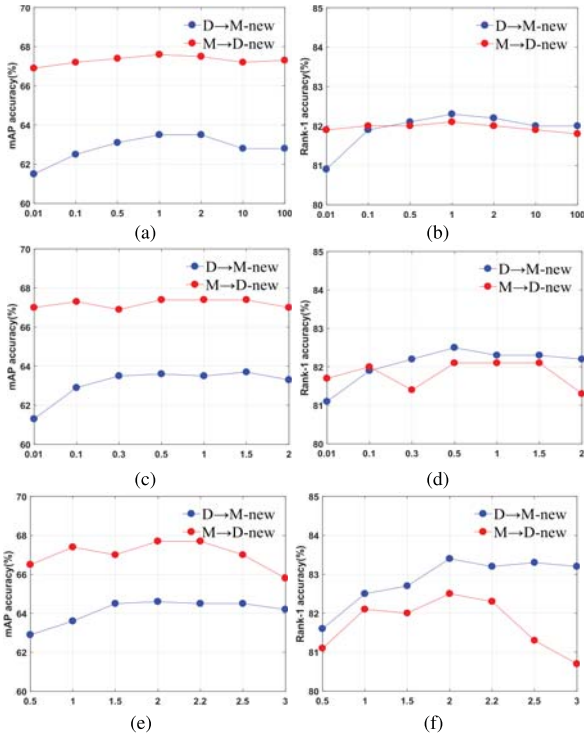


Fig. 6. The result of Rank-1 and mAP under the changes of different parameters. D→M-new denotes DukeMTMC→Market1501-new and M→D-new denotes Market1501→DukeMTMC-new. (a) shows the result of λ_1,(b) shows the result of λ_2,(c) shows the result of λ_3.

F. Parameter Selection and Analysis

The proposed method contains three hyperparameters λ_1, λ_2 and λ_3. According to Eq. (11), they adjust $L_{b,ce1}(D^1_{b,c})$,

$L_{b,ce3}(D^2_{b,c})$ and $L_{b,ce4}(\boldsymbol{E}_m)$ respectively. In hyperparameter analysis, two parameters are fixed to analyze the effect of another parameter on experimental performance.

1) The influence of λ_1: Fig.6(a) shows the change in model performance when $\lambda_1 \in [0.01, 100]$. On D→M-new, when $\lambda_1 = 1$, the proposed method achieves the highest recognition accuracy of Rank-1 and mAP, and when $\lambda_1 > 1$, the performance of the proposed method decreases slightly. For M→D-new, when $\lambda_1 = 1$, the proposed method also achieves the highest recognition accuracy of Rank-1 and mAP. Therefore, $\lambda_1 = 1$ is the optimal choice.

2) The influence of λ_2: Fig.6(b) shows the effect of different values of λ_2 on the recognition accuracy of Rank-1 and mAP on D→M-new and M→D-new. When $\lambda_2 \in [0.01, 0.5]$, the recognition accuracy of Rank-1 and mAP obtained by the proposed method on the two tasks shows an overall improvement. When $\lambda_2 \in [0.5, 2]$, the recognition accuracy of Rank-1 and mAP obtained by the proposed method on the two tasks decreases. Therefore, $\lambda_2 = 0.5$ is the optimal choice.

3) The influence of λ_3: The hyperparameters λ_1 and λ_2 are fixed. λ_3 is taken values within the range of $[0.5, 3]$. On D→M-new and M→D-new, the changes in the recognition accuracy of Rank-1 and mAP with different values of are shown in Fig. 6(c). When $\lambda_3 = 2$, the proposed method achieves the best performance on both D→M-new and M→D-new. Therefore, it is reasonable to set λ_3 to 2.

V. Conclusion

To get rid of the dependence of pseudo-label prediction-based domain-adaptive methods on the reliability of pseudo labels, this paper designs a Transformer framework for domain alignment at body part level. This framework aggregates the local features from the same body part by the Transformer to obtain the classification token for the body part, and uses it as the global representation of different body parts. Additionally, an adversarial strategy embedded in Transformer layer is designed. This strategy makes full use of the different structures and morphologies of different pedestrian body parts (e.g., head, torso, and legs) to achieve pedestrian body-part-level domain alignment. Additionally, the proposed method does not need to perform pseudo-label prediction for target samples, and gets rid of the influence of noisy labels on recognition performance. The proposed method shows excellent performance on the datasets that are in line with the actual scene settings, which proves that the proposed method has strong applicability.

Acknowledgment

The source codes of this paper will be available at https://github.com/lhf12278/BPDA.

References

[1] K. Wang, P. Wang, C. Ding, and D. Tao, "Batch coherence-driven network for part-aware person re-identification," *IEEE Trans. Image Process.*, vol. 30, pp. 3405–3418, 2021.

[2] 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.

[3] H. Tang, C. Yuan, Z. Li, and J. Tang, "Learning attention-guided pyramidal features for few-shot fine-grained recognition," *Pattern Recognit.*, vol. 130, Oct. 2022, Art. no. 108792.

[4] Y. Dai, J. Liu, Y. Sun, Z. Tong, C. Zhang, and L.-Y. Duan, "IDM: An intermediate domain module for domain adaptive person Re-ID," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2021, pp. 11864–11874.

[5] Y. Dai, J. Liu, Y. Bai, Z. Tong, and L.-Y. Duan, "Dual-refinement: Joint label and feature refinement for unsupervised domain adaptive person re-identification," *IEEE Trans. Image Process.*, vol. 30, pp. 7815–7829, 2021.

[6] H. Tang, Z. Li, Z. Peng, and J. Tang, "BlockMix: Meta regularization and self-calibrated inference for metric-based meta-learning," in *Proc. 28th Int. ACM Multimedia*, 2020, pp. 610–618.

[7] T. Zhang, L. Xie, L. Wei, Y. Zhang, B. Li, and Q. Tian, "Single camera training for person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 7, 2020, pp. 12878–12885.

[8] H. Tang and K. Jia, "Discriminative adversarial domain adaptation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 4, 2020, pp. 5940–5947.

[9] F. Yang *et al.*, "Asymmetric co-teaching for unsupervised cross-domain person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 7, 2020, pp. 12597–12604.

[10] 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. 1–15.

[11] Y. Zhai *et al.*, "AD-cluster: Augmented discriminative clustering for domain adaptive person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9018–9027.

[12] K. Zeng, M. Ning, Y. Wang, and Y. Guo, "Hierarchical clustering with hard-batch triplet loss for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13654–13662.

[13] F. Zhao, S. Liao, G.-S. Xie, J. Zhao, K. Zhang, and L. Shao, "Unsupervised domain adaptation with noise resistible mutual-training for person re-identification," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 526–544.

[14] Y. Luo, C. Zhang, M. Zhao, H. Zhou, and J. Sun, "Where, what, whether: Multi-modal learning meets pedestrian detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 14065–14073.

[15] 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.

[16] 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.

[17] 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. 7195–7204.

[18] Y. Tian, X. Peng, L. Zhao, S. Zhang, and D. N. Metaxas, "CR-GAN: Learning complete representations for multi-view generation," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, 2018, pp. 942–948.

[19] 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.

[20] Y.-J. Li, C.-S. Lin, Y.-B. Lin, and Y.-C.-F. Wang, "Cross-dataset person re-identification via unsupervised pose disentanglement and adaptation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7918–7928.

[21] 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/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8079–8088.

[22] 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.

[23] H. Li, Y. Chen, D. Tao, Z. Yu, and G. Qi, "Attribute-aligned domain-invariant feature learning for unsupervised domain adaptation person re-identification," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 1480–1494, 2021.

[24] H. Li, N. Dong, Z. Yu, D. Tao, and G. Qi, "Triple adversarial learning and multi-view imaginative reasoning for unsupervised domain adaptation person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 5, pp. 2814–2830, May 2022.

[25] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 5998–6008.

[26] H. Chen *et al.*, "Pre-trained image processing transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12299–12310.

[27] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 213–229.

[28] E. Xie *et al.*, "Segmenting transparent objects in the wild with transformer," in *Proc. 13th Int. Joint Conf. Artif. Intell.*, Aug. 2021, pp. 1194–1200.

[29] P. Sun *et al.*, "TransTrack: Multiple object tracking with transformer," 2020, *arXiv:2012.15460*.

[30] S. He, H. Luo, P. Wang, F. Wang, H. Li, and W. Jiang, "TransReID: Transformer-based object re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 15013–15022.

[31] S. Lai, Z. Chai, and X. Wei, "Transformer meets part model: Adaptive part division for person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 4133–4140.

[32] K. Zhu *et al.*, "AAformer: Auto-aligned transformer for person re-identification," 2021, *arXiv:2104.00921*.

[33] Z. Ma, Y. Zhao, and J. Li, "Pose-guided inter- and intra-part relational transformer for occluded person re-identification," in *Proc. 29th ACM Int. Conf. Multimedia*, 2021, pp. 1487–1496.

[34] Y. Li, J. He, T. Zhang, X. Liu, Y. Zhang, and F. Wu, "Diverse part discovery: Occluded person re-identification with part-aware transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Oct. 2021, pp. 2898–2907.

[35] S. Liao and L. Shao, "TransMatcher: Deep image matching through transformers for generalizable person re-identification," 2021, *arXiv:2105.14432*.

[36] T. Liang *et al.*, "CMTR: Cross-modality transformer for visible-infrared person re-identification," 2021, *arXiv:2110.08994*.

[37] W. Ge, C. Pan, A. Wu, H. Zheng, and W.-S. Zheng, "Cross-camera feature prediction for intra-camera supervised person re-identification across distant scenes," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 3644–3653.

[38] Y. Luo, L. Zheng, T. Guan, J. Yu, and Y. Yang, "Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2507–2516.

[39] H. Li, K. Xu, J. Li, and Z. Yu, "Dual-stream reciprocal disentanglement learning for domain adaptation person re-identification," *Knowl.-Based Syst.*, vol. 251, Sep. 2022, Art. no. 109315.

[40] 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.

[41] 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.

[42] 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.

[43] X. Wang, G. Doretto, T. Sebastian, J. Rittscher, and P. Tu, "Shape and appearance context modeling," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, Oct. 2007, pp. 1–8.

[44] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 13001–13008.

[45] H. Robbins and S. Monro, "A stochastic approximation method," *Ann. Math. Statist.*, vol. 22, no. 3, pp. 400–407, Sep. 1951.

[46] A. Paszke *et al.*, "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, Dec. 2019, pp. 8026–8037.

[47] 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, pp. 1487–1495.

[48] S. Xuan and S. Zhang, "Intra-inter camera similarity for unsupervised person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 11926–11935.

[49] T. Liu, Y. Lin, and B. Du, "Unsupervised person re-identification with stochastic training strategy," 2021, *arXiv:2108.06938*.

[50] H. Chen, B. Lagadec, and F. Bremond, "ICE: Inter-instance contrastive encoding for unsupervised person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 14960–14969.

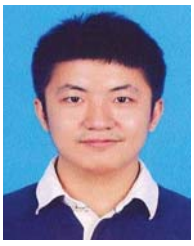
[51] S. Choi, T. Kim, M. Jeong, H. Park, and C. Kim, "Meta batch-instance normalization for generalizable person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 3425–3435.

[52] K. Zhou, Y. Yang, Y. Qiao, and T. Xiang, "Domain generalization with MixStyle," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021, pp. 1–15.

[53] Y. Ge, F. Zhu, D. Chen, R. Zhao, and H. Li, "Self-paced contrastive learning with hybrid memory for domain adaptive object Re-ID," in *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, 2020, pp. 1–13.

[54] Y. Zhai, Q. Ye, S. Lu, M. Jia, R. Ji, and Y. Tian, "Multiple expert brainstorming for domain adaptive person re-identification," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2020, pp. 594–611.

[55] H. Li, J. Pang, D. Tao, and Z. Yu, "Cross adversarial consistency self-prediction learning for unsupervised domain adaptation person re-identification," *Inf. Sci.*, vol. 559, pp. 46–60, Jun. 2021.



Yiming Wang received the B.E. degree from the College of Computer Science, Chongqing University of Posts and Telecommunications, China, in 2014. He is currently pursuing the Ph.D. degree with the College of Automation, Chongqing University, China. His research interests include image processing, computer vision, fault diagnosis, and reliability analysis.

Guanqiu Qi received the Ph.D. degree in computer science from Arizona State University in 2014. He is currently an Assistant Professor with the Computer Information Systems Department, State University of New York at Buffalo State. His research interests include deep learning, machine learning, and image processing. He also spans 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.

Shuang Li received the B.E. degree from the College of Software Engineering, Chongqing University of Posts and Telecommunications, China, in 2019. He is currently pursuing the master's degree with the College of Information Engineering and Automation, Kunming University of Science and Technology, China. His research interests include machine learning and computer vision.

Yi Chai received the B.E. degree from the Department of Electronic Engineering, National University of Defense Technology, Changsha, China, in 1982, and the M.S. and Ph.D. degrees from the Department of Automation, Chongqing University, Chongqing, China, in 1994 and 2001, respectively. He is currently a Professor and a Ph.D. Advisor at Chongqing University. His research interests include nonlinear dynamic systems, signal processing, information fusion, fault detection and diagnosis, and intelligence systems.

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 at 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.