Logical Relation Inference and Multiview Information Interaction for Domain Adaptation Person Re-Identification

Shuang Li, Fan Li, Jinxing Li, *Member, IEEE*, Huafeng Li, Bob Zhang, *Senior Member, IEEE*, Dapeng Tao, *Member, IEEE*, and Xinbo Gao, *Senior Member, IEEE*

***Abstract*— Domain adaptation person re-identification (Re-ID) is a challenging task, which aims to transfer the knowledge learned from the labeled source domain to the unlabeled target domain. Recently, some clustering-based domain adaptation Re-ID methods have achieved great success. However, these methods ignore the inferior influence on pseudo-label prediction due to the different camera styles. The reliability of the pseudo-label plays a key role in domain adaptation Re-ID, while the different camera styles bring great challenges for pseudo-label prediction. To this end, a novel method is proposed, which bridges the gap of different cameras and extracts more discriminative features from an image. Specifically, an intra-to-intermechanism is introduced, in which samples from their own cameras are first grouped and then aligned at the class level across different cameras followed by our logical relation inference (LRI). Thanks to these strategies, the logical relationship between simple classes and hard classes is justified, preventing sample loss caused by discarding the hard samples. Furthermore, we also present a multiview information interaction (MvII) module that takes features of different images from the same pedestrian as patch tokens, obtaining the global consistency of a pedestrian that contributes to the discriminative feature extraction. Unlike the existing clustering-based methods, our method employs a two-stage framework that generates reliable pseudo-labels from the views of the intracamera and intercamera, respectively, to differentiate the camera styles, subsequently increasing its robustness. Extensive experiments on several benchmark datasets show that the proposed method outperforms a wide range of state-of-the-art methods. The source code has been released at https://github.com/lhf12278/LRIMV.**

***Index Terms*— Domain adaptation, logical relation inference (LRI), multiview information interaction (MvII), person re-identification (Re-ID).**

I. Introduction

PERSON re-identification (Re-ID) is the technology of targeting the related pedestrian according to the given probe image. Currently, supervised person Re-ID methods [1], [2], [3], [4], [5], [6], [7], [8], [9] are widely learned based on the same dataset, which does achieve satisfactory performances. However, when such models trained on the source dataset are deployed directly to an unknown/new target dataset, dramatic performance degradation is unavoidable due to the domain offsets between the source domain and the target domain [10]. To address this problem, domain adaptation person Re-ID methods have been proposed.

Among them, clustering-based pseudo-label prediction is an effective and popular method [11], [12], [13], [14], [15], [16], [17], [18], which predicts the pseudo-labels as the guidance of model training on the target dataset. Despite the fact that clustering-based approaches do outperform those based on other techniques, they also suffer from limitations.

Specifically, as shown in Fig. 1, a pair of images from different identities in the same camera enjoy a larger similarity than those from the same identity in different cameras [19], which is mainly caused by the different camera styles. Obviously, if the training samples in the target domain are directly clustered without considering the influence of the different camera styles, different samples belonging to different pedestrians captured by the same camera could be classified into the same category, while samples of the same pedestrian captured by different cameras would inversely be classified to different categories, subsequently generating noise labels and greatly creating an inferior influence.

Although CAIL [19] alleviates the impact of different camera styles on cross-camera pedestrian retrieval tasks through introduce camera-aware neighborhood invariance, it does not increase intraclass similarity and reduce interclass similarity at the instance level. MetaCam-Dsce [20], CCSE [21], and STS [22] alleviate the adverse impact of camera style on pedestrian retrieval through meta-learning, camera style translation, and camera irrelevant matrix, respectively. However, similar to

Manuscript received 28 August 2022; revised 16 March 2023; accepted 26 May 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 62272133, Grant 61906162, Grant 62276120, and Grant 61966021; in part by the Shenzhen Colleges and Universities Stable Support Program under Grant GXWD20220811170100001; in part by the Shenzhen Science and Technology Program under Grant RCBS20200714114910193; and in part by the Yunnan Fundamental Research Projects under Grant 202301AV070004. *(Corresponding authors: Jinxing Li; Huafeng Li.)*

Shuang Li and Xinbo Gao are with the Chongqing Key Laboratory of Image Cognition, Chongqing University of Posts and Telecommunications, Chongqing 400065, China (e-mail: shuangli936@gmail.com; gaoxb@cqupt.edu.cn).

Fan Li and Huafeng Li are with the Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 6505000, China (e-mail: lifan_kust@163.com; lhfchina99@kust.edu.cn).

Jinxing Li is with the Harbin Institute of Technology, Shenzhen 518055, China, and also with the Shenzhen Key Laboratory of Visual Object Detection and Recognition, Shenzhen 518055, China (e-mail: lijinxing158@gmail.com).

Bob Zhang is with the PAMI Research Group, Department of Computer and Information Science, University of Macau, Macau 999078, China (e-mail: bobzhang@um.edu.mo).

Dapeng Tao is with the Fist Laboratory, School of Information Science and Engineering, Yunnan University, Kunming 650091, China (e-mail: dapeng.tao@gmail.com).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TNNLS.2023.3281504.

Digital Object Identifier 10.1109/TNNLS.2023.3281504

Fig. 1. Similarities of a pair of images from the intracamera and intercamera. Similarities of different pedestrians from the same camera are even larger than that of the same pedestrian from different cameras.

Fig. 2(a), these methods directly cluster mixed samples from different cameras, and noise pseudo-label will be generated due to differences in camera. The work which is the most similar to our method is IICS [23]. To prohibit the degradation in pseudo-label estimation across cameras due to the distribution discrepancy among cameras, IICS also computes the sample similarity via two stages (intracamera and intercamera) and optimizes the model through two kinds of pseudo-labels generated in two stages. However, the similarity of samples belonging to different pedestrians in the same camera is possibly higher than that of the same pedestrians from different cameras. Directly mixing all samples for clustering would result in inaccurate estimation. In addition, the pseudo-labels of the intracamera generated by the intercamera stage may be inconsistent with that generated by the intracamera stage. As a result, the more accurate pseudo-labels generated by the intracamera stage are not effectively utilized.

To further alleviate the gap in the camera styles, an intra-to-intermechanism is introduced to conduct the pseudo-label prediction followed by two stages. As shown in Fig. 2(b), the clustering is first performed on the target samples from the same camera to roughly predict labels for each sample. By following our logical relation inference (LRI), labels belonging to hard samples are further finetuned. Specifically, as displayed in Fig. 2(c)(1), assume A, B, C, and D are four different samples, and A and B (C and D) belong to the same class. If A and D belong to the same class, A, B, C, and D then enjoy the same class. Except for this membership-based inference, the non-membership-based inference is also introduced. Let E, F, G, and H be four different groups, and F and H belong to the same class. Since E and F come from the same camera, so do G and H. If E is classified to G and H, while E is different from F, E only enjoys the same class as G. Compared with existing methods that directly apply the distance metric for the pseudo-label prediction, our strategy exploits the logical relation between different instances or groups, such that more accurate pseudo-labels for hard samples are predicted.

Besides this, due to the differences in viewpoints and pedestrian postures, an image can only present partial appearances, where the features extracted from different images of the same identity encounter diversity, which not only brings challenges to pedestrian identity matching, but also produces noise labels for the sake of incompleteness in the features. Obviously, if the model can extract multiview features from a single image, this issue can be effectively alleviated. Therefore, here we propose a multiview feature interaction mechanism, which utilizes the information interaction of patch tokens during the generation of class tokens by a Transformer. Specifically, we use different images of the same pedestrian (including images captured from the same/different views) as the patch token inputs for the Transformer layer. Information interaction and transportation between different patch tokens are then carried out through multihead attention and multilayer perceptron (MLP) in a Transformer layer. Thanks to these strategies, the tokens after information interaction are used for label prediction to guide the feature extraction of every single image.

Our main contributions can be summarized as follows.

1) An intra-to-intermechanism for pseudo-label prediction is proposed, which generates the clustering-based intra- and intercamera pseudo-labels stage by stage, efficiently reducing the risk of introducing noise labels during the straight clustering on intercamera samples due to the camera style differences.
2) Instead of neglecting the hard samples like many existing methods, a logical relational inference mechanism is proposed, such that their pseudo-labels are relatively determined by other classes with high confidence of intracamera relation, which contributes to the full information exploitation in the training set.
3) An information interaction mechanism is further proposed, which realizes the multiview information extraction across different images with the same identity. Specifically, multiview information extraction from every single image is conducted through the self-distillation mode, by which its extracted feature enjoys much more complementary information associated with different viewpoints and pedestrian postures.

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

II. Related Work

The proposed method in this article first uses source-domain data to obtain a pretraining model, and the model is then fine-tuned using the target-domain data. Therefore, the proposed method belongs to domain adaptation person re-ID. Domain adaptation person Re-ID can be roughly categorized into feature distribution alignment-based methods [22], [24], [25], [26], [27], [28], [29], [30], [31], [32], [33], image-style translation-based methods [21], [34], [35], [36], [37], [38], [39], [40], [41], [42], [43], [44], and clustering-based methods [11], [12], [13], [14], [15], [16], [17], [18].

A. Feature Distribution Alignment-Based Methods

Feature distribution alignment-based methods conduct the distribution alignment [24], [25], [26], [27] between the source domain and the target domain to transfer the knowledge from the source domain to the target domain, realizing the domain adaptation. For instance, Qi et al. [28] employed a camera-aware domain adaptation framework to realize the distribution alignment between different domains. Li et al. [29] designed the cross-adversarial consistency self-prediction

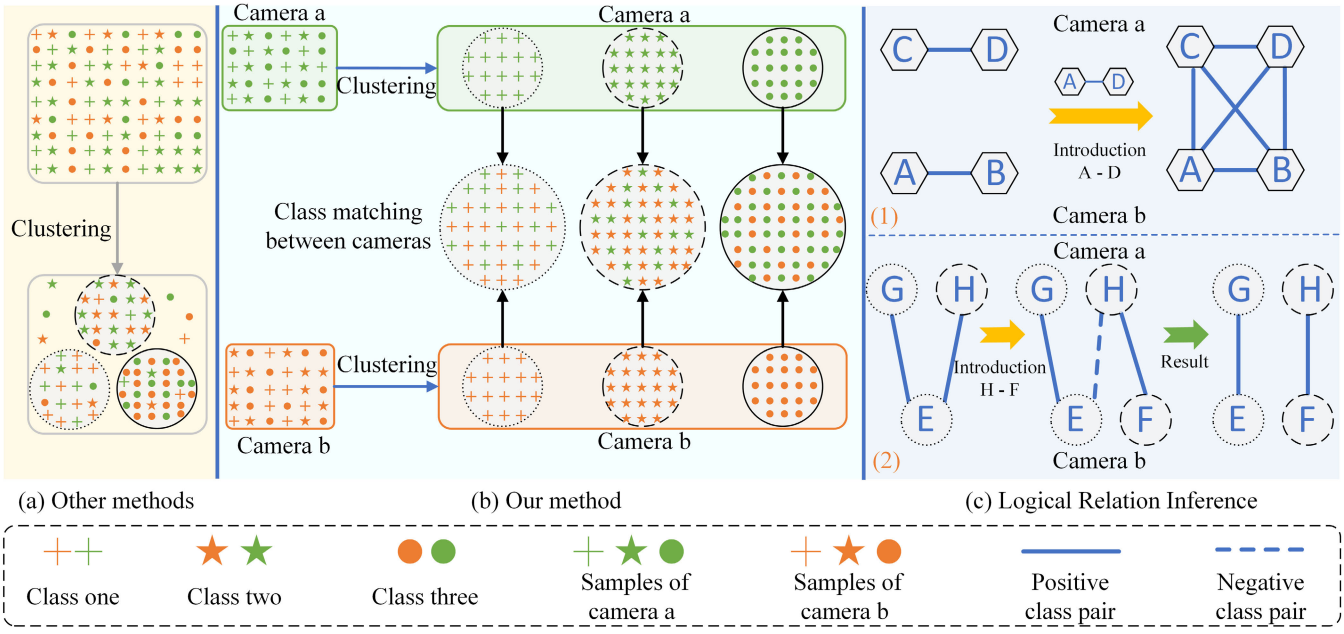

Fig. 2. Differences between other methods and our proposed method. (a) Existing methods directly cluster mixed samples from different cameras. (b) Our method achieves the clustering from the intracamera to the intercamera. (c) LRI.

learning strategy, which helps generate domain-invariant features. Li et al. [30] proposed to utilize the stability and complementarity of pedestrian attributes and corresponding low-level visual features to guide the learning of attribute-aligned domain-invariant features. Additionally, CBN [31] forces the image data of all cameras to fall onto the same subspace by camera-based batch normalization to align the distribution between different cameras. PREST [32] mitigates the adverse influence of different styles among cameras by employing a classifier with a gradient inversion layer to learn view-invariant representation. Attributes are important semantic information for pedestrians. To effectively exploit them, Wang et al. [33] introduced the transferable joint attribute identity deep learning (TJ-AIDL) to learn attribute semantics and identify feature representation space at the same time, aligning feature distribution. However, feature distribution alignment-based methods also encounter inferior influence on discriminative feature extraction, which results in relatively worse performances.

B. Image-Style Translation-Based Methods

Image-style translation [21], [34], [35], [36], [37], [38]-based methods generally try to adapt to the different image styles of different domains by translating images through GANs, which mainly translate labeled images of the source domain to the target domain, to realize the domain-invariant feature learning. To make the model adaptive to different camera styles, CamStyle [39] introduces the camera style adaptation so that labeled training images can be style-transferred to each camera. Zhong et al. [40] proposed a hetero-homogeneous learning (HHL) method to transfer the style of an image from one camera to that of another camera, which is then used as the negative sample compared with the source samples. SPGAN [41] attempts to retain the identity clues during the image-style translation by two heuristic constraints of maintaining self-similarity of person identity and introducing dissimilarity of domains. To overcome the influence of different camera styles, Zhong et al. [42] carry out camera-level image style translation to achieve the constraint of camera invariance. More precisely, the differences in different domains mainly come from lighting conditions, resolutions, and camera styles. To alleviate them, ATNET [43] achieves accurate style migration at the factor level by sensing different factors of image styles. Considering that the above methods do not take the influence of pedestrian posture on image feature representation during image-style translation into account, Li et al. [44] realized pose disentanglement to achieve the purpose of extracting domain-invariant features unaffected by different postures. Image-style translation methods consider differences in both camera distribution and inter-class relations. However, the performance of these methods is heavily constrained by the quality of generated images.

C. Clustering-Based Methods

Clustering-based methods aim to explore the relationship between samples in the target domain, generate pseudo-labels to fine-tune the model, and transfer the knowledge from the source domain to the target domain. Fan et al. [11] first proposed to generate pseudo-labels by conducting clustering in the target domain. Fu et al. [12] proposed to use the potential similarity of unlabeled samples for automatic clustering (from local to global) from different views and then assigned labels to these independent clusters for training. To reduce the impact of noise pseudo-labels, MMT [13] attempts to refine hard pseudo-labels offline and soft pseudo-labels online and then soft-refined pseudo-labels in the target domain in an alternate

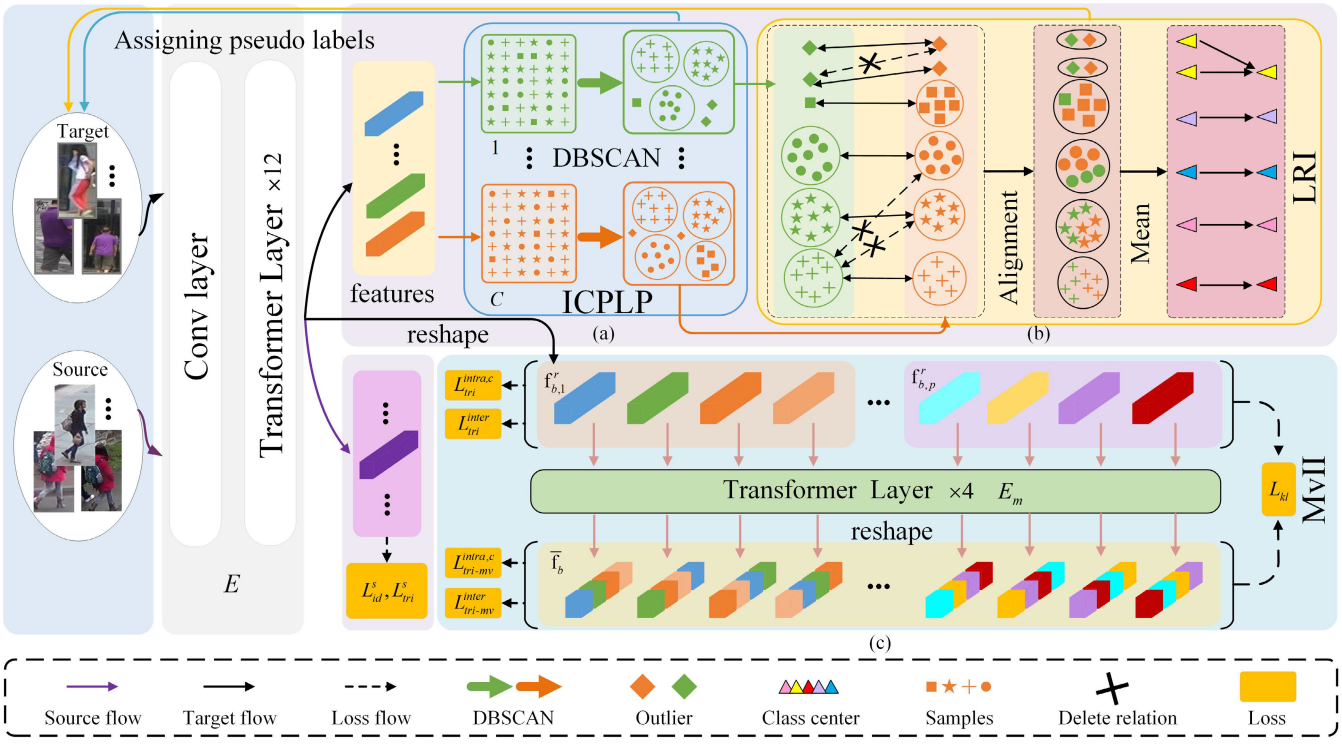

Fig. 3. Overview of our method. (a) ICPLP is used to enhance the discrimination of the intracamera features and provide reliable intracamera pseudo-labels for our proposed LRI. (b) LRI is used to generate intercamera pseudo-labels by interclass alignment, HSI, and ScF based on the intracamera clustering results. (c) Additionally, we also propose the MvII module, which allows us to gain more complementary and discriminative features.

training model. Zheng et al. [14] proposed a group-aware label transfer (GLT) algorithm, which enables the online interaction and mutual promotion of pseudo-label prediction and representation learning. Zheng et al. [15] also proposed to evaluate the reliability of pseudo-labels, suppressing the contribution of noise samples. Li and Zhang [16] considered visual similarity and temporal consistency to predict multiple categories of labels to ensure the quality of label prediction. Networks with different architectures usually have different advantages. Zhai et al. [17] introduced a regularization scheme on expert authority to adapt to the heterogeneity of experts learned in different architectures and improve the recognition ability of the domain adaptive Re-ID model. To improve the discriminability of features, Zhai et al. [18] proposed a new augmented discriminant cluster technology to expand the samples of different categories in the target domain. Pseudo-label-based methods generally generate pseudo-labels by clustering, which explores the class relations in the target domain. Despite that these methods can achieve better performances, most of them do not fully consider the difference in camera distributions, subsequently contaminating by noise pesudo-labels and misguiding the model training. In this article, the LRI and multiview information interaction (MvII) are both introduced, efficiently tackling the aforementioned problems.

III. Proposed Method

The pipeline of our proposed method is shown in Fig. 3. After pretraining the encoder E on the source domain, the model is transferred to the target domain through our proposed LRI and MvII. In the pretraining phase, being similar to many existing methods, we exploit the cross-entropy loss and soft triplet loss to update the encoder E, such that the discriminative features can be extracted. Due to the missing labels of the target set, LRI is introduced to gain the pseudo-labels more robustly. Furthermore, MvII achieves the information interaction in one batch, obtaining more complementary multiview information even from a single image.

A. Learning in the Source Domain

The method proposed in this article aims to train the model with labeled source-domain data and unlabeled target-domain data, so that the model can adapt to the target domain and thus can extract strong discriminant pedestrian features from the pedestrian images in the target domain. To make the network enjoy better initialization, we first pretrain the network through the labeled samples from the source domain. In the pretraining process, the pretrained Vision Transformer based on ImageNet is used as the backbone, followed by a fully connected layer, that is, identity classifier W_{id}. Given the labeled source-domain sample $S = \{(x_{s,i}, y_{s,i})|_{i=1}^{n_s}\}$, where n_s represents the total number of pedestrian images in the source domain, $x_{s,i}$ represents the ith pedestrian image in the source domain and $y_{s,i}$ represents its corresponding identity label, we can pretrain the network using following functions:

$$\boldsymbol{L}_{id}^{s} = -\frac{1}{n_b}\sum_{i=1}^{n_b} q_{s,i}\log\left(\boldsymbol{W}_{id}\left(\boldsymbol{E}\left(x_{s,i}\right)\right)\right) \tag{1}$$

$$L_{\text{tri}}^{s} = \frac{1}{n_b}\sum_{i=1}^{n_b}\log\left[1+\exp\left(\|\boldsymbol{E}\left(x_{s,i}\right)-\boldsymbol{E}\left(x_{s,i}^{p}\right)\|_2^2\right.\right.$$

$$-\|\boldsymbol{E}(x_{s,i}) - \boldsymbol{E}(x_{s,i}^n)\|_2^2)\big] \tag{2}$$

where n_b represents the batch size, $x_{s,i}^n$ and $x_{s,i}^p$ are the hard-negative and hard-positive samples of $x_{s,i}$, respectively, and $q_{s,i} \in \mathbb{R}^{K\times 1}$ is a one-hot vector (only the element at $y_{s,i}$ is 1).

B. Intracamera Pseudo-Label Prediction (ICPLP)

The pseudo-label prediction has a large influence on the discriminative learning of the encoder E. As analyzed above, neglecting the differences among different cameras does result in an inaccurate estimation, while labels predicted in the same camera enjoy robustness. Thus, here we first conduct pseudo-label prediction for the intracamera samples. Specifically, assume training samples in the cth camera as $T^c = \{x_{t,i}^c\}_{i=1}^{n_t^c}$. By applying E to these samples, features are obtained and then measured by the Euclidean distance and Jaccard distance [45]. The pseudo-label $\{y_{t,i}^c\}_{i=1}^{n_t^c}$ belonging to the ith sample in the cth camera is learned by DBSCAN [46], where n_t^c is the total number of samples in the cth camera. Since there are no intercamera differences, these labels are relatively accurate. Using the soft triplet loss $\{L_{\text{tri}}^{\text{intra},c}\}_{c=1}^{C}$ (where C is the number of cameras), E is finetuned to be adaptive to our target domain camera by camera

$$L_{\text{tri}}^{\text{intra},c} = \frac{1}{n_b}\sum_{i=1}^{n_b}\log\big[1+\exp\big(\|\boldsymbol{f}_{t,i}^c - \boldsymbol{f}_{t,i}^{c,p}\|_2^2 - \|\boldsymbol{f}_{t,i}^c - \boldsymbol{f}_{t,i}^{c,n}\|_2^2\big)\big] \tag{3}$$

where $f_{t,i}^c = E(x_{t,i}^c)$, n_b represents the batch size, and $f_{t,i}^{c,n}$ and $f_{t,i}^{c,p}$ are the hard-negative and hard-positive samples of $f_{t,i}^c$, respectively. And the positive (negative) relationship of the sample pair of $f_{t,i}^{c,p}(f_{t,i}^{c,n})$ and $f_{t,i}^c$ is obtained according to the predicted pseudo-label. Note that some outliers (hard samples) are not taken into account in this phase.

C. Logical Relation Inference

Despite the fact that pseudo-labels in the intracamera do achieve robustness, labels in different cameras are also inconsistent. To tackle this problem, LRI is proposed, which is composed of interclass alignment, hard sample inference (HSI), and similar-class fusion (ScF).

1) Interclass Alignment: Through clustering, samples in T^c are grouped as $T^c = \{\mathbf{x}_{t,i}^c, y_{t,i}^c\}_{i=1}^{n_t^c}$. Note that to avoid performance loss caused by discarding training samples, we regard some outliers (hard samples) as independent classes, so we will take them into account at this stage. Without the loss of generality, let $\mathbf{F}_{c1} = [\mathbf{f}_{t,1}^{c1}, \mathbf{f}_{t,2}^{c1}, \ldots, \mathbf{f}_{t,n_t^{c1}}^{c1}]$ and $\mathbf{F}_{c2} = [\mathbf{f}_{t,1}^{c2}, \mathbf{f}_{t,1}^{c2}, \ldots, \mathbf{f}_{t,n_t^{c2}}^{c2}]$ be the extracted features in the $c1$th and $c2$th cameras, respectively. Euclidean distances between each pair can be represented as

$$\mathbf{A}_{c1,c2}(i,j) = \mathbf{D}(\mathbf{F}_{c1},\mathbf{F}_{c2})(i,j) = \|\mathbf{f}_{t,i}^{c1} - \mathbf{f}_{t,j}^{c2}\|_2^2 \times \big(i = 1,2,\ldots,\mathbf{n}_t^{c1}, j = 1,2,\ldots,\mathbf{n}_t^{c2}\big) \tag{4}$$

where $\mathbf{D}$ represents the operator of calculating the distance between different column vectors of the matrix. The element $\mathbf{A}_{c1,c2}(i,j)$ on the ith row and jth column in $\mathbf{A}_{c1,c2}$ means the distance between $\mathbf{f}_{t,i}^{c1}$ and $\mathbf{f}_{t,j}^{c2}$. In this article, we empirically set a threshold thres. If $\mathbf{A}_{c1,c2}(i,j) <$ thres, then $\mathbf{x}_{t,i}^{c1}$ and $\mathbf{x}_{t,j}^{c2}$ enjoy the same identity. Subsequently, we can get that the groups $G_{y_{t,i}^{c1}}^{c1}$ and $G_{y_{t,j}^{c2}}^{c2}$ that $\mathbf{x}_{t,i}^{c1}$ and $\mathbf{x}_{t,j}^{c2}$ belonging to enjoy the same category.

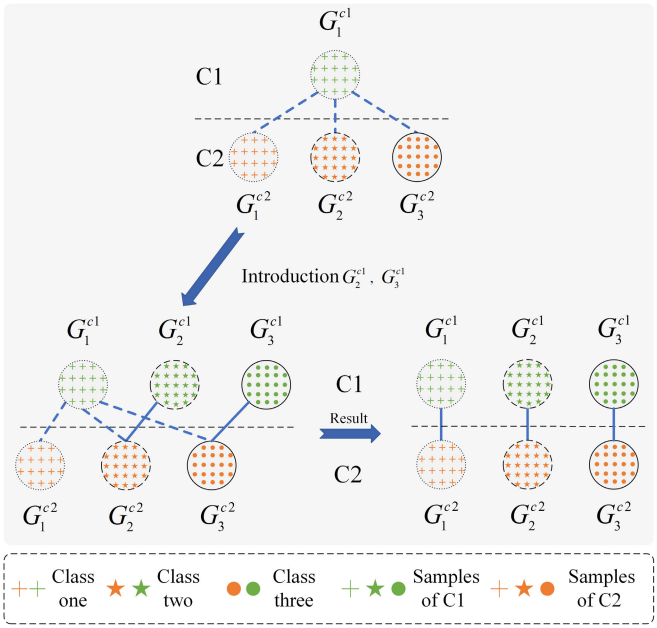

Fig. 4. $c1$,$c2$ represent different cameras. G_1^{c1} and G_1^{c2} (G_2^{c1} and G_2^{c2}, $G_3^{c1}$and G_3^{c2}) represent the image sets with the same pedestrian captured by camera $c1$ and $c2$. By introducing G_2^{c1} and G_3^{c1} according to the strategy of HSI, it is inferred that G_1^{c1} is only aligned with G_1^{c2}.

2) Hard Sample Inference: In fact, it is inevitable to encounter the case that a group in one camera is similar to multiple groups in the other cameras by following the strategy in the interclass alignment. For instance, G_1^{c1} may be simultaneously aligned with G_1^{c2}, G_2^{c2}, and G_3^{c2}. Of course, it is difficult to decide which identity G_1^{c1} belongs to. Here, we name samples like that in G_1^{c1} as the hard samples.

In this article, we propose a simple but efficient strategy to address these hard samples, the detailed process is shown in Fig. 4. Specifically, assume that G_2^{c1} and G_2^{c2}, as well as G_3^{c1} and G_3^{c2} belong to their own identities. Since G_1^{c1} is completely different from G_2^{c1} and G_3^{c1}, G_1^{c1} is also dissimilar to G_2^{c2}, and G_3^{c2}. Therefore, G_1^{c1} and G_1^{c2} belong to the same class. By following this processing, we finally cluster training samples into K groups $P = \{P_1, P_2, \ldots, P_K\}$. Note that if a cluster does not meet this assumption, we do not take corresponding samples in it into account at the current stage.

3) Similar-Class Fusion: Although K groups are clustered, it is also possible that P_i and P_j may belong to the same pedestrian. Because these independent classes (outliers of ICPLP) may be able to align the independent classes of other cameras to form a new class in this process of interclass alignment. This may lead to the situation of assigning different identities to samples belonging to the same pedestrians in the pseudo-label prediction results. Here, we additionally propose an ScF strategy to further combine similar classes into a single one. Mathematically, the class-center vectors are used

to achieve this purpose

$$\mathbf{m}_k = \frac{1}{n_k}\sum_{i=1}^{n_k} E(\mathbf{x}_i \in P_k) \tag{5}$$

where n_k is the number of samples in P_k and $\mathbf{x}_i$ is the input image. Denote $\mathbf{M} = [\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_K]$. The distance of the class-center vectors via $\mathbf{M}$ is then obtained

$$\begin{aligned}\mathbf{J}(i, j) &= \mathbf{D}(\mathbf{M}, \mathbf{M})(i, j) \\ &= ||\mathbf{m}_i - \mathbf{m}_j||_2^2 (i = 1, 2, \ldots, \mathbf{K}, j = 1, 2, \ldots, \mathbf{K}).\end{aligned} \tag{6}$$

If $\mathbf{J}(i, j) < \text{thres}$, then P_i and P_j can be fused as a single class. Finally, all training samples are classified into $P = \{P_1, P_2, \ldots, P_{\bar{K}}\}$, where $\bar{K} \leqslant K$. According to these pseudo-labels, the encoder E is further finetuned by the soft triplet loss $L_{\text{tri}}^{\text{inter}}$

$$L_{\text{tri}}^{\text{inter}} = \frac{1}{n_b}\sum_{i=1}^{n_b} \log\left[1 + \exp\left(\|\boldsymbol{f}_{t,i} - \boldsymbol{f}_{t,i}^{p}\|_2^2 - \|\boldsymbol{f}_{t,i} - \boldsymbol{f}_{t,i}^{n}\|_2^2\right)\right] \tag{7}$$

where $f_{t,i} = E(x_{t,i})$, n_b represents the batch size, and $f_{t,i}^{n}$ and $f_{t,i}^{p}$ are the hard-negative and hard-positive samples of $f_{t,i}$, respectively. And the positive (negative) relationship of the sample pair of $f_{t,i}^{p}$ ($f_{t,i}^{n}$) and $f_{t,i}$ is obtained according to the predicted pseudo-label.

D. Multiview Information Interaction

Apart from the inferior influence caused by the difference in camera styles, a person's image captured from different viewpoints also suffers from incomplete feature extraction. For instance, a pedestrian captured in the front may be more discriminative than that captured in the background viewpoint. Of course, obtaining multiview information from a single image through E contributes to more discriminative feature extraction. Here, a module named MvII is presented. In detail, we regard different features belonging to the same identity to be the patch token. By introducing the Transformer layers, these tokens are interacted with each other, so that complementary information is transported to each feature.

Assume that there are p pedestrians in a batch, which can be represented as $\mathbf{f}_b = [\mathbf{f}_{b,1}, \mathbf{f}_{b,2}, \ldots, \mathbf{f}_{b,n_b}] \in \mathbb{R}^{n_b \times d}$, where n_b is the size of a batch and d is the dimension of features. According to the identity belongings, $\mathbf{f}_b$ is reshaped to $\mathbf{f}_b^r = [\mathbf{f}_{b,1}^r, \mathbf{f}_{b,2}^r, \ldots, \mathbf{f}_{b,p}^r] \in \mathbb{R}^{p \times (n_b/p) \times d}$, where p is the number of identities in this batch. By forwarding it into a subnetwork E_m constructed by four transformer layers, the MvII is achieved for each image. By reshaping the outputs from E_m to the size of $n_b \times d$, features $\bar{\mathbf{f}}_b = [\bar{\mathbf{f}}_{b,1}, \bar{\mathbf{f}}_{b,2}, \ldots, \bar{\mathbf{f}}_{b,n_b}] \in \mathbb{R}^{n_b \times d}$ are obtained. Being similar to $L_{\text{tri}}^{\text{inter}}$, another triplet loss $L_{\text{tri-mv}}$ is exploited whose inputs are $\mathbf{f}_b$

$$\begin{aligned}L_{\text{tri-mv}} = \frac{1}{n_b}\sum_{i=1}^{n_b} \\ \times \log\left[1 + \exp\left(\|\bar{\mathbf{f}}_{b,i} - \bar{\mathbf{f}}_{b,i}^{p}\|_2^2 - \|\bar{\mathbf{f}}_{b,i} - \bar{\mathbf{f}}_{b,i}^{n}\|_2^2\right)\right]\end{aligned} \tag{8}$$

where n_b represents the batch size, and $\bar{\mathbf{f}}_{b,i}^{n}$ and $\bar{\mathbf{f}}_{b,i}^{p}$ are the hard-negative and hard-positive samples of $\bar{\mathbf{f}}_{b,i}$, respectively.

Although the aforementioned analysis encourages the networks to learn multiview information, it is impractical to input multiple images of the same person in the testing phase. To tackle this problem, we further introduce KL diversity to achieve self-supervised distillation

$$L_{kl} = \frac{1}{n_b}\sum_{i=1}^{n_b} S\left(\bar{\mathbf{f}}_{b,i}\right) \log \frac{S\left(\bar{\mathbf{f}}_{b,i}\right)}{S\left(\mathbf{f}_{b,i}\right)} \tag{9}$$

where $S(\cdot)$ denotes the softmax. By considering the intercamera and intracamera, as well as the soft triplet loss and KL diversity loss, the loss functions in MvII are, respectively,

$$\begin{aligned}L_{\text{mv}}^{\text{intra},c} &= L_{\text{tri-mv}}^{\text{intra},c} + L_{kl}^{\text{intra},c}, \quad c = 1, \ldots, C \\ L_{\text{mv}}^{\text{inter}} &= L_{\text{tri-mv}}^{\text{inter}} + L_{kl}^{\text{inter}}\end{aligned} \tag{10}$$

where inputs in (10) are all from MvII and labels are obtained by intracamera or intercamera pesudo-label predictions. It is worth noting that the MvII is only used for auxiliary training in the training process and does not participate in the testing phase. Therefore, we use feature $\mathbf{f}_b$ for pedestrian retrieval during the testing phase.

E. Overall Training

By taking the aforementioned analysis into account, our objective function is

$$L = \sum_{c=1}^{C} L_{\text{tri}}^{\text{intra},c} + \mu_1 L_{\text{tri}}^{\text{inter}} + \mu_2\left(\sum_{c=1}^{C} L_{\text{mv}}^{\text{intra},c} + L_{\text{mv}}^{\text{inter}}\right) \tag{11}$$

where μ_1 and μ_2 are nonnegative parameters to tradeoff the importance of the different modules.

IV. Experiments

A. Datasets and Evaluation Metrics

1) Datasets: Being similar to existing methods, three benchmarks including Market1501 [61], DuMTMC-reID (Duke) [62], and MSMT17 [41] are used for quantitative evaluation. To demonstrate the robustness of our proposed method, Market1501 and MSMT17 are further modified to new versions by following [63], in which most pedestrians are captured by only a camera. Here, we denote these modifications as Market1501-new and MSMT17-new, respectively. The details of these datasets are shown in Table I.

Market1501 contains 32 668 pedestrian pictures of 1501 pedestrians. These data are collected from six nonoverlapping cameras on the campus of Tsinghua University. There are 12 936 images in the training set for training, which is composed of 751 pedestrian images, and each pedestrian is captured by multiple cameras. In the test phase, the test set contains 19 732 images composed of 750 pedestrians, and the query set contains 3368 images.

Dukemtmc-Reid contains 36 411 images captured by eight nonoverlapping cameras. The dataset is a subset of Dukemtmc [64] and is used for the person-rerecognition task. The training set contains 16 522 images composed of 702 pedestrians, and each pedestrian is captured by multiple cameras. The query set and the gallery set used for the test, respectively,

TABLE I

SETTINGS OF DIFFERENT PERSON RE-ID DATASETS IN PERFORMANCE COMPARISON. PED: NUMBER OF PEDESTRIANS; IMG: NUMBER OF IMAGES; CAM: NUMBER OF CAMERAS

Datasets	Ped	Training		Gallery (Testing)		Probe (Testing)		Cam
		Ped	Img	Ped	Img	Ped	Img	
Market1501	1501	751	12936	750	19732	750	3368	6
Duke	1812	702	16522	1110	17661	702	2228	8
MSMT17	4101	1041	32621	3060	82161	3060	11659	15
Market1501-new	1367	617	3197	750	19732	750	3368	6
MSMT17-new	2831	1790	15356	1041	29721	1041	2900	15

TABLE II

RESULTS OF MAP AND CMC (%) OBTAINED BY OUR PROPOSED METHOD AND THE STATE-OF-THE-ART RE-ID METHODS ON DUKE→MARKET1501 AND MARKET1501→DUKE TASK."R1," "R5," AND "R10" DENOTE RANK-1, RANK-5, AND RANK-10, RESPECTIVELY. THESE RESULTS ARE COPIED FROM THEIR PAPERS

Methods	Reference	Duke→Market1501				Market1501→Duke			
		R1	R5	R10	mAP	R1	R5	R10	mAP
ECN [42]	CVPR'19	75.1	87.6	91.6	43.0	63.3	75.8	80.4	40.4
PDA-Net [44]	ICCV'19	75.2	86.3	90.2	47.6	63.2	77.0	82.5	45.1
PCB-PAST [47]	ICCV'19	78.4	-	-	54.6	72.4	-	-	54.3
SSG [12]	ICCV'19	80.0	90.0	92.4	58.3	73.0	80.6	83.2	53.4
AADFL [30]	TIFS'20	71.8	85.9	90.1	39.6	64.1	77.2	81.4	43.1
UDATP [48]	PR'20	75.8	89.5	93.2	53.7	68.4	80.1	83.5	49.0
ECN-GPP [49]	TPAMI'20	84.1	92.8	95.4	63.8	74.0	83.7	87.4	54.4
ACT [50]	AAAI'20	80.5	-	-	60.6	72.4	-	-	54.5
AD-Cluster [18]	CVPR'20	86.7	94.4	96.5	68.3	72.6	82.5	85.5	54.1
MMT [13]	ICLR'20	87.7	94.9	96.9	71.2	78.0	88.8	92.5	65.1
CBN [31]	ECCV'20	72.7	85.8	90.7	43.0	58.7	74.1	78.1	38.2
JVTC+ [16]	ECCV'20	86.8	95.2	97.1	67.2	80.4	89.9	92.2	66.5
CAIL [19]	ECCV'20	88.1	94.4	96.2	71.5	79.5	88.3	91.4	65.2
NRMT [51]	ECCV'20	87.8	94.6	96.5	71.7	77.8	86.9	89.5	62.2
MEB-Net [17]	ECCV'20	89.9	96.0	97.5	76.0	79.6	88.3	92.2	66.1
SpCL [52]	NeurIPS'20	90.3	96.2	97.7	76.7	82.9	90.1	92.5	68.8
HCN [53]	TCSVT'21	90.2	-	-	70.2	78.9	-	-	57.3
GCMT [54]	IJCAI'21	90.6	96.3	97.7	77.1	81.1	91.1	93.9	67.8
Dual-Refifinement [55]	TIP'21	90.9	96.4	97.7	78.0	82.1	90.1	92.5	67.7
UNRN [15]	AAAI'21	91.9	96.1	97.8	78.1	82.0	90.7	93.5	69.1
OPLGHCD [56]	ICCV'21	91.5	96.3	-	80.0	82.2	89.7	-	70.1
GLT [14]	CVPR'21	92.2	96.5	97.8	79.5	82.0	90.2	92.8	69.2
TALMIR [57]	TCSVT'22	73.1	86.3	-	40.0	63.5	76.6	-	41.3
MSC-GDC [58]	TCSVT'22	90.1	95.7	97.5	76.1	80.1	89.9	92.3	66.4
ICE [59]	TMM'22	90.8	95.8	97.2	73.8	80.2	88.5	91.6	66.4
LRCC [60]	TMM'22	91.2	96.5	97.9	78.1	83.1	90.8	92.9	69.2
Ours		**93.1**	**97.6**	**98.5**	**83.8**	**85.5**	**92.5**	**94.5**	**75.6**

contain 2228 images of 702 pedestrians and 17 661 images of 1110 identities, of which the gallery set contains interference images composed of 408 pedestrians. For the convenience of expression, Duke is used to express Dukemtmc-Reid.

MSMT17 is a large dataset that is closer to the real scene. It contains 126 441 images composed of 4101 pedestrians, which are collected by 15 nonoverlapping cameras (including 12 outdoor cameras and three indoor cameras). Among them, the training set contains 32 621 images composed of 1041 pedestrians, and each pedestrian is captured by multiple cameras. The test set contains 93 820 images composed of 3060 pedestrians, of which 11 659 pedestrian images construct the query set, and the remaining 82 161 pedestrian images construct the gallery set.

Market1501-New uses the pedestrian images of market1501 to simulate the street in the real scene. According to the idea that the pedestrian of adjacent cameras are not all the same, the protocol installs cameras at each crossroad and assumes that only 25% of the pedestrians under each camera enter the adjacent cameras. According to this idea, the training set contains 3197 pedestrian images composed of 617 identities. The test set still follows the original protocol of Market1501, so the gallery set and query set are composed of 19 732 images and 3368 images of 750 pedestrians, respectively.

MSMT17-New uses the pedestrian images of MSMT17 to simulate the cross street in the real scene. The protocol takes the test set in the original MSMT17 as the training set and the original training set as the test set. Being similar to Market1501-new, the reorganized training set consists of 15 356 images composed of 1790 pedestrians. The new test set consists of 32 621 images of 1041 pedestrians, of which

TABLE III

RESULTS OF MAP AND CMC (%) OBTAINED BY OUR PROPOSED METHOD AND THE STATE-OF-THE-ART RE-ID METHODS ON DUKE→MSMT17 AND MARKET1501→MSMT17 TASK. "R1," "R5," AND "R10" DENOTE RANK-1, RANK-5, AND RANK-10, RESPECTIVELY. THESE RESULTS ARE COPIED FROM THEIR PAPERS

Methods	Reference	Duke→MSMT17				Market1501→MSMT17			
		R1	R5	R10	mAP	R1	R5	R10	mAP
ECN [42]	CVPR'19	30.2	41.5	46.8	10.2	25.3	36.3	42.1	8.5
SSG [12]	ICCV'19	32.2	-	51.2	13.3	31.6	-	49.6	13.2
AADFL [30]	TIFS'20	38.6	50.8	56.1	14.0	30.5	42.6	48.8	11.4
ECN-GPP [49]	TPAMI'20	42.5	55.9	61.5	16.0	40.4	53.1	58.7	15.2
MMT [13]	ICLR'20	50.1	63.9	69.8	23.3	49.2	63.1	68.8	22.9
CBN [31]	ECCV'20	35.4	-	-	13.0	26.8	-	-	9.6
JVTC+ [16]	ECCV'20	52.9	70.5	75.9	27.5	48.6	65.3	68.2	25.1
CAIL [19]	ECCV'20	51.7	64.0	68.9	24.3	43.7	56.1	61.9	20.4
NRMT [51]	ECCV'20	45.2	57.8	63.3	20.6	43.7	56.5	62.2	19.8
SpCL [52]	NeurIPS'20	53.1	65.8	70.5	26.5	53.7	65.0	69.8	26.8
Dual-Refifinement [55]	TIP'21	55.0	68.4	73.2	26.9	53.3	66.1	71.5	25.1
UNRN [15]	AAAI'21	54.9	67.3	70.6	26.2	52.4	64.7	69.7	25.3
OPLGHCD [56]	ICCV'21	56.1	-	-	29.3	54.9	-	-	28.4
HCN [53]	TCSVT'21	58.7	-	-	29.9	55.1	-	-	27.0
GLT [14]	CVPR'21	59.5	70.1	74.2	27.7	56.6	67.5	72.0	26.5
TALMIR [57]	TCSVT'22	39.0	51.5	-	14.2	30.9	43.5	-	11.2
ICE [59]	TMM'22	51.6	-	-	24.3	50.9	-	-	24.0
LRCC [60]	TMM'22	55.6	67.5	72.1	27.4	53.6	65.7	70.6	25.7
MSC-GDC [58]	TCSVT'22	59.8	71.6	77.3	30.4	56.6	67.9	72.2	28.1
Ours		**60.9**	**73.5**	**78.2**	**35.6**	**58.0**	**70.6**	**75.2**	**32.3**

TABLE IV

RESULTS OF MAP AND CMC (%) OBAINED BY DIFFERENT METHODS ON COMPLEX SCENARIOS. "R1," "R5," AND "R10" DENOTE RANK-1, RANK-5, AND RANK-10, RESPECTIVELY. THESE RESULTS ARE COPIED FROM PAPER [63]

Methods	Duke→Market1501-new				Duke→MSMT17-new				Market1501→MSMT17-new			
	R1	R5	R10	mAP	R1	R5	R10	mAP	R1	R5	R10	mAP
SPCL [52]	14.1	26.1	33.0	5.6	19.8	31.7	37.7	8.8	18.8	30.4	36.6	9.0
ACT [50]	51.5	67.0	72.5	26.0	15.2	24.2	29.4	6.4	9.5	18.3	24.1	4.0
UDATP [48]	56.6	72.2	77.5	30.6	23.3	35.1	41.0	9.8	14.5	25.0	31.2	6.0
MEB-NET [17]	57.3	73.0	79.1	33.4	33.9	46.5	52.2	15.9	26.1	37.3	43.5	12.0
MMT [13]	59.7	75.1	80.9	33.7	39.0	51.0	57.8	17.3	33.9	47.5	54.9	15.4
Dual-Refifinement [55]	56.1	70.2	76.2	34.3	28.0	40.0	45.4	12.1	27.2	39.2	46.1	12.3
Ours	**76.6**	**87.7**	**92.0**	**54.5**	**57.2**	**69.7**	**75.3**	**31.6**	**57.2**	**70.1**	**75.0**	**30.9**

the query set contains 2900 pedestrian images, and the gallery set contains 29 721 images.

2) Evaluation Metrics: Cumulative matching characteristics (CMC) [65], [66] and mean average precision (mAP) [61] as two objective evaluation indicators are used to evaluate the performance of the proposed method and the comparative methods.

B. Implementation Details

ViT-BoT [67] pretrained on the ImageNet [68] is exploited as the backbone. It is implemented by PyTorch [69] on one GPU of NVIDIA RTX3090 with CUDA version 11. For the input, each image is resized to 256×128, which is divided into 16×16 patches using the sliding window strategy. We also introduce random horizontal flipping, padding with 10 pixels, color jitter, random cropping, and random erasing for data argumentation. For each batch, the size is set to 128, containing 32 identities and four images associated with each identity. In the training phase for the source domain and the target domain, the maximums of epochs, learning rates, and strides of the sliding window are, respectively, set to 120/200, 0.008/0.004, and 12/14. Besides this, the values of μ_1 and μ_2 are empirically set to 1.8 and 1.4, respectively, and the value of thres is set to 0.118 (0.115) when Duke (Market1501) serves as the source domain. Referring to training phases for both the source domain and target domain, we adopt SGD as the optimizer. When the number of epochs is less than 5, the warm-up strategy [70] is utilized to tune the learning rate. Otherwise, the cosine annealing is used.

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

To verify the effectiveness of our proposed method, we compared it with state-of-the-art methods, including.

1) *Feature Distribution Alignment-Based Methods:* AADFL [30], CBN [31], and TALMIR [57].
2) *Image-Style Translation-Based Methods:* ECN [42], PDA-Net [44], and ECN-GPP [49].
3) *Pseudo-Label-Based Methods:* PCB-PAST [47], SSG [12], UDATP [48], ACT [50], AD-Cluster [18], MMT [13], JVTC+ [16], CAIL [19], NRMT [51], MEB-Net [17], SpCL [52], HCN [53], MSC-GDC [58], GCMT

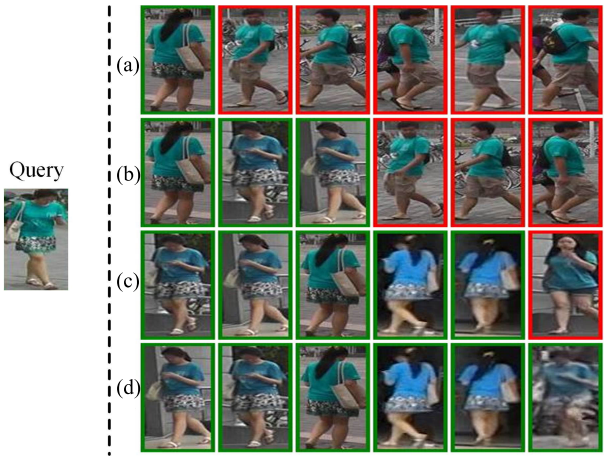

Fig. 5. Retrieval results under different ablation settings on Market1501. (a)–(d) "B," "B+ICPLP," "B+ICPLP+LRI," and "B+ICPLP+LRI+MvII," respectively.

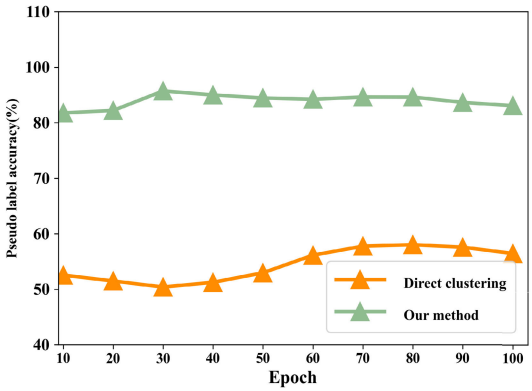

Fig. 6. Compare the pseudo-label accuracy of the proposed method and direct clustering (DBSCAN).

[54], Dual-Refinement [55], UNRN [15], OPLGHCD [56], ICE [59], LRCC [60], and GLT [14].

From Table II, we can see that our proposed method achieves competitive results compared with other approaches. Specifically, in contrast to GLT, which attains the best performance in the state-of-the-art, our presented strategy obtains 0.9%/4.3% enhancements on the Rank-1/mAP in the Duke→Market1501 task. Referring to Market1501→Duke, there are also 3.5%/6.4% improvements, demonstrating the superiority of our proposed method.

To further verify the effectiveness and generality of our method, we conduct experiments on the tasks of Market1501→MSMT17 and Duke→MSMT17. Since the number of samples in Market1501 and Duke is much smaller than that in MSMT17, these extended experiments are more challenging and practical. As shown in Table III, our methods can improve the Rank-1 of the best compared method MSC-GDC [58] from 59.8% to 60.9% and mAP from 30.4% to 35.6% on Duke→MSMT17. On Market1501→MSMT17, Rank-1 and mAP obtained by our method is 1.4% and 4.2% higher than suboptimal method MSC-GDC [58]. The above results confirm the effectiveness of the proposed method.

Since Market1501, Duke, and MSMT17 are constructed manually, each pedestrian has a high possibility of appearing under multiple cameras. However, in real-world applications, a large number of samples only appear under one camera. Thus, here we also conduct experiments on Market1501-new and MSMT17-new, which meet the aforementioned requirement. As shown in Table IV, our presented method achieves 76.4%/53.7%, 57.2%/31.6%, and 57.1%/30.9% on Rank-1/mAP in the tasks of Duke→Market1501-new, Duke→MSMT17-new, and Market1501→MSMT17-new, respectively, which are much larger than other clustering-based methods. Specifically, the Rank-1 accuracy of our proposed method exceeds the suboptimal method MMT [13] by 16.9%, 18.2%, and 23.3%, respectively, on tasks Duke→Market1501-new, Duke→MSMT17-new, and Market1501→MSMT17-new, respectively. The mAP accuracy also exceeds the suboptimal method MMT [13] by 20.8%, 14.3%, and 15.5%, respectively. Obviously, our proposed method does enjoy robustness in real-world applications.

D. Ablation Study

In this section, we carefully analyze the effects of each component including ICPLP, LRI, and MvII.

1) Effectiveness of ICPLP: In this article, the ViT-BoT trained only with source domain data is used as the baseline. On this basis, ICPLP is added to the baseline to obtain "Baseline + ICPLP," which is compared with the baseline to verify the effectiveness of ICPLP. As shown in Table V, ICPLP improves the recognition accuracy of rank-1 from 69.0% (66.2%) to 78.9% (76.6%), and mAP from 43.3% (48.6%) to 55.2% (61.2%) on Duke→Market1501 (Market1501→Duke). Furthermore, when we remove the ICPLP module from "B+ICPLP+LRI+MvII" to get "B+LRI+MVII," each sample is regarded as an independent class in the training phase. As a result, the Rank-1/mAP decrease by 6.6%/23.1% (9.0%/16%) on the Duke→Market1501 (Market1501→Duke) task, respectively. These show that pseudo-labels generated through ICPLP contribute to the discriminative feature extraction in the target domain.

2) Effectiveness of LRI: To alleviate the camera style bias, LRI is used to assign pseudo-labels to samples of intercameras, to guide the model to extract domain-invariant features. As shown in Table V, compared with the "Baseline+ICPLP," the Rank-1/mAP recognition accuracy obtained by "Baseline+ICPLP+LRI" is improved by 13.9%/27.8% (7.4%/12.8%), respectively. This result strongly demonstrates the effectiveness of LRI.

To more finely demonstrate the effectiveness of LRI, on the basis of "Baseline+ICPLP+LRI," we remove the HSI from LRI to obtain "Baseline+ICPLP+LRI (w/o HSI)." As shown in Table V, the values of Rank-1 decrease by 0.4%/0.3% in Duke→Market1501/Market1501→Duke. This is because, in the absence of HSI, our proposed method lacks a difficult sample label for training, meeting valuable information loss. Similarly, when we remove ScF from LRI, the corresponding performance of "Baseline+ICPLP+LRI (w/o ScF)" decreases by 2.4%/4.2% (0.7%/1.8%). The main reason is that, in this

TABLE V

RESULTS OF CMC AND MAP (%) OBTAINED BY OUR METHOD UNDER DIFFERENT CASES. "B" DENOTES BASELINE. "R1," "R5," AND "R10" DENOTE RANK-1, RANK-5, AND RANK-10, RESPECTIVELY

Methods	Duke→Market1501				Market1501→Duke			
	R1	R5	R10	mAP	R1	R5	R10	mAP
B	69.0	83.5	87.8	43.3	66.2	78.4	83.0	48.6
B +ICPLP	78.9	88.9	92.4	55.2	76.6	86.4	89.3	61.2
B +ICPLP+MvII	79.8	91.0	94.3	60.0	79.0	88.2	90.8	64.1
B +LRI+ MvII	86.5	93.4	95.8	60.7	76.5	86.0	89.3	59.6
B + ICPLP +LRI(w/o HSI)	92.4	97.2	98.2	81.8	83.7	91.6	94.0	74.0
B + ICPLP +LRI(w/o ScF)	90.4	96.3	98.2	78.8	83.3	91.4	93.4	72.2
B + ICPLP +LRI	92.8	97.4	98.4	83.1	84.0	92.3	94.0	74.0
B + ICPLP +LRI+MVII	**93.1**	**97.6**	**98.5**	**83.8**	**85.5**	**92.5**	**94.5**	**75.6**

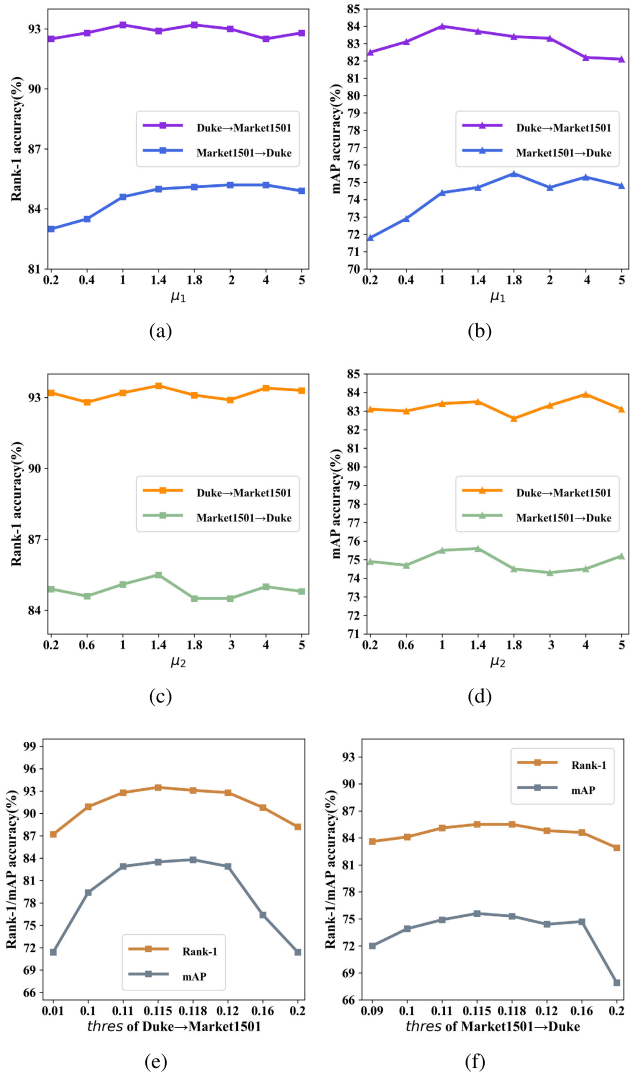

Fig. 7. Results of Rank-1 and mAP with different values of μ_1, μ_2, and thres on Duke→Market1501 and Market1501→Duke tasks. (a) Effect of μ_1 on Rank-1. (b) Effect of μ_1 on mAP. (c) Effect of μ_2 on Rank-1. (d) Effect of μ_2 on mAP. (e) Effect of thres on Rank-1/mAP on Duke→Market1501. (f) Effect of thres on Rank-1/mAP on Market1501→Duke.

case, our proposed method fails to assign correct pseudo-labels to samples of the hard class and the similar class.

3) Effectiveness of MvII: To further improve the discrimination of pedestrian features, MvII is added to "Baseline+ICPLP+LRI" to obtain "Baseline+ICPLP+LRI+MvII." In Table V, the Rank-1/mAP metrics are improved from 92.8%/83.0% to 93.5%/83.5% on the Duke→Market1501 and from 84.0%/74.0% to 85.5%/75.6% on the Market1501→Duke. Thus, MvII does promote the model to extract multiview features from a single pedestrian image, which avoids the inaccurate prediction of pseudo-labels caused by incomplete features. In addition, compare to "B+ICPLP," the Rank-1/mAP of "B+ICPLP+MvII" also has been improved, reaching 79.8%/60.0% and 79.0%/64.1% on the Duke→Market1501 and Market1501→Duke tasks, respectively. This also proves that different samples with the same identity from the same view still have complementary and discriminative information.

Fig. 5 further displays some examples obtained from our proposed method under different ablation settings. It is easy to observe that each technique does provide a contribution. To prove the effectiveness of the intra-to-intermechanism proposed in this article, we additionally produced a comparison experiment on the accuracy of pseudo-label with direct clustering (DBSCAN [46]) on the Market1501→Duke task. As shown in Fig. 6, the accuracy of pseudo-labels is recorded every 10 epochs. Experiments show that the accuracy of our method is higher than direct clustering.

TABLE VI

COMPARISONS OF COMPUTATIONAL COSTS OF DIFFERENT METHODS ON MARKET1501→DUKE

Methods	Params/M	FLOPs/G	Train time/h
ECN [42]	826.58	2821.74	≈ 84.9
JVTC+ [16]	816.53	2823.12	≈ 87.3
MEB-NET [17]	52.26	15.26	≈ 10.0
Our	114.00	17.90	≈ 10.5

E. Training Cost Analysis

Compared with the existing domain-adaptive methods, the training cost of our method meets the requirement. Quantitatively, we compared the Params, FLOPs, and Train time of ECN [42], JVTC+ [16], MEB-NET [17], and our method. The details are shown in Table VI. It is worth noting that we conduct ECN, JVTC+, MEB-NET, and our method on the Market1501→Duke task in the same software and hardware environment. As shown In Table VI, the Params, FLOPs, and Train time of ECN and JVTC+ are much higher than our methods. The reason is that ECN and JVTC+ introduce an additional style transfer model, CamStyle [71], to expand

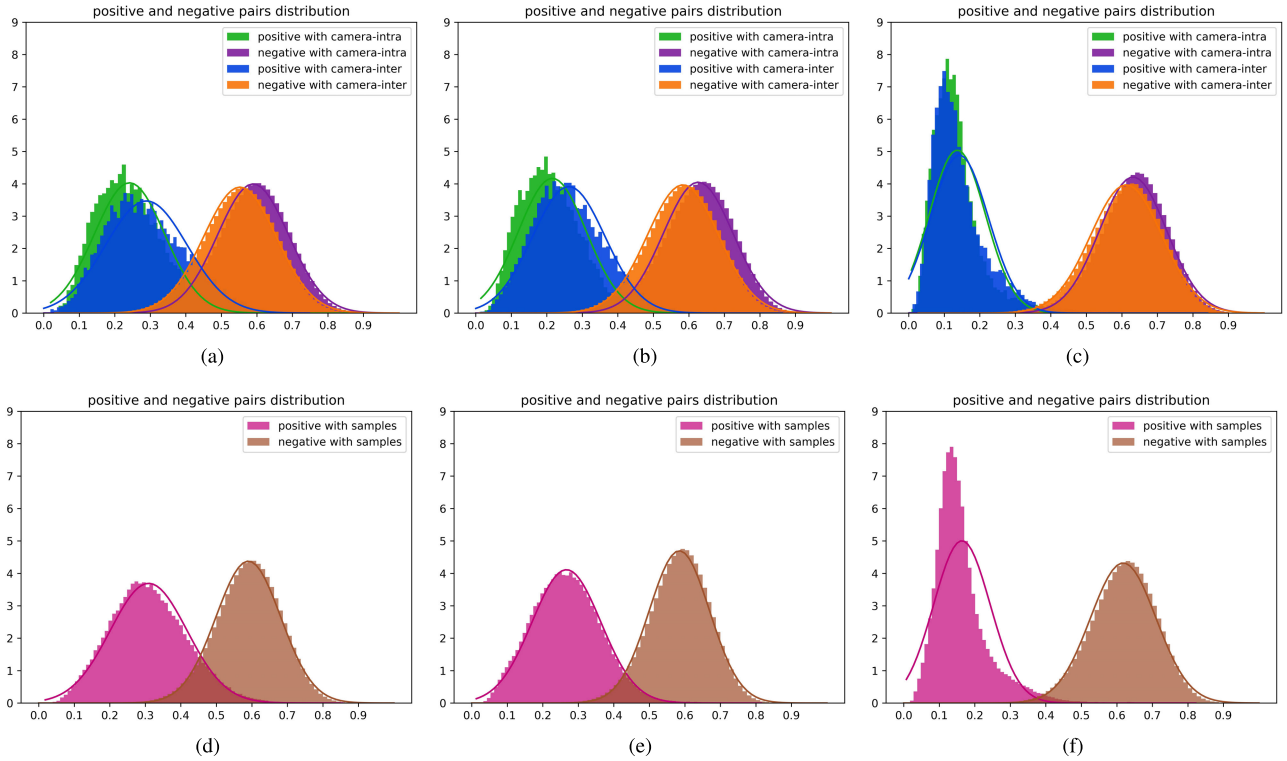

Fig. 8. (a)–(c) Feature distance distribution of positive/negative pairs from the intracameral and intercamera on baseline, intracamera learning, intercamera learning, respectively. (d)–(f) Feature distance distribution of positive/negative pairs from the training set on baseline, intracamera learning, intercamera learning, respectively.

the training data and transfer the image style from one camera to another. However, our method removes this stage, greatly reducing the complexity. Compared with MEB-NET, the number of parameters in our method is about twice as large, since our method uses the vision transformer as the backbone. However, the FLOPs and Train time of our method are similar to that of MEB-NET.

F. Parameter Analysis

Three hyperparameters μ_1, μ_2, and thres are empirically selected in our article. To demonstrate their influences, we conduct experiments by fixing one hyperparameter while changing another one. During this process, all experimental results are generated on Duke→Market1501 and Market1501→Duke tasks.

1) Influence of μ_1: In (11), the hyperparameter μ_1 mainly regulates the role of $L_{\text{tri}}^{\text{inter}}$. This loss finetunes the model through the pseudo-labels predicted by the logical relation influence to make the model adapt to different camera styles. Fig. 7(a) and (b) shows the impact of different values on Rank-1 and mAP on Duke→Market1501 and Market1501→Duke tasks. From them, we can find that when $\mu_1 \in [0.2, 1.8]$, the recognition accuracy of Rank-1 and mAP on the two tasks of the model has the potential to improve as a whole. When $\mu_1 \in [1.8, 5]$, Rank-1 and mAP on the two tasks meet the degradation.

2) Influence of μ_2: The role of the hyperparameter μ_2 is to adjusts $L_{\text{mv}}^{\text{intra},c}$ and $L_{\text{mv}}^{\text{inter}}$ in (11). The loss $L_{\text{mv}}^{\text{intra},c}$ and $L_{\text{mv}}^{\text{inter}}$ encourage the model to infer multiview information from a single pedestrian image from the aspects of intracamera and intercamera, respectively. When the hyperparameter μ_1 is set to 1.8, the value μ_2 is taken within the range of [5, 0.2]. The changes of Rank-1 and mAP are shown in Fig. 7(c) and (d) on Duke→Market1501 and Market1501→Duke tasks with different values of μ_2. It can be seen that when μ_2 is 1.4, the proposed method can obtain the best performance on both Duke→Market1501 and Market1501→Duke tasks.

3) Influence of Thres: The hyperparameter thres is a threshold value used to determine whether the sample pair (class center pair) belongs to the same identity in the LRI. Fig. 7(e) and (f) shows that the Rank-1/mAP is affected by the change of hyperparameter thres on Duke→Market1501 and Market1501→Duke tasks, when μ_1 and μ_2 are set to 1.4 and 1.8. For the Duke→Market1501 task, when the thres is 0.118, we can see that the proposed method can obtain the best performance. For Market1501→Duke task, the Rank-1 and mAP reach the peak value when the thres is 0.115. To sum up, when the source domain is Duke/Market1501, the threshold thres is set to 0.118/0.115, and both tasks can achieve the best performance.

G. Distribution Visualization

To more precisely investigate the influence of our method, we also visualize the distance distribution of positive/negative pairs. Obviously, Fig. 8(a) shows a large gap between distributions of the intracamera and intercamera. However, the overlapped area between the positive pairs and negative pairs is inversely large. On this basis, if the clustering is performed directly on all samples, images from various cameras belonging to different identities may be assigned to the same class, generating noise pseudo-labels and resulting in performance degradation. By contrast, in Fig. 8(b), for the

sake of finetuning on the intracamera samples, the difference between positive pairs and negative pairs indeed achieves a slight rise. Furthermore, thanks to the intercamera learning, Fig. 8(c) displays that the overlapping area of the distribution of positive sample pairs and negative sample pairs tends to be minimized, while the distribution differences between the intercamera and intracamera are removed. Fig. 8(d)–(f) further demonstrates the significance of our introduced intracamera learning and intercamera learning strategies. It is easy to see that, by introducing these two strategies into the model one by one, the proposed method gradually gains strong discriminant invariant-camera features.

V. Conclusion and Future Works

Pseudo-label noise is the main reason that hinders the clustering-based domain adaptation person Re-ID methods. To reduce its negative impact, a novel method based on LRI and from MvII is proposed. LRI not only removes the differences from the multiple cameras, but also achieves pseudo-labels of the hard samples, gaining much more accurate label predictions. Furthermore, to enjoy more complementary and global-consistent information from a single image, MvII is introduced through which each image is regarded as a patch token and interacted in a batch, contributing to discriminative feature extraction. Comparative experimental results confirm the effectiveness and superiority of our proposed method.

However, the proposed method still needs pairs of positive pedestrian image samples across cameras to participate in training. For different cameras with a long distance, the pedestrians captured under their cameras usually do not have the same identity. Therefore, in the future, we will explore how to use unpaired pedestrian image samples across cameras to overcome the negative impact of camera style changes for the better practical value of domain adaptation person Re-ID.

References

[1] Z. Zhang, C. Lan, W. Zeng, X. Jin, and Z. Chen, "Relation-aware global attention for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3183–3192.

[2] 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.

[3] L. Wu, Y. Wang, J. Gao, M. Wang, Z. Zha, and D. Tao, "Deep coattention-based comparator for relative representation learning in person re-identification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 2, pp. 722–735, Feb. 2021.

[4] Z. Wei, Y. Xi, N. Wang, and X. Gao, "Flexible body partition-based adversarial learning for visible infrared person re-identification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 9, pp. 4676–4687, Sep. 2022.

[5] H. Tan, X. Liu, B. Yin, and X. Li, "MHSA-Net: Multihead self-attention network for occluded person re-identification," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Mar. 21, 2022, doi: 10.1109/TNNLS.2022.3144163.

[6] K. Zhu, H. Guo, S. Liu, J. Wang, and M. Tang, "Learning semantics-consistent stripes with self-refinement for person re-identification," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Mar. 17, 2022, doi: 10.1109/TNNLS.2022.3151487.

[7] P. Chen, X. Xu, and C. Deng, "Deep view-aware metric learning for person re-identification," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 620–626.

[8] F. Zheng et al., "Pyramidal person re-identification via multi-loss dynamic training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8506–8514.

[9] 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.

[10] 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.

[11] H. Fan, L. Zheng, C. Yan, and Y. Yang, "Unsupervised person re-identification: Clustering and fine-tuning," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 14, no. 4, pp. 1–18, Nov. 2018.

[12] 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. 6111–6120.

[13] 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.*, 2020. [Online]. Available: https://openreview.net/forum?id=rJlnOhVYPS

[14] K. Zheng, W. Liu, L. He, T. Mei, J. Luo, and Z. Zha, "Group-aware label transfer for domain adaptive person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 5306–5315.

[15] K. Zheng, C. Lan, W. Zeng, Z. Zhang, and Z.-J. Zha, "Exploiting sample uncertainty for domain adaptive person re-identification," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2021, pp. 3538–3546.

[16] J. Li and S. Zhang, "Joint visual and temporal consistency for unsupervised domain adaptive person re-identification," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 483–499.

[17] 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. (ECCV)*, 2020, pp. 594–611.

[18] 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.

[19] C. Luo, C. Song, and Z. Zhang, "Generalizing person re-identification by camera-aware invariance learning and cross-domain mixup," in *Proc. Eur. Conf. Comput. vision(ECCV)*, 2020, pp. 224–241.

[20] F. Yang et al., "Joint noise-tolerant learning and meta camera shift adaptation for unsupervised person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 4853–4862.

[21] Y. Lin, Y. Wu, C. Yan, M. Xu, and Y. Yang, "Unsupervised person re-identification via cross-camera similarity exploration," *IEEE Trans. Image Process.*, vol. 29, pp. 5481–5490, 2020.

[22] T. Liu, Y. Lin, and B. Du, "Unsupervised person re-identification with stochastic training strategy," *IEEE Trans. Image Process.*, vol. 31, pp. 4240–4250, 2022.

[23] S. Xuan and S. Zhang, "Intra-inter camera similarity for unsupervised person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 11921–11930.

[24] Y. Su, Y. Li, W. Nie, D. Song, and A. Liu, "Joint heterogeneous feature learning and distribution alignment for 2D image-based 3D object retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 10, pp. 3765–3776, Oct. 2020.

[25] Z. Zhuang, L. Wei, L. Xie, H. Ai, and Q. Tian, "Camera-based batch normalization: An effective distribution alignment method for person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 1, pp. 374–387, Jan. 2022.

[26] J. Zhang, J. Liu, B. Pan, and Z. Shi, "Domain adaptation based on correlation subspace dynamic distribution alignment for remote sensing image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 11, pp. 7920–7930, Nov. 2020.

[27] L. Zhou, J. Luo, X. Gao, W. Li, B. Lei, and J. Leng, "Selective domain-invariant feature alignment network for face anti-spoofing," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 5352–5365, 2021.

[28] 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.

[29] 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.

[30] 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.

[31] Z. Zhuang et al., "Rethinking the distribution gap of person re-identification with camera-based batch normalization," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2020, pp. 140–157.

[32] H. Zhang, H. Cao, X. Yang, C. Deng, and D. Tao, "Self-training with progressive representation enhancement for unsupervised cross-domain person re-identification," *IEEE Trans. Image Process.*, vol. 30, pp. 5287–5298, 2021.

[33] 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.

[34] Y. Sheng, "Asymmetric CycleGAN for unpaired image-to-image translation based on dual attention module," in *Proc. 3rd Int. Academic Exchange Conf. Sci. Technol. Innov. (IAECST)*, Dec. 2021, pp. 726–730.

[35] J. Wang, C. Jin, W. Zhao, S. Liu, and X. Lv, "An unsupervised methodology for musical style translation," in *Proc. Int. Conf. Comput. Intell. Secur. (CIS)*, 2019, pp. 216–220.

[36] M. Li, R. Xi, and M. Hou, "Generate novel image styles using weighted hybrid generative adversarial nets," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2018, pp. 1–8.

[37] P. Zhang, B. Zhang, D. Chen, L. Yuan, and F. Wen, "Cross-domain correspondence learning for exemplar-based image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5142–5152.

[38] Y. Zhou, R. Jiang, X. Wu, J. He, S. Weng, and Q. Peng, "BranchGAN: Unsupervised mutual image-to-image transfer with a single encoder and dual decoders," *IEEE Trans. Multimedia*, vol. 21, no. 12, pp. 3136–3149, Dec. 2019.

[39] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang, "Camera style adaptation for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5157–5166.

[40] 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.

[41] 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.

[42] 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.

[43] J. Liu, Z. 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.

[44] Y. Li, C. Lin, Y. Lin, and Y. 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.

[45] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with k-reciprocal encoding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3652–3661.

[46] M. Ester et al., "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. Int. Conf. Knowl. Discovery Data Mining (KDD)*, 1996, pp. 226–231.

[47] 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. 8221–8230.

[48] L. Song et al., "Unsupervised domain adaptive re-identification: Theory and practice," *Pattern Recognit.*, vol. 102, Jun. 2020, Art. no. 107173.

[49] Z. Zhong, L. Zheng, Z. Luo, S. Li, and Y. Yang, "Learning to adapt invariance in memory for person re-identification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 8, pp. 2723–2738, Aug. 2021.

[50] 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.

[51] 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.

[52] 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 *Advances in Neural Information Processing Systems*, vol. 33, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds. Curran, 2020, pp. 11309–11321. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/821fa74b50ba3f7cba1e6c53e8fa6845-Paper.pdf

[53] Z. Zhang, Y. Wang, S. Liu, B. Xiao, and T. S. Durrani, "Cross-domain person re-identification using heterogeneous convolutional network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 1160–1171, Mar. 2022.

[54] X. Liu and S. Zhang, "Graph consistency based mean-teaching for unsupervised domain adaptive person re-identification," in *Proc. 30th Int. Joint Conf. Artif. Intell.*, Aug. 2021, pp. 874–880.

[55] Y. Dai, J. Liu, Y. Bai, Z. Tong, and L. 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.

[56] Y. Zheng et al., "Online pseudo label generation by hierarchical cluster dynamics for adaptive person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 8351–8361.

[57] 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.

[58] Z. Pang, J. Guo, Z. Ma, W. Sun, and Y. Xiao, "Median stable clustering and global distance classification for cross-domain person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 5, pp. 3164–3177, May 2022.

[59] Y. Li, H. Yao, and C. Xu, "Intra-domain consistency enhancement for unsupervised person re-identification," *IEEE Trans. Multimedia*, vol. 24, pp. 415–425, 2022.

[60] X. Song and Z. Jin, "Robust label rectifying with consistent contrastive-learning for domain adaptive person re-identification," *IEEE Trans. Multimedia*, vol. 24, pp. 3229–3239, 2022.

[61] 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.

[62] 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. 3774–3782.

[63] 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.

[64] 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. (ECCV)*. Cham, Switzerland: Springer, 2016, pp. 17–35.

[65] R. Zhao, W. Ouyang, and X. Wang, "Person re-identification by salience matching," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2528–2535.

[66] 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.

[67] 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. 14993–15002.

[68] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[69] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, vol. 32, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. Curran, 2019. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf

[70] 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.

[71] 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.