# Triple Adversarial Learning and Multi-View Imaginative Reasoning for Unsupervised Domain Adaptation Person Re-Identification

Huafeng Li, Neng Dong, Zhengtao Yu, Dapeng Tao, *Member, IEEE*, and Guanqiu Qi

***Abstract*— Due to the importance of practical applications, unsupervised domain adaptation (UDA) person re-identification (re-ID) has attracted increasing attention. However, most of existing methods often lack the multi-view information reasoning and ignore the domain discrepancy of the pedestrian images with the same identity, which constrain the further improvement of recognition performance. So, this paper proposes a triple adversarial learning and multi-view imaginative reasoning network (TAL-MIRN) for UDA person re-ID, which consists of a multi-view imaginative reasoning module (IRM) and a triple adversarial learning module (TALM). IRM makes the classified pedestrian identity features from a single-view image extracted by a feature encoder consistent with the classification results of the aggregated multi-view pedestrian identity features, so the strong multi-view imaginative reasoning ability of the feature encoder is obtained. TALM is composed by the adversarial learning between the camera classifier and feature encoder, adversarial learning of joint distribution alignment, and adversarial learning of the difference between two classifiers used in classification. In particular, the domain-invariant features at camera level are guaranteed by the adversarial learning between the feature extractor and camera classifier. The joint alignment of identity and domain is achieved by the competition between the feature extractor and classifier integrated with identity and domain. The discriminability and robustness of the learned features are enhanced by playing a MinMax game between two different identity classifiers. Furthermore, a simple normalization operation named as cross normalization (CN) is proposed to increase both modeling and generalization capability of the proposed TAL-MIRN across multiple domains. The proposed TAL-MIRN is applied to five benchmark datasets, and the comparative experimental results confirm its superiority over the state-of-the-art methods. The related source codes is available at https://github.com/lhf12278/TALM-IRM.**

***Index Terms*— Person re-identification, unsupervised domain adaptation, multi-view information reasoning, triple adversarial learning.**

## I. Introduction

PERSON re-identification (re-ID) as an important intelligent surveillance technique aims to extract and identify people of interest from the images captured by non-overlapping cameras. Benefitting from the recent advancement of deep learning and big data, person re-ID has drawn widespread attention from both academia and industry. Some high-performance person re-ID methods have been proposed [1]–[10], but their performance relies on the supervised learning with a large amount of labeled samples. However, it is extremely expensive to manually label such a large amount of samples. When the model trained by supervised learning is directly applied to any unlabeled target dataset, its performance may be unsatisfactory due to the domain shift between the labeled training dataset and unlabeled target dataset [11]–[13].

As a popular and effective solution, unsupervised domain adaptation (UDA) is used to solve the domain shift issue of person re-ID. The recently published UDA-based person re-ID solutions involve self-labeling [14]–[18], transferring image styles [19]–[21], and extracting domain-invariant features [22]–[27]. These solutions generally focus on how to extract domain-invariant features for pedestrian identity matching. However, they do not have the ability of imagination and reasoning like people. So, it is difficult to extract the complementary features shown in different views from a single image. As shown in Fig. 1, the appearance of the same pedestrian varies considerably in different camera views. Existing person re-ID methods tend to extract the shared pedestrian information of the images captured from different camera views to facilitate the identity matching. For example, one pedestrian wears a backpack. If the backpack does not appear in all the captured images of this pedestrian, the backpack related information may not be extracted from the captured

Manuscript received October 27, 2020; revised March 11, 2021; accepted July 17, 2021. Date of publication July 26, 2021; date of current version May 5, 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 61966021, Grant 61772455, and Grant 61562053; in part by the National Key Research and Development Plan Project under Grant 2018YFC0830105 and Grant 2018YFC0830100; in part by the Yunnan Provincial Major Science and Technology Special Plan Projects: Digitization Research and Application Demonstration of Yunnan Characteristic Industry under Grant 202002AD080001; in part by the Yunnan Natural Science Funds under Grant 2018FY001(-013) and Grant 2019FA-045; and in part by the Yunnan University Natural Science Funds under Grant 2018YDJQ004. This article was recommended by Associate Editor Q. Tian. *(Huafeng Li and Neng Dong contributed equally to this work.) (Corresponding authors: Dapeng Tao; Zhengtao Yu.)*

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

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

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

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSVT.2021.3099943.

Digital Object Identifier 10.1109/TCSVT.2021.3099943

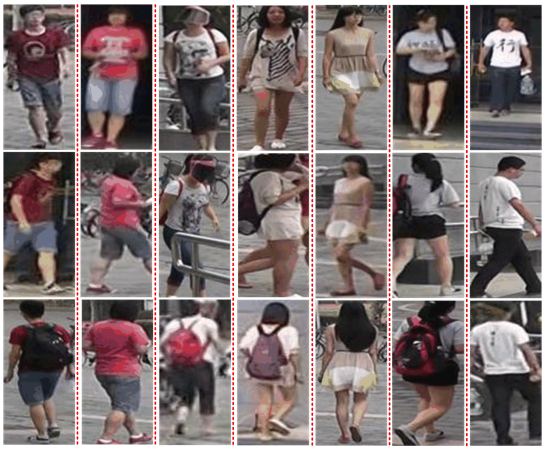

Fig. 1. The appearance of various pedestrians captured from different camera views. The images of each column show the same pedestrian captured from three different camera views.

pedestrian images. Due to the label constraints, the relevance of multiple views can be guaranteed in supervised person re-ID. However, the features extracted from each view may be different or incomplete. So, it is difficult to match the pedestrian identity across multiple cameras in UDA person re-ID without the multi-view imaginative reasoning ability.

Inspired by the imaginative reasoning ability of people, the appearance of a pedestrian captured from one camera view can be used to imaginatively reason the appearance of the same pedestrian in another camera view. The application of imaginative reasoning in the feature extraction network helps to obtain the comprehensive descriptions of pedestrian appearance, so the extracted pedestrian features are conducive to person re-ID [28]. In addition, the domain-invariance and discriminability of the extracted pedestrian features as two key factors also affect the recognition performance improvement of person re-ID. Since the target-domain samples are unlabeled and the identities do not have any overlapping between target and source domains, it is challenging to align the features of each pedestrian captured from different camera views in person re-ID. Existing methods achieve the alignment of source and target domains to extract domain-invariant features. However, these methods ignore the impact of the domain difference (identity-level domain discrepancy) among the images of the same pedestrian captured from different camera views on the recognition performance.

So, this paper proposes a novel triple adversarial learning and multi-view imaginative reasoning network (TAL-MIRN). As shown in Fig. 2, an imaginative reasoning module (IRM) is first constructed to obtain the comprehensive descriptions of pedestrian appearance, and then a triple adversarial learning module (TALM) is developed to alleviate the domain shift between the target and source domains. Particularly, the multi-view imaginative reasoning ability of the feature encoder of TAL-MIRN is obtained by making the classified pedestrian identity features from a single-view image extracted by the feature encoder consistent with the classification results of the aggregated multi-view pedestrian identity features.

The developed TALM contains a domain-invariant feature extraction (DIFE) sub-module, a joint distribution alignment of identity and domain (JDAID) sub-module, and a feature discriminability and robustness improvement (FDRI) sub-module. In DIFE, the camera classifier and feature encoder are trained in an adversarial manner. The domain-invariant features at camera level are obtained by forcing the camera classifier to misclassify the image features extracted from different camera views into the same additional category. In traditional adversarial learning-based domain-invariant feature extraction, adversarial learning is carried out between the feature extractor and domain classifier to achieve the distribution alignment of source and target domains. In TALM, adversarial learning is implemented between the camera classifier and feature extractor at camera level, which only eliminates the inconsistency of camera information. According to the guidance of camera labels, it is conducive to the retention of discriminative features. In DIFE, the domain alignment of discriminative features is only achieved at camera level, which does not guarantee the identity-level alignment of the same features simultaneously.

To alleviate the issues caused by the unrealized identity-level alignment, a novel adversarial learning strategy is developed based on a classifier integrated with identity and domain to achieve the joint distribution alignment of both identity and domain in JDAID. The integrated classifier can differentiate both identity and domain information of pedestrian images at the same time. Inspired by existing UDA solutions [29], [30], a novel adversarial learning is developed in FDRI to further enhance the discriminability and robustness of learning features, in which two classifiers are assigned to the features of each pedestrian image, and adversarial learning is applied to the two classifiers. One identity classifier is used in DIFE, and the other classifier integrated with identity and domain is used in JDAID after removing the domain classification neurons.

Moreover, to further promote both the discriminability of the learned features and the generalization capability of the learned model, a novel normalization method named as cross normalization (CN) is developed by performing both instance normalization (IN) [31] and batch normalization (BN) [32] on a single feature map and cross-concatenating the normalized results of different groups of feature maps. Compared with the instance batch normalization (IBN) [33], CN can apply the integrated IN and BN to each feature map, so the generalization capability of the proposed re-ID method is further improved.

There are three main contributions of the proposed method.

- To reason the multi-view appearance information from a single-view pedestrian image, the TAL-MIRN is endowed with the multi-view imaginative reasoning ability by making the pedestrian identity prediction results of the features extracted by the feature encoder consistent with the corresponding prediction results of the aggregated multi-view features.
- The classifier integrated with identity and domain and feature encoder compete with each other to achieve the joint alignment of identity and domain. In addition, adversarial

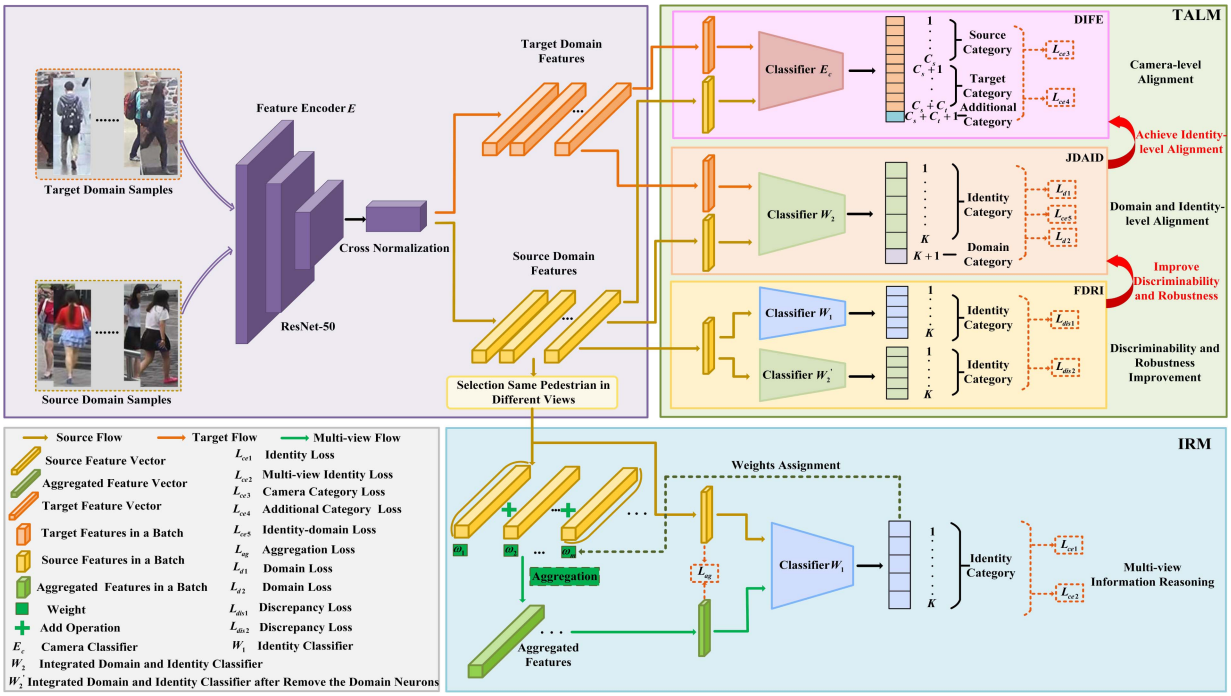


Fig. 2. An overview of the proposed TAL-MIRN. The architecture of the proposed TAL-MIRN mainly consists of two modules, IRM and TALM. TALM contains three sub-modules, DIFE, JDAID, and FDRI. CN is used to improve the generalization ability of TAL-MIRN. The imaginative reasoning ability of TAL-MIRN is obtained by making the predicted pedestrian identity of the features extracted by the feature encoder consistent with the corresponding prediction results of the aggregated multi-view features. Subsequently, source and target data are used to optimize TALM, which guarantees the domain invariance of the learned features and further improves the discriminability and robustness of the learned features. $\boldsymbol{W}_1$ and $\boldsymbol{W}_2'$ represent two different identity classifiers, and $\boldsymbol{E}_C$ denotes a camera-style network.

learning is conducted between the identity classifier and integrated classifier obtained after removing the domain classification neurons to improve the discriminability of the learned features.

- A simple and effective normalization method named as CN is proposed to improve the cross-domain generalization ability of the proposed method, in which IN and BN are integrated.

The rest of the paper is organized as follows: Section II reviews the related work; Section III discusses the proposed TAL-MIRN for UDA person Re-ID; Section IV analyzes the comparative experimental results; and Section V concludes this paper.

## II. Related Work

### A. Unsupervised Self-Labeling Person Re-ID

Unsupervised person re-ID solutions were proposed to improve the generalization ability of person re-ID models. Existing unsupervised person re-ID solutions focus on learning views-invariant features from the unlabeled target domain [34], [35]. However, due to the lack of pairwise-label guidance, the performance of existing unsupervised person re-ID solutions is still unsatisfactory. According to the recent discovery that the label estimation can play a positive role in improving the performance of existing unsupervised person re-ID solutions, the self-label estimation has attracted considerable attention in person re-ID [15], [16], [18], [36]–[39]. Specifically, Lin *et al.* [37] developed a bottom-up clustering approach for unsupervised person re-ID. Li *et al.* [38] proposed a new concept of soft tracklet labelling to explore the inherent space-time visual correlation for unsupervised person re-ID. To address the issues caused by lacking pairwise-label guidance in unsupervised re-ID, Yu *et al.* [39] developed the soft multi-label learning to explore the potential label information. Zhao *et al.* [40] proposed a noise resistible mutual-training method to suppress the noise in the predicted pseudo labels which seriously affects the model training. To mitigate the negative effects of noisy pseudo labels, Ge *et al.* [41] proposed an unsupervised framework with mutual mean-teaching. Zhai *et al.* [42] proposed a new augmented discriminative clustering solution to achieve the pedestrian pseudo-label prediction.

These solutions achieve better performance than domain-invariant feature extraction solutions, when all the target-domain samples participating in pseudo-label prediction have the corresponding positive samples. However, in practice, it is common that the isolated samples do not have any corresponding positive samples. As a negative effect on the prediction of pseudo labels, the isolated samples may cause the performance of the related solutions in real-world scenes is inferior to the corresponding performance on public datasets. Compared with these methods, the performance of the proposed solution greatly reduces the dependence on the paired positive samples, so the practicability of the proposed solution is improved.

### B. Style Transfer-Based Person Re-ID

As a well-known method, image-to-image translation is applied to UDA person re-ID [20], [43], which benefits from the powerful generative ability of

Cycle-GAN [44]. In particular, Zhong *et al.* [45] proposed a hetero-homogeneous learning method based on camera style transfer to achieve both camera invariance and domain connection. Wang *et al.* [46] presented a transferable attribute-identity deep network to improve the extendibility of person re-ID model. Ren *et al.* [20] developed a camera style adaptation framework to narrow the gaps between different domains, in which the intrinsic local structure of target domain was explored by a soft-labeling method to further reduce the gaps. Huang *et al.* [24] presented the suppression of background shift for the generative adversarial network to generate a style-consistent image with the suppressed background in UDA person re-ID.

However, these methods are difficult to ensure that the identity information of the original images does not shift during the image-to-image translation. To alleviate this issue, Wei *et al.* [43] first used a segmentation network to extract the person-related areas from a pedestrian image, and then employed an identity loss to ensure the accuracy of the identity cues in the transferred pedestrian images. Deng *et al.* [47] developed a generative adversarial network for similarity preservation to keep the identity information from being tampered during style transfer. Although the above methods can preserve the underlying identity information during image-to-image translation, the preservation of the visual cues associated with the identity is still an open problem. Moreover, the style-transfer based re-ID models need to transfer the style of training samples from source domain to target domain to alleviate the domain shift issue between source and target domains, which seriously affects the practical applications of these models in real-world scenes. The proposed method does not use style transfer to solve the domain shift issue. Therefore, compared with the style-transfer based person re-ID methods, the applicability of the proposed method is effectively improved.

### C. Domain-Adaptive Person Re-ID

Based on the research results a few years ago, the good recognition performance of person re-ID can be achieved by learning domain-adaptive or domain-invariant features on the labeled source domain and transferring them to target domain. Due to the promising performance, dictionary learning has received considerable attention in image processing and pattern recognition [48]–[51] in the past few years. Peng *et al.* [50] applied a novel dictionary learning method to UDA person re-ID, in which the dictionary space was first decomposed into semantic, latent discriminative, and latent background attributes, and then the predicted semantic attributes and discriminative latent attributes were treated as the features in similarity measuring. Qi *et al.* [52] developed an unsupervised joint subspace and dictionary learning for cross-domain person re-ID, which can effectively alleviate the domain shift between source and target domains by jointly learning both cross-view and cross-domain variations. However, the dictionary learning based methods cannot effectively explore the discriminative information contained in large-scale data, so their recognition performance is far blow people's expectations.

Deep learning can alleviate the deficiency of dictionary learning, so it has attracted wide attention in domain-invariant feature extraction. Particularly, Song *et al.* [53] presented a domain-invariant mapping network for unsupervised person re-ID. A meta-learning pipeline was used, in which a subset of source dataset was employed to make the learned features domain-invariant. To improve the performance of UDA person re-ID, Qi *et al.* [25] solved two main challenges of person re-ID (the data distribution discrepancy between source and target domains and the lack of label information in target domain) from the perspective of representation learning, and Liu *et al.* [54] addressed the domain gap between source and target domains by using a novel adaptive transfer network. Yang *et al.* [23] proposed a patch-based discriminative feature learning method to improve the discriminability and scalability of the learned features. Wu *et al.* [55] developed a camera-aware similarity consistency learning approach to learn the similarity-consistent domain-invariant features for UDA person re-ID. Zhong *et al.* [56] utilized three types of underlying invariance to learn the domain-adaptive features and introduced an exemplary memory to store the target-domain features. Compared with the previous two categories, domain-adaptive methods may have higher practicability in real-world scenes, because they get rid of the dependence on positive target-domain samples, and do not need any additional models for assistance. However, due to lacking the fine-tuning of pseudo-labels and assistance of additional models (such as style transfer model), domain-adaptive methods often show poor recognition performance on public datasets. The proposed method is a domain-adaptive person re-ID method, but it is more effective than existing methods. The details and performance of the proposed method will be discussed in the following sections.

## III. The Proposed Method

Although the significant progress has been made in UDA pedestrian re-ID, the performance of existing solutions is still far from satisfaction. As a reason, existing feature extraction methods only directly extract features from a single image, but cannot imaginatively reason multi-view features from a single image. In addition, both the low discriminability of the learned features and domain discrepancy at identity level are also another two key factors affecting the recognition performance. To address these issues, this paper proposes a novel TAL-MIRN to obtain discriminative domain-invariant features.

### A. Overview

As shown in Fig. 2, the proposed method has two main components, IRM and TALM. IRM uses the labeled source samples to train, which can imaginatively reason the aggregated features from a single-view pedestrian image. TALM consists of three sub-modules, DIFE, JDAID, and FDRI. DIFE extracts domain-invariant features at camera level. JDAID achieves the joint distribution alignment of identity and domain. FDRI

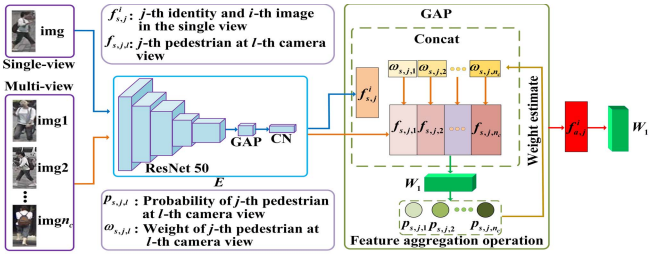


Fig. 3. The illustration of IRM module. The feature aggregated by each multi-view feature $\boldsymbol{f}_{s,j,l}$ is classified to fit the corresponding part of the single-view feature $\boldsymbol{f}^i_{s,j}$ for realizing the back propagation from multi-view to single-view features, which endows the IRM module with the imaginative reasoning ability.

guarantees the discriminability and robustness of ambiguous appearance features. In addition, a simple normalization named as cross normalization (CN) is proposed and integrated into the proposed network framework.

### *B. Imaginative Reasoning Module (IRM)*

As a key factor, the comprehensive descriptions of a pedestrian's appearance are conducive to improving the recognition performance. The feature encoder used in IRM can imaginatively reason the aggregated multi-view features from a single-view pedestrian image. As a result, the extracted features from a single-view image can incorporate the information from other single-view images. Specifically, as shown in Fig. 3, this module contains a feature encoder $\boldsymbol{E}$ and a feature aggregation operation. The feature aggregation operation aggregates the features of the same pedestrian obtained from different camera views. The feature encoder $\boldsymbol{E}$ is a ResNet-50 [57] network with the parameters pre-trained on ImageNet [58], and the stride of the last spatial downsampling operation is set to 1 as the backbone. In addition, global average pooling (GAP) and CN layers are added to the backbone. The identity classifier $\boldsymbol{W}_1$ is composed of a $K$-dimensional fully-connected layer, where $K$ indicates the number of pedestrians in source domain.

Supervised training is first performed on the labeled source domain, which allows the encoder $\boldsymbol{E}$ to extract discriminative features. In particular, given the labeled source-domain sample set $\boldsymbol{X}_s = \{x^i_s\}^{N_s}_{i=1}$ and unlabeled target-domain sample set $\boldsymbol{X}_t = \{x^i_t\}^{N_t}_{i=1}$, where $N_s$ and $N_t$ represent the total number of samples in the source and target domains respectively, the features are extracted by using $\boldsymbol{E}$ and sent to the identity classifier $\boldsymbol{W}_1$ to obtain an identity score first. Then, $\boldsymbol{E}$ and $\boldsymbol{W}_1$ are optimized by using the cross-entropy loss of both the identity score and its corresponding label. The above process can be formulated as follows:

$$\boldsymbol{L}_{ce1}(\boldsymbol{W}_1, \boldsymbol{E}) = -\frac{1}{n_b}\sum_{i=1}^{n_b}\sum_{k=1}^{K}\hat{\mathbb{1}}_{[k=y^i_s]}\log p(\boldsymbol{W}_1(\boldsymbol{E}(x^i_s))), \quad (1)$$

where $\hat{\mathbb{1}}_{[k=y^i_s]}$ is the indicator function, $n_b$ is the batch size, and $\boldsymbol{Y}_s = \{y^i_s\}^{N_s}_{i=1}$ is ground truth label set of $K$ pedestrians. In UDA person re-ID, there is no any identity overlap between source and target domains, so the model trained on source domain may suffer from overfitting when it is applied to target domain. To solve the overfitting issue, the label smoothing is used to define the indicator function $\hat{\mathbb{1}}_{[k=y^i_s]}$ [59] as follows:

$$\hat{\mathbb{1}}_{[k=y^i_s]} = \begin{cases} 1-\varepsilon\dfrac{K-1}{K}, & k = y^i_s \\ \dfrac{\varepsilon}{K}, & \text{otherwise}, \end{cases} \quad (2)$$

where $\varepsilon$ as a constant is set to 0.1.

To endow IRM with the strong multi-view imaginative reasoning ability, the classified single-view features of the same pedestrian are made consistent with the classification results of the corresponding aggregated multi-view features. However, the features of the same pedestrian from different views have different discrimination abilities. For the same pedestrian under different camera views, the less shared information exists, the identity-related information may have greater differences. In the feature aggregation, high weights are assigned to the features of the same pedestrian with great differences under different camera views, which is beneficial to improve the performance of IRM. (In fact, when the appearance features of the same pedestrian show great differences under different views, the related information captured from different views is likely to be complementary. High weights are assigned to the complementary information.) So, a weighted aggregation strategy is introduced, which adaptively assigns the corresponding weight to each image according to its discriminability, and no any additional network parameter is involved.

As illustrated in Fig. 3, the $i$-th image feature $\boldsymbol{f}^i_{s,j}$ with identity label $j$ (i.e. $j = y^i_s$) only contains the pedestrian information shown in the current view. To improve the identification performance, IRM is applied to reason multi-view image features from single-view image features. As discussed earlier, different appearance features of the same pedestrian are captured in different camera views, which are usually complementary. The feature $\boldsymbol{f}_{s,j,l}$ of any image of the $j$-th pedestrian at $l$-th camera view is sent to the identity classifier, and then the probability $p_{s,j,l}$ belonging to this pedestrian is obtained. The feature $\boldsymbol{f}_{s,j,l}$ with a high confidence probability $p_{s,j,l}$ often implies that it has strong discrimination ability and carries less complementary information. (If the features of the same pedestrian captured in different views are not the shared features, the pedestrians from different views have a low probability of being identified as the same pedestrian due to the low similarity of the captured features, and vice versa.) A small weight is assigned to such feature, and vice versa. Thus, its weight $\omega_{s,j,l}$ can be determined as follows:

$$\omega_{s,j,l} = 1 - p_{s,j,l}. \quad (3)$$

According to the learned weight $\omega_{s,j,l}$, the average weighted features of multi-view pedestrian images are obtained and then concatenated with one single-view image feature $\boldsymbol{f}^i_{s,j}$. A GAP operation is performed on the aggregated features to obtain the final feature $\boldsymbol{f}^i_{a,j}$, which contains multi-view pedestrian information as follows:

$$\boldsymbol{f}^i_{a,j} = \mathrm{GAP}(\mathrm{Concat}(\boldsymbol{f}^i_{s,j}, \frac{1}{n_c}\sum_{l=1}^{n_c}\omega_{s,j,l}\boldsymbol{f}_{s,j,l})), \quad (4)$$

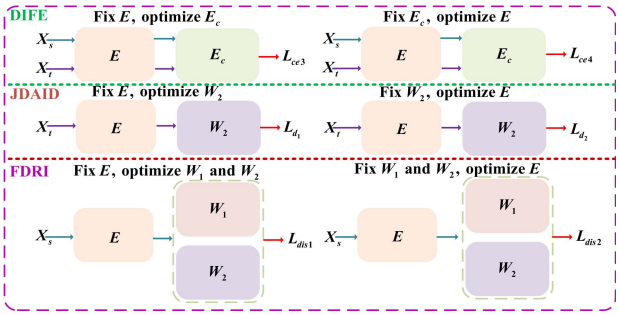


Fig. 4. The illustration of TALM module. The first adversarial learning process is performed between $\boldsymbol{E}$ and $\boldsymbol{E}_c$ in DIFE, which ensures the extracted features are domain-invariant. The second adversarial learning process is performed between $\boldsymbol{E}$ and $\boldsymbol{W}_2$ to achieve the joint distribution alignment of identity and domain. The third adversarial learning process is performed between $\boldsymbol{E}$ and $\boldsymbol{W}_1, \boldsymbol{W}_2$ to improve the discriminability and robustness of the learned features. Three adversarial learning sub-modules are processed simultaneously.

where $n_c$ denotes the number of multi-view images of the $j$-th person. Different from the average strategy used in [28], a weight $\omega_{s,j,l}$ predicted from the probability $p_{s,j,l}$ is assigned to image feature $\boldsymbol{f}_{s,j,l}$ to adaptively adjust its role (when the value of $\omega_{s,j,l}$ is large, $\boldsymbol{f}_{s,j,l}$ plays an important role in feature aggregation). Similar to the supervised learning of source-domain features, the cross-entropy loss is used to ensure the discriminability of the aggregated feature $\boldsymbol{f}^i_{a,j}$ as follows:

$$\boldsymbol{L}_{ce2}(\boldsymbol{W}_1, \boldsymbol{E}) = -\frac{1}{n_b}\sum_{i=1}^{n_b}\sum_{k=1}^{K}\hat{\mathbb{1}}_{[k=y^i_s]}\log p(\boldsymbol{W}_1(\boldsymbol{f}^i_{a,j})). \quad (5)$$

Minimizing the loss functions $\boldsymbol{L}_{ce1}(\boldsymbol{W}_1, \boldsymbol{E})$, $\boldsymbol{L}_{ce2}(\boldsymbol{W}_1, \boldsymbol{E})$ can make the classification results of the aggregated pedestrian identity features consistent with the classified single-view pedestrian identity features. To further facilitate the extraction of multi-view image features, the model is optimized by minimizing the following aggregation ($l_2$)-loss.

$$\boldsymbol{L}_{ag}(\boldsymbol{E}) = \|\boldsymbol{f}^i_{s,j} - \boldsymbol{f}^i_{a,j}\|_2. \quad (6)$$

Benefiting from the above design, the encoder $\boldsymbol{E}$ is able to imaginatively reason multi-view image information from a single-view image.

### C. Triple Adversarial Learning Strategy

In the proposed method, DIFE and JDAID sub-modules learn the domain-invariant features, and FDRI sub-module improves the discriminability and robustness of the learned features. As shown in Fig. 4, all three sub-modules use adversarial learning, so they are named as a triple adversarial learning strategy.

*1) DIFE Sub-Module:* In practice, each camera has its own imaging style, which as a main factor affects the domain shift between the images captured by different cameras. To obtain domain-invariant features, a camera style network $\boldsymbol{E}_c$ is first applied to the features extracted by the feature encoder $\boldsymbol{E}$. Then, adversarial learning is carried out between $\boldsymbol{E}_c$ and $\boldsymbol{E}$ to align the features from different camera views. Existing camera-aware domain adaptation methods [25] utilize adversarial learning to make the camera classifier classify the training samples into different classes with equal probability. Compared with existing methods, the proposed method can achieve more effective feature alignment at domain level by classifying training samples into the same additional category. As a newly added category, the additional category is different from all the previous categories. After domain alignment, all the features are classified into this additional category. Particularly, the dimension of the fully-connected layer is set to $C_s + C_t + 1$, where $C_s$ and $C_t$ represent the number of cameras in source and target domains, respectively.

Given the source-domain image $x^i_s$ and target-domain image $x^i_t$, $\boldsymbol{E}_c$ is trained to minimize the following loss, so $\boldsymbol{E}_c$ can correctly recognize the corresponding camera IDs of $x^i_s$ and $x^i_t$.

$$\begin{aligned}\boldsymbol{L}_{ce3}(\boldsymbol{E}_c) = &-\frac{1}{n_b}\sum_{i=1}^{n_b}\sum_{c=1}^{C_s}\mathbb{1}_{[c=y^c_s]}\log p(\boldsymbol{E}_c(\boldsymbol{E}(x^i_s))) \\ &-\frac{1}{n_b}\sum_{i=1}^{n_b}\sum_{c=1}^{C_t}\mathbb{1}_{[c=y^c_t]}\log p(\boldsymbol{E}_c(\boldsymbol{E}(x^i_t))). \quad (7)\end{aligned}$$

When $\boldsymbol{E}$ is fixed, the camera IDs of $\boldsymbol{E}(x^i_t)$ and $\boldsymbol{E}(x^i_s)$ need to be distinguished for $\boldsymbol{E}_c$. However, when $\boldsymbol{E}_c$ is updated, $\boldsymbol{E}$ is updated to ensure the features $\boldsymbol{E}(x^i_t)$ and $\boldsymbol{E}(x^i_s)$ extracted by $\boldsymbol{E}$ can be classified into the $(C_s + C_t + 1)$-th category after passing $\boldsymbol{E}_c$. Given $y^c_c = C_s + C_t + 1$, the loss function specialized for updating $\boldsymbol{E}$ is shown as follows:

$$\begin{aligned}\boldsymbol{L}_{ce4}(\boldsymbol{E}) = &-\frac{1}{n_b}\sum_{i=1}^{n_b}\sum_{c=1}^{C_s+C_t+1}\mathbb{1}_{[c=y^c_c]}\log p(\boldsymbol{E}_c(\boldsymbol{E}(x^i_s))) \\ &-\frac{1}{n_b}\sum_{i=1}^{n_b}\sum_{c=1}^{C_s+C_t+1}\mathbb{1}_{[c=y^c_c]}\log p(\boldsymbol{E}_c(\boldsymbol{E}(x^i_t))). \quad (8)\end{aligned}$$

The source- and target-domain features can be projected into an aligned space through continuous gaming and optimization of $\boldsymbol{E}$ and $\boldsymbol{E}_c$.

*2) JDAID Sub-Module:* DIFE sub-module achieves the feature alignment at camera level (i.e, domain alignment of samples captured by different cameras). Except the differences in camera style, other factors can also cause the domain discrepancy. So, the features with the same identity from difference camera views are not guaranteed to be aligned. According to the recent research of UDA, the joint optimization of category and domain classifiers can effectively alleviate this issue [60], [61]. A novel adversarial strategy is developed to achieve the alignment of the learned features at identity level (i.e. the distribution alignment of images with the same identity). Specifically, the domain and identity classifiers are integrated. So, the joint distribution alignment of both identity and domain can be achieved simultaneously in the proposed adversarial learning.

Different from $\boldsymbol{W}_1$, $\boldsymbol{W}_2$ as an integrated classifier consists of a fully-connected layer with $(K + 1)$-dimension, and the $(K + 1)$-th dimension is the domain category. This paper assumes the probability of $\boldsymbol{W}_2$ in classifying the $i$-th image features into the corresponding identity categories is $\theta$, and

the probability of $\boldsymbol{W}_2$ in classifying the domain information of $i$-th image into the $K+1$-th category is $1-\theta$. Given $\bar{\boldsymbol{y}}_s^i = [\underbrace{0, 0, \cdots, \theta}_{i}, \cdots 0, 1-\theta]$ as the joint label of the $i$-th image, the encoder $\boldsymbol{E}$ and $\boldsymbol{W}_2$ can be optimized by minimizing the following cross-entropy equation.

$$\boldsymbol{L}_{ce5}(\boldsymbol{W}_2, \boldsymbol{E}) = -\Big(\frac{1}{n_b}\sum_{i=1}^{n_b}\sum_{k=1}^{K}\bar{\boldsymbol{y}}_{s[k]}^i \log p_{id}(\boldsymbol{W}_2(\boldsymbol{E}(x_s^i))) + \bar{\boldsymbol{y}}_{s[K+1]}^i \log p_{do}(\boldsymbol{W}_2(\boldsymbol{E}(x_s^i)))\Big), \quad (9)$$

where $\bar{\boldsymbol{Y}}_s = [\bar{\boldsymbol{y}}_s^1, \bar{\boldsymbol{y}}_s^2, \ldots, \bar{\boldsymbol{y}}_s^K]$, $\bar{\boldsymbol{y}}_{s[k]}^i$ denotes the $k$-th element in $\bar{\boldsymbol{y}}_s^i$, $p_{id}$ is the probability of $x_s^i$ belonging to a specific identity, and $p_{do}$ is the probability of $x_s^i$ belonging to the domain category. Due to the domain shift between source and target domains, $\boldsymbol{W}_2$ cannot categorize target-domain samples into the same domain class of source-domain samples as follows:

$$\boldsymbol{L}_{d_1}(\boldsymbol{W}_2) = \frac{1}{n_t}\sum_{j=1}^{n_t}\log p_{do}(\boldsymbol{W}_2(\boldsymbol{E}(x_t^j))), \quad (10)$$

where $n_t$ represents the number of target-domain samples in a mini-batch. To further align the feature distribution of source and target domains, the features extracted by the encoder $\boldsymbol{E}$ confuse the integrated classifier $\boldsymbol{W}_2$ as follows:

$$\boldsymbol{L}_{d_2}(\boldsymbol{E}) = -\frac{1}{n_t}\sum_{j=1}^{n_t}\log p_{do}(\boldsymbol{W}_2(\boldsymbol{E}(x_t^j))). \quad (11)$$

The above adversarial learning process not only ensures the discriminability of the learned features, but also further enhances the domain invariance of the learned features, which are conducive to improving the performance of UDA person re-ID.

*3) FDRI Sub-Module:* To further improve the discriminability and robustness of the learned domain-invariant features, two different classifiers $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ are used at the same time to recognize pedestrian identities. As mentioned in the above discussion, these two classifiers are learned in different ways, so the discriminability of the learned features can be improved from different perspectives. When the two classifiers give the consistent classification results for the same image, the learned features are robust and discriminative. However, due to the difference in the output dimension of the two classifiers, they cannot be used together directly. To this end, domain neurons are first moved from $\boldsymbol{W}_2$, then the classifier $\boldsymbol{W}_2'$ is obtained. After that, classifiers $\boldsymbol{W}_1$ and $\boldsymbol{W}_2'$ can be optimized by adversarial learning. Specifically, after updating the encoder $\boldsymbol{E}$, the two classifiers $\boldsymbol{W}_1$ and $\boldsymbol{W}_2'$ are updated to classify the pedestrian images of the same identity into different individual classes according to the different focus of two classifiers. Meanwhile, $\boldsymbol{W}_1$ and $\boldsymbol{W}_2'$ can correctly identify the pedestrian images with the same identity. After updating $\boldsymbol{W}_1$ and $\boldsymbol{W}_2'$, the encoder $\boldsymbol{E}$ is updated in turn. The corresponding loss functions are formulated as follows:

$$\boldsymbol{L}_{dis_1}(\boldsymbol{W}_1, \boldsymbol{W}_2') = -\frac{1}{n_b}\sum_{i=1}^{n_b}[d(\boldsymbol{W}_1(\boldsymbol{E}(x_s^i)) - \boldsymbol{W}_2'(\boldsymbol{E}(x_s^i)))], \quad (12)$$

$$\boldsymbol{L}_{dis_2}(\boldsymbol{E}) = \frac{1}{n_b}\sum_{i=1}^{n_b}[d(\boldsymbol{W}_1(\boldsymbol{E}(x_s^i)) - \boldsymbol{W}_2'(\boldsymbol{E}(x_s^i)))], \quad (13)$$

where $d(\cdot)$ denotes the $L_1$-norm distance metric.

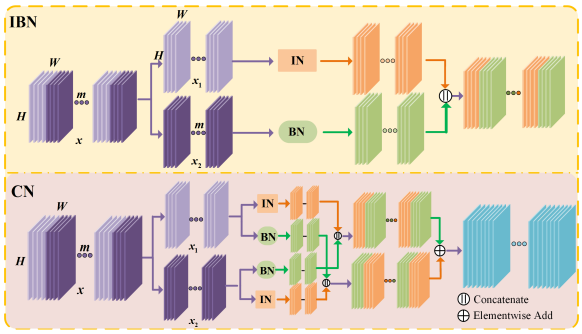


Fig. 5. Comparison between the traditional instance-batch normalization (IBN) and the proposed cross normalization (CN).

### D. Network Structure and Optimization

*1) Network Structure:* The proposed method uses ResNet-50 pre-trained on ImageNet as the backbone following a GAP, which resizes features to 2,048-dimensional vectors. Moreover, the spatial downsampling operation is used as the last stride in the backbone network, and the last stride is set to 1 according to the suggestion mentioned in [62]. Due to the lack of identity labels in target domain, it is difficult for DIFE and JDAID sub-modules to completely align the domain distribution of target-domain samples at identity level. To improve both the modeling and generalization abilities from source domain to target domain, IBN as a simple and effective method integrates IN and BN. However, only one type of normalization is performed on one feature map in IBN. In IBN, parts of feature maps use IN, and the remaining parts use BN. Therefore, the advantages of IN and BN cannot be reflected in a feature map at the same time, which constrains the further optimization of the model generalization ability. So, a novel normalization named as cross normalization (CN) is integrated into the backbone.

Specifically, as shown in Fig. 5, the feature maps are first divided into two equal parts (i.e. $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$). Different from IBN, the proposed solution first performs IN and BN on $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ respectively, and meanwhile the proposed solution also performs BN and IN on $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ respectively. Then the final normalization result is obtained as follows:

$$\hat{\boldsymbol{x}} = \frac{1}{2}(\text{Concat}(\text{IN}(\boldsymbol{x}_1), \text{BN}(\boldsymbol{x}_2)) + \text{Concat}(\text{BN}(\boldsymbol{x}_1), \text{IN}(\boldsymbol{x}_2))), \quad (14)$$

where "Concat" is short for "concatenate". IN$(\cdot)$ and BN$(\cdot)$ denote the instance normalization operation and the batch

normalization operation, respectively. This design can integrate the advantages of IN and BN in an effective way, and further improve the generalization ability of the proposed model from source domain to target domain. In FDRI sub-module, the camera style network $\boldsymbol{E}_c$ consists of six Conv-BN-LeakyRelu blocks and one fully-connected layer with dimension $C_s + C_t + 1$, where $C_s$ is the number of cameras in source domain, and $C_t$ is the number of cameras in target domain.

*2) Optimization:* In the training process, the total loss function is formalized as follows:

$$\begin{aligned}\boldsymbol{L} = {} & \boldsymbol{L}_{ce1}(\boldsymbol{W}_1, \boldsymbol{E}) + \boldsymbol{L}_{ce2}(\boldsymbol{W}_1, \boldsymbol{E}) + \boldsymbol{L}_{ce5}(\boldsymbol{W}_2, \boldsymbol{E}) + \lambda_1 \boldsymbol{L}_{ag}(\boldsymbol{E}) \\ & + \lambda_2(\boldsymbol{L}_{ce3}(\boldsymbol{E}_c) + \boldsymbol{L}_{ce4}(\boldsymbol{E})) + \lambda_3(L_{d_1}(\boldsymbol{W}_2) + L_{d_2}(\boldsymbol{E})) \\ & + \lambda_4(\boldsymbol{L}_{dis_1}(\boldsymbol{W}_1, \boldsymbol{W}_2') + \boldsymbol{L}_{dis_2}(\boldsymbol{E})), \end{aligned} \tag{15}$$

where $\lambda_1$, $\lambda_2$, $\lambda_3$, and $\lambda_4$ are hyperparameters. During the training process, the identity loss $\boldsymbol{L}_{ce1}(\boldsymbol{W}_1, \boldsymbol{E})$ and identity-domain loss $\boldsymbol{L}_{ce5}(\boldsymbol{W}_2, \boldsymbol{E})$ are used to endow the feature encoder $\boldsymbol{E}$ with the initial ability to extract discriminative and robust features. Furthermore, the losses $\boldsymbol{L}_{ce2}(\boldsymbol{W}_1, \boldsymbol{E})$ and $\boldsymbol{L}_{ag}(\boldsymbol{E})$ are minimized to facilitate IRM to generate the multi-view image features. In DIFE sub-module, the loss functions $\boldsymbol{L}_{ce3}(\boldsymbol{E}_c)$ and $\boldsymbol{L}_{ce4}(\boldsymbol{E})$ are alternately minimized to make the features domain-invariant. At the same time, losses $\boldsymbol{L}_{d_1}(\boldsymbol{W}_2)$ and $\boldsymbol{L}_{d_2}(\boldsymbol{E})$ are performed to further align the data distribution at identity level. In addition, two discrepancy losses $\boldsymbol{L}_{dis_1}(\boldsymbol{W}_1, \boldsymbol{W}_2')$ and $\boldsymbol{L}_{dis_2}(\boldsymbol{E})$ in FDRI sub-module can further improve the discriminability and robustness of the learned domain-invariant features. For testing, the feature cosine similarity [63] between the probe and gallery image pairs are first measured and then the ranking list is obtained based on the measured cosine similarity scores. The optimization procedure of the proposed solution is summarized as Algo. 1.

**Algorithm 1** Triple Adversarial Learning and Multi-View Imaginative Reasoning Network

**Input:** Source-domain images $\boldsymbol{X}_s = \{x_s^i\}_{i=1}^N$, the corresponding identity label space $\boldsymbol{Y}_s = \{y_s^i\}_{i=1}^K$, and camera identity space $\boldsymbol{Y}_s^C = \{y_s^c\}_{c=1}^{C_s}$. Target-domain images $\boldsymbol{X}_t = \{x_t^i\}_{i=1}^{N_t}$ and the corresponding camera label space $\boldsymbol{Y}_t^C = \{y_t^c\}_{c=1}^{C_t}$.

**Output:** The trained encoder $\boldsymbol{E}$.

**Sampling:** Sampling a batch samples from source and target domains.

**Optimization:**

**Step I**: **for** $iter = 1, \cdots, Iteration_1$ **do**
Update $\boldsymbol{E}$, $\boldsymbol{W}_1$, $\boldsymbol{W}_2$ by Eq.(1) and Eq.(9).
**end for**
**Step II**: **for** $iter = 1, \cdots, Iteration_2$ **do**
Fix $\boldsymbol{E}$, update $\boldsymbol{W}_1$, $\boldsymbol{W}_2$ and $\boldsymbol{E}_c$ by Eq.(1), Eq.(5), Eq.(7), Eq.(9), Eq.(10) and Eq.(12).
Fix $\boldsymbol{W}_1$, $\boldsymbol{W}_2$ and $\boldsymbol{E}_c$, update $\boldsymbol{E}$ by Eq.(1), Eq.(5), Eq.(6), Eq.(8), Eq.(9), Eq.(11) and Eq.(13).
**end for**

### *E. Discussion*

In this paper, DIFE sub-module is used to achieve the domain alignment at camera level. However, the domain alignment at camera level does not guarantee that the domain at identity level is also aligned. The domain alignment at identity level (involving identity information) is a finer-gained domain alignment than the domain alignment at camera level. JDAID sub-module can achieve the identity-level alignment. Since the target-domain samples are unlabeled, JDAID can only facilitate the identity-level alignment in target domain, but cannot completely achieve the identity-level alignment of all target-domain samples. Although JDAID can also achieve the global domain alignment across source and target domains, it cannot completely achieve the fine-grained camera-level alignment without the use of camera labels. DIFE can alleviate the deficiency of JDAID in achieving identity-level alignment on target domain due to lacking the labels of target data.

The proposed CN integrates IN and BN in an appropriate manner. According to existing work, IN can reduce the image style variations and improve the generalization ability of the proposed solution on target domain, and BN can improve the learning ability of the proposed solution in image recognition [33], [64]. In the developed method, CN takes advantage of IN and BN. Therefore, CN has the generalization ability as IN, which plays a certain compensation role in alleviating the issue caused by the incomplete alignment of target-domain samples by JDAID and DIFE at identity level. In addition, CN uses the same method as IN to reduce the impact of image style differences. However, IN performs normalization operations on a single image to reduce the image style change, without considering the domain shift caused by the style differences between different cameras. So, like IN, CN cannot completely eliminate the camera style change between different images. In the proposed method, DIFE and JDAID can alleviate the shortcomings of CN in eliminating camera style change. In summary, CN, DIFE, and JDAID play a complementary role in eliminating style change and realizing sample domain alignment.

## IV. Experiments

### *A. Datasets and Evaluation Metrics*

*1) Datasets:* The proposed solution is compared with the state-of-the-art methods to demonstrate its effectiveness. All the experiments are conducted on three large-scale person re-ID datasets Market1501 [65], Duke [66], and MSMT17 [43] (used as both source- and target-domain datasets) to implement six tasks, Market1501 → Duke, Duke → Market1501, MSMT17 → Market1501, MSMT17 → Duke, Duke → MSMT17, and Market1501 → MSMT17. Due to the relatively small amount of data, GRID [67] and PRID [68] are only used as the target-domain datasets. Comparative experiments are also conducted on Market1501→PRID, Market1501→GRID, Duke→PRID and Duke→GRID to illustrate the superiority of the proposed method over existing methods. In these settings,

TABLE I
SETTINGS OF DIFFERENT PERSON RE-ID DATASETS IN PERFORMANCE COMPARISON. S1: SETTING 1; S2: SETTING 2; #ID: NUMBER OF IDENTITIES; CAMS: NUMBER OF CAMERAS; #IMG: NUMBER OF IMAGES

| Datasets | #ID | Training | | Gallery(Test) | | Query(Test) | | Cams |
|---|---|---|---|---|---|---|---|---|
| | | #ID | #Img | #ID | #Img | #ID | #Img | |
| Market1501 | 1,501 | 751 | 12,936 | 750 | 19,732 | 750 | 3,368 | 6 |
| Duke | 1,812 | 702 | 16,522 | 1,110 | 17,661 | 702 | 2,228 | 8 |
| MSMT17 | 4,101 | 1,041 | 32,621 | 3,060 | 93,820 | 3,060 | 11,659 | 15 |
| PRID(S1) | 749 | 100 | 200 | 649 | 649 | 100 | 100 | 2 |
| GRID(S1) | 1,025 | 125 | 250 | 900 | 900 | 125 | 125 | 2 |
| PRID(S2) | 749 | 400 | 500 | 349 | 349 | 100 | 100 | 2 |
| GRID(S2) | 1,025 | 525 | 650 | 500 | 500 | 125 | 125 | 2 |

A→B means dataset A is used as the labeled source domain and dataset B is used as the unlabeled target domain.

**Market1501** contains 32,668 labeled pedestrian images of 1,501 identities captured from six camera views, in which 12,936 images of 751 identities are used for training and the remaining ones are used for testing.

**Duke** contains 36,411 images of 1,404 pedestrians captured by eight cameras without any interference images. All images in Duke were collected in winter. According to the standard protocol used in [66], 16,522 images of 702 pedestrians (702 pedestrians out of 1,404 pedestrians) and 19,889 images of 1,110 pedestrians (the remaining 702 pedestrians out of 1,404 pedestrians + additional 408 pedestrians) are used for training and testing, respectively. The 408 pedestrian images that were only collected from one camera view are used as interference.

**MSMT17** contains 126,441 pedestrian images of 4,101 identities captured by 15 cameras, including 12 outdoor and three indoor cameras. Compared with Market1501 and Duke, MSMT17 is more challenging for pedestrian matching, because the collection of pedestrian images lasted four days, experienced different weather conditions, and involved complex scenes and various brightness. According to the split setting used in [43], 32,621 labeled pedestrian images of 1,041 identities are used for training, and the remaining 93,820 pedestrian images of 3,060 identities are used for testing.

**PRID** contains 934 identities captured from two different camera views (A and B). 385 identities appeared in camera view A and 749 identities appeared in camera view B. Only 200 identities appeared in both camera views A and B. According to the protocol used in [80], [81], 100 pairs of pedestrian images were randomly selected for training. The remaining 100 identities only have one image in each camera review. The images of the 100 identities were selected from camera view A as the probe set. All the images of the remaining 649 identities in camera view B are used as the gallery set.

**GRID** consists of 250 pedestrian image pairs (total 500 images) captured from six disjoint camera views. Each pedestrian only appears in two camera views. There is only one image of each pedestrian in one camera view. According to the settings used in [3], [78], [81], 250 pedestrian images of 125 identities were randomly selected for training, and another 250 images of the remaining 125 identities and 775 interference images from GRID are used for testing. The interference images are isolated pedestrian images. All the pedestrians in the interference images only appear in one camera view.

In comparative experiments, the above settings of PRID and GRID are used to test the impact of individual sample size on the performance of each comparative method when the sample size of the same pedestrian captured from different camera views is relatively small (named as "setting 1"). After that, 300 and 400 interference images are randomly selected from the testing sets of PRID and GRID respectively, and added to the corresponding training sets to test the impact of interference images on the performance of each comparative method (named as "setting 2"). More details about training and testing are presented in Tab. I.

*2) Evaluation Metrics:* In the comparative experiments, both cumulative match characteristic (CMC) and mean average precision (mAP) are used to evaluate the performance, and the single-query evaluation protocol is applied to all the datasets.

### B. Implementation Details

The training was conducted on the pytorch platform using one GTX 2080Ti GPU. In addition, data augmentation was performed by random image flipping and cropping, and all images were resized to $256 \times 128$. Adam optimizer [82] was employed, and the batch size $n_b$ was set to 32, and half of samples in a batch are from source domain and the remaining ones are from target domain. The training epochs were set to 140. The first 120 epochs performed supervised learning on source domain, and the last 20 epochs jointly trained the other modules except baseline. Hyperparameters $\lambda_1$, $\lambda_2$, $\lambda_3$, and $\lambda_4$ were set to 0.015, 0.2, 0.01, and 0.1 in the comparative experiments, respectively. According to the settings used in [62], the initial learning rates of the feature encoder $\boldsymbol{E}$ and classifiers $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ were set to $3 \times 10^{-6}$, which increased linearly from $3 \times 10^{-6}$ to $3 \times 10^{-4}$ after the first 10 epochs, and decayed to $3 \times 10^{-5}$ and $3 \times 10^{-6}$ after the 40th and 70th epoches, respectively. For the camera style network $\boldsymbol{E}_c$, the learning rate was set to $3.5 \times 10^{-5}$. All comparative experiments in this paper adopted the above learning rate settings.

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

The proposed method was compared with the state-of-the-art UDA person re-ID methods to confirm its effectiveness on the tasks of Market1501→Duke and Duke→Market1501. The state-of-the-art UDA person re-ID methods can be classified into three main categories: (i) the methods with pseudo label prediction, including CAMEL [69], PUL [14], PCB-PAST [16], SSG [36], UDAP [70], DECAMEL [34], SHRED+ktCUDA [71], MMT [41], and DG-Net++ [72]; (ii) the methods with style transfer, including CameraStyle [73], SPGAN [47], PTGAN [43], HHL [45], IPGAN [74], SBSGAN [24], ATNet [54], ECN [56], PDA-Net [75], CSGLP [20], and LVRP [76]; (iii) the methods without pseudo label

TABLE II

COMPARISON OF THE PROPOSED METHOD WITH THE STATE-OF-THE-ART UDA METHODS ON MARKET1501 AND DUKE. THE CMC AND MAP RATES (%) OF EACH METHOD ARE LISTED. WHEN TESTING IS PERFORMED ON MARKET1501, DUKE IS USED AS SOURCE DOMAIN, AND VICE VERSE. "–" DENOTES NOT REPORTED. THE BEST RESULTS ARE SHOWN IN BOLD

| Methods | Duke→ Market1501 | | | Market1501→ Duke | | |
|---|---|---|---|---|---|---|
| | Rank-1 | Rank-5 | mAP | Rank-1 | Rank-5 | mAP |
| | Methods with Pseudo Label Prediction | | | | | |
| CAMEL(ICCV'17) [69] | 54.50 | 73.10 | 26.30 | 40.30 | 57.60 | 19.80 |
| PUL(TCCA'18) [14] | 44.70 | 59.10 | 20.10 | 30.40 | 44.50 | 16.40 |
| PCB-PAST(ICCV'19) [16] | 78.38 | – | 54.62 | 72.35 | – | 54.26 |
| SSG(ICCV'19) [36] | 80.00 | 90.00 | 58.30 | 73.00 | 80.60 | 53.40 |
| UDAP(PR'20) [70] | 75.80 | 89.50 | 53.70 | 68.40 | 80.10 | 49.00 |
| DECAMEL(TPAMI'20) [34] | 60.24 | – | 32.44 | – | – | – |
| SHRED+ktCUDA(WACV'20) [71] | 68.60 | – | 49.40 | 58.70 | – | 40.90 |
| MMT(ICLR'20) [41] | 87.70 | 94.90 | 71.20 | 78.00 | 88.88 | 65.10 |
| DG-Net++(ECCV'20) [72] | 82.10 | 90.20 | 61.70 | 78.90 | 87.80 | 63.80 |
| | Methods with Style Transfer | | | | | |
| CameraStyle(CVPR'18) [73] | 58.50 | 78.20 | 27.40 | 48.40 | 62.50 | 25.10 |
| SPGAN(CVPR'18) [47] | 57.70 | 75.80 | 26.70 | 46.40 | 62.30 | 26.20 |
| PTGAN(CVPR'18) [43] | 38.60 | 57.30 | 15.70 | 27.40 | 43.60 | 13.50 |
| HHL(ECCV'18) [45] | 62.20 | 78.80 | 31.40 | 46.90 | 61.00 | 27.20 |
| IPGAN(ICA'19) [74] | 57.20 | 76.00 | 28.00 | 47.00 | 62.80 | 27.00 |
| SBSGAN(ICCV'19) [24] | 58.50 | – | 27.30 | 53.50 | – | 30.80 |
| ATNet(CVPR'19) [54] | 55.70 | 73.20 | 25.60 | 45.10 | 59.50 | 24.90 |
| ECN(CVPR'19) [56] | 75.10 | 87.60 | 43.00 | 63.30 | 75.80 | 40.40 |
| PDA-Net(ICCV'19) [75] | 75.20 | 86.30 | 47.60 | 63.20 | 77.00 | 45.10 |
| CSGLP(TIFS'20) [20] | 61.20 | 77.50 | 31.50 | 47.80 | 62.30 | 27.10 |
| LVRP(TMM'20) [76] | 63.90 | 81.10 | 33.90 | 36.30 | 54.00 | 17.90 |
| | Methods without Pseudo Label Prediction and Style Transfer | | | | | |
| TJ-AIDL(CVPR'19) [46] | 58.20 | 74.80 | 26.50 | 44.30 | 59.60 | 23.00 |
| PAUL(CVPR'19) [23] | 66.70 | – | 36.80 | 56.10 | – | 35.70 |
| CASCL(ICCV'19) [55] | 64.70 | 80.20 | 35.60 | 51.50 | 71.70 | 30.50 |
| UCDA-CCE(ICCV'19) [25] | 64.30 | – | 34.50 | 55.40 | – | 36.70 |
| CFSM(AAAI'19) [77] | 61.20 | – | 28.30 | 49.80 | – | 27.30 |
| ECN(CVPR'19) [56] | 58.00 | 69.90 | 27.70 | 39.70 | 53.00 | 23.60 |
| SSAE(PR'20) [78] | 60.70 | – | 26.60 | 50.20 | – | 28.10 |
| CaNE(WACV'20) [79] | 57.20 | 73.00 | 27.40 | – | – | – |
| DG-Net++(ECCV'20) [72] | 52.20 | 70.70 | 28.60 | 53.20 | 68.70 | 36.30 |
| **Proposed** | **73.08** | **86.34** | **39.95** | **63.53** | **76.62** | **41.34** |

TABLE III

COMPARISON OF THE PROPOSED METHOD WITH THE STATE-OF-THE-ART UDA METHODS, WHEN MSMT17 IS USED AS SOURCE DOMAIN, AND MARKET1501 AND DUKEMTMC-REID ARE USED AS TARGET DOMAIN. THE CMC AND MAP RATES (%) OF EACH METHOD ARE LISTED. "–" DENOTES NOT REPORTED. THE BEST RESULTS ARE SHOWN IN BOLD

| Methods | MSMT17→ Market1501 | | | MSMT17→ Duke | | |
|---|---|---|---|---|---|---|
| | Rank-1 | Rank-5 | mAP | Rank-1 | Rank-5 | mAP |
| CASCL(ICCV'19) [55] | 65.40 | 80.60 | 35.50 | 59.30 | 73.20 | 37.80 |
| MAR(CVPR'19) [39] | 67.70 | 81.90 | 40.00 | 67.10 | 79.80 | 48.00 |
| CaNE(WACV'20) [79] | 59.10 | 75.40 | 30.30 | 60.70 | 74.70 | 39.10 |
| **Proposed** | **74.55** | **87.55** | **42.94** | **68.35** | **80.89** | **48.67** |

prediction and style transfer, including TJ-AIDL [46], PAUL [23], CASCL [55], UCDA-CCE [25], CFSM [77], ECN (without camera style transfer) [56], SSAE [78], CaNE [79], and DG-Net++ [72]. The comparative results on Market1501 and Duke are listed in Tab. II.

As shown in Tab. II, the methods with pseudo label prediction achieved good performance. For example, MMT obtained 87.7% (78.0%) Rank-1 accuracy and 71.2% (65.1%) mAP on Duke→Market1501 (Market1501→Duke), because each sample in target domain participating in training has the corresponding positive samples. The performance of the proposed method is inferior to the latest pseudo-label prediction-based methods. However, due to the presence of interference samples, the performance of pseudo-label prediction-based methods varies in real-world scenes and may be far below the corresponding performance shown in Tab. II. The proposed method does not need any fine-tuning with the predicted pseudo-labels. So, its performance is not restricted to the number of positive sample pairs, which means the proposed method has good practicability. Unlike the style transfer

TABLE IV

COMPARISON OF THE PROPOSED METHOD WITH THE STATE-OF-THE-ART UDA METHODS ON THE TASKS OF MARKET1501→MSMT17 AND DUKEMTMC-REID→MSMT17. THE CMC AND MAP RATES (%) OF EACH METHOD ARE LISTED. "–" DENOTES NOT REPORTED. THE BEST RESULTS ARE SHOWN IN BOLD

| Methods | Market1501→MSMT17 | | | Duke→MSMT17 | | |
|---|---|---|---|---|---|---|
| | Rank-1 | Rank-5 | mAP | Rank-1 | Rank-5 | mAP |
| PTGAN(CVPR'18) [43] | 10.20 | – | 2.90 | 11.80 | – | 3.30 |
| ECN(CVPR'19) [56] | 25.30 | 36.30 | 8.50 | 30.20 | 41.50 | 10.00 |
| **Proposed** | **30.87** | **43.53** | **11.24** | **39.04** | **51.51** | **14.22** |

learning-based methods, such as SPGAN, PTGAN, HHL, IPGAN, SBSGAN, ATNet, CSGLP, LVRP, CameraStyle, ECN and PDA-Net, the proposed method does not need to transfer image style from source domain to target domain, so its recognition performance does not rely on the quality of transferred images. Furthermore, the image style transfer from the labeled source domain to target domain is extremely time-consuming, which severely reduces the recognition efficiency.

According to Tab. II, the recognition rate of the proposed method is only slightly lower than the corresponding ones of the optimal style transfer-based methods (ECN, PDA-Net) on Duke→Market1501, and greatly outperforms other style transfer methods such as TJ-AIDL, HHL, ATNet, and CSGLP. Except the pseudo label prediction-based methods, the proposed method outperforms all the other methods on Market1501→Duke. In addition, the proposed method is a domain-invariant feature extraction method (without pseudo label prediction and style transfer). Compared with the state-of-the-art domain-invariant feature extraction method PAUL, the proposed method improves the Rank-1 recognition rate from 66.70% to 73.08% and the mAP from 36.80% to 39.75% on Duke→Market1501, respectively. The proposed method also achieves good recognition performance on Market1501→Duke. Compared with the suboptimal method PAUL, the Rank-1 recognition rate and mAP of the proposed method are improved by 7.43% and 5.64%, respectively. So, the comparative results confirm the effectiveness of the proposed method and its superiority over other methods. As shown in Tab. II, if the style transfer is removed, the performance of ECN drops significantly, and its recognition rate is far below the corresponding one obtained by the proposed method.

To evaluate the superiority of the proposed method comprehensively, a challenging dataset MSMT17 is used as source domain, two datasets Market1501 and Duke are used as target domain. The performance of the proposed method is compared with three competitive methods, CaNE, CASCL and MAR [39]. As shown in Tab. III, the proposed method can improve the best Rank-1 recognition accuracy and mAP obtained by the comparative method MAR from 67.70% to 74.55% and from 40.00% to 42.94% on MSMT17→Market1501, respectively. On MSMT17→Duke, the proposed method achieves 68.35% Rank-1 accuracy and 48.67% mAP, which outperform the Rank-1 accuracy and mAP obtained by MAR by 1.25% and 0.67% respectively. The comparative results demonstrate the effectiveness and extendibility of the proposed method.

To further verify the effectiveness and extendibility of the proposed algorithm, comparative experiments are conducted on Duke → MSMT17 and Market1501 → MSMT17. These two tasks are more challenging than the previous tasks. Although the data in Duke and Market1501 is much less than that in MSMT17, the corresponding setting of comparative experiments is closer to the real-world scenes. As shown in Tab. IV, the proposed method is compared with two state-of-the-art methods, PTGAN and ECN. The comparative results indicate that the proposed method significantly outperforms PTGAN and ECN on all three types of recognition accuracy. Specifically, compared with the second best results obtained by ECN, the proposed model improves Rank-1 accuracy (mAP) from 25.30%(8.50%) to 30.87%(11.24%) and from 30.20%(10.00%) to 39.04%(14.22%) on Duke → MSMT17 and Market1501 → MSMT17, respectively. As the main reason, the proposed method can imagine multi-view pedestrian information from a single camera view image. Therefore, more discriminative features can be learnt for UDA person re-ID.

### *D. Further Discussion*

Clustering-based pseudo label prediction has been widely studied in recent years. The pseudo label prediction based person re-ID methods can achieve good performance on public large-scale datasets, such as Market1501 (as shown in Tab. II). Pseudo labels are first predicted on the target-domain dataset, and then supervised model training is applied to the samples with pseudo labels. The promising performance may only be achieved, when each target-domain sample participating in training has the corresponding positive samples. However, this requirement is obviously not consistent with real-world scenes. The large-scale datasets collected from real-world scenes often contain many isolated samples, and the identities in these samples are matched, which brings great challenges to the accurate prediction of pseudo labels. In addition, cluster-based pseudo-label prediction also has certain requirements on the number of samples of the same pedestrian. If the number of samples with the same identity from different camera views is extremely small, it may also bring great challenges to pseudo-label prediction.

Compared with these methods, the proposed method has stronger practicability. To demonstrate the practicability, GRID and PRID containing interference images are used as target dataset, and one of Market1501 and Duke is used as source dataset. The comparative methods include the latest pseudo-label prediction-based methods, UDAP [70],

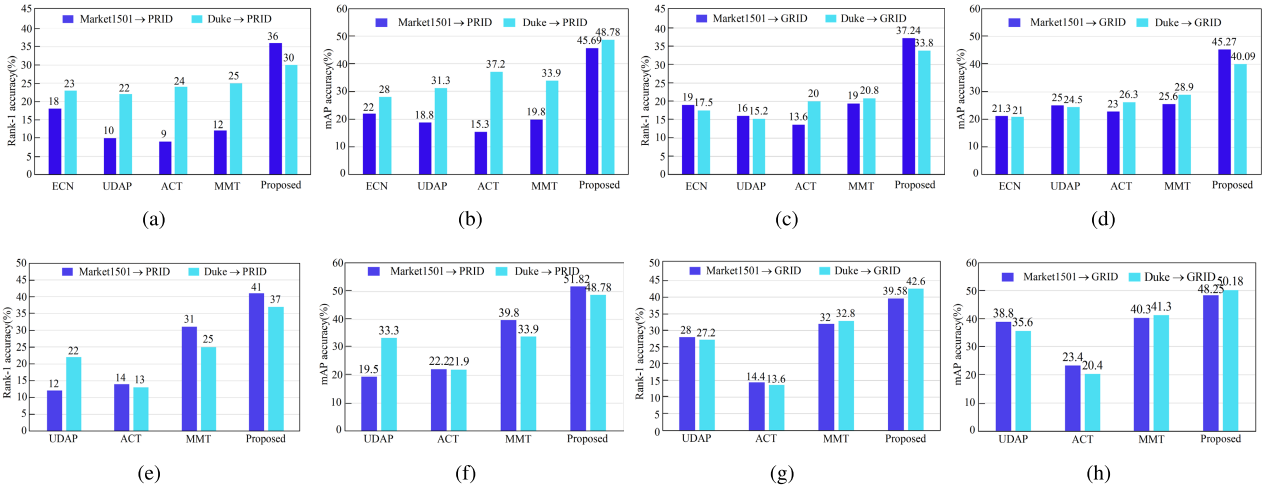


Fig. 6. Comparison of the proposed method with the state-of-the-art unsupervised self-labeling-based and camera style adaptation-based methods on PRID and GRID. The CMC and mAP rates (%) of each method are illustrated. The results shown in (a)–(d) are obtained on setting 1, and the results shown in (e)–(h) are obtained on setting 2.

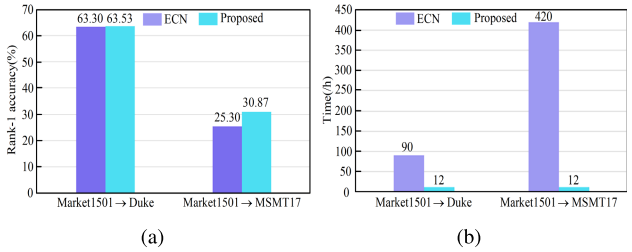


Fig. 7. Comparison between ECN and the proposed method. (a) The CMC rates (%) comparison (b) The time-consuming comparison of model training.

ACT [15], and MMT [41], and the style transfer-based methods, PTGAN [43], ATNet [54], and ECN [56]. In the comparative experiments, "setting 1" and "setting 2" introduced in subsection of Datasets and Evaluation Metrics are used to test the performance of each re-ID method.

As shown in Fig. 6, under Setting 1(Setting 2), UDAP, ACT, and MMT (RestNet50 as the backbone), which have high performance on Market1501 and Duke, only achieve no more than 12%/25% (31%/25%) recognition rate of Rank-1 on Market1501 → PRID/Duke → PRID, while the proposed method can achieve more than 36%/30% (41%/37%) recognition accuracy of Rank-1. On Market1501→GRID/Duke→GRID, the Rank-1 recognition rates of UDAP, ACT, and MMT are less than 20%/21% (32%/33%), but the proposed method obtains more than 37%/33% (39%/42%) Rank-1 recognition accuracy. Since the number of pedestrian images with the same identity in PRID and GRID is scarce, it is challenging for cluster-based pseudo-label prediction. Therefore, the performance of UDAP, ACT, and MMT is poor on PRID and GRID.

In addition, when some interference samples are moved from testing set to training set, the performance of some methods is improved. Compared with these methods, the performance of the proposed method is better. The interference images in testing set bring great challenges to the matching of pedestrian identities. When the number of interference images decreases, the corresponding performance improves accordingly. Interference images considerably affect the correct prediction of the pseudo-labels generated by clustering. However, the proposed method does not need to predict any pseudo-labels by clustering. The interference images in the training set do not have much impact on the performance of the recognition algorithm. So, the proposed method shows better performance than UDAP, ACT and MMT.

Although ECN achieves the promising performance on Duke→Market1501, it can only obtain 18.0% Rank-1 accuracy on PRID and 19.0% Rank-1 accuracy on GRID when Market1501 is used as source domain on setting 1. Since the number of training samples of PRID (GRID) is small, the performance of the style transfer models trained on this dataset is poor, which results in poor image quality after style transfer. Training a re-ID model with these transferred low-quality samples definitely results in the low performance of the trained model. In contrast, the proposed method does not reply on the training sample size of the target-domain dataset, so it is more practical.

As shown in the previous discussion, style transfer based methods have low computational efficiency, which affects their practicability. The proposed method is compared with ECN on Market1501 → Duke and Market1501 → MSMT17 to confirm its practicability. As shown in Fig. 7, the proposed method takes about 12 hours to train the re-ID model. Since ECN needs to transfer image style from source domain to target domain by using CamStyle model [73], it takes about 90 hours to train the re-ID model. The training will last longer when a larger-scale training dataset such as MSMT17 is transferred. Compared with ECN, the proposed method not only takes less time to train the model, but also achieves better recognition performance. Compared with the style transfer-based methods, the proposed method neither requires any pre-trained model, nor needs to transfer image style from source domain to target domain. So, the proposed method is more efficient. In addition, the data in listed Tabs. III∼V

TABLE V

ABLATION STUDY. THE IMPACT OF EACH MODULE ON THE PROPOSED FRAMEWORK. "B" INDICATES THE SUPERVISED LEARNING ON LABELED SOURCE DOMAIN. THE CMC AND MAP RATES (%) ARE LISTED

| Methods | Duke→ Market1501 | | Market1501→ Duke | |
|---|---|---|---|---|
| | Rank-1 | mAP | Rank-1 | mAP |
| Proposed w/B(baseline) | 43.28 | 21.61 | 30.25 | 16.33 |
| Proposed w/B+BN | 61.93 | 31.30 | 42.81 | 24.79 |
| Proposed w/B+IBN | 61.99 | 30.98 | 43.40 | 25.58 |
| Proposed w/B+BIN | 61.57 | 31.10 | 43.65 | 25.73 |
| Proposed w/B+CN | 62.97 | 32.11 | 44.82 | 26.87 |
| Proposed w/B+CN+IRM | 65.29 | 35.14 | 49.64 | 31.58 |
| Proposed w/B+CN+IRM+DIFE | 70.55 | 38.97 | 60.17 | 39.80 |
| Proposed w/B+CN+IRM+DIFE+JDAID | 72.61 | 39.84 | 62.57 | 41.10 |
| Proposed w/B+CN+IRM+DIFE+JDAID+FDRI | **73.08** | **39.95** | **63.53** | **41.34** |

demonstrates that the proposed method is more competitive than the comparative methods in recognition performance.

### *E. Ablation Study*

The proposed method consists of CN, IRM, DIFE, JDAID and FDRI sub-modules. A series of experiments are conducted to demonstrate the effectiveness of each sub-module. The corresponding results are presented in Tab. V. CN, IRM, DIFE, JDAID and FDRI sub-modules are first removed from the proposed method, and then the obtained model is used as the baseline method. After the baseline method is trained on the labeled source domain with the identity loss, it is directly applied to the target dataset. The Rank-1 recognition rates on Market1501 and Duke are only 43.28% and 30.25%, respectively. The performance of the proposed method significantly drops.

*1) Effectiveness of CN:* To illustrate the advantages of CN over BN, IBN and BIN, BN, IBN and BIN are used to replace CN in the proposed method respectively, and the newly obtained methods are compared with the proposed method. As shown in Tab. V, the new model only with BN (IBN, BIN) is named as "Proposed w/B+BN (IBN, BIN)" ("w/B" means with baseline). According to Tab. V, the recognition performance of "Proposed w/B+CN" is higher than the corresponding recognition performance of "Proposed w/B+BN", "Proposed w/B+IBN" and "Proposed w/B+BIN". In addition, compared to the "*baseline*", the mAP and Rank-1 obtained by "Proposed w/B+CN" are improved by 10.5% and 19.69% on Duke→ Market1501, respectively. Similarly, the mAP and Rank-1 obtained by "Proposed w/B+CN" increases from 16.33% to 26.87% and from 30.25% to 44.82%, respectively. The above results confirm that CN can enable the proposed method to have strong generalization ability across different domains.

*2) Effectiveness of IRM:* As shown in Tab. V, after IRM is added to "Proposed w/B+CN", Rank-1 recognition performance on Market1501 and Duke is improved by 2.32% and 4.82%, respectively. Meanwhile, mAP is also improved from 32.11% to 35.14% on Duke→Market1501, and from 26.87% to 31.58% on Market1501→Duke. "Proposed w/B+CN" is only trained with $\boldsymbol{L}_{ce1}(\boldsymbol{W}_1, \boldsymbol{E})$. Except $\boldsymbol{L}_{ce1}(\boldsymbol{W}_1, \boldsymbol{E})$, $\boldsymbol{L}_{ce2}(\boldsymbol{W}_1, \boldsymbol{E})$ and $\boldsymbol{L}_{ag}(\boldsymbol{E})$ are added to the proposed model. So, the proposed method can extract more discriminative multi-view features from a single-view image. The above results confirm $\boldsymbol{L}_{ce2}(\boldsymbol{W}_1, \boldsymbol{E})$ and $\boldsymbol{L}_{ag}(\boldsymbol{E})$ can play a positive role in extracting discriminative features.

*3) Effectiveness of DIFE:* As shown in Tab. V, "Proposed w/B+CN+IRM+DIFE" significantly improves "Proposed w/B+CN+IRM", after DIFE is added to "Proposed w/B+CN+IRM". Specifically, "Proposed w/B+CN+IRM+DIFE" outperforms the Rank-1 accuracy/mAP of "Proposed w/B+CN+IRM" by 5.26%/3.83% on Duke → Market1501 and 10.53 %/8.22% on Market1501 → Duke, respectively. Since DIFE and IRM have a certain complementary effect, they can improve the discriminability of features captured from different perspectives. IRM can imagine multi-view pedestrian features according to a single camera view image, while DIFE can suppress the domain bias between source and target datasets. So, DIFE and IRM can significantly improve the recognition performance of the proposed model together.

*4) Effectiveness of JDAID:* As shown in Tab. V, JDAID sub-module plays an energetic role in improving the performance of "Proposed w/B+CN+IRM+DIFE". Specifically, JDAID sub-module improves mAP and Rank-1 recognition performance on Market1501 by 0.87% and 2.06%. Meanwhile, both mAP and Rank-1 recognition performance on Duke is also improved from 39.80% to 41.10% and from 60.17% to 62.57%, respectively. The main reason is that JDAID sub-module can align the distribution at identity level. So, the domain gap at identity level is narrowed.

*5) Effectiveness of FDRI:* As shown in Tab. V, "Proposed w/B+CN+IRM+DIFE+JDAID+FDRI" further improves the recognition performance when FDRI sub-module is added to "Proposed w/B+CN+IRM+DIFE+JDAID". So, FDRI sub-module can effectively improve the discriminability and robustness of the learned features. In fact, using two different classifiers to identify the identity of the same image is equivalent to identifying the identity of the images from different perspectives, which is conducive to extracting the features associated with identity. The comparative results shown in Tab. V confirm the effectiveness of triple adversarial learning in improving feature robustness.

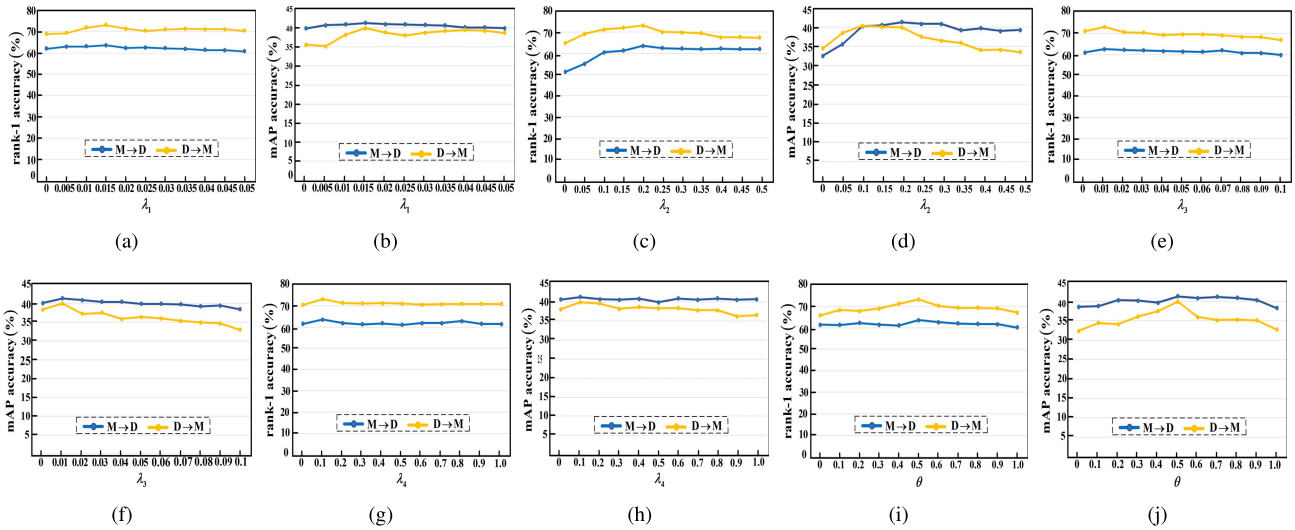


Fig. 8. The effect analysis on hyperparameters $\lambda_1$, $\lambda_2$, $\lambda_3$, $\lambda_4$, and $\theta$. When a hyperparameter is evaluated, the other hyperparameters are fixed at the optimal values. (a) The effect of $\lambda_1$ on Rank-1 (b) The effect of $\lambda_1$ on mAP, (c) The effect of $\lambda_2$ on Rank-1, (d) The effect of $\lambda_2$ on mAP, (e) The effect of $\lambda_3$ on Rank-1, (f) The effect of $\lambda_3$ on mAP, (g) The effect of $\lambda_4$ on Rank-1, (h) The effect of $\lambda_4$ on mAP, (i) The effect of $\theta$ on Rank-1, and (j) The effect of $\theta$ on mAP. M and D denote the datasets Market1501 and Duke, respectively.

### *F. Parameter Analysis*

The proposed model contains four hyperparameters $\lambda_1$, $\lambda_2$, $\lambda_3$, and $\lambda_4$, which are used to control the relative importance of loss items $\boldsymbol{L}_{ag}(\boldsymbol{E})$, $(\boldsymbol{L}_{ce3}(\boldsymbol{E}_c), \boldsymbol{L}_{ce4}(\boldsymbol{E}))$, $(\boldsymbol{L}_{d_1}(\boldsymbol{W}_2), \boldsymbol{L}_{d_2}(\boldsymbol{E}))$, and $(\boldsymbol{L}_{dis_1}(\boldsymbol{W}_1, \boldsymbol{W}_2), \boldsymbol{L}_{dis_2}(\boldsymbol{E}))$, respectively. In addition, a hyperparameter $\theta$ is set to balance the proportion of identity and domain information in JDAID sub-module. Fig. 8 shows the analysis of how each hyperparameter affects the proposed model in the learning process. It should be noted that when a hyper-parameters is analyzed, the other hyper-parameters are fixed. All hyperparameter settings remain the same in all the experiments of this paper.

*1) The Effect of $\lambda_1$:* The hyperparameter $\lambda_1$ is used to control the importance of $\boldsymbol{L}_{ag}(\boldsymbol{E})$. This loss item further guarantees that the features extracted from a single perspective have the related multi-view information on the basis of identity constraints. As shown in Fig. 8 (a) and (b), Rank-1 and mAP accuracy are improved when $\lambda_1$ increases from 0 to 0.015, and the best performance of the proposed model is reached on both tasks Duke→ Market1501 and Market1501→Duke when $\lambda_1 = 0.015$. Moreover, when $\lambda_1 > 0.015$, the performance of the proposed model begins to degrade, which indicates $\lambda_1 = 0.015$ is an optimal value for the proposed model.

*2) The Effect of $\lambda_2$:* Fig. 8 (c) and (d) show the effect of parameter $\lambda_2$ on the performance of the proposed model. When $\lambda_2 \in [0.1, 0.3]$, the performance of the proposed model is relatively stable, and a high recognition rate of Rank-1 and mAP is achieved on both Duke→Market1501 and Market1501→Duke. However, when $\lambda_2 = 0$ or $\lambda_2 > 0.3$, the recognition rate of Rank-1 and mAP on Duke→Market1501 and Market1501→Duke drops drastically, which further confirms the validity of DIFE and the rationality of $\lambda_2 = 0.2$.

*3) The Effect of $\lambda_3$:* The hyperparameter $\lambda_3$ is used to control the effects of $\boldsymbol{L}_{d_1}(\boldsymbol{W}_2)$ and $\boldsymbol{L}_{d_2}(\boldsymbol{E})$. As shown in Fig. 8 (e) and (f), the recognition performance increases when the value of $\lambda_3$ increases from 0 to 0.01. When $\lambda_3 = 0.01$, the recognition rate of Rank-1 and mAP reaches the peak on both Market1501 and Duke datasets, which indicates that $\lambda_3 = 0.01$ is an optimal value.

*4) The Effect of $\lambda_4$:* In order to further improve the discriminability and robustness of the learned features, the FDRI is designed and parameter $\lambda_4$ is used to control its importance. As shown in Fig. 8 (g) and (h), the performance of the proposed method remains relatively stable when $\lambda_4$ increases from 0 to 1. When $\lambda_4 = 0.1$, the performance of the proposed model slightly exceeds the corresponding performance achieved at other values of $\lambda_4$. So, $\lambda_4$ is set to 0.1 in all the experiments.

*5) The Effect of $\theta$:* Different from the above four hyperparameters, $\theta$ represents the probability of being classified into a specific category. Fig. 8 (i) and (j) show the effect of $\theta$, when its value increases from 0 to 1. $\theta = 0$ means that the extracted features only contain domain information, and $\theta = 1$ indicates the features only contains identity information. The effect analysis shown in Figs. 8 (i) and (j) confirms the performance of the proposed method is optimal when $\theta = 0.5$.

## V. Conclusion

In this paper, a novel triple adversarial learning is proposed and a multi-view imaginative reasoning network is constructed for UDA person re-ID. In the proposed method, the developed CN can improve the generalization ability of the re-ID model from one domain to another domain. The proposed IRM can imagine and reason multi-view features from the input single-view image. In addition, the proposed DIFE, JDAID and FDRI sub-modules not only reduce the negative impact of domain shift, but also ensure the discriminability and robustness of the learned features. The results of comparative experiments conducted on widely-used benchmark datasets confirm that the proposed method can significantly outperform other

competitive UDA person re-ID methods. Besides, the ablation study demonstrates that each sub-module in the proposed model is useful. In future, the improvement of the proposed method will be further explored to satisfy the requirements of person re-ID in more complex real-world scenes.

## REFERENCES


[1] Z. Yu *et al.*, "Progressive transfer learning for person re-identification," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, 2019, pp. 4220–4226.

[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, Jun. 2020.

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

[4] D. Tao, Y. Guo, B. Yu, J. Pang, and Z. Yu, "Deep multi-view feature learning for person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 10, pp. 2657–2666, Oct. 2018.

[5] Y. Huang, S. Lian, S. Zhang, H. Hu, D. Chen, and T. Su, "Three-dimension transmissible attention network for person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 12, pp. 4540–4553, Dec. 2020.

[6] X. Liu, S. Bi, S. Fang, and A. Bouridane, "Bayesian inferred self-attentive aggregation for multi-shot person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 10, pp. 3446–3458, Oct. 2020.

[7] C. Shen *et al.*, "Sharp attention network via adaptive sampling for person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 10, pp. 3016–3027, Oct. 2019.

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

[9] L. Wu, R. Hong, Y. Wang, and M. Wang, "Cross-entropy adversarial view adaptation for person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 7, pp. 2081–2092, Jul. 2020.

[10] L. Wu, Y. Wang, H. Yin, M. Wang, L. Shao, and B. Lovell, "Few-shot deep adversarial learning for video-based person re-identification," *IEEE Trans. Image Process.*, vol. 29, pp. 1233–1245, Mar. 2020.

[11] Y. Luo, T. Liu, D. Tao, and C. Xu, "Decomposition-based transfer distance metric learning for image classification," *IEEE Trans. Image Process.*, vol. 23, no. 9, pp. 3789–3801, Sep. 2014.

[12] Y. Luo, Y. Wen, T. Liu, and D. Tao, "Transferring knowledge fragments for learning distance metric from a heterogeneous domain," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 4, pp. 1013–1026, Apr. 2019.

[13] Y. Luo, H. Hu, Y. Wen, and D. Tao, "Transforming device fingerprinting for wireless security via online multitask metric learning," *IEEE Internet Things J.*, vol. 7, no. 1, pp. 208–219, Jan. 2020.

[14] 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, doi: 10.1145/3243316.

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

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

[17] 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.

[18] L. Qi, L. Wang, J. Huo, Y. Shi, and Y. Gao, "Progressive cross-camera soft-label learning for semi-supervised person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 9, pp. 2815–2829, Sep. 2020.

[19] 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. (CVPR)*, Jun. 2018, pp. 5157–5166.

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

[21] 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.

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

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

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

[25] 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. 8080–8089.

[26] G. Qi, G. Hu, X. Wang, N. Mazur, Z. Zhu, and M. Haner, "EXAM: A framework of learning extreme and moderate embeddings for person re-ID," *J. Imag.*, vol. 7, no. 1, p. 6, Jan. 2021.

[27] 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, Nov. 2021.

[28] X. Jin, C. Lan, W. Zeng, and Z. Chen, "Uncertainty-aware multi-shot knowledge distillation for image-based object re-identification," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 11165–11172.

[29] K. Saito, K. Watanabe, Y. Ushiku, and T. Harada, "Maximum classifier discrepancy for unsupervised domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 3723–3732.

[30] 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. 2502–2511.

[31] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4105–4113.

[32] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn. (ICML)*, no. 37, 2015, pp. 448–456.

[33] X. Pan, P. Luo, J. Shi, and X. Tang, "Two at once: Enhancing learning and generalization capacities via IBN-Net," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 464–479.

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

[35] H. Wang, S. Gong, and T. Xiang, "Unsupervised learning of generative topic saliency for person re-identification," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2014, pp. 1–11, doi: 10.5244/C.28.48.

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

[37] Y. Lin, X. Dong, L. Zheng, Y. Yan, and Y. Yang, "A bottom-up clustering approach to unsupervised person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 8738–8745.

[38] M. Li, X. Zhu, and S. Gong, "Unsupervised tracklet person re-identification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 7, pp. 1770–1782, Jul. 2020.

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

[40] 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.

[41] 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)*, 2019, pp. 1–15.

[42] 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. 9021–9030.

[43] 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. (CVPR)*, Jun. 2018, pp. 79–88.

[44] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2223–2232.

[45] 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.

[46] 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. (CVPR)*, Jun. 2018, pp. 2275–2284.

[47] 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. (CVPR)*, Jun. 2018, pp. 994–1003.

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

[49] S. Li, M. Shao, and Y. Fu, "Person re-identification by cross-view multi-level dictionary learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 12, pp. 2963–2977, Oct. 2018.

[50] P. Peng, Y. Tian, T. Xiang, Y. Wang, M. Pontil, and T. Huang, "Joint semantic and latent attribute modelling for cross-class transfer learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 7, pp. 1625–1638, Jul. 2018.

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

[52] L. Qi, J. Huo, X. Fan, Y. Shi, and Y. Gao, "Unsupervised joint subspace and dictionary learning for enhanced cross-domain person re-identification," *IEEE J. Sel. Topics Signal Process.*, vol. 12, no. 6, pp. 130–146, Oct. 2018.

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

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

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

[56] 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.

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

[58] 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 (CVPR)*, Jun. 2009, pp. 79–88.

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

[60] H. Tang and K. Jia, "Discriminative adversarial domain adaptation," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2020, pp. 5940–5947.

[61] L. Tran, K. Sohn, X. Yu, X. Liu, and M. Chandraker, "Gotta adapt'em all: Joint pixel and feature-level domain adaptation for recognition in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2672–2681.

[62] 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, p. 1.

[63] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Deep metric learning for person re-identification," in *Proc. 22nd Int. Conf. Pattern Recognit.*, Aug. 2014, pp. 34–39.

[64] X. Jin, C. Lan, W. Zeng, Z. Chen, and L. Zhang, "Style normalization and restitution for generalizable person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3425–3435.

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

[66] 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)*, 2016, pp. 17–35.

[67] C. C. Loy, C. Liu, and S. Gong, "Person re-identification by manifold ranking," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2013, pp. 3567–3571.

[68] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof, "Person re-identification by descriptive and discriminative classification," in *Proc. Scand. Conf. Image Anal.*, 2011, pp. 91–102.

[69] H. Yu, A. Wu, and W. Zheng, "Cross-view asymmetric metric learning for unsupervised person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern (CVPR)*, Oct. 2017, pp. 994–1002.

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

[71] D. Kumar, P. Siva, P. Marchwica, and A. Wong, "Unsupervised domain adaptation in person re-ID via k-reciprocal clustering and large-scale heterogeneous environment synthesis," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 2645–2654.

[72] Y. Zou, X. Yang, Z. Yu, B. V. K. 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.

[73] 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.

[74] J. Liu *et al.*, "Identity preserving generative adversarial network for cross-domain person re-identification," *IEEE Access*, vol. 7, pp. 114021–114032, 2019.

[75] 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. 7919–7929.

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

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

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

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

[80] L. Zhang, T. Xiang, and S. Gong, "Learning a discriminative null space for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1239–1248.

[81] 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.

[82] D. P. Kingma and J. Lei Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–15.

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

**Neng Dong** received the M.S. degree in pattern recognition and intelligent system from Kunming University of Science and Technology, Yunnan, China, in 2021. He is currently pursuing the Ph.D. degree in computer science and technology with Nanjing University of Science and Technology. His research interests include machine learning and computer vision.

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

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

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