# Improving the Style Adaptation for Unsupervised Cross-Domain Person Re-identification

Wenyuan Zhang, Li Zhu*, Lu Lu
*School of Software Engineering*
*Xi'an Jiaotong University*
Xi'an, China
zhangwy919@stu.xjtu.edu.cn, zhuli@xjtu.edu.cn, woshilulu1996@stu.xjtu.edu.cn

***Abstract*—Most existing person re-identification (Re-ID) methods are based on supervised learning, in which a large amount of labeled data are required for training. However, it remains a challenge task for adapting a model trained in a labeled source domain to an unlabeled target domain, due to the domain gap. To alleviate this problem, we design an unsupervised person style transfer adaptation pipeline for the task of unsupervised domain adaptation (UDA) Re-ID. Following the pipeline, we first apply an image translator to generate style-transferred images. To preserve the ID-related information after translation, we introduce the intra-class similarity and inter-domain diversity, which are crucial properties for Re-ID. In this way, a Cross-domain Similarity Generative Adversarial Network (CSGAN) is proposed to bridge the domain gap. CSGAN is learned by jointly optimizing an image translator and a domain-invariant feature representation network (DIFRN), which constrains the CSGAN to maintain the intra-class similarity and inter-domain diversity during image-image translation. Comparison with current competitive methods demonstrates that the effectiveness of the proposed method under the setting of unsupervised domain adaptation.**

***Index Terms*—person re-identification, generative adversarial network, unsupervised domain adaptation**

## I. Introduction

Person re-identification (Re-ID) is a cross-camera image retrieval task, which aims to find a target person across non-overlapping camera views by using a probe pedestrian image. In recent years, with the widespread adoption of convolutional neural networks (CNN), many person Re-ID works focus on supervised learning and lead to impressive achievements [1]–[4].

In spite of the satisfactory improvements, there still remains several issues hindering the applications of Re-ID. First, the supervised learning Re-ID methods require abundant manually labeled images, which are prohibitively expensive and sometimes impossible to collect in the real-world scenarios. This scalability limitation severely reduces the applicability of existing supervised Re-ID methods. In addition, another challenge we observed is that, there is a significant performance drop when the models are directly applied to unseen domains, *i.e.*, training and testing on different datasets. The reason of performance drop is that there exists domain gap between different datasets, since the images from different datasets are captured by different cameras containing significant variant scenes, *e.g.,* illumination, viewpoints, backgrounds, human poses, and so on. This issue indicates the poor domain generalizability of supervised learning models.

In this paper, we consider the problem of unsupervised domain adaptation (UDA) in the cross-domain scenario. Unsupervised domain adaptation means that learning a model for the target domain when provided with a fully annotated source domain and an unlabeled target domain. Nevertheless, in generic UDA, most existing methods assume that the source and target domains are in the same label space [5], [6]. Therefore, they have limitations to be applied to Re-ID, where the classes (person identities) from the source and target domains are entirely different. Recently, several image-level domain translation Re-ID works [7], [8] have been proposed based on Generative Adversarial Network (GAN). These approaches transfer the labeled images from source domain to the target domain, so that the translated images and target domain images share similar styles. Then the style-transferred images and their associated labels are used for supervised learning in the target domain. However, these current methods either require auxiliary segmentation annotations or lack effective person feature representations.

To address the problems above, we introduce a person style transfer adaptation pipeline to help improve the performance of cross-domain Re-ID. Our method is motivated from three aspects. First, under the setting of UDA, our model should translate the labeled images from source domain to target domain without any annotation. Second, during the style transfer, two crucial properties for Re-ID ought to be preserved, which are the intra-class similarity and inter-domain diversity. The intra-class similarity constrains the underlying identity information of pedestrian foreground to be remained during the translation, while the inter-domain diversity renders the ID of translated images to be different from any of images in the target dataset, due to the prior knowledge that the source dataset and target dataset have totally different classes (person identities). Third, to obtain discriminative feature embeddings, a latent feature space for better representing the pedestrian images is learned by combining global and local information, so that the visual cues associated with the identity of pedestrian images could be preserved during the image translation.

Based on the motivations described above, a Cross-domain

*Corresponding author.

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

Similarity Generative Adversarial Network (CSGAN) is proposed to reduce the domain gap. We utilize CSGAN to conduct the image-image translation in an unsupervised manner. To better maintain the two properties, a cross-domain triplet loss is introduced to optimize CSGAN by maximizing the diversity across identities in different domains and increasing the similarity within each identity. In addition, to possess a effective and discriminative feature representation, we design a domain-invariant feature representation network (DIFRN), which is a multi-branch network architecture including one global and two local branches. With the proposed DIFRN, detailed information is captured and represented, which is quite helpful for identifying person images with slight difference. To train the CSGAN and DIFRN simultaneously, a joint optimization process is presented. By integrating the motivations above, our proposed method improves the style transfer procedure and perfects the quality of translated person images.

Our main contributions are summarized as bellow:

- Based on the person style transfer adaptation pipeline, we propose a novel CSGAN to learning mappings between different datasets for domain adaptation Re-ID.
- We propose a DIFRN to capture both global and local information of pedestrian images, which is a discriminative feature representation network. We combine an image translator with the DIFRN to jointly optimize the CSGAN. In this way, CSGAN maintains the intra-class similarity and inter-domain diversity during the image-image translation, which benefits the quality of generated images.
- The experimental results demonstrate the effectiveness of our proposed CSGAN. By using our proposal, the performance of unsupervised domain adaptation Re-ID obtains the remarkable improvements.

## II. Related Work

### A. Person Re-Identification

**Supervised learning for person re-identification.** Recent existing person Re-ID methods are dominated by supervised learning. They are mainly based on feature learning or metric learning. In terms of feature learning based methods, they focus on extracting features to describe the query images and the gallery images [3], [4], [9], [10]. Li *et al.* [1] use STN to localize body parts and extract local features. Sun *et al.* [4] propose a PCB and refined part pooling, which is a strong part baseline. As for the metric learning methods, their goal is to find a similarity metric for comparing features [2], [11], [12]. Hermans *et al.* [2] use triplet loss and effectively improve the performance of Re-ID. Although above supervised methods achieve significant progress, they result in performance drop in realistic person Re-ID deployment where no such a large labeled training set is available.

**Unsupervised learning for person re-identification.** To alleviate the above limitations, some approaches [13]–[17] are proposed based on unsupervised learning. These works are divided into three categories: designing hand-craft features [18]–[21], exploiting cross-view information to extract discriminative features [13]–[15] or refining the Re-ID model by unsupervised clustering unlabeled images into different classes [16], [22]. Fan *et al.* [16] propose a progressive unsupervised learning method, which obtains the pseudo labels for unlabeled images by pedestrian clustering, instance selection and fine-tuning model successively. However, since the absence of specific identity labels, these unsupervised learning approaches still have a few limitations and cannot achieve comparable performance as supervised-based methods.

### B. Unsupervised Domain Adaptation for Person Re-Identification

Due to the poor performance of unsupervised learning on single dataset, many domain adaptation Re-ID algorithms are developed to overcome previous drawbacks. These approaches leverage the models trained in the labeled source domain and adapt them to the unlabeled target domain [7], [8], [13], [23]–[27]. Recent UDA based Re-ID models mainly utilize domain alignment [13], [24], [25], [27] or image-synthesis [7], [8], [26] to reduce the domain gap between different datasets. Domain alignment methods require other auxiliary annotation as assistants to improve the generalization of models. TJ-AIDL [13] simultaneously learns an attribute-semantic and identity-discriminative feature representation space which can be adapted to any new target domain. EANet [25] relies on pose estimation and part segmentation to enhance alignment and improve model generalization. Therefore, these approaches suffer from the requirement of collecting auxiliary annotation and have limited applicability in real-world deployments.

Another research direction of unsupervised domain adaptation Re-ID is applying GAN [28] as a way of image generation and data augmentation. Zheng *et al.* [29] first introduce GAN to Re-ID and generate pedestrian images by using DCGAN [30]. Then PTGAN [7] and SPGAN [8] are proposed, both of which apply CycleGAN [31] to image-to-image translation and style transfer for Re-ID. However, PTGAN introduces a segmentation net to extract the mask on person images, which relies on additional annotation. Although SPGAN does not use auxiliary information, our proposed model differs from the SPGAN in both network architecture and loss function.

We aim to maintain the intra-class similarity and the inter-domain diversity by learning discriminative feature representations with multiple granularities. Thus stable style-transferred person images are generated to compensate the gap between different domains, which finally improves the performance of Re-ID under the setting of UDA.

## III. Proposed Method

In this section, we first illustrate the pipeline that we proposed, and then describe its major components in detail. Finally, we introduce the overall loss function and optimization procedure of the proposed approach.

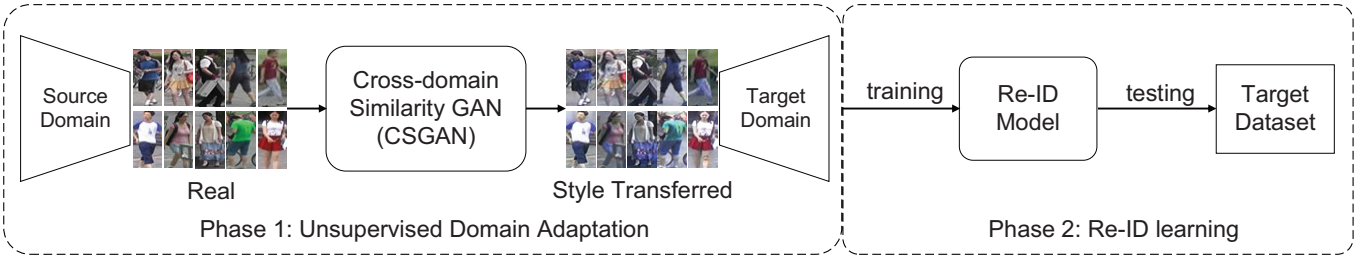


Fig. 1. The person style transfer adaptation pipeline consists of two phases: transferring image style and learning Re-ID model. In first phase, the labeled images from source domain are transferred into target domain style. Then in the second phase, we train a Re-ID model with the transferred images and test on the target dataset. Note that, in first phase we train CSGAN without annotation information.

### A. Overview

The person style transfer adaptation framework is shown in Fig.1, which consists of two phases: transferring image style and learning Re-ID model. In first phase, the labeled images from source domain are transferred into target domain style. Then in the second phase, we train a Re-ID model with the style transferred images and test on the target dataset. Note that, in first phase we train CSGAN without annotation information, *i.e.*, unsupervised learning.

The proposed CSGAN includes two modules: 1) an image translator for style transfer, which learns mapping functions between two domains, and 2) a domain-invariant feature representation network (DIFRN), which is designed to map all images, including real images and generated images into a latent space. The feature embeddings produced by DIFRN is used to compute the cross-domain triplet loss, which constrains the learning procedure of image translator to preserve the intra-class similarity and the inter-domain diversity. By jointly optimizing the image translator and DIFRN, the CSGAN is able to generate high-quality cross-domain person images. Then we employ CSGAN to translate images from source domain to target domain. These images are used for learning Re-ID model in the second phase of the framework.

### B. The Image Translator for Style Transfer

Given two datasets: a labeled dataset $\{(x_a^i, y_a^i)\}_{i=1}^m$ from source domain $\mathcal{A}$ and an unlabeled dataset $\{x_b^j\}_{j=1}^n$ from target domain $\mathcal{B}$. We use CycleGAN [31] as the image translator, which learns mapping functions between two domains. It contains two generators $G : \mathcal{A} \rightarrow \mathcal{B}$ and $F : \mathcal{B} \rightarrow \mathcal{A}$, where they map sample images from one domain to the other. Meanwhile, CycleGAN introduces two adversarial discriminators $D_{\mathcal{A}}$ and $D_{\mathcal{B}}$ to distinguish real images and fake (style-transferred) images.

The loss function of CycleGAN can be formulated as,

$$\begin{aligned}\mathcal{L}_{cyc}(G, F, D_{\mathcal{A}}, D_{\mathcal{B}}) = &\mathcal{L}_{GAN}(G, D_{\mathcal{B}}, \mathcal{A}, \mathcal{B}) \\ &+\mathcal{L}_{GAN}(F, D_{\mathcal{A}}, \mathcal{B}, \mathcal{A}) \\ &+\lambda\mathcal{L}_{rec}(G, F),\end{aligned} \tag{1}$$

where $\mathcal{L}_{GAN}(G, D_{\mathcal{B}}, \mathcal{A}, \mathcal{B})$ and $\mathcal{L}_{GAN}(F, D_{\mathcal{A}}, \mathcal{B}, \mathcal{A})$ are the adversarial loss functions for the generators $G, F$ and the discriminators $D_{\mathcal{B}}, D_{\mathcal{A}}$. $\mathcal{L}_{rec}(G, F)$ is the cycle-consistent reconstruction loss function, which forces each image can be reconstructed after a cycle mapping, and $\lambda$ controls the relative importance between the adversarial loss and cycle-consistent reconstruction loss. More details about CycleGAN can be accessed in [31].

In addition to the adversarial loss and cycle-consistent loss, we use the identity constraint loss [32] to preserve the color of person images for image generation. The identity constraint loss is defined as,

$$\begin{aligned}\mathcal{L}_{idc}(G, F) = &\mathbb{E}_{x_a \sim p_{\mathcal{A}}} \|F(x_a) - x_a\|_1 \\ &+\mathbb{E}_{x_b \sim p_{\mathcal{B}}} \|G(x_b) - x_b\|_1 ,\end{aligned} \tag{2}$$

Finally, the objective function of style transfer is given as follow,

$$\mathcal{L}_{style} = \mathcal{L}_{cyc} + \alpha\mathcal{L}_{idc}, \tag{3}$$

where $\alpha$ is hyper-parameter to control the importance of identity constraint loss function. In all our experiments, we empirically set $\lambda = 10$ in Eq. (1) and $\alpha = 5$ in Eq. (3).

### C. Domain-Invariant Feature Representation Network

We aim to further improve the ability to maintain the intra-class similarity and inter-domain diversity so that the person identity information can be preserved in the image translation stage. To this end, a domain-invariant feature representation network (DIFRN) is proposed, and we train the CycleGAN and DIFRN in a joint manner. We integrate ID-related information with various granularities, in order that the DIFRN can extract both global and local feature representations from person images. The architecture of DIFRN is illustrated in Fig. 2.

In view of the significantly heavy memory consumption and computational costs of CycleGAN, we employ a lightweight and efficient network architecture for DIFRN. In our work, the MobileNetV2 [33] is utilized as the backbone of DIFRN. The MobileNetV2 is a lightweight CNN with competitive performance compared to commonly used architectures such as ResNet-50 [34].

As shown in Fig. 2, we fine-tune the structure of MobileNetV2 by dividing the subsequent part after inverted residual blocks into three branches. We also remove the last classifier of MobileNetV2, considering that our purpose is learning discriminative feature representations, in stead of classification.

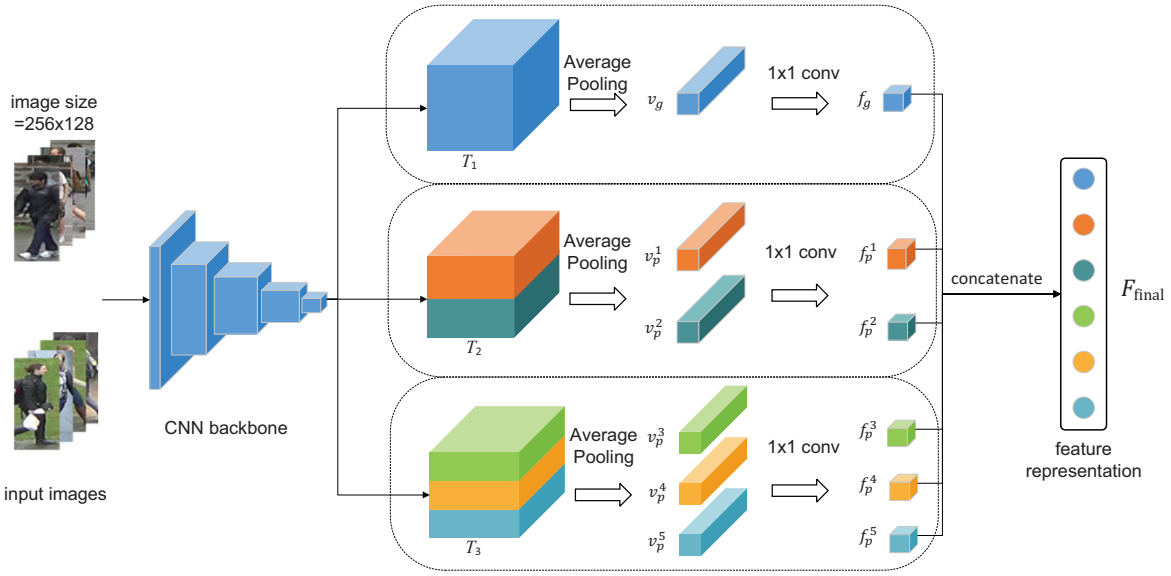


Fig. 2. The structure of the proposed Domain-Invariant Feature Representation Network (DIFRN). The DIFRN consists of one branch for global feature representations and two branches for local feature representations. For the global branch, we use the same setting as the MobileNetV2, except that after global average pooling (GAP) we employ a $1 \times 1$ convolution layer to reduce the dimension of column vector $v_g$ from 1280-dim to 256-dim. For the local branches, called part-2 branch and part-3 branch, to extract features with multiple granularities, we divide feature maps $T_2$ and $T_3$ into 2 and 3 stripes in horizontal orientation, respectively. Then GAP is applied on each part. Finally all the dimension-reduced features are concatenated together as the final feature descriptor of the input images.

The DIFRN consists of one branch for global feature representations and two branches for local feature representations. In the upper branch, called global branch, we use the same setting as original MobileNetV2 [33], except that after global average pooling (GAP) we employ a $1 \times 1$ convolution layer to reduce the dimension of column vector $v_g$ from 1280-dim to 256-dim.

As shown in Fig. 2, for the middle branch, we uniformly split the output feature map of backbone, $T_2$, into 2 stripes horizontally. For the third branch, the feature map $T_3$ is divided into 3 strides in horizontal orientation. Thus, these two branches are named as part-2 branch and part-3 branch, respectively. Then we conduct GAP on each stripe to learn local feature representations. Similar to the operation in global branch, a following $1 \times 1$ convolution layer is implemented to obtain the 256-dim local feature embeddings $\{f_p^i\}_{i=1}^5$.

Finally, to learn the discriminative feature representations with different granularities, all the dimension-reduced features are concatenated together as the final feature descriptor of the input images. Combining both global and local information can improve the comprehensiveness of learned feature representations, and finally benefit the process of style transfer by our proposed cross-domain triplet loss.

### *D. Cross-domain Triplet Loss Function*

In Section I, two properties: intra-class similarity and inter-domain diversity are described. Here we introduce the cross-domain triplet loss to maintain the two crucial properties during the image-image translation.

As shown in Fig. 3, we train DIFRN with the cross-domain triplet loss in a triplet manner. The triplet of images is denoted as $T =< I^a, I^p, I^n >$, where $I^a$ denotes the anchor image, $I^p$ indicates the image of the same person, $I^n$ means the negative sample. Given a source dataset image $x_a$, a target dataset image $x_b$ and generators $G : \mathcal{A} \to \mathcal{B}$ and $F : \mathcal{B} \to \mathcal{A}$, the style-transferred images are expressed as $G(x_a), F(x_b)$. DIFRN maps images $x_a, x_b, G(x_a), F(x_b)$ into a latent space. We denote the corresponding feature embeddings extracted by DIFRN as $\phi(x_a), \phi(x_b), \phi(G(x_a))$ and $\phi(F(x_b))$. The intra-class similarity represents the distance of images before and after translation, written as,

$$d_{intra} = D(\phi(x_a), \phi(G(x_a))) + D(\phi(x_b), \phi(F(x_b))), \quad (4)$$

where $D$ is the distance function. The inter-domain diversity indicates the distance between images from different domains. Based on the prior that person images from different Re-ID datasets are of different identities, the inter-domain diversity can be formulated as,

$$d_{inter} = D(\phi(x_a), \phi(F(x_b))) + D(\phi(x_b), \phi(G(x_a))). \quad (5)$$

**Triplet selection.** Due to the motivation of unsupervised domain adaptation, we train the CSGAN in an unsupervised manner. To select triplet without using any annotation information, the positive pair is generated by image translation, considering the label of the translated image should be the same as its corresponding source image, even if style changes. For the negative pair, since the source and target datasets have completely different classes, the negative pair consists of a source

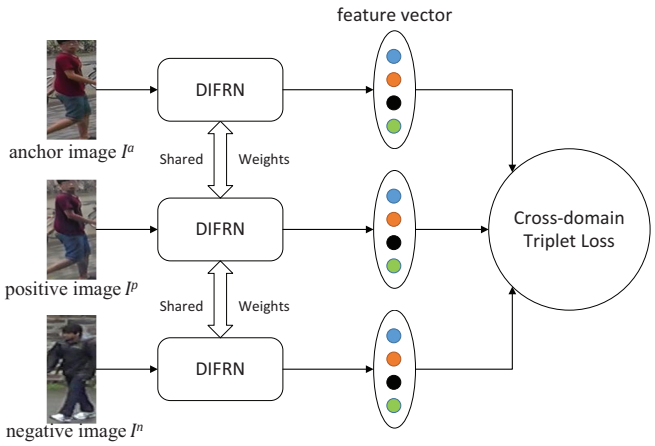


Fig. 3. The illustration of cross-domain triplet loss. The input images includes an anchor image $I^a$, a positive image $I^p$ and a negative image $I^n$.

image and a target image. Therefore, for source domain $\mathcal{A}$, the triplet is denoted as $T_{\mathcal{A}} =< x_a, G(x_a), F(x_b) >$. Similarly, we have the target domain triplet $T_{\mathcal{B}} =< x_b, F(x_b), G(x_a) >$.

The goal of cross-domain triplet loss is to improve the identity discriminability by pulling images of the same identity closer and pushing images from different identities away. The cross-domain triplet loss $L_{cdtl}$ is formulated as,

$$\mathcal{L}_{cdtl} = \frac{1}{n}\sum[m + d_{intra} - d_{inter}]_+, \tag{6}$$

where $m$ denote a predefined margin, $n$ is the number of image triplets in a training batch. $L_{cdtl}$ constrains the distance between positive image pairs to be less than negative pairs by a predefined margin. In our implementation, we use the squared Euclidean distance as distance metric function $D(\cdot)$.

Under the constraint of $L_{cdtl}$, we pull the anchor image of a specific identity from source dataset and its corresponding translated image closer, and push the anchor image and negative images from target dataset away. On one hand, $L_{cdtl}$ renders the ID-related information to remain after image translation. On the other hand, it ensures that the translated person images are dissimilar to the person of target dataset.

**Overall objective function.** Finally, we optimize the CSGAN by the combination of CycleGAN and DIFRN to better promoting the domain generalization. The overall loss function of CSGAN is shown as,

$$\mathcal{L}_{cs} = \mathcal{L}_{style}(G, F, D_A, D_B) + \beta\mathcal{L}_{cdtl}(G, F, N), \tag{7}$$

where $\beta$ is the loss weight to trade off the influence between the style transfer loss and the cross-domain triplet loss, and $N$ represents the proposed domain-invariant feature representation network. The optimization process of CSGAN can be written as,

$$G^*, F^*, N^* = \arg\min_{G,F,N}\max_{D_A,D_B}\mathcal{L}_{cs}(G, F, D_A, D_B, N). \tag{8}$$

### E. Re-ID Feature Learning Model

After translating pedestrian images, the second step is training Re-ID model. For the task of domain adaptation $\mathcal{A} \to \mathcal{B}$, where $\mathcal{A}$ is the labeled source dataset and $\mathcal{B}$ is the unlabeled target dataset, we apply the generator $G : \mathcal{A} \to \mathcal{B}$ to generate the style-transferred dataset $\{(G(x_a^i), y_a^i)\}_{i=1}^m$.

Given the translated images and corresponding ID labels, a cross-domain Re-ID model is trained in a supervised manner. We utilize the Part-based Convolutional Baseline (PCB) [4] as the baseline model, which learns a convolutional descriptor consisting of several part-level features. The details can be accessed in Section IV-B.

## IV. Experiments

### A. Datasets

We evaluate our proposed work on two person Re-ID datasets: Market-1501 [9] and DukeMTMC-reID [29], [35], which are both large-scale datasets. Market-1501 contains 1,501 identities and 32,668 bounding boxes collected from 6 camera viewpoints. The training set includes 12,936 images of 751 identities, and the testing set contains 19,732 images of 750 identities. There are 3,368 query images which are hand-drawn from 750 identities in testing set. DukeMTMC-reID contains 1,812 identities from 8 cameras. Of all the 1,812 identities, only 1,404 identities appear under more than two cameras. Following the setting in [29], the dataset is divided into two parts: 16,522 images of 702 identities for training, and 19,989 images of the other 702 identities as testing set. In testing set, there are totally 19,989 images, including 2,228 query images of 702 identities and 17,661 gallery images.

### B. Implementation Details

**Cross-domain Similarity Generative Adversarial Networks.** We implement our method to translate the image styles between two datasets. In our work, the CycleGAN and DIFRN are trained simultaneously. For the DIFRN, we employ the MobileNetV2 [33] as the backbone. To possess the discriminative ability at the beginning of CSGAN training process, we initialize the DIFRN by pretraining it on the annotated source dataset $\mathcal{A}$, which is better than training from the scratch. The input images are all resized to $384 \times 128$. During training, we only use random flipping and random cropping as data augmentation. The Adam optimizer [36] is used to train CSGAN, with a learning rate = 0.0002 and the momentum terms $\beta_1 = 0.5$, $\beta_2 = 0.999$. We set the batch size to 1 and train the model for 8 epochs. For all the experiments, the hyper-parameters are set as follow: $\lambda = 10, \alpha = 5, \beta = 3$ and $m = 0.5$.

**Re-ID model training.** We adopt PCB [4] as our Re-ID feature learning model structure, in which ResNet-50 [34] is used as backbone. Following the setting in [4], all input images are resized to $384 \times 128$. We only employ random flipping as data augmentation. During testing, we calculate the cosine distance between the query images and all gallery images, then the ranking results are used compute the CMC and mAP. In our experiments, we set batch size=32 and train PCB model by

TABLE I
EFFECTIVENESS OF THE STYLE TRANSFER LOSS FUNCTION AND THE CROSS-DOMAIN TRIPLET LOSS FUNCTION IN OUR PROPOSED CSGAN.

| Methods | DukeMTMC-reID→Market-1501 | | | | Market-1501→DukeMTMC-reID | | | |
|---|---|---|---|---|---|---|---|---|
| | rank-1 | rank-5 | rank-10 | mAP | rank-1 | rank-5 | rank-10 | mAP |
| Direct Transfer | 51.5 | 69.7 | 75.8 | 23.7 | 39.2 | 55.6 | 62.0 | 21.5 |
| CSGAN w/$\mathcal{L}_{style}$ | 56.4 | 72.6 | 78.8 | 24.5 | 43.9 | 60.1 | 65.3 | 23.1 |
| CSGAN w/$\mathcal{L}_{style} + \mathcal{L}_{cdtl}$ (m=0.5) | **61.9** | **78.8** | **84.4** | **29.7** | **47.8** | **63.5** | **67.2** | **26.3** |

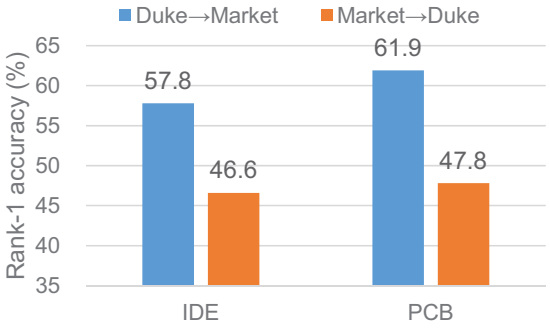


Fig. 4. Comparison of using different Re-ID feature learning models, including IDE [38] and PCB [4]. Duke→Market denotes using Duke as the source dataset and Market as the target dataset. Market→Duke, on the contrary, means Market as the source dataset and Duke as the target dataset.

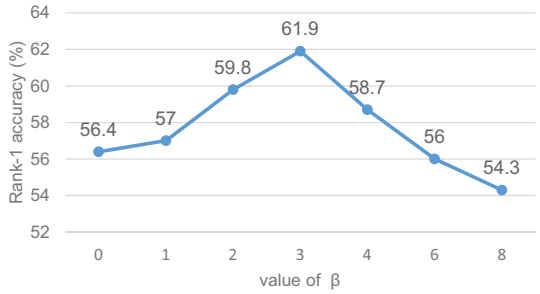


Fig. 5. Analysis of the parameter $\beta$. $\beta$ is a weight to balance the similarity metric constraint with other constrains in overall loss function. The model is trained on DukeMTMC-reID and tested on Market-1501.

60 epochs. The SGD optimizer is used with momentum=0.9 and weight decay= $5 \times 10^{-4}$. The backbone model is pre-trained on ImageNet [37]. Specifically, the initial learning rate for fine-tune layers is 0.05, while newly added classifier layers use 0.005. After 40 epochs, all learning rates are multiplied by 0.1.

### C. Performance Evaluation

**Effectiveness of the CSGAN.** The performance results of our proposed method is shown in Table I. When training a Re-ID model on the source dataset and directly deploying the learned model on the target dataset without any domain adaptation operation, the rank-1 accuracy is 51.5% and 39.2% on Market-1501 and DukeMTMC-reID. Comparing with direct transfer, if only using the style transfer loss function $L_{style}$ for optimizing CSGAN, we obtain +4.7% and +1.6% improvements for rank-1 accuracy and mAP when adopting Market-1501 as source domain and DukeMTMC-reID as target domain. Moreover, when adding the cross-domain triplet loss $\mathcal{L}_{cdtl}$ in CSGAN, we observe further improvements over the results only using style transfer loss. For example, the performance gains +5.5% (from 56.4% to 61.9%) and +3.9% (from 43.9% to 47.8%) in rank-1 accuracy, when tested on Market-1501 and DukeMTMC-reID, respectively. These results indicate the effectiveness of the proposed CSGAN on domain adaptation.

**Analysis of the integration of the image translator and domain-invariant feature representation network.** In our work, there are tow components, an image translator, *i.e.*, CycleGAN and a domain-invariant feature representation network. The two components collaborate with each other in this way: the image translator provides the DIFRN with target domain style images, and the DIFRN forces the CycleGAN to maintain the intra-class similarity and the inter-domain diversity throughout the image generation. In other words, during training, the DIFRN guides the image translation of CycleGAN and the CycleGAN strengthens the discriminability of DIFRN. As shown in Table I, experimental results suggest that the joint optimization of the two components is critical and significant.

**Comparison of different Re-ID feature learning models.** Given the same translated images, we compare two different Re-ID feature learning models: IDE [38] and PCB [4]. As illustrated in Fig. 4, under the same setting, we notice that PCB model outperforms IDE model by $+4.1\%$ and $+1.2\%$ in rank-1 accuracy for Duke→Market and Market→Duke, respectively. This result indicates that a robust feature learning model also facilitates the performance of domain adaptation Re-ID.

### D. Parameters Analysis

Here we conduct additional experiments to evaluate the parameter sensitivity, and results are illustrated in Fig. 5 and Table II.

**The impact of the weight $\beta$.** $\beta$ is a weight to trade off the importance of the proposed cross-domain triplet loss in overall loss function. As shown in Fig. 5, we can clearly see that compared with $\beta = 0$, when gradually increasing the value of $\beta$, our proposed cross-domain triplet loss enhances the Re-ID accuracy. Nonetheless, when $\beta$ grows to 6 or 8,

TABLE II
ANALYSIS OF THE PARAMETER $m$, WHICH IS A PREDEFINED MARGIN.

| $m$ | Duke→Market | | Market→Duke | |
|---|---|---|---|---|
| | Rank-1 | mAP | Rank-1 | mAP |
| 0.1 | 57.3 | 27.6 | 44.1 | 24.7 |
| 0.3 | 59.7 | 28.7 | 46.0 | 25.8 |
| 0.5 | **61.9** | **29.7** | **47.8** | **26.3** |
| 1.0 | 58.1 | 28.2 | 45.5 | 24.6 |
| 2.0 | 57.5 | 27.4 | 44.3 | 23.3 |

$\mathcal{L}_{cdtl}$ has a larger weight in overall objective function, and the performance drops dramatically, even inferior to the result when $\beta = 0$, which implies that an over-large $\beta$ compromises the performance of CSGAN. Based on the results in Fig. 5, we set $\beta$ to 3 in all experiments.

**The influence of margin $m$.** The parameter $m$ in Eq. (6) is a predefined margin, which represents the threshold between $d_{intra}$ and $d_{inter}$. As illustrated in Table II, we conduct experiments on Market-1501 and DukeMTMC-reID to investigate the influence of $m$. Using a lower value of $m$ leads to a narrower threshold, which makes the negative images and positive images too close to be distinguished in the latent space. When increasing $m$, the performance can get improvements. However, the accuracy degrades if the margin is too large. We observe that the best result is obtained when $m$ is set to 0.5. Note that a small or large margin has limitation on improving the performance.

### *E. Comparison with State-of-the-art Methods*

To better evaluate the effectiveness of our proposed method, we compare with the state-of-the-art methods on Market-1501 and DukeMTMC-reID. The compared approaches are categorized into three groups, *i.e.*, hand-crafted methods: local maximal occurrence(LOMO) [39] and bag-of-words(BoW) [9], unsupervised learning methods: UMDL [40], PUL [16] and CAMEL [15], and unsupervised domain adaptation methods: PTGAN [7], SPGAN [8], MMFA [24] and TJ-AIDL [13].

**Comparison of DukeMTMC-reID→Market-1501.** Table III presents the results when DukeMTMC-reID as source dataset and Market-1501 as target dataset. We first compare with two hand-crafted feature based methods which do not require transfer learning on the source dataset and target dataset. Both these two hand-crafted methods fail to produce considerable results and they are far behind the transfer learning based methods. Then we compare with unsupervised domain adaptation methods, which use labeled source data to initialize the model and learn with unlabeled target data. For the task of Duke→Market, our proposed approach achieves rank-1 accuracy=61.9% and mAP=29.7%, clearly outperforming other methods, which is $+10.4\%, +5.2\%$ and $+3.7\%$ higher than SPGAN [8], MMFA [24] and TJ-AIDL [13], respectively. Although BUC [22] achieves the same result in rank-1 accuracy, our proposed method is superior in rank-5, rank-10 accuracy and mAP.

TABLE III
COMPARISON WITH THE STATE-OF-THE-ART METHODS ON MARKET-1501.

| Methods | DukeMTMC-reID→Market-1501 | | | |
|---|---|---|---|---|
| | R-1 | R-5 | R-10 | mAP |
| LOMO [39] | 27.2 | 41.6 | 49.1 | 8.0 |
| BoW [9] | 35.8 | 52.4 | 60.3 | 14.8 |
| UMDL [40] | 34.5 | 52.6 | 59.6 | 12.4 |
| PUL [16] | 45.5 | 60.7 | 66.7 | 20.5 |
| CAMEL [15] | 54.5 | - | - | 26.3 |
| BUC [22] | **61.9** | 73.5 | 78.2 | 29.6 |
| PTGAN [7] | 38.6 | - | 66.1 | - |
| SPGAN [8] | 51.5 | 70.1 | 76.8 | 22.8 |
| MMFA [24] | 56.7 | 75.0 | 81.8 | 27.4 |
| TJ-AIDL [13] | 58.2 | 74.8 | 81.1 | 26.5 |
| Ours(CSGAN) | **61.9** | **77.8** | **82.4** | **29.7** |

TABLE IV
COMPARISON WITH THE STATE-OF-THE-ART METHODS ON DUKEMTMC-REID.

| Methods | Market-1501→DukeMTMC-reID | | | |
|---|---|---|---|---|
| | R-1 | R-5 | R-10 | mAP |
| LOMO [39] | 12.3 | 21.3 | 26.6 | 4.8 |
| BoW [9] | 17.1 | 28.8 | 34.9 | 8.3 |
| UMDL [40] | 18.5 | 31.4 | 37.6 | 7.3 |
| PUL [16] | 30.0 | 43.4 | 48.5 | 16.4 |
| BUC [22] | 40.4 | 52.5 | 58.2 | 22.1 |
| PTGAN [7] | 27.4 | - | 50.7 | - |
| SPGAN [8] | 41.1 | 56.6 | 63.0 | 22.3 |
| MMFA [24] | 45.3 | 59.8 | 66.3 | 24.7 |
| TJ-AIDL [13] | 44.3 | 59.6 | 65.0 | 23.0 |
| Ours(CSGAN) | **47.8** | **63.5** | **67.2** | **26.3** |

**Comparison of Market-1501→DukeMTMC-reID.** Table IV shows the results when we using Market-1501 as the source dataset and testing on DukeMTMC-reID. Compared to the state-of-the-art approaches, our method obtains rank-1 accuracy= 47.8% and mAP= 26.8%. The rank-1 accuracy is $+6.7\%, +2.5\%$ and $+3.5\%$ higher than SPGAN [8],MMFA [24] and TJ-AIDL [13], respectively.

## V. CONCLUSION

In this paper, we present a person style transfer adaptation pipeline for unsupervised domain adaptation Re-ID. Based on the pipeline, we bridge the domain gap by employ an image translator to transfer the labeled images from source domain into target domain style. We further introduce that the intra-class similarity and the inter-domain diversity should be maintained after image translation. To this end, a Domain-Invariant Feature Representation Network (DIFRN) is proposed to extract the discriminative feature representations by integrating

global and fine-grained features. By the joint optimization of CycleGAN and DIFRN, we propose the CSGAN to generate high-quality style-transferred images for domain adaptation. Experimental results show the effectiveness of our method in the task of reducing the Re-ID domain gap. In future work, we will explore the new method to study the camera invariance for cross-domain person Re-ID.

## REFERENCES


[1] Y. Guo and N.-M. Cheung, "Efficient and deep person re-identification using multi-level similarity," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2335–2344, 2018.

[2] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *ArXiv*, vol. abs/1703.07737, 2017.

[3] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, and X. Tang, "Spindle net: Person re-identification with human body region guided feature decomposition and fusion," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 907–915, 2017.

[4] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling," in *ECCV*, 2017.

[5] Y. Ganin and V. S. Lempitsky, "Unsupervised domain adaptation by backpropagation," *ArXiv*, vol. abs/1409.7495, 2014.

[6] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2962–2971, 2017.

[7] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer gan to bridge domain gap for person re-identification," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 79–88, 2017.

[8] W. Deng, L. Zheng, G. Kang, Y. Yang, Q. Ye, and J. Jiao, "Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 994–1003, 2017.

[9] L. Zheng, L. Shen, L. Tian, S. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015.

[10] L. Wu, Y. Wang, J. Gao, and X. Li, "Where-and-when to look: Deep siamese attention networks for video-based person re-identification," *IEEE Transactions on Multimedia*, vol. 21, pp. 1412–1424, 2018.

[11] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, "Person re-identification by multi-channel parts-based cnn with improved triplet loss function," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1335–1344, 2016.

[12] M. M. Kalayeh, E. Basaran, M. Gokmen, M. E. Kamasak, and M. Shah, "Human semantic parsing for person re-identification," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1062–1071, 2018.

[13] J. Wang, X. Zhu, S. Gong, and W. Li, "Transferable joint attribute-identity deep learning for unsupervised person re-identification," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2275–2284, 2018.

[14] Y. Suh, J. Wang, S. Tang, T. Mei, and K. M. Lee, "Part-aligned bilinear representations for person re-identification," in *ECCV*, 2018.

[15] H.-X. Yu, A. Wu, and W.-S. Zheng, "Cross-view asymmetric metric learning for unsupervised person re-identification," *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 994–1002, 2017.

[16] H. Fan, L. Zheng, and Y. Yang, "Unsupervised person re-identification: Clustering and fine-tuning," *TOMCCAP*, vol. 14, pp. 83:1–83:18, 2017.

[17] J. Song, Y. Yang, Y.-Z. Song, T. Xiang, and T. M. Hospedales, "Generalizable person re-identification by domain-invariant mapping network," in *CVPR*, 2019.

[18] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2360–2367, 2010.

[19] G. Lisanti, I. Masi, A. D. Bagdanov, and A. del Bimbo, "Person re-identification by iterative re-weighted sparse ranking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, pp. 1629–1642, 2015.

[20] C. Yan, M. Luo, W. Liu, and Q. Zheng, "Robust dictionary learning with graph regularization for unsupervised person re-identification," *Multimedia Tools and Applications*, vol. 77, pp. 3553–3577, 2017.

[21] E. Kodirov, T. Xiang, and S. Gong, "Dictionary learning with iterative laplacian regularisation for unsupervised person re-identification," in *BMVC*, 2015.

[22] Y. Lin, X. Dong, L. Zheng, Y. Yan, and Y. Yang, "A bottom-up clustering approach to unsupervised person re-identification," in *AAAI*, 2019.

[23] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang, "Camera style adaptation for person re-identification," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5157–5166, 2017.

[24] S. Lin, H. Li, C.-T. Li, and A. C. Kot, "Multi-task mid-level feature alignment network for unsupervised cross-dataset person re-identification," in *BMVC*, 2018.

[25] H. Huang, W. Yang, X. Chen, X. Zhao, K. Huang, J. Lin, G. Huang, and D. Du, "Eanet: Enhancing alignment for cross-domain person re-identification," *ArXiv*, vol. abs/1812.11369, 2018.

[26] Z. Zhong, L. Zheng, S. Li, and Y. Yang, "Generalizing a person retrieval model hetero- and homogeneously," in *ECCV*, 2018.

[27] J. Lv, W. Chen, Q. Li, and C. Yang, "Unsupervised cross-dataset person re-identification by transfer learning of spatial-temporal patterns," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7948–7956, 2018.

[28] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, "Generative adversarial nets," in *NIPS*, 2014.

[29] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by gan improve the person re-identification baseline in vitro," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[30] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *CoRR*, vol. abs/1511.06434, 2015.

[31] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2242–2251, 2017.

[32] Y. Taigman, A. Polyak, and L. Wolf, "Unsupervised cross-domain image generation," *ArXiv*, vol. abs/1611.02200, 2016.

[33] M. Sandler, A. G. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Inverted residuals and linear bottlenecks: Mobile networks for classification, detection and segmentation," *ArXiv*, vol. abs/1801.04381, 2018.

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2015.

[35] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *European Conference on Computer Vision workshop on Benchmarking Multi-Target Tracking*, 2016.

[36] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014.

[37] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and F.-F. Li, "Imagenet: A large-scale hierarchical image database," *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

[38] L. Zheng, Y. Yang, and A. G. Hauptmann, "Person re-identification: Past, present and future," *ArXiv*, vol. abs/1610.02984, 2016.

[39] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2197–2206, 2014.

[40] P. Peng, T. Xiang, Y. Wang, M. Pontil, S. Gong, T. Huang, and Y. Tian, "Unsupervised cross-dataset transfer learning for person re-identification," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1306–1315, 2016.