Style Transfer with Adversarial Learning for Cross-Dataset Person Re-identification

Furong Xu[1], Bingpeng Ma[1⋆], Hong Chang[2], Shiguang Shan[2], and Xilin Chen[2]

[1] School of Computer and Control Engineering, University of Chinese Academy of Sciences, Beijing, 100049, China
[2] Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing 100190, China
xufurong17@mails.ucas.ac.cn, bpma@ucas.ac.cn, {changhong, sgshan, xlchen}@ict.ac.cn

Abstract. Person re-identification (ReID) has witnessed great progress in recent years. Existing approaches are able to achieve significant performance on the single dataset but fail to generalize well on another datasets. The emerging problem mainly comes from style difference between two datasets. To address this problem, we propose a novel style transfer framework based on Generative Adversarial Networks (GAN) to generate target-style images. Specifically, we get the style representation by calculating the Garm matrix of the three-channel original image, and then minimize the Euclidean distance of the style representation on different images transferred by the same generator while image generation. Finally, the labeled source dataset combined with the style-transferred images are all used to enhance the generalization ability of the ReID model. Experimental results suggest that the proposed strategy is very effective on the Market-1501 and DukeMTMC-reID.

Keywords: Person re-identification · Style transfer · Adversarial learning

1 Introduction

Person re-identification (ReID) [31] is a challenging image retrieval task, which amis to match persons in the different places/time under non-overlapping cameras. It is obvious diversity between different images of the same identity because of various viewpoint, body structures, and occlusion (see Fig. 1). Although existing many supervised deep learning approaches [28, 22] have achieved significant performance on the single dataset, it is worth noting that manually annotating a new dataset is expensive and impractical.

Many approaches perform well on one person dataset but fail to generalize well to another dataset [7]. Difference between different datasets mainly comes from light conditions and backgrounds (we call it style), which results in poor

⋆ Corresponding author.

Fig. 1. Examples of person re-identification datasets. The left is selected from the Market-1501 and the right is selected from the DukeMTMC-reID.

performance under cross-dataset ReID. For example, as shown in Fig. 1, Market-1501 dataset is collected in summer with strong lighting, but the DukeMTMC-reID is collected in winter. Therefore, the seasonal difference causes different styles of dressing. A single dataset often has its own style characteristics, and the model is easy to overfit on the style of the dataset. Considering this matter, some approaches [21, 26] employ unsupervised dictionary learning to acquire a dataset-shared space, Xiao et al. [24] design a Domain Guided Dropout (DGD) classification model to learn generic feature representations with multiple datasets. Recently, several works [6, 23] adopt domain adaptation method based on CycleGAN [34] to translate the labeled source dataset to target domain, and then the transferred images are used to train ReID models.

As reported in [9], the images can be composed of the content and the style. Therefore, the features extracted by the ReID model will consist of both the style representation and the content representation, any poor representation will affect the final accuracy. To ensure performance of the cross-dataset ReID, it is quite important to extract the dataset-invariant content representation, as well as to obtain the consistent style representation between the source dataset and the target dataset. Some works [23, 21, 26] have committed to learning the shared features across all datasets. However, none of these methods focuses on pushing the style consistency between datasets.

To explicitly address these issues of the cross-dataset ReID, we propose a novel style-transferrable framework based on CycleGAN to generate target-style images for learning discriminative features. Generally speaking, the style representation can be extracted by looking at the spatial correlation of the values within a given feature map, which is available by calculating Gram matrix of the feature map, and the feature map is usually the output of the network middle layer [9, 10]. Inspired by their style representation way, we calculate the Gram matrix of the three-channel original image to get the style representation, rather than the feature map of the network middle layer, because the original image contains all the style information. In the image generation process, we constrain the Euclidean distance of the style representation to be minimized, the constrained objects are different images generated by the same generator. We call

this structure Style-consistent, Identity-consistent, Cycle-consistent Generative Adversarial Network(SICGAN). After we get the transferred target-style images, considering that ID-related information is preserved between the original images and the corresponding generated images, we use the style-transferred image together with the original image as a training set to train the ReID model in supervised learning, which enables the model to learn discriminative features. Extensive experimental results show that our method can achieve consistent and competitive ReID accuracy on two large-scale datasets.

2 Related Work

In this section, we first review some related works in person ReID, especially some methods about single-dataset and cross-dataset, then we introduce a few image generation works related to our method, such as style transfer and image-image translation.

2.1 Person ReID

Single-dataset ReID. Single-dataset ReID means that both training and testing are in the same dataset. Recently, deep learning based person ReID approaches [18, 1, 29] have achieved great success through simultaneously learning the person representation and similarity within one network on a single-dataset. These methods usually learn the ID-discriminative Embedding (IDE) feature [31] via training a deep classification network. In addition, some works try to introduce the pair-wise contrastive loss [5], triplet loss [14, 27] and quadruplet loss [4] to further enhance the IDE feature. There are currently some methods that achieve good performance in single-dataset. AlignedReID [28] introduces a feature matching method to align different body parts. [3] uses manifold-based affinity learning for postprocessing. These deep learning methods can achieve considerable results in large-scale datasets [17, 20].

Cross-dataset ReID. Cross-dataset ReID is very close to the realistic scenarios, which trains on one dataset and tests on another. Many transfer learning methods [25, 24, 11] have been adopted for cross-dataset ReID in the hope that labelled data from other datasets can provide transferable identity-discriminative information for the given target dataset. Peng et al. [21] use unsupervised multi-task dictionary learning(UMDL) to learn discriminative representation. Yu et al. [26] use unsupervised camera-specific projection to get the shared space. Generative Adversarial nets(GAN) [12] has also gained promotion in the ReID task recently. Deng et al. [6] propose a similarity preserving generative adversarial network (SPGAN) for domain adaptation. Wei et al. [23] propose a Person Transfer Generative Adversarial Network (PTGAN) to bridge the domain gap between different datasets which relieves the expensive costs of annotating new training samples. Different from [6, 23], our SICGAN use extra style consistency constraints to ensure the style-transferred images can be applied to cross-dataset model training.

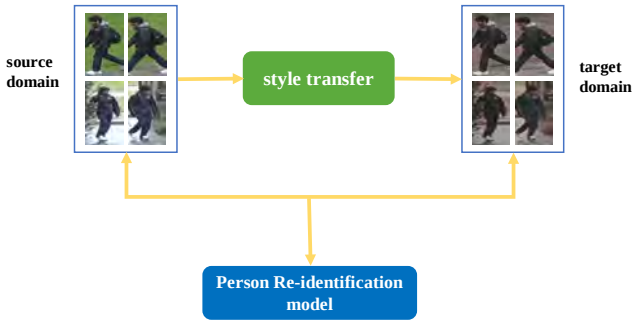

Fig. 2. Pipeline of our framework. First we transfer the style of the target dataset to the source dataset, and then use the style-tarnsferred images together with the source dataset to train the ReID model.

2.2 Image Generation

Style Transfer. Style transfer is the task of generating a new image, whose style is equal to a style image and whose content is equal to a content image. Some works [9, 10] have a clear definition of the style and content representation of an image. In [9], Gatys et al. study how it is actually possible to transfer artistic style from one painting to another picture using convolutional neural networks. Frigo et al. [8] propose an effective style transfer algorithm fully based on traditional texture synthesis method. Inspired by [9, 10], we use Gram matrix at the image level to represent the style of image.

Image-Image translation. Recently, there are some methods [34, 16] based on Generative Adversarial Networks(GANs) [12] for image-level domain translation. In [34], Zhu et al. present an approach to translate an image from a source domain to a target domain in the absence of paired examples. Image analogy [15] aims to learn a mapping between a pair of source images and target stylised images in a supervised manner. Aytar et al. [2] use a weight-sharing strategy to learn a common representation across domains. Different from the previous methods which mainly consider the quality of the generated samples, some works [6, 23] aims at using the style-transferred samples to improve the performance of ReID. However, our method introduce extra style consistency constraints during image generation for cross-dataset ReID.

3 Our Proposed Method

Our method transfers the style of the target dataset to the source dataset, and then uses the source dataset combined with the style-transferred images to learn discriminative feature for ReID. The framework of the proposed approach is shown in Fig. 2, including image generation and feature learning.

Fig. 3. Visual examples of image-image translation. The left five columns map DukeMTMC-reID images to the Market-1501 style, and the right five columns map Market-1501 images to the DukeMTMC-reID style. From top to bottom: (a) original image, (b) output of CycleGAN, (c) output of SICGAN. Images produced by SICGAN are further constrained by style loss.

3.1 CycleGAN Revisit

CycleGAN, as one of the image generation technique, has achieved good performance in many tasks, such as collection style transfer, object transfiguration, season transfer and photo enhancement. It mainly consists of two generator-discriminator pairs, $\{G, D_T\}$ and $\{F, D_S\}$. Generator G is used to translate the source dataset to the target-style. Discriminator D_T is used to determine the real and fake of the target dataset and the G generated images. Similarly, F and D_S complete the reverse transfer and discrimination. The adversarial loss of generator G and discriminator D_T is:

$$L_{Tadv} = E_{y\sim p_y}[(D_T(y) - 1)^2] + E_{x\sim p_x}[(D_T(G(x)))^2] \tag{1}$$

The adversarial loss of generator F and discriminator D_S is:

$$L_{Sadv} = E_{x\sim p_x}[(D_S(x) - 1)^2] + E_{y\sim p_y}[(D_S(F(y)))^2] \tag{2}$$

Where p_x and p_y denote the sample distributions in the source and target dataset.

For the purpose of reducing the space of possible mapping functions, CycleGAN also introduces a cycle-consistent loss, which attempts to recover the original image after a cycle of translation and reverse translation. The cycle-consistent loss is:

$$L_{cyc}(G, F) = E_{x\sim p_x}[||F(G(x)) - x||_1] + E_{y\sim p_y}[||G(F(y)) - y||_1] \tag{3}$$

Apart from cycle-consistent loss and adversarial loss, CycleGAN proposes an additional identity loss to encourage the mapping to preserve color composition between the input and output. The identity mapping loss is:

$$L_{identity}(G, F) = E_{x\sim p_x}[||F(x) - x||_1] + E_{y\sim p_y}[||G(y) - y||_1] \tag{4}$$

As shown in Fig. 3(b), CycleGAN can generate relatively realistic images and preserve most ID-related information of original images, which improved the performance of cross-dataset ReID to a certain extent. However, there is still a certain style bias between the generated images and the corresponding dataset, and we can further improve performance by reducing style differences.

3.2 SICGAN

Generally speaking, the images generated by CycleGAN are realistic and preserve a large amount of ID-related information about the original image, but the change in style is not very obvious. As shown in Fig. 3(b), CycleGAN slightly fades the background and changes the lighting. Considering that the decrease in performance of cross-dataset is mainly due to style divergence between different datasets, in order to generate images that can improve the performance of the ReID model, the following requirements should be met: 1) the style-transferred images are realistic. 2) preserve ID-related information between the style-transferred images and the original images, so that they can be used to train the ReID models as the same ID. 3) the style of style-transferred images is consistent with the target dataset. CycleGAN can solve the first two requirements, therefore, we focus on the style transfer between the source dataset and the target dataset. For the purpose of generating target-style images, we propose Style-consistent, Identity-consistent, Cycle-consistent Generative Adversarial Network(SICGAN). Our SICGAN based on CycleGAN pay attention to the style consistency constraints of different images generated by the same generator.

For style constraints, we first need to extract the style representation of the image. Gatys et al. [9, 10] found that the extraction of style representation can be done by looking at the spatial correlation of the values within a given feature map, the feature map is usually the output of the network middle layer. Inspired by their style representation manner, we calculate the Gram matrix of the three-channel original image to get the style representation. Different from [9, 10], we believe that the original image contains more style information than feature map, which guarantees a better extraction of the style representation of the image. Specifically, in the image generation process, we constrain the Euclidean distance of the style representation from different images generated by the same generator to be minimized. The style loss is:

$$L_{style}(G,F) = \frac{1}{c*m*n}[||M(F(x)) - M(F(y))||_2 + ||M(G(x)) - M(G(y))||_2] \tag{5}$$

where $M(.)$ is used to get the Gram matrix, c is the number of channel of the image, m is the height of the image, and n is the width of the image. In our setting, they are 3, 256 and 128 respectively. In order to get the Gram matrix of fake image *F(x), F(y), G(x), G(y)*, the two-dimensional pixel value of each channel is expanded into a one-dimensional vector by row. Then each channel vector of a image is dot-producted each other to obtain a $c*c$ matrix. The specific calculation formula is as follows:

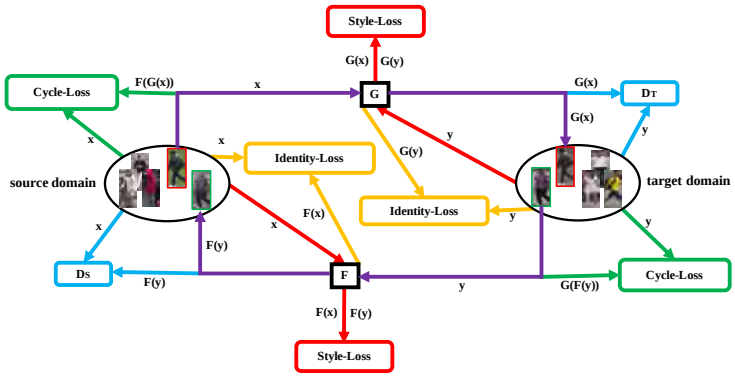

Fig. 4. Our SICGAN consists of two generator-discriminator pairs, $\{G, D_T\}$ and $\{F, D_S\}$. G is used to generate the target-style image, F is the opposite. The adversarial loss, cycle-consistent loss and identity loss here are the same as those used in CycleGAN, our style loss is used to constrain the style representation of different images generated by the same generator, such as G(x) and G(y).

$$M_{ij} = \sum_{k} A_{ik} A_{jk} \tag{6}$$

where A_{ik} represents the k-th pixel value of the i-th channel in channel vector. During the data generation phase, the overall loss function can be written as:

$$L = L_{Tadv} + L_{Sadv} + \lambda_1 L_{cyc}(G, F) + \lambda_2 L_{identity}(G, F) + \lambda_3 L_{style}(G, F) \tag{7}$$

where λ_1, λ_2 and λ_3 correspond to the weight of cycle-consistent loss, identity loss and style loss respectively. The image generation process is shown in Fig. 4. To get images that meet the requirements, the parameters of generators G, F and discriminators D_S, D_T are adversarially optimized step by step, and style loss is only used to update the parameters of generator. The image generated by SICGAN is shown in Fig. 3(c). When testing, target-style and source-style images are generated by well-trained generators G and F.

3.3 Feature Learning

Following the pipeline [6], feature learning is the second step for cross-dataset person ReID. But Deng et al. only use the transferred images for feature learning. Considering that the generated image preserves the ID-related information of the original image, we employ the labeled original images and style-transferred images together to learn discriminative features, which makes full use of the original image information and reduce the impact of generating less realistic images. Our pipeline is shown in Fig. 2. Similar to [31], we train a classification

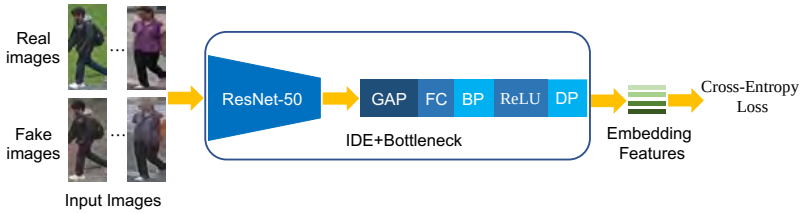

Fig. 5. ReID network structure with bottleneck. Compared to IDE, fully-connected(FC), batch normalization(BN), ReLu activation function(ReLu), and dropout(DP) are newly added layers.

network named IDE for ReID embedding learning. The base model is ResNet-50 [13], we retain layer4 of ResNet-50 and its previous structure, adding the global average pooling(GAP) layer and fully-connected(FC) layer whose output dimension is the number of training identities. When testing, given an input image of query, we can extract the 2,048-dim feature after GAP to calculate Euclidean distance from the gallery.

Bottleneck. In order to confirm the performance of the generated image does not depend on the model, we also adopt an improved version of the classification network. Similar to [33], we add a bottleneck after the layer4 of ResNet-50. The bottleneck includes 1) global average pooling(GAP), 2) fully connected layer(FC), 3) batch normalization(BN), 4) ReLu activation function(ReLu), 5) dropout(DP), which used to fuse the learned features. The network structure after adding the bottleneck is shown in Fig. 5. It differs from IDE in that the FC, BN, ReLu, and DP layer of the network structure are newly added. Generally speaking, its performance is higher than IDE.

4 Experiment

4.1 Dataset

We evaluate our method on the Market-1501 and DukeMTMC-reID, because both datasets are large-scale and have different styles. Sample images of the two datasets are show in Fig. 1. Their amount of data can train deep learning methods, and different styles can strongly prove the effectiveness of our method. For both datasets, we use rank-1, rank-5, rank-10 accuracy and mean average precision(mAP) for result evaluation.

Market-1501[30] is collected from six cameras in front of a supermarket in Tsinghua University, and contains 32,668 annotated bounding boxes of 1,501 identities. For evaluation, we employe the same settings as [30], 12,936 images from 751 identities are used for training, and 19,732 images from 750 identities plus some distractors form the gallery. Moreover, 3,368 hand-drawn bounding boxes from 750 identities are used as queries to retrieve the corresponding person images in the gallery. We use the single-query evaluation in our experiment.

DukeMTMC-reID[32] is a ReID version of the DukeMTMC dataset, which recorded outdoors on the Duke University campus with 8 synchronized cameras. It contains 34,183 image boxes of 1,404 identities. Similar to the division of Market-1501, the dataset contains 16,522 training images from 702 identities, 2,228 query images from another 702 identities and 17,661 gallery images.

4.2 Experiment Settings

The entire experiment includes two steps. Firstly, SICGAN generates realistic images that perserve the ID-related information of the source images and have the same style as the target dataset. Secondly, the original images and the generated images are used together as a training set to train the ReID model. In the second step of training, the generated images and the corresponding original images maintain the same ID.

SICGAN model. We implement SICGAN with PyTorch to achieve image-image translation. To prove the validity of our method, we adopt the network structure of CycleGAN released by its authors, which consists of a series of building block based on ResNet [13]. When training SICGAN model, we use the training set of Market-1501 and DukeMTMC-reID as transfer objects, the inputs images are resized to 256×128. In all experiment, we empirically set $\lambda_1 = 10, \lambda_2 = 5, \lambda_3 = 5$ in Equ. (7). We make the initial learning rate 0.0002, batch size is 1, and the model stop training after 10 epochs. For generator and discriminator, we use Adam optimizer with the default hyper-parameters($\eta = 10^{-8}, \beta_1 = 0.9, \beta_2 = 0.999$). During the testing procedure, we employ the Generator G to get fake DukeMTMC-reID images of Market-1501 style, the Generator F to get fake Market-1501 images of DukeMTMC-reID style.

ReID model. To test the effectiveness of the generated images, we train a classification network named IDE [31] for ReID embedding learning. The input images are resized to 256×128, and random horizontal flipping for data augmentation are employed during training. We use ResNet-50 as backbone, in which the last fully connected layer has 751 and 702 units for Market-1501 and DukeMTMC-reID, respectively. We set the batch size to 32 and use the SGD(momentum=0.9, weight_decay=5e-4, nesterov=True) optimizer to train ReID model. The learning rate starts with 0.001 for ResNet-50 base layers and 0.01 for the full connected layers, and divided by 10 after 40 epochs. We train 60 epochs in total. In testing, we extract the output of the global average pooling(GAP) layer as image descriptor (2,048-dim) and use the Euclidean distance to compute the similarity between images. For ReID network structure with bottleneck, we have the same settings with IDE. Additionally, the output units of FC in bottleneck is 512, the dropout probability is 0.5 and the learning rate of new layer starts with 0.01.

4.3 Evaluation

Evaluate the Effect of CycleGAN. In our experiment, there is a large performance drop when directly using a source-trained model on the target dataset.

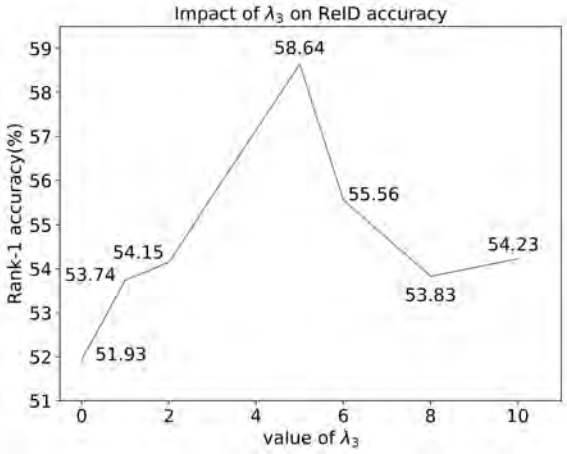

Fig. 6. λ_3 (Equ. 7) *v.s* ReID Rank-1 accuracy of IDE model on Market-1501. A larger λ_3 means larger weight of style loss.

As show in Table 1, the ReID model trained and tested on Market-1501 achieves 79.16 % (IDE(Market)) in rank-1 accuracy, but drops to 43.34% when trained on DukeMTMC-reID and tested on Market-1501(IDE(Real)).We can see the same phenomenon in Table 2. Following the pipeline of [6], we first translate the labeled images from the source dataset to the target dataset and then use the translated images to train ReID models. We achieve significant performance improvements, training with generated fake DukeMTMC-reID of Market-1501 style(*e.g.*,Table 1(IDE(Fake))) improved by 7.51% and 3.17% in rank-1 accuracy and mAP compared to direct testing (IDE(Real)). A similar improvement can be observed on DukeMTMC-reID, after the Market-1501 is translated by CycleGAN, the performance gain is +6.35% and +2.75%. This shows that the image generated by CycleGAN has a certain effect on accuracy improvement of the cross-dataset ReID. It is consistent with the experiments reported in [6].

Evaluate the Effect of SICGAN. On top of the CycleGAN baseline, we add style loss to optimize generators. The effectiveness of the proposed SICGAN can be seen in Table 1 and Table 2. Compared with CycleGAN, our SICGAN leads to +2.56% and +4.24% in rank-1 accuracy and mAP on Market-1501, and +2.6% and +2.96% on DukeMTMC-reID. The coherent improvement suggests that style consistency constraints on the different samples generated by the same generator are necessary. Our SICGAN generate the target-style images, so that the ReID model can learn the style representation feature of the target dataset. Example of transferred images by SICGAN are show in Fig. 3(c).

Evaluate the Effect of training together. Wei et al. [6] only use the generated images to train the ReID model. However, in view of the following fact: 1) the most ID-related information is retained during image generation, 2) there is always a gap between the generated image and the real image. In view of the

Table 1. Comparison of various methods on the Market-1501 dataset. The Market in parentheses indicates that the model is trained using the training set of Market-1501, the Real indicates the training set using the original DukeMTMC-reID, and the Fake indicates the training set using the style-transferred DukeMTMC-reID.

Tab Generate	ReID	top1	top5	top10	mAP
-	IDE(Market)	79.16	87.32	89.28	59.34
-	IDE(Real)	43.34	61.52	68.97	17.31
-	Bottleneck(Real)	46.76	66.27	72.26	20.39
CycleGAN	IDE(Fake)	50.85	66.92	73.29	20.48
CycleGAN	IDE(Real+Fake)	51.93	68.85	76.33	22.85
SICGAN	IDE(Fake)	53.41	70.81	77.19	24.72
SICGAN	IDE(Real+Fake)	58.64	74.94	81.79	26.68
SICGAN	Bottleneck(Real+Fake)	**60.18**	**75.92**	**82.21**	**29.00**

Table 2. Comparison of various methods on the DukeMTMC-reID dataset. The Duke in parentheses indicates that the model is trained using the training set of DukeMTMC-reID, the Real indicates the training set using the original Market-1501, and the Fake indicates the training set using the style-transferred Market-1501.

Generate	ReID	top1	top5	top10	mAP
-	IDE(Duke)	70.94	80.87	84.65	53.49
-	IDE(Real)	36.47	52.51	58.79	19.30
-	Bottleneck(Real)	38.01	53.05	59.91	21.83
CycleGAN	IDE(Fake)	42.82	55.79	61.98	22.05
CycleGAN	IDE(Real+Fake)	43.96	58.25	63.60	23.65
SICGAN	IDE(Fake)	45.42	61.23	66.53	25.01
SICGAN	IDE(Real+Fake)	48.42	63.57	69.27	27.85
SICGAN	Bottleneck(Real+Fake)	**49.82**	**65.66**	**70.86**	**29.23**

above-mentioned facts, we use the original images and the style-transferred fake images together as training set to train the ReID model. As show in Table 1 and Table 2, for CycleGAN, we gains +1.08% and +2.37% in rank-1 accuracy and mAP on Market-1501, +1.14% and +1.6% on DukeMTMC-reID. For SICGAN, the gains are 5.23% and 1.96% on Market-1501, 3% and 2.84% on DukeMTMC-reID. When training together, the images generated by SICGAN combined with the original images will get a more significant performance improvement, indicating that the style of images generated by SICGAN is closer to the target images, and the two styles of images together help to learn discriminative features.

Comparison of different feature learning methods. In this paper, we use two feature learning methods. For comparison with [6], we use the more common IDE [31]. In order to better prove that generated images by SICGAN can effectively improve performance, and do not depend on the selected ReID model, we employ an improved version of the IDE as well. After adding the

Table 3. Results on the Market-1501 dataset.

Methods	top1	top5	top10	mAP
BOW [30]	35.8	52.4	60.3	14.8
LOMO [19]	27.2	41.6	49.1	8.0
UMDL [21]	34.5	52.6	59.6	12.4
PUL [7]	45.5	60.7	66.7	20.5
CAMEL [26]	54.5	-	-	26.3
Direct transfer	43.34	61.52	68.97	17.31
SPGAN [6]	51.5	70.1	76.8	22.8
SICGAN(Fake)	53.41	70.81	77.19	24.72
SICGAN(Real+Fake)	58.64	74.94	81.79	26.68
SICGAN(Real+Fake+Bottleneck)	**60.18**	**75.92**	**82.21**	**29.00**

bottleneck show in Fig. 5, we achieve better accuracy as [33]. As show in Table 1 and Table 2, our network achieves 60.18% rank-1 accuracy and 29.00% mAP on Market-1501, 49.82% rank-1 accuracy and 29.23% mAP on DukeMTMC-reID, both higher than IDE, which is consistent with direct testing (Bottleneck(Real)).

Sensitivity of SIGAN to key parameters. The weight λ_3 of style loss in Equ. 7 is a key parameter. If λ_3=0, the style loss is not back propagated. If λ_3 gets larger, the weight of style differences in loss calculation increases. We do experiment to verify the impact of λ_3, and results are shown in Fig. 6. When increasing λ_3 to 5, we have much superior accuracy. It indicates that the style loss and the identity loss are equally important.

4.4 Comparison with State-of-the-art Methods

We compare the proposed method with the state-of-the-art unsupervised learning method and cross-dataset methods on Market-1501 and DukeMTMC-reID in Table 3 and Table 4, respectively.

Market-1501. On Market-1501, we first compare our results with two hand-crafted features, *i.e.*, Bag-of-Words (BoW) [30] and local maximal occurrence (LOMO) [19]. Those two hand-crafted features are directly applied on test dataset without any training process, their inferiority can be clearly observed. We also compare existing unsupervised methods, including the Clustering-based Asymmetric MEtric Learning (CAMEL) [26], UMDL [21], and the Progressive Unsupervised Learning (PUL) [7]. In addition, we compare to data augmentation method for cross-dataset ReID, such as SPGAN [6]. In the single-query setting, our SICGAN achieves 60.18% rank-1 accuracy and 29.00% mAP. It outperforms the second best method (CAMEL) by +5.68% in rank-1 accuracy and +2.70% in mAP. In order to compare with the similar method SPGAN, we use the same settings, using IDE to train the generated image, we get 53.41% and 24.72% in rank-1 accuracy and mAP with fewer model parameters, both higher than SPGAN. The comparisons indicate the competitiveness of the proposed method on Market-1501.

Table 4. Results on the DukeMTMC-reID dataset.

Methods	top1	top5	top10	mAP
BOW [30]	17.1	28.8	34.9	8.3
LOMO [19]	12.3	21.3	26.6	4.8
UMDL [21]	18.5	31.4	37.6	7.3
PUL [7]	30.0	43.4	48.5	16.4
Direct transfer	36.47	52.51	58.79	19.30
SPGAN [6]	41.1	56.6	63.0	22.3
SICGA(Fake)	45.42	61.23	66.53	25.01
SICGAN(Real+Fake)	48.42	63.57	69.27	27.85
SICGAN(Real+Fake+Bottleneck)	**49.82**	**65.66**	**70.86**	**29.23**

DukeMTMC-reID. On DukeMTMC-reID, we compare the proposed method with BoW [30], LOMO [19], UMDL [21], PUL [7] and SPGAN [6] under the single-query setting. Our proposed method achieve 49.82% in rank-1 accuracy and 29.23% in mAP. Compared with the second best method, *i.e.*, SPGAN, our results are +4.32% higher in rank-1 accuracy and +2.71% in mAP under the same settings. Therefore, the superiority of SICGAN can be concluded.

5 Conclusions

This paper focuses on generating target-style images for cross-dataset person ReID. Models trained on one dataset often failed to generalize well on another due to style diversity. To achieve improved performance in the new dataset, we propose a novel style-transferrable framework to generate target-style images for style adaption. Specifically, we get the style representation by calculating the Gram matrix of the three-channel original image, and then minimize the Euclidean distance of the style representation on different images transferred by the same generator during image generation. Finally, the labeled source dataset combined with style-transferred images are both used for training the ReID models to enhance the generalization ability. We show that the images generated by SICGAN allow the model to learn more discriminative features and yield consistent improvement on different ReID model.

Acknowledgment

This work is supported in part by National Basic Research Program of China (973 Program): 2015CB351802, and Natural Science Foundation of China (NSFC): 61390501, 61876171 and 61572465.

References

1. Ahmed, E., Jones, M., Marks, T.K.: An improved deep learning architecture for person re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 3908–3916 (2015)
2. Aytar, Y., Castrejon, L., Vondrick, C., Pirsiavash, H., Torralba, A.: Cross-modal scene networks. arXiv preprint arXiv:1610.09003 (2016)
3. Bai, S., Bai, X., Tian, Q.: Scalable person re-identification on supervised smoothed manifold. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 2530–2539 (2017)
4. Chen, W., Chen, X., Zhang, J., Huang, K.: Beyond triplet loss: a deep quadruplet network for person re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 403–412 (2017)
5. Chung, D., Tahboub, K., Delp, E.J.: A two stream siamese convolutional neural network for person re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 1983–1991 (2017)
6. Deng, W., Zheng, L., Ye, Q., Kang, G., Yang, Y., Jiao, J.: Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person reidentification. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 994–1003 (2018)
7. Fan, H., Zheng, L., Yang, Y.: Unsupervised person re-identification: clustering and fine-tuning. arXiv preprint arXiv:1705.10444 (2017)
8. Frigo, O., Sabater, N., Delon, J., Hellier, P.: Split and match: Example-based adaptive patch sampling for unsupervised style transfer. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 553–561 (2016)
9. Gatys, L.A., Ecker, A.S., Bethge, M.: A neural algorithm of artistic style. arXiv preprint arXiv:1508.06576 (2015)
10. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 2414–2423 (2016)
11. Geng, M., Wang, Y., Xiang, T., Tian, Y.: Deep transfer learning for person reidentification. arXiv preprint arXiv:1611.05244 (2016)
12. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in Neural Information Processing Systems. pp. 2672–2680 (2014)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
14. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person reidentification. arXiv preprint arXiv:1703.07737 (2017)
15. Hertzmann, A., Jacobs, C.E., Oliver, N., Curless, B., Salesin, D.H.: Image analogies. In: Conference on Computer Graphics and Interactive Techniques. pp. 327–340 (2001)
16. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 1125–1134 (2017)
17. Ji, R., Liu, H., Cao, L., Liu, D., Wu, Y., Huang, F.: Toward optimal manifold hashing via discrete locally linear embedding. IEEE Transactions on Image Processing pp. 5411–5420 (2017)

18. Li, W., Zhao, R., Xiao, T., Wang, X.: Deepreid: Deep filter pairing neural network for person re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 152–159 (2014)
19. Liao, S., Hu, Y., Zhu, X., Li, S.Z.: Person re-identification by local maximal occurrence representation and metric learning. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 2197–2206 (2015)
20. Liu, H., Ji, R., Wang, J., Shen, C.: Ordinal constraint binary coding for approximate nearest neighbor search. IEEE Transactions on Pattern Analysis & Machine Intelligence (2018). https://doi.org/10.1109/TPAMI.2018.2819978
21. Peng, P., Xiang, T., Wang, Y., Pontil, M., Gong, S., Huang, T., Tian, Y.: Unsupervised cross-dataset transfer learning for person re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 1306–1315 (2016)
22. Wang, G., Yuan, Y., Chen, X., Li, J., Zhou, X.: Learning discriminative features with multiple granularities for person re-identification. arXiv preprint arXiv:1804.01438 (2018)
23. Wei, L., Zhang, S., Gao, W., Tian, Q.: Person transfer gan to bridge domain gap for person re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 79–88 (2017)
24. Xiao, T., Li, H., Ouyang, W., Wang, X.: Learning deep feature representations with domain guided dropout for person re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 1249–1258 (2016)
25. Yosinski, J., Clune, J., Bengio, Y., Lipson, H.: How transferable are features in deep neural networks? In: Advances in Neural Information Processing Systems. pp. 3320–3328 (2014)
26. Yu, H.X., Wu, A., Zheng, W.S.: Cross-view asymmetric metric learning for unsupervised person re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 994–1002 (2017)
27. Yu, R., Dou, Z., Bai, S., Zhang, Z., Xu, Y., Bai, X.: Hard-aware point-to-set deep metric for person re-identification. arXiv preprint arXiv:1807.11206 (2018)
28. Zhang, X., Luo, H., Fan, X., Xiang, W., Sun, Y., Xiao, Q., Jiang, W., Zhang, C., Sun, J.: Alignedreid: Surpassing human-level performance in person re-identification. arXiv preprint arXiv:1711.08184 (2017)
29. Zhao, L., Li, X., Wang, J., Zhuang, Y.: Deeply-learned part-aligned representations for person re-identification. In: IEEE International Conference on Computer Vision. pp. 3219–3228 (2017)
30. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: A benchmark. In: IEEE International Conference on Computer Vision. pp. 1116–1124 (2015)
31. Zheng, L., Yang, Y., Hauptmann, A.G.: Person re-identification: Past, present and future. arXiv preprint arXiv:1610.02984 (2016)
32. Zheng, Z., Zheng, L., Yang, Y.: Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In: IEEE International Conference on Computer Vision. pp. 3754–3762 (2017)
33. Zhong, Z., Zheng, L., Zheng, Z., Li, S., Yang, Y.: Camera style adaptation for person re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 5157–5166 (2018)
34. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: IEEE International Conference on Computer Vision. pp. 2223–2232 (2017)