Image-Image Domain Adaptation with Preserved Self-Similarity and Domain-Dissimilarity for Person Re-identification

Weijian Deng[†], Liang Zheng[‡], Guoliang Kang[‡], Yi Yang[‡], Qixiang Ye[†], Jianbin Jiao[†*]
[†]University of Chinese Academy of Sciences [‡]University of Technology Sydney
dengweijian16@mails.ucas.ac.cn, liangzheng06@gmail.com

Abstract

Person re-identification (re-ID) models trained on one domain often fail to generalize well to another. In our attempt, we present a "learning via translation" framework. In the baseline, we translate the labeled images from source to target domain in an unsupervised manner. We then train re-ID models with the translated images by supervised methods. Yet, being an essential part of this framework, unsupervised image-image translation suffers from the information loss of source-domain labels during translation.

Our motivation is two-fold. First, for each image, the discriminative cues contained in its ID label should be maintained after translation. Second, given the fact that two domains have entirely different persons, a translated image should be dissimilar to any of the target IDs. To this end, we propose to preserve two types of unsupervised similarities, 1) self-similarity of an image before and after translation, and 2) domain-dissimilarity of a translated source image and a target image. Both constraints are implemented in the similarity preserving generative adversarial network (SPGAN) which consists of a Siamese network and a CycleGAN. Through domain adaptation experiment, we show that images generated by SPGAN are more suitable for domain adaptation and yield consistent and competitive re-ID accuracy on two large-scale datasets.

1. Introduction

This paper considers domain adaptation in re-ID. The re-ID task aims at searching for the relevant images to the query. In our setting, the source domain is fully annotated, while the target domain does not have ID labels. In the community, domain adaptation re-ID are gaining increasing popularity, because 1) of the expensive labeling process and 2) when models trained on one dataset are directly used on another, the re-ID accuracy drops dramatically [6] due to *dataset bias* [41]. As a result, current fully supervised, single-domain re-ID methods may be limited in real world scenarios, where domain-specific labels are not available.

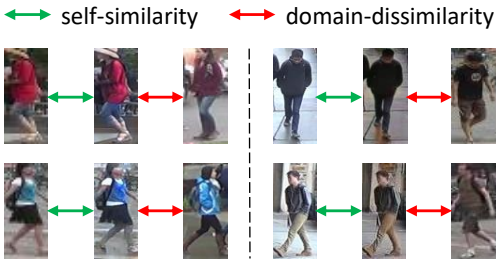

Figure 1: Illustration of self-similarity and domain-dissimilarity. In each triplet, left: a source-domain image, middle: a source-target translated version of the source image, right: an arbitrary target-domain image. We require that 1) a source image and its translated image should contain the same ID, *i.e.*, self-similarity, and 2) the translated image should be of a different ID with any target image, *i.e.*, domain dissimilarity. Note: the source and target domains contain entirely different IDs.

A common strategy for this problem is unsupervised domain adaptation (UDA). But this line of methods assume that the source and target domains contain the same set of classes. Such assumption does not hold for person re-ID because different re-ID datasets usually contain entirely different persons (classes). In domain adaptation, a recent trend consists in image-level domain translation [18, 4, 28]. **In the baseline approach**, two steps are involved. First, labeled images from the source domain are transferred to the target domain, so that the transferred image has a similar style with the target domain. Second, the style-transferred images and their associated labels are used in supervised learning in the target domain. In literature, commonly used style transfer methods include [27, 22, 46, 53]. In this paper, we use CycleGAN [53] following the practice in [27, 18].

In person re-ID, there is a distinct yet unconsidered requirement for the baseline described above: the visual content associated with the ID label of an image should be preserved after image-image translation. In our scenario, such

*Corresponding Author

Figure 2: Pipeline of the "learning via translation" framework consisting of two steps. First, we translate the labeled images from a source domain to a target domain in an unsupervised manner. Second, we train re-ID models with the translated images using supervised feature learning methods. The major contribution consists in the first step, *i.e.*, similarity preserving image-image translation.

visual content usually refers to the underlying (latent) ID information for a foreground pedestrian. To meet this requirement tailored for re-ID, we need additional constraints on the mapping function. In this paper, we propose a solution to this requirement, motivated from two aspects. First, a translated image, despite of its style changes, should contain the same underlying identity with its corresponding source image. Second, in re-ID, the source and target domains contain two entirely different sets of identities. Therefore, a translated image should be different from any image in the target dataset in terms of the underlying ID.

This paper introduces the Similarity Preserving cycle-consistent Generative Adversarial Network (SPGAN), an unsupervised domain adaptation approach which generates images for effective target-domain learning. SPGAN is composed of a Siamese network (SiaNet) and a CycleGAN. Using a contrastive loss, the SiaNet pulls close a translated image and its counter part in the source, and push away the translated image and any image in the target. In this manner, the contrastive loss satisfies the specific requirement in person re-ID. Note that, the added constraint is unsupervised, *i.e.,* the source labels are not used during domain adaptation. During training, in each mini-batch (batch size = 1), a training image is firstly used to update the Generator (of CycleGAN), then the Discriminator (of CycleGAN), and finally the convolutional layers in SiaNet. Through the coordination between CycleGAN loss and the SiaNet loss, we are able to generate samples which not only possess the style of the target domain and but also preserve their underlying ID information.

Using SPGAN, we are able to create a dataset on the target domain in an unsupervised manner. The dataset inherits the labels from the source domain and thus can be used in supervised learning in the target domain. The contributions of this work are summarized below:

- Minor contribution: we present a "learning via translation" baseline for domain adaptation in person re-ID.
- Major contribution: we introduce SPGAN to improve the baseline. SPGAN works by preserving the underlying ID information during image-image translation.

2. Related Work

Image-image translation. Image-image translation aims at constructing a mapping function between two domains. A representative mapping function is the conditional GAN [20], which using paired training data produces impressive image-to-image transition results. However, the paired training data is often difficult to acquire. Unsupervised image-image translation is thus more applicable since data collection is easier. To tackle unpaired settings, a cycle consistency loss is introduced by [22, 46, 53]. In [3], an unsupervised distance loss is proposed for one side domain mapping. In [27], a general framework is proposed by making a shared latent space assumption. Our work aims to find a mapping function between source domain and target domain, and we more concerned with similarity preserving translation.

Neural style transfer [12, 23, 43, 21, 5, 24, 19, 25] is another strategy of image-image translation, which aims at replicating the style of one image, while our work focuses on learning the mapping function between two domains, rather than two images.

Unsupervised domain adaptation. Our work relates to unsupervised domain adaptation (UDA) where no labeled target images are available during training. In this community, some methods aim to learn a mapping between source and target distributions [37, 13, 9, 38]. Correlation Alignment (CORAL) [38] propose to match the mean and covariance of two distributions. Recent methods [18, 4, 28] use an adversarial approach to learn a transformation in the pixel space from one domain to another. Other methods seek to find a domain-invariant feature space [34, 31, 10, 30, 42, 11, 2]. Long *et al*. [30] and Tzeng *et al*. [42] use the Maximum Mean Discrepancy (MMD) [15] for this purpose. Ganin *et al*. [11] and Ajakan *et al*. [2] introduce a domain confusion loss to learn domain-invariant features. Different from the settings in this paper, most of the UDA methods assume that class labels are the same across domains, while different re-ID datasets contain entirely different person identities (classes). Therefore, the approaches mentioned above can not be utilized directly for domain adaptation in re-ID.

Unsupervised re-ID. Hand-craft features [32, 14, 7, 33, 26, 49] can be directly employed for unsupervised re-ID. All these methods focus on feature design, but the rich information from the distribution of samples in the dataset has not been fully exploited. Some methods are based on saliency statistics [48, 44]. In [47], K-means clustering is used for learning a unsupervised asymmetric metric. For unsupervised domain adaptation re-ID, the authors of [35] propose an asymmetric multi-task dictionary learning method to transfer the view invariant representation learned on source data to target data.

Recently, several works focus on label estimation of the unlabeled target dataset. Ye *et al*. [45] use graph matching for cross-camera label estimation. Fan *et al*. [6] propose a progressive method based on the iterations between K-means clustering and IDE [50] fine-tuning. Liu *et al.* [29] employ a reciprocal search process to refine the estimated labels. Our work aims to learn re-ID models that can be utilized directly to target domain, and can potentially cooperate with label estimation methods in model initialization.

3. Proposed Method

3.1. Baseline Overview

Given an annotated dataset $\mathcal{S}$ from source domain and unlabeled dataset $\mathcal{T}$ from target domain, our goal is to use the labeled source images to train a re-ID model that generalizes well to target domain. Figure 2 presents a pipeline of the "learning via translation" framework, which consists of two steps, *i.e.,* source-target image translation for training data creation, and supervised feature learning for re-ID.

- **Source-target image translation.** Using a generative function $G(\cdot)$ that translates the annotated dataset $\mathcal{S}$ from the source domain to target domain in an unsupervised manner, we "create" a labeled training dataset $G(\mathcal{S})$ on the target domain. In this paper, we use CycleGAN [53], following the practice in [27, 18].
- **Feature learning.** With the translated dataset $G(\mathcal{S})$ that contains labels, feature learning methods are applied to train re-ID models. Specifically, we adopt the same setting as [50], in which the rank-1 accuracy and mAP on the fully-supervised Market-1501 dataset is 75.8% and 52.2%.

The focus of this paper is to improve Step 1, so that with better training samples, the overall re-ID accuracy can be improved. The experiment will validate the proposed Step 2 ($G_{sp}(\cdot)$) on several feature learning methods. A brief summary of different methods considered in this paper is presented in Table 1. We denote the method "Direct Transfer" as directly using the training set $\mathcal{S}$ instead of $G(\mathcal{S})$ for model learning. This method yields the lowest accuracy because the style difference between the source and target is not resolved (to be shown in Table 2). Using CycleGAN and SPGAN to generate a new training set, which is more style-consistent with the target, respectively yields improvement.

Method	Train. Set	Test Set	Accuracy
Supervised	$\mathcal{T}_{train}$	$\mathcal{T}_{test}$	+++++
Direct Transfer	$\mathcal{S}_{train}$	$\mathcal{T}_{test}$	++
CycleGAN (basel.)	$G(\mathcal{S}_{train})$	$\mathcal{T}_{test}$	+++
SPGAN	$G_{sp}(\mathcal{S}_{train})$	$\mathcal{T}_{test}$	++++

Table 1: A brief summary of different methods considered in this paper. "G" and "G_{sp}" denote the Generator in CycleGAN and SPGAN, respectively. $\mathcal{S}_{train}, \mathcal{T}_{train}, \mathcal{T}_{test}$ denote the training set of the source dataset, the training set and testing set of the target dataset, respectively.

3.2. SPGAN: Approach Details

3.2.1 CycleGAN Revisit

CycleGAN introduces two generator-discriminator pairs, $\{G, D_{\mathcal{T}}\}$ and $\{F, D_{\mathcal{S}}\}$, which map a sample from source (target) domain to target (source) domain and produce a sample which is indistinguishable from those in the target (source) domain, respectively. For generator G and its associated discriminator $D_{\mathcal{T}}$, the adversarial loss is,

$$\begin{aligned}\mathcal{L}_{\mathcal{T}adv}(G, D_{\mathcal{T}}, p_x, p_y) =& \mathbb{E}_{y\sim p_y}[(D_{\mathcal{T}}(y) - 1)^2] \\ &+ \mathbb{E}_{x\sim p_x}[(D_{\mathcal{T}}(G(x))^2],\end{aligned} \tag{1}$$

where p_x and p_y denote the sample distributions in the source and target domain, respectively. For generator F and its associated discriminator $D_{\mathcal{S}}$, the adversarial loss is,

$$\begin{aligned}\mathcal{L}_{\mathcal{S}adv}(F, D_{\mathcal{S}}, p_y, p_x) =& \mathbb{E}_{x\sim p_x}[(D_{\mathcal{S}}(x) - 1)^2] \\ &+ \mathbb{E}_{y\sim p_y}[(D_{\mathcal{S}}(F(y))^2].\end{aligned} \tag{2}$$

Considering there exist infinitely many alternative mapping functions due to the lack of paired training data, CycleGAN introduces a cycle-consistent loss, which attempts to recover the original image after a cycle of translation and reverse translation, to reduce the space of possible mapping functions. The cycle-consistent loss is,

$$\begin{aligned}\mathcal{L}_{cyc}(G, F) =& \mathbb{E}_{x\sim p_x}[\|F(G(x)) - x\|_1] \\ &+ \mathbb{E}_{y\sim p_y}[\|G(F(y)) - y\|_1].\end{aligned} \tag{3}$$

Apart from the cycle-consistent loss, adversarial loss, we use the target domain identity constraint as an auxiliary for image-image translation. Target domain identity constraint was introduced by [40] to regularize the generator to be the identity matrix on samples from target domain, written as,

$$\begin{aligned}\mathcal{L}_{ide}(G, F, p_x, p_y) =& \mathbb{E}_{x\sim p_x}\|F(x) - x\|_1 \\ &+ \mathbb{E}_{y\sim p_y}\|G(y) - y\|_1.\end{aligned} \tag{4}$$

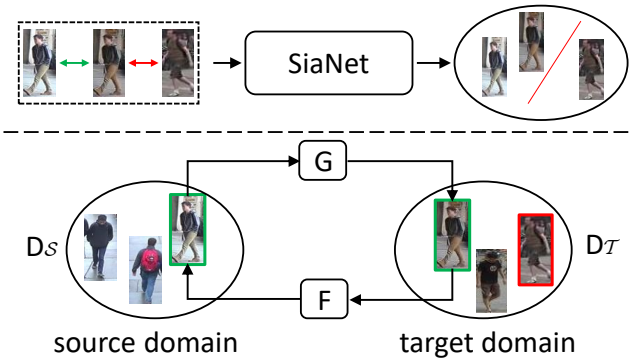

Figure 3: SPGAN consists of two components: a SiaNet (top) and CycleGAN (bottom). CycleGAN learns mapping functions G and F between two domains, and the SiaNet learns a latent space that constrains the learning procedure of mapping functions.

As mentioned [53], generators G and F may change the color of output images without L_{ide}. In experiment, we observe that the model may generate unreal results without L_{ide} (Fig. 4(b)). This is undesirable for re-ID feature learning. By turning on the identity loss, the color of the input and output can be preserved (see Section 4.3).

3.2.2 SPGAN

Applied in person re-ID, similarity preserving is an essential function to generate improved samples for domain adaptation. As analyzed in Section 1, we aim to preserve the ID-related information for each translated image. We emphasize that such information should not be the background or image style, but should be underlying and latent. To fulfill this goal, we integrate a SiaNet with CycleGAN, as shown in Fig 3. During training, CyleGAN is to learn a mapping function between two domains, and SiaNet is to learn a latent space that constrains the learning procedure of mapping function.

Similarity preserving loss function. We utilize the contrastive loss [16] to train SiaNet,

$$\mathcal{L}_{con}(i, x_1, x_2) = (1-i)\{\max(0, m-d)\}^2 + id^2, \quad (5)$$

where x_1 and x_2 are a pair of input vectors, and i represents the binary label assigned to this pair. $i = 1$ if x_1 and x_2 are positive pair; $i = 0$ if x_1 and x_2 are negative pair. m is the margin that defines the separability in the embedding space. When $m = 0$, the loss of the negative training pair is not back-propagated in the system. When $m > 0$, both positive and negative sample pairs are considered. A larger m means that the loss of negative training samples has a higher weight in back propagation. d denotes the Euclidean distance between two input vectors: $d(x_1, x_2) = \|x_1 - x_2\|_2$.

Training image pair selection. In Eq. 5, the contrastive loss uses binary labels of input image pairs. The design of the pair similarities reflects the "self-similarity" and "domain-dissimilarity" principles. Note that, *we select training pairs in an unsupervised manner*, so that we use the contrastive loss without additional annotations.

Formally, CycleGAN has two generators, *i.e.*, generator G which maps source-domain images to the style of the target domain, and generator F which maps target-domain images to the style of the source domain. Suppose two samples denoted as $x_{\mathcal{S}}$ and $x_{\mathcal{T}}$ come from the source domain and target domain, respectively. Given G and F, we define two positive pairs: 1) $x_{\mathcal{S}}$ and $G(x_{\mathcal{S}})$, 2) $x_{\mathcal{T}}$ and $F(x_{\mathcal{T}})$. In either image pair, the two images contain the same person; the only difference is that they have different styles. In the learning procedure, we encourage the whole network to pull these two images close.

On the other hand, for generator G and F, we also define two types of negative training pairs: 1) $G(x_{\mathcal{S}})$ and $x_{\mathcal{T}}$, 2) $F(x_{\mathcal{T}})$ and $x_{\mathcal{S}}$. Such design of negative training pairs is based on the prior knowledge that datasets in different re-ID domains have entirely different sets of IDs. As a result, a translated image should be of different ID from any target image. In this manner, the network pushes two dissimilar images away. Training pairs are shown in Fig. 1. Some positive pairs are also shown in (a) and (d) of each column in Fig. 4.

Overall objective function. The final SPGAN objective can be written as,

$$\mathcal{L}_{sp} = \mathcal{L}_{\mathcal{T}adv} + \mathcal{L}_{\mathcal{S}adv} + \lambda_1 \mathcal{L}_{cyc} + \lambda_2 \mathcal{L}_{ide} + \lambda_3 \mathcal{L}_{con}, \quad (6)$$

where $\lambda_t, t \in \{1, 2, 3\}$ controls the relative importance of four objectives. The first three losses belong to the CycleGAN formulation [53], and the contrastive loss induced by SiaNet imposes a new constraint on the system.

SPGAN training procedure. In the training phase, SPGAN are divided into three components which are learned alternately, the generators, discriminators, and SiaNet. When the parameters of two components are fixed, the parameters of the third component is updated. We train the SPGAN until the convergence or the maximum iterations.

3.3. Feature Learning

Feature learning is the second step of the "learning via translation" framework. Once we have the style-transferred dataset $G(\mathcal{S})$ composed of the translated images and their associated labels, the feature learning step is the same as supervised methods. Since we mainly focus on Step 1 (source-target image translation), we adopt the baseline ID-discriminative Embedding (IDE) specified in [50]. We employ ResNet-50 [17] as the base model and only modify the output dimension of the last fully-connected layer to the

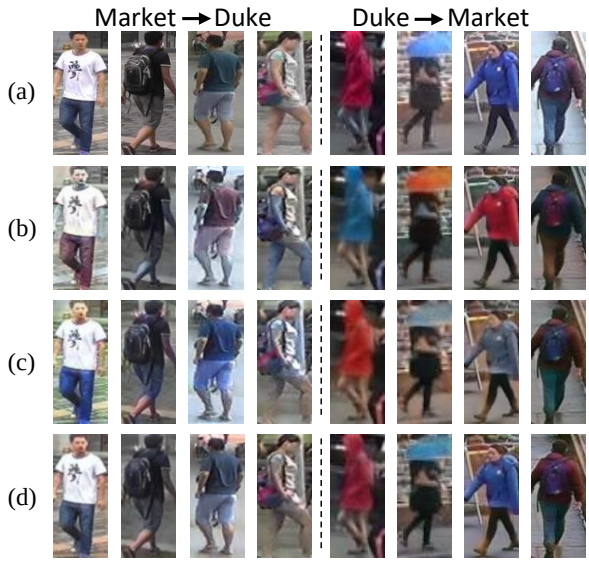

Figure 4: Visual examples of image-image translation. The left four columns map Market images to the Duke style, and the right four columns map Duke images to the Market style. From top to bottom: (a) original image, (b) output of CycleGAN, (c) output of CycleGAN + L_{ide}, and (d) output of SPGAN. Images produced by SPGAN have the target style while preserving the ID information in the source.

number of training identities. During testing, given an input image, we can extract the 2,048-dim Pool5 vector for retrieval under the Euclidean distance.

Local Max Pooling. To further improve re-ID performance on the target dataset $\mathcal{T}$, we introduce a feature pooling method named as local max pooling (LMP). It works on a well-trained IDE model and can reduce the impact of noisy signals incurred by the fake examples. In the original ResNet-50, global average pooling (GAP) is conducted on Conv5. In our proposal (Fig. 5), we first partition the Conv5 feature maps to P parts horizontally with one pixel overlap. Then, we conduct global max/avg pooling on each part. Finally, we concatenate the output of global max pooling (GMP) or GAP of each part as the final feature representation. The procedure is nonparametric, and can be directly used in the testing phase. In the experiment, we will compare local max pooling and local average pooling, and demonstrate the superiority of the former (LMP).

4. Experiment

4.1. Datasets

We select two large-scale re-ID datasets for experiment, *i.e.*, **Market-1501** [49] and **DukeMTMC-reID** [36, 51]. Market-1501 is composed of 1,501 identities, 12,936 training images and 19,732 gallery images (with 2,793 distractors). It is split into 751 identities for training and 750 identities for testing. Each identity is captured by at most 6 cameras. All the bounding boxes are produced by DPM [8]. DukeMTMC-reID is a re-ID version of the DukeMTMC dataset [36]. It contains 34,183 image boxes of 1,404 identities: 702 identities are used for training and the remaining 702 for testing. There are 2,228 queries and 17,661 database images. For both dataset, we rank-1 accuracy and mAP for re-ID evaluation [49]. Sample images of the two datasets are shown in Fig. 6.

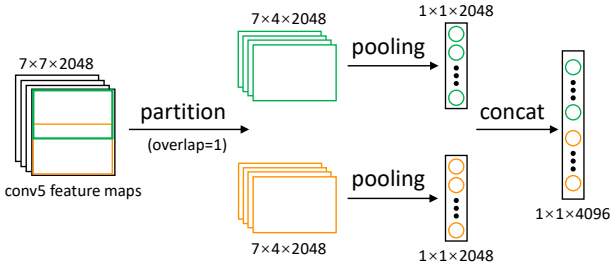

Figure 5: Illustration of LMP. We partition the feature map into $P(P = 2)$ parts horizontally with one pixel overlap. We conduct global max/avg pooling on each part and concatenate the feature vectors as the final representation.

Figure 6: Sample images of (upper left:) DukeMTMC-reID dataset, (lower left:) Market-1501 dataset, (upper right:) Duke images which are translated to Market style, and (lower right:) Market images translated to Duke style. We use SPGAN for image-image translation.

4.2. Implementation Details

SPGAN training and testing. We use Tensorflow [1] to train SPGAN using the training images of Market-1501 and DukeMTMC-reID. Note that, we do not use any ID annotation during the learning procedure. In all experiment, we empirically set $\lambda_1 = 10, \lambda_2 = 5, \lambda_3 = 2$ in Eq. 6

Methods	DukeMTMC-reID					Market-1501				
	rank-1	rank-5	rank-10	rank-20	mAP	rank-1	rank-5	rank-10	rank-20	mAP
Supervised Learning	66.7	79.1	83.8	88.7	46.3	75.8	89.6	92.8	95.4	52.2
Direct Transfer	33.1	49.3	55.6	61.9	16.7	43.1	60.8	68.1	74.7	17.0
CycleGAN (basel.)	38.1	54.4	60.5	65.9	19.6	45.6	63.8	71.3	77.8	19.1
CycleGAN (basel.) + L_{ide}	38.5	54.6	60.8	66.6	19.9	48.1	66.2	72.7	80.1	20.7
SPGAN ($m = 0$)	37.7	53.1	59.5	65.6	20.0	49.2	66.9	74.0	80.0	20.5
SPGAN ($m = 1$)	39.5	55.0	61.4	67.3	21.0	48.7	65.7	73.0	79.3	21.0
SPGAN ($m = 2$)	41.1	56.6	63.0	69.6	22.3	51.5	70.1	76.8	82.4	22.8
SPGAN ($m = 2$) + LMP	**46.4**	**62.3**	**68.0**	**73.8**	**26.2**	**57.7**	**75.8**	**82.4**	**87.6**	**26.7**

Table 2: Comparison of various methods on the target domains. When tested on DukeMTMC-reID, Market-1501 is used as source, and vice versa. "Supervised learning" denotes using the full ID labels on the corresponding target dataset. "Direct Transfer" means directly applying the source-trained model on the target domain (see Section 3.1). By varying m specified in Eq. 5, the sensitivity of SPGAN to the relative importance of the positive and negative pairs is shown. When local max pooling (LMP) is applied, the number of parts is set to 6. We use IDE [50] for feature learning.

and $m = 2$ in Eq. 5. With an initial learning rate 0.0002, and model stop training after 5 epochs. During the testing procedure, we employ the Generator G for Market-1501 $\rightarrow$ DukeMTMC-reID translation and the Generative F for DukeMTMC-reID $\rightarrow$ Market-1501 translation. The translated images are used for training re-ID models.

For CycleGAN, we adopt the architecture released by its authors. For SiaNet, it contains 4 convolutional layers, 4 max pooling layers and 1 fully connected (FC) layer, configured as below. (1) Conv. 4×4, stride = 2, #feature maps = 64; (2) Max pooling 2×2, stride = 2; (3) Conv. 4×4, stride = 2, #feature maps = 128; (4) Max pooling 2×2, stride = 2; (5) Conv. 4×4, stride = 2, feature maps = 256; (6) Max pool 2×2, stride = 2; (7) Conv. 4×4, stride = 2, #feature maps = 512; (8) Max pooling 2×2, stride = 2; (9) FC, output dimension = 128.

Feature learning for re-ID. Following [50], we train a classification network for re-ID embedding learning, named ID-discriminative Embedding (IDE). Specifically, ResNet-50 [17] pretrained on ImageNet is used for fine-tuning on the translated training set. We modify the output of the last fully-connected layer to 751 and 702 for Market-1501 and DukeMTMC-reID, respectively. We use mini-batch stochastic gradient descent to train the CNN model on a Tesla K80 GPU. Training parameters such as batch size, maximum number epochs, momentum and gamma are set to 16, 50, 0.9 and 0.1, respectively. The initial learning rate is set as 0.001, and decay to 0.0001 after 40 epochs.

4.3. Evaluation

Comparison between supervised learning and direct transfer. The supervised learning method and the direct transfer method are specified in Table 1. When comparing the two methods in Table 2, we can clearly observe a large performance drop when directly using a source-trained model on the target domain. For instance, the ResNet-50 model trained and tested on Market-1501 achieves 75.8% in rank-1 accuracy, but drops to 43.1% when trained on DukeMTMC-reID and tested on Market-1501. A similar drop can be observed when DukeMTMC-reID is used as the target domain, which is consistent with the experiments reported in [6]. The reason behind the performance drop is the bias of data distributions in different domains.

The effectiveness of the "learning via translation" baseline using CycleGAN. In this baseline domain adaptation approach (Section 3.1), we first translate the label images from the source domain to the target domain and then use the translated images to train re-ID models. As shown in Table 2, this baseline framework effectively improves the re-ID performance in the target dataset. Compared with the direct transfer method, the CycleGAN transfer baseline gains $+2.5\%$ and $+2.1\%$ improvements in rank-1 accuracy and mAP on Market-1501. When tested on DukeMTMC-reID, the performance gain is +5.0% and +2.9% in rank-1 accuracy and mAP, respectively. Through such an image-level domain adaptation method, effective domain adaptation baselines can be learned.

The impact of the target domain identity constraint. We conduct experiment to verify the influence of the identity loss on performance in Table 2. We arrive at mixed observations. On the one hand, on DukeMTMC-reID, compared with the CycleGAN baseline, CycleGAN + L_{ide} achieves similar rank-1 accuracy and mAP. On the other hand, on Market-1501, CycleGAN + L_{ide} gains $+2.5\%$ and 1.6% improvement in rank-1 accuracy and mAP, respectively. The reason is that Market-1501 has a larger inter-camera variance. When translating Duke images to the Market style, the translated image may be more prone to translation errors induced by the camera variances. Therefore, the identity loss is more effective when Market is the target

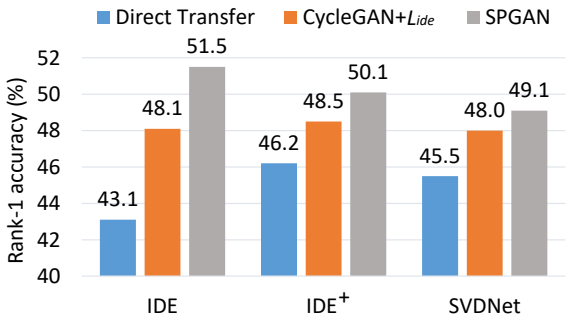

Figure 7: Domain adaptation performance with different feature learning methods, including IDE (Section 3.3), IDE$^+$ [52], and SVDNet [39]. Three domain adaptation methods are compared, *i.e.*, direct transfer, CycleGAN with identity loss, and the proposed SPGAN. The results are on Market-1501.

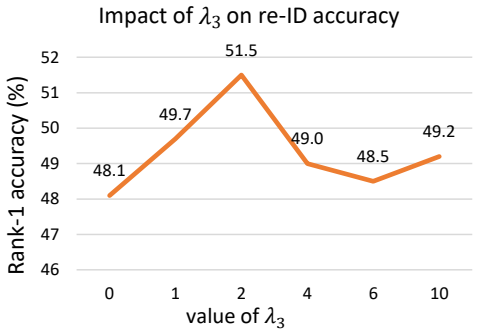

Figure 8: λ_3 (Eq. 6) *v.s* re-ID accuracy. A larger λ_3 means larger weight of similarity preserving constraint.

domain. Considering that the performance never drops, we use the target domain identity constraint as an auxiliary tool for image-image translation. As shown in Fig. 4, this loss helps CycleGAN prevent from generating strangely colored results.

The effectiveness of the proposed SPGAN. On top of the CycleGAN baseline, we replace CycleGAN with SPGAN ($m = 2$). The effectiveness of the proposed similarity preserving constraint can be seen in Table 2. Compared with Cycle + L_{ide}, on DukeMTMC-reID, the similarity preserving constraint leads to $+2.6\%$ and $+2.4\%$ improvement over CycleGAN + L_{ide} in rank-1 accuracy and mAP, respectively. On Market-1501, the gains are $+3.4\%$ and $+2.1\%$ in rank-1 accuracy and mAP, respectively. The working mechanism of SPGAN consists in preserving the underlying visual cues associated with the ID labels. The consistent improvement suggests that this working mechanism is critical for generating suitable samples for training in the target domain. Examples of translated images by SPGAN are shown in Fig. 6.

Comparison of different feature learning methods. In Step 2, we evaluate three feature learning methods, *i.e.*, IDE [50] (described in Section 3.3), IDE* [52], and SVDNet [39], as shown in Fig. 7. An interesting observation is that, while IDE$^+$ and SVDNet are superior to IDE under the scenario of "Direct Transfer", the three learning methods are basically on par with each other when using training samples generated by SPGAN. A possible explanation is that many of the generated samples are imperfect, which has a larger effect on those better learning schemes.

Sensitivity of SPGAN to key parameters. The margin m defined in Eq. 5 is a key parameter. If $m = 0$, the loss of negative pairs is not back propagated. If m gets larger, the weight of negative pairs in loss calculation increases. We conduce experiment to verify the impact of m, and results are shown in Table 2. When turning off the contribution of negative pairs in Eq. 5, ($m = 0$), SPGAN only marginally improve the accuracy on Market-1501, and even compromises the system on Duke. When increasing m to 2, we have much superior accuracy. It indicates that the negative pairs are critical to the system.

Moreover, we evaluate the impact of λ_3 in Eq. 6 on Market-1501. λ_3 controls the relative importance of the proposed similarity preserving constraint. As shown in Fig. 9, the proposed constraint is proven effective when compared to $\lambda_3 = 0$, but a larger λ_3 does not bring more gains in accuracy. Specifically, $\lambda_3 = 2$ yields the best accuracy.

Local max pooling further improves the transfer performance. We apply the LMP on the Conv5 layer to mitigate the influence of noise. Note that LMP is directly adopted in the feature extraction step for testing without fine-tuning. We empirically study how the number of parts and the pooling mode affect the performance. Experiment is conducted on SPGAN. The performance of various numbers of parts ($P = 1, 2, 3, 6$) and different pooling modes (max or average) is provided in Table 3. When we use average pooling and $P = 1$, we have the original GAP used in ResNet-50. From these results, we speculate that with more parts, a finer partition leads to higher discriminative descriptors and thus higher re-ID accuracy.

Moreover, we test LMP on supervised learning and domain adaptation scenarios with three feature learning methods, *i.e.*, IDE [50], IDE$^+$ [52], and SVDNet [39]. As shown in Fig. 9, LMP does not guarantee stable improvement on supervised learning as observed in "IDE$^+$" and SVDNet. However, when applied in the scenario of domain adaptation with SPGAN, LMP yields improvement over IDE, IDE$^+$, and SVDNet. The superiority of LMP probably lies in that max pooling filters out some noisy signals in the descriptor induced by SPGAN.

4.4. Comparison with State-of-the-art Methods

We compare the proposed method with the state-of-the-art unsupervised learning methods on Market-1501 and DukeMTMC-reID in Table 4 and Table 5, respectively.

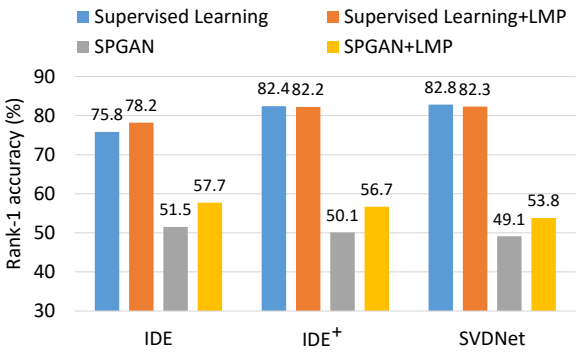

Figure 9: Experiment of LMP ($P = 6$) on scenarios of supervised learning and SPGAN. Three feature learning methods are compared, *i.e.*, IDE [50], IDE$^+$ [52], and SVDNet [39]. The results are on Market-1501.

#parts	mode	dim	DukeMTMC-reID		Market-1501	
			rank-1	mAP	rank-1	mAP
1	Avg	2048	41.1	22.3	51.5	22.8
	Max		44.3	25.0	55.7	21.8
2	Avg	4096	42.3	23.3	54.4	25.0
	Max		45.6	25.5	57.3	26.2
3	Avg	6144	43.1	23.6	54.9	25.5
	Max		45.5	25.6	57.4	26.4
6	Avg	12288	44.1	24.4	55.9	26.0
	Max		**46.4**	**26.2**	**57.7**	**26.7**

Table 3: Performance of various pooling strategies with different numbers of parts (P) and pooling modes (maximum or average) over SPGAN. The best results are in bold.

Market-1501. On Market-1501, we first compare our results with two hand-crafted features, *i.e.*, Bag-of-Words (BoW) [49] and local maximal occurrence (LOMO) [26]. 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) [47], the Progressive Unsupervised Learning (PUL) [6], and UMDL [35]. The results of UMDL are reproduced by Fan *et al.* [6]. In the single-query setting, we achieve rank-1 accuracy = 51.5% and mAP = 22.8%. It outperforms the second best method [6] by +6.0% in rank-1 accuracy. In the multiple-query setting, we arrive at rank-1 accuracy = 57.0% and mAP = 27.1%, which is +2.5% higher than CAMEL [47]. The comparisons indicate the competitiveness of the proposed method on Market-1501.

DukeMTMC-reID. On DukeMTMC-reID, we compare the proposed method with BoW [49], LOMO [26], UMDL [35], and PUL [6] under the single-query setting (there is no multiple-query setting in DukeMTMC-reID). The result obtained by the proposed method is rank-1 accuracy = 41.1%, mAP = 22.3%. Compared with the second best method, *i.e.,* PUL [6], our result is +11.1% higher in rank-1 accuracy. Therefore, the superiority of SPGAN can be concluded.

Methods	Market-1501				
	Setting	Rank-1	Rank-5	Rank-10	mAP
Bow [49]	SQ	35.8	52.4	60.3	14.8
LOMO [26]	SQ	27.2	41.6	49.1	8.0
UMDL [35]	SQ	34.5	52.6	59.6	12.4
PUL [6]*	SQ	45.5	60.7	66.7	20.5
Direct transfer	SQ	43.1	60.8	68.1	17.0
Direct transfer	MQ	47.9	65.5	73.0	20.6
CAMEL [47]	MQ	54.5	-	-	26.3
SPGAN	SQ	51.5	70.1	76.8	22.8
SPGAN	MQ	57.0	73.9	80.3	27.1
SPGAN+LMP	SQ	**57.7**	**75.8**	**82.4**	**26.7**

Table 4: Comparison with state of the art on Market-1501. * denotes unpublished papers. "SQ" and "MQ" are the single-query and multiple-query settings, respectively.

Methods	DukeMTMC-reID			
	Rank-1	Rank-5	Rank-10	mAP
Bow[49]	17.1	28.8	34.9	8.3
LOMO[26]	12.3	21.3	26.6	4.8
UMDL[35]	18.5	31.4	37.6	7.3
PUL[6]*	30.0	43.4	48.5	16.4
Direct transfer	33.1	49.3	55.6	16.7
SPGAN	41.1	56.6	63.0	22.3
SPGAN+LMP	**46.4**	**62.3**	**68.0**	**26.2**

Table 5: Comparison with state of the art on DukeMTMC-reID. * denotes unpublished papers.

5. Conclusion

This paper focuses on domain adaptation in person re-ID. When models trained on one dataset are directly transferred to another dataset, the re-ID accuracy drops dramatically due to dataset bias. To achieve improved performance in the new dataset, we present a "learning via translation" baseline for domain adaptation, characterized by 1) unsupervised image-image translation and 2) supervised feature learning. We further propose that the underlying (latent) ID information for the foreground pedestrian should be preserved after image-image translation. To meet this requirement tailored for re-ID, we introduce the unsupervised self-similarity and domain-dissimilarity for similarity preserving image generation (SPGAN). We show that SPGAN better qualifies the generated images for domain adaptation and yields consistent improvement over the CycleGAN.

References

[1] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. A. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, and X. Zheng. Tensorflow: A system for large-scale machine learning. In *OSDI*, 2016. 5

[2] H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, and M. Marchand. Domain-adversarial neural networks. *arXiv preprint arXiv:1412.4446*, 2014. 2

[3] S. Benaim and L. Wolf. One-sided unsupervised domain mapping. In *NIPS*, 2017. 2

[4] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *CVPR*, 2017. 1, 2

[5] T. Q. Chen and M. Schmidt. Fast patch-based style transfer of arbitrary style. *arXiv preprint arXiv:1612.04337*, 2016. 2

[6] H. Fan, L. Zheng, and Y. Yang. Unsupervised person re-identification: Clustering and fine-tuning. *arXiv preprint arXiv:1705.10444*, 2017. 1, 3, 6, 8

[7] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*, 2010. 2

[8] P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *CVPR*, 2008. 5

[9] B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars. Unsupervised visual domain adaptation using subspace alignment. In *ICCV*, 2013. 2

[10] Y. Ganin and V. S. Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, 2015. 2

[11] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. S. Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 2016. 2

[12] L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. In *CVPR*, 2016. 2

[13] B. Gong, Y. Shi, F. Sha, and K. Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *CVPR*, 2012. 2

[14] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV*, 2008. 2

[15] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. J. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 2012. 2

[16] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *CVPR*, 2006. 4

[17] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 4, 6

[18] J. Hoffman, E. Tzeng, T. Park, and J.-Y. Zhu. Cycada: Cycle-consistent adversarial domain adaptation. *arXiv preprint arXiv:1711.03213*, 2017. 1, 2, 3

[19] X. Huang and S. J. Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, 2017. 2

[20] P. Isola, J. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 2

[21] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016. 2

[22] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim. Learning to discover cross-domain relations with generative adversarial networks. In *ICML*, 2017. 1, 2

[23] C. Li and M. Wand. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *ECCV*, 2016. 2

[24] Y. Li, C. Fang, J. Yang, Z. Wang, X. Lu, and M. Yang. Diversified texture synthesis with feed-forward networks. In *CVPR*, 2017. 2

[25] Y. Li, N. Wang, J. Liu, and X. Hou. Demystifying neural style transfer. In *IJCAI*, 2017. 2

[26] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, 2015. 2, 8

[27] M. Liu, T. Breuel, and J. Kautz. Unsupervised image-to-image translation networks. In *NIPS*, 2017. 1, 2, 3

[28] M. Liu and O. Tuzel. Coupled generative adversarial networks. In *NIPS*, 2016. 1, 2

[29] Z. Liu, D. Wang, and H. Lu. Stepwise metric promotion for unsupervised video person re-identification. In *ICCV*, 2017. 3

[30] M. Long, Y. Cao, J. Wang, and M. I. Jordan. Learning transferable features with deep adaptation networks. In *ICML*, 2015. 2

[31] M. Long, G. Ding, J. Wang, J. Sun, Y. Guo, and P. S. Yu. Transfer sparse coding for robust image representation. In *CVPR*, 2013. 2

[32] B. Ma, Y. Su, and F. Jurie. Covariance descriptor based on bio-inspired features for person re-identification and face verification. *Image Vision Comput.*, 2014. 2

[33] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato. Hierarchical gaussian descriptor for person re-identification. In *CVPR*, 2016. 2

[34] S. Motiian, M. Piccirilli, D. A. Adjeroh, and G. Doretto. Unified deep supervised domain adaptation and generalization. In *ICCV*, 2017. 2

[35] P. Peng, T. Xiang, Y. Wang, M. Pontil, S. Gong, T. Huang, and Y. Tian. Unsupervised cross-dataset transfer learning for person re-identification. In *CVPR*, 2016. 3, 8

[36] 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. 5

[37] K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting visual category models to new domains. In *ECCV*, 2010. 2

[38] B. Sun, J. Feng, and K. Saenko. Return of frustratingly easy domain adaptation. In *AAAI*, 2016. 2

[39] Y. Sun, L. Zheng, W. Deng, and S. Wang. SVDNet for pedestrian retrieval. In *ICCV*, 2017. 7, 8

[40] Y. Taigman, A. Polyak, and L. Wolf. Unsupervised cross-domain image generation. *ICLR*, 2016. 3

[41] A. Torralba and A. A. Efros. Unbiased look at dataset bias. In *CVPR*, 2011. 1

[42] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014. 2

[43] D. Ulyanov, V. Lebedev, A. Vedaldi, and V. S. Lempitsky. Texture networks: Feed-forward synthesis of textures and stylized images. In *ICML*, 2016. 2

[44] H. Wang, S. Gong, and T. Xiang. Unsupervised learning of generative topic saliency for person re-identification. In *BMVC*, 2014. 3

[45] M. Ye, A. J. Ma, L. Zheng, J. Li, and P. C. Yuen. Dynamic label graph matching for unsupervised video re-identification. In *ICCV*, 2017. 3

[46] Z. Yi, H. Zhang, P. Tan, and M. Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *ICCV*, 2017. 1, 2

[47] H. Yu, A. Wu, and W. Zheng. Cross-view asymmetric metric learning for unsupervised person re-identification. In *ICCV*, 2017. 3, 8

[48] R. Zhao, W. Ouyang, and X. Wang. Unsupervised salience learning for person re-identification. In *CVPR*, 2013. 3

[49] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015. 2, 5, 8

[50] L. Zheng, Y. Yang, and A. G. Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016. 3, 4, 6, 7, 8

[51] Z. Zheng, L. Zheng, and Y. Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *ICCV*, 2017. 5

[52] Z. Zhong, L. Zheng, D. Cao, and S. Li. Re-ranking person re-identification with k-reciprocal encoding. 2017. 7, 8

[53] J. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *ICCV*, 2017. 1, 2, 3, 4