

Similarity-preserving Image-image Domain Adaptation for Person Re-identification

Weijian Deng, Liang Zheng, Qixiang Ye, Yi Yang and Jianbin Jiao

Abstract—Person re-identification (re-ID) models often fail to generalize well to new domains. We propose a “learning via translation framework based on the Generative Adversarial Network (GAN). It consists of two components, *i.e.*, 1) translating the labeled source images to style of the target domain, and 2) learning a re-ID model for testing on the target domain using the translated images. Typically, source-target translation suffers from information loss with respect to the discriminative cues that form human identity. To this end, we propose a similarity-preserving generative adversarial network (SPGAN) and its upgraded version, end-to-end SPGAN (eSPGAN). SPGAN improves the first component of the framework. It enforces two heuristic constraints in an unsupervised manner, 1) preserving self-similarity of human identity, and 2) introducing domain dissimilarity, such that the source images preserve the discriminative cues while being transferred to the target style. In comparison, eSPGAN seamlessly integrates the two components of the framework. During its end-to-end training, feature learning guides image translation to preserve the underlying identity information of an image. Meanwhile, image translation improves feature learning by providing identity-preserving training samples of the target domain style. Experiment on two large-scale datasets shows that both SPGAN and eSPGAN obtain state-of-the-art domain adaptation results.

Index Terms—Person Re-Identification, Domain Adaptation, Learning via Translation



1 INTRODUCTION

THIS article studies the domain adaptation problem in person re-ID under a “learning via translation” framework. In our setting, the source domain is fully annotated with identity labels, and the target domain does not have any ID labels. In the community, domain adaptation of re-ID is gaining increasing popularity, because of 1) the expensive labeling process and 2) when models trained on one dataset are directly used on another, the re-ID accuracy drops dramatically [1] due to dataset bias [2].

A commonly used strategy to above-mentioned problems is unsupervised domain adaptation (UDA). But this line of methods usually assumes that the source and target domains contain the same set of classes. This assumption does not hold in person re-ID because different re-ID datasets usually contain entirely different persons (classes). In UDA, a recent trend is image-level domain translation [3], [4], [5], which motivates us to explore a “learning via translation” framework. The framework consists of two components. First, labeled images from the source domain are translated to the target domain, so the translated images and images from the target domain share similar styles, *e.g.*, backgrounds, resolutions, and light conditions. Second, the style-translated images and their associated labels are used for supervised learning in the target domain. In literature, commonly used image-level translation methods include [6], [7], [8], [9]. In

our work, we adopt CycleGAN in the baseline.

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 during image-image translation. In our scenario, such visual content usually refers to the underlying (latent) ID information for a foreground pedestrian. To meet this requirement tailored for re-ID, we first propose a heuristic solution, named Similarity Preserving Generative Adversarial Network (SPGAN). Then, we further study the relation between feature learning and image translation, and propose eSPGAN, an upgrade version of SPGAN.

SPGAN is motivated by 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. SPGAN is composed of an Siamese network (SiaNet) and a CycleGAN. Using a contrastive loss, the SiaNet pulls close a translated image and its counterpart 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 re-ID. Note that, the added constraints are unsupervised, *i.e.*, the source labels are not used in source-target image translation. Through the coordination between CycleGAN and SiaNet, we are able to generate samples which not only possess the style of target domain but also preserve their underlying ID information from the source domain.

Essentially, SPGAN focuses only on improving the first component of the “learning via translation” framework, *i.e.*, source-target image translation, which is actually independent of the feature learning component. Thus, the impact of image translation on feature learning and the reverse

-
- W. Deng and L. Zheng are with the Research School of Computer Science, Australian National University, CBR, Australia.
E-mail: dengwj16@gmail.com, liangzheng06@gmail.com
 - Y. Yang is with Centre for Artificial Intelligence, University of Technology Sydney, NSW, Australia.
E-mail: yi.yang@uts.edu.au
 - Q. Ye, and J. Jiao are with the University of Chinese Academy of Sciences, Beijing, China.
E-mail: qxeye@ucas.ac.cn, jiaojb@ucas.ac.cn.
 - Corresponding authors: J. Jiao and L. Zheng.



Fig. 1. Pipeline of the “learning via translation” framework. First, we translate the labeled images from a source domain to a target domain. Second, we train re-ID models with the translated images using supervised feature learning methods. SPGAN is only used to improve the first component of the framework, while eSPGAN simultaneously learns the two components in an end-to-end-manner.

remains unknown. A natural question then arises: can these two components be jointly optimized, so that they could benefit each other?

In light of this question, we propose eSPGAN by seamlessly integrating the two components into an end-to-end training system. In eSPGAN, there exists a mutually beneficial interactive loop between image translation and re-ID feature learning. Thus, the translated images are better suited for the re-ID task, leading to higher re-ID accuracy. More specifically, feature learning guides image translation to preserve the identity of images during translation; in return, image translation delivers the knowledge of how a person looks like on the target domain to feature learning. During training, we alternately optimize the two components, so that knowledge and constraint of both components are gradually transferred to each other.

This paper extends our previous conference paper [10] in several aspects. Primarily, we integrate the two components of “learning via translation framework into an end-to-end system, yielding eSPGAN. In eSPGAN, we discover the mutually benefit between image translation and re-ID feature learning. In addition, insightful analyses of the visual changes conducted by the image translation are provided. Also, the difference from other similarity-preserving generation methods is discussed. Finally, we present significant extensions in the experiment to validate the effectiveness of our methods: 1) we report higher results of baseline methods and SPGAN with our latest implementations; 2) we extensively investigate eSPGAN.

Overall, the contributions of this study are mainly in the following four aspects:

- To address the domain adaptation in person re-ID, we present a “learning via translation framework. We further introduce SPGAN, a heuristic method, to preserve the underlying ID information during source-target image translation. SPGAN better qualifies the translated images and produces competitive domain adaptation accuracy.
- We report the mutual benefit between generative image translation and discriminative feature learning. Inspired by this, we propose eSPGAN, an upgraded version of SPGAN, by simultaneously optimizing image translation and feature learning for the domain adaptive person re-ID. In eSPGAN, there exists a beneficial interactive loop between image translation and re-ID feature learning. Thus, the translated images are better suited for re-ID feature learning, leading to higher re-ID accuracy.

- We provided insightful analyses of the “style change introduced by image translation. We find that “style change involves various factors, such as illumination and color composition. This helps us take a closer look at the “style transfer and gives a better understanding of the dataset bias.
- As a minor contribution, we propose a local max pooling (LMP) scheme as a post-processing step. LMP is tailored for the domain adaptation scenario, and consistently improves over SPGAN and eSPGAN.

The remainder of this paper is organized as follows. Related work is presented in Section 2. Section 3 describes SPGAN and eSPGAN. In Section 4, the experimental results are presented and analyzed. Section 5 concludes the paper.

2 RELATED WORK

Image-image translation. Image-image translation aims at learning a mapping function between two domains. As a representative image-image translation method, the “pixel2pixel” framework uses input-output pairs for learning a mapping from input to output images. In practice, the paired training data is often difficult to acquire and hence the unpaired image-image translation is often more applicable. To tackle the unpaired setting, a cycle consistency loss is introduced by DiscoGAN [6], DualGAN [7], and CycleGAN [8]. Benaim *et al.* [11] propose an unsupervised distance loss for one side domain mapping. Liu *et al.* [12] propose a general framework by making a shared latent space assumption that the corresponding images in two domains are mapped to the same latent code. Recently, some methods [9], [13] have been proposed to learn the relations among multiple domains. In this work, while we aim to find mapping functions between the source domain and target domain, our primary focus is similarity-preserving mapping.

Neural style transfer [14], [15], [16], [17], [18], [19], [20] is another strategy of image-image translation, which aims at rendering the content of an image in the style of another image. Gatys *et al.* [21] employ an optimization process to match feature statistics in layers of a convolutional network. The optimization is replaced by a feed-forward neural network in [14], [15], [16]. Huang *et al.* [19] propose a AdaIN layer for arbitrary style transfer. Unlike the neural style transfer, our work focuses on learning the mapping function between two domains, rather than two images.

Unsupervised domain adaptation. Our work is related to unsupervised domain adaptation (UDA). Within this community, a portion of methods aim to learn a mapping

between source and target distributions [22], [23], [24], [25]. As a representative UDA method, Correlation Alignment (CORAL) [25] matches the mean and covariance of two distributions.

Other methods seek to find a domain-invariant feature space [26], [27], [28], [29], [30], [31], [32]. Long *et al.* [29] use the Maximum Mean Discrepancy (MMD) [33] for this purpose. Ganin *et al.* [31] and Ajakan *et al.* [32] introduce a domain confusion loss to learn domain-invariant features. In addition, several approaches estimate the labels of unlabeled samples [34], [35], [36], [37]. The estimated labels are then used to learn the optimal classifier. Zhang *et al.* [37] propose a progressive method to select a set of pseudo-labeled target samples. Sener *et al.* [36] use the K-nearest neighbors to predict the labels of target samples.

Recent methods [3], [4], [5] use an adversarial approach to learn a transformation in the pixel space from one domain to another. The CYCADA [3] maps samples across domains at both pixel level and feature level. We note that most of the UDA methods assume that class labels are the same across domains. However, the setting in this paper is different, because different re-ID datasets contain entirely different person identities (classes). Therefore, the approaches mentioned above cannot be utilized directly for domain adaptation in person re-ID.

Unsupervised person re-ID. Unsupervised person re-ID approaches leverage hand-craft features [38], [39], [40], [41], [42], [43] or learning based features [44], [45] as representation. Hand-craft features can be directly employed in the unsupervised setting, but they do not fully exploit data distribution and fail to perform well on large-scale datasets. Some methods are based on saliency statistics [44], [45]. Yu *et al.* [46] use K-means clustering to learn an unsupervised asymmetric metric. Peng *et al.* [47] propose an asymmetric multi-task dictionary learning for cross-data transfer. Wang *et al.* [48] utilize additional attribute annotations to learn a feature representation space for the unlabeled target dataset.

Several works focus on label estimation of unlabeled target dataset [1], [49], [50], [51]. Fan *et al.* [1] propose a progressive method based on the iterations between K-means clustering and IDE [52] fine-tuning. Ye *et al.* [49] use graph matching for cross-camera label estimation. Liu *et al.* [50] employ a reciprocal search process to refine the estimated labels. Wu *et al.* [51] propose a dynamic sampling strategy for one-shot video-based re-ID. Our work seeks to learn re-ID models that can be utilized directly on the target domain and can potentially cooperate with label estimation methods in the model initialization.

Recently, some Generative Adversarial Network (GAN) based methods are applied to explore domain adaptive re-ID models. The most recent HHL approach [53] enforces cameras invariance and domain connectedness simultaneously for learning more generalizable embeddings on the target domain. PTGAN [54], a concurrent work, adopts CycleGAN [8] to generate training samples on the target domain. The common characteristic of PTGAN and our SPGAN lies in that they both consider the similarity between the generated and original image. The key difference is that PTGAN requires the foreground mask using an extra segmentation step, while SPGAN leverages two unsupervised heuristic constraints to preserve the identity of translated images.

3 PROPOSED METHOD

For unsupervised domain adaptation in person re-ID, we are provided with an annotated dataset $\mathcal{S} = \{(x_i^s, y_i^s)\}_{i=1}^{n_s}$ of n_s labeled images associated with $|\mathcal{C}_s|$ identities from the source domain and an unlabeled dataset $\mathcal{T} = \{x_i^t\}_{i=1}^{n_t}$ of n_t unlabeled images associated with $|\mathcal{C}_t|$ identities from the target domain. Note that the label space of the source domain \mathcal{C}_s is totally different from that in the target domain \mathcal{C}_t , *i.e.*, $\mathcal{C}_s \cap \mathcal{C}_t = \emptyset$. The goal of this paper is to use both the labeled source images and the unlabeled target images to train a re-ID model that generalizes well on the target domain. Briefly, in Section 3.1, we introduce the “learning via translation” framework. In Section 3.2, we revisit SPGAN. In Section 3.4, we extend the SPGAN to eSPGAN to comprehensively study the relation between source-target image translation and feature learning.

3.1 Learning via Translation

The “learning via translation” framework shown in Fig. 1. This framework consists of two components, *i.e.*, source-target image translation for training data creation, and supervised feature learning for person 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.
- **Feature learning.** With the translated dataset $G(\mathcal{S})$ that contains labels, supervised feature learning methods can be applied to train re-ID models.

In the baseline, source-target image translation is achieved by CycleGAN. For feature learning, we adopt several existing methods, such as identity discriminative embedding (IDE+) [52] and part-based convolutional baseline (PCB) [55].

As analyzed in Section 1, we aim to preserve the ID-related cues for each translated image. We emphasize that the ID information should not be the background or image style, but should be underlying and latent. To this end, **SPGAN** focuses on improving the image translation component of Fig. 1, so as to improve the re-ID accuracy. **eSPGAN** integrates image translation and feature learning into an end-to-end training system, and yields higher re-ID accuracy.

3.2 SPGAN

SPGAN mainly consists of SiaNet and CycleGAN, as shown in Fig. 2. During the training, CycleGAN aims to learn mapping functions between source and target domains, and SiaNet learns a latent space that constrains the learning of mapping functions.

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 that is indistinguishable from those in the target (source) domain, respectively. The overall objective of CycleGAN can be written as,

$$\mathcal{L}_{cyc}(G, F, D_{\mathcal{T}}, D_{\mathcal{S}}) = \mathcal{L}_{adv}(G, D_{\mathcal{T}}) + \mathcal{L}_{adv}(F, D_{\mathcal{S}}) + \alpha \mathcal{L}_{rec}(G, F), \quad (1)$$

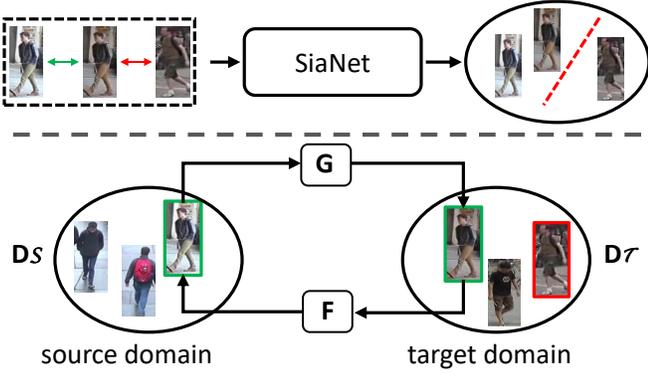


Fig. 2. SPGAN consists of two components: SiaNet (top) and CycleGAN (bottom). CycleGAN learns mapping functions G and F between the two domains, and SiaNet constrains the learning of mapping functions using two heuristic similarity-preserving losses.

where L_{adv} denotes the standard adversarial loss [56], L_{rec} represents the cycle consistency loss [8], and α controls the relative importance of the cycle-consistent loss. We would like to refer the readers to the CycleGAN [8] for more details about these loss functions.

In the experiment, we observe in Fig. 4 (b) that the model may change the color composition of the input image. This is undesirable for re-ID feature learning. Thus, we introduce the inside-domain identity constraint [57] as an auxiliary for image translation. Inside-domain identity constraint is introduced to regularize the generator to be an identity matrix on samples from the expected domain, written as:

$$\mathcal{L}_{ide}(G, F) = \mathbb{E}_{x^s \sim p_{data(S)}} \|F(x^s) - x^s\|_1 + \mathbb{E}_{x^t \sim p_{data(T)}} \|G(x^t) - x^t\|_1, \quad (2)$$

where $p_{data(S)}$ and $p_{data(T)}$ denote the sample distributions in the source and target domain, respectively.

Similarity preserving loss function. We utilize the contrastive loss [58] to train the SiaNet M ,

$$\mathcal{L}_{con}(i, x_1, x_2) = (1 - i) \{\max(0, m - d)\}^2 + id^2, \quad (3)$$

where x_1 and x_2 form a pair of input vectors, d denotes the Euclidean distance between the normalized embeddings of the two input vectors, and i represents the binary label of the pair. $i = 1$ if x_1 and x_2 are a positive pair; $i = 0$ if x_1 and x_2 are a negative pair. $m \in [0, 2]$ is the margin that defines the separability of the negative pair in the embedding space. When $m = 0$, the loss of the negative training pair is not backpropagated in the system. When $m > 0$, both positive and negative sample pairs are considered. A larger m means the loss of negative training samples has a higher impact in back-propagation.

Training data construction. In Eq. 3, the contrastive loss uses binary labels of input image pairs. In this article, we design these image pairs to reflect the proposed ‘‘self-similarity’’ and ‘‘domain-dissimilarity’’ principles. Note that, *training pairs are constructed in an unsupervised manner*, so that we use the contrastive loss without additional annotations.

- *self similarity.* Suppose two samples denoted as x^s and x^t come from the source domain and target domain, respectively. Given G and F , we define two positive

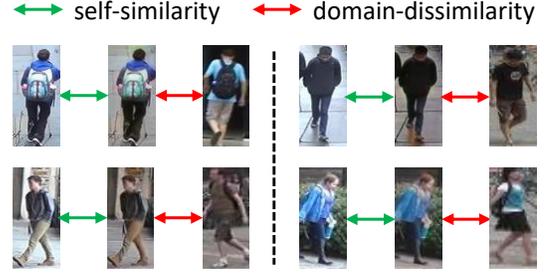


Fig. 3. Illustration of self-similarity and domain-dissimilarity. In each triplet, left: a source-domain image, middle: a source-target translated version of the source image, right: an arbitrary target-domain image. We require that 1) a source image and its translated image should contain the same ID, *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. Best viewed in color.

pairs: 1) x^s and $G(x^s)$, 2) x^t and $F(x^t)$. 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 SiaNet M to pull these two images close.

- *domain dissimilarity.* For generators G and F , we also define two types of negative training pairs: 1) $G(x^s)$ and x^t , 2) $F(x^t)$ and x^s . This design of negative training pairs is based on the prior knowledge that datasets in different re-ID domains have entirely different sets of IDs. Thus, a translated image should be of a different ID from any target image. In this manner, the network M pushes two dissimilar images away. Training pairs are shown in Fig. 3. Some positive pairs are also shown in (a) and (d) of each column in Fig. 4.

Overall objective of SPGAN. The overall objective function of SPGAN can be written as,

$$\mathcal{L}_{sp}(G, F, D_T, D_S, M) = \mathcal{L}_{cyc}(G, F, D_T, D_S) + \beta \mathcal{L}_{ide}(G, F) + \gamma \mathcal{L}_{con}(G, F, M), \quad (4)$$

where the first two losses belong to the CycleGAN formulation [8], the parameters β and γ control the relative importance of the identity loss of CycleGAN and the proposed contrastive constraint. In other words, the contrastive loss induced by SiaNet imposes a new constraint on the GAN system. The optimization process of SPGAN is,

$$G^*, F^*, M^* = \arg \min_{G, F, M} \max_{D_T, D_S} \mathcal{L}_{sp}(G, F, D_T, D_S, M). \quad (5)$$

Training procedure of SPGAN. In practice, we replace the standard adversarial loss in Eq. 1 by the least-squares loss [8], [59] to make training more stable. Specifically, for the adversarial loss $\mathcal{L}_{adv}(G, D_T)$ in Eq. 1, we train the G to minimize $\mathbb{E}_{x^s \sim p_{data(S)}} [D_T(G(x^s) - 1)^2]$ and train the D_T to minimize $\mathbb{E}_{x^s \sim p_{data(S)}} [D_T(G(x^s))^2] + \mathbb{E}_{x^t \sim p_{data(T)}} [D_T(G(x^t) - 1)^2]$.

There are three parts in SPGAN, generators, discriminators, and SiaNet. They are optimized alternately during training. When the parameters of any one part are updated, the parameters of the remaining two parts are fixed. We train SPGAN until convergence or reaching maximum iterations.

3.3 Feature Learning

Feature learning is the second component of the “learning via translation” framework. Once we have the style-transferred dataset $G(S)$ composed of translated images and their associated labels, the feature learning step is the same as supervised methods. We adopt the baseline ID-discriminative Embedding (IDE+) following the practice in [52], [60]. Given an annotated dataset $G(S)$, IDE+ aims to learn a model C by $|\mathcal{C}_s|$ -way classification with a cross-entropy loss. This corresponds to,

$$\mathcal{L}_c(C) = -\mathbb{E}(G(x^s), y^s) \sum_{k=1}^{|\mathcal{C}_s|} \mathbb{1}_{[k=y^s]} \log \left(\sigma(C^{(k)}(G(x^s))) \right), \quad (6)$$

where σ denotes the softmax activation function.

3.4 End-to-end SPGAN (eSPGAN)

SPGAN only focuses on preserving the image identity information during image translation. It is independent of the subsequent feature learning component of “learning via translation” framework. We believe the two components of the framework could benefit each other if jointly trained: 1) feature learning could guide image translation to generate identity-preserving images without heuristic constraints; 2) a stronger image translator will generate more beneficial samples for feature learning, leading to more robust person descriptors for the target domain. To this end, this article further studies the inherent relation between these two components. Specifically, we propose eSPGAN by merging the two components into an end-to-end training system.

3.4.1 Objective

eSPGAN is a unified system. It translates images to the target domain and learns re-ID features simultaneously. Following the idea of learning via translation, eSPGAN consists of two models: an image translator and a feature learner (Fig. 1). The image translator translates source images to the style of the target domain, and the feature learner learns discriminative embeddings that can be used on the target domain. Note that feature learner is differentiable with respect to the elements in the translated image $G(x^s)$. Thus, the whole system can be trained end-to-end.

Overall objective of eSPGAN. On the top of CycleGAN, we adopt the feature learner as the supervisor of the image translation. We alternately optimize the feature learner and the image translator, 1) when training the feature learner, we keep the image translator fixed, and learn a model C by $|\mathcal{C}_s|$ -way classification; 2) when training the image translator, we keep the feature learner fixed, and use its re-ID accuracy as the guidance. The feature learner will propagate a supervision signal (Eq. 6) to update the image translator, so that the translated images could be classified correctly by the former. Namely, the visual content associated with the identity information of an image is preserved. The overall objective function of eSPGAN can be written as,

$$\begin{aligned} \mathcal{L}_{esp}(G, F, D_{\mathcal{T}}, D_{\mathcal{S}}, C) = & \mathcal{L}_{cyc}(G, F, D_{\mathcal{T}}, D_{\mathcal{S}}) \\ & + \beta \mathcal{L}_{ide}(G, F) \\ & + \lambda \mathcal{L}_c(G, C), \end{aligned} \quad (7)$$



Fig. 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 + \mathcal{L}_{ide} , (d) output of SPGAN, and (e) output of eSPGAN, respectively. We observe some visual changes after image translation, such as resolution, illumination, color, and background. Moreover, SPGAN and eSPGAN have the characteristic of preserving underlying semantics of input images. Thus, their translated images will share some visual similarities with the original images. Best viewed in color.

where the first two losses belong to the CycleGAN formulation [8]. The parameter λ controls the relative importance of the feature learner constraint $\mathcal{L}_c(C)$. The optimization process of eSPGAN is,

$$G^*, F^*, C^* = \arg \min_{G, F, C} \max_{D_{\mathcal{T}}, D_{\mathcal{S}}} \mathcal{L}_{esp}(G, F, D_{\mathcal{T}}, D_{\mathcal{S}}, C). \quad (8)$$

Training procedure of eSPGAN. Similar to SPGAN, we also use the least-squares loss [8], [59] for training the generator-discriminator pairs. The proposed eSPGAN adopts an alternate optimization procedure. There are three parts in eSPGAN: generators, discriminators and feature learner (IDE+). While updating the parameters of a single part, the parameters of the other two parts are fixed. We train eSPGAN until convergence or the maximum iterations is reached.

3.4.2 Discussions on eSPGAN

In this section, we present a comprehensive discussion on eSPGAN. We first introduce the knowledge transfer mechanism of eSPGAN. Then, we analyze the crucial factor that determines whether the end-to-end system can work effectively. Moreover, the difference from other similarity-preserving methods is provided. Finally, we fully compare eSPGAN and SPGAN.

1. Bidirectional knowledge transfer. The optimizing procedure of eSPGAN can be regarded as transferring knowledge between the two components. The knowledge transfer is bidirectional: the feature learner tells the image translator *how to preserve the identity of an image*; the image

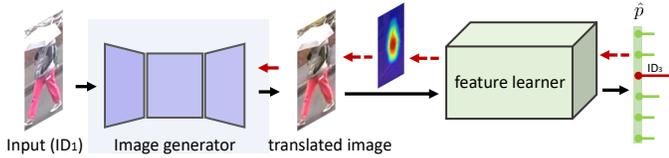


Fig. 5. Illustration of transferring knowledge of the person identity from the feature learner to the image translator. In this example, the identity of the original image is ID_1 , but its corresponding translated image is miss-classified to ID_3 by the feature learner. Namely, the identity similarity between the original image and its translated image is not preserved. To solve this problem, feature learner directly backpropagates the gradients to the input pixels of the translated image, and further updates the image translator (the red arrow). Thus, the image translator is guided to preserve person identity during translation. Note that the feature learner is fixed when we train the image translator.



Fig. 6. Illustration of transferring knowledge of the target domain from image translator to feature learner. Images with green boxes and orange boxes are on the source domain and target domain, respectively. Image translator delivers the knowledge of how a person looks like on the target domain to feature learner. Thus, the feature learner learns a domain invariant feature space by using translated images and source images for training.

translator provides *what a person from source domain looks like in target domain* for the feature learner.

(a) *Feature learner guides image translator.* Feature learner has the ability to distinguish between different identities, so it serves as a guide for image translator. During training, the translated image passes through the feature learner with fixed parameters and computes the classification loss, corresponding to Eq. 6. The feature learner then backpropagates the gradients to the input pixels of the translated image, and further updates the image translator. Thus, the image translator is guided to translate images that benefit the classification of the feature learner. As we can interpret, the translated image preserves its visual content associated with its identity.

In an example shown in Fig. 5, the translated image is misclassified by the feature learner because its identity is somehow lost during translation. In this case, the feature learner backpropagates a supervision signal to guide the training of the image translator, so that the translated image can be correctly classified. Namely, the visual content associated with the identity of an image is preserved after image translation.

(b) *Image translator strengthens feature learner.* As shown in Fig. 6, image translator translates images from the source domain to the target domain, *i.e.*, image translator creates a training dataset with labels in the target domain for feature learner. Based on this, the feature learner can learn discriminative person embeddings for the target domain.

2. Maintaining the discriminative ability of feature learner. To provide beneficial knowledge for image translator,

feature learner has to maintain its discrimination ability. Several techniques and issues are described below.

i) *Pre-training the feature learner on the source dataset.* The feature learner adopts ResNet-50 [61] pre-trained on ImageNet [62] as the backbone. We first fine-tune the feature learner on the labeled source dataset \mathcal{S} , such that it has discriminative ability at the beginning of eSPGAN training.

ii) *Real data regularization.* There exist some poorly translated images, especially at the early epochs of eSPGAN training. By “poorly translated image”, we mean two types of images. First, the image translator fails to generate high-quality images from the source to the target domain. Second, the identity of a translated image is largely lost. These poorly translated images are likely to be misclassified by feature learner and produce relatively large losses. This might cause the instability problem that affects the learning of eSPGAN.

To this end, We also use the source images when training eSPGAN. In practice, besides the batch of translated source images, we sample another batch of unaltered source images for training the feature learner. This practice guarantees that the feature learner will not be led to divergence by the poorly translated images. Moreover, using both the source and the translated images allows feature learner to learn domain invariant person embeddings. Namely, the learned feature is effective for both the source and target domains.

In late training epochs, the image translator has the ability to improve the poorly translated images based on the gradient computed by feature learner. Thus, the translated images usually have high quality and largely preserve person identities. At this stage, their effectiveness for learning desirable feature at the target domain is understandable.

3. Different from other similarity-preserving methods. There are some existing methods that also focus on the similarity-preserving property of generated images [3], [63], [64], [65]. For example, CYCADA [3] and Pose-transfer [63] both propose to utilize a model that is pre-trained on real images to preserve the semantics of generated images. SP-AEN [64] uses pre-trained AlexNet [66] to preserve perceptual information of an generated image. SRN [65] also adopts pre-trained FaceNet model to maintain the identity of the recovered face image. These existing methods all keep the pre-trained model fixed, *i.e.*, the parameters of the pre-trained model are not updated during training. Under this case, these methods can be viewed as the content loss [14] in the style transfer.

Departing from these methods, **we actually find that pre-trained feature learner should be updated during training.** We speculate the reason is that pre-trained feature learner only contains the semantic knowledge about the source domain, and it is not effective in classifying target-style images. As a consequence, a translated image might still be mis-classified by the pre-trained model even if it has successfully preserved its identity during translation.

In the person re-ID community, there are two end-to-end methods [67], [68] for the low-resolution re-ID task. Our work is inherently different from both works in several essential aspects. First, in CSR-GAN [67], the loss of re-ID does not influence the super-resolution network. In comparison, in eSPGAN the re-ID model and the image translator are well-aligned and have the impact on each other. Second, super-resolution network in SING [68] and image translator have the

substantial difference in loss function and network structure. For example, SING learns to do super resolution with ground truths in the form of low-resolution and high-resolution pairs. However, the image translator does not have paired data, and it learns to map domains from data distributions in an unsupervised manner. Thus, eSPGAN and SING [68] have significant different working mechanisms, making them completely different end-to-end systems.

4. eSPGAN vs. SPGAN. To understand the differences between SPGAN and eSPGAN, we thoroughly compare them in three aspects:

i) Network architecture. SPGAN only focuses on learning an image translator, and the re-ID feature is separately learned. In comparison, eSPGAN consists of both the image translator and the re-ID feature learner and learns them in an end-to-end manner. Thus, SPGAN is not end-to-end trainable, but eSPGAN is.

ii) Working mechanism. Both aiming at similarity-preserving image translation, SPGAN enforces this property by two unsupervised heuristic constraints, while eSPGAN does so by optimally facilitating the re-ID model learning. eSPGAN seamlessly integrates image translation and re-ID model learning, which allows us to gradually leverage the knowledge of the two components to learn discriminative embeddings for the target domain. In the experiment, we also applied the two unsupervised heuristic constraints to eSPGAN, but this does not bring any improvement.

iii) Training procedure. The alternative training procedures of eSPGAN and SPGAN appear similar from a high-level perspective. However, they use significantly different loss functions and architectures, and as such their training logistics are significantly different.

3.5 Local Max Pooling

After describing SPGAN (Section 3.2) and eSPGAN (Section 3.4), this article also introduces a useful technique for person re-ID under the domain adaptation setting, named local max pooling (LMP). LMP is not used in training; it works on a well-trained re-ID model, and is used for feature extraction of the query and gallery images. This method can reduce the impact of noisy signals incurred by fake translated images.

Specifically, in the original ResNet-50, global average pooling (GAP) is conducted on the last Convolution layer (Conv5). In the LMP (Fig. 7), we first partition the Conv5 feature maps to P horizontal parts, and then conduct global max pooling (GMP) or global average pooling (GAP) on each part. Finally, we concatenate the output of GMP or GAP of each horizontal part as the final feature representation. This procedure is non-parametric, 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. Moreover, we will show that LMP is useful under the domain adaptation setting and does not yield improvement under the normal setting where training and testing are conducted on the same domain.

4 EXPERIMENTAL EVALUATION

4.1 Datasets

We evaluate the proposed methods on two large-scale datasets, *i.e.*, Market-1501 [43] and DukeMTMC-reID [69],

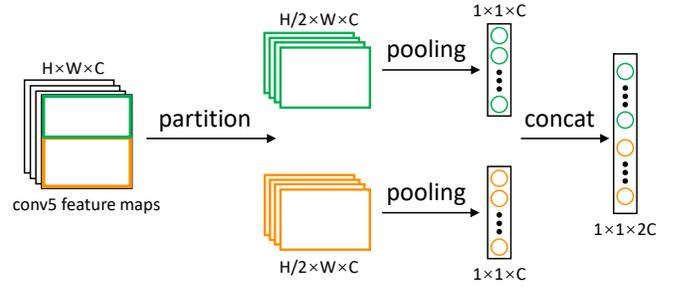


Fig. 7. Illustration of LMP. We partition the feature map into P ($P = 2$ in this example) parts horizontally. We conduct global max/ average pooling on each part and concatenate the resulting feature vectors as the final representation.



Fig. 8. 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 the source-target image translation. We observe that image resolution, illumination, color, and background are changed. Best viewed in color.

and investigate the components of our methods in details.

DukeMTMC-reID consists of 34,183 bounding boxes of 1,404 identities. There are 16,522 images from 702 identities for training, 2,228 query images from another 702 identities and 17,661 gallery images for testing. Each identity is captured by at most 8 cameras. DukeMTMC-reID is denoted as Duke for short.

Market-1501 contains 12,936 training images and 19,732 gallery images (with 2,793 distractors) detected by DPM [70]. It is split into 751 identities for training and 750 identities for testing. There are 3,368 hand-drawn bounding boxes from the 750 identities used as queries. Each identity is captured by at most 6 cameras. We also denote Market-1501 as Market for short. Sample images of two datasets are shown in Fig. 8.

Evaluation protocol. We adopt rank-1 accuracy for re-ID evaluation, which counts the percentage of queries that successfully retrieve a true match at rank 1. Besides, since multiple true positives should be returned for each query bounding box, we adopt the mean average precision (mAP) for re-ID evaluation. For Market and Duke, we use the

evaluation packages provided by the [43] and [69]. If not specified, the re-ID results in this paper are reported under the single-query setting.

4.2 Implementation Details

Feature learning method. To learn the re-ID model, we adopt IDE+ [52] as the feature learning method. For IDE+, we employ the training strategy in [60]. We adopt ResNet-50 [61] pre-trained on ImageNet [62] as the backbone network. All the images are resized to 256×128 . During training, we adopt random flipping and random cropping as data augmentation methods. Dropout probability is set to 0.5. The initial learning rate is set to 0.001 for the layers in the backbone network, and to 0.01 for the remaining layers. The learning rate is decayed by 10 after 40 epochs. We use mini-batch SGD to train IDE+ on a Tesla K80 GPU in a total of 60 epochs. Training parameters such as batch size, momentum, and gamma are set to 16, 0.9, and 0.1, respectively. We do not fine-tune the batch normalization [71] layers. During testing, given an input image, we extract the 2,048-dim Pool5 vector for retrieval under the Euclidean distance.

SPGAN training and testing. SPGAN consists of CycleGAN and SiaNet. For CycleGAN, we adopt the architecture released by its authors [8]. We use instance normalization [72] for generators but no normalization for the discriminators. For SiaNet, it contains 3 convolutional layers, 3 max pooling layers and 2 fully connected (FC) layers, 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) Max pooling 2×2 , stride = 2; (8) FC, output dimension = 256; 9) FC, output dimension = 128.

SPGAN is an unsupervised method, *i.e.*, we do not use any ID annotation during the training. In all experiment, we empirically set $\beta = 5$, $\gamma = 2$ in Eq. 4, $m = 2$ in Eq. 3, and $\alpha = 10$ in Eq. 1. The input images are resized to 256×128 . During training, two data augmentation methods, random flipping and random cropping, are employed. We use the Adam optimizer [73] with a batch size of 1, and the β_1 and β_2 are set to 0.5 and 0.999, respectively. The initial learning rate is 0.0002, and we stop training after 6 epochs. During testing, we employ the Generator G for the source (Market) \rightarrow target (Duke) image translation and the Generator F for the target (Duke) \rightarrow source (Market) image translation.

With translated images, we use three strategies to learn a re-ID model: (1) using translated images as training data; (2) using original images and translated images as training data; (3) using translated images to fine-tune the model trained on source images. The results of the three methods are nearly the same, and we adopt the third one to train re-ID models in all the experiment.

eSPGAN training and testing. eSPGAN consists of two models: an image translator and a feature learner. In this paper, we adopt CycleGAN as the image translator and IDE+ as the feature learner, and follow their original architectures. The input images are all resized to 256×128 . Besides, the feature learner is pre-trained on the source dataset following the above setting of feature learning method. During the training eSPGAN, we use two data augmentations: random

flipping and random cropping. We set the batch size to 16. For image translator, we use Adam optimizer. The learning rate is 0.0001 at the first 10 epochs and linearly decays to 0 in the remaining 5 epochs. For feature learner, we use mini-batch SGD in a total of 15 epochs. The initial learning rate is set to 0.001 for the layers in the backbone network, and to 0.01 for the remaining layers. The learning rate is decayed by 10 after 10 epochs. For all the experiment, we set hyper-parameters following CycleGAN for simplicity. Besides, we set the $\lambda = 5$ in Eq. 7. Note that the image translator (CycleGAN) and the feature learner (IDE+) are trained end-to-end, so they share the same image preprocessing procedure. Specifically, we normalize the image with the same mean (0.5, 0.5, 0.5) and standard deviation (0.5, 0.5, 0.5) for both the image translator and the feature learner. At the test time, the re-ID model produced by the feature learner is directly used for the target dataset.

4.3 Baseline Evaluation

In this section, we evaluate the direct transfer method and the “learning via translation” baseline.

Dataset bias in re-ID. To demonstrate the influence of the dataset bias, we report the results of the supervised learning method and the direct transfer method in Table 1. The supervised learning method is trained and tested on the same domain, which defines the upper bound of our system. In the direct transfer, we train a re-ID model on the source domain and directly deploy the resulting model on the target domain without any domain adaptation technique. We clearly observe a large performance drop when directly using a source-trained model on the target domain. For example, the IDE+ model trained and tested on Market achieves a rank-1 accuracy of 85.1%, but drops to 48.1% when trained on Duke and tested on Market. A similar drop can be observed when Duke is used as the target domain, which is consistent with the experiment reported in [1]. The reason behind the performance drop is the large difference between data distributions in different domains.

Effectiveness of the “learning via translation” baseline. We use CycleGAN as the baseline for source-target image translation. It is worth noting that CycleGAN does not involve any identity-preserving technique. As shown in Table 1, the baseline effectively improves over the direct transfer method on the target dataset. For example, comparing with the direct transfer, the CycleGAN baseline gains +3.5% improvement in rank-1 accuracy on Market.

Moreover, when we adopt the inside-domain identity loss in CycleGAN, we observe some further improvement. When tested on Duke and Market, the rank-1 accuracy gains brought by adding the identity loss are +2.3% and +1.4%, respectively. We speculate that the inside-domain identity loss constrains the mapping functions, such that some original semantics are preserved in the translated images. To some extent, the effectiveness of the inside-domain identity loss suggests the necessity of preserving image content. Overall, considering the results of the baselines using CycleGAN and CycleGAN + inside-domain identity loss, we conclude that the “learning via translation” baseline is effective in domain adaptation. However, comparing with SPGAN and eSPGAN, its effectiveness is limited without learning the identity-preserving property.

TABLE 1

Comparison of various methods on the target domains. When tested on Duke, Market is used as the source dataset, and vice versa. ‘‘Supervised Learning’’ denotes using labeled training images on the corresponding target dataset. ‘‘Direct Transfer’’ means directly applying the source-trained model on the target domain. ‘‘Direct Transfer (ColorJitter)’’ means using images of randomly increased/decreased brightness, contrast, and saturation during training. When local max pooling (LMP) is applied, the number of parts is set to 8. We use IDE + [52] for feature learning.

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	76.5	87.5	91.1	93.6	58.9	85.1	94.4	96.6	97.8	66.3
Direct Transfer	38.4	54.3	61.0	66.1	22.0	48.1	66.3	73.1	79.0	21.2
Direct Transfer (ColorJitter)	41.0	56.4	63.0	67.5	23.0	51.0	67.6	73.5	80.5	22.1
CycleGAN (basel.)	40.2	56.7	62.8	68.2	22.4	51.6	68.1	75.8	81.5	22.3
CycleGAN (basel.) + L_{ide}	42.5	58.5	64.3	69.3	23.1	53.0	70.2	77.6	82.4	23.5
SPGAN ($m = 2$)	44.3	61.2	66.0	71.1	24.6	54.6	72.4	79.7	84.2	25.1
SPGAN ($m = 2$) + LMP	47.1	63.8	70.0	74.2	26.1	57.2	74.0	82.1	86.4	27.4
eSPGAN	47.9	61.9	67.1	73.2	26.1	59.5	76.0	82.2	88.2	28.9
eSPGAN+ LMP	52.6	66.3	71.7	76.2	30.4	63.6	80.1	86.1	90.1	31.7

Style change. Our methods perform distribution alignment in raw pixel space - translating source data to the ‘‘style of a target domain. The ‘‘style change is complex and abstract; it involves many various factors. For example, from the visual examples in Fig. 4 and Fig. 8, we observe that resolution, illumination, color, and background are changed, as well as other changes that are harder to describe.

We provide a quantitative analysis of changes in two example visual factors: illumination and color. In Fig. 9, we visualize channel-wise histograms of translation examples in the LAB color space. We choose the LAB color space because it relates closely to how human vision works [74]. We observe all methods introduce a distribution shift of the ‘‘L channel (the distribution moves towards left), making more areas of image dark. Moreover, all the compared methods introduce the distribution shifts in ‘‘A and ‘‘B channels, too. These shifts correspond to the color composition changes. For example, CycleGAN introduces the largest distribution shifts in channels ‘‘A and ‘‘B; it changes the color from blue to red. The histogram shifts in the LAB space demonstrate that both illumination and color composition are the changed factors introduced by image translation.

To further study the impact of illumination, we have newly added a data augmentation method (‘‘colorjitter’) to the direct transfer baseline. ‘‘Colorjitter uses images of randomly increased/decreased brightness, contrast, and saturation. As shown in Table 1, ‘‘colorjitter augmentation brings about some improvement over the direct transfer baseline. This indicates illumination is a factor that causes the dataset bias between Duke and Market. However, even with ‘‘colorjitter, the direct transfer baseline is still lower than CycleGAN + L_{ide} , SPGAN and eSPGAN. It means that considering only the illumination during image translation is insufficient. More importantly, it suggests that manually designing how certain factors should be changed is not optimal. In fact, SPGAN and eSPGAN not only consider multiple factors (e.g., color composition, background, and some indescribable ones), but also change them in an automatic manner, which is much more effective than manually designed changes.

In addition, the ‘‘style is an abstract and comprehensive notion, and it is non-trivial to list and define all the related factors. Thus, we cannot manually specify certain factor changes for the image translator to learn. In comparison, our GAN-based method looks at the global data distribution,

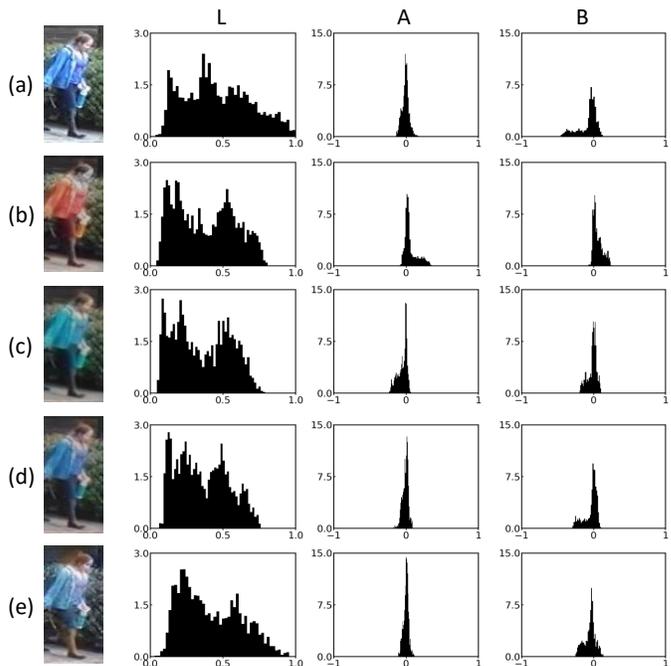


Fig. 9. Global histograms of images in LAB color space. From top to bottom: (a) original image, outputs of (b) CycleGAN, (c) CycleGAN + Lide, (d) SPGAN and (e) eSPGAN, respectively. The LAB space expresses color with three values: L^* for the luminance from black (0) to white (1), A^* from green (-1) to red (+1), and B^* from blue (-1) to yellow (+1). The histogram shifts in three channels demonstrate that illumination and color composition are the changed factors introduced by image translation.

such that image style is optimized/changed automatically.

4.4 Evaluation of SPGAN

On top of the ‘‘learning via translation’’ baseline, we replace CycleGAN with SPGAN and leave the feature learning component unchanged. In this section, we present a step-by-step evaluation and analysis of SPGAN.

SPGAN effect. On top of the ‘‘learning via translation’’ baseline, we replace CycleGAN with SPGAN ($m = 2$). The effectiveness of the proposed similarity preserving constraint can be seen in Table 1. On Duke, the similarity preserving constraint leads to +1.8% and +1.5% improvements over CycleGAN + L_{ide} in rank-1 accuracy and mAP, respectively. On Market, the performance gains are +1.6% and 1.6%.

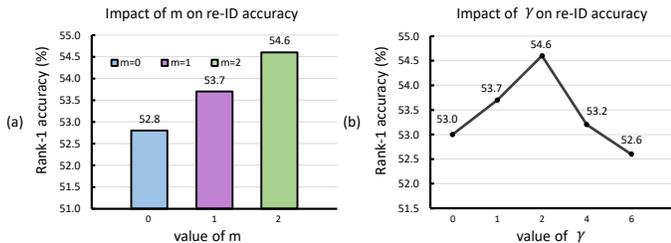


Fig. 10. The impact of the hyper-parameters of SPGAN on the re-ID rank1 accuracy. (a): the impact of m in Eq. 3, a larger m means that the loss of negative training samples has a higher weight in back-propagation. (b): the impact of γ in Eq. 4, a larger γ means a larger weight of similarity preserving constraint. The results are on Market.

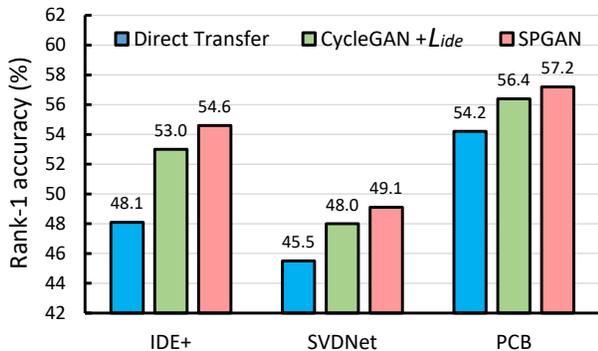


Fig. 11. Domain adaptation performance with different feature learning methods, including IDE+ [52], SVDNet [75], PCB [55]. Three domain adaptation methods are compared, *i.e.*, direct transfer, CycleGAN with identity loss, and the proposed SPGAN. The results are on Market.

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 re-ID models in the target domain. Examples of translated images by SPGAN are shown in Fig. 8.

Sensitivity of SPGAN to key hyper-parameters. SPGAN has two parameters that affect the re-ID accuracy, *i.e.*, m in Eq. 3 and γ in Eq. 4. We conduct the experiment to analyze the impact of the m and γ on Market, and results are shown in Fig. 10.

First, $m \in [0, 2]$ is the margin that defines the separability of negative pairs in the embedding space. If $m = 0$, the loss of the negative pairs is not back-propagated. If m gets larger, the weight of negative pairs in loss calculation increases. When turning off the contribution of negative pairs ($m = 0$), SPGAN only marginally improves the accuracy on Market. When increasing m to 2, we have much superior accuracy. It indicates that the negative pairs are critical to the system.

Second, γ controls the relative importance of the proposed similarity preserving constraint. As shown in Fig. 10 (b), the proposed constraint is proven effective when compared to $\gamma = 0$, but a larger γ does not bring more gains in re-ID accuracy. Specifically, $\gamma = 2$ yields the best accuracy.

Comparison of different feature learning methods. Given the same translated images, we evaluate three feature learning methods, *i.e.*, IDE+ [52], SVDNet [75], PCB [55]. We choose Market as the target dataset and duke as the source dataset, and results are shown in Fig. 11. Under the domain adaptation setting, we observe that better feature

TABLE 2
Comparison of eSPGAN and Naïve eSPGAN on Market and Duke datasets. Rank-1 accuracy (%) and mAP (%) are shown.

Methods	DukeMTMC-reID		Market-1501	
	Rank-1	mAP	Rank-1	mAP
CycleGAN + L_{ide}	42.5	23.1	53.0	23.5
Naïve eSPGAN	44.3	24.4	55.1	24.9
eSPGAN	47.9	26.1	59.5	28.9

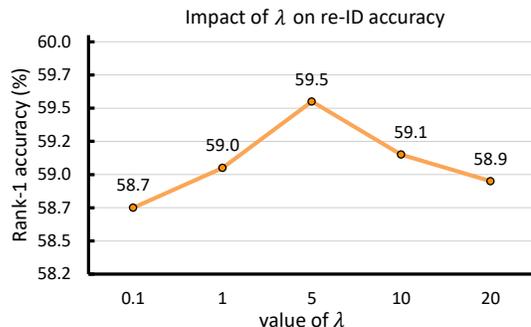


Fig. 12. Sensitivity of eSPGAN to key parameter λ in Eq. 7. A larger λ means that the feature learner has a greater influence on the image translator. The results are on Market.

learning methods lead to higher direct transfer results. For example, PCB achieves higher accuracy than IDE+ under the supervised setting on Market (92.3% vs. 85.1%), and its direct transfer accuracy is also higher than that of IDE+ (54.2% vs. 48.1%). As shown in Fig. 11, the SPGAN gains consistent improvement with three different feature learning methods. Compared with the very high direct transfer accuracy (54.2%) of PCB, the “learning via translation” framework baseline (CycleGAN + L_{ide}) gains +2.2% improvement, and the SPGAN gains +3.0% improvement.

4.5 Evaluation of eSPGAN

eSPGAN effect. An evaluation of eSPGAN is shown in Table 1. eSPGAN adopts CycleGAN + L_{ide} as the image translator. Compared with CycleGAN + L_{ide} , eSPGAN further gains +6.5 % in rank-1 accuracy on the Market dataset. We also observe the significant improvement on Duke dataset, the rank-1 accuracy increases from 42.5% to 47.9%. Moreover, eSPGAN greatly improves the performance of direct transfer, the rank-1 accuracy on Market and Duke increases from 48.1% and 38.4% to 59.5% and 47.9%, respectively. The experimental results strongly indicate that eSPGAN can effectively leverage the knowledge of image translation and feature learner to learn more discriminative embeddings for the target domain. Examples of translated images by eSPGAN are shown in Fig. 4.

Naïve eSPGAN. By this we mean that the parameters of feature learner will not be updated during training. Thus, the image translator naïvely utilizes a pre-trained source model to guide its translation procedure. Note that Naïve eSPGAN only aims to learn an image translator. As analyzed in Section 3.4.2, many existing methods adopt this way to preserve the similarity of the generated image [3], [63], [64], [65]. In Table 2, we compare eSPGAN with Naïve eSPGAN. We can observe that Naïve eSPGAN can improve the accuracy of the

TABLE 3

Performance of eSPGAN after source-target adaptation on the **source** dataset. Rank-1 accuracy (%) and mAP (%) are shown.

Methods	DukeMTMC-reID		Market-1501	
	Rank-1	mAP	Rank-1	mAP
Supervised Learning	76.5	58.9	85.1	66.3
eSPGAN	76.1	57.7	84.6	65.4

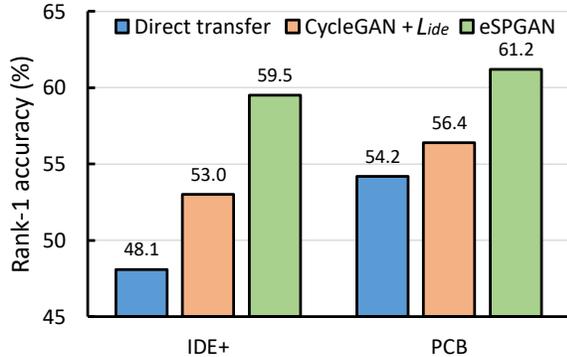


Fig. 13. eSPGAN performance with different feature learning methods, including IDE⁺ [52], PCB [55]. Three domain adaptation methods are compared, *i.e.*, direct transfer, CycleGAN with identity loss, and eSPGAN. The results are on Market.

baseline (CycleGAN + L_{ide}). However, the accuracy of Naïve eSPGAN is still much lower than eSPGAN. For example, eSPGAN obtains a much higher rank-1 accuracy than Naïve eSPGAN (47.9% vs. 44.3%) on Duke. This suggests that the parameters of pre-trained feature learner should be updated during training, so that the knowledge of feature learner and image translator can be gradually transferred to each other.

Analysis of the hyper-parameter of eSPGAN. λ in Eq. 7 is an important parameter of eSPGAN, which defines the influence of the feature learner on the image translator. To further analyze the effect of λ , we vary it from 0.1 to 20 to evaluate the performance of eSPGAN on Market. The rank-1 accuracies when using different λ are plotted in Fig. 12. In our system, when the λ is set to 5, we can obtain the best re-ID accuracy. Note that setting the λ to 0 means the feature learner has no influence on the image translator.

Analysis of the different forms of the feature learner. eSPGAN consists of an image translator and a feature learner. The feature learner is crucial for the image translator to generate similarity-preserving images, *i.e.*, the translated image maintains its visual contents that associated with the identity information. We analyze two forms of the feature learner, *i.e.*, IDE⁺ [52], PCB [55]. We choose Market as the target dataset and duke as the source dataset and report results in Fig. 13. Under the domain adaptation setting, we observe that eSPGAN gains consistent improvement with two feature learning methods. For example, when using PCB as feature learning method, eSPGAN gains +4.8% improvement over the CycleGAN + L_{ide} .

Domain invariant person embeddings. As discussed in Section 3.4.2, we also use source images when training eSPGAN. This practice ensures feature learner will not be led to divergence by the poorly translated images. In addition, using both source and translated images for feature learning leads to domain invariant person embeddings. To validate

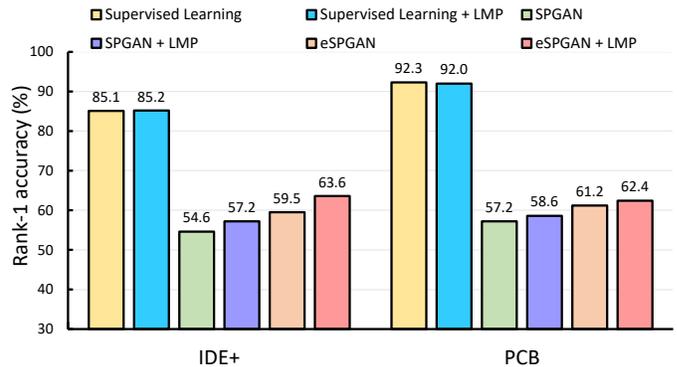


Fig. 14. The experiment of LMP ($P = 8$) on scenarios of supervised learning and domain adaptation with SPGAN and eSPGAN. Two feature learning methods are compared, *i.e.*, IDE⁺ [52], and PCB [55]. The results are on Market.

this, We also report the accuracy of eSPGAN after source-target adaptation on the **source** dataset in Table 3. We observe eSPGAN slightly decreases the rank-1 accuracy on the source dataset after source-target adaptation. Moreover, compared with the direct transfer baseline, eSPGAN significantly improves the performance on the target dataset. Thus, eSPGAN can learn person embeddings that are effective for both the source and target datasets.

Impact of two heuristic constraints. We further investigate the impact of two heuristic constraints of SPGAN on eSPGAN. We add the two heuristic constraints to eSPGAN during training, and report results in Table 4. We can observe that two heuristic constraints do not improve the performance of eSPGAN. This is because the feature learner provides the sufficiently informative and accurate constraint for the image translator. Thus, eSPGAN does not adopt two heuristic constraints during training.

Local max pooling. We apply the LMP on the last convolution layer of a re-ID model to mitigate the influence of noise. Note that LMP is directly adopted in the feature extraction step for testing without any fine-tuning. In Table 1, we can observe that LMP ($P=8$) can improve the accuracy of SPGAN and eSPGAN. With the help of LMP ($P=8$), SPGAN obtains +2.6% improvement on Market in rank-1 accuracy. LMP also improves the rank-1 accuracy of eSPGAN from 59.5% to 63.6% on Market. We empirically study how the number of parts and the pooling mode affect the accuracy. The experiment is conducted on eSPGAN. The performance of various numbers of parts ($P = 1, 2, 4, 8$) and different pooling modes (max or average) is provided in Table 5. When using 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 two feature learning methods, *i.e.*, IDE⁺ [52] and PCB [55]. As shown in Fig. 14, LMP does not guarantee stable improvement on supervised learning as observed in “IDE⁺” and PCB. However, when applied in the scenario of domain adaptation, LMP yields consistent improvement over IDE⁺ and PCB. We believe that the superiority of LMP probably lies in that it could filter out some detrimental signals in the descriptor induced by

TABLE 4

Impact of two heuristic constraints on eSPGAN. Rank-1 accuracy (%) and mAP (%) are shown.

	Training w/ heuristic constraints?	DukeMTMC-reID		Market-1501	
		Rank-1	mAP	Rank-1	mAP
eSPGAN		47.9	26.1	59.5	28.9
eSPGAN	✓	47.5	26.2	59.6	28.6

TABLE 5

Performance of various pooling strategies with different numbers of parts (P) and pooling modes (maximum or average) over eSPGAN.

#parts	mode	dim	DukeMTMC-reID		Market-1501	
			rank-1	mAP	rank-1	mAP
1	Avg	2048	47.9	26.1	59.5	28.9
	Max		50.7	28.1	62.6	30.2
2	Avg	4096	50.1	27.6	61.3	29.8
	Max		51.9	28.5	62.9	30.5
4	Avg	8192	50.1	28.0	62.5	30.1
	Max		52.4	29.0	63.2	30.9
8	Avg	16384	51.5	28.9	63.2	31.0
	Max		52.6	29.6	63.6	31.7

unsatisfied translated images.

4.6 Comparison with State-of-the-art Methods

Finally, we compare SPGAN and eSPGAN with the state-of-the-art unsupervised learning methods on Market and Duke in Table 6 and Table 7, respectively.

Market as the target domain. On Market, we first compare the proposed methods with two hand-crafted features, *i.e.*, bag-of-Words (BoW) [43] and local maximal occurrence (LOMO) [42]. These two hand-crafted features are directly applied to the target dataset without any training process, their inferiority can be clearly observed. We also compare with existing unsupervised learning methods, including the clustering-based asymmetric metric learning (CAMEL) [46], the Progressive Unsupervised Learning (PUL) [1], and UMDL [47]. For UMDL, we use the results reproduced by Fan *et al.* [1]. Moreover, we compare the proposed methods with recent domain adaptation methods of re-ID, *i.e.*, PTGAN [54], TJ-AIDL [48], and HHL [53]. In the multiple-query setting, SPGAN and eSPGAN arrive at rank-1 accuracy = 58.0% and 63.5%, respectively. The accuracy of SPGAN is 3.5% higher than CAMEL [46]. In the single-query setting, SPGAN achieves 54.6% in rank-1 accuracy, and eSPGAN achieves 59.5%. We can observe that SPGAN outperforms many other methods. With the help of LMP ($P=8$), SPGAN is comparable with recent work TJ-AIDL [48]. Moreover, eSPGAN outperforms TJ-AIDL [48] by 1.3%, which indicates that it is beneficial to jointly optimize feature learner and image translator. With the help of LMP ($P=8$), eSPGAN achieves a new state-of-the-art rank-1 accuracy=63.6%, which is 1.4% higher than the second best method HHL [53]. The comparisons indicate the competitiveness of SPGAN and eSPGAN on Market.

Duke as the target domain. On Duke, we compare the results with BoW [43], LOMO [42], UMDL [47], and PUL [1] under the single-query setting (there is no multiple-query setting in DukeMTMC-reID). We also compare with recent

TABLE 6

Comparison with the state-of-the-art methods on Market. “SQ” and “MQ” are the single-query and multiple-query settings, respectively.

Methods	Market-1501				
	Setting	Rank-1	Rank-5	Rank-10	mAP
Bow [43]	SQ	35.8	52.4	60.3	14.8
LOMO [42]	SQ	27.2	41.6	49.1	8.0
UMDL [47]	SQ	34.5	52.6	59.6	12.4
PUL [1]	SQ	45.5	60.7	66.7	20.5
Direct transfer	SQ	48.1	66.3	73.1	21.2
Direct transfer	MQ	52.3	70.1	77.2	25.0
CAMEL [46]	MQ	54.5	-	-	26.3
TJ-AIDL [48]	SQ	58.2	74.8	81.1	26.5
PTGAN [54]	SQ	38.6	-	66.1	-
HHL [53]	SQ	62.2	78.8	84.0	31.4
SPGAN	SQ	54.6	71.4	79.1	25.1
SPGAN	MQ	58.0	74.7	83.2	29.6
SPGAN+LMP	SQ	57.2	74.0	82.1	27.4
eSPGAN	SQ	59.5	76.0	82.2	28.9
eSPGAN	MQ	63.5	81.1	87.3	34.5
eSPGAN+LMP	SQ	63.6	80.1	86.1	31.7

TABLE 7

Comparison with the state-of-the-art methods on Duke under the single-query setting.

Methods	DukeMTMC-reID			
	Rank-1	Rank-5	Rank-10	mAP
Bow [43]	17.1	28.8	34.9	8.3
LOMO [42]	12.3	21.3	26.6	4.8
UMDL [47]	18.5	31.4	37.6	7.3
Direct transfer	38.4	54.3	61.0	22.0
PUL [1]	30.0	43.4	48.5	16.4
PTGAN [54]	27.4	-	50.7	-
TJ-AIDL [48]	44.3	59.6	65.0	23.0
HHL [53]	46.9	61.0	66.7	27.2
SPGAN	44.3	61.2	66.0	24.6
SPGAN+LMP	47.1	63.8	70.0	26.1
eSPGAN	47.9	61.9	67.1	26.1
eSPGAN+LMP	52.6	66.3	71.7	29.6

domain adaptation methods of re-ID, *i.e.*, PTGAN [54], TJ-AIDL [48], and HHL [53]. The result obtained by SPGAN is rank-1 accuracy = 44.3%, mAP = 24.6%, which is competitive with the recent work TJ-AIDL [48]. With the help of LMP ($P=8$), SPGAN is comparable with HHL [53]. Moreover, eSPGAN gains rank-1 accuracy=47.9%, which is +1% higher than HHL [53]. With the help of LMP ($P=8$), eSPGAN achieves a new state-of-the-art rank-1 accuracy=52.6%. Therefore, the superiority of SPGAN and eSPGAN can be concluded.

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” framework characterized by 1) unsupervised image-image translation and 2) supervised feature learning. We 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 propose a similarity preserving generative adversarial network (SPGAN) and its upgraded version, eSPGAN. Both aiming

at similarity preserving, SPGAN enforces this property by heuristic constraints, while eSPGAN does so by leveraging the discriminative knowledge of the re-ID model. We show that SPGAN and eSPGAN better qualify the generated images for domain adaptation and achieve state-of-the-art results on two large-scale person re-ID datasets. In the future, we plan to further study the relation between generative and discriminative learning, and improve our method for more general applications in visual understanding.

REFERENCES

- [1] H. Fan, L. Zheng, and Y. Yang, "Unsupervised person re-identification: Clustering and fine-tuning," *arXiv preprint arXiv:1705.10444*, 2017.
- [2] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.
- [3] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. A. Efros, and T. Darrell, "Cycada: Cycle-consistent adversarial domain adaptation," in *International Conference on Machine Learning*, 2018.
- [4] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan, "Unsupervised pixel-level domain adaptation with generative adversarial networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.
- [5] M. Liu and O. Tuzel, "Coupled generative adversarial networks," in *Advances in Neural Information Processing Systems*, 2016.
- [6] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim, "Learning to discover cross-domain relations with generative adversarial networks," in *International Conference on Machine Learning*, 2017.
- [7] Z. Yi, H. Zhang, P. Tan, and M. Gong, "Dualgan: Unsupervised dual learning for image-to-image translation," in *IEEE International Conference on Computer Vision*, 2017.
- [8] J. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," *IEEE International Conference on Computer Vision*, 2017.
- [9] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.
- [10] W. Deng, L. Zheng, Q. Ye, G. Kang, Y. Yang, and J. Jiao, "Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.
- [11] S. Benaim and L. Wolf, "One-sided unsupervised domain mapping," in *Advances in Neural Information Processing Systems*, 2017.
- [12] M. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," in *Advances in Neural Information Processing Systems*, 2017.
- [13] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *European Conference on Computer Vision*, 2018.
- [14] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European Conference on Computer Vision*, 2016.
- [15] C. Li and M. Wand, "Precomputed real-time texture synthesis with markovian generative adversarial networks," in *European Conference on Computer Vision*, 2016.
- [16] D. Ulyanov, V. Lebedev, A. Vedaldi, and V. S. Lempitsky, "Texture networks: Feed-forward synthesis of textures and stylized images," in *International Conference on Machine Learning*, 2016.
- [17] T. Q. Chen and M. Schmidt, "Fast patch-based style transfer of arbitrary style," *arXiv preprint arXiv:1612.04337*, 2016.
- [18] Y. Li, C. Fang, J. Yang, Z. Wang, X. Lu, and M. Yang, "Diversified texture synthesis with feed-forward networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.
- [19] X. Huang and S. J. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *IEEE International Conference on Computer Vision*, 2017.
- [20] Y. Li, N. Wang, J. Liu, and X. Hou, "Demystifying neural style transfer," in *IJCAI*, 2017.
- [21] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.
- [22] K. Saenko, B. Kulis, M. Fritz, and T. Darrell, "Adapting visual category models to new domains," in *European Conference on Computer Vision*, 2010.
- [23] B. Gong, Y. Shi, F. Sha, and K. Grauman, "Geodesic flow kernel for unsupervised domain adaptation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.
- [24] B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars, "Unsupervised visual domain adaptation using subspace alignment," in *IEEE International Conference on Computer Vision*, 2013.
- [25] B. Sun, J. Feng, and K. Saenko, "Return of frustratingly easy domain adaptation," in *Proceedings of AAAI Conference on Artificial Intelligence*, 2016.
- [26] S. Motiian, M. Piccirilli, D. A. Adjeroh, and G. Doretto, "Unified deep supervised domain adaptation and generalization," in *IEEE International Conference on Computer Vision*, 2017.
- [27] M. Long, G. Ding, J. Wang, J. Sun, Y. Guo, and P. S. Yu, "Transfer sparse coding for robust image representation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.
- [28] Y. Ganin and V. S. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *International Conference on Machine Learning*, 2015.
- [29] M. Long, Y. Cao, J. Wang, and M. I. Jordan, "Learning transferable features with deep adaptation networks," in *International Conference on Machine Learning*, 2015.
- [30] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell, "Deep domain confusion: Maximizing for domain invariance," *arXiv preprint arXiv:1412.3474*, 2014.
- [31] 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.
- [32] H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, and M. Marchand, "Domain-adversarial neural networks," *arXiv preprint arXiv:1412.4446*, 2014.
- [33] 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.
- [34] M. Chen, K. Q. Weinberger, and J. Blitzer, "Co-training for domain adaptation," in *Advances in Neural Information Processing Systems*, 2011.
- [35] M. Rohrbach, S. Ebert, and B. Schiele, "Transfer learning in a transductive setting," in *Advances in Neural Information Processing Systems*, 2013.
- [36] O. Sener, H. O. Song, A. Saxena, and S. Savarese, "Learning transferrable representations for unsupervised domain adaptation," in *Advances in Neural Information Processing Systems*, 2016.
- [37] W. Zhang, W. Ouyang, W. Li, and D. Xu, "Collaborative and adversarial network for unsupervised domain adaptation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.
- [38] 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.
- [39] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *European Conference on Computer Vision*, 2008.
- [40] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.
- [41] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato, "Hierarchical gaussian descriptor for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.
- [42] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.
- [43] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *IEEE International Conference on Computer Vision*, 2015.
- [44] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.
- [45] H. Wang, S. Gong, and T. Xiang, "Unsupervised learning of generative topic saliency for person re-identification," in *BMVC*, 2014.
- [46] H. Yu, A. Wu, and W. Zheng, "Cross-view asymmetric metric learning for unsupervised person re-identification," in *IEEE International Conference on Computer Vision*, 2017.

- [47] 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 *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.
- [48] J. Wang, X. Zhu, S. Gong, and L. Wei, "Transferable joint attribute-identity deep learning for unsupervised person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.
- [49] M. Ye, A. J. Ma, L. Zheng, J. Li, and P. C. Yuen, "Dynamic label graph matching for unsupervised video re-identification," in *IEEE International Conference on Computer Vision*, 2017.
- [50] Z. Liu, D. Wang, and H. Lu, "Stepwise metric promotion for unsupervised video person re-identification," in *IEEE International Conference on Computer Vision*, 2017.
- [51] Y. Wu, Y. Lin, X. Dong, Y. Yan, W. Ouyang, and Y. Yang, "Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.
- [52] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian, "Mars: A video benchmark for large-scale person re-identification," in *European Conference on Computer Vision*. Springer, 2016, pp. 868–884.
- [53] Z. Zhong, L. Zheng, S. Li, and Y. Yang, "Generalizing a person retrieval model hetero- and homogeneously," in *European Conference on Computer Vision*, 2018.
- [54] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer GAN to bridge domain gap for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.
- [55] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling," in *European Conference on Computer Vision*, 2018.
- [56] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, "Generative adversarial networks," in *Advances in Neural Information Processing Systems*, 2014.
- [57] Y. Taigman, A. Polyak, and L. Wolf, "Unsupervised cross-domain image generation," in *International Conference on Learning Representations*, 2016.
- [58] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2006.
- [59] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, "Least squares generative adversarial networks," in *IEEE International Conference on Computer Vision*, 2017.
- [60] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang, "Camera style adaptation for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.
- [61] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.
- [62] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and F. Li, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, 2015.
- [63] J. Liu, B. Ni, Y. Yan, P. Zhou, S. Cheng, and J. Hu, "Pose transferrable person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.
- [64] L. Chen, H. Zhang, J. Xiao, W. Liu, and S.-F. Chang, "Zero-shot visual recognition using semantics-preserving adversarial embedding network," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.
- [65] F. Shiri, X. Yu, F. Porikli, R. Hartley, and P. Koniusz, "Identity-preserving face recovery from stylized portraits," *International Journal of Computer Vision*, 2019.
- [66] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012.
- [67] Z. Wang, M. Ye, F. Yang, X. Bai, and S. Satoh, "Cascaded srgan for scale-adaptive low resolution person re-identification." in *International Joint Conference on Artificial Intelligence*, 2018.
- [68] J. Jiao, W.-S. Zheng, A. Wu, X. Zhu, and S. Gong, "Deep low-resolution person re-identification," in *Proceedings of AAAI Conference on Artificial Intelligence*, 2018.
- [69] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by gan improve the person re-identification baseline in vitro," in *IEEE International Conference on Computer Vision*, 2017.
- [70] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.
- [71] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning*, 2015.
- [72] D. Ulyanov, A. Vedaldi, and V. S. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," *CoRR*, vol. abs/1607.08022, 2016.
- [73] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations*, 2014.
- [74] S. Palacio, J. Folz, J. Hees, F. Raue, D. Borth, and A. Dengel, "What do deep networks like to see?" in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.
- [75] Y. Sun, L. Zheng, W. Deng, and S. Wang, "SVDNet for pedestrian retrieval," in *IEEE International Conference on Computer Vision*, 2017.