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

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

Abstract—This article studies the domain adaptation problem in person re-identification (re-ID) under a "learning via translation" framework, consisting of two components, 1) translating the labeled images from the source to the target domain in an unsupervised manner; 2) learning a re-ID model using the translated images. The objective is to preserve the underlying human identity information after image translation, so that translated images with labels are effective for feature learning on the target domain. To this end, we propose a similarity preserving generative adversarial network (SPGAN) and its end-to-end trainable version, eSPGAN. Both aiming at similarity preserving, SPGAN enforces this property by heuristic constraints, while eSPGAN does so by optimally facilitating the re-ID model learning. More specifically, SPGAN *separately* undertakes the two components in the "learning via translation" framework. It first preserves two types of unsupervised similarity, namely, self-similarity of an image before and after translation, and domain-dissimilarity of a translated source image and a target image. It then learns a re-ID model using existing networks. In comparison, eSPGAN *seamlessly* integrates image translation and re-ID model learning. During the end-to-end training of eSPGAN, re-ID learning guides image translation to preserve the underlying identity information of an image. Meanwhile, image translation improves re-ID learning by providing identity-preserving training samples of the target domain style. In the experiment, we show that identities of the fake images generated by SPGAN and eSPGAN are well preserved. Based on this, we report the new state-of-the-art domain adaptation results on two large-scale person re-ID datasets.

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

◆

1 INTRODUCTION

THIS article considers the domain adaptation in person re-ID. The re-ID task aims at searching for images of the same person to the query. 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 assume 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. In the **baseline approach**, two components are involved. 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,

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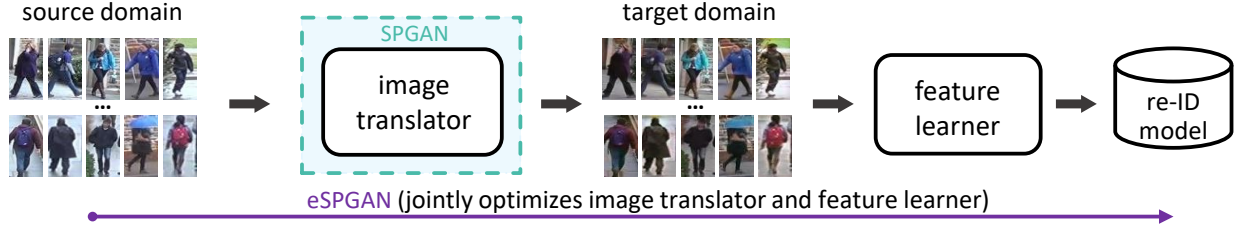 E-mail: yi.yang@uts.edu.au

commonly used image-level translation methods include [6], [7], [8], [9]. In this paper, we adopt CycleGAN in the baseline.

The objective of this research is two-fold: 1) the visual cues associated with the ID label of an image should be preserved during image-image translation. In our scenario, the visual cues refer to the underlying ID information for the foreground pedestrians. 2) The translated images can be used to learn a discriminative model for the target domain.

To achieve this objective, we first propose a heuristic solution, named similarity preserving cycle-consistent generative adversarial network (SPGAN). In our method, SPGAN is only used to improve the first component in the baseline, *i.e.*, image-image translation. That is to say, in our first solution, image-image translation and re-ID feature learning are *separately* performed. 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 image-image translation. During training, in each mini-batch, a training image is firstly used to update the Generator (of CycleGAN), then the Discriminator (of CycleGAN), and finally the layers in SiaNet. Through the coordination between CycleGAN and SiaNet, we are able to generate samples which not only possess the style of target

Fig. 1. Pipeline of the "learning via translation" framework. 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. SPGAN is only used to improve the first component of the framework, while eSPGAN seamlessly integrates the two components into a single system.

domain but also preserve their underlying ID information from the source domain.

Essentially, in SPGAN image-image translation and re-ID feature learning are separately performed. SPGAN focuses on improving the first component of the "learning via translation" framework, *i.e.*, image-image translation, while is actually independent of the feature learning component. Thus, the impact of image translation on feature learning and the reverse 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 extend SPGAN to an end-to-end version, *i.e., eSPGAN*, which *seamlessly integrates* the two components into a joint training system. eSPGAN is expected to leverage the mutual benefit of the two components, so that more discriminative embeddings on the target domain can be learned. 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. Compared with SPGAN, eSPGAN is end-to-end trainable. The unique feature of eSPGAN is that eSPGAN undertakes similarity preserving by simultaneously optimizing the feature learning component, so that the generated images better benefit feature learning, while SPGAN undertakes similarity preserving by heuristic constraints. We will show in the experiment that in eSPGAN, both components, *i.e.*, image-image translation and feature learning, are improved as a result of joint training.

The contribution of this paper is three-fold:

- Under the "learning via translation" framework, we introduce SPGAN to preserve the underlying ID information during image-image translation. SPGAN alone produces competitive domain adaptation accuracy in person re-ID.
- We further propose eSPGAN, an end-to-end trainable version of SPGAN, by simultaneously optimizing image translation and feature learning. eSPGAN gradually leverages the knowledge of the two components to learn more discriminative embeddings for the target domain. We show that eSPGAN is superior to SPGAN in terms of both image translation and feature learning.
- As a minor contribution, we propose a local max pooling (LMP) scheme as a post-processing step.

We demonstrate that the LMP add-on consistently improves over SPGAN and eSPGAN.

The rest of this article is organized as follows. The related work is presented in Section 2. Section 3 describes SPGAN and eSPGAN. In Section 4 , the experimental results are presented and analyzed. Finally, we conclude the article in Section 5.

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, "pixel2pixel" framework uses input-output pairs for learning a mapping from input to output images. In practice, considering that the paired training data is often difficult to acquire and the unpaired image-image translation is thus more applicable. To tackle the unpaired setting, a cycle consistency loss is introduced by DiscoGAN [6], DualGAN [7], and CycleGAN [8]. Benaim *et al.* [10] propose an unsupervised distance loss for one side domain mapping. Liu *et al.* [11] 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], [12] are 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, and we are more concerned with similarity-preserving mapping.

Neural style transfer [13], [14], [15], [16], [17], [18], [19] is another strategy of image-image translation, which aims at rendering a content image in the style of another image. Gatys *et al.* [20] 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 [13], [14], [15]. Huang *et al.* [18] 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. With the settings that the source domain is fully labeled and the target domain does not contain labels, our work is related to unsupervised domain adaptation (UDA). In this community, a portion of methods aim to learn a mapping between source and target distributions [21], [22], [23], [24]. As a representative UDA method, Correlation Alignment (CORAL) [24] matches the mean and covariance of two distributions.

Many other representative methods seek to find a domain-invariant feature space [25], [26], [27], [28], [29], [30], [31].

Long *et al.* [28] uses the Maximum Mean Discrepancy (MMD) [32] for this purpose. Ganin *et al.* [30] and Ajakan *et al.* [31] introduce a domain confusion loss to learn domain-invariant features. In addition, several approaches estimate the labels of unlabeled samples [33], [34], [35], [36]. The estimated labels are then used to learn the optimal classifier. Zhang *et al.* [36] propose a progressive method to select a set of pseudo-labeled target samples. Sener *et al.* [35] 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 re-ID.

Unsupervised person re-ID. Unsupervised person re-ID approaches leverage hand-craft features [37], [38], [39], [40], [41], [42] or learning based features [43], [44] 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 [43], [44]. Yu *et al.* [45] use K-means clustering to learn an unsupervised asymmetric metric. Peng *et al.* [46] propose an asymmetric multi-task dictionary learning for cross-data transfer. Wang *et al.* [47] 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], [48], [49], [50]. Fan *et al.* [1] propose a progressive method based on the iterations between K-means clustering and IDE [51] fine-tuning. Ye *et al.* [48] use graph matching for cross-camera label estimation. Liu *et al.* [49] employ a reciprocal search process to refine the estimated labels. Wu *et al.* [50] propose a dynamic sampling stragy for one-shot video-based re-ID. Our work seeks to learn re-ID models that can be utilized directly to the target domain and can potentially cooperate with label estimation methods in model initialization.

Recently, some Generative Adversarial Network (GAN) based methods are applied to explore domain adaptive re-ID models. The most recent HHL approach [52] enforces cameras invariance and domain connectedness simultaneously for learning more generalizable embeddings on the target domain. PTGAN [53], 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 image and its original image. The difference lies in 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 $\mathcal{S} = \{(x_i^s, y_i^s)\}_{i=1}^{n_s}$
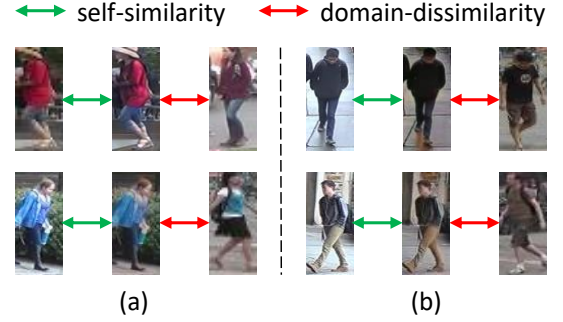

Fig. 2. Illustration of self-similarity and domain-dissimilarity. In each triplet, left: a source-domain image, middle: a source-target translated version of the source image, right: an arbitrary target-domain image. We require that 1) a source image and its translated image should contain the same ID, *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.

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 3.3, we extend the SPGAN to eSPGAN to comprehensively study the relation between source-target image translation and feature learning.

3.1 Learning via Translation

In this article, we adopt a "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, we adopt CycleGAN for source-target image translation. For feature learning, we adopt several existing methods, such as identity discriminative embedding (IDE+) [51] and part-based convolutional baseline (PCB) [54]. For the proposed methods, **SPGAN** focuses on improving source-target image translation, so as to improve re-ID accuracy. **eSPGAN** extends SPGAN by investigating the relation between image-translation and feature learning, and higher re-ID accuracy can be achieved. Note that, in both SPGAN and eSPGAN, we use existing methods for feature learning, such as [51], [54].
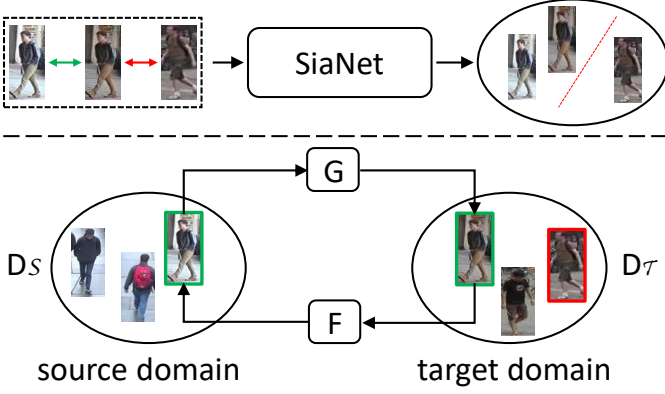
Fig. 3. SPGAN consists of two components: an SiaNet (top) and a CycleGAN (bottom). CycleGAN learns mapping functions G and F between the two domains, and SiaNet constraints the learning of mapping functions using two heuristic similarity-preserving losses.

3.2 SPGAN

SPGAN is designed to improve the image translation component of Fig. 1. The output images of SPGAN are used for the feature learning component. Our motivation is that similarity preserving is essential for generating improved fake samples for domain adaptation. 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 fulfill this goal, we integrate an SiaNet with CycleGAN, as shown in Fig. 3. During the training, CyleGAN is to learn mapping functions between two domains, and SiaNet is to learn a latent space that constrains the learning of mapping function.

3.2.1 CycleGAN revisit.

CycleGAN learns mappings between two domains. It 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. For generator G and its associated discriminator $D_{\mathcal{T}}$, the adversarial loss is

$$
\begin{aligned}
\mathcal{L}_{adv}(G, D_{\mathcal{T}}) = & \mathbb{E}_{x^s \sim p_{\text{data}(\mathcal{S})}}[\log(1 - D_{\mathcal{T}}(G(x^s)))] \\
& + \mathbb{E}_{x^t \sim p_{\text{data}(\mathcal{T})}}[\log(D_{\mathcal{T}}(x^t)],
\end{aligned} \quad (1)
$$

where $p_{\text{data}(\mathcal{S})}$ and $p_{\text{data}(\mathcal{T})}$ 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

$$
\begin{aligned}
\mathcal{L}_{adv}(F, D_{\mathcal{S}}) = & \mathbb{E}_{x^t \sim p_{\text{data}(\mathcal{T})}}[\log(1 - D_{\mathcal{S}}(F(x^t)))] \\
& + \mathbb{E}_{x^s \sim p_{\text{data}(s)}}[\log(D_{\mathcal{S}}(x^s)],
\end{aligned} \quad (2)
$$

To encourage the translated image to preserve the content of its input image during the translation, CycleGAN introduces a cycle-consistent loss, which attempts to reconstruct the original image after a cycle of translation and reverse translation. The cycle-consistent loss is

$$
\begin{aligned}
\mathcal{L}_{rec}(G, F) = & \mathbb{E}_{x^s \sim p_{\text{data}(\mathcal{S})}}[\|F(G(x^s)) - x^s\|_1] \\
& + \mathbb{E}_{x^t \sim p_{\text{data}(\mathcal{T})}}[\|G(F(x^t)) - x^t\|_1].
\end{aligned} \quad (3)
$$

The overall CycleGAN objective can be written as,

$$
\begin{aligned}
\mathcal{L}_{cyc}(G, F, D_{\mathcal{T}}, D_{\mathcal{S}}) = & \mathcal{L}_{adv}(G, D_{\mathcal{T}}) + \mathcal{L}_{adv}(F, D_{\mathcal{S}}) \\
& + \alpha \mathcal{L}_{rec}(G, F),
\end{aligned} \quad (4)
$$

where α controls the relative importance of the cycle-consistent loss.

Apart from cycle-consistent loss and adversarial loss, we use the inside-domain identity constraint [55] 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

$$
\begin{aligned}
\mathcal{L}_{ide}(G, F) = & \mathbb{E}_{x^s \sim p_{\text{data}(\mathcal{S})}}\|F(x^s) - x^s\|_1 \\
& + \mathbb{E}_{x^t \sim p_{\text{data}(\mathcal{T})}}\|G(x^t) - x^t\|_1.
\end{aligned} \quad (5)
$$

As mentioned in [8], generators G and F may change the color of input images without L_{ide}. In the experiment, we observe in Fig. 4 (b) that the model may generate unreal results without L_{ide}. This is undesirable for re-ID feature learning. Thus, we use L_{ide} to preserve the color composition between the input and output.

3.2.2 Similarity preserving

Similarity preserving loss function. We utilize the contrastive loss [56] to train the SiaNet M:

$$
\mathcal{L}_{con}(i, x_1, x_2) = (1 - i)\{\max(0, m - d)\}^2 + id^2, \quad (6)
$$

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 $m = 0$, 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 weight in backpropagation.

Training data construction. In Eq. 6, 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.

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. First, we describe pair construction for **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 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 the SiaNet M to pull these two images close.

Then, we describe **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. 2. 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,

$$
\begin{aligned}
\mathcal{L}_{sp}(G, F, D_{\mathcal{T}}, D_{\mathcal{S}}, M) =& \mathcal{L}_{cyc}(G, F, D_{\mathcal{T}}, D_{\mathcal{S}}) \\
& + \beta \mathcal{L}_{ide}(G, F) \\
& + \gamma \mathcal{L}_{con}(G, F, M),
\end{aligned} \quad (7)
$$

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,

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

Training procedure of SPGAN. 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.

3.2.3 Feature Learning

Feature learning is the second component of the "learning via translation" framework. Once we have the style-transferred dataset $G(\mathcal{S})$ 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 [51], [57]. Given an annotated dataset $G(\mathcal{S})$, IDE+ aims to learn a model C by $|\mathcal{C}_s|$-way classification with a cross-entropy loss. This corresponds to,

$$
\mathcal{L}_c(C) = -\mathbb{E}(G(x^s), y^s) \sum_{k=1}^{|\mathcal{C}_s|} \mathbb{1}_{[k=y^s]} \log \left(\sigma(C^{(k)}(G(x^s))) \right),
\quad (9)
$$

where σ denotes the softmax activation function.

3.3 End-to-end SPGAN (eSPGAN)

SPGAN focuses on preserving the image identity information during source-target image translation. It is independent of the subsequent feature learning component. In fact, we believe that the two components could benifit each other if jointly trained. On the one hand, feature learning could guide image translation to generate identity-preserving images without heuristic constraints. On the other hand, a stronger image translator will make the learned person descriptors more robust on the target domain. To this end, this article further studies the inherent relation between these two components. Specifically, we extend SPGAN to its end-to-end version, named "eSPGAN", by merging the image-image translation component and the feature learning component into a unified system.
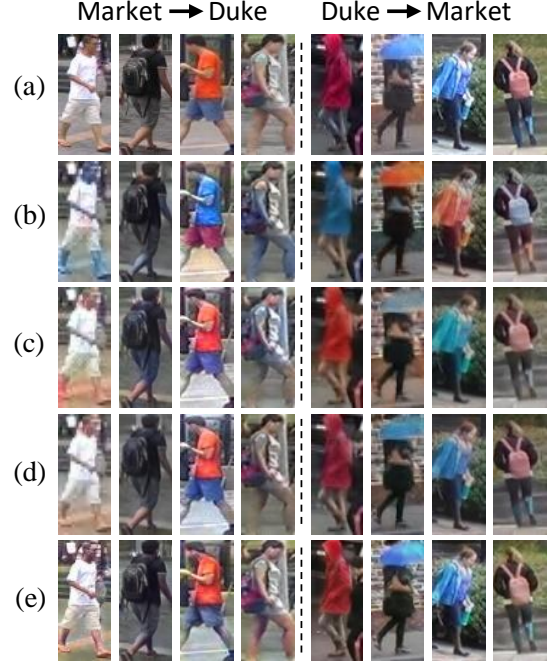

Fig. 4. Visual examples of image-image translation. The left four columns map Market images to the Duke style, and the right four columns map Duke images to the Market style. From top to bottom: (a) original image, (b) output of CycleGAN, (c) output of CycleGAN + L_{ide}, (d) output of SPGAN, and (e) output of eSPGAN. We observe that images generated by SPGAN and eSPGAN are of the target style and well preserve the identity information. Best viewed in color.

3.3.1 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 feature learner and image translator, 1) when training feature learner, we keep image translator fixed, and learn a model C by $|\mathcal{C}_s|$-way classification; 2) when training image translator, we keep feature learner fixed, and use feature learner to guide image translator. The feature learner will propagate a supervision signal (Eq. 9) 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,

$$
\begin{aligned}
\mathcal{L}_{esp}(G, F, D_{\mathcal{T}}, D_{\mathcal{S}}, C) =& \mathcal{L}_{cyc}(G, F, D_{\mathcal{T}}, D_{\mathcal{S}}) \\
& + \beta \mathcal{L}_{ide}(G, F) \\
& + \lambda \mathcal{L}_c(G, C),
\end{aligned} \quad (10)
$$

where the first two losses belong to the CycleGAN formulation [8]. The parameter λ controls the relative importance of
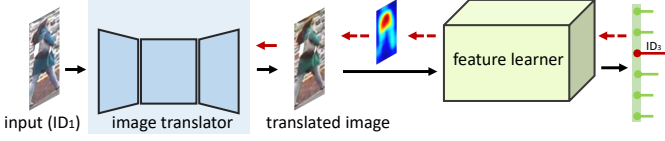
Fig. 5. Illustration of transferring knowledge of the person identity from the feature learner to the image translator. In this example, the identity of the original image is ID_1, but its corresponding translated image is miss-classified to ID_3 by the feature learner. Namely, the identity similarity 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.

the feature learner constraint. The optimization process of eSPGAN is,

$$G^*, F^*, C^* = \arg \min_{G,F,C} \max_{D_{\mathcal{T}}, D_{\mathcal{S}}} \mathcal{L}_{esp}(G, F, D_{\mathcal{T}}, D_{\mathcal{S}}, C). \quad (11)$$

Training procedure of eSPGAN. There are three parts in eSPGAN, generators, discriminators and feature learner (IDE+). 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 eSPGAN until convergence or reaching maximum iterations.

3.3.2 Discussions on eSPGAN

eSPGAN seamlessly integrates the two components of "learning via translation" framework into a unified system. Jointly optimizing the two components is critical and non-trivial. In principle, the two components should collaborate with each other effectively: the image translator provides the feature learner with effective training images of the target domain style, while the feature learner forces the image translator to preserve the image identity.

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

(i) 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. 9. 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.

(ii) Image translator strengthens feature learner. The feature learner aims at learning a re-ID model that can be used for the target domain. As shown in Fig. 6, the image translator translates images from the source domain to target domain, *i.e.* image translator creates a training dataset with labels in the target domain for feature learner. The feature learner will progressively gain knowledge of the target domain by utilizing the translated images to learn discriminative person embeddings.

The proposed eSPGAN adopts an alternate optimization procedure. We alternately optimize the feature learner and the image translator, so that their knowledge can be progressively transferred to each other. In our method, when we optimize one component, the parameters of the other component are fixed.

Maintaining the discriminative ability of the feature learner. To provide beneficial knowledge for the image translator, the feature learner has to maintain its discrimination ability. Several techniques and issues are described below.

(i) Pretraining the feature learner. We initialize the feature learner by training it on the annotated source dataset $\mathcal{S}$. In this manner, the feature learner has a decent discriminative ability at the beginning of eSPGAN training.

(ii) Real data regularization. In addition to source pretraining, we also use the source images when training eSPGAN. In this case, the source images and the translated images are both used. This practice allows feature learner to maintain its knowledge of the source domain during training. Moreover, because the feature learner adopts both translated images and source images for training, it is thus able to learn domain invariant person embeddings. Namely, the learned feature is effective for both the source and target domains.

(iii) How do poorly translated images affect eSPGAN? 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 the feature learner and their influence on the feature learner should be considered. We observe that the poorly translated images are not detrimental for the discriminative ability of the feature learner. This is consistent with previous findings [58], [59]. Moreover, feature learner also uses real images for training. This practice guarantees

that the learning procedure will not be led to divergence by the poorly translated images.

In late training epochs, the image translator will improve the poorly translated images based on the gradient generated by the feature learner. Therefore, at this stage, the translated images usually have relatively high quality and largely preserve the person identities. So their effectiveness is understandable at this stage. Based on the above discussions, we are able to use translated images to train the feature learner in both early and late training epochs.

Comparison with other similarity-preserving methods. There are some existing methods that also focus on the similarity-preserving property of generated images [3], [60], [61]. For example, CYCADA [3] and Pose-transfer [60] both propose to utilize a model that is pre-trained on real images to preserve the semantics of generated images. SP-AEN [61] uses pre-trained AlexNet [62] to preserve perceptual information of an generated 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 [13] in the style transfer. Departing from these competing methods, **we actually find that pre-trained feature learner should be updated during training.** We speculate that the reason is that the pre-trained model only contains the knowledge about the source domain, so the feature learner is not effective in classifying target-style images. As a consequence, a translated image might still be misclassified by the pre-trained model even if it has successfully preserved its identity during translation.

Another related method is PTGAN [53]. It aims to preserve the foreground content of a person image using an extra segmentation step. We could see that PTGAN adopts a hand-crafted constraint (foreground should be maintained) to preserve the identity. In comparison, for eSPGAN, the identity is such preserved as to benefit the ultimate feature learning step for person re-ID. Considering the superior performance of eSPGAN to PTGAN (Table 6 and Table 7), we speculate that end-to-end learning is very beneficial to generate task-related training samples.

eSPGAN vs. SPGAN. eSPGAN and SPGAN share the same objective that the person identity of the translated images should be preserved. Thus, they both focus on generating identity-preserving images, *i.e.*, an image should preserve content associated with its ID after image translation. SPGAN focuses on improving the first component of "learning via translation" framework using two unsupervised heuristic constraints. In comparison, eSPGAN seamlessly integrates the two components into a joint training system. eSPGAN leverages the feature learner to guide image translator to generate similarity-preserving images. Note that eSPGAN does not adopt the two unsupervised constraints of SPGAN, because the feature learner provides sufficiently informative and accurate constraint for the image translator. In our experiment, we show that adding the two heuristic constraints on eSPGAN does not improve the performance.

Image-level domain adaptation. This paper considers the unsupervised domain adaptation (UDA) in person re-ID, where different datasets contain entirely different persons (classes). In general UDA, a recent trend is image-level domain translation [3], [4], [5], which motivates us to alleviate
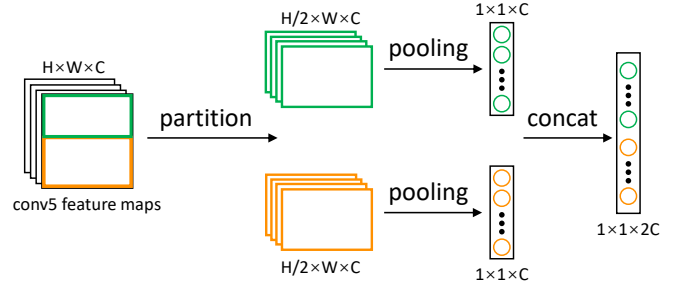

Fig. 7. Illustration of LMP. We partition the feature map into P ($P = 2$ in this example) parts horizontally. We conduct global max/ average pooling on each part and concatenate the resulting feature vectors as the final representation.

the re-ID dataset bias under the "learning via translation" framework. In this article, we focus on preserving the similarity of the translated image, and this mechanism can be applied to general domain adaptation and other visual understanding tasks.

3.4 Local Max Pooling

After describing SPGAN (Section 3.2) and eSPGAN (Section 3.3), 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 [42] and DukeMTMC-reID [63], and investigate the components of our methods in details.

DukeMTMC-reID is a re-ID protocol of the DukeMTMC dataset [64]. It contains 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. We denote DukeMTMC-reID as Duke for short.

Market-1501 contains 12,936 training images and 19,732 gallery images (with 2,793 distractors) detected by DPM [65]. It is split into 751 identities for training and 750 identities for

Duke images   Duke images to Market style

Market images   Market images to Duke style

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 unpaired image-image translation. Best viewed in color.

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 [42] and [63]. 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+ [51] as the feature learning method. For IDE+, we employ the training strategy in [57]. 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. We adopt ResNet-50 [66] as the backbone network. 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 CNN models on a Tesla K80 GPU in a total of 60 epochs. Training parameters such as batch size, maximum number of epoch, momentum, and gamma are set to 16, 50, 0.9, and 0.1, respectively. We do not finetune the batch normalization [67] 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 Cycle-GAN and SiaNet. For CycleGAN, we adopt the architecture released by its authors [8]. We use instance normalization [68] 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. 7, $m = 2$ in Eq. 6, and $\alpha = 10$ in Eq. 4. 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 [69] 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 dataset (Market) $\rightarrow$ the target dataset (Duke) translation and the Generator F for target dataset (Duke) $\rightarrow$ the source dataset (Market) 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 model 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 batch size to 16. For image translator, we use the 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. 10. 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.

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. When local max pooling (LMP) is applied, the number of parts is set to 8. We use IDE + [51] 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
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**

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. In the baseline domain adaptation approach (Section 1), we first translate the labeled images from the source domain to the target domain and then use the translated images to train re-ID models. Note that this process does not involve any identity-preserving technique. In our baseline, we adopt CycleGAN for source-target image translation and report the results in Table 1. 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. When tested on Duke, the performance gain of CycleGAN is +1.8% and +0.4% in rank-1 accuracy and mAP, respectively. Moreover, when we adopt the identity loss in CycleGAN, we observe some further improvement. When tested on Duke and Market, the performance gain brought by adding the identity loss is +2.3% and +1.4%, respectively. We speculate that the identity loss constrains the mapping functions, such that some original semantics are preserved in the translated images. To some extent, the effectiveness of the identity loss suggests the necessity of preserving image content. Overall, considering the results of the baselines using CycleGAN and CycleGAN+identity loss, we conclude that the "learning via translation" baseline is effective in domain adaptation. However, comparing with the proposed method, its effectiveness is limited without learning the identity-preserving property.

4.4 Evaluation of SPGAN

On top of the "learning via translation" baseline, we replace CycleGAN with SPGAN and leave the feature learning
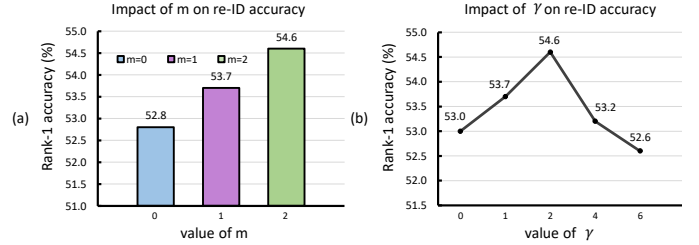

Fig. 9. The impact of the hyper-parameters of SPGAN on the re-ID rank1 accuracy. (a): the impact of m in Eq. 6, a larger m means that the loss of negative training samples has a higher weight in back-propagation. (b): the impact of γ in Eq. 7, a larger γ means a larger weight of similarity preserving constraint. The results are on Market.
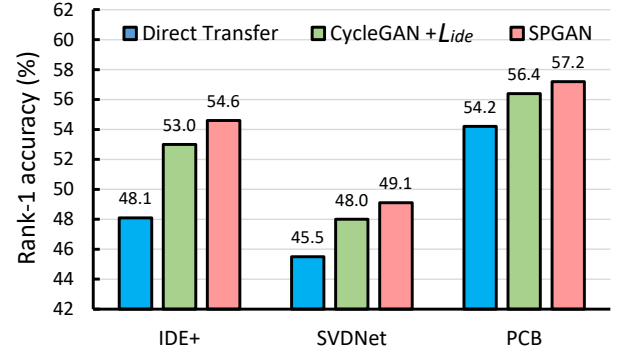

Fig. 10. Domain adaptation performance with different feature learning methods, including IDE$^+$ [51], SVDNet [70], PCB [54]. Three domain adaptation methods are compared, *i.e.*, direct transfer, CycleGAN with identity loss, and the proposed SPGAN. The results are on Market.

component unchanged. In this section, we present 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%. 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.

TABLE 2
Comparision 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

TABLE 3
Performance of eSPGAN after source-target adaptation on the **source** dataset. Rank-1 accuracy (%) and mAP (%) are shown.

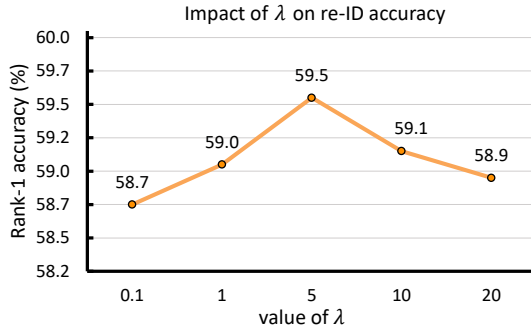| Methods	DukeMTMC-reID		Market-1501	
	Rank-1	mAP	Rank-1	mAP
Supervised Learning	76.5	58.9	85.1	66.3
eSPGAN	76.1	57.7	84.6	65.4

Fig. 11. Sensitivity of eSPGAN to key parameter λ in Eq. 10. A larger λ means that the feature learner has a greater influence on the image translator. The results are on Market.

Sensitivity of SPGAN to key hyper-parameters. SPGAN has two parameters that affect the re-ID accuracy, *i.e.*, m in Eq. 6 and γ in Eq. 7. We conduct the experiment to analyze the impact of the m and γ on Market, and results are shown in Fig. 9.

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. 9 (b), the proposed constraint is proven effective when compared to $\gamma = 0$, but a larger γ does not bring more gains in 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$^+$ [51], SVDNet [70], PCB [54]. We choose Market as the target dataset and duke as the source dataset, and results are shown in Fig. 10. Under the domain adaptation setting, we observe that better feature 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. 10, 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 seamlessly merges the image-image translation component and the feature learning component into a unified system. In this section, we perform step-by-step studies 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. As analyzed in Section 3.3.2, many existing methods adopt this way to preserve the similarity of the generated image. In Table 2, we compare eSPGAN with Naïve eSPGAN. We can observe that Naïve eSPGAN can improve the accuracy of the 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. 10 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. 11. 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 maintain its visual contents that associated with the identity information. We analyze two forms of the feature learner, *i.e.*, IDE$^+$ [51], PCB [54]. We choose Market as the target dataset and duke as the source dataset and report
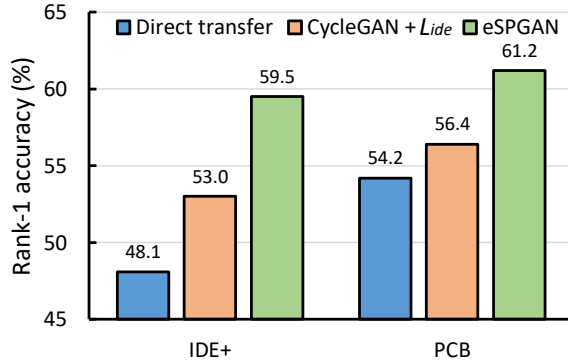
Fig. 12. eSPGAN performance with different feature learning methods, including IDE$^+$ [51], PCB [54]. Three domain adaptation methods are compared, i.e., direct transfer, CycleGAN with identity loss, and eSPGAN. The results are on Market.
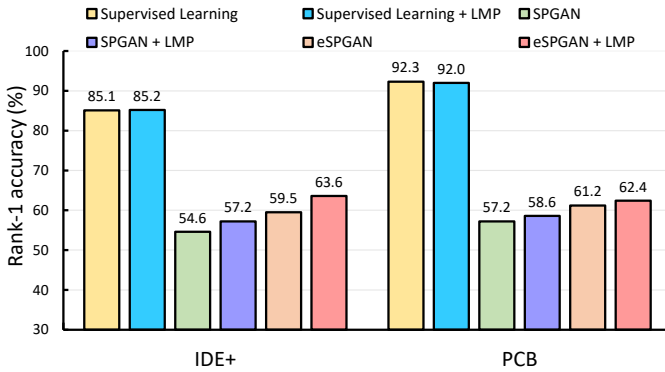

Fig. 13. The experiment of LMP ($P = 8$) on scenarios of supervised learning and domain adaptation with SPGAN and eSPGAN. Two feature learning methods are compared, i.e., IDE$^+$ [51], and PCB [54]. The results are on Market.

TABLE 4
Impact of two heuristic constraints on eSPGAN. Rank-1 accuracy (%) and mAP (%) are shown.

	Training w/ heuristic constraints?	DukeMTMC-reID Rank-1	mAP	Market-1501 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 rank-1	mAP	Market-1501 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**

results in Fig. 12. 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}.

Impact of the real sample regularization. As discussed in Section 3.3.2, we also use source images when training eSPGAN. This practice ensures the feature learner will not be led to divergence by the poorly translated images. Moreover, the feature learner adopts both source images and translated images as training images, which makes it learn domain invariant person embeddings. We report the performance of eSPGAN after source-target adaptation on the **source** dataset in Table 3. We can observe that eSPGAN slightly decreases the rank-1 accuracy on the source dataset after source-target adaptation. Besides, compared with the direct transfer baseline, eSPGAN significantly improve 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 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$^+$ [51] and PCB [54]. As shown in Fig. 13, 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 improvement over IDE$^+$ and PCB. The superiority of LMP probably lies in that max pooling filters out some detrimental signals in the descriptor induced by fake 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. The comparisons are shown in Table 6 and Table 7, respectively.

Market-1501 as target domain. On Market-1501, we first compare the proposed methods with two hand-crafted features, i.e., bag-of-Words (BoW) [42] and local maximal

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 [42]	SQ	35.8	52.4	60.3	14.8
LOMO [41]	SQ	27.2	41.6	49.1	8.0
UMDL [46]	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 [45]	MQ	54.5	-	-	26.3
TJ-AIDL [47]	SQ	58.2	74.8	81.1	26.5
PTGAN [53]	SQ	38.6	-	66.1	-
HHL [52]	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 [42]	17.1	28.8	34.9	8.3
LOMO [41]	12.3	21.3	26.6	4.8
UMDL [46]	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 [53]	27.4	-	50.7	-
TJ-AIDL [47]	44.3	59.6	65.0	23.0
HHL [52]	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**

occurrence (LOMO) [41]. 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) [45], the Progressive Unsupervised Learning (PUL) [1], and UMDL [46]. 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 [53], TJ-AIDL [47], and HHL [52]. 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 [45]. 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 [47]. Moreover, eSPGAN outperforms TJ-AIDL [47] 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 [52]. The comparisons indicate the competitiveness of SPGAN and eSPGAN on Market.

DukeMTMC-reID as target domain. On DukeMTMC-reID, we compare the results with BoW [42], LOMO [41], UMDL [46], and PUL [1] under the single-query setting (there is no multiple-query setting in DukeMTMC-reID). We also compare with recent domain adaptation methods of re-ID, *i.e.*, PTGAN [53], TJ-AIDL [47], and HHL [52]. The result obtained by SPGAN is rank-1 accuracy = 44.3%, mAP = 24.6%, which is competitive with the recent work TJ-AIDL [47]. With the help of LMP (P=8), SPGAN is comparable with HHL [52]. Moreover, eSPGAN gains rank-1 accuracy=47.9%, which is +1% higher than HHL [52]. 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 end-to-end trainable 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 the state-of-the-art results on two large-scale person re-ID datasets. In the future, we plan to further 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] S. Benaim and L. Wolf, "One-sided unsupervised domain mapping," in *Advances in Neural Information Processing Systems*, 2017.

[11] M. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," in *Advances in Neural Information Processing Systems*, 2017.

[12] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *European Conference on Computer Vision*, 2018.

[13] 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.

[14] C. Li and M. Wand, "Precomputed real-time texture synthesis with markovian generative adversarial networks," in *European Conference on Computer Vision*, 2016.

[15] 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.

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

[17] 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.

[18] X. Huang and S. J. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *IEEE International Conference on Computer Vision*, 2017.

[19] Y. Li, N. Wang, J. Liu, and X. Hou, "Demystifying neural style transfer," in *IJCAI*, 2017.

[20] 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.

[21] K. Saenko, B. Kulis, M. Fritz, and T. Darrell, "Adapting visual category models to new domains," in *European Conference on Computer Vision*, 2010.

[22] 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.

[23] B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars, "Unsupervised visual domain adaptation using subspace alignment," in *IEEE International Conference on Computer Vision*, 2013.

[24] B. Sun, J. Feng, and K. Saenko, "Return of frustratingly easy domain adaptation," in *Proceedings of AAAI Conference on Artificial Intelligence*, 2016.

[25] 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.

[26] 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.

[27] Y. Ganin and V. S. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *International Conference on Machine Learning*, 2015.

[28] M. Long, Y. Cao, J. Wang, and M. I. Jordan, "Learning transferable features with deep adaptation networks," in *International Conference on Machine Learning*, 2015.

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

[30] 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.

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

[32] 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.

[33] M. Chen, K. Q. Weinberger, and J. Blitzer, "Co-training for domain adaptation," in *Advances in Neural Information Processing Systems*, 2011.

[34] M. Rohrbach, S. Ebert, and B. Schiele, "Transfer learning in a transductive setting," in *Advances in Neural Information Processing Systems*, 2013.

[35] 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.

[36] 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.

[37] 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.

[38] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *European Conference on Computer Vision*, 2008.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

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

[45] 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.

[46] 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.

[47] 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.

[48] 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.

[49] Z. Liu, D. Wang, and H. Lu, "Stepwise metric promotion for unsupervised video person re-identification," in *IEEE International Conference on Computer Vision*, 2017.

[50] 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.

[51] 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.

[52] Z. Zhong, L. Zheng, S. Li, and Y. Yang, "Generalizing a person retrieval model hetero- and homogeneously," in *European Conference on Computer Vision*, 2018.

[53] 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.

[54] 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.

[55] Y. Taigman, A. Polyak, and L. Wolf, "Unsupervised cross-domain image generation," in *International Conference on Learning Representations*, 2016.

[56] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2006.

[57] 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.

[58] Z. Dai, Z. Yang, F. Yang, W. W. Cohen, and R. Salakhutdinov, "Good semi-supervised learning that requires a bad GAN," in *Advances in Neural Information Processing Systems*, 2017.

[59] Y.-X. Wang, R. Girshick, M. Hebert, and B. Hariharan, "Low-Shot Learning from Imaginary Data," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[60] 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.

[61] 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.

[62] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012.

[63] 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.

[64] 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.

[65] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[66] 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.

[67] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning*, 2015.

[68] D. Ulyanov, A. Vedaldi, and V. S. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," *CoRR*, vol. abs/1607.08022, 2016.

[69] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations*, 2014.

[70] Y. Sun, L. Zheng, W. Deng, and S. Wang, "SVDNet for pedestrian retrieval," in *IEEE International Conference on Computer Vision*, 2017.