

Received July 25, 2019, accepted August 4, 2019, date of publication August 8, 2019, date of current version August 28, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2933910

Identity Preserving Generative Adversarial Network for Cross-Domain Person Re-Identification

JIALUN LIU¹, WENHUI LI¹, HONGBIN PEI, YING WANG, FENG QU, YOU QU, AND YUHAO CHEN

College of Computer Science and Technology, Jilin University, Changchun 130012, China

Corresponding author: Wenhui Li (liwh@jlu.edu.cn)

This work was supported in part by the Science and Technology Development Plan of Jilin Province under Grant 20170204020GX, in part by the Development and Reform Commission of Jilin Province under Grant 2019C054-2, and in part by the National Natural Science Foundation of China under Grant 5180523.

ABSTRACT In this paper, we study the domain adaptive person re-identification(re-ID) problem: train a re-ID model on the labeled source domain and test it on the unlabeled target domain. It's known challenging due to the feature distribution bias between the source domain and target domain. The previous methods directly reduce the bias by image-to-image style translation between the source and the target domain in an unsupervised manner. However, these methods only consider the rough bias between the source domain and the target domain but neglect the detailed bias between the source domain and the target camera domains (divided by camera views), which contain critical factors influencing the testing performance of re-ID model. In this work, we particularly focus on the bias between the source domain and the target camera domains. To overcome this problem, a multi-domain image-to-image translation network, termed Identity Preserving Generative Adversarial Network (IPGAN) is proposed to learn the mapping relationship between the source domain and the target camera domains. IPGAN can translate the styles of images from the source domain to the target camera domains and generate many images with styles of target camera domains. Then the re-ID model is trained with the translated images generated by IPGAN. During the training of the re-ID model, we aim to learn the discriminative feature. We design and train a novel re-ID model, termed IBN-reID, in which Instance and Batch Normalization block (IBN-block) are introduced. Experimental results on Market-1501, DukeMTMC-reID and MSMT17 show that the images generated by IPGAN are more suitable for cross-domain re-ID. Very competitive re-ID accuracy is achieved by our method.

INDEX TERMS Person re-identification, domain adaptation, style transfer, unsupervised learning.

I. INTRODUCTION

As one of the most challenging problems in the field of computer vision [1]–[6], the person re-identification(re-ID) task aims at matching the relevant images across no-overlap camera views at different locations and times. It plays a crucial role in law enforcement and video surveillance for public safety, such as long-term cross camera tracking and video retrieval. To enhance the re-ID performance, the existing re-ID works mostly attempt to learn discriminative feature embedding representation for better measuring the similarities between person image pairs. With the recent developments in deep learning, deeply-learning representations provide high discriminative feature learning ability. Deep

learning methods dominate the re-ID community. Many deep learning-based methods [9]–[12] are proposed for the re-ID task. However, all of above-mentioned re-ID methods are in supervised learning program, in which, those models are trained and tested on the same labeled dataset(domain). There are many fundamental limitations: those works based on supervised learning need a certain number of high quality training samples and large-scale manual labels to learn an optimal feature extractor. It is expensive and impractical to collect such scale manual labelling in the real-world. Moreover, those re-ID models often fail to generalize well from a dataset to a new one, because of the feature distribution bias between two different datasets.

A promising strategy to solve this problem is unsupervised domain adaptation. Domain adaptation strategy defines two

The associate editor coordinating the review of this manuscript and approving it for publication was Lefei Zhang.

domains, source domain (a dataset with fully labeled information) and target domain (a dataset without label information). In order to transfer information from source domain to target domain, most adaptation methods make a closed-set assumption, i.e., the source and target domains share entirely same classes. However, such assumption does not hold for re-ID because the unsupervised domain adaptation re-ID is an open-set problem where the classes (persons) from two domains are completely different. Therefore, the traditional unsupervised domain adaptation method is not applicable to re-ID.

Recently, advanced re-ID methods based on unsupervised domain adaptation are proposed. These works reduce the bias between the source and target domain on the image-level [13], [14] or the feature-level [24]. However, these methods only take into consideration the general gap between the source and the target domain and **ignore the detailed gap between the source domain and the target camera domains**. As Fig 1 shows, images captured by different cameras often have many obviously different styles because the types of camera and scenes for image acquisition are different. In the real world, the feature distribution of one camera view may greatly differ from the feature distribution of the other ones. Therefore, it is improper to take the target domain as a whole. Intuitively, a complete domain should be divided into several camera domains(sub-domains) according to the number of cameras. For example, DukeMTMC-reID and Market1501 have 8 and 6 camera domains, respectively. A transfer model which only captures the overall data bias between the two domains will face tough performance tests. The better solution of domain adaptation for re-ID is to reduce the bias between source domain and each camera domain in target domain.

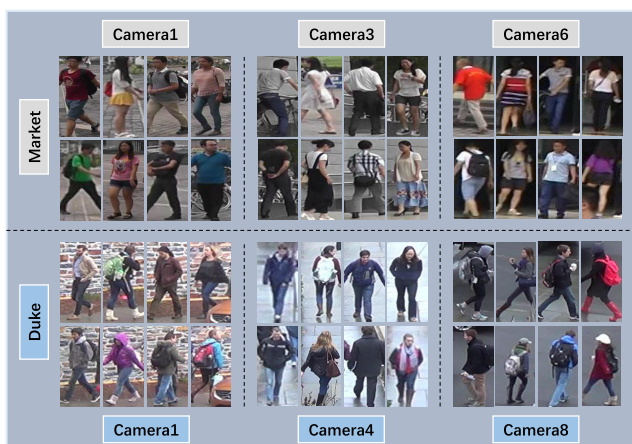


FIGURE 1. Example images from Market-1501 and DukeMTMC-reID, respectively.

Based on the above intuition, we present the “multi-style-transfer for re-ID” framework. The overall framework is shown in Fig 2. It consists of two pieces: 1) style transfer between the source domain and the target camera domains.

2) learning a discriminative re-ID model by the generated images as inputs. Firstly, in order to reduce the gaps between two different domains, we translate the styles of images from the source domain to the target camera domains. We can generate many images with styles of target camera domains. Then, we train the reID model with the generated images. For the first part, we propose a multi-domain image-to-image style translation approach, termed Identity Preserving Generative Adversarial Network (IPGAN). Different from learning the mapping relationship between the source and target domain, IPGAN is designed to learn the mapping relationship between the source domain and the target camera domains. For the second part, we present the IBN-reID model. It's a simple but robust feature extractor. The IBN-reID model can learn more discriminative feature than the baseline model.

For the style transfer model IPGAN, IPGAN consists of a generator G , a domain discriminator D_{dom} , and a semantic discriminator D_{sem} . The inputs of G are images and domain labels. The inputs of D_{dom} are the fake images generated by G and the domain labels. G learns to flexibly translate the input images into the corresponding target camera domains. D_{dom} learns to distinguish between real and fake images and minimize the domain classification loss for the known label. The inputs of D_{sem} are the style-translated images from source domain and the identity labels of images from the source domain. We design the D_{sem} to implement the semantic constraint that the identity information of images from source domain keeps invariance before and after translation. Such a semantic discriminator brings more challenges to a generator in GAN framework. Through the IPGAN, we are able to generate images which not only possess the style of target camera domains but also preserve the original identity information. Using IPGAN, we can create a new dataset, which inherits the labels from source domain and has the style of target camera domains. Then, we train the re-ID model on the new dataset in a supervised way to test on the target dataset.

For the re-ID feature learning model, we introduce the Instance and Batch Normalization block (IBN-block) into baseline model inspired by [22]. We name the new re-ID model as IBN-reID model. The model can eliminate appearance variance in shallow layers, and holds discrimination of the learned features in deep layers. In the model, instance normalization and batch normalization are integrated. Instance normalization provides visual and appearance invariance, while batch normalization accelerates training and preserves discriminative features. The IBN-reID model has a better generalization ability than the baseline model [14], especially in the case of deploying the model trained on source domain to target domain without any domain adaptation. In experiments, we train the IBN-reID model on the style-translated DukeMTMC-reID training set and test it on the Market-1501 testing set. The proposed models achieve competitive performances, rank-1 accuracy = 57.2%, mAP = 28.0%.

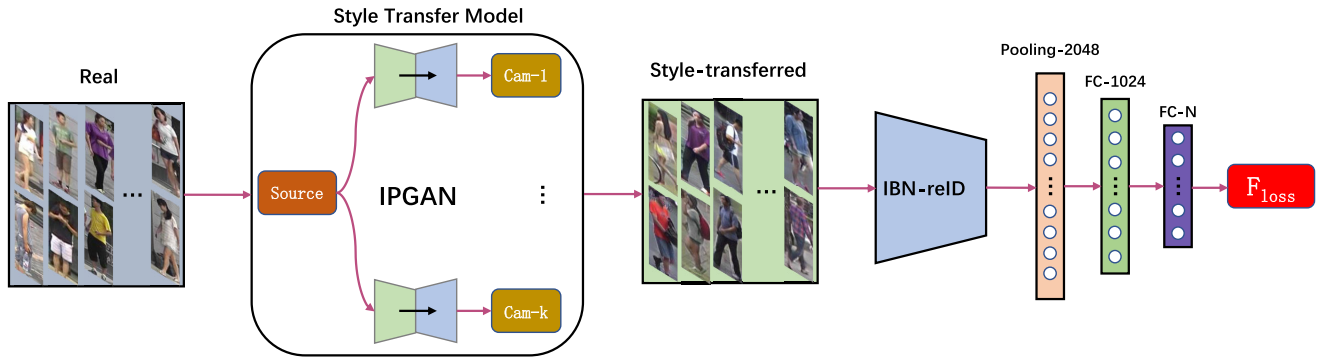


FIGURE 2. The pipeline of our method consists of two steps. First, the IPGAN is learned from the training data between source domain and target camera domains. For each image from source domain, we can use the trained IPGAN to generate images which inherit the original identity semantic information and fit the style of target camera domains. Second, we train IBN-reID model with the style-transferred (green boxes) images in a supervised way.

In summary, this paper has the following contributions:

- 1) We present the “multi-style-transfer for re-ID” framework for unsupervised domain adaptation in re-ID.
- 2) We propose IPGAN, a novel and efficient unsupervised multi-domain image-to-image translation network.
- 3) We present IBN-reID model which intuitively induces appearance invariance into the re-ID model.
- 4) We do many experiments on three large-scale datasets: Market-1501, DukeMTMC-reID and MSMT17. We prove that the proposed method achieves very competitive re-ID accuracy and it is efficient and applied.

II. RELATED WORK

A. GENERATIVE ADVERSARIAL NETWORKS

Generative adversarial network (GAN) [33] has shown remarkable performance improvement in various computer vision tasks, especially image-to-image translation, in recent years. For image-to-image translation task, pix2pix [25] uses a conditional GAN to learn mappings from input to output images by combining adversarial loss and L_1 loss. However, this method needs paired data to train its model. For an unpaired image-to-image translation, several methods are proposed [23], [26]–[29]. UNIT [28] combines variational autoencoders [29] and CoGAN [23], in which the two generators share same weights. CycleGAN [26] and DiscoGAN [27] use a cycle consistency to preserve key attributes. However, all the aforementioned frameworks only consider the mapping from source domain to target domain. Unlike the old frameworks, we propose a new one which can translate images from source domain to every target camera domain using only a single model and be used to improve the performance of cross domain person re-ID.

B. UNSUPERVISED DOMAIN ADAPTATION RE-ID

Most existing re-ID models are based on supervised learning [8], [31], [32], [37]. These models suffer from poor scalability in the real-world environment. Recent works [15], [34], [39] employ deep learning technology to

estimate the labels of unlabeled target dataset. Reference [15] proposes an unsupervised approach which utilizes K-means to offer fake-labels for the unlabeled samples and train the re-ID model with those samples interactively. Reference [34] uses K-reciprocal nearest neighbors to get proximate label information for unsupervised video re-ID. DAREID [39] proposes three assumptions: covariate assumption, Separately Probabilistic Lipschitzness assumption and weight ratio. Concretely, DAREID [39] is a process of clustering in target domain then assigns pseudo-labels to target domain images, iteratively. However, They will be unstable, if there are large intra-class variations and small inter-class variations in the target domain. Those methods cannot obtain the perfect clustering results. A few unsupervised domain adaptation for person re-ID methods has proposed [13], [16], [24], [38], [40]. Based on CycleGAN [26], PTGAN [13] and SPGAN [14] translate images from the source domain to the target domain which works by keeping the underlying identity information during translation. Then, the translated images are utilized to train re-ID model in a supervised way. The above methods attempt to reduce the bias between the source and the target domain on image space and feature space, however they all ignore the divergence of image style caused by target camera domains. Camstyle [36] considers the intra-domain image variations caused by target cameras and learn discriminative representations of target domain. But this method cannot keep the same identity information between original images and translated images because identify semantic constraint is not considered. Besides, the above approaches have limited scalability in handling multiple domains since different models need to be trained on each pair of domains. ECN [40] investigates into the intra-domain variations of the target domain, including three variations: exemplar-invariance, camera-invariance and neighborhood-invariance. ECN [40] only considers the intra-domain variations of the target domain. Images from source domain are just used to learn the basic human attributes. Different from the general unsupervised person re-ID, DIMN [38] proposes a novel deep re-ID model with strong generalizable ability, which learns a

mapping between a person image and its identity classifier using a single shot. There is no information from target domain in the learning process of DIMN [38]. ECN [40] and DIMN [38] ignore the mapping relationship between the source domain and target domain. Unlike existing method, our method explicitly considers the difference between the source domain and the target camera domains, and we can learn the relations among multiple domains using only a single model.

III. PRELIMINARY

In this section, we first describe the cross-domain re-ID task by giving the baseline in Section III-A. We then briefly review the StarGAN [21] in Section III-B.

A. BASELINE RE-ID MODEL

In the community of unsupervised domain adaptation in re-ID, a source domain $\{X_s, I\}$ with full identity label information is provided. Each person image x_s is associated with an only identity label i . In addition, we are provided with an unlabeled target domain X_t . The identity information of the target domain is not available. Our goal is to learn a transferable re-ID model using the labeled source domain, which generalizes well on the target testing set.

We use ResNet-50 [30] as a backbone network for re-ID feature learning. We use two FC layers to replace the final 1,000-dim fully connected(FC) layer after the Pooling-5 layer. The dimensions of the two FC layers are 1,024 and N , where N is the number of classes in the dataset. The cross-entropy loss function is used to optimize the model parameters. Our training process follows the ID discriminative embedding (IDE) [9].

B. STARGAN REVISIT

StarGAN [21] is a novel and efficient generative adversarial network for multi-domain image-to-image translations that learns the mapping relations among multiple domains using only a single model. To achieve this goal, the generator translates the input sample x into an output image y conditioned on the target domain label c . The target domain label c is generated randomly so that the generator can translate the input image flexibly. The loss functions of StarGAN [21] include three aspects.

1) ADVERSARIAL LOSS

To make the generated images indistinguishable from real images, the adversarial loss is defined as:

$$\mathcal{L}_{adv}(x, c) = \mathbb{E}_x [\log D_{\mathcal{T}}(x)] + \mathbb{E}_{x,c} [\log (1 - D_{\mathcal{T}}(G(x, c)))], \quad (1)$$

where x and c are the input image and target domain label, respectively. G generates an image $G(x, c)$ to fake $D_{\mathcal{T}}(x)$. The term $D_{\mathcal{T}}(x)$ is a probability distribution over sources.

2) DOMAIN CLASSIFICATION LOSS

The goal of StarGAN is to translate x to an output image y which is classified as the target domain c . For this, a classifier is added on the top of the discriminator. The domain classification loss of real/fake images is defined as:

$$\mathcal{L}_{dom}^r(x, c') = \mathbb{E}_{x,c'} [-\log D_{dom}(c'|x)], \quad (2)$$

$$\mathcal{L}_{dom}^f(x, c) = \mathbb{E}_{x,c} [-\log D_{dom}(c|G(x, c))]. \quad (3)$$

where $\mathcal{L}_{dom}^r(x, c')$ and $\mathcal{L}_{dom}^f(x, c)$ are the domain classification loss of real and fake images, respectively. The term x and c' are the real image and the original domain label, respectively. The term $G(x, c)$ and c are the generated image and the target domain label, respectively. The term $D_{dom}(c'|x)$ and $D_{dom}(c|G(x, c))$ represent the probability distribution over domain labels computed by the discriminator. By minimizing Eq.2 and Eq.3, the discriminator learns to classify a real image x to its original domain c' and a fake image $G(x, c)$ to target domain c .

3) RECONSTRUCTION LOSS

To guarantee that translated images preserve the content of their original images, StarGAN introduces a cycle-consistent loss [26], [27] which attempts to recover the original image after a cycle of translation and reverse translation. it's defined as:

$$\mathcal{L}_{rec}(x, c, c') = \mathbb{E}_{x,c,c'} [\|x - G(G(x, c), c')\|_1], \quad (4)$$

where the $L1$ norm is used as reconstruction loss. The generator G takes in the translated image $G(x, c)$ and attempts to reconstruct the original image x conditioned on the original domain label c' .

More details can be found in StarGAN [21].

IV. PROPOSED METHOD

In this section, The two components of our framework, IPGAN and IBN-reID model, are described in Section IV-A and Section IV-B, respectively.

A. IPGAN

In the real-world, different camera views have different camera styles and the bias as shown in Fig 1. In this paper, we define each camera domain in target domain as a sub-domain. The distribution of one sub-domain may greatly differ from the distribution of the others. Thus, it is improper to take the target domain as a whole. A better way to smooth the bias between source and target domain is to translate images from source domain to each target camera domain. To this end, we propose the Identity Preserving Generative Adversarial Network (IPGAN) based on StarGAN [21]. Our method can ensure that the transferred image has a similar style as the style in target camera domain. The method is also able to keep the identity information of images from source domain during the translation. IPGAN consists of a style transfer model $G(x, c)$, a domain discriminator D_{dom} , and a semantic discriminator D_{sem} , as illustrated in Fig 3. The

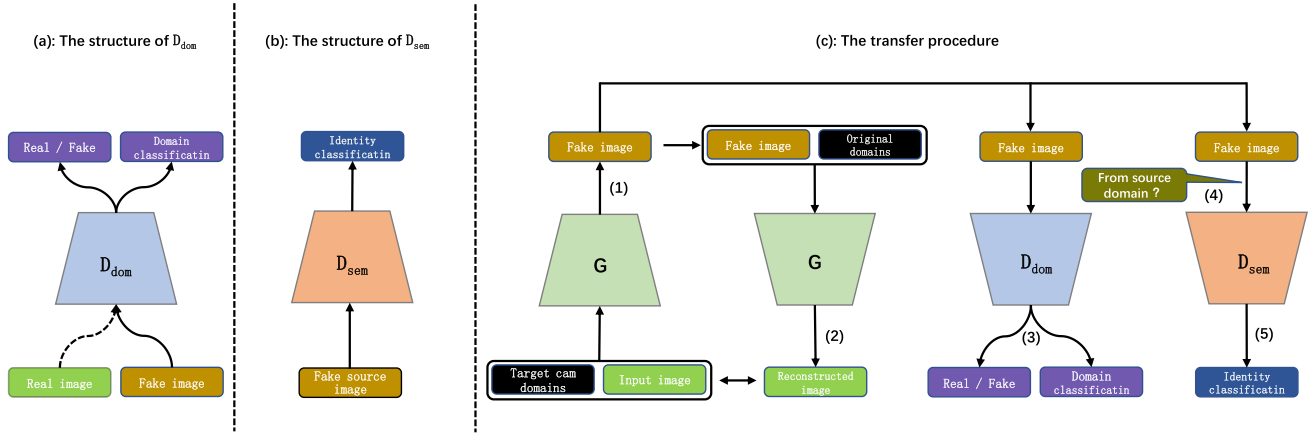


FIGURE 3. Overview of IPGAN, consisting of three modules, a generator G , a domain discriminator D_{dom} , and a semantic discriminator D_{sem} . (a) and (b) show the structures of D_{dom} and D_{sem} , respectively. (c) displays the procedure of domain adaptation. (1) The generator G learns to focus on the target camera domain labels to perform image-to-image translation and generate fake images. (2) The generator G attempts to reconstruct the original images using the fake images and the original domain labels. (3) The fake images are fed into the domain discriminator D_{dom} . D_{dom} learns to distinguish between real and fake images and minimize the domain classification loss for the known label. (4),(5) If the fake images are generated based on images from source domain images, input it into the semantic discriminator D_{sem} for minimizing the identity semantic loss.

inputs of $G(x, c)$ are images and domain labels. The inputs of D_{dom} are the fake images generated by $G(x, c)$ and the domain labels. The inputs of D_{sem} are the style-translated images from source domain and the identity labels of images from the source domain. Compared to identity labels, camera labels can be obtained along with surveillance videos without tedious and expensive manual labeling. There is no identity information of the target set during image-image translation. Thus, IPGAN is an unsupervised domain adaptation method.

1) DIFFERENCE FROM StarGAN

Here we elaborate the difference between existing StarGAN [21] and the proposed IPGAN.

- 1) The task of the two methods is different. StarGAN [21] is proposed for the multi-domain image-to-image translation, while IPGAN is designed specifically for cross-domain re-ID and more effective in reducing the bias between two different re-ID datasets.
- 2) The goal of the two methods is different. StarGAN [21] focuses on the low-level appearance changes. IPGAN focuses on not only the low-level appearance changes but also the high-level semantic information.
- 3) The structure of two methods is different. The structure of IPGAN and StarGAN [21] are $\{G, D_{dom}, D_{sem}\}$ and $\{G, D_{dom}\}$, respectively. IPGAN has a novel semantic discriminator D_{sem} , to achieve the constraint that the identity information of images from source domain keeps invariance before and after translation. We will describe D_{sem} detailedly in the following.

IPGAN has the following features:

2) SOURCE TO TARGET CAMERA DOMAINS IMAGE-IMAGE TRANSLATION

As analyzed in the above, we aim to further reduce the feature distribution bias between two domains by learning

the mapping relationship between the source domain and target camera domains. The source dataset and images captured by different cameras in target dataset can be considered as independent domains. In this step, IPGAN inherits the feature of multi-domain image-to-image translation from StarGAN [21]. Without the identity semantic constraint, we train a generator model that can translate images among the $L + 1$ domains which consist of L camera domains from target dataset and one source domain. In respect of G and D , the objective functions are given as

$$\mathcal{L}_{\mathcal{D}} = -\mathcal{L}_{adv} + \lambda_{dom} \mathcal{L}_{dom}^r, \quad (5)$$

$$\mathcal{L}_G = \mathcal{L}_{adv} + \lambda_{dom} \mathcal{L}_{dom}^f + \lambda_{rec} \mathcal{L}_{rec}, \quad (6)$$

where λ_{dom} and λ_{rec} are two hyper-parameters, that tradeoff the importance of domain classification loss and reconstruction loss. Specifically, following the training strategy in [21], the generator G contains 2 convolutional layers, 6 residual blocks and 2 transposed convolution layers. The discriminator D_{dom} has the same structure as PatchGANs [25].

3) IDENTITY SEMANTIC CONSTRAINT LOSS FUNCTION

The another goal of IPGAN is that although the style of images in source domain is changed, the translated image should preserve the same identity information with its corresponding original image. Note that the identity information is not the low-level image appearance information but high-level semantic information which cannot be displayed on the image. The identity information is underlying and latent. We hope to preserve the identity information of images from the source domain during image-image translation. To fulfill this goal, we design a semantic discriminator D_{sem} , which is used for identity preserving. It's a key part in IPGAN. The consistency on person identity is important for person re-ID training. The identity semantic constraint loss



FIGURE 4. Visual examples of images from source domain to target camera domains translation. The left seven columns map DukeMTMC-reID to Market-1501 six camera domain styles. The right nine columns map Market-1501 to DukeMTMC-reID eight camera domain styles.

function is:

$$\mathcal{L}_{sem}(x_s, y) = \mathbb{E}_{x_s, y}[-\log D_{sem}(y|G(x_s, c))], \quad (7)$$

where the term x_s and c are the source domain images and the target camera domain labels, respectively. G generates an image $G(x_s, c)$ conditioned on both the input source domain images x_s and the target camera domain label c . x_s corresponds to a identity label y . The structure of D_{sem} is a classifier. In order to learn the high-level semantic information, we pretrain the D_{sem} with the images from source domain. During training IPGAN, in a min-batch, the D_{sem} aims to classify a generated $G(x_s, c)$ to its corresponding original identity label y . The identity semantic constraint loss is only used to optimize the generator.

The above two parts are integrated in a framework IPGAN. Its structure is illustrated in Fig 3. The generator G maps source domain images to the styles of target camera domains. The discriminator D_{dom} is used to distinguish real and fake images and judge the domain that the translated images belong to. The D_{sem} enforces the images from source domain keeping identity information after translation. Note that D_{sem} is a pretrained classifier with source domain training set and only the identity information from source domain is applied.

Finally, the overall IPGAN objective function can be written as:

$$\mathcal{L}_{\mathcal{D}_{IPGAN}} = -\mathcal{L}_{adv} + \lambda_{dom} \mathcal{L}_{dom}^r, \quad (8)$$

$$\begin{aligned} \mathcal{L}_{\mathcal{G}_{IPGAN}} = & \mathcal{L}_{adv} + \lambda_{dom} \mathcal{L}_{dom}^f \\ & + \lambda_{rec} \mathcal{L}_{rec} + \lambda_{sem} \mathcal{L}_{sem}, \end{aligned} \quad (9)$$

where λ_{sem} is a hyper-parameter just like λ_{dom} and λ_{rec} . It tradeoffs the importance of identity semantic loss. We use $\lambda_{dom} = 1$, $\lambda_{rec} = 10$, and $\lambda_{sem} = 1$ in our experiments.

With the learned IPGAN model, for an image in source domain, we generate $L + 1$ fake samples whose styles are similar to $L + 1$ domains. Meanwhile the identity information is kept during the image-to-image translation. Note that the $L + 1$ samples contain a fake image whose style is same as its original style and the fake image should be ignored. Finally, we train re-ID model with the style transferred images in a supervised way. Fig 4 provides some examples generated by IPGAN.

B. IBN-RE-ID MODEL

To address the problem of the appearance gap between source and target camera domains, a intuitive way is to introduce appearance invariance into re-ID models. Inspired by [22], we design a novel re-ID model, termed IBN-reID, by introducing the IBN-block. The characteristic of IBN-reID is that batch normalization and instance normalization are utilized together in the network. The advantage of batch normalization is preserving discrimination between individual samples by deep features, but the drawback is that it makes CNNs vulnerable to appearance transforms. On the contrary, instance normalization eliminates individual contrast, but diminishes useful information at the same time. IBN-block integrates Instance normalization and batch normalization. Thus, it can eliminate appearance variance and maintains discrimination of the learned features.

To our best knowledge, our IBN-reID is the first attempt to introduce IBN-block into re-ID model. In IBN-reID, the appearance variance is mainly reflected in shallow features, and the change of content information is reflected in deep features. IBN-reID has better generalization ability than the baseline model, especially in the case of deploying the

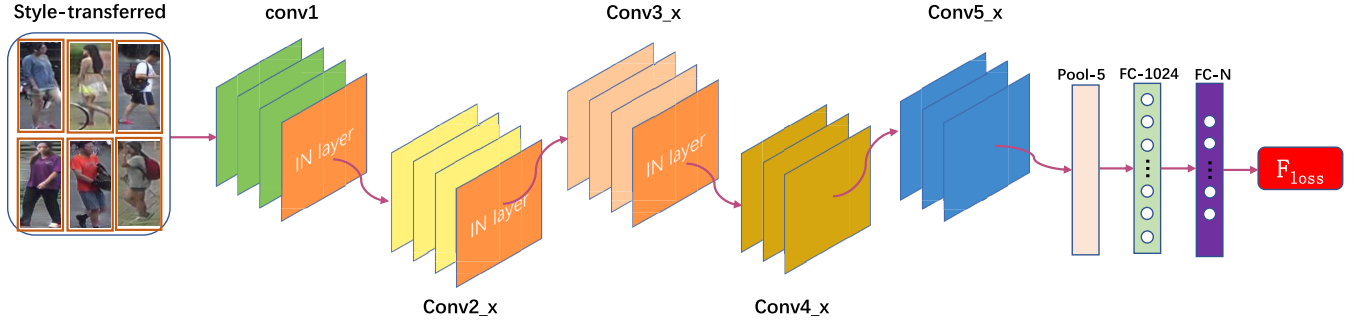


FIGURE 5. We add three IN layers after the first convolution layer (conv1) and the first two convolution groups (conv2_x, conv3_x). Then we utilize the images which fit the style of target camera domains to train the IBN-reID model. The F_{loss} is cross-entropy loss.

TABLE 1. The information of Market-1501 [20], DukeMTMC-reID [18], [19] and MSMT17 [13].

Dataset	Identities	Train set	Gallery	Query	Cameras
Market [20]	1,501	12,936	19,732	3,368	6
Duke [18], [19]	1,404	16,522	17,611	2,228	8
MSMT17 [13]	4,101	32,621	82,161	11,659	15

model trained on source domain to target domain without any domain adaptation.

To construct IBN-reID, we use ResNet-50 [30] as base model that consists of four groups of residual blocks. Then, we add three IN layers after the first convolution layer (conv1) and the first two convolution groups (conv2_x, conv3_x), respectively. Its structure is shown in Fig 5. The latter layers are the same as baseline model (Section III-A)

V. EXPERIMENT

In this section, we will first describe the experimental setting in Section V-A. Then, we show the comparisons to the state-of-the-art methods in Section V-B. Finally, the ablation study is analyzed in Section V-C.

A. EXPERIMENTAL SETTING

1) DATASETS

Three widely used benchmark datasets have been chosen for our experiments, Market-1501 [20], DukeMTMC-reID [18], [19] and MSMT17 [13], because these datasets 1) are large-scale and 2) camera labels for each image are available. The information of Market-1501 [20], DukeMTMC-reID [18], [19] and MSMT17 [13] are shown in Table 1,

- 1) The **Market-1501** [20] dataset contains 32,668 images from 1501 identities collected from 6 cameras. All of the images are produced by deformable part mode (DPM) [35]. The dataset is split into two non-overlapping parts: 12,936 images from 751 identities for training and 19,732 images from 750 for testing. Moreover, 3,368 query images from 750 identities are used to retrieve the matching persons in the gallery.
- 2) The **DukeMTMC-reID** [18], [19] is also a large-scale re-ID dataset which is collected from 8 cameras.

It contains 16,522 training images from 702 identities, 2,228 query images from another 702 identities and 17,661 gallery images in total 36411 images belonging to 1404 identities.

- 3) The **MSMT17** [13] dataset is collected from 15 cameras: 12 outdoor cameras and 3 indoor cameras. In total, it contains 126,441 bounding boxes, 4,101 identities. The training set contains 32,621 bounding boxes of 1,041 identities. The query set contains 11,659 bounding boxes. The gallery set contains 82,161 bounding boxes.

2) EVALUATION METRICS

For three datasets, we followed the standard evaluation settings for performing a fair comparison with existing methods. We use two evaluation metrics to evaluate the performance of our proposed method. The first one is the Cumulative Matching Characteristic (CMC). Considering re-ID as a ranking problem, CMC represents the probability that a query image appears in different sized candidate lists. We use rank-1, rank-5, rank-10 accuracy to represent the CMC curve. The onther one is the mean average precision (mAP), which is calculated as the mean value of average precision across all queries. The two evaluation metrics can reflect the performance more comprehensively in both precision and recall rate.

3) IMPLEMENTATION DETAILS

We use Pytorch to train IPGAN on NVIDIA GeForce GTX Titan xp GPU using the training set of Market-1501, DukeMTMC-re-ID and MSMT17. We use a single model to learn the mapping between source domain and target camera domains. The architecture of generator and the discriminator D_{dom} are the same as StarGAN [21]. the generator contains 2 convolutional layer, 6 residual blocks and 2 transposed convolution layers, while the discriminator is the same as PatchGANs [25]. For the generator network, we use instance normalization in all layers except the last output layer. For the discriminator D_{dom} network, we use Leaky ReLU with a negative slope of 0.01. The structure of discriminator D_{sem} is similar to the baseline model. The weighs of D_{sem} are fixed

TABLE 2. Comparison with state-of-the-art on Market-1501 and DukeMTMC-reID. The results based on the methods proposed by this paper are in bold.

Methods	Duke → Market				Market → Duke			
	rank-1	rank-5	rank-10	mAP	rank-1	rank-5	rank-10	mAP
LOMO [7]	27.2	41.6	49.1	8.0	12.3	21.3	26.6	4.8
BOW [20]	35.8	52.6	60.3	14.8	17.1	28.8	34.9	8.3
UMDL [16]	34.5	52.6	59.6	12.4	18.5	31.4	37.6	7.3
PUL [15]	45.5	60.7	66.7	20.5	30.0	43.4	48.5	16.4
CAMEL [17]	54.5	-	-	26.3	-	-	-	-
PTGAN [13]	38.6	-	66.1	-	27.4	-	50.7	-
SPGAN [14]	51.5	70.1	76.8	22.8	41.1	56.6	63.0	22.3
TJ-AIDL [24]	58.2	74.8	81.1	26.5	44.3	59.6	65.0	23.0
IPGAN+Baseline	56.4	75.6	82.5	25.6	46.8	62.0	67.9	25.7
IPGAN+IBN-reID	57.2	76.0	82.7	28.0	47.0	63.0	68.1	27.0

during training IPGAN. The input images are resized to 128×64 . The model is trained about 200 epochs. The learning rate is 0.0001 in the first 100 epochs and linearly reduces to zero in the last 100 epochs. for each image in the source training set, we generate L style-transferred images (the number of target cameras). These fake images are regarded as a new training set which is used to train re-ID model.

B. COMPARISON WITH STATE-OF-THE-ART METHODS

1) STATE-OF-THE-ART METHODS

We compare the proposed method with the state-of-the-art unsupervised learning methods, including two hand-crafted methods, three unsupervised methods, and three unsupervised domain adaptation methods. Table 2 presents the comparison when tested on Market-1501 and DukeMTMC-reID, respectively. Our method achieves **57.2%/47.0% rank-1 accuracy and 28.0%/27.0% mAP accuracy** by using DukeMTMC-reID/Market-1501 as source set and tested on Market-1501/DukeMTMC-reID.

- 1) **LOMO** [7] is a traditional discriminative method, which proposes an effective feature representation called Local Maximal Occurrence(LOMO), and a metric learning method called Cross-view Quadratic Discriminant Analysis(XQDA). We train the model on source training set by using LOMO as the feature representation and XQDA as the metric, then test on target testing set.
- 2) **BoW** [20] is a traditional method which proposes an unsupervised Bag-of-Words descriptor, in which local features are quantized to visual words using a pretrained codebook. An image is represented by a visual word histogram weighted by TF-IDF scheme. We train this model using the public code implemented by its authors.
- 3) **UMDL** [16] is an unsupervised multi-task dictionary learning method which is able to learn a dataset-shared, but target-data-biased representation. We implement the experiment following the original paper by using Market-1501 [20] and DukeMTMC-reID [18], [19] as experimental inputs.
- 4) **PUL** [15] is an unsupervised method. Standard k-means clustering is performed on the CNN features, followed by a selection operation to choose reliable samples which are used to fine-tune the original model. We train

this model using the public code implemented by its authors.

- 5) **CAMEL** [17] proposes an unsupervised asymmetric metric learning method which finds a shared space where view-specific bias is alleviated. We use its code supplied by its authors and set its parameters following the suggestions in its original paper.
- 6) **PTGAN** [13] is an unsupervised domain adaptation method which integrates the foreground mask of a person's image with CycleGAN. We train this model using the public code implemented by its authors.
- 7) **SPGAN** [14] is an unsupervised domain adaptation method which is composed of a Siamese network and a CycleGAN. We use its code supplied by its authors to train the model.
- 8) **TJ-AIDL** [24] is an unsupervised domain adaptation method which learns an attribute-semantic and identity-discriminative feature representation space transferrable to any unseen target domain. We use the results supplied by the original paper.

2) COMPARING WITH HAND-CRAFTED FEATURES

Overall, we can observe that the proposed method obtains much better performance than the hand-crafted features methods: Bag-of-Words(BoW) [20] and local maximal occurrence (LOMO) [7]. For example, LOMO [7] achieves 27.2%/12.3% rank-1 accuracy when using DukeMTMC-reID/Market-1501 as source set and tested on Market-1501/DukeMTMC-reID. Our methods improve the rank-1 accuracy by +30% (57.2-27.2), +34.7%(47-12.3) on Market-1501/DukeMCMT-ReID. Their inferiority is obvious. Both of them fail to achieve competitive results.

3) COMPARING WITH UNSUPERVISED METHODS

We compare our results with three unsupervised approaches including CAMEL [17], PUL [15], and UMDL [16]. These methods obtain better results than hand-crafted methods. For example, PUL [15] achieves 45.5% rank-1 accuracy and 20.5% mAP accuracy by using DukeMTMC-reID as source domain and Market-1501 as target domain. Overall, we can observe that our results significantly outperform

TABLE 3. We use the baseline model for feature learning. Methods comparison using Duke/Market as source, and Market/Duke as target. “Supervised learning” denotes training and testing on target domain, simultaneously. “Direct Transfer” means directly applying the baseline re-ID model trained by using images from source domain on the target domain. CycleGAN is only used to translate images from source domain to target domain, without camera based adaptation. “Baseline + CycleGAN” means that the baseline re-ID model is trained by images generated by CycleGAN and then applied on target domain. StarGAN and IPGAN are used to translate images from the source domain to the target camera domains with camera based adaptation. “Baseline + StarGAN/IPGAN” means that the baseline re-ID model is trained by images generated by StarGAN/IPGAN and then applied on target domain.

Methods	Duke → Market				Market → Duke			
	rank-1	rank-5	rank-10	mAP	rank-1	rank-5	rank-10	mAP
Supervised Learning	84.8	93.7	96.2	65.3	72.6	84.5	87.6	52.0
Baseline+Direct Transfer	44.3	61.9	69.7	18.4	30.2	45.1	51.5	16.1
Baseline+CycleGAN [14]	49.9	67.1	74.2	22.6	39.2	54.8	60.5	20.1
Baseline+StarGAN [21]	54.3	74.7	81.5	24.4	44.1	60.0	65.8	21.9
Baseline+IPGAN	56.4	75.6	82.5	25.6	46.8	62.0	67.9	25.7

TABLE 4. Performance evaluation of the proposed method when tested on MSMT17.

Methods	Src.	MSMT17			
		rank-1	rank-5	rank-10	mAP
PTGAN [13]	Market	10.2	-	24.4	2.9
IPGAN+Baseline	Market	16.8	26.4	31.7	5.5
IPGAN+IBN-reID	Market	17.7	26.9	32.7	5.8
PTGAN [13]	Duke	11.8	-	27.4	3.3
IPGAN+Baseline	Duke	20.1	30.0	34.8	6.4
IPGAN+IBN-reID	Duke	21.1	31.4	37.0	6.8

CAMEL [17], PUL [15], and UMDL [16], and achieve the highest scores in all the cases.

a: COMPARING WITH UNSUPERVISED DOMAIN ADAPTATION METHODS

We also compared three unsupervised domain adaptation approaches, including PTGAN [13], SPGAN [14] and TJ-AIDL [24]. Comparing with those methods, when tested on Market-1501, the proposed method achieves **rank-1 accuracy = 57.2% that is closed to the result of TJ-AIDL [24] and the best mAP = 28.0%**. When tested on DukeMTMC-reID, the result achieves **the best rank-1 accuracy = 47.0% and mAP = 27.0%**. The proposed method acquires a very competitive re-ID accuracy on “Market→Duke” and “Duke→Market”.

In order to further verify the effectiveness of the proposed method, we also evaluate the proposed approach on MSMT17 [13] which is a newly released dataset. The differences in pedestrian styles are more obvious in MSMT17. It is a more challenging dataset. As shown in Table 4, when use Market/Duke as source domains, our method obtains the better performance than PTGAN [13]. For example, when use Duke as the source domain, **the rank-1 and mAP accuracy are 21.1% and 6.8%, respectively**. The performance surpasses PTGAN [13] by 9.3% in rank-1 accuracy and by 3.5% in mAP.

C. FURTHER ANALYSIS AND DISCUSSIONS

We perform an ablation study to demonstrate (1) the importance of “translation between the source domain and the tar-

get camera domains” (2) the impact of the identity semantic constraint and (3) the impact of the “IBN-block”.

1) BASELINE ACCURACY

As shown in Table 3. The re-ID feature extractor is the baseline model(section 3.1). When both trained and tested on the target dataset in supervised way, excellent accuracy can be achieved. However, when trained on the source dataset and directly tested on the target dataset, the performance drops significantly. For instance, the baseline model trained and tested on Market-1501 achieves 84.8% in item of rank-1 accuracy, but drops to 44.3% when trained on DukeMTMC-reID training set and tested on Market-1501 testing set. The reason is the *bias* of data distributions between different domains.

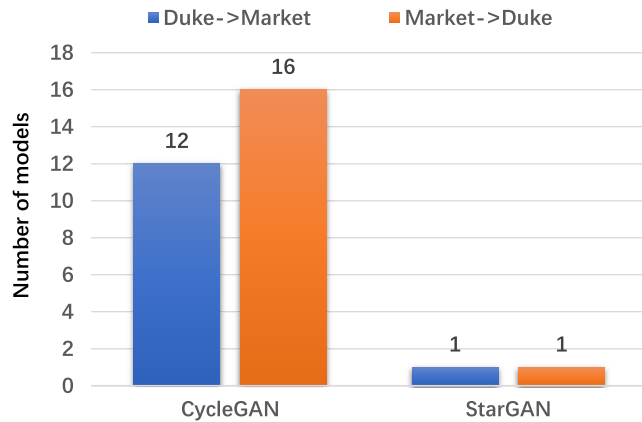
2) THE IMPACT OF “SOURCE DOMAIN TO TARGET CAMERA DOMAINS IMAGE-IMAGE TRANSLATION”

In this subsection, we aim to verify that the better solution of domain adaptation for re-ID is to reduce the bias between the source domain and the target camera domains. Firstly, following the work of [14], without regard to the adaptation between the source domain and the target camera domains, we use CycleGAN [26] to translate the labeled images from the source domain to the target domain and then train the baseline re-ID model with the translated images. As shown in Table 3, the third method “Baseline + CycleGAN” shows that when the re-ID model is trained on the style translated DukeMTMC-reID training set and tested on the Market-1501 testing set, rank-1 accuracy is 49.9% and mAP accuracy is 22.6%. When the re-ID model is trained on the style translated Market-1501 training set and tested on the DukeMTMC-reID testing set, rank-1 accuracy is 39.2% and mAP accuracy is 20.1%.

Secondly, we consider the bias between source domain and target camera domains. We use StarGAN [21] to implement style transfer from source domain to target camera domains, rather than CycleGAN [26]. Compared to CycleGAN [26], StarGAN [21] is more efficient. As shown in Fig 6, when translated from Market-1501(source domain) to DukeMTMC-reID(8 camera domains), CycleGAN [26] needs 16 pairs of $\{G, D\}$ (CycleGAN has two generator-

TABLE 5. Comparison of various domain adaptation methods over Baseline model and IBN-reID model. The results of “IBN-reID” are in bold.

Methods	Duke → Market				Market → Duke			
	rank-1	rank-5	rank-10	mAP	rank-1	rank-5	rank-10	mAP
Baseline+Direct Transfer	44.3	61.9	69.7	18.4	30.2	45.1	51.5	16.1
Baseline+StarGAN	54.3	74.7	81.5	24.4	44.1	60.0	65.8	21.9
Baseline+IPGAN	56.4	75.6	82.5	25.6	46.8	62.0	67.9	25.7
IBN-reID+Direct Transfer	45.7	62.8	70.8	19.8	32.4	47.4	54.6	17.3
IBN-reID+StarGAN	56.0	74.6	81.5	25.5	44.6	60.0	66.0	22.3
IBN-reID+IPGAN	57.2	76.0	82.7	28.0	47.0	63.0	68.1	27.0

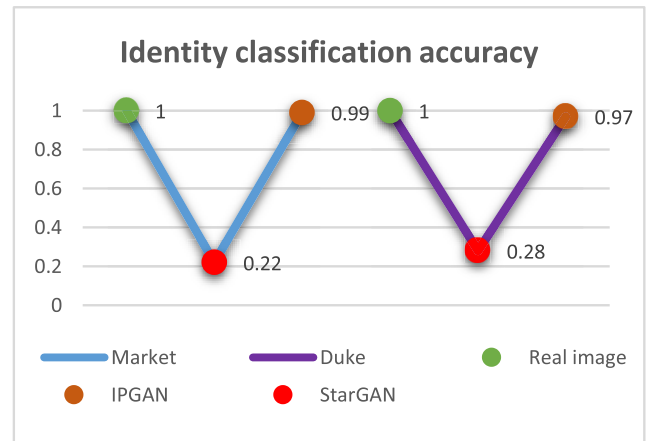
**FIGURE 6.** The number of models used by CycleGAN and StarGAN to complete style transfer from the source domain to the target camera domains. When they are used to implement multi-domain image-to-image translation, it shows that StarGAN is more efficient than CycleGAN. Therefore, StarGAN is used to implement multi-domain image-to-image translation from source domain to target camera domains. CycleGAN is only used to implement style transfer from source domain to target domain.

discriminator pairs). When translated from DukeMTMC-reID(source domain) to Market-1501(6 camera domains), it needs 12 pairs of $\{G, D\}$ on Market-1501. StarGAN [21] only uses two pairs of $\{G, D_{dom}\}$ to complete them. In this paper, CycleGAN [26] is only used to complete style transfer from source domain to target domain. Results are shown in Table 3. Compared to the “Baseline + CycleGAN”, the “Baseline + StarGAN” improves by +4.4% in item rank-1 accuracy and by +1.8% in item of mAP when tested on Market-1501. When tested on DukeMTMC-reID, the performance improves by +4.9% in item of rank-1 accuracy and by +1.8% in item of mAP. Through such a source to target camera domains adaptation method, effective improvement can be achieved.

3) THE IMPACT OF IPGAN

Compared to StarGAN [21], IPGAN was the same efficiency as StarGAN [21]

but owns identity semantic constraint. We conduct an experiment to verify the influence of the identity semantic constraint. As in Fig 7, we displayed the difference of images translated by StarGAN and IPGAN. We use DukeMTMC-reID/Market-1501 training set to train two classifiers which are used to predict the identities of translated

**FIGURE 7.** The classification accuracy of new Market-1501/DukeMTMC-reID dataset generated by StarGAN is 0.22/0.28. The classification accuracy of new Market-1501/DukeMTMC-reID dataset generated by IPGAN is 0.99/0.97.

images. The classification accuracy of DukeMTMC-reID/Market-1501 training set, after translated by StarGAN, is only 28.0%/22.0%. But, when DukeMTMC-reID/Market-1501 training set is translated by IPGAN, the classification accuracy is 97.0%/99.0%. The fake images generated by IPGAN keep almost the same identity information as the original, real images. However, most of the images generated by StarGAN lose identity semantic information. The performance of re-ID is showed in Table 3. “Baseline + IPGAN” was a greater performance than “Baseline + StarGAN” in item of rank-k and mAP accuracy.

4) THE IMPACT OF “IBN-BLOCK”

To further improve re-ID performance on target dataset, we propose IBN-re-ID model. From Fig 8, we can find that when we add the IBN-block in the baseline reID model, there is no bad impact on the ability of the network to learn the feature. The training loss is rapidly declining. As in Table 5, compared to methods with the baseline model when we use IBN-reID model as feature extractor, the rank-k and mAP accuracy are improved in a varying degree. More obviously “IBN-reID + Direct Transfer” gains rank-1 accuracy with 32.4% and mAP with 17.3% when tested on DukeMTMC-reID, surpassing the “Baseline + Direct Transfer” by +2.2% and +1.2%, respectively. The loss The similar improvement is obtained when tested on Market-1501. However,

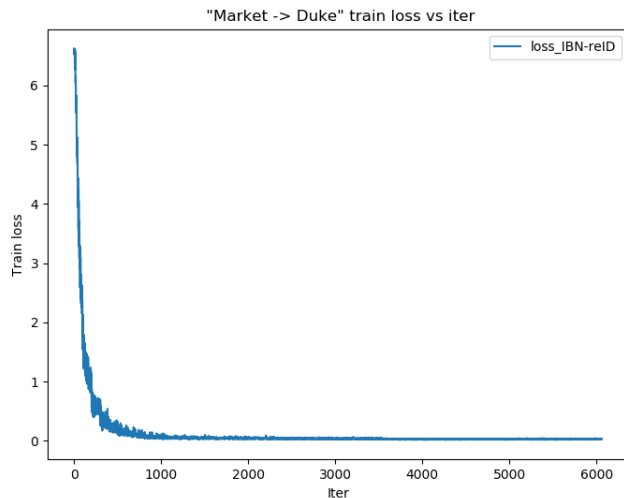


FIGURE 8. The IBN-reID model loss of training on Market-1501 but testing on DukeMTMC-reID.

compared with “Baseline + StaGAN/IPGAN”, the models with IBN-reID, i.e., “IBN-reID + StarGAN/IPGAN”, get slight improvements in term of rank-1 accuracy. The main reason is that the *bias* between source and target camera domains is significantly reduced by StarGAN and IPGAN. Thus, in IBN-reID, the instance normalization only provides limited help which eliminates appearance variance in shallow layers. This weakens the generalization capacity of IBN-reID. Even so, **the mAP values have been significantly improved and the models get the best results.**

VI. CONCLUSION

In this paper, we first propose IPGAN, a novel and efficient unsupervised domain adaptation method for person re-identification. IPGAN is used to translate images from source domain to target camera domains with identities preserving. We also propose IBN-reID model for re-ID task. IBN-reID has better generalization ability than the baseline models because it can eliminate appearance variance and maintain discrimination of the learned features. Experiments on the Market-1501, DukeMTMC-reID and MSMT17 datasets show that the proposed framework can achieve more competitive performance.

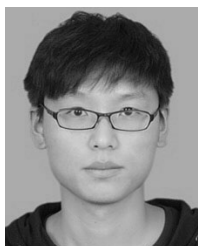
ACKNOWLEDGMENT

Jialun Liu would like to thank Zhun Zhong and Xiaohang Li for helpful discussions and encouragement.

REFERENCES

- [1] Y. Chen, F. Zhang, and W. Zuo, “Deep image annotation and classification by fusing multi-modal semantic topics,” *KSH Trans. Internet Inf. Syst.*, vol. 12, no. 1, pp. 392–412, Jan. 2018.
- [2] F. Luo, B. Du, L. Zhang, L. Zhang, and D. Tao, “Feature learning using spatial-spectral Hypergraph discriminant analysis for hyperspectral image,” *IEEE Trans. Cybern.*, vol. 49, no. 7, pp. 2406–2419, Jul. 2019.
- [3] S. Pang, J. J. del Coz, Z. Yu, O. Luaces, and J. Díez, “Deep learning and preference learning for object tracking: A combined approach,” *Neural Process. Lett.*, vol. 47, no. 3, pp. 859–876, Jun. 2018.
- [4] S.-J. Wang, H.-L. Chen, W.-J. Yan, Y.-H. Chen, and X. Fu, “Face recognition and micro-expression recognition based on discriminant tensor subspace analysis plus extreme learning machine,” *Neural Process. Lett.*, vol. 39, no. 1, pp. 25–43, 2014.
- [5] L. Zhang, L. Zhang, B. Du, J. You, and D. Tao, “Hyperspectral image unsupervised classification by robust manifold matrix factorization,” *Inf. Sci.*, vol. 485, pp. 154–169, Jun. 2019.
- [6] F. Zhu, Q. Liu, S. Zhong, and Y. Yang, “Human-level moving object recognition from traffic video,” *Comput. Sci. Inf. Syst.*, vol. 12, no. 2, pp. 787–799, 2015.
- [7] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, “Person re-identification by local maximal occurrence representation and metric learning,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2197–2206.
- [8] L. Zhang, T. Xiang, and S. Gong, “Learning a discriminative null space for person re-identification,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1239–1248.
- [9] L. Zheng, Y. Yang, and A. G. Hauptmann, “Person re-identification: Past, present and future,” 2016, *arXiv:1610.02984*. [Online]. Available: <https://arxiv.org/abs/1610.02984>
- [10] Y. Sun, L. Zheng, W. Deng, and S. Wang, “SVDNet for Pedestrian Retrieval,” 2017, *arXiv:1703.05693*. [Online]. Available: <https://arxiv.org/abs/1703.05693>
- [11] A. Hermans, L. Beyer, and B. Leibe, “In defense of the triplet loss for person re-identification,” 2017, *arXiv:1703.07737*. [Online]. Available: <https://arxiv.org/abs/1703.07737>
- [12] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, “Beyond part models: Person retrieval with refined part pooling,” 2017, *arXiv:1711.09349*. [Online]. Available: <https://arxiv.org/abs/1711.09349>
- [13] L. Wei, S. Zhang, W. Gao, and Q. Tian, “Person transfer GAN to bridge domain gap for person re-identification,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 79–88.
- [14] W. Deng, L. Zheng, G. Kang, Y. Yang, Q. Ye, and J. Jiao, “Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, vol. 1, no. 2, pp. 994–1003.
- [15] H. Fan, L. Zheng, C. Yan, and Y. Yang, “Unsupervised person re-identification: Clustering and fine-tuning,” *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 14, no. 4, p. 83, Oct. 2018.
- [16] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1306–1315.
- [17] H.-X. Yu, A. Wu, and W.-S. Zheng, “Cross-view asymmetric metric learning for unsupervised person re-identification,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 994–1002.
- [18] Z. Zheng, L. Zheng, and Y. Yang, “Unlabeled samples generated by gan improve the person re-identification baseline *in vitro*,” 2017, *arXiv:1701.07717*. [Online]. Available: <https://arxiv.org/abs/1701.07717>
- [19] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, “Performance measures and a data set for multi-target, multi-camera tracking,” in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 17–35.
- [20] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, “Scalable person re-identification: A benchmark,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1116–1124.
- [21] 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,” 2017, *arXiv:1711.09020*. [Online]. Available: <https://arxiv.org/abs/1711.09020>
- [22] X. Pan, P. Luo, J. Shi, and X. Tang, “Two at once: Enhancing learning and generalization capacities via IBN-net,” 2018, *arXiv:1807.09441*. [Online]. Available: <https://arxiv.org/abs/1807.09441>
- [23] M.-Y. Liu and O. Tuzel, “Coupled generative adversarial networks,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 469–477.
- [24] J. Wang, X. Zhu, S. Gong, and W. Li, “Transferable joint attribute-identity deep learning for unsupervised person re-identification,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2275–2284.
- [25] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” 2017, *arXiv:1611.07004*. [Online]. Available: <https://arxiv.org/abs/1611.07004>
- [26] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” 2017, *arXiv:1703.10593*. [Online]. Available: <https://arxiv.org/abs/1703.10593>

- [27] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim, "Learning to discover cross-domain relations with generative adversarial networks," 2017, *arXiv:1703.05192*. [Online]. Available: <https://arxiv.org/abs/1703.05192>
- [28] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 700–708.
- [29] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," 2013, *arXiv:1312.6114*. [Online]. Available: <https://arxiv.org/abs/1312.6114>
- [30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.
- [31] F. Wang, W. Zuo, L. Lin, D. Zhang, and L. Zhang, "Joint learning of single-image and cross-image representations for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1288–1296.
- [32] T. Xiao, H. Li, W. Ouyang, and X. Wang, "Learning deep feature representations with domain guided dropout for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 1249–1258.
- [33] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.
- [34] Z. Liu, D. Wang, and H. Lu, "Stepwise metric promotion for unsupervised video person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2448–2457.
- [35] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.
- [36] Z. Zhong, L. Zheng, Z. Zhenga, S. Li, and Y. Yang, "CamStyle: A novel data augmentation method for person re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 3, pp. 1176–1190, Mar. 2018.
- [37] A. Subramaniam, M. Chatterjee, and A. Mittal, "Deep neural networks with inexact matching for person re-identification," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 2667–2675.
- [38] J. Song, Y. Yang, Y.-Z. Song, T. Xiang, and T. M. Hospedales, "Generalizable person re-identification by domain-invariant mapping network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 719–728.
- [39] L. Song, C. Wang, L. Zhang, B. Du, Q. Zhang, C. Huang, and X. Wang, "Unsupervised domain adaptive re-identification: Theory and practice," 2018, *arXiv:1807.11334*. [Online]. Available: <https://arxiv.org/abs/1807.11334>
- [40] Z. Zhong, L. Zheng, Z. Luo, S. Li, and Y. Yang, "Invariance matters: Exemplar memory for domain adaptive person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 598–607.



JIALUN LIU received the B.S. degree from the College of Computer Science and Technology, Jilin University, Changchun, China, where he is currently pursuing the Ph.D. degree. His research interests include person re-identification and domain adaptation.



WENHUI LI is currently a Professor and a Doctoral Supervisor with the College of Computer Science and Technology, Jilin University, Changchun, China. His research interests include computer vision image processing and virtual reality.



HONGBIN PEI received the B.S. and M.S. degrees from the College of Computer Science and Technology, Jilin University, in 2012 and 2015, respectively, where he is currently pursuing the Ph.D. degree. His research interests include network data analysis and spatiotemporal data mining with applications for health informatics and intelligent transportation.



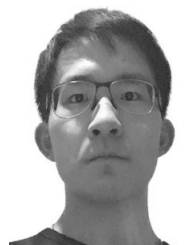
YING WANG received the Ph.D. degree from Jilin University, in 2012, where she is currently with the College of Computer Science and Technology. Her research interests include image processing, computer vision, pattern recognition, and driving environment perception.



FENG QU received the B.S. degree from the College of Computer Science and Technology, Jilin University, Changchun, China, where he is currently pursuing the Ph.D. degree. His research interests include computer vision and driving environment perception.



YOU QU received the B.S. degree from the College of Computer Science and Technology, Jilin University, Changchun, China, where he is currently pursuing the Ph.D. degree. His research interests include image processing and computer vision.



YUHAO CHEN received the B.S. and M.S. degrees from the College of Computer Science and Technology, Jilin University, where he is currently pursuing the Ph.D. degree. His research interests include computer vision and driving environment perception.

...