SBSGAN: Suppression of Inter-Domain Background Shift for Person Re-Identification

Yan Huang† Qiang Wu† JingSong Xu† Yi Zhong§
†School of Electrical and Data Engineering, University of Technology Sydney, Australia
§School of Information and Electronics, Beijing Institute of Technology, China
{yanhuang.uts, zhongyim2m}@gmail.com, {Qiang.Wu, JingSong.Xu}@uts.edu.au

Abstract

Cross-domain person re-identification (re-ID) is challenging due to the bias between training and testing domains. We observe that if backgrounds in the training and testing datasets are very different, it dramatically introduces difficulties to extract robust pedestrian features, and thus compromises the cross-domain person re-ID performance. In this paper, we formulate such problems as a background shift problem. A Suppression of Background Shift Generative Adversarial Network (SBSGAN) is proposed to generate images with suppressed backgrounds. Unlike simply removing backgrounds using binary masks, SBSGAN allows the generator to decide whether pixels should be preserved or suppressed to reduce segmentation errors caused by noisy foreground masks. Additionally, we take ID-related cues, such as vehicles and companions into consideration. With high-quality generated images, a Densely Associated 2-Stream (DA-2S) network is introduced with Inter Stream Densely Connection (ISDC) modules to strengthen the complementarity of the generated data and ID-related cues. The experiments show that the proposed method achieves competitive performance on three re-ID datasets, i.e.*, Market-1501, DukeMTMC-reID, and CUHK03, under the cross-domain person re-ID scenario.*

1. Introduction

The task of person re-identification (re-ID) is to match the identities of a person under non-overlapped camera views [10, 39, 18, 41, 24]. Most existing methods assume that the training and testing images are captured from the same scenario. However, this assumption is not guaranteed in many applications. For instance, person images captured from two different campuses have distinct illumination condition and background (BG) (*e.g.*, Market-1501 [38] and DukeMTMC-reID [29, 41] datasets). In this situation, the bias between data distributions on two domains becomes large. Directly training a classifier from one dataset (*i.e.*, source domain) often produces a degraded performance when testing is conducted on another dataset (*i.e.*, target domain). Therefore, it is important to investigate solutions for such a cross-domain issue. For person re-ID, the domain adaption solutions have drawn attention in recent years [4, 7, 35, 36, 42].

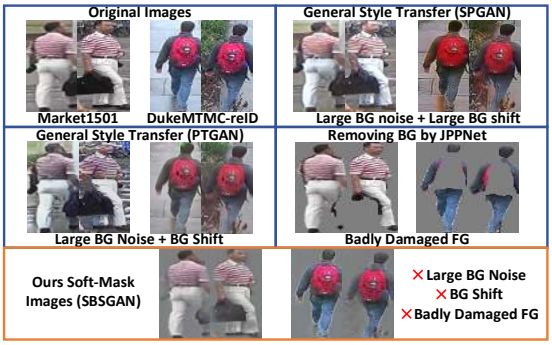


Figure 1. Comparison between different input images for cross-domain person re-ID. Images from Market-1501 and DukeMTMC-reID show distinct BG shift. Images generated by SPGAN [7] and PTGAN [36] do not suppress the BG noise, and have the BG shift problem. The hard-mask solution, *i.e.*, JPP-Net [25] damages the FG. Our SBSGAN takes all the impacts into consideration.

Recent cross-domain person re-ID methods usually adopt (or resort to) Generative Adversarial Network (GAN) to learn the domain variants [4, 7, 36, 42]. These approaches can be categorized into two main types: 1) general inter-domain style transfer [4, 7, 36]; 2) inter-camera style transfer [42]. All of them may perform well on certain cases, *i.e.*, domain style changes or camera style changes. However, they do not consider to remove or suppress BGs for reducing domain gaps. For instance, when a network is trained based on limited BG information presented in a source domain, such network may not well distinguish es-

sential pedestrian features against noise caused by BG variations in a target domain. Unfortunately, BGs in the target domain is normally very different from the source domain. In this paper, we formulate this problem as a BG shift problem which may significantly degrade the overall performance of cross-domain person re-ID.

One possible solution to sort out BG shift is to directly remove BGs using foreground (FG) masks in a hard manner (*i.e*., applying the binary masks on original images) [9, 17, 30, 33]. However, it is observed that methods, such as JPPNet [25] and Mask-RCNN [1, 13], specifically being designed for removing BG may damage the FG information. By simply removing BGs, this hard manner solution does improve the performance of cross-domian person re-ID to a certain extent (see Table 2). At the same time, it can be seen that this is still an open problem. "Is there a way to better suppress BG shift to improve cross-domain re-ID performance?" This paper makes the first effort to generate images while BGs being suppressed moderately instead of completely removing the BGs in a hard manner.

To address the problem above, a Suppression of BG Shift Generative Adversarial Network (SBSGAN) is proposed. Compared with hard-mask solutions, images generated by the proposed SBSGAN can be regarded as FG images, with BG being suppressed moderately. The generated images by SBSGAN can be called as *soft-mask images*. In addition, previous works [7, 36] show that keeping the consistency of image style between domains can improve the performance of cross-domain person re-ID. Such an idea is also integrated into our SBSGAN to further reduce the domain gap. Fig. 1 shows images selected from two different person re-ID datasets. The BGs are quite different. A model trained on one dataset may easily be biased on another one due to the BG shift problem mentioned above. Images generated by recent cross-domain re-ID approaches, such as SPGAN [7] and PTGAN [36] still present some undesirable results. If we directly use FG masks obtained by JPPNet [25] to zero out BGs, the FG can be badly damaged by the noisy masks. On the contrary, every pixel in our generated images are preserved in a soft manner. Fig. 1 shows that our SBSGAN generates visually better images which can further reduce the domain gap caused by BG shift.

In order to enhance FG information and better integrate ID-related cues into the network, we propose a Densely Associated 2-Stream (DA-2S) network. This work is to argue that certain context information, *e.g*., companions and vehicles in BG may also provide ID-related cues. Both images with suppressed BGs (our generated images) and images with full BGs are respectively fed into the two individual streams of DA-2S. Unlike previous 2-stream methods (*e.g*., [2, 5, 40]), we propose Inter-Stream Densely Connection (ISDC) modules as new components used between the two streams of DA-2S. With ISDCs, more gradients produced by the final objective function can participate to strengthen the relationship between signals coming from two different streams in the back-propagation.

The contributions of this paper can be summarized in three-fold. 1) BG shift is comprehensively investigated as an impact on cross-domain person re-ID. A SBSGAN is proposed to make the first effort by generating soft-mask images in order to reduce domain gaps. Compared with previous methods, BGs are mitigated rather than completely removed in our generated images. 2) A DA-2S CNN network with the proposed ISDC components is presented to facilitate complementary information between our generated data and more ID-related cues from the BG. 3) A comprehensive experiment is given to show the effectiveness of our soft-mask images in reducing domain gaps as well as the DA-2S model for cross-domain person re-ID.

2. Related Work

Recently, followed by image-to-image translation approaches (*e.g*., CycleGAN [43] and StarGAN [6]), some researches focus on the inter-domain style transfer to reduce domain gaps for person re-ID. Deng *et al*. [7] proposed SPGAN to transfer general image style between domains. Wei *et al*. [36] introduced PTGAN to transfer the body pixel values and generate new BGs with the similar statistic distribution of the target domain. Unlike SPGAN, PTGAN explicitly considered the BG shift problem between domains. However, PTGAN overlooked the fact that BGs should be suppressed rather than retained, because the BG shift may degrade the cross-domain re-ID performance. In addition to the inter-domain style transfer, Zhong *et al*. [42] proposed to transfer the style of images between cameras to reduce the domain gap by using StarGAN [6]. A synthetic dataset was proposed to generalize illumination between different light conditions for cross-domain person re-ID in [4]. Cycle-consistency translation of GAN was employed to retain identities of the synthetic dataset. Unlike the these approaches, our SBSGAN concentrates on the BG shift problem by generating soft-mask images amongst different domains. We also take the style consistency into consideration to further reduce the domain gap.

To deal with the BG shift problem, one possible solution is to completely remove BGs using the binary body mask obtained by semantic segmentation or human parsing methods. Currently, methods such as Mask-RCNN [13] and JPPNet [25] can obtain body masks with the pre-trained model on large-scale datasets, *e.g*., MS COCO [26] and LIP [25]. However, masks obtained by these methods often contained errors due to reasons such as low-resolution person images and highly dynamic person poses. Directly using the noisy masks may further jeopardize the cross-domain re-ID performance. Instead, we make the first effort to suppress the BG noise by generating soft-mask images. Previous work

such as [22] embeded the concept of 'soft' to learn more informative features by using probability maps of different body parts on the feature level. Our SBSGAN focuses on the data level that tries to deal with the BG shift problem for cross-domain person re-ID.

2-Stream Models have been used in many computer vision tasks [2, 5, 31, 40, 41, 19]. Generally, the learning objective of 2-stream models are categorized into two types. One verified inputs of the two individual streams belonging to the same or different classes, *e.g.*, [2, 40, 41, 19] in person re-ID and [31] in face recognition. The other type tried to enrich the representation by considering the complementarity between the inputs, *e.g.*, [5] in person search. We follow the latter type and propose a DA-2S model. Unlike the above-mentioned 2-Stream models, ISDC is introduced between two individual streams of our DA-2S to strengthen the inter-stream relationship and explore a stronger complementarity between input images.

3. SBSGAN for Soft-Mask Image Generation

3.1. Objective Functions in SBSGAN

There are two tasks in the generator (G) of SBSGAN. The main task is to generate soft-mask images with suppressed BGs. The auxiliary task is to generate inter-domain style-transferred images (retain BG) to normalize the style of soft-mask images across all the training domains. Our discriminator (D) is used to distinguish the real and fake images, and classify these images to their corresponding domains. Fig. 2 shows the proposed SBSGAN.

Specifically, given an input image (*e.g.*, $I_{\mathbb{D}_s}$) from source domain $\mathbb{D}_s$, G can generate its corresponding soft-mask image $I_{\bar{\mathbb{D}}}$ by $G(I_{\mathbb{D}_s}, \bar{\mathbb{D}}) \rightarrow I_{\bar{\mathbb{D}}}$. G takes both the input image (*e.g.*, $I_{\mathbb{D}_s}$) and an indicator (*e.g.*, $\bar{\mathbb{D}}$) as inputs. In addition, G can also transfer the style of $I_{\mathbb{D}_s}$ to the k-th ($k \neq s$) target domain $\mathbb{D}_k$ via $G(I_{\mathbb{D}_s}, \mathbb{D}_k) \rightarrow I_{\mathbb{D}_k}$. The proposed SBSGAN supports multi-domain data as inputs. If there are K domains in training, then, all $I_{\mathbb{D}_k} (k \in [1, K] \cap k \neq s)$ and the input image $I_{\mathbb{D}_s}$ will be used to normalize the style of $I_{\bar{\mathbb{D}}}$, ensuring it is consistent across all the K domains. Several loss functions are involved to train SBSGAN.

(1) ID Constraint Loss. The ID constraint (IDC) loss was proposed to preserve the underlying image information (*e.g.*, color) for data generation [32]. We use IDC loss to preserve the color of person images for the auxiliary style-transferred image generation. The IDC loss is defined as follows:

$$\mathcal{L}_{idc} = \mathbb{E}_{I_{\mathbb{D}_s}, \mathbb{D}_k} \left[\| G(I_{\mathbb{D}_s}, \mathbb{D}_k) - I_{\mathbb{D}_s} \|_1 \right]. \quad (1)$$

We observe that without the IDC loss, G may change the color of input images. Consequently, the color of generated soft-mask images are changed (see Fig. 5) when the auxiliary style-transferred images are directly applied to the soft-mask images for normalizing the style of them (see Eq. 4).

(2) Reconstruction Loss. We apply a reconstruction (REC) loss to ensure the content between an input image and its corresponding generated image remains unchanged. REC loss is a conventional objective function for the domain-to-domain image style transfer [6, 7, 36, 43]. In our soft-mask image (or style-transferred image) generation, the image content of the FG (or FG+BG) should be kept with the input image. We only expect the domain-related parts being changed by the G. The REC loss is given as follows:

$$\mathcal{L}_{rec} = \mathbb{E}_{I_{\mathbb{D}_s}, \mathbb{D}_k \vee \bar{\mathbb{D}}} \left[\left\| G(G(I_{\mathbb{D}_s}, \mathbb{D}_k \vee \bar{\mathbb{D}}), \mathbb{D}_s) - I_{\mathbb{D}_s} \right\|_1 \right], \quad (2)$$

where $\vee$ is 'or' operator.

(3) BG Suppression Loss. We propose a BG Suppression (BGS) loss to suppress BG in data generation. The BGS loss also can preserve the FG color information of the generated soft-mask images. Therefore, part of functions between IDC loss and BGS loss are similar, but concentrate on generating different types of data. The BGS loss is formulated as follows:

$$\mathcal{L}_{bgs} = \mathbb{E}_{I_{\mathbb{D}_s}, \bar{\mathbb{D}}} \left[\left\| I_{\mathbb{D}_s} \odot M(I_{\mathbb{D}_s}) - G(I_{\mathbb{D}_s}, \bar{\mathbb{D}}) \right\|_2 \right]. \quad (3)$$

An auxiliary body mask $M(I_{\mathbb{D}_s})$ is used to suppress BG of the input image $I_{\mathbb{D}_s}$. L_2 distance is applied to minimize the loss. The JPPNet [25] is employed to extract $M(I_{\mathbb{D}_s})$. We find that masks obtained by JPPNet often contain segmentation errors. However, our SBSGAN is robust to the segmentation errors in the data generation process (see Fig. 4).

(4) Style Consistency Loss. The Style Consistency (SC) Loss is proposed to encourage the style of soft-mask images (particular the part of FG) to be consistent across all the input domains, by which the domain gap of soft-mask images can be further reduced. The SC loss is given as follows:

$$\begin{aligned} \mathcal{L}_{sc} = \mathbb{E}_{I_{\mathbb{D}_s}, \bar{\mathbb{D}}, \mathbb{D}_k} [& \left\| G(I_{\mathbb{D}_s}, \bar{\mathbb{D}}) - I_{\mathbb{D}_s} \odot M(I_{\mathbb{D}_s}) \right\|_1 + \\ & \sum_{k=1, k \neq s}^{K} \left\| G(I_{\mathbb{D}_s}, \bar{\mathbb{D}}) - G(I_{\mathbb{D}_s}, \mathbb{D}_k) \odot M(I_{\mathbb{D}_s}) \right\|_1]. \end{aligned} \quad (4)$$

We first transfer the style of $I_{\mathbb{D}_s}$ to all the other $K-1$ domains. Then, $I_{\mathbb{D}_s}$ and all its corresponding style-transferred images are used to encourage the style of $G(I_{\mathbb{D}_s}, \bar{\mathbb{D}})$ being consistent across all the K domains.

Apart from the above-mentioned loss functions, we add the conventional adversarial loss ($\mathcal{L}_{adv}$) [11] of GAN to distinguish real and fake images in training. Also, $\mathcal{L}_{cls}^r$ [6] is used to classify the source domains of real images for optimizing D, and $\mathcal{L}_{cls}^f$ [6] is used to classify the target domains of fake images for optimizing G. Since the style of $I_{\bar{\mathbb{D}}}$ is normalized across all the K domains, a uniform distribution (*i.e.*, $\frac{1}{K}$) over the K domains is assigned as the target domain of $I_{\bar{\mathbb{D}}}$.

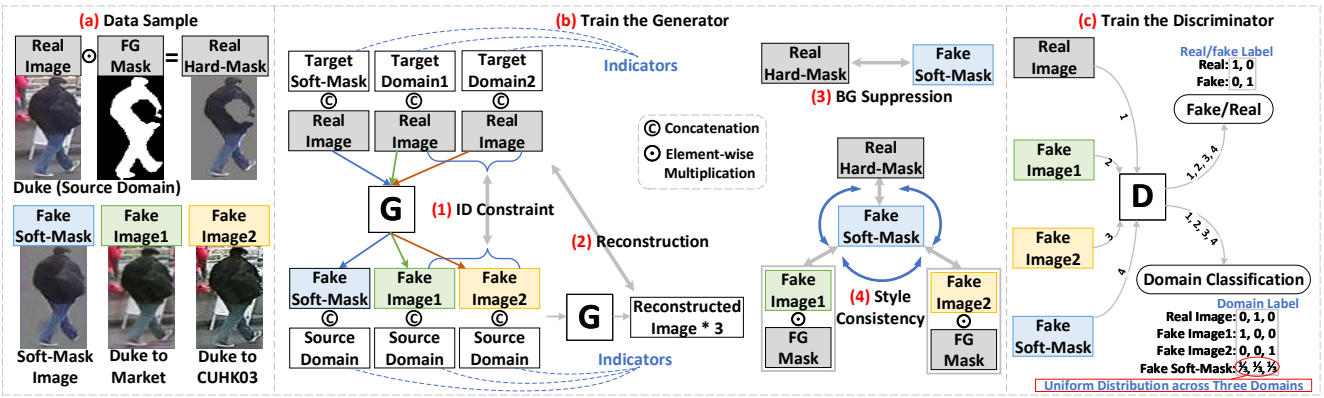


Figure 2. Overview of SBSGAN. Three domains are used as an example, including DukeMTMC-reID (source domain), Market-1501 (target domain1), and CUHK03 (target domain2). (a) shows an input image from DukeMTMC-reID. The FG mask is obtained by JPPNet. The generated soft-mask image and inter-domain style-transferred images are listed in the second row. (b) G takes both input images and indicators as the inputs. All images (real/fake) participate in the different training process of G to optimize different loss functions (1)-(4). (c) All the real and fake images are used to optimize the real/fake classification and domain classification losses in D.

Finally, the objective functions of G and D are respectively given as follows:

$$\mathcal{L}_D = \mathcal{L}_{adv} + \mathcal{L}^r_{cls}, \tag{5}$$

$$\begin{aligned}\mathcal{L}_G =& \mathcal{L}_{adv} + \mathcal{L}^f_{cls} + \lambda_{rec}\mathcal{L}_{rec} + \\ & \lambda_{idc}\mathcal{L}_{idc} + \lambda_{bgs}\mathcal{L}_{bgs} + \lambda_{sc}\mathcal{L}_{sc},\end{aligned} \tag{6}$$

where λ is hyper-parameter to control the importance of different loss functions. We empirically set $\lambda_{rec} = 10$ and $\lambda_{idc} = \lambda_{bgs} = \lambda_{sc} = 5$ in all our experiments.

3.2. Indicators in Data Generation

The proposed SBSGAN supports multi-domain images as inputs. In experiments, images from three domains/datasets are used in training. When images are fed into G, an indicator is concatenated after each image on the dimension of channel to let G knows which kind of image should be generated. A 3D tensor $\mathbb{D}$ is used as the indicator (see Fig. 2). The height and width of $\mathbb{D}$ equal to the input image. There are K channels in $\mathbb{D}$. For the auxiliary style-transferred image generation, $\mathbb{D}$ is denoted as $\mathbb{D}_k$; all values in the k-th channel of $\mathbb{D}_k$ are set to be one, and other values in the remaining $K-1$ channels are set to be zero. For the soft-mask image generation, $\mathbb{D}$ is denoted as $\bar{\mathbb{D}}$; all values of $\bar{\mathbb{D}}$ are set to be $\frac{1}{K}$.

3.3. Network Architecture

Adapted from [43], given an input image, we use two down-sampling convolutional layers followed by six residual blocks [14] in G. Then, unlike [43], two branches (without parameters sharing) are respectively used for generating soft-mask images and auxiliary style-transferred images followed by the output of the last residual block. Each branch contains two up-sampling transposed convolutional layers with the stride of 2. For D, we use the PatchGAN [21, 43] structure.

4. Densely Associated 2-Stream Network

The main contribution of this paper is to deal with the cross-domain person re-ID task from a brand new perspective, *i.e.*, suppression of the inter-domain BG shift. Moreover, to make use of helpful background cues, a DA-2S network is proposed. We argue that the context information, *e.g.*, companions and vehicles in BG is also useful in cross-domain person re-ID. Therefore, our DA-2S network is used to enrich person representations by using both our soft-mask images and the image after general inter-domain style transfer. Fig. 3 shows the DA-2S network. A pair of input images (a soft-mask image and its style-transferred image to the target domain) is fed into two ImageNet-trained Densenet-121 [16] networks (without parameters sharing). It can be observed that the companion in white clothes is regarded as BG being suppressed in the soft-mask image. To use the companion as an ID-related cue, a style-transferred images is fed into the second stream without suppressed BGs. To strengthen the complementarity of the two inputs, ISDC is proposed after the first pooling layer and every Dense Block. Specifically, the input information of each ISDC module is accumulated from both the outputs of the two streams as well as the previous ISDC module. Thus, the output of each ISDC module is defined as:

$$O^{ISDC}_n = \delta(\mathcal{F}(y \cdot O^{ISDC}_{n-1} \oplus [O^{S1}_n, O^{S2}_n], \{\mathbf{W}_n\})), \tag{7}$$

where $S1$ and $S2$ respectively represent the two streams, O^{S1}_n and O^{S2}_n are their respective output after the first pooling layer or each Dense Block, $n \in [1, 4]$ represents the in-

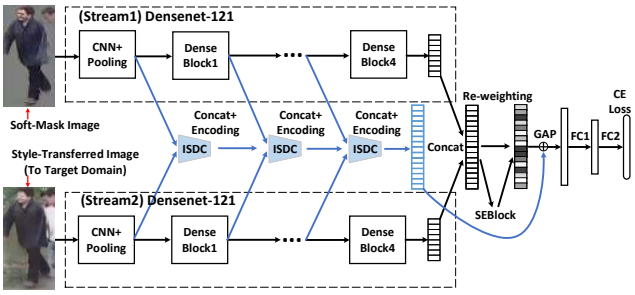


Figure 3. Overview of DA-2S. ISDC, GAP, FC, and CE respectively represent Inter-Stream Densely Connection, Global Average Pooling, Fully-Connected layer, and Cross-Entropy loss.

dex of ISDC modules, $\mathcal{F}$ is a CNN encoder parameterized by $\mathbf{W}_n$, $\oplus$ is element-wise summation, $[\cdot]$ refers to concatenation along channel dimension, y indicates whether this is the first (*i.e.*, $n = 1$) ISDC module between the two streams. If $n = 1$, $y = 0$, it refers to the first ISDC module. If $n \in [2, ..., 4]$, $y = 1$, element-wise summation is used to transfer the knowledge from one previous ISDC module to the other. δ denotes ReLU [27]. Also, Batch Normalization (BN) [20] is used before each ReLU activation function.

We re-weigh the final output of the two Densenet-121 backbone networks (after concatenation by channel) using SEBlock [15] to emphasize informative features and suppress useless ones. The output of the last ISDC is directly connected to the re-weighted feature maps by an element-wise summation. Hence, gradients produced by objective function can be directly used to update parameters of layers connected to ISDC modules. Then, a global pooling is used followed by a fully-connected layer (FC1), BN, and ReLU. Another fully-connected layer (FC2) is used with N neurons, where N is the number of training identities. At last, a cross-entropy loss is adopted by casting the training process as an ID classification problem. Notably, we use DenseNet-121 as the backbone network because in each layer it takes all preceding feature maps as input to strengthen the gradients received by all preceding layers. The proposed ISDC module is also designed to strengthen the gradients produced by the inter-streams connections. We aim to verify whether the proposed ISDC module is still workable even with dense gradients being existed in the two individual streams (refer to Table 3).

5. Experiments

In this section, comprehensive evaluations (qualitative and quantitative) are carried out to verify the effectiveness of SBSGAN and DA-2S for cross-domain person re-ID. In the qualitative evaluation, we verify the effectiveness of soft-mask images generated by SBSGAN. In the quantitative evaluation, we evaluate our soft-mask images and DA-2S for cross-domain person re-ID. Our experiment is mainly conducted on Market-1501→ DukeMTMC-reID (using Market-1501 [38] for training and DukeMTMC-reID [29, 41] for testing), since both datasets have fixed training/testing splits. In addition, other results are given on three widely used person re-ID datasets, including Market-1501, DukeMTMC-reID, and CUHK03 [24].

Table 1. Person re-ID datasets for evaluations.

Dataset	Train		Gallery (Test)		Query (Test)	
	#ID	#Img	#ID	#Img	#ID	#Img
Market [38]	751	12,936	750	19,732	750	3,368
Duke [29, 41]	702	16,522	702	17,661	702	2,228
CUHK03 [24]	1,367	13,009	100	987	100	100

5.1. Person Re-ID Datasets

Table 1 lists the training/testing settings of the three datasets. In the testing set, all query images are used to retrieve corresponding person images in the galley set. CUHK03 contains two image settings: one is annotated by hand-drawn bounding boxes, the other one is produced by a person detector. We only use and report the result of detected images which is more challenging. For all datasets, we use the single-query evaluation. The conventional rank-n accuracy and mean Average Precision (mAP) are used as evaluation protocols [38].

5.2. Implementation Details

SBSGAN. All images of the three datasets ($K = 3$) are used to train the proposed SBSGAN. Only weak domain labels are used. Input images and their corresponding body masks are resized to 256×128. Adam [23] is used with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. The batchsize is set to 16. To train G, $\frac{K+1}{16}$ images of each mini-batch are randomly selected for soft-mask images generation as well as the auxiliary style-transferred images generation. The remaining images in a mini-batch are used for the general style transfer to stabilize the performance of data generation in G. We initially set the learning rate to 0.0001 for both G and D, and model stops training after 5 epochs. We perform one G update after five D updates as in [12]. In testing, an indicator (*i.e.*, $\bar{\mathbb{D}}$) and an original image (*i.e.*, $I_{\mathbb{D}_s}$) are concatenated for the soft-mask image generation. Notably, there is no need to use any FG or body mask in testing.

DA-2S. Both soft-mask and style-transferred images (to the target domain) are used to train DA-2S (refer to Section 4). The soft-mask images are generated by the proposed SBSGAN. PTGAN [36] is used to get the general style-transferred images as the input to DA-2S. The batchsize is set to 50. Input images are resized to 256×128 with random horizontal flipping. The SGD is used with momentum 0.9. The initial learning rate is set to 0.1 and decay to 0.01 after 40 epochs. We stop training after the

Figure 4. Comparison between hard-mask and soft-mask images. Images are selected from three different person re-ID datasets. The original images are listed in the first row. The second and the third rows respectively show hard-mask images by Mask-RCNN [1, 13] and JPPNet [25]. The last row shows our soft-mask images generated by the proposed SBSGAN.

60-th epoch. A reduction rate of 16 is used for SEBlock as in [15]. A dropout layer with the rate of 0.5 is inserted after FC1 (see Fig. 3) to reduce the risk of over-fitting. The FC1 has 512 neurons. According to the number of training identities, we set FC2 to have 751, 702, and 1,367 neurons when training is conducted on Market-1501, DukeMTMC-reID, and CUHK03 respectively. For each convolutional layer of ISDC, the kernel size=3, and padding=1. In addition, we use stride=2 for the first three ISDC modules and stride=1 for the last ISDC module. The number of channels is doubled by each ISDC. Finally, 2,048 channels are obtained after four ISDC modules. In testing, original images of the target domain and their corresponding soft-mask images are used as the inputs of DA-2S. We extract 2,048-dim CNN features for each testing image after the GAP layer. The Euclidean distance is used to compute the similarity between query and gallery images.

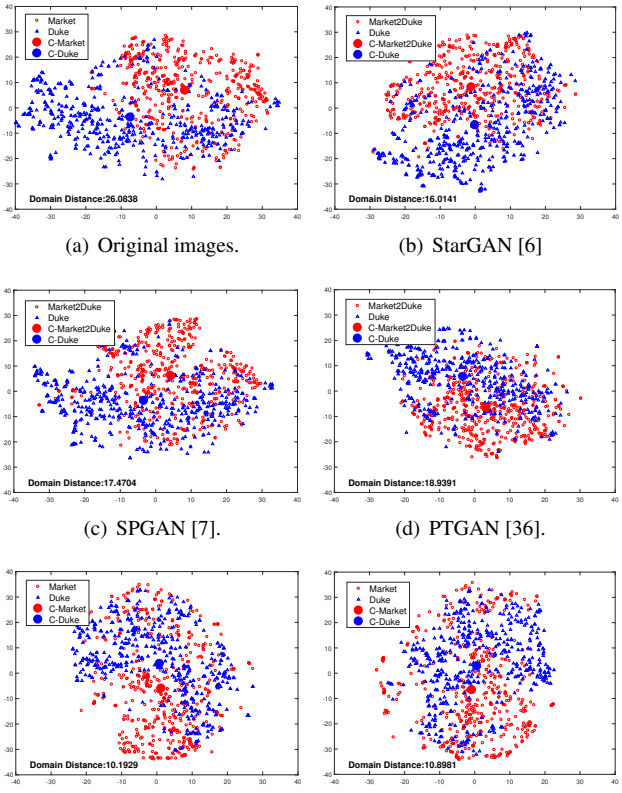
5.3. Qualitative Evaluation

Soft-Mask Images Are Better Than Hard-Mask Images in Suppression of BG Shift. In Fig. 4, we compare our soft-mask images with the hard-mask images. The hard-mask images are respectively obtained by JPPNet [25] and Mask-RCNN [1, 13]. Both methods have shown compelling performance in person parsing or object instance segmentation. However, we find that the two methods cannot perform well in the segmentation of body from the BG on existing person re-ID datasets. It can be observed in Fig. 4 that when people carry objects (*e.g.*, bags), these objects are regarded as BGs and removed by noisy FG masks with segmentation errors. However, such features are significant to person re-ID, which should be retained rather than removed. In our soft-mask images, important cues such as bags and body parts can be well generated and retained. This is because we do not directly utilize the binary body mask on original images to remove the BGs. Although we use the mask obtained by JPPNet (the third row in Fig. 4) to suppress the BG (refer to Eq. 3) in data generation, our images visually show better results. This phenomenon also shows that the proposed SBSGAN is robust to the noisy masks in the data generation.

Figure 5. The effectiveness of different loss functions.

The Effectiveness of Loss Functions in SBSGAN. The proposed SBSGAN jointly optimizes over several loss functions (see Eq. 5 and Eq. 6). Fig. 5 shows images generated by SBSGAN using different loss functions. We elaborate on the effectiveness of $\mathcal{L}_{idc}$, $\mathcal{L}_{bgs}$, and $\mathcal{L}_{sc}$. The others are conventional GAN-based loss functions, and their effectiveness is already verified by several previous works [3, 6, 12, 21, 32, 43]. It can be observed in Fig. 5 that when $\mathcal{L}_{idc}$ and $\mathcal{L}_{bgs}$ are removed, the color information of original images cannot be well preserved. In addition, the BG cannot be well suppressed. By only removing $\mathcal{L}_{sc}$, SBSGAN can generate soft-mask images which are close to our objective. The $\mathcal{L}_{sc}$ is proposed to encourage the style of generated soft-mask images being consistent (refer to Section 3). Apart from the qualitative comparison in Fig. 5, a quantitative evaluation can be found in Table 2 to further verify the effectiveness of $\mathcal{L}_{sc}$.

Reducing the BG Shift Is Effective to Reduce Domain Gaps: Visualization of Data Distributions Between Two Domains. We visualize the domain distance using different types of data, including the popular style-transferred images, hard-mask images, and our soft-mask images. Three recently published methods SPGAN [7], PTGAN [36], and StarGAN [6] are used to transfer the im-



(a) Original images. (b) StarGAN [6]

(c) SPGAN [7]. (d) PTGAN [36].

(e) Hard-mask images (JPPNet). (f) Soft-mask images (Ours).

Figure 6. Data visualization. 5000 images are randomly selected from Market-1501 and DukeMTMC-reID to learn data distributions via the Barnes-Hut t-SNE [34], respectively. Another 200 images of each domain are used for visualization. The red circle and blue triangle respectively represent images belonging to Market-1501 and DukeMTMC-reID. The center points (*i.e.*, 'C-') are shown using their corresponding domain color. Domain distance (*i.e.*, L_1 distance) is given between center points.

age style from Market-1501 to DukeMTMC-reID, respectively. Fig. 6 shows the result. Compared with the general style-transferred results, the hard-mask and soft-mask images can reduce the domain gap by a large margin. This phenomenon verifies the effectiveness of reducing domain gaps by considering the BG shift problem. The domain distance of hard-mask images is on par with our soft-mask images (10.19 *vs*. 10.90). However, compared with hard-mask images, our soft-mask images show better performance in cross-domain person re-ID (*e.g.*, rank-1: 43.3% *vs*. 38.6%, see Table 2). Naturally, it is unfair to directly compare the domain distance between soft-mask and hard-mask images. This is because many pixel values of hard-mask images are simply zeroed out, which makes approximately half the information of hard-mask images already being discarded in the comparison. Our soft-mask images suppress the BGs rather than simply removing them.

Table 2. Baseline performance of cross-domain person re-ID. Market-1501 is for training and DukeMTMC-reID is for testing.

Training Data	mAP	R-1	R-5	R-10
Original	17.7	33.5	49.3	55.1
Hard-mask Images				
Mask-RCNN [1, 13]	20.6	37.5	53.4	59.1
JPPNet [25]	21.5	38.6	54.3	60.0
Style-transferred Images				
PTGAN [36]	22.7	42.9	58.0	64.2
SPGAN [7]	**22.8**	42.0	57.9	64.1
StarGAN [6]	21.6	39.8	53.4	59.9
Soft-mask Images (Ours)				
Soft-mask w/o $\mathcal{L}_{sc}$	21.2	41.7	56.3	62.7
Soft-mask	22.3	**43.3**	**58.2**	**64.4**
Soft-mask$_{2-Domains}$	**23.5**	**44.2**	**59.5**	**65.3**

Table 3. Ablation study of DA-2S. Market-1501 is used for training and DukeMTMC-reID is used for testing. The baseline does not use SEBlocks and any ISDC modules. We also try to add SEBlocks to every ISDC module to re-weight the output of ISDC in the middle layers (denoted as ISDC-SE). The DA-2S† (DA-2S‡) means only using the style-transferred images (soft-mask images) as the inputs of the 2-stream network.

Methods	mAP	R-1
Basel.	28.8	50.2
Basel.+SEBlock	28.9	50.5
Basel.+SEBlock+ISDC-SE	**30.4**	**51.5**
Basel.+SEBlock+ISDC (DA-2S)	**30.8**	**53.5**
DA-2S† (2*Style-transfer)	28.4	49.6
DA-2S‡ (2*Soft-mask)	27.0	51.5

5.4. Quantitative Evaluation

Soft-mask Images *vs*. Other Types of Images. The popular IDE model [7, 41] with ImageNet-trained DenseNet-121 as backbone network is adopted to compare our soft-mask images with the general style-transferred images and hard-mask images. Table 2 lists the performance. By directly using the original images for cross-domain learning, the performance is inferior (mAP: 17.7%, rank-1: 33.5%). A clear performance improvement is achieved by simply removing BGs from both training and testing images using masks obtained by JPPNet and Mask-RCNN, respectively. However, the performance of our soft-mask images outperforms the hard-mask images by +4.7% in rank-1 accuracy (43.3% *vs*. 38.6%). This is because hard-mask images often involve segmentation errors. General style-transferred results such as PTGAN and SPGAN achieve competitive performance. However, our soft-mask images obtain the best rank-1 accuracy (43.3%), which shows their effectiveness by considering the BG shift problem in cross-domain person re-ID. In addition, without $\mathcal{L}_{sc}$, images generated by SBSGAN can satisfy the visual requirement (see Fig. 5), but the cross-domain re-ID performance is dropped by 1.1% in mAP and 1.6% in rank-1 accuracy. This is because we use $\mathcal{L}_{sc}$ to normalize the style of soft-mask images across multiple domains, by which the inter-domain gap can

Table 4. Comparison with state-of-the-art methods. M, C, and D respectively represent Market-1501, CUHK03, and DukeMTMC-reID. X→Y means training is conducted on X and testing is conducted on Y.

Methods	M→ D		M→ C		D→ M		D→ C		C→ M		C→ D	
	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1
UMDL [28] *CVPR16*	7.3	18.5	-	-	12.4	34.5	-	-	-	-	-	-
CAMEL [37] *ICCV17*	-	-	-	-	26.3	54.5	-	-	-	-	-	-
PUL [8] *TOMM18*	16.4	30.0	-	-	20.5	45.5	-	-	18.0	41.9	12.0	23.0
PTGAN [36] *CVPR18*	-	27.4	-	26.9	-	38.6	-	24.8	-	31.5	-	17.6
SPGAN$_{LMP}$ [7] *CVPR18*	26.4	**46.9**	-	-	26.9	58.1	-	-	-	-	-	-
TJ-AIDL [35] *CVPR18*	23.0	44.3	-	-	26.5	58.2	-	-	-	-	-	-
HHL [42] *ECCV18*	**27.2**	**46.9**	-	-	**31.4**	**62.2**	-	-	**29.8**	**56.8**	**23.4**	**42.7**
DA-2S (Ours)	**30.8**	**53.5**	**32.5**	**42.2**	**27.3**	**58.5**	**27.3**	**33.7**	**28.5**	**57.6**	**27.8**	**47.7**

be further reduced. Since SPGAN and PTGAN only support images of two domains as inputs, we also train our SBSGAN in the same way instead of using images from three domains. Without interference from images of the third domain (*i.e.*, CUHK03), we obtain performance gains by Soft-mask$_{2-Domains}$ (mAP: 23.5%, rank-1: 44.2%). However, we still use multiple domains as inputs in all the other experiments to generate soft-mask images. This is because we can train only one model instead of multiple models between any two domains.

Ablation Study of DA-2S. An ablation study of our DA-2S network is given in Table 3. Without SEBlock and ISDC (*i.e.*, baseline), we achieve 28.8% in mAP and 50.2% in rank-1 accuracy. By using SEBlock (similar to [5]), the performance is improved from 50.2% to 50.5% in rank-1 accuracy. To strengthen the inter-stream relationship, the baseline+SEBlock+ISDC produces the best performance (mAP: 30.8%, rank-1: 53.5%), demonstrating the effectiveness of the proposed ISDC modules. If we add SEBlock to every ISDC modules (ISDC-SE), the performance is dropped by 2% in rank-1 accuracy. This is because additional SE-Blocks produce more parameters which can potentially increase the risk of over-fitting. Moreover, we also change the inputs of our 2-stream DA-2S to style-transferred images or soft-mask images only (*i.e.*, the network receives two style-transferred images or two soft-mask images). The results demonstrate that the combination of the two types of images is better than using them independently.

Comparison With State-of-the-Art Methods. We compare our method with several recently published state-of-the-art approaches, including three unsupervised methods, *i.e.*, UMDL [28], CAMEL [37], and PUL [8], and four cross-domain re-ID approaches, *i.e.*, PTGAN [36], SPGAN+LMP [7], TJ-AIDL [35], and HHL [42]. For a fair comparison, all the selected cross-domain methods (including our method) use images from one domain/dataset for training the re-ID model and the other domain/dataset for testing; no extra training images or strong labels are used from the target domain.

Table 4 lists the comparison results. It is clear to see that our DA-2S method achieves very competitive performance. For instance, on M→D, our method outperforms the state-of-the-art method HHL by +3.6% in mAP and +6.6% in rank-1 accuracy; on C→D, our performance is higher by +4.4% in mAP and +5.0% in rank-1 accuracy. Compared with our method, the HHL achieves the best performance on D→M and competitive performance on C→M. However, HHL uses extra camera labels in the target domain. Specifically, N times images are generated according to the number of cameras to learn about the camera invariant features. This inherently limits its expansibility to the large camera networks (*e.g.*, $N = 100$), where the training data should be increased by N (*e.g.*, 100) times. Amongst all the methods, only PTGAN gives the performance on M→C and D→C. Under the same experimental setting, our DA-2S outperforms PTGAN by a large margin (+15.3% and +8.9% in rank-1 accuracy) when training is respectively conducted on Market-1501 and DukeMTMC-reID, and testing is conducted on CUHK03.

6. Conclusion

In this paper, we verify that the BG shift problem can be considered to reduce domain gaps for cross-domain person re-ID. SBSGAN is proposed to generate soft-mask images with the BG being suppressed. Compared with hard-mask solutions, soft-mask images are able to suppress the BG in a moderate way. Compared with general inter-domain style-transferred approaches, soft-mask images can further reduce the domain gap by considering the BG shift problem. A DA-2S model is introduced along with the proposed ISDC module to make use of helpful background cues. Experiment results demonstrate the effectiveness of our method in both the qualitative and quantitative evaluations.

Acknowledgment

This research is supported by an Australian Government Research Training Program Scholarship.

References

[1] Waleed Abdulla. Mask r-cnn for object detection and instance segmentation on keras and tensorflow. https://github.com/matterport/Mask_RCNN, 2017.

[2] Ejaz Ahmed, Michael Jones, and Tim K Marks. An improved deep learning architecture for person re-identification. In *CVPR*, 2015.

[3] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *ICML*, 2017.

[4] Slawomir Bak, Peter Carr, and Jean-François Lalonde. Domain adaptation through synthesis for unsupervised person re-identification. In *ECCV*, 2018.

[5] Di Chen, Shanshan Zhang, Wanli Ouyang, Jian Yang, and Ying Tai. Person search via a mask-guided two-stream cnn model. In *ECCV*, 2018.

[6] Yunjey Choi, Minje Choi, and Munyoung Kim. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *CVPR*, 2018.

[7] Weijian Deng, Liang Zheng, Guoliang Kang, Yi Yang, Qixiang Ye, and Jianbin Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person reidentification. In *CVPR*, 2018.

[8] Hehe Fan, Liang Zheng, Chenggang Yan, and Yi Yang. Unsupervised person re-identification: Clustering and fine-tuning. *TOMM*, 2018.

[9] Michela Farenzena, Loris Bazzani, Alessandro Perina, Vittorio Murino, and Marco Cristani. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*, 2010.

[10] Shaogang Gong, Marco Cristani, Shuicheng Yan, and Chen Change Loy. *Person re-identification*. Springer, 2014.

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014.

[12] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *NeurIPS*, 2017.

[13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[15] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, 2018.

[16] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *CVPR*, 2017.

[17] Yan Huang, Hao Sheng, and Zhang Xiong. Person re-identification based on hierarchical bipartite graph matching. In *ICIP*, 2016.

[18] Yan Huang, Hao Sheng, Yanwei Zheng, and Zhang Xiong. Deepdiff: Learning deep difference features on human body parts for person re-identification. *Neurocomputing*, 2017.

[19] Yan Huang, Jingsong Xu, Qiang Wu, Zhedong Zheng, Zhaoxiang Zhang, and Jian Zhang. Multi-pseudo regularized label for generated data in person re-identification. *TIP*, 2018.

[20] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015.

[21] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017.

[22] Mahdi M Kalayeh, Emrah Basaran, Muhittin Gökmen, Mustafa E Kamasak, and Mubarak Shah. Human semantic parsing for person re-identification. In *CVPR*, 2018.

[23] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[24] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, 2014.

[25] Xiaodan Liang, Ke Gong, Xiaohui Shen, and Liang Lin. Look into person: Joint body parsing & pose estimation network and a new benchmark. *TPAMI*, 2018.

[26] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[27] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *ICML*, 2010.

[28] Peixi Peng, Tao Xiang, Yaowei Wang, Massimiliano Pontil, Shaogang Gong, Tiejun Huang, and Yonghong Tian. Unsupervised cross-dataset transfer learning for person re-identification. In *CVPR*, 2016.

[29] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *ECCV*, 2016.

[30] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. Mask-guided contrastive attention model for person re-identification. In *CVPR*, 2018.

[31] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deeply learned face representations are sparse, selective, and robust. In *CVPR*, 2015.

[32] Yaniv Taigman, Adam Polyak, and Lior Wolf. Unsupervised cross-domain image generation. In *ICLR*, 2016.

[33] Maoqing Tian, Shuai Yi, Hongsheng Li, Shihua Li, Xuesen Zhang, Jianping Shi, Junjie Yan, and Xiaogang Wang. Eliminating background-bias for robust person re-identification. In *CVPR*, 2018.

[34] Laurens van der Maaten. Accelerating t-sne using tree-based algorithms. *JMLR*, 2014.

[35] Jingya Wang, Xiatian Zhu, Shaogang Gong, and Wei Li. Transferable joint attribute-identity deep learning for unsupervised person re-identification. In *ICCV*, 2018.

[36] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *CVPR*, 2018.

[37] Hongxing Yu, Ancong Wu, and Weishi Zheng. Cross-view asymmetric metric learning for unsupervised person re-identification. In *ICCV*, 2017.

[38] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015.

[39] Zhedong Zheng, Xiaodong Yang, Zhiding Yu, Liang Zheng, Yi Yang, and Jan Kautz. Joint discriminative and generative learning for person re-identification. In *CVPR*, 2019.

[40] Zhedong Zheng, Liang Zheng, and Yi Yang. A discriminatively learned cnn embedding for person reidentification. *TOMM*, 2017.

[41] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *ICCV*, 2017.

[42] Zhun Zhong, Liang Zheng, Shaozi Li, and Yi Yang. Generalizing a person retrieval model hetero-and homogeneously. In *ECCV*, 2018.

[43] Junyan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *CVPR*, 2017.