Object and background disentanglement for unsupervised cross-domain person re-identification

Yuehua Zhu, Cheng Deng *, Huanhuan Cao, Hao Wang
School of Electronic Engineering, Xidian University, Xi’an 710071, China

ABSTRACT

Person re-identification (re-ID) is an important task in many application fields. While most previous works have conducted feature embedding under the supervision of a prior information, it is also true that data collection and annotation in real-world scenarios are very expensive. Moreover, although researchers have borrowed transferring knowledge to alleviate the dependence on labeling, the bias among various domains remains an open problem. Accordingly, motivated by the observation that reducing the style and background differences between domains can promote the generalization capability of the learning model, this paper proposes an unsupervised model, namely, person component decomposition and synthesis generative adversarial network (PCDS-GAN), to minimize the distribution gap among multiple person re-ID datasets. More specifically, we first disentangle the pedestrian image into foreground, background and style features, then use these features to synthesize person images with various backgrounds from the target domain. Finally, the synthesized images are used to train person re-ID models. Comprehensive experiments demonstrate that our model can effectively reduce the domain gap, and also outperforms state-of-the art methods on the Market-1501 and CUHK03 benchmarks.

ARTICLE INFO

Article history:
Received 2 December 2019
Revised 17 February 2020
Accepted 17 April 2020
Available online 22 April 2020
Communicated by Chenping Hou

Keywords:
Person re-identification
Cross-domain
Generative adversarial network

1. Introduction

Generally speaking, person re-identification (re-ID) involves matching people across non-overlapping camera views. This process plays an important role in many applications, including video surveillance, pedestrian retrieval, and human-computer interaction.

Although considerable progress in this field has been made, re-ID remains a challenging problem due to the effects of background clutter, lighting conditions, occlusion, etc. To address such issues, existing methods typically focus on extracting robust features or finding measures of similarity. For instance, semantic features from different regions of the body can be captured via a multi-stage region of interest (ROI) pooling pipeline [1]. Hierarchical LSTM [2] (HLSTM) enables more complex representation of visual data to capture information at different scales. Local Maximal Occurrence (LOMO) learns a stable feature representation invariant to deal with changes between different cameras [3]. Moreover, in [4], a semantics-aware image representation is learned to capture the intrinsic structural information of the persons. An enhanced triplet loss function [5] has also been proposed to learn a distance measure between two pedestrian images. Furthermore, some works focusing on subspace learning [6,7] to learn more discriminative representation by mapping raw image data into a projection space. For instance, an approach known as DVAML [8] uses metric learning to project the image pairs into different feature subspaces according to their view information.

Since label collection is often expensive or even impossible in real-world scenarios, several unsupervised models have been developed for this purpose. For instance, in [9,10], the authors aimed to discover and utilize the spatial-temporal patterns in re-ID datasets to guide the models to incrementally predict IDs. PUL [11] introduces a progressive unsupervised learning method to transfer pretrained deep representations to unseen domains. Moreover, to eliminate the bias between different datasets, SPGAN [12] synthesizes identity-preserving training samples of the target domain style to train the re-ID model.

However, the aforementioned methods ignore the influence of the background discrepancy between domains. In fact, deep neural networks (DNNs) are easily biased by similar backgrounds and the domain gap caused by background difference will lead to the poor performance of unsupervised cross-domain person re-ID. Since existing datasets typically contain multiple images of humans against similar backgrounds that are captured by a small number of cameras, the background-bias phenomenon can easily be
observed when analyzing the two popular datasets from Fig. 1. As shown in Fig. 1, the in-domain backgrounds are always similar, but the cross-domain backgrounds are quite different; therefore, the learning models, especially DNNs trained on one dataset usually perform poorly on other datasets with different backgrounds. We argue that such a phenomenon is a major drawback for existing cross-domain re-ID and was not thoroughly investigated before.

Accordingly, we propose a novel background-transferable person re-identification framework that greatly enhances the cross-domain re-ID model in this paper. We present a component decomposition and synthesis generative adversarial network (PCDS-GAN) to synthesize identity-preserving training samples with the background of the target domain. The main contributions of this paper are threefold:

- A novel PCDS-GAN model is proposed to synthesize source labeled images with target domain-specific information (backgrounds) in order to bridge the domain gap and thus facilitate more effective domain adaptation.
- To fill arbitrary missing regions on synthesized images, a hole-filling module (HFM) is further proposed, which makes our synthesized images more realistic.
- We report the competitive accuracy achieved by our method compared to state-of-art methods on two large-scale person re-ID datasets, Market-1501 and CUHK03.

2. Related work

This work is closely related to descriptor learning in person re-ID and image generation with GAN. In this section, we briefly summarize representative works dealing with these two categories.

2.1. Person re-ID

Along with the rapid development of deep learning (DL) [13], many DL-based methods [14–18] have been proposed; these can be divided into supervised and unsupervised categories. In [19], a person detection and an identification model are integrated into one framework. A two-stream model was proposed so that the global and local features [20] could be jointly learned. Moreover, inspired by the progress in pose estimation [21], some works [22,1] have proposed several pose-based person re-ID methods. It must be noted that these supervised methods reveal that backgrounds are significantly influential on person identification. Therefore, Song et al. [23] leverages the binary masks of images obtained using a popular segmentation model [24] to guide the re-ID model so that it can learn more discriminative features of the human body. By contrast, in this work, we adopt masks to produce a foreground human part and a background part, which are used to randomly synthesize cross-domain images, and then train a re-ID model with these synthesized images using unsupervised transfer learning.

As for unsupervised person re-ID, earlier works [25–27] have mostly leveraged hand-crafted features directly. However, some more recent works [9,10] have utilized spatial-temporal information existing in re-ID datasets to guide the model in predicting IDs incrementally. With the settings that the source domain is fully labeled and the target domain does not contain labels, many unsupervised domain adaptation (UDA) have been proposed. For example, the approach HHL [28] enforces camera invariance and domain connectedness simultaneously to enable more generalizable embeddings to be learned on the target domain. PTGAN [29], adopts a translation model, namely CycleGAN [30], to synthesize training samples of the target domain style. Similarly, SPGAN [12] proposes a framework based on unsupervised image-image translation to train the re-ID model with the underlying (latent) ID information of the foreground pedestrian. These GAN-based UDA methods mainly focus on transferring the style of the target domain to reduce the domain gap, which neglect background bias between domains. Unlike abovementioned methods, we disentangle an image into different parts and then synthesize background-transferred samples to improve cross-domain re-ID performance.

2.2. Generative adversarial networks

The GAN [31] framework has been widely used in various tasks recently, including natural style transfer, super-resolution, image completion, image-to-image translation, etc. A conditional adversarial network was introduced to learn the mapping function from input to output images [32]. However, this method requires paired training data, which makes it unsuitable for many tasks. In an attempt to tackle the task of unpaired image-to-image translation, cycle consistency loss is proposed to train unpaired data [30]. Moreover, in [33], the authors proposed an AdaIN layer for arbitrary style transfer. In this work, we employ GANs to align the distribution of the translated images with the real images in the target domain without constraint of paired data, which simultaneously enables the generation of diverse outputs from a given source domain image. On the other hand, in deep learning research, learning an interpretable latent space representation has been a prevalent focus, especially in the field of generative model [34,35]. Thus, a lot of works which focus on the disentanglement of object appearance [36] or style [37] have been proposed. In contrast, we disentangle the pedestrian image into latent spaces of foreground, background and style, and three types of disentangled features in their corresponding latent space are learned.

3. The proposed framework

Huang et al. [34] proposed a shared latent space assumption. While they assumed that there is only one specific space, we postulate that only part of the latent space (body) can be shared across domains, whereas the other parts (style and background) are domain-specific. This is a more reasonable approach for dealing with cross-domain mapping for person re-ID tasks.

The architecture of our proposed PCDS-GAN is illustrated in Fig. 2. Our model consists of a foreground convolution architecture $E_f$, a background convolution architecture $E_b$, a style convolution architecture $E_s$, a U-Net $U$ which constitutes HFM, and a decoder $G$ for domains $S$ and $T$. Given labeled dataset $S = \{(x_i, y_i, m_i)\}_{i=1}^{n_i}$ of $n_i$ images from the source domain and unlabeled dataset $T = \{(\tilde{x}_i, \tilde{m}_i)\}_{i=1}^{n_T}$ of $n_T$ unlabeled images from the target domain,
PCDS-GAN first disentangles the image into two parts using a mask \( m \), i.e., a person foreground image \( x_f \) and background image \( x_b \). We then extract the features of the body image, the raw background image and the style of the original image \( x \), which are represented as \( f \), \( b \) and \( q \) respectively. Finally, these features can be decoded into the original image. When the raw background image is filled appropriately after the segmentation and the feature map of the body is exchanged, the translated images \( x^{t \rightarrow s} \) and \( x^{s \rightarrow t} \) can be synthesized.

### 3.1. Hole-Filling Module (HFM)

Without the use of a sophisticated GAN, we can clip out the persons using a pedestrian detector in the source domain and add the foreground directly to the background of the target domain. However, the following two problems will occur if this approach is adopted. First, directly superimposing the data over a combination of two different distributions makes the synthesized data being affected by artifacts, which is not conducive for the cross-domain feature learning model. Second, simply superimposing pedestrians on the background will result in a large number of holes appearing on the background.

Since GANs are inherently capable of synthesizing unseen data, our PCDS-GAN can synthesize realistic images that are closer to the distribution of the target domain. The HFM is an indispensable component of our model, as shown in Fig. 3. The goal of the HFM is to fill in pixels of the scene occupied by the foreground in \( x \). While we tried to use \( x_0 \) as the direct input to the U-Net, we found that the texture and color information of the synthesized background were not realistic in Fig. 3 (b). Accordingly, after considering that noise provides high-frequency gradients that are useful for texture synthesis [38], we utilize a background image where the foreground is filled with Gaussian noise as the input to U-Net, and Fig. 3 (a) shows more visually satisfactory result.

More visual results are shown in Fig. 4. Comparing Fig. 4 (b) and Fig. 4 (d), it can be seen that using HFM can greatly improve the quality of the synthesized images, which conducts to the training of the re-ID model by reducing the noise in the data. In addition, comparing Fig. 4 (c) and Fig. 4 (d) can show the benefits of using noise in HFM.

Finally, we can get the foreground \( x_f \), the original background \( x_b \) and the filled background \( x_{bu} \), which can be denoted as:

\[
\begin{align*}
    x_f &= x \otimes m, \\
    x_0 &= x \otimes (1 - m), \\
    x_{bu} &= U(x \otimes (1 - m) + \mathcal{N}(0, 1) \otimes m),
\end{align*}
\]

where \( U \) represents the U-Net and \( \otimes \) denotes element-wise multiplication. \( \mathcal{N}(0, 1) \) and \( m \) refer to the normal Gaussian noise and foreground mask respectively. We multiply each background mask \((1 - m)\) pixelwise across all three channels of \( x \) to obtain the original masked background images.

### 3.2. PCDS-GAN

Fig. 5 shows an overview of our model and its learning process. As shown in Fig. 5 (a), the latent feature of each auto-encoder is factorized into a style feature \( q^d \), a foreground feature \( f^d \) and a background feature \( b^d \), i.e., \((q^d, b^d, f^d) = (E^d_f(x^f), E^d_b(x^b), E^d_f(x^f))\). It should be here that \( d \in (s, t) \) indicates domain. To obtain a self-reconstructed image, \( G^d \) can be utilized to synthesize the original image \( \hat{x}^d = G^d(f^d, b^d, q^d) \).

Image-to-image translation is performed by swapping cross-domain foregrounds and filling the in-domain background, as illustrated in Fig. 5 (b). For example, in order to translate an image \( x^t \in T \) to \( T \), we first extract the foreground feature \( f^t \), along with the filled background feature \( \mathcal{F}^t = E^s_b(x^b) \) and randomly extract a style latent feature \( q^t \) from a prior distribution \( \mathcal{N}(0, 1) \). To create the composite image containing the source foreground, target background and style, we then use \( G^t \) to synthesize the final translated image \( x^{t \rightarrow s} = G^t(f^t, \mathcal{F}^t, q^t) \).
3.2.1. Bidirectional reconstruction loss and GAN loss

To ensure that the encoders and decoders are inverse, we design losses that encourage reconstruction in both the image and latent directions. Given domain self-reconstructed image \( x^s \), we should thus be able to reconstruct it after encoding and decoding. The loss is formulated as

\[
L_{xs rec} = \mathbb{E}_s x^s / C_0 / x^s - x^t / C_1
\]

Given a set of latent features (style, background or foreground) sampled from the latent distribution during the translation process, we should be able to reconstruct it after decoding and encoding. The loss is formulated as

\[
L_{xs rec} = \mathbb{E}_s x^s / C_0 / x^s - x^t / C_1
\]

We note the other loss terms \( L_{xs rec}, L_{xt rec}, L_{xs rec} \) and \( L_{xs rec} \) are defined in a similar manner. The style reconstruction loss \( L_{xs rec} \) is reminiscent of the latent reconstruction loss used in prior works [39–42], which encourages diverse outputs given different style features. The foreground reconstruction loss \( L_{xs rec} \) and the background reconstruction loss \( L_{xs rec} \) encourage the translated image to preserve the semantic content of arbitrary paired images.

GANs are employed to match the distribution of translated images with the target data distribution of the background and illumination. The loss is defined as follows:

\[
\mathcal{L}_{GAN} = \mathbb{E}_{x^t \sim \mathbb{P}_T} \left[ \log (1 - D^t (x^t)) \right] + \mathbb{E}_{x \sim \mathbb{P}_S} \left[ \log D^t (x) \right],
\]

where \( D^t \) is a discriminator that aims to distinguish between the translated and real images in \( T \). The discriminator \( D^t \) and loss \( \mathcal{L}_{GAN} \) are defined similarly.
3.2.2. Identity-preserving loss
As mentioned in [12], GANs may change the color of input images without identity-preserving loss, which is undesirable for re-ID feature learning. To encourage the translated image to preserve the ID-related cues and transfer the intersecting background during the process of translation, an additional identity loss is introduced in the proposed PCDS-GAN, which is beneficial to subsequent cross-domain re-ID feature learning.

\[ L_{id} = E_{`} \left( \left\| \left( x^{d} - x^{e} \right) \otimes m \right\|_1 \right) \]

(5)

In our experiments, we observe in Fig. 4 (b) that adding the two deterministic constraints to PCDS-GAN greatly improves the visual effect of synthesized images compared to Fig. 4 (a), while \( L_{id} \) is defined in a similar manner.

3.2.3. Overall objective of PCDS-GAN
Finally, we incorporate all abovementioned losses into a joint function as

\[
\text{arg min}_{\theta} \max_{\gamma} L_{\text{GAN}}^{\text{src}} + L_{\text{GAN}}^{\text{trg}} + \lambda_s \left( L_{\text{rec}}^{\text{src}} + L_{\text{rec}}^{\text{trg}} \right) \\
+ \lambda_d \left( L_{\text{adv}}^{\text{src}} + L_{\text{adv}}^{\text{trg}} + L_{\text{adv}}^{\text{rec}} + L_{\text{adv}}^{\text{rec}} + L_{\text{adv}}^{\text{rec}} + L_{\text{adv}}^{\text{rec}} \right) \\
+ \lambda_i \left( L_{\text{id}}^{\text{src}} + L_{\text{id}}^{\text{trg}} + L_{\text{id}}^{\text{rec}} \right),
\]

(6)

where the hyper-parameters \( \lambda_s, \lambda_d, \lambda_i \) are weights that control the importance of different types of losses.

3.3. Training with Synthesized Images
Given the translated dataset with labels from source domain, supervised feature learning methods can be applied to train the re-ID models. To reduce the noise caused by the image synthesis process, we add a soft labeling scheme to the cross-entropy loss for the synthesized samples. The cross-entropy loss with label smoothing regularization (LSR) is derived as follows:

\[ L_{\text{LSR}} = -(1 - \varepsilon) \log (p(y)) + \frac{\varepsilon}{N} \sum_{n=1}^{N} \log (p(k)), \]

(7)

where \( n \in \{1, 2, \ldots, N\} \) denotes the pre-defined classes of the training data, \( N \) is the number of source domain classes, and \( y \) is the ground truth. \( \varepsilon \in [0, 1] \) denotes our confidence in the sample with ground truth, which is a hyper-parameter that can be adjusted according to the quality of the synthesized images.

3.4. Network architecture

3.4.1. Foreground and background content encoder
Our content encoder consists of several strided convolutional layers to downsample the input, along with several residual blocks to further process it. All convolutional layers are followed by Instance Normalization (IN) [43]. The two content features are then concatenated together and fed into a 1-by-1 convolution network to reduce the number of dimensions. In addition, our U-Net is an encoder-decoder neural net with skip connections.

3.4.2. Style encoder
The style encoder includes several strided convolutional layers, followed by a global average pooling layer and a fully connected layer. We do not use IN layers in the style encoder, since IN removes the original feature mean and variance, which represent important style information.

3.4.3. Decoder
The decoder uses an MLP to produce a set of AdaIN parameters from the style feature. The concatenated content features are then processed via residual blocks with AdaIN layers, then finally decoded to the image space by means of upsampling and convolutional layers.

3.4.4. Discriminator
We employ the multi-scale discriminators proposed by [44] to guide the generators to synthesize both realistic details and correct global structure.

3.5. Datasets and evaluation protocols
We evaluate our method on two public datasets: Market-1501 [45] and CUHK03 [46]. These two datasets make up two source—target pairs: Market-1501→CUHK03.

Market-1501 contains 32,668 labeled images of 1,501 identities collected from six camera views. The dataset is split into two fixed parts: 12,936 images from 751 identities for training and 19,732 images from 750 identities for testing. In testing, 3,368 hand-drawn images from 750 identities are used as queries to retrieve the matching persons in the database. Single-query evaluation is used.

CUHK03 contains 1,467 identities, each of which is captured from two cameras. There are 14,096 person images in total. Following [47], we adopt the new training/testing protocol to split the dataset into two balanced parts: 767 identities are allocated to the training set, while the remaining 700 identities are allocated to the testing set.

3.6. Implementation details

3.6.1. PCDS-GAN training and testing
All experiments are implemented using an NVIDIA TITAN XP GPU with 12 GB memory. The input images are all resized to 256 x 256. We use the Adam optimizer with \( \beta_1 = 0.5, \beta_2 = 0.999 \), and an initial learning rate of \( 10^{-4} \). We also use a batch size of 1 and set the loss weights to \( \lambda_s = 10, \lambda_d = 10, \lambda_i = 1 \). The learning rate is halved every 15000 iterations, and training is stopped after 180000 iterations.

In the process of image generation, arbitrary paired samples \((x', m', y')\), \((x, m')\) are taken as inputs. We employ the encoder \( E_f \) to extract the source foreground content, and further employ the encoder \( E_f, E_s \) for the target dataset to extract the filled background content and style vector. We then feed them into the generator \( G \) to synthesize the translated image \( x' \), accompanied with label \( y' \).

Having discovered that the synthesized images are distorted when the masks of the input images are sparse, we calculate the proportion of the foreground in all masks and the statistical results are shown in Fig. 7. There are a small number of low-occupied masks in each dataset, which may not truly capture the foreground of the person. It is obvious that images that do not contain a human body have a highly adverse effect on the training of the re-ID model. Accordingly, to obtain the full foreground, a mask replacement strategy is further proposed. As long as the proportion of human body area in the source domain mask is less than the parameter \( \alpha \), we will set the value of \( m' \) to be all 1.

Moreover, given a sample from the source domain, \( K \) sample images from the target domain can be arbitrarily picked to generate K images. As shown in Fig. 6, we can achieve the purpose of data augmentation, which is complementary to various data augmentation methods including random flipping and random erasing. Random flipping is a technique commonly used in CNN training to improve the robustness to image flipping and object
translation. Random erasing is designed to enable invariance to occlusions. Unlike the above methods, our data augmentation strategy expands the background diversity in order to promote the improved extraction of human features by the re-ID model.

3.6.2. Feature learning method

Our goal is to train the re-ID model using translated images. We adopt IDE+ as the feature learning method using translated images. For IDE+, we employ the architecture and training strategy outlined in [48]. We set $\epsilon = 0.1$ in Eq. 7. All images are resized to 256 × 128. Two data augmentation methods, namely random cropping and random horizontal flipping, are employed during training. The dropout probability is set to 0.5. In most of person re-ID works, ResNet-50 [49] is used as the features extraction network. We thus adopt it as our backbone. The core of our method is to study the impact of background-transferred samples on cross-domain re-ID, so we do not pay much attention to the impact of different backbones on features extraction and focus on one commonly used backbone for a fair comparison. The learning rate begins at 0.01 for ResNet-50 base layers and 0.1 for the two newly added full connected layers. We use the SGD solver to train the re-ID model and set the batch size to 128. The learning rate is divided by 10 after 40 epochs; we train 50 epochs in total. In testing, we extract the output of the Pool-5 layer for use as the image descriptor (2,048-dim) and use the Euclidean distance to compute the similarity between images on the test set of the target domain.

4. Experiments

4.1. Baseline evaluation

In this section, we evaluate the influence of domain bias and further demonstrate that our PCDS-GAN is capable of overcoming this bias through a series of data synthesis strategies. Results are presented in Table 1.

The supervised learning method is trained and tested on the same domain, which yields good results without domain bias. In the direct transfer, a re-ID model is trained on the source domain,
then directly applied to the target domain without the use of any domain adaptation technique. We can observe a large performance drop when a source-trained model is used directly on the target domain. For example, the re-ID model trained and tested on Market-1501 achieves a rank-1 accuracy of 85.2%; however, this drops to 40.1% when trained on CUHK03 and tested on Market-1501. A similar drop occurs when CUHK03 is used as the target domain.

In our baseline, we adopt PCDS-GAN for source-target image translation. As shown in Table 1, the baseline achieves substantial improvement over the direct transfer method on the target dataset. For example, compared with direct transfer, the PCDS-GAN baseline gains 5% improvement in rank-1 accuracy on Market-1501. On CUHK03, moreover, the performance gain of PCDS-GAN is 4.5%, while gains of 3.8% can be observed in rank-1 accuracy and mAP, respectively. To eliminate the impact of low confidence mask, the mask replacement strategy is introduced into PCDS-GAN in the sample generation stage. When tested on Market-1501 and CUHK03, the rank-1 performance gain brought about by setting the threshold $\alpha = 0.3$ is 0.5% and 0.3%, respectively. Moreover, when we apply data augmentation and set the value $K = 2$, some further improvements can be observed: the rank-1 performance achieved are 46.0% and 12.1% for Market-1501 and CUHK03, respectively.

4.2. Parameter analysis

4.2.1. Analysis of the hyper-parameter $\alpha$

Although the object segmentation technique has already been perfected, in complex scenes, a phenomenon still arises such that the human body cannot be properly segmented, which leads to segmentation errors in the human body region. Such poorly segmented body will lose a lot of discriminative human information and this is why we usually opt not to split the human body directly when training a re-ID model, as indicated by [23]. Different with [23], which combines segmented body images with complete images to extract more discriminative pedestrian features, we propose to utilize a parameter $\alpha$ as noted in Section 3.6) to screen out low-quality masks. Specifically, a proportion of human body area in the source domain mask $m^s$ will be compared with the threshold $\alpha$. Once the $\alpha$ is more than the proportion, that is the ratio of $A \rightarrow B$ among which $A$ means the sum of the number of pixels of the body area in the binary mask $m^b$ and $B$ means the sum of the number of all pixels, we will reset the value of mask $m^b$ to be all 1, which retains the complete source domain image. The results of differing $\alpha$ values on Market-1501 and CUHK03 are shown in Fig. 8 (a) and Fig. 8 (b), respectively.

As the value of $\alpha$ increases, the performance of the model also continues to improve, which implies that our mask replacement strategy can effectively preserve the unclear human bodies. However, when the value of $\alpha$ is too large, too much of the source domain foreground is retained, such that less background of the target domain appears in synthesized images, which causes the re-ID model to deteriorate. This phenomenon fully demonstrates the effectiveness of reducing the domain bias by transferring the background. In summary, it is necessary to strike a balance between maintaining human integrity and background transfer. Positive background transfer can effectively improve the generalization performance of the cross-domain re-ID model when the intact human body area is preserved as much as possible. Our method achieves the best performance when $\alpha$ is set to 0.3 for both datasets.

4.2.2. Analysis of the hyper-parameter $K$

Our PCDS-GAN can flexibly generate pedestrian images with arbitrary target domain background, as shown in Fig. 6. Thus, we propose another important parameter, namely $K$, which is conducive to the generation of more samples with variable background changes. Specifically, for each image in the source domain, we can randomly sample $K$ images from the target domain to synthesize images of the same person in different scenes. The experiment results with varying $K$ are presented in Fig. 8 (c) and Fig. 8 (d) for Market-1501 and CUHK03 respectively. Best performance is achieved when $K = 2$, which suggests transferring background can be used as a complement to data augmentation techniques. It can also be concluded that the performance will not necessarily increase given more samples. For this reason, we can assume that forcing the model to learn features of an unchanged character in a lot of different scenes is likely to cause overfitting, and then damage the performance of the model.

4.3. Comparison results and discussions

The above experiments have demonstrated that the proposed PCDS-GAN can achieve satisfactory performance. The next step is to compare the proposed model with state-of-the-art methods on two popular re-ID datasets. Results of these comparisons are shown in Tables 2 and 3.

4.3.1. Results on Market-1501

To verify that the background-transferred samples synthesis are more effective than other sample generation methods, we compare our method with recent domain adaptation works SPGAN [12], HHL [28] and PTGAN [29], which mainly focus on reducing domain gap by transferring style. As shown in Table 2, our model

<table>
<thead>
<tr>
<th>Methods</th>
<th>CUHK03 → Market-1501</th>
<th>Market-1501 → CUHK03</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>rank-1</td>
<td>rank-5</td>
</tr>
<tr>
<td>Supervised Learning</td>
<td>85.2</td>
<td>93.2</td>
</tr>
<tr>
<td>Direct Transfer</td>
<td>40.1</td>
<td>57.4</td>
</tr>
<tr>
<td>PCDS-GAN ($\alpha = 0. K = 1$)</td>
<td>45.1</td>
<td>62.4</td>
</tr>
<tr>
<td>PCDS-GAN ($\alpha = 0.3, K = 1$)</td>
<td>45.6</td>
<td>62.8</td>
</tr>
<tr>
<td>PCDS-GAN ($\alpha = 0.3, K = 2$)</td>
<td>46.0</td>
<td>63.2</td>
</tr>
</tbody>
</table>
achieves the top performance by 46.0% on rank-1 and 20.8% on mAP. Compared to SPGAN and HHL, rank-1 results increase by 1.5% and 4.8% respectively, and mAP increases by 1.3% and 0.5%. Compared to PTGAN, moreover, our method achieves an improvement in rank-1 of 1.4%. We also compare our model with the existing unsupervised learning methods PUL [11] and UMDL [50] as well as the traditional methods BoW [45] and LOMO [3], and our results show better superiority. The comparison results suggest that transferring background is competitive to reduce the domain gap, which indicates that background interference will affect the pedestrian feature extraction of current deep learning model. From the results, we can also conclude that existing data collection in re-ID research community is flawed because the camera lacks a mechanism to focus on specific person regions. Moreover, although the segmentation technique can remove the background of a pedestrian image, it still can generate segmentation errors when meeting complex scenes. Therefore, once the two issues are addressed, we may not need to achieve the goal of domain adaptation from the perspective of transferring background.

4.3.2. Results on CUHK03
For transfer learning and domain adaption methods, most works do not have the training setup Market-1501—CUHK03.

Table 2
Comparison of various methods on Market-1501.

<table>
<thead>
<tr>
<th>Methods</th>
<th>CUHK03 → Market-1501</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>rank-1</td>
</tr>
<tr>
<td>BoW [45]</td>
<td>35.8</td>
</tr>
<tr>
<td>LOMO [3]</td>
<td>27.2</td>
</tr>
<tr>
<td>UMDL [50]</td>
<td>34.5</td>
</tr>
<tr>
<td>PUL [11]</td>
<td>41.9</td>
</tr>
<tr>
<td>PTGAN [29]</td>
<td>38.6</td>
</tr>
<tr>
<td>HHL [28]</td>
<td>42.2</td>
</tr>
<tr>
<td>SPGAN [12]</td>
<td>44.5</td>
</tr>
<tr>
<td>PCDS-GAN</td>
<td><strong>46.0</strong></td>
</tr>
</tbody>
</table>

Accordingly, based on existing reported results, we conduct comparisons with several methods including the SPGAN [12], PUL [11], LOMO [3] and BoW [45]. As shown in Table 3, our method outperforms the comparison unsupervised methods by a slight margin in terms of both rank-1 accuracy and mAP, further demonstrating the superiority of the proposed method.

Table 3
Comparison of various methods on CUHK03.

<table>
<thead>
<tr>
<th>20pt Methods</th>
<th>Market-1501 → CUHK03</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>rank-1</td>
</tr>
<tr>
<td>BoW [45]</td>
<td>2.1</td>
</tr>
<tr>
<td>LOMO [3]</td>
<td>0.6</td>
</tr>
<tr>
<td>PUL [11]</td>
<td>7.6</td>
</tr>
<tr>
<td>SPGAN [12]</td>
<td>11.1</td>
</tr>
<tr>
<td>PCDS-GAN</td>
<td><strong>12.1</strong></td>
</tr>
</tbody>
</table>

5. Conclusion
In this paper, we first propose a background-transferable deep network framework, named PCDS-GAN, which transfers various backgrounds from one dataset to another. We further show that, by using image generation technique based on a synthesized background, our method has the potential to solve the domain background bias among existing re-ID datasets. Furthermore, our approach is complementary to current data augmentation methods, which can be utilized to enhance the feature learning model. Extensive experimental results show that the proposed method is effective and achieves state-of-the-art results.

Declaration of Competing Interest
The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.
CRediT authorship contribution statement

Yuehua Zhu: Methodology, Software, Writing - original draft.
Cheng Deng: Conceptualization, Software, Writing - review & editing.
Huanhuan Cao: Validation, Investigation.
Hao Wang: Writing - review & editing.

References


Yuehua Zhu is now a Ph.D. student at the School of Electronic Engineering, Xidian University, Xi’an, China. His research interests focus on computer vision, pattern recognition, and multi-modal analysis.

Cheng Deng (M’11) received the B.E., M.S., and Ph.D. degrees in signal and information processing from Xidian University, Xi’an, China. He is currently a Full Professor with the School of Electronic Engineering at Xidian University. He is the author and coauthor of more than 70 scientific articles at top venues, including IEEE T-NLS, T-IP, T-CYB, T-MM, T-CSVT, Neurocomputing, ICCV, CVPR, ICMI, NeuIPS, KDD, AAAI. His research interests include computer vision, pattern recognition, and deep learning.
Huanhuan Cao received the B.E. degree from the School of Electronic Engineering, Xidian University, Xi'an, China, in 2017. She is currently pursuing the degree of Master at the School of Electronic Engineering, Xidian University. Her research interests focus on computer vision, deep metric learning, and person re-identification.

Hao Wang received the B.E degree in Electronic and Information Engineering from Hangzhou Dianzi University, China, in 2015. He is currently pursuing his Ph.D. degree at School of Electronic Engineering, Xidian University. His main research interests include human action recognition, video understanding, and zero shot learning.