PGAN: Part-based Nondirect Coupling Embedded GAN for Person Re-identification

Yue Zhang*
Institute of Information Science, Beijing Jiaotong University, Beijing Key Laboratory of Advanced Information Science and Network Technology

Yi Jin*
School of Computer and Information Technology, Beijing Jiaotong University

Jianqiang Chen
School of Mechanical Engineering, University of Guizhou

Shichao Kan, Yigang Cen
Institute of Information Science, Beijing Jiaotong University, Beijing Key Laboratory of Advanced Information Science and Network Technology

Qi Cao
School of Computing Science, University of Glasgow

***Abstract*—The block-based representation learning method has been proven to be a very effective method for person re-identification (Re-ID), but the features extracted by the existing block-based approach tend to have a high correlation among different blocks. Also, these methods perform less well for persons with large posture changes. Thus, Part-based Nondirect Coupling (PNC) representation learning method is proposed by introducing a similarity measure loss to constrain features of different blocks. Moreover, Part-based Nondirect Coupling Embedded GAN (PGAN) method is proposed, which aims to extract more common features of different postures of a same person. In this way, the extracted features of the network are robust for posture changes of a person, and there are no auxiliary pose information and additional computational cost required in the test stage. Experimental results on public datasets show that our proposed method achieves good performances, especially, it outperforms the state-of-the-art GAN-based methods for person Re-ID.**

■ **PERSON RE-IDENTIFICATION** (Re-ID) is an important application in security and intelligent surveillance system. Re-ID aims to search for the same person under different cameras by the characteristics of the person's body shape, appearance, and pose. Initially, person Re-ID methods

mainly based on traditional algorithms, including manually extract visual features and similarity measures. Different from the traditional methods, the deep learning method can automatically extract better features and learn better similarity metrics, which significantly improves the performance of Re-ID. However, persons in real scenes are often obscured by moving targets or static objects. Also, the distance from the person to the camera is not fixed, which will cause low-resolution objects and extensive scale variations. Furthermore, a same person may have a large deformation under different cameras when he is in different postures. Different persons' dress, posture, body shape, and appearance may be very similar. Thus, person Re-ID is still a hot issue in recent years. Both the accuracy and speed are needed to be further explored.

Representation learning [1], [2], [3] is the most commonly used method of person recognition because it is fast for training and easy for convergence, which treats the recognition task as a classification task or a verification task. The methods based on representation learning can extract features with discernment to separate different persons, but they are not performing well for a same person with large posture changes. Specifically, when different persons wear similar clothes, the similarity between these two persons may be higher than the similarity of a same person under different poses. The current popular methods for solving such problems are the block-based method and the align-based method. But these methods perform less well to identify the same person in different poses.

In order to reduce the influence of the pose difference and body occlusion, generative adversarial network (GAN) was employed in the field of Re-ID. It is well known that the GAN can be applied to generate various poses of a person under the supervision of various pose-related information. During the training process, when GAN generates another pose according to the features of the input image, the extracted features are demanded to exist in different pose images. So the extracted features of the GAN network are robust for posture changes of a person (pose-unrelated) at the test stage. [4], [5], [6] applied GAN to generate new images as a supplement of the training set, which failed to extract pose-unrelated features because they trained GAN and recognition networks separately. The FDGAN [7] pays attention to features extraction, but its encoder adopts the siamese structure that extracts the same features for the paired samples, which fails to capture some discriminative detail features.

In light of the above observations, we propose a novel Part-based Nondirect Coupling Embedded GAN (PGAN) method. Different from the symmetrical structure (encoding and decoding networks) of the previous GAN, we employ the block-based representation learning network to replace the encoder of the original generator. The block features obtained by current methods may be highly correlated and are not complementary to each other during the test stage. For this, we propose a Part-based Nondirect Coupling (PNC) representation learning method, which calculates the similarity of any two blocks and similarity minimization. The proposed PNC is used as our encoder of the GAN to obtain more distinguishable features. Then, the image decoder can generate new person images based on the pose information and the input person features. But the fake images by existed image decoder have some problems such as lacking some necessary personal details (e.g. color, outline, and logo). To solve these problems, Bilinear interpolation and Deconvolution methods (BD) are fused for upsampling operations, and we employ SSIM loss to improve the generated image quality. In the discriminator stage, identity discriminator and pose discriminator are utilized [7]. The code is available at `https://github.com/IvyYZ/PGAN`.

The contributions of this paper are as follows:

- A new Patial-based Nondirect Coupling (PNC) representation learning method is developed, which employs metric loss to minimize the feature similarity of different parts of an image. It can not only obtain the distinguishing features of different identities, but also provide richer information to GAN such that the generated images contain more details and to be more real.
- A new decoder that combines the bilinear interpolation feature maps and the deconvolution feature maps is proposed, which can improve the quality of the generated images.

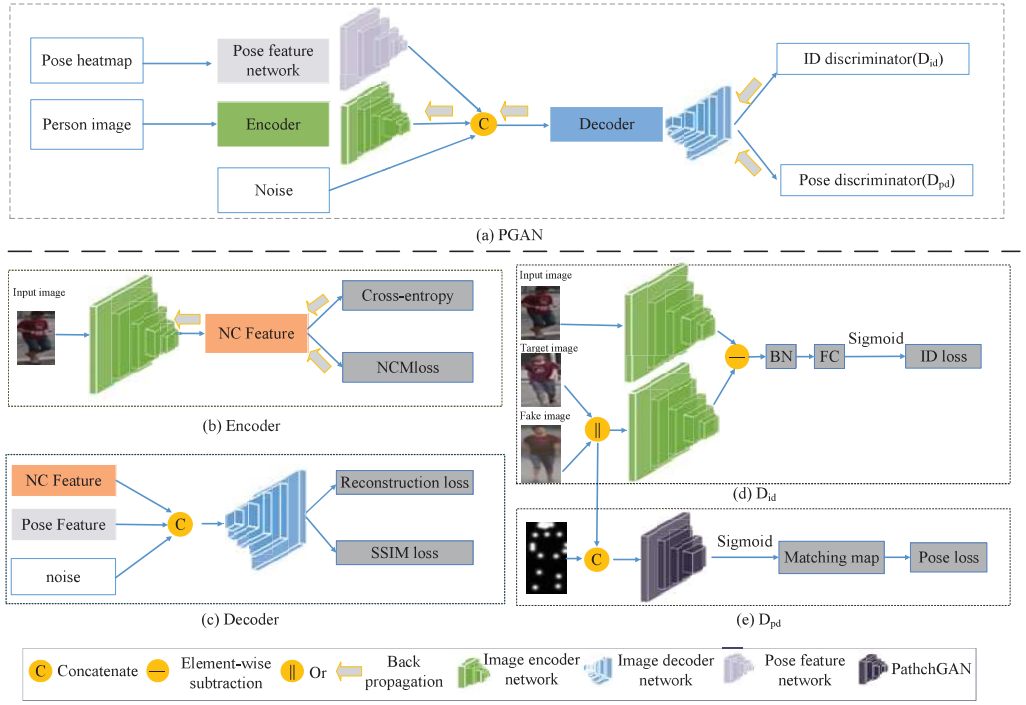

Figure 1. A schematic overview of PGAN Network. (a) The whole framework of PGAN. (b) The structure of encoder. (c) The structure of decoder. (d) The structure of D_{id}. (e) The structure of D_{pd}.

- Additionally, we improve our performance by introducing the SSIM (the Structural Similarity Index) loss [8] in the generator loss. It can not only improve the convergence speed, but also can generate a more realistic image.

- We propose a novel PGAN method by combined the PNC representation learning network with the improved GAN network. It can extract identity-related and pose-unrelated features without additional computational complexity in the testing phase.

The rest of this paper is organized as follows. In Section "RELATED WORK", we discuss representation learning and GAN for person re-identification. The idea, framework, and details of the proposed PGAN are represented in Section "PROPOSED METHOD". In Section "EXPERIMENTAL RESULTS", datasets and the experimental results are presented. Section "CONCLUSION" concludes the paper.

RELATED WORK

With the development of deep learning, representation learning has become a common strategy for person Re-ID, which transfers the person Re-ID task into a classification task or verification task rather than computing the similarity of images. Instead of handcrafted features, Li *et al.* [1] adopt CNN features for Re-ID, Kan *et al.* [9], [10] fused handcrafted feature information into CNNs for image retrieval and Re-ID. But the features extracted by these methods are global features, which lack discriminative detail features. For this, some part-based representation methods were proposed, [2] is based on human body region guided multi-stage feature decomposition. In addition, HA-CNN [11] combined local features and global features, PCB[2] and person alignment [12] applied position alignment in Re-ID. These methods effectively reduced the errors of similarity matching. Although representation learning is easy to convergence, the features extracted by the existing block-based approach tend to have a high

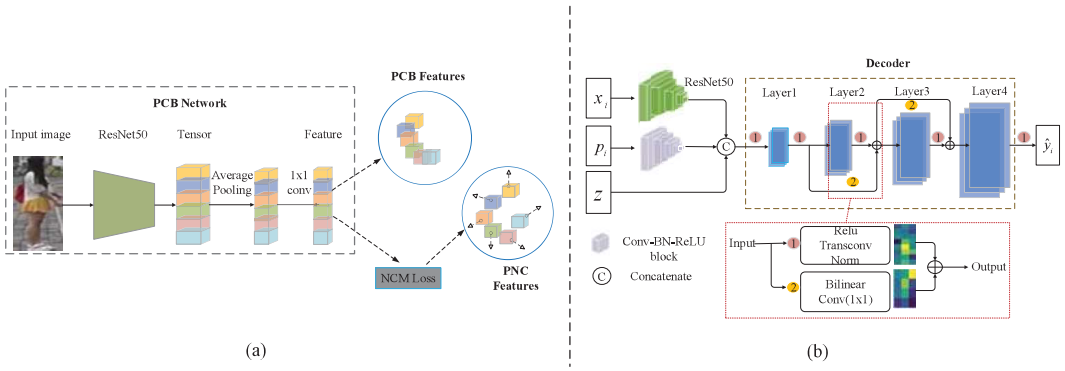

Figure 2. (a) Encoder. Comparison of feature extraction processes of PCB and our PNC methods. (b) The main process of decoder.

correlation among different blocks. Also, it paid more attentions to extract discriminative features of different pedestrians without considering a same pedestrian with large posture changes.

Generative adversarial network (GAN) has gained great attentions in various fields, since it was proposed in 2014. In the field of person recognition, Zheng *et al.* [4] firstly used the GAN network to generate persons to expand the data set. Then Wei *et al.* [13] developed a pose transferrable person Re-ID framework, which utilizes the generated images as hard negative samples of the training set to enhance Re-ID model. None of these methods focused on extracting more discriminative features for person recognition. PNGAN [13] includes a generator and a discriminator, which merges the features of the generated image with the original image features for feature diversity, and it is time-consuming during the testing phase. Although FDGAN in [7] paid attention to extract features, it fails to capture discriminative detail features. [14] proposed an asymmetrically GAN to make the generated images of higher quality. Motivated by these works, we combine the representation learning and GAN to extract robust features for person Re-ID. Different from FDGAN adopt the siamese network as the encoder, we adopt our proposed PNC representation learning method, which can extract more different features. In the decoder phase, instead of the deconvolution method for upsampling operation in FDGAN, we combine the bilinear interpolation method with the deconvolution method to perform upsampling, so that the generated images contain more details.

PROPOSED METHOD

In this paper, we propose a Part-based Nondirect Coupling (PNC) representation learning method, which adopts a novel Nondirect Coupling Metric loss (NCM loss) to minimize the similarity of any two blocks, such that extracted block features are low correlation and more discriminative. Then, in order to recognize the same person in different poses, we embed the proposed PNC network into the pose-related GAN network (PGAN). As same as the GAN consists of three modules, our PGAN has an encoder module, a decoder module, and two discriminator modules D_{pd} and D_{id}. The overall framework is described in Figure 1. Because the key issue of Re-ID is extracting robust features under different poses and backgrounds of a same person, we employ the GAN structure in the training stage to train the feature extraction network, i.e., the Encoder in Figure 1, such that the encoder can learn intrinsic features of a same person under different poses. After the encoder is trained, the features extracted by the encoder can be devoted to Re-ID. In Figure 1, the encoder and decoder modules consist an image generator. The discriminator module is responsible for distinguishing real ground-truth images and fake generated images (appearance and pose). In the testing stage, we only exploit the encoder to extract the features without additional computational cost. From Figure 1, we can see that the extracted features in the encoder are not only affected by the forward propagation of the

encoder, but also back propagations. Back propagation in the encoder encourages the network to extract discerning, rich features of different images. Also, in the decoder and discriminators, it serves the network to learn pose-unrelated features.

Network Architecture

Encoder In this work, the proposed PNC representation learning network is used as the encoder of the generator. It includes a basic network and two losses, as shown in Figure 1(b). As same as PCB network [2], it employs ResNet50 as the backbone network, the obtained three-dimensional tensor is equally divided into six blocks in the horizontal direction before the pooling layer, as shown in Figure 2(a). For each block, average pooling and an 1x1 convolution operation are performed. After the above operations, we can get a 256-dimensional column vector for each block. It should be noted that during the testing and GAN stages, the image features are obtained by concatenating the features of the six blocks.

But the block features obtained by the PCB network tend to have a high correlation among different blocks. So we propose the PNC representation learning method, which can get partial independent features by importing Nodirect Couple Metric (NCM) loss to constrain features of each block. Specifically, we calculate the cosine similarity between any two blocks of an image. By constraining the similarity between two block features, the network can learn different features from different blocks. For the convenience of comparison, we draw the feature extraction schematic diagram of PCB and PNC together, as shown in Figure 2(a).

The pose feature extraction network consists of a 5-block Conv-BN-ReLU sub-network, as shown in the purple part of Figure 2(b). Here, Figure 2(b) illustrates the decoder network structure in detail, the left part represents the input of the decoder network, the right part represents the decoder network structure. The pose map is represented by a map of 18 channels[7], which is obtained by the OpenPose detection toolkit. Then the pose map is encoded by a 5-block Conv-BN-ReLU sub-network to obtain a 128-dimensional pose feature vector.

In Figure 2(b), the input images are denoted as $X = \{x_i\}_{i=1}^{N}$, the target images are $Y = \{y_i\}_{i=1}^{N}$, and the generated images are represented as $\hat{Y} = \{\hat{y_i}\}_{i=1}^{N}$, where N is the number of the images. The pose map corresponding to the target map is expressed as $P = \{p_i\}_{i=1}^{N}$. z represents the noise. The concatenated image features, target pose features, and an additional 256-dimensional noise vector are concatenated as inputs of the decoder.

Decoder As shown in Figure 2(b), the concatenated features are input into the decoder, which is the process of upsampling. This process includes four feature map layers, which obtained by combined the deconvolution and the bilinear interpolation methods. Specifically, the deconvolution module is encoded by a Relu-Transconv-Norm sub-network (the upper part in the red dashed box of Figure 2(b)). The bilinear interpolation module is encoded by a Bilinear-Conv sub-network. Here, an 1x1 convolution layer is added following the bilinear interpolation layer to obtain a same channel number with the deconvolution module. In the decoder, only two layers of bilinear interpolation are added for fast convergence. In layer 1, only the concatenated features are input. The output of layer 1 is input into the deconvolution module of layer 2. In layer 3 and layer 4, the inputs are the feature map fusion of the Bilinear interpolation and Deconvolution (BD) results.

In addition, we introduce the SSIM (the structural similarity index)[8] loss, which measures the similarity between two feature vectors obtained by the target image and the generated image. SSIM evaluates images according to the fact that the HVS (the human visual system) is sensitive to changes in local structure. Also, it can measure the similarity according to the illumination, contrast, and structure. This is very valuable for generating a more realistic image.

Discriminator In the discriminant module, two discriminant losses are used. One is the identity discriminant loss, another is the pose discriminant loss. Identity discriminant network D_{id} adopts the ResNet50 as a backbone network to encode the input image, but it does not share weights with the classification network. The network structure of the identity discriminator is shown in Figure 1(d). We set the label as true

if input and target are a same person and false otherwise. According to these operations, the distances among different poses of the same ID can be reduced. Another discriminator is used to determine if the generated map has the same pose as the target, i.e., pose discriminant network D_{pd}. The pose discriminator adopts the PatchGAN [7], as shown in Figure 1(e), where the confidence map represents the matching degree between the input image and pose map.

Loss Functions

In the encoder, the NCM loss and the cross-entropy loss are adopted. The three-dimensional tensor obtained from the input image x_i is equally divided into M blocks in the horizontal direction. In order to reduce the feature redundancy among different blocks of an image, NCM loss is proposed to ensure the features of each part are independent. Specifically, we firstly use the cosine distance to measure the similarity of any two blocks, $d_{cos} = \frac{<(f_i^s)^T, f_i^t>}{\|f_i^s\| \|f_i^t\|}$, where f_i^s (f_i^t) represents the $s-th$ ($t-th$) part features of the $i-th$ identity person separately. Then the similarity is forced to approach a constant c by l_1 norm. The similarity loss of any two block features is as follows:

$$L_d = \frac{1}{C_M^2} \sum_{s=1}^{M} \sum_{t=s+1}^{M} \left\| \frac{<(f_i^s)^T, f_i^t>}{\|f_i^s\| \|f_i^t\|} - c \right\|_1 . \tag{1}$$

Here, C_M^2 denotes the permutation and combination. c represents the constant, in this paper, $c = 0$.

In the training stage, the weight matrix w_i^o of $o-th$ embedding feature may be a zero matrix. In order to avoid $\boldsymbol{W}$ be a zero matrix, we add a regularization term for $\boldsymbol{W}$, as follows:

$$L_{weight} = \sum_{o=1}^{M} \left\| (w_i^o)^T w_i^o - 1 \right\|_2 \tag{2}$$

w_i^o (with $1 \leq o \leq M$) is the row vectors of $\boldsymbol{W}$. The NCM loss is described as follows:

$$L_{nc} = L_d + \lambda_w L_{weight} \tag{3}$$

The above objective function L_{nc} is a linear function. The linear object is not iterable, so it will not be added to the whole loss function.

The cross-entropy loss is computed in each block according to the predicted ID ($\hat{I}_i$) and real label (I_i) of each block. The average of the M blocks' losses is used as the classification loss of the input x_i. As shown in the following equation:

$$\begin{aligned} L_v = \frac{1}{M} \sum_{s=1}^{M} - [\, & I_i^s log \hat{I}_i^s \\ & + \left((1 - I_i^s)\, log(1 - \hat{I}_i^s) \right)] \end{aligned} \tag{4}$$

The main mission of the generator is to generate an image $\hat{y}_i$ that is similar to the target image y_i. Thus we employ the reconstruction loss and SSIM loss to constrain the decoder. Here the reconstruction loss is used to minimize the differences between the generated image $\hat{y}_i$ and the target image y_i. It is described as follows:

$$L_{re} = \frac{1}{hw} \left\| y_i - \hat{y}_i \right\|_1 \tag{5}$$

Here h, w represents the height and width of the image y_i.

In our model, the calculation formula of SSIM is the same as [8]. The features is divided into Q blocks. Thus the loss function for SSIM can be defined as:

$$L_s(Q) = \frac{1}{N} \sum_{q \epsilon Q} 1 - SSIM(q) \tag{6}$$

In the discriminator loss D_{pd}, the adversarial loss of the generated image $\hat{y}_i$ (target image y_i) and the target pose map p_i is described as follows:

$$\begin{aligned} L_{pd} = \max_{D_{pd}} (& E_{\hat{y}_i \in Z} [\log D_{pd}([p_i, \hat{y}_i])] \\ & + E_{y_i \in Y} [\log (1 - D_{pd}([p_i, y_i]))]) , \end{aligned} \tag{7}$$

where Y represents the real image distribution and Z represents the generated image distribution.

In the discriminator D_{id}, the adversarial loss of the generated image $\hat{y}_i$ (target image y_i) and the input image x_i is as follows:

$$\begin{aligned} L_{id} = \max_{D_{id}} (& E_{\hat{y}_i \in Z} [\log D_{id}([x_i, \hat{y}_i])] \\ & + E_{y_i \in Y} [\log (1 - D_{id}([x_i, y_i]))]) \end{aligned} \tag{8}$$

The whole loss function is:

$$L = L_v + \lambda_{re} L_{re} + \lambda_s L_s + \lambda_{id} L_{id} + \lambda_{pd} L_{pd} \tag{9}$$

Here, λ_{re}, λ_{ssim}, λ_{id}, λ_{pd} represent the weights of the loss functions. In our model, we set $\lambda_{re} = 100$, $\lambda_{ssim} = 10$, $\lambda_{id} = 10$, $\lambda_{pd} = 10$, $\lambda_w = 0.01$.

EXPERIMENTAL RESULTS

Comparison With the State-of-the-Art Methods

We performed the experiment on one 1080Ti GPU and PyTorch 0.3.1 platform. Two evaluation metrics are used to measure our proposed method. The first one is top-1, top-5 and top-10 of CMC accuracy. Another one is mean Average Precision (mAP) to perform the evaluation on Market1501 [15], DukeMTMC-reID [16] and CUHK03 [1] datasets.

For the three public datasets, the proposed method is compared with 5 existing non-GAN methods and 6 state-of-the-art GAN-based methods. For Market-1501 dataset, as shown in Table. 1, our method achieves the top-1 accuracy of 93.0% and mAP of 79.8%, which outperforms [7] with 2.1% and 2.5% for mAP and top-1, respectively. With the re-ranking scheme, our proposed PGAN obtains the top-1 accuracy of 94.0% and mAP of 88.3%. In addition, we conducted a multi-query test, as shown in Table. 2, and we get Top1 accuracy as 98.9% and 88.9% mAP. It is much better than the other methods.

On the DukeMTMC-reID dataset, our PGAN obtains mAP of 66.3%, which outperforms the other methods. With the re-ranking scheme, our PGAN method outperforms the state-of-the-art methods by a large margin, where improving the top-1 accuracy from 80.5% to 82.6% compared with [11] and mAP from 64.5% to 75.6% compared with [7].

For the CUHK03 dataset, our top-1 accuracy is better than other methods. But we get 88.6% mAP, which is slightly lower than [7]. After re-ranking, the mAP and top-1 obtained by the proposed PGAN are all improved significantly. Our method achieves the top-1 accuracy of 96.9% and mAP of 96.0%. The outstanding performance demonstrates the importance of the combination of pose-unrelated representation method with GAN method.

Compare images generated with different methods

Similar to the FDGAN network [7], our proposed PGAN, is more concerned with the characteristics of the person itself, with less attention to the background, so the background of the generated image looks vaguer. As shown above the dashed line in Figure 3, we have listed the generated maps of four persons, which are denoted as (a) (b) (c) (d). Each person contains three different poses. In Figure 3(a), we can see that the color of the images generated by our method (the fourth line) is better than FDGAN (the third line), especially for the first images in the third and fourth lines. Also, in Figure 3(c), we can see that the bag in the images generated by our method is much better than FDGAN. The images generated by our method have a clear strap. In general, the images generated by our method contain more details and the overall shape of the person is more complete (the arms and legs in Figure 3(b)-(d)). The generated images further verify that the features extracted by our PGAN method are discriminative. In addition, we show some failure cases, as shown below the dashed line in Figure 3. In these cases, our model ignores some necessary details of persons because it pays too much attention to the background.

Rank results

In this section, we list some person retrieval samples in Figure 4. The first image of each row is the query image and other images are "Top1 to Top10" in the rank list. It can be seen that the accuracy is obviously improved by the PGAN method. In Figure 4, for the (a), (b) and (c), the top10 images include some wrong images (red box). It can be seen that the red boxes in the results of our PGAN is less than the others. In addition, even the wrong matching images are also very similar to the query image, they obtained by our method are ranked at the end of the row. It shows that our method is very effective in reducing the similar distances of different poses of the same person and sorting the correct images forward. In Figure 4(d), the images rank from top1 to top10 are completely correct, and

Table 1. Performance comparison with the state-of-the-art methods on three public datasets. The CMC scores (%) at Rank 1 and mAP (%) are reported.

Methods	Market1501		DukeMTMC		CUHK03	
	mAP	top-1	mAP	top-1	mAP	top-1
BoW+kissme[15]	20.8	44.4	12.2	25.1	6.4	6.4
MSCAN[17]	57.5	80.3	-	-	-	74.2
PAN[12]	63.4	82.8	51.5	71.6	34.0	36.3
MaskRe-ID[3]	75.3	90.0	61.9	78.9	-	88.8
HA-CNN[11]	75.7	91.2	63.8	**80.5**	44.4	41.0
DeformGAN[18]	61.3	80.6	-	-	-	-
LSRO[4]	66.1	84.0	58.6	76.8	77.4	73.1
Multi-pseudo[6]	67.5	85.8	58.6	76.8	87.5	85.4
Pose-transfer[5]	68.9	87.7	56.9	78.5	30.5	33.8
PNGAN[13]	72.6	89.4	53.2	73.6	-	79.8
FDGAN[7]	77.7	90.5	64.5	80.0	**91.3**	92.6
Our method (PGAN)	**79.8**	**93.0**	**66.3**	80.4	88.6	**93.0**
Our method (PGAN+re-rank)	**88.3**	**94.0**	**75.6**	**82.6**	**96.0**	**96.9**

Table 2. Multiple query results are reported on Market-1501 dataset. The CMC scores (%) at Rank 1 and mAP (%) are reported.

Methods	mAP	top-1
MSCAN[17]	66.7	86.8
MaskRe-ID[3]	82.3	93.3
HA-CNN[11]	82.8	93.8
LSRO[4]	68.5	85.1
Multi-pseudo[6]	77.9	89.9
PNGAN[13]	80.2	92.9
Our method (PGAN)	**88.9**	**98.9**

Table 3. Component analysis of the proposed PGAN on Market-1501 dataset. The CMC scores (%) at Rank 1, 5, 10 are reported.

Methods	mAP	top-1	top-5	top-10
PCB[2]	77.3	92.4	97.0	97.9
PNC	74.3	91.1	96.6	97.8
PNC+GAN	78.6	92.4	97.3	98.4
PNC+GAN +Bilinear	79.4	92.8	97.3	98.1
PNC+GAN +Bilinear +SSIM	**79.8**	**93.0**	**97.4**	**98.4**

the error rates of other methods are very high, which further illustrates the effectiveness of our proposed PGAN method.

Ablation study

In this work, we conduct a series of experiments on the Market-1501 dataset. After training the PNC classification network, we obtain 74.3% mAP, and the accuracy of top-1, top-5, top-10 is 91.1%, 96.6%, and 97.8% respectively, as shown in Table 3. The accuracy of the top-1 and mAP of PNC is lower than the PCB. Note that, we do not add data alignment here. After training a lot of group parameters, the PNC results were still lower than [2]. There are two main reasons, one is related to the initialization of different versions of PyTorch. Another reason is that features extracted by PNC may contain more background due to the feature-independent constraints. After combining with the GAN network, we got better results comparing with PNC and PCB. This is because, during the training process of GAN, the network pays more attentions to the characteristics of pedestrians, which will filter out most of the background information. Furthermore, we conduct an experimental comparison of each improvement in PGAN, as shown in Table 3. Compared with PNC, the top-1 accuracy and mAP of PNC+GAN were significantly improved. Also, it is better than PCB. After imposing the bilinear interpolation method, the mAP and accuracy are continuously increasing. Besides, we also add SSIM loss in the training stage. The bottom line of Table 3 shows the effectiveness of the SSIM loss, we get 79.8% mAP and 93.0% top-1 accuracy. In our experiment, the SSIM loss is positively correlated with the classification loss, which means that the more accurate the classification, the better the quality of the generated image.

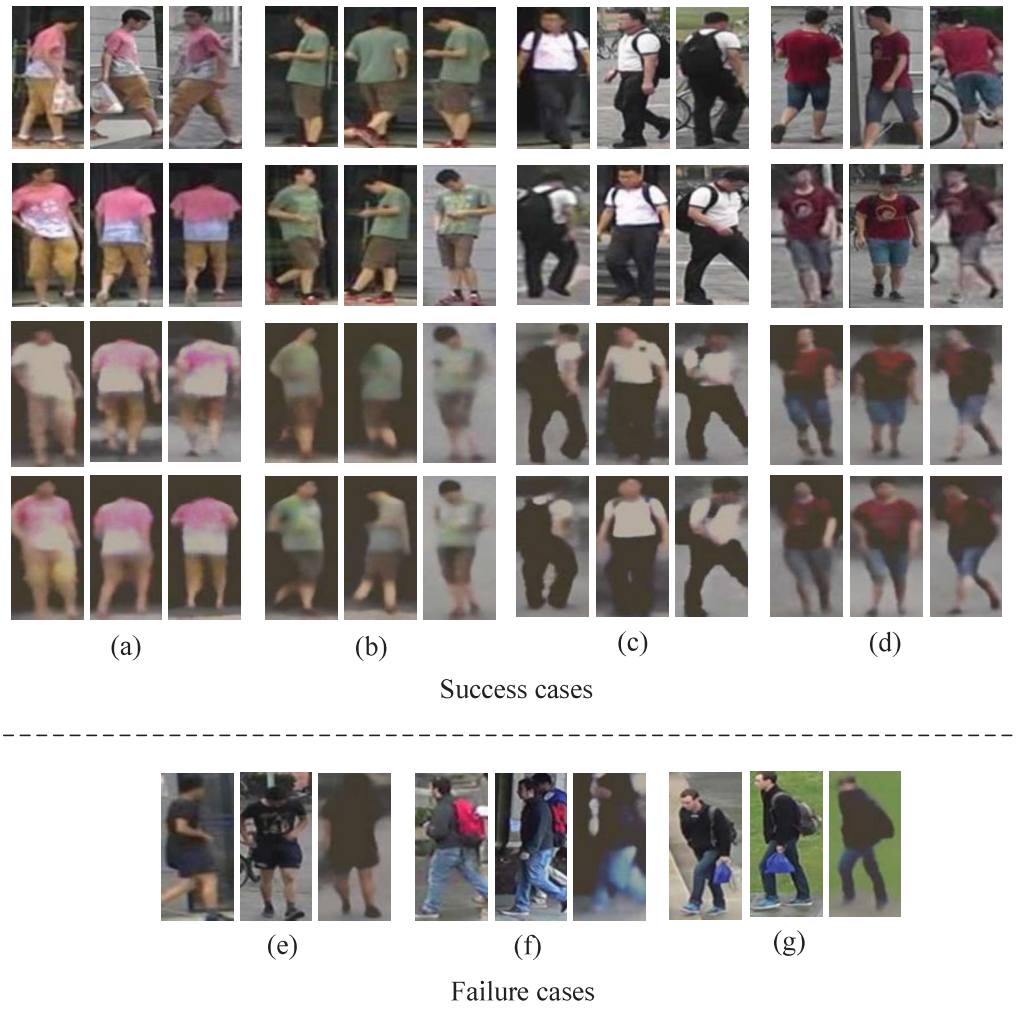

Figure 3. Generated person samples. Images above the dashed line are some success cases, generated images by our proposed PGAN (the bottom row) on Market-1501 and comparing with FDGAN (the third row), input images (the first row), and ground-truth images (the second row). Below the dashed line are some failure cases, each person contains an input image, a target image, and a generated image.It can be seen from the last image of (e)-(g) that the generated images lack of details of the persons such as the bag shape and color, the arms, face or the logo on the cloth.

CONCLUSIONS

In this paper, we proposed a novel Part-based Nondirect Coupling (PNC) representation learning method to get richer features than some common block-based approach for generator. We also proposed a novel PGAN network by embedding the PNC representation learning method into the improved pose-related GAN network. The extracted features can distinguish different images and identify the same persons without additional computational complexity in the testing phase. Experiments on three benchmarks demonstrated that our proposed method achieved good performances. Especially, it outperformed the state-of-the-art GAN-based methods for person Re-ID. The generated new pose images also include more detail information and higher quality than the existing GAN-based Re-ID methods.

Figure 4. Person retrieval samples. For each query, the 1st to 4th row are the results of PCB, PNC, FDGAN and our PGAN, respectively. The first image of each row is the query image and other images are "Top1 to Top10" in the rank list. The green (red) boxes denote the positive (negative) images with the query images.

ACKNOWLEDGMENT

This work was supported in part by the National Key R&D Program of China 2019YFB2204200, in part by the Central Universities under Grant 2019YJS039, in part by the National Natural Science Foundation of China under Grant 61872034 and 61972030, in part by the Beijing Municipal Natural Science Foundation under Grant 4202055, in part by the Natural Science Foundation of Guizhou Province under Grant [2019]1064, in part by the Science and Technology Program of Guangzhou under grant 201804010271. (The first two authors (Yue Zhang and Yi Jin) contribute equally. Corresponding author: Yigang Cen.)

REFERENCES

1. W. Li, R. Zhao, T. Xiao, X. Wang. "Deepreid: Deep filter pairing neural network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 152-159, 2014.
2. Y. Sun, L. Zheng, Y. Yang, Q. Tian, S. Wang. "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in *Proc. Euro. Conf. Comput. Vis.*, pp. 480-496, 2018.
3. L. Qi, J. Huo, L. Wang, Y. Shi, Y. Gao. "Maskreid: A mask based deep ranking neural network for person re-identification," arXiv preprint arXiv:1804.03864, 2018.
4. Z. Zheng, L. Zheng, Y. Yang. "Unlabeled samples generated by gan improve the person re-identification baseline in vitro," in *Proc. IEEE Conf. Comput. Vision*, pp. 3754-3762, 2017.
5. J.Liu, B. Ni, Y. Yan, P. Zhou, S. Cheng, J. Hu. "Pose transferrable person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 4099-4108, 2018.
6. Y. Huang, J. Xu, Q. Wu, Z. Zheng, Z. Zhang, J. Zhang. "Multi-pseudo regularized label for generated data in person re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 3, pp. 1391-1403, 2018.

7. Y. Ge, Z. Li, H. Zhao, G. Yin, S. Yi, X. Wang. "FD-GAN: Pose-guided feature distilling GAN for robust person re-identification," *Adv. neural inf. proces. syst.*, pp. 1222-1233, 2018.
8. H. Zhao, O. Gallo, I. Frosio, J. Kautz. "Loss functions for image restoration with neural networks," *IEEE Trans. Computational Imaging*, vol. 3, no. 1, pp, 47-57, 2016.
9. S. Kan, Y. Cen, Z. He, Z. Zhang, L. Zhang, Y. Wang. "Supervised Deep Feature Embedding with Hand Crafted Feature," *IEEE Trans. Image Process.*, vol.28, no. 12, pp. 5809-5823. 2019.
10. S. Kan, L. Zhang, Z. He, Y. Cen, S. Chen, J. Zhou. "Metric Learning-Based Kernel Transformer With Triplets and Label Constraints for Feature Fusion," *Pattern Recogn.*, vol. 15, no. 8, 2019.
11. W. Li, X. Zhu, S. Gong. "Harmonious attention network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 2285-2294, 2018.
12. Z. Zheng, L. Zheng, Y. Yang. "Pedestrian alignment network for large-scale person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 10, pp. 3037-3045, 2019.
13. X. Qian, Y. Fu, T. Xiang, W. Wang, J. Qiu, Y. Wu, X. Xue. "Pose-normalized image generation for person re-identification," in *Proc. Euro. Conf. Comput. Vis.*, pp. 650-667, 2018.
14. Y. Li, S. Tang, R. Zhang, Y. Zhang, J. Li, S. Yan, "Asymmetric GAN for Unpaired Image-to-image Translation", *IEEE Trans. Image Process.*, vol. 28, no. 12, pp. 5881-5896, 2019.
15. L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, Q. Tian. "Scalable person re-identification: A benchmark," In *Proc. IEEE Int. Conf. Comput. Vision*, pp. 1116-1124, 2015.
16. E. Ristani, F. Solera, R. Zou, R. Cucchiara, C. Tomasi. "Performance measures and a data set for multi-target, multi-camera tracking," in *Proc. Euro. Conf. Comput. Vis.*, pp. 17-35, 2016.
17. D. Li, X. Chen, Z. Zhang, K. Huang. "Learning deep context-aware features over body and latent parts for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 384-393, 2017.
18. A. Siarohin, E. Sangineto, S. Lathuilière, N. Sebe. "Deformable gans for pose-based human image generation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 3408-3416, 2018.

Yue Zhang is currently working toward the Ph.D. degree at the Institute of Information Science, Beijing Jiaotong University. Her main research interests include machine learning, person re-identification, metric learning, and deep learning. Contact her at zhangyuede@bjtu.edu.cn.

Yi Jin is an associate professor with the School of Computer and Information Technology, Beijing Jiaotong University. Her research interests include image processing and signal processing. She received the Ph.D. degree in signal and information processing from the Institute of Information Science, Beijing Jiaotong University, in 2010. She was a Visiting Scholar with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, from 2013 to 2014. (SM'06-M'13), contact her at yjin@bjtu.edu.cn.

Jianqiang Chen is an associate professor with the School of Mechanical Engineering, University of Guizhou. His research interests include fault diagnosis, defect inspection, image understanding etc. He received the Master degree in computer science from the Zhejiang University, Hangzhou, China, in 2002. Contact him at 200875600@qq.com.

Shichao Kan is currently working toward the Ph.D. degree at the Institute of Information Science, Beijing Jiaotong University. His research interests include general image retrieval, metric learning, image-to-image translation, large-scale image retrieval, object search, object detection, and deep learning. Contact him at 16112062@bjtu.edu.cn.

Yigang Cen is a professor in the Institute of Information Science, Beijing Jiaotong University. His research interests include computer vision, image understanding/processing, IOT etc. He received the Ph.D. degree in control theory and control engineering from the Huazhong University of Science Technology, Wuhan, China, in 2006. In 2006, he joined the Signal Processing Centre, School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, as a Research Fellow. From 2014 to 2015, he was a Visiting Scholar with the Department of Computer Science, University of Missouri, Columbia, MO, USA. (M'09), contact him at ygcen@bjtu.edu.cn.

Qi Cao is an assistant professor with the School of Computing Science, University of Glasgow, Singapore. His research interests include image processing, signal processing, virtual reality (VR) and augmented reality (AR). He obtained his PhD degree from Nanyang Technological University (NTU), Singapore in 2007. Contact him at Qi.Cao@glasgow.ac.uk.