A Novel Pedestrian Reidentification Method Based on a Multiview Generative Adversarial Network

JIFENG GUO ©, ZHIQI PANG ©, MING YU ©, PEIJIAO XIE ©, AND DAN LIU ©
College of Information and Computer Engineering, Northeast Forestry University, Harbin 150040, China

Corresponding author: Dan Liu (1370890813@qq.com)

This work was supported in part by the Basic Scientific Research Projects of Central Universities under Grant 2572018BH07, and in part by the Natural Science Foundation of Heilongjiang Province under Grant LH2019C003.

ABSTRACT

Emerging deep learning (DL) techniques have greatly improved pedestrian reidentification (PRI) performance. However, the existing DL-based PRI methods cannot learn robust feature representations owing to the single view of query images and the limited number of extractable features. Inspired by generative adversarial networks (GANs), this paper proposes a novel PRI method based on a pedestrian multiview GAN (PmGAN) and a classification recognition network (CRN). The PmGAN consists of three generators and one multiclass discriminator. The three generators produce pedestrian images from the front, side and back, while the multiclass discriminator determines whether the input image is a real image or a generated image. In addition to expanding the existing pedestrian datasets, the PmGAN can generate pedestrian images from front, side and back views based on a given query image and thereby increase the feature semantic space of the query image. To verify the performance of our method, the PmGAN was compared with mainstream pedestrian image generation models, and then the proposed method was contrasted with mainstream PRI methods. The results show that the proposed PmGAN greatly improved the performance of mainstream PRI methods. For example, the combination of the PmGAN and Pyramidal increased the mean average precision (mAP) on three common datasets by 1.2% on average. The research findings provide new insights into the application of multiview generation in PRI tasks.

INDEX TERMS

Deep learning (DL), generative adversarial networks (GANs), image generation, pedestrian reidentification (PRI), classification recognition network (CRN).

I. INTRODUCTION

Traditionally, pedestrians are reidentified in two steps: manually extracting features from pedestrian images, e.g., using a color histogram and a histogram of oriented gradients (HOG) [1], and learning the measurement matrix via similarity measurement methods such as the large margin nearest-neighbor (LMNN) [2] and cross-view quadratic discriminant analysis (XQDA) [3]. In recent years, emerging deep learning (DL) techniques have been widely applied to pedestrian reidentification (PRI) and have achieved far better results than traditional approaches.

At present, DL-based PRI methods can be divided into two steps. Specially, the first step is feature vector extraction. In the early studies, feature vector extraction is mainly based on global images [4]–[7]. By treating PRI as a classification task, these methods learn pedestrian features through network training. Then, the pedestrian features are extracted from the original image with a convolutional neural network (CNN) and judged to determine whether they belong to the same pedestrian. For instance, Geng et al. [4] designed a network model containing a classification subnetwork and a verification subnetwork: the classification subnetwork predicts the image identity (ID) and trains the network model using classification errors, enabling the model to extract pedestrian features effectively; the verification subnetwork judges whether the extracted features belong to the same pedestrian. Despite being stable and easy to train, this type of method performs poorly with a large number of images and tends to ignore some details of pedestrians.

To overcome the defects, a second type of DL-based PRI methods, which focus on the local features of pedestrians,
have emerged [8]–[11]. Initially, the local features were extracted via image segmentation [8]. If two original images are not aligned, however, it is difficult to compare each pair of one image with the corresponding part in the other image. To ensure image alignment, some scholars have aligned pedestrians in advance via a priori strategies, such as human pose estimation [9], key point extraction from human skeletons [10], and a multiple granularity network [11]. These alignment strategies help to extract more details and improve the PRI effect, but they increase computational overhead.

The second step is metric learning [12]–[17]. The main idea is to reduce the distance between pedestrian images with the same ID and increase the distance between those with different IDs. For example, Varior et al. [12] trained a pair of twin networks with the contrast loss and imported a pair of labeled images into the network pair. If the two images have the same ID, they make a positive sample pair with the label y=1; otherwise, they make a negative sample pair with the label y=0. Then, the contrast loss was minimized to reduce the distance between each positive sample pair and increase the distance between each negative sample pair. Similar methods include the triplet loss [13]–[15], the quadruplet loss [16], and the group similarity learning coupling conditional random field (CRF) with a deep neural network (DNN) [17].

Based on the analysis of the existing data set, it is found that pedestrians with the same identity will have different appearance features under different shooting angles. Through visual analysis of the images reidentified by these types of methods, it is learned that the distance between an image and the target image, and thus the identification probability of that image, is negatively correlated with their similarity view. This obviously limits the performance of DL-based PRI methods using metric learning. In order to mitigate the appearance change caused by the change of perspective, a multiview generation model based on GAN is introduced into the existing PRI model.

Recently, generative adversarial networks (GANs) [18] have achieved excellent results in image generation [19]–[21], image recovery [22], image dehazing [23], [24] and style transfer [25], [26] thanks to their powerful image processing capabilities. In addition, GAN is widely used in multiview generation, and it has been able to generate higher-quality images. In view of the GAN’s multiview generation ability, this paper introduces GAN into PRI tasks. Therefore, this paper proposes a novel PRI method based on the pedestrian multiview GAN (PmGAN) and a classification recognition network (CRN). To improve the PRI performance, the PmGAN generates query images from different views to offset the feature loss of query images with the same view.

The main contributions of this paper are as follows:

1. A novel PRI method was designed based on the PmGAN, which consists of three generators and one multiclass discriminator. The three generators produce pedestrian images from the front, side, and back, while the multiclass discriminator determines whether the input image is a real image or a generated image.

2. Monte Carlo search (MCS) [27] and an attention mechanism were incorporated in the generators, providing each generator with enough semantic features and adding semantic details to the final generated images. The view label was introduced such that the discriminator could discriminate multiview images.

3. The proposed PRI method was verified through experiments on three mainstream datasets.

II. RELATED WORK

A. GAN

The original GAN [18] encompasses a generator G and a discriminator D. The generator generates data G(z) based on a random noise z, while the discriminator receives either real data or generated data and judges whether the input is real data or generated data. The generator and the discriminator are trained alternately until reaching the Nash equilibrium [28]. The objective function of the GAN can be expressed as:

$$\min_G \max_D V(D, G) = E_{x \sim p_{\text{data}}}(x) \log D(x) + E_{z \sim p_{\text{z}}}(z) \log(1 - D(G(z)))$$  \(1\)

where \(p_{\text{data}}(x)\) and \(p_{\text{z}}(z)\) are the probability distributions of the real data and random noise, respectively, and \(G(z)\) and \(D(x)\) are the generator functions and discriminator functions, respectively. \(E\) is the mathematical expectation.

After the addition of condition variable \(y\) to the inputs of G and D, the original GAN becomes a conditional GAN (CGAN) [19]. In other words, the generator receives \(y\) and \(z\) at the same time, while the discriminator receives \(y\) and \(z\) as real data/generated data. The objective function of the CGAN can be described as:

$$\min_G \max_D V(D, G) = E_{x \sim p_{\text{data}}(x)} \log D(x|y) + E_{z \sim p_{\text{z}}(z)} \log(1 - D(G(z)|y)))$$  \(2\)

where \(p_{\text{data}}(x)\), \(p_{\text{z}}(z)\) and \(y\) have the same definitions as those in formula (1), and \(y\) is a condition variable in any form.

The objective function of the original GAN can be converted into the Jensen–Shannon (JS) divergence [29] between the generated data and the real data. As a result, the original GAN is unstable in the training process; the better the discriminator is trained, the more severe the vanishing gradient problem [30]. To solve the instability, the original objective function was modified with the Wasserstein distance into the Wasserstein GAN (WGAN) [20]. The objective of the WGAN can be expressed as:

$$\min_G \max_{f, \|f\|_{L^1} \leq 1} E_{x \sim p}(f(x)) - E_{z \sim q}(f(G(z)))$$  \(3\)

where \(f(x)\) is a discriminator function subjected to Lipschitz constraints [31].
B. MULTIVIEW GENERATION
Multiview generation aims to create images from other views based on an image with a single view. The main tools of multiview generation include the variational autoencoder (VAE) [32] and GANs [18].

Early studies relied on three-dimensional (3D) modeling to generate images from different views [33]–[35]. For example, Choy et al. [34] adopted the 3D to two-dimensional (2D) mapping mechanism to obtain 3D features from 2D data. Later, Zhao et al. [36] designed the VariGAN to produce high-resolution multiview images from coarse to fine. Recently, Tang et al. [37] extended the VariGAN to the SelectionGAN, which further expands the generated space and performs well in the translation of satellite images and street view maps.

The above models can generate high-quality images after being trained with a large amount of data or semantic segmentation images. Their requirement of very large training set limits their flexibility.

To make up for the lack of pedestrian features in a single view, in this study, a generation model consisting of three generators was developed. During the PRI, the model can produce images from the front, side, and back views at the same time.

C. PEDESTRIAN REIDENTIFICATION
The task of PRI is to use computer vision technology to judge whether there are specific pedestrians in images or video sequences. Early PRI technology mainly relied on manual extraction of pedestrian features. With the rise of deep learning, PRI based on deep learning have been widely studied. PRI is mainly divided into two stages: feature vector extraction and metric learning.

The main task of the feature vector extraction stage is to extract pedestrian feature from pedestrian images. Early research was mainly based on global images for feature extraction. The advantage of this method is that it is stable and easy to train. The disadvantage is that it has poor performance in large-scale scenarios. So in subsequent studies, researchers gradually realized the importance of local features. Sun et al. [38] not only divided the image into several parts, but also considered the continuity of information between each part, and the index of Rank-1 of this method on Market-1501 [39] exceeded 92%. Wang et al. proposed the Multiple Granularity Network (MGN) [11], which effectively combined local and global features, and its performance was highly competitive at that time. Huang et al. [40] extracted three body regions from pedestrian images and set fragment Learning method to optimize the contribution feedback module. Finally, the Rank-1 index of this method on CUHK03 [41] reached 93.5%. Zheng et al. [42] proposed a novel coarse-to-fine pyramid model that combines not only local and global information, but also the gradual clues between them. This method achieved the latest results at that time. Liu et al. [43] proposed a novel PRI method, namely, Adversarial Erasing Attention (AEA), to use adversarial method to mine the complete features of differences. Guo et al. [44] propose a method for group-shuffling dual random walks with label smoothing. This method is highly competitive in CUHK03, Market-1501 and DukeMTMC data sets.

Metric learning’s task is to measure the similarity of extracted features to obtain pedestrian images with higher similarity to query images. Specific methods are the triplet loss [13]–[15], the quadruplet loss [16] and so on.

III. PmGAN-BASED PRI
This section introduces the overall framework and workflow of the proposed method and then details the two subnetworks, namely, the PmGAN and CRN.

A. OVERALL FRAMEWORK OF OUR METHOD
The proposed PRI method covers two subnetworks: a PmGAN and a CRN. The latter was designed to verify the effectiveness of the PmGAN. The overall framework of the proposed method is illustrated in Figure 1, where $P_1$-$P_3$ are real images, and $P_1'$-$P_3'$ are generated images.

Our method involves a three-stage training process and a two-stage test process.

1) TRAINING PROCESS
First, the front, side, and back images were selected from the original training set and merged into the training set of the PmGAN. These images were grouped based on the original pedestrian IDs so that each group contains all three views and several other views. The training set was used to train the PmGAN, which contains three generators and one multiclass discriminator. The three generators produce pedestrian images from the front, side and back, while the multiclass discriminator determines whether the input image is a real image or a generated image. The generator and the discriminator were trained alternately until achieving convergence.

Second, the trained generators were employed to produce images from other views based on an image with a given view. The generated images were given the same ID label as the original image. Next, the labeled images were added to the original training set, creating an expanded training set.

Third, the expanded training set was adopted to train the CRN. With the aim of validating the PmGAN, the CRN was designed based on ResNet-50 [45]. In the CRN, the number of neurons of the classification layer is adjustable according to the specific training set.

2) TEST PROCESS
First, the data set used in this paper includes two parts during the test: query set and test set. This paper selects a pedestrian image from the query set. For a selected query image, the pedestrian images with three views (front, side, and back) were produced by the trained multiview generator model. The generated images and the original image were sent to the trained CRN. The output of the fully connected layer, which is right before the classification layer, was taken
as the extracted pedestrian features. Based on the maximum principle, the four extracted eigenvectors were fused into the eigenvector of the query image. At this point, the feature vector after fusion contains richer pedestrian features and can better represent the pedestrian.

Second, the CRN was called to extract the features of all the pedestrian images in the test set, and then the Euclidean distances between the features of the query image and those extracted from the test set were calculated. Then, the images were sorted in ascending order of their Euclidean distance.

B. PmGAN

In the PmGAN, three generators $G_1$, $G_2$, and $G_3$ are responsible for generating pedestrian images with different views, and the multiclass discriminator $D$ is responsible for distinguishing between real and generated images.

The training process can be viewed as adversarial learning in which $G_i$ ($i = 1, 2, 3$) aims to generate images with view $I_i$ ($I_i$ is front, side, or back), and $D$ aims to distinguish the real image from generated image. $G$ and $D$ were trained alternatively. All images were processed to the size of $128 \times 256$.

1) MULTIVIEW GENERATOR MODEL

Generators $G_1$, $G_2$, and $G_3$ have the same structure but different parameters. As an example, the structure generator $G_1$ is explained as follows.

As shown in Figure 2, generator $G_1$ contains two small subgenerators $E_1$ and $F_1$ and an attention mechanism based on a two-layer CNN. It takes three steps for generator $G_1$ to generate a new image with a front view.

First, a real image with a given view was input into subgenerator $E_1$, which output a coarse-grained image. Second, the MCS [24] was performed to sample the coarse-grained image six times, creating a larger semantic generation space with six samples $J_1$-$J_6$. After that, the attention mechanism extracted the features from the six samples. Third, the features and the original image were imported to subgenerator $F_1$ to produce a fine-grained image.

The subgenerator $F_1$ was designed by adding a convolutional layer (Cov)-Bayes network (BN)-leaky rectified linear unit (ReLU) residual block [46] to U-Net [47]. The U-Net helps to perverse the features of the original image in the generated image, and the skip connection [48] between the encoder and decoder can transmit the textural features of the original images between high-level feature layers. The structure of subgenerator $F_1$ is explained in Table 1.

During the training, the pedestrian images with a given view were input into the three generators. Then, each generator produced a pedestrian image with another view under the guidance of the discriminator and updated their parameters.

2) MULTICLASS DISCRIMINATOR

The function of the multiclass discriminator is to distinguish the real image from the generated images with different views. The multiclass discriminator was designed based on the CGAN discriminator. The input is a real image or a generated image, and the view label is $I_i$ ($I_i$ is front, side or back). The output is the probability of the input being a real image. The discriminator is essentially a classifier. As shown in Figure 3, the discriminator contains five convolutional layers and three residual blocks. Each convolutional layer has an activation layer and a BN layer. The last activation layer uses the ReLU function, while the other activation layers use the LReLU function.

3) OBJECTIVE FUNCTION

The adversarial loss of the PmGAN was calculated via the objective function of the WGAN. The multiclass discriminator of the PmGAN was subjected to Lipschitz constraints on the global scale according to the physical meaning of the matrix spectral norm [49]; the length of any vector after the matrix transform is smaller than or equal to the length of the product between the vector and the matrix spectral
The structure of generator $G_1$.

| Layer Type          | Input              | Output            |
|---------------------|--------------------|-------------------|
| Downsampling        | Conv (64, 1, 4x4), BN, LReLU | (128,256,3) | (128,256,64) |
|                     | Conv (128, 2, 4x4), BN, LReLU | (128,256,64) | (64,128,128) |
|                     | Conv (256, 2, 4x4), BN, LReLU | (64,128,128) | (32,64,256)  |
|                     | Conv (512, 2, 4x4), BN, LReLU | (32,64,256) | (16,32,512)  |
| Residual block      | Conv (512, 1, 4x4), BN, LReLU | (16,32,512) | (16,32,512)  |
|                     | Conv (512, 1, 4x4), BN, LReLU | (16,32,512) | (16,32,512)  |
|                     | Conv (512, 1, 4x4), BN, LReLU | (16,32,512) | (16,32,512)  |
|                     | Conv (512, 1, 4x4), BN, LReLU | (16,32,512) | (16,32,512)  |
|                     | Conv (512, 1, 4x4), BN, LReLU | (16,32,512) | (16,32,512)  |
| Upsampling          | DeConv (256, 2, 4x4), BN, LReLU | (16,32,512) | (32,64,256)  |
|                     | DeConv (128, 2, 4x4), BN, LReLU | (32,64,256) | (64,128,128) |
|                     | DeConv (64, 2, 4x4), BN, LReLU | (64,128,128) | (128,256,64) |
|                     | Conv (3, 1, 4x4), BN, LReLU | (128,256,64) | (128,256,3)  |

where $\sigma(W)$ is the spectral norm of the weight matrix, $x$ is the input vector of the current layer, and $\delta$ is the variation in $x$.

Then, the pixelwise mean squared error (pMSE) [50] and perception loss [51] were introduced to retain the features of the original pedestrian images in the generated images and improve the visual satisfaction of the generated images. The pMSE can be defined as:

$$L_{\text{pMSE}} = \frac{1}{WH} \sum_{x=1}^{W} \sum_{y=1}^{H} (I_{x,y} - G_\theta(I_{x,y}'))^2$$  \hspace{1cm} (5)

where $I_{x,y}$ and $I_{x,y}'$ are the pixel values at $(x, y)$ in the generated image and the real image, respectively; $W$ and $H$ are the height and width of the image, respectively; and $\theta$ is the generator parameter.

The pMSE calculates the image difference on a pixel basis. Via this calculation, the generated image was too smooth in texture and poor in visual perception. To solve the defects, the perception loss, which is capable of enhancing visual satisfaction, was included in the objective function:

$$L_{\text{pl}} = \frac{1}{W_i H_i} \sum_{x=1}^{W_i} \sum_{y=1}^{H_i} (\phi_{i,j}(I) - \phi_{i,j}(G_\theta(I')))^2$$  \hspace{1cm} (6)

where $\phi_{i,j}$ is the feature map between the $i$th max pooling layer and the $j$th convolutional layer in the pretrained VGG19 network [52]; $I$ and $I'$ are the generated image and the real image, respectively; and $W_{i,j}$ and $H_{i,j}$ are the width.
and height of each feature map in the VGG19 network, respectively.

The overall cost function can be expressed as:

$$L_{\text{total}} = L_{\text{WGAN}} + \alpha L_{\text{pMSE}} + \beta L_{\text{pl}}$$ (7)

where $L_{\text{WGAN}}$ is the adversarial loss of the WGAN, $L_{\text{pMSE}}$ is the pixelwise mean squared error, $L_{\text{pl}}$ is the perception loss, and $\alpha$ and $\beta$ are the hyperparameters for proportional control ($\alpha = 0.05$).

C. CRN

Based on ResNet-50, the proposed CRN has a simple structure and is easy to train. The purpose of CRN design in this paper is to extract the feature of the input pedestrian image, and the ultimate goal is to test the effectiveness of PmGAN in PRI task. Here, the CRN is pretrained on the ImageNet dataset. Then, the number of neurons in the final classification layer was adjusted as per the pedestrian class in the training set. Finally, the CRN was finetuned based on the training set. During testing, the CRN extracts the features from the query image, calculates the Euclidean distance between the features of the query image and those from the test set, and ranks the images in ascending order of their Euclidean distance. In the experiment with PmGAN, pedestrian images from three perspectives were first generated by PmGAN, then input the four images into CRN for feature extraction; then the four feature vectors extracted are fused according to the maximum value principle.

The architecture of the CRN, in which the number of neurons in the “FC” layer was adjusted according to the pedestrian class in the training set, is shown in Table 2. Therefore, the CRN is essentially a classifier, and it will output a “category” in the training phase and the test phase.

### Table 2. The detailed architecture of the CRN.

| Layer name | Conv (64, 4×4) |
|------------|----------------|
| Conv1      | 7×7, 64, S2    |
|            | 3×3 max pool, S2 |
| Conv2      | 1×1 64         |
|            | 3×3 64         |
|            | 1×1 256        |
| Conv3      | 1×1 128        |
|            | 3×3 128        |
|            | 3×3 256        |
|            | 1×1 512        |
| Conv4      | 1×1 256        |
|            | 3×3 256        |
|            | 1×1 1024       |
| Conv5      | 1×1 1024       |
|            | 3×3 1024       |
|            | 1×1 2048       |
| FC         | FC             |

The CRN was trained by the cross-entropy loss of a single sample:

$$L(\hat{y}, y) = -\sum_{i=1}^{k} y_i \log \hat{y}_i$$ (8)

where $k$ is the number of classes, $y$ is the actual label of the sample, and $\hat{y}$ is the predicted value of the output.

IV. EXPERIMENTS AND ANALYSIS

This section mainly compares the generated image quality and application effect of the PmGAN with those of mainstream PRI methods and then contrasts our method with mainstream PRI methods.

A. EXPERIMENTAL SETUP

Our experiments were conducted on three authoritative PRI datasets, namely, Market-1501 [39], DukeMTMC-reID [53], and CUHK03 [41].
The Market-1501 dataset contains a total of 32,668 images of 1,501 pedestrians with different IDs. The dataset was split into a training set with 12,936 images of 751 pedestrians with different IDs and a test set with 19,732 images of 750 pedestrians with different IDs.

The DukeMTMC-reID dataset contains a total of 34,183 images of 1,404 pedestrians with different IDs. The dataset was split into a training set with 16,522 images of 702 pedestrians with different IDs and a test set with 17,661 images of 702 pedestrians with different IDs.

The CUHK03 dataset contains a total of 14,097 images of 1,467 pedestrians with different IDs. The dataset was split into a training set with 13,132 images of 1,367 pedestrians with different IDs and a test set with 965 images of 100 pedestrians with different IDs. The purpose of choosing this division method was to provide the PmGAN with more optional IDs to conform with the test protocol.

Based on the three public datasets, a combined dataset was generated to train the PmGAN, including 1,705 groups of images. Since each group of PmGAN training images must contain front, back and side images of pedestrians with the same ID, we selected a total of 1705 groups of images that met the requirements from the three data sets. The remaining 1115 groups of images did not meet the requirements of the PmGAN for training set, so they were not used for training PmGAN. In each group, the images depicted a pedestrian with the same ID from the front, side, back and several other views.

The similarity between each generated image and the real image was evaluated at the pixel level by two metrics: the peak signal-to-noise ratio (PSNR) and the structural similarity (SSIM) index [54]. Regarding the high-level feature space, each generated image was also evaluated by the Inception Score (IS) [55] and the Fréchet Inception Distance (FID) [56]. In addition, the PRI performance was assessed by two common criteria, namely, the mean average precision (mAP) and Rank-1.

The multiview generator model, the multiclass discriminator, and the CRN were all processed by the Adam optimizer [57], where the first and second momentum terms β1 and β2 were set to 0.9 and 0.99, respectively; the learning rate of the PmGAN was set to 0.01; and the learning rate of the CRN was set to 0.005. Finally, the server configuration is as follows: an E5-2620 v4 @ 2.10 GHz CPU and an NVIDIA Tesla V100 32 G GPU.

**B. CONTRASTIVE EXPERIMENT ON MULTIVIEW GENERATION**

This subsection compares the PmGAN with mainstream image generation models in terms of the generated image quality and application effect.

1) CONTRASTIVE MODELS

Pix2pix [58]: Pix2pix adds image x as a condition variable to the generator and discriminator and guides the generator to produce image y in a new domain. Pix2pix uses the GAN to translate images. In this experiment, pix2pix was adopted to map a pedestrian image with a single view into image y. Pix2pix consists of a generator and a discriminator. The generator uses a U-net structure, which consists of 15 layers (8 convolutional layers and 7 deconvolution layers). The discriminator uses the PatchGAN architecture, including a convolutional network with 6 layers.

BiGAN [59]: The BiGAN combines the original GAN with a subgenerator E that can map data to the feature space and acquires the ability to learn meaningful feature representation. Two data pairs (G(z), z) and (x, E(x)) are input into discriminator D, where G(z) and E(x) are the outputs of G and E, respectively; x is the original data; and z is a random variable. The BiGAN consists of three subnetworks: an encoder, a decoder and a discriminator. The encoder is composed of convolution layers, BN layers and activation layers. The decoder is composed of deconvolution layers, BN layers and activation layers. The discriminator also uses a convolutional network, but it has one more encoding to pair with the image than the Pix2pix discriminator.

IntroVAE [60]: The IntroVAE can assess the quality of the generated samples and improve itself. Without adding discriminators, The IntroVAE generates high-resolution images, rather than the blurry composite images of the VAE. The IntroVAE consists of an encoder, which is composed of convolution layers, and a generator. The generator is similar to a decoder and is composed of deconvolution layers. They were trained against each other. The output of the encoder is used as the input of the generator, and the final generator outputs the generated image. The encoder causes the encoding of the generated image to deviate from the encoding of the real image.

VariGAN [36]: The VariGAN, which integrates the GAN with the VAE, can generate high-quality images from coarse to fine without producing serious artifacts. This model has achieved good results on the DeepFashion dataset [61]. The VariGAN consists of three modules: a coarse image generator, which consists of two encoders, fully connected layers and a decoder; fine image generator; and a conditional discriminator. The encoders share the weights. The fine image generator is a dual-path U-Net, which is composed of two encoders, skip connections and a decoder. In addition, the two encoders also share the weight. The conditional discriminator in this paper uses six convolutional layers.

SelectionGAN [37]: Using the multichannel attention selection mechanism, the SelectionGAN divides image translation into two stages: generating coarse-grained images and generating fine-grained images with the rich details from the said mechanism. The SelectionGAN consists of two generators, a discriminator and a multi-channel attention selection module composed of multi-scale spatial pooling and multi-channel attention selection. The generator uses the U-Net architecture and the discriminator uses the PatchGAN architecture.
2) COMPARISON OF GENERATED IMAGE QUALITY

The quality of the images generated by the PmGAN and the above models was compared both qualitatively and quantitatively. Figure 4 provides the multiview images produced by different models.

As shown in Figure 4, the images generated by the SelectionGAN and PmGAN were much clearer than those generated by the other models. The PmGAN learned more details of the pedestrian, such as their clothes and backpacks, than the SelectionGAN. In addition, the PmGAN reduced the distortion of the background features (e.g., street scenes and steps) in comparison to the other models.

Next, the generated images were evaluated using the PSNR, SSIM, IS, and FID (Table 3). The similarity between a generated image and the real image was positively correlated with the PSNR, SSIM, and IS and negatively correlated with the FID.

As shown in Table 3, the images generated by the PmGAN had better scores than those of other models, as evidenced by all four indices. The SelectionGAN was the second best-performing model. Compared with those of the SelectionGAN, the images produced by the PmGAN had a 2.21% higher PSNR, roughly the same SSIM, a 1.98% higher IS, and a 1.56% lower FID.

3) COMPARISON OF THE APPLICATION EFFECT

This section tests the performance of the CRN on three different datasets, namely, Market-1501, DukeMTMC-reID and CUHK03, and compares the application effects of the VariGAN, SelectionGAN and PmGAN on the CRN performance. The three models were selected because they can generate high-quality images and extract representative features.

In the training phase, the VariGAN, SelectionGAN, and PmGAN were adopted separately to generate multiview pedestrian images. These images were added to the original dataset, creating three expanded datasets with the expansion ratio of 1:1 (original image: real image). Then, the CRN was trained with the three expanded datasets, respectively.

To determine the effect of each expanded dataset on CRN performance, the experimental results of the CRN trained with the original training set (baseline method $R_0$) were also given. The test phase was divided into three stages to compare the three models step by step.

First, the performance of $R_0$ was tested without using a multiview generation model to make up for the features of the query image. As shown in Table 4, the performances ($R_1$, $R_2$, and $R_3$) of the CRN trained on the datasets expanded by the VariGAN, SelectionGAN, and PmGAN were tested separately without using any multiview generation model to make up for the features of the query image. Therefore, the difference between $R_1$, $R_2$, $R_3$ and $R_0$ was that the training sets used by $R_1$, $R_2$ and $R_3$ were respectively obtained by the VariGAN, SelectionGAN and PmGAN, expanding the original dataset. In the test phase, $R_1$, $R_2$, $R_3$ and $R_0$ followed the same steps; that is, they did not use any multiview generation model to make up for the features of the query image.

Third, the performances (VariGAN+$R_1$, SelectionGAN+$R_2$, and PmGAN+$R_3$) of the CRN trained on the datasets expanded by the VariGAN, SelectionGAN, and PmGAN were tested separately using a multiview generation model to make up for the features of the query image. First, the query image was input into the generation model to produce the pedestrian images with three views. Then, the generated images and the original image were entered into the CRN for feature extraction, and the extracted features were fused into the final query feature by the maximum principle. Finally, the similarity between the final query feature and all image features in the test set was calculated. The experimental results are shown in Table 4.

As shown in Table 4, $R_1$, $R_2$ and $R_3$ outperformed $R_0$, and $R_3$ achieved the best performance. Therefore, the multiview generation model improved the CRN performance by expanding the training set.

In addition, VariGAN+$R_1$, SelectionGAN+$R_2$ and PmGAN+$R_3$ performed better than $R_1$, $R_2$ and $R_3$. The best performance was achieved by PmGAN+$R_3$. The edge of PmGAN+$R_3$ over VariGAN+$R_1$ and SelectionGAN+$R_2$ was greater than that of $R_3$ over $R_1$ and $R_2$. Therefore, the multiview generation model produced query images with multiple views to improve the eigenvectors of pedestrian images and optimize the CRN performance in the test phase. It can also be seen that the proposed PmGAN is superior to the mainstream generative models in executing PRI tasks.

C. COMPARISON WITH MAINSTREAM METHODS

Finally, the proposed method (PmGAN+$R_3$) was compared with the mainstream methods on the three datasets of Market-1501, DukeMTMC-reID and CUHK03.
The contrastive methods included the traditional manual method (XQDA + local maximal occurrence (LOMO) [3]), a DL method based on global features (PAN [7]), DL methods based on local features (Spindle [10], PCB [38], Pose-transfer [58], MGN [11], and Pyramidal [40]), GSDRWLS [44], and a method based on metric learning (CRF+DNN [17]). The results of these methods and our method are compared in Table 5.
the average test time of VariGAN+R₁, SelectionGAN+R₂ and PmGAN+R₃ is longer than that of R₀, R₁, R₂ and R₃. Similarly, the average test time of PmGAN+PCB, PmGAN+MGN and PmGAN+Pyramidal is longer than that of PCB, MGN and Pyramidal. In which the time cost of PmGAN + R₃ is 6.56% higher than that of R₃; The time cost of PmGAN+MGN is 5.62% higher than that of PCB; The time cost of PmGAN+PCB is 3.98% higher than that of MGN; The time cost of PmGAN+Pyramidal is 3.93% higher than that of Pyramidal; On the Market-1501 dataset and DukeMTMC-reID dataset, the average test time of VariGAN+R₁, SelectionGAN+R₂ and PmGAN+R₃ was very close to the test time of R₀, R₁, R₂ and R₃. Similarly, the average test time of PmGAN+PCB, PmGAN+MGN and PmGAN+Pyramidal was very close to the test time of PCB, MGN and Pyramidal (The gap is less than 1%). It is equal to say that the time cost of introducing multiview generation model is less than time error of causing by hardware system. It can be concluded that although the time cost increases after the introduction of PmGAN, the proportion of the increased time cost in the total cost decreases with the increase of the number of images in the test set. For larger data sets, PmGAN can significantly improve the performance of PRI system at a minimal time cost.

V. CONCLUSION

Our PmGAN-based PRI method can generate images from three fixed views from a given pedestrian image. This function improves the network performance in the training phase and promotes the features of the query image with multiple views in the test phase. In this way, the PRI network can achieve much better performance than other methods. In addition, the proposed PmGAN is highly flexible. This model can be coupled with existing PRI methods, resulting in a surge in PRI performance. The effectiveness of the PmGAN was fully verified through experiments on three common datasets. Future research will aim to solve the problems of occlusion and lighting and give full play to the advantages of multiview generation in PRI tasks.

REFERENCES

[1] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2005, pp. 886–893.
[2] K. Q. Weinberger and L. K. Saul, “Fast solvers and efficient implementations for distance metric learning,” in Proc. 25th Int. Conf. Mach. Learn. (ICML), Jun. 2008, pp. 886–893.
[3] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, “Person re-identification by attribute and identity learning,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 1160–1167.
[4] M. Gong, Y. Wang, T. Xiang, and Y. Tian, “Deep transfer learning for person re-identification,” 2016, arXiv:1611.05244. [Online]. Available: http://arxiv.org/abs/1611.05244
[5] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, Z. Hu, C. Yan, and Y. Yang, “Improving person re-identification by attribute and identity learning,” Pattern Recognit., vol. 95, pp. 151–161, Nov. 2019.
[6] L. Zheng, Y. Yang, and A. G. Hauptmann, “Person re-identification: Past, present and future,” 2016, arXiv:1610.02984. [Online]. Available: http://arxiv.org/abs/1610.02984

TABLE 5. Rank-1 (%) and mAP (%) obtained by our method and the mainstream methods on three datasets.

| Method               | Market-1501 | DukeMTMC-reID | CUHK03  |
|----------------------|-------------|---------------|---------|
|                      | Rank (%)    | mAP (%)       | Rank (%)|
| XQDA+LOMO [3]        | 43.8        | 22.2          | 30.7    | 17.0 | 12.8 | 11.5 |
| PAN [7]              | 82.8        | 63.4          | 71.6    | 51.5 | 36.3 | 34.0 |
| Spindle [10]         | 76.9        | -             | -       | -    | -    | -    |
| PCB [38]             | 92.4        | 77.3          | 81.9    | 65.3 | 61.3 | 54.2 |
| Pose-transfer [58]   | 87.6        | 68.9          | 78.5    | 56.9 | 41.6 | 38.7 |
| MGN [11]             | 95.7        | 86.8          | 87.7    | 76.4 | 66.8 | 66.0 |
| Pyramidal [40]       | 95.7        | 88.2          | 89.0    | 78.9 | 78.9 | 74.8 |
| CRF+DNN [17]         | 93.5        | 81.6          | 84.9    | 69.5 | -    | -    |
| GSDRWSLs [44]        | 94.0        | 85.6          | 83.9    | 69.1 | 96.2 | 96.2 |
| PmGAN+R₀             | 90.2        | 76.7          | 80.6    | 64.5 | 61.8 | 58.6 |
| PmGAN+MGN            | 95.9        | 88.3          | 88.8    | 79.4 | 66.8 | 69.9 |
| PmGAN+Pyramidal      | 95.9        | 86.9          | 89.9    | 80.3 | 79.1 | 75.7 |
| PmGAN+* GSDRWSLs     | 94.1        | 87.2          | 84.2    | 71.2 | 96.1 | 96.4 |

TABLE 6. Time cost(second) obtained by different methods on three datasets.

| Method               | Market-1501 | DukeMTMC-reID | CUHK03  |
|----------------------|-------------|---------------|---------|
|                      | Rank (%)    | mAP (%)       | Rank (%)|
| R₀                   | 609.16      | 530.45        | 29.65   |
| R₁                   | 621.73      | 525.16        | 29.69   |
| R₂                   | 616.04      | 532.38        | 30.33   |
| R₃                   | 608.14      | 533.96        | 28.96   |
| VariGAN+R₁           | 613.47      | 534.89        | 32.43   |
| SelectionGAN+R₂      | 619.32      | 531.31        | 33.08   |
| PmGAN+R₂             | 617.82      | 527.34        | 30.86   |
| PCB [38]             | 782.43      | 701.43        | 39.84   |
| MGN [11]             | 897.32      | 776.53        | 44.21   |
| Pyramidal [40]       | 1194.72     | 1049.23       | 58.72   |
| PmGAN+PCB            | 787.65      | 698.34        | 42.08   |
| PmGAN+MGN            | 904.65      | 771.39        | 45.97   |
| PmGAN+Pyramidal      | 1189.43     | 1054.90       | 61.03   |

The following can be observed from Table 5:

(1) Among the mainstream methods, the DL methods based on local features, especially the Pyramidal, were superior to other methods.

(2) The combination of the PmGAN and simple CRN R₃ achieved an effect comparable to those of the mainstream methods such as the PCB. The combination of the PmGAN and Pyramidal achieved the best performance.

(3) The PmGAN improved the performance of the PRI network, especially in terms of the mAP. Compared with the original method, the combination of the PmGAN and Pyramidal increased the mAP by 1.4% on the Market-1501 dataset, 1.3% on the DukeMTMC-reID dataset, and 0.9% on the CUHK03 dataset.

Due to the introduction of PmGAN, the performance of PRI system is significantly improved. However, the time cost is also increased. Therefore, in order to test the running time of the system, we implemented ten experiments under the same conditions and then take the average value of the ten results as the final result, in which time is measured in seconds (s), as shown in Table 6.

As can be seen from Table 6, on the CUHK03 dataset, due to the introduction of multiview generation model,
I. Grosse, P. Bernaola-Galván, P. Carpena, R. Román-Roldán, J. Oliver, and H. E. Stanley, “Analysis of symbolic sequences using the Jensen-Shannon divergence,” *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 65, no. 4, Mar. 2002, Art. no. 041905.

J. Guo et al.: Novel PRI Method Based on a Multiview GAN
J. Guo et al.: Novel PRI Method Based on a Multiview GAN

[52] M. Mateen, J. Wen, S. Song, and Z. Huang, “Fundus image classification using VGG-19 architecture with PCA and SVD,” Symmetry, vol. 11, no. 1, p. 1, Dec. 2018.

[53] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, “Performance measures and a data set for multi-target, multi-camera tracking,” in Proc. Eur. Conf. Comput. Vis., 2016, pp. 17–35.

[54] A. Hore and D. Ziou, “Image quality metrics: PSNR vs. SSIM,” in Proc. 20th Int. Conf. Pattern Recognit., Aug. 2010, pp. 2366–2369.

[55] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, “Improved techniques for training GANs,” in Proc. Adv. Neural Inf. Process. Syst., 2016, pp. 2234–2242.

[56] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, “Gans trained by a two time-scale update rule converge to a local Nash equilibrium,” in Proc. Adv. Neural Inf. Process. Syst., 2017, pp. 6626–6637.

[57] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” 2014, arXiv:1412.6980. [Online]. Available: http://arxiv.org/abs/1412.6980

[58] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-Image translation with conditional adversarial networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 1125–1134.

[59] J. Donahue, P. Krähenbühl, and T. Darrell, “Adversarial feature learning,” 2016, arXiv:1605.09782. [Online]. Available: http://arxiv.org/abs/1605.09782

[60] H. Huang, R. He, Z. Sun, and T. Tan, “IntroVAE: Introspective variational autoencoders for photographic image synthesis,” in Proc. Adv. Neural Inf. Process. Syst., 2018, pp. 52–63.

[61] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, “DeepFashion: Powering robust clothes recognition and retrieval with rich annotations,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 1096–1104.

[62] J. Liu, B. Ni, Y. Yan, P. Zhou, S. Cheng, and J. Hu, “Pose transference person re-identification,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 4099–4108.

JIFENG GUO was born in Yichun, Heilongjiang, China, in 1974. He received the D.Eng. degree. He is currently an Associate Professor and the Director of the Information and Computer Science and Technology Experiment Center, Northeast Forestry University. He has published more than 30 academic articles, including ten articles in EI. He is currently engaged in artificial intelligence, intelligent control research, and teaching.

ZHIQI PANG was born in Dezhou, Shandong, China, in 1998. He is currently pursuing the master’s degree with Northeast Forestry University. He has coauthored two academic articles, one of which was included in SCI. His research interests include machine learning, deep learning, and computer vision.

MING YU was born in Harbin, Heilongjiang, China, in 1981. He received the master’s degree in computer application technology from Northeast Forestry University, in 2006. As the chief editor of three books, he has published more than ten articles, and holds more than ten patents. He is mainly engaged in the field of intelligent research.

PEIJIAO XIE was born in Wenchang, Hainan, China, in 1997. She is currently pursuing the master’s degree with Northeast Forestry University. She worked as a Research and Development Engineer with Beijing Gengmei Technology Company, for six months. Her research interests include deep learning and computer vision.

DAN LIU was born in Harbin, Heilongjiang, China, in 1976. She received the Ph.D. degree in mechanical design and theory from Northeast Forestry University, in 2012. Since 2015, she has been an Assistant Professor of Software Engineering with Northeast Forestry University. She is the author of five books and more than ten articles. Her research interests include big data analysis and artificial intelligence.

* * *

J. Guo et al.: Novel PRI Method Based on a Multiview GAN

[53] M. Mateen, J. Wen, S. Song, and Z. Huang, “Fundus image classification using VGG-19 architecture with PCA and SVD,” Symmetry, vol. 11, no. 1, p. 1, Dec. 2018.

[54] A. Hore and D. Ziou, “Image quality metrics: PSNR vs. SSIM,” in Proc. 20th Int. Conf. Pattern Recognit., Aug. 2010, pp. 2366–2369.

[55] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, “Improved techniques for training GANs,” in Proc. Adv. Neural Inf. Process. Syst., 2016, pp. 2234–2242.

[56] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, “Gans trained by a two time-scale update rule converge to a local Nash equilibrium,” in Proc. Adv. Neural Inf. Process. Syst., 2017, pp. 6626–6637.

[57] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” 2014, arXiv:1412.6980. [Online]. Available: http://arxiv.org/abs/1412.6980

[58] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-Image translation with conditional adversarial networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 1125–1134.

[59] J. Donahue, P. Krähenbühl, and T. Darrell, “Adversarial feature learning,” 2016, arXiv:1605.09782. [Online]. Available: http://arxiv.org/abs/1605.09782

[60] H. Huang, R. He, Z. Sun, and T. Tan, “IntroVAE: Introspective variational autoencoders for photographic image synthesis,” in Proc. Adv. Neural Inf. Process. Syst., 2018, pp. 52–63.

[61] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, “DeepFashion: Powering robust clothes recognition and retrieval with rich annotations,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 1096–1104.

[62] J. Liu, B. Ni, Y. Yan, P. Zhou, S. Cheng, and J. Hu, “Pose transference person re-identification,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 4099–4108.

JIFENG GUO was born in Yichun, Heilongjiang, China, in 1974. He received the D.Eng. degree. He is currently an Associate Professor and the Director of the Information and Computer Science and Technology Experiment Center, Northeast Forestry University. He has published more than 30 academic articles, including ten articles in EI. He is currently engaged in artificial intelligence, intelligent control research, and teaching.

ZHIQI PANG was born in Dezhou, Shandong, China, in 1998. He is currently pursuing the master’s degree with Northeast Forestry University. He has coauthored two academic articles, one of which was included in SCI. His research interests include machine learning, deep learning, and computer vision.

MING YU was born in Harbin, Heilongjiang, China, in 1981. He received the master’s degree in computer application technology from Northeast Forestry University, in 2006. As the chief editor of three books, he has published more than ten articles, and holds more than ten patents. He is mainly engaged in the field of intelligent research.

PEIJIAO XIE was born in Wenchang, Hainan, China, in 1997. She is currently pursuing the master’s degree with Northeast Forestry University. She worked as a Research and Development Engineer with Beijing Gengmei Technology Company, for six months. Her research interests include deep learning and computer vision.

DAN LIU was born in Harbin, Heilongjiang, China, in 1976. She received the Ph.D. degree in mechanical design and theory from Northeast Forestry University, in 2012. Since 2015, she has been an Assistant Professor of Software Engineering with Northeast Forestry University. She is the author of five books and more than ten articles. Her research interests include big data analysis and artificial intelligence.

* * *= * ==