Multi-task GANs for View-Specific Feature Learning in Gait Recognition

Yiwei He, Junping Zhang, *Member, IEEE*, Hongming Shan, Liang Wang, *Senior Member, IEEE*

***Abstract*—Gait recognition is of great importance in the fields of surveillance and forensics to identify human beings since gait is the unique biometric feature that can be perceived efficiently at a distance. However, the accuracy of gait recognition to some extent suffers from both the variation of view angles and the deficient gait templates. On one hand, the existing cross-view methods focus on transforming gait templates among different views, which may accumulate the transformation error in a large variation of view angles. On the other hand, a commonly used Gait Energy Image (GEI) template loses temporal information of a gait sequence. To address these problems, this paper proposes Multi-task Generative Adversarial Networks (MGANs) for learning view-specific feature representations. In order to preserve more temporal information, we also propose a new multi-channel gait template, called Period Energy Image (PEI). Based on the assumption of view angle manifold, MGANs can leverage adversarial training to extract more discriminative features from gait sequences. Experiments on OU-ISIR, CASIA-B and USF benchmark datasets indicate that compared with several recently published approaches, PEI+MGANs achieves competitive performance and is more interpretable to cross-view gait recognition.**

***Index Terms*—Gait recognition, Cross-view, Generative Adversarial Networks, Surveillance**

I. Introduction

DIFFERENT from other biometric features such as human faces, fingerprints, and irises which are usually obtained at a close distance, gait is the unique biometric feature that can identify humans at a far distance. However, the performance of gait recognition [1] suffers from various exterior factors including clothing [2], walking speed [3], low resolution [4] and so on. Among these factors, the change of view angles greatly influences the generalization ability of gait recognition models. For example, when a person walks across a camera located at a fixed position, the gait appearance of the person may vary along walking directions, making a formidable barricade in recognizing the human under the cross-view case.

To solve this problem, some researchers [5]–[8] proposed to learn transformations or projections between different view angles in cross-view gait recognition. Specifically, the View Transform Model (VTM) [9]–[12] transforms gait templates such as Gait Energy Image (GEI) [13] from one view to another. However, VTM requires predicting each pixel value of GEI independently, which is time-consuming and inefficient. To reduce the computational time, an auto-encoder based model [14] is used to reconstruct GEI and extract view-invariant features. In order to achieve view transformations, these two methods reconstruct gait templates via transitional view angles. In this way, however, the reconstruction error may be accumulated if there is a large view variation between two view angles.

The recently published Generative Adversarial Networks (GANs) interpolate facial poses or age variations along a low-dimensional manifold [15], [16]. It has the ability to model data distribution to improve the performance of different vision tasks such as super-resolution [17] and inpainting [18]. However, original GANs methods generate images from random noise, lacking features that can preserve identity information, which is undesirable for cross-view gait recognition.

In order to overcome the shortcomings mentioned above, this paper proposes Multi-task Generative Adversarial Networks (MGANs) to learn view-specific features from gait templates. Further, we propose a new multi-channel gait template, named Period Energy Image (PEI), which is a generalization of GEI. The PEI template can maintain more temporal and spatial information compared with other templates such as GEI and Chrono-Gait Image (CGI) [19], [20]. Extensive experiments on three gait benchmark datasets indicate that our MGANs model with PEI achieves competitive performance in cross-view gait recognition compared with several recently published approaches.

The training structure of the proposed MGANs models is illustrated in Fig. 1. Inspired by the recent success of deep networks for cross-view gait recognition [21], the convolutional neural network is utilized in our model. PEI is first encoded as a view-specific feature in a latent space by the encoder. Then, a view transform layer transforms the feature from one view to another. Finally, a modified GANs structure is trained with both pixel-wise loss and multi-task adversarial loss. In addition, a view-angle classifier is trained with cross-entropy loss to predict the view angle of the PEI in the testing phase.

The rest of this paper is organized as follows. Related work is reviewed in Section II. Section III presents the PEI template and explain the proposed MGANs model. Experimental results are analyzed in Section IV. Discussion and conclusion are given in Sections V and VI, respectively.

This work has been partially funded by the National Natural Science Foundation of China (No. 61673118) and Shanghai Pujiang Program (No. 16PJD009).

Y. He and J. Zhang are with the Shanghai Key Laboratory of Intelligent Information Processing and the School of Computer Science, Fudan University, Shanghai, 200433, China. Tel.: +86-21-55664503, Fax: +86-21-65654253, Emails: {heyw15, jpzhang}@fudan.edu.cn.

H. Shan is with the Department of Biomedical Engineering, the Center for Biotechnology and Interdisciplinary Studies, Rensselaer Polytechnic Institute, Troy, NY, 12180. E-mail: shanh@rpi.edu

L. Wang is with the Center for Research on Intelligent Perception and Computing (CRIPAC), National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing, 100190, P. R. China. Emails: wangliang@nlpr.ia.ac.cn.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2018.2844819, IEEE Transactions on Information Forensics and Security

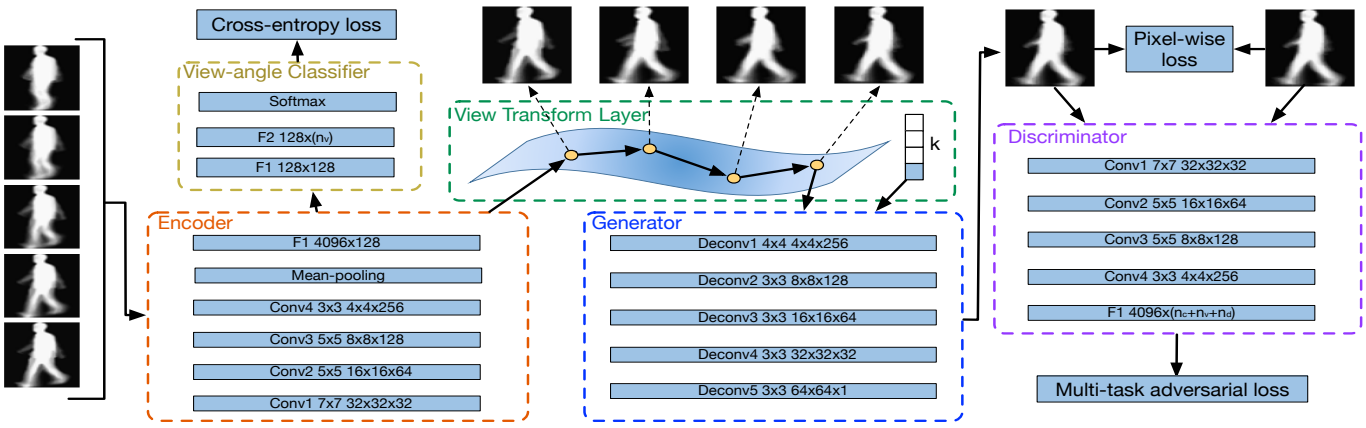


Fig. 1. The training structure of our proposed multi-task generative adversarial networks. PEI is encoded as a view-specific feature in a latent space. A view-angle classifier is used to predict view angle of the feature. The feature is then transformed gradually in a view transform layer. A generator produces gait images by utilizing the transformed feature and one-hot channel vector. Pixel-wise loss, multi-task adversarial loss and cross-entropy loss are calculated for training.

II. Related work

A. Cross-view gait recognition

The approaches of cross-view gait recognition can be divided into three categories. The first category devotes to reconstructing the 3D structure of a person through a set of gait images taken from multi-view cameras [22]–[24]. Constrained by the strict environmental requirement and expensive computational cost, however, it is less applicable in practice. The second category is to extract the hand-crafted view-invariant features from gait images to represent a person [25]–[28]. Due to the strong nonlinear relationship between view angles and gait images, extracting such view-invariant features from images is a challenge. As a result, the hand-crafted view-invariant features cannot generalize well in the condition of a large view variation [21].

Most of the state-of-the-art methods belonging to the third category directly learn transformations or projections of gaits in different view angles. For example, Makihara et al. proposed a View Transformation Model (VTM) to transform gait templates from one view to another [10]. In their work, Singular Value Decomposition (SVD) was used to compute the projection matrix and view-invariant features for each GEI. Further, a truncated SVD was proposed to overcome the overfitting problem of the original VTM [11]. Moreover, they refined a VTM-based method to learn a nonlinear transformation between different view angles [9] by employing support vector regression.

Different from VTM-based methods, Canonical Correlation Analysis (CCA) based approaches project the gait templates from multiple view angles onto a latent space with maximal correlation [5], [8], [12], [29]. For example, Bashir et al. modeled the correlation of gait sequences from different view angles using CCA [5]. Kusakunniran et al. claimed that there might exist some local correlations in GEIs of different view angles [12]. Xing et al. proposed Complete Canonical Correlation Analysis (C3A) to overcome the shortcomings of CCA when directly dealing with two sets of high dimensional features [8]. In their method, the original CCA was decomposed into two stable eigenvalue decomposition problems to avoid inconsistent projection directions between Principle Component Analysis (PCA) and CCA.

However, CCA-based methods assume that view angles are known in advance. Therefore, Hu et al. proposed an alternative View-invariant Discriminative Projection (ViDP) to project the gait templates onto a latent space without knowing the view angles [7].

Recently, deep neural networks based gait recognition methods were introduced in [2], [21], [30]–[32]. The CNN-based method proposed by Wu et al. [21] automatically recognized the discriminative features to predict the similarity given a pair of gait images. The model they used is opaque on how the view variation affects the similarities between different samples. Instead of using silhouette images which are sensitive to the clothing and carrying variations, a pose-based temporal-spatial network [2] is proposed to extract dynamic and static information from the key-point of human bodies [33]. The experimental results show that it may be a challenge for a pose-based method to extract discriminative information from key-points of a human body.

However, there are some shortcomings remained in the existing methods. For example, VTM-based methods suffer from error accumulation stemming from large view variations. CCA-based methods and ViDP only model the linear correlation between features. CNN-based method lacks the interpretability of view variations. In order to overcome these shortcomings, our MGANs model benefits from the feature transformation in a latent space and the nonlinear deep model. Different from directly predicting the similarity given a pair of samples as in [21], our method learns view-specific features by utilizing prior knowledge about the view angles. This greatly facilitates the understanding of the view variation to the learned features.

B. Generative adversarial networks

Recently, Generative Adversarial Networks (GANs) [34] were introduced as a novel way to model data distributions.

Specifically, GANs are a pair of neural networks consisting of a generator G and a discriminator D. In the original GANs, the generator G generates fake data from a distribution of $P_{\boldsymbol{z}}$. The goal of the discriminator D is to distinguish fake data from real data $\boldsymbol{x}$. We assume that the distribution of real data is P_{data}. Both the generator and discriminator are iteratively optimized against each other in a minimax game as follows [34]:

$$\min_{\theta_G}\max_{\theta_D}\ \mathbb{E}_{\boldsymbol{x}\sim P_{data}}\big[\log D(\boldsymbol{x})\big]+\mathbb{E}_{\boldsymbol{z}\sim P_{\boldsymbol{z}}}\big[\log\left(1-D(G(\boldsymbol{z}))\right)\big] \tag{1}$$

where θ_G and θ_D are the parameters of G and D, respectively.

However, the training of original GANs suffers from the problems of low quality, instability and mode collapse. Several variants of GANs were thus introduced to solve these problems. For example, WGANs [35], [36] and DCGANs [15] were proposed to improve the stability of training and to alleviate mode collapse.

Research on original GANs has also focused on utilizing supervised information. For example, conditional GANs [37] was proposed to generate samples by providing label information. Various vision problems such as super-resolution [17] and inpainting [18] were advanced based on conditional GANs.

Recent researches on GANs are capable of interpolating facial poses or age variations along a low-dimensional manifold [15], [16]. In order to capture the manifold of view angles and model the distribution of gait images, we also introduce the GANs into our model. The structure of GANs in our proposed model is composed of one generator and several sub-discriminators. Each sub-discriminator is responsible to ensure that the generated gait images belong to a certain domain such as a view angle domain, an identity domain, or a channel domain of gait images.

Note that the recently published GaitGAN [38] also introduced the GANs to learn view-invariant features. The proposed two discriminators were used to ensure that the generated gait images are realistic and the identity can be preserved. There are two main differences between their work and ours. The first is that their work directly transformed the gait template from arbitrary view angles to the side view angle without utilizing the assumption of view angle manifold. The second is that the two discriminators proposed in their work are mutually independent, whereas different discriminators will share the weights of the network in our method.

III. Method

In this section, we first give an overview of our method for cross-view gait recognition. Then, we describe a novel gait template called Period Energy Image (PEI). We also formulate the model of our proposed Multi-task Generative Adversarial Networks (MGANs). Finally, we introduce the objective functions of our methods.

A. Method overview

The pipeline of the recognition process is shown in Fig. 2. Given a probe template $\boldsymbol{x}^p$ in view angle p and n gallery templates $\{\boldsymbol{x}_1^g, \boldsymbol{x}_2^g, ..., \boldsymbol{x}_n^g\}$ in view angle g, the identities of the gallery templates are defined as $\{y_1^g, y_2^g, ..., y_n^g\}$. The goal

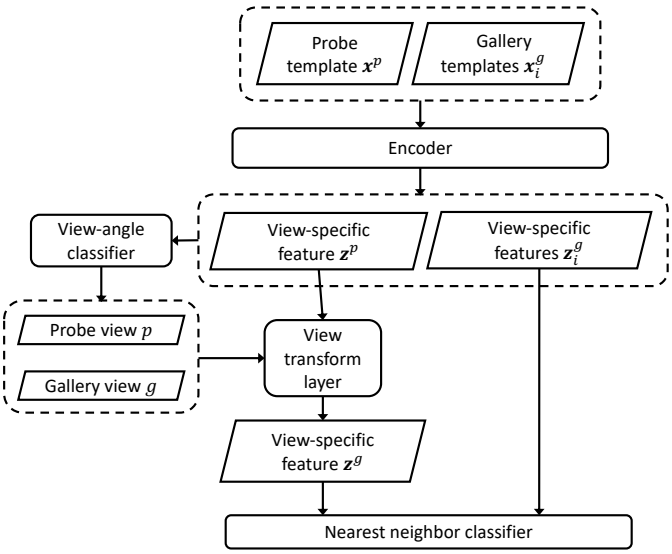


Fig. 2. The pipeline of the recognition process. probe template $\boldsymbol{x}^p$ and gallery templates $\boldsymbol{x}_i^g$ are encoded as view-specific features $\boldsymbol{z}^p$ and $\boldsymbol{z}_i^g$. Then, view-specific feature $\boldsymbol{z}^p$ are transformed to $\boldsymbol{z}^g$ by a view transform layer. A nearest neighbor classifier is then used to recognize the identity of the probe template $\boldsymbol{x}^p$.

of cross-view gait recognition is to recognize the identity of $\boldsymbol{x}^p$. The gait template $\boldsymbol{x}^p$ is first encoded as a view-specific feature $\boldsymbol{z}^p = E(\boldsymbol{x}^p)$ in a latent space, where E is an encoder. Then, a view-angle classifier predicts the view angles of both probe and gallery templates. After that, the view transform layer V is used to transform z^p from view p to g based on function $\boldsymbol{z}^g = V(\boldsymbol{z}^p, p, g)$. The gait templates $\boldsymbol{x}_i^g$ are also encoded as features $\boldsymbol{z}_i^g = E(\boldsymbol{x}_i^g)$ in the latent space, where $i \in \{1, 2, ..., n\}$. The identity of $\boldsymbol{x}^p$ is recognized as $y_{i^*}^g$ by using the nearest neighbor classifier, where

$$i^* = \underset{i\in\{1,2,...,n\}}{\operatorname{argmin}} \|\boldsymbol{z}^g - \boldsymbol{z}_i^g\|_2, \tag{2}$$

and $\|\cdot\|_2$ is ℓ_2 norm.

B. Period energy image

Among several state-of-the-art gait templates [13], [19], [39], [40], GEI [13] averages the gait sequence in one period. It can preserve spatial information but lose temporal information. Chrono-Gait Image (CGI) [19] utilizes the multi-channel technology to merge temporal information into one gait template. However, some spatial information in different time domains may be confused. To take advantages of both GEI and CGI templates, we propose a generalization of GEI, i.e., Period Energy Image (PEI). The flowchart of building a 5-channel PEI is shown in Fig. 3.

Template construction In order to capture the temporal information in a gait sequence, a gait period detection method proposed by Wang et al. [19] is employed to construct a periodic signal from silhouette images. In their method, the amplitude of the tth frame in a gait sequence is represented as r_t, which is defined by the normalized average width of the leg region. After that, they utilized multi-channel mapping functions to construct the Chrono-Gait Image (CGI) [19]. In

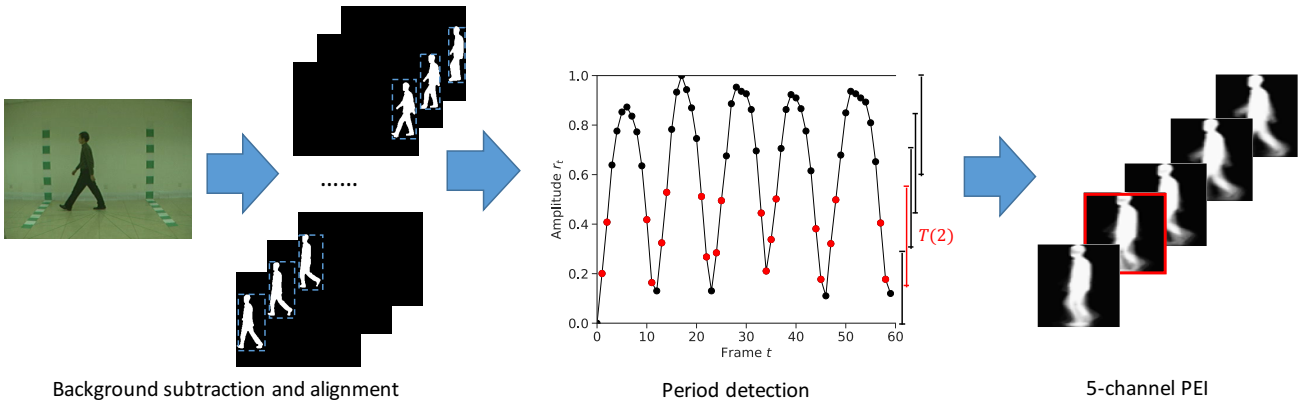


Fig. 3. Flowchart of building a 5-channel PEI template from raw video frames. Silhouette images are constructed by background subtraction and center alignment of the human body. The PEI templates are then built based on the amplitude of each frame in a gait sequence. The silhouette images that contributed to the second channel of PEI are colored in red.

TABLE I
COMPARISON AMONG OUR PROPOSED PEI, GEI AND CGI IN COMPUTATIONAL AND SPATIAL COMPLEXITIES.

Complexity	GEI	CGI	PEI
Computational	$\mathcal{O}(Nwh)$	$\mathcal{O}(Nwhn_c)$	$\mathcal{O}(\sum_{k=1}^{n_c} N_k wh)$
Spatial	$\mathcal{O}(wh)$	$\mathcal{O}(n_c wh)$	$\mathcal{O}(n_c wh)$

PEI, frames in a gait sequence are also mapped to different channels of a gait image according to the amplitudes of frames. Each channel of PEI is the average of several frames in a whole sequence. More specifically, to construct the kth channel of a PEI ($k = 1, 2, ..., n_c$), we average the silhouette images that correspond to a certain range of amplitude $T(k)$ as follows:

$$PEI_k = \frac{1}{N_k} \sum_{r_t \in T(k)} B_t, \tag{3}$$

where

$$T(k) = \Big[\frac{k}{n_c+1} - \frac{m}{2}, \frac{k}{n_c+1} + \frac{m}{2}\Big]. \tag{4}$$

Here m denotes the size of the sliding window, N_k is the number of silhouette images covered by $T(k)$ and B_t is the silhouette image of the tth frame. The spatial information of the entire gait sequence is divided into n_c possible overlapping windows. If we set both n_c and m to 1, there will be only one channel and the range of amplitude covered by this channel will become $T(1) = [0, 1]$, which makes PEI degenerate to GEI.

Spatial and computational complexities We analyze the spatial and computational complexities of PEI, GEI and CGI, as shown in Table I. The height and width of gait templates are represented as w and h. The length of a whole sequence and the number of frames in the kth channel are N and N_k, respectively. The number of channels is n_c. It is obvious that like CGI, the spatial complexity of PEI is affected by the number of channels. In addition, the computational complexity of PEI is determined by both m and n_c, because m decides the number of frames in each channel. Note that if we choose $m = \frac{1}{n_c}$, the computational complexity of PEI is equal to that of GEI because $\sum_{k=1}^{n_c} N_k = N$. However, we generally choose a larger m to ensure that the silhouette noise could be reduced, which resulting in higher computational cost than GEI. In this way, our proposed PEI can preserve more spatial and temporal information as the number of channels increases.

C. Multi-task generative adversarial network

Our proposed MGANs model consists of five components: an encoder that encodes the gait templates as view-specific features in a latent space, a view-angle classifier that predicts the view angles of the view-specific features, a view transform layer that transforms the view-specific features from one view to another, a generator that generates the gait images from the view-specific features, and a discriminator that discriminates whether the generated gait images belong to certain domains or distributions. In this subsection, we detail each component as follows.

Encoder: In order to obtain a view-specific feature for recognition, a convolutional neural network is adopted as the encoder in our model. The structure of the encoder is shown in Fig. 1. The input $\boldsymbol{x}^u$ to the encoder is our proposed PEI template in view angle u. The size of the $\boldsymbol{x}^u$ is $64 \times 64 \times n_c$ where n_c is the number of channels in PEI. We use temporal pooling to aggregate temporal information in gait templates. Temporal pooling is commonly used to summarize several video frames into one feature vector in previous literature [41]. In our method, we treat each channel of PEI as one frame and aggregate the temporal information across all channels. Therefore, each channel of PEI is independently fed to the encoder. Mean-pooling is chosen as the implementation of temporal pooling in our method, which is the same as in [41]. We use four convolutional layers followed by batch-normalization layers to build the encoder. Each component in MGANs uses LeakyReLU as the nonlinear activation function. The negative slope of LeakyReLU is set as 0.01. Instead of using the max pooling layer, convolutional layers with a stride size of 2 are adopted. The number of filters is increased by a factor of 2 from 32 to 256.

View-angle classifier: It should be noted that in the testing phase, the view transform layer requires the view angles of

probe and gallery templates. Therefore, we employ a view-angle classifier to predict the view angles of gait templates. As shown in Fig. 1, the classifier consists of two fully connected layers and one softmax layer. The classifier takes the view-specific feature of the gait template as input and predicts its view angle. The prediction task can be seen as a classification problem which viewing each view angle as one class. Since this classifier relies on the effective feature of gait template, we start to train it once other four components (the encoder, the view transform layer, the generator and the discriminator) are done.

View transform layer: Assuming that gait images with view variations lie on a low-dimensional manifold, samples moving along this manifold can achieve transformation between different view angles while preserving its identity. Therefore, the view transformation from angle u to v can be formulated as

$$\boldsymbol{z}^v = \boldsymbol{z}^u + \sum_{i=u}^{v-1} \boldsymbol{h}_i \tag{5}$$

where $\boldsymbol{h}_i$ is the view transform vector from view i to $i+1$. Different from those methods that transform gait templates in transitional view angles, our model directly transforms view-specific features in the latent space, which, in turn, reduces the accumulation reconstruction error. Specifically, we implement the view transformation in the form of a fully connected layer without bias parameter. The weight parameter of the fully connected layer is $H = [\boldsymbol{h}_1, \boldsymbol{h}_2, ..., \boldsymbol{h}_{n_v}]$, where n_v is the number of views. By properly encoding the view angles u and v as a vector representation $\boldsymbol{e}^{uv} = [e_1^{uv}, e_2^{uv}, ..., e_{n_v}^{uv}]^T$ where $e_i^{uv} \in \{0, 1\}$, the view transformation can be written as

$$\boldsymbol{z}^v = \boldsymbol{z}^u + H\boldsymbol{e}^{uv}. \tag{6}$$

Generator: The output of the view transform layer is then fed into the generator. The goal of the generator is to generate PEI which is indistinguishable from the real PEI. In our experiments, however, we cannot achieve stable training of GANs to generate PEI due to its high dimensions. We thus randomly select one channel of PEI to address this issue in the training phase. By concatenating $\boldsymbol{z}^v$ and the one-hot representation $\boldsymbol{k}$ of channel label, we define the input of the generator as $[\boldsymbol{z}^v, \boldsymbol{k}]$, As in Fig. 1, the generator is composed of five deconvolutional layers followed by batch-normalization layers.

Discriminator: The inputs of the discriminator are the generated gait image $\boldsymbol{x}^v$ and the real gait image $\hat{\boldsymbol{x}}^v$. It is noted that the view v and the channel label k of $\boldsymbol{x}^v$ and $\hat{\boldsymbol{x}}^v$ are same. In the original GANs model, the output of the discriminator is a scalar which aims to discriminate whether the generated image is real. In MGANs, however, the output of the discriminator is a vector with $n_v + n_c + n_d$ dimension, where n_v is the number of view angles, n_c is the number of channels in PEI and n_d is the number of subjects in the training set. We assign different tasks to each scalar output of the discriminator. In this way, there are a total of $n_v + n_c + n_d$ tasks which share the weights of network except for the last layer of the discriminator. The tasks can be concluded as three types as follows. 1) Discriminating whether $\boldsymbol{x}^v$ is in the view v for the first n_v tasks. 2) Discriminating whether $\boldsymbol{x}^v$ satisfies the distribution of gait images in channel k for the next n_c tasks. 3) Discriminating whether $\boldsymbol{x}^v$ preserves the identity information of the input $\boldsymbol{x}^u$ for the last n_d tasks. In the training phase, the last n_d of these tasks are employed to preserve the identity information of the generated images. This is because the real images with certain identity only participate one of the last n_d tasks. In the testing phase, we can just use the encoder and the view transform layer to generate view-specific features.

D. Objective function

In our cross-view gait recognition method, three losses are employed as follows:

Pixel-wise loss: In order to enhance the ability to preserve the identity information in the view-specific features, we first minimize the pixel-wise reconstruction error between the fake images and the real images:

$$\min_{E,V,G} \mathcal{L}_p = \mathbb{E}\left\| G\big(V(E(\boldsymbol{x}^u), v, u), \boldsymbol{k}\big) - \hat{\boldsymbol{x}}^v \right\|_1 \tag{7}$$

where $\|\cdot\|_1$ denotes ℓ_1 norm. E, V and G represent the encoder, the view transform layer and the generator respectively.

Multi-task adversarial loss: Using the same token as in pixel-wise loss, the networks D, E, V, G can be trained by the following multi-task adversarial loss:

$$\min_{E,V,G} \max_{D} \mathcal{L}_a = \mathbb{E}\big[\boldsymbol{s}^T \log D(\hat{\boldsymbol{x}}^v)\big] + \mathbb{E}\left[\boldsymbol{s}^T \log\Big(1 - D\big(G(V(E(\boldsymbol{x}^u), v, u), \boldsymbol{k})\big)\Big)\right] \tag{8}$$

where $\boldsymbol{s} \in \{0, 1\}^{n_v + n_c + n_d}$ is one-hot encoding vector, represented by concatenating the one-hot representation of v, k and the identity of gait image. It is noted that the output of the discriminator is a vector with the same dimension as $\boldsymbol{s}$ and each scalar output corresponds to one adversarial task. Each value in $\boldsymbol{s}$ decides which adversarial tasks the gait images should join in.

The final objective function of E, V, G and D is described as:

$$\min_{E,V,G} \max_{D} \mathcal{L} = \mathcal{L}_p + \gamma \mathcal{L}_a \tag{9}$$

The parameter γ plays a trade-off between the pixel-wise loss and multi-task adversarial loss.

Cross-entropy loss: In our method, the view angle prediction is considered as a classification problem. Let C represent the view-angle classifier, then the cross-entropy loss is defined as follows:

$$\min_{C} \mathcal{L}_v = \mathbb{E}[\boldsymbol{u}^T \log C(\boldsymbol{x}^u)] \tag{10}$$

where $\boldsymbol{u}$ is the one-hot representation of u.

IV. Experiments

A. Experimental setup

First, we introduce three datasets used in our experiments and then describe the training details. We also state the evaluation metric.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2018.2844819, IEEE Transactions on Information Forensics and Security

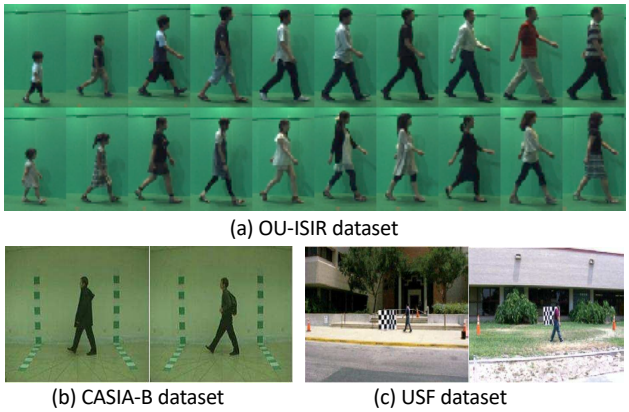


Fig. 4. Some examples from the three benchmark gait datasets [42]–[44].

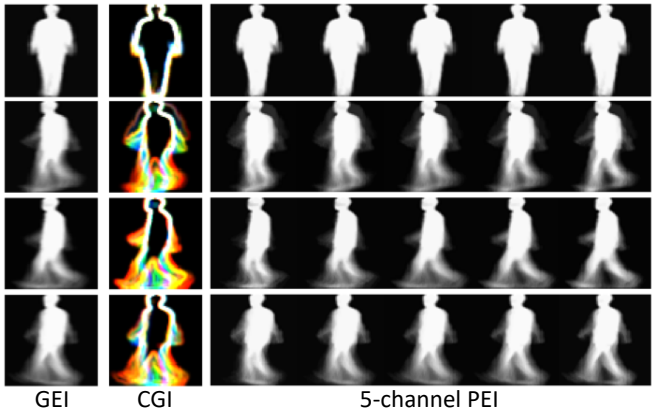


Fig. 5. GEI, CGI and PEI of one subject under four different view angles are shown. The templates are constructed by gait sequences in the CASIA-B dataset.

1) Datasets: Three widely used gait datasets are selected in our experiments, including the OU-ISIR, CASIA-B and USF. Some examples from the three datasets are illustrated in Fig. 4.

OU-ISIR [42] is a gait dataset with large population. It is made up of 4007 subjects (2135 males and 1872 females) with ages ranging from 1 to 94 years and has 4 different view angles $55°$, $65°$, $75°$ and $85°$. Our experimental setting is same as in [21]. Consequently, there are totally 3836 subjects used in the subsequent experiments. Note that the original silhouettes have already been cropped and aligned. We directly use the given silhouettes to construct the gait templates.

CASIA-B [43] is a widely used cross-view gait dataset. It contains 124 different subjects and 11 different view angles $0°$, $18°$, $36°$, $54°$, $72°$, $90°$, $108°$, $126°$, $144°$, $162°$ and $180°$. Each subject has six sequences in normal walking (NM01-NM06), two sequences in coats (CL01-CL02) and two sequences with bags (BG01-BG02) in each view. Obviously, CASIA-B has a larger range of view angles but relatively less number of subjects compared with OU-ISIR.

USF [44] is also a commonly used gait dataset. There are totally 122 different subjects in USF. The gait sequences of each subject are sampled under the effects of five covariates, which is closer to a practical scenario. The covariates include view angles, shoe types, walking surface, carrying conditions and time instants. In our experiment, the first probe set is used to evaluate the performance of our method because it is the only probe set that is different from the gallery set by view angle.

2) Training details: We employ the training strategy proposed in [36]. The parameter γ in Equation (9) is empirically set as 1×10^{-5}. The weights of each layer are initialized by using a Gaussian distribution with a mean of zero and a standard deviation of 0.01. All the bias terms are initialized to zero. The weight decay is set as 1.5×10^{-4}. We update the parameters with RMSProp [45] and set mini-batch size as 128. The learning rate is set as 5×10^{-5}. For each mini-batch, we feed pairs of samples with same identities to the networks. All gait templates are scaled to 64×64 pixels. For the CASIA-B and USF datasets, we apply data augmentation to the input data of the encoder. For each channel in gait templates, we first randomly shift $0 - 8$ pixel along the horizontal axis and then resize the images by $0.8 - 1.0$ with zero padding. The dimension of the view-specific feature is set as 128 on the CASIA-B and USF datasets and 512 on the OU-ISIR dataset because we find that a small dimension is hard to converge on the OU-ISIR dataset. The model is implemented by the PyTorch. We provide the source code on Github[1].

3) Evaluation metrics: Rank-1 identification accuracy is calculated in most cases by the nearest neighbor classifier in the latent space. For each feature in the probe set, we search the nearest feature in the gallery set based on the Euclidean distance and then check whether they have the same identity.

B. Effect of PEI

To evaluate the performance of the proposed PEI template, we compare PEI with Gait Energy Image (GEI) [13] and Chrono-Gait Image (CGI) [19] by using our proposed MGANs model. Some examples of GEI, CGI and PEI are shown in Fig. 5. Each channel of PEI is visualized independently. The number of channels n_c and the window size m are set to 5 and 1 respectively in the following experiments.

For the OU-ISIR dataset, we apply five-fold cross-validation on the whole dataset. For the CASIA-B dataset, the first 62 subjects are used for training and the remaining subjects are used for testing. In the testing phase, the first four normal walking gait sequences (NM01-NM04) are put into the gallery set and the others (NM05-NM06) into the probe set. For the USF dataset, we randomly split subjects into the gallery and probe sets, each of which contains 61 subjects. The model is trained five times repeatedly, and the average accuracy is reported.

It can be seen from Table II that MGANs achieve a significant improvement on the CASIA-B and USF datasets, while the improvement on the OU-ISIR dataset is deficient. This may be caused by a shorter average sequence length of gait in the OU-ISIR dataset. Note that we have not reported the performance of CGI on the OU-ISIR dataset because the number of gait cycles in this dataset is too small to effectively construct CGI.

[1]https://github.com/zztant/MGANs

TABLE II
AVERAGE RANK-1 ACCURACIES (%) OF DIFFERENT GAIT TEMPLATES ON THE OU-ISIR, CASIA-B AND USF GAIT DATASETS. THE NUMBER OF CHANNELS AND THE WINDOW SIZE OF PEI ARE SET AS 5 AND 1, RESPECTIVELY.

	PEI+MGANs	GEI+MGANs	CGI+MGANs
OU-ISIR	93.1	**93.2**	/
CASIA-B	**74.6**	70.1	63.5
USF	**94.7**	93.3	90.1

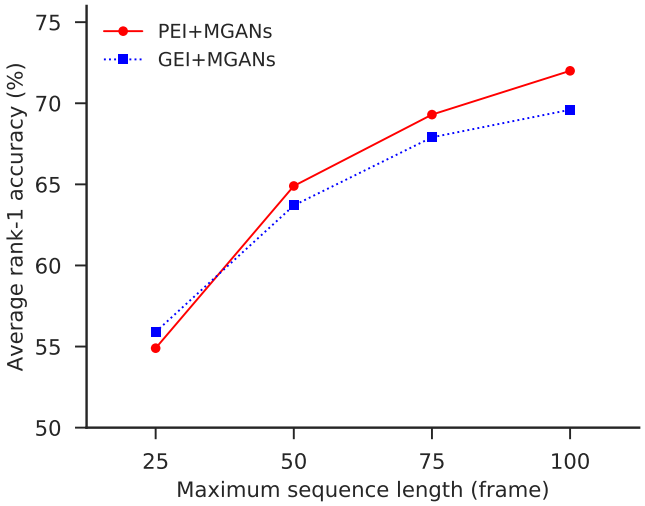


Fig. 6. The effect of the sequence length on the performance of PEI and GEI on the CASIA-B dataset.

In order to prove our claim that the sequence length affects the performance of PEI, we perform the experiment on the CASIA-B dataset which has generally more than one gait cycle. By controlling the maximum sequence length, we evaluate the performance of PEI. As shown in Fig. 6, our proposed PEI with MGANs can outperform GEI with MGANs when there are moderate frames (more than 50 in our experiment) in gait sequences. In addition, as the number of frames increases, the performance gap becomes larger. It indicates that PEI is able to preserve more temporal information.

We conclude the following three points to explain the reason why PEI is better than GEI if the dataset has more frames. First, GEI is able to reduce the effect of noise by averaging the silhouette images. Similarly, each channel of PEI is built in the same way as GEI to reduce the silhouette noise. If the data have more frames, each channel of PEI can incorporate more silhouette images, which can reduce noise in each channel. Second, when the sequence length and the window size are sufficient to suppress noise in each channel, more channels can preserve more temporal information. Third, the improvement of performance is indeed at the cost of higher computational and spatial complexity, which has been shown in Subsection III-B.

C. Effect of MGANs

In this subsection, we evaluate the MGANs model on the CASIA-B dataset. The experimental setup is the same as in Subsection IV-B.

The performance of view prediction using our classifier and the influence of the classifier are shown in Table III. Except for 90° and 108°, our classifier achieves 90% or higher accuracy on view prediction. This is because that the real difference between these view angles on the CASIA-B dataset is subtle. It is obvious that the performance of view prediction influences the subsequent view transformation, and thus the final performance of gait recognition. We also show its influence in Table III. As a comparison, the recognition accuracies under the ground truth probe views are also reported in the same table. From the results, it can be seen that compared with using the ground truth probe views, the degradation of recognition performance is subtle.

We then report the rank-1 accuracies among all view angles in Table IV. In order to verify effectiveness and significance of adversarial training, we train the networks with and without adversarial loss with PEI templates. It is not difficult to see that when trained only with pixel-wise loss or adversarial loss, our method achieves acceptable results. It is also noted that without pixel-wise loss, the model still works for recognition, which produces an effective way of adversarial training for supervised learning. After training the network with both mentioned losses, we obtain the best average recognition performance. It experimentally justifies that the adversarial training produces a positive effect on the generalization of the model.

In the next section, we compare our PEI+MGANs with several state-of-the-art methods in the recent literature.

D. Evaluation of comparison

1) Experimental result on OU-ISIR: We only compare our method with CNN-based method [21] because this method is better than others in most of the gait recognition datasets. We use the experimental setting described in Section IV-B which is the same as in [21]. For a fair comparison with the CNN-based method, we only report the average rank-1 accuracies of MGANs with GEI template. The score of the CNN-based method is directly taken from the original paper. The results are listed in Table V. When both using GEI as gait template, we achieve 93.2% average rank-1 accuracies which have a slight improvement compared with the CNN-based method.

2) Experimental result on CASIA-B: The performance of cross-view gait recognition under the normal walking condition is reported. Several state-of-the-art methods are chosen, including Auto-encoder [14], KCCA [46], C3A [8], ViDP [7], VTM-SVR [9] and CNN [21]. The recognition accuracy of the probe views 54°, 90° and 126° on the CASIA-B dataset with the first 62 training subjects are shown in Fig 7. By comparing the accuracies under all gallery views excluding identical-view cases, we can see that our proposed method outperforms others under different probe views. When the view variation is large, our method can still handle well while others show the sharper decreases of recognition accuracies. With the first 74 training subjects, the average recognition accuracies of our proposed method under the probe views 54°, 90° and 126° are listed in Table VI. It can be seen from Table VI that our method outperforms KCCA, C3A and ViDP. The result demonstrates

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2018.2844819, IEEE Transactions on Information Forensics and Security

TABLE III
CLASSIFICATION ACCURACIES OF THE VIEW-ANGLE CLASSIFIER ON THE CASIA-B DATASET. PROBE VIEWS: 0° TO 180°. WE ALSO COMPARE THE PERFORMANCE UNDER GROUND TRUTH PROBE VIEW WITH PREDICTED PROBE VIEW TO SHOW HOW THE VIEW-ANGLE CLASSIFIER AFFECT OUR RECOGNITION PERFORMANCE.

Probe view	0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°
Acc. of view-angle classifier	95.1	95.9	97.5	99.2	96.7	86.0	88.3	99.2	95.7	98.3	91.7
Acc. with predicted view	63.7	73.7	79.4	81.6	75.9	71.2	73.8	80.3	80.3	77.1	63.4
Acc. with ground truth view	63.9	73.5	79.0	81.3	76.3	71.4	73.9	80.3	80.3	77.1	63.6

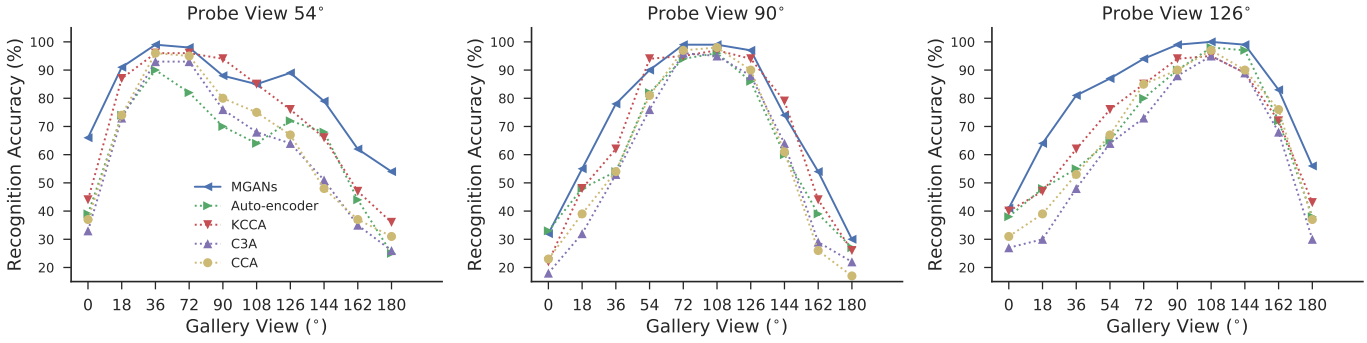


Fig. 7. Comparison with Auto-encoder [14], KCCA [46], C3A [8] and CCA [47] under the probe views 54°, 90° and 126°. Models are trained with the first 62 subjects in normal walking condition.

TABLE IV
RANK-1 ACCURACIES (%) OF DIFFERENT EXPERIMENTAL SETTINGS ON THE CASIA-B DATASET IN NORMAL WALKING CONDITION. MODELS ARE TRAINED WITH THE FIRST 62 SUBJECTS. PX: TRAINED ONLY WITH PIXEL-WISE LOSS. AD: TRAINED ONLY WITH ADVERSARIAL LOSS. PROBE VIEWS: 0° TO 180° WITHOUT IDENTICAL-VIEW CASES.

A + B	PEI + MGANs	PEI + PX	PEI + AD	GEI + MGANs	CGI + MGANs
0°	63.7	**65.8**	53.4	58.0	54.0
18°	**73.7**	73.7	66.2	65.9	63.1
36°	**79.4**	76.0	74.3	73.0	68.2
54°	**81.6**	79.4	75.0	77.1	68.9
72°	**75.9**	73.1	69.8	73.1	63.2
90°	**71.2**	64.6	64.0	67.3	56.0
108°	**73.8**	72.2	70.3	69.9	61.8
126°	80.3	**81.2**	78.4	78.6	69.6
144°	80.3	**82.1**	74.5	77.9	69.6
162°	**77.1**	74.6	68.9	71.9	71.2
180°	**63.4**	59.8	55.0	58.4	52.5
Ave.	**74.6**	73.0	68.2	70.1	63.5

TABLE V
RANK-1 ACCURACIES (%) ON THE OU-ISIR DATASET. WE APPLY FIVE-FOLD CROSS-VALIDATION ON THE WHOLE DATASET. THE STANDARD DEVIATIONS ARE ALSO REPORTED.

Probe View	Gallery View	GEI+MGANs	CNN [21]
55°	65°	**99.4±0.1**	98.3±0.1
55°	75°	**96.1±0.3**	96.0±0.1
55°	85°	77.9±0.4	**80.5±0.4**
65°	55°	**97.7±0.1**	96.3±0.2
65°	75°	**98.5±0.1**	97.3±0.0
65°	85°	**84.4±0.5**	83.3±0.3
75°	55°	**94.8±0.1**	94.2±0.2
75°	65°	**98.9±0.1**	97.8±0.2
75°	85°	**86.4±0.3**	85.1±0.1
85°	55°	86.9±0.6	**90.0±0.5**
85°	65°	**97.4±0.3**	96.0±0.3
85°	75°	**99.5±0.1**	98.4±0.1
Average		**93.2**	92.4

TABLE VI
RANK-1 ACCURACIES (%) UNDER THE PROBE VIEWS 54°, 90° AND 126°, EXCLUDING IDENTICAL-VIEW CASES. WE COMPARE OUR METHOD WITH C3A [8], VIDP [7], VTM-SVR [9] AND CNN [21] . MODELS ARE TRAINED WITH THE FIRST 74 SUBJECTS.

	54°	90°	126°	Average
PEI+MGANs	84.2	72.3	83.0	79.8
C3A [8]	75.7	63.7	74.8	71.4
ViDP [7]	64.2	60.4	65.0	63.2
VTM-SVR [9]	55.0	46.0	54.0	51.0
CNN [21]	**94.6**	**88.3**	**93.8**	**92.2**

that our view-specific feature is superior to view-invariant feature in other methods except for CNN [21].

There are several possible reasons to explain why [21] obtains better performance on the CASIA-B dataset. First, since there is no the fully-connected layer in their models, the number of adjustable parameters is greatly reduced, making the model work well on the small dataset. Second, they utilized the 3D convolution layer and the feature map averaging to extract temporal information directly from the frame sequences. Third, they noticed that matching local features at the bottom layers is helpful to obtain better performance after analyzing the influence of different network structures. Such a matching strategy cannot be employed directly in our proposed method due to the limitation of the encoder-decoder structure of our model.

Although the performance is better, the CNN-based method is a black-box model to deal with the view variation for cross-view gait recognition. Compared with the CNN-based method, one advantage of our method is that we can learn view-specific features by using prior knowledge about the view

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2018.2844819, IEEE Transactions on Information Forensics and Security

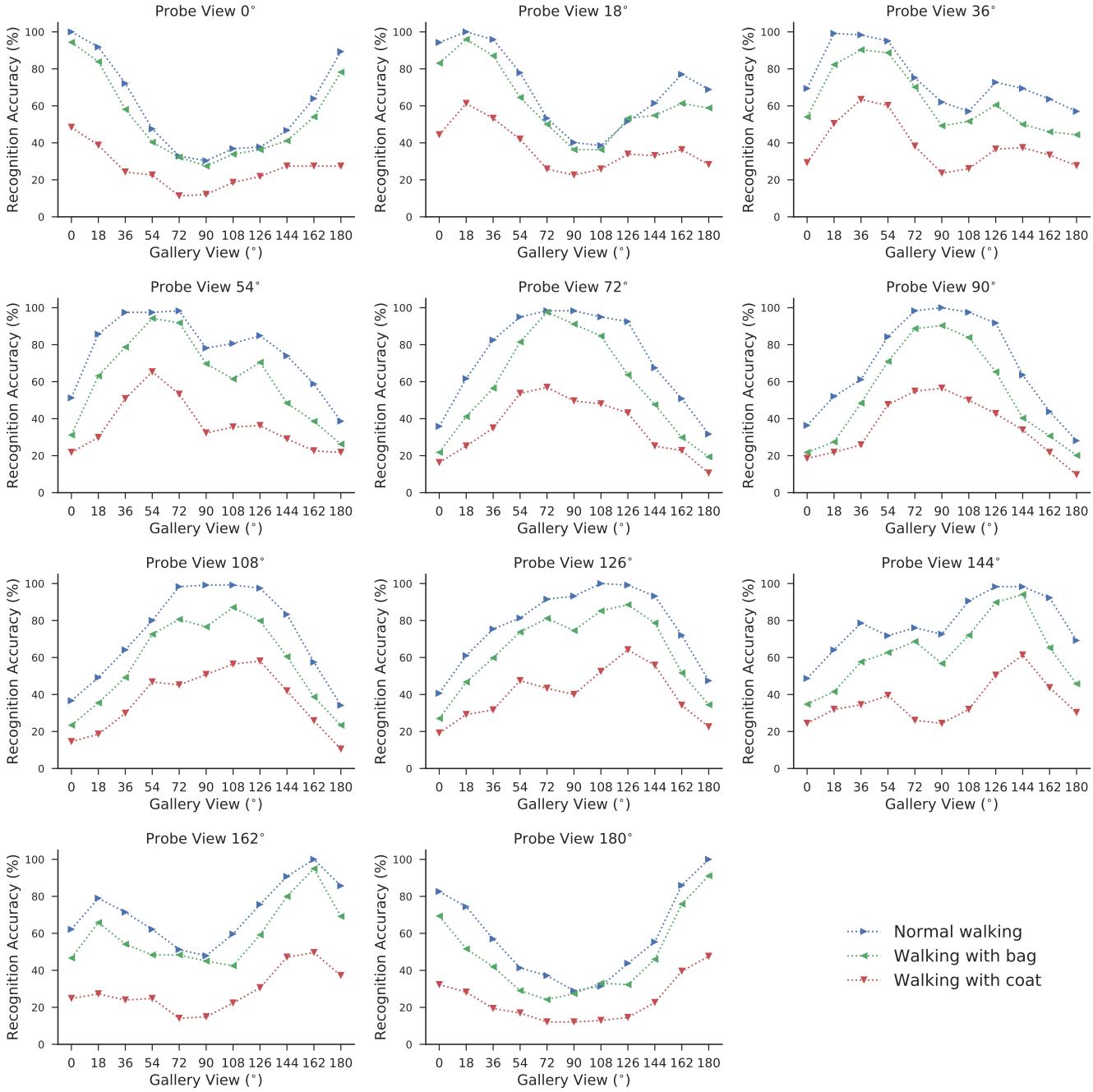


Fig. 8. There are three different walking conditions on the CASIA-B dataset including normal walking (NM), walking in coats (CL) and walking with bags (BG). The figure shows rank-1 accuracies (%) by three different probe sets NM, CL and BG. In the training phase, we use the first 62 subjects to train our MGANs model.

angles and utilize view manifold to model the view variation. Therefore, when discriminating whether the two gait images from different view angles have the same identity, the proposed method that operates view-specific features precisely from one view to another view, is more straightforward and interpretable to the features regarding the view variation, rather than just learning a similarity of a pair of gait images as CNN-based model did. In addition, since the generator in our model has the ability to generate gait images from the view-specific features, the generated gait images can facilitate understanding of the view variation from learned features in our model, whereas the CNN-based method can only produce the similarity of a pair of gait images.

Finally, we analyze the performance of our model under the variations of different walking conditions. There are a total of three different walking conditions on the CASIA-B dataset including normal walking (NM), walking in coats (CL) and walking with bags (BG). In order to train the model which is able to handle these variations, we not only feed the normal walking gait templates but also the gait templates with clothing and carrying variations to the encoder. Therefore, the input of the encoder includes gait templates in all walking conditions.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2018.2844819, IEEE Transactions on Information Forensics and Security

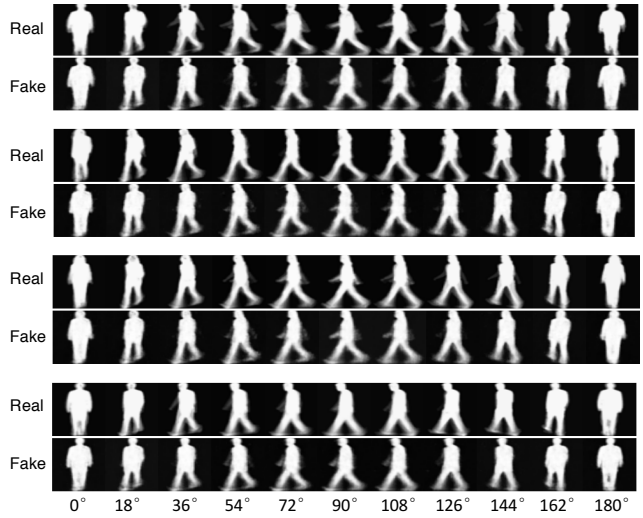


Fig. 9. Four real gait images from the testing set in 90° are used as inputs of the encoder. The fake gait images from all view angles based on these four gait images are visualized. For each example, the real images are also provided to compare at the first and second columns, respectively.

The model then reconstructs gait images in normal walking. In this manner, the view-specific features in our model can be robust to the clothing and carrying variations. In the testing phase, gait templates in normal walking are put into the gallery set. The rank-1 accuracies in each probe view are shown in Fig 8. The results indicate that the carrying variation has a negative impact on the performance. It is obvious that clothing variations lead to a worse performance compared with carrying variation.

3) Experimental result on USF: We compare our method with CNN [21] and EGG [48] under the condition of the same number of training and testing subjects on the USF dataset. CNN [21] is the best previous method which achieves 96.7±0.5% average rank-1 accuracy and EGG [48] achieves 93%. The accuracies of our method with PEI and GEI templates are 94.7±2.2% and 93.3±1.7% respectively which is also superior to EGG [48] but slightly inferior to CNN [21].

E. Gait generation

One of the advantages of our proposed MGANs model is that we can generate gait images which are understandable to humans according to the learned view-specific features. Given a gait template, after obtaining its view-specific feature, gait images under different view angles can be generated by using the proposed view transform layer and the generator. In Fig. 9, four real images from the testing set in 90° are used as inputs of the encoder. We then show the fake images from all view angles based on these four gait images. When comparing real and fake images, we have the following conclusions. First, we see that the fake images generated by our model are transformed from 90° to all other view angles successfully. This means that view transformation in the view transform layer works well in our method. Second, we see that the fake images from 54° to 128° are more similar to the real images than the fake images from 0° to 36° and 144° to 180°. Therefore, it is reasonable that our model can achieve better performance when the view variation is relatively small because the fake images are actually the visualization of the view-specific features.

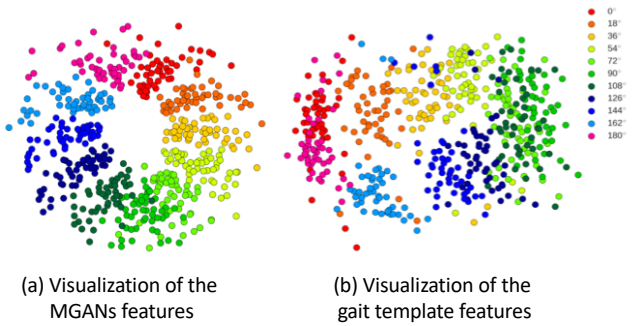


Fig. 10. Two-dimensional visualization of features by using PCA. Different colors represent different view angles. (a) MGANs features (best viewed in color). (b) Gait template features.

F. Feature visualization

The main difference between our model and previous models in the literature is that features learned by our model are view-specific, while other models prefer view-invariant features. By learning view-specific features, our model has the ability to capture the view manifold embedded in the data space and transform features to the same view angle by the view transform layer defined in Equation 5. In this subsection, we first train the model on the CASIA-B dataset and then visualize MGANs features of the testing samples in a two-dimensional subspace based on PCA. For comparison, we project the original gait templates onto the subspace by performing PCA. Their visualization results are shown in Fig. 10, where different colors represent different view angles and each circle represents one sample. It can be seen that our MGANs features are well separated according to the ground truth view angles, while the original gait template features confuse the samples from different view angles. In addition, MGANs features show the structure of the view manifold, which proves the assumption of the manifold in the view transform layer.

V. Discussion

In the field of cross-view gait recognition, there is still a lack of public datasets which have both large population and large view variations. This factor affects the possibility to train a unified and reliable model for practical scenarios in video surveillance. For example, if there are view angles not existing in the training set, the view-angle classifier may predict a biased view angle. In addition, it is expensive to obtain the labeled gait sequences. In this section, we summarize several possible ways to improve the practicality and effectiveness of our proposed model as follows.

Dynamic information: In the architecture of our MGANs model, the temporal pooling operation is directly added after the convolutional network to capture temporal information contained in the PEI template. However, this operation ignores

the order of channels and loses dynamic information of gait to some degree. Recent studies of gait recognition and person re-identification use Long-Short Term Memory (LSTM) or Gated Recurrent Units (GRU) to preserve the dynamic information between consecutive video frames [31]. It is worth investigating how to apply LSTM, GRU or any other similar modules to our PEI templates.

Semi-supervised learning: Generative adversarial networks have provided a possible way to utilize unlabeled data. However, how to utilize unlabeled samples for semi-supervised learning is still an open problem in the field of machine learning. Previous work such as [49] uses unpaired data for domain adaptation, which provides us with a possible way to utilize unlabeled gait sequences in our model. We will explore to enhance our method by exploiting unlabeled information in future work.

Pose-based method: Recent works on pose estimation [33], [50] makes it possible to use the dynamic parameters of human bodies for gait recognition. The positions of body-joints extracted in their methods provide us with features which are insensitive to clothing and carrying variations [2]. We believe that fusing such features with gait templates such as GEI and PEI is valuable to cross-view gait recognition.

VI. Conclusion

This paper has proposed a new gait template called PEI which is an extension of GEI to enrich the spatial and temporal information in cross-view gait recognition. Instead of performing view transformation step by step in latent and data spaces alternatively, we directly transform the view-specific features in the latent space when facing large view variations. By taking advantage of adversarial training to model the distribution of different domains, more representative features are extracted in our model. Experimental results have indicated that compared with other published methods, our MGANs model achieves the competitive performance and better interpretability of view variation on the OU-ISIR, USF, and CASIA-B benchmark datasets. It is worth noting that the existing gait dataset such as OU-ISIR did not collect from the forensic environment. And MGANs can only learn view-specific features from PEI. In the future, we will study how to extract view-invariant features, which is likely to be used in forensic applications, from gait sequence.

ACKNOWLEDGMENT

The authors would like to thank the reviewers and associate editor for their comments and constructive suggestions.

References

[1] P. Connor and A. Ross, "Biometric recognition by gait: A survey of modalities and features," *Comput. Vis. Image Understanding*, vol. 167, pp. 1–27, 2018.

[2] R. Liao, C. Cao, E. B. Garcia, S. Yu, and Y. Huang, "Pose-based temporal-spatial network (PSTN) for gait recognition with carrying and clothing variations," in *Proc. Chinese Conf. on Biometric Recog.*, pp. 474–483, 2017.

[3] J. Kovač, V. Štruc, and P. Peer, "Frame-based classification for cross-speed gait recognition," *Multimedia Tools and Appl.*, pp. 1–23, 2017.

[4] J. Zhang, J. Pu, C. Chen, and R. Fleischer, "Low-resolution gait recognition," *IEEE Trans. Syst., Man, Cybern., Cybern.*, vol. 40, no. 4, pp. 986–996, 2010.

[5] K. Bashir, T. Xiang, and S. Gong, "Cross view gait recognition using correlation strength.," in *Proc. Brit. Mach. Vis. Conf.*, pp. 1–11, 2010.

[6] Y. Makihara, A. Suzuki, D. Muramatsu, X. Li, and Y. Yagi, "Joint intensity and spatial metric learning for robust gait recognition," in *Proc. IEEE Comput. Vis. Pattern Recog.*, pp. 5705–5715, 2017.

[7] M. Hu, Y. Wang, Z. Zhang, J. J. Little, and D. Huang, "View-invariant discriminative projection for multi-view gait-based human identification," *IEEE Trans. Inf. Forensics Security*, vol. 8, no. 12, pp. 2034–2045, 2013.

[8] X. Xing, K. Wang, T. Yan, and Z. Lv, "Complete canonical correlation analysis with application to multi-view gait recognition," *Pattern Recog.*, vol. 50, pp. 107–117, 2016.

[9] W. Kusakunniran, Q. Wu, J. Zhang, and H. Li, "Support vector regression for multi-view gait recognition based on local motion feature selection," in *Proc. IEEE Comput. Vis. Pattern Recog.*, pp. 974–981, 2010.

[10] Y. Makihara, R. Sagawa, Y. Mukaigawa, T. Echigo, and Y. Yagi, "Gait recognition using a view transformation model in the frequency domain," in *Proc. Eur. Conf. Comput. Vis.*, pp. 151–163, Springer, 2006.

[11] W. Kusakunniran, Q. Wu, H. Li, and J. Zhang, "Multiple views gait recognition using view transformation model based on optimized gait energy image," in *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 1058–1064, 2009.

[12] W. Kusakunniran, Q. Wu, J. Zhang, H. Li, and L. Wang, "Recognizing gaits across views through correlated motion co-clustering," *IEEE Trans. Image Process.*, vol. 23, no. 2, pp. 696–709, 2014.

[13] J. Man and B. Bhanu, "Individual recognition using gait energy image," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 2, pp. 316–322, 2006.

[14] S. Yu, H. Chen, Q. Wang, L. Shen, and Y. Huang, "Invariant feature extraction for gait recognition using only one uniform model," *Neurocomputing*, vol. 239, pp. 81–93, 2017.

[15] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2016.

[16] G. Antipov, M. Baccouche, and J.-L. Dugelay, "Face aging with conditional generative adversarial networks," in *Proc. IEEE Int. Conf. on Image Process.*, pp. 2089–2093.

[17] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Comput. Vis. Pattern Recog.*, pp. 2414–2423, 2016.

[18] R. Yeh, C. Chen, T. Y. Lim, M. Hasegawa-Johnson, and M. N. Do, "Semantic image inpainting with deep generative models," in *Proc. IEEE Comput. Vis. Pattern Recog.*, 2017.

[19] C. Wang, J. Zhang, L. Wang, J. Pu, and X. Yuan, "Human identification using temporal information preserving gait template," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2164–2176, 2012.

[20] C. Wang, J. Zhang, J. Pu, X. Yuan, and L. Wang, "Chrono-gait image: A novel temporal template for gait recognition," in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2010.

[21] Z. Wu, Y. Huang, L. Wang, X. Wang, and T. Tan, "A comprehensive study on cross-view gait based human identification with deep CNNs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 2, pp. 209–226, 2017.

[22] G. Zhao, G. Liu, H. Li, and M. Pietikainen, "3D gait recognition using multiple cameras," in *Proc. 7th Int. Conf. Automatic Face and Gesture Recognition*, pp. 529–534, 2006.

[23] G. Ariyanto and M. S. Nixon, "Model-based 3D gait biometrics," in *Proc. Int. Joint Conf. Biometrics*, pp. 1–7, 2011.

[24] R. Bodor, A. Drenner, D. Fehr, O. Masoud, and N. Papanikolopoulos, "View-independent human motion classification using image-based reconstruction," *J. Image Vis. Comput.*, vol. 27, no. 8, pp. 1194–1206, 2009.

[25] M. Goffredo, I. Bouchrika, J. N. Carter, and M. S. Nixon, "Self-calibrating view-invariant gait biometrics," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 40, no. 4, pp. 997–1008, 2010.

[26] W. Kusakunniran, Q. Wu, J. Zhang, Y. Ma, and H. Li, "A new view-invariant feature for cross-view gait recognition," *IEEE Trans. Inf. Forensics Security*, vol. 8, no. 10, pp. 1642–1653, 2013.

[27] J. Han, B. Bhanu, and A. K. Roy-Chowdhury, "A study on view-insensitive gait recognition," in *Proc. IEEE Int. Conf. Image Process.*, vol. 3, pp. III–297, 2005.

[28] F. Jean, R. Bergevin, and A. B. Albu, "Computing and evaluating view-normalized body part trajectories," *J. Image Vis. Comput.*, vol. 27, no. 9, pp. 1272–1284, 2009.

[29] C. Luo, W. Xu, and C. Zhu, "Robust gait recognition based on partitioning and canonical correlation analysis," in *Proc. IEEE Int. Conf. on Imaging Syst. and Tech.*, pp. 1–5, 2015.

[30] K. Shiraga, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi, "Geinet: View-invariant gait recognition using a convolutional neural network," in *Proc. IEEE Int. Conf. on Biometrics*, pp. 1–8, 2016.

[31] Y. Feng, Y. Li, and J. Luo, "Learning effective gait features using LSTM," in *Proc. IEEE Int. Conf. on Pattern Recog.*, pp. 320–325, 2016.

[32] N. Takemura, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi, "On input/output architectures for convolutional neural network-based cross-view gait recognition," *IEEE Trans. Circuits Syst. Video Technol.*, pp. 1–1, 2017.

[33] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE Comput. Vis. Pattern Recog.*, p. 7, 2017.

[34] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst*, pp. 2672–2680, 2014.

[35] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. Int. Conf. on Mach. Learn.*, pp. 214–223, 2017.

[36] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved training of Wasserstein GANs," pp. 5769–5779, 2017.

[37] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[38] S. Yu, H. Chen, E. B. G. Reyes, and N. Poh, "GaitGAN: Invariant gait feature extraction using generative adversarial networks," in *Proc. IEEE Comput. Soc. Workshop Biometrics*, 2017.

[39] T. H. Lam, K. H. Cheung, and J. N. Liu, "Gait flow image: A silhouette-based gait representation for human identification," *Pattern Recog.*, vol. 44, no. 4, pp. 973–987, 2011.

[40] P. Arora and S. Srivastava, "Gait recognition using gait Gaussian image," in *Proc. IEEE Int. Conf. on Sign. Proces. and Integr. Networks*, pp. 791–794, 2015.

[41] N. McLaughlin, J. Martinez del Rincon, and P. Miller, "Recurrent convolutional network for video-based person re-identification," in *Proc. IEEE Comput. Vis. Pattern Recog.*, pp. 1325–1334, 2016.

[42] H. Iwama, M. Okumura, Y. Makihara, and Y. Yagi, "The OU-ISIR gait database comprising the large population dataset and performance evaluation of gait recognition," *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 5, pp. 1511–1521, 2012.

[43] S. Yu, D. Tan, and T. Tan, "A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition," in *IEEE Int. Conf. Pattern Recog.*, vol. 4, pp. 441–444, 2006.

[44] S. Sarkar, P. J. Phillips, Z. Liu, I. R. Vega, P. Grother, and K. W. Bowyer, "The humanID gait challenge problem: Data sets, performance, and analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 2, pp. 162–177, 2005.

[45] T. Tieleman and G. Hinton, "Lecture 6.5-RMSProp: Divide the gradient by a running average of its recent magnitude," *COURSERA: Neural Networks for Machine Learning*, vol. 4, no. 2, 2012.

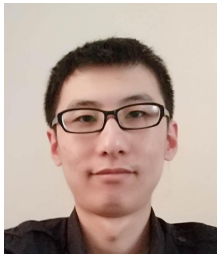
[46] P. L. Lai and C. Fyfe, "Kernel and nonlinear canonical correlation analysis," *World Scientific Int. J. Neural Syst.*, vol. 10, no. 05, pp. 365–377, 2000.

[47] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Comput.*, vol. 16, no. 12, pp. 2639–2664, 2004.

[48] H. Hu, "Enhanced Gabor feature based classification using a regularized locally tensor discriminant model for multiview gait recognition," *IEEE Trans. Biomed. Circuits Syst.*, vol. 23, no. 7, pp. 1274–1286, 2013.

[49] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017.

[50] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2D human pose estimation: New benchmark and state of the art analysis," in *Proc. IEEE Comput. Vis. Pattern Recog.*, 2014.



Yiwei He received the B.S. degree from the Nanjing Audit University, China, in 2015. He is currently pursuing the Master's degree in Computer Science at Fudan University. His research interests include machine learning, computer vision and gait recognition.

Junping Zhang (M'05) received the B.S. degree in automation from Xiangtan University, Xiangtan, China, in 1992. He received the M.S. degree in control theory and control engineering from Hunan University, Changsha, China, in 2000. He received his Ph.D. degree in intelligent system and pattern recognition from the Institute of Automation, Chinese Academy of Sciences, in 2003. He is a professor at School of Computer Science, Fudan University since 2011. His research interests include machine learning, image processing, biometric authentication, and intelligent transportation systems. He has been an associate editor of the IEEE INTELLIGENT SYSTEMS since 2009 and the IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS since 2010. He has widely published in highly ranked international journals such as IEEE TPAMI and IEEE TNNLS, and leading international conferences such as ICML and ECCV.

Hongming Shan received the B.S. degree from Shandong University of Technology, China, in 2011 and obtained a Ph.D. degree from Fudan University, China, in 2017. He is currently a post-doc at Rensselaer Polytechnic Institute, USA. His research interests include machine/deep learning, computer vision, dimensionality reduction, and biomedical imaging.

Liang Wang (SM'09) received both the BEng and MEng degrees from Anhui University in 1997 and 2000, respectively, and the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences (CASIA) in 2004. Currently, he is a full professor of the Hundred Talents Program at the National Lab of Pattern Recognition, CASIA. He has widely published in highly ranked international journals such as IEEE TPAMI and IEEE TIP, and leading international conferences such as CVPR, ICCV, and ICDM. He is a senior member of the IEEE, and an IAPR Fellow.