Pattern Recognition 87 (2019) 179–189

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

GaitGANv2: Invariant gait feature extraction using generative adversarial networks

Shiqi Yu[a,*], Rijun Liao[a], Weizhi An[a], Haifeng Chen[a], Edel B. García[b], Yongzhen Huang[c,d], Norman Poh[e]

[a] College of Computer Science and Software Engineering, Shenzhen University, China
[b] Advanced Technologies Application Center (CENATAV), Cuba
[c] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, China
[d] Watrix technology limited co. ltd, China
[e] Trust Stamp, USA

ARTICLE INFO

Article history:
Received 2 March 2018
Revised 21 August 2018
Accepted 16 October 2018
Available online 18 October 2018

Keywords:
Gait recognition
Generative adversarial networks
Invariant feature

ABSTRACT

The performance of gait recognition can be adversely affected by many sources of variation such as view angle, clothing, presence of and type of bag, posture, and occlusion, among others. To extract invariant gait features, we proposed a method called GaitGANv2 which is based on generative adversarial networks (GAN). In the proposed method, a GAN model is taken as a regressor to generate a canonical side view of a walking gait in normal clothing without carrying any bag. A unique advantage of this approach is that, unlike other methods, GaitGANv2 does not need to determine the view angle before generating invariant gait images. Indeed, only one model is needed to account for all possible sources of variation such as with or without carrying accessories and varying degrees of view angle. The most important computational challenge, however, is to address how to retain useful identity information when generating the invariant gait images. To this end, our approach differs from the traditional GAN in that GaitGANv2 contains two discriminators instead of one. They are respectively called fake/real discriminator and identification discriminator. While the first discriminator ensures that the generated gait images are realistic, the second one maintains the human identity information. The proposed GaitGANv2 represents an improvement over GaitGANv1 in that the former adopts a multi-loss strategy to optimize the network to increase the inter-class distance and to reduce the intra-class distance, at the same time. Experimental results show that GaitGANv2 can achieve state-of-the-art performance.

1. Introduction

Gait is a behavioural biometric modality with a great potential for person identification because of its unique advantages such as being contactless, hard to fake and passive in nature, i.e., it requires no explicit cooperation from the subjects. Furthermore, the gait features can be captured at a distance in uncontrolled scenarios. Therefore, gait recognition is a very valuable technique in video surveillance, with a wide-ranging applications. This is particular so since many surveillance cameras have already been installed in major cities around world. Therefore, by continually improving its accuracy, the gait recognition technology will certainly add to the repertoire of tools available for crime prevention and forensic identification. For this reason, gait recognition is and will become an ever more important research topic in the computer vision community.

Unfortunately, automatic gait recognition remains a challenging task because it suffers from many potential sources of variation that can alter the human appearance drastically, such as, but not limited to aspects such as viewpoint, clothing, and objects being carried. These variations can affect the recognition accuracy greatly. Among these sources of variation, view angle is one of the most common one because we can not control the walking directions of subjects in real applications, and that is the central focus of our work here.

As a proof of concept, we shall consider variability in conditions of consisting of view angle, choice of clothing and type of objects being carried by the subject. The proposed generative adversarial networks (GAN) can handle all these variations *simultaneously* by using only one model. GAN acts as a regressor which takes an gait image captured with any combination of the above-mentioned

[*] Corresponding author.
E-mail address: shiqi.yu@szu.edu.cn (S. Yu).

https://doi.org/10.1016/j.patcog.2018.10.019
0031-3203/© 2018 Elsevier Ltd. All rights reserved.

sources of variation and then transforms it into a canonical side view image The method can do so without any knowledge regarding the factors that contribute to the gait variability. The most important computational challenge, however, is to address how to retain useful identity information when generating the canonical, invariant gait images.

The rest of the paper is organized as follows. Section 2 presents the state-of-the-art literature in gait recognition that deals with invariance in gait recognition. Section 3 describes the proposed method. Experiments and evaluation are presented in Section 4. The last section, Section 5, gives the conclusions and identifies future work.

2. Related work

To reduce the effect of different kinds of variations is what is concerned about by most gait recognition methods. Early literature such as [1] uses static body parameters measured from gait images as a kind of view-invariant feature. Kale et al. [2] used the perspective projection model to generated side view features from arbitrary views. Unfortunately, the relation between two views is hard to be modelled by a simple linear function, which is achieved via the perspective projection model.

Some other researchers [3,4] tried to build a 3D model for different human bodies so that any arbitrary 2D view can be generated by projecting the 3D model at any desirable angle. Unfortunately, this method usually requires multiple calibrated cameras installed in a fully-controlled environment and subjects to be co-operative, that is, they are told to walk in a particular direction. In [5], Tang et al. proposed a gait partial similarity matching method that assumed a 3D project shares commons view surfaces at different views. A 3D human body model can be built based on the silhouettes, and then improve silhouettes can be obtained from the 3D model.

In order to attain more robustness with respect to view-angle variation, the most commonly used model is arguably the view transformation model (VTM) which transforms a gait feature from one view to another view. Makihara et al. [6] designed a VTM named as FD-VTM that works in the frequency-domain. Different from FD-VTM, RSVD-VTM proposed in [7] operates in the spatial domain. It uses reduced singular value decomposition (SVD) to construct a VTM and then produces an optimal Gait Energy Image (GEI) feature vector based on linear discriminant analysis (LDA). RSVD-VTM achieves good results. Motivated by the capability of robust principal component analysis (RPCA) for feature extraction, Zheng et al. [8] achieved a robust VTM via RPCA for view invariant feature extraction. By considering view transformation as a regression problem, Kusakunniran et al. [9] used elastic net as means of achieving a sparse VTM-based regression model. VTM can also be achieved using canonical correlation analysis(CCA). Bashir et al. [10] formulated a gaussian process classification framework to estimate view angle in the probe set, then used CCA to model the correlation of gait sequences from different, arbitrary views. Luo et al. [11] proposed a gait recognition method based on partitioning and CCA. They separated a GEI image into 5 non-overlapping parts, and for each part they used CCA to model the correlation. In [12], Xing et al. also used CCA; but they reformulated the traditional CCA so that it can deal with a high-dimensional matrix, and reduced the computational burden in view-invariant feature extraction. Last but not least, Lu et al. [13] proposed a method that can handle arbitrary walking directions by using cluster-based averaged gait images. However, if there is no view with similar walking direction in the gallery set, the recognition rate will decrease.

For most VTM-related methods, a view transformation model [6–9,14,15] can only transform one specific view angle to another one. The model heavily depends on the accuracy of view angle estimation. Furthermore, in order to transform gait images from any arbitrary view angle to a specific view, a lot of models are needed. To overcome this limitation, recently researchers have tried to achieve view invariance using only one model. For instance, Hu et al. [16] proposed a method named as ViDP which extracts view invariant features using a linear transform. Hu [17] also applied regularized local tensor discriminant analysis (RLTDA) which can capture nonlinear manifolds as a means to achieving dimensionality reduction. However, the method is sensitive to initialization. A similar method based on the tensor representation can also be find in [18]. Instead of using a linear transformation, Wu et al. [19] trained deep convolution neural networks for any view pairs; thus achieved a high recognition accuracy.

Besides variation in view, clothing can also change the human body appearance as well as shape greatly. Some clothes, such as long overcoats, can occlude the leg motion. Carrying condition is another factor which can affect feature extraction since it is not easy to separate the carried object from a human body just from the image information. In the literature, there are a few methods that can achieve clothing invariance in gait recognition, unlike its view invariant counterpart. In [20] , clothing invariance is achieved by dividing the human body into 8 parts, each of which is subject to discrimination analysis. In [21], Guan et al. proposed a random subspace method (RSM) for clothing-invariant gait recognition by combining multiple inductive biases for classification. One recent method named as SPAE in [22] can extract invariant gait features using only a single model that can handle angle, clothing and carry conditions. Wu et al. [19] adopted convolutional neural networks (CNNs) to optimize the feature extraction and achieved state-of-the-art performance.

Recently, thanks to the advancements in human pose estimation, it is now possible to use joints information from a raw image for gait recognition. In addition, some methods focus on pose-based gait recognition [23–26]. The rationale for this is that human joints are invariant to objects being carried and the choice of clothing. Some researchers have studied the problem as a pose-based gait recognition, like Liang et al. [23] who used skeleton data provided by low-cost Kinect sensors. In [24], instead of using binary silhouette to describe each frame, Feng et al. used a human-body joint heatmap. They fed the joint heatmap of consecutive frames to a Long Short Term Memory (LSTM), which is a kind of recurrent neural network, to extract the gait features. Liao et al. [25] proposed a pose-based gait recognition method and used the posed-based temporal-spatial network (PTSN) to extract the invariant features. The more recent work in [26] employ the estimated 3D pose for gait recognition to improve the robustness to view variation. However, those pose-based methods are still challenging due to the fact that a pose inherently contains very little information about the subject identity.

In this paper, we propose to use generative adversarial networks (GAN) as a means to robustly recognising gait against adverse factors such as view angle, clothing and carrying condition *simultaneously* using only a single model. GAN is inspired by the two person zero-sum game in Game Theory, developed by Goodfellow et al. [27] in 2014. The result of the theory is a model that is composed of one generative model G and another discriminative model D. While the generative model captures the distribution of the training data, the discriminative model is a second classifier that determines whether the input is real or generated. In order to optimize the parameters of these two models, the problem can be cast as a minimax two-player game. In this competitive learning scenario, the generative model attempts to produce a realistic image from an input random vector z. As we know the early GAN model is too flexible in generating image. In [28], Mirza et al. fed

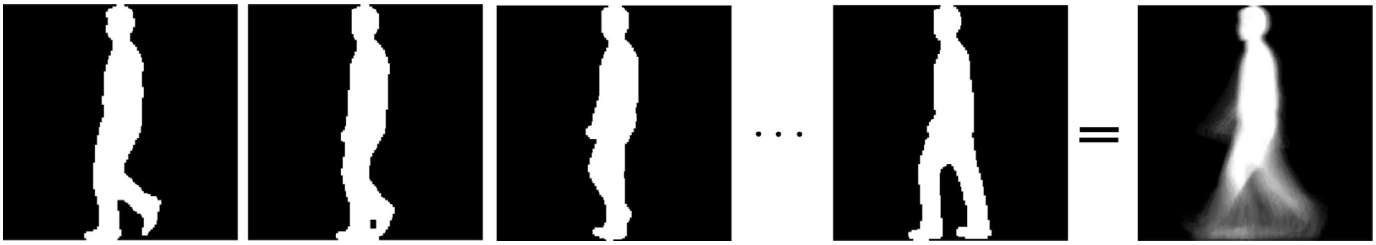


Fig. 1. A gait energy image (the right most one) is produced by averaging all the silhouettes (all the remaining images on the left) in one gait cycle.

a conditional parameter y into both the discriminator and generator as additional input layer to increase the constraint. Meanwhile, Denton et al. proposed a method using a cascade of convolutional networks within a Laplacian pyramid framework to generate images in a coarse-to-fine fashion. Unfortunately, early GANs are not only hard to train, but its generator often produces nonsensical outputs. To overcome these limitations, Radford et al. proposed the Deep Convolutional GAN [29] which contains a series of strategies such as using fractional-strided convolutions and batch normalization. This makes GAN more stable in training. Recently, Yoo et al. [30] presented an image-conditional generation model which contains a vital component named domain-discriminator. This discriminator ensures that a generated image is relevant to its input image. Furthermore, this method proposes domain transfer using GANs at the pixel level; and is subsequently known as pixel-level domain transfer GAN, or PixelDTGAN in [30].

3. Proposed method

To reduce the effect of variations, we propose to use GAN as a regressor to generate an *invariant* canonical gait image. The generated canonical image contains a subject's gait viewed from the side, wearing a normal (standardized) cloth but without carrying anything. Any gait image appearing from any arbitrary poses is converted to the above canonical view because it contains richer information about the gait dynamics. While this is intuitively appealing, a key challenge that must be addressed is to preserve the human identification information in the generated gait images.

The GaitGANv2 model is trained to generate a canonical gait image (normal clothing and without carrying objects at the side view) using a sufficiently large training data set. In the test phase, a gait image is sent to the GAN model and an invariant gait image that contains human identification information is generated. The difference between the proposed method and most other GAN related methods is that the generated image here can help to improve the discriminant capability, not just generating a gait image that appears to be realistic. The most important challenge here is to preserve human identification when generating a realistically-looking gait image. Compared with the previous work, that is GaitGANv1 in [31], GaitGANv2 adopts a multi-loss strategy to optimize the network to enlarge the inter-class distance whilst reduce the intra-class distance, at the same time.

3.1. Gait energy image

The gait energy image [32] is a popular feature representation for gait. It is produced by averaging all the silhouettes in a single gait cycle, as illustrated in Fig. 1. GEI is well known for its robustness to noise and its efficient computation. The pixel values in a GEI can be interpreted as the probability of pixel positions in GEI being occupied by a human body over one gait cycle. According to the success of GEI in gait recognition, we take GEI as the input and target image of our method. The silhouettes and energy images used in the experiments are produced in the same way as those described in [33].

3.2. Generative adversarial networks for pixel-level domain transfer

Generative adversarial network (GAN) [27] is a branch of unsupervised machine learning, which is implemented by a system of two neural networks competing against each other in a zero-sum game framework. A generative model G that captures the data distribution. A discriminative model D then takes either a real data from the training set or a fake image generated from model G and estimates the probability of its input having come from the training data set rather than the generator. In the GAN for image data, the eventual goal of the generator is to map a small dimensional space z to a pixel-level image space with the objective that the generator can produce a realistic image given an input random vector z. Both G and D could be a non-linear mapping function. In the case where G and D are defined by multilayer perceptrons, the entire system can be trained with back propagation.

The input of the generative model can be an image instead of a noise vector. GAN can realize pixel-level domain transfer between input image and target image such as PixelDTGAN proposed by Yoo et al. [30]. PixelDTGAN can transfer a visual input into different forms which can then be visualized through the generated pixel-level image. In this way, it simulates the creation of mental images from visual scenes and objects that are perceived by the human eyes. In that work, the authors defined two domains, a source domain and a target domain. The two domains are connected by a semantic meaning. For instance, the source domain is an image of a dressed person with variations in pose and the target domain is an image of the person's shirt. So PixelDTGAN can transfer an image from the source domain which is a photo of a dressed person to the pixel-level target image of shirts. Meanwhile the transferred image should look realistic yet preserving the semantic meaning. The framework consists of three important parts as illustrated in Fig. 2. While the real/fake discriminator ensures that the generated images are realistic, the domain discriminator, on the other hand, ensures that the generated images contain semantic information.

The first important component is a pixel-level converter which are composed of an encoder for semantic embedding of a source image and a decoder for producing a target image. The encoder and decoder are implemented by convolution neural networks. However, training the converter is not straightforward because the target is not deterministic. Consequently, on the top of converter, additional loss function is needed to constrain the target image produced. Therefore, Yoo et al. connected a separate network named domain discriminator on top of the converter. The domain discriminator takes a pair of a source image and a target image as input, and is trained to produce a scalar probability of whether the input pair is associated or not. The loss function L_A^D in [30] for the domain discriminator D_A is defined as

$$L_A^D(I_S, I) = -t \cdot log[D_A(I_S, I)] + (t-1) \cdot log[1 - D_A(I_S, I)],$$

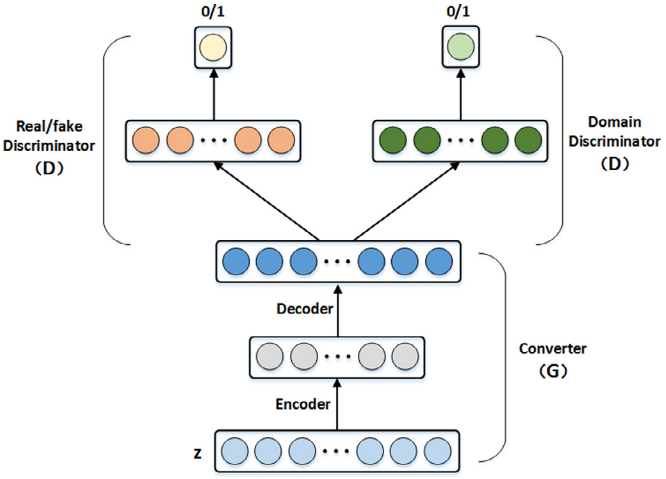


Fig. 2. The framework of PixelDTGAN [30], which consists of three important parts, one converter and two discriminators.

$$s.t. \quad t = \begin{cases} 1 & if \quad I = I_T \\ 0 & if \quad I = \hat{I}_T \\ 0 & if \quad I = I_T^-. \end{cases} \tag{1}$$

where I_S is the source image, I_T is the ground truth target, I_T^- the irrelevant target, and $\hat{I}_T$ is the generated image from converter.

Another component is the real/fake discriminator which similar to the traditional GAN in that it is supervised by the labels of real or fake, in order for the entire network to produce realistic images. Here, the discriminator produces a scalar probability to indicate if the image is real or not. The discriminator's loss function L_R^D, according to [30], takes the form of binary cross entropy:

$$L_R^D(I) = -t \cdot log[D_R(I)] + (t-1) \cdot log[1 - D_R(I)],$$
$$s.t. \quad t = \begin{cases} 1 & if \quad I \in \{I^i\} \\ 0 & if \quad I \in \{\hat{I}^i\}. \end{cases} \tag{2}$$

where $\{I^i\}$ contains real training images and $\{\hat{I}^i\}$ contains fake images produced by the generator.

Labels are given to the two discriminators, and they supervise the converter to produce images that are realistic while keeping the semantic meaning.

3.3. GaitGANv2: GAN for gait recognition

Inspired by the pixel-level domain transfer in PixelDTGAN, GaitGANv2 is proposed to transform the gait data from any view, clothing and carrying conditions to the side view with normal clothing and without carrying objects. Additionally, identification information is preserved.

We set the GEIs at all the viewpoints in normal walking, with clothing and carrying variations as the source and the GEIs of normal walking at 90° (side view) as the target, as shown in Fig. 3. The converter contains an encoder and a decoder as shown in Fig. 4.

There are two discriminators. The first one is a real/fake discriminator which is trained to predict whether an image is real or not. The structure of the real/fake discriminator is the same with that in [30]. If the input GEI is from a real gait image at 90° view in normal walking, the discriminator will output 1. Otherwise, it will output 0. The domain discriminator in [30] has been adopted to a identification discriminator in the proposed method. The difference is that a multiple loss strategy is involved to identification. The loss strategy is described in the following part.

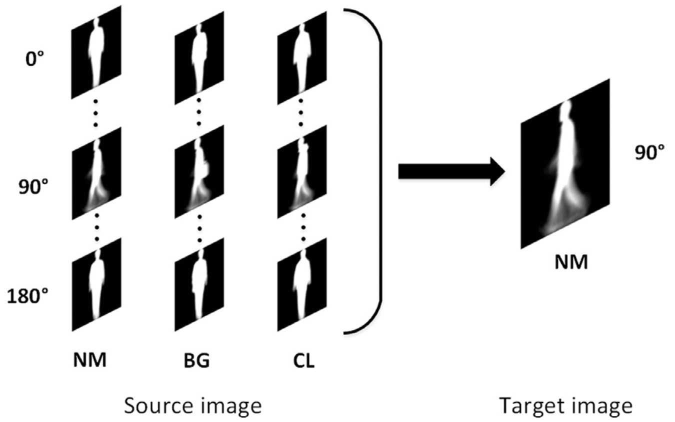


Fig. 3. The source and the target images. 'NM' stands for the normal condition; 'BG' for carrying a bag; and 'CL' for dressing in a coat as defined in CASIA-B dataset.

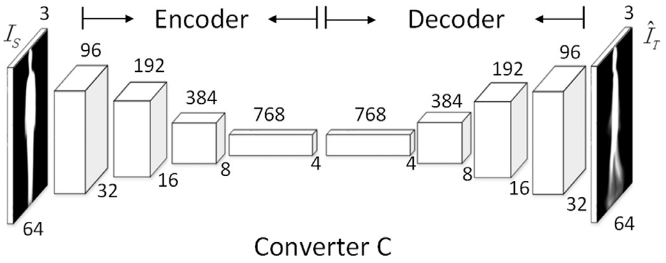


Fig. 4. The structure of the converter which transforms the source images to a target one as shown in Fig. 3.

3.4. Loss function

In order to generate GEIs with large inter-class variations and reduce the intra-class variations, which is preserving identification information, we adopt a multi-loss strategy to optimize the GaitGANv2 network. The softmax loss and the contrastive loss are employed in the proposed method. In our previous work [31], to preserve the identification information, identification discriminator which is similar to the domain discriminator in [30] is involved. The identification discriminator takes a source image and a target image as input, and is trained to produce a scalar probability of whether the input pair is the same person. If the two input images are from the same subject, the output should be 1. If they are input images belonging to two different subjects, the output should be 0. Likewise, if the input is a source image and the target one is generated by the converter, the discriminator function should output 0. This identification discriminator has made a great contribution to preserve the identification information, however, it does not directly constrain the generated GEIs as we need to use the generated GEIs as feature in the gait recognition. For the purpose of better preserving identification information and increasing the accuracy of recognition, we use the multi-loss strategy to constrain the generated GEIs rather than identification discriminator.

3.4.1. Softmax loss

The class labels give strong supervised information to help learning the discriminant features. The softmax loss is the most commonly used in classification tasks in neural networks [34,35]. The softmax loss can promote the generated GEIs with large inter-class variations, it is defined as:

$$Loss_S = -\sum_{i=1}^{m} \log \frac{e^{W_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^{n} e^{W_j^T x_i + b_j}} \tag{3}$$

where $x_i \in \mathbb{R}^d$ is the input of the loss layer, and it is generated from the ith GEI that belongs to the y_ith class. d, $W \in \mathbb{R}^{d \times n}$ and $b \in \mathbb{R}^d$ denote the GEI dimension, last connected layer and bias term, respectively.

3.4.2. Contrastive loss

The softmax loss increases the inter-class distance, but it has not a strong capacity to reduce the variations of identical human gaits. Many researchers employ another loss function to constrain the feature, such as Liao et al. [25] use the CTLoss and Wen et al. [36] adopt the center loss. In order to extract powerful features, we adopt an additional contrastive loss, which is helpful not only to increase the inter-class distance, but also can reduce the intra-class variations. The contrastive loss is defined as:

$$Loss_C = \frac{1}{2} y \|f_i - f_j\|_2^2 + \frac{1}{2}(1-y) max(\lambda - \|f_i - f_j\|_2^2, 0) \qquad (4)$$

where f_i and f_j are generated from two input GEIs, $y = 1$ when the two inputs are from the same subject, then the f_i and f_j will to be close. $y = 0$ means that the two inputs are from different subjects. In this case, the distance of f_i and f_j is limited to be larger than margin λ.

3.4.3. Fusion of loss functions

As in [25,36], the joint contrastive loss and the softmax loss are employed for constrain the quality of the generated GEIs. If only the softmax loss is employed, the learned features could cause a large intra-class variations. So it is necessary to fuse the two loss functions with the two learning objectives. The fusion is given in Eq. (5).

$$\begin{aligned} L &= Loss_S + Loss_C \\ &= -\sum_{i=1}^{m} \log \frac{e^{W_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^{n} e^{W_j^T x_i + b_j}} \\ &\quad + \frac{1}{2} y \|f_i - f_j\|_2^2 + \frac{1}{2}(1-y) max(\lambda - \|f_i - f_j\|_2^2, 0) \end{aligned} \qquad (5)$$

4. Experiments and analysis

4.1. Datasets

To evaluate the proposed method, two datasets are involved. One is CASIA-B with 124 subjects and another is OU-ISIR Large Population Dataset with 4007 subjects.

CASIA-B gait dataset [33] is one of the popular public gait datasets which has been widely used to evaluate different gait recognition methods. It was created by the Institute of Automation, Chinese Academy of Sciences in January 2005. It consists of 124 subjects (31 females and 93 males) captured from 11 views. The view range is from 0° to 180° with 18° interval between two nearest views. There are 11 views for each subject, as shown in Fig. 5. There are 6 sequences for normal walking (“nm”), 2 sequences for walking with a bag (“bg”) and 2 sequences for walking in a coat (“cl”).

OU-ISIR Large Population Dataset [37] gait dataset is a very large dataset which contains 4007 subjects ranging from 1 to 94 years old. The OU-ISIR dataset contains 4 views(55°, 65°, 75°, 85°) and it includes two sequences under the normal walking conditions. It allows us to investigate the upper limit of gait recognition performance in a more statistically reliable way. Fig. 6 shows some samples from OU-ISIR dataset.

4.2. Experimental design

In our experiments using CASIA-B dataset, the three types of gait data including “nm”, “bg” and “cl” are all involved. We put the six normal walking sequences, two sequences with coat and two sequences containing walking with a bag of the first 62 subjects into the training set and the remaining 62 subjects into the test set. In the test set, the first 4 normal walking sequences of each subjects are put into the gallery set and the others into the probe set as it is shown in Table 1. There are four probe sets to evaluate different kind of variations.

Table 1
The experimental design for CASIA-B dataset.

Training set		ID: 001–062, nm01-nm06, bg01, bg02, cl01, cl02
Gallery set		ID: 063–124, nm01-nm04
Probe set	ProbeNM	ID: 063–124, nm05, nm06
	ProbeBG	ID: 063–124, bg01, bg02
	ProbeCL	ID: 063–124, cl01, cl02
	ProbeALL	ID: 063–124, nm05, nm06, bg01, bg02, cl01, cl02

Table 2
Details of the encoder. The first four layers of encoder is the same as real/fake discriminator. After Conv.4, the real/fake and identification discriminator connect Conv.5 to output binary value.

Layers	Number of filters	Filter size	Stride	Batch norm	Activation function
Conv.1	96	4 × 4 × {1,1,2}	2	N	L-ReLU
Conv.2	192	4 × 4 × 96	2	Y	L-ReLU
Conv.3	384	4 × 4 × 192	2	Y	L-ReLU
Conv.4	768	4 × 4 × 384	2	Y	L-ReLU

Table 3
Details of the decoder. F denotes fractional-stride.

Layers	Number of filters	Filter size	Stride	Batch norm	Activation function
F-Conv.1	768	4 × 4 × 384	1/2	Y	L-ReLU
F-Conv.2	384	4 × 4 × 192	1/2	Y	L-ReLU
F-Conv.3	192	4 × 4 × 96	1/2	Y	L-ReLU
F-Conv.4	96	4 × 4 × 1	1/2	N	Tanh

We also evaluate the proposed method on OU-ISIR Large Population dataset and the model is the same as in CASIA-B used in the experiment. We apply five-fold cross-validation on the OU-ISIR and divide all the subjects into five sets randomly. We keep one set for testing and four sets for training in each run. In the training phase, the target image is 85° GEI which is closest the 90° GEI among the four views. The GAN model is trained using CASIA-B data firstly, and then finetuned using the training set of OU-ISIR. In each test set, the first sequence is put into gallery set and the rest sequence is put into probe set.

4.3. Model parameters

In the experiments, we used a similar setup to that of [30], which is shown in Fig. 4. The converter is a unified network that is end-to-end trainable but we can divide it into two parts, an encoder and a decoder. The encoder part is composed of four convolutional layers to abstract the source into another space which should capture the personal attributes of the source as well as possible. Then the result feature z is fed into the decoder in order to construct a relevant target through the four decoding layers. Each decoding layer conducts fractional stride convolutions, where the convolution operates in the opposite direction. The details of the encoder and decoder structures are shown in Table 2 and Table 3. The F-Conv layers in Table 3 have fractional-strides which is 1/2 in our experiments. F-Conv layers can upsample the inputs. The structure of the real/fake discriminator is similar to the encoder's first four convolution layers. The layers of the discriminators are all convolution layers.

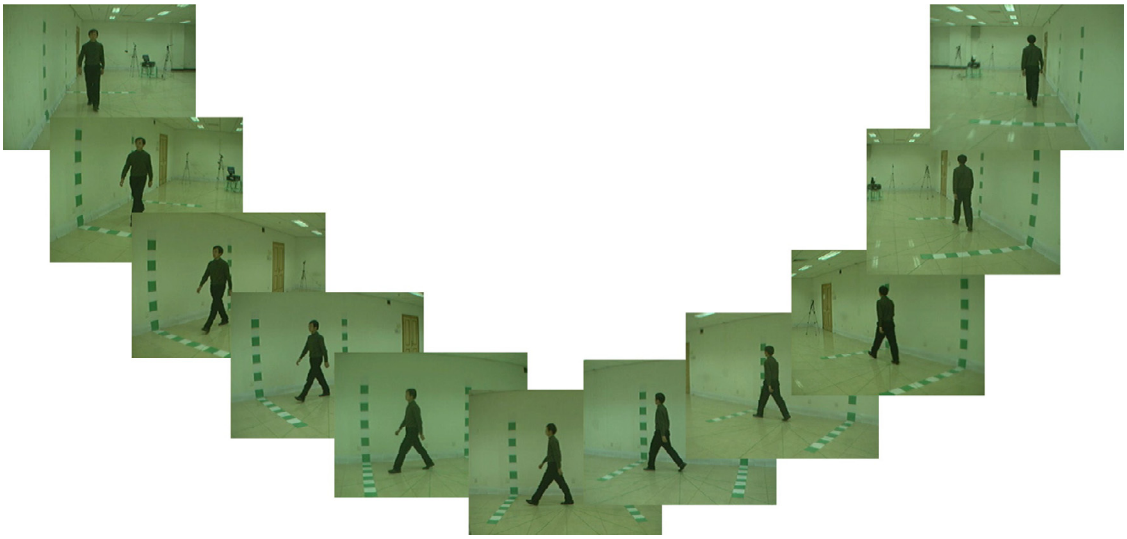


Fig. 5. Normal walking sequences at 11 views from CASIA B dataset.

Table 4
The recognition Rates of ProbeNM, training with sequences containing three conditions.

		Probe angle θ_p (normal walking #5-6)										
		0	18	36	54	72	90	108	126	144	162	180
Gallery angle θ_g (normal #1-4)	0	100.0	80.65	59.68	46.77	29.84	32.26	29.03	35.48	38.71	66.94	79.84
	18	83.87	98.39	95.97	73.39	48.39	47.58	42.74	51.61	66.94	73.39	70.97
	36	60.48	90.32	95.97	90.32	79.03	66.13	62.10	63.71	66.94	60.48	50.81
	54	38.71	69.35	91.13	95.97	92.74	86.29	84.68	81.45	81.45	54.84	33.87
	72	25.81	46.77	67.74	88.71	97.58	95.16	94.35	91.13	70.97	41.13	25.00
	90	28.23	42.74	61.29	84.68	96.77	97.58	95.97	89.52	70.97	36.29	21.77
	108	24.19	41.13	57.26	82.26	95.16	95.97	97.58	95.97	77.42	36.29	23.39
	126	31.45	50.81	64.52	82.26	89.52	91.13	95.97	99.19	97.58	58.87	29.03
	144	37.90	59.68	70.97	73.39	68.55	67.74	82.26	95.97	100.00	77.42	46.77
	162	66.13	75.00	69.35	59.68	39.52	38.71	45.97	63.71	89.52	99.19	79.84
	180	83.87	62.10	49.19	35.48	27.42	26.61	27.42	33.06	55.65	83.06	99.19

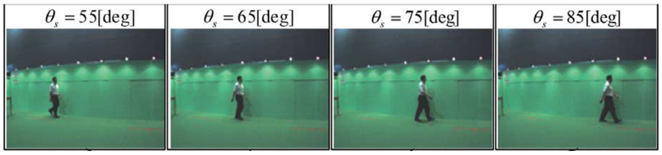


Fig. 6. Sample of 4 views from OU-ISIR dataset.

4.4. Experimental results on CASIA-B dataset

To evaluate the robustness of the proposed GaitGANv2, three kinds of variations have been evaluated, and they are view, clothing, and carrying variations. The experimental results on CASIA-B dataset are shown inTables 4–6. In the experiments of Table 4, the first four normal sequences at a specific view are put into the gallery set, and the last two normal sequences at another view are put into the probe set. Since there are 11 views in the dataset, there are 121 pairs of combinations. In each table, each row corresponds to a view angle of the gallery set, whereas each column corresponds to the view angle of the probe set. The recognition rates of these combinations are listed in Table 4. For the results in Table 5, the main difference with those in Table 4 are the probe sets. The probe data contains images of people carrying bags, and the carrying conditions are different from that of the gallery set. The probe sets for Table 6 contain gait data with coats.

4.5. Comparisons with GEI+PCA, SPAE and GaitGANv1

Since GEIs are used as input to extract invariant features, we first compare the proposed GaitGANv2 with GEI+PCA [32], SPAE [22] and GaitGANv1 [31]. The experiment protocols in terms of the gallery and probe sets for GEI+PCA and SPAE are exactly the same as those presented in Table 1. Due to limited space, we only list 4 probe angles with a 36° interval. Each row in this figure represents a probe angle. The compared angles are 36°, 72°, 108° and 144°. The first column of Fig. 7 compares the recognition rates of the proposed with GEI+PCA, SPAE and GaitGANv1 at different probe angles in normal walking sequences. The second column shows the comparison with different carrying conditions, and the third shows the comparison with different clothing. As illustrated in Fig. 7, the proposed GaitGANv2 outperforms GEI+PCA at all probe angle and gallery angle pairs. Meantime, its performance sometimes is similar to that of SPAE and better than SPAE at most of the time. The results show that the proposed method can extract the gait feature and achieves state-of-the-art performance.

We also compared the recognition rates without view variation. This can be done by taking the average of the rates on the diagonal of Tables 4, 5 and 6. The corresponding average rates of GEI+PCA and SPAE are also obtained in the same manner. The results are shown in Fig. 8. When there is no clothing variation, the proposed method achieve a high recognition rate which is better than GEI+PCA and SPAE.

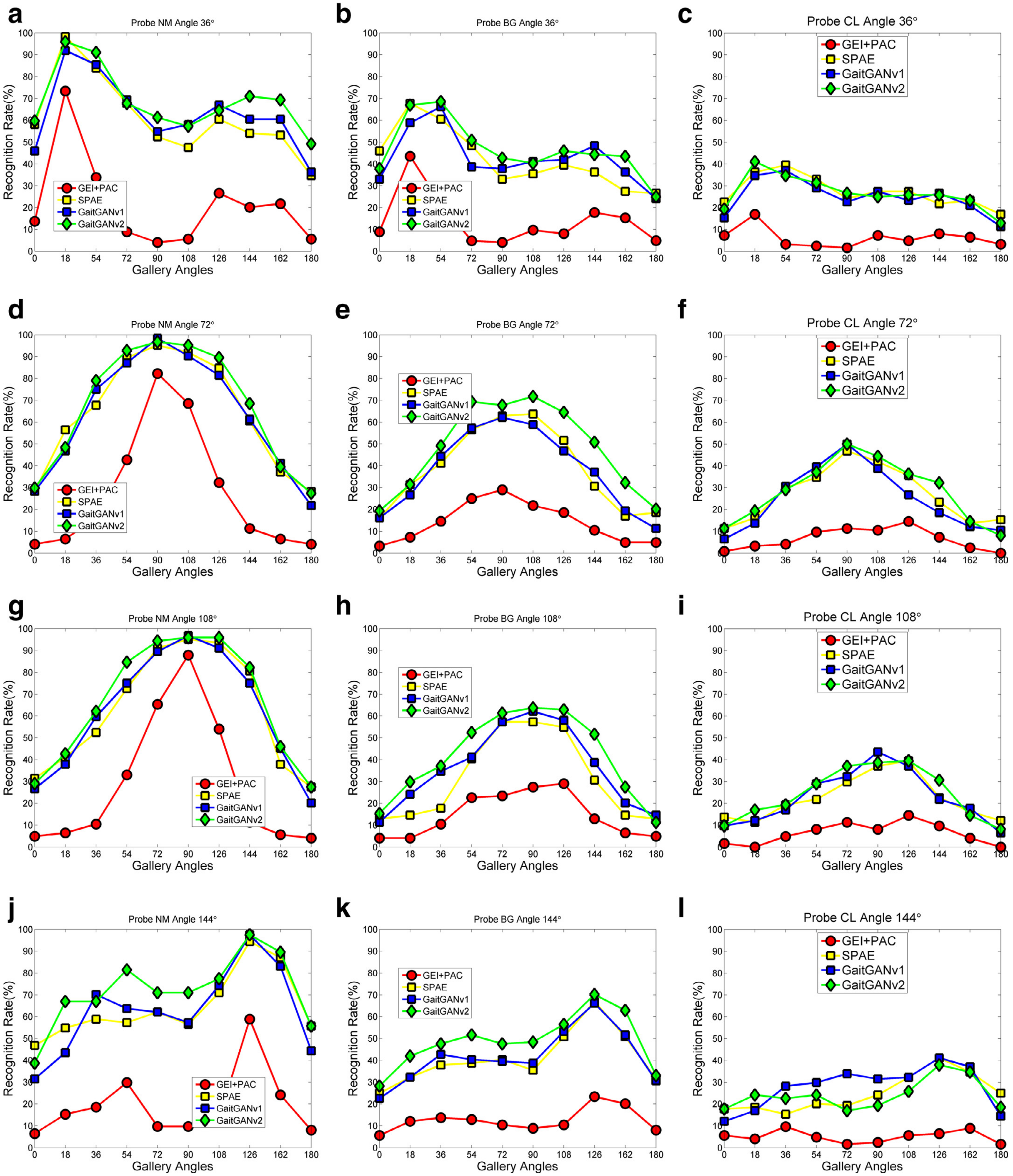


Fig. 7. Comparisons with GEI+PCA [32], SPAE [22] and GaitGANv1 [31] at different probe angle. Each row represents a probe angle and each column represents different conditions probe sequences. The blue lines are achieved by proposed method. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 5
The recognition rates of ProbeBG, training with sequences containing three conditions.

		Probe angle θ_p (walking with bag #1-2)										
		0	18	36	54	72	90	108	126	144	162	180
Gallery angle θ_g (normal #1-4)	0	84.68	57.26	37.90	21.77	19.35	13.71	15.32	19.35	28.23	47.58	59.68
	18	65.32	82.26	66.94	42.74	31.45	30.65	29.84	38.71	41.94	58.87	45.97
	36	38.71	67.74	74.19	66.94	49.19	41.94	37.10	52.42	47.58	45.97	33.87
	54	33.06	54.03	68.55	79.64	69.35	52.42	52.42	59.68	51.61	35.48	22.58
	72	20.97	33.06	50.81	60.48	72.58	63.71	61.29	63.71	47.58	28.23	13.71
	90	22.58	29.03	42.74	52.42	67.74	70.97	63.71	62.90	48.39	29.03	12.90
	108	22.58	25.00	40.32	56.45	71.77	70.16	65.32	66.94	56.45	29.03	15.32
	126	22.58	35.48	45.97	55.65	64.52	57.26	62.90	78.23	70.16	39.52	21.77
	144	34.68	41.13	44.35	47.58	50.61	44.35	50.61	67.74	74.19	51.61	35.48
	162	49.19	52.23	43.55	37.90	32.25	25.81	27.42	39.52	62.90	79.03	54.84
	180	62.10	38.71	25.00	17.74	20.16	15.32	11.29	13.71	33.06	56.45	77.42

Table 6
The recognition Rates of ProbeCL, training with sequences containing three conditions.

		Probe angle θ_p (walking with coat #1-2)										
		0	18	36	54	72	90	108	126	144	162	180
Gallery angle θ_g (normal #1-4)	0	33.06	22.58	19.35	15.32	11.29	9.68	9.68	15.32	17.74	20.16	26.61
	18	27.42	45.97	41.13	25.81	19.35	17.74	16.94	22.58	24.19	27.42	24.19
	36	25.00	37.90	44.35	44.35	29.03	24.19	19.35	27.42	22.58	22.58	15.32
	54	18.55	25.00	34.68	44.35	37.10	29.84	29.03	29.84	24.19	16.94	9.68
	72	19.35	25.81	31.45	45.97	55.65	39.52	37.10	32.26	16.94	17.74	8.87
	90	14.52	23.39	26.61	40.32	50.00	43.55	38.71	30.65	19.35	16.13	8.87
	108	12.10	16.13	25.00	39.52	44.35	36.29	45.97	38.71	25.81	18.55	8.06
	126	17.74	19.35	25.81	32.26	36.29	33.06	39.52	44.35	37.90	23.39	12.90
	144	16.94	22.58	25.81	30.65	32.26	25.81	30.65	40.32	40.32	31.45	18.55
	162	29.84	23.39	23.39	20.97	14.52	8.06	14.52	23.39	34.68	41.13	26.61
	180	25.81	15.32	12.90	12.90	8.06	5.65	8.06	13.71	18.55	24.19	33.06

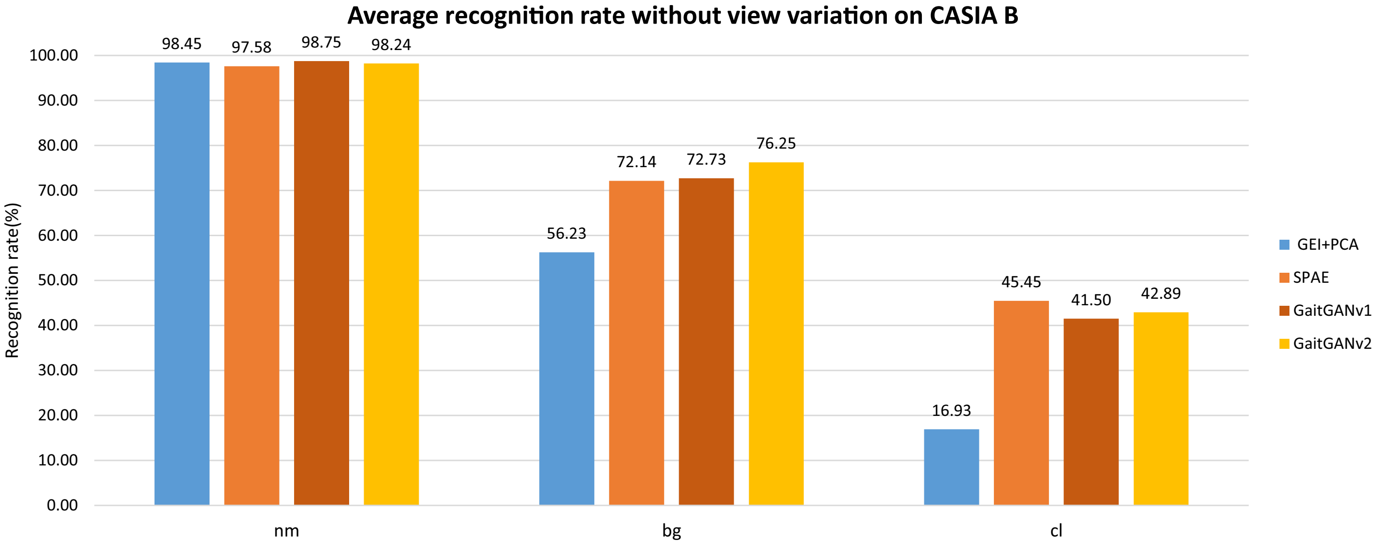


Fig. 8. The average recognition rates without view variation with GEI+PCA, SPAE and GaitGANv1 at three conditions.

4.6. Comparisons with state-of-the-art

In order to better analyse the performance of the proposed method, we further compare the proposed GaitGANv2 with additional state-of-the-art methods including FD-VTM [6], RSVD-VTM [7], RPCA-VTM [8], R-VTM [9], GP+CCA [10] , C3A [12], SPAE [22] and GaitGANv1 [31]. The probe angles selected are 54°, 90° and 126° as in experiments of those methods. The experimental results are listed in Fig. 9. From the results we can find that the proposed method outperforms others when the angle difference between the gallery and the probe is large. This shows that the model can handle large viewpoint variation well. When the viewpoint variation is not large enough, the proposed method can also improve the recognition rate obviously.

In Table 7, the experimental results of C3A [12], ViDP [16], CNN [19], SPAE [22], GaitGAN [31] and the proposed method are listed. Here we want to emphasis that the proposed method obtains similar results using only one generative model for any views, and for clothing or carrying condition variations, simultaneously. Meanwhile, this method is the first use of GAN for gait recognition and the experimental results show that GAN is feasible for gait recognition under significant variations.

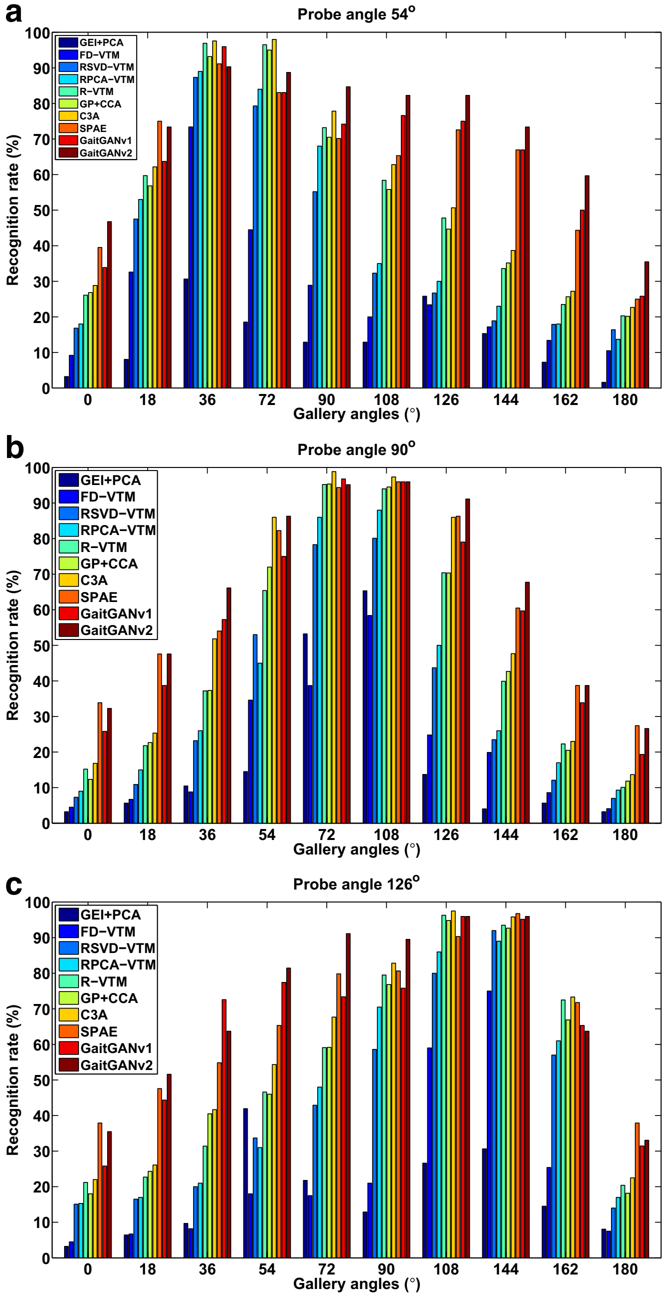


Fig. 9. Comparisons with existing methods at probe angles (a)54°, (b)90° and (c)126°. The gallery angles are the rest 10 angles except the corresponding probe angle.

Table 7
Average recognition rates at probe angles 54°, 90° and 126°. The gallery angles are the rest 10 angles except the corresponding probe angle. The values in the right most column are the averages rate at the three probe angles 54°, 90° and 126°.

Method	Probe angle			
	54°	90°	126°	Average
C3A [12]	56.64%	54.65%	58.38%	56.56%
ViDP [16]	64.2%	60.4%	65.0%	63.2%
CNN [19][a]	77.8%	64.9%	76.1%	72.9%
SPAE [22]	63.31%	62.1%	66.29%	63.9%
GaitGANv1 [31]	64.52%	58.15%	65.73%	62.8%
GaitGANv2	71.69%	64.76%	70.16%	68.87%

[a] Models are trained with GEIs of the first 24 subjects.

Table 8
Experimental results on OU-ISIR dataset.

Probe angle	Gallery angle				Average	
	55°	65°	75°	85°	CNN [19]	GaitGANv2
55°	–	94.4%	93.2%	88.2%	91.6%	**91.9**%
65°	94.9%	–	96.2%	94.0%	92.3%	**95.0**%
75°	91.7%	95.5%	–	95.9%	92.4%	**94.4**%
85°	91.7%	95.5%	96.5%	–	**94.8**%	94.6%

Table 9
The recognition rate for 4 identical views on OU-ISIR dataset.

View	55°	65°	75°	85°
NN [37]	84.7%	86.6%	86.9%	85.7%
CNN [19]	98.8%	98.9%	98.9%	98.9%
GaitGANv2	96.3%	96.7%	96.5%	95.9%

4.7. Experimental results on OU-ISIR dataset

OU-ISIR dataset is also used to evaluate the proposed method. In the experiments, the five-fold cross-validation is involved. The recognition rates of experiments on OU-ISIR dataset is shown in Tables 8 and 9. There is view variation in the experiments of Table 8 and the proposed method outperforms CNN [19]. The results in Table 9 are for identical views. Our results are also much better than the baseline reported by the dataset authors [37]. However, the results reported by CNN [19] are better. From the results it can be shown that the proposed method has an obvious advantage on gait recognition with view variation. That means the GAN model can generated better feature which is robust to variations.

5. Conclusions and future work

In this paper, we applied GaitGANv2 which is a variant of generative adversarial networks, PixelDTGAN, adopted to deal with variations in viewpoint, clothing and carrying conditions simultaneously in gait recognition. Extensive experiments on two large datasets show that the GaitGANv2 can transform gait images obtained from any viewpoint to the side view and remove the variations of clothings and carrying without the need to estimate the subject's view angle, clothing type and carrying condition beforehand. Experimental results show that the recognition rate of proposed model is comparable to that of the state-of-the-art methods. Indeed, GaitGANv2 is shown to be promising for practical applications in video surveillance.

There are however, a number of limitations which need to be addressed in future work. The proposed method shows that GAN can be used to eliminate the variations and keep identification information of subjects. If we use more complex and powerful networks, better performance should be achieved. We believe that better GAN technologies will further improve gait recognition in future. Besides gait recognition, different recognition and classification problems under pose and other variations could also benefit from that.

Acknowledgment

The authors would like to thank Dr. Xianglei Xing for his support on experimental results of some methods. The work is supported by the Science Foundation of Shenzhen (Grant No. JCYJ20150324141711699 and 20170504160426188).

References

[1] A.Y. Johnson, A.F. Bobick, A multi-view method for gait recognition using static body parameters, in: Proceedings of the 3rd International Conference on Audio and Video Based Biometric Person Authentication, 2001, pp. 301–311.

[2] A. Kale, A.K.R. Chowdhury, R. Chellappa, Towards a view invariant gait recognition algorithm, in: Proceedings of the IEEE Conference on Advanced Video and Signal Based Surveillance, 2003, pp. 143–150.

[3] G. Zhao, G. Liu, H. Li, M. Pietikainen, 3d gait recognition using multiple cameras, in: Proceedings of the International Conference on Automatic Face and Gesture Recognition, 2006, pp. 529–534.

[4] G. Ariyanto, M.S. Nixon, Model-based 3d gait biometrics, in: Proceedings of the International Joint Conference on Biometrics, 2011, pp. 1–7.

[5] J. Tang, J. Luo, T. Tjahjadi, F. Guo, Robust arbitrary-view gait recognition based on 3d partial similarity matching, IEEE Trans. Image Process. 26 (1) (2017) 7–22.

[6] Y. Makihara, R. Sagawa, Y. Mukaigawa, T. Echigo, Y. Yagi, Gait recognition using a view transformation model in the frequency domain, in: Proceedings of the ECCV, 2006, pp. 151–163.

[7] W. Kusakunniran, Q. Wu, H. Li, J. Zhang, Multiple views gait recognition using view transformation model based on optimized gait energy image, in: Proceedings of the ICCV Workshops, 2009, pp. 1058–1064.

[8] S. Zheng, J. Zhang, K. Huang, R. He, T. Tan, Robust view transformation model for gait recognition, in: Proceedings of the ICIP, 2011, pp. 2073–2076.

[9] W. Kusakunniran, Q. Wu, J. Zhang, H. Li, Gait recognition under various viewing angles based on correlated motion regression, IEEE TCSVT 22 (6) (2012) 966–980.

[10] K. Bashir, T. Xiang, S. Gong, Cross-view gait recognition using correlation strength, in: Proceedings of the BMVC, 2010.

[11] C. Luo, W. Xu, C. Zhu, Robust gait recognition based on partitioning and canonical correlation analysis, in: Proceedings of the IEEE International Conference on Imaging Systems and Techniques, 2015.

[12] X. Xing, K. Wang, T. Yan, Z. Lv, Complete canonical correlation analysis with application to multi-view gait recognition, Pattern Recognit. 50 (2016) 107–117.

[13] J. Lu, G. Wang, P. Moulin, Human identity and gender recognition from gait sequences with arbitrary walking directions, IEEE TIFS 9 (1) (2014) 51–61.

[14] X. Ben, W. Meng, R. Yan, K. Wang, An improved biometrics technique based on metric learning approach, Neurocomputing 97 (2012) 44–51.

[15] X. Ben, P. Zhang, W. Meng, R. Yan, M. Yang, W. Liu, H. Zhang, On the distance metric learning between cross-domain gaits, Neurocomputing 208 (2016) 153–164.

[16] M. Hu, Y. Wang, Z. Zhang, J.J. Little, D. Huang, View-invariant discriminative projection for multi-view gait-based human identification, IEEE TIFS 8 (12) (2013) 2034–2045.

[17] H. Hu, Enhanced Gabor feature based classification using a regularized locally tensor discriminant model for multiview gait recognition, IEEE Trans. Circuits Syst. Video Technol. 23 (7) (2013) 1274–1286.

[18] X. Ben, P. Zhang, R. Yan, M. Yang, G. Ge, Gait recognition and micro-expression recognition based on maximum margin projection with tensor representation, Neural Comput. Appl. 27 (8) (2016) 2629–2646.

[19] Z. Wu, Y. Huang, L. Wang, X. Wang, T. Tan, A comprehensive study on cross-view gait based human identification with deep CNNS, IEEE TPAMI 39 (2) (2017) 209–226.

[20] M.A. Hossain, Y. Makihara, J. Wang, Y. Yagi, Clothing-invariant gait identification using part-based clothing categorization and adaptive weight control, Pattern Recognit. 43 (6) (2010) 2281–2291.

[21] Y. Guan, C.T. Li, Y. Hu, Robust clothing-invariant gait recognition, in: Proceedings of the Eighth International Conference on Intelligent Information Hiding and Multimedia Signal Processing, 2012, pp. 321–324.

[22] S. Yu, H. Chen, Q. Wang, L. Shen, Y. Huang, Invariant feature extraction for gait recognition using only one uniform model, Neurocomputing 239 (2017) 81–93.

[23] Y. Liang, C.T. Li, Y. Guan, Y. Hu, Gait recognition based on the golden ratio, Eurasip J. Image Video Process. 2016 (1) (2016) 22.

[24] Y. Feng, Y. Li, J. Luo, Learning effective gait features using lstm, in: International Conference on Pattern Recognition, 2017, pp. 325–330.

[25] R. Liao, C. Cao, E.B. Garcia, S. Yu, Y. Huang, Pose-based temporal-spatial network (PTSN) for gait recognition with carrying and clothing variations, in: Proceedings of the 12th Chinese Conference on Biometric Recognition, Springer, 2017, pp. 474–483.

[26] W. An, R. Liao, S. Yu, Y. Huang, P.C. Yuen, Improving gait recognition with 3d pose estimation, in: Proceedings of the 13th Chinese Conference on Biometric Recognition (CCBR), 2018, pp. 137–147.

[27] I.J. Goodfellow, J. Pougetabadie, M. Mirza, B. Xu, D. Wardefarley, S. Ozair, A. Courville, Y. Bengio, Z. Ghahramani, M. Welling, Generative adversarial nets, Adv. Neural Inf. Process. Syst. 3 (2014) 2672–2680.

[28] M. Mirza, S. Osindero, Conditional generative adversarial nets, Comput. Sci. (2014) 2672–2680.

[29] A. Radford, L. Metz, S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks, Comput. Sci. (2015).

[30] D. Yoo, N. Kim, S. Park, A.S. Paek, I.S. Kweon, Pixel-level domain transfer, in: Proceedings of ECCV, 2016.

[31] S. Yu, H. Chen, E.B.G. Reyes, P. Norman, GaitGAN: invariant gait feature extraction using generative adversarial networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 30–37.

[32] J. Han, B. Bhanu, Individual recognition using gait energy image, IEEE TPAMI 28 (2) (2006) 316–322.

[33] S. Yu, D. Tan, T. Tan, A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition, in: Proceedings of the ICPR, 2006, pp. 441–444.

[34] K. Zhang, Y. Huang, Y. Du, L. Wang, Facial expression recognition based on deep evolutional spatial-temporal networks, IEEE Trans. Image Process. 26 (9) (2017) 4193–4203.

[35] Y. Sun, X. Wang, X. Tang, Deep learning face representation from predicting 10,000 classes, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1891–1898.

[36] Y. Wen, K. Zhang, Z. Li, Y. Qiao, A discriminative feature learning approach for deep face recognition, in: Proceedings of the European Conference on Computer Vision, Springer, 2016, pp. 499–515.

[37] H. Iwama, M. Okumura, Y. Makihara, Y. Yagi, The OU-ISIR gait database comprising the large population dataset and performance evaluation of gait recognition, IEEE Trans. Inf. Forensics Secur. 7, Issue 5 (2012) 1511–1521.

Shiqi Yu received his B.E. degree in computer science and engineering from the Chu Kochen Honors College, Zhejiang University in 2002, and Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences in 2007. He worked as an assistant professor and then as an associate professor in the Shenzhen Institutes of Advanced Technology, Chinese Academy of Science from 2007 to 2010. Currently, he is an associate professor in the College of Computer Science and Software Engineering, Shenzhen University, China. He especially focuses on image classification and related research topics.

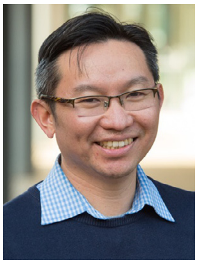
Rijun Liao received his B.S. degree from the College of Physics and Energy, Shenzhen University, China in 2015. He is currently a master student in the College of Computer Science and Software Engineering, Shenzhen University, China. His research interests include biometrics, computer vision and deep learning.

Weizhi An received her B.S. degree from the College of Computer Science and Software Engineering, Shenzhen University, China in 2016. She is currently a master student in the College of Computer Science and Software Engineering, Shenzhen University, China. Her research interests include biometrics, computer vision and deep learning.

Haifeng Chen received his B.S. degree in computer science and engineering from Qufu Normal University, China, in 2013, and his master degree from the College of Computer Science and Software Engineering, Shenzhen University, China, in 2017. His research interests include computer vision and deep learning.

Edel B. Garcia Reyes is graduated of Mathematic and Cybernetic from University of Havana, in 1986 and received the Ph.D. in Technical Sciences at the Technical Military Institute "Jose Marti" of Havana, in 1997. At the moment, he is working as a researcher in the Advanced Technologies Application Center. Dr. Edel has focused his researches on digital image processing of remote sensing data, biometrics and video surveillance. He has participated as member of technical committees and experts groups and has been reviewer for different events and journals as Pattern Recognition Letter, Journal of Real-Time Image Processing, etc. Dr. Edel worked in the Cuban Institute of Geodesy and Cartography (1986-1995) and in the Enterprise Group GeoCuba (1995-2001) where he was the head of the Agency of the Centre of Data and Computer Science of Geocuba - investigation and Consultancy (1998-2001).

Yongzhen Huang received the B.E. degree from the Huazhong University of Science and Technology in 2006 and the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences (CASIA) in 2011. In July 2011, he joined the National Laboratory of Pattern Recognition (NLPR), CASIA, where he is currently an associate professor. He has published more than 50 papers in the areas of computer vision and pattern recognition at international journals such as IEEE Transactions on Pattern Analysis and Machine Intelligence, International Journal of Computer Vision, IEEE Transactions on Systems, Man, and Cybernetics, IEEE Transactions on Circuits and Systems for Video Technology, IEEE Transactions on Multimedia, and conferences such as CVPR, ICCV, NIPS, and BMVC. His current research interests include pattern recognition, computer vision, and machine learning.



Norman Poh currently serves as CSO for Truststamp Europe and Data Scientist for BJSS London. He holds a PhD in Machine Learning and Information Fusion from IDIAP research institute, École Polytechnique Fédérale de Lausanne (EPFL), Switzerland. He is passionate about machine learning with applications to biometric person recognition, healthcare, forensics, financial forecasting, and other practical data intensive areas, where he published more than 100 peer-reviewed publications, including 5 best paper awards. Previously, he was a Senior Lecturer at University of Surrey where he conducted research as principal investigator of two personal fellowship/grant schemes, i.e., Swiss NSF Advanced Researcher Award and Medical Research Council's New Investigator Research Grant. He was named Researcher of the Year, University of Surrey in 2011.