Neurocomputing 339 (2019) 245–254

Contents lists available at ScienceDirect

Neurocomputing

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

# Learning view invariant gait features with Two-Stream GAN

Yanyun Wang[a,b], Chunfeng Song[b,c], Yan Huang[b,c,d], Zhenyu Wang[a,*], Liang Wang[b,c,d]

[a] School of Control and Computer Engineering, North China Electric Power University, Beijing 102206, China
[b] National Laboratory of Pattern Recognition (NLPR), Center for Research on Intelligent Perception and Computing (CRIPAC), Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing 100190, China
[c] University of Chinese Academy of Sciences (UCAS), Beijing 100190, China
[d] Center for Excellence in Brain Science and Intelligence Technology (CEBSIT), Beijing 100190, China

ARTICLE INFO

*Article history:*
Received 20 September 2018
Revised 23 December 2018
Accepted 11 February 2019
Available online 16 February 2019

Communicated by dr Yongmin Li

*Keywords:*
Gait recognition
Cross-veiw
Two-Stream GAN

ABSTRACT

Gait recognition is an important yet challenging problem in computer vision. The changing view of gait is one of the most challenging factors, which could greatly affect the accuracy of cross-view gait recognition. In this paper, we propose a Two-Stream Generative Adversarial Network (TS-GAN) for cross-view gait recognition. For any view of gait representations, GAN can restore it to the corresponding standard view, to learn view invariant gait features. To achieve this goal, TS-GAN has two streams : (1) the global-stream can learn global contexts, and (2) the part-stream can learn local details. We combine the two streams to learn final identities. Moreover, we add a pixel-wise loss along with the generators of GAN to restore the gait details in pixel-level. We evaluate the proposed method on two widely used gait databases: CASIA-B and OU-ISIR. Experiment results show that our approach outperforms the compared state-of-the-art approaches.

© 2019 Elsevier B.V. All rights reserved.

## 1. Introduction

Gait refers to the posture of people walking, which consists of the relative movement of body parts. Gait recognition is to identify people by their unique gait. Early medical researches have proved that human gait has 24 consistent components [1,2], thus gait is relatively stable. In addition, gait is a biometric feature that can be extracted from a long distance. Whereas other biometric features, such as human faces, fingerprints and irises, usually need to be captured from a near distance, which could be limited to some practical applications. Gait recognition has a wide range of applications such as intelligent monitoring, access control, medical diagnosis and human-computer interactions.

In the past few years, many datasets have been established by some organization, such as CASIA-B [3], OU-ISIR [4], OUMVLP [5] and so on. At the same time, many approaches [6–10] have been proposed to address the problem of gait recognition. Among various factors that greatly affect the accuracy of gait recognition in practical applications, the change of view is the most difficult one. When the view changes, the profiles of a person captured by the camera can vary dramatically. The view transform models [10] and the view invariant features [7,8] have been proposed to solve the cross-view gait recognition in feature level. Recently, Yu et al. [9] tried to use the Generative Adversarial Networks (GAN) as the view transform model to learn view invariant gait features in input level. Their method could learn global features well, but they do not model the local area that lead to the absence of detail information. Local features have undergone major changes when the view changes, making the gait features fuzzy.

In this paper, we propose a Two-Stream Generative Adversarial Network (TS-GAN) to learn both part-aware and global-aware view invariant gait features in a pixel-level manner within a unified model, as shown in Fig. 1. Firstly, TS-GAN can transfer gait images with different views to the gait images with the standard view, i.e., 90°. Since the views of gait vary from 0°to 180°, we choose 90° as the standard view, and it is easier to restore the image from other views to 90°. Moreover, 90° is one of the best views to show discriminative information in gait. Secondly, considering that typical GAN used in [9] can not restore the details in pixel-level, to address this problem, we adopt a pixel-wise loss on the generator, aiming at learning accurate view transformation in pixel-level. Taking into account that the absence of local modeling may lead to ambiguous gait representation, we add the part-stream in our TS-GAN. In addition, we design an identity discriminator following the TS-GAN to learn final view invariant gait features.

* Corresponding author.
*E-mail addresses:* yanyunwang@ncepu.edu.cn (Y. Wang), chunfeng.song@nlpr.ia.ac.cn (C. Song), yhuang@nlpr.ia.ac.cn (Y. Huang), zywang@ncepu.edu.cn (Z. Wang), wangliang@nlpr.ia.ac.cn (L. Wang).

https://doi.org/10.1016/j.neucom.2019.02.025
0925-2312/© 2019 Elsevier B.V. All rights reserved.

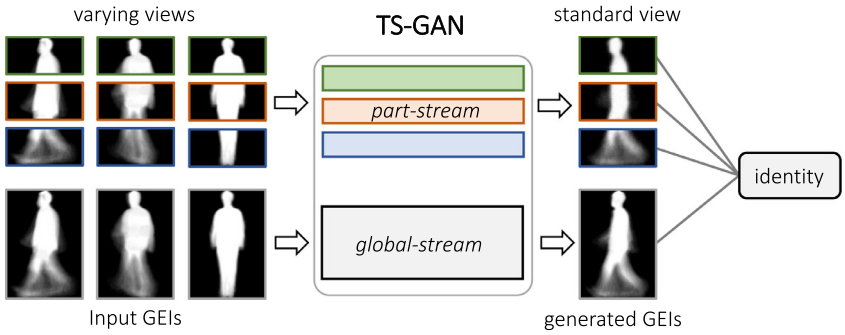



**Fig. 1.** The pipeline of the proposed Two-Stream GAN. TS-GAN can transfer the input gait image with any view to an image that has a standard view. The two streams can learn view invariant gait features from the global and part streams, separately. The global-stream can learn global features and the part-stream can learn detail information. Therefore, fusion of the two streams can learn better gait features.

We summarize this paper's contributions as follows.

- We present a Two-Stream GAN (TS-GAN) to learn view invariant gait features via combining the features from both the global and part streams. It can simultaneously transfer the gait images from the global and part streams, maintaining both of the identical and local features.
- We add a pixel-wise loss along with TS-GAN to restore the gait details in pixel-level. The loss is very helpful for accurate gait image transformation and making the generator more stable in the training phase.
- We introduce the multi-scale context-aware unit in the generator, to learn a closer image distribution. Experimental results have shown the effectiveness.

The rest of this paper is organized as follows. The next section presents related work of cross-view gait recognition. Section 3 details the proposed method. Section 4 describes experiment results, and Section 5 concludes this paper.

## 2. Related work

In this section, we first review recent research progress of gait recognition methods, which include traditional methods and deep learning methods. Then summarize the GAN based approaches and global and local aware feature learning approaches related with this work.

### 2.1. Traditional gait recognition methods

There are numbers of methods proposed for gait recognition. According to different methods of feature extraction, the methods of gait recognition can be roughly divided into two categories: model-based methods and non-model methods.

The model-based approaches usually model the movement of a person or the structure of a human body to extract the features. There are some typical model-based methods [11–14]. Those approaches have great advantages for solving the occlusion problem. Due to the fact that the model-based methods are completely based on the motion which could help to correlate the state variations from the past movements to current one, it can contribute for resolving the occlusion problem.

The non-model approaches establish the relationships between adjacent frames by estimating the relative features of position, velocity, shape, color, etc., many of which are based on human contours. The typical non-model methods [15–20] have been proposed to address gait recognition task. Recently, several methods based VTM had been proposed [10,21,22] to implement view transformation. Xing et al. [23] proposed the canonical correlation analysis (C3A) algorithm to overcome the computational complexity of multi-view gait recognition. Hu et al. [24] proposed a view-invariant discriminative projection (ViDP) method to improve the discriminative ability of multi-view gait features. The advantage of this method is that the gait features can directly be matched without knowing the view. Muramatsu et al. [25] calculated multiple scores that measured the consistency of the multiple transformed features and original features whether the target subjects are same.

### 2.2. Deep learning based gait recognition methods

However, the above methods have limitations in accuracy, especially in large scale datasets. With the breakthrough of deep learning in many areas, more and more researchers begin to introduce this technique into gait recognition, and some advanced methods are proposed, such as [26–28]. One of the pioneer methods proposed by Wu et al. [8] tried to learn the similarity of gait pairs with the deep Convolutional Neural Networks (CNNs). Yu et al. [29] proposed an auto-encoder model for invariant gait extraction. The feature extracted by the method is robust to view, clothing and carrying condition variation. Whereas Shiraga et al. [7] proposed to directly learn gait features from the Gait Energy Image (GEI) with a CNN based classifier. Although this method has high accuracy in OU-ISIR database, its accuracy is slightly lower in other datasets.

### 2.3. GAN based methods

Since Goodfellow et al. [30] proposed the concept of Generative Adversarial Network, GAN has become a hot topic. Some improved variants of GAN are proposed, such as [31]. Inspired by the success of GAN, there are some GAN based applications. Huang et al. [32] proposed a two-pathway GAN to learn both the global and local textures for face recognition. Yu et al. [9] adopted a GAN based model to learn view invariant features for gait recognition. The contribution of this method is to introduce GAN into the field of gait recognition.

### 2.4. Global and local feature learning

Global and local feature learning have been implemented in many areas, such as person re-identification, face recognition and soft-biometrics [33–35]. Samaria and Fallside [33] proposed to use hidden markov models to solve face identification problem. In order to improve the attribute recognition for small-size training data with poor quality images, Wang et al. [34] proposed a Joint Recurrent Learning (JRL) model for exploring attribute context and correlation. Zhao et al. [35] proposed an end-to-end Grouping Recurrent Learning (GRL) model to detect precise body region via body region proposal followed by the feature extraction unit from detected regions.

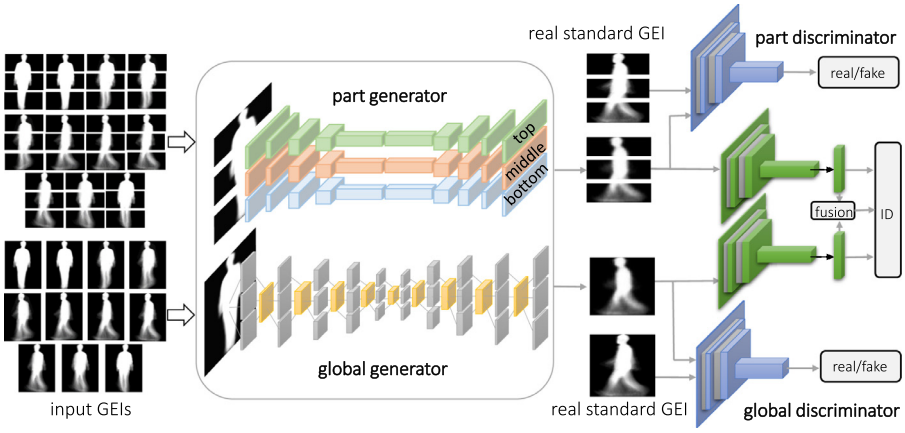



**Fig. 2.** The framework of TS-GAN. It takes the GEIs with any view as its inputs, and transforms them into GEIs with a standard view, e.g., 90°. The TS-GAN has two streams, i.e., the global stream and the part stream. It should be noted that the generator of global stream has multiple MSCAN stages, details refer to the text. The part stream is composed with three parts, i.e., the upper part (containing head and shoulders), the middle part (containing the chest and stomach) and the lower part (containing the legs and feet). Each stream has its own discriminator and identity classifier to jointly learn the view-invariant and identical features. In addition, we combine the two streams so as to learn better gait features.

## 3. Proposed method

The proposed Two-Stream GAN (TS-GAN) takes the GEIs with any view as its inputs, and transforms them into GEIs with standard view. As shown in Fig. 2, the TS-GAN has two streams, i.e., the global stream and the part stream. Each stream has its own discriminator and identity classifier to jointly learn the view-invariant and identical features. In addition, we fuse the features at the top layers of the two streams for obtaining better performance. We will describe the details of each component as follows.

### 3.1. Generative adversarial networks

In general, the basic idea of GAN [30] can be regarded as a two-player game. GAN can learn the mapping from the input variable to the output. If the input variable follows a normal distribution, then a generating network $G(z; \theta_g)$ can be obtained. A discriminator $D(x; \theta_d)$ is connected to the backend of the generator. It randomly selectes the real samples and the generated data as its input. The discriminator is a two-way classification network, which can distinguish whether the input sample is real data or generated data by the generator. In the adversarial phase, the generator tries to generate samples more like the real samples, so that the discriminator can not distinguish if it is not real. Then the discriminator is enhanced to distinguish the fake data. In practice, the generator and discriminator are usually implemented with Convolutional Neural Networks (CNNs).

### 3.2. Two-Stream GAN

We propose a Two-Stream GAN (TS-GAN) to jointly learn view-invariant features from both global and part streams. In our model, we take the Gait Energy Image (GEI) [36] as the default input. In general, GEI can be computed via averaging all the images in the gait sequence, which contains both the shape and motion information. Examples are shown in Fig. 3. GEI has been widely used because of its simplicity, effectiveness and robustness over the other gait templets, e.g., the Gait Entropy Image (GEnI) [37], Gait Flow Image (GFI) [38], and Chrono Gait Image (CGI) [39]. The main purpose of the proposed method is to convert GEI images of different views into standard view GEI images and then identify them.

Different from the previous GAN based method [9], a two-stream model is adopted to recover both the global and local details. In addition, we introduce the pixel-wise loss at the end of the generators for pixel-level constraints. The whole framework is shown in Fig. 2.

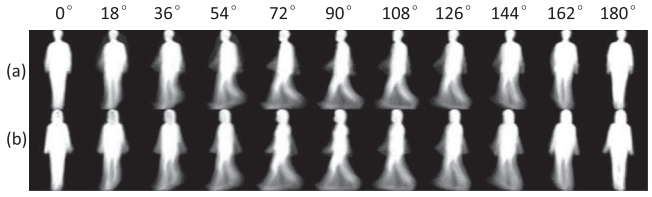



**Fig. 3.** Example GEIs of two subjects. These GEIs come from the CASIA-B dataset, with 11 views.

### 3.3. Global stream

The global stream is designed based on DC-GAN [31], which consists of a convolutional generator and a discriminator. For better performance, we introduce the Multi-Scale Context-Aware Network (MSCAN) to the generator. MSCAN is first proposed in [40] and has been proved to be effective for person re-identification. Different from basic convolution layer with only one scale filter, MSCAN has multiple filters with different scales. Thus, at each convolutional stage, the features with various scales can be extracted and combined, as shown in Fig. 4. In this paper, three convolutional scales are implemented at each stage. In this way, the generator can catch the gait features with different scales at each stage. Detailed structure of the global stream GAN is listed in Table 1.

### 3.4. Part stream

Although the global stream can learn to convert the full gait images, it may ignore some parts, e.g., the part of foot in gait image may be difficult to convert, due to the fast moving of the subject. One possible solution for this is to convert each part separately. Following the work of [41] in person ReID, we adopt rigid body parts to learn part-based feature. To this end, we evenly split the original gait image into three parts: the upper part (containing head and shoulders), middle part (containing the chest and stomach) and the lower part (containing the legs and feet), as shown in Fig. 2. We hope to restore the details of the three parts in local stream. Therefore, each part can be well learned by a separate GAN

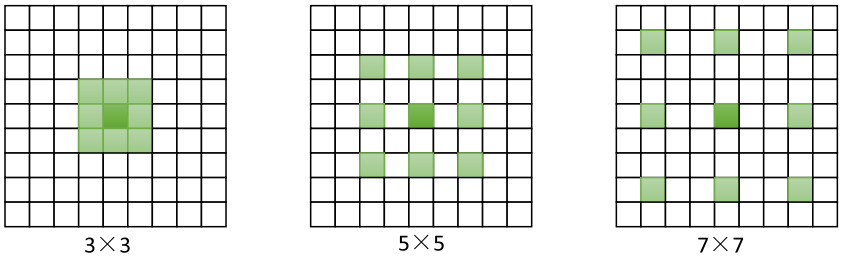



**Fig. 4.** Illustration of the multi-scale context-aware network. At each convolutional stage, there are three scaled filters. Though these filters are with the same kernel size, they have different dilations which can make different filter scales. Details of each MSCAN stage can refer to Table 2.

**Table 1**
Model architecture of the global-stream GAN and part-stream GAN. K: Kernel size, S: Stride, P: Padding, Rate: Dilated rate of dilated convolution, Output: -*-*- is the each convolution channel number(each layer have three convolution with different dilated rate besides the first layer and last layer, and the three convolution have the same channel number), width, height, respectively.

| Global-stream GAN | | | | | | Part-stream GAN | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Layer | K | S | P | Rate | Output | Layer | K | S w/h | P w/h | Output |
| image | – | – | – | – | 1*126*126 | Image | – | – | – | 1*42*126 |
| conv1 | 5 | 2 | 2 | 1 | 48*63*63 | conv1 | 5 | 2 | 2 | 48*19*63 |
| conv2 | 3 | 2 | 1 | 1/2/3 | 32*32*32 | conv2 | 3 | 2 | 1 | 64*10*32 |
| conv3 | 3 | 2 | 1 | 1/2/3 | 32*16*16 | conv3 | 3 | 2/1 | 1 | 64*10*16 |
| conv4 | 3 | 2 | 1 | 1/2/3 | 32*8*8 | conv4 | 3 | 2 | 1/2 | 64*6*8 |
| conv5 | 3 | 2 | 1 | 1/2/3 | 32*4*4 | conv5 | 3 | 2 | 1/2 | 64*4*4 |
| deconv1 | 3 | 2 | 0 | 1/2/3 | 32*9*9 | deconv1 | 3 | 2 | 0 | 64*9*9 |
| deconv2 | 3 | 2 | 1 | 1/2/3 | 32*17*17 | deconv2 | 3 | 2 | 1/3 | 64*13*17 |
| deconv3 | 3 | 2 | 1 | 1/2/3 | 32*33*33 | deconv3 | 3 | 2/1 | 1 | 64*13*33 |
| deconv4 | 3 | 2 | 2 | 1/2/3 | 32*63*63 | deconv4 | 3 | 2 | 2/3 | 64*21*63 |
| deconv5 | 4 | 2 | 1 | 1 | 1*126*126 | deconv5 | 4 | 2 | 1 | 1*42*126 |

**Table 2**
Experimental results of each probe views on CASIA-B. For each probe view, we compare it with the samples in gallery and average the 10 cross-view results as its result.

| Probe | PCA | Pixel-wise | Global-stream | Part-stream | Two-stream |
|---|---|---|---|---|---|
| 0° | 17.6 | 41.7 | 40.8 | 36.5 | **47.7** |
| 18° | 21.5 | 61.5 | 57.9 | 49.0 | **65.6** |
| 36° | 20.1 | 66.5 | 63.3 | 51.5 | **70.4** |
| 54° | 22.5 | 67.0 | 66.0 | 53.3 | **71.6** |
| 72° | 26.7 | 61.6 | 59.3 | 48.7 | **68.3** |
| 90° | 25.7 | 54.5 | 55.1 | 41.7 | **57.1** |
| 108° | 27.9 | 65.3 | 62.1 | 51.7 | **67.6** |
| 126° | 25.7 | 67.7 | 64.0 | 55.6 | **70.9** |
| 144° | 22.0 | 69.4 | 64.0 | 50.9 | **69.5** |
| 162° | 21.8 | 60.4 | 57.0 | 49.5 | **66.3** |
| 180° | 16.8 | 47.5 | 42.9 | 37.8 | **50.0** |
| Mean | 22.6 | 60.3 | 57.5 | 47.8 | **64.1** |

stream. Detailed structure of the part stream GAN is also listed in Table 1.

### 3.5. GAN Loss

The entropy loss is added to the discriminator to determine whether the input of the discriminator is a real or generated gait image. This loss can facilitate the generator to generate an image close to the real gait image with a standard view. The loss function of the real-or-fake discriminator is as follows:

$$L_D^{R/F} = -((1-y)\log(1-D_{R/F}(G(I))) + y\log D_{R/F}(x)) \quad (1)$$

Where $x$ is the real data, and $y$ is the label for the discriminator to recognize the real or fake input, i.e., $y$ equals to 1 for the real input and 0 for the generated input. Following the same training strategy with [9,31], we train the generators and discriminators iteratively.

### 3.6. Pixel-wise constraint

We add the pixel-wise loss to restrain the generators of both the global and three part streams. This loss is very similar with the FCN [42] model, which can accurately constrain the image generated by the generator at the pixel level. Although this constraint may have a slight side effect on the distribution learned by the generator, it is helpful for accurate gait image transformation and making the generator more stable in training phase. The pixel-wise loss of one image can be denoted as follows:

$$L_{pixel} = \sum_{i=1}^{w}\sum_{j=1}^{h}\left|x' - x\right| \quad (2)$$

Where $w$ and $h$ are the width and height of the image, respectively, and $x'$ is the generated image.

### 3.7. Identity discriminator

Besides the discriminators of TS-GAN, we add the identity discriminator to recognize the pedestrian's identity. We combine the global and local streams to train a total identity discriminator, which can learn better features than the two separated features. For the fusion of global and local stream, we proposed two different network architectures.

**Early-fusion.** The first one is early fusion, we extract global features and part features from the last layer of generators of global stream and part stream to concat in channel, then send them to a total identity discriminator.

**Late-fusion.** The second one is late fusion, two identity discriminators are respectively connected to the generators of global

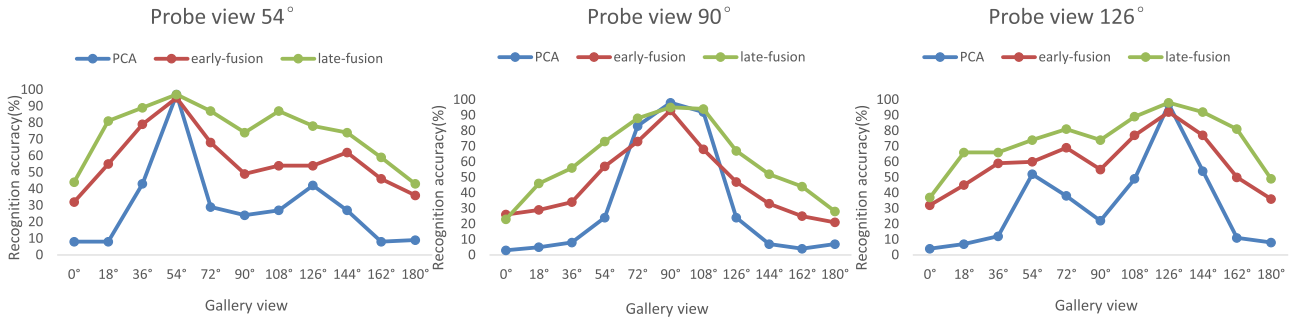



**Fig. 5.** Comparison of baseline PCA, early-fusion and late-fusion under the probe views 54°, 90° and 126°.

stream and part stream, as shown in Fig. 2. At the top of each discriminator is a soft-max classifier. We first extract features separately from the penultimate layer of the global-stream identity discriminator and part-stream identity discriminator, then we concatenate the global-stream and the part-stream features from channel direction. Finally we send the fusion feature to the total soft-max classifier to learn identity. In our experiments, we first train the two separate classifiers, then train the fusion one. In the testing phase, we extract the features from the last layer of the classifiers and compare the similarities between the probe and gallery samples to recognize human identities.

We select the popular and mostly used gait recognition approach proposed in [6] which applies Principal Component Analysis (PCA) to GEI as the baseline method. The performance of the two fusion methods and baseline are described in Fig. 5. We can see that late fusion is better than early fusion and baseline. Therefore, in the subsequent experiments, we use the method of late fusion to fuse global stream and part stream.

### 3.8. Objective function

The above three kinds of losses can be corporated jointly. The objective function used in this paper is the weighted sum of these three losses, and the weights of them are varying at each training phase. The formulation of total loss is as follows:

$$L_{all} = \partial_1 L_{GAN} + \partial_2 L_{pixel} + \partial_3 L_{ID} \tag{3}$$

where $L_{GAN}$, $L_{pixel}$ and $L_{ID}$ are the losses of the Two-Stream GAN, pixel-wise constraint and the identity discriminator, $\partial_{\#}$ is the hyper-parameter. In the pre-training phase, we train the generator to learn general view transformation, regardless of identity, so we set $\partial_1 = 0,\ \partial_2 = 1,\ \partial_3 = 0$. We alternately and iteratively train the generator and discriminator when training GAN. We first set $\partial_1 = 1,\ \partial_2 = 0.01,\ \partial_3 = 0$ to train the generator of GAN model. We then set $\partial_1 = 1,\ \partial_2 = 0,\ \partial_3 = 0$ to train the discriminator of GAN model. Note that above training is to learn a general view transform model which has none identity related information. Thus when training the identity discriminator, we set $\partial_1 = 0,\ \partial_2 = 0.1,\ \partial_3 = 1$.

## 4. Experiments

We evaluate the proposed method and compare it with several state-of-the-art methods on two large-scale currently available gait databases: CASIA-B [3] and OU-ISIR [4].

### 4.1. Datasets

#### 4.1.1. CASIA-B [3]

This gait database was established by the Institute of Automation, Chinese Academy of Sciences(CASIA). It provides the gait videos, silhouette squences and GEI images. There are 124 subjects in this dataset, and each of them has 10 gait sequences. For each subject, 6 sequences were taken under natural walking conditions (NM#1-6), 2 sequences were taken when walking with a bag (BG#1-2), and 2 sequences were taken when walking with a coat (CL#1-2). There are 11 views for each condition and viewing angles range from 0° to 180° with an adjacent interval view of 18°. The examples are shown in Fig. 6.

#### 4.1.2. OU-ISIR [4]

The OU-ISIR gait database was created by the Institute of Scientific and Industrial Research (ISIR), Osaka University (OU). The examples are shown in Fig. 6. There are 2,135 male subjects and 1,872 female subjects (in total 4,007) in this database. It should be noted that the age of subjects ranges from 1 to 94 years. Each subject has 2 gait sequences as gallery and probe, both of which were taken under normal walking conditions. There are 4 views in each sequence, i.e., 55°, 65°, 75°, and 85°. Following the experiment in [7], we select a subset with 1,912 subjects for evaluation. The first 956 subjects are in the training set and the left 956 are for testing.

### 4.2. Implementation details

(a) Model parameters: The detailed network structures of the global-stream and part-stream generators are shown in

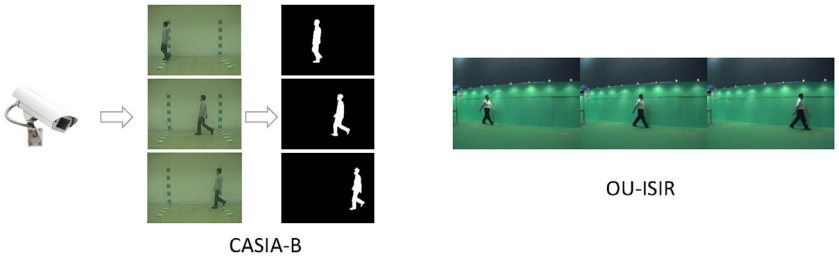



**Fig. 6.** Examples of the CASIA-B and OU-ISIR datasets.

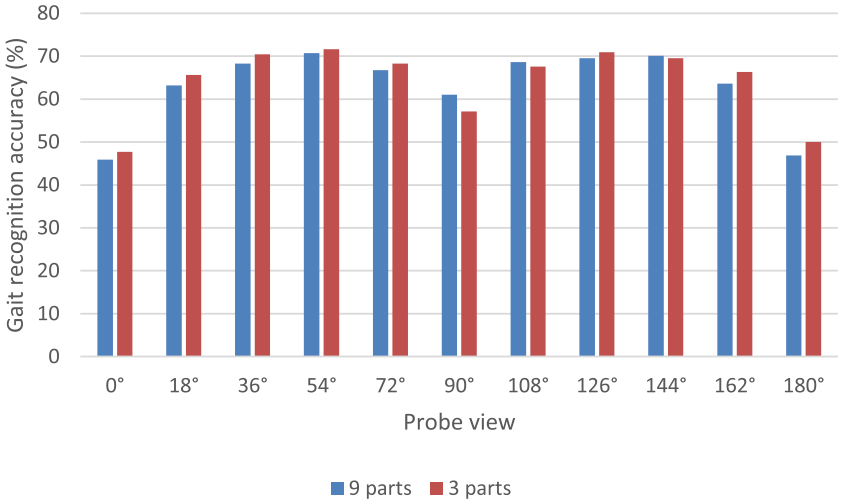



**Fig. 7.** Comparison of different part divisions. We evaluate our method with three or nine parts.

Table 1 and, respectively. The global stream takes a gait image with a size of 126 × 126 as input, whereas the part stream takes cropped parts with a size of 42 × 126.

(b) Training: Our experiments are conducted on the Caffe framework. The batch gradient descent method is adopted to minimize the objective loss. We set the size of each batch to 32, the momentum to 0.9, and a weight decay of 0.0005. When training the generator, the initial learning rate is $10^{-5}$, whereas tuning it to 0.001 when training the discriminators and 0.01 for training the identity discriminator. As described in Section 3.10, we train each phase roughly $2 \times 10^4$ iterations.

### 4.3. Effectiveness of Two-Stream GAN

We first evaluate the effectiveness of our method on the CASIA-B dataset. As this work focuses on cross-view gait recognition, we implement the experiments on the NM sequences with all the views. The first 74 subjects under normal walking conditions, i.e., NM#1-6 are in the training set, the left 50 subjects are in the testing set. In the testing phase, we set the NM#1-4 as gallery, NM#5-6 as probe. In detail, we use a total of 4884 training samples in CASIA-B dataset. In order to find a suitable way to divide the body GEI into parts, we exploit two different manners. One is to divide the GEI evenly into three parts from top to bottom and the other one is to further divide the GEI into nine parts from left to right. We perform this evaluation on CASIA-B dataset. The experimental results are listed in Fig. 7. The experimental results show that the division of GEI (part-stream) into more parts can not increase the performance greatly. Because the gait energy image is different from other images, the division of images from other fields into more parts may focus on more local information, such as colors, textures, and so on, while gait energy image do not contains such information. Therefore, dividing the gait energy image into 3 parts can maintain enough local details. In order to verify the effectiveness of multi-scale context-aware unit, we implement experiments to evaluate the performance of the global-stream model with or without the multi-scale unit. The experimental results are listed in Fig. 8. It can be concluded that the performance of GAN with multi-scale context-aware unit has been obviously improved. Because the multi-scale context-aware network can learn gait features form context with different scales, these features contain more scale-aware information, which may be critical for improving recognition accuracy. We use the two-stream generators with pixel-wise loss to generate the gait images with the standard view, and report the result in Table 2. The experiment result means Two-Stream GAN is more effective than two-stream

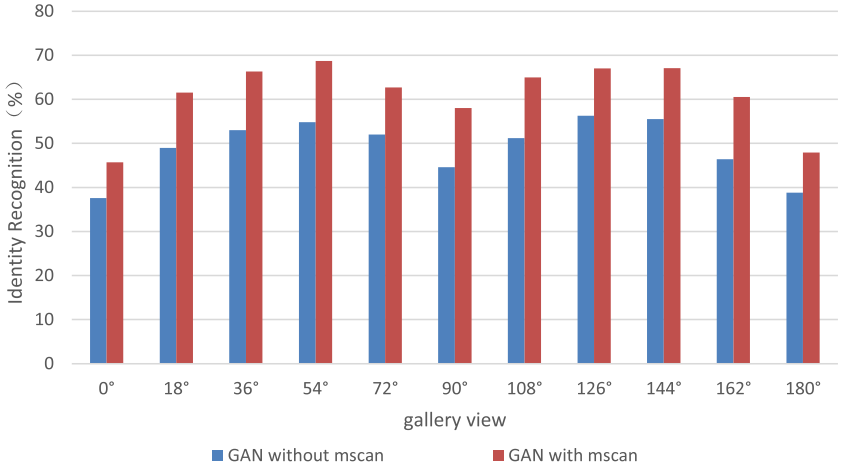



**Fig. 8.** Comparison of the models with or without the multi-scale context-aware unit. This experiment is implemented with the global-stream GAN.

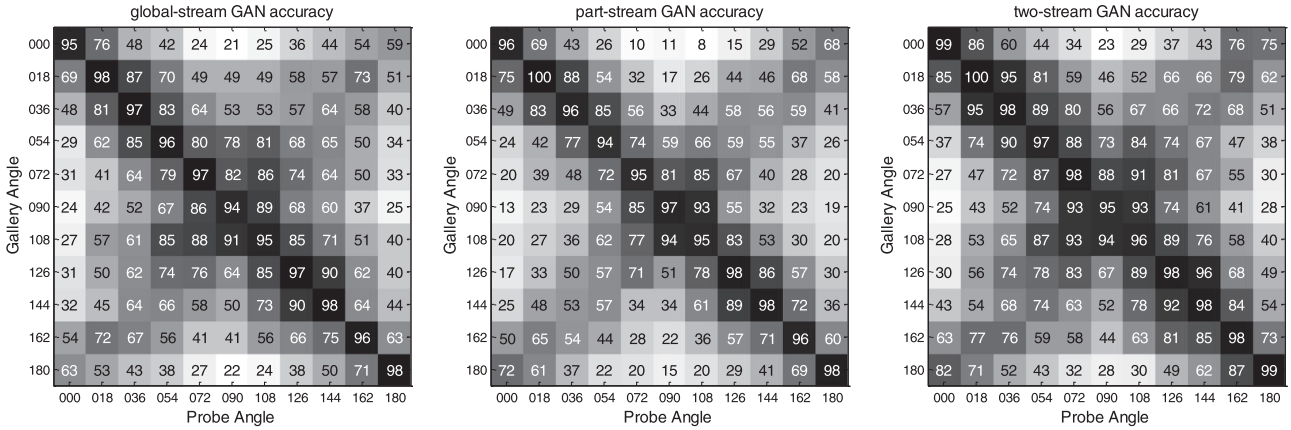



**Fig. 9.** The cross-view gait recognition results of global-stream, part-stream, and two-stream. The models are trained on the first 74 subjects on CASIA-B, and the models are tested on the left subjects.

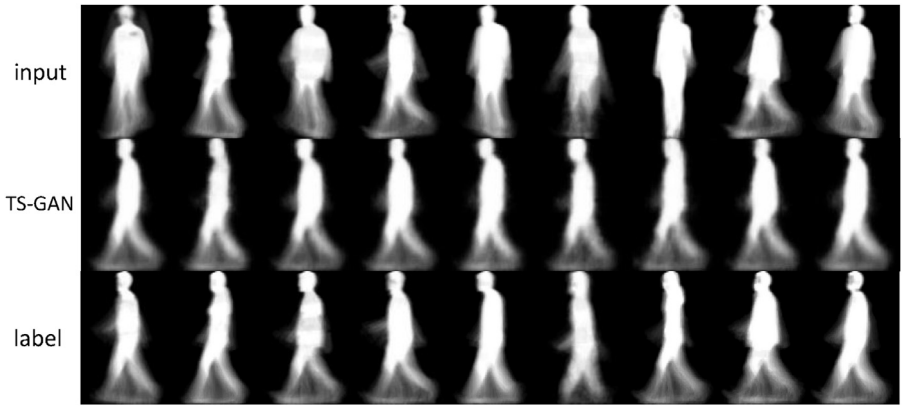



**Fig. 10.** Visualization of the generated images by TS-GAN. The first and the last rows are the input images and labels, respectively. It can be seen that the generated images by TS-GAN look satisfying.

generators with pixel-wise loss. Because GAN is very powerful, it can learn the distribution of the real data through alternating iterative training between generator and discriminator. Therefore, it is reasonable that our TS-GAN model with pixel-wise loss can achieve better performance than the two-stream generators model with pixel-wise loss.

To test the cross-view performance of each stream, we train the two streams separately, and report the results of them. Furthermore, we jointly train the two streams via fusing the features from both the global and part streams, the result becomes better than that of single stream. The experimental results are listed in Table 2. It is obvious that our proposed methods achieve much better performance than the baseline method. Although the feature of the part stream is a slight weaker than the global one, the fusion feature of the two streams outperforms the single-stream feature. This result shows that it is effective to use GAN to learn view-invariant gait features, and fusing the any global and the part streams can further boost the performance. The performance of Two-Stream GAN is better than that of the single-stream GAN, because the global stream could learn to transform the view of the whole image, while the part-stream GAN could learn the local details which are difficult to transform. It also indicates that the fusion of two streams can make full use of both global and local features to achieve better performance.

The view-to-view results of three feature models are shown in Fig. 9. We can draw three main conclusions from these results: (a) Though a powerful GAN is adopted to learn to transfer the gait image, e.g., GEI, the accuracies will become better when the view difference becomes smaller. It shows that the cross-view gait recognition is still challenging. (b) Fusing features can intergrate the advantages from each stream, especially for the case of large view differences. We also visualize the generated gait images of TS-GAN for intuitive understanding. As shown in Fig. 10, the TS-GAN can transform GEIs of any views to a standard view. It can be found that the images generated by the TS-GAN are clear and look satisfying.

### 4.4. Comparison with state-of-the-art methods

We further verify the performance of our method via comparing with the state-of-the-art cross-view gait methods on CASIA-B and OU-ISIR.

### 4.5. Comparison on CASIA-B

We compare the TS-GAN with several recently proposed state-of-the-art methods, including C3A [23], ViDP [24], SPAE [29] and GaitGAN [9]. We follow the same experimental setting with GaitGAN [9], which trains the model with the first 62 subjects and tests on the left subjects. The experimental results are listed in Table 3. It proves that our method outperforms the compared methods and is 3.4 percentage points higher than GaitGAN [9] on average, because of the introduction of pixel-wise loss and the part-stream which could learn detail informations.

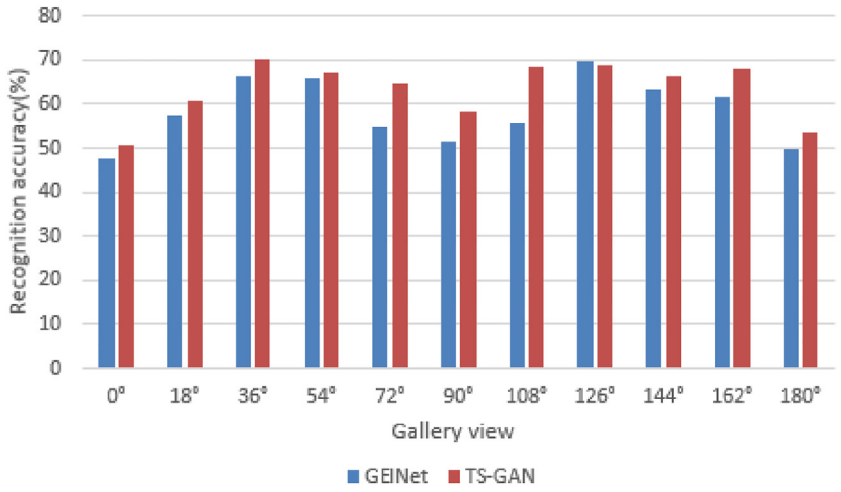



**Fig. 11.** The comparison of our method with GEINet [7] on the CASIA database.

**Table 3**
Comparison with the state-of-the-art methods on CASIA-B. All the models are trained on the first 62 subjects, and tested on the left subjects.

| Method | Probe angle | | | Mean |
|---|---|---|---|---|
| | 54° | 90° | 126° | |
| C3A [23] | 56.6 | 54.7 | 58.3 | 56.6 |
| ViDP [24] | 64.2 | 60.4 | 65.0 | 63.2 |
| SPAE [29] | 63.3 | 62.1 | 66.3 | 63.9 |
| GaitGAN [9] | 64.5 | 58.2 | 65.7 | 62.8 |
| **Our method** | **67.3** | **63.1** | **68.2** | **66.2** |

**Table 4**
Comparison with the state-of-the-art methods on OU-ISIR. All the models are trained on the first 956 subjects, and tested on the left subjects.

| Gallery view | Method | Probe view | | | |
|---|---|---|---|---|---|
| | | 55° | 65° | 75° | 85° |
| 55° | wQVTM [10] | – | 78.3 | 64 | 48.6 |
| | TCM+ [25] | – | 79.9 | 70.8 | 54.5 |
| | GEINet [7] | (94.7) | 93.2 | 89.1 | 79.9 |
| | **Our method** | **(91.2)** | **88.7** | **84.7** | **76.5** |
| 65° | wQVTM | 81.5 | – | 79.2 | 67.5 |
| | TCM+ | 81.7 | – | 79.5 | 70.2 |
| | GEINet | 93.7 | (95.1) | 93.8 | 90.6 |
| | **Our method** | **90.2** | **(93.7)** | **91.7** | **87.4** |
| 75° | wQVTM | 70.2 | 80.0 | – | 78.2 |
| | TCM+ | 71.9 | 80.0 | – | 79.0 |
| | GEINet | 90.1 | 94.1 | (95.2) | 93.8 |
| | **Our method** | **85.9** | **91.9** | **(92.2)** | **91.1** |
| 85° | wQVTM | 51.1 | 68.5 | 79.0 | – |
| | TCM+ | 53.7 | 73.0 | 79.4 | – |
| | GEINet | 81.4 | 91.2 | 94.6 | (94.7) |
| | **Our method** | **78.6** | **87.7** | **90.1** | **(93.1)** |

### 4.6. Comparison on OU-ISIR

As this dataset contains only 4 views, it is easier to achieve high accuracies. The experiment settings are the same as GEINet [7], and we considered 7648 training sample in OU-ISIR dataset. We compare with the famous methods of the view transformation model [10] and transformation consistency measures [25] on this dataset. VTM, TCM and GEINet are the well-known and classic methods. The view-to-view results are listed in Table 4. The proposed TS-GAN performs better than VTM [10] and TCM [25] in almost all cross-view conditions, while TS-GAN performs slightly worse than the pioneer method GEINet [7]. It demonstrates the effectiveness of the proposed Two-Stream GAN for cross-view gait recognition. For more comparison with GEINet, we also re-implement the GEINet method on CASIA-B dataset. The experimental setup is the same as described in Section 4.3. We compare the performance of GEINet and our TS-GAN in Fig. 11. It shows that the performance of TS-GAN is better than GEINet on CASIA-B. By analyzing the results of the two experiments on OU-ISIR and CASIA-B, we can conclude that TS-GAN is more stable on the database with large view conversion difference, while GEINet has a good performance on the dataset with large number of ID.

### 4.7. Discussion

The comparison experiments of TS-GAN with and without pixel-wise loss have proven the effectiveness of GAN based view transformation. As a powerful generation model, the learnt distribution by GAN is more close to the distribution of the true gait data, thus GAN is a suitable selection for gait view conversion. However, our method also has some limitations. We will discuss those challenges from two main aspects. Firstly, compare with the GEINet method, our method performs well on the recognitions with large view difference, while performs worse on the dataset with large amount IDs such as OU-ISIR. We need to increase the generalization ability of our method for such large-scale dataset. Secondly, the recognition accuracy of cross-view conditions are still not satisfying which need further improvement for the GAN unit. We will explore to address those challenges in our future works.

## 5. Conclusion and future work

In this paper, we have proposed a two-stream generative adversarial network (TS-GAN) for cross-view gait recognition. The proposed TS-GAN can learn view invariant gait features from both of the global and part streams, indicating both the identical and local details. We also have added a pixel-wise loss along with TS-GAN to restore the gait details in pixel-level. In addition, we have introduced the multi-scale context-aware unit in the identity discriminator to learn better gait features. We have evaluated the proposed method on two large gait databases: CASIA-B and OU-ISIR. Experimental results have shown that our approach is effective and outperforms the compared state-of-the-art methods.

In the future, we will use this model to deal with more challenging problems in gait recognition field, such as gait recognition

with varying speeds, different clothing conditions and carrying bags. At present, there are few large gait databases related to speed varying. Therefore, we will consider establishing such a speed-related gait database, and further improve our methods. In addition, we will try to explore the more sophisticated fusion manner of TS-GAN to further improvement the performance.

## Acknowledgement

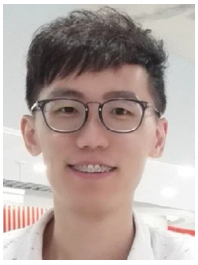
This work is jointly supported by National Key Research and Development Program of China (2016YFB1001000), National National Science Foundation of China (61573139, 61525306, 61633021, 61721004, 61420106015), Capital Science and Technology Leading Talent Training Project (Z181100006318030), the Fundamental Research Funds for the Central Universities (2018ZD05) and Beijing Science and Technology Project (Z181100008918010)

## References

[1] M.P. Murray, A.B. Drought, R.C. Kory, Walking patterns of normal men, J. Bone Joint Surg. Am. 46 (2) (1964) 335.
[2] M.P. Murray, Gait as a total pattern of movement., Am. J. Phys. Med. Rehabil. 46 (1) (1967) 290.
[3] 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 International Conference on Pattern Recognition, 2006, pp. 441–444.
[4] 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 (5) (2012) 1511–1521.
[5] N. Takemura, Y. Makihara, D. Muramatsu, T. Echigo, Y. Yagi, Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2001, p. 726.
[6] L. Wang, T. Tan, H. Ning, W. Hu, Silhouette analysis-based gait recognition for human identification, IEEE Trans. Pattern Anal. Mach. Intell. 25 (12) (2003) 1505–1518.
[7] K. Shiraga, Y. Makihara, D. Muramatsu, T. Echigo, Y. Yagi, GEINet: view-invariant gait recognition using a convolutional neural network, in: Proceedings of the International Conference on Biometrics, 2016, pp. 1–8.
[8] Z. Wu, Y. Huang, L. Wang, X. Wang, T. Tan, A comprehensive study on cross-view gait based human identification with deep CNNs, IEEE Trans. Pattern Anal. Mach. Intell. 39 (2) (2017) 209–226.
[9] S. Yu, H. Chen, E.B.G. Reyes, N. Poh, GaitGAN: Invariant gait feature extraction using generative adversarial networks, in: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 532–539.
[10] Y. Makihara, R. Sagawa, Y. Mukaigawa, T. Echigo, Y. Yagi, in: Gait recognition using a view transformation model in the frequency domain, European Conference on Computer Vision 5 (2006) 151–163.
[11] W.E.L. Grimson, Gait analysis for recognition and classification, in: Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition, 2002, pp. 148–155.
[12] D. Cunado, M.S. Nixon, J.N. Carter, Using gait as a biometric, via phase-weighted magnitude spectra 1206 (1206) (1997) 93–102.
[13] S.L. Dockstader, M.J. Berg, A.M. Tekalp, Stochastic kinematic modeling and feature extraction for gait analysis, IEEE Trans. Image Process. 12 (8) (2003) 962–976.
[14] J. Han, B. Bhanu, Performance prediction for individual recognition by gait, Pattern Recognit. Lett. 26 (5) (2005) 615–624.
[15] L. James J, B. Jeffrey E, Recognizing people by their gait : the shape of motion, J. Comput. Vis. Res. 1 (2) (1998) 1.
[16] J.P. Foster, M.S. Nixon, A. Prgel-Bennett, Automatic gait recognition using area-based metrics, Pattern Recognit. Lett. 24 (14) (2003) 2489–2497.
[17] K. Bashir, T. Xiang, S. Gong, Cross view gait recognition using correlation strength, in: Proceedings of the British Machine Vision Conference, 2010, pp. 1–11.
[18] A.A. Kale, N. Cuntoor, V. Krger, Gait-based recognition of humans using continuous HMMs, in: Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition, 2002, p. 0336.
[19] H. Murase, Moving object recognition in eigenspace representation : gait analysis and lip reading, Pattern Recognit. Lett. 17 (95) (1996) 155–162.
[20] P.S. Huang, C.J. Harris, M.S. Nixon, Recognizing humans by gait via parametric canonical space 13 (4) (1998) 359–366.
[21] 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 IEEE International Conference on Computer Vision Workshops, 2010, pp. 1058–1064.
[22] W. Kusakunniran, Q. Wu, J. Zhang, H. Li, Support vector regression for multi-view gait recognition based on local motion feature selection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 974–981.
[23] X. Xing, K. Wang, T. Yan, Z. Lv, Complete canonical correlation analysis with application to multi-view gait recognition, Pattern Recognit. 50 (C) (2016) 107–117.
[24] M. Hu, Y. Wang, Z. Zhang, J.J. Little, D. Huang, View-invariant discriminative projection for multi-view gait-based human identification, IEEE Trans. Inf. Forensics Secur. 8 (12) (2013) 2034–2045.
[25] D. Muramatsu, Y. Makihara, Y. Yagi, Cross-view gait recognition by fusion of multiple transformation consistency measures, Biom. Iet 4 (2) (2015) 62–73.
[26] T. Wolf, M. Babaee, G. Rigoll, Multi-view gait recognition using 3D convolutional neural networks, in: Proceedings of the IEEE International Conference on Image Processing, 2016, pp. 4165–4169.
[27] X. Chen, J. Weng, W. Lu, J. Xu, Multi-gait recognition based on attribute discovery, IEEE Trans. Pattern Anal. Mach. Intell. 27 (99) (2018) 1.
[28] N. Takemura, Y. Makihara, D. Muramatsu, T. Echigo, Y. Yagi, On input/output architectures for convolutional neural network-based cross-view gait recognition, IEEE Trans. Circuits Syst. Video Technol. 27 (99) (2017) 1.
[29] S. Yu, H. Chen, Q. Wang, L. Shen, Y. Huang, Invariant feature extraction for gait recognition using only one uniform model, Neurocomputing 239 (C) (2017) 81–93.
[30] I. Goodfellow, J. Pougetabadie, M. Mirza, B. Xu, D. Wardefarley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, Adv. Neural Inf. Process. Syst. 27 (2014) 2672–2680.
[31] A. Radford, L. Metz, S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks, arxiv (2015) 1.
[32] R. Huang, S. Zhang, T. Li, R. He, Beyond face rotation: Global and local perception GAN for photorealistic and identity preserving frontal view synthesis, in: IEEE International Conference on Computer Vision, 2017, pp. 2458–2467.
[33] F. Samaria, F. Fallside, Face identification and feature extraction using hidden Markov models, in: Proceedings of the Image Processing Theory and Applications, 1993, pp. 295–298.
[34] J. Wang, X. Zhu, S. Gong, W. Li, Attribute recognition by joint recurrent learning of context and correlation, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 531–540.
[35] X. Zhao, L. Sang, G. Ding, X. Jin, Grouping attribute recognition for pedestrian with joint recurrent learning, in: Proceedings of the International Joint Conference on Artificial Intelligence, 2018, p. 1.
[36] J. Man, B. Bhanu, Individual recognition using gait energy image, IEEE Trans. Pattern Anal. Mach. Intell. 28 (2) (2005) 316–322.
[37] B. Khalid, T. Xiang, S. Gong, Gait recognition using gait entropy image, in: Proceedings of the International Conference on Crime Detection and Prevention, 2009, pp. 1–6.
[38] T.H.W. Lam, K.H. Cheung, J.N.K. Liu, Gait flow image: a silhouette-based gait representation for human identification, Pattern Recognit. 44 (4) (2011) 973–987.
[39] C. Wang, J. Zhang, J. Pu, X. Yuan, L. Wang, Chrono-gait image: a novel temporal template for gait recognition, in: Proceedings of the European Conference on Computer Vision, 2010, pp. 257–270.
[40] D. Li, X. Chen, Z. Zhang, K. Huang, Learning deep context-aware features over body and latent parts for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 7398–7407.
[41] D. Cheng, Y. Gong, S. Zhou, J. Wang, N. Zheng, Person re-identification by multi-channel parts-based CNN with improved triplet loss function, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1335–1344.
[42] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431–3440.

**Yanyun Wang** received her B.Sc. degree in ShanDong University of Technology, China, in 2016. She is currently pursuing the Master's degree at North China Electric Power University, China. Her research interests include pattern recognition, computer vision and gait recognition.

**Chunfeng Song** received his B.Sc. degree in QiLu University Of Technology, China, in 2012 and M.Sc. degree in North China Electric Power University, China, in 2016. He is now a Ph.D. candidate working in the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing, China. His research interests include image segmentation, data clustering, gait recognition and deep learning.



**Yan Huang** received the B.Sc. degree from University of Electronic Science and Technology of China (UESTC) in 2012, and the Ph.D.degree from University of Chinese Academy of Sciences (UCAS) in 2017. Since July 2017, He has joined the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA) as an assistant professor. His research interests include machine learning and pattern recognition. He has published papers in the leading international journals and conferences such as IEEE TPAMI, IEEE TIP, NIPS, ICCV and CVPR.

**Zhenyu Wang** received his B.Sc. degree from National Defense University of Science and Technology, China, in 1997 and M.Sc. degree from Xi'an Jiaotong University, China, in 2003 and Ph.D. degree from National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, China, in 2007. He is currently an Associate Professor with the School of Control and Computer Engineering, North China Electric Power University, Beijing, China. His research interests include pattern recognition, computer vision and machine learning.

**Liang Wang** received both the B.Eng. and M.Eng. 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. From 2004 to 2010, he was a research assistant at Imperial College London, United Kingdom, and Monash University, Australia, a research fellow at the University of Melbourne, Australia, and a lecturer at the University of Bath, United Kingdom, respectively. Currently, he is a full professor of the Hundred Talents Program at the National Lab of Pattern Recognition, CASIA. His major research interests include machine learning, pattern recognition, and computer vision. He has widely published in highly ranked international journals such as IEEE Transactions on Pattern Analysis and Machine Intelligence and IEEE Transactions on Image Processing, and leading international conferences such as CVPR, ICCV, and ICDM. He is a Fellow of the IEEE, and an IAPR Fellow.