VT-GAN: View Transformation GAN for Gait Recognition Across Views

1st Peng Zhang
Global Big Data and Technologies Centre
University of Technology Sydney
Sydney, Australia
pengzhang.sdu@gmail.com

2nd Qiang Wu, 3rd Jingsong Xu
Global Big Data and Technologies Centre
University of Technology Sydney
Sydney, Australia
{Qiang.Wu, Jingsong.Xu}@uts.edu.au

***Abstract*—Recognizing gaits without human cooperation is of importance in surveillance and forensics because of the benefits that gait is unique and collected remotely. However, change of camera view angle severely degrades the performance of gait recognition. To address the problem, previous methods usually learn mappings for each pair of views which incurs abundant independently built models. In this paper, we proposed a View Transformation Generative Adversarial Networks (VT-GAN) to achieve view transformation of gaits across two arbitrary views using only one uniform model. In specific, we generated gaits in target view conditioned on input images from any views and the corresponding target view indicator. In addition to the classical discriminator in GAN which makes the generated images look realistic, a view classifier is imposed. This controls the consistency of generated images and conditioned target view indicator and ensures to generate gaits in the specified target view. On the other hand, retaining identity information while performing view transformation is another challenge. To solve the issue, an identity distilling module with triplet loss is integrated, which constrains the generated images inheriting identity information from inputs and yields discriminative feature embeddings. The proposed VT-GAN generates visually promising gaits and achieves promising performances for cross-view gait recognition, which exhibits great effectiveness of the proposed VT-GAN.**

I. Introduction

Gait refers to the locomotion of human body, which depicts one's unique walking style and has been proved potential for person identification. Distinct from other popular biometric modalities such as face, iris, fingerprint, vein and lip-print, the uniqueness of gait is the fact that it can be collected without awareness of the observed subject, *i.e.*, it can be captured remotely without prior informed consent and without physical contact. This brings some competitive benefits as a biometric trait, *e.g.*, gait is unobtrusive and hard to hide as well as pretend. Considering the advantages, gait recognition is significant in applications of forensics and surveillance security.

However, gait recognition is still a challenging because of various factors in practical scenarios, for instance, walking direction (or view) variation [15], non-constant walking speed [29], bear loading [28], appearance of or different type of carrying bags [20], changes of dressing [32], *etc*. Among these factors, the most representative one is view variation because the observed subject may appear and walk along different route under one deployed camera in real world. It attracts wide interest in recent works and boosts the accuracy of gait recognition. However, previous literatures mainly focus on distilling features, which rarely visualizes differences between gaits across different views. In this work, we also target on the cross-view gait recognition task and further try to explicitly compares the distinctiveness of gaits from various views.

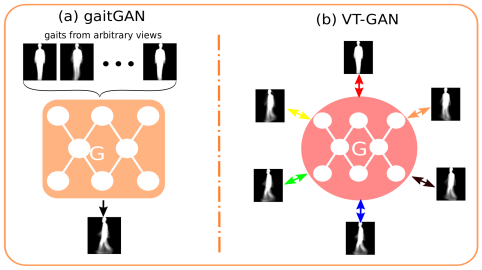


Fig. 1. Comparison between gaitGAN and the proposed VT-GAN. (a) The former normalizes GEIs from arbitrary views to a reference one, and (b) the latter directly transform gait images between any pair of views.

To address gait recognition across views, previous research experiences three stages which are model-based methods, mapping-based methods and deep learning-based methods. In the early stage, researchers mainly focus on body measurement based on human body model, *e.g.*, extracting view invariant features by measuring geometric parameters of human body [3], or reconstructing 3D body using depth sensors and then re-projecting to 2D in arbitrary view [37]. However, the former is not accurate and the latter is too complicated. In addition, the depth sensors are not practical in real scenarios due to its higher expenses. In contrast, gait recognition using visual RGB sensors is more practical since the fact that surveillance cameras are deployed everywhere. Thus, approaches using 2D images are arising in the middle stage, which correlate gait templates across views by linear mapping such as regressing gait templates from one view to another view [15], [16], [21] or projecting gait templates across views into a common feature space via asymmetric mapping [1], [2], [21], [31]. These approaches are popular in past years because of their effectiveness and computational efficiency. However, this kind of approaches can only process gaits across views in pair, which will yield multiple models with the increase of views.

In addition, these methods rely on view estimation, though it is not a problem nowadays. With development of deep neural networks, it is fashion to distill view invariant features using convolutional neural networks (CNN) [11], [26], [30]. This kind of methods significantly improve the accuracy of cross-view gait recognition. However, CNN is a black-box, which is hard to interpret visually. In comparison, view transformation using encoder-decoder model [34] or generative adversarial network (GAN) [9], [33] as regressor exactly solves the interpretable problem, which transforms gait templates from various views to a canonical one only using a single model. However, it can only transform gaits to a specific one, which is hard to visualize view relationship between pairs from any two views. If we want to transform gaits across any two views, a lot of models are needed, *i.e.*, the number of views. In contrast, this paper tries to visualize the view transformation relationship between any two views using only one model and thus achieve cross-view gait recognition explicitly. The difference between current GAN-based view normalization and the proposed VT-GAN is shown in Fig. 1.

Recently, image translation [4], [13], [18] using conditional GAN becomes one of the hottest topics in computer vision, which transforms images from source domain to target domain conditioned on specific requirements such as style transfer, image impainting, image editor, *etc*. For instance, some recent works try to generate persons in specific pose, synthesize faces in conditioned expression, hair style or even gender. In specific, the popular starGAN [4] can simultaneously process several attributes using only a single model, which demonstrates great success in face editing. Inspired by idea of starGAN, the paper proposes to perform view transformation between gaits from any two views name view transformation GAN (VT-GAN). As in Fig. 2, the proposed VT-GAN architecture consists of a generator, a discriminator and a similarity preserver. The generator $\mathbf{G}$ takes the gait templates from source view and the corresponding reference view indicator as input, and generates images in the reference view. And, the discriminator $\mathbf{D}$ learns to not only distinguish if the input images are real but also classify the synthesized images to its corresponding view class. To preserve the identity information and make the synthesized gaits discriminative, we impose two regularization terms, i.e., identity preserving term and identity discriminative term. The former forces the generated gaits close to the target reference one and the latter pulls feature embedding of generated images extracted by a similarity preserver $\mathbf{\Phi}$ close to their corresponding positive samples in target view and push the negatives far away with a margin. Through such a supervised training strategy, the synthesized gait image (on the target view) created by generator can be distinguished against the same/different subject under the same view (i.e. target view). In this way, we achieved identity-preserved view transformation by a single model and depicts relationships across any two views.

To conclude, the proposed VT-GAN brings following benefits,

- A novel view-transformation framework for cross-view gait recognition named VT-GAN is presented. Distinct from existing GAN-based approaches which normalizes gaits to a unified view, the proposed framework takes both gaits in source view and a reference view indicator as input and generates gaits in the reference view controlled by a real/fake discriminator and a view classifier. In this way, the proposed VT-GAN can achieve view transformation across any two views using a single model.
- We integrate an identity preserver to the whole framework, which attempts to preserve the identity information while performing view transformation. This is beneficial to gait recognition.

II. Related Works

In this section, we review literatures in two fields that are related to our work, *i.e.*, cross-view gait recognition and conditional generative adversarial networks.

A. Cross-view Gait Recognition

The development of approaches for cross-view gait recognition includes three stages, *i.e.*, model-based methods, mapping-based methods and deep learning-based methods. In the early stage, approaches are focused on constructing and analysing human body model, *i.e.*, body parameters measurement [3] and 3D reconstruction [37]. The former is simple but suffers poor performances. And in contrast, the 3D-based model achieved promising performance, but it is expensive as well as complicated since multiple calibrated cameras are needed. After that, mapping-based approaches using 2D images are prevailing to address the cross-view gait recognition problem. These methods either learn to map gaits from one view to another view, *i.e.*, VTM-based (view transformation model) models [15], [16], [23], [39], or asymmetrically project gaits from different views to a shared space, *i.e.*, CCA-like (canonical correlation analysis) models [2], [21], [31]. Though this kind of approaches achieve promising performance, they heavily rely on view estimation. Moreover, these approaches only perform gait recognition across a couple of views which would yield abundant of models with increasing number of views. Recently, approaches using neural network achieves significant performances. For instance, Wu *et al.* [30] proposed to learn differences between gaits from arbitrary views with CNN. Shiraga *et al*. [26] proposed GEINet based on CNN framework which demonstrate effectiveness when view changes are small. However, these models suffer interpretation problem due to the black-box characteristics of CNN. More recently, GAN is applied to gait recognition which try to explicitly visualize the synthesized GEIs while keeping competitive performances, *e.g.* gaitGAN [33] and MGANs [9]. These models normalized gaits from different views into a reference one which behaves as a regressor but only uses one single model. Rather than view normalization, the proposed method attempts to synthesize gaits of arbitrary views from a typical view. It is useful to predict unknown views of a typical identities.

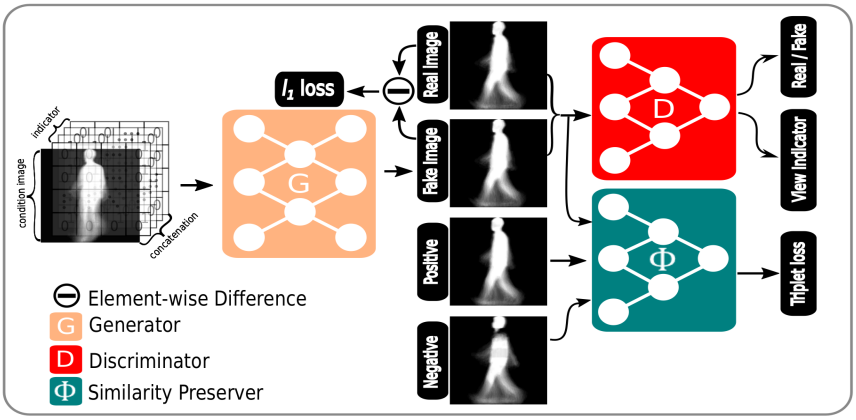


Fig. 2. The architecture of the proposed VT-GAN. It consists of three modules, *i.e.*, generator $\mathbf{G}$, discriminator $\mathbf{D}$ and similarity preserver $\mathbf{\Phi}$. Generator $\mathbf{G}$ inputs the concatenation of source/condition gait image from arbitrary view and target view indicator, and synthesizes gait image from target view. Discriminator $\mathbf{D}$ learns to distinguish between synthesized gait image and real gait image and classify real gait image to its corresponding view. Similarity preserver $\mathbf{\Phi}$ learns to pull positive gait pairs together and push negative gait pairs away.

B. Conditional Generative Adversarial Networks

Generative adversarial networks (GAN) [6] has been actively studied since it was proposed and incurs many variations. One of the most popular is conditional GANs [22] which are usually conditioned on class labels [22], text description [24] and specific attributes [4], [18], [27]. In specific, Choi *et al*. [4] proposed starGAN which achieves multiple domain transferring conditioned on provided domain information using a single model. They achieved remarkable performances in facial attribute transfer and facial expression generation. Inspired by this, we presented a novel framework which treats view angle as attribute and tries to synthesize arbitrary views controlled by provided view indicator. Different from starGAN, an identity-preserving module is integrated to our framework, which tries to keep the identity consistency while performing view transfer. By coordinating losses of GAN and imposed identity-related losses, the proposed framework thus can synthesize view-guided identity-preserving gaits for cross-view gait recognition.

III. View Transformation Generative Adversarial Network

The proposed View Transformation Generative Adversarial Network (VT-GAN) aims to transform gaits from one specific view to arbitrary views using a single model. The overall framework of the proposed VT-GAN is shown in Fig. 2 which consists of a generator $\mathbf{G}$, a discriminator $\mathbf{D}$ and an identity preserver $\mathbf{\Phi}$. The generator $\mathbf{G}$ takes the gaits x_{src} from source view θ and the target view indicator c as input, and generate gaits in the target view ϑ, *i.e.*, $\mathbf{G}(x_{src}, c) \mapsto x_{dst}$. In the proposed VT-GAN, we use one-hot vector to denote the view indicator where the target view is assigned 1 and other views are assigned 0. Thus, x_{src} from the source view θ can be translated to x_{dst} from arbitrary view ϑ specified by indicator c. And, the discriminator $\mathbf{D}$ learns to not only distinguish if the input images are real but also classify the synthesized images to its corresponding domain. That is, the discriminator produces probability distributions for both real/fake discrimination and view classification [4], $\mathbf{D} : x \mapsto \{\mathbf{D}_{dis}(x), \mathbf{D}_{cls}(x)\}$, where x is the input to discriminator. To preserve the identity information and make the synthesized gaits discriminative, we impose two regularization terms, *i.e.*, identity preserving term and identity discriminative term. The former forces the generated gaits close to the target reference one and the latter pulls feature embedding of generated images extracted by a similarity preserver $\mathbf{\Phi}$ [5] close to their corresponding positive samples in target view and push the negatives far away with a margin.

Adversarial Loss. We apply adversarial learning in our training process to constrain the output of $\mathbf{G}$ visually similar to the reference gait x_{dst} from view ϑ. The adversarial loss is defined as

$$\begin{aligned}\mathcal{L}_{adv}(\mathbf{G}, \mathbf{D}_{dis}, p_\theta, p_\vartheta, p_c) = \mathbb{E}_{x_{dst}\sim p_\vartheta(x_{dst})}[\log \mathbf{D}_{dis}(x_{dst})] \\ + \mathbb{E}_{x_{src}\sim p_\theta(x_{src}), c\sim p_c(c)}[\log(1 - \mathbf{D}_{dis}(\mathbf{G}(x_{src}, c)))],\end{aligned} \tag{1}$$

where p_θ and p_ϑ are sample distributions in source view θ and target view ϑ, respectively. p_c denotes the distribution of view indicator. $\mathbf{G}$ tries to generate gaits $\mathbf{G}(x_{src}, c)$ in target view ϑ conditioned on gaits x_{src} from source view θ and the corresponding view indicator c, while $\mathbf{D}_{dis}$ aims to distinguish between generated gait image $\mathbf{G}(x_{src}, c)$ and real gait image x_{dst}. $\mathbf{G}$ tries to minimize the objective while $\mathbf{D}_{dis}$ competes against it which maximizes the objective, *i.e.*, $\min_{\mathbf{G}} \max_{\mathbf{D}_{dis}} \mathcal{L}_{adv}(\mathbf{G}, \mathbf{D}_{dis}, p_\theta, p_\vartheta, p_c)$.

It is worth noting that the generated gait image cannot be mapped back to the source one. This is because the person may carry a bag or wear a coat whose style is unpredictable in source view and our aim is to synthesize gait image in target view as well as normal walking condition.

View Classification Loss. For a given gait image x_{src}, one of our aim is to translate it to another view specified by view indicator c. To fulfill the condition, an auxiliary view

classifier is added on the top of $\mathbf{D}$ as in [4] and incurs a view classification loss when iteratively optimizing $\mathbf{G}$ and $\mathbf{D}$. In specific, $\mathbf{D}$ learns to classify the real GEI x_{dst} to its corresponding view while $\mathbf{G}$ tries to synthesize gait image $\mathbf{G}(x_{src}, c)$ that can be classified to the view specified by indicator c. To optimize $\mathbf{D}$, we minimize the following loss function

$$\mathcal{L}_{cls}^{D}(\mathbf{D}_{cls}, p_{\vartheta}, p_c) = \mathbb{E}_{x_{dst}\sim p_{\vartheta}(x_{dst}), c\sim p_c(c)}[-\log \mathbf{D}_{cls}(c|x_{dst})] \tag{2}$$

where $\mathbf{D}_{cls}(c|x_{dst})$ is the probability distribution of input gait image over view indicator computed by $\mathbf{D}$. It is worth noting that $\mathbf{D_{cls}}$ is learned from real gaits in target view. When optimizing $\mathbf{G}$, generated gait image and its corresponding view indicator are used. The loss function is defined as

$$\mathcal{L}_{cls}^{G}(\mathbf{G}, p_{\theta}, p_c) = \mathbb{E}_{x_{src}\sim p_{\theta}(x_{src}), c\sim p_c(c)}[-\log \mathbf{D}_{cls}(c|\mathbf{G}(x_{src}, c))]. \tag{3}$$

By minimizing the loss function, $\mathbf{G}$ attempts to synthesize gait image that can be classified to the view specified by c.

Reconstruction Loss. To make the generated gait image similar to the real target at low frequencies, we additionally add an reconstruction loss, *i.e.*, we use an ℓ_1 loss to minimize the reconstruction error between the generated gait image and real target gait image which can be denoted as

$$\mathcal{L}_{id}(\mathbf{G}, p_{\theta}, p_c) = \mathbb{E}_{x_{src}\sim p_{\theta}(x_{src}), c\sim p_c(c)}[\|x_{dst} - \mathbf{G}(x_{src}, c)\|_1]. \tag{4}$$

It has been proved by previous works that ℓ_1 distance [5], [13], [18] helps regularize the adversarial training process, and preserve appearance consistency.

Identity-preserving Loss. Since our ultimate goal is to achieve cross-view gait recognition, identity preserving becomes essential while translate gait image in source view to target views. As analysed before, we do not only bridge view gaps between gaits in source and target view but also make them identity-distinguishable while performing gait view translation. Thus, exploring the potential and underlying identity-related information rather than only image style is significant. To achieve this, we integrate a three-stream network $\mathbf{\Phi}$ (also be called identity preserver) to regularize the generation process as shown in Fig. 2. That is, Φ learns to distinguish real gait images in target view by identity while $\mathbf{G}$ tries to generate gait images from source view that satisfy the classification criterion. Thus, we decompose the loss function into two parts: a similarity preserving loss of real gait images for $\mathbf{\Phi}$ optimization and another similarity preserving loss of fake gait images for $\mathbf{G}$ optimization. We use triplet loss [25] to train the identity preserver module, *i.e.*,

$$\mathcal{L}_{tri}(x_{anc}, x_{pos}, x_{neg}) = \max\{d(x_{anc}, x_{pos}) - d(x_{anc}, x_{neg}) + \rho, 0\} \tag{5}$$

where x_{anc}, x_{pos} and x_{neg} are three normalized embeddings which denotes anchor, positive and negative sample, respectively. In specific, the embeddings are extracted by $\mathbf{\Phi}$. $d(\cdot, \cdot)$ denotes the Euclidean distance between two input elements, and $\rho \geq 0$ represents the margin between classes in the embedding space. We set $\rho = 0.8$ empirically in our experiment. By minimizing the loss, $d(x_{anc}, x_{pos})$ tends to be 0 and $d(x_{anc}, x_{neg})$ tends to be greater than $d(x_{anc}, d_{pos})$ with a margin ρ. When the condition is fulfilled, the loss becomes 0 and no gradient would be back-propagated. When optimizing $\mathbf{\Phi}$, embeddings of real GEIs in target view are utilized, the loss function is denoted as $\mathcal{L}_{tri}^{\Phi}$. On the other hand, the anchor x_{anc} is replaced to embedding of translated GEI $\mathbf{G}(x_{src}, c)$ when optimizing $\mathbf{G}$. The loss function loss function is thus denoted as $\mathcal{L}_{tri}^{G}$. The main difference between $\mathcal{L}_{tri}^{\Phi}$ and $\mathcal{L}_{tri}^{G}$ are where the anchor embedding is from, *i.e.*, real gait images or translated gait images.

Overall Objective. As analysed, our aim is to transform gait image from one specific view to arbitrary views and meanwhile preserve the identity information. To achieve the aim, the above losses work cooperatively and the overall objective function is

$$\mathcal{L} = \mathcal{L}_{adv} + \lambda_1 \mathcal{L}_{cls}^{D} + \lambda_2 \mathcal{L}_{id} + \lambda_3 \mathcal{L}_{tri}^{G} \tag{6}$$

where $\lambda_t, t \in \{1, 2, 3\}$ are trade-off hyper-parameters which balance contribution of four losses. In practice, we set $\lambda_1 = 1$ and $\lambda_2 = \lambda_3 = 10$ in our experiments.

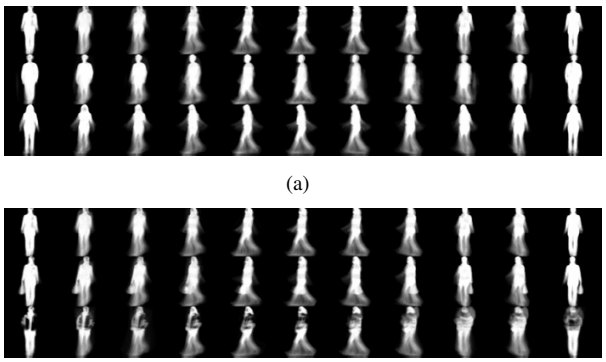
Training Strategies. As in Fig. 2, the proposed VT-GAN includes three modules, *i.e.*, identity preserver $\mathbf{\Phi}$, discriminator $\mathbf{D}$ and generator $\mathbf{G}$. In the training phase, we alternatively optimize each component. This is, we update parameters of one module while keeping parameters of the other two modules fixed. The discriminator learns to distinguish whether the input gait image is generated and classify it to its corresponding view. The similarity preserver learns to separate the input gait image by its identity. And, the generator tries to generate gaits to satisfy the requirements of both discriminator and identity preserver. We do not stop the training procedure until the objective is converged or maximum iterations are reached.

IV. Experiments

In this section, we describe details about the experimental setup, implementation settings and evaluate the proposed VT-GAN on cross-view gait recognition task.

A. Experiment Settings

Dataset. The CASIA gait database B [36] is adopted for evaluation of view transformation and cross-view gait recognition in our experiment. This dataset is widely used for gait recognition under view change because of its abundant gait sequences and view variations, *i.e.*, the database involves of 110 gait sequences from 11 views (varying from 0° to 180° with a interval of 18°) per subject and totally 124 subjects. For each view, the subject is collected 10 times of which six are in normal conditions (*i.e.*, the subject walks in tight cloth without bearing load, denoting as $NM\#01-06$), two are with a bag (denoting as $BG\#01-02$) and two are in a coat (denoting as $CL\#01-02$). It is worth to pointing out that both silhouettes of gait sequences and corresponding gait energy images (GEIs) [7] are provided in the dataset. GEI is the average image of gait silhouettes in a walking cycle which



(a)

(b)

Fig. 3. An example of GEIs from 11 views. Each row includes GEIs of the same subject from 0° to 180° with 18° interval. (a) shows GEIs of three different subjects in terms of three rows. (b) exhibits GEIs in three different conditions of the same subject, *i.e.*, NN, BG and CL.

has shown effectiveness in previous works [2], [7], [21], [30], [32]. Thus, GEI is adopted to represent gait patterns in our paper. Fig. 3 lists some examples of GEIs from 11 views in which (a) shows GEIs from different subjects in NM condition and (b) shows GEIs in different conditions from the same subject.

Setup. To compare fairly, we follow the experiment setup in gaitGAN [33]. The dataset is equally separated into two groups, one involves of first 62 subjects that is used for training the proposed VT-GAN and another includes the rest 62 subjects that is used for cross-gait recognition evaluation. When training the proposed VT-GAN, all GEIs of the 62 subjects are utilized. In the evaluation stage, four GEIs in normal walking condition from each view form the gallery set, *i.e.*, $NM\#01-04$,. And, three challenging probe sets are built, ProbeNM ($NM\#04-05$), ProbeBG ($BG\#01-02$), ProbeCL ($CL\#01-02$). It is worth noting that the proposed VT-GAN not only performs view transformation, but also normalizes other variations. Thus, GEIs in ProbeBG and ProbeCL will be normalized to standard NM condition while performing view transformation.

Evaluate Metric. As VTM-based methods [15], [16], [23] that gait features from gallery view are mapped to probe view and then conduction distance measurement, we perform cross-gait recognition the same way, *i.e.*, we first conduct view transformation by the proposed VT-GAN and then distance measurement. In our experiment, we first perform linear discriminative analysis [38] to reduce dimension and then compute Euclidean distance. Identification performance measured by nearest neighbour classifier is reported.

B. Implementation Details.

Network Architecture. There are three modules in the proposed VT-GAN, generator $\mathbf{G}$, discriminator $\mathbf{D}$ and identity preserver $\mathbf{\Phi}$. Adapted from [4], $\mathbf{G}$ includes two Conv-InstanceNorm-ReLU blocks for down-sampling, three residual blocks [8], and two TransposeConv-InstanceNorm-ReLu blocks for up-sampling. $\mathbf{D}$ follows theory of PtachGAN [13] to discriminate real/fake on local patches which is benefit to sharp the synthesized image. It consists of four Conv-LeakyReLu blocks and another two convolution layers with the stride of in parallel, where one is for real/fake discrimination and the other is for view classification. $\mathbf{\Phi}$ is adapted from [33] which includes three convolution layers together with max-pooling with 2×2 kernels and a stride of 2. In addition, 4×4 filters with stride 2 are involved in all convolution and transposed convolution layers above except residual blocks.

Training Details. The model is trained using Adam [14] with $\beta_1 = 0.5$, $\beta_2 = 0.999$ and an initial learning rate of 0.0001. We train the model for 20k iterations which the learning rate keeps fixed during first 10k iterations and then linearly decay to 0. Considering the GPU resources, the batch size is set to 100. In our experiments, we empirically set trade-off parameters in objective Eq. (6) as $\lambda_1 = 1$, $\lambda_2 = 10$ and $\lambda_3 = 10$. When training the model, we alternatively optimize generator $\mathbf{G}$, discriminator $\mathbf{D}$ and identity preserver $\mathbf{\Phi}$. In specific, we perform a single optimisation of generator $\mathbf{G}$ after every five optimisations of discriminator $\mathbf{D}$ and $\mathbf{\Phi}$.

C. Quantitative Analysis

In this section, we evaluate the proposed VT-GAN on cross-view gait recognition task and analyse the effect of identity preserving loss to recognition accuracy.

Effect of View Changes. In order to analyse how view variation affects the performance of gait recognition, we evaluate the proposed VT-GAN on three challenging probe sets, *i.e.*, ProbeNM, ProbeBG and ProbeCL, and report the result between any two views. Fig. 4 shows results on each probe view versus all views except the identical one. From the figure, it is easy to observe that view variation significantly affects the performance of gait recognition. In specific, 1) view difference degrades the identification accuracy and larger one causes worse performance, 2) performance would improve near the symmetric view in terms of centre view 90° when the probe view is fixed, 3) carrying a bag or wearing a coat also lower the recognition accuracy and wearing a coat results in worst performances.

It is reasonable for above conclusions. This is because 1) view difference incurs heterogeneous data distribution and unaligned information structure, 2) gaits from symmetric views can be mutually mirror-reflected, and 3) carrying a bag or wearing a coat severely change silhouette shape of human body and occlude the walking traits. Though the proposed VT-GAN weakens the effect of these factors, they still degrade the identification accuracy.

Performance Comparison. As in [33], we select three probe view 54°, 90° and 126° which are identical for cross-view gait recognition to compare performances with the state-of-the-art methods. In this section, we compared some representative approaches such as PCA [7], FD-VTM [19], TSVD-VTM [15], R-VTM [17], C3A [31], SPAE [34], ViDP [10], GaitGAN [33], gaitGANv2 [35] and MGAN [9]. Fig. 5 compares cross-view gait recognition performances between each probe view and other rest views. From the figure, it is clear that the results

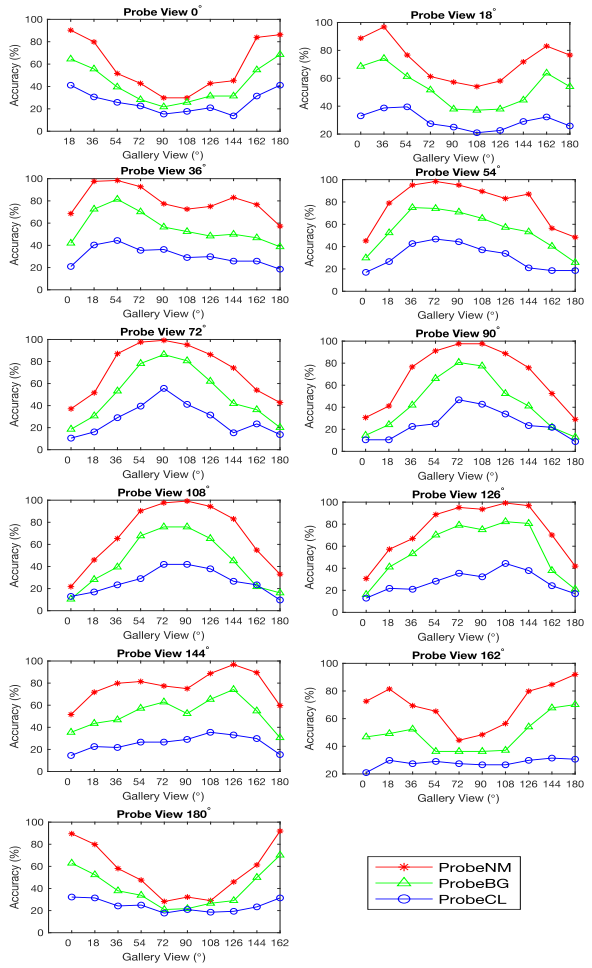


Fig. 4. Comparison of accuracies on the three probe subsets at the 11 probe views. For each probe view, accuracies of the rest views are excluded. (Best view in Acrobat PDF.)

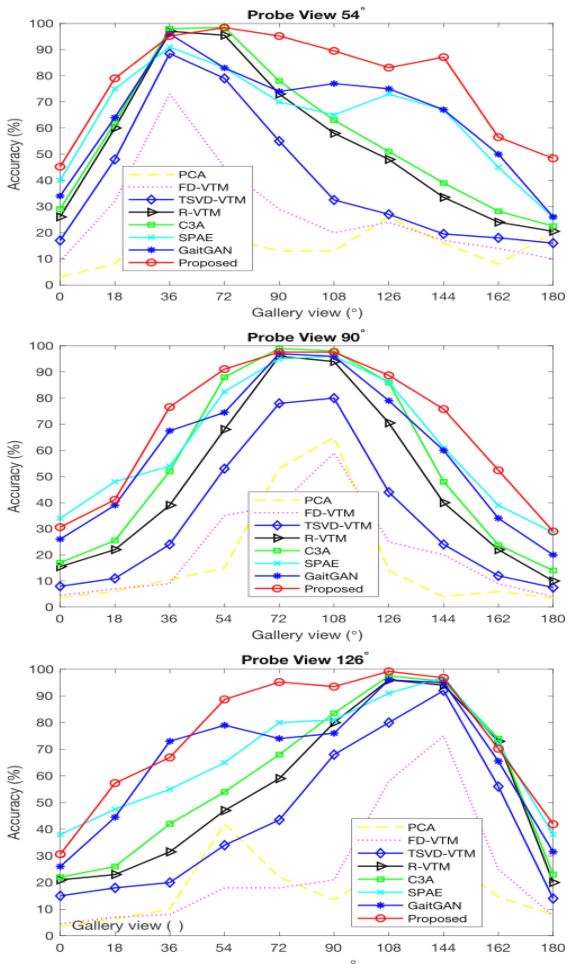


Fig. 5. Comparison with previous methods at three representative probe views on ProbeNM. (Best view in Acrobat PDF.)

support conclusions in *Effect of View Changes*. Moreover, we can observe the proposed VT-GAN improves the matching performance to some extent. Table I reports the average identification accuracies of the three probe views against the rest views. From the table, it is easy to see that the proposed VT-GAN outperforms other recent GAN-based works, *i.e.*, we improve 12%, 10% and 10% in terms of the three probe views compared with GaitGAN. Along with Fig. 5, the proposed VT-GAN achieves promising performances.

Effect of Identity-preserving Loss. In this part, we train the vanilla VT-GAN and pruned VT-GAN, respectively. Compared vanilla VT-GAN, the pruned VT-GAN does not include the identity preserver $\mathbf{\Phi}$ which is denoted as VT-GAN (*w/o* $\mathbf{\Phi}$). Table II reports results of the two models. It is easy to observe that the identity preserver does benefit to cross-view gait recognition. In addition, the VT-GAN can achieve equivalent performances even without this module, because the reconstruction loss keeps appearance similarity while performing view translation.

TABLE I
COMPARISON WITH RECENT WORKS ON CROSS-VIEW GAIT RECOGNITION ON PROBENM. AVERAGE IDENTIFICATION ACCURACIES EXCEPT THE CORRESPONDING VIEW ARE REPORTED.

Method	Probe View			
	54°	90°	126°	Average
ViDP [10]	0.64	0.60	0.65	0.63
SPAE [34]	0.63	0.62	0.66	0.64
GaitGAN [33]	0.65	0.58	0.66	0.63
MGAN [9]	**0.77**	0.67	**0.79**	**0.74**
GaitGANv2 [35]	0.72	0.65	0.73	0.70
Proposed	**0.77**	**0.68**	0.76	**0.74**

TABLE II
EVALUATION OF EFFECTIVENESS OF THE IDENTITY PRESERVER IN VT-GAN.

Method	Probe View			
	54°	90°	126°	Average
GaitGAN [33]	0.65	0.58	0.66	0.63
VT-GAN (*w/o* $\mathbf{\Phi}$)	0.75	0.64	0.73	0.71
VT-GAN	0.77	0.68	0.76	0.74

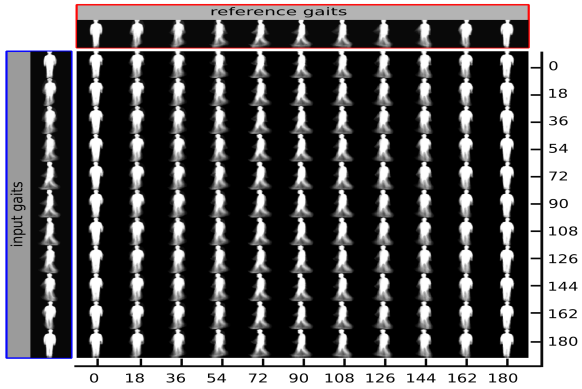


Fig. 6. Visualization of synthesized gaits between any two views. Images in blue box are input gaits, and the ones in red box are reference gaits. The rest 11×11 images are synthesized gaits conditioned on the corresponding input image and view indicator.

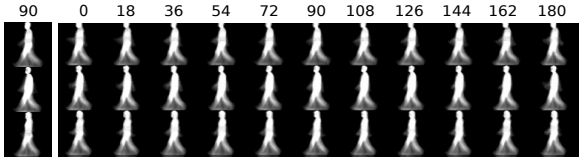


Fig. 7. Visualization of synthesized gaits of three different subject. Each row includes gaits from the same subject. And, the first column is reference gaits in $90°$, the rest rows are synthesized gaits from input gaits in $0°$ to $180°$.

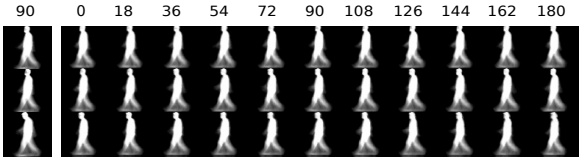


Fig. 8. Visualization of synthesized gaits of the same subject in three distinct conditions, *i.e.*, normal walking, carrying a bag and wearing a coat corresponding to three rows, respectively. The first column includes reference gaits of the same subject in $90°$ and the rest columns consist of synthesized gaits from various views.

D. Qualitative Analysis

As analysed, one benefit of the proposed VT-GAN is that it provides visual interpretation for cross-view gait recognition. As in Fig. 6, we visualize synthesized gaits between any two views. From the figure, we can see that 1) the VT-GAN synthesizes visually promising images which is hard to distinguish from real images, 2) The VT-GAN successfully achieves view transformation between any two views, 2) The generated images are appearance similar to the target gaits in gallery set. Fig. 7 shows the synthesized gaits in $90°$ from various views of three different subjects in normal walking condition. It is easy to observe that smaller view difference results in more similar artifact, *i.e.*, it is easy to generate appearance similar results when the view of probe gaits is near $90°$ in the figure. Fig. 8 shows the synthesized gaits $90°$ from various views of the same subject different walking conditions. It is obvious that carrying a bag or wearing a coat cause poorer synthesized gaits. In Fig. 9, we also lists some bad samples caused by poor

Fig. 9. An example of some bad samples and their corresponding translated gaits. For each pair, the left is the input gaits and the right is the synthesized gaits.

segmentation or limited silhouettes and their corresponding synthesized artifacts by VT-GAN. These samples undoubtedly degrade the identification accuracy. However, the proposed VT-GAN also tries to translate them and makes the generated gaits look normally. This hints that the VT-GAN can correct minor GEI faults caused by poor segmentation and incomplete walking cycle. We suggest that this is consistent to the cases such as carrying a bag or wearing a coat.

E. Further Analysis

Relation to GaitGAN. As in Fig. 1, GaitGAN aims to normalize gaits in various views to a reference one, *i.e.*, $90°$. Thus, only gaits from $90°$ are used to train the discriminator. If we want to map gaits in the identical view to arbitrary views, multiple models should be built. Different from GaitGAN, the proposed VT-GAN could translate gaits to arbitrary view using only one single model benefiting from the view indicator input. This means that view indicator constrains the generation process and force the generator to synthesize gaits in view corresponding to the specific indicator. Considering this, GaitGAN is a special case of the proposed VT-GAN, *i.e.*, we set the view indicator to $90°$. In another aspect, the proposed VT-GAN adopts distinct framework with GaitGAN because of their different goal. GaitGAN is adapted from classical GAN model while the proposed VT-GAN is derived from conditional GAN which takes view indicator as the condition. This brings out different discriminator design, *i.e.*, GaitGAN adopts traditional real/fake discriminator while the proposed VT-GAN adapts the discriminator and makes it not only distinguish the artifacts but also discriminate view of input gaits.

We also studied the influences of reference view, *i.e.*, we changed the view indicator and translated gaits of arbitrary views to each reference view. Fig. 6 compares average accuracies of all the 121 view pairs with respect to different reference views. From the figure, we can observe that side view $90°$ adopted in GaitGAN does not achieve the highest performance. In contrast, two peak identification accuracies appear at $54°$ and $126°$. The result verifies the view symmetry theory in Sec. IV-C because $54°$ and $126°$ are the center of two half sphere separated by $90°$.

Potential Benefits of VT-GAN. In addition to above benefits, the proposed VT-GAN brings other applications such as data augmentation and unseen view prediction. As we know, lack of data restricts the performance of CNN models. In gait recognition, most of existing datasets only contains limited samples with minor view variation. It is insufficient to train large-scale models for complicated scenarios. The proposed VT-GAN provides a solution to the problem because it can

synthesize gaits of arbitrary views from one specific view. Moreover, it can synthesize gaits of arbitrary views from other existing datasets which only contain minor view variation. This significantly enlarges the volume of data. In another aspect, the proposed VT-GAN can synthesize gaits of one specific subject in unseen views. It is beneficial to predict subject's identity appearing in the view at the first time.

V. Conclusion

This paper studies view transformation problem in gait recognition and propose to achieve view-to-view transformation via a single model, *i.e.*, the proposed VT-GAN. Instead of normalizing to one unified view as previous GAN-based works, VT-GAN achieves to translate gaits between any two views only using a single model. In specific, the proposed VT-GAN includes three modules, a generator, a discriminator and a similarity preserver. The generator attempts to generate gaits in reference view from source view conditioned on a view indicator and fool the real/fake discriminator while the discriminator tires to distinguish whether its input is artificial and meanwhile constrain it to its corresponding view. The similarity preserver supervises the generation process which forces to synthesize gaits sharing same identity with its positive sample in the reference view. Both qualitative and quantitative analysis are conducted on cross-view gait recognition task, the experiment results show promising performances of the proposed VT-GAN. Furthermore, the proposed VT-GAN brings other usages such as data augmentation [12] and unseen view prediction which will be researched in further study.

References

[1] X. Ben, C. Gong, P. Zhang, R. Yan, Q. Wu, and W. Meng. Coupled bilinear discriminant projection for cross-view gait recognition. *IEEE TCSVT*, 2019.

[2] X. Ben, P. Zhang, Z. Lai, R. Yan, X. Zhai, and W. Meng. A general tensor representation framework for cross-view gait recognition. *Pattern Recognition*, 90:87–98, 2019.

[3] A. F. Bobick and A. Y. Johnson. Gait recognition using static, activity-specific parameters. In *CVPR*, volume I, pages 423–430, 2001.

[4] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *CVPR*, pages 8789–8797, 2018.

[5] W. Deng, L. Zheng, G. Kang, Y. Yang, Q. Ye, and J. Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person reidentification. In *CVPR*, pages 994–1003, 2018.

[6] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, pages 2672–2680, 2014.

[7] J. Han and B. Bhanu. Individual recognition using gait energy image. *IEEE TPAMI*, (2):316–322, 2006.

[8] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[9] Y. He, J. Zhang, H. Shan, and L. Wang. Multi-task gans for view-specific feature learning in gait recognition. *IEEE TIFS*, 14(1):102–113, 2019.

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

[11] Y. Huang, H. Sheng, Y. Zheng, and Z. Xiong. Deepdiff: Learning deep difference features on human body parts for person re-identification. *Neurocomputing*, 241:191–203, 2017.

[12] Y. Huang, J. Xu, Q. Wu, Z. Zheng, Z. Zhang, and J. Zhang. Multi-pseudo regularized label for generated data in person re-identification. *IEEE TIP*, 28(3):1391–1403, 2019.

[13] P. Isola, J. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, pages 1125–1134, 2017.

[14] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[15] W. Kusakunniran, Q. Wu, H. Li, and J. Zhang. Multiple views gait recognition using view transformation model based on optimized gait energy image. In *ICCV Workshops*, pages 1058–1064, 2009.

[16] W. Kusakunniran, Q. Wu, J. Zhang, and H. Li. Support vector regression for multi-view gait recognition based on local motion feature selection. In *CVPR*, pages 974–981, 2010.

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

[18] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, and L. Van Gool. Pose guided person image generation. In *NIPS*, pages 406–416, 2017.

[19] Y. Makihara, R. Sagawa, Y. Mukaigawa, T. Echigo, and Y. Yagi. Gait recognition using a view transformation model in the frequency domain. In *ECCV*, pages 151–163, 2006.

[20] Y. Makihara, A. Suzuki, D. Muramatsu, X. Li, and Y. Yagi. Joint intensity and spatial metric learning for robust gait recognition. In *CVPR*, pages 5705–5715, 2017.

[21] A. Mansur, Y. Makihara, D. Muramatsu, and Y. Yagi. Cross-view gait recognition using view-dependent discriminative analysis. In *IJCB*, 2014.

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

[23] D. Muramatsu, Y. Makihara, and Y. Yagi. View transformation model incorporating quality measures for cross-view gait recognition. *IEEE TC*, 46(7):1602–1615, 2016.

[24] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text to image synthesis. *arXiv preprint arXiv:1605.05396*, 2016.

[25] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, pages 815–823, 2015.

[26] K. Shiraga, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi. Geinet: View-invariant gait recognition using a convolutional neural network. In *IJB*, 2016.

[27] Z. Shu, E. Yumer, S. Hadap, K. Sunkavalli, E. Shechtman, and D. Samaras. Neural face editing with intrinsic image disentangling. In *CVPR*, pages 5444–5453, 2017.

[28] S. Singh and K. Biswas. Biometric gait recognition with carrying and clothing variants. In *ICPRMI*, pages 446–451, 2009.

[29] R. Tanawongsuwan and A. Bobick. Modelling the effects of walking speed on appearance-based gait recognition. In *CVPR*, volume 2, pages II–II, 2004.

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

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

[32] L. Yao, W. Kusakunniran, Q. Wu, J. Zhang, and Z. Tang. Robust gait recognition under unconstrained environments using hybrid descriptions. In *DICTA*, pages 1–7, 2017.

[33] S. Yu, H. Chen, E. B. G. Reyes, and N. Poh. Gaitgan: Invariant gait feature extraction using generative adversarial networks. In *CVPR Workshops*, pages 532–539, 2017.

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

[35] S. Yu, R. Liao, A. Weizhi, C. Haifeng, E. B. García, Y. Huang, and N. Poh. Gaitganv2: Invariant gait feature extraction using generative adversarial networks. *Pattern Recognition*, 87:179–189, 2019.

[36] S. Yu, D. Tan, and T. Tan. A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In *ICPR*, volume 4, pages 441–444, 2006.

[37] G. Zhao, G. Liu, H. Li, and M. Pietikainen. 3d gait recognition using multiple cameras. In *FGR*, pages 529–534, 2006.

[38] W. Zhao, A. Krishnaswamy, R. Chellappa, D. L. Swets, and J. Weng. Discriminant analysis of principal components for face recognition. In *Face Recognition*, pages 73–85. 1998.

[39] S. Zheng, J. Zhang, K. Huang, R. He, and T. Tan. Robust view transformation model for gait recognition. In *ICIP*, pages 2073–2076, 2011.