VN-GAN: Identity-preserved Variation Normalizing GAN for Gait Recognition

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*—Gait is recognized as a unique biometric characteristic to identify a walking person remotely across surveillance networks. However, the performance of gait recognition severely suffers challenges from view angle diversity. To address the problem, an identity-preserved Variation Normalizing Generative Adversarial Network (VN-GAN) is proposed for learning purely identity-related representations. It adopts a coarse-to-fine manner which firstly generates initial coarse images by normalizing view to an identical one and then refines the coarse images by injecting identity-related information. In specific, Siamese structure with discriminators for both camera view angles and human identities is utilized to achieve variation normalization and identity preservation of two stages, respectively. In addition to discriminators, reconstruction loss and identity-preserving loss are integrated, which forces the generated images to be the same in view and to be discriminative in identity. This ensures to generate identity-related images in an identical view of good visual effect for gait recognition. Extensive experiments on benchmark datasets demonstrate that the proposed VN-GAN can generate visually interpretable results and achieve promising performance for gait recognition.**

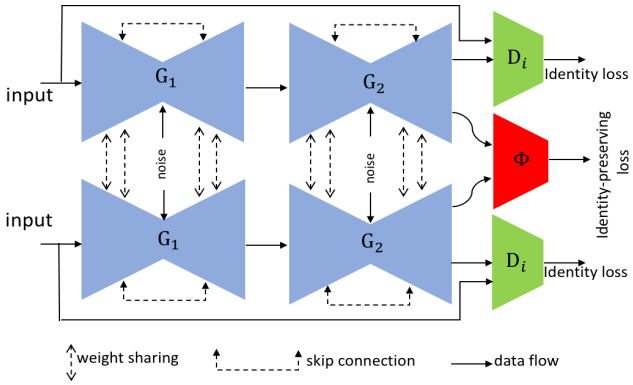

Fig. 1. Overview of Siamese structure of the proposed VN-GAN. It includes two-stream networks where each stream consists of the proposed coarse-to-fine design in Fig. 2. Noting that the variation discriminator $\mathbf{D}_v$ and variation classifier $\mathbf{P}$ in Stage-I are not shown in the figure, because the figure only aims to show the Siamese design. Details of each branch is illustrated in Fig. 2.

I. Introduction

In the field of surveillance, gait is regarded as one of the most potential biometric features because of its characteristic that gait is collected remotely without target subject's cooperation. This brings many benefits such as contactless and hard to disguise, which is identical to other biometrics, *i.e.*, face, iris, fingerprint, *etc*. Since the irreplaceable advantages, gait recognition attracted widely research interests and rendered lots of research outcomes in the past decade [4], [14], [20], [23], [26], [36]. Unfortunately, gait recognition retains challengeable in practical scenarios. This is because gait recognition suffers great challenges incurred by uncontrolled surveillance scenarios such as illumination variation, view difference, change of cloth, various types of carrying conditions and so on. Among them, view difference [7], [23], [36] has been treated as the most typical yet challenging issue because of random walking direction of a pedestrian, which seriously restricts the performance of gait recognition. In this paper, we mainly tackle the view variation issue and then extend to other challenges. To solve the issue, researchers proposed many approaches which can be categorized into two aspects, *i.e.,* feature-based methods and learning-based methods. The former focuses on describing gaits with view invariant patterns such as static body parameters [7], 3D gait model and motion trajectory estimation [2]. These methods either suffer high computational cost or limited performance, which is mostly studied in early literatures. In contrast, the latter concentrates on learning the latent relationship between gaits across views which is popular in recent researches. Two representative frameworks are view transformation model (VTM) [20], [21], [26], [28] and coupled subspace learning (CSL) [4]–[6], [35]. VTM aims to transform gaits from one view to another view by learning a regressor while CSL tries to map gaits from different views into a latent space in which view-specific feature is suppressed. Learning-based methods achieve great success for cross-view gait recognition which significantly improves the identification accuracy. However, these methods heavily depend on accurate view estimation and require to learn mutually independent models for each pair of views which incurs abundant models and causes great computation costs. Recently, some works attempt to extract view invariant features using a single model such as ViDP [15] and CNN [34] which achieve promising performances. However, the learned features are hard to explain because of black-box characteristics of neural network. With the rise of

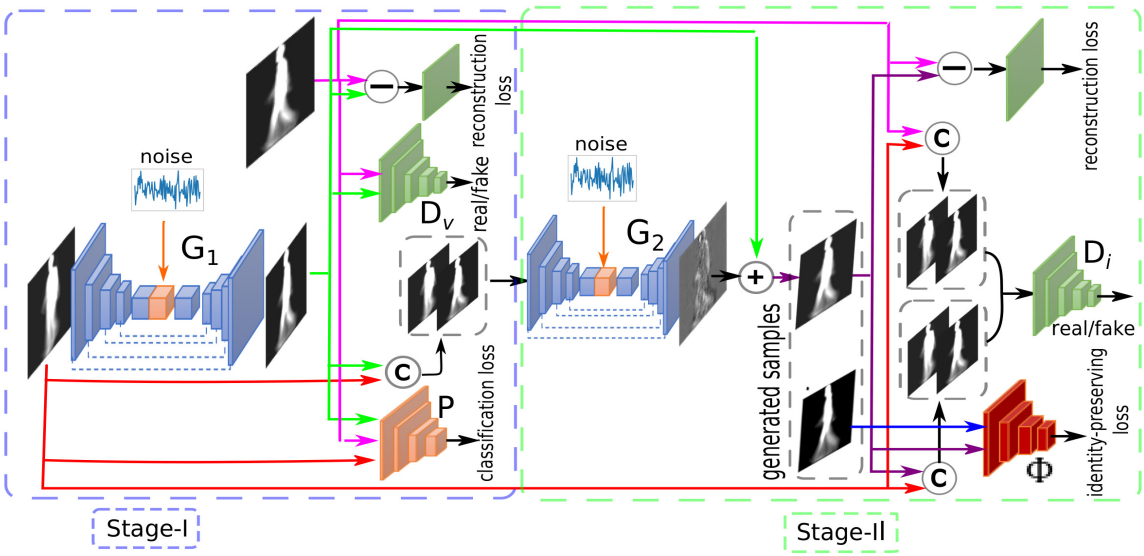

Fig. 2. Overview of the coarse-to-fine structure. It is sub-branch of the proposed VN-GAN which consists of two stages: coarse gait image generation and refinement. In the framework, ⓒ denotes channel concatenation and ⊖ denotes image difference, ⊕ denotes image summation.

GAN, it is available to normalize gaits to an identical view such as gaitGAN [36] and MGANs [14]. Moreover, this GAN-based methods not only achieve competitive performances but also generate visually realistic results that makes the model more interpretable. Therefore, we propose a two-stage framework based on GAN, which normalizes gait templates, *i.e.*, gait energy images (GEIs), from arbitrary views to a specific one such as 90° in a coarse-to-fine manner.

As in Fig. 2, the proposed framework includes two stages, *i.e.*, coarse generation which generates GEIs in a target view guided by a variation discriminator and a variation classifier, and image refinement which refines results in the previous stage and generate differences between results in Stage-I and target images. As illustrated in [25], [40], the coarse-to-fine design makes the task easier since each stage simply focuses on one task, *i.e.*, view transformation and identity injecting. At Stage-I, we attempt to transform gait templates from various views into a identical one. To achieve the goal, one variation discriminator and one variation classifier are integrated. The variation discriminator competes with the generator and forces it to synthesize images by following the same statistic distribution of real gaits on the target view, and the variation classifier regularizes the generating process by classifying the generated gait templates to its view classes. At Stage-II, we tries to embed identity information to the coarse results in Stage-I. In this stage, we utilize identity discriminator [11] instead of common real/fake discriminator [12] because we aim to distinguish whether the normalized gait and the input gait belong to the same person. Meanwhile, identity classifier is integrated during training which constrains the generator and forces the generated gait in target view belong to the class of input gait. By combining the two stages, the input gait from arbitrary views are normalized to the one of same identity with input but from an identical view.

In another aspect, the proposed VN-GAN adopts the popular Siamese structure as shown in Fig. 1, which introduces an identity preserving regularization for Stage-II. The former ensures the generated gait of the same person in the same view, and the latter pulls generated gaits of the same person close and pushes generated gaits of different persons away. In this manner, the generated gaits are discriminative for gait recognition. In each stage, we iteratively optimize view/identity discriminator, view/identity classifier and generator. By coordinating view-related losses in Stage-I and identity-related losses in Stage-II, the proposed VN-GAN is capable to generate view-normalized and identity-preserved gait templates. It is worth noting that the normalization process is also effective to other variations such as presence of or different types of a bag or wearing a coat.

To summarize, the contributions of this paper are i) We propose a novel framework, *i.e.*, VN-GAN, to achieve identity-preserved gait normalization for variation invariant gait recognition. Unlike existing methods, we separate the task into two stages which achieves view (or other variations) normalization and identity injecting, respectively. The coarse-to-fine design makes identity-preserved gait normalization easier. ii) We integrate the coarse-to-fine design into Siamese structure which achieves not only competitive gait recognition performance but also provides normalized gaits in visual quality that makes cross-view gait recognition visually interpretable.

II. Related Works

In this section, we first make a brief survey on gait recognition, especially the cases that are robust to view changes. Then, we introduce some GAN variations that are closely related to our method.

Gait Recognition. Approaches for gait recognition that are robust to factors such as view changing, cloth difference and carrying conditions can be roughly segmented into two groups, *i.e.*, feature-based methods and learning-based methods. The former tries to represent gait using view invariant traits. For instance, approaches in [7], [18] tried to extract view invariant features by analysing static body parameters or activity-specific (*i.e.*, stride) parameters from gait images. However, these methods suffer poor performances. Ariyanto *et al.* [2] proposed to model human lower legs using a structural 3D model and then extracting kinematics trajectories. Approaches of this kind improve the performance. Unfortunately, they are usually impractical due to the fact that 3D reconstruction requires multiple calibrated cameras. The latter attempts to correlate gaits from distinct views or learn distinctive features by data-driven manner. Representative models of this category are VTM, CSL and CNN. VTM acts like a regressor, which tried to correlate gaits from two distinct views by learning mappings between them, *e.g.*, [20], [21], [23], [26], [28]. Distinctly, CSL tries to mitigate view bias by asymmetrically mapping gaits from different views into a shared subspace such as [3], [5], [6], [35]. However, VTM and CSL are heavily influenced by view estimation and require to train for each pair of views. With the rise of deep learning, CNN-based models achieves promising performances such as [32], [34]. These methods usually learn identity discriminative features using a neural network in a data-driven manner. Unfortunately, it is hard to be interpreted the learned features due to the black-box characteristics of CNN. To overcome the issue, Yu *et al.* [36] and He *et al.* [14] proposed to normalize gaits across views by GAN because of its strong style transfer ability. These methods achieve competitive performances. And notably, these approaches also provide good visualization by synthesizing images. Because of the benefits, we build our methods based on GAN theory.

Generative Adversarial Networks. Generative adversarial networks [12] shows potential applications in computer vision since it is proposed, *e.g.*, image generation [12], [16], [29], [30], style transfer [25], [33], [39], [41], super-resolution reconstruction [24], face editing [8], [9], *etc*. Generally, it includes a generator and a discriminator where the generator tires to synthesize fake images to fool the discriminator. Many GAN variations are developed, especially in style transfer or image translation which are closely to our methods. For instance, Isola *et al*. [17] achieved image to image translation by pixel level mapping using a conditional GAN. Zhu *et al.* [41] proposed to integrate cycle-consistent loss to GAN framework and thus translate images from different distributions unsupervisedly. Ma *et al.* [25] proposed to synthesize person in specific pose using GAN by a divide-and-conquer strategy. Deng *et al*. [10] utilized cycleGAN to perform identity-preserved domain transfer and then achieve cross-dataset person re-identification. In gait recognition, GaitGAN [36] and MGANs [14] are two precursory works. GaitGAN normalizes gaits in all views conditioned on a real/fake discriminator and an identification discriminator. MGANs achieves view transformation by adding a view transformation layer. Both of them achieve gait view normalization using a classical GAN model. Different from them, the proposed approach 1) adopts a coarse-to-fine manner which makes identity-preserved gait normalization easier and 2) integrates the coarse-to-fine design to the popular Siamese structure which synthesizes more discriminative images for gait recognition.

III. Variation Normalizing Generative Adversarial Network

Our task is to reduce the effect of view variation for gait recognition by normalizing gait images to a referential condition, *i.e.*, the observed subject walks in tight-fitting clothes without carrying objects and is collected in a side view. This requires to preserve identity while performing view normalization. To make the task easier, we adopt a divide-and-conquer strategy which generates coarse view-normalized gaits in Stage-I and implants identity information in Stage-II. In addition, we integrate the two-stage design into the Siamese structure, which forces the synthesized gaits discriminative. The overall framework of the proposed VN-GAN and the two-stage design are shown in Fig. 1 and Fig. 2, respectively.

A. Stage-I: Variation Normalization

At Stage-I, we aims to normalize gaits from arbitrary views or other variations to a standard condition. As in Fig. 2, this stage includes a generator $\mathbf{G_1}$, a variation discriminator $\mathbf{D}_v$ and a view preserver $\mathbf{P}$.

Generator $\mathbf{G_1}$. we utilize a U-Net-like architecture [25], [41] to generate coarse gaits in target view. Specially, we first use several stacked convolutional layers to extract feature embeddings in bottleneck from source images x_{src} in arbitrary views. To simulate variations among intra-class samples, we integrate a random noise z in the bottleneck layer. After that, the feature embedding is decoded using a set of stacked deconvolutional layers which is symmetric to the encoder. For sake of information loss, skip connections between encoder and decoder are utilized in the U-Net to propagate gait information directly from source to target. The generated coarse gait in Stage-I is denoted as $x_{dst}^{(1)}$ or $\mathbf{G}_1(x_{src}, z)$.

Variation Discriminator $\mathbf{D_v}$. In GANs, discriminator competes against generator to distinguish whether the input is generated or not. In Stage-I, we adopt a variation discriminator $\mathbf{D}_1$ that forces the generated gaits $x_{dst}^{(1)}$ obey the distribution of target samples $\mathbf{x}_{dst}$ of the same subject in reference view. That is, $\mathbf{D}_v$ takes the synthesize image $x_{dst}^{(1)}$ and real image $\mathbf{x}_{dst}$ in reference view as input, and discriminates whether the input is sampled from distribution of the reference view.

Variation Classifier $\mathbf{P}$. To ensure view consistency between generated image $x_{dst}^{(1)}$ and target image x_{dst}, we regularize the generation process by a variation classifier $\mathbf{P}$. It classifies the generated gaits and real gaits in reference view into the same class and distinguishes real gaits from other views rather than the reference one to another classes. In practice, the variation classifier $\mathbf{P}$ can be pre-trained beforehand since it is only related to real samples. However, we integrate

it to our proposed framework and train it alternatively with $\mathbf{G}_1$ and $\mathbf{D}_v$ following a end-to-end manner. When training the generator $\mathbf{G}_1$, gradients from $\mathbf{P}$ are back-propagated to jointly optimize $\mathbf{G}_1$ with generative loss, and regularize to generate gaits on the reference view, *i.e.*, the walking person dresses tightly and walks along a side view.

To force the synthesized gaits near the ground truth, we impose l_1 distance between generated image $x_{dst}^{(1)}$ and target images x_{dst} in generation process, which is defined as

$$\ell_{id-1}(x_{src}, x_{dst}) = \|x_{dst}^{(1)} - x_{dst}\|_1 \tag{1}$$

Therefore, objective functions for generator $\mathbf{G}_1$, discriminator $\mathbf{D}_v$ and variation discriminator $\mathbf{P}$ in Stage-I are defined respectively,

$$\begin{aligned}
\ell_{G_1} &= \ell_{bce}(\mathbf{D}_v(\mathbf{G}_1(x_{src}, z), 1)) + \alpha\ell_{id-1}(\mathbf{G}_1(x_{src}, z), x_{dst}) \\
&\quad + \beta\ell_{bce}(\mathbf{P}(\mathbf{G}_1(x_{src}, z)), 1), \\
\ell_{D_v} &= \ell_{bce}(\mathbf{D}_v(\mathbf{G}_1(x_{src}, z)), 0) + \ell_{bce}(\mathbf{D}_v(x_{dst}), 1), \\
\ell_P &= \ell_{bce}(\mathbf{P}(x_{src}), 0) + \ell_{bce}(\mathbf{P}(x_{dst}), 1),
\end{aligned} \tag{2}$$

where ℓ_{bce} denotes binary cross-entropy loss, α and β are trade-off parameters of ℓ_{id} loss and pose regularization loss ℓ_P, respectively. α controls the level of similarity between generated artifacts and real targets at low frequencies. Large α incurs blurry results since ℓ_1 loss encourages to average all possible cases, while small α causes artifacts due to domination of adversarial loss at training phase [25]. β determines the contribution of view regularization, which also averages all possible cases and blurs results specified by target view if it is too large.

In Stage-I, the output of $\mathbf{G}_1$ captures the global structure information specified by the target view, as shown in Fig. 2, which does normalize the source gaits from arbitrary views into the specific one. Though Stage-I encourages the appearance between generated gait and the real target looks similar, it cannot guarantee the identity discriminability of the generated gaits under the target distribution. Thus, we refine the result in Stage-II using another GAN by imposing a contrastive loss.

B. Stage-II: Identity Implantation

Since Stage-I has synthesized appearance similar results with real ground truth in both global structure and view angle, Stage-II will target on making the generated results discriminative on target distribution, which pulls the generated gaits of the same identity closer while pushes that of different identities further.

Generator $\mathbf{G}_2$. Similar to [25], we also adopt a variant of DCGAN [30] conditioned on the output of Stage-I as our base model. Considering that the initial output of Stage-I is globally similar to the target, we propose to simultaneously narrow their gaps and incorporate discriminative information by generating appearance difference map. In particular, we use similar U-Net architecture as in Stage-I but different input and regularization terms to generate the difference map. It means that we input the concatenation of source image x_{src} and initial output $x_{dst}^{(1)}$ in Stage-I to $\mathbf{G}_2$ rather than x_{src}. This speeds up model training and reduces computation cost since we learns the difference map from initial reasonable results rather than from scratch.

Identity Discriminator $\mathbf{D}_i$. Different from Stage-I, we do not directly output the target result but difference map which refines the coarse result $x_{dst}^{(1)}$. Thus, a identity discriminator $\mathbf{D}_i$ which discriminate pairs' artifact, *i.e.*, fake pair $(x_{dst}^{(2)}, x_{src})$ v.s. real pair (x_{dst}, x_{src}), are utilized as shown in Fig. 2. This makes $\mathbf{D}_i$ not only recognize the difference between synthesized and natural images but also the identity distinction between $x_{dst}^{(2)}$ and x_{dst}. In another words, the pairwise input encourages $\mathbf{G}_2$ to synthesize identity-preserving result with target. This is beneficial to recognition tasks.

Identity Preserver $\mathbf{\Phi}$. To keep identity discriminability of synthesized gaits, we integrate an identity preserver $\mathbf{\Phi}$ in Stage-II. Benefits of the Siamese structure, we impose a identity-preserving loss to end of $\mathbf{\Phi}$ which pulls the synthesized gaits of the same subject together and push the synthesized gaits of different subjects away. In this work, we adopt the contrastive loss which is defined as

$$\ell_\Phi(x_1, x_2, y) = \frac{1}{2}\{yd^2 + (1-y)[max(0, \rho - d)]^2\}, \tag{3}$$

where d denotes the Euclidean distance between normalized input embeddings of two generated gaits x_1 and x_2 from two-branch of the Siamese structure, which is defined as $d(x_1, x_2) = \|\Phi(x_1) - \Phi(x_2)\|_2$. And, y denotes the binary label of the input pair which indicates x_1 and x_2 are positive pair if $y = 1$ and 0 otherwise. $\rho \geq 0$ is the margin that determines the separability of generated samples in the embedding space defined by Φ. If $\rho = 0$, only gradient of positive pairs will be back-propagated and thus it dominates the training process. While $\rho > 0$, both negative and positive pairs are taken into account.

Reconstruction loss. As stated in [11], [17], it is helpful to reduce blurriness and generate human-perceivable images if regularising generation process by ℓ_1 distance between fake and real images. Therefore, we introduce a reconstruction loss by minimizing the ℓ_1 distance between the synthesized gaits $x_{dst}^{(2)}$ in Stage-II and their corresponding real targets x_{dst},

$$\ell_{id-2}(x_{src}, x_{dst}) = \mathbb{E}_{x_{src} \sim p_{data}(x_{src})}\|x_{dst}^{(2)} - x_{dst}\|_1 \tag{4}$$

where $x_{dst}^{(2)}$ represents output of Stage-II which is the sum of initial result in Stage-I and the generated difference map in Stage-II, denoting as $x_{dst}^{(2)} = x_{dst}^{(1)} + \mathbf{G}_i([x_{src}; x_{dst}^{(1)}], z)$, and $[\cdot;\cdot]$ represents tensor concatenation.

As mentioned above, losses on generator, identity discriminator and identity preserver work collaboratively to generate identity-preserved gaits in Stage-II. Thus, losses for $\mathbf{G}_2$ and $\mathbf{D}_i$ are defined respectively as

$$\begin{aligned}
\ell_{G_2} &= \ell_{bce}(\mathbf{D}_i([x_{src}; x_{dst}^{(2)}]), 1) + \alpha\ell_{id-2} + \gamma\ell_\Phi \\
\ell_{D_i} &= \ell_{bce}(\mathbf{D}_i([x_{src}; x_{dst}]), 1) + \ell_{bce}(\mathbf{D}_i([x_{src}; x_{dst}^{(2)}]), 0)
\end{aligned} \tag{5}$$

where α and γ are trade-off parameters which balance the contribution of reconstruction loss and identity-preserving loss. In practice, we alternatively optimize generator $\mathbf{G}_2$, identity discriminator $\mathbf{D}_i$ and identity preserver $\mathbf{\Phi}$ by minimizing losses ℓ_{G_2}, ℓ_{D_i} and ℓ_{Φ}, respectively.

C. Network Architecture

In this section, we summarize architecture of the proposed VN-GAN. In both stages, we adopt the U-net structure [31] with skipped connections to construct our generators, *i.e.*, each generator includes an encoder of N Convolution-InstanceNorm-LeakyReLu blocks to down-sample the input images into bottleneck and a decoder of N Deconvolution-InstanceNorm-ReLu blocks to up-sample to images from reference distribution. In our experiment, we set $N = 4$. It worth noting that random noise is imposed to concatenate with features from encoder in the bottleneck and an additional Convolution-InstanceNorm is performed before the concatenated feature transmitting to decoder. In addition, skip connections between encoders and decoders are applied which can be seen in Fig. 1 and Fig. 2. For discriminators in both stages, we utilize the structure of PatchGAN [17] which models high-frequencies of image. It is critical to prevent much more smoothness caused by reconstruction loss. At Stage-I, variation classifier $\mathbf{P}$ consists of one Convolution-ReLu layers, three Convolution-BatchNorm-LeakyReLU layers and one convolution layer. At Stage-II, identity preserver $\mathbf{\Phi}$ shares the same structure with $\mathbf{P}$ except an additional fully connected layer. In the framework, all convolution/deconvolution layers contain 4×4 filters and number of filters are doubled/halved with each block except last convolution layers which are 1 for both variation discriminator and variation classifier. In addition, strides are set to 2 for all convolution layers except last layer of discriminator, variation classifier and identity preserver, which are set 1, 4 and 4, respectively.

IV. Experiments

In this section, we evaluate the proposed VN-GAN on widely-used gait database, *i.e.*, CASIA(B) gait dataset [38], against various influencing factors such as view differences, changes of carry conditions and clothing variation.

A. Experiment Setup

Dataset. CASIA(B) gait dataset includes about 13640 sequences of 124 subjects from 11 views, *i.e.*, $0^\circ, 18^\circ, \ldots, 180^\circ$ with 18° interval. For each view, there are 10 sequences of which six are collected under normal walking (*NM01-NM06*), two are captured under walking with a bag (*BG01-BG02*) and the rest two are taken under walking with a coat (*CL01-CL02*). Noting that the popular spatio-temporal template for gait, *i.e.*, gait energy image (GEI) [13], formed by aligned silhouettes in a walking cycle has been provided with the original videos. It has been proved noise robustness and computation efficiency by previous works [6], [13], [23], [34]. Fig. 3 lists an example of GEIs from 11 views of different people, carry conditions and cloth. Therefore, we directly use the provided GEIs as our input to perform gait recognition.

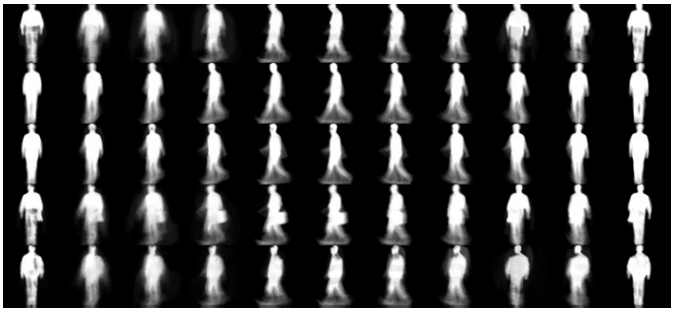

Fig. 3. An example of GEIs from 11 views, *i.e.*, $0^\circ, 18^\circ, \cdots, 180^\circ$ from left to right, and images at each column stand for GEIs from the same view angle. The first three rows compares GEIs from three different people in normal walking pattern. The last two rows includes the same subject with the first row but different walking conditions, *i.e.*, the subject carries a bag in the forth row and wear a coat in the last row.

Settings. In all experiments, we follow the setting in [36], [37] which splits the first 62 people to form training set and the rest 62 people to form testing set. When performing training, all samples of the first 62 people are used to normalize the GEIs from arbitrary views into a reference one, *i.e.*, 90°. In testing phase, four of normal walking sequences (Gallery: *NM01 – NM04*) are taken as gallery set. And, three challenging probe sets against different influencing factors are built, *i.e.* the rest two normal walking sequences (*i.e.*, ProbeNM: *NM05, NM06*), two walking with a bag (*i.e.*, ProbeBG: BG01, BG02), two walking with a coat (*i.e.*, ProbeCL: *CL01, CL02*). In our experiments, all GEIs are resized into 64×64 pixels.

Training Details. We use Tensorflow [1] to train the proposed VN-GAN on the training set. In particular, we utilize Adam optimizer [19] with parameters $\beta_1 = 0.5, \beta_2 = 0.999$ and set initial learning rate to 0.0002. We respectively train each stage with a mini-batch of size 100 for $5k$ iterations. Empirically, the balance parameters are set $\alpha = 10, \beta = 5, \gamma = 10$. For the contrastive loss, we empirically set margin $\rho = 0.8$. In the training process of Stage-I, we alternatively optimize Generator $\mathbf{G}_1$, discriminator $\mathbf{D}_v$ and variation classifier $\mathbf{P}$. However, the variation classifier $\mathbf{P}$ can also be pretrained ahead our model training. When the training of Stage-I is finished, we fixed the variables in Stage-I and alternatively optimize generator $\mathbf{G}_2$, identity discriminator $\mathbf{D}_i$ and identity preserver $\mathbf{\Phi}$. As in Fig. 2, Stage-II takes output of Stage-I as input and generates difference maps, which refines the results of Stage-I.

Evaluation Metrics. Nearest Neighbour Classifier (NNC) based on Euclidean distance is adopted to evaluate performances of the proposed VN-GAN. Before performing distance measurement, linear discriminant analysis [27] is firstly utilized to pre-process the generated GEIs. Recognition accuracies at rank-1 place are reported on all the probes.

TABLE I
IDENTIFICATION ACCURACY ON PROBENM SUBSET.

Acc. (%)		Probe View										
		0	18	36	54	72	90	108	126	144	162	180
Gallery View	0	99.2	86.3	66.1	58.9	40.3	41.1	46.0	54.8	52.4	72.6	81.5
	18	92.7	100.0	95.2	80.6	66.1	62.9	68.5	62.9	67.7	78.2	66.1
	36	73.4	92.7	96.8	96.0	83.9	75.8	78.2	79.0	78.2	69.4	56.5
	54	49.2	76.6	92.7	97.6	92.7	86.3	89.5	81.5	82.3	64.5	46.8
	72	46.8	63.7	84.7	91.1	97.6	96.0	92.7	87.1	68.5	58.9	41.9
	90	41.9	56.5	80.6	87.1	96.0	96.0	95.2	88.7	78.2	56.5	37.9
	108	38.7	57.3	71.8	84.7	93.5	95.2	96.8	96.8	87.9	62.1	39.5
	126	44.4	59.7	75.0	83.9	87.9	87.1	93.5	96.8	94.4	76.6	50.0
	144	49.2	54.8	71.0	80.6	79.8	75.8	89.5	96.0	99.2	86.3	51.6
	162	68.5	75.0	68.5	62.1	51.6	54.8	62.9	73.4	85.5	98.4	84.7
	180	83.9	63.7	53.2	42.7	37.1	37.1	40.3	51.6	53.2	83.9	98.4

TABLE II
IDENTIFICATION ACCURACY ON PROBEBG SUBSET.

Acc. (%)		Probe View										
		0	18	36	54	72	90	108	126	144	162	180
Gallery View	0	74.2	61.3	50.0	32.3	32.3	21.0	22.6	28.2	37.1	48.4	60.5
	18	63.7	81.5	72.6	53.2	46.0	32.3	33.9	35.5	42.7	51.6	41.9
	36	48.4	73.4	83.9	69.4	54.0	47.6	37.9	47.6	50.8	47.6	29.0
	54	36.3	52.4	71.8	78.2	68.5	53.2	56.5	68.5	49.2	41.9	25.0
	72	32.3	48.4	62.1	67.7	79.8	66.1	62.1	69.4	50.8	39.5	21.8
	90	26.6	38.7	58.9	61.3	75.0	71.0	68.5	66.9	52.4	37.1	24.2
	108	28.2	45.2	59.7	62.9	75.0	71.8	75.0	71.8	53.2	40.3	21.0
	126	32.3	46.0	51.6	59.7	58.9	54.0	64.5	79.0	65.3	50.8	20.2
	144	33.1	47.6	50.0	49.2	53.2	43.5	55.6	71.0	74.2	54.0	32.3
	162	46.0	54.0	45.2	37.9	32.3	31.5	31.5	43.5	49.2	69.4	57.3
	180	58.1	52.4	42.7	25.8	28.2	21.0	19.4	31.5	45.2	52.4	74.2

B. Evaluation

In this section, we evaluate performances of the proposed VN-GAN on three challenging probe sets, *i.e.*, ProbeNM, ProbeBG and ProbeCL, which covers variations against view, clothing and carrying conditions.

Effects of Influencing Factors. We measure recognition accuracies between each view pair in this section. The results are reported in Table I – Table III in terms of three influencing factors in CASIA(B) gait dataset. There are totally 121 pairs of view combination in the table and we measured recognition accuracy across views of each pair. Table I lists the results on ProbeNM, *i.e.*, for each view pair, we take *NM01-04* in one view as gallery and *NM05-06* in the other view as query. It responds to the normal walking conditions. Table II and Table III report the results on ProbeBG and ProbeCL, respectively, which use the same gallery set but different probe set compared to Table I. From the three tables, it is easy to observe that recognition accuracy is severely influenced by view changes. In specific, larger view difference incurs poorer performance. However, the performance increases at the symmetrical view such as 54° *vs*. 126°. Thus, there two recognition accuracy peaks except the case that probe view is 90° because 90° is in the middle. Fig. 4 also verifies the conclusion.

In another aspect, we can observe that carry condition and wearing a coat also degrade the recognition accuracies drastically. And, wearing a coat causes worst performance among all the mentioned variations. This is reasonable because 1) wearing a coat changes the shape of human silhouettes, and 2) wearing a coat also occludes human body which causes incomplete motion patterns. This conclusion is also supported by Fig. 4.

Comparison with The State-of-the-art. In this part, we compare performances with some state-of-the-art methods on cross-view gait recognition, *i.e.*, GEI+PCA [13], FD-VTM [26], TSVD-VTM [20], R-VTM [22], C3A [35], SPAE [37], GaitGAN [36], ViDP [15] and MGAN [14]. Typically, we select three probe views, *i.e.*, 54°, 90° and 126°. Fig. 5 compares the performances of each probe angle against the rest view angles. From the figure, we can observe that view changes significantly affect recognition accuracy and larger view difference causes worse performance. This is consistent

TABLE III
IDENTIFICATION ACCURACY ON PROBECL SUBSET.

Acc. (%)		Probe View										
		0	18	36	54	72	90	108	126	144	162	180
Gallery View	0	31.5	25.8	22.6	15.3	11.3	12.9	16.1	16.1	22.6	17.7	21.8
	18	28.2	32.3	32.3	24.2	22.6	21.8	20.2	24.2	18.5	25.0	15.3
	36	23.4	33.9	39.5	39.5	37.1	29.8	29.0	30.6	28.2	26.6	13.7
	54	20.2	27.4	32.3	39.5	36.3	35.5	32.3	29.0	30.6	23.4	12.9
	72	17.7	26.6	31.5	32.3	50.8	40.3	36.3	32.3	29.8	26.6	12.9
	90	19.4	27.4	32.3	41.9	50.8	52.4	41.9	41.1	25.8	23.4	17.7
	108	22.6	25.0	29.8	37.1	45.2	41.1	46.8	37.1	30.6	29.8	17.7
	126	25.0	26.6	31.5	33.1	31.5	31.5	33.9	33.9	34.7	27.4	14.5
	144	28.2	25.8	25.0	22.6	25.0	25.0	35.5	36.3	42.7	36.3	18.5
	162	25.0	23.4	22.6	16.1	21.0	16.9	20.2	22.6	31.5	33.9	25.0
	180	21.8	20.2	20.2	12.9	8.9	11.3	16.1	15.3	14.5	19.4	22.6

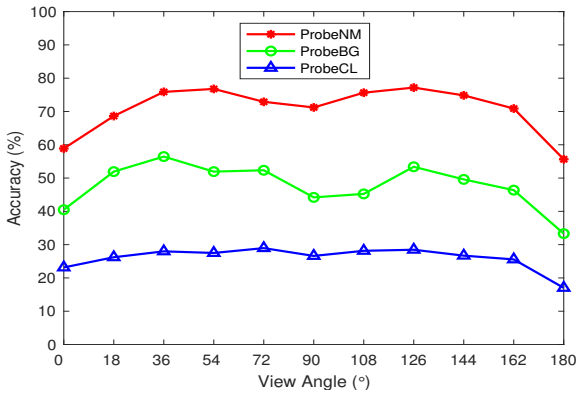

Fig. 4. Average accuracies except the identical view on three probe sets.

with the result in Table I-III. In another aspect, it is easy to see that the proposed VN-GAN significantly outperforms the state-of-the-art methods in most cases. Table IV compares average accuracies of the three probe view against other views in recent literatures. It is easy to see that the proposed VN-GAN achieve promising performances. In specific, it improves 56% (from 19% to 75%) compared with baseline, and 12% (from 63% to 75%) compared with GaitGAN in terms of average accuracy of the three probe view. In addition, the proposed VT-GAN synthesize GEIs in reference view which can provide visual interpretation to the result. As in Fig. 6, the generated GEIs are in good visual quality and looks similar with the corresponding target in appearance.

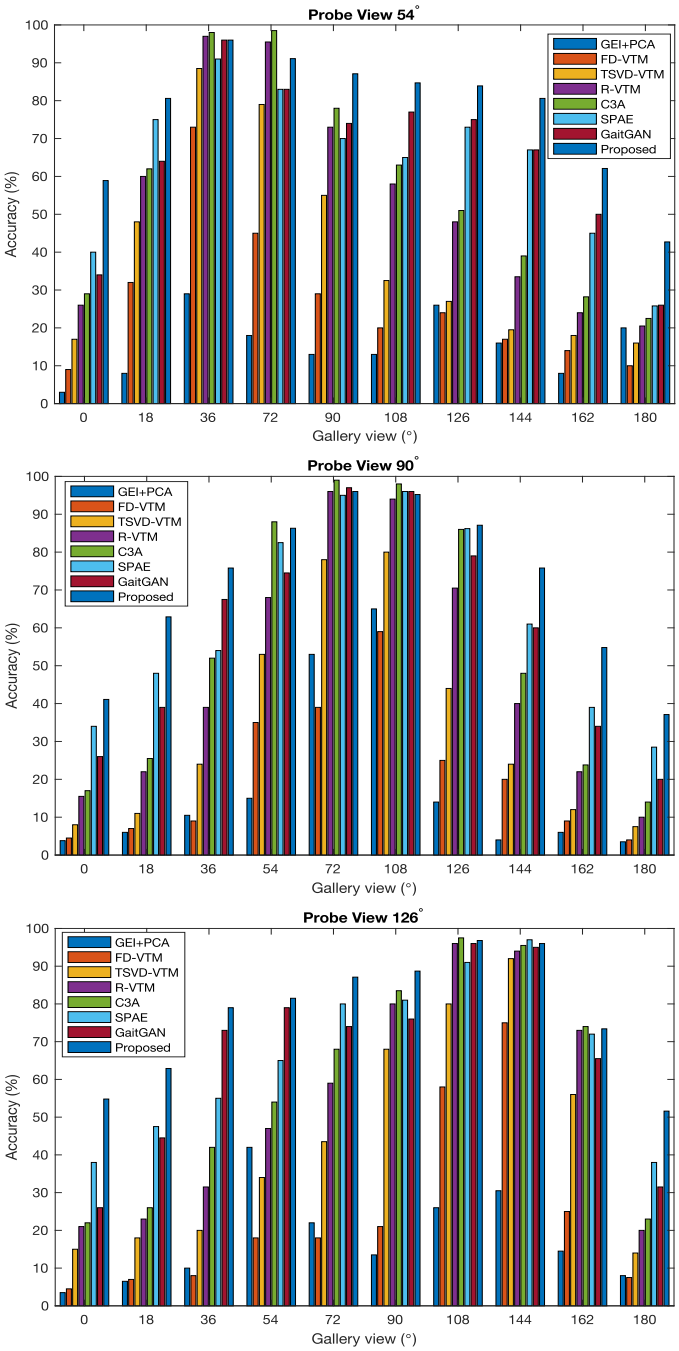

Fig. 5. Comparison with the state-of-the-art approaches on ProbeNM subset at three distinct probe views $54°, 90°$ and $126°$.

C. Ablation Study

This section studies how the two-stage design affects the recognition accuracies. In the proposed VN-GAN, Stage-I attempts to normalize GEIs from arbitrary views to the reference one and Stage-II aims to implant identity information. Hence, the study is equivalent to evaluate the effect of variation normalization and identity preserving. To explore the effect of each stage, we train Stage-I alone (denoting as VT-GAN(s1)) and the whole VT-GAN framework, respectively. We report

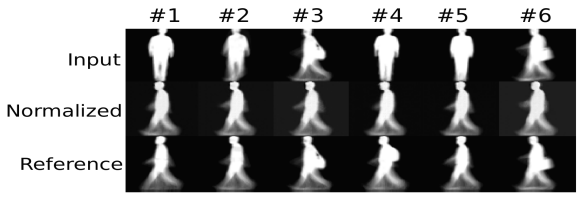

Fig. 6. An example of normalized gaits of six subjects. The three rows stand for input gaits, generated gaits and reference gaits, respectively. Each column represents one subject.

TABLE IV
COMPARISON OF AVERAGE ACCURACIES OF 10 GALLERY VIEWS (THE CORRESPONDING VIEW IS EXCLUDED.) ON PROBENM WITH PROBE VIEW $54°, 90°$ AND $126°$.

Method	Probe View			
	54°	90°	126°	Average
Baseline [38]	0.16	0.23	0.17	0.19
ViDP [15]	0.64	0.60	0.65	0.63
C3A [35]	0.57	0.55	0.58	0.57
SPAE [37]	0.63	0.62	0.66	0.64
GaitGAN [36]	0.65	0.58	0.66	0.63
MGANs [14]	0.77	0.67	0.79	0.74
Proposed	0.77	0.71	0.77	0.75

the results in Table V. From the table, we can observe that each stage achieves promising performance compared to the state-of-the-art and contributes to the ultimate results together.

TABLE V
AVERAGE ACCURACIES OF 10 GALLERY VIEWS (THE CORRESPONDING VIEW IS EXCLUDED.) ON PROBENM WITH 5 DISTINCT PROBE VIEWS. 'S1' DEMOTES TRAINING WITH STAGE-I OF THE PROPOSED VN-GAN.

Method	Probe View					
	18°	54°	90°	126°	162°	Average
GaitGAN [36]	0.54	0.65	0.58	0.66	0.54	0.59
MGANs [14]	0.66	0.77	0.67	0.79	0.72	0.72
VN-GAN(s1)	0.65	0.75	0.66	0.74	0.67	0.69
VN-GAN	0.69	0.77	0.71	0.77	0.71	0.73

V. CONCLUSION

In this paper, we propose a novel framework, *i.e.*, VN-GAN, to address view angle diversity problem in gait recognition. Typically, the proposed VN-GAN is inspired by GAN theory and tries to normalize gaits under various conditions in a coarse-to-fine manner. In Stage-I, A GAN variant that includes a variation discriminator $\mathbf{D}_v$ and a variation classifier $\mathbf{P}$ is used to achieves view normalization, and in Stage-II, another GAN variant that consists of a identity discriminator $\mathbf{D}_i$ and a identity preserver $\mathbf{\Phi}$ is utilized to implant identity information. Moreover, we integrate the two-stage design to the Siamese structure and preserve identity using identity-preserving loss. It is worth to noting that the proposed VN-GAN is also effective to other variations such as dressing changes and carrying variation. Comprehensive experiments on CASIA(B) gait dataset exhibit that the proposed VN-GAN benefits to gait

recognition against various influencing factors. Furthermore, the proposed VN-GAN generates visually promising gaits in reference view which provides intuitional interpretation to the experimental results. We believe that the proposed VN-GAN also benefits to other biometric recognition tasks.

REFERENCES

[1] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al. Tensorflow: a system for large-scale machine learning. In *OSDI*, volume 16, pages 265–283, 2016.

[2] G. Ariyanto and M. S. Nixon. Model-based 3d gait biometrics. In *IJCB*, pages 1–7, 2011.

[3] K. Bashir, T. Xiang, and S. Gong. Cross view gait recognition using correlation strength. In *BMVC*, pages 1–11, 2010.

[4] X. Ben, C. Gong, P. Zhang, X. Jia, Q. Wu, and W. Meng. Coupled patch alignment for matching cross-view gaits. *IEEE TIP*, 2019.

[5] 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.

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

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

[8] H. Chang, J. Lu, F. Yu, and A. Finkelstein. Pairedcyclegan: Asymmetric style transfer for applying and removing makeup. In *CVPR*, pages 40–48, 2018.

[9] 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, 2018.

[10] 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*, volume 1, page 6, 2018.

[11] Y. Ge, Z. Li, H. Zhao, G. Yin, S. Yi, X. Wang, and H. Li. Fd-gan: Pose-guided feature distilling gan for robust person re-identification. In *NIPS*, 2018.

[12] 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.

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

[14] 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.

[15] 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.

[16] 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.

[17] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks, 2017.

[18] A. Y. Johnson and A. F. Bobick. A multi-view method for gait recognition using static body parameters. In *AVBPA*, pages 301–311, 2001.

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

[20] 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.

[21] 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.

[22] 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.

[23] W. Kusakunniran, Q. Wu, J. Zhang, H. Li, and L. Wang. Recognizing gaits across views through correlated motion co-clustering. *IEEE TIP*, 23(2):696–709, 2014.

[24] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. P. Aitken, A. Tejani, J. Totz, Z. Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, volume 2, page 4, 2017.

[25] 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.

[26] 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.

[27] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K.-R. Mullers. Fisher discriminant analysis with kernels. In *Proceedings of the 1999 IEEE signal processing society workshop on neural networks for signal processing*, pages 41–48, 1999.

[28] 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.

[29] X. Qu, X. Wang, Z. Wang, L. Wang, and L. Zhang. Perceptual-dualgan: Perceptual losses for image to image translation with generative adversarial nets. In *IJCNN*, pages 1–8, 2018.

[30] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[31] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCI*, pages 234–241, 2015.

[32] N. Takemura, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi. On input/output architectures for convolutional neural network-based cross-view gait recognition. *IEEE TCSVT*, 2017.

[33] L. Wei, S. Zhang, W. Gao, and Q. Tian. Person transfer gan to bridge domain gap for person re-identification. In *CVPR*, pages 79–88, 2018.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] Z. Zhai and J. Zhai. Identity-preserving conditional generative adversarial network. In *IJCNN*, pages 1–5, 2018.

[40] H. Zhang, T. Xu, H. Li, S. Zhang, X. Huang, X. Wang, and D. Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *ICCV*, pages 5907–5915, 2017.

[41] J. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *CVPR*, pages 2223–2232, 2017.