A Novel View Synthesis Approach based on View Space Covering for Gait Recognition

Rijun Liao[a], Weizhi An[b], Zhu Li[a,*], and Shuvra S. Bhattacharyya[c]

[a] *Department of Computer Science and Electrical Engineering,
University of Missouri-Kansas City, MO, USA.*
[b] *Department of Computer Science and Engineering, University of Texas at Arlington, TX, USA.*
[c] *Department of Electrical and Computer Engineering and UMIACS,
University of Maryland, College Park, MD, USA.*

Abstract

Gait recognition has proven to be effective for long-distance human recognition. View angle, one form of the gait variations, can change the human appearance greatly and reduce its performance. For most existing gait datasets, the angle interval between the two nearest views is large. This means that the angle does not cover the entire view space and prevent better view-invariant feature extraction for CNN. Additionally, the angles between cameras and people vary widely in typical camera deployments for monitoring people. In this paper, we, therefore, propose a novel view synthesis approach based on view space covering to deal with the challenge of large-angle interval. Specifically, a Dense-View GEIs Set (DV-GEIs) is introduced to expand this view approach, from 0° to 180° with 1° interval. GEI is a popular feature representation for gait, which can be obtained by aligning human silhouettes and averaging them in a gait cycle. In order to synthesize DV-GEIs set, Dense-View GAN (DV-GAN) is proposed to model the gait attribute distribution and generate new GEIs with various views. DV-GAN consists of a generator, discriminator, and monitor, where the monitor is designed to preserve human identification and

[⋆]

[*] Corresponding author

Email addresses: `rlyfv@mail.umkc.edu` (Rijun Liao), `weizhi.an@mavs.uta.edu` (Weizhi An), `lizhu@umkc.edu` (Zhu Li), `ssb@umd.edu` (and Shuvra S. Bhattacharyya)

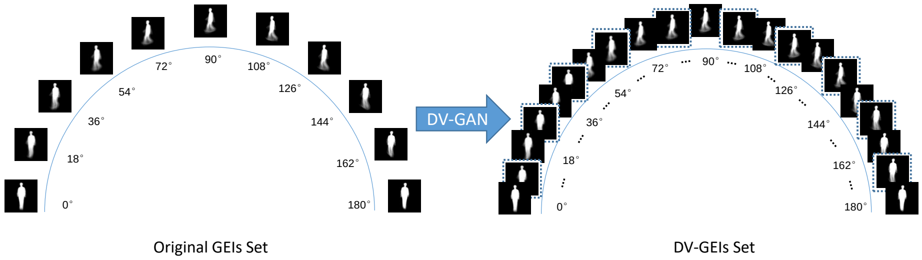


Figure 1: Dense-View GEIs Set (DV-GEIs): view space covering to lighten the burden of view-invariant feature extraction for a CNN and make the feature more discriminative, view angle from 0° to 180° with 1° interval. DV-GAN is proposed to synthesize realistic samples with various view angle conditions. (Sample images from CASIA-B dataset [1])

view information. Compared with our previous work DV-GAN-pre, we add a center for each object in the monitor to improve the discriminative capability of synthesized images during the modeling process. The proposed method is evaluated on the CASIA-B and OU-ISIR dataset. The experimental results show that view space covering is an effective way to light the burden of view-invariant feature extraction for CNN and make the feature more discriminative. We believe the idea of view space covering will further improve the development of gait recognition.

Keywords: Gait Recognition, View Space Covering, Dense-View GEIs,
Dense-View GAN

1. Introduction

1.1. Motivation

Gait is a popular type of biometric feature for human identification. Gait recognition is a technology that can recognize human identity by human's walking pattern. It has many potential applications in video surveillance and public safety. This is because gait can identify subjects at longer distances compared with other features like face, iris, palmprint, and fingerprint. In addition, gait provides a unique possibility to identify a subject without people's cooperation, which can make a great contribution to catching criminals. Therefore, gait recognition is an important area of study for researchers.

However, gait recognition is often challenging in real applications. This is because there are many potential sources of variation that can change the human shape drastically, such as view, clothing, and carrying a bag. Such variation can have a strong negative influence on gait recognition performance. The view angle is one of the most common sources of variation. This is because it is difficult to synchronize changes in view with the direction in which subjects are walking. Moreover, the angles between cameras and people vary widely in typical camera deployments for monitoring people.

Most existing gait datasets [1, 2, 3] are limited in the variety of view conditions that are covered. Typical datasets exhibit large angular distances between the two nearest views. For example, the interval between the two closest views in the CASIA-B dataset [1] is 18°. This dataset includes 11 views from 0° to 180°, as shown in the left image of Figure 1. As additional examples, interval angles on the OU-ISIR [3] and OU-MVPL [2] datasets are 10° and 15°, respectively. OU-MVPL [2] has 14 views and OU-ISIR [3] has only 4 views. A limited number of view angles has a negative influence on view-invariant feature extraction. This problem can be overcome with a larger number of view angles, meaning a smaller interval (e.g., 1°) between closest views. However, it is very challenging to collect this type of data manually.

In this paper, we propose a novel gait energy image (GEI) view synthesis solution based on view space covering to deal with the challenge of large-angle interval. The goal of DV-GEIs is to cover the whole gait view space, and further extract better view-invariant feature from dense view sets.

1.2. Method Overview and Contributions

An overview of the proposed method is shown in Figure 2, which is based on a GAN-CNN framework. Given a gait dataset, the training GEIs set is used to train a DV-GAN for gait attribute distribution modeling and sampling. A large number of GEIs with various views is synthesized to cover the entire view space

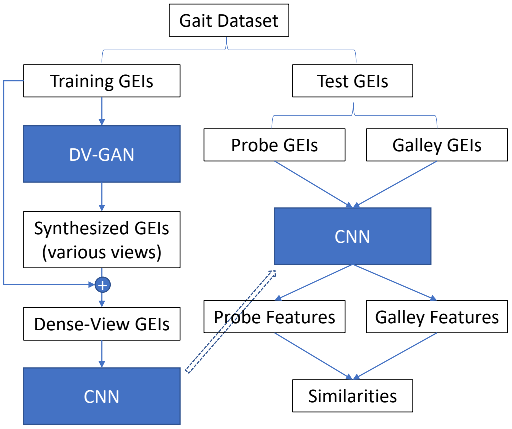


Figure 2: Overview of our proposed GAN-CNN based framework. Dense-View GAN (DV-GAN) is proposed to model the gait attribute distribution and synthesize new GEIs with various views. Then, both synthesized GEIs and original training GEIs are combined to obtain the Dense-View GEIs set, which is used to train a deep CNN for extracting view-invariant feature. In the inference stage, probe and gallery features are used to measure the similarity between the gallery and probe GEIs, and then predict the human IDs label.

from the trained DV-GAN model. Then, both synthesized GEIs and the original training GEIs are combined to obtain the Dense-View GEIs set, which is used to train a deep CNN for extracting view-invariant feature. In the inference stage, probe GEIs and gallery GEIs are used to extract probe and gallery features from the trained CNN model. Because the number of subjects will be changed in the inference stage. A classifier can not be designed to directly decide the IDs, so the similarities between the probe and gallery features are evaluated to predict the human IDs by the nearest neighbor algorithm.

A preliminary version [4] of this work was published in the IEEE International Joint Conference on Biometrics (IJCB) 2020. We denote our preliminary work [4] DV-GEIs as DV-GEIs-pre and this work as DV-GEIs, DV-GAN as DV-GAN-pre, and this work as DV-GAN. We extend our work in two aspects. 1) One is the extension of DV-GAN, we extend our monitor by adding an additional center for each object during synthesizing to minimize the intra-class distances of synthesized images, which enables the synthesized GEI to not only

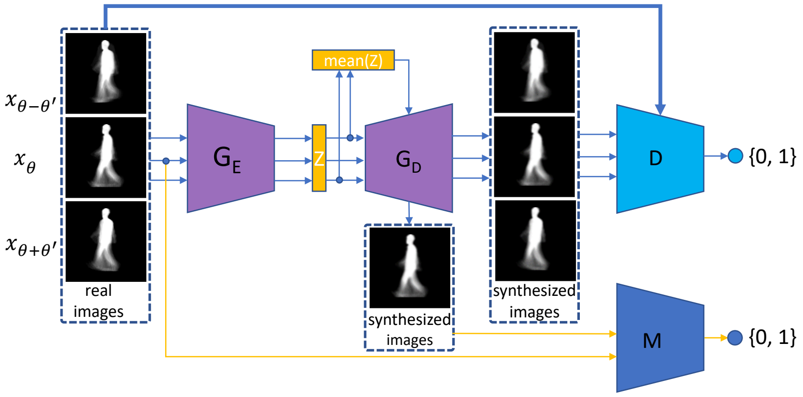


Figure 3: The overall system diagram of Dense-View GAN (DV-GAN), which includes the generator G, discriminator D and monitor M. The generator G consists of an encoder G_E and a decoder G_D. DV-GAN is proposed to model the gait attribute distribution.

looks realistic but also has more discriminative capability. 2) In our previous work, Dense-View GEIs (DV-GEIs) set covers the whole view space only under normal walking conditions. In order to enable gait recognition can deal with multi-variations, we extend DV-GEIs set under carrying a bag and wearing coat conditions. That is, DV-GEIs set also covers all kinds of view angles in the perspective of carrying a bag or clothing condition.

In summary, our method in this paper has the following contributions:

- We propose a novel gait energy image (GEI) view synthesis solution based on view space covering, which provides a much denser sampling to lighten the burden of view-invariant feature extraction for CNN and make the feature more discriminative. Specifically, a Dense-View GEIs Set (DV-GEIs) is introduced to expand this view approach, from 0° to 180° with 1° interval, as shown in Figure 1.
- A GAN-CNN based framework is proposed to improve gait robustness to view variation. DV-GAN is used to solve the key issue of building the DV-GEIs dataset, while a CNN is employed for feature extraction and prediction of human IDs, as shown in Figure 2. One advantage of the

GAN-CNN framework is that although DV-GAN is required to synthesize images at training time (left part of Figure 2), DV-GAN is not needed again during the inference stage (right part of Figure 2), which enhances its utility in real applications.

- A novel Dense-View GAN (DV-GAN) is proposed to model the gait view attribute distribution and develop the perspective space to cover the gait view space. Unlike a traditional GAN, which mainly consists of a generator and discriminator, DV-GAN includes an additional monitor, which not only can maintain human identification and view information very well, but also improves the discriminative capability of synthesized images, as shown in Figure 3.

The rest of this paper is organized as follows. Section 2 introduces the latest methods that handle with variances in gait recognition. The proposed method is presented in Section 3, including the structure of DV-GAN and how to cover the gait view space. Experiments and evaluations are presented in Section 4. The last section, Section 5, illustrates conclusions.

2. Related Work

In this section, we will give a brief review of existing gait recognition methods. The recent methods of gait recognition can be roughly divided into two categories, namely template-based and sequence-based. In addition, we will briefly review some researches based on synthesized samples to improve original performance.

2.1. Template-based methods

One common pipeline of template-based methods are 1) get the human silhouettes by background subtraction in each frame sequence; 2) create a gait template by aligning the silhouettes; 3) extract invariant feature through some machine learning or deep learning approaches. 4) and then compute the similarities between each of the two invariant features. GEI+PCA [5] is one of the classic methods, it has a good performance when there are no obvious variations. But it is difficult to deal with view angle variations.

Previous methods generally divide this pipeline into two parts, template creation, and template matching. Gait Energy Image (GEI) template [5] and Chrono-Gait Image (CGI) template [6] are very popular gait template features. In template matching methods, the most common method is the view transformation model (VTM) which can transform gait template features from one view to another view, reducing the effect of view variation. Yasushi *et al.* introduced FD-VTM [7] method to extract frequency-domain features of the volume by Fourier analysis. To further improve the performance of the VTM, RSVD-VTM [8] was proposed by Worapan *et al.* which employed Linear Discriminant Analysis (LDA) and Singular Value Decomposition (SVD) to optimize the obtained GEI feature vectors. In order to deal with large intra-class variations, Zheng *et al.* proposed RPCA-VTM [9] to establish a robust view transformation model by using robust principal component analysis. Zheng *et al.* [9] found out a shared linear correlated low-rank subspace has a positive influence on robust to viewing angle variation.

Above VTM-based methods have made big progress in dealing with the cross-view problem in gait recognition. However, a view transformation model of those methods can only convert a specific angle to another one. And its performance of the model depends heavily on the accuracy of the view angle estimation. In addition, they need to know the angles of the probe and gallery before extracting gait features. This means that a lot of models are needed because each view needs one model, which led to some challenges in the real application.

To deal with the limitation of each angle needs one model, some researchers have achieved view-invariant transformation only using one model. For instance, Hu *et al.* [10] proposed a view-invariant discriminative projection (ViDP) method to improve the discriminative ability by iteratively learning the low dimensional geometry and finding the optimal projection. What's more, Hu *et al.* [7] capture nonlinear manifolds and reduce dimensional by combining the enhanced Gabor (EG) representation of GEI and the regularized local tensor discriminant analysis (RLTDA) method. However, the method is sensitive to initialization. In the following years, Yu *et al.* used only one uniform model to extract view-invariant feature. SPAE [11] was proposed by him to synthesize the gait feature in a progressive way by stacked multi-layer auto-encoders. In addition, Yu *et al.* also proposed GaitGAN [12] and GaitGANv2 [13] to transform any view gait into the side view gait by using only a uniform model. A GAN model is taken as a regressor at their proposed methods [12, 13] to create a canonical side view of a walking gait in normal clothing without wearing a coat and carrying bag condition. However, the side view transform strategy will collapse when the view variance is large.

Recently, a very solid piece of work [14] contributes new knowledge to the cross-view gait recognition task. Ben *et al.* [14] proposed coupled patch alignment (CPA) and multi-view patch alignment (MPA) to handle gait recognition across two or more views. Moreover, CPA and MPA produce more favorable results than the state-of-the-art. Then, they further developed higher-order tensor-based methods [15, 16], and discovered the optimal matrix subspace where the GEIs across views are aligned in both horizontal and vertical coordinates.

2.2. Sequence-based methods

Sequence-based methods directly employ a sequence of human silhouettes or other human features based on video as input data rather than a template feature. In 2017, Liao *et al.* [17] proposed a pose-based temporal-spatial network (PTSN) extract the temporal-spatial features from a sequence of 2D human pose coordinates. To further improve its robustness to view variation, they [18, 19] generated directly 3D human pose coordinates from a single RBG frame and extract invariant feature from them. Human pose coordinates have an advantage that it is invariant to human appearance compared with human silhouettes. For providing a platform to study human pose information, An *et al.* [20] created a large-scale human pose-based gait database (OUMVLP-Pose) by using deep learning-based pose estimation algorithms. Pose-based methods are robust to human shape, but their performance still needs to be improved because human pose coordinates have not enough information compared with human silhouettes.

Because human silhouettes have rich information compared with human pose coordinates, Wu *et al.* [21] used CNN to extract gait feature from a sequence of human silhouettes and achieved high performance. Different from [21] which uses continuous human silhouettes, Chao *et al.* [22] introduced Gaitset network to further improve gait recognition performance based on unordered silhouettes set. In order to make the use of the local gait feature, GaitPart [23] was proposed by Fan *et al.* Rather than using the human silhouettes as input data, Zhang *et al.* proposed GaitNet [24] to explicitly disentangle pose and appearance features from RGB image, and then LSTM-based integration of pose features would produce the gait feature.

In order to tackle the problem of view variation, there are some popular works for view-invariant action recognition. An Unsupervised AttentioN Transfer (UANT) approach was proposed by Ji *et al.* [25], which can transfer attention from one selected reference view to arbitrary views. In addition, Ji *et al.* [26] proposed a View-guided Skeleton CNN (VS-CNN) which divides full-circle views (360°) into four view groups and learns four invariant features from corresponding to four view groups. Different from the approaches of [25, 26] that transfer or group views, we synthesize samples with arbitrary views through GAN, and provide a much denser sampling to lighten the burden for CNN to seek a common feature space in various views.

Above sequence-based methods can achieve high performance in gait recognition. However, the price of these methods has a high computational cost because they need to deal with videos with a large number of images. This will bring challenges to real-time and low-cost applications. In contrast, the gait energy image (GEI) [5] is a very popular feature representation for gait recognition because of its efficient computation and robustness to noise. It can be obtained by aligning human silhouettes and averaging them in a gait cycle, as shown in Figure 4. The pixel value in the GEI can be represented as the probability that the human body occupies the pixel location in the GEI during the gait cycle. Because of the successful experience of GEI in gait recognition, GEI is employed in our proposed method as the input and target data, the same as GaitGAN [12] and SPAE [11].

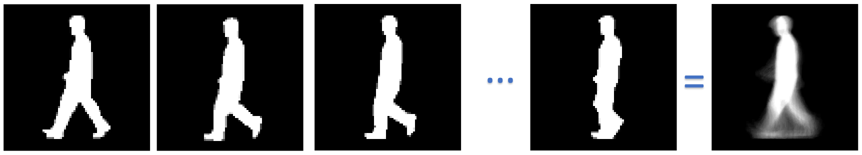


Figure 4: A gait energy image (GEI) [5] is obtained by aligning human silhouettes and averaging them.

2.3. Sample Synthesis and View Space Covering

Recently, some researchers proposed novel approaches to synthesize samples and improve original performance in some specific tasks. For examples, Chen *et al.* [27] used GAN to generate noise samples and use them in image blind denoising. Qian *et al.* [28] combined human pose and GAN to synthesize human image in a specific pose for person re-identification. Those methods can greatly improve its original performance. However, this idea has not been achieved in the gait recognition task, because it needs to find a suitable solution to synthesize samples with different conditions. One popular based on GAN work, GaitGAN [12], has used GAN to transform any view GEI into the side view GEI, which effectively improves gait robustness. In contrast, our proposed DV-GAN is to model the gait attribute distribution and generate samples with various view angles, rather than view transformation.

A recent work [29] provides a change to synthesize samples with various view angles to cover the whole gait view space. In [29], authors model the face attribute distribution and produce latent vectors that can capture the semantic information of facial expressions by autoencoder. And then generate a series of different view angle human faces from *left face* to *right face* by linear transformation $z = \alpha z_p + (1 - \alpha) z_q$ in latent space. Based on this idea, we propose

DV-GAN to synthesize the gait images with different view angles with cover the whole view space, and further improve its robustness to view variation. Unlike the above view transformation methods, we generate gait features with dense views to cover the whole view space and improve the recognition rate on the cross-view condition.

3. GAN-CNN Based for View Space Covering

In this paper, we propose a GAN-CNN based framework to cover the whole view space and further learn better invariant features for gait recognition. View space is covered by trained Dense-View GAN (DV-GAN). When dealing with samples with limited view angles, training samples are used to train a DV-GAN to model the gait attribute distribution. The trained DV-GAN is utilized to solve the key issue of building a training dataset Dense-View GEIs (DV-GEIs) set with various views, and then CNN is employed for view-invariant feature extraction from DV-GEIs set. The overview of our proposed method can be seen in Figure 2. In this section, we will give the detail of the structure of DV-GAN and how to synthesize DV-GEIs set to cover the view space.

3.1. Dense-View GEIs Set (DV-GEIs)

The purpose of the DV-GEIs set is to cover the view space from angle p to angle q, as shown in Figure 5. Here, we denote the views of input GEI x_p and x_q for angles p and q. The GEI feature sampling from existing features follows the equation below:

$$x' = \{G_D(z)|z = \alpha z_p + (1 - \alpha)z_q\} \tag{1}$$

where $z_p = G_E(x_p), z_q = G_E(x_q)$, latent space features z_p, z_q are encoded by encoder G_E which keep the characteristic of gait attribute. The interpolation is defined by linear transformation $z = \alpha z_p + (1 - \alpha)z_q$, where $\alpha \in [0, 1]$, and then z is fed to decoder G_D to generate new GEIs.

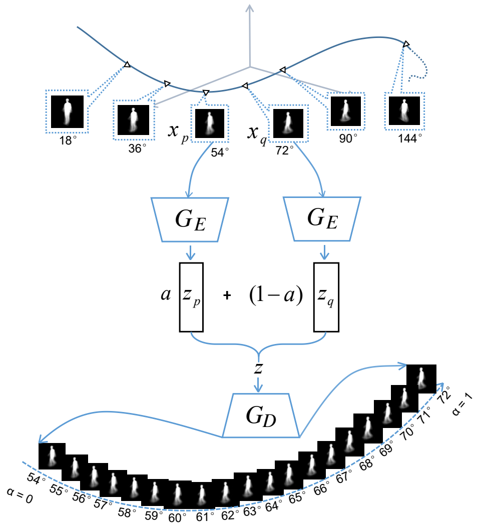


Figure 5: The workflow of view space covering. The latent space z_p, z_q are encoded by encoder G_E with two input GEIs x_p, x_q. A large number of views GEIs from p angle to q angle are synthesized by decoding the latent space z, where $z = \alpha z_p + (1-\alpha) z_q$, $\alpha \in [0, 1]$.

3.2. Dense-View GAN (DV-GAN)

Unlike the traditional GAN [30, 28] which usually consists of one generator and one discriminator, our DV-GAN model has three neural networks: generator G, monitor M and discriminator D, the overall diagram of DV-GAN as shown in Figure 3.

3.2.1. Generator

Given an input GEI x, and a target GEI $\hat{x}$, where $x = \hat{x}$. The generator can reconstruct GEI and model gait attribute distribution in latent space and develop the perspective space. The network is inspired by the pixels to pixels level idea [31] to reconstruct the GEI image, that is adding L_1 norm loss to make sure the output GEI $\hat{x} = G(z, x)$ is the same as input x, the generator loss function is defined as:

$$\min_{E,G} L_{L1}(G(z), x) \tag{2}$$

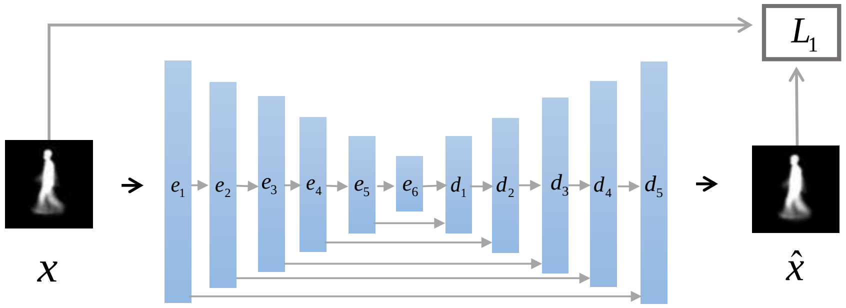


Figure 6: The U-net architecture generator of DV-GAN for modelling gait view space distribution.

The network architecture of generator adds skip connections between two layers to form a U-Net [32] architecture. This is because U-Net [32] structure allows low-level information to shortcut across the high-level information and effectively boost the quality of the synthesized images, as shown in Figure 6. In order to use latent space to cover the view space, as shown in Figure 5, we define U-Net into two parts, encoder $G_E\{e_1\}$ and decoder $G_D\{e_2, e_3, e_4, e_5, e_6, d_1, d_2, d_3, d_4, d_5\}$. The feature map of e_1 layer will be defined as latent space z. This is because the U-Net architecture network does not allow decoding latent space if other layers' feature map as latent space z. For example, if output feature of e_2 is defined as latent space z, then $G_E\{e_1, e_2\}$ and $G_D\{e_3, e_4, e_5, e_6, d_1, d_2, d_3, d_4, d_5\}$ will be as encoder and decoder respectively. It allows us to do linear interpolation $z = \alpha G_E(x_p) + (1-\alpha)G_E(x_q)$, but it does not allow us to decode latent space z, because the calculation of feature map of d_5 requires the feature map of e_1 in U-Net structure, while decoder G_D does not include e_1 layer.

3.2.2. Discriminator

The discriminator network D is designed to make sure the generated GEIs are more realistic. A pair of a real GEI x and a synthesized GEI $\hat{x}$ are taken as input data and is trained to recognize whether a GEI is real or not. If the input GEI is from a real GEI, the discriminator output value is 1, otherwise 0. The discriminator could make sure the synthesized GEI is more and more similar to the original GEI. The objective function of the discriminator is:

$$\min_{G}\max_{D}\mathbb{E}_{x,z\ \ p_{data(x,z)}}[logD(x,x)]+\mathbb{E}_{x,z\ \ p_{data(x,z)}}[1-logD(x,G(z))] \qquad (3)$$

3.2.3. Monitor

The monitor is different from traditional GAN, which is designed to preserve human identification information and view information, as shown in Figure 7. The monitor has three input GEIs $x_{\theta-\theta'}, x_\theta$ and $x_{\theta+\theta'}$. In the training process, first, a GEI $\hat{x}_\theta$ will be synthesized by decoding the latent feature z, where latent feature z is the mean value of encoded features ($z_{\theta-\theta'}$ and $z_{\theta+\theta'}$) of two input GEIs ($x_{\theta-\theta'}$ and $x_{\theta+\theta'}$). The monitor will then create a scalar probability to indicate if the synthesized GEI $\hat{x}_\theta$ is the same as the original GEI x_θ or not, so the view information and identification information would be preserved in the training processing, the equation as follows:

$$\min_{G}\max_{D}\mathbb{E}_{x_\theta,z\ \ p_{data(x_\theta,z)}}[logD(x_\theta,x_\theta)]+\mathbb{E}_{x_\theta,z\ \ p_{data(x_\theta,z)}}[1-logD(x_\theta,\hat{x}_\theta)]$$
$$where \quad \hat{x}_\theta = G(\frac{1}{2}E(x_{\theta-\theta'})+\frac{1}{2}E(x_{\theta+\theta'})) \qquad (4)$$

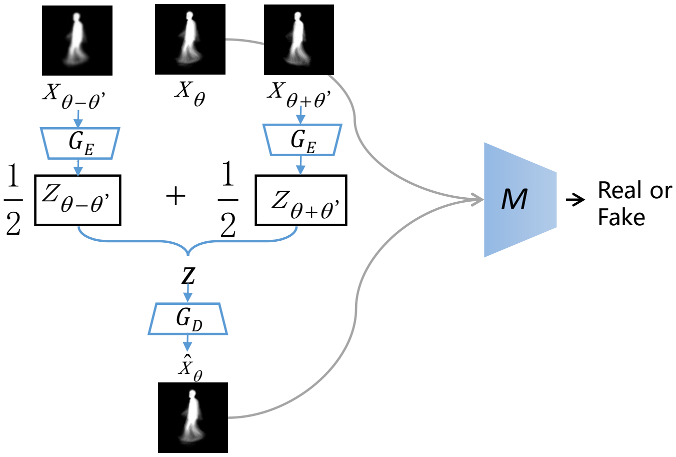


Figure 7: The structure of the real/fake monitor. Monitor will identify if the generate image $\hat{x}_\theta$ is same as the original image x_θ or not, to preserve human identification and view information.

To improve the discriminative capability of synthesized images. We are inspired by center loss [33] and extend our monitor by assigning an additional center c_{yi} for each object $\hat{x}_{\theta i}$ in the monitor, as shown in the Equation 5, where m represents training batch size, and the subscript i represents the ith sample in the training batch, $\hat{x}_{\theta i}$ is the ith synthesized GEI that belongs to the lth class, c_{li} is the l_ith class center of synthesized GEI. If the synthesized images $\hat{x}_{\theta i}$ belong to the same person yi, then synthesized images would close to the center c_{yi}. From Equation 5, it would reduce the intra-class distance, and the inter-class distance is enlarged as a consequence of this optimization. This is because each class has its own center, and there is a distance between different centers. In the training stage, samples will be close to their own center, so inter-class distance is also enlarged. The adversarial training framework provides considerable flexibility in the composition of two constraints and improves the discriminant capability of synthesized images.

$$\sum_{i=1}^{m} ||\hat{x}_{\theta i} - c_{yi}||_2^2 \tag{5}$$

4. Experiments and Analysis

4.1. Datasets

The proposed method is evaluated on CASIA-B dataset [1] with 11 views at 18° interval between two closest views and OU-ISIR Large Population Dataset [3] with 4 views at 10° interval, respectively.

CASIA-B gait dataset [1] is one of the popular public gait databases and it was created by the Institute of Automation, Chinese Academy of Sciences in January 2005. It consists of 124 subjects in total, including 31 females and 93 males. Each subject has 10 sequences, 6 sequences of normal walking (NM), 2 sequences of walking with a bag (BG), and 2 sequences of walking with a coat (CL), which provide a platform to analyze gait in the real environment.

Some experiments are also performed on OU-ISIR Large Population Dataset [3] with only 4 views at 10° interval between two closest views (55°,65°,75° and 85°). OU-ISIR is a very large data set, containing 4007 persons, ranging in age from 1 to 94 years old. Under normal walking conditions, each subject has two walking sequences. It allows us to research the upper limit of gait recognition performance in a more statistically reliable way.

4.2. Experimental Setting

For better comparison with SPAE [11] and GaitGAN [12], our experimental setting is exactly like theirs. The training set contains the first 62 subjects, and the test set contains the rest of the subjects. In the test set, the gallery set contains the first 4 normal sequences. The probe set contains three different conditions: the rest 2 normal sequences, 2 walking sequences with a bag, and 2 walking sequences with a coat condition, respectively, as shown in Table 1. We only generate images in the training set, not in the test set in our proposed method. One advantage is that although DV-GAN is required to synthesize images at training time, DV-GAN is not needed again during the inference stage, which enhances its utility in real applications.

Table 1: Experimental setting on CASIA-B dataset (NM: normal walking, BG: walking with a bag, CL: walking with a coat).

Training	Test	
	Gallery Set	Probe Set
ID: 001-062 NM01-NM06 BG01-BG02,CL01-CL02	ID: 063-124 NM01-NM04	ID: 063-124 NM05-NM06 BG01-BG02,CL01-CL02

The setting of the OU-ISIR [3] dataset is similar to that of the CASIA-B dataset. In the experiment, we randomly divide all subjects into five groups, reserve one group for testing, and four groups for training to synthesize samples. In each test set, the first sequence is put into the gallery set, and the remaining sequences are put into the probe set.

4.3. Implementation Details of DV-GAN

We follow the idea of the network structure of Isola *et al.* [31] and propose the DV-GAN network. *Pix2pix* [31] is an open code deal with the challenge of image-to-image translation. The network architecture of generator and discriminator can be seen in Table 2 and Table 3. To preserve human identification and view information, the monitor is proposed in our DV-GAN. The network architecture of the monitor (Table 4) is the same as that of the discriminator, but their input data settings are different. The number of input images in the discriminator is 2, while the number of input images in the monitor is 3. In the CASIA-B experiment, we set the $\theta' = 18^\circ$ in the Equation 4, where $\theta \in \{18^\circ,36^\circ,54^\circ,72^\circ,90^\circ,108^\circ,126^\circ,144^\circ,172^\circ\}$.

Table 2: Network Architecture of the Generator

Layers	Number of filters	Filter size	Stride	Batch norm	Activation function
Conv.1	64	5× 5	2	N	L-ReLU
Conv.2	128	5× 5	2	Y	L-ReLU
Conv.3	256	5× 5	2	Y	L-ReLU
Conv.4	512	5× 5	2	Y	L-ReLU
Conv.5	512	5× 5	2	Y	L-ReLU
Conv.6	512	5× 5	2	Y	L-ReLU
Deconv.1	512	5× 5	2	Y	ReLU
Deconv.2	512	5× 5	2	Y	ReLU
Deconv.3	256	5× 5	2	Y	ReLU
Deconv.4	128	5× 5	2	Y	ReLU
Deconv.5	64	5× 5	2	Y	ReLU
Deconv.6	64	5× 5	2	N	Tanh

After the DV-GAN models the gait attribution distribution, the next step is to synthesize dense view samples. DV-GEIs with 1° interval will be generated by trained DV-GAN. We synthesize GEI samples from 0° to 180° with 1° interval

Table 3: Network Architecture of the Discriminator.

Layers	Number of filters	Filter size	Stride	Batch norm	Activation function
Conv.1	64	5× 5	2	N	L-ReLU
Conv.2	128	5× 5	1	Y	L-ReLU

Table 4: Network Architecture of the Monitor.

Layers	Number of filters	Filter size	Stride	Batch norm	Activation function
Conv.1	64	5× 5	2	N	L-ReLU
Conv.2	128	5× 5	1	Y	L-ReLU

by linear transformation $z = \alpha z_p + (1 - \alpha) z_q$ and decoder latent space $G_D(z)$. Follow the Equation 1, we set $\alpha \in \{\frac{1}{18}, \frac{2}{18}, \cdots, \frac{17}{18}\}$, where the angle set of z_p and z_q is $\{(0°, 18°), (18°, 36°), \cdots, (162°, 180°)\}$. So we can get a large number of view angle GEIs that do not exist in the original dataset. DV-GEIs set is formed by combining the synthesized GEIs and original GEIs, which being fed into CNN to extract invariant features.

4.4. Implementation Details of Feature Extraction

To extract view-invariant feature from view space covering DV-GEIs set. We follow the idea of CNN architecture of Pan *et al.* [34] and multi-loss function of PoseGait [19] and use a simple CNN to extract feature. The network of PoseGait [19] can effectively extract gait dynamic and static information from the human pose sequence. The network architecture of the CASIA-B dataset can be seen in Table 5, the number of convolutional layers of the OU-ISIR dataset is four more than that of CASIA-B because the number of the subject of OU-ISIR is 4007 which is more than that of CASIA-B.

We follow the idea of multi-loss function by Wen *et al.* [33] which used the softmax loss and center loss jointly to supervise the learning of deep convolutional neural networks. The multi-loss function is shown in Equation 6. The softmax loss can contribute to pulling apart different GEIs and it can enlarge the inter-class dispersion. The center loss is useful to minimize the intra-class variation and keep the features of different classes separable.

$$L = L_S + \gamma L_c = -\sum_{i=1}^{m} \log \frac{e^{W_{l_i}^T \hat{x}_i + b_{l_i}}}{\sum_{j=1}^{n} e^{W_j^T \hat{x}_i + b_j}} + \frac{\gamma}{2} \sum_{i=1}^{m} ||\hat{x}_i - c_{li}||_2^2 \quad (6)$$

Table 5: Network Architecture of the CNN.

Layers	Number of filters	Filter size	Stride	Activation function
Conv.1	32	3× 3	1	P-ReLU
Conv.2	64	3× 3	1	P-ReLU
Pooling.1	N	2× 2	2	N
Conv.3	64	3× 3	1	P-ReLU
Conv.4	64	3× 3	1	P-ReLU
Eltwise.1	Sum operation between Pooling.1 and Conv.4			
Conv.5	128	3× 3	1	P-ReLU
Pooling.2	N	2× 2	2	N
Conv.6	128	3× 3	1	P-ReLU
Conv.7	128	3× 3	1	P-ReLU
Eltwise.2	Sum operation between Pooling.2 and Conv.7			
Conv.8	128	3× 3	1	P-ReLU
Conv.9	128	3× 3	1	P-ReLU
Eltwise.3	Sum operation between Eltwise.2 and Conv.9			
Conv.10	128	3× 3	1	P-ReLU
FC.1	512	N	N	N

where m is the number of training batch size, the subscript i represents the ith sample in m, $\hat{x}_i \in \mathbb{R}^d$ is the ith GEI feature that belongs to the l_ith class. d, $W \in \mathbb{R}^{d \times n}$ and $b \in \mathbb{R}^d$ denote the feature dimension, last connected layer and bias term, respectively. $c_{li} \in \mathbb{R}^d$ is the l_ith class center of gait features. We set

$\gamma = 0.008$ in the experiment.

4.5. Experimental Results on CASIA-B dataset

The gait recognition results on CASIA-B dataset are shown in Table 6, 7 and 8. The recognition rates of normal walking, carrying condition, and clothing variations with different view angles are listed in the three tables. In our experiments, the first 4 normal walking sequences are put into the gallery set. The probe set contains three different conditions, the rest 2 normal sequences, 2 walking sequences with a bag, and 2 walking sequences with a coat condition, respectively. In these tables, each row represents a gallery view, and each column represents a probing view, resulting in each table has 121 combinations. To make it easy to understand the rules of results, we bold the highest results for each column in Table 6, 7 and 8. From these tables, we can see that the performance will be higher when the view angle difference between gallery and probe is small. The overall performance when probe set is on the normal sequences is better than that of on walking with a bag and walking with a coat.

Table 6: Recognition rates when the probe data is normal walking (NM) on CASIA-B dataset.

		Probe set view (NM05, NM06)										
		0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°
Gallery set view (NM01-NM04)	0°	**100.0**	93.55	81.45	56.45	47.58	40.32	41.13	51.61	60.48	77.42	93.55
	18°	95.97	**100.0**	**98.39**	85.48	67.74	58.06	62.10	66.94	77.42	88.71	83.06
	36°	79.03	**100.0**	97.58	95.97	84.68	70.16	79.84	87.10	89.52	83.87	67.74
	54°	54.84	88.71	95.97	96.77	95.16	89.52	90.32	84.68	83.06	71.77	50.00
	72°	37.90	69.35	90.32	**97.58**	**99.19**	99.19	98.39	93.55	87.90	65.32	35.48
	90°	37.90	57.26	82.26	92.74	**99.19**	99.19	98.39	91.94	78.23	55.65	38.71
	108°	35.48	56.45	79.03	89.52	**99.19**	**100.0**	99.19	96.77	92.74	70.97	40.32
	126°	41.13	75.00	80.65	85.48	95.16	95.16	**100.0**	**98.39**	98.39	79.84	54.03
	144°	63.71	79.03	84.68	80.65	73.39	74.19	90.32	97.58	**99.19**	96.77	76.61
	162°	89.52	91.13	80.65	65.32	63.71	58.87	69.35	80.65	96.77	**100.0**	97.58
	180°	95.16	83.06	72.58	48.39	44.35	40.32	44.35	52.42	75.81	94.35	**100.0**

4.6. Effectiveness of View Space Covering with Dense Sampling

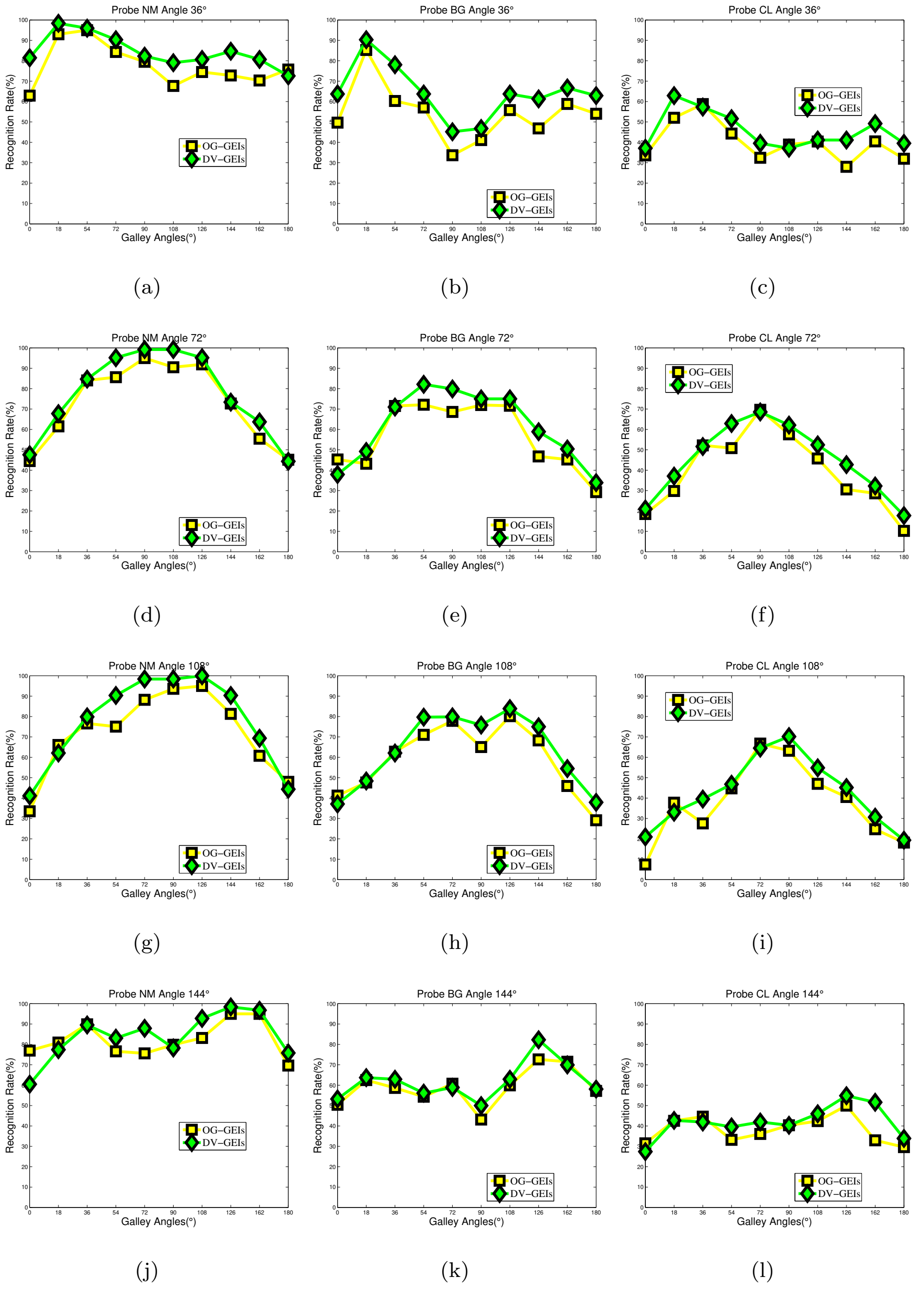
In order to demonstrate that view space covering with dense sampling can contribute to better view-invariant feature extraction, we compare with another

Table 7: Recognition rates when the probe data is with a bag (BG) on CASIA-B dataset.

		Probe set view (BG05, BG06)										
		0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°
Gallery set view (NM01-NM04)	0°	**93.55**	81.45	63.71	51.61	37.90	26.61	37.10	43.55	53.23	67.74	82.26
	18°	80.65	**94.35**	90.32	77.42	49.19	45.16	48.39	53.23	63.71	70.16	64.52
	36°	58.87	87.90	**92.74**	**87.90**	70.97	52.42	62.10	71.77	62.90	58.87	55.65
	54°	33.33	58.54	78.05	86.99	82.11	73.17	79.67	66.67	56.10	45.53	33.33
	72°	26.61	44.35	63.71	81.45	**88.71**	**83.06**	79.84	69.35	58.87	37.10	25.81
	90°	27.42	36.29	45.16	66.94	79.84	**83.06**	75.81	66.94	50.00	31.45	23.39
	108°	25.81	35.48	46.77	67.74	75.00	79.84	79.03	72.58	62.90	37.10	19.35
	126°	33.06	48.39	63.71	70.16	75.00	73.39	**83.87**	**87.90**	82.26	64.52	29.84
	144°	46.77	58.06	61.29	62.90	58.87	54.84	75.00	85.48	**87.10**	84.68	54.03
	162°	63.41	74.80	66.67	55.28	50.41	45.53	54.47	60.16	69.92	**91.06**	82.11
	180°	79.03	70.97	62.90	41.13	33.87	33.06	37.90	43.55	58.06	78.23	**91.94**

Table 8: Recognition rates when the probe data is with a coat (CL) on CASIA-B dataset.

		Probe set view (CL05, CL06)										
		0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°
Gallery set view (NM01-NM04)	0°	**58.06**	49.19	37.10	25.81	20.97	16.94	20.97	22.58	27.42	41.94	39.52
	18°	42.74	**71.77**	62.90	54.03	37.10	33.06	33.06	37.10	42.74	53.23	37.10
	36°	32.26	61.29	**66.94**	55.65	51.61	40.32	39.52	45.16	41.94	37.90	28.23
	54°	20.97	44.35	57.26	**66.94**	62.90	54.03	46.77	50.00	39.52	29.03	26.61
	72°	16.13	33.87	51.61	55.65	66.94	66.13	64.52	54.84	41.94	28.23	22.58
	90°	16.94	25.81	39.52	54.03	**68.55**	**73.39**	70.16	55.65	40.32	24.19	23.39
	108°	11.29	28.23	37.10	46.77	62.10	67.74	**70.97**	**60.48**	45.97	23.39	16.94
	126°	18.55	30.65	41.13	45.97	52.42	48.39	54.84	59.68	54.84	33.06	23.39
	144°	22.58	33.87	41.13	45.97	42.74	40.32	45.16	58.06	**59.68**	47.58	36.29
	162°	33.87	48.39	49.19	33.87	32.26	24.19	30.65	37.90	51.61	**67.74**	45.16
	180°	43.55	37.90	39.52	25.00	17.74	11.29	19.35	23.39	33.87	47.58	**59.68**



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

(j) (k) (l)

Figure 8: Effectiveness on View Space Covering. Training data of OG-GEIs model is employing original GEIs set with large interval between the two nearest views. Training data of DV-GEIs model is employing the view covering DV-GEIs set. Each row represents a probe angle. From left column to right column are NM, BG and CL condition respectively.

experiment OG-GEIs result, as shown in Figure 8. OG-GEIs model is trained by using original GEIs set with a larger interval between the two nearest views. while the DV-GEIs model is trained by using the proposed dense sampling DV-GEIs set. Each row represents a probe angle, we only list 4 probe angles (36°, 72°, 108° and 144°) with a 36 interval because of limited space. The first column shows the comparison in normal walking sequences, while the second and third column shows the comparison on carrying condition and clothing condition respectively.

As illustrated in Figure 8, it is clear that the performance of DV-GEIs is better than that of OG-GEIs at many points not only in normal walking (NM) condition, but also in carrying condition (BG) and clothing condition (CL). This shows that our view covering dense view sampling is an effective way to extract better view-invariant feature and enhance the robustness for gait recognition.

In addition, in order to further show the effectiveness of dense view space covering, we evaluate the average accuracy of the generated DV-GEI set on different degree intervals. The average accuracy is computed by the average of all recognition rates excluding identical-view cases from Table 6, 7 and 8, as shown in Figure 9. The proposed approach of view space covering is to synthesize samples that do not exist on the original dataset, and add generated samples into the orignial dataset to extract better view-invariant feature. Specifically, the interval of two nearest views on CASIA-B is 18°, we synthesize 17 GEIs and add them into the CASIA-B dataset. The angle set on CASIA-B is $\theta \in$ {0°,18°,36°,54°,72°,90°,108°,126°,144°,172°,180°}, the interval degrees are set as $\{2, 1, 1/2, 1/6, 1/18\}$ times of 18° in the experiment, that is 36°, 18°, 9°, 3° and 1° interval. From 9, the gap of performance from 3° interval to 1° interval is not obvious, but it can still a little bit with 1° interval in terms of normal walking and carrying bag conditions. Generally, the finer the granularity of view angles, the better the recognition performance.

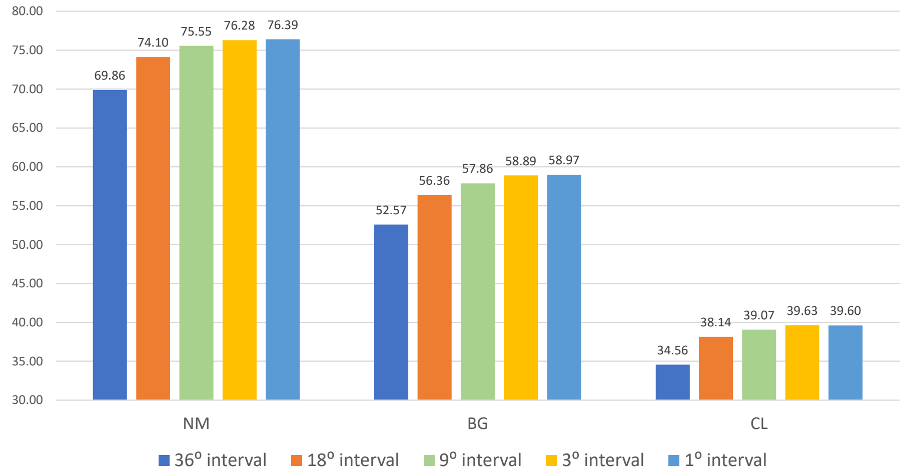


Figure 9: The average accuracy for the probe data being NM, BG, and CL with different degree intervals on CASIA-B dataset. NM: normal walking, BG: walking with a bag, CL: walking with a coat.

4.7. Visualization of View Covering GEIs

To see the quality of view covering generated GEIs, we synthesize some view samples, as shown in Figure 10. In our above experiment, we generate the GEIs by using two adjacent GEIs with 18° interval between the two closest views in CASIA-B dataset. Therefore, we can get various GEIs with 1° interval. The difference of GEIs between the adjacent angle (1°) is difficult to be distinguished in vision. Besides, it does not provide ground truth GEIs in the original dataset to compare the synthesized GEIs. Therefore, in order to visualize the transformation obviously and have ground truth for comparison, we use two GEIs (0° and 90°) with large 90° interval to generate some sample GEIs. That is, in the linear transformation $z = \alpha z_p + (1-\alpha) z_q$, we set linear ratio $\alpha \in \{\frac{18}{90}, \frac{36}{90}, \frac{54}{90}, \frac{72}{90}\}$, and the angle set of z_p and z_q is $\{(0^\circ, 90^\circ)\}$.

To better show the synthesized quality of DV-GAN, we compare our DV-GEIs with another type of synthesized GEIs which direct view morphing by linear interpolation of two GEI images. That is, the second row GEIs are generated by equation $\hat{x} = \alpha x_p + (1-\alpha) x_q$, where $\alpha \in \{\frac{18}{90}, \frac{36}{90}, \frac{54}{90}, \frac{72}{90}\}$. From

Figure 10, we can see that synthesized GEIs by direct view morphing have obvious ghost, while synthesized GEIs by DV-GANs is very similar to ground truth although they are synthesized by two images with 90° interval. It will be more similar to the original GEI if using images with a smaller angle interval to synthesize. The comparison shows that our DV-GAN can synthesize realistic GEI with any angle condition, and our generated GEI can preserve human identification and view information very well.

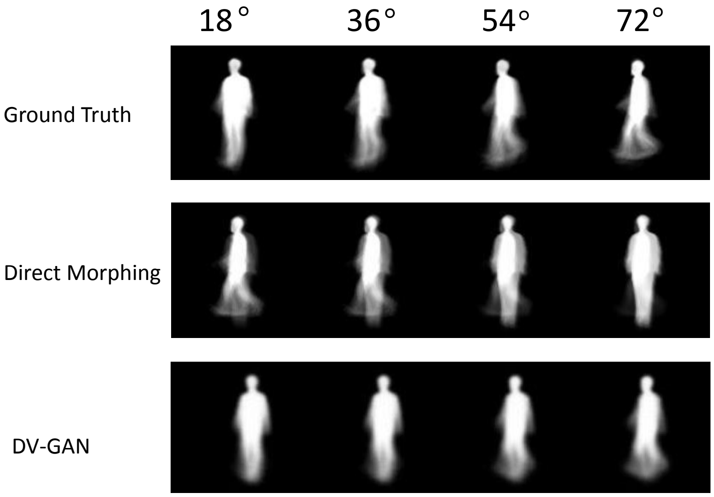


Figure 10: Visualization of view covering synthesized GEIs. Second-row GEIs are synthesized by linear interpolation of two GEI images. Third-row GEIs are synthesized by the proposed DV-GAN. Two types of synthesized GEIs are generated by 0° and 90° two original GEIs.

4.8. Comparison with GEI template methods

To further evaluate the proposed method, we compare proposed method with recent GEI-based template methods. Including SPAE [11], GaitGAN [12], GaitGANv2 [13], GaitSet-GEI [22] and our previous work DV-GEIs-pre [4]. For the reasonableness of comparison, these methods are all based on the GEI template, and their experimental setting are also based on Table 1. In the Table 9, we can see the mean accuracy (76.4%) of proposed method is much better than that of SPAE [11] (59.3%), GaitGAN [12] (57.2%) and GaitGANv2 [13] (63.1%) on the

Table 9: Comparison with GEI-based template approaches at average accuracy (%) on CASIA-B dataset. Excluding identical-view cases. (GaitSet-GEI [22]: GEI is fed into the GaitSet.)

Training Subjects	Gallery View NM:01-04	0°-180°											
	Probe View NM:05-06	0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	Mean
62	SPAE [11]	50.0	58.1	61.0	63.3	64.0	62.1	62.3	66.3	64.4	54.5	46.7	59.3
	GaitGAN [12]	41.9	53.5	63.0	64.5	63.1	58.1	61.7	65.7	62.7	54.1	40.6	57.2
	GaitGANv2 [13]	48.1	61.9	68.7	71.7	66.7	64.8	66.0	70.2	71.6	58.9	46.1	63.1
	DV-GEIs-pre [4]	64.5	76.2	81.3	80.8	77.1	72.6	74.4	78.9	80.6	75.6	63.7	75.1
	DV-GEIs	63.1	79.4	84.6	79.8	77.0	72.6	77.4	80.3	84.0	78.5	63.7	**76.4**
74	GaitSet-GEI [22]	-	-	-	-	-	-	-	-	-	-	-	80.4
	DV-GEIs-pre [4]	71.0	86.4	91.4	89.6	80.4	80.1	82.5	90.1	90.4	85.3	70.5	83.4
	DV-GEIs	72.9	85.9	89.3	87.1	83.7	81.7	82.8	87.3	91.3	87.1	74.9	**84.0**
62	Probe Angle BG:01-02	0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	Mean
	SPAE [11]	34.0	38.6	42.1	42.7	39.0	32.8	31.3	39.9	41.0	35.7	32.3	37.2
	GaitGAN [12]	28.5	35.2	42.7	34.4	38.0	33.5	36.2	44.8	41.8	33.3	23.6	35.6
	GaitGANv2 [13]	37.2	43.4	46.6	46.0	47.6	41.5	41.2	48.5	48.8	42.2	31.6	43.1
	DV-GEIs	47.5	59.6	64.2	66.3	61.3	56.7	63.4	63.3	61.8	57.5	47.0	**59.0**
74	GaitSet-GEI [22]	-	-	-	-	-	-	-	-	-	-	-	**68.1**
	DV-GEIs	58.3	71.6	79.1	71.5	63.6	56.7	57.7	73.7	74.4	69.4	58.4	66.8
62	Probe Angle CL:01-02	0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	Mean
	SPAE [11]	21.5	25.4	27.3	28.1	26.9	22.2	22.3	26.3	24.8	21.5	19.6	24.2
	GaitGAN [12]	9.8	15.2	24.8	25.0	24.7	19.9	22.7	24.5	27.7	18.0	11.9	20.4
	GaitGANv2 [13]	20.7	23.1	26.6	30.8	28.2	23.0	24.4	27.4	24.2	21.9	16.0	24.2
	DV-GEIs	30.2	43.3	43.4	43.1	43.6	41.9	40.0	40.3	41.4	38.7	29.9	**39.6**
74	GaitSet-GEI [22]	-	-	-	-	-	-	-	-	-	-	-	40.8
	DV-GEIs	36.4	51.5	51.1	49.1	44.9	46	47.7	46.2	44.2	41.2	32.6	**44.6**

normal walking condition (NM). In addition, we can see the proposed method can also achieve a better performance on carrying a bag (BG) and wearing a coat (CL) condition. These methods [11, 12, 13] aim to transform any GEI into the side GEI, while ours synthesized dense view samples aim to cover the whole view space. The comparison shows that gait view space covering has a positive impact on solving the cross-view challenge.

The method of GaitSet [22], GaitNet [24] and GaitPart [23] can achieve very high performance when it uses the human walking image sequence as an input feature. But the performance of GaitSet [22] would be decreased dramatically when using GEI as an input feature (drop from 95.0% to 80.4%). One reason for this is that the human silhouette sequence has richer information than GEI.

Here, we compare our method with GaitSet-GEI [22] which uses GEI as an input feature and be fed into the GaitSet, not compare with those based on human walking image sequence methods. Because our method is based on the GEI template, so it is fair to compare with those based on the GEI template method. The comparison as shown in Table 9, the number of training subjects is 74, and the rest of the subjects is for the testing. The comparison between our method and GaitSet-GEI [22] shows that our performance is better than that of GaitSet-GEI [22] on the normal walking condition on clothing condition. In addition, the comparison between DV-GEIs-pre [4] and DV-GEIs illustrates that center loss in the monitor can improve the discriminative capability of synthesized images and boost recognition rate.

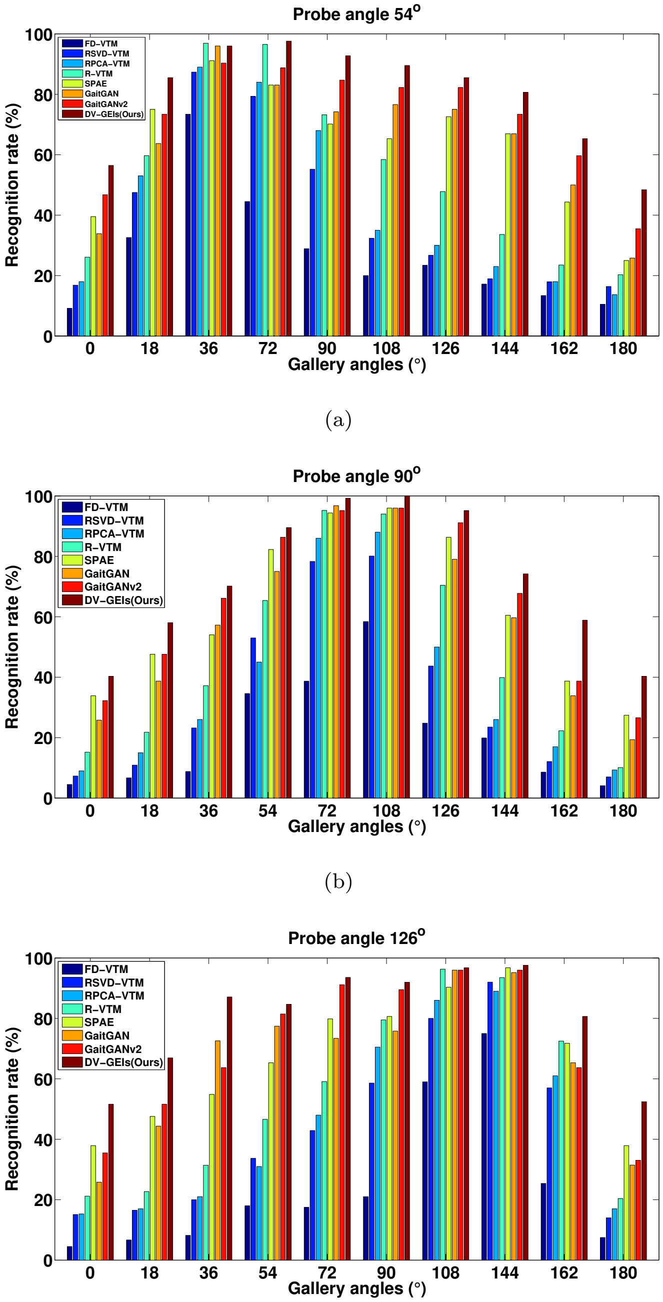
4.9. Comparison with VTM methods

To illustrate the contribution of space covering on the cross-view problem, we also compare with view transformation model (VTM) methods when probe angle at 54°, 90° and 126° respectively, as shown in Figure 11. Compared methods are FD-VTM [35], RSVD-VTM [8], RPCA-VTM [9], R-VTM [36], SPAE [11], GaitGAN [12] and GaitGANv2 [13]. These methods all try to transform the gait features from one view to another, and our method is to synthesize more samples to cover the entire view space.

From Figure 11, it is clear that the performance of the proposed DV-GEIs method is better than other methods, especially when the view angle difference between gallery and probe is large. The comparison shows that dense view space covering can handle larger viewpoint changes well. Besides, when the viewpoint changes are not large enough, this method can also significantly improve the recognition rate.

4.10. Experimental results on OU-ISIR dataset

We also do some experiments on OU-ISIR dataset [3]. The original view of OU-ISIR dataset [3] has 4 views, 55°, 65°, 75° and 85°, respectively. Because of the start angle from 55°, and end with 85° on the OU-ISIR dataset. So



(a)

(b)

(c)

Figure 11: Comparisons with view transformation model methods when probe view at (a)54°, (b)90° and (c)126° condition.

we synthesize DV-GEIs data from 55° to 85° with 1° interval between the two closest views. The result of experiments on the OU-ISIR dataset is shown in Table 10. There are 16 recognition rates in this table, each row represents an angle of the gallery group, and each column represents an angle of the probe group. the overall performance is the best when the probe angle is the same as the galley angle.

Table 10: Cross-view recognition result on OU-ISIR dataset.

Probe angle	Gallery angle			
	55°	65°	75°	85°
55°	97.0	95.2	94.0	90.1
65°	95.7	96.7	95.0	94.3
75°	92.2	96.0	97.5	95.3
85°	91.6	95.8	97.0	97.8

We compare our recognition rate with OG-GEIs, DeepCNN [21], GaitGANv2[13] and DV-GEIs-pre [4], as shown in Table 10. From this table, we can see that the performance of DV-GEI is better than the benchmark reported by the author of the dataset [21, 13] when the probe angle is 55° and 75°. It should be noticed that DeepCNN [21] is operating on the sequence of human silhouettes instead of GEI, which contains richer information. But our performance can still achieve high performance. In addition, our new result is also better than our previous work DV-GEIs-pre [4] when probe angle is 55°, 75° and 85°, which shows again that center loss in the monitor further improve the quality of synthesized images and recognition rate. In addition, the accuracy of view covering DV-GEIs is better than that of OG-GEIs (trained by original GEIs set), which shows again that gait view space covering has a positive impact on solving the cross-view problem.

Table 11: Comparison with other approaches with average accuracy (%) on OU-ISIR dataset. Excluding identical view cases. Training data of OG-GEIs model is employing original GEIs set. Training data of DV-GEIs model is employing the view covering DV-GEIs set.

Methods	Probe angle			
	55°	65°	75°	85°
DeepCNN [21]	91.6	92.3	92.4	**94.8**
GaitGANv2 [13]	91.9	95.0	94.4	94.6
OG-GEIs	92.0	93.8	94.0	93.8
DV-GEIs-pre [4]	92.6	**95.1**	94.4	94.6
DV-GEIs	**93.1**	95.0	**94.5**	**94.8**

5. Conclusions and Future Work

In this paper, view space covering with dense sampling is introduced to improve the gait recognition performance. We synthesize GEI samples with various views, ranging from 0° to 180° with 1° interval to fill the GEI view space. DV-GEIs is synthesized by the proposed DV-GAN network. It consists of three parts, generator, discriminator, and monitor. The new monitor not only can maintain human identification and view information very well but also improves the discriminative capability of synthesized images. The experimental results show that dense view space covering can lighten the burden of view-invariant extraction for CNN and make the feature more discriminative compared with the original datasets.

With the development of synthesized sample technology, the generated images will be more and more realistic. We believe the idea of view covering synthesized samples by DV-GAN not only can enhance robustness to view variation but also deal with other variations, such as synthesizing different types of bags and different types of clothes. Such generalization has the potential to make a great contribution to gait research with the simulation of real environments. Eventually, the generalization may help to further improve the development of gait recognition technology and enhance practical applications.

In the future we will explore subspace indexing on Grassmann manifold [37] framework to have an even lighter and faster inference time engine for gait recognition, with DV-GAN as a samples generator.

Acknowledgment

This work was supported in part by the U.S. Air Force Office of Scientific Research under the DDIP Program.

References

References

[1] S. Yu, D. Tan, T. Tan, A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition, in: the 18th International Conference on Pattern Recognition, 2006, pp. 441–444.

[2] 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, IPSJ Transactions on Computer Vision and Applications 10 (1) (2018) 4.

[3] 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 Transactions on Information Forensics and Security 7 (5) (2012) 1511–1521.

[4] R. Liao, W. An, S. Yu, Z. Li, Y. Huang, Dense-view geis set: View space covering for gait recognition based on dense-view gan, in: 2020 IEEE International Joint Conference on Biometrics, IEEE, 2020, pp. 1–9.

[5] J. Han, B. Bhanu, Individual recognition using gait energy image, IEEE transactions on pattern analysis and machine intelligence 28 (2) (2005) 316–322.

[6] C. Wang, J. Zhang, L. Wang, J. Pu, X. Yuan, Human identification using temporal information preserving gait template, IEEE transactions on pattern analysis and machine intelligence 34 (11) (2011) 2164–2176.

[7] H. Hu, Enhanced gabor feature based classification using a regularized locally tensor discriminant model for multiview gait recognition, IEEE transactions on circuits and systems for video technology 23 (7) (2013) 1274–1286.

[8] W. Kusakunniran, Q. Wu, H. Li, J. Zhang, Multiple views gait recognition using view transformation model based on optimized gait energy image, in: 2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops, IEEE, 2009, pp. 1058–1064.

[9] S. Zheng, J. Zhang, K. Huang, R. He, T. Tan, Robust view transformation model for gait recognition, in: IEEE International Conference on Image Processing, 2011, pp. 2073–2076.

[10] M. Hu, Y. Wang, Z. Zhang, J. J. Little, D. Huang, View-invariant discriminative projection for multi-view gait-based human identification, IEEE Transactions on Information Forensics & Security 8 (12) (2013) 2034–2045.

[11] S. Yu, Q. Wang, L. Shen, Y. Huang, View invariant gait recognition using only one uniform model, in: International Conference on Pattern Recognition, 2017, pp. 889–894.

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

[13] S. Yu, R. Liao, W. An, H. Chen, E. B. G. Reyes, Y. Huang, N. Poh, GaitGANv2: Invariant gait feature extraction using generative adversarial networks, Pattern recognition 87 (2019) 179–189.

[14] X. Ben, C. Gong, P. Zhang, X. Jia, Q. Wu, W. Meng, Coupled patch alignment for matching cross-view gaits, IEEE Transactions on Image Processing 28 (6) (2019) 3142–3157.

[15] X. Ben, C. Gong, P. Zhang, R. Yan, Q. Wu, W. Meng, Coupled bilinear discriminant projection for cross-view gait recognition, IEEE Transactions on Circuits and Systems for Video Technology 30 (3) (2019) 734–747.

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

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

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

[19] R. Liao, S. Yu, W. An, Y. Huang, A model-based gait recognition method with body pose and human prior knowledge, Pattern Recognition 98 (2020) 107069.

[20] W. An, S. Yu, Y. Makihara, X. Wu, C. Xu, Y. Yu, R. Liao, Y. Yagi, Performance evaluation of model-based gait on multi-view very large population database with pose sequences, IEEE Transactions on Biometrics, Behavior, and Identity Science.

[21] Z. Wu, Y. Huang, L. Wang, X. Wang, T. Tan, A comprehensive study on cross-view gait based human identification with deep cnns, IEEE Transactions on Pattern Analysis & Machine Intelligence 39 (2) (2017) 209–226.

[22] H. Chao, Y. He, J. Zhang, J. Feng, Gaitset: Regarding gait as a set for cross-view gait recognition, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2019, pp. 8126–8133.

[23] C. Fan, Y. Peng, C. Cao, X. Liu, S. Hou, J. Chi, Y. Huang, Q. Li, Z. He, Gaitpart: Temporal part-based model for gait recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 14225–14233.

[24] Z. Zhang, L. Tran, X. Yin, Y. Atoum, X. Liu, J. Wan, N. Wang, Gait recognition via disentangled representation learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 4710–4719.

[25] Y. Ji, Y. Yang, H. T. Shen, T. Harada, View-invariant action recognition via unsupervised attention transfer (uant), Pattern Recognition 113 (2021) 107807.

[26] Y. Ji, Y. Yang, F. Shen, H. T. Shen, W.-S. Zheng, Arbitrary-view human action recognition: A varying-view rgb-d action dataset, IEEE Transactions on Circuits and Systems for Video Technology.

[27] J. Chen, J. Chen, H. Chao, M. Yang, Image blind denoising with generative adversarial network based noise modeling, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3155–3164.

[28] X. Qian, Y. Fu, T. Xiang, W. Wang, J. Qiu, Y. Wu, Y.-G. Jiang, X. Xue, Pose-normalized image generation for person re-identification, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 650–667.

[29] X. Hou, L. Shen, K. Sun, G. Qiu, Deep feature consistent variational autoencoder, in: 2017 IEEE Winter Conference on Applications of Computer Vision, 2017, pp. 1133–1141.

[30] I. J. Goodfellow, J. Pougetabadie, M. Mirza, B. Xu, D. Wardefarley, S. Ozair, A. Courville, Y. Bengio, Z. Ghahramani, M. Welling, Generative adversarial nets, Advances in Neural Information Processing Systems 3 (2014) 2672–2680.

[31] P. Isola, J.-Y. Zhu, T. Zhou, A. A. Efros, Image-to-image translation with conditional adversarial networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1125–1134.

[32] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, Springer, 2015, pp. 234–241.

[33] Y. Wen, K. Zhang, Z. Li, Y. Qiao, A discriminative feature learning approach for deep face recognition, in: European conference on computer vision, 2016, pp. 499–515.

[34] L. Pan, R. Liao, Z. Li, S. S. Bhattacharyya, Dynamic, data-driven hyperspectral image classification on resource-constrained platforms, in: International Conference on Dynamic Data Driven Application Systems, Springer, 2020, pp. 320–327.

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

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

[37] X. Wang, Z. Li, D. Tao, Subspaces indexing model on grassmann manifold for image search, IEEE Trans. Image Process. 20 (9) (2011) 2627–2635. `doi:10.1109/TIP.2011.2114354`.
URL `https://doi.org/10.1109/TIP.2011.2114354`