2024 18th International Conference on Automatic Face and Gesture Recognition (FG)

ViewDiffGait: View Pyramid Diffusion for Gait Recognition

Rijun Liao[1], Zhu Li[1], Shuvra S. Bhattacharyya[2], George York[3]
[1]Department of Computer Science and Electrical Engineering, University of Missouri-Kansas City, MO, USA.
[2] Department of Electrical and Computer Engineering and UMIACS,
University of Maryland, College Park, MD, USA.
[3] Department of Electrical and Computer Engineering and UAS Research Center,
US Air Force Academy, Colorado Springs, CO, USA.

2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG) | 979-8-3503-9494-8/24/$31.00 ©2024 IEEE | DOI: 10.1109/FG59268.2024.10581996

***Abstract*— View transformation is crucial for gait recognition. Most existing methods use a view transformation models (VTM) or generative models (VAE or GAN) to achieve transformation. These approaches commonly adopt a paradigm of transforming a gait feature from one view to another. However, most existing methods attempt to use a single or multiple, large-view transformation model to directly transform a source view image to the target view image. Such transformations usually suffer from precision problems under large viewpoint variations due to the lack of fine view prediction. To overcome this challenge, we introduce a novel framework, *ViewDiffGait*, employing a view pyramid structure and diffusion models. *ViewDiffGait* is formulated as an iterative refinement generation task in a biologically interpretable way, capable of generating more accurate lateral view images from coarse to fine. Unlike the typical diffusion model that directly adds and removes Gaussian noise in the original image, the *ViewDiffGait* diffusion process involves a view pyramid structure to capture fine view transformations. The diffusion process adds view noise from the pyramid top to the bottom, while the denoising process removes view noise from the pyramid bottom to the top. We conducted extensive experiments on the CASIA-B and OUMVLP datasets, demonstrating that *ViewDiffGait* can generate more realistic images, remove variations effectively, and lead to high performance in real applications.**

keywords: Gait Recognition, Diffusion Model, View Pyramid, View Noise Removing

I. Introduction

A. Motivation

With the outbreak of the novel coronavirus 2019 (COVID-19), it has become imperative to develop biometric technologies to address various concerns arising from a similar virus event. Biometric technologies usually have two categories, contact and non-contact biometrics. Contact biometrics such as fingerprints and palm prints will obviously speed up the spread of the virus. For non-contact biometric technology, face recognition [20] is one of the mature biometric technologies. But identifying subjects becomes challenging when people are wearing masks. Iris recognition also faces challenges when wearing anti-virus glasses. What is more, due to the close-range collection of iris data, it also brings the risk of personnel touching the device.

Compared with the above biometrics, gait biometric has the following advantages, 1) long-distance human identification 2) no user action and cooperation required. It is particularly suitable for impeding the spread of COVID-19, monitoring people [19], video surveillance, crime prevention, and forensic identification.

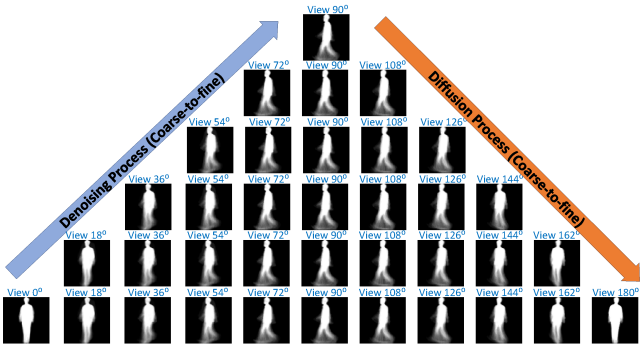

Fig. 1. *ViewDiffGait* employs a view pyramid structure and diffusion models. It involves adding view noise from the pyramid top to the bottom during diffusion and denoising by removing view noise from the bottom to the top.

However, automatic gait recognition faces challenges due to various sources of variation, including viewpoint, clothing, and carried objects, potentially compromising recognition accuracy. Among these, view angle stands out as a common challenge, given the uncontrollable walking directions in real-world scenarios—this forms the central focus of our work.

To enhance robustness against view-angle variation, existing methods often employ View Transformation Model (VTM) [14], [6], [24] to transform gait features from one view to another. While effective in cross-view recognition, VTM are limited as each model can only handle a specific view angle. Some recent approaches seek view invariance using a single generative model (VAE [21] or GAN [22], [7], [8]), as shown in Figure 2. They achieve improved performance but still face precision issues, particularly under large viewpoint variations. This is because of the lack of fine view prediction under large viewpoint variations. To address these challenges, we draw inspiration from the state-of-the-art diffusion models and introduce a novel framework, *ViewDiffGait* , which incorporates a view pyramid structure for an iterative refinement generation task.

B. Contributions

In summary, our major contributions are:

- A novel framework *ViewDiffGait* for gait recognition is proposed based on the denoising diffusion model. This framework, formulated as an iterative refinement generation task, interprets biological processes, enabling

979-8-3503-9494-8/24/$31.00 ©2024 IEEE

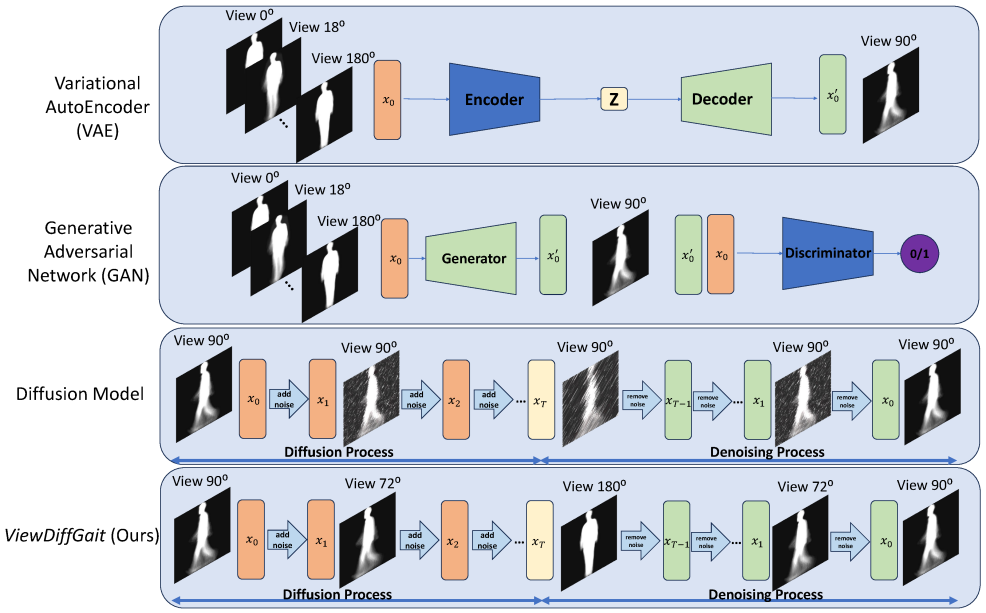

Fig. 2. Structure of three generative models. Existing methods usually lack fine view prediction and lead to suffering from precision problems under large viewpoints. Proposed *ViewDiffGait* is formulated as an iterative refinement generation task in a biologically interpretable way, capable of generating more accurate lateral view images from coarse to fine.

the generation of more accurate lateral view images from coarse to fine. To the best of our knowledge, we are the first ones to use the diffusion model in gait recognition.

- Unlike the traditional diffusion model which directly adds and removes Gaussian noise in the image (Figure 2), *ViewDiffGait* 's diffusion process incorporates a view pyramid structure. This approach involves adding view noise from the pyramid top to the bottom during diffusion, and denoising by removing view noise from the bottom to the top, as shown in Figure 1.
- We evaluated our proposed method on both popular CASIA-B dataset [23] and OUMVLP dataset [15] and achieved a high recognition rate in comparison to recent advanced methods. Our experiment results demonstrate the significant improvement of the proposed method in varied situations, validating its potential for practical gait recognition applications.

II. Related Works

In this section, we present a concise overview of the development and categories of existing gait recognition models, along with on the structure of three generative models.

Gait recognition can be broadly categorized into two groups: model-based [9], [11], [10] and appearance-based [2], [4], [3]. Model-based approaches extract features by modeling the human body structure and analyzing movement patterns of different body parts, often utilizing human skeletons as initial input features. Examples include PTSN [9], PoseGait [11], and PoseMapGait [10]. However, these methods may suffer from low performance due to limited input information. In contrast, appearance-based methods extract human silhouettes as initial input features, achieving higher recognition rates.

Appearance-based methods can be further divided into two categories: template-based approaches [22], [7], [8] and sequence-based approaches [2], [4], [3]. Sequence-based approaches use a video clip of human silhouettes as input data, such as GaitPart [4], GaitSet [2], GaitGL [12], LidarGait [16], and OpenGait [3]. While achieving high performance, these methods are often time-consuming in the reference stage due to the need for a sequence of human silhouettes as input data. In contrast, template-based approaches are faster in the reference stage, requiring only one template feature instead of a video clip. Gait Energy Image (GEI) templates, produced by averaging all silhouettes in a single gait cycle, are popular features in template-based approaches due to their low computational cost and relatively high recognition rate.

To design a fast model in the reference stage and improve real-world applications, this paper primarily focuses on GEI-based research. Previous studies extensively explored GEI templates, introducing VTM models to enhance robustness against view-angle variation. Makihara et al. [14] designed FD-VTM, operating in the frequency domain. RSVD-VTM [6] operates in the spatial domain, using reduced singular value decomposition (SVD) and linear discriminant analysis (LDA) to construct a VTM and generate an optimal GEI feature vector. Zheng et al. [24] achieved a robust VTM via RPCA for view-invariant feature extraction, with the inspired by the robust principal component analysis (RPCA) for feature extraction,

Most VTM-related methods [14], [6], [24] can only transform one specific view angle to another, requiring numerous models. With the development of generative models, some researchers use a single model to transform any view angle

based on auto Variational AutoEncoder [21] and Generative Adversarial Networks [22], [7], [8]. While achieving improved performance, these methods still face precision issues under large viewpoint variations due to the lack of fine view prediction.

Inspired by the diffusion model [18], [5], we propose a novel framework, *ViewDiffGait*, employing a view pyramid structure and diffusion models. *ViewDiffGait* is formulated as an iterative refinement generation task in a biologically interpretable way, generating more accurate lateral view images from coarse to fine.

The structure of three generative models is illustrated in Figure 2. VAE structure encompasses one encoder and one decoder. GAN structure involves a generator and a discriminator. However, VAE has the weakness of limited sample quality, and GAN has the weakness of mode collapse. In contrast, the diffusion model structure includes diffusion and denoising processes. Compared with VAE and GAN, diffusion models directly estimate data likelihood, avoiding mode collapse, providing a clear generative process, and offering more stable training.

The diffusion model can directly add and remove noise from the GEI image, as shown in Figure 2. However, this strategy is similar to existing VAE-based or GAN-based methods, lacking fine view prediction under large viewpoint variations. This paper introduces a view pyramid strategy to achieve fine view prediction. This view pyramid strategy perfectly matches the structure of diffusion models, involving adding view noise from the pyramid top to the bottom during diffusion and denoising by removing view noise from the bottom to the top. Considering the advantages and structure of diffusion models, the diffusion model is used in our proposed *ViewDiffGait* framework over other generative models.

III. Methodology

Unlike the typical diffusion models [18], [5] which involve iteratively adding random Gaussian noise for T steps, and iteratively denoising the Gaussian noise, as shown in Figure 2 Diffusion Model. The proposed *ViewDiffGait* framework introduces a paradigm of view pyramid noise adding and removing, which is an innovative approach to iterative refinement generation tasks, particularly for generating more realistic images resembling lateral view images. *ViewDiffGait* includes diffusion process and denoising process. The diffusion process is to add view noise to form the view pyramid, while denoising process is to remove view noise from view paramid, as shown in Figure 1, 3.

A. View Pyramid and View Noise

View Pyramid: It plays a crucial role in the *ViewDiffGait* framework, serving as a mechanism for view transformation. Specifically designed for datasets such as CASIA-B [23], the view pyramid encompasses a combination of view transformations. In this case, the input view is set at 90°, and the output views vary from 0° to 180°. The interval degree between each nearby pair of views, denoted as $x_\alpha \rightarrow x_\beta$, is either 0° or 18°. The diagram of view pyramid as shown in Figure 1, mathematical expression as followings:

$$
\begin{aligned}
&\mathrm{x}_{90^\circ} \rightarrow x_{72^\circ} \rightarrow x_{54^\circ} \rightarrow x_{36^\circ} \rightarrow x_{18^\circ} \rightarrow x_{0^\circ} \\
&x_{90^\circ} \rightarrow x_{90^\circ} \rightarrow x_{72^\circ} \rightarrow x_{54^\circ} \rightarrow x_{36^\circ} \rightarrow x_{18^\circ} \\
&x_{90^\circ} \rightarrow x_{90^\circ} \rightarrow x_{90^\circ} \rightarrow x_{72^\circ} \rightarrow x_{54^\circ} \rightarrow x_{36^\circ} \\
&x_{90^\circ} \rightarrow x_{90^\circ} \rightarrow x_{90^\circ} \rightarrow x_{90^\circ} \rightarrow x_{72^\circ} \rightarrow x_{54^\circ} \\
&x_{90^\circ} \rightarrow x_{90^\circ} \rightarrow x_{90^\circ} \rightarrow x_{90^\circ} \rightarrow x_{90^\circ} \rightarrow x_{72^\circ} \\
&x_{90^\circ} \rightarrow x_{90^\circ} \rightarrow x_{90^\circ} \rightarrow x_{90^\circ} \rightarrow x_{90^\circ} \rightarrow x_{90^\circ} \\
&x_{90^\circ} \rightarrow x_{90^\circ} \rightarrow x_{90^\circ} \rightarrow x_{90^\circ} \rightarrow x_{90^\circ} \rightarrow x_{108^\circ} \\
&x_{90^\circ} \rightarrow x_{90^\circ} \rightarrow x_{90^\circ} \rightarrow x_{90^\circ} \rightarrow x_{108^\circ} \rightarrow x_{126^\circ} \\
&x_{90^\circ} \rightarrow x_{90^\circ} \rightarrow x_{90^\circ} \rightarrow x_{108^\circ} \rightarrow x_{126^\circ} \rightarrow x_{144^\circ} \\
&x_{90^\circ} \rightarrow x_{90^\circ} \rightarrow x_{108^\circ} \rightarrow x_{126^\circ} \rightarrow x_{144^\circ} \rightarrow x_{162^\circ} \\
&x_{90^\circ} \rightarrow x_{108^\circ} \rightarrow x_{126^\circ} \rightarrow x_{144^\circ} \rightarrow x_{162^\circ} \rightarrow x_{180^\circ}
\end{aligned}
$$

The purpose of the view pyramid diffusion is to facilitate a step-by-step, view-wise transformation in a coarse-to-fine manner. This approach aligns with an iterative refinement generation task, contributing to a biologically interpretable progression. The ultimate goal is to generate more realistic images that closely resemble lateral view images, enhancing the quality and interpretability of the generated content.

View Noise Definition: In the diffusion process from x_α to x_β, the generation of x_β involves the addition of $noise_{\beta-\alpha}$ to x_α. This step, illustrated in Figure 4, is a departure from the typical diffusion model of adding random Gaussian noise. The $noise_{\beta-\alpha}$ is determined by the difference between a series of view β images and a series of view α images, both belonging to the same identity.

The choice of utilizing view-specific noise and incorporating it into the diffusion process is a meaningful contribution of the *ViewDiffGait* diffusion process. This departure from traditional noise addition allows for a more tailored and context-aware generation of images, resulting in a finer and more accurate representation of lateral views.

B. ViewDiffGait *Diffusion Process*

The *ViewDiffGait* diffusion process presents a novel view pyramid diffusion to iterative refinement adding view noise in a biologically interpretable manner. It focuses on view-wise, coarse-to-fine transformation steps.

Provide a GEI image, denoted as x_0 with 90° view. The diffusion process involves iteratively introducing view noise for a total of T steps. The forward trajectory, which corresponds to initiating the diffusion process from the gait distribution x_0 and proceeding through T diffusion steps, is described by the following equation:

$$q(x_t|x_{t-1}) := \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I) \tag{1}$$

Here, t represents the step number, and β_t signifies the variance schedule. This procedure, known as the "diffusion process" or "forward process," establishes a Markov chain that systematically introduces view noise to the data in accordance with the variance schedule β_t.

For clarity, starting with the initial GEI x_0, one random view noise is added to produce x_1. Subsequently, noise is

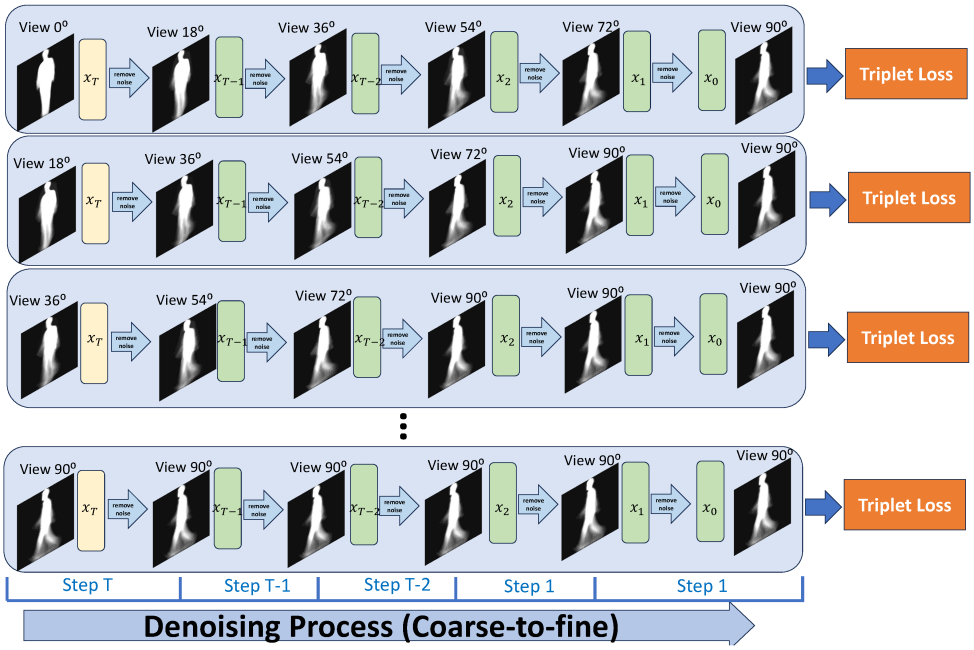

Fig. 3. The overview of the proposed method framework *ViewDiffGait*. It constructs a diffusion process and a denoising process. The diffusion process is to add a random combination view set in the training phase, while denoising process as an iterative refinement generation task and generate real image (Lateral View) from coarse to fine. And the triplet loss is to increase the inter-gait distance and to reduce the intra-gait distance.

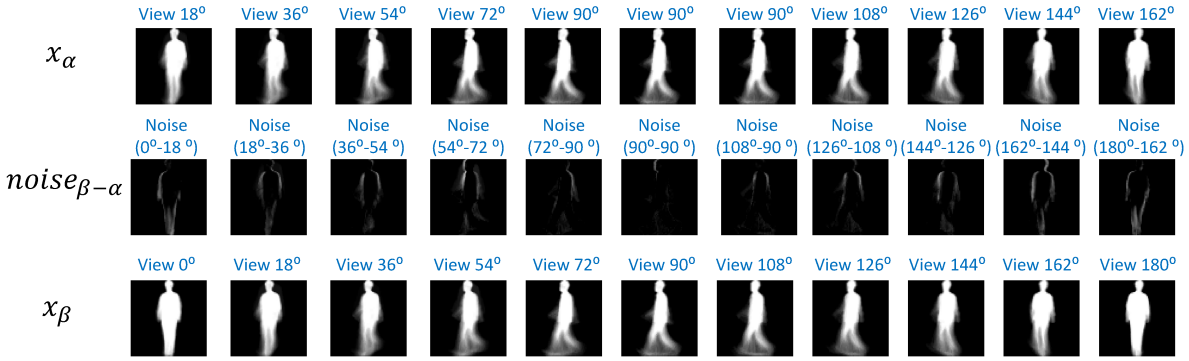

Fig. 4. The View Noise Diffusion stands as a distinctive feature, deviating from the typical Gaussian Noise Diffusion approach. This module integrates the view pyramid structure into the diffusion process, strategically adding and removing view noise in a hierarchical manner.

added again to obtain x_2, and this process is repeated for a total of T steps, as shown in the view pyramid. As a result, x_0 evolves into x_T, represented as:

$$x_0 \rightarrow x_1 \rightarrow x_2 \cdots \rightarrow x_{T-1} \rightarrow x_T \tag{2}$$

C. ViewDiffGait *Denoising Process*

The backward denoising process aims to transfer x_T with any view angle into target image x_0 by denoising process, the equation is as follows.

$$x_T \rightarrow x_{T-1} \cdots \rightarrow x_2 \rightarrow x_1 \rightarrow x_0 \tag{3}$$

The view pyramid will be reversed with the process of denoising, as shown in followings:

$$\mathrm{x}_{90^\circ} \leftarrow x_{72^\circ} \leftarrow x_{54^\circ} \leftarrow x_{36^\circ} \leftarrow x_{18^\circ} \leftarrow x_{0^\circ}$$
$$x_{90^\circ} \leftarrow x_{90^\circ} \leftarrow x_{72^\circ} \leftarrow x_{54^\circ} \leftarrow x_{36^\circ} \leftarrow x_{18^\circ}$$
$$x_{90^\circ} \leftarrow x_{90^\circ} \leftarrow x_{90^\circ} \leftarrow x_{72^\circ} \leftarrow x_{54^\circ} \leftarrow x_{36^\circ}$$
$$x_{90^\circ} \leftarrow x_{90^\circ} \leftarrow x_{90^\circ} \leftarrow x_{90^\circ} \leftarrow x_{72^\circ} \leftarrow x_{54^\circ}$$
$$x_{90^\circ} \leftarrow x_{90^\circ} \leftarrow x_{90^\circ} \leftarrow x_{90^\circ} \leftarrow x_{90^\circ} \leftarrow x_{72^\circ}$$
$$x_{90^\circ} \leftarrow x_{90^\circ} \leftarrow x_{90^\circ} \leftarrow x_{90^\circ} \leftarrow x_{90^\circ} \leftarrow x_{90^\circ}$$
$$x_{90^\circ} \leftarrow x_{90^\circ} \leftarrow x_{90^\circ} \leftarrow x_{90^\circ} \leftarrow x_{90^\circ} \leftarrow x_{108^\circ}$$
$$x_{90^\circ} \leftarrow x_{90^\circ} \leftarrow x_{90^\circ} \leftarrow x_{90^\circ} \leftarrow x_{108^\circ} \leftarrow x_{126^\circ}$$
$$x_{90^\circ} \leftarrow x_{90^\circ} \leftarrow x_{90^\circ} \leftarrow x_{108^\circ} \leftarrow x_{126^\circ} \leftarrow x_{144^\circ}$$
$$x_{90^\circ} \leftarrow x_{90^\circ} \leftarrow x_{108^\circ} \leftarrow x_{126^\circ} \leftarrow x_{144^\circ} \leftarrow x_{162^\circ}$$
$$x_{90^\circ} \leftarrow x_{108^\circ} \leftarrow x_{126^\circ} \leftarrow x_{144^\circ} \leftarrow x_{162^\circ} \leftarrow x_{180^\circ}$$

Specifically, given the prior probability $q(x_{t-1}|x_t)$, aim to learn the posterior probability $p_\theta(x_{t-1}|x_t)$,

$$p_\theta(x_{t-1}|x_t) := \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \tag{4}$$

where the μ_θ and Σ_θ represent image mean and variance, respectively. Experimentally, set the $\Sigma_\theta(x_t, t) = \sigma_t^2 I = \beta_t$.

The x_{t-1} can be get from x_t.

In the typical denoising process, several different methods [13] are likely to parameterize p_θ, including the prediction of mean value μ_θ, the prediction of the noise $\mathbf{z}$, and the prediction of variance σ_t. In our proposal, we use a neural network $f_\theta(x_t, t)$ to predict x_{t-1}, instead of directly predicting the μ_θ. To optimize the model, a mean squared error loss can be used to match $f_\theta(x_t, t)$ and x_{t-1}: $L_t = ||f_\theta(x_t, t) - x_{t-1}||^2$. The step t is randomly selected at each training iteration. Finally, the generated images x_0 would be added to the triplet loss to increase the inter-gait distance and reduce the intra-gait distance.

In our experiment, we utilized the Euclidean distance measure for triplet loss evaluation with a margin of 1, ensuring effective discrimination between anchor, positive, and negative instances. Employing two NVIDIA GeForce GTX 1080 with 12GB of memory expedited computations. We set T as 5 on the training with CASIA-B dataset.

IV. Experiments

A. Datasets

CASIA-B[23]: A widely utilized gait dataset, CASIA-B comprises 124 subjects, each contributing 10 sequences. These sequences encompass 6 instances of normal walking (NM), 2 sequences involving walking with a bag (BG), and 2 sequences involving walking with a coat (CL). Each sequence captures gait from 11 distinct views $\{0°, 18°, \cdots, 180°\}$, as visually represented in Figure 4. The experimental setup for training and testing involves allocating the initial 62 subjects to the training set, while the remaining subjects constitute the test set. Within the test set, the gallery set is formed by retaining the last 4 normal walking sequences for each subject. The probe set comprises the remaining 2 normal walking sequences, as shown in Table I.

TABLE I
EXPERIMENTAL SETTING ON CASIA-B DATASET. NM: NORMAL WALKING.

Training	Testing	
	Gallery Set	Probe Set
ID: 001-062 Seqs: NM01-NM06	ID: 063-124 Seqs: NM01-NM04	ID: 063-124 Seqs: NM05-NM06

OU-MVLP [15]: Comprising a diverse population of 10,307 individuals (5,114 males and 5,193 females) spanning a wide age range from 2 to 87 years, the OU-MVLP dataset captures gait images from 14 view angles, covering the ranges of 0°-90° and 180°-270° at intervals of 15-deg azimuth angles . Each subject is associated with 28 sequences, incorporating 2 sequences (indexed as Seq#00 and Seq#01) for each of the 14 camera views. The first 5153 subjects contribute to the training set, while the remaining 5154 subjects are designated for testing. During testing, sequences with index Seq#01 are designated as the gallery set, while those with index Seq#00 form the probe set, as shown in Table II.

TABLE II
EXPERIMENTAL SETTING ON THE OU-MVLP [15] DATASET.

Training	Test	
	Gallery Set	Probe Set
5153 subjects Seq#00,Seq#01	5154 subjects Seq#01	5154 subjects Seq#00

B. Experimental Analysis on CASIA-B Dataset

In this subsection, we provide a comprehensive comparison between the proposed *ViewDiffGait* and recent advanced methods, encompassing three typical categories: GEI-based methods, skeleton-based methods, and silhouette-based methods, as outlined in Table III, III.

Comparison with Skeleton-based Methods: We evaluate the performance against skeleton-based methods, including PTSN [9], PTSN-3D [1], PoseGait [11], and PoseMapGait [10]. From Table III, it is evident that our proposed method achieves the highest mean accuracy (87.5%). This surpasses the best existing method, PoseMapGait [10], by a substantial margin of 11.8% (75.7%).

Comparison with GEI-based Methods: The comparison generative model methods, including SPAE [21], GaitGANv2 [22], DV-GEIs-pre [8], and DV-GEIs [7], are based on existing generative models (VAE or GAN). The results in Table III demonstrate that our proposed method achieves the highest mean accuracy (87.5%). This outperforms the best existing method, DV-GEIs [10], by a significant margin of 11.1% (76.4%). This showcases the superiority of the diffusion generative model over traditional VAE and GAN approaches.

Comparison with Silhouette-based Methods: Silhouette-based methods, represented by GaitSet [2] and GaitPart [4], are comparable in performance with our proposed method. The probe view is under $36°$, $54°$, $72°$, the performance gap is minimal, as shown in Figure IV. It's important to note that silhouette-based methods typically utilize a sequence of silhouettes as input data, usually employing 30 frames, while our proposed method leverages only one gait feature frame. Notably, when the silhouette-based method GaitSet [2] uses the same input (GEI) as ours, its accuracy decreases to 80.4%, while our method consistently achieves high accuracy at 89.9%.

Moreover, we compare the reference time with GaitSet [2] and GaitPart [4], as shown in Table IV. Test environment in a single GeForce GTX 1080, 12GB, CUDA version: 10.2, PyTorch vesrion: 1.9.1. The results clearly indicate that our proposed method exhibits faster processing speed than silhouette-based methods. This underscores that *ViewDiffGait* not only achieves high performance in cross-view conditions but also operates at a faster pace. This balance between performance and speed positions *ViewDiffGait* as a promising solution for real-world gait recognition applications.

C. Ablation Study

In this section, we systematically examine the integral components of our proposed gait recognition framework – the View Pyramid Module and the View Noise Diffusion Module. Through a detailed analysis, we elucidate the specific contributions and benefits that these modules bring to the overall system.

Analysis of View Pyramid Module: The View Pyramid Module is designed to capture fine view transformations by employing a hierarchical pyramid structure. Each level of the pyramid refines the view prediction, allowing for a gradual and detailed adjustment from coarse to fine views. This intricate mechanism proves to be invaluable, especially under large viewpoint variations. The absence of the View Pyramid Module, as demonstrated by the *DiffGait* method, results in a direct transformation without the intermediary steps provided by the pyramid ($T = 1$ in Equation 2). From Figure 5, we can see the significance of the View Pyramid Module becomes apparent in its ability to enhance precision and generate more accurate lateral view images. It acts as a sophisticated guidance system, ensuring that the gait feature transformation is not only efficient but also attuned to the subtleties of different viewpoints.

Analysis of View Noise Diffusion Module: The View Noise Diffusion (defined in Section III-A) Module stands as a distinctive feature in our framework, deviating from the typical Gaussian Noise Diffusion approach. This module integrates the view pyramid structure into the diffusion process, strategically adding and removing view noise in a hierarchical manner. This dynamic diffusion mechanism plays a pivotal role in refining the gait image generation task. In contrast to conventional methods that directly add and remove Gaussian noise to and from the original image, as demonstrated by the *DiffGait-Gaussian* method. The comparison of result can be seen in From Figure 5, showcases a remarkable improvement in performance with view noise diffusion. The View Noise Diffusion Module not only aids in denoising but also facilitates the generation of more realistic images by capturing fine view transformations. This nuanced approach ensures that the recognition framework is robust and effective, particularly in scenarios characterized by significant viewpoint variations.

In essence, both the View Pyramid Module and the View Noise Diffusion Module contribute substantially to the success of our proposed *ViewDiffGait* framework. They collectively address challenges associated with precision, viewpoint variations, and overall recognition performance, making our approach a robust and sophisticated solution in the realm of gait recognition.

D. Experimental Analysis on OU-MVLP Dataset

In order to robustly demonstrate the effectiveness of our proposed *ViewDiffGait*, we meticulously conduct an extensive experimental analysis, pitting it against state-of-the-art methods on the challenging OU-MVLP dataset. Our evaluation spans two key categories: GEI-based methods and silhouette-based methods, each meticulously compared and contrasted in Table V and Table VI.

Comparison with GEI-based Methods: A meticulous examination of *ViewDiffGait* against leading GEI-based methods: GEINet [17], GaitGANv2 [22], DV-GEIs-pre [8], and DV-GEIs [7], reveals a noteworthy superiority. While GaitGANv2, DV-GEIs-pre, and DV-GEIs did not originally perform experiments on the OU-MVLP dataset, we conducted tailored experiments based on their network details to ensure fairness. The results, meticulously presented in Table V, unequivocally showcase that our proposed method attains the highest mean accuracy (77.8%). This stellar performance significantly outperforms DV-GEIs [7] by an impressive margin of 10.6% (67.2%). The stark advancement accentuates the unparalleled effectiveness of the diffusion generative model in surpassing traditional GEI-based methodologies.

Comparison with Silhouette-based Methods: Silhouette-based methods, represented by GaitSet [2] and GaitPart [4], stand as formidable contenders with performance comparable to our proposed *ViewDiffGait*. Our evaluation, meticulously conducted for probe views under 30°, 45°, 210°, and 225° (as illustrated in Table VI), reveals a noteworthy aspect. Despite the typical utilization of a sequence of silhouette images, usually comprising 30 frames, in silhouette-based methods, our proposed approach surpasses performance expectations with minimal performance gaps.

Furthermore, the assessment of reference time, meticulously detailed in Table VI, solidifies the superiority of *ViewDiffGait* over silhouette-based methods. In a single GeForce GTX 1080 environment with 12GB memory, CUDA version 10.2, and PyTorch version 1.9.1, our proposed method exhibits a processing speed nearly three times faster than its counterparts. This impressive efficiency underscores *ViewDiffGait* not only as a high-performing solution in cross-view conditions but also as an expeditious option. The harmonious balance between performance and speed reinforces *ViewDiffGait*'s standing as a promising and superior solution for real-world gait recognition applications.

V. Conclusions and Future Work

In this paper, we introduced a groundbreaking framework, *ViewDiffGait*, aimed at addressing the critical challenge of precision in gait recognition under significant viewpoint variations. Leveraging a unique combination of a view pyramid structure and diffusion models, *ViewDiffGait* emerges as a transformative solution, capable of generating highly accurate lateral view images through an iterative refinement generation process. Unlike conventional large-view transformation models, our approach ensures fine view prediction, mitigating precision issues commonly encountered in gait recognition.

The extensive experiments conducted on the CASIA-B and OUMVLP datasets underscore the effectiveness of *ViewDiffGait*. The results reveal its capacity to generate

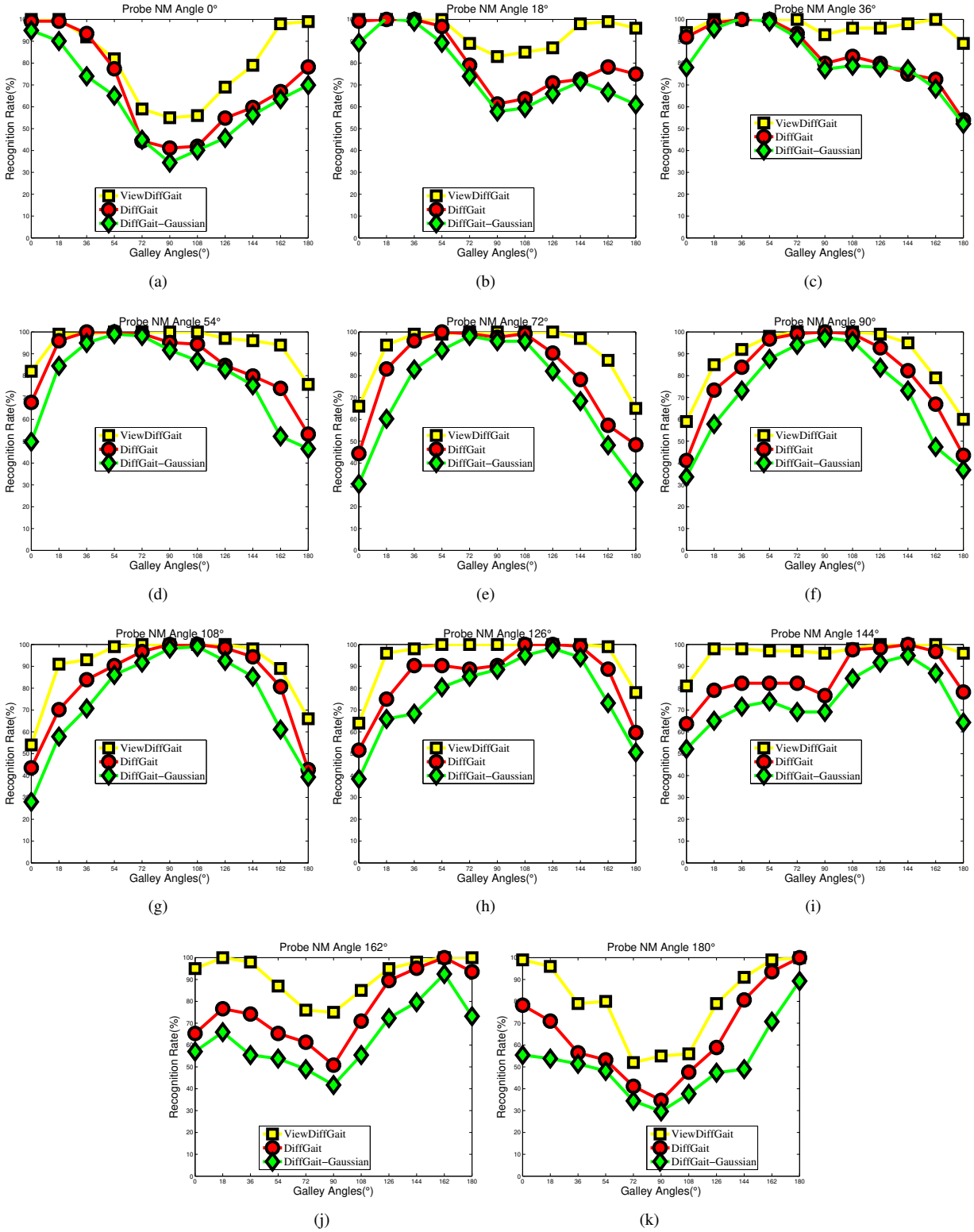

Fig. 5. Ablation Study Experiment. *ViewDiffGait*: with View Pyramid and with View Noise. *DiffGait*: without View Pyramid and with View Noise. *DiffGait-Gaussian*: without View Pyramid and without View Noise.

TABLE III

RANK-1 ACCURACY (%) ON CASIA-B UNDER ALL VIEW ANGLES, EXCLUDING IDENTICAL-VIEW CASE.

Train/Test Subjects	Input Feature	Gallery View NM:01-04 / Probe View NM:05-06	0°-180°											
			0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	Mean
62/62	Skeletons	PTSN [9]	34.5	45.6	49.6	51.3	52.7	52.3	53	50.8	52.2	48.3	31.4	47.4
		PTSN-3D [1]	38.7	50.2	55.9	56	56.7	54.6	54.8	56	54.1	52.4	40.2	51.9
		PoseGait [11]	48.5	62.7	66.6	66.2	61.9	59.8	63.6	65.7	66	58	46.5	60.5
		PoseMapGait [10]	59.9	76.2	81.7	83.1	76.8	76.1	76.3	81.1	79.6	75.4	66.1	75.7
	GEI	*ViewDiffGait* (Ours)	77.0	90.6	94.7	92.7	87.7	83.0	86.1	92.1	93.5	88.3	76.5	**87.5**
62/62	GEI	SPAE [21]	50.0	58.1	61.0	63.3	64.0	62.1	62.3	66.3	64.4	54.5	46.7	59.3
		GaitGANv2 [22]	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 [8]	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 [7]	63.1	79.4	84.6	79.8	77.0	72.6	77.4	80.3	84.0	78.5	63.7	76.4
		ViewDiffGait (Ours)	77.0	90.6	94.7	92.7	87.7	83.0	86.1	92.1	93.5	88.3	76.5	**87.5**
74/50	GEI	DV-GEIs-pre [8]	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 [7]	72.9	85.9	89.3	87.1	83.7	81.7	82.8	87.3	91.3	87.1	74.9	84.0
		GaitSet [2]	-	-	-	-	-	-	-	-	-	-	-	80.4
		ViewDiffGait (Ours)	78.9	93.7	96.6	94.3	90.7	86.7	89.0	93.5	96.1	90.9	78.6	**89.9**

TABLE IV

RANK-1 ACCURACY (%) ON CASIA-B UNDER ALL VIEW ANGLES, EXCLUDING IDENTICAL-VIEW CASE.

Train/Test Subjects	Input Feature	Gallery View NM:01-04 / Probe View NM:05-06	0°-180°				Inference Time Total (s)	Inference Time Per Sequence (ms)
			36°	54°	72°	Mean		
74/50	Silhouettes	GaitSet [2]	99.4	96.9	93.6	96.6	66.86	12.19
		GaitPart [4]	99.3	98.5	94.0	97.3	93.68	17.08
	GEI	*ViewDiffGait* (Ours)	96.6	94.3	90.7	93.9	**44.99**	**8.2**

TABLE V

RANK-1 ACCURACY (%) ON OUMVLP UNDER 14 PROBE VIEWS EXCLUDING IDENTICAL-VIEW CASES.

	Probe View														
Methods	0°	15°	30°	45°	60°	75°	90°	180°	195°	210°	225°	240°	255°	270°	Mean
GEINet [17]	23.2	38.1	48	51.8	47.5	48.1	43.8	27.3	37.9	46.8	49.9	45.9	45.7	41	42.5
GaitGANv2 [22]	46.8	55.3	60.5	59.6	57.6	57.1	56.2	49.5	54.5	57.5	60.5	55.4	57.5	55.4	55.9
DV-GEIs-pre [8]	51.3	63.7	69.6	71.4	66.2	68.5	66.1	57.9	64.9	69.5	69.2	64.7	65.6	65.6	65.3
DV-GEIs [7]	53.5	67.5	72.1	73.8	68.7	70.5	68.5	56.9	67.3	70.5	71.8	66.5	68.1	64.7	67.2
ViewDiffGait (Ours)	61.3	77.3	84.5	85.2	78.7	79.9	78.6	67.5	77.1	83.2	83.9	77.5	77.9	76.5	**77.8**

TABLE VI

RANK-1 ACCURACY (%) ON OUMVLP UNDER 4 PROBE VIEWS EXCLUDING IDENTICAL-VIEW CASES.

Train/Test Subjects	Input Feature	Gallery View: Seq#01 / Probe View: Seq#00	0°-360°					Inference Time Total (min)	Inference Time Per Sequence (ms)
			30°	45°	210°	225°	Mean		
5153/5154	Silhouettes	GaitSet [2]	90	90.1	89	89.2	89.6	57.32	25.69
		GaitPart [4]	90.8	91	90	90.1	90.5	68.33	30.63
	GEI	*ViewDiffGait* (Ours)	84.5	85.2	83.2	83.9	84.2	**23.56**	**10.56**

realistic images, effectively remove variations, and achieve high performance in real-world applications. Our method not only surpasses many existing GEI-based approaches but also stands shoulder to shoulder with state-of-the-art silhouette-based methods, showcasing its versatility and superiority.

Moreover, *ViewDiffGait* exhibits an impressive processing speed, outpacing silhouette-based methods. This efficiency further establishes *ViewDiffGait* as not just a high-performing solution for cross-view conditions but also an expeditious option. The harmonious balance between performance and speed positions *ViewDiffGait* as a promising and superior choice for practical gait recognition applications.

In the future, our research can explore and harness the potential of diffusion models and view pyramids within silhouette-based methods. Such as exploring the integration of temporal information to capture gait dynamics over time, potentially improving recognition accuracy and performance. Or investigate enhancements to the view pyramid structure, exploring ways to optimize its configuration for improved accuracy and robustness across diverse gait patterns. This extension aims to enhance the capabilities of gait recognition systems, addressing challenges and unlocking new possibilities in the evolving landscape of biometric identification. We anticipate that further innovations and refinements in these directions will contribute to the continual advancement of gait recognition technology, providing robust and efficient solutions for real-world applications.

VI. ACKNOWLEDGMENTS

This work is supported in part with grants from NSF 2148382. The views expressed in this article, book, or presentation are those of the author and do not necessarily reflect the official policy or position of the United States Air Force Academy, the Air Force, the Department of Defense, or the U.S. Government. Approved for public release: distribution unlimited. PA#: USAFA-DF-2024-237.

REFERENCES

[1] W. An, R. Liao, S. Yu, Y. Huang, and P. C. Yuen. Improving gait recognition with 3d pose estimation. In *Chinese Conference on Biometric Recognition*, pages 137–147. Springer, 2018.

[2] H. Chao, Y. He, J. Zhang, and J. Feng. Gaitset: Regarding gait as a set for cross-view gait recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8126–8133, 2019.

[3] C. Fan, J. Liang, C. Shen, S. Hou, Y. Huang, and S. Yu. Opengait: Revisiting gait recognition towards better practicality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9707–9716, 2023.

[4] C. Fan, Y. Peng, C. Cao, X. Liu, S. Hou, J. Chi, Y. Huang, Q. Li, and Z. He. Gaitpart: Temporal part-based model for gait recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14225–14233, 2020.

[5] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[6] W. Kusakunniran, Q. Wu, H. Li, and J. Zhang. Multiple views gait recognition using view transformation model based on optimized gait energy image. In *IEEE 12th International Conference on Computer Vision Workshops*, pages 1058–1064, 2009.

[7] R. Liao, W. An, Z. Li, and S. S. Bhattacharyya. A novel view synthesis approach based on view space covering for gait recognition. *Neurocomputing*, 453:13–25, 2021.

[8] R. Liao, W. An, S. Yu, Z. Li, and 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 (IJCB)*, pages 1–9. IEEE, 2020.

[9] R. Liao, C. Cao, E. B. Garcia, S. Yu, and Y. Huang. Pose-based temporal-spatial network (ptsn) for gait recognition with carrying and clothing variations. In *Chinese Conference on Biometric Recognition*, pages 474–483. Springer, 2017.

[10] R. Liao, Z. Li, S. S. Bhattacharyya, and G. York. Posemapgait: A model-based gait recognition method with pose estimation maps and graph convolutional networks. *Neurocomputing*, 501:514–528, 2022.

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

[12] B. Lin, S. Zhang, and X. Yu. Gait recognition via effective global-local feature representation and local temporal aggregation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14648–14656, 2021.

[13] C. Luo. Understanding diffusion models: A unified perspective. *arXiv preprint arXiv:2208.11970*, 2022.

[14] Y. Makihara, R. Sagawa, Y. Mukaigawa, T. Echigo, and Y. Yagi. Gait recognition using a view transformation model in the frequency domain. In *European Conference on Computer Vision*, pages 151–163, 2006.

[15] D. M. T. E. Y. Y. Noriko Takemura, Yasushi Makihara. Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition. *IPSJ Trans. on Computer Vision and Applications*, 10(4):1–14, 2018.

[16] C. Shen, C. Fan, W. Wu, R. Wang, G. Q. Huang, and S. Yu. Lidargait: Benchmarking 3d gait recognition with point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1054–1063, 2023.

[17] K. Shiraga, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi. Geinet: View-invariant gait recognition using a convolutional neural network. In *2016 international conference on biometrics (ICB)*, pages 1–8. IEEE, 2016.

[18] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.

[19] J. Tao and Y.-P. Tan. A probabilistic approach to incorporating domain knowledge for closed-room people monitoring. *Signal Processing: Image Communication*, 19(10):959–974, 2004.

[20] X. Yang, Z. Wang, H. Wu, L. Jiao, Y. Xu, and H. Chen. Stable and compact face recognition via unlabeled data driven sparse representation-based classification. *Signal Processing: Image Communication*, 111:116889, 2023.

[21] 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.

[22] S. Yu, R. Liao, W. An, H. Chen, E. B. G. Reyes, Y. Huang, and N. Poh. Gaitganv2: Invariant gait feature extraction using generative adversarial networks. *Pattern recognition*, 87:179–189, 2019.

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

[24] S. Zheng, J. Zhang, K. Huang, R. He, and T. Tan. Robust view transformation model for gait recognition. In *18th IEEE International Conference on Image Processing*, pages 2073–2076, 2011.