論文 / 著書情報
Article / Book Information

Title	Diffusion Pretraining for Gait Recognition in the Wild
Author	Wei Ming Neo, Koichi Shinoda, Tat-Jen Cham
Journal/Book name	2025 IEEE International Conference on Image Processing (ICIP), , , pp. 1295 - 1300
Pub. date	2025, 9
DOI	https://doi.org/10.1109/ICIP55913.2025.11084665
Copyright	(c)2025 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.
Note	This file is author (final) version.

DIFFUSION PRETRAINING FOR GAIT RECOGNITION IN THE WILD

Wei Ming Neo[⋆†] *Koichi Shinoda*[†] *Tat-Jen Cham*[⋆]

[⋆] Nanyang Technological University, Singapore
[†] Institute of Science Tokyo

ABSTRACT

Recently, diffusion models have garnered much attention for their remarkable generative capabilities. Yet, their application for representation learning remains largely unexplored. In this paper, we explore the potential of diffusion models to pretrain the backbone of a deep learning model for a specific application—gait recognition in the wild. To do so, we condition a latent diffusion model on the output of a gait recognition model backbone. Our pretraining experiments on the Gait3D and GREW datasets reveal an interesting phenomenon: diffusion pretraining causes the gait recognition backbone to separate gait sequences belonging to different subjects further apart than those belonging to the same subjects. Subsequently, our transfer learning experiments on Gait3D and GREW show that the pretrained backbone can serve as an effective initialization for the downstream gait recognition task, improving gait recognition accuracies by as much as 7.9% on Gait3D and 4.2% on GREW.

Index Terms— Diffusion Models, Gait Recognition, Representation Learning

1. INTRODUCTION

Gait, the unique manner in which a person walks, offers a way to identify individuals, alongside the more common biometric modalities like fingerprints and irises. With the emergence of deep learning, advancements in computer vision architectures, and the collection of well-labelled gait datasets, deep gait recognition has become an increasingly popular area of research over the last few years [1]. Consequently, owing to the efforts of many previous works, it has become possible to achieve impressive accuracy performance on existing controlled datasets such as CASIA-B [2] and OU-MVLP [3]. However, when these techniques are applied to recently released in-the-wild gait datasets such as Gait3D [4] and GREW [5], their performance pales in comparison, highlighting their limited applicability to unconstrained settings. As a result, recent studies [4, 5, 6, 7, 8] have begun shifting their focus towards addressing the more challenging problem of gait recognition in the wild.

Diffusion models [9, 10, 11, 12] have recently emerged as a dominant paradigm in generative modeling, surpassing traditional approaches like variational autoencoders (VAEs), flow-based models, and generative adversarial networks (GANs). Their stability during training and ability to produce high-quality, realistic outputs—spanning images, videos, and audio—have driven widespread adoption across academia and industry. However, amidst the hype surrounding their generative capabilities, other promising directions of diffusion models remain relatively unexplored.

One particularly promising direction involves utilizing diffusion models for representation learning. Conditional diffusion models, for instance, can generate specific data based on provided conditions, often leveraging features from discriminative models [13, 14]. This begs the question: if diffusion models can generate precise outputs when conditioned on discriminative features, could the process be reversed? Specifically, could a model be trained to extract discriminative features for accurate input reconstruction using the same diffusion loss?

Despite growing interest in applying diffusion-based representations to common tasks like image classification [15, 16, 17], their potential for more specialized applications, such as gait recognition, remains largely untapped. Moreover, prior studies often neglect finetuning the learnt representations on downstream tasks, overlooking the pretraining potential of diffusion training.

Considering the limited research on diffusion-based representation learning in gait recognition, particularly for challenging in-the-wild scenarios, we propose a diffusion-based approach to pretrain the backbone of a gait recognition model by using its output as a condition for a latent diffusion model. We then initialize the gait recognition model with the pretrained backbone and perform transfer learning on the downstream gait recognition task. To evaluate this approach, we conducted extensive applicability studies with multiple existing gait recognition models, including GaitGL [6], GaitPart [18], GaitSet [19], SMPLGait w/o 3D [4], and GaitBase [7], using two in-the-wild datasets, Gait3D [4] and GREW [5].

Our finding reveals that during diffusion pretraining, the gait recognition model backbone, regardless of its architecture, learns to separate gait sequences belonging to different subjects further apart than those belonging to the same subject, even without explicit supervision. This results in a steady improvement in gait recognition performance, demonstrating

the potential of diffusion pretraining for learning discriminative features. Subsequently, when initialized with the pretrained backbone and further finetuned, the gait recognition model outperforms its trained-from-scratch counterpart by as much as 7.9% on Gait3D and 4.2% on GREW. This remains the case even when the number of supervised training iterations is significantly reduced by as much as 89% on Gait3D and 70% on GREW.

To the best of our knowledge, we are first to explore diffusion training for representation learning in gait recognition and demonstrate its effectiveness as a pretraining approach for the gait recognition task. We provide the results of our extensive ablation studies in the Supplementary Materials[1].

2. RELATED WORK

Gait Recognition: With the advent of deep learning, gait recognition has evolved into a task of extracting discriminative features from gait sequences and projecting them into embeddings that can be compared using distance metrics such as Euclidean or cosine distances. These gait sequences typically come in the form of either silhouettes or skeletons, with silhouette-based recognition being more dominant [1].

State-of-the-art research in deep gait recognition primarily focuses on designing robust backbone networks to maximize the extraction of meaningful gait features. [6, 7, 8, 19, 18, 20]. Inspired by the field of person re-identification, techniques such as horizontal pyramid matching [21] and batch normalization neck [22] have been adapted to further enhance gait recognition accuracy. In parallel, self-supervised learning approaches [23, 24] have gained traction, particularly contrastive learning frameworks that treat augmented versions of the same gait sequence as positive pairs and sequences from different individuals as negative pairs.

While many of these methods demonstrate impressive performance on controlled datasets such as CASIA-B [2] and OU-MVLP [3], their efficacy diminishes when applied to recently introduced in-the-wild datasets [4, 5]. These datasets introduce realistic complexities, including occlusions, varying camera viewpoints, and inconsistent lighting conditions, making them a critical testbed for robust gait recognition. Enhancing the generalization of gait recognition models to perform reliably in such unconstrained environments remains an open and pressing research challenge. To this end, OpenGait [7] has recently been introduced. Several previous works that precede the introduction of in-the-wind datasets have been reproduced, trained, and evaluated on the in-the-wild datasets, serving as baselines for further research in gait recognition in the wild.

Diffusion: Diffusion models [9], particularly Denoising Diffusion Probabilistic Models [10], have become a prominent area of research due to their generative prowess. A typical diffusion process comprises two phases: the forward and reverse phases. In the forward phase, a data sample is progressively corrupted into near-pure noise by iteratively adding noise across a series of timesteps based on a predefined noise schedule. Conversely, in the reverse phase, a model iteratively removes noise from a noisy sample, reconstructing a clean sample. By training to predict the added noise in the noisy data at any timestep of the forward process, diffusion models learn a mapping from random noise to the data manifold, enabling their generative capability.

To achieve more precise generation, conditional diffusion models extend this process by incorporating additional prompts to learn conditional data distributions. Various enhancements have been proposed to improve the sample quality of these models, with classifier-free guidance [25] emerging as a simple yet effective technique.

To address the computational and memory demands associated with high-resolution data, latent diffusion models [12] have been introduced. These models first encode data into a lower-dimensional latent space before applying the diffusion process, significantly reducing resource consumption while maintaining generative quality. Additionally, training efficiency has been further improved through advancements such as optimized noise schedulers [10, 11, 12] and timestep weighting strategies [26, 27], which prioritize specific noise ranges during the diffusion process to enhance convergence. While diffusion models initially gained traction in image generation, they have proven versatile across domains, including video and audio.

Despite their impressive generative capabilities, their potential for learning meaningful representations remains underexplored. In the field of image recognition, Hudson et al. [16] demonstrate that diffusion models can learn strong semantic representations that support downstream classification tasks. However, the applicability to other domains remains uncertain, and the impact of finetuning these representations on downstream tasks has yet to be thoroughly investigated.

3. PROPOSED METHOD

In this section, we introduce our proposed method, which consists of two stages: (1) pretraining with diffusion and (2) transfer learning on the downstream gait recognition task.

3.1. Diffusion Pretraining

To learn gait representations via diffusion, we employ an end-to-end trainable conditional latent diffusion model. Our architecture (Fig. 1) consists of three main components—an encoder $\mathcal{E}$, a denoiser $\mathcal{D}$, and a gait feature extractor $\mathcal{G}$. The encoder first compresses an input silhouette sequence into a latent representation, which is then added with random noise and denoised via the denoiser. Concurrently, relevant features from the input silhouette sequence are extracted via the gait

[1] https://dx.doi.org/10.60864/k68n-qf22

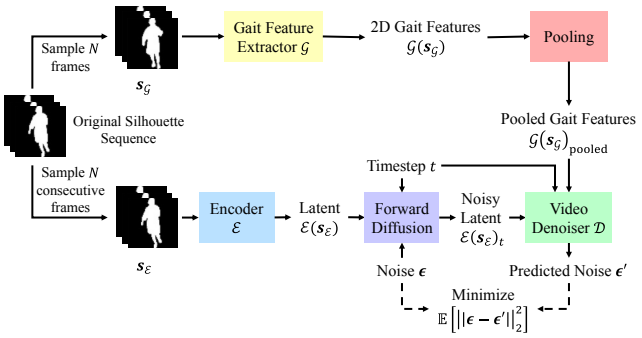

Fig. 1. Proposed architecture. Only the gait feature extractor and video denoiser are trained during diffusion pretraining.

feature extractor and these features are pooled and passed on as a condition to aid the denoiser in the denoising process for a more precise reconstruction of the input silhouette sequence.

Encoder: Considering its small size and decent encoding ability, we use the open-sourced Tiny AutoEncoder for Stable Diffusion (TAESD) [28] as our encoder.

Denoiser: For the denoiser, given the lack of open-sourced pretrained gait silhouette sequence diffusion models, we adapt a recent video diffusion model [14] for our use case.

Gait Feature Extractor: The gait feature extractor is derived from the backbone of a deep gait recognition model. To evaluate the proposed approach, we adopt various existing backbones, including GaitGL [6], GaitPart [18], GaitSet [19], SMPLGait w/o 3D [4], and GaitBase [7].

Pooling Method: Since the denoiser requires one- dimensional tensor conditions, the two-dimensional gait features are pooled. In particular, mean pooling is used, following our ablation study. The pooled features are then concatenated with timestep conditions and used to scale and bias the activations within the denoiser layers [29].

Input Pretreatment : To allow the gait extractor to focus learning on gait information, the autoencoder $\mathcal{E}$ and gait feature extractor $\mathcal{G}$ are presented with subsequences sampled from the same silhouette sequence via different sampling algorithms. We denote this pair of input subsequences as $\boldsymbol{s} = (\boldsymbol{s}_{\mathcal{E}}, \boldsymbol{s}_{\mathcal{G}})$.

As a video diffusion model is used for the denoiser, N frames are sampled consecutively from the silhouette sequence to serve as $\boldsymbol{s}_{\mathcal{E}}$. As for $\boldsymbol{s}_{\mathcal{G}}$, we sample N frames using the sampling algorithm employed by the authors who proposed the respective backbones [6, 18, 19, 4, 7]. We fixed $N = 30$ throughout the study.

To increase the generalizability of the model, $\boldsymbol{s}_{\mathcal{E}}$ and $\boldsymbol{s}_{\mathcal{G}}$ are augmented separately with a composition of RandomAffine, RandomPerspective, RandomHorizontalFlip, RandomPartDilate [24] and RandomPartBlur.

Noise Scheduler and Loss Weighting Strategy: For representation learning, prioritizing medium noise has been shown to be more effective than the high- or low-noise focus typical of generative tasks. We adopt an inverted cosine noise scheduler [16] and implement a medium noise prioritization strategy. Specifically, we modify the Min-SNR weighting strategy [26], which downweighs losses from low noise levels, to also downweigh losses from high noise levels.

Loss Function : To pretrain our gait recognition model via diffusion, we use the $L2$ noise prediction loss, which can be summarized as:

$$L_{\text{diffusion}} = ||\boldsymbol{\epsilon} - \mathcal{D}(\mathcal{E}(\boldsymbol{s}_{\mathcal{E}})_t, t, \mathcal{G}(\boldsymbol{s}_{\mathcal{G}})_{\text{pooled}})||_2^2 \quad (1)$$

where $\mathcal{E}(\boldsymbol{s}_{\mathcal{E}})_t$ is the noised latent representation fed to the denoiser $\mathcal{D}$, at timestep $t \in (0, 1000]$ of the forward diffusion process, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ is the random noise added, and $\mathcal{G}(\boldsymbol{s}_{\mathcal{G}})_{\text{pooled}}$ is the pooled gait feature condition. During diffusion pretraining, the denoiser and gait feature extractor are trained from scratch while the pretrained TAESD encoder is kept frozen. A higher learning rate is assigned to the gait feature extractor to enable it to better guide the denoiser during training, following the work of Hudson et al. [16].

3.2. Transfer Learning

Once the gait feature extractor is trained by diffusion, we evaluate it on the downstream gait recognition task. We replicate the remaining parts of the gait recognition model accordingly and initialize the weights of the gait backbone with the ones learnt during diffusion pretraining. The untrained parameters are initialized based on the settings provided by OpenGait.

During transfer learning, each gait recognition model is trained using the standard triplet loss, L_{triplet}, for identification. We use cosine distance, as it yields better results for existing works (Supplementary Material). Additionally, some methods, such as GaitBase and SMPLGait w/o 3D, also incorporate a smoothed identity loss, L_{ID}, by having another module predict the identity of each gait sequence to enhance the gait recognition performance. In this case, the net loss is the sum of the triplet loss and the reweighted identity loss:

$$L_{\text{net}} = L_{\text{triplet}} + 0.1 \cdot L_{\text{ID}} \quad (2)$$

4. EXPERIMENTS

4.1. Experimental Setup

With the focus on practical gait recognition, two datasets meant for gait recognition in the wild, namely, Gait3D [4] and GREW [5], were chosen. For pretraining, we used AdamW with warmup and cosine annealing with a learning rate of $1e^{-4}$ for the denoiser and $5e^{-4}$ for the gait feature extractor. The models were trained for 120k iterations with a batch size of 64 for Gait3D and 128 for GREW. More specific hyperparameter settings are detailed in the Supplementary Materials.

Table 1. Rank-1 accuracy on Gait3D and GREW. GaitPart on GREW is excluded due to training instability with cosine distance. Train iterations: X + Y denotes Xk diffusion pretraining followed by Yk transfer learning. For GaitBase on GREW, transfer learning iterations are 90k (without augmentation) and 120k (with augmentation).

Method	Gait3D				GREW			
	Rank-1 Accuracy (%) ✗ Data Aug.	Rank-1 Accuracy (%) ✓ Data Aug.	r	Train Iter. ($\times 10^3$)	Rank-1 Accuracy (%) ✗ Data Aug.	Rank-1 Accuracy (%) ✓ Data Aug.	r	Train Iter. ($\times 10^3$)
Reproduced Baseline								
GaitGL	29.2	32.4	-	180	54.0	58.4	-	250
GaitPart	31.2	38.7	-	180	-	-	-	-
GaitSet	42.2	47.8	-	180	48.1	53.1	-	250
SMPLGait w/o 3D	45.5	42.9	-	180	47.6	52.1	-	250
GaitBase	56.5	65.8	-	60	58.1	61.8	-	180
Diffusion Pretraining + Transfer Learning with Frozen Backbone								
GaitGL	17.0	-	0.0	120 + 60	32.2	-	0.0	120 + 125
GaitPart	18.8	-	0.0	120 + 60	-	-	-	-
GaitSet	23.5	-	0.0	120 + 60	33.5	-	0.0	120 + 125
SMPLGait w/o 3D	30.7	-	0.0	120 + 60	36.0	-	0.0	120 + 125
GaitBase	35.0	-	0.0	120 + 60	40.5	-	0.0	120 + 90
Diffusion Pretraining + Transfer Learning with Finetuning of Backbone								
GaitGL	34.4 $_{\uparrow 5.2}$	34.4 $_{\uparrow 2.0}$	0.1	120 + 60	56.3 $_{\uparrow 2.3}$	58.6 $_{\uparrow 0.2}$	1.0	120 + 125
GaitPart	35.7 $_{\uparrow 4.5}$	41.7 $_{\uparrow 3.0}$	0.5	120 + 60	-	-	-	-
GaitSet	45.0 $_{\uparrow 2.8}$	49.9 $_{\uparrow 2.1}$	0.5	120 + 60	52.0 $_{\uparrow 3.9}$	55.4 $_{\uparrow 2.3}$	1.0	120 + 125
SMPLGait w/o 3D	53.4 $_{\uparrow 7.9}$	60.7 $_{\uparrow 17.8}$	0.5	120 + 60	51.8 $_{\uparrow 4.2}$	54.1 $_{\uparrow 2.0}$	1.0	120 + 125
GaitBase	62.3 $_{\uparrow 5.8}$	69.7 $_{\uparrow 3.9}$	1.0	120 + 60	58.5 $_{\uparrow 0.4}$	62.0 $_{\uparrow 0.2}$	0.5	120 + 90/120

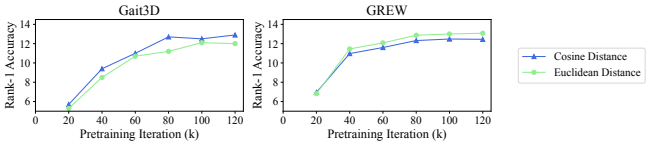

Fig. 2. Rank-1 accuracy curves during diffusion pretraining on Gait3D and GREW (GaitSet).

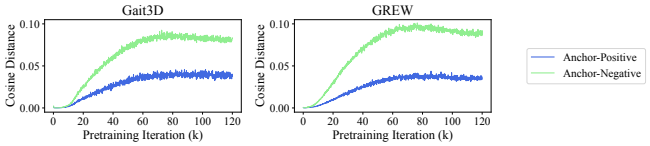

Fig. 3. Mean cosine distance of anchor-positive pair and anchor-negative pair during diffusion pretraining on Gait3D and GREW (GaitSet).

For transfer learning, we trained and evaluated on the same dataset that was used during diffusion pretraining. We used the rank-1 gait recognition accuracy as our main evaluation metric. For GREW, we submitted the results to the official website for evaluation.

To evaluate the effectiveness of our proposed method, we reproduced the results of the corresponding gait recognition models on Gait3D and GREW trained solely via the supervised objective. The performance of each reproduced baseline, with and without data augmentation, is presented in Table 1.

4.2. Diffusion Pretraining Results

To evaluate whether the gait feature extractor learns useful features during diffusion pretraining, we measured gait recognition performance at various pretraining checkpoints (Fig. 2). This was done by inputting the test set's silhouette sequences into the extractor and determining similarity using cosine or Euclidean distances between the output gait features. For brevity, we present results for GaitSet pretrained on Gait3D and GREW, with findings for other models included in the Supplementary Materials.

We observed a steady improvement in gait recognition performance during pretraining, indicating that the gait feature extractor learns meaningful features for gait recognition, even while only reconstructing inputs at this stage.

To investigate further, we recorded the mean cosine distance between anchor-positive and anchor-negative pairs within a batch during pretraining (Fig. 3). Interestingly, the difference in distance between these pairs increased and stabilized, regardless of the dataset or gait feature extractor architecture. This suggests that diffusion pretraining inherently encourages separation between anchor-positive and anchor-negative pairs, despite the absence of explicit supervisory signals, explaining the observed improvement in gait recognition performance.

4.3. Transfer Learning Results

Frozen Backbone: With the backbones of the various gait recognition models pretrained via diffusion, we evaluated their performance on downstream tasks by freezing the pretrained backbones and training only the remaining layers via supervised learning.

Table 1 shows that transfer learning with pretrained backbones achieves up to 75% of baseline accuracy, confirming that diffusion pretraining extracts useful gait features. However, these features are less discriminative than those from full supervision, likely due to the inclusion of irrelevant reconstruction features. Despite this, diffusion pretraining captures key discriminative features, highlighting its potential as a pretraining approach.

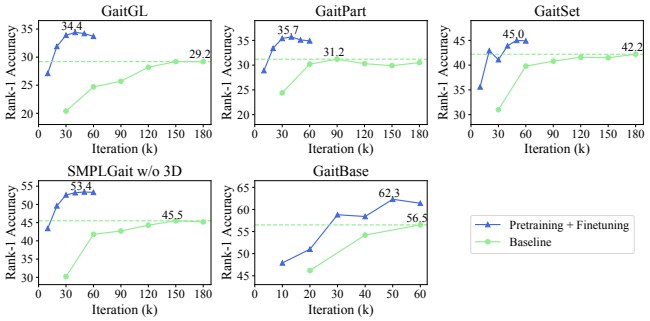

Fig. 4. Rank-1 gait recognition accuracy curves when no data augmentation is applied during supervised training (Gait3D).

Table 2. Rank-1 accuracy on GREW with further reduction in supervised training iterations.

Method	GREW		
	Rank-1 Accuracy (%)		Train Iter.
	✗ Data Aug.	✓ Data Aug.	($\times 10^3$)
GaitGL	54.3	56.1	120 + 75
GaitSet	51.1	53.3	120 + 75
SMPLGait w/o 3D	50.3	51.8	120 + 75

Finetuning of the Backbone: We finetuned the pretrained backbones to assess potential improvements. During finetuning, we found that the learning rate ratio between pretrained and untrained layers, r, is a key hyperparameter. We attempted $r \in \{0.1, 0.5, 1.0\}$ for Gait3D and GREW and applied a small weight decay ($5e^{-5}$) to mitigate overfitting on GREW. Table 1 shows the best finetuning results obtained with the corresponding value of r used. Results for other r values are in the Supplementary Materials.

As shown in Table 1, all models initialized with diffusion-pretrained backbones outperformed their trained-from-scratch counterparts, even with significantly fewer supervised training iterations. Excluding the anomalous case for SMPLGait w/o 3D which deteriorated with data augmentation on Gait3D, rank-1 accuracy improved by up to 7.9% on Gait3D and 4.2% on GREW. Moreover, Fig. 4 highlights that with diffusion-pretrained backbones, models surpassed supervised baselines within just as little as 20k iterations—an 89% reduction in supervised training iterations. On GREW, GaitGL, GaitSet, and SMPLGait w/o 3D maintained competitive performance after being finetuned for only 30% of the baselines' training iterations (Table 2). Notably, diffusion pretraining mitigated poor initialization issues, reinforcing its value as a strong starting point for the downstream gait recognition task.

5. CONCLUSION

In summary, we propose a diffusion pretraining approach for gait recognition in the wild. By conditioning a denoiser on the output of a gait feature extractor, we can pretrain the extractor to capture relevant gait features. Initializing the gait recognition model with the pretrained backbone and finetuning it on the downstream gait recognition task allows it to outperform its supervised learning counterpart while requiring far less supervised training time. We hope this work inspires further exploration of diffusion models for representation learning across gait recognition and beyond.

6. ACKNOWLEDGEMENT

This work was supported by JSPS KAKENHI Grant Number JP23H00490.

7. REFERENCES

[1] Alireza Sepas-Moghaddam and Ali Etemad, "Deep gait recognition: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 1, pp. 264–284, 2022.

[2] Shiqi Yu, Daoliang Tan, and Tieniu Tan, "A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition," in *18th international conference on pattern recognition (ICPR'06)*. IEEE, 2006, vol. 4, pp. 441–444.

[3] Noriko Takemura, Yasushi Makihara, Daigo Muramatsu, Tomio Echigo, and Yasushi Yagi, "Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition," *IPSJ transactions on Computer Vision and Applications*, vol. 10, pp. 1–14, 2018.

[4] Jinkai Zheng, Xinchen Liu, Wu Liu, Lingxiao He, Chenggang Yan, and Tao Mei, "Gait recognition in the wild with dense 3d representations and a benchmark," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 20228–20237.

[5] Zheng Zhu, Xianda Guo, Tian Yang, Junjie Huang, Jiankang Deng, Guan Huang, Dalong Du, Jiwen Lu, and Jie Zhou, "Gait recognition in the wild: A benchmark," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 14789–14799.

[6] Beibei Lin, Shunli Zhang, Ming Wang, Lincheng Li, and Xin Yu, "Gaitgl: Learning discriminative global-local feature representations for gait recognition," *arXiv preprint arXiv:2208.01380*, 2022.

[7] Chao Fan, Junhao Liang, Chuanfu Shen, Saihui Hou, Yongzhen Huang, and Shiqi Yu, "Opengait: Revisiting gait recognition towards better practicality," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 9707–9716.

[8] Chao Fan, Saihui Hou, Yongzhen Huang, and Shiqi Yu, "Exploring deep models for practical gait recognition," *arXiv preprint arXiv:2303.03301*, 2023.

[9] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *International conference on machine learning*. PMLR, 2015, pp. 2256–2265.

[10] Jonathan Ho, Ajay Jain, and Pieter Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[11] Alexander Quinn Nichol and Prafulla Dhariwal, "Improved denoising diffusion probabilistic models," in *International conference on machine learning*. PMLR, 2021, pp. 8162–8171.

[12] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10684–10695.

[13] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al., "Imagen video: High definition video generation with diffusion models," *arXiv preprint arXiv:2210.02303*, 2022.

[14] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet, "Video diffusion models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 8633–8646, 2022.

[15] Kevin Clark and Priyank Jaini, "Text-to-image diffusion models are zero shot classifiers," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[16] Drew A Hudson, Daniel Zoran, Mateusz Malinowski, Andrew K Lampinen, Andrew Jaegle, James L McClelland, Loic Matthey, Felix Hill, and Alexander Lerchner, "Soda: Bottleneck diffusion models for representation learning," *arXiv preprint arXiv:2311.17901*, 2023.

[17] Alexander C Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak, "Your diffusion model is secretly a zero-shot classifier," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 2206–2217.

[18] Chao Fan, Yunjie Peng, Chunshui Cao, Xu Liu, Saihui Hou, Jiannan Chi, Yongzhen Huang, Qing Li, and Zhiqiang 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.

[19] Hanqing Chao, Yiwei He, Junping Zhang, and Jianfeng Feng, "Gaitset: Regarding gait as a set for cross-view gait recognition," in *Proceedings of the AAAI conference on artificial intelligence*, 2019, vol. 33, pp. 8126–8133.

[20] Chunfeng Song, Yongzhen Huang, Yan Huang, Ning Jia, and Liang Wang, "Gaitnet: An end-to-end network for gait based human identification," *Pattern recognition*, vol. 96, pp. 106988, 2019.

[21] Yang Fu, Yunchao Wei, Yuqian Zhou, Honghui Shi, Gao Huang, Xinchao Wang, Zhiqiang Yao, and Thomas Huang, "Horizontal pyramid matching for person re-identification," in *Proceedings of the AAAI conference on artificial intelligence*, 2019, vol. 33, pp. 8295–8302.

[22] Hao Luo, Wei Jiang, Youzhi Gu, Fuxu Liu, Xingyu Liao, Shenqi Lai, and Jianyang Gu, "A strong baseline and batch normalization neck for deep person re-identification," *IEEE Transactions on Multimedia*, vol. 22, no. 10, pp. 2597–2609, 2019.

[23] Yiqun Liu, Yi Zeng, Jian Pu, Hongming Shan, Peiyang He, and Junping Zhang, "Selfgait: A spatiotemporal representation learning method for self-supervised gait recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 2570–2574.

[24] Chao Fan, Saihui Hou, Jilong Wang, Yongzhen Huang, and Shiqi Yu, "Learning gait representation from massive unlabelled walking videos: A benchmark," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[25] Jonathan Ho and Tim Salimans, "Classifier-free diffusion guidance," *arXiv preprint arXiv:2207.12598*, 2022.

[26] Tiankai Hang, Shuyang Gu, Chen Li, Jianmin Bao, Dong Chen, Han Hu, Xin Geng, and Baining Guo, "Efficient diffusion training via min-snr weighting strategy," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 7441–7451.

[27] Tim Salimans and Jonathan Ho, "Progressive distillation for fast sampling of diffusion models," *arXiv preprint arXiv:2202.00512*, 2022.

[28] Ollin Boer Bohan, "Tiny autoencoder for stable diffusion," https://github.com/madebyollin/taesd, 2023.

[29] Prafulla Dhariwal and Alexander Nichol, "Diffusion models beat gans on image synthesis," *Advances in neural information processing systems*, vol. 34, pp. 8780–8794, 2021.