On Denoising Walking Videos for Gait Recognition

Dongyang Jin[1*], Chao Fan[2,1*], Jingzhe Ma[3,1], Jingkai Zhou[4,5], Weihua Chen[5], and Shiqi Yu[1†]
[1] Department of Computer Science and Engineering, Southern University of Science and Technology, China
[2] National Engineering Laboratory for Big Data System Computing Technology, Shenzhen University, China
[3] Shenzhen Polytechnic University, China [4] Zhejiang University, China [5] Alibaba Group, China
12332451@mail.sustech.edu.cn, chaofan996@szu.edu.cn, jingzhema@szpu.edu.cn,
fs.jingkaizhou@gmail.com, kugang.cwh@alibaba-inc.com, yusq@sustech.edu.cn

Abstract

To capture individual gait patterns, excluding identity-irrelevant cues in walking videos, such as clothing texture and color, remains a persistent challenge for vision-based gait recognition. Traditional silhouette- and pose-based methods, though theoretically effective at removing such distractions, often fall short of high accuracy due to their sparse and less informative inputs. Emerging end-to-end methods address this by directly denoising RGB videos using human priors. Building on this trend, we propose DenoisingGait, a novel gait denoising method. Inspired by the philosophy that "what I cannot create, I do not understand", we turn to generative diffusion models, uncovering how they partially filter out irrelevant factors for gait understanding. Additionally, we introduce a geometry-driven Feature Matching module, which, combined with background removal via human silhouettes, condenses the multi-channel diffusion features at each foreground pixel into a two-channel direction vector. Specifically, the proposed within- and cross-frame matching respectively capture the local vectorized structures of gait appearance and motion, producing a novel flow-like gait representation termed Gait Feature Field, which further reduces residual noise in diffusion features. Experiments on the CCPG, CASIA-B, and SUSTech1K datasets demonstrate that DenoisingGait achieves a new SoTA performance in most cases for both within- and cross-domain evaluations. Code is available at https://github.com/ShiqiYu/OpenGait.*

1. Introduction

Pedestrian gait, as captured in walking videos, uniquely conveys identifiable traits through body shape and limb movement, making it an effective biometric feature. This application, known as gait recognition, differs from methods like face, fingerprint, and iris recognition by enabling non-intrusive human identification from a distance, without requiring the subject's active cooperation [35]. Additionally, gait is challenging to disguise or obscure, making it suitable for security applications in unconstrained environments, such as criminal suspect tracking and retrieval [25].

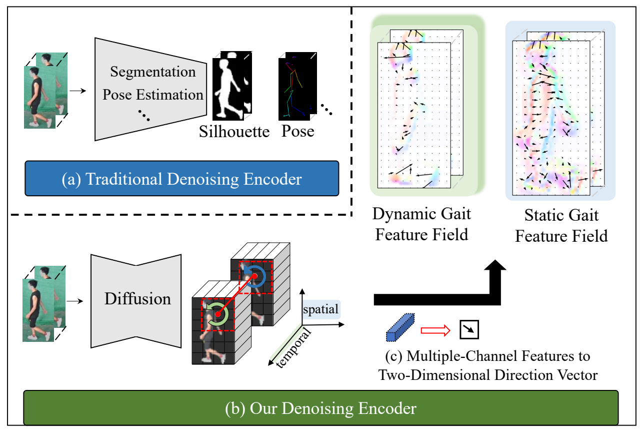


Figure 1. The proposed knowledge-driven denoising, derived from generative diffusion models, and the geometry-driven denoising, enforced by Feature Matching.

To minimize the influence of irrelevant cues like background and texture, many gait recognition methods rely on predefined representations extracted from walking videos. The most commonly used ones include the binary silhouettes [2, 7, 10, 21, 34], skeleton coordinates [11, 20, 41], SMPL models [18, 56], and body parsing images [57, 60]. As illustrated in Figure 1, the gait representation extraction can be regarded as a *denoising encoder*, which can enhance the subsequent gait learning process through either a two-stage or end-to-end training manner. In contrast, some recent works [50, 55] go beyond explicit gait representations, focusing instead on directly denoising RGB videos through

*Equal contribution.
†Corresponding author.

human priors, using techniques such as image reconstruction [55] and feature smoothness [50]. In this study, we extend the scope of gait denoising research by proposing DenoisingGait, a method that combines knowledge-driven and geometry-driven denoising to improve gait recognition.

Guided by the philosophy that "what I cannot create, I do not understand", researchers [4, 6, 54] are increasingly exploring the use of generative diffusion models [14, 29] for representation learning. In this study, we observe that by carefully manipulating the timestep t [1], diffusion models [29] can selectively filter out gait-irrelevant cues in RGB videos. This finding aligns with prior studies [16, 28, 59] indicating that different timesteps of diffusion models contribute to the feature reconstruction at varying levels of granularity. However, even with a suitable timestep t, diffusion outputs [39] remain closely tied to RGB details [53], making it still not clean enough for gait recognition.

To address this issue, DenoisingGait introduces a geometry-driven *Feature Matching* module to further reduce RGB-encoded (noise-prone) features. Building on the background removal via gait silhouettes, our core idea is to condense the multi-channel diffusion features at each foreground pixel into a two-dimensional direction vector. This design draws inspiration partly from the classical SIFT descriptor [23], which formulates image as robust locality vectors, and partly from optical flow estimation [46], which encodes motion as dense temporal directions. As illustrated in Figure 1, the proposed within-frame and cross-frame matching mechanisms respectively assign orientations to neighboring locations based on feature similarity along the spatial and temporal dimension, thus effectively capturing the vectorized characteristics of gait appearance and motion. This work terms these new representations as static and dynamic *Gait Feature Fields*, due to their visual similarity with optical flow field. Interestingly, we observe that the magnitude of direction vector in the static gait feature field can, in many cases, reflect image texture intensity. Therefore, DenoisingGait applies random zero-padding to regions with high magnitude, thus promoting the learning of texture-invariant gait features.

Overall, DenoisingGait refines gait features through two key mechanisms: knowledge-driven denoising, derived from generative diffusion models, and the geometry-driven denoising, enforced by the proposed *Feature Matching* module. Experiments on CCPG [17], CASIA-B* [52] and SUSTech1K [32] datasets validate the effectiveness of DenoisingGait and its components across both within- and cross-domain evaluations. This work contributes to gait recognition research in three primary ways:

- To our knowledge, DenoisingGait introduces one of the first diffusion-based approaches for gait recognition, demonstrating the potential of diffusion models for effective gait representation learning.
- Gait feature field presents a novel and effective recognition-driven spatiotemporal gait representation.
- By integrating the proposed knowledge- and geometry-driven denoising mechanisms, in most cases, DenoisingGait sets a new state-of-the-art performance on several commonly used RGB gait datasets [17, 32, 52] for both within- and cross-domain evaluations.

[1]The generative process of diffusion models is typically a Markov chain, where each latent z_t is generated only from the previous latent z_{t+1}.

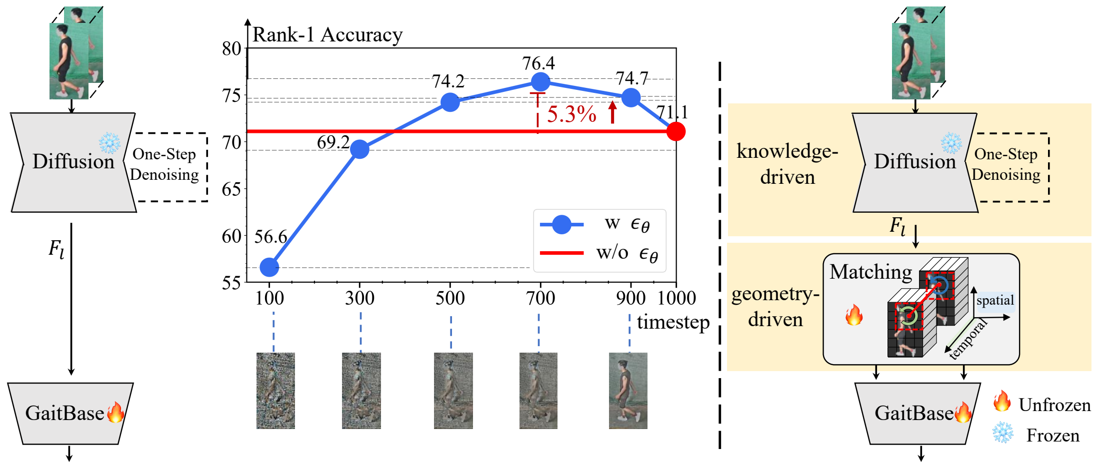
2. Related Work

2.1. Gait Recognition

Gait recognition aims to extract gait features that are robust to background and clothing textures. Avoiding these challenges in RGB videos has been a major focus in recent research [2, 7, 8, 11, 18–21, 41, 50, 55–57, 60]. Existing approaches can be broadly divided into two strategies: hard-denoising [2, 7, 8, 11, 18, 20, 21, 41, 56, 57, 60] and soft-denoising [19, 50, 55]. Hard-denoising methods employ algorithms like segmentation [2, 7, 8, 21], pose estimation [11, 20, 41], and 3D parameter estimation [18, 56] to explicitly isolate gait-relevant features, thus mitigating interference from irrelevant cues. While effective in refining gait representations, these methods may also strip away structural details essential for identity recognition. On the other hand, soft-denoising methods [19, 50, 55] rely on task-specific modules informed by human priors to suppress non-gait information within RGB images. Despite progress, removing gait-irrelevant factors remains challenging. In this paper, we explore this issue from a novel perspective by proposing a method that integrates both knowledge-driven and geometry-driven denoising for enhanced gait representation extraction.

2.2. Diffusion Models for Representation Learning

Diffusion models are generative models that learn to reverse a transformation from data to Gaussian noise [14], following either a Markov [14] or non-Markov [38] process. Latent Diffusion Models (LDMs) [29], commonly referred to as Stable Diffusion (SD), extend traditional diffusion models by operating in latent space, greatly improving efficiency. Recent studies have shown that SD models can serve as knowledge providers, contributing perceptual insights that enhance the generalization capabilities of some discriminative tasks, such as object detection [3, 12], semantic segmentation [1, 42, 48], and object classification [5, 49], *etc.* Unlike other discriminative tasks, gait recognition requires robust generalization to handle variations in environments, clothing, and other factors. Diffusion models have the potential to enhance gait representation extraction by effectively filtering out gait-irrelevant cues [16, 28, 59]. Therefore, we proposed knowledge-



(a) Baseline (b) Rank-1 Accuracy varying timestep t in Baseline (c) DenoisingGait

Figure 2. (a) A simple baseline on diffusion models for gait representation learning. (b) The rank-1 accuracy of our baseline with varying timestep t. (c) The pipeline of the proposed DenoisingGait.

driven approach is based on LDMs to coarsely filter out such gait-irrelevant cues in RGB videos by carefully manipulating the denoising timestep t.

2.3. Feature Vectorization for Gait Recognition

Feature vectorization has made substantial progress in gait recognition by capturing body structural details, which helps gait models reduce the impact of gait-irrelevant variations. [22, 40]. Similarly, optical flow is widely used for capturing pixel-wise motion across consecutive frames, effectively analyzing walking dynamics by tracking temporal movement patterns unique to individuals [13, 47, 51]. Inspired by the strengths of these traditional techniques, we propose a geometry-driven *feature matching* module that aims to leverage their core insights for gait representation refinement. Unlike previous works, which typically rely on predefined descriptors or handcrafted motion features, our *feature matching module* is a learnable, task-driven module. This allows it to better capture both the local structural details and the personalized motion details of pedestrians, contributing to improved gait representation extraction.

3. Method

This work aims to denoise walking videos for gait recognition through two proposed mechanisms: knowledge-driven denoising, leveraging generative diffusion models, and the geometry-driven denoising, enabled by a novel Feature Matching module. In this section, we first examine the application of diffusion models for gait denoising, followed by an introduction to the proposed DenoisingGait framework and its implementation details.

3.1. Diffusion Model for Gait Denoising

Diffusion Model Formulation. Given an image $x \sim p(x)$, the Latent Diffusion [29] projects it into a latent space by an efficient image encoder $\mathcal{E}$, *i.e.*, $z = \mathcal{E}(x)$. Then, the diffusion *forward* process [14] gradually diffuses the latent z into Gaussian noise z_T via a Markov Chain:

$$z_t = \sqrt{1-\beta_t} z_{t-1} + \sqrt{\beta_t}\epsilon, \tag{1}$$

where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ and $\{\beta_1, ..., \beta_T\}$ presents a fixed variance schedule of noising scales with T timesteps ($z_0 = z$).

For the *reverse* process, the Stable Diffusion [29] trains a time-conditional UNet [30] $\epsilon_\theta(z_t, t)$ to predict a variant of its input z_t. The objective can be simplified to the following equation [29]:

$$\mathcal{L}_{\text{LDM}} = \mathbb{E}_{z,\epsilon,t}[\|\epsilon - \epsilon_\theta(z_t, t)\|_2^2], \tag{2}$$

with t is uniformly sampled from $\{1, ..., T\}$.

During the inference, a Gaussian noise $\hat{z}_T$ will be sampled and then iteratively denoised by $\epsilon_\theta(\hat{z}_t, t)$ with the timestep ranging from T to 0. The last $\hat{z}_0$ can be decoded to image space with a single pass through the decoder $\mathcal{D}$.

Gait Recognition using Diffusion Models. Following the configures of Latent Diffusion [29], this section makes a pioneering step investigating the performance of diffusion-based gait representation learning.

Given a gait sequence $\{I_l \in \mathbb{R}^{H\times W\times 3} | l = 1, ..., L\}$, we first utilize the encoder $\mathcal{E}$ to project each frame into the latent space, followed by a one-step denoising using the pre-trained Stable Diffusion model ϵ_θ [29]:

$$F_l = \epsilon_\theta(\mathcal{E}(I_l), t), \tag{3}$$

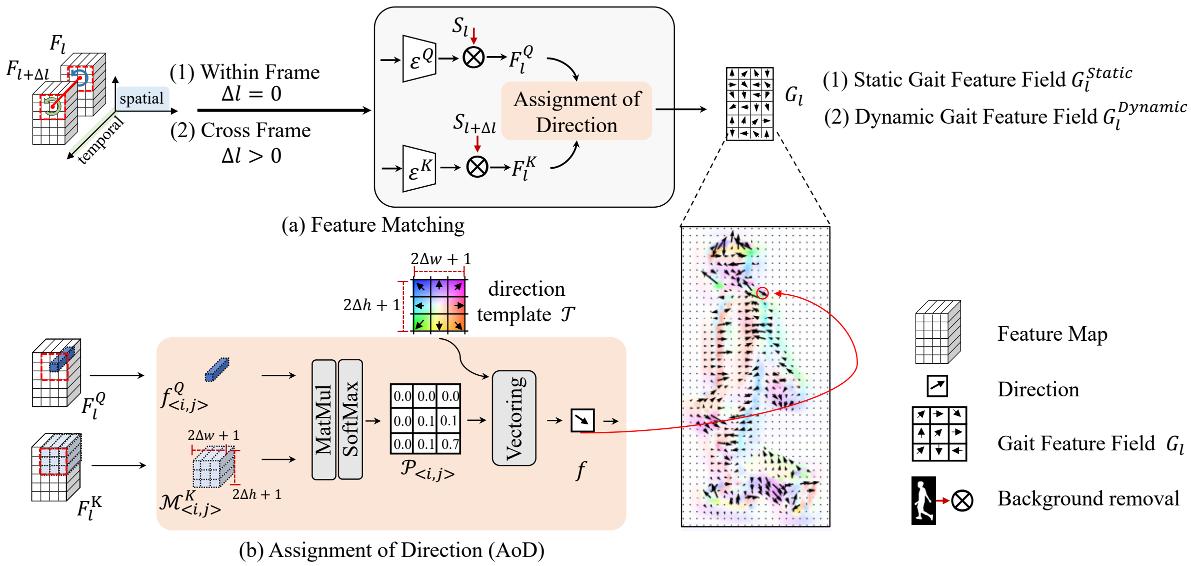


Figure 3. (a) our Feature Matching module. (b) Assignment of Direction (AoD). Background removal operation uses pre-extracted silhouettes to mask background regions.

where t represents the timestep, and $F_l \in \mathbb{R}^{\frac{H}{d} \times \frac{W}{d} \times 4}$ denotes the resulting diffusion features (with d as the downsampling factor introduced by the image encoder $\mathcal{E}$).

As shown in Figure 2 (a), we feed the denoised F_i into the popular gait recognition model GaitBase [9] to establish a baseline for subsequent analysis[2]. Here, we temporarily set aside experimental details (in Sec. 4) to focus on the performance trend in the most challenging full cloth-changing case from the CCPG [17]. Additional experiments on SUSTech1K [33] are included in Supplementary Material.

When the timestep t is gradually decreased from $T = 1,000$ to 100 with an interval of 200, as shown in Figure 2 (b), the rank-1 accuracy of our baseline initially rises and then falls, and the peak is at $t = 700$. The baseline shows a significant improvement of 5.3% over the red line in Figure 2 at the peak, and the red line represents the baseline without the denoising effects of ϵ_θ in Eq. 3. To our understanding, these findings reveal the following:

- Related works [16, 28, 59] indicate that in diffusion models, early timesteps ($t \to T$) tend to capture the overall shape of an image, while later timesteps ($t \to 0$) refine finer details. Figure 2 (b) illustrates a similar trend here: emphasizing an appropriate level of overall shape improves gait recognition accuracy, while excessive detail refinement leads to performance degradation. This suggests that the diffusion model ϵ_θ assists gait denoising by selectively filtering out RGB details that are not pertinent to gait information in the input image.
- Even at the optimal timestep $t = 700$, diffusion features decoded by $\mathcal{D}$ retain substantial RGB-encoded texture and color information.

Therefore, beyond the denoising driven by diffusion knowledge, we introduce an additional geometry-driven *feature matching* module, completing the proposed DenoisingGait framework, as shown in Figure 2 (c).

3.2. DenoisingGait

Building on the diffusion features F_l, the feature matching module aims to condense its multi-channel features at each pixel into a two-dimensional direction vector, thus reducing the expression of RGB-encoded noises while highlighting the local vectorized features of gait appearance and motion.

As shown in Figure 3 (a), the within-frame and cross-frame matching follow a similar workflow:

$$\begin{aligned} F_l^Q &= S_l \cdot \mathcal{E}^Q(F_l), \\ F_l^K &= S_{l+\Delta l} \cdot \mathcal{E}^K(F_{l+\Delta l}), \\ G_l &= \mathcal{A}(F_l^Q, F_l^K) \end{aligned} \tag{4}$$

where Δl denotes the temporal receptive scope and $\Delta l = 0$ for within-frame matching and $\Delta l > 0$ for cross-frame matching. The input $F_l \in \mathbb{R}^{\frac{H}{d} \times \frac{W}{d} \times 4}$ and $S_l \in \mathbb{R}^{\frac{H}{d} \times \frac{W}{d} \times 1}$ respectively represent the diffusion features and resized silhouette. The modules $\mathcal{E}^Q$ and $\mathcal{E}^K$ are convolutional stacks, each containing four layers with identical architectures but independently trained weights (the input and output channel are respectively set to 4 and C). Here, $\mathcal{A}$ handles the assignment of directions (AoD) based on the query feature $F_l^Q \in \mathbb{R}^{\frac{H}{d} \times \frac{W}{d} \times C}$ and key feature $F_l^K \in \mathbb{R}^{\frac{H}{d} \times \frac{W}{d} \times C}$. The resulting output $G_l \in \mathbb{R}^{\frac{H}{d} \times \frac{W}{d} \times 2}$ represent the gait feature

[2]Only the input channel is modified to match the output channel of ϵ_θ.

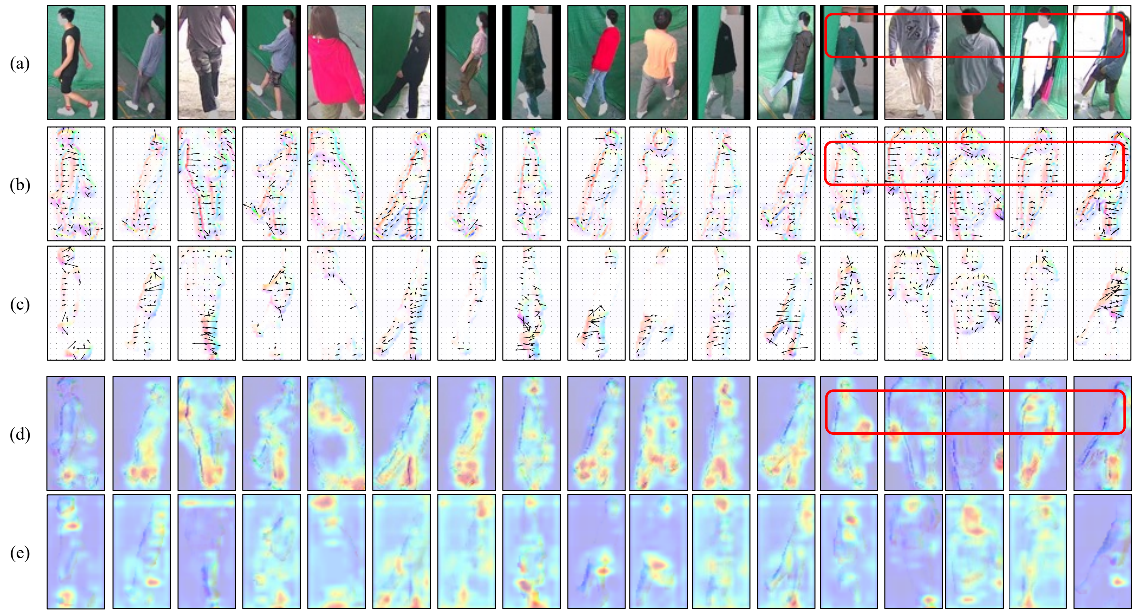


Figure 4. (a) Raw RGB images. (b) Static gait feature field, G^{Static}. (c) Dynamic gait feature field, G^{Dynamic}. (d) Activation focus on G^{Static}. (e) Activation focus on G^{Dynamic}. (For optimal viewing, please refer to the color version and zoom in.)

field, where each pixel corresponds to a two-dimensional direction vector.

Next we focus on the AoD operation $\mathcal{A}$. For clarity, the following formulation ignores the subscript of l (frame index) in Eq. 4. For a single pixel $<i, j>$ within the query feature F^Q, represented as $f^Q_{<i,j>} \in \mathbb{R}^{1\times C}$, we can identify its set of neighboring pixels within the key feature F^K as shown in Figure 3 (b), :

$$\mathcal{M}^K_{<i,j>} = \{f^K_{<\hat{i},\hat{j}>} | \hat{i} \in \{i - \Delta h, ..., i + \Delta h\}, \hat{j} \in \{j - \Delta w, ..., j + \Delta w\}\}, \tag{5}$$

For simplicity, we treat $\mathcal{M}^K_{<i,j>}$ as a matrix with elements arranged in raster order, *i.e.*, from top to bottom and left to right. Thus, $\mathcal{M}^K_{<i,j>}$ has a shape of $(2\Delta h + 1)(2\Delta w + 1) \times C$.

Then, we compute the similarity distribution between $f^Q_{<i,j>}$ with its neighboring pixels:

$$\mathcal{P}_{<i,j>} = \text{SoftMax}(f^Q_{<i,j>}(\mathcal{M}^K_{<i,j>})^\top) \tag{6}$$

where $\mathcal{P}_{<i,j>} \in \mathbb{R}^{1\times(2\Delta h+1)(2\Delta w+1)}$ and the SoftMax function is conducted along the last dimension.

To determine the final direction vector, we introduce a fixed direction template $\mathcal{T}$ as follows:

$$\mathcal{T} = \{[\hat{i}, \hat{j}] | \hat{i} \in \{-\Delta h, ..., \Delta h\}, \hat{j} \in \{-\Delta w, ..., \Delta w\}\}, \tag{7}$$

Obviously, there is an element-wise correspondence between the direction template $\mathcal{T}$ and neighboring pixels $\mathcal{M}^K_{<i,j>}$ (in Eq. 5). So we also treat $\mathcal{T}$ as a matrix with elements arranged in raster order, *i.e.*, from top to bottom and left to right. Thus, $\mathcal{T}$ has a shape of $(2\Delta h+1)(2\Delta w+1)\times 2$.

Now, we assign pixel $<i, j>$ with a direction vector:

$$G_{<i,j>} = \mathcal{P}_{<i,j>}\mathcal{T}, \tag{8}$$

During each inference step, we perform both within-frame ($\Delta l = 0$) and cross-frame matching ($\Delta l > 0$) in parallel, resulting in both the static gait feature filed G^{Static} and the dynamic gait feature field G^{Dynamic}.

Interestingly, we observe that the magnitude of direction vector in the static gait feature field, $\|G^{\text{Static}}_{<i,j>}\|_2$, often reflects image texture intensity. To enhance texture-invariant gait feature learning, we design a *texture suppression* operation that applies random zero-padding to high-magnitude pixels during training with a probability p:

$$G^{\text{Static}}_{<i,j>} = \begin{cases} \mathbf{0}, & \text{if} \quad \|G^{\text{Static}}_{<i,j>}\|_2 > m \\ G^{\text{Static}}_{<i,j>}, & \text{otherwise} \end{cases}. \tag{9}$$

This operation encourages DenoisingGait to recognize that texture elements are unreliable, thereby promoting texture-free learning of gait features.

As illustrated in Figure 2 (c), the static and dynamic gait feature fields are fed in parallel into the subsequent GaitBase [9]. To fit its multi-branch inputs, we adopt the widely-used high-level attention fusion strategy proposed by Fan et al. [11]. Consistent with recent works [7–9, 24, 43, 44, 56, 61], the training of DenoisingGait is driven by a combination of triplet loss and cross-entropy loss.

Table 1. Within-domain Evaluation on CCPG [17] (CL: full cloth-changing, UP: up-changing, DN: pant-changing, and BG: bag-changing).

Input	Model	Venue	Gait Evaluation Protocol					ReID Evaluation Protocol				
			CL	UP	DN	BG	Mean	CL	UP	DN	BG	Mean
Sils	GaitSet [2]	TPAMI'22	60.2	65.2	65.1	68.5	64.8	77.5	85.0	82.9	87.5	83.2
	GaitPart [7]	CVPR'20	64.3	67.8	68.6	71.7	68.1	79.2	85.3	86.5	88.0	84.8
	GaitBase [9]	CVPR'23	71.6	75.0	76.8	78.6	75.5	88.5	92.7	93.4	93.2	92.0
	DeepGaitV2 [8]	Arxiv	78.6	84.8	80.7	89.2	83.3	90.5	**96.3**	91.4	96.7	93.7
Flow	GaitBase^f	CVPR'23	70.0	74.5	77.7	77.5	74.9	82.4	88.9	90.9	91.5	88.4
Sils + Parsing	XGait [58]	MM'24	72.8	77.0	79.1	80.5	77.4	88.3	91.8	92.9	94.3	91.9
Sils + Parsing + Flow	MultiGait++ [15]	AAAI'25	83.9	**89.0**	86.0	91.5	87.6	-	-	-	-	-
Sils + Skeleton	BiFusion [27]	MTA'23	62.6	67.6	66.3	66.0	65.6	77.5	84.8	84.8	82.9	82.5
	SkeletonGait++ [11]	AAAI'24	79.1	83.9	81.7	89.9	83.7	90.2	95.0	92.9	96.9	93.8
RGB	BigGait [50]	CVPR'24	82.6	85.9	87.1	93.1	87.2	89.6	93.2	95.2	97.2	93.8
RGB+Sils	GaitEdge [19]	ECCV'22	66.9	74.0	70.6	77.1	72.2	73.0	83.5	82.0	87.8	81.6
	DenoisingGait	Ours	**84.0**	88.0	**90.1**	**95.9**	**89.5**	**91.8**	95.8	**96.4**	**98.7**	**95.7**

3.3. Understanding Gait Feature Field

Figure 4 displays the static and dynamic gait feature fields, G^{Static} and G^{Dynamic}, along with their activation focuses [31].

As observed, the static gait feature field G^{Static} (Figure 4 (b)) reveals a gradient-like representation of gait appearance, with each pixel's direction vector mostly oriented toward neighboring regions of high visual similarity:

- In related works [36, 45, 46], similar local vectorized features are employed through supervised learning to capture local image depth or structural characteristics. Although the proposed DenoisingGait is driven solely by identity signals, we assume these local details may effectively populate G^{Static} when they enhance human identification, guided by similar geometric constraints.
- As evidenced by the red boxes in Figure 4, both the representation visualization and activation focus of G^{Static} avoid texture-rich regions, thus encouraging texture-invariant gait feature learning.

As observed in Figure 4 (c) and (e), the dynamic gait feature field G^{Dynamic} captures a flow-like representation of gait motion, with each pixel's direction vector primarily aligned with moving body parts. But unlike traditional optical flow, the dynamic gait feature field G^{Dynamic} is fully recognition-oriented, with a focused objective to capture fine-grained gait motions for human identification. Additional video examples are provided in the **Supplementary Material**.

4. Experiment

4.1. Dataset

In our experiments, we used three widely recognized clothing-changing and multi-view gait datasets: CCPG [17], CASIA-B*[52], and SUSTech1K[32], chosen for their provision of RGB images. Among these, CCPG [17] serves as the primary benchmark due to its extensive and diverse variations in clothing. This dataset includes a wide range of coats, pants, and bags in various colors and styles, with faces and shoes masked to simulate real-world challenges for cloth-changing gait recognition.

Table 2. Implementation details. The batch size indicates the number of the IDs and the sequences per ID.

DataSet	Batch Size	Schedule	Frames	Steps
CCPG	(8, 4)	(20k, 40k, 50k)	20	60k
CASIA-B*	(8, 4)	(15k, 25k, 35k)	20	40k
SUSTech1K	(8, 4)	(15k, 25k, 35k)	20	40k

All implementations strictly adhere to the protocols set by dataset publishers. For performance reporting, we primarily follow the gait evaluation protocols for multi-view settings, using rank-1 accuracy as the main evaluation metric unless otherwise noted.

4.2. Implementation Details

(1) All images are processed with Pad-and-Resize [50] to keep body proportion; (2) We employ SD 1.5 [29] as the diffusion model for Figure 2; (3) (H, W, d) = (768, 384, 8), $(\Delta h, \Delta w)$ = (3, 3), m=0.5 for Section 3; (4) The main hyper-parameters are listed in Table 2; (5) The SGD optimizer with an initial learning rate of 0.1 and weight decay of 0.0005 is utilized; (6) During training, we adopt an ordered sampling strategy.

4.3. Experimental Results

Within-domain Evaluation. To show its superiority, DenoisingGait has been compared with various SoTA methods, including silhouette-based [2, 7–9, 21], multimodal-based [11, 15, 19, 27], and RGB-based methods [50].

As shown in Table 1, DenoisingGait achieves superior performance across both gait recognition and ReID evaluation protocols on the CCPG [17] dataset. Under the gait recognition protocal, DenoisingGait shows improvements

Table 3. Within-domain Evaluation on CASIA-B* [19, 52] and SUSTech1K [32] (Abbreviations: NM for normal, BG for bag, CL for clothing, CR for carrying, UB for umbrella, UN for uniform, OC for occlusion, and NT for night).

Input	Model	Venue	CASIA-B*			SUSTech1K									
			NM	BG	CL	NM	BG	CL	CR	UB	UN	OC	NT	R-1	R-5
Sils	GaitSet [2]	TPAMI'22	92.3	86.1	73.4	69.1	68.2	37.4	65.0	63.1	61.0	67.2	23.0	65.0	84.8
	GaitPart [7]	CVPR'20	96.2	91.5	78.7	62.2	62.8	33.1	59.5	57.2	54.8	57.2	21.7	59.2	80.8
	GaitBase [9]	CVPR'23	96.5	91.5	78.0	81.5	77.5	49.6	75.8	75.5	76.7	81.4	25.9	76.1	89.4
	DeepGaitV2 [8]	Arxiv	94.3	90.0	78.6	86.5	82.8	49.2	80.4	83.3	81.9	86.0	28.0	80.9	91.9
Sils+Skeleton	BiFusion [27]	MTA'23	93.0	78.1	68.3	69.8	62.3	45.4	60.9	54.3	63.5	77.8	33.7	62.1	83.4
	SkeletonGait++ [11]	AAAI'24	—	—	—	85.1	82.9	46.6	81.9	80.8	82.5	86.2	47.5	81.3	95.5
Sils+Parsing+Flow	MultiGait++ [15]	AAAI'25	—	—	—	92.0	89.4	50.4	87.6	89.7	89.1	93.4	45.1	87.4	95.6
RGB	BigGait [50]	CVPR'24	**100.0**	99.6	90.5	96.1	**97.0**	73.2	**97.2**	96.0	93.2	**99.3**	**85.3**	**96.2**	**98.7**
RGB+Sils	DenoisingGait	Ours	**100.0**	**99.9**	**91.5**	**98.4**	96.3	**79.0**	95.3	**97.1**	**94.7**	**99.3**	69.5	95.4	98.4

Table 4. Cross-domain Evaluation, in which all methods are trained on CCPG and tested on CASIA-B* and SUSTech1K.

Trained on **CCPG**

Model	CASIA-B*			SUSTech1K				
	NM	BG	CL	NM	BG	CL	UM	Overall
GaitSet [2]	47.4	40.9	25.8	11.5	14.5	8.2	11.0	12.8
GaitBase [9]	59.1	52.7	30.4	16.6	19.7	9.7	13.8	17.3
BigGait [50]	77.4	71.5	33.6	60.7	57.2	**43.7**	41.1	56.4
Ours	**83.9**	**76.1**	**34.8**	**66.9**	**59.7**	37.3	**45.7**	**59.1**

of +1.4% on CL subset, +2.1% on UP subset, +2.9% on DN subset, +2.8% on BG subset, and +2.3% on average across all subsets. These results highlight DenoisingGait's effectiveness in learning robust static and dynamic gait features by filtering out color and texture elements. The visualization of gait feature fields and their activation focus in Figure 4 also illustrate the robustness of DenoisingGait.

Table 3 shows the within-domain evaluation on CASIA-B* [52] and SUSTech1K [32]. In most cases, DenoisingGait outperforms other SoTA methods. However, its performance under night (NT) conditions on SUSTech1K [32] is somewhat limited, due to low-quality silhouettes in such settings. More details are provided in the Supplementary Material. Even so, DenoisingGait achieves notable gains over other silhouette-based methods, demonstrating its robustness.

The CASIA-B*[52] and SUSTech1K[32] datasets include limited clothing variations and do not mask faces or shoes. As noted by CCPG [17] dataset, evaluating RGB-based methods on these datasets may not accurately reflect real-world gait recognition performance, as algorithms can focus on the consistently visible features, such as face and shoes, which remain unchanged across samples. We align with this viewpoint and therefore recommend the CCPG [17] dataset as the primary benchmark for evaluating RGB-based gait recognition methods. Additionally, this work introduces cross-domain evaluation to further validate the effectiveness of DenoisingGait, as described below.

Table 5. Ablation study on the knowledge-driven denoising: Diffusion (Ours) *vs.* No Denoising (RGB Input) *vs.* Color Denoising (Gray Image Input) *vs.* DINOv2 and *vs.* Sils+Flow.

Input	Feature Matching	CCPG				
		CL	UP	DN	BG	R1
Diffusion Features	✓	**84.0**	**88.0**	**90.1**	**95.9**	**89.5**
	×	76.4	79.8	85.1	91.2	83.1
RGB Image	✓	77.9	83.0	87.7	94.4	85.8
	×	62.2	68.4	75.7	82.5	72.2
Gray Image	✓	75.6	81.4	86.2	92.0	83.8
DINOv2 Features	✓	78.4	83.2	85.7	92.3	84.9
Sils+Flow	N/A	79.5	83.1	84.0	84.6	82.8

Table 6. Ablation study on *feature matching* module.

Within-frame Matching	Cross-frame Matching	CCPG				
		CL	UP	DN	BG	R1
×	×	76.4	79.8	85.1	91.2	83.1
×	✓	68.4	73.6	80.9	87.9	77.7
✓	×	82.0	86.5	89.9	95.4	88.5
✓	✓	**84.0**	**88.0**	**90.1**	**95.9**	**89.5**

Cross-domain Evaluation. As shown in Table 4, cross-domain experiments (trained on CCPG, tested on CASIA-B* and SUSTech1K) further validate the generalization capability of DenoisingGait. We evaluated several SoTA methods based on silhouettes and RGB images. In most cases, DenoisingGait outperforms other SoTA methods. Specifically, it improves by +6.5% in NM, +4.6% in BG, and +1.2% in CL scenarios on CASIA-B*, and achieves an overall improvement of +2.7% on SUSTech1K. Although its CL performance has not reached the highest, it remains highly competitive. More cross-domain experiments can be found in the **Supplementary Material**.

4.4. Ablation Study

All ablation experiments are conducted on CCPG [17].

Knowledge-driven Denoising by Diffusion Model. Table 5 presents the results of incorporating DenoisingGait with various denoising knowledge: generative diffusion (Ours), no denoising (RGB input), color denoising (gray image input), DINOv2 representation [26], and traditional gait representation like silhouette and optical flow. Compared to RGB images, Table 5 demonstrates that the diffusion model effectively filters out gait-irrelevant cues in the videos. The performance improvement over grayscale images further emphasizes that the diffusion model not only removes color information but also filters out other non-essential cues. While DINOv2 excels in representation learning, it is not specifically designed for denoising and thus show no improvements for DenoisingGait. We also include two gait modalities, silhouette and optical flow. DenoisingGait outperforms GaitBase^{s+f} with improvements of +4.5% in the CL subset, +4.9% in UP, +6.1% in DN, +11.3% in BG, and +6.7% on average across all subsets. These results highlight the advantages of using the diffusion model as a knowledge-driven denoising module.

Geometry-driven Denoising by feature matching. Table 6 shows that both within-frame and cross-frame feature matching contribute to the improvements of DenoisingGait. The visualization and discussion of the respective roles of within-frame and cross-frame matching for gait description can be found in Sec. 3.3. It is worth mentioning that using only cross-frame matching, *i.e.*, taking only the dynamic gait feature fiedl G^{Dynamic} as input (similar to GaitBase^f [9], which uses the optical flow as input in Table 3), result in inferior performances. This aligns with the fact that static appearance features are crucial for gait description.

Table 7 presents additional ablation results on the within-frame feature matching mechanism. We conducted experiments to assess the impact of background removal (as defined in Eq.4) and texture suppression (as described in Eq.9). It is important to note that cross-frame feature matching was kept consistent throughout these experiments, and when background removal was excluded, cross-frame feature matching was also omitted. Without the background removal operation, recognition performance drops noticeably, indicating that irrelevant background factors interfere with our geometry-driven denoising process, thereby damaging gait feature extraction. With the background removal operation, we found that applying the texture suppression operation significantly improves the Rank-1 accuracy across all scenarios, especially with a +3.0% improvement in CL scenarios. This performance improvement indicates that the texture suppression operation effectively enhances DenoisingGait's texture-invariant gait feature learning.

Table 7. More ablation results on *Within-frame feature matching*.

Eq. 9, Texture Suppression	Eq. 4, Background Removal	CCPG				
		CL	UP	DN	BG	R1
×	×	73.1	78.1	84.3	91.5	81.8
✓	×	70.0	75.4	82.3	88.4	79.0
×	✓	81.0	86.5	89.7	95.7	88.2
✓	✓	**84.0**	**88.0**	**90.1**	**95.9**	**89.5**

5. Challenges and Conclusions

Challenges. Recent studies [50] highlight a pioneering trend of incorporating general knowledge of visual world into gait recognition. Related works often leverages large vision models, including discriminative models like DINOv2 [26] and generative models such as the Latent Diffusion Models [29] we use. However, these large models demand extensive training data, high computational costs, and substantial storage, all while advancing rapidly. Effectively utilizing and fairly comparing these diverse models, with due consideration of these costs, remains an unexplored issue for gait recognition. Additionally, fully extracting useful features from walking videos while filtering out identity-irrelevant factors to create reliable gait pattern descriptions continues to be a persistent challenge for RGB-based gait methods [37, 50, 55]. In many cases, the proportion of cross-clothing pairs in the training data plays a critical role [50], yet most of existing gait datasets contain only limited clothing changes. **Supplementary Material** shows that DenoisingGait encounters a similar issue.

Conclusions. The differences among pedestrians in videos are subtle since a video is in a very high dimensional space and with many noises. It makes gait recognition a very challenging task and is attracting increasing attentions. In our method, gait feature extraction is treated as a denoising process. This is achieved through the innovative use of diffusion model for knowledge-driven denoising and geometry-driven denoising by feature matching. In addition to achieving a new state-of-the-art, the success of using a generative model for a challenging discriminative task highlights the potential of generative models. Furthermore, the geometry-driven denoising based on feature matching transforms features into gait feature fields, effectively masking gait-irrelevant cues while enhancing both structural characteristics and motion features. We hope our work can inspire more researchers and encourage the exploration of additional gait denoising approaches in the future.

Acknowledgements

We thank Dingqiang Ye for his insightful discussions and help during the course of this work. This work was supported by the National Natural Science Foundation of China under Grant 62476120, and the Shenzhen International Research Cooperation Project under Grant GJHZ20220913142611021.

References

[1] Tomer Amit, Tal Shaharbany, Eliya Nachmani, and Lior Wolf. Segdiff: Image segmentation with diffusion probabilistic models. *arXiv preprint arXiv:2112.00390*, 2021. 2

[2] Hanqing Chao, Kun Wang, Yiwei He, Junping Zhang, and Jianfeng Feng. Gaitset: Cross-view gait recognition through utilizing gait as a deep set. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3467–3478, 2022. 1, 2, 6, 7

[3] Shoufa Chen, Peize Sun, Yibing Song, and Ping Luo. Diffusiondet: Diffusion model for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 19830–19843, 2023. 2

[4] Jooyoung Choi, Jungbeom Lee, Chaehun Shin, Sungwon Kim, Hyunwoo Kim, and Sungroh Yoon. Perception prioritized training of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11472–11481, 2022. 2

[5] Kevin Clark and Priyank Jaini. Text-to-image diffusion models are zero shot classifiers. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[6] Kamil Deja, Anna Kuzina, Tomasz Trzcinski, and Jakub Tomczak. On analyzing generative and denoising capabilities of diffusion-based deep generative models. *Advances in Neural Information Processing Systems*, 35:26218–26229, 2022. 2

[7] 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*, pages 14225–14233, 2020. 1, 2, 5, 6, 7

[8] Chao Fan, Saihui Hou, Yongzhen Huang, and Shiqi Yu. Exploring deep models for practical gait recognition. *arXiv preprint arXiv:2303.03301*, 2023. 2, 6, 7

[9] 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 (CVPR)*, pages 9707–9716, 2023. 4, 5, 6, 7, 8

[10] Chao Fan, Saihui Hou, Junhao Liang, Chuanfu Shen, Jingzhe Ma, Dongyang Jin, Yongzhen Huang, and Shiqi Yu. Opengait: A comprehensive benchmark study for gait recognition towards better practicality. *arXiv preprint arXiv:2405.09138*, 2024. 1

[11] Chao Fan, Jingzhe Ma, Dongyang Jin, Chuanfu Shen, and Shiqi Yu. Skeletongait: Gait recognition using skeleton maps. In *Proceedings of the AAAI conference on artificial intelligence*, pages 1662–1669, 2024. 1, 2, 5, 6, 7

[12] Haoyang Fang, Boran Han, Shuai Zhang, Su Zhou, Cuixiong Hu, and Wen-Ming Ye. Data augmentation for object detection via controllable diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1257–1266, 2024. 2

[13] Yuxiang Feng, Jiabin Yuan, and Lili Fan. Gaitfusion: Exploring the fusion of silhouettes and optical flow for gait recognition. In *Artificial Neural Networks and Machine Learning – ICANN 2023*, pages 88–99, Cham, 2023. Springer Nature Switzerland. 3

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2, 3

[15] Dongyang Jin, Chao Fan, Weihua Chen, and Shiqi Yu. Exploring more from multiple gait modalities for human identification. *arXiv preprint arXiv:2412.11495*, 2024. 6, 7

[16] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9492–9502, 2024. 2, 4

[17] Weijia Li, Saihui Hou, Chunjie Zhang, Chunshui Cao, Xu Liu, Yongzhen Huang, and Yao Zhao. An in-depth exploration of person re-identification and gait recognition in cloth-changing conditions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13824–13833, 2023. 2, 4, 6, 7

[18] Xiang Li, Yasushi Makihara, Chi Xu, Yasushi Yagi, Shiqi Yu, and Mingwu Ren. End-to-end model-based gait recognition. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 1, 2

[19] Junhao Liang, Chao Fan, Saihui Hou, Chuanfu Shen, Yongzhen Huang, and Shiqi Yu. Gaitedge: Beyond plain end-to-end gait recognition for better practicality. In *Computer Vision – ECCV 2022*, 2022. 2, 6, 7

[20] Rijun Liao, Chunshui Cao, Edel B Garcia, Shiqi Yu, and Yongzhen 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. 1, 2

[21] Beibei Lin, Shunli Zhang, and Xin 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. 1, 2, 6

[22] Zongyi Liu and S. Sarkar. Improved gait recognition by gait dynamics normalization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(6):863–876, 2006. 3

[23] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60:91–110, 2004. 2

[24] Jingzhe Ma, Dingqiang Ye, Chao Fan, and Shiqi Yu. Pedestrian attribute editing for gait recognition and anonymization. *arXiv preprint arXiv:2303.05076*, 2023. 5

[25] Mark S Nixon and John N Carter. Automatic recognition by gait. *Proceedings of the IEEE*, 94(11):2013–2024, 2006. 1

[26] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 8

[27] Yunjie Peng, Kang Ma, Yang Zhang, and Zhiqiang He. Learning rich features for gait recognition by integrating skeletons and silhouettes. *Multimedia Tools and Applications*, 83(3):7273–7294, 2024. 6, 7

[28] Zhiyuan Ren, Minchul Kim, Feng Liu, and Xiaoming Liu. Tiger: Time-varying denoising model for 3d point cloud generation with diffusion process. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9462–9471, 2024. 2, 4

[29] 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*, pages 10684–10695, 2022. 2, 3, 6, 8

[30] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, 2015. 3

[31] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 6

[32] Chuanfu Shen, Chao Fan, Wei Wu, Rui Wang, George Q. Huang, and Shiqi Yu. Lidargait: Benchmarking 3d gait recognition with point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1054–1063, 2023. 2, 6, 7

[33] Chuanfu Shen, Chao Fan, Wei Wu, Rui Wang, George Q Huang, and Shiqi 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. 4

[34] Chuanfu Shen, Beibei Lin, Shunli Zhang, Xin Yu, George Q Huang, and Shiqi Yu. Gait recognition with mask-based regularization. In *2023 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10. IEEE, 2023. 1

[35] Chuanfu Shen, Shiqi Yu, Jilong Wang, George Q Huang, and Liang Wang. A comprehensive survey on deep gait recognition: Algorithms, datasets, and challenges. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2024. 1

[36] Xiaoyu Shi, Zhaoyang Huang, Weikang Bian, Dasong Li, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, Jifeng Dai, and Hongsheng Li. Videoflow: Exploiting temporal cues for multi-frame optical flow estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12469–12480, 2023. 6

[37] Chunfeng Song, Yongzhen Huang, Yan Huang, Ning Jia, and Liang Wang. Gaitnet: An end-to-end network for gait based human identification. *Pattern recognition*, 96:106988, 2019. 8

[38] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. 2

[39] Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14453–14463, 2023. 2

[40] Dacheng Tao, Xuelong Li, Xindong Wu, and Stephen J Maybank. General tensor discriminant analysis and gabor features for gait recognition. *IEEE transactions on pattern analysis and machine intelligence*, 29(10):1700–1715, 2007. 3

[41] Torben Teepe, Ali Khan, Johannes Gilg, Fabian Herzog, Stefan Hörmann, and Gerhard Rigoll. Gaitgraph: graph convolutional network for skeleton-based gait recognition. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 2314–2318. IEEE, 2021. 1, 2

[42] Junjiao Tian, Lavisha Aggarwal, Andrea Colaco, Zsolt Kira, and Mar Gonzalez-Franco. Diffuse attend and segment: Unsupervised zero-shot segmentation using stable diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3554–3563, 2024. 2

[43] Rui Wang, Chuanfu Shen, Chao Fan, George Q Huang, and Shiqi Yu. Pointgait: Boosting end-to-end 3d gait recognition with point clouds via spatiotemporal modeling. In *2023 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10. IEEE, 2023. 5

[44] Rui Wang, Chuanfu Shen, Manuel J Marin-Jimenez, George Q Huang, and Shiqi Yu. Cross-modality gait recognition: Bridging lidar and camera modalities for human identification. In *2024 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–11. IEEE, 2024. 5

[45] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, and Dacheng Tao. Gmflow: Learning optical flow via global matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8121–8130, 2022. 6

[46] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, Fisher Yu, Dacheng Tao, and Andreas Geiger. Unifying flow, stereo and depth estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2, 6

[47] Jian Xu, Hai Li, and Shujuan Hou. Attention-based gait recognition network with novel partial representation pgofi based on prior motion information. *Digital Signal Processing*, 133:103845, 2023. 3

[48] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2955–2966, 2023. 2

[49] Xingyi Yang and Xinchao Wang. Diffusion model as representation learner. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18938–18949, 2023. 2

[50] Dingqiang Ye, Chao Fan, Jingzhe Ma, Xiaoming Liu, and Shiqi Yu. Biggait: Learning gait representation you want by large vision models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1, 2, 6, 7, 8

[51] Hongyi Ye, Tanfeng Sun, and Ke Xu. Gait recognition based on gait optical flow network with inherent feature pyramid. *Applied Sciences*, 13(19):10975, 2023. 3

[52] 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)*, pages 441–444. IEEE, 2006. 2, 6, 7

[53] Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. *arXiv preprint arXiv:2410.06940*, 2024. 2

[54] Zhongqi Yue, Jiankun Wang, Qianru Sun, Lei Ji, Eric I Chang, Hanwang Zhang, et al. Exploring diffusion timesteps for unsupervised representation learning. *arXiv preprint arXiv:2401.11430*, 2024. 2

[55] Ziyuan Zhang, Luan Tran, Feng Liu, and Xiaoming Liu. On learning disentangled representations for gait recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 1, 2, 8

[56] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2, 5

[57] Jinkai Zheng, Xinchen Liu, Shuai Wang, Lihao Wang, Chenggang Yan, and Wu Liu. Parsing is all you need for accurate gait recognition in the wild. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 116–124, 2023. 1, 2

[58] Jinkai Zheng, Xinchen Liu, Boyue Zhang, Chenggang Yan, Jiyong Zhang, Wu Liu, and Yongdong Zhang. It takes two: Accurate gait recognition in the wild via cross-granularity alignment. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 8786–8794, 2024. 6

[59] Yixuan Zhu, Ao Li, Yansong Tang, Wenliang Zhao, Jie Zhou, and Jiwen Lu. Dpmesh: Exploiting diffusion prior for occluded human mesh recovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1101–1110, 2024. 2, 4

[60] Shinan Zou, Chao Fan, Jianbo Xiong, Chuanfu Shen, Shiqi Yu, and Jin Tang. Cross-covariate gait recognition: A benchmark. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7855–7863, 2024. 1, 2

[61] Shinan Zou, Jianbo Xiong, Chao Fan, Chuanfu Shen, Shiqi Yu, and Jin Tang. A multi-stage adaptive feature fusion neural network for multimodal gait recognition. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2024. 5