On Denoising Walking Videos for Gait Recognition

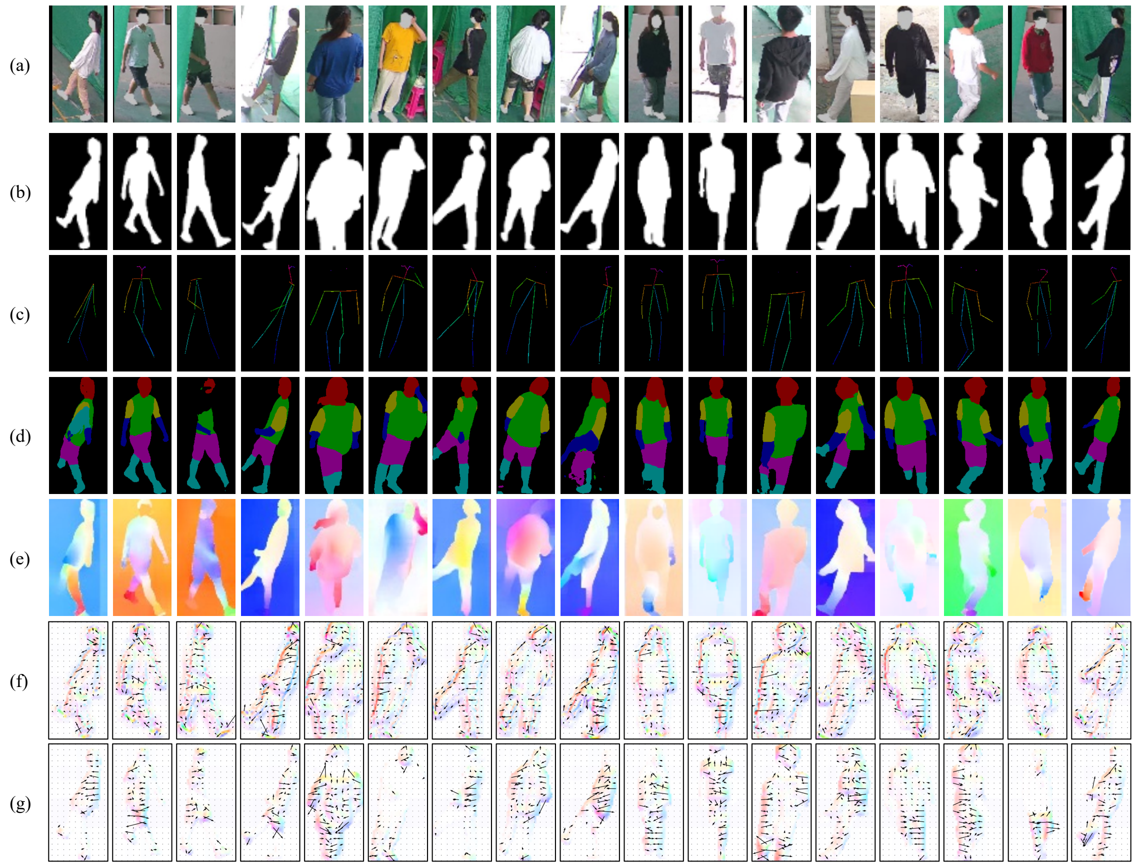

Figure 5. (a) Raw RGB images. (b) Silhouette images. (c) Skeleton images. (d) Human parsing images. (e) Optical flow images. (f) Static gait feature field, G^{Static}. (g) Dynamic gait feature field, G^{Dynamic}. (For optimal viewing, please refer to the color version and zoom in.)

6. Supplementary Material

In this section, we first provide more details of Gait Feature Field. Then more experimental results under both the within and crossdomain scenarios are presented. Some related issues in rebuttal are attached as well.

6.1. Understanding Gait Feature Field More

As illustrated in Figure 5, there are various vision modalities commonly used for gait description, including (but not limited to) binary silhouettes, skeleton coordinates, human parsing, and optical flow images. Typically derived from RGB videos, these modalities are provided by third-party tasks designed to express specific physical meanings, such as separating background from body regions or capturing joint-level and pixel-level walking movements. While these modalities effectively exclude gait-unrelated cues, it is important to note that their definitions are not explicitly tailored for identifying individuals based on gait. Many end-to-end works [15, 20, 25] argued this point and highlighted the superiority of global optimization in directly extracting gait characteristics from pedestrian videos.

In this study, the comparison between traditional optical flow and our dynamic gait feature field, both designed to capture pixel-level temporal dynamics, highlights a significant distinction between gait representations generated by third-party tasks and those specifically optimized for gait recognition. As illustrated in Figure 5 (e), optical flow images effectively depict smooth dynamics across the entire body. In contrast, the proposed dynamic gait feature field, illustrated in Figure 5 (g), adaptively adjusts the scale of movements at the pixel level. Because its learning process is entirely driven by recognition supervision, we hypothesize that the dynamic gait feature field effectively amplifies identity-related movements while suppressing those unrelated to identity.

Similarly, our static gait feature field, shown in Figure 5 (f), captures the vectorized local details of gait ap-

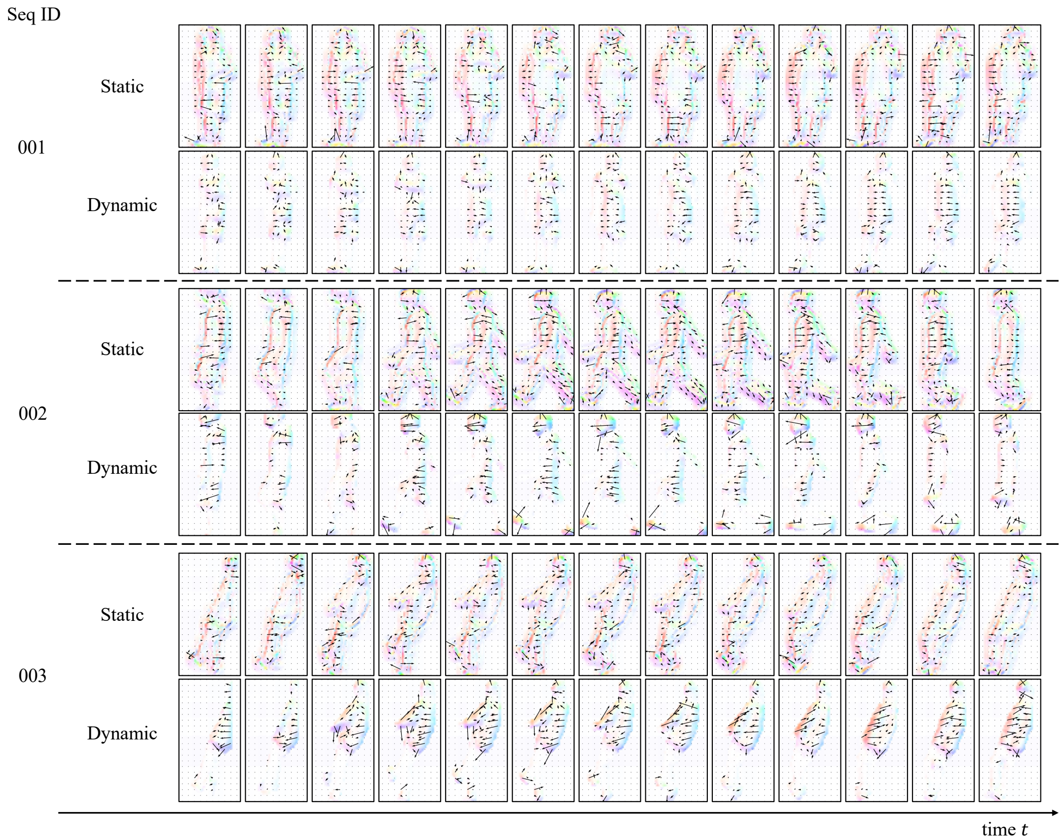

Figure 6. The sequences of gait feature field, G

Table 8. Within-domain evaluation on CCPG [9] (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
Skeleton	GaitGraph2 [16]	CVPRW'22	5.0	5.3	5.8	6.2	5.6	5.0	5.7	7.3	8.8	6.7
	GaitTR [24]	ES'23	15.7	18.3	18.5	17.5	17.5	24.3	28.7	31.1	28.1	28.1
	SkeletonGait [6]	AAAI'24	29.0	34.5	37.1	33.3	33.5	43.1	52.9	57.4	49.9	50.8
Sils	GaitSet [1]	TPAMI'22	60.2	65.2	65.1	68.5	64.8	77.5	85.0	82.9	87.5	83.2
	GaitPart [3]	CVPR'20	64.3	67.8	68.6	71.7	68.1	79.2	85.3	86.5	88.0	84.8
	AUG-OGBase [9]	CVPR'23	52.1	57.3	60.1	63.3	58.2	70.2	76.9	80.4	83.4	77.7
	GaitBase [5]	CVPR'23	71.6	75.0	76.8	78.6	75.5	88.5	92.7	93.4	93.2	92.0
	DeepGaitV2 [4]	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 [26]	MM'24	72.8	77.0	79.1	80.5	77.4	88.3	91.8	92.9	94.3	91.9
Sils + Flow	GaitBase^{s+f}	CVPR'23	79.5	83.1	84.0	84.6	82.8	90.2	93.9	94.2	93.3	92.9
Sils + Skeleton	BiFusion [12]	MTA'23	62.6	67.6	66.3	66.0	65.6	77.5	84.8	84.8	82.9	82.5
	SkeletonGait++ [6]	AAAI'24	79.1	83.9	81.7	89.9	83.7	90.2	95.0	92.9	96.9	93.8
RGB	AP3D [7]	ECCV'20	53.4	57.3	69.7	91.4	67.8	62.6	67.6	82.0	97.3	77.4
	PSTA [18]	ICCV'21	42.2	52.2	60.3	84.5	59.8	51.9	62.0	72.3	94.1	70.1
	PiT [23]	TII'22	41.0	47.6	64.3	91.0	61.0	49.1	56.2	78.0	96.9	70.1
	BigGait [20]	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 [10]	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**

Table 9. More cross-domain evaluation, where all methods are trained on one dataset and tested on the remaining two datasets.

(a) Trained on **CCPG** [9]

Model	Test Set										
	CASIA-B*				SUSTech1K						
	NM	BG	CL	Overall	NM	BG	CL	UB	UM	OC	Overall
GaitSet [1]	47.4	40.9	25.8	38.0	11.5	14.5	8.2	9.7	11.0	11.4	12.8
GaitBase [5]	59.1	52.7	30.4	47.4	16.6	19.7	9.7	11.8	13.8	16.8	17.3
AP3D [7]	53.7	46.2	11.9	37.3	**68.1**	52.4	36.2	42.6	38.3	**65.9**	55.3
PSTA [18]	49.7	42.3	8.8	33.6	51.4	37.8	25.7	33.8	26.8	52.5	40.6
BigGait [20]	77.4	71.5	33.6	60.8	60.7	57.2	**43.7**	48.5	41.1	63.6	56.4
Ours	**83.9**	**76.1**	**34.8**	**64.9**	66.9	**59.7**	37.3	**55.0**	**45.7**	64.0	**59.1**

(c) Trained on **SUSTech1K** [14]

Model	Test Set								
	CASIA-B*				CCPG				
	NM	BG	CL	Overall	CL	UP	DN	BG	Overall
GaitSet	63.3	50.8	26.4	46.8	14.0	**23.7**	20.3	43.2	25.3
GaitBase	73.1	61.2	**28.2**	54.2	**16.8**	21.7	**26.0**	42.7	**26.8**
AP3D	56.7	48.1	15.3	40.0	5.5	7.9	13.9	35.1	15.6
PSTA	31.2	27.7	10.6	23.2	3.7	5.7	9.5	26.5	11.4
BigGait	**91.1**	**85.8**	18.7	**65.2**	4.5	11.5	11.9	**45.5**	18.4
Ours	87.0	81.6	21.1	63.2	5.5	11.0	15.4	45.3	19.3

(b) Trained on **CASIA-B*** [22]

Model	Test Set											
	SUSTech1K							CCPG				
	NM	BG	CL	UB	UM	OC	Overall	CL	UP	DN	BG	Overall
GaitSet	13.6	13.8	7.2	10.3	10.3	11.5	12.8	**10.6**	16.4	17.2	24.9	17.3
GaitBase	19.2	16.7	8.1	12.0	14.5	15.6	15.6	**10.6**	18.1	**21.4**	28.7	19.7
AP3D	60.3	44.2	29.3	42.6	49.5	56.3	48.3	2.1	2.9	3.9	6.1	3.8
PSTA	47.4	33.2	19.9	25.5	33.0	43.4	34.6	1.7	1.9	3.4	5.0	3.0
BigGait	68.6	62.8	**36.9**	60.3	55.6	**68.9**	**64.8**	7.5	**19.5**	14.2	**43.0**	**24.6**
Ours	**69.8**	**63.5**	**36.9**	**64.4**	**57.1**	68.2	63.9	6.2	13.0	13.8	34.7	16.9

pearance, with high-magnitude pixels predominantly concentrated along the body's edge regions. This phenomenon aligns with the characteristics of human silhouettes and parsing images, as these edge regions effectively convey body and part shape features essential for gait understanding. Moreover, the vector-valued nature of the static gait feature field makes it more informative than traditional appearance-based gait modalities.

In summary, DeonisingGait effectively extracts recognition-oriented features by leveraging the proposed knowledge- and geometry-driven gait denoising priors.

6.2. More Experimental Results

More Within-domain Evaluation on CCPG. In Table 8, in addition to the content in the main text, we include several video-based ReID methods, including AP3D [7], PSTA [18], and PiT [23]. Compared to these methods, DenoisingGait outperforms them considerably, e.g., +30.6% for cloth-changing (CL), +30.7% for up-changing (UP), +20.4% for pant-changing (DN), and +4.5% for bag-changing (BG) scenarios. Accordingly, we consider that DenoisingGait can effectively remove gait-irrelevant cues from RGB videos and extract robust identity representations.

More Cross-domain Evaluation. In addition to comparing DenoisingGait with several state-of-the-art (SoTA) gait recognition methods, we include two video-based ReID methods, AP3D [7] and PSTA [18], as references. Table 9 presents cross-domain experiments conducted on CCPG, CASIA-B*, and SUSTech1K, where the model is trained on a certain dataset and evaluated on the other two.

The results reveal phenomena similar to those reported in previous studies [20]. Specifically, Table 9 illustrates how DenoisingGait's cross-domain performance varies depending on the training dataset. When trained on CCPG, DenoisingGait demonstrates strong adaptability to unseen datasets, outperforming both video-based ReID methods [7, 18], silhouette-based methods [1, 5], and the RGB-based method [20].

However, when trained on CASIA-B* or SUSTech1K, DenoisingGait encounters challenges in cross-dressing scenarios on CCPG, particularly in settings such as CL, UP, and DN. Table 9 presents the cross-domain experiments conducted on CCPG, CASIA-B*, and SUSTech1K, where the model is trained on one dataset and tested on the other two datasets.

This limitation, fortunately, can be addressed with more diverse training data. Compared to CASIA-B* and SUSTech1K, CCPG offers a broader range of outfit variations. As shown in Table 9 (a), training on CCPG allows DenoisingGait to develop more robust gait representations. In summary, the distribution of training data influences learned representations. Greater cross-dressing diversity improves performance in such scenarios, though achieving strong cross-dressing capability with limited diversity remains an open challenge.

6.3. Related Issues in Rebuttal

Q1: Comparison to Multi-Inputs. We developed multi-input (RGB+Sil) GaitBase [5] and BigGait [20], where

GaitBase uses silhouette-masked RGB and BigGait replaces the learned mask with silhouettes. Apart from this, the settings remain consistent with the original GaitBase and BigGait. As shown in Table 10, DenoisingGait remains the best performance on CCPG, while RGB+Sil BigGait performs even worse. We suspect that this drop may be due to the strong shape priors within silhouettes, which could prevent BigGait from learning better features from DINOv2.

Table 10. Comparison to multi-inputs on CCPG

CCPG	Input type	CL	UP	DN	BG	R1
GaitBase [5]	RGB+Sil	74.4	80.1	87.1	93.2	83.7
BigGait [20]	RGB	82.6	85.9	87.1	93.1	87.2
BigGait [20]	RGB+Sil	78.0	82.0	86.5	92.8	84.8
GaitEdge [10]	RGB+Sil	66.9	74.0	70.6	77.1	72.2
DenoisingGait	RGB+Sil	**84.0**	**88.0**	**90.1**	**95.9**	**89.5**

Q2: Justification for timestep *t*=700 in knowledge-driven denoising. Much evidence suggests that early timesteps ($t \to T$) in diffusion models mainly capture overall shapes, while later timesteps ($t \to 0$) focus on refining details [2, 8, 13, 17]. Based on this, we set timestep t=700 to retain overall shape features and partially mitigate identity-unrelated RGB details, as validated in Figure 2 (b). As shown in Table 11, more experiments on SUSTech1K confirm the effectiveness of timestep t=700, consistent with observations from CCPG in Figure 2 (b). Here, we focus on the challenging cloth-changing (CL) cases on both CCPG (Figure 2 (b)) and SUSTech1K (Table 11), where the color and texture of cloth become significant noise for gait recognition.

Table 11. Comparing Rank-1 Accuracy of different timestep t in Baseline on SUSTech1K.

SUSTech1K	t=1000	t=700	t=500	t=300	t=100
NM-cases	97.5	97.4	**97.6**	97.2	96.7
CL-cases	68.6	**76.5**	75.4	74.8	69.1
Mean-R1	94.6	**95.1**	**95.1**	94.4	93.5

Q3: About multi-timestep input. For the multi-timestep input, we test t={700, 500} and t={700, 500, 300} on CCPG. The Mean-R1 Accuracy improved by +0.6% and +0.9%, respectively, while the time costs increased to 2× and 3×.

Q4: NT (night)-cases on SUSTech1K. The NT-cases silhouette quality of SUSTech1K is poor. Table 12 presents experiments conducted on SUSTech1K. In this case, DenoisingGait outperforms GaitBase by +43.6%, showing its robustness under low-visibility conditions. Once we enhance the SUSTech1K silhouette quality (denoted by *, in Table 12), DenoisingGait improves from 69.5% to 90.2%, surpassing BigGait's 85.3%. Meanwhile, GaitBase improves from 25.9% to 68.9%.

Table 12. Comparing Rank-1 Accuracy on SUSTech1K.

SUSTech1K	GaitBase	GaitBase*	Ours	Ours*	BigGait
NT-cases	25.9	68.9	69.5	**90.2**	85.3
Mean-R1	76.1	85.2	95.4	**97.5**	96.2

Q5: The latent space F_l and Gait Feature Field can be noisy. Traditional gait inputs can also be noisy, *e.g.*, silhouette and parsing images often retains clothing shapes and even background, especially in in-the-wild imagery. DenoisingGait is designed to progressively remove identity-unrelated cues. While F_l only partially mitigate RGB noises, Feature Matching is further introduced to enhance denoising. We believe DenoisingGait's advantage is not from texture or color, as evidenced by:
a) In Table 8, it outperforms BigGait, while the BigGait [21] significantly surpasses ReID methods, despite the latter having full access to color and texture cues.
b) In Table 5, with denoising via Diffusion Features and Feature Matching, DenoisingGait achieves the best performance.
c) Feature and activation visualizations (Figure 4) further support this conclusion.
Q6: Vectors Pointing out of Body. These vectors are mainly located within the dynamic G^{Dynamic} (shown in Figure 4), revealing body movements (videos are in Figure 6). The directions are totally determined by neighboring visual similarity. Section 3.3 and 6.1 can provide more understandings.
Q7:About Global Matching. Integrating global matching into DenoisingGait yielded a slight 0.3% improvement. We assume that Local Matching, widely used in related works [11, 19], allows the CNN head to capture both local and global cues.
Q8: Visualize BG cases. As shown in Figure 7, DenoisingGait is still robust in this case (activations are not on BG regions, below).

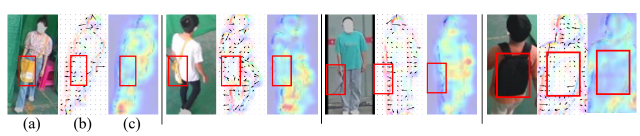

Figure 7. (a) Raw RGB image. (b) Static gait feature field, G^{Static}. (c) Activation focus on G^{Static}.

References

[1] 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. 2, 3

[2] Sherry X Chen, Yaron Vaxman, Elad Ben Baruch, David Asulin, Aviad Moreshet, Kuo-Chin Lien, Misha Sra, and

Pradeep Sen. Tino-edit: Timestep and noise optimization for robust diffusion-based image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6337–6346, 2024. 4

[3] 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. 2

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

[5] 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. 2, 3, 4

[6] 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. 2

[7] Xinqian Gu, Hong Chang, Bingpeng Ma, Hongkai Zhang, and Xilin Chen. Appearance-preserving 3d convolution for video-based person re-identification. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 228–243. Springer, 2020. 2, 3

[8] 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. 4

[9] 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, 3

[10] 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, 4

[11] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60:91–110, 2004. 4

[12] 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. 2

[13] 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. 4

[14] 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. 3

[15] 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. 1

[16] Torben Teepe, Johannes Gilg, Fabian Herzog, Stefan Hörmann, and Gerhard Rigoll. Towards a deeper understanding of skeleton-based gait recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1569–1577, 2022. 2

[17] 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. 4

[18] Yingquan Wang, Pingping Zhang, Shang Gao, Xia Geng, Hu Lu, and Dong Wang. Pyramid spatial-temporal aggregation for video-based person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12026–12035, 2021. 2, 3

[19] 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. 4

[20] 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, 3, 4

[21] 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. 4

[22] 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. 3

[23] Xianghao Zang, Ge Li, and Wei Gao. Multidirection and multiscale pyramid in transformer for video-based pedestrian retrieval. *IEEE Transactions on Industrial Informatics*, 18 (12):8776–8785, 2022. 2, 3

[24] Cun Zhang, Xing-Peng Chen, Guo-Qiang Han, and Xiang-Jie Liu. Spatial transformer network on skeleton-based gait recognition. *Expert Systems*, page e13244, 2023. 2

[25] 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

[26] 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. 2