BigGait: Learning Gait Representation You Want by Large Vision Models

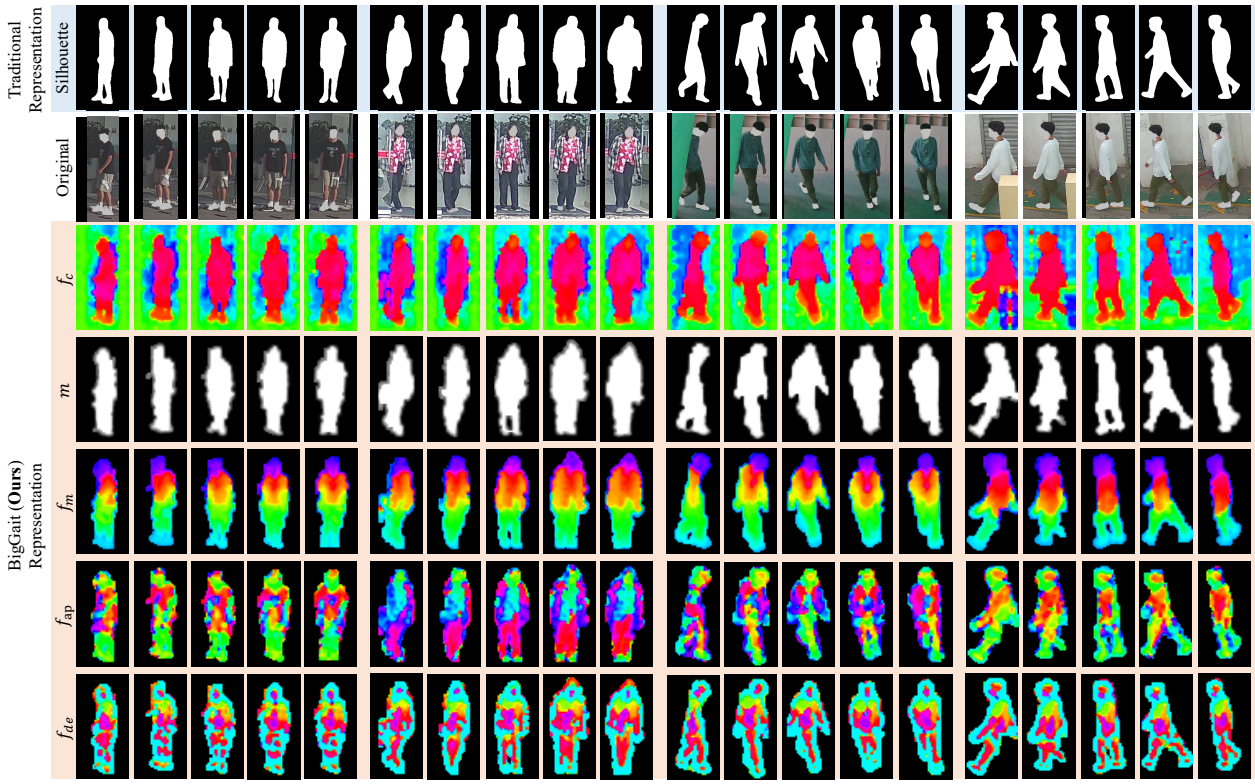

Figure 7. Visualization of intermediate representations. This figure is a supplement for Fig. 5.

7. Supplementary Material

In this section, we first provide the details of the Pad-and-Resize trick used for preserving body aspect ratio. Then more experimental results under both the within and cross-domain scenarios are presented. Finally, we conducted more visualization to understand the learned characteristics. Some related issues in rebuttal are attached as well.

7.1. Pad-and-Resize

Large vision models tend to directly resize the RGB frame within different bounding boxes into a fixed size, thereby completely or temporarily losing the authenticity of body proportions crucial for gait description. Comparing (a) with (b) in Tab.5, this lack of attention to body structure can lead to a significant performance drop. Therefore, this paper adopts a straightforward strategy, *i.e.*, conducting two-side padding or cutting horizontally to force the frame to be a predetermined aspect ratio of $2:1$, and then resizing it into a fixed resolution of 448×224, all while maintaining the original aspect ratio of body parts. Intuitively, the presence of double-sided black regions shown in Fig.2 (a) contributes to the preservation of limb ratios.

7.2. More Experimental Results

Within-domain Evaluation. Tab. 8 presents more within-domain performance comparisons. We observe that BigGait exhibited impressive performance on various kinds of datasets, including CCPG [7] with abundant clothing variations, CASIA-B* [14] focusing ups-clothing changes, and nearly cloth-unchanging SUSTech1K [11]. Compared with video-based ReID methods [6, 13], BigGait significantly outperforms them on the CCPG [7] comprising rich and challenging cloth-changes, and achieves similar performance on other two datasets [11, 14]. Compared with silhouette-based methods [1, 4], BigGait surpasses them on all of these datasets [7, 11, 14]. Moreover, Tab. 9 provides more within-domain results under ReID evaluation protocols on CCPG. The above results indicate that BigGait can

Table 8. **Within-domain Evaluation.** Rank-1 accuracy on the four popular benchmark datasets under the within-domain task: BigGait *v.s.* recent SoTA methods.

Type	Model	Within-domian Evaluation																
		CCPG [7]				CASIA-B* [14]			SUSTech1K [11]						CCGR [17]			
		CL	UP	DN	BG	NM	BG	CL	Normal	Umbrella	Uniform	Clothing	Night	Overall	R-1^{hard}	R-1^{easy}	R-5^{hard}	R-5^{easy}
Gait Recognition	GaitSet [1]	60.2	65.2	65.1	68.5	92.3	86.1	73.4	71.4	66.2	63.8	39.4	24.0	67.1	25.3	35.3	46.7	58.9
	GaitBase [4]	71.6	75.0	76.8	78.6	96.5	91.5	78.0	80.9	74.8	76.3	47.2	26.4	75.8	31.3	43.8	51.3	64.4
Video-based ReID	AP3D [6]	53.4	57.3	69.7	91.4	99.8	99.4	87.6	94.4	95.3	91.6	**82.7**	**89.4**	**96.8**	70.2	82.5	83.0	92.4
	PSTA [13]	42.2	52.2	60.3	84.5	98.2	96.5	54.2	92.9	92.1	83.2	72.3	79.9	93.6	74.5	85.0	86.2	93.7
Ours	BigGait	**82.6**	**85.9**	**87.1**	**93.1**	**100.0**	**99.6**	**90.5**	**96.1**	**96.0**	**93.2**	73.3	85.3	96.2	**77.1**	**86.2**	**87.9**	**94.3**

Table 9. **Within-domain Evaluation on CCPG.** This is a supplement for Tab.3, providing more results under ReID Evaluation Protocol.

Input	Model	Venue	Gait Evaluation Protocol					ReID Evaluation Protocol				
			CL	UP	DN	BG	Mean	CL	UP	DN	BG	Mean
Skeleton	GaitGraph2 [12]	CVPRW'22	5.0	5.3	5.8	6.2	5.6	5.0	5.7	7.3	8.8	6.7
	Gait-TR [16]	ES'23	15.7	18.3	18.5	17.5	17.5	24.3	28.7	31.1	28.1	28.1
	MSGG [9]	MTA'23	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 [2]	CVPR'20	64.3	67.8	68.6	71.7	68.1	79.2	85.3	86.5	88.0	84.8
	AUG-OGBase [7]	CVPR'23	52.1	57.3	60.1	63.3	58.2	70.2	76.9	80.4	83.4	77.7
	GaitBase [4]	CVPR'23	71.6	75.0	76.8	78.6	75.5	88.5	92.7	93.4	93.2	92.0
	DeepGaitV2 [3]	Arxiv	78.6	84.8	80.7	89.2	83.3	**90.5**	**96.3**	91.4	96.7	93.7
Parsing	GaitBase^{p}	CVPR'23	59.1	62.1	66.8	68.1	64.0	75.9	81.3	86.5	87.5	82.8
Parsing+Sils	GaitBase^{p+s}	CVPR'23	73.6	76.2	79.1	79.2	77.0	89.3	91.9	93.0	94.3	92.1
Skeleton+Sils	SkeletonGait++ [5]	AAAI'24	79.1	83.9	81.7	89.9	83.7	90.2	95.0	92.9	96.9	**93.8**
RGB+Sils	GaitEdge [8]	ECCV'22	66.9	74.0	70.6	77.1	72.2	73.0	83.5	82.0	87.8	81.6
RGB	AP3D [6]	ECCV'20	53.4	57.3	69.7	91.4	67.8	62.6	67.6	82.0	**97.3**	77.4
	PSTA [13]	ICCV'21	42.2	52.2	60.3	84.5	59.8	51.9	62.0	72.3	94.1	70.1
	PiT [15]	TII'22	41.0	47.6	64.3	91.0	61.0	49.1	56.2	78.0	96.9	70.1
	BigGait	Ours	**82.6**	**85.9**	**87.1**	**93.1**	**87.2**	89.6	93.2	**95.2**	97.2	**93.8**

extract robust gait patterns on different kinds of gait datasets and on different evaluation protocols.

CCGR Evaluation. CCGR [17] is a recently released well-labeled gait dataset consisting of over 1.5 million sequences, which has 970 subjects with 33 views and 53 walking conditions. We evaluate the performance of BigGait trained on CCGR under various tasks. As shown in Tab. 8 and Tab. 10, BigGait trained on CCGR surpasses all SoTA methods under the within and cross-domain tasks. Comparing Tab. 11 and Tab. 12, BigGait presents comparable performance with video-based ReID methods under the single-covariate task, and more outstanding performance under the mixed-covariate task than SoTA methods. Based on these results, we consider that BigGait learns robust gait representation to resist various covariates.

7.3. More Visualizations

To better understand the representation learned by BigGait, more visualizations are provided in Fig. 7 and Fig. 8.

Intermediate Feature Maps. Fig. 7 created by the PCA method exhibits a supplement for Fig. 5 and shows more intermediate feature maps. As we can see, all-purpose features f_c produced by the upstream DINOv2 are dominated by the separation of foreground and background regions accompanied by noisy spots. The mask branch in BigGait can automatically infer the foreground mask m from f_c in an unsupervised manner. Compared with silhouettes, m only presents the coarse-grained approximation of body segmentation. After masking the background regions, f_c becomes f_m and displays a parsing-like representation, *i.e.*, purple head, red abdomen, yellow arm, green leg, and blue shoe.

However directly using the all-purpose features f_m can result in inferior performances, as shown in Tab. 5 (c). We consider that all-purpose features f_m also contain gait-unrelated noise in foreground regions, like the noisy spots in the background regions of f_c. To alleviate this problem, the Gait Representation Extractor (GRE) is designed to extract effective gait representations from f_m while excluding gait-unrelated noise, as mentioned in Sec. 3. Specifically, GRE converts f_m into f_{ap} and f_{de}, with f_{ap} inheriting features by a linear transformation and showing body shape representation with high-frequency texture noise, and f_{de} embodying highly consistent skeleton-like structure representation by deploying soft geometric constraints to denoise most high-frequency texture noise.

Activation Maps. Fig. 8 obtained by the Grad-CAM [10]

Table 10. **Cross-domain Evaluation.** This table is a supplement for Tab. 4, in which all methods are trained on CCGR and tested on three unseen datasets.

Model	Test Set									
	CCPG				CASIA-B*			SUSTech1K		
	CL	UP	DN	BG	NM	BG	CL	Clothing	Night	Overall
GaitSet	7.6	13.4	16.2	30.1	32.8	22.2	12.9	13.9	16.9	21.6
GaitBase	5.8	10.2	15.7	26.3	22.1	14.1	7.5	13.4	16.5	27.6
AP3D	9.8	18.2	25.4	54.7	60.6	55.4	19.7	59.8	48.7	71.3
PSTA	10.1	17.9	22.0	52.7	31.4	27.8	14.2	55.0	40.3	65.4
BigGait	**20.8**	**38.2**	**31.9**	**83.6**	**93.1**	**91.8**	**61.7**	**73.8**	**76.8**	**88.1**

Table 11. **Single-Covariate Evaluation**: R-1^{easy} accuracy (%) with excluding identical-view cases on CCGR dataset.

Gallery: Normal 1					
Publication			CVPR'23	ICCV'21	Ours
Type	Covariate	Abbr.	GaitBase	PSTA	BigGait
Carrying	Book	BK	65.7	96.7	94.8
	Bag	BG	64.9	96.1	94.1
	Heavy Bag	HVBG	60.0	95.4	93.6
	Box	BX	61.5	95.6	93.5
	Heavy Box	HVBX	58.7	95.3	93.6
	Trolley Case	TC	64.1	94.2	93.0
	Umbrella	UB	47.2	89.5	85.4
	Average	-	60.3	**94.7**	92.6
Clothing	Thick Coat	CL	40.4	88.6	88.7
Road	Up Ramp	UTR	60.3	90.3	91.7
	Down Ramp	DTR	60.5	93.2	93.3
	Up Stair	UTS	54.9	92.0	92.5
	Down Stair	DTS	54.0	93.3	93.1
	Bumpy Road	BM	63.3	93.4	93.2
	Curved Road	CV	70.0	94.4	93.6
	Soft Road	SF	66.0	93.7	93.1
	Average	-	61.3	**92.9**	**92.9**
Speed	Normal 1	NM1	76.6	97.6	96.1
	Fast	FA	47.2	94.8	91.3
	Stationary	ST	32.0	92.3	88.1
	Average	-	51.9	**94.9**	91.8
Walking Style	Normal 2	NM2	75.3	97.5	95.8
	Confident	CF	64.9	96.1	94.3
	Freedom	FD	57.1	93.7	93.6
	Multi-person	MP	24.0	51.1	47.8
	Average	-	55.3	**84.6**	82.9

Table 12. **Mixed-Covariate Evaluation**: R-1^{easy} accuracy (%) with excluding identical-view cases on CCGR dataset. We use "-" to connect the mixed covariates.

Gallery: Normal 1				
Publication		CVPR'23	ICCV'21	Ours
Category	Covariate	GaitBase	PSTA	BigGait
Two Mixed	CL-UB	25.2	73.4	70.7
	HVBX-BG	52.1	94.5	92.7
	BG-TC	58.1	93.3	92.4
	SF-CL	36.1	82.0	87.0
	UTR-BX	51.0	88.9	90.2
	DTR-BK	55.1	93.1	93.3
	DTS-HVBX	42.6	91.9	92.4
	UTS-BG	46.8	90.5	91.6
	BM-CL	35.2	79.6	86.0
	CV-HVBX	61.0	94.3	93.3
	CL-CF	39.2	88.4	88.6
	Average	45.7	88.2	**88.9**
Three Mixed	CL-UB-BG	23.4	71.3	69.1
	BX-BG-CL	35.1	82.9	84.1
	BG-TC-CL	34.3	81.8	85.0
	SF-UB-BG	36.4	83.5	82.2
	UTR-HVBX-CL	31.8	77.0	83.9
	DTR-BK-BG	49.2	92.3	92.9
	DTS-HVBX-CL	26.4	74.9	83.4
	UTS-BG-CL	25.1	76.0	84.9
	BM-CL-BG	33.0	78.4	85.2
	CV-BX-BG	58.8	93.7	93.1
	UB-BG-FA	28.0	83.0	79.2
	Average	34.7	81.3	**83.9**
Four Mixed	CL-UB-BG-FA	16.2	67.5	63.8
	BM-CL-BG-BX	32.2	76.3	83.5
	BG-TC-CL-CV	38.0	79.4	86.4
	DTR-BK-BG-CL	32.2	78.8	86.9
	DTS-BX-CL-BG	25.6	73.2	82.7
	SF-UB-BG-CL	20.6	66.7	69.3
	BG-TC-CL-ST	11.7	66.8	66.5
	UTS-UB-BG-CL	15.8	61.6	70.3
	Average	24.0	71.3	**76.2**
Five Mixed	BG-TC-CL-CV-UB	34.1	69.6	**73.8**
	UTR-BG-CL-BX-CV	31.3	73.8	**84.8**

algorithm exhibits a supplement for Fig. 6 and shows more activation maps on BigGait and video-based ReID methods [6, 13, 15]. The visualization insights are the same as in Fig. 6.

7.4. Related Issues in Rebuttal

Q1: BigGait's Representation can be noisy. We argue a representation should be validated by the performance statistics drawn from a large test set, instead of one or two visual examples. Traditional gait representations could be noisy too, *e.g.*, silhouette includes clothing shapes and/or segmentation errors on in-the-wild imagery. BigGait does demonstrate superiority in cross-clothing and cross-domain tasks, where background/accessory are different between training vs testing data. More discussions are in Sec. 4.2.

Q2: Background info inflates BigGait's performance. We believe BigGait's superiority is not due to the background, from 3 observations. 1) Compared Tab. 5 (a) with (d), including backgrounds in BigGait harms its performance. 2) In Tab. 3 and 4, BigGait outperforms ReID methods by large margins though the latter can see full backgrounds. 3) The visualization of activation maps in Fig. 6

Figure 8. Visualization of activation maps. Unlike video-based ReID methods, BigGait focus on robust gait pattern rather than background and clothing texture noises. This figure is a supplement for Fig. 6.

further reflects that BigGait focuses on foreground regions.

Q3: How to handle color noises? BigGait regards texture noises as high-frequency signals. Thus we assume that here color noises refer to low-frequency ones. Thanks to the fitting power of neural networks, training models with cross-clothing pairs can partially learn the immunity to these noises (texture and color). But the red box of Fig. 6 shows that it is the high-frequency that still heavily impacts existing RGB-based methods. Hence, we consider high-frequency textures as the primary challenge, and meanwhile, look forward to further improvements brought by color-specific designs. Thanks for providing this insight. The revision will discuss it in detail.

Q4: Why not directly using f_{de}? Some gait-related features like the body shape may be partially damaged by the geometrical constraints of the denoising branch, while carefully preserved by the appearance branch. In light of this, we choose to fuse f_{de} and f_{ap} by attention mechanisms, as shown in Fig. 3 and supported by Tab. 5.

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 (TPAMI)*, 44 (7):3467–3478, 2022. 1, 2

[2] 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 *Computer Vision and Pattern Recognition (CVPR)*, pages 14225–14233, 2020. 2

[3] Chao Fan, Saihui Hou, Yongzhen Huang, and Shiqi Yu. Exploring Deep Models for Practical Gait Recognition. *arXiv preprint arXiv:2303.03301*, 2023. 2

[4] Chao Fan, Junhao Liang, Chuanfu Shen, Saihui Hou, Yongzhen Huang, and Shiqi Yu. OpenGait: Revisiting Gait Recognition Towards Better Practicality. In *Computer Vision and Pattern Recognition (CVPR)*, pages 9707–9716, 2023. 1, 2

[5] Chao Fan, Jingzhe Ma, Dongyang Jin, Chuanfu Shen, and Shiqi Yu. SkeletonGait: Gait Recognition Using Skeleton Maps. In *Association for the Advancement of Artificial Intelligence (AAAI)*, page [in press], 2024. 2

[6] Xinqian Gu, Hong Chang, Bingpeng Ma, Hongkai Zhang, and Xilin Chen. Appearance-Preserving 3D Convolution for Video-based Person Re-identification. In *European Conference on Computer Vision (ECCV)*, pages 228–243, 2020. 1, 2, 3

[7] 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 *Computer Vision and Pattern Recognition (CVPR)*, pages 13824–13833, 2023. 1, 2

[8] 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 *European Conference on Computer Vision (ECCV)*, pages 375–390, 2022. 2

[9] 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

[10] 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 *International Conference on Computer Vision (ICCV)*, pages 618–626, 2017. 2

[11] Chuanfu Shen, Chao Fan, Wei Wu, Rui Wang, George Q. Huang, and Shiqi Yu. LidarGait: Benchmarking 3D Gait Recognition With Point Clouds. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1054–1063, 2023. 1, 2

[12] Torben Teepe, Johannes Gilg, Fabian Herzog, Stefan Hörmann, and Gerhard Rigoll. Towards a Deeper Understanding of Skeleton-based Gait Recognition. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1569–1577, 2022. 2

[13] Yingquan Wang, Pingping Zhang, Shang Gao, Xia Geng, Hu Lu, and Dong Wang. Pyramid Spatial-Temporal Aggregation for Video-based Person Re-Identification. In *International Conference on Computer Vision (ICCV)*, pages 12026–12035, 2021. 1, 2, 3

[14] 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)*, pages 441–444, 2006. 1, 2

[15] Xianghao Zang, Ge Li, and Wei Gao. Multi-direction and Multi-scale Pyramid in Transformer for Video-based Pedestrian Retrieval. *IEEE Transactions on Industrial Informatics*, 18(12):8776–8785, 2022. 2, 3

[16] Cun Zhang, Xing-Peng Chen, Guo-Qiang Han, and Xiang-Jie Liu. Spatial Transformer Network on Skeleton-based Gait Recognition. *Expert Systems*, 40(6):e13244, 2023. 2

[17] Shinan Zou, Chao Fan, Jianbo Xiong, Chuanfu Shen, Shiqi Yu, and Jin Tang. Cross-Covariate Gait Recognition: A Benchmark. In *Association for the Advancement of Artificial Intelligence (AAAI)*, page [in press], 2024. 2