BarbieGait: An Identity-Consistent Synthetic Human Dataset with Versatile Cloth-Changing for Gait Recognition

Supplementary Material

7. Details of Kinematic Motion Matching

Our kinematic matching process begins by extracting bone rotations from the input 3D keypoint sequence. Since the source 3D pose data only provides independent joint positions without a pre-defined articulated structure (i.e., no rigged skeleton or skinning rig), we construct a local coordinate system for each "virtual bone segment" (i.e., a conceptual bone defined by two joints) based on joint-to-joint geometric relationships, enabling the extraction of pose parameters required for motion retargeting.

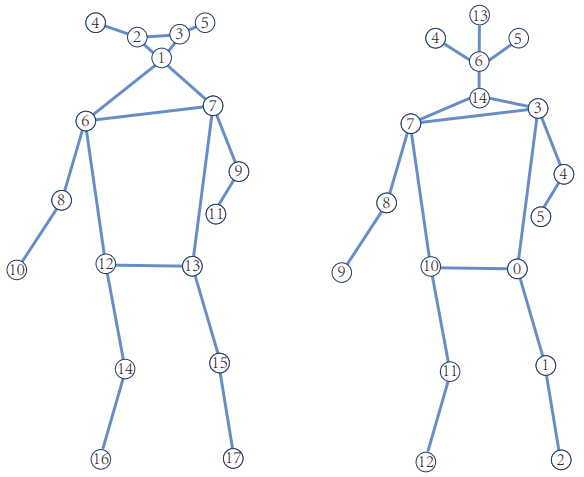
Specifically, for a bone defined by a parent joint J_p and a child joint J_c, its primary axis is given by the vector $v = J_c - J_p$. To resolve the inherent twist ambiguity around this axis, we form a reference plane using adjacent joints to obtain a stable and reproducible secondary axis direction. This procedure uniquely determines the local coordinate system of each bone.

Based on these dynamically constructed local coordinate systems, we compute the world-space rotation quaternion of each bone for both the source and target skeletons, denoted as Q_s and Q_t in Algorithm 1. By further composing the child bone's world rotation with the inverse world rotation of its parent, we obtain the local rotation $\Delta Q_t(k_p)$, which characterizes the hierarchical relative motion of the source pose. All local rotations are then assembled according to the skeletal topology to form a complete hierarchical pose representation. Finally, these local rotations are applied to the target skeleton to accomplish the retargeting process via hierarchical bone rotation.

A more complete treatment of the analytical quaternion computation and the Blender implementation details underlying this kinematic matching pipeline can be found in [1, 2].

8. More Cloth-Changing Experiments

We additionally evaluate GaitCLIF on established cloth-changing benchmarks, including HybridGait [6], OU-ISIR [19], CCVID [10], and MEVID [5]. As shown in Table 9, GaitCLIF consistently improves performance under cloth-changing settings over DeepGaitV2 across both gait recognition and video Re-identification benchmarks. Specifically, it gains of 4.77 and 3.50 points on HybridGait and OU-ISIR, respectively, and further improves performance by 3.90 and 4.11 points on CCVID and MEVID, respectively. For gait recognition on HybridGait and OU-ISIR, we follow OpenGait [7], using 64×44 silhouettes and sampling 30 frames from each sequence during training. For video Re-identification on CCVID and MEVID, we follow CCVID [10], using 128×88 RGB images and sampling 8 frames from each sequence during training.



(a) COCO-17 (b) Barbie-17

Figure 5. The pose format we used in our experiments. (a) COCO-17 format (b)Barbie-17 format.

Table 9. Performance comparison of DeepGaitV2 and GaitCLIF on additional cloth-changing benchmarks.

Methods	Gait Recognition		Video Re-ID	
	HybridGait	OU-ISIR	CCVID	MEVID
DeepGaitV2 [9]	55.34	90.54	93.18	69.94
GaitCLIF (ours)	**60.11**	**94.04**	**97.08**	**74.05**

9. Additional Ablation Studies

9.1. Effectiveness of Each Module

Due to space limitations, Table 4 reports only the averaged performance on BarbieGait. A more comprehensive evaluation under all nine clothing conditions (THK1–THK9) is provided in Table 10. As shown in the table, both GON-P3D and GON-FC contribute positively to cross-clothing robustness, each improving the baseline to varying degrees. The improvements are consistent across all settings and are particularly evident in the challenging THK7–THK9 conditions, highlighting the strong robustness and generalizability of our design.

GON-P3D	GON-FC	THK1		THK2		THK3		THK4		THK5		THK6		THK7		THK8		THK9		AVG	
		R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP
×	×	85.4	72.4	81.0	68.8	76.7	64.8	76.1	64.4	72.8	59.9	63.1	54.2	50.9	44.8	55.3	46.9	47.9	42.3	67.7	57.6
√	×	86.1	71.3	81.8	68.0	78.3	64.3	78.2	64.2	75.3	59.8	64.7	54.2	53.6	45.8	58.9	47.2	51.2	43.6	69.8	57.6
×	√	83.8	71.7	80.1	68.6	76.8	65.2	76.9	65.2	73.1	60.7	65.8	56.6	53.4	46.8	59.7	50.3	53.3	46.4	69.2	59.1
√	√	**88.1**	**74.2**	**84.8**	**71.5**	**82.0**	**68.3**	**82.3**	**68.5**	**80.1**	**65.1**	**71.6**	**60.5**	**62.8**	**53.5**	**67.5**	**55.1**	**61.1**	**51.9**	**75.6**	**63.2**

Table 10. Ablation study of each module under different clothing conditions (THK1-THK9). We report Rank-1 (R1) accuracy and mean Average Precision (mAP) for each variant.

Norm Type	THK1		THK2		THK3		THK4		THK5		THK6		THK7		THK8		THK9		AVG	
	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP
BN	83.2	71.8	78.6	68.0	74.2	63.8	73.6	63.2	69.1	58.3	60.9	53.6	47.8	43.0	54.7	46.9	45.9	41.6	65.3	56.7
IN	75.4	67.2	71.0	63.7	66.2	59.9	65.1	58.9	61.0	55.5	52.3	49.1	43.0	41.6	46.6	44.3	38.0	37.5	57.6	53.1
LN	76.9	61.7	71.9	58.4	66.7	54.6	66.3	54.3	61.9	50.4	53.5	45.8	43.3	38.4	50.1	42.0	41.1	35.8	59.1	49.0
GON	**88.1**	**74.2**	**84.8**	**71.5**	**82.0**	**68.3**	**82.3**	**68.5**	**80.1**	**65.1**	**71.6**	**60.5**	**62.8**	**53.5**	**67.5**	**55.1**	**61.1**	**51.9**	**75.6**	**63.2**

Table 11. Comparison of common normalization methods (BN, IN, LN) and our proposed GON across different clothing thickness levels.

9.2. Ablations of the type of Normalization

To further examine the effectiveness of our GON module, we conduct an additional ablation comparing it with commonly used normalization strategies, including Instance Normalization (IN), Batch Normalization (BN), and Layer Normalization (LN). By replacing GON with each standard normalization type while keeping all other components unchanged, we obtain a clear comparison in Table 11 that highlights the effects of *Gait-Oriented Normalization* in challenging cloth-changing scenarios.

10. Enhancing the Upstream Pose Estimator

As an informative representation of human motion, 2D pose provides a stable and clothing-invariant description of gait through a fixed set of keypoints. From Table 3, the best-performing pose-based methods already surpass the best silhouette-based methods, indicating the strong potential of keypoint representations for cross-clothing gait recognition. However, current 2D pose estimation models [14, 23, 24] are predominantly trained on action-oriented and motion-oriented datasets [3, 13, 15, 18, 22]. These datasets contain diverse activities and large motion amplitudes but involve only a limited number of subjects and lack clothing variation. In particular, they do not provide large-scale cross-clothing sequences for the same identity, making them insufficient as upstream supervision for cross-clothing gait recognition.

To enhance both the identity preservation capability and the generalization ability of upstream pose estimators under clothing variations, we extract one image every 10 frames from each sequence in BarbieGait, resulting in a 953K-image training set, which is significantly larger than the commonly used MS COCO [17] dataset with 150K images. Beyond its scale, BarbieGait also provides richer gait-specific motion patterns and identity-consistent clothing variations, aligning more closely with the requirements of downstream gait tasks.

For a fair comparison with MS COCO–based methods, we ensure that our pose estimator predicts the same number of keypoints and maintains comparable body semantics. As shown in Figure 5, both the COCO-17 format (the standard 17-keypoint layout used in MS COCO) and our Barbie-17 format (the 17-keypoint layout defined in BarbieGait) contain the same total number of joints. Following the keypoint selection strategy adopted in [16], we keep the COCO-style semantics for the 12 skeletal joints of the limbs and torso, while replacing the original COCO head keypoints with 5 mesh-derived head landmarks that provide more stable and anatomically reliable supervision.

For training, we adopt ViTPose-H [24] implemented in MMPose [4] as our backbone. The model is initialized with MAE [11] pre-trained weights and trained with default MMPose settings: an input size of 256×192, the AdamW [20] optimizer with a learning rate of 1e-3, UDP [12] post-processing, a batch size of 512, and 20 training epochs with learning rate decay at epochs 8 and 16.

Ultimately, by deploying our retrained pose estimation model, we achieve substantial performance improvements on publicly available RGB gait datasets, including CCPG and SUSTech1K. By incorporating our retrained pose model into the SkeletonGait [8] pipeline, we obtain new state-of-the-art performance on these real-world benchmarks, as shown in Table 12.

Figure 6. The Illustration of our diverse clothing. BarbieGait includes a variety of hairstyles, clothing, shoes, and carried objects, introducing significant clothing variations for gait recognition under cloth-changing conditions.

Methods	Pose Type	Upstream Model	CCPG								SUSTech1K									
			CL		UP		DN		Mean		NM	BG	CL	CA	UM	UN	OC	NG	Overall	
			R1 & mAP								R1								R1	R5
GaitTR [25]	COCO-17	HRNet [23]	24.3	9.7	28.7	16.1	31.1	16.4	28.0	14.1	33.3	31.5	21.0	30.4	22.7	34.6	44.9	23.5	72.6	56.0
GaitGraph [21]	COCO-17	HRNet [23]	5.0	2.4	5.7	4.0	7.3	4.2	6.0	3.5	22.2	18.2	6.8	18.6	13.4	19.2	27.3	16.4	18.6	40.2
SkeletonGait [8]	COCO-17	HRNet [23]	52.4	20.8	65.4	35.8	72.8	40.3	63.5	32.3	55.0	51.0	24.7	49.9	42.3	52.0	62.8	43.9	50.1	72.6
DPGait [16]	COCO-17	ViTPose [24]	70.7	—	82.4	—	84.2	—	79.1	—	—	—	—	—	—	—	—	—	—	—
Ours	Barbie-17	ViTPose [24]	**76.9**	**40.0**	**84.9**	**56.7**	**87.2**	**57.7**	**83.0**	**51.5**	**64.3**	**50.6**	**39.9**	**59.0**	**52.6**	**61.9**	**69.8**	**40.1**	**59.7**	**80.4**

Table 12. BarbieGait improves upstream pose estimation, enabling better generalization to real-world dataset CCPG and SUSTech1K.

11. Additional Visualization

11.1. Visualization of Synthetic Images

We also present additional synthetic images from BarbieGait in Figure 7. We select two subjects, each shown in four different scenes, including indoor scenes and outdoor scenes under varying lighting conditions. We simulate **lighting variations** as realistically as possible by setting appropriate lighting conditions in both indoor and outdoor environments. For example, indoors, we simulated incandescent lighting conditions, while outdoors, we mimicked sunlight variations. In addition, the subjects naturally interact with scene objects, resulting in realistic **occlusions**. The realistic scene and lighting simulation, combined with real gait data sources, ensure the validity and authenticity of BarbieGait as a synthetic gait dataset.

11.2. Visualization of Heatmaps

Figure 8 provides qualitative insights into the impact of clothing. As clothing complexity increases, crucial gait information is obscured while irrelevant clothing details are introduced in silhouette-based methods. If the model focuses excessively on clothing, it fails to capture critical gait features. Comparing Figure 8(b) and 8(c), DeepGaitV2 focuses more on the clothing areas, while GaitCLIF concentrates on unoccluded and discriminative areas such as body joints and edges. This demonstrates that GaitCLIF effectively removes clothing stylization and guides the model to focus more on gait information that is independent of clothing. Pose-based methods are less affected by clothing than silhouette-based methods but provide limited useful information. This requires the model to focus more on fine-grained features. As shown in (e), SkeletonGait mainly activates the entire skeleton map, while with the help of

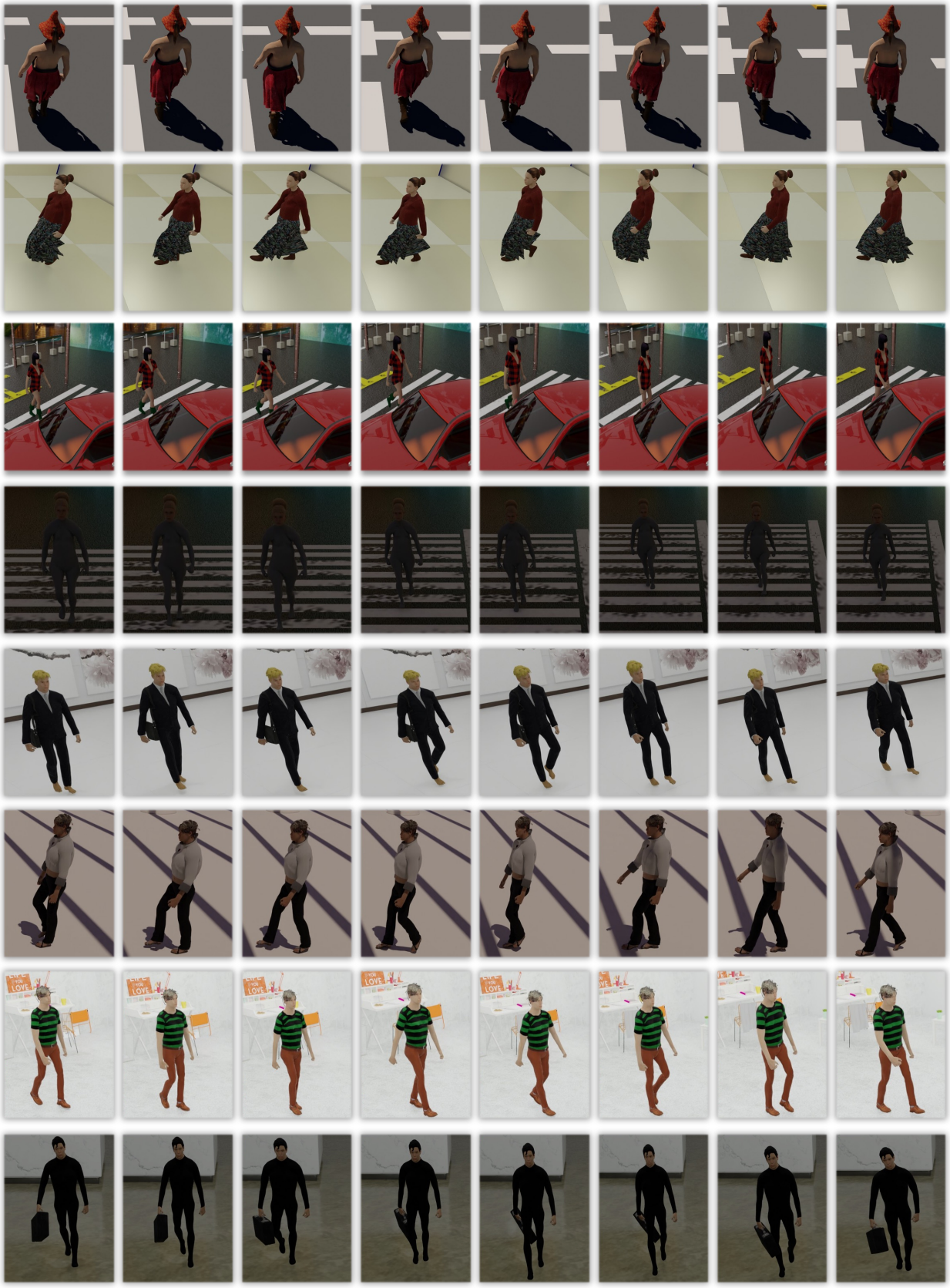


Figure 7. The illustration of our synthesized images. Our synthetic images are rendered in different scenes, realistic lighting conditions, diverse clothing conditions, and natural occlusions.

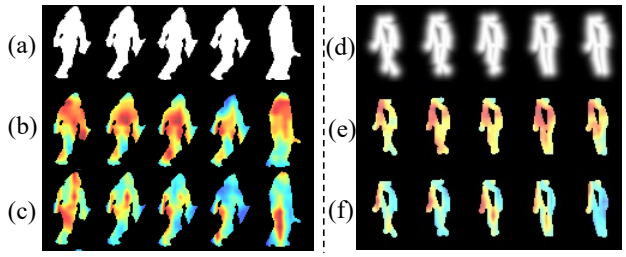


Figure 8. Visualization of heatmaps in Silhouette-based (a)-(c) and Pose-based methods (d)-(f). (b) and (e) show activation heatmaps of DeepGaitV2 and SkeletonGait overlaid on the silhouette. (c) and (f) show the effect with GaitCLIF.

GaitCLIF, (f) shows a shift toward dynamic joint regions, learning more discriminative gait features.

11.3. Visualization of Diverse Clothing

In this section, we provide additional visualizations of the diverse clothing used in BarbieGait, as shown in Figure 6. The wardrobe includes a wide range of apparel and appearance attributes, covering various hairstyles, tops, pants, skirts, shoes, and carried objects. Each category contains roughly 100 individual items, enabling more than 200,000 theoretically valid clothing combinations.

During outfit generation, we follow common real-world dressing conventions and further apply manual filtering to remove implausible combinations as well as cases exhibiting mesh–cloth penetration (i.e., garments intersecting with the body surface). These measures ensure that the clothing variations in BarbieGait are both realistic and suitable for gait recognition studies under cloth-changing conditions.

References

[1] https://www.rokoko.com/insights/ace-retargeting-in-blender-with-this-simple-workflow-i-the-ultimate-retargeting-guide. 1

[2] https://github.com/Mwni/blender-animation-retargeting. 1

[3] Mykhaylo Andriluka, Umar Iqbal, Eldar Insafutdinov, Leonid Pishchulin, Anton Milan, Juergen Gall, and Bernt Schiele. Posetrack: A benchmark for human pose estimation and tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5167–5176, 2018. 2

[4] MMPose Contributors. Openmmlab pose estimation toolbox and benchmark. https://github.com/open-mmlab/mmpose, 2020. 2

[5] Daniel Davila, Dawei Du, Bryon Lewis, Christopher Funk, Joseph Van Pelt, Roderic Collins, Kellie Corona, Matt Brown, Scott McCloskey, Anthony Hoogs, et al. Mevid: Multi-view extended videos with identities for video person re-identification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1634–1643, 2023. 1

[6] Yilan Dong, Chunlin Yu, Ruiyang Ha, Ye Shi, Yuexin Ma, Lan Xu, Yanwei Fu, and Jingya Wang. Hybridgait: A benchmark for spatial-temporal cloth-changing gait recognition with hybrid explorations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1600–1608, 2024. 1

[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*, pages 9707–9716, 2023. 1

[8] 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, 3

[9] 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. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 1

[10] Xinqian Gu, Hong Chang, Bingpeng Ma, Shutao Bai, Shiguang Shan, and Xilin Chen. Clothes-changing person re-identification with rgb modality only. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1060–1069, 2022. 1

[11] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 2

[12] Junjie Huang, Zheng Zhu, Feng Guo, and Guan Huang. The devil is in the details: Delving into unbiased data processing for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5700–5709, 2020. 2

[13] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013. 2

[14] Tao Jiang, Peng Lu, Li Zhang, Ningsheng Ma, Rui Han, Chengqi Lyu, Yining Li, and Kai Chen. Rtmpose: Real-time multi-person pose estimation based on mmpose. *arXiv preprint arXiv:2303.07399*, 2023. 2

[15] Angjoo Kanazawa, Jason Y Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5614–5623, 2019. 2

[16] Wenpeng Lang, Saihui Hou, and Yongzhen Huang. Beyond sparse keypoints: Dense pose modeling for robust gait recognition. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 669–678, 2025. 2, 3

[17] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 2

[18] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5442–5451, 2019. 2

[19] Y. Makihara, H. Mannami, A. Tsuji, M.A. Hossain, K. Sugiura, A. Mori, and Y. Yagi. The ou-isir gait database comprising the treadmill dataset. *IPSJ Trans. on Computer Vision and Applications*, 4:53–62, 2012. 1

[20] Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. *arXiv preprint arXiv:1904.09237*, 2019. 2

[21] 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. 3

[22] Timo Von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European conference on computer vision (ECCV)*, pages 601–617, 2018. 2

[23] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364, 2020. 2, 3

[24] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in Neural Information Processing Systems*, 35:38571–38584, 2022. 2, 3

[25] 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. 3