MMGait: Towards Multi-Modal Gait Recognition

Supplementary Material

6. Related Work

6.1. Gait Recognition Methods

Gait recognition methods are broadly classified into two categories: silhouette-based methods [7, 18, 19, 21, 22, 35, 50] and pose-based methods [13, 14, 23]. Silhouette-based methods rely on binary gait silhouettes, which are widely used due to their effectiveness in capturing spatio-temporal information. For example, GaitSet [3] models gait as unordered frame sets using set pooling. More recent works like GaitBase [10] and DeepGaitV2 [9] revisit architectural design, introducing efficient backbone networks tailored for gait recognition. Additionally, some studies [31, 32] emphasize modeling specific motion patterns to enhance the temporal discriminability of gait features. Pose-based methods, on the other hand, utilize structural representations of the human body based on joint coordinates. PoseGait [29] leverages 3D human poses as input to CNN-based architectures to extract discriminative features. GaitGraph [41], GaitGraph2 [42], and GPGait [12] adopt graph convolutional networks (GCNs) to explicitly model spatial dependencies between joints. Furthermore, methods like SkeletonGait [11] and GaitHeat [13] represent joint positions through heatmaps, preserving fine-grained spatial details and enabling a more comprehensive encoding of body shape and motion. In addition, several emerging modalities have recently been explored for gait recognition. BigGait [45] and BiggerGait [46] leverage large vision models to extract powerful visual representations from RGB inputs. EdinoGait [4] exploits the capability of large vision models to address event-based gait recognition, demonstrating clear advantages under low-light conditions. LidarGait++ [38] further introduces effective local representation techniques for point-cloud-based gait recognition.

More recently, multi-modal gait recognition has gained increasing attention by integrating complementary modalities [44]. MMGaitFormer [5] fuses silhouettes and 2D poses through spatial–temporal modules to enhance cross-source alignment, while MultiGait++ [24] provides a simple yet strong baseline and systematically compares diverse fusion strategies. TriGait [39] introduces a tri-branch hybrid fusion framework to jointly exploit the complementary cues of silhouettes and poses. LiCAF [6] further proposes an effective LiDAR-camera fusion scheme to obtain robust cross-sensor gait representations.

6.2. Gait Recognition Benchmark

Existing gait recognition benchmarks predominantly rely on silhouette or pose modalities, exemplified by indoor datasets such as CASIA-B [47], OU-MVLP [40], CCPG [28], and CCGR [52], as well as in-the-wild datasets like Gait3D [49] and GREW [51]. While effective in controlled settings, these modality choices limit scalability to diverse sensing environments and raise potential privacy concerns. Recent advances in hardware have broadened the sensing landscape, with LiDAR emerging as a promising modality due to its resilience to lighting variations, background clutter, and occlusions. Representative efforts such as LidarGait [37], which introduces the SUSTech1K dataset, and FreeGait [16], which extends LiDAR-based recognition to unconstrained outdoor conditions, highlight the potential of structural 3D cues for robust gait analysis. However, these efforts still explore only a small portion of the sensing modalities used in real-world systems, leading to fragmented progress across isolated modal combinations and lacking a unified basis for comparison. A truly comprehensive multi-modal benchmark remains absent, underscoring the need for a unified and diverse dataset to advance gait recognition in realistic settings.

6.3. Unified Models for Identity Recognition

Recent efforts in person re-identification increasingly focus on building unified models capable of handling diverse modalities and tasks. Instruct-ReID++ [17] takes a significant step toward universal identity retrieval by employing natural-language instructions to guide a single model across varied scenarios, enabling task-adaptive behavior through instruction tuning and adaptive losses. Complementary to this task-unified perspective, the All-in-One (AIO) framework [27] adopts a frozen pre-trained backbone with modality-specific designs to extract consistent identity features from RGB, infrared, sketch, and text inputs. ReID5o [53] further expands modality unification by introducing a five-modality benchmark and a unified encoder with multi-expert routing, achieving flexible cross-modal retrieval across arbitrary modality pairs. Our work aligns with this direction but focuses specifically on unifying diverse sensing modalities within a single framework.

7. More Details in MMGait

7.1. Data Process Pipeline

We constructed separate processing pipelines for each sensor, ultimately producing 12 distinct modalities for evaluation. To the best of our knowledge, MMGait is the first large-scale gait recognition dataset that covers such a comprehensive range of modalities. The details of the processing pipelines are summarized below:

RGB Camera: We use a pedestrian tracking network [48] to extract bounding boxes. Instance segmentation is applied to obtain silhouettes, followed by pose estimation to extract both 2D and 3D pose representations [1, 2, 30, 43]. Additionally, we employ the V2E [20] algorithm to convert RGB videos into event-based representations. Cropped event sequences are then generated by scaling the RGB-based bounding boxes according to the resolution ratio between the event and RGB videos.

Depth Camera: As pedestrian tracking is challenging directly on depth data, and the RGB and depth cameras are integrated in a shared device, we computed the spatial offset between the RGB and depth cameras to adjust the bounding box coordinates obtained from the RGB videos, and used the transformed boxes to crop the corresponding depth-based gait sequences.

IR Camera: Pedestrian tracking and instance segmentation are performed independently on infrared videos. Both the cropped IR images and the corresponding silhouettes demonstrate high visual quality.

LiDAR Scanner: The raw LiDAR scans are processed into clean human walking point clouds and their corresponding projection maps. Specifically, we first retain points within a predefined Region of Interest (ROI) that captures the walking subject. Ground removal [26] is then performed to remove floor points, followed by denoising algorithms [8] to filter out scattered noise and irrelevant objects. The projection of point clouds into 2D depth maps is implemented based on the OpenGait codebase [10, 37].

4D Radar System: As the radar sensor captures mainly moving targets and there are no additional dynamic objects indoors, the point clouds are restricted to those falling within the specified ROI. The projection of point clouds into 2D depth maps is also implemented based on the OpenGait codebase [10, 37].

7.2. Visualization

We present supplementary visualization examples, as shown in Figure 7 and Figure 8, encompassing all sensor modalities, ten viewpoints, and the three walking conditions: normal walking (NM), walking with a backpack (BG), and walking with changed clothing (CL). These samples illustrate the diversity and high quality of the collected gait sequences across different sensing configurations and walking scenarios.

Privacy Statement. All data collection procedures followed ethical guidelines, and all participants provided written informed consent for research purposes. The dataset is released strictly for academic research, and any form of misuse or unauthorized application is explicitly prohibited.

8. Experimental Setup

To comprehensively evaluate the effectiveness of **MMGait**, we design three categories of experiments: *Single-Modal Recognition*, *Cross-Modal Recognition*, and *Multi-Modal Recognition*. These experiments aim to systematically assess the recognition capability of each modality, cross-modal retrieval performance, and the complementary relationships among different modalities under a unified and fair evaluation protocol.

All input visual modalities are standardized to a spatial resolution of 64×64. Unless otherwise stated, all configurations follow the default settings of OpenGait [10] and FastPoseGait [33]. For visual and pose modalities, we adopt a batch configuration of $(p, k, l) = (8, 8, 30)$, where p denotes the number of identities, k the number of sequences per identity, and l the number of frames. For LiDAR and Radar point cloud modalities, we use a batch size of $(8, 8, 10)$ and uniformly sample 512 points per frame. All experiments are conducted on a workstation equipped with eight NVIDIA GeForce RTX 3090 GPUs.

8.1. Single-Modal Recognition

For baseline evaluations on silhouette and pose modalities, we adopt the original default configurations of each method. For extended analysis, we apply different modeling strategies. For image-based modalities, we employ GaitBase [10] trained for 60,000 iterations using SGD [36]. For pose input, we use GPGait++ [34] trained for 40,000 iterations with the Adam optimizer [25]. For point cloud modalities, we adopt LidarGait++ [38], trained for 40,000 iterations using SGD [36].

8.2. Cross-Modal Retrieval

Inspired by CL-Gait [15], we adopt a two-stream architecture for cross-modal gait recognition. We use the GaitBase [10] framework, where Stage 1 contains modality-specific parameters, while Stages 2-4 share parameters across modalities. Both branches output features of dimension 16×256.

Loss Function. We train the network using a combination of symmetric cross-modality triplet loss and cross-entropy loss, equally weighted. To enhance cross-modal feature alignment, we modify the triplet formulation by selecting the anchor from one modality and the positive/negative samples from the other modality:

$$L_{\text{cross-triplet}} = \frac{1}{2}(L_{\text{triplet}}(A_{\text{modal1}}, P_{\text{modal2}}, N_{\text{modal2}}) + L_{\text{triplet}}(A_{\text{modal2}}, P_{\text{modal1}}, N_{\text{modal1}})). \quad (6)$$

We further apply independent cross-entropy losses to

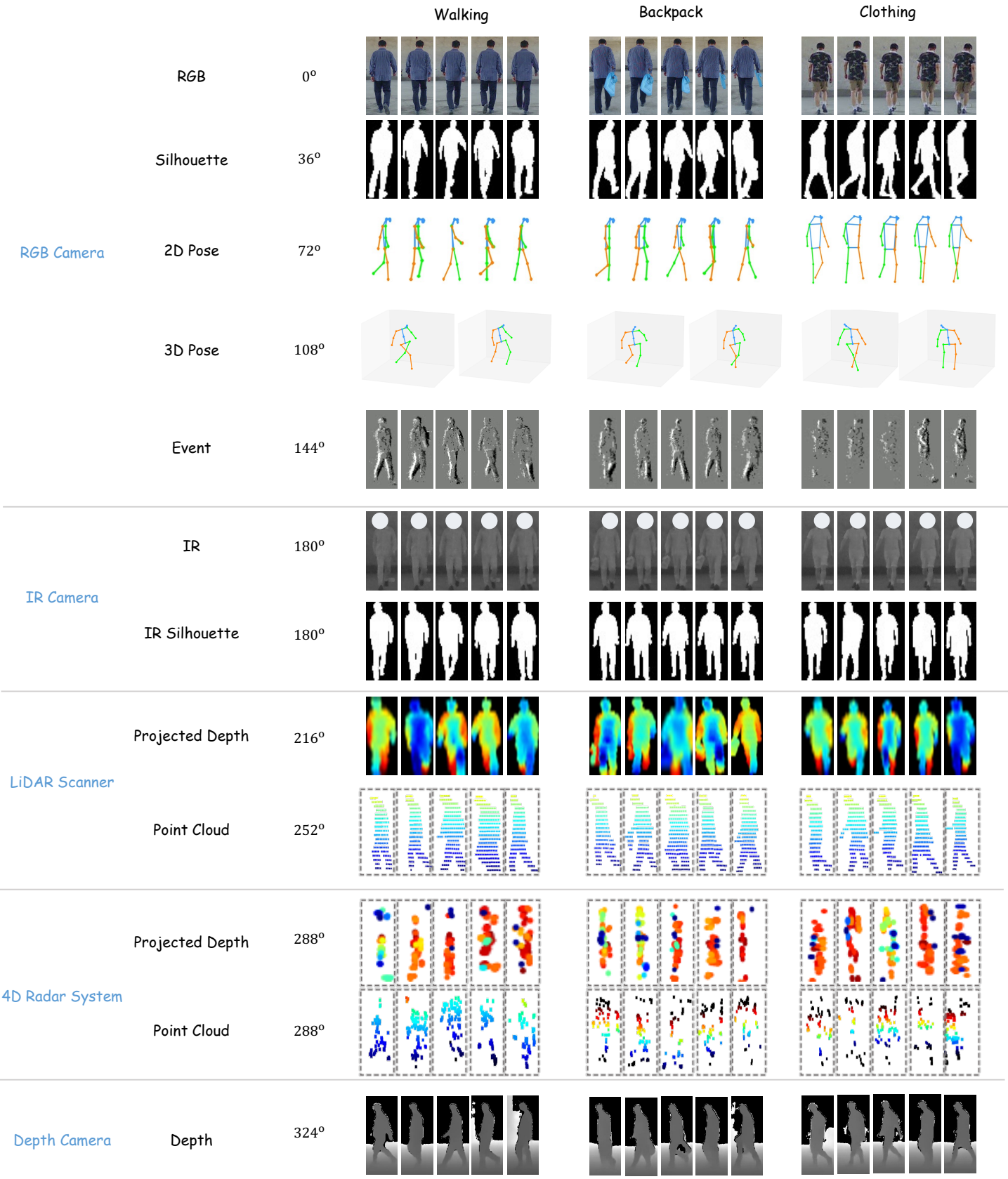

Figure 7. Visualization of gait sequences across all sensor modalities, ten viewpoints, and three walking conditions (NM, BG, CL).

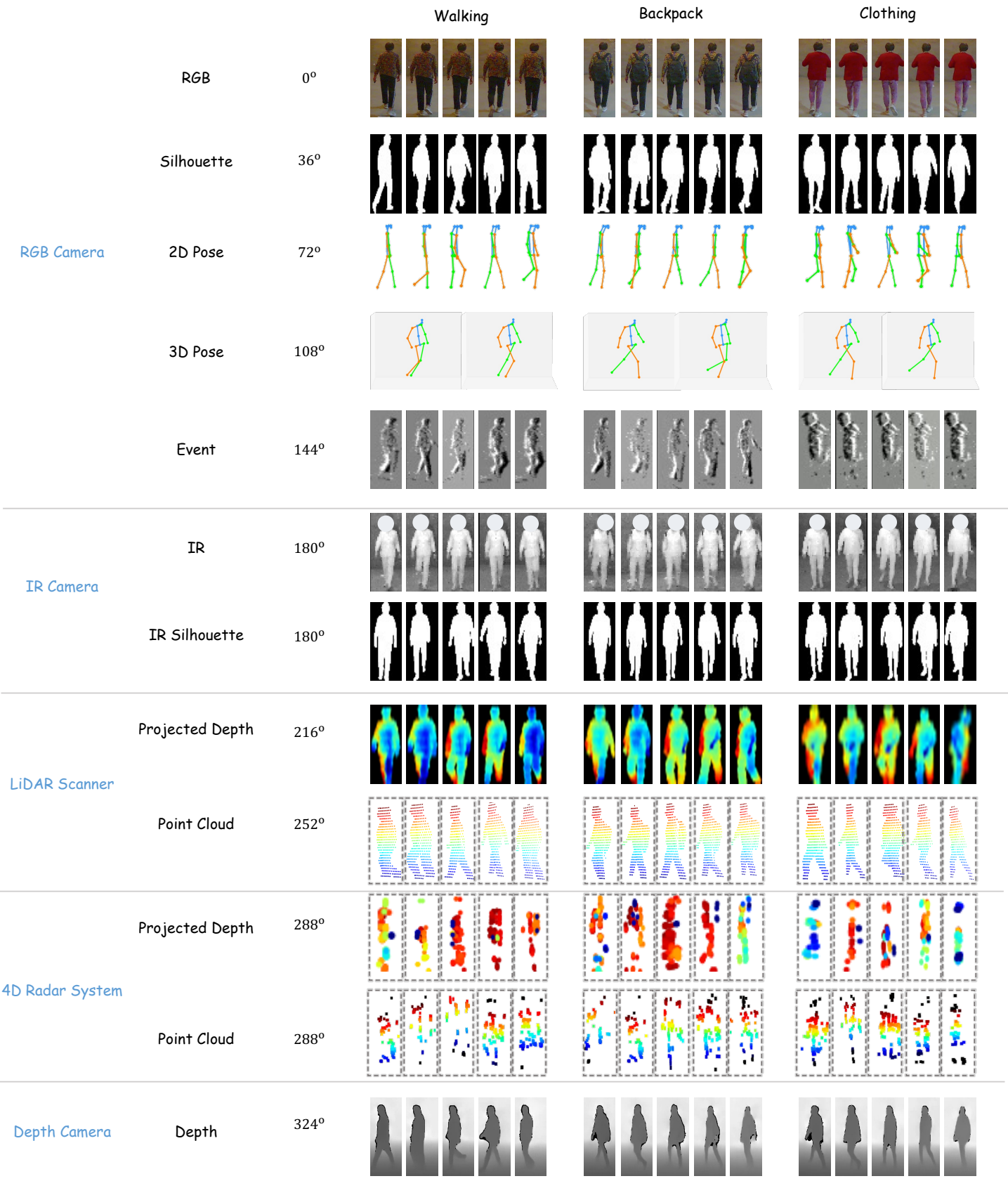

Figure 8. Visualization of gait sequences across all sensor modalities, ten viewpoints, and three walking conditions (NM, BG, CL).

Task Setting	Modality	Parameters (M)	FLOPs (G)
Single-Modal (GaitBase)	RGB (Sil.)	7.821504	51.67659827
	RGB (RGB)	7.822656	51.81815603
	RGB (2D Pose)	7.822080	51.74737715
	RGB (Event)	7.822656	51.81815603
	IR (Sil.)	7.821504	51.67659827
	IR (IR)	7.822656	51.81815603
	Depth	7.822656	51.81815603
	LiDAR (Projected Depth)	7.822656	51.81815603
	4D Radar (Projected Depth)	7.822656	51.81815603
Cross-Modal (Two-Stream)	RGB (Sil.) ↔ IR (Sil.)	10.820736	103.3531965
	RGB (Sil.) ↔ Depth	10.821888	103.4947543
	RGB (Sil.) ↔ LiDAR (Projected Depth)	10.821888	103.4947543
	RGB (Sil.) ↔ 4D Radar (Projected Depth)	10.821888	103.4947543
	IR (Sil.) ↔ Depth	10.821888	103.4947543
	IR (Sil.) ↔ LiDAR (Projected Depth)	10.821888	103.4947543
	IR (Sil.) ↔ 4D Radar (Projected Depth)	10.821888	103.4947543
	Depth ↔ LiDAR (Projected Depth)	10.823040	103.6363121
	Depth ↔ 4D Radar (Projected Depth)	10.823040	103.6363121
	LiDAR (Projected Depth) ↔ 4D Radar (Projected Depth)	10.823040	103.6363121
Multi-Modal (MultiGait++)	RGB (Sil.) + RGB (Event)	11.920448	106.2864364
	RGB (Sil.) + RGB (Pose)	11.919872	106.2156575
	RGB (Sil.) + Depth	11.920448	106.2864364
	RGB (Sil.) + LiDAR (Projected Depth)	11.920448	106.2864364
Omni Multi-Modal (OmniGait)	RGB (Sil., RGB, 2D Pose, Event), IR (Sil., IR), Depth, LiDAR (Projected Depth), 4D Radar (Projected Depth)	9.963266	1054.108566

Table 9. Parameters and FLOPs for different models evaluated in our experiments.

identity predictions from each modality:

$$L_{\text{ce}}^{\text{modal}} = -\sum_{i=1}^{c} y_i \log(\hat{y}_i), \tag{7}$$

$$L_{\text{ce}} = \frac{1}{2}\left(L_{\text{ce}}^{\text{modal1}} + L_{\text{ce}}^{\text{modal2}}\right), \tag{8}$$

where c denotes the number of identity classes, y_i is the ground-truth one-hot label, and $\hat{y}_i$ is the predicted probability for class i.

The final objective is computed as:

$$L_{\text{total}} = \frac{1}{2}\left(L_{\text{cross-triplet}} + L_{\text{ce}}\right). \tag{9}$$

8.3. Multi-Modal Recognition

We follow the two-stream fusion strategy proposed in MultiGait++ [24] and investigate several representative visual modality combinations. Besides, for the fusion of LiDAR point clouds and Radar point clouds, we adopt a dual-stream LidarGait++ architecture without parameter sharing. Each modality is modeled independently, and the resulting features are concatenated before being fed into a shared fully connected layer and BNNeck. The final fused feature has a dimension of 31×256. All of the models are trained for 60,000 iterations using SGD.

8.4. Paramers Comparison

Table 9 presents the parameter counts and FLOPs for all modality configurations evaluated in our study. The Omni Multi-Modal entry corresponds to our proposed OmniGait model. Despite supporting every task across the Single-Modal, Cross-Modal, and Multi-Modal settings with a single unified architecture, OmniGait remains remarkably lightweight, requiring only 9.96M parameters. This highlights the efficiency and scalability of our design, enabling broad modality coverage without incurring significant computational overhead.

8.5. Cross-Dataset Evaluation on SUSTech1K

To further evaluate the generalization capability of OmniGait, we conduct a cross-dataset evaluation by directly transferring the model trained on MMGait to SUSTech1K without any fine-tuning. This setting is particularly challenging due to significant domain gaps, including differences in sensor configuration, data distribution, environmental conditions, and identity diversity. All results are obtained under a strict zero-shot transfer protocol.

The detailed cross-dataset results on SUSTech1K are reported in Table 10. Under the strict zero-shot transfer

Input Modality	Probe Sequence								Overall	
	Normal	Bag	Clothing	Carrying	Umbrella	Uniform	Occlusion	Night	Rank1	Rank5
Lidar Depth	13.67	10.81	4.88	7.29	1.14	7.06	10.88	9.55	7.77	17.38
RGB	43.15	30.07	20.52	32.45	27.93	24.33	32.31	44.66	32.03	53.21
Silhouette	47.24	44.68	26.28	41.78	39.42	42.04	42.98	18.08	42.65	62.15
RGB+Silhouette	63.93	54.39	35.98	52.71	50.98	48.86	46.97	32.86	53.07	70.96

Table 10. Cross-dataset evaluation on SUSTech1K. OmniGait is trained on MMGait and directly evaluated without fine-tuning.

setting (trained on MMGait and evaluated without fine-tuning), OmniGait achieves 7.77% Rank-1 accuracy with LiDAR depth input. Using RGB input improves the performance to 32.03%, while silhouette input further boosts it to 42.65%, indicating that shape-based representations exhibit stronger cross-domain robustness than appearance-only cues.

When fusing RGB and silhouette modalities, the performance is substantially improved to 53.07% Rank-1 and 70.96% Rank-5. The gain from multi-modal fusion is consistent across all probe conditions. These results demonstrate that OmniGait is capable of learning transferable representations, and multi-modal integration effectively enhances robustness under distribution shift.

9. Discussion

The experimental results on MMGait reveal several important observations that offer insights and future directions for multimodal gait recognition research:

(1) Cross-modal retrieval remains highly challenging. Although certain similar modalities achieve relatively strong cross-modal retrieval performance, most modality pairs still exhibit substantial difficulty, particularly under cross-clothing conditions. A key challenge ahead is how to enable models to simultaneously improve cross-modal alignment and cross-covariate robustness, which remain two conflicting objectives.

(2) Multi-modal fusion provides substantial benefits. Our results show that the complementary information across modalities is crucial for identity recognition. For example, combining RGB silhouettes with LiDAR projected depth leads to a 19.7% improvement in challenging cross-clothing scenarios. This improvement stems from the complementary strengths of the two modalities: LiDAR provides stable geometric depth cues, while RGB supplies richer and more discriminative shape information. Their combination significantly enhances robustness under varying appearance conditions.

(3) Omni Multi-Modal Recognition presents both opportunities and challenges. The Omni Multi-Modal Gait Recognition task is inherently challenging, as it requires a unified framework to handle heterogeneous sensing modalities (e.g., RGB, IR, Depth, LiDAR) while simultaneously supporting diverse retrieval paradigms, including single-modal recognition, multi-modal fusion, and cross-modal retrieval. The large domain gaps across modalities, discrepancies in data distributions, and modality-specific noise patterns make it difficult to learn a shared representation that is both discriminative and modality-invariant. In practice, a unified model often sacrifices single-modal optimality compared to modality-specific counterparts, and its cross-covariate robustness remains constrained under real-world variations. To establish a feasible starting point for this challenging setting, we introduce OmniGait as a baseline framework. In the current implementation, 3D point cloud data are projected into depth maps before being fed into the network, which helps reduce domain discrepancies between geometric and image-based modalities and enables shared backbone processing. While this design simplifies cross-modal alignment, it inevitably discards part of the intrinsic geometric structure. Directly modeling raw 3D point clouds within a unified omni-modal architecture could therefore be a promising direction for future research, potentially allowing richer geometric cues to be preserved.

Limitations: MMGait does not enforce strict temporal synchronization across modalities, as heterogeneous sensors in real-world deployments naturally differ in sampling rates, exposure cycles, and hardware triggering pipelines. Nevertheless, to approximate synchronization across modalities, we made the following efforts: (1) Frame-Level Synchronization: RGB and Depth are inherently aligned, as they are captured from the same device, ensuring frame-level synchronization for these two modalities. (2) Sequence-Level Synchronization: For other modalities, we perform sequence-level synchronization by recording the start and end timestamps of each device's recording session, enabling approximate temporal alignment across modalities during preprocessing.

References

[1] Qingyuan Cai, Xuecai Hu, Saihui Hou, Li Yao, and Yongzhen Huang. Disentangled diffusion-based 3d human pose estimation with hierarchical spatial and temporal denoiser. In *Proceedings of the AAAI conference on artificial intelligence*, pages 882–890, 2024. 2

[2] Qingyuan Cai, Linxin Zhang, Xuecai Hu, Saihui Hou, and

Yongzhen Huang. Fastddhpose: Towards unified, efficient, and disentangled 3d human pose estimation. *arXiv preprint arXiv:2512.14162*, 2025. 2

[3] Hanqing Chao, Yiwei He, Junping Zhang, and Jianfeng Feng. Gaitset: Regarding gait as a set for cross-view gait recognition. In *Proceedings of the AAAI conference on artificial intelligence*, pages 8126–8133, 2019. 1

[4] Liaogehao Chen, Zhenjun Zhang, and Yaonan Wang. Edino-gait: Transferring large visual models to event-based vision for enhancing gait recognition. *IEEE Transactions on Multimedia*, 2025. 1

[5] Yufeng Cui and Yimei Kang. Multi-modal gait recognition via effective spatial-temporal feature fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17949–17957, 2023. 1

[6] Yunze Deng, Haijun Xiong, and Bin Feng. Licaf: Lidar-camera asymmetric fusion for gait recognition. In *2024 IEEE International Conference on Image Processing (ICIP)*, pages 2424–2430. IEEE, 2024. 1

[7] Huanzhang Dou, Pengyi Zhang, Yuhan Zhao, Lu Jin, and Xi Li. Clash: Complementary learning with neural architecture search for gait recognition. *IEEE Transactions on Image Processing*, 2024. 1

[8] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, pages 226–231, 1996. 2

[9] Chao Fan, Saihui Hou, Yongzhen Huang, and Shiqi Yu. Exploring deep models for practical gait recognition. *arXiv preprint arXiv:2303.03301*, 2023. 1

[10] 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, 2

[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

[12] Yang Fu, Shibei Meng, Saihui Hou, Xuecai Hu, and Yongzhen Huang. Gpgait: Generalized pose-based gait recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19595–19604, 2023. 1

[13] Yang Fu, Saihui Hou, Shibei Meng, Xuecai Hu, Chunshui Cao, Xu Liu, and Yongzhen Huang. Cut out the middleman: Revisiting pose-based gait recognition. In *European Conference on Computer Vision*, pages 112–128. Springer, 2024. 1

[14] Hongji Guo and Qiang Ji. Physics-augmented autoencoder for 3d skeleton-based gait recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19627–19638, 2023. 1

[15] Wenxuan Guo, Yingping Liang, Zhiyu Pan, Ziheng Xi, Jianjiang Feng, and Jie Zhou. Camera-lidar cross-modality gait recognition. In *European Conference on Computer Vision*, pages 439–455. Springer, 2024. 2

[16] Xiao Han, Yiming Ren, Peishan Cong, Yujing Sun, Jingya Wang, Lan Xu, and Yuexin Ma. Gait recognition in large-scale free environment via single lidar. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 380–389, 2024. 1

[17] Weizhen He, Yiheng Deng, Yunfeng Yan, Feng Zhu, Yizhou Wang, Lei Bai, Qingsong Xie, Rui Zhao, Donglian Qi, Wanli Ouyang, et al. Instruct-reid++: Towards universal purpose instruction-guided person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 1

[18] Saihui Hou, Chenye Wang, Wenpeng Lang, Zhengxiang Lan, and Yongzhen Huang. Gaitsnippet: Gait recognition beyond unordered sets and ordered sequences. *arXiv preprint arXiv:2508.07782*, 2025. 1

[19] Saihui Hou, Chenye Wang, Aoqi Li, Jilong Wang, Liang Wang, and Yongzhen Huang. Gaitasset: In defense of regarding gait as a set. *IEEE Transactions on Information Forensics and Security*, 20:12301–12316, 2025. 1

[20] Yuhuang Hu, Shih-Chii Liu, and Tobi Delbruck. v2e: From video frames to realistic dvs events. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1312–1321, 2021. 2

[21] Panjian Huang, Saihui Hou, Chunshui Cao, Xu Liu, and Yongzhen Huang. Vocabulary-guided gait recognition. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*. 1

[22] Panjian Huang, Saihui Hou, Junzhou Huang, and Yongzhen Huang. Learning a unified template for gait recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12459–12469, 2025. 1

[23] Xiaohu Huang, Xinggang Wang, Zhidianqiu Jin, Bo Yang, Botao He, Bin Feng, and Wenyu Liu. Condition-adaptive graph convolution learning for skeleton-based gait recognition. *IEEE Transactions on Image Processing*, 32:4773–4784, 2023. 1

[24] Dongyang Jin, Chao Fan, Weihua Chen, and Shiqi Yu. Exploring more from multiple gait modalities for human identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4120–4128, 2025. 1, 5

[25] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 2

[26] Seungjae Lee, Hyungtae Lim, and Hyun Myung. Patchwork++: Fast and robust ground segmentation solving partial under-segmentation using 3d point cloud. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 13276–13283. IEEE, 2022. 2

[27] He Li, Mang Ye, Ming Zhang, and Bo Du. All in one framework for multimodal re-identification in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17459–17469, 2024. 1

[28] 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. 1

[29] Rijun Liao, Shiqi Yu, Weizhi An, and Yongzhen Huang. A model-based gait recognition method with body pose and

human prior knowledge. *Pattern Recognition*, 98:107069, 2020. 1

[30] Yi Liu, Lutao Chu, Guowei Chen, Zewu Wu, Zeyu Chen, Baohua Lai, and Yuying Hao. Paddleseg: A high-efficient development toolkit for image segmentation. *arXiv preprint arXiv:2101.06175*, 2021. 2

[31] Kang Ma, Ying Fu, Dezhi Zheng, Chunshui Cao, Xuecai Hu, and Yongzhen Huang. Dynamic aggregated network for gait recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22076–22085, 2023. 1

[32] Kang Ma, Ying Fu, Chunshui Cao, Saihui Hou, Yongzhen Huang, and Dezhi Zheng. Learning visual prompt for gait recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 593–603, 2024. 1

[33] Shibei Meng, Yang Fu, Saihui Hou, Chunshui Cao, Xu Liu, and Yongzhen Huang. Fastposegait: A toolbox and benchmark for efficient pose-based gait recognition. *arXiv preprint arXiv:2309.00794*, 2023. 2

[34] Shibei Meng, Yang Fu, Saihui Hou, Xuecai Hu, Chunshui Cao, Xu Liu, and Yongzhen Huang. From fastposegait to gpgait++: Bridging the past and future for pose-based gait recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 2

[35] Guozhen Peng, Yunhong Wang, Yuwei Zhao, Shaoxiong Zhang, and Annan Li. Glgait: a global-local temporal receptive field network for gait recognition in the wild. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 826–835, 2024. 1

[36] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951. 2

[37] 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. 1, 2

[38] Chuanfu Shen, Rui Wang, Lixin Duan, and Shiqi Yu. Lidargait++: Learning local features and size awareness from lidar point clouds for 3d gait recognition. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6627–6636, 2025. 1, 2

[39] Yan Sun, Xueling Feng, Xiaolei Liu, Liyan Ma, Long Hu, and Mark S Nixon. Trigait: hybrid fusion strategy for multimodal alignment and integration in gait recognition. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 7(1):82–94, 2024. 1

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

[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

[42] 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. 1

[43] 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

[44] Likai Wang, Ruize Han, and Wei Feng. Combining the silhouette and skeleton data for gait recognition. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 1

[45] Dingqiang Ye, Chao Fan, Jingzhe Ma, Xiaoming Liu, and Shiqi Yu. Biggait: Learning gait representation you want by large vision models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 200–210, 2024. 1

[46] Dingqiang Ye, Chao Fan, Zhanbo Huang, Chengwen Luo, Jianqiang Li, Shiqi Yu, and Xiaoming Liu. Biggergait: Unlocking gait recognition with layer-wise representations from large vision models. *arXiv preprint arXiv:2505.18132*, 2025. 1

[47] 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. 1

[48] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. In *European Conference on Computer Vision*, pages 1–21. Springer, 2022. 2

[49] Jinkai Zheng, Xinchen Liu, Wu Liu, Lingxiao He, Chenggang Yan, and Tao Mei. Gait recognition in the wild with dense 3d representations and a benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20228–20237, 2022. 1

[50] 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. 1

[51] Zheng Zhu, Xianda Guo, Tian Yang, Junjie Huang, Jiankang Deng, Guan Huang, Dalong Du, Jiwen Lu, and Jie Zhou. Gait recognition in the wild: A benchmark. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 14789–14799, 2021. 1

[52] 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

[53] Jialong Zuo, Yongtai Deng, Mengdan Tan, Rui Jin, Dongyue Wu, Nong Sang, Liang Pan, and Changxin Gao. Reid5o:

Achieving omni multi-modal person re-identification in a single model. *arXiv preprint arXiv:2506.09385*, 2025. 1