Supplementary Material for VM-Gait: Multi-Modal 3D Representation Based on Virtual Marker for Gait Recognition

Zhao-Yang Wang[1] Jiang Liu[2] Jieneng Chen[1†] Rama Chellappa[1†]
[1]Johns Hopkins University [2]Advanced Micro Devices, Inc.

Abstract

In this supplementary material, we provide experiment results on the BRIAR datasets and a comprehensive overview of the configuration parameters, implementation details, network details of our proposed VM-Gait. Additionally, we analyze the influence of temporal stride and visualize examples of silhouettes, 3D mesh, and virtual markers with occlusion and noise from the Gait3D dataset, thereby enhancing the contextual understanding of our model's performance. Lastly, we discuss potential future works that include developing silhouette encoder model, multi-modal fusion model and a large-scale multi-modal gait dataset.

1. Experiment Results on the BRIAR Datasets

To demonstrate the robustness of VM-Gait in real-world scenarios involving long-range and high-altitude conditions, we conducted additional experiments using the BRIAR dataset. The results are shown in Table 1, VM-Gait significantly outperforms other methods. The system's ability to provide valuable complementary information is particularly evident when silhouette masks become less accurate at greater distances, underscoring VM-Gait's effectiveness in real-world environments.

Dataset: BRAIR [1]	Metrics (%)		
Methods	R-1	R-5	R-20
SMPLGait w/o 3D	32.4	65.1	90.1
GaitPart	33.8	63.6	85.5
GaitSet	34.9	68.8	88.4
GaitBase	36.9	68.2	90.3
DeepGaitV2	50.6	77.8	98.3
VM-Gait	**52.0**	**78.1**	**98.6**

Table 1. Gait Recognition Results on the BRAIR [1] dataset. We conduct experiments with input size: (64 × 44). The VM-Gait method stands out among state-of-the-art methods. VM-Gait is able to provide valuable complementary information when silhouette masks become less accurate at greater distances,

2. Analysis of Fusion Methods

To investigate the impact of different fusion methods, we performed an ablation study comparing three commonly used fusion techniques for feature integration. The results, presented in Table 2, reveal that concatenating silhouette features with virtual marker features is both a straightforward and effective approach.

Fusion	R-1	R-5	mAP	mINP
Concatenate	**75.4**	**87.5**	**66.4**	**39.5**
Add	74.4	86.9	65.5	39.1
Attention	70.5	84.5	61.1	35.0

Table 2. The ablation study for the fusion on the Gait3D dataset.

3. Implement Details of the VM-Gait Framework

In this section, we present the implementation details of the VM-Gait framework applied to the Gait3D, OUMVLP-Mesh and BRIAR datasets. The configuration parameters for Gait3D are presented in Table 5, whereas the configuration parameters for OUMVLP-Mesh are detailed in Table 6. The configuration parameters for BRIAR dataset are detailed in Table 7. It's worth noting that our framework implementation is built upon the OpenGait [2] codebase to ensure flexibility and adaptability.

4. Additional Network Details of the VM-Gait Framework

In this section, we present additional details and hyperparameters utilized within the VM-Gait Framework. Specifically, we elaborate on the hyperparameter details of the PST-Transformer for virtual markers in the 3D branch in Table 4. We also provide the hyperparameters of the PST-Transformer for mesh vertices inputs in the 3D branch. The results of these comparisons are discussed in the main paper's ablation study section.

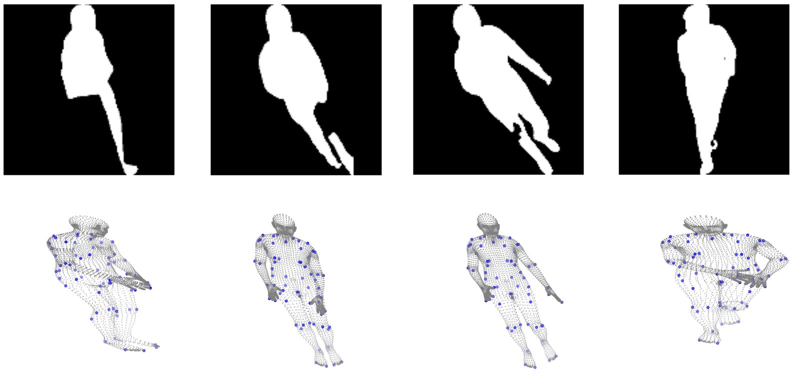

Figure 1. Example of a partial silhouette sequence, exhibiting occlusion and noise, along with the corresponding 3D mesh and virtual markers.

Temporal Stride	R-1	R-5	mAP	mINP
3	**75.4**	**87.5**	**66.4**	**39.5**
4	74.0	87.0	65.9	39.3
5	75.2	86.6	65.7	38.7

Table 3. The ablation study evaluates the influence of temporal stride on the Gait3D dataset.

5. Visualization of the Gait Recognition

We present gait representation examples with occlusion and noise extracted from Gait3D. These examples showcase the silhouettes along with their corresponding 3D mesh vertices and virtual markers, as depicted in Figure 1. Notably, when image sequences suffer from occlusion and noise, the segmented silhouettes tend to possess incomplete information, potentially impacting the performance of gait recognition. Despite the inherent challenges in the ill-posed problem of 3D representation reconstruction from videos, 3D representation reconstruction algorithms consider spatial-temporal features, enabling the generated 3D mesh and virtual markers to offer plausible shapes from videos. These plausible shapes, in turn, may furnish complementary information to enhance the performance of gait recognition.

6. Analysis of the influence of temporal stride

The temporal stride refers to the interval at which the transformer processes temporal information. This parameter determines how many time frames are considered at each step of the PST transformer operation. The findings are presented in Table 3. A larger temporal stride results in more loss of temporal information. Balancing the temporal stride with available GPU memory is crucial.

Hyper-parameters	VM	Meshes	skeletons
spatial radius	0.05	0.05	0.05
n samples	4	430	1
spatial stride	4	430	1
temporal kernel size	3	3	3
temporal stride	3	3	3
dim	8	8	8
dim_head	8	8	8

Table 4. Hyper-parameters of PST-Transformer for virtual markers, mesh vertices and skeletons inputs in the 3D branch.

7. FutureWork and Discussion

While we introduced a multi-modal framework featuring novel gait representations, there are still aspects that can be improved. In this section, we highlight some subsequent works that are worth further exploration including developing silhouette encoder model, multi-modal fusion model and a large-scale multi-modal gait dataset.

7.1. Developing Silhouette Encoder Model

Developing the silhouette encoder model can substantially improve silhouette and multi-modal gait recognition performance. Currently, state-of-the-art silhouette-based approaches employ the CNN architecture. Investigating the use of transformer architecture has the potential to further improve recognition performance. The self-attention mechanism in transformers allows the model to focus on different parts of the input sequence, enabling it to discern subtle patterns and variations in the gait cycle.

7.2. Developing Multi-Modal Fusion Model

When different features are extracted from different gait representations, there is a compelling need for research into the development of effective fusion techniques to fuse these features. This integration can be accomplished through a range of methods, including concatenation, weighted averaging, or more advanced approaches based on attention mechanisms.

7.3. Developing Large-Scale Multi-Modal Gait Datasets

The development of comprehensive large-scale multi-modal gait datasets holds the potential to significantly advance the development of multi-modal gait recognition systems. These datasets offer a wealth of diverse information from various modalities, fostering improvements in the accuracy, robustness, and overall effectiveness of multi-modal gait recognition technology. Moreover, the development of long-distance gait datasets holds significant value. Developing datasets that focus on gait recognition from extended distances contributes to a more realistic and practical understanding of gait analysis, enhancing the applicability of such systems in real-world scenarios.

References

[1] David Cornett, Joel Brogan, Nell Barber, Deniz Aykac, Seth Baird, Nicholas Burchfield, Carl Dukes, Andrew Duncan, Regina Ferrell, Jim Goddard, et al. Expanding accurate person recognition to new altitudes and ranges: The briar dataset. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 593–602, 2023. 1, 4

[2] 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

[3] Xiang Li, Yasushi Makihara, Chi Xu, and Yasushi Yagi. Multi-view large population gait database with human meshes and its performance evaluation. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 4(2):234–248, 2022. 4

[4] 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. 3

data_cfg	dataset_name: Gait3D [4]
evaluator_cfg	sampler: frames_all_limit: 720 metric: euc transform: - type: BaseSilCuttingTransform
loss_cfg	- loss_term_weight: 1.0 margin: 0.2 type: TripletLoss - loss_term_weight: 1.0 scale: 16 type: CrossEntropyLoss
model_cfg	model: VM-Gait backbone_cfg: channels: - 64 - 128 - 256 - 512 layers: - 1 - 4 - 4 - 1 SeparateFCs: in_channels: 520 out_channels: 256 parts_num: 16 SeparateBNNecks: class_num: 3000 in_channels: 256 parts_num: 16 bin_num: - 16
optimizer_cfg	lr: 0.1 momentum: 0.9 solver: SGD weight_decay: 0.0005
scheduler_cfg	gamma: 0.1
trainer_cfg	log_iter: 100 total_iter: 80000 sampler: batch_size: - 32 - 4 frames_num_fixed: 30 sample_type: fixed_ordered transform: - type: Compose trf_cfg: - type: RandomPerspective prob: 0.2 - type: BaseSilCuttingTransform - type: RandomHorizontalFlip prob: 0.2 - type: RandomRotate prob: 0.2

Table 5. Configuration parameters for the Gait3D [4] dataset

data_cfg	dataset_name: OUMVLP-Mesh [3]
evaluator_cfg	sampler: frames_all_limit: 720 metric: euc transform: - type: BaseSilCuttingTransform
loss_cfg	- loss_term_weight: 1.0 margin: 0.2 type: TripletLoss - loss_term_weight: 1.0 scale: 16 type: CrossEntropyLoss
model_cfg	model: VM-Gait backbone_cfg: channels: - 64 - 128 - 256 - 512 layers: - 1 - 1 - 1 - 1 SeparateFCs: in_channels: 520 out_channels: 256 parts_num: 16 SeparateBNNecks: class_num: 5153 in_channels: 256 parts_num: 16 bin_num: - 16
optimizer_cfg	lr: 0.1 momentum: 0.9 solver: SGD weight_decay: 0.0005
scheduler_cfg	gamma: 0.1
trainer_cfg	log_iter: 100 total_iter: 120000 sampler: batch_size: - 32 - 8 frames_num_fixed: 30 sample_type: fixed_ordered transform: - type: Compose trf_cfg: - type: RandomPerspective prob: 0.2 - type: BaseSilCuttingTransform - type: RandomHorizontalFlip prob: 0.2 - type: RandomRotate prob: 0.2

Table 6. Configuration parameters for the OUMVLP-Mesh [3] dataset

data_cfg	dataset_name: BRIAR [1]
evaluator_cfg	sampler: frames_all_limit: 720 metric: euc transform: - type: BaseSilCuttingTransform
loss_cfg	- loss_term_weight: 1.0 margin: 0.2 type: TripletLoss - loss_term_weight: 1.0 scale: 16 type: CrossEntropyLoss
model_cfg	model: VM-Gait backbone_cfg: channels: - 64 - 128 - 256 - 512 layers: - 1 - 4 - 4 - 1 SeparateFCs: in_channels: 520 out_channels: 256 parts_num: 16 SeparateBNNecks: class_num: 273 in_channels: 256 parts_num: 16 bin_num: - 16
optimizer_cfg	lr: 0.1 momentum: 0.9 solver: SGD weight_decay: 0.0005
scheduler_cfg	gamma: 0.1
trainer_cfg	log_iter: 100 total_iter: 100000 sampler: batch_size: - 32 - 4 frames_num_fixed: 30 sample_type: fixed_ordered transform: - type: Compose trf_cfg: - type: BaseSilCuttingTransform

Table 7. Configuration parameters for the BRIAR [1] dataset