ENGINEERING Information Technology & Electronic Engineering
https://www.academax.com/EITEE; https://www.engineering.org.cn
https://ieeexplore.ieee.org/xpl/RecentIssue.jsp?punumber=11344065
E-mail: eng.itee@zju.edu.cn

Engineering Journals
Available on IEEE Xplore®

Research Article
https://doi.org/10.1631/ENG.ITEE.2026.0053

# SigmaGait: silhouette-guided motion-augmentation module for depth-based gait recognition

Changxin YE[1], Jingqi LI[1], Xuqian XUE[1], Xianye BEN[2], Hongming SHAN[3], Junping ZHANG[1]✉

[1]College of Computer Science and Artificial Intelligence, Fudan University, Shanghai 200438, China
[2]School of Information Science and Engineering, Shandong University, Qingdao 266237, China
[3]Institute of Science and Technology for Brain-inspired Intelligence, Fudan University, Shanghai 200433, China

**Abstract:** Precise depth data derived from light detection and ranging (LiDAR) have been shown to be effective in improving gait recognition performance. Those depth data, however, are coarse and lack clear segmentation boundaries, which often results in increased ambiguity in interpreting precise human motion and body shape, potentially limiting the model's ability to learn discriminative gait-related features. To address this challenge, we propose a novel silhouette-guided motion-augmentation module for depth-based gait recognition, namely, SigmaGait. SigmaGait leverages the clear boundaries and homogeneous foregrounds of silhouettes to enhance the model's awareness of part-level motions and fine-grained appearance features. Furthermore, we explore a pseudo-depth generation approach that leverages accessible red–green–blue (RGB) data for scenarios where LiDAR sensors are unavailable. By employing off-the-shelf skinned multi-person linear (SMPL) models, we synthesize pseudo-depth maps from estimated human meshes, bridging the performance gap between camera and LiDAR-reliant gait recognition. We empirically evaluate our approach on real and synthetic datasets including cloth-changing benchmark for person re-identification and gait recognition (CCPG), SUSTech1K, and FreeGait. Our model achieves state-of-the-art performance on real LiDAR data, and our pseudo-depth maps improve accuracy on 2D datasets despite their synthetic origin.

**Key words:** Gait recognition; Depth; Skinned multi-person linear (SMPL); Light detection and ranging (LiDAR); Pseudo-depth

## 1 Introduction

Gait recognition has emerged as a compelling biometric technology due to its non-invasive and long-range capability of identifying individuals based on their walking patterns. While traditional camera-based approaches relying on two-dimensional (2D) silhouettes have demonstrated strong performance in constrained settings, their representational capacity remains inherently limited. These 2D silhouettes only offer flattened projections of the human body, discarding critical depth-related information such as spatial configurations, body proportions, and relative positioning of limbs. This is in stark contrast to how humans perceive and recognize gait—through a naturally three-dimensional (3D) understanding of posture, body shape, and motion from different viewpoints.

These limitations have driven the exploration of 3D representations, particularly those derived from light detection and ranging (LiDAR) sensors, as a means to more explicitly capture structural and spatial details that are otherwise ambiguous in 2D. Recent studies like LidarGait (Shen et al., 2023) and FreeGait (Han et al., 2024) have demonstrated the feasibility of using LiDAR-based point clouds and projected range views—a form of depth map—for accurate gait recognition in large-scale outdoor settings. Compared to silhouettes, depth maps contain richer spatial information, using the red–green–blue (RGB) spectrum to show depth fluctuations of the entire body. However, they often lack clear segmented boundaries, leading to increased ambiguity in interpreting human motion and body shape. Depth maps projected from point clouds are also coarse, due to the limited LiDAR beam count (Ahn et al., 2025). This further degrades appearance, making it more difficult to accurately identify body-background boundaries. In contrast, silhouettes offer sharper and more distinct shape boundaries, thus preserving precise 2D body shapes and subtle motion at a higher resolution.

✉ Junping ZHANG, jpzhang@fudan.edu.cn
Xianye BEN, https://orcid.org/0000-0001-8083-3501
Hongming SHAN, https://orcid.org/0000-0002-0604-3197
Junping ZHANG, https://orcid.org/0000-0002-5924-3360
CLC number: TP391.41

Received: Feb. 25, 2026; Revision accepted: July 10, 2026;
Crosschecked: July 22, 2026

© The Authors 2026. Published by Zhejiang University Press Co., Ltd. This is an open access article distributed under the terms of the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0/)

While depth information has shown strong potential in gait recognition, its adoption in gait recognition remains limited. LiDAR sensors, while powerful, are not yet widely available or easily integrated into existing surveillance infrastructures (Guo et al., 2024). The widespread availability of RGB video cameras ensures that RGB-based data remain a highly relevant and widely adopted approach in gait recognition. This highlights the need for alternative solutions—such as synthesized depth maps—to allow depth-aware gait recognition in environments where LiDAR is not available.

We propose a silhouette-guided motion-augmentation module for our depth-based gait recognition, namely SigmaGait. Our model consists of two branches: a fusion branch for extracting appearance features by fusing the clearly segmented silhouettes with the spatially rich depth maps, and the other Sigma branch that uses the homogeneous foreground of silhouettes to guide the augmentation of part-based motions via our Sigma module.

Although fusing silhouettes with other modalities to obtain more discriminative gait representations has become common, their integration with LiDAR-based gait representations remains underexplored in recent works (Shen et al., 2023, 2025; Han et al., 2024), possibly due to their perceived redundancy and limited informativeness. Drawing inspiration from prior works (Zheng et al., 2022b; Li JQ et al., 2023b; Wang M et al., 2023; Jin et al., 2025a) that use cross-frame feature differences to model dynamic motion, our Sigma module relies on neighboring frame differences to capture temporal changes across body parts. Rather than allowing the model to freely learn implicit features from raw depth input, our module employs silhouette-guided attention to adaptively select body regions that are critical to gait recognition—independent of depth features. By aligning these attended regions with their counterparts in the depth modality, our SigmaGait model can learn localized depth dynamics that correspond to identity-specific motion patterns.

To compensate for the lack of LiDAR in real-world scenarios, prior work (Guo et al., 2024) has explored monocular depth estimation from RGB images as a substitute for depth input. While effective for general scene understanding, such method struggles to capture the fine-grained geometry of articulated human motion (Fig. 1). Based on this observation, we propose rendering pseudo-depth maps (PDmaps) from skinned multi-person linear (SMPL) meshes, which are generated from RGB videos using off-the-shelf SMPL inference models. The results can be viewed in Fig. 2.

SMPL inference incorporates human body priors, which fixes certain issues mentioned above. However, generating accurate 3D meshes from 2D images is still an objectively difficult task, due to the lack of explicit depth information and the ambiguity of projecting 3D structures onto a 2D space. As a result, PDmaps are best viewed as complementary information for silhouettes, unlike LiDAR-generated range views that can serve as the main modality. The silhouette guidance featured in our Sigma module can remedy this issue, as it uses the more accurate silhouettes to guide the relatively less accurate PDmaps.

Our main contributions are summarized as follows:

1. We propose a pipeline to generate PDmaps from SMPL meshes. This provides a novel way of synthesizing depth information from RGB data.

2. We propose a Sigma module, which leverages the clear boundaries and homogeneous foreground of silhouettes to enhance the model's awareness of subtle part-level motions and fine-grained appearance features.

3. We run a series of experiments comparing our method with different models and modalities. With the guidance of silhouettes, our SigmaGait method can achieve a boost of +5.5 percentage points (PPs) on the FreeGait dataset when using pure range views. When using PDmaps, our method has a boost of +4.7 PPs on CCPG compared to pure silhouette data.

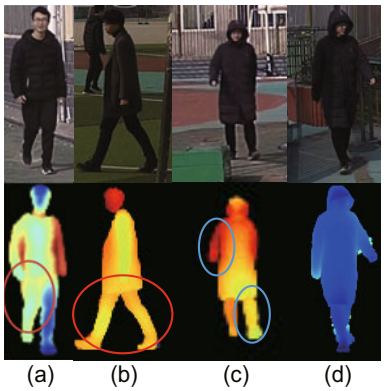


**Fig. 1 Issues with monocular depth estimation. (a, b) Limb regions that span multiple depth planes are incorrectly rendered with uniform color values. (c) Hands and feet at similar gait phases show mismatched depth values. (d) The depth map exhibits low coloration variation across the body, yielding limited depth information**

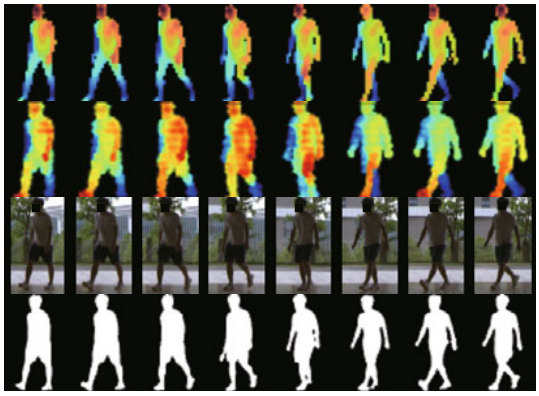

**Fig. 2 Comparison of gait representations: from top to bottom, PDmaps (ours), LiDAR range views, RGB, and silhouettes**

## 2 Related works

### 2.1 Silhouette-based gait recognition

Silhouette images have been the most popular form of 2D representation in recent years. Early works focused more on modeling local and global features, as well as aggregating spatial-temporal gait cues through handcrafted templates or simple pooling strategies. Wang KJ et al. (2021) proposed a gait template fusion strategy that integrates multiple silhouette-based representations into a unified gait

energy image (GEI) to enhance robustness under clothing and carrying variations. GaitSet (Chao et al., 2019, 2022) treats the gait sequence as an unordered set and employs a maximum function to compress the sequence of frame-level spatial features. GaitPart (Fan et al., 2020) proposes a micro-motion capture module that learns local body details from silhouette sequences. GaitGL (Lin et al., 2022) uses global and local convolution layers to capture different levels of features independently. Recently, researchers have shifted their focus toward stronger temporal modeling and more robust feature learning (Zheng et al., 2022b; Li JQ et al., 2023a; Ma et al., 2023; Wang L et al., 2023). For example, DyGait (Wang M et al., 2023) uses a dynamic augmentation module to calculate temporal differences and model dynamical motion. DeepGaitV2 (Fan et al., 2023a) challenges the trend of shallow models and gains strong performance on the challenging Gait3D (Zheng et al., 2022a) and GREW (Zhu et al., 2021) datasets. GaitGCI (Dou et al., 2023) addresses the issue of models focusing on confounding regions rather than true gait patterns. It uses counterfactual reasoning to guide attention toward body-relevant and discriminative regions and away from shortcut cues.

### 2.2 SMPL-based gait recognition

Early model-based gait recognition approaches focused on using SMPL parameters, particularly pose and shape vectors, as direct input features. Notable works such as human mesh recovery (HMR) network (Li X et al., 2020) and MvModelGait (Li X et al., 2021) aim to reconstruct accurate SMPL meshes from RGB video by leveraging multi-view inputs to reduce the inherent ambiguity of monocular viewpoints. These recovered SMPL parameters are then fed into temporal or spatial encoders such as long short-term memories (LSTMs) or convolutional neural networks (CNNs) for gait recognition. Building upon this paradigm, Gait3D (Zheng et al., 2022a) introduces the first large-scale in-the-wild gait dataset with dense 3D annotations, including SMPL parameters, 3D meshes, and silhouettes. While the SMPL parameters generated in Gait3D are relatively noisy due to the challenges of real-world capture, the authors incorporated them as supplementary input to a two-branch network, alongside silhouettes, in an effort to enhance recognition performance in complex scenarios.

Recent studies have highlighted the limitations of using raw SMPL parameters for gait recognition. While compact, these parameters are too abstract to capture dynamic topological or spatio-temporal cues. To address this, newer methods shift focus to SMPL-reconstructed meshes. HybridGait (Dong et al., 2024) projects SMPL meshes into 2D views, combining geometric and appearance features. VM-Gait (Wang ZY et al., 2025) introduces virtual markers—sparse and topologically meaningful points derived from SMPL meshes—processed via a point-based Transformer and fused with silhouette features. This fusion achieves strong performance across benchmarks, demonstrating the value of lightweight 3D cues alongside 2D representations. Despite the advantages of SMPL meshes, SMPL mesh-based gait recognition methods remain relatively sparse in the literature, with HybridGait and VM-Gait currently representing recent explorations in this direction.

### 2.3 LiDAR-based gait recognition

LidarGait (Shen et al., 2023) pioneers the usage of LiDAR in gait recognition, by building a LiDAR-based dataset SUSTech1K. The LidarGait model relies on image-based range views instead of using raw point clouds for input. This work proves that depth information is essential, as range-view models have significantly higher performance when compared to silhouette-only models. FreeGait dataset, released by Han et al. (2024), is also collected using LiDAR sensors. Improving upon LidarGait, HMRNet incorporates the input of raw point clouds via the hierarchical adaptive cross-representation mapping (ACM) module. However, both methods neglect to include silhouettes and instead choose to focus solely on the depth modality.

CL-Gait (Guo et al., 2024) identifies the need for camera–LiDAR cross-modality gait recognition, and applies contrastive pre-training to align the silhouette and depth modalities. A pipeline for pseudo-depth generation is further proposed via monocular depth estimated from a single RGB image, in the case of missing depth data. Monocular depth estimation is effective for general scene understanding; however, such methods struggle to capture the fine-grained geometry of articulated human motion (Fig. 1). Although CL-Gait achieves promising results under the cross-modality setting between LiDAR depth and pseudo-depth, its evaluation for pseudo-depth does not follow the standard probe-gallery identification protocol commonly used in gait recognition. Therefore, the results are not directly comparable with mainstream gait recognition methods.

## 3 Methods

### 3.1 Model overview

In this work, we propose SigmaGait, a novel two-stream framework that integrates depth maps and silhouettes for gait recognition. As illustrated in Fig. 3, there are two parts in our method. First is the PDmap pipeline, in which we use the world-grounded humans with accurate motion (WHAM) (Shin et al. 2024) SMPL inference model. We generate SMPL parameters and project PDmaps from SMPL meshes. Second is our model SigmaGait, which employs a two-stream convolutional encoder, with separate streams for silhouettes and depth maps. We adopt the design of DeepGaitV2 (Fan et al., 2023a) as our backbone, and feed depth maps and silhouettes into separate stacked pseudo-3D (P3D) encoders, enabling effective low-level feature extraction. After feature extraction, we apply conditional positional embedding (CPE) (Chu et al., 2021) before inputting our features into two branches of fusion: our silhouette-guided motion-augmentation (Sigma) module designed to extract dynamic motion features, and the other fusion based on the attention fusion used in SkeletonGait++ (Fan et al., 2024) for extracting static features (Jin et al., 2025b). We then take the output of both branches, and temporal pooling is used, before finally being input into separate fully-connected layers, a common embedding scheme used by gait recognition models (Chao et al., 2019; Fan et al., 2020, 2023a, 2024; Lin et al., 2022; Gao et al., 2023).

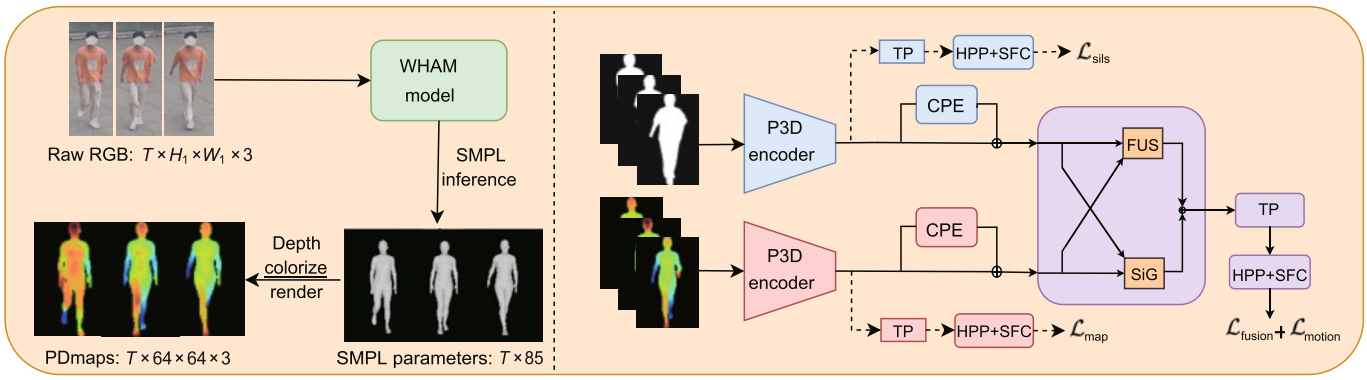


**Fig. 3 Overview of our PDmap pipeline (left) and SigmaGait network (right). TP is the temporal-pooling, SFC is the separate fully-connected layer, FUS is the fusion, SiG is the Sigma module, $\mathcal{L}_{\text{sils}}$ is the silhouette loss, $\mathcal{L}_{\text{map}}$ is the depth map encoder loss, $\mathcal{L}_{\text{fusion}}$ is the fusion loss, and $\mathcal{L}_{\text{motion}}$ is the silhouette-guided attention loss. $\oplus$ is the vector addition**

### 3.2 PDmap pipeline

First, we generate SMPL meshes from RGB videos for datasets that lack meshes (cloth-changing benchmark for person re-identification and gait recognition (CCPG) and SUSTech1K (Shen et al., 2023)). For our SMPL inference model, we choose WHAM (Shin et al., 2024), which is a state-of-the-art (SoTA) model at the time of this writing. WHAM is also heavily focused on predicting 3D human trajectories from videos captured using a moving camera, which includes large numbers of walking humans. This aligns well with our gait recognition setting and mitigates the errors introduced by discrepancies in data distribution.

For an RGB video sequence of $\boldsymbol{V} \in \mathbb{R}^{T \times H_1 \times W_1 \times 3}$, where $T$ is the number of frames, and $H_1$ and $W_1$ are the height and width of the frame, respectively, we extract SMPL features using the WHAM model and obtain the SMPL parameters $\boldsymbol{P} \in \mathbb{R}^{T \times 85}$. Once the SMPL parameters are obtained, we generate PDmaps from the corresponding SMPL meshes through mesh rasterization. Specifically, we render the $Z$-buffer of each mesh and apply perceptual colormaps to convert the scalar depth values into three-channel RGB pseudo-depth representations, which are directly used as gait representations. This yields a PDmap sequence of $\boldsymbol{D} \in \mathbb{R}^{T \times H_1 \times W_1 \times 3}$.

Compared to monocular image depth generation, SMPL inference incorporates human body priors, enforcing plausible limb positions and preventing missing limbs, while also excluding clothing, thereby reducing clothing-related noise. Despite these human body prior reinforcements, generating accurate 3D meshes from 2D images is still an objectively difficult task, due to the lack of explicit depth information and the ambiguity of projecting 3D structures onto a 2D space. From a single viewpoint, it is difficult to accurately infer hidden body parts, resolve self-occlusions, or capture fine-grained pose variations. This challenge extends to video-based inference as well, where per-frame SMPL predictions often suffer from temporal inconsistency, resulting in jittery or unstable mesh sequences without explicit temporal smoothing or tracking mechanisms. As a result, our PDmaps, despite an improvement upon monocular depth generation, still face minor temporal inconsistencies and spatial inaccuracies. Therefore, we rely on the silhouette data, which we consider higher-quality and more accurate, to guide the PDmaps during the training process.

### 3.3 SigmaGait network structure

1. 2-Stream encoder. Employing DeepGaitV2 as our model backbone, we rely on four stacks of P3D modules to effectively encode the silhouettes and depth maps. Instead of using full 3D convolutions, which are computationally expensive, P3D (Fan et al., 2023a) decomposes 3D convolutions into separate 2D spatial and one-dimensional (1D) temporal convolutions (as illustrated in Fig. 4), which are more computationally efficient. Given an input of $\boldsymbol{S} \in \mathbb{R}^{T \times H_1 \times W_1 \times 1}$ silhouette sequences and $\boldsymbol{D} \in \mathbb{R}^{T \times H_1 \times W_1 \times 3}$ depth-map sequences, the encoders extract feature maps of $\boldsymbol{F}^{\text{s}} \in \mathbb{R}^{T \times H_2 \times W_2 \times C}$ and $\boldsymbol{F}^{\text{d}} \in \mathbb{R}^{T \times H_2 \times W_2 \times C}$. Here, $C$ denotes the channel size. Before performing fusion and silhouette-guided motion augmentation, we apply CPE, achieved by a 1D convolutional neural network Conv1D() with a 3×1 kernel, to the temporal dimension. The 1D CPE is added:

$$\boldsymbol{F}^{\text{s}}_{\text{enc}} = \boldsymbol{F}^{\text{s}} + \text{Conv1D}(\boldsymbol{F}^{\text{s}}), \tag{1}$$

allowing the model to learn temporal position information. Here, $\boldsymbol{F}^{\text{s}}_{\text{enc}}$ denotes encoder features extracted from silhouette sequences. The same process is repeated for the depth branch, resulting in encoder features extracted from depth sequences $\boldsymbol{F}^{\text{d}}_{\text{enc}}$.

2. Fusion. Our fusion branch directly adopts the attention-fusion mechanism used in SkeletonGait++ (Fan et al., 2024) for extracting static features. Specifically, the silhouettes and depth features are first concatenated along the channel dimension and then fed into a lightweight attention module to model cross-modality interactions. The resulting attention responses are normalized by a softmax operation to generate modality-specific weights, which indicate the relative contribution of each branch. Finally, the silhouettes and depth features are recalibrated by their corresponding weights and fused through a weighted summation. While SkeletonGait++ applies this attention fusion after the first P3D layer, we perform the same fusion operation after the last P3D layer, allowing each modality to first undergo deeper temporal–spatial feature extraction before cross-modality fusion.

In our setting, this design allows the model to adaptively

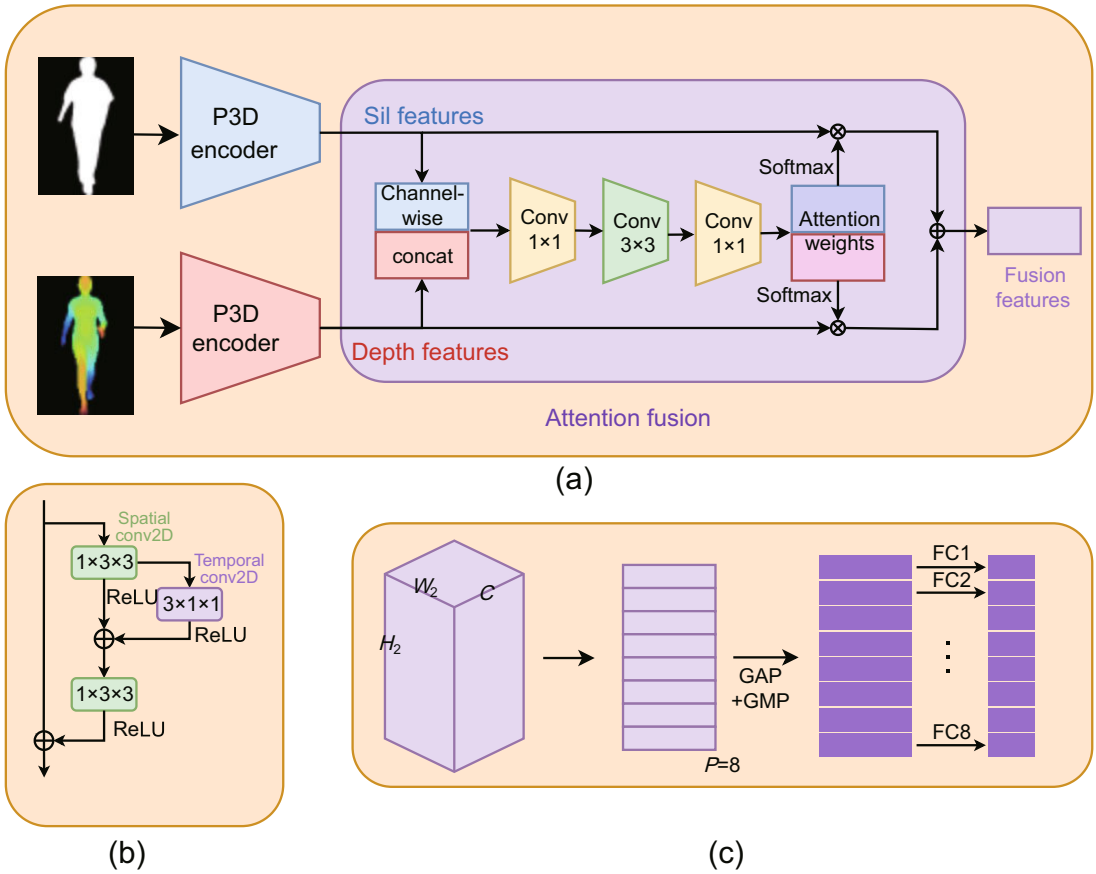


**Fig. 4 Diagram of attention fusion (a), P3D (b), and HPP (c) modules. GAP and GMP denote the global-average-pooling and global-max-pooling, respectively, ReLU is the rectified linear unit, FC is the fully-connected layer, Sil is silhouettes, concat is concatenate, ⊕ is the vector addition, and ⊗ is the vector multiplication**

integrate the complementary strengths of both modalities: Depth maps provide rich spatial and body-structure cues but often suffer from coarse or ambiguous boundaries, while silhouettes offer clean foreground segmentation and clear body contours. Therefore, fusing depth maps with well-segmented silhouettes helps preserve depth-aware structural information while enhancing the model's sensitivity to fine-grained appearance and boundary-related features.

3. Silhouette-guided motion augmentation. Depth maps offer richer spatial details than silhouettes by capturing full-body depth variations, but they often lack clear segmented boundaries, making motion and structure interpretation more difficult. The same body part may appear dissimilar across frames because of depth variations, while different body parts at similar depths may appear overly similar. Coupled with the fact that body parts may appear in various angles and move in different directions, the model may not be able to correctly track the subtle movement changes of specific parts, making localized depth dynamics challenging to learn.

In contrast, silhouettes are filled with a homogeneous color to represent the foreground, varying only in shape. Thus, performing cross-frame attention on silhouettes facilitates more reliable part-specific tracking, compared to depth maps where the same parts may vary in color.

As illustrated in Fig. 5, first we project the $t^{\text{th}}$ frame feature $\boldsymbol{f}_t^{\text{s}}$ to $\boldsymbol{k}$ and $(t+1)^{\text{th}}$ frame feature $\boldsymbol{f}_{t+1}^{\text{s}}$ to $\boldsymbol{q}$, where

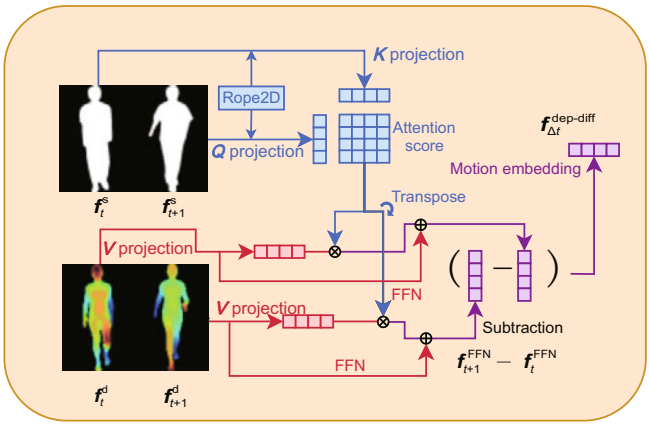


**Fig. 5 Diagram of our Sigma module. Note that the query ($Q$), key ($K$), values ($V$) are projected from silhouettes and depth features and not from raw images**

$\boldsymbol{F}_{\text{enc}}^{\text{s}} = [\boldsymbol{f}_t^{\text{s}} \,|\, t = 1, 2, \ldots, T]$, and calculate the $\boldsymbol{qk}$ correlation matrix as commonly done in cross-attention. With the $\boldsymbol{qk}$ correlation matrix producing an attention map that emphasizes regions undergoing positional change, we apply this map to $\boldsymbol{v}_t$, where $\boldsymbol{v}_t$ is projected from $\boldsymbol{f}_t^{\text{d}}$, thereby extracting depth features from adaptively attended body regions $\boldsymbol{f}_t^{\text{atten}}$. The same process is repeated for $\boldsymbol{v}_{t+1}$.

$$\boldsymbol{f}_t^{\text{atten}} = \text{softmax}\left(\frac{\boldsymbol{q}\boldsymbol{k}^{\text{T}}}{\sqrt{d_{\text{k}}}}\right)\boldsymbol{v}_t, \tag{2}$$

$$\boldsymbol{f}_{t+1}^{\text{atten}} = \text{softmax}\left(\frac{\boldsymbol{k}\boldsymbol{q}^{\text{T}}}{\sqrt{d_{\text{k}}}}\right)\boldsymbol{v}_{t+1}, \tag{3}$$

where $d_{\text{k}}$ denotes the feature dimension of key.

The final frame feature $\boldsymbol{f}_t^{\text{FFN}}$ is acquired through the feed-forward network (FFN) and layer normalization (LN), which is commonly found in Transformers.

$$\boldsymbol{f}_t^{\text{FFN}} = \text{LN}(\boldsymbol{f}_t^{\text{atten}} + \text{FFN}(\boldsymbol{f}_t^{\text{atten}})), \tag{4}$$

$$\boldsymbol{f}_{t+1}^{\text{FFN}} = \text{LN}(\boldsymbol{f}_{t+1}^{\text{atten}} + \text{FFN}(\boldsymbol{f}_{t+1}^{\text{atten}})). \tag{5}$$

Ideally, matching shapes that correspond to the same body parts will yield higher correlation scores and thus higher attention weights. Since the height and width dimensions are flattened into a single spatial dimension, the attention mechanism is invariant to positional distance, allowing effective tracking of moving body parts across frames. Applying the attention map to the depth features helps us effectively track key parts even if their color changes.

Finally, we perform max pooling to aggregate the attended depth features and perform neighboring frame subtraction:

$$\boldsymbol{f}_{\Delta t}^{\text{dep-diff}} = \boldsymbol{f}_{t+1}^{\text{FFN}} - \boldsymbol{f}_t^{\text{FFN}}. \tag{6}$$

The subtraction of neighboring frames gives us the depth fluctuation of adaptively attended parts $\boldsymbol{f}_{\Delta t}^{\text{dep-diff}}$, effectively modeling local motion dynamics.

Compared to the attention-fusion module, our Sigma module is more focused on using cross-attention to find relations between features of neighboring frames, whereas the attention fusion focuses mainly on intra-frame cross-modal representation.

4. Horizontal pyramid pooling (HPP). HPP (Chao et al., 2022) is a feature aggregation module commonly used in gait recognition to capture body-part-level representations. Since human gait contains discriminative cues from different horizontal body regions, such as the head, torso, legs, and feet, HPP divides the feature map into multiple horizontal strips at different scales. At each pyramid scale, the feature map is split into several horizontal bins, and pooling (global-average-pooling + global-max-pooling) is applied within each bin to obtain local part features, as illustrated in Fig. 4. By combining coarse global features and fine-grained part features, HPP helps the model learn both the overall body structure and detailed motion patterns.

### 3.4 Training objective

Triplet loss and softmax loss are common loss objectives used in gait recognition. Our model is guided by four triplet losses: two for the silhouette ($\mathcal{L}_{\text{sils}}$) and depth map encoder ($\mathcal{L}_{\text{map}}$) and two for the fusion ($\mathcal{L}_{\text{fusion}}$) and silhouette-guided attention embeddings ($\mathcal{L}_{\text{motion}}$). Our training objective features a combined loss term for encoder ($\mathcal{L}_{\text{enc}}$), joint ($\mathcal{L}_{\text{joint}}$), and softmax ($\mathcal{L}_{\text{softmax}}$) losses:

$$\mathcal{L}_{\text{enc}} = \mathcal{L}_{\text{sils}} + \mathcal{L}_{\text{map}}, \tag{7}$$

$$\mathcal{L}_{\text{joint}} = \mathcal{L}_{\text{fusion}} + \mathcal{L}_{\text{motion}}, \tag{8}$$

$$\mathcal{L} = \mathcal{L}_{\text{enc}} + \mu\mathcal{L}_{\text{joint}} + \gamma\mathcal{L}_{\text{softmax}}, \tag{9}$$

where $\mu$ and $\gamma$ denote the corresponding loss weights.

## 4 Experiments

### 4.1 Datasets

We conduct experiments on CCPG (Zou et al., 2024), SUSTech1K (Shen et al., 2023), and FreeGait (Han et al., 2024). For datasets with RGB videos, namely, CCPG and SUSTech1K, we apply our pipeline to generate PDmaps.

CCPG provides diverse covariate conditions, such as clothing and carrying variations, making it suitable for evaluating the effectiveness of PDmaps under appearance changes.

SUSTech1K provides both RGB videos and LiDAR range views, allowing direct comparison between synthesized PDmaps and sensor-captured depth data under relatively controlled indoor conditions.

FreeGait is an unconstrained LiDAR-based benchmark collected in open public environments such as subway exits and school gates. Compared with SUSTech1K, FreeGait involves more complex backgrounds, unscripted walking trajectories, longer capture distances, and frequent occlusions, making it a more challenging benchmark.

Table 1 displays the key statistics of the evaluated datasets. The train/test splits follow the standard partition lists provided by OpenGait to ensure consistency with existing gait recognition protocols.

**Table 1 Datasets in use**

| Dataset | Train set | | Test set | | Collection situation |
|---|---|---|---|---|---|
| | $N_{\text{ID}}$ | $N_{\text{Seq}}$ | $N_{\text{ID}}$ | $N_{\text{Seq}}$ | |
| SUSTech1K | 200 | 5988 | 850 | 19 228 | Constrained |
| CCPG | 100 | 8187 | 100 | 8095 | Constrained |
| FreeGait | 500 | 7659 | 695 | 4262 | Real-world |

$N_{\text{ID}}$ and $N_{\text{Seq}}$ present the numbers of identities and sequences, respectively

### 4.2 Implementation details

Our experiments follow the established protocol used in Chao et al. (2019), Fan et al. (2020, 2023a, 2024), and Lin et al. (2022), using the OpenGait (Fan et al., 2023b) framework.

1. For CCPG and SUSTech1K, we use off-the-shelf WHAM for SMPL inference to generate SMPL parameters. For WHAM model, we keep the modifications to a minimum and only replace the minimum frame smoothing with a frame interpolation function, without adjusting the model. We use pytorch3d.render to render the SMPL parameters into depth maps.

2. For inputs, our silhouettes and depth maps are aligned and resized to 1×64×44 and 3×64×44, respectively. Our model requires modalities during training and inference stages. Depth maps can be either range views or PDmaps.

3. For the attention in our silhouette-guided motion-augmentation module, we have four heads with 256 dimensions for each head. We incorporate 2D rotary positional encoding (RoPE) (Heo et al., 2024) to allow the attention mechanism to effectively capture relative spatial dependencies across height and width. For the HPP, part count is set to 16.

Table 2 displays the training milestones and total training steps used for training. We use the stochastic gradient descent

**Table 2 Training configurations for different datasets**

| Dataset | Batch size | Milestone ($\times 10^4$) | Number of steps ($\times 10^4$) |
|---|---|---|---|
| SUSTech1K | (8, 8, 10) | (3, 4, 5) | 6 |
| CCPG | (8, 16, 30) | (2, 4, 5) | 6 |
| FreeGait | (32, 4, 10) | (3, 4, 5) | 6 |

(SGD) optimizer with an initial learning rate of 0.1, momentum of 0.9, and weight decay of 0.0005. We use FP16 mixed-precision training to reduce memory consumption and improve training efficiency.

### 4.3 Result analysis

1. SUSTech1K. The SUSTech1K dataset includes a wide range of experimental settings, such as varying camera viewpoints, clothing, and environmental conditions. This diversity makes it a comprehensive benchmark for evaluating model robustness across different real-world scenarios and testing generalization under varied conditions. Additionally, it provides both LiDAR and camera data, making it a suitable benchmark for comparing our PDmaps against depth information obtained directly from LiDAR sensors.

As shown in Table 3, in the SUSTech1K-PDmap setting, SigmaGait surpasses camera-based SoTA approaches, demonstrating that synthesized PDmaps can effectively introduce depth-aware structural cues into RGB-based gait recognition. Although real LiDAR data naturally provide a more precise sensor-captured depth, PDmaps offer a practical and low-cost alternative in scenarios where LiDAR sensors are unavailable or infeasible to deploy, making depth-enhanced gait recognition more accessible in real-world applications. When using a real LiDAR dataset, our SigmaGait model achieves SoTA performance, primarily because of the addition of silhouette data. Our model is less competitive on the clothing scenario, which is also a challenging scenario for silhouette-based models. As such, our model is still prone to appearance-based biases induced by variations in clothing.

2. CCPG. The CCPG dataset includes a diverse set of modalities such as silhouette, parsing, skeleton, and RGB, making it a suitable benchmark for evaluating the effectiveness of our proposed PDmap modality in comparison to existing representations.

As shown in Table 4, SigmaGait outperforms all other SoTA models in terms of rank-1 accuracy under the clothing (CL) and up (UP) conditions. Under bag (BG) and down (DN) conditions, our model lags slightly behind BigGait, but still outperforms other SoTA methods. Compared with DeepGaitV2, which uses only silhouettes, SigmaGait achieves improvements of +6.6 PPs, +4.3 PPs, +5.8 PPs, and +2.0 PPs under the CL, UP, DN, and BG conditions, respectively. Compared with other 2-stream models like SkeletonGait++, our model achieves improvements of +4.3 PPs on average. This demonstrates that the performance gain does not simply come from a stronger backbone, but from the complementary information introduced by PDmaps. While silhouettes mainly capture body contours and foreground shape, PDmaps provide additional depth-aware cues related to 3D body structure and part-level spatial relationships. The consistent improvements across different conditions therefore verify that pseudo-depth representations can effectively supplement silhouette-based gait features.

**Table 3 Rank-1 accuracy on the SUSTech1K dataset under various probe conditions: normal (NM), bag (BG), clothing (CL), carrying (CAR), umbrella (UMB), uniform (UNI), occlusion (OCC), and night (NI)**

| Model | Input | Rank-1 accuracy (%) | | | | | | | | Overall (%) |
|---|---|---|---|---|---|---|---|---|---|---|
| | | NM | BG | CL | CAR | UMB | UNI | OCC | NI | |
| GaitBase (Fan et al., 2023b) | Sil | 83.1 | 79.3 | 51.0 | 77.0 | 77.3 | 77.3 | 83.5 | 26.7 | 77.5 |
| DeepGaitV2 (Fan et al., 2023a) | Sil | 83.5 | 79.5 | 46.3 | 76.8 | 79.2 | 78.5 | 81.2 | 27.3 | 77.4 |
| SkeletonGait++ (Fan et al., 2024) | Sil+Ske | 85.1 | 82.9 | 46.6 | 81.9 | 80.8 | 82.5 | 86.2 | **47.5** | 81.3 |
| SigmaGait (ours) | Sil+PDmaps | **88.7** | **86.2** | **53.3** | **85.0** | **84.7** | **87.3** | **88.4** | 38.6 | **84.7** |
| LidarGait (Shen et al., 2023) | RV | 91.9 | 88.6 | 75.3 | 89.0 | 67.6 | 81.2 | 94.7 | 90.0 | 86.8 |
| HMRNet (Han et al., 2024) | RV+PC | 92.7 | 92.3 | 79.6 | 90.3 | 83.1 | 86.2 | 95.2 | 90.4 | 90.2 |
| LidarGait++ (Shen et al., 2025) | PC | 94.2 | 93.9 | **79.7** | 92.4 | 91.5 | 91.9 | 96.6 | 92.2 | 92.7 |
| SigmaGait (ours) | Sil+RV | **96.4** | **96.3** | 78.2 | **95.1** | **93.7** | **95.5** | **97.9** | **92.6** | **94.9** |

Sil denotes silhouettes, Ske denotes skeleton, RV denotes depth images from the LiDAR range views, and PC denotes point clouds. The best results are in bold

**Table 4 Comparison of rank-1 accuracy across different modalities on CCPG**

| Model | Input | Rank-1 accuracy (%) | | | | AVG (%) |
|---|---|---|---|---|---|---|
| | | CL | UP | DN | BG | |
| BigGait (Ye et al., 2024) | RGB | 82.6 | 85.9 | **87.1** | **93.1** | 87.2 |
| DeepGaitV2 (Fan et al., 2023a) | Sil | 78.6 | 84.8 | 80.7 | 89.2 | 83.3 |
| GaitC$^3$I (Wang JL et al., 2025) | Sil | 81.5 | 86.6 | 83.6 | 90.7 | 85.6 |
| BiFusion (Peng et al., 2024) | Sil+Ske | 62.6 | 67.6 | 66.3 | 66.0 | 65.6 |
| SkeletonGait++ (Fan et al., 2024) | Sil+Ske | 79.1 | 83.9 | 81.7 | 89.9 | 83.7 |
| SigmaGait (ours) | Sil+PDmaps | **85.2** | **89.1** | 86.5 | 91.2 | **88.0** |

The various conditions are clothing (CL), up (UP), down (DN), and bag (BG). AVG denotes the average score. Sil denotes silhouettes and Ske denotes skeleton. The best results are in bold

3. FreeGait. FreeGait is designed to evaluate gait recognition in open and unconstrained public environments, and contains more complex real-world variations, including dynamic backgrounds, unscripted walking trajectories, longer capture distances, and frequent occlusions. These factors increase the difficulty of recognition and make FreeGait a more challenging benchmark for evaluating model robustness in practical scenarios.

As shown in Table 5, SigmaGait achieves SoTA performance, outperforming existing LiDAR-based methods that rely on point clouds or range-view representations alone. This demonstrates the advantage of incorporating silhouettes into depth-based gait recognition: range views provide 3D structural cues, while silhouettes offer clearer foreground boundaries and discriminative body shape information, resulting in a more complete gait representation.

**Table 5 Comparisons with methods on FreeGait**

| Model | Input | Accuracy (%) | mAP (%) |
|---|---|---|---|
| GaitSet (Chao et al., 2019) | Sil | 57.1 | 64.0 |
| GaitBase (Fan et al., 2023b) | Sil | 62.6 | 68.6 |
| LidarGait (Shen et al., 2023) | RV | 74.2 | 80.7 |
| HMRNet (Han et al., 2024) | RV+PC | 80.8 | 86.5 |
| LidarGait++ (Shen et al., 2025) | PC | 82.0 | 87.2 |
| SigmaGait (ours) | Sil+RV | **87.5** | **91.1** |

mAP denotes the mean average precision, Sil denotes silhouettes, RV denotes depth images from the LiDAR range views, and PC denotes point clouds. The best results are in bold

## 4.4 Ablation studies

1. Single modality. We analyze the rank-1 accuracy of single modalities such as silhouette, skeleton, optical flow, parsing, and PDmaps. The purpose of this experiment is to assess the standalone discriminative ability of each modality and provide a baseline for evaluating multimodal complementarity. By comparing silhouette-only, PDmaps-only, and silhouette–PDmaps results with other single/combined modalities, we can determine whether PDmaps provide additional useful information beyond silhouettes and whether the proposed fusion strategy effectively exploits this complementary information.

As shown in Table 6, PDmaps exhibit relatively limited standalone discriminative ability compared with skeleton maps, optical flow (Jin et al., 2025b), and parsing. This suggests that pseudo-depth representations are better suited to complement silhouette features rather than serve as an independent modality. This relatively weak standalone performance may be attributed to inaccuracies accumulated during the RGB-to-SMPL parameter estimation process, which can introduce errors in body geometry and local limb structures. In addition, SMPL-based reconstruction relies on generalized human body priors and may not fully preserve fine-grained identity cues, such as gender-specific body shape variations, hairstyle details, or other subtle appearance characteristics. These limitations may also explain the remaining performance gap between synthesized PDmaps and real LiDAR range views.

However, despite these imperfections, PDmaps bring a substantially larger performance improvement when combined with silhouettes. This indicates that PDmaps contain modality-specific complementary cues that are not explicitly captured by silhouette representations. In contrast, the smaller gains obtained from skeleton maps and optical flow may result from their considerable information overlap with silhouettes, as these modalities mainly encode structural patterns. PDmaps, by introducing additional depth-aware information, provide stronger complementarity and serve as an effective supplementary representation for silhouette-based gait recognition.

**Table 6 Comparison of different modalities on the CCPG dataset**

| Backbone | Input | Rank-1 accuracy (%) | | | | AVG (%) |
|---|---|---|---|---|---|---|
| | | CL | UP | DN | BG | |
| DeepGaitV2 | Sil | 78.6 | 84.8 | 80.7 | 89.2 | 83.3 |
| SkeletonGait | Ske | 40.4 | 48.5 | 53.0 | 61.7 | 50.9 |
| DeepGaitV2$^{P}$ | P | 69.6 | 75.8 | 75.8 | 83.3 | 76.1 |
| DeepGaitV2$^{F}$ | F | 68.8 | 73.3 | 75.0 | 76.8 | 73.5 |
| DeepGaitV2 | PDmaps | 39.1 | 43.1 | 47.2 | 40.8 | 42.6 |
| SkeletonGait++ | Sil+Ske | 79.1 | 83.9 | 81.7 | 89.9 | 83.7 |
| MultiGait++ | Sil+P+F | 83.9 | 89.0 | 86.0 | **91.5** | 87.6 |
| SigmaGait | Sil+PDmaps | **85.2** | **89.1** | **86.5** | 91.2 | **88.0** |

P and F for MultiGait++ (Jin et al., 2025b) and DeepGaitV2 (Fan et al., 2023a) denote parsing and optical flow, respectively. The various conditions are clothing (CL), up (UP), down (DN), and bag (BG). AVG denotes the average score. The best results are in bold

2. Fusion/Sigma modules. We evaluate the effectiveness of the fusion module and the proposed Sigma module, and analyze their individual and combined contributions to verify how each component improves the performance of our gait model.

As shown in Table 7, our model experiences a performance drop when branches are removed. Our silhouette-guided motion-augmentation module does not provide as much performance gain as using static fusion methods, which is consistent with other research works (Cui and Kang, 2023; Han et al., 2024; Jin et al., 2025b), as appearance remains the main contributor to accurate gait recognition.

**Table 7 Contributions of fusion and Sigma modules on CCPG, SUSTech1K-LiDAR (SUS), and FreeGait**

| Fusion | Sigma | Rank-1 accuracy (%) | | |
|---|---|---|---|---|
| | | CCPG | SUS | FreeGait |
| | ✓ | 69.8 | 80.8 | 78.8 |
| ✓ | | 87.5 | 94.1 | 84.0 |
| ✓ | ✓ | **88.0** | **94.9** | **87.5** |

Best results are in bold

A substantial performance improvement is observed on the more unconstrained FreeGait dataset (+3.5 PPs), highlighting the effectiveness of the Sigma module in complex real-world scenarios where longer LiDAR capture ranges can degrade the quality of range-view representations. Compared with SUSTech1K-LiDAR (+0.8 PPs), which is captured at a range of approximately 8 m, FreeGait includes longer-range LiDAR data of up to 15 m, making the projected range views coarser and less structurally coherent. In this setting, silhouette guidance becomes more beneficial by providing

clearer body boundaries and more stable shape cues to complement the degraded depth representations.

3. Sigma-guided components. We demonstrate the necessity of the silhouette-guided and cross-frame attention components featured in our silhouette-guided motion-augmentation (Sigma) module. To improve the reliability of this component study, we conduct experiments across four train/test splits (the original split plus three randomly generated ones) and report the mean±standard deviation values. Specifically, we evaluate the Sigma module under the following settings:

(1) Using other forms of guidance such as depth-guided depth, depth-guided silhouette, and silhouette-guided silhouette.

(2) We replace the cross-frame attention with intra-frame attention.

As shown in Table 8, both silhouette-guided and cross-frame attention consistently improve recognition performance across different train/test split configurations, demonstrating the reliability of the proposed Sigma design. Without cross-frame attention, the module degrades to self-attention on the spatial dimension, losing its motion awareness. Under depth-guided depth or silhouette-guided silhouette setting, our module loses its cross-modality attention, and performance drops. When depth is used to guide the silhouette branch, the fine-grained contour information of silhouettes is constrained by the comparatively coarse structural cues encoded in depth maps, which may limit the model's attention to discriminative shape details.

**Table 8 Component study on FreeGait range views. We examine the silhouette-guided and cross-frame attention in our Sigma module across four different train/test split configurations**

| Guidance | Cross/Intra-frame | Mean±Std (%) |
|---|---|---|
| RV-guided RV | Cross-frame | 83.8±2.1 |
| RV-guided Sil | Cross-frame | 83.5±2.5 |
| Sil-guided Sil | Cross-frame | 82.8±2.8 |
| Sil-guided RV | Intra-frame | 83.4±2.3 |
| Sil-guided RV | Cross-frame | **84.3±2.4** |

Std denotes the standard deviation. RV stands for range views, and is the depth input for the FreeGait dataset. The best result is in bold

In contrast, the silhouette-guided depth strategy achieves superior performance and aligns with our core design principle. By leveraging the precise boundary and structural priors of silhouettes to refine depth representations, the model enhances cross-modal feature interaction and improves recognition robustness.

4. Attention map visualization. We retrieve the attention map from the silhouette-guided motion-augmentation module for visualization to better understand what the Sigma module attends to. The attention maps are processed after subtraction to highlight regions with stronger relative attention values, allowing the spatial focus of the Sigma module to be more clearly observed.

In Fig. 6, attention maps are visualized using four adjacent frames across 12 subjects. Each labeled group contains three different sequences, with each row showing a four-frame segment from one sequence. Figs. 6a and 6b are collected from the SUSTech1K-LiDAR dataset, while Figs. 6c and 6d are from the CCPG-PDmaps. We overlay the attention maps onto the silhouettes for clearer visualization, as the colorized depth maps may interfere with the color-coded attention map. Since the depth maps are approximately aligned with the silhouettes, overlaying the attention maps on silhouettes can still reflect the corresponding attended regions in the depth features. It should be noted that this is only for visualization purposes; in the actual model, the attention is applied to the depth features.

As shown in Figs. 6a and 6b, the attention is mainly concentrated around the knee and upper-thigh regions. In contrast, Figs. 6c and 6d exhibit more distributed attention across multiple body parts. Specifically, Fig. 6c shows relatively stronger focus around the shoulders and arms, while Fig. 6d places more prominent attention around the feet. We observe that for range-view data (Figs. 6a and 6b), the model places more attention weight on the upper thighs and knee, compared to PDmaps (Figs. 6c and 6d) where shoulders and body are more attended.

We think that this difference reflects the modality-specific reliability of the depth representations. Since LiDAR range views directly capture lower-limb depth variations, the model tends to focus more on highly discriminative regions such as the knees and upper thighs. In contrast, PDmaps are generated from SMPL-based reconstruction and may provide smoother structural representations, making them less sensitive to subtle fine-grained motion changes. As a result, the attention may shift toward other informative regions, such as the shoulders, arms, torso, or feet.

## 5 Conclusions

In this paper, we propose a silhouette-guided motion-augmentation model that enhances the motion and appearance awareness of depth-based gait recognition. Our experiments reveal the limitations of LiDAR-only methods and demonstrate the effectiveness of combining silhouette and depth modalities. We also propose a novel method to supplement 2D data with pseudo-depth representations to enhance gait recognition performance. Our method involves generating SMPL parameters from raw RGB image sequences and then synthesizing 2D PDmaps from the SMPL meshes. Experimental results demonstrate that the incorporation of PDmaps yields significant performance improvements over silhouette-only baselines. This validates the effectiveness of our synthesized depth information, and highlights its potential as a practical alternative for scenarios lacking access to LiDAR sensors.

Despite its effectiveness, the proposed method still has several limitations that can be explored in future work. The primary limitation lies in its dependence on SMPL estimation quality, since challenging cases such as occlusion or low-quality RGB inputs may lead to inaccurate mesh reconstruction and affect the reliability of the generated pseudo-depth representations. In addition, large clothing variations and severe or long-term occlusions remain challenging conditions for gait recognition in general, as they may substantially alter or obscure discriminative body cues. The current framework is also

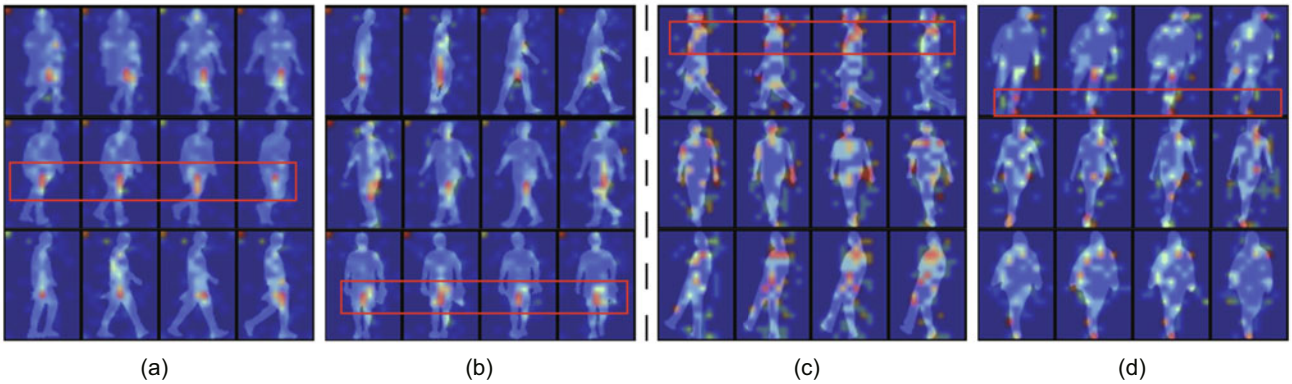


**Fig. 6 Attention map visualization. The attention maps are processed after subtraction to highlight regions with stronger relative attention responses, allowing the spatial focus of the Sigma module to be more clearly observed. Examples (a) and (b) are from SUSTech1K-LiDAR, (c) and (d) are from CCPG**

designed for paired silhouette–depth input during training and inference. Future work will explore more robust SMPL estimation, cross-modality alignment, and missing-modality adaptation to improve robustness and support more flexible inference when either one or both modalities are available.

## Acknowledgments

This work was supported by the National Natural Science Foundation of China (Nos. 62576103, 62322111, and 62271289) and the Natural Science Fund for Distinguished Young Scientists of Shandong Province (No. ZR2024JQ007).

## Author contributions

Changxin YE conceived and designed the study, conducted the experiments and analyses, prepared the visualizations, and drafted and revised the paper. Jingqi LI contributed to the conceptualization and methodological design of the study and reviewed and edited the paper. Xuqian XUE and Xianye BEN reviewed and edited the paper. Hongming SHAN contributed to writing the original draft and supervised the research. Junping ZHANG supervised the research and reviewed and edited the paper.

## Conflict of interest

Junping ZHANG is an editorial board member of *ENGINEERING Information Technology & Electronic Engineering*, and he was not involved with the peer review process of this paper. All the authors declare that they have no conflict of interest.

## Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

## Declaration on the use of generative AI tools

During the preparation of this work, the authors used ChatGPT to improve language. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

## References

Ahn J, Nakashima K, Yoshino K, et al., 2025. Gait sequence upsampling using diffusion models for single LiDAR sensors. IEEE/SICE Int Symp on System Integration, p.658-664. https://doi.org/10.1109/SII59315.2025.10870999

Chao HQ, He YW, Zhang JP, et al., 2019. GaitSet: regarding gait as a set for cross-view gait recognition. Proc 33rd AAAI Conf on Artificial Intelligence, p.8126-8133. https://doi.org/10.1609/aaai.v33i01.33018126

Chao HQ, Wang K, He YW, et al., 2022. GaitSet: cross-view gait recognition through utilizing gait as a deep set. *IEEE Trans Pattern Anal Mach Intell*, 44(7):3467-3478. https://doi.org/10.1109/TPAMI.2021.3057879

Chu XX, Tian Z, Zhang B, et al., 2021. Conditional positional encodings for vision Transformers. Proc 11th Int Conf on Learning Representations.

Cui YF, Kang YM, 2023. Multi-modal gait recognition via effective spatial-temporal feature fusion. IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.17949-17957. https://doi.org/10.1109/CVPR52729.2023.01721

Dong YL, Yu CL, Ha RY, et al., 2024. HybridGait: a benchmark for spatial-temporal cloth-changing gait recognition with hybrid explorations. Proc 38th AAAI Conf on Artificial Intelligence, p.1600-1608. https://doi.org/10.1609/aaai.v38i2.27926

Dou HZ, Zhang PY, Su W, et al., 2023. GaitGCI: generative counterfactual intervention for gait recognition. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.5578-5588. https://doi.org/10.1109/CVPR52729.2023.00540

Fan C, Peng YJ, Cao CS, et al., 2020. GaitPart: temporal part-based model for gait recognition. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.14213-14221. https://doi.org/10.1109/CVPR42600.2020.01423

Fan C, Hou SH, Huang YZ, et al., 2023a. Exploring deep models for practical gait recognition. https://arxiv.org/abs/2303.03301

Fan C, Liang JH, Shen CF, et al., 2023b. OpenGait: revisiting gait recognition towards better practicality. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.9707-9716. https://doi.org/10.1109/CVPR52729.2023.00936

Fan C, Ma JZ, Jin DY, et al., 2024. SkeletonGait: gait recognition using skeleton maps. Proc 38th AAAI Conf on Artificial Intelligence, p.1662-1669. https://doi.org/10.1609/aaai.v38i2.27933

Gao JQ, Li JQ, Shan HM, et al., 2023. Forget less, count better: a domain-incremental self-distillation learning benchmark for lifelong crowd counting. *Front Inform Technol Electron Eng*, 24(2):187-202. https://doi.org/10.1631/FITEE.2200380

Guo WX, Liang YP, Pan ZY, et al., 2024. Camera-LiDAR cross-modality gait recognition. Proc 18th European Conf on Computer Vision, p.439-455. https://doi.org/10.1007/978-3-031-72754-2_25

Han X, Ren YM, Cong PS, et al., 2024. Gait recognition in large-scale free environment via single LiDAR. Proc 32nd ACM Int Conf on Multimedia, p.380-389. https://doi.org/10.1145/3664647.3681166

Heo B, Park S, Han D, et al., 2024. Rotary position embedding for vision Transformer. Proc 18th European Conf on Computer Vision, p.289-305. https://doi.org/10.1007/978-3-031-72684-2_17

Jin DY, Fan C, Ma JZ, et al., 2025a. On denoising walking videos for gait recognition. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.12347-12357. https://doi.org/10.1109/CVPR52734.2025.01152

Jin DY, Fan C, Chen WH, et al., 2025b. Exploring more from multiple gait modalities for human identification. Proc 39th AAAI Conf on Artificial Intelligence, p.4120-4128. https://doi.org/10.1609/aaai.v39i4.32432

Li JQ, Zhang YZ, Shan HM, et al., 2023a. GaitCoTr: improved spatial-temporal representation for gait recognition with a hybrid convolution-Transformer framework. IEEE Int Conf on Acoustics, Speech and Signal Processing, p.1-5. https://doi.org/10.1109/ICASSP49357.2023.10096602

Li JQ, Gao JQ, Zhang YZ, et al., 2023b. Motion matters: a novel motion modeling for cross-view gait feature learning. IEEE Int Conf on Acoustics, Speech and Signal Processing, p.1-5. https://doi.org/10.1109/ICASSP49357.2023.10096571

Li X, Makihara Y, Xu C, et al., 2020. End-to-end model-based gait recognition. Proc 15th Asian Conf on Computer Vision, p.3-20. https://doi.org/10.1007/978-3-030-69535-4_1

Li X, Makihara Y, Xu C, et al., 2021. End-to-end model-based gait recognition using synchronized multi-view pose constraint. Proc IEEE/CVF Int Conf on Computer Vision Workshops, p.4089-4098. https://doi.org/10.1109/ICCVW54120.2021.00456

Lin BB, Zhang SL, Wang M, et al., 2022. GaitGL: learning discriminative global-local feature representations for gait recognition. https://arxiv.org/abs/2208.01380

Ma K, Fu Y, Zheng DZ, et al., 2023. Dynamic aggregated network for gait recognition. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.22076-22085. https://doi.org/10.1109/CVPR52729.2023.02114

Peng YJ, Ma K, Zhang Y, et al., 2024. Learning rich features for gait recognition by integrating skeletons and silhouettes. *Multim Tools Appl*, 83(3):7273-7294. https://doi.org/10.1007/s11042-023-15483-x

Shen CF, Chao F, Wu W, et al., 2023. LidarGait: benchmarking 3D gait recognition with point clouds. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.1054-1063. https://doi.org/10.1109/CVPR52729.2023.00108

Shen CF, Wang R, Duan LX, et al., 2025. LidarGait++: learning local features and size awareness from LiDAR point clouds for 3D gait recognition. Proc Computer Vision and Pattern Recognition Conf, p.6627-6636. https://doi.org/10.1109/CVPR52734.2025.00621

Shin S, Kim J, Halilaj E, et al., 2024. WHAM: reconstructing world-grounded humans with accurate 3D motion. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.2070-2080. https://doi.org/10.1109/CVPR52733.2024.00202

Wang JL, Hou SH, Guo XD, et al., 2025. GaitC$^3$I: robust cross-covariate gait recognition via causal intervention. *IEEE Trans Circ Syst Video Technol*, 35(8):8057-8070. https://doi.org/10.1109/TCSVT.2025.3545210

Wang KJ, Liu LL, Ding XN, et al., 2021. A partition approach for robust gait recognition based on gait template fusion. *Front Inform Technol Electron Eng*, 22(5):709-719. https://doi.org/10.1631/FITEE.2000377

Wang L, Liu B, Liang FF, et al., 2023. Hierarchical spatio-temporal representation learning for gait recognition. IEEE/CVF Int Conf on Computer Vision, p.19582-19592. https://doi.org/10.1109/iccv51070.2023.01799

Wang M, Guo XD, Lin BB, et al., 2023. DyGait: exploiting dynamic representations for high-performance gait recognition. Proc IEEE/CVF Int Conf on Computer Vision, p.13378-13387. https://doi.org/10.1109/ICCV51070.2023.01235

Wang ZY, Liu J, Chen JN, et al., 2025. VM-Gait: multi-modal 3D representation based on virtual marker for gait recognition. IEEE/CVF Winter Conf on Applications of Computer Vision, p.5326-5335. https://doi.org/10.1109/WACV61041.2025.00520

Ye DQ, Fan C, Ma JZ, et al., 2024. BigGait: learning gait representation you want by large vision models. IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.200-210. https://doi.org/10.1109/cvpr52733.2024.00027

Zheng JK, Liu XC, Liu W, et al., 2022a. Gait recognition in the wild with dense 3D representations and a benchmark. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.20196-20205. https://doi.org/10.1109/CVPR52688.2022.01959

Zheng JK, Liu XC, Gu XY, et al., 2022b. Gait recognition in the wild with multi-hop temporal switch. Proc 30th ACM Int Conf on Multimedia, p.6136-6145. https://doi.org/10.1145/3503161.3547897

Zhu Z, Guo X, Yang T, et al., 2021. Gait recognition in the wild: a benchmark. Proc IEEE/CVF Int Conf on Computer Vision, p.14769-14779. https://doi.org/10.1109/ICCV48922.2021.01452

Zou SN, Fan C, Xiong JB, et al., 2024. Cross-covariate gait recognition: a benchmark. Proc 38th AAAI Conf on Artificial Intelligence, p.7855-7863. https://doi.org/10.1609/aaai.v38i7.28621