# Walking Further: Semantic-aware Multimodal Gait Recognition Under Long-Range Conditions

**Zhiyang Lu[1], Wen Jiang[1], Tianren Wu[1], Zhichao Wang[1], Changwang Zhang[2], Siqi Shen[1], Ming Cheng[1]✉**

[1]Fujian Key Laboratory of Sensing and Computing for Smart Cities, Xiamen University
[2]OPPO Research Institute
chm99@xmu.edu.cn, changwangzhang@foxmail.com

## Abstract

Gait recognition is an emerging biometric technology that enables non-intrusive and hard-to-spoof human identification. However, most existing methods are confined to short-range, unimodal settings and fail to generalize to long-range and cross-distance scenarios under real-world conditions. To address this gap, we present **LRGait**, the first LiDAR-Camera multimodal benchmark designed for robust long-range gait recognition across diverse outdoor distances and environments. We further propose **EMGaitNet**, an end-to-end framework tailored for long-range multimodal gait recognition. To bridge the modality gap between RGB images and point clouds, we introduce a semantic-guided fusion pipeline. A CLIP-based Semantic Mining (SeMi) module first extracts human body-part-aware semantic cues, which are then employed to align 2D and 3D features via a Semantic-Guided Alignment (SGA) module within a unified embedding space. A Symmetric Cross-Attention Fusion (SCAF) module hierarchically integrates visual contours and 3D geometric features, and a Spatio-Temporal (ST) module captures global gait dynamics. Extensive experiments on various gait datasets validate the effectiveness of our method.

**Code** — https://github.com/O-VIGIA/LRGait.git

## 1 Introduction

In practical applications such as intelligent surveillance and remote identity verification, gait recognition has emerged as a promising biometric technique owing to its non-intrusive nature and robustness over long distances(Fan et al. 2023, 2025; Sepas-Moghaddam and Etemad 2022; Zhu et al. 2021; Zheng et al. 2022). Recent advances have demonstrated strong performance in controlled environments. *However, they are limited to short-range and unimodal settings, and are unexplored under long-range and multimodal conditions.*

Most publicly available gait datasets—such as CASIA-Series (Yu, Tan, and Tan 2006; Tan et al. 2006; Song et al. 2022), OU-MVLP (Takemura et al. 2018a), and Gait3D (Zheng et al. 2022)—are predominantly composed of RGB videos collected within 15 meters, thereby limiting their applicability to real-world surveillance scenarios that demand long-range, cross-distance recognition. Recently, SUSTech1K (Shen et al. 2023) introduced a large-scale LiDAR-Camera multimodal benchmark, laying the groundwork for multimodal gait analysis. FreeGait (Han et al. 2024) further advances this direction by capturing gait data in unconstrained outdoor environments. However, SUSTech1K is confined to ranges below 12 meters, and FreeGait includes samples at most 25 meters, underscoring the need for datasets enabling long-range (e.g., 50m) multimodal recognition. Moreover, these datasets lack cross-distance samples per identity, which impedes evaluation under cross-distance cross-view scenarios (e.g., 50m→10m). To address these limitations, we introduce LRGait, a long-range cross-distance multimodal gait dataset. It captures synchronized RGB and LiDAR gait sequences across five scopes and eight viewpoints (see Fig. 1), encompassing variations in illumination, weather, and carried objects. A comparative overview of existing gait datasets is provided in Table1.

Despite recent progress, effectively harnessing the complementary strengths of LiDAR and RGB modalities remains a significant challenge due to the intrinsic modality gap (Bai et al. 2025; Chen et al. 2025; Li et al. 2025; Fan et al. 2025). Although multimodal gait datasets have emerged, most approaches still rely on unimodal pipelines (Shen et al. 2023; Han et al. 2024; Sepas-Moghaddam and Etemad 2022). LiCAF (Deng, Xiong, and Feng 2024) introduces cross-attention for asymmetric fusion, yet its straightforward fusion strategy struggles to bridge the modality discrepancy. Furthermore, current methods (Han et al. 2024; Shen et al. 2023; Cui and Kang 2023) typically adopt depth maps from point clouds or RGB-based silhouettes as pretreatment inputs, resulting in the loss of fine-grained details. *This limitation is particularly exacerbated in long-range or nighttime scenarios, where sparse point clouds and blurred RGB images offer degenerate solutions during pretreatment.*

To address these challenges, we propose EMGaitNet, an end-to-end semantic-guided framework that directly exploits raw RGB videos and point cloud sequences for multimodal long-range gait recognition . Specifically, we propose a CLIP-based Semantic Mining (SeMi) module that extracts body-part-aware semantic cues. Furthermore, a Semantic-Guided Alignment (SGA) module is designed to reconstruct and align 2D3D cross-modal features by leveraging semantic features. Moreover, a Symmetric Cross-Attention Fusion (SCAF) module is devised to integrate 2D3D features through the cross-

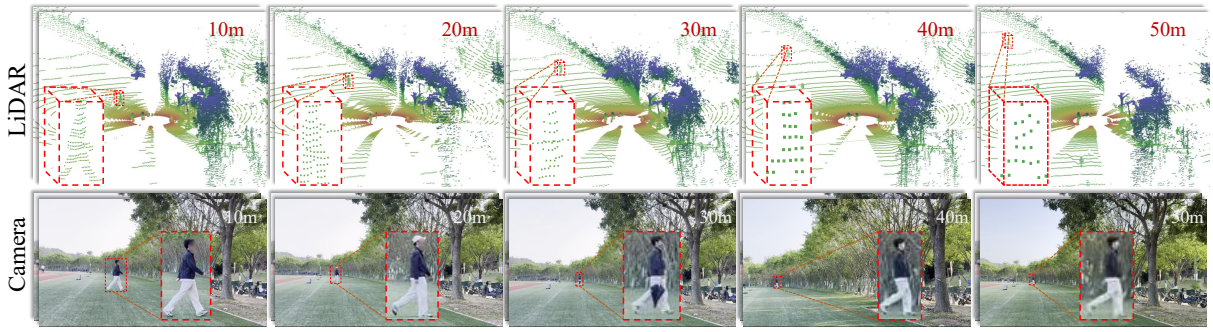



Figure 1: Visualization of our proposed multimodal dataset LRGait at various and long-range distances.

attention module hierarchically. Finally, a Spatio-Temporal (ST) module is employed to aggregate global gait dynamics across both spatial and temporal domains. Our contributions are as follows:

- We introduce **LRGait**, the first multimodal gait recognition dataset explicitly designed for long-range and cross-distance scenarios. It comprises synchronized LiDAR and RGB data captured at 5 scopes ($10m \sim 50m$) from 8 viewpoints. To simulate real-world situations, diverse appearances (e.g., carrying, attire) and environmental variables (e.g., illumination, weather) are incorporated.
- We propose **EMGaitNet**, an end-to-end framework that directly inputs raw RGB videos and point cloud sequences for multimodal gait recognition. EMGaitNet incorporates a CLIP-based Semantic Mining (SeMi) module to extract body-part-aware semantics, which guide cross-modal feature alignment in a Semantic-Guided Alignment (SGA) module. Furthermore, we design a Symmetric Cross-Attention Fusion (SCAF) module for deep hierarchical fusion, and a Spatio-Temporal (ST) module to capture global gait dynamics.

## 2 Related Works

### 2.1 Gait Recognition Methods

Gait recognition methods are broadly categorized into 2D and 3D representations. 2D appearance-based models (Chao et al. 2019; Fan et al. 2020; Huang et al. 2021; Liang et al. 2022) depend on silhouettes but are sensitive to segmentation quality and appearance changes. Model-based methods (Li et al. 2021; Liao et al. 2020; Teepe et al. 2021) capture gait via pose but suffer from pose estimation errors. Both paradigms lack 3D geometric cues critical for robustness in unconstrained settings. To address this, 3D-based methods have been proposed. Some reconstruct 3D meshes from RGB (Zheng et al. 2022), while others use LiDAR-derived depth or range images (Shen et al. 2023; Ahn et al. 2022), often losing geometric details during projection. Recent multimodal approaches (Deng, Xiong, and Feng 2024; Cui and Kang 2023) integrate 2D and 3D cues for improved performance, yet rely on pre-processed inputs, limiting end-to-end learning. In contrast, our EMGaitNet directly processes raw RGB and LiDAR data, leveraging semantic guidance for end-to-end cross-modal alignment and fusion.

### 2.2 Gait Recognition Benchmark

Gait datasets can be broadly categorized into three types: in-the-lab (Yu, Tan, and Tan 2006; Tan et al. 2006; Song et al. 2022; Iwama et al. 2012; Takemura et al. 2018b; Shen et al. 2023), synthetic (Dou et al. 2021), and outdoor/in-the-wild (Mu et al. 2021; Zheng et al. 2022; Han et al. 2024). In-the-lab datasets like the CASIA series support controlled evaluation but are limited to short ranges ($\leq$10m) and RGB modality, restricting their utility for data-driven multimodal models. SUSTech1K (Shen et al. 2023) addresses some of these issues by incorporating LiDAR and offering large-scale, multi-view, multimodal data, yet it remains constrained to 12m. Synthetic datasets ease annotation costs but face domain gaps and limited real-world applicability. Outdoor datasets such as FreeGait(Han et al. 2024) enable evaluation in unconstrained conditions, yet their maximum range remains below 25m. We argue that gait recognition is theoretically feasible at ranges beyond 50m. To this end, we introduce LRGait, a long-range, multimodal dataset designed to explore and push the boundaries of distance-aware gait recognition.

## 3 Long-Range Gait Benchmark

### 3.1 Overall

The LRGait dataset is collected using a mobile robot equipped with a 128-beam LiDAR and a monocular RGB camera, capturing synchronized multimodal data. It comprises 5,280 gait sequences from 101 subjects (79 males and 22 females), totaling over 209,000 frames of raw point clouds and RGB images, along with corresponding depth maps and silhouettes. Each subject is recorded walking at distances ranging from 10m to 50m, enabling the study of gait recognition across varying ranges. To simulate various illuminations, 31 subjects were recorded under both day and night environments. To ensure ethical data collection, all participants provided informed consent, and facial regions in RGB images were anonymized via blurring.

### 3.2 Data Collection

The LRGait dataset was collected over a four-week period across three diverse outdoor scenes, covering four distinct weather conditions: sunny, cloudy, overcast, and rainy. We employed an industrial-grade RGB camera and a 128-beam Ouster LiDAR sensor to capture synchronized video and

| Dataset | Sensor | Viewpoint | Distance | Outdoor | LR | CD | D&N |
|---|---|---|---|---|---|---|---|
| CASIA-B (Yu, Tan, and Tan 2006) | Camera | 11 | 2m ∼ 4m | ✖ | ✖ | ✖ | ✖ |
| CASIA-C (Tan et al. 2006) | Camera | 1 | N/A | ✔ | ✖ | ✖ | ✖ |
| TUM-GAID (Hofmann et al. 2014) | RGB-D | 1 | 3.6m | ✖ | ✖ | ✖ | ✖ |
| SZTAKI-LGA (Benedek et al. 2016) | LiDAR | 1 | N/A | ✔ | ✖ | ✖ | ✖ |
| OU-MVLP (Takemura et al. 2018a) | Camera | 14 | 8m | ✖ | ✖ | ✖ | ✖ |
| GREW (Zhu et al. 2021) | Camera | 882 | N/A | ✔ | ✖ | ✖ | ✖ |
| Gait3D (Zheng et al. 2022) | Camera | 39 | N/A | ✔ | ✖ | ✖ | ✖ |
| CASIA-E (Song et al. 2022) | Camera | 26 | 8m | ✖ | ✖ | ✖ | ✖ |
| CCPG (Li et al. 2023) | Camera | 10 | N/A | ✖ | ✖ | ✖ | ✖ |
| SUSTech1K (Shen et al. 2023) | LiDAR&Camera | 12 | 8m ∼ 12m | ✖ | ✖ | ✖ | ✔ |
| FreeGait (Han et al. 2024) | LiDAR&Camera | 3 | 25m | ✔ | ✖ | ✖ | ✔ |
| LRGait(Ours) | LiDAR&Camera | 8 | 10/20/30/40/50m | ✔ | ✔ | ✔ | ✔ |

Table 1: Comparison with public datasets for gait recognition, where "LR" refers to the long-range distance exceeding 30m, "CD" denotes cross-distance retrieval, and "D&N" represents day and night during data collection.

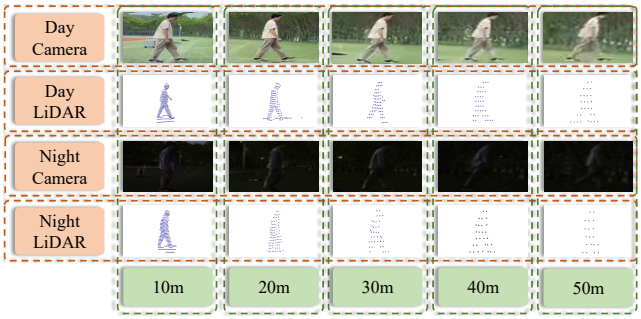
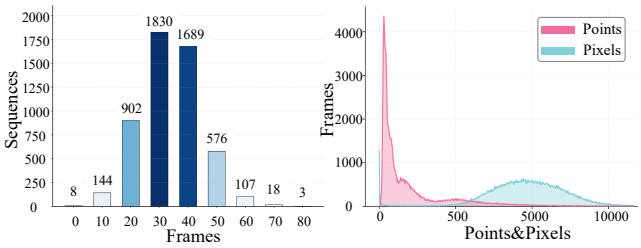



Figure 2: Visualizations under daytime and nighttime.




(a) Sequences over frames (b) Frames over points/pixels

Figure 3: Statistics of the proposed LRGait.

point cloud sequences at 30Hz and 10Hz, respectively. The modalities were temporally aligned during post-processing to ensure accurate cross-modal correspondence. During data collection, each subject was instructed to walk within five various scopes (10, 20, 30, 40, and 50m). To mimic real-world variations, participants were randomly assigned to carry props such as suitcases, umbrellas, and hats. Additionally, for a subset of participants, gait data were recorded under both daytime and nighttime conditions to support research in all-day gait recognition, as shown in Fig. 2. To enable multi-view analysis, gait sequences were captured from eight different viewpoints for each walking distance, facilitating robust cross-view and cross-distance gait recognition.

### 3.3 Annotations and Representations

For the camera-based data, we first employed 2D object detection(Ge et al. 2021) and tracking(Zhang et al. 2022) models to extract gait sequences, followed by silhouette generation using a segmentation model(Xie et al. 2021). In challenging conditions such as nighttime or at long distances ($\geq$40m), where the targets are small and visually degraded, we manually annotated 500 frames to fine-tune the detector. For the LiDAR-based data, existing 3D detection frameworks struggle to localize pedestrians reliably at long ranges due to the sparsity of the point cloud. To overcome this limitation, we manually labeled 4,500 frames with pedestrian bounding boxes across various scenes and distances, and used them to train a 3D detection model(Lang et al. 2019) tailored to long-range scenarios. To address inaccuracies such as false positives and missed detections, we manually corrected the corresponding frames. Upon completion, the entire dataset was thoroughly verified by three expert annotators.

### 3.4 Statistics and Evaluation Metrics

Fig. 3 illustrates the data distribution of the LRGait dataset. The evaluation protocol adopts the cross-view recognition setting, following the standard used in SUSTech1K(Shen et al. 2023), in which probe features are matched against gallery features across various views. Probe sets are partitioned by distance to assess attribute influence in cross-view retrieval. Rank-1 and Rank-5 accuracies are used as evaluation metrics. Refer to the supplementary materials for statistical details.

## 4 Semantic-Guided Multimodal Gait Recognition

### 4.1 Problem Definition

We design an end-to-end multimodal gait recognition framework that directly takes RGB video

$$I = \left\{ I_i^j | i = 1, 2, \ldots, m; j = 1, 2, \ldots, n \right\} \quad (1)$$

and point cloud sequences

$$P = \left\{ P_i^j | i = 1, 2, \ldots, m; j = 1, 2, \ldots, n \right\} \quad (2)$$

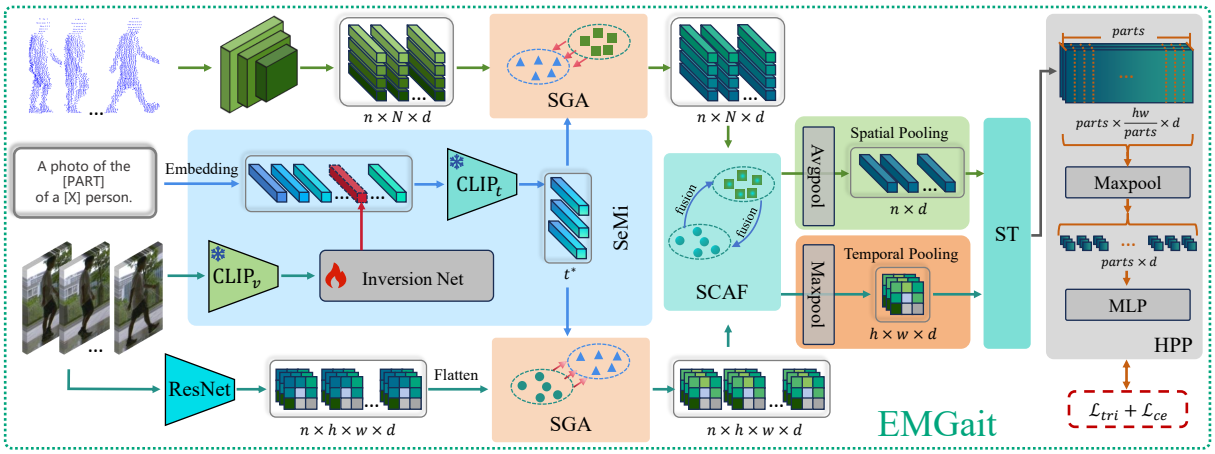



Figure 4: Illustration of the proposed framework.

as input, where $m$ is the number of gait sequences and $n$ is the number of frames per sequence. Specifically, $I_i^j \in \mathbb{R}^{H\times W\times 3}$ denotes an RGB frame with height $H$ and width $W$, and $P_i^j \in \mathbb{R}^{N\times 3}$ is the corresponding LiDAR point cloud containing $N$ number of points. Our goal is to learn discriminative multimodal gait representations directly from the raw RGB and point cloud data:

$$F = G_\theta\left(I, P\right), \tag{3}$$

where $F = \{F_i | i = 1, 2, \ldots, m\}$ denotes the learned gait features, and $G_\theta$ is our proposed multimodal framework.

## 4.2 Feature Extraction

For the camera modality, we adopt the lightweight ResNet9 backbone from OpenGait(Fan et al. 2023) to extract frame-wise visual features, formulated as:

$$F_{i,j}^{2d} = \text{ResNet}_9\left(I_i^j\right), \tag{4}$$

where $F_{i,j}^{2d} \in \mathbb{R}^{h\times w\times d}$ denotes the spatial feature map for the $j$-th frame in the $i$-th sequence, with $h$, $w$, and $d$ being the height, width, and channel dimensions, respectively. For the LiDAR modality, we adopt a PointGNN-based backbone(Shi and Rajkumar 2020) to extract discriminative 3D features, which mitigates the adverse effects of point cloud sparsity. Specifically, given a point cloud sequence $P_i^j$ and its corresponding feature representation $F_{i,j}^{3d}$ (initialized as coordinates), we first construct a local neighborhood for each point:

$$A_{i,j}[k] = \left\{P_i^j[k], \mathcal{N}_{P_i^j}\left(P_i^j[k]\right)\right\}, \tag{5}$$

where $k \in \{1, 2, \ldots, N\}$, $P_i^j[k]$ denotes the $k$-th point, and $\mathcal{N}_{P_i^j}\left(P_i^j[k]\right)$ retrieves its spatial neighbors in $P_i^j$. The adjacency relationship is defined based on the cosine similarity between the feature vectors of points, defined as:

$$\cos(F_{i,j}^{3d}[k], F_{i,j}^{3d}[u]) = \frac{F_{i,j}^{3d}[k] \cdot F_{i,j}^{3d}[u]}{\|F_{i,j}^{3d}[k]\|_2 \cdot \|F_{i,j}^{3d}[u]\|_2}, \tag{6}$$

where $F_{i,j}^{3d}[k] \in \mathbb{R}^d$ denotes the feature vector of $P_i^j[k]$. Using the similarity scores, we construct a local graph by selecting the TopK most similar points from the entire point set as neighbors:

$$\mathcal{N}_{P_i^j}(P_i^j[k]) = \underset{u\neq k}{\text{TopK}}\left(\cos(F_{i,j}^{3d}[k], F_{i,j}^{3d}[u])\right). \tag{7}$$

We compute the edge features in the local graph as:

$$e_{k,u} = \text{Concat}\left[P_i^j[u] - P_i^j[k], F_{i,j}^{3d}[k], F_{i,j}^{3d}[u]\right]. \tag{8}$$

Here, Concat denotes the concatenation operation along the feature dimension. Subsequently, a multi-layer perceptron (MLP) layer is applied to introduce nonlinearity, followed by feature aggregation within each local neighborhood to update the point features from the previous layer:

$$F_{i,j}^{3d}[k] = \underset{u\in\mathcal{N}_{P_i^j}\left(P_i^j[k]\right)}{\text{Maxpool}}\left(\text{MLP}\left(e_{k,u}\right)\right). \tag{9}$$

By stacking multiple such graph convolution layers, the model progressively captures both local and global geometric patterns, yielding the final representation $F_{i,j}^{3d} \in \mathbb{R}^{N\times d}$. Our overall framework is illustrated in Fig. 4.

## 4.3 CLIP-Based Semantic Mining

The modality gap between 2D images and 3D point clouds frequently results in suboptimal fusion during feature integration. Moreover, background noise introduces ineffective fusion, exemplified by the blending of LiDAR point clouds with irrelevant RGB background areas. To address this, we propose a CLIP-based Semantic Mining (SeMi) module that leverages semantic cues to align cross-modal features and enhance regional correspondence. Specifically, we construct explicit body-part prompts and feed them into the CLIP text encoder CLIP$_t$ to obtain semantic cues. A general template—"A photo of the [PART] of a [X] person"—is instantiated with body-part terms, using a predefined list ["head", "arms", "torso", "legs", "feet"] to generate fine-grained semantic descriptions, which are then tokenized into embeddings $t \in \mathbb{R}^{5\times l\times d}$. However, these class-level semantics fall

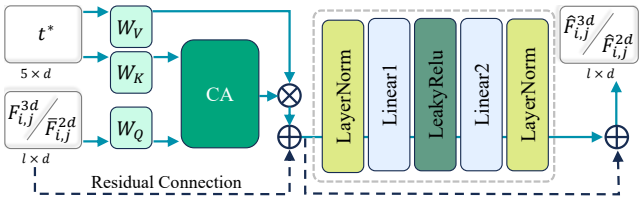



Figure 5: Details of the proposed SGA module.

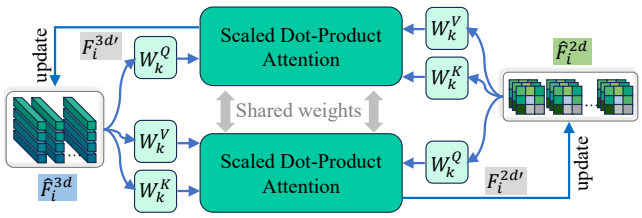



Figure 6: Details of the proposed SCAF module.

short in capturing instance-level nuances for gait recognition. Inspired by PromptSG(Yang et al. 2024), we replace [X] with visual embeddings derived from the inversion net to generate identity-aware descriptions. Specifically, for an input image $I_i^j$, we utilize CLIP's visual encoder to extract the global visual embedding:

$$v = \text{CLIP}_v\left(I_i^j\right), \tag{10}$$

where $v \in \mathbb{R}^{1\times d}$. We then employ an inversion network to map the visual feature $v$ from visual space into the text space, formulated as: $v^* = F_{inv}(v)$. The transformed feature $v^*$ is leveraged as a pseudo token to replace the [X] placeholder in the prompt for fine-grained and identity-aware semantic token embedding. The modified prompts are then fed into the text encoder $\text{CLIP}_t$ to extract body-part-aware semantic features, denoted as $t^* \in \mathbb{R}^{5\times d}$.

### 4.4 Semantic-Guided Alignment Module

We leverage multi-grained semantic cues as an intermediate bridge to mitigate the modality gap between RGB images and point cloud data, while simultaneously suppressing feature noise, as shown in Fig. 5. Specifically, we design the feature alignment module based on a cross-attention mechanism. Given a 2D feature map $F_{i,j}^{2d}$, we flatten its spatial dimensions to obtain $\bar{F}_{i,j}^{2d} \in \mathbb{R}^{hw\times d}$ that match the format of the 3D feature $F_{i,j}^{3d}$. Taking $\bar{F}_{i,j}^{2d}$ as the query and the semantic features $t^*$ as both key and value, we perform cross-attention fusion, formally defined as:

$$\text{CA}\left(\bar{F}_{i,j}^{2d}, t^*\right) = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V, \tag{11}$$

where $Q = \bar{F}_{i,j}^{2d}W_Q, K = t^*W_K, V = t^*W_V$, with $W_Q, W_K, W_V \in \mathbb{R}^{d\times d}$ denoting learnable projection matrices. Furthermore, we integrate residual connections and a Feed Forward Network (FFN) following the attention output:

$$\tilde{F}_{i,j}^{2d} = \text{LayerNorm}\left(\bar{F}_{i,j}^{2d} + \text{CA}(\bar{F}_{i,j}^{2d}, t^*)\right), \tag{12}$$

$$\hat{F}_{i,j}^{2d} = \text{LayerNorm}\left(\tilde{F}_{i,j}^{2d} + \text{FFN}(\tilde{F}_{i,j}^{2d})\right). \tag{13}$$

Here, $\text{FFN}(\cdot)$ consists of two MLP layers with LeakyReLU activation, and the refined 2D features are denoted by $\hat{F}_{i,j}^{2d} \in \mathbb{R}^{hw\times d}$. Similarly, we employ semantic features $t^*$ to align the point cloud features, yielding the refined $\hat{F}_{i,j}^{3d} \in \mathbb{R}^{N\times d}$.

### 4.5 Symmetric Cross-Attention Fusion Module

After cross-modal alignment, the features from different modalities are projected into a shared latent space, where a symmetric multi-head cross-attention is applied to integrate complementary information, as presented in Fig. 6. Specifically, we construct a symmetric dual-stream module, where the refined image features $\hat{F}_{i,j}^{2d}$ and point cloud features $\hat{F}_{i,j}^{3d}$ alternately serve as query and attend to each other as key and value, enabling bidirectional alignment and mutual information fusion. For each fusion layer, the attention-based update of the image stream is computed as:

$$F_{i,j}^{2d'} = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O, \tag{14}$$

where $h$ denotes the number of attention heads, and $W^O \in \mathbb{R}^{hd_h\times d}$ represents the output projection matrix. The operation for each attention head is defined as follows:

$$\text{head}_k = \text{Softmax}\left(\frac{Q_kK_k^\top}{\sqrt{d_h}}\right)V_k, \tag{15}$$

$$Q_k = F_{i,j}^{2d}W_k^Q, \quad K_k = F_{i,j}^{3d}W_k^K, \quad V_k = F_{i,j}^{3d}W_k^V, \tag{16}$$

where $W_k^Q, W_k^K, W_k^V \in \mathbb{R}^{d\times d_h}$ are learnable projection matrices. In parallel, the point cloud stream is updated analogously with shared weights, yielding $F_{i,j}^{3d'}$. The refined features are subsequently propagated to the next stage of hierarchical fusion by updating $\hat{F}_{i,j}^{2d} = F_{i,j}^{2d'}$ and $\hat{F}_{i,j}^{3d} = F_{i,j}^{3d'}$.

### 4.6 Spatial Temporal Module

To capture temporal dynamics, we apply Temporal Pooling over the sequence of the fused sequence features $\hat{F}_i^{2d} = \{\hat{F}_{i,j}^{2d}\}_{j=1}^n$, where $\hat{F}_i^{2d} \in \mathbb{R}^{n\times h\times w\times d}$. Meanwhile, Spatial Pooling is performed on the point cloud feature sequence $\hat{F}_i^{3d} = \{\hat{F}_{i,j}^{3d}\}_{j=1}^n$ to obtain global spatial features, where $\hat{F}_i^{3d} \in \mathbb{R}^{n\times N\times d}$. These operations are formulated as:

$$F_i^{tp} = \underset{n}{\text{Maxpool}}\left(\hat{F}_i^{2d}\right), \tag{17}$$

$$F_i^{sp} = \underset{N}{\text{Avgpool}}\left(\hat{F}_i^{3d}\right), \tag{18}$$

where $F_i^{tp} \in \mathbb{R}^{h\times w\times d}$ represents the aggregated temporal feature, and $F_i^{sp} \in \mathbb{R}^{n\times d}$ denotes the global spatial feature. We leverage the CA to integrate the spatio-temporal features:

$$\tilde{F}_i^{sp} = \text{CA}\left(F_i^{sp}, F_i^{tp}\right) + F_i^{sp}, \tag{19}$$

$$F_i^{fusion} = \text{MLP}\left(\text{CA}\left(F_i^{tp}, \tilde{F}_i^{sp}\right) + F_i^{tp}\right), \tag{20}$$

where $F_i^{fusion} \in \mathbb{R}^{h \times w \times d}$. Subsequently, the Horizontal Pyramid Pooling (HPP) module(Fan et al. 2025) is employed to perform part-based matching.

## 5 Training and Inference

Following prior works(Fan et al. 2023, 2025), we optimize the model using a combination of triplet loss and cross-entropy loss. The objective functions are formulated as:

$$\mathcal{L} = \alpha \mathcal{L}_{tri} + \beta \mathcal{L}_{ce}, \tag{21}$$

where $\alpha = 1.0, \beta = 2.0$ in default. During inference, the L2 Euclidean distance between features is used to measure the similarity between the probe and gallery samples.

## 6 Experiments

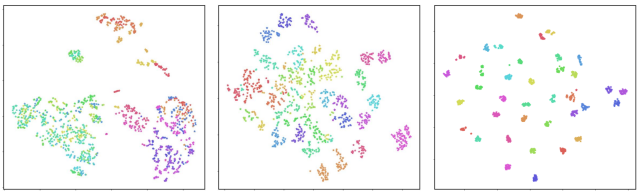
### 6.1 Implementation Details

In SUSTech1K(Shen et al. 2023), we apply Farthest Point Sampling (FPS) to downsample each point cloud (pc) frame to 512 points to facilitate end-to-end model input. For LRGait, following the same protocol, RGB images (rgb) and silhouettes (sil) are resized to 64×64. Each point cloud frame is downsampled to 256 points via FPS, and the corresponding depth projections (depth) are also resized to 64×64. In FreeGait, we follow the same preprocessing as HMRNet(Han et al. 2024). The model is trained for 40,000 epochs using the Adam optimizer with a weight decay of 0.0005. The initial learning rate is set to 0.0003 for the SUSTech1K and FreeGait datasets, and 0.00005 for the LRGait dataset. A MultiStepLR scheduler is employed to decay the learning rate by a factor of 0.1 at the 15,000th and 30,000th epochs. During each training epoch, 10 RGB frames and their corresponding point cloud frames are randomly sampled for end-to-end training. All experiments are implemented and evaluated on two NVIDIA RTX 3090 GPUs.

### 6.2 Results and Analysis

We compare our approach against advanced unimodal and multimodal methods. The proposed EMGait achieves superior performance across all benchmarks, establishing new state-of-the-art results not only on the long-range and cross-distance LRGait dataset but also on the large-scale, multi-view, and multi-variable SUSTech1K, as well as the highly challenging FreeGait dataset.

**SUSTech1K** Table 2 provides a detailed comparison of the SUSTech1K dataset. Specifically, the rgb and pc correspond to the raw inputs captured by the camera and LiDAR, respectively. The sil and depth indicate the preprocessed representations obtained from RGB and point cloud inputs(Fan et al. 2023; Shen et al. 2025). Meanwhile, ps and fl denote higher-level semantic cues extracted from rgb, representing human parsing and optical flow(Jin et al. 2025). On the Overall Rank-1 metric, our EMGaitNet achieves an accuracy of 96.0%, establishing a new state-of-the-art (SOTA) by outperforming the second-best method by 2.2%. This highlights EMGaitNet's capability to not only capture 2D cues such as human shape and contours, but also extract 3D representations involving body dimensions and skeletal topology. Compared to the previous SOTA unimodal end-to-end approach LidarGait++(Shen et al. 2025), EMGaitNet improves Overall Rank-1 by 3.6%. Remarkably, it reaches 99.6% accuracy under occlusion conditions, evidencing its robust understanding of multimodal priors such as body structure and shape.



(a) LidarGait++ (b) Gaitbase (c) Ours

Figure 7: Comparison of t-SNE visualizations.

**LRGait** We conduct a comparative study of multiple approaches on our proposed LRGait dataset, with the day-10m (D-10) setting designated as the gallery. EMGaitNet delivers SOTA performance overall and cross-distance retrieval under day conditions, as presented in Table 3. Notably, it reaches 74.4% accuracy at a distance of day-50m (D-50), outperforming the second-best method by a significant margin of 14.0%. These results affirm the effectiveness of our end-to-end multimodal framework in integrating the 3D geometric structures from point clouds with the semantic cues from RGB images, enabling robust performance across long-range and cross-distance. Fig. 7 shows advancement of our EMGaitNet in the t-SNE (Maaten and Hinton 2008) manner. In contrast, under night (N) conditions, the performance of nearly all methods deteriorates markedly, largely due to the severe domain shift between day and night. Consequently, developing multimodal domain-adaptive approaches for cross-distance gait recognition under day-night conditions is a crucial and promising research direction.

**FreeGait** Table 4 presents the results on FreeGait. Our EMGaitNet outperforms LiDARGait++(Shen et al. 2025) by 3.9%, 3.4%, and 2.1% on Rank-1, Rank-5, and mAP, respectively, demonstrating its ability to learn highly discriminative and robust multimodal gait features.

### 6.3 Ablation Study

To validate the effectiveness of each component in our proposed EMGaitNet, we conduct comprehensive ablation studies on the LRGait dataset, as shown in Table 5. Starting from a baseline that directly fuses RGB and point cloud features, we observe a Rank-1 accuracy of 52.3%, indicating the inherent challenge in multimodal fusion under long-range conditions. Introducing the Semantic-Guided Alignment (SGA) module significantly boosts accuracy to 58.5%, highlighting its critical role in mitigating modality gaps between 2D and 3D features. Further integrating the CLIP-based Semantic Mining (SeMi) module improves performance to 64.2%, demonstrating the importance of human semantic cues in guiding feature representation. Finally, the inclusion of the Spatio-Temporal (ST) module increases the Rank-1 accuracy to 68.9%, demonstrating the benefit of modeling global gait dynamics.

| Methods | Modality | Probe Sequence (Rank-1 Accuracy %) | | | | | | | | Overall | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | NM | BG | CL | CR | UB | UN | OC | NT | R-1 | R-5 |
| GaitSet(Chao et al. 2019) | sil | 69.1 | 68.3 | 37.4 | 65.0 | 63.1 | 67.2 | 61.0 | 23.0 | 65.0 | 84.8 |
| GaitBase(Fan et al. 2023) | sil | 81.3 | 77.3 | 49.6 | 75.7 | 75.4 | 76.7 | 81.4 | 25.8 | 76.0 | 89.1 |
| SimpleView(Goyal et al. 2021) | depth | 72.3 | 68.8 | 57.2 | 63.3 | 49.2 | 79.7 | 62.5 | 66.5 | 64.8 | 85.8 |
| LidarGait(Shen et al. 2023) | depth | 91.8 | 88.6 | 74.6 | 89.0 | 67.5 | 80.9 | 94.5 | 90.4 | 86.8 | 96.1 |
| PointTransformer(Zhao et al. 2021) | pc | 53.2 | 48.1 | 32.0 | 43.2 | 39.1 | 47.9 | 41.8 | 47.1 | 44.4 | 76.7 |
| PointNet++(Qi et al. 2017) | pc | 82.5 | 78.7 | 58.7 | 76.1 | 74.0 | 85.4 | 75.8 | 74.8 | 77.1 | 94.1 |
| LidarGait++(Shen et al. 2025) | pc | 94.2 | 93.9 | 79.7 | 92.4 | 91.5 | **96.6** | 91.9 | **92.2** | 92.7 | 98.2 |
| HMRNet(Han et al. 2024) | pc+depth | 92.7 | 92.3 | 79.6 | 90.3 | 83.1 | 95.2 | 86.2 | 90.4 | 90.2 | 97.5 |
| MMGaitFormer(Cui and Kang 2023) | depth+sil | 94.3 | 93.7 | 80.0 | 91.8 | 84.0 | 88.7 | 95.7 | 86.0 | 91.1 | 98.2 |
| LiCAF(Deng, Xiong, and Feng 2024) | depth+sil | 95.8 | 95.7 | **82.7** | 94.5 | 89.3 | 93.6 | 96.6 | 88.7 | 93.9 | 98.8 |
| MultiGait++(Jin et al. 2025) | sil+ps+fl | 92.0 | 89.4 | 50.4 | 87.6 | 89.7 | 89.1 | 93.4 | 45.1 | 87.4 | 95.6 |
| EMGaitNet(Ours) | pc+rgb | **98.2** | **96.4** | 81.7 | **96.2** | **94.9** | 93.6 | **99.6** | 88.1 | **96.0** | **99.0** |

Table 2: Comparison of cross-view Rank-1 accuracy under various conditions on the SUSTech1K dataset. The best results are highlighted in **bold**, while the second-best entries are underlined.

| Methods | Modality | Probe Sequence (Rank-1 Accuracy %) | | | | | | | | Overall | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | D-20 | D-30 | D-40 | D-50 | N-20 | N-30 | N-40 | N-50 | R-1 | R-5 |
| GaitBase(Fan et al. 2023) | sil | 67.9 | 53.9 | 48.5 | 33.8 | 41.6 | **33.4** | 19.8 | 13.2 | 46.8 | 77.8 |
| LidarGait(Shen et al. 2023) | depth | 26.1 | 14.6 | 13.2 | 10.8 | 18.5 | 14.6 | 10.6 | 9.6 | 15.7 | 51.3 |
| LidarGait++(Shen et al. 2025) | pc | 32.1 | 24.4 | 18.8 | 12.6 | 23.1 | 13.2 | 11.7 | 11.2 | 20.9 | 59.0 |
| HMRNet(Han et al. 2024) | pc+depth | 58.7 | 53.1 | 50.3 | 44.8 | 22.5 | 19.3 | 11.8 | 8.6 | 41.1 | 71.7 |
| MMGaitFormer(Cui and Kang 2023) | depth+sil | 72.4 | 70.2 | 58.1 | 62.7 | 40.3 | 26.8 | 15.5 | 7.6 | 57.1 | 80.4 |
| LiCAF(Deng, Xiong, and Feng 2024) | depth+sil | 74.8 | 71.6 | 60.4 | 65.3 | **42.5** | 27.8 | 14.6 | 9.9 | 59.6 | 82.9 |
| EMGaitNet (Ours) | pc+rgb | **88.5** | **82.4** | **80.8** | **74.4** | 38.2 | 31.7 | **21.9** | **17.1** | **68.9** | **85.8** |

Table 3: Comparison of cross-distance cross-view Rank-1 accuracy under various settings on the proposed LRGait dataset.

| Methods | Modality | R-1 | R-5 | mAP |
|---|---|---|---|---|
| GaitSet(Chao et al. 2019) | sil | 57.1 | 71.9 | 64.0 |
| GaitBase(Fan et al. 2023) | sil | 62.6 | 75.3 | 68.6 |
| LidarGait(Shen et al. 2023) | depth | 74.2 | 88.8 | 80.7 |
| PointNet++(Qi et al. 2017) | pc | 59.3 | 81.2 | 69.3 |
| LidarGait++(Shen et al. 2025) | pc | 82.0 | 93.6 | 87.2 |
| HMRNet(Han et al. 2024) | pc+depth | 80.8 | 93.6 | 86.5 |
| EMGaitNet(Ours) | pc+rgb | **85.2** | **96.8** | **89.0** |

Table 4: Comparisons with SOTA methods on FreeGait.

| Baseline | SeMi | SGA | ST | Overall R-1 | Overall R-5 |
|---|---|---|---|---|---|
| ✓ | | | | 52.3 | 70.2 |
| ✓ | | ✓ | | 58.5 | 75.9 |
| ✓ | ✓ | ✓ | | 64.2 | 80.7 |
| ✓ | ✓ | ✓ | ✓ | **68.9** | **85.8** |

Table 5: Ablation studies on the LRGait dataset.

## 7 Conclusion and Future Work

In this paper, we introduce LRGait, the first large-scale multi-modal gait dataset designed for long-range and cross-distance scenarios, along with EMGaitNet, a semantic-guided, end-to-end multimodal fusion framework. LRGait comprises 101 subjects, with gait sequences captured across distances ranging from 10m to 50m, under diverse lighting, weather, and environmental conditions. Compared with existing public gait datasets—none of which exceed 25m in collection range—LRGait significantly expands the distance frontier, with the ambition of advancing effective gait recognition beyond 50m. Our proposed EMGaitNet achieves SOTA performance on LRGait, SUSTech1K, and FreeGait, demonstrating the effectiveness of leveraging both raw RGB videos and LiDAR point clouds in a unified end-to-end pipeline.

While our method performs robustly under daytime conditions on LRGait—achieving over 74% Rank-1 accuracy within the scope of 50m—it still faces challenges in nighttime scenarios, leaving substantial room for future improvement. We hope our contributions inspire further research toward more scalable and resilient multimodal gait recognition systems, ultimately enabling gait analysis to go farther, across longer distances and more diverse conditions.

## References


Ahn, J.; Nakashima, K.; Yoshino, K.; Iwashita, Y.; and Kurazume, R. 2022. 2V-Gait: Gait recognition using 3D LiDAR robust to changes in walking direction and measurement distance. In *IEEE/SICE International Symposium on System Integration*, 602–607. IEEE.

Bai, Y.; Ji, Y.; Cao, M.; Wang, J.; and Ye, M. 2025. Chat-based Person Retrieval via Dialogue-Refined Cross-Modal Alignment. In *PProceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3952–3962.

Benedek, C.; Gálai, B.; Nagy, B.; and Jankó, Z. 2016. Lidar-based gait analysis and activity recognition in a 4d surveillance system. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(1): 101–113.

Chao, H.; He, Y.; Zhang, J.; and Feng, J. 2019. Gaitset: Regarding gait as a set for cross-view gait recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 8126–8133.

Chen, W.; Liu, Y.; Chen, B.; Su, J.; Zheng, Y.; and Lin, L. 2025. Cross-modal causal relation alignment for video question grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24087–24096.

Cui, Y.; and Kang, Y. 2023. Multi-modal gait recognition via effective spatial-temporal feature fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17949–17957.

Deng, Y.; Xiong, H.; and Feng, B. 2024. Licaf: Lidar-Camera Asymmetric Fusion For Gait Recognition. In *IEEE International Conference on Image Processing*, 2424–2430. IEEE.

Dou, H.; Zhang, W.; Zhang, P.; Zhao, Y.; Li, S.; Qin, Z.; Wu, F.; Dong, L.; and Li, X. 2021. Versatilegait: a large-scale synthetic gait dataset with fine-grainedattributes and complicated scenarios. *arXiv preprint arXiv:2101.01394*.

Fan, C.; Hou, S.; Liang, J.; Shen, C.; Ma, J.; Jin, D.; Huang, Y.; and Yu, S. 2025. OpenGait: A Comprehensive Benchmark Study for Gait Recognition towards Better Practicality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Fan, C.; Liang, J.; Shen, C.; Hou, S.; Huang, Y.; and Yu, S. 2023. Opengait: Revisiting gait recognition towards better practicality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9707–9716.

Fan, C.; Peng, Y.; Cao, C.; Liu, X.; Hou, S.; Chi, J.; Huang, Y.; Li, Q.; and He, Z. 2020. Gaitpart: Temporal part-based model for gait recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14225–14233.

Ge, Z.; Liu, S.; Wang, F.; Li, Z.; and Sun, J. 2021. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*.

Goyal, A.; Law, H.; Liu, B.; Newell, A.; and Deng, J. 2021. Revisiting point cloud shape classification with a simple and effective baseline. In *International Conference on Machine Learning*, 3809–3820. PMLR.

Han, X.; Ren, Y.; Cong, P.; Sun, Y.; Wang, J.; Xu, L.; and Ma, Y. 2024. Gait Recognition in Large-scale Free Environment via Single LiDAR. In *Proceedings of the ACM International Conference on Multimedia*, 380–389.

Hofmann, M.; Geiger, J.; Bachmann, S.; Schuller, B.; and Rigoll, G. 2014. The TUM Gait from Audio, Image and Depth (GAID) database: Multimodal recognition of subjects and traits. *Journal of Visual Communication and Image Representation*, 25(1): 195–206.

Huang, X.; Zhu, D.; Wang, H.; Wang, X.; Yang, B.; He, B.; Liu, W.; and Feng, B. 2021. Context-sensitive temporal feature learning for gait recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 12909–12918.

Iwama, H.; Okumura, M.; Makihara, Y.; and Yagi, Y. 2012. The OU-ISIR gait database comprising the large population dataset and performance evaluation of gait recognition. *IEEE Transactions on Information Forensics and Security*, 7(5): 1511–1521.

Jin, D.; Fan, C.; Chen, W.; and Yu, S. 2025. Exploring more from multiple gait modalities for human identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 4120–4128.

Lang, A. H.; Vora, S.; Caesar, H.; Zhou, L.; Yang, J.; and Beijbom, O. 2019. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12697–12705.

Li, W.; Hou, S.; Zhang, C.; Cao, C.; Liu, X.; Huang, Y.; and Zhao, Y. 2023. 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*, 13824–13833.

Li, X.; Makihara, Y.; Xu, C.; and Yagi, Y. 2021. End-to-end model-based gait recognition using synchronized multi-view pose constraint. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4106–4115.

Li, Y.; Xing, Y.; Lan, X.; Li, X.; Chen, H.; and Jiang, D. 2025. AlignMamba: Enhancing Multimodal Mamba with Local and Global Cross-modal Alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24774–24784.

Liang, J.; Fan, C.; Hou, S.; Shen, C.; Huang, Y.; and Yu, S. 2022. Gaitedge: Beyond plain end-to-end gait recognition for better practicality. In *European Conference on Computer Vision*, 375–390. Springer.

Liao, R.; Yu, S.; An, W.; and Huang, Y. 2020. A model-based gait recognition method with body pose and human prior knowledge. *Pattern Recognition*, 98: 107069.

Maaten, L. v. d.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(Nov): 2579–2605.

Mu, Z.; Castro, F. M.; Marin-Jimenez, M. J.; Guil, N.; Li, Y.-R.; and Yu, S. 2021. ReSGait: The real-scene gait dataset. In *IEEE International Joint Conference on Biometrics*, 1–8. IEEE.

Qi, C. R.; Yi, L.; Su, H.; and Guibas, L. J. 2017. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in Neural Information Processing Systems*, 30.

Sepas-Moghaddam, A.; and Etemad, A. 2022. Deep gait recognition: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1): 264–284.

Shen, C.; Fan, C.; Wu, W.; Wang, R.; Huang, G. Q.; and Yu, S. 2023. Lidargait: Benchmarking 3d gait recognition with point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1054–1063.

Shen, C.; Wang, R.; Duan, L.; and Yu, S. 2025. LidarGait++: Learning Local Features and Size Awareness from LiDAR Point Clouds for 3D Gait Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6627–6636.

Shi, W.; and Rajkumar, R. 2020. Point-gnn: Graph neural network for 3d object detection in a point cloud. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1711–1719.

Song, C.; Huang, Y.; Wang, W.; and Wang, L. 2022. CASIA-E: A large comprehensive dataset for gait recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3): 2801–2815.

Takemura, N.; Makihara, Y.; Muramatsu, D.; Echigo, T.; and Yagi, Y. 2018a. Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition. *IPSJ Transactions on Computer Vision and Applications*, 10(1): 4.

Takemura, N.; Makihara, Y.; Muramatsu, D.; Echigo, T.; and Yagi, Y. 2018b. Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition. *IPSJ Transactions on Computer Vision and Applications*, 10(1): 4.

Tan, D.; Huang, K.; Yu, S.; and Tan, T. 2006. Efficient night gait recognition based on template matching. In *International Conference on Pattern Recognition*, volume 3, 1000–1003. IEEE.

Teepe, T.; Khan, A.; Gilg, J.; Herzog, F.; Hörmann, S.; and Rigoll, G. 2021. Gaitgraph: Graph convolutional network for skeleton-based gait recognition. In *IEEE International Conference on Image Processing*, 2314–2318. IEEE.

Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J. M.; and Luo, P. 2021. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34: 12077–12090.

Yang, Z.; Wu, D.; Wu, C.; Lin, Z.; Gu, J.; and Wang, W. 2024. A pedestrian is worth one prompt: Towards language guidance person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17343–17353.

Yu, S.; Tan, D.; and Tan, T. 2006. A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In *International Conference on Pattern Recognition*, volume 4, 441–444. IEEE.

Zhang, Y.; Sun, P.; Jiang, Y.; Yu, D.; Weng, F.; Yuan, Z.; Luo, P.; Liu, W.; and Wang, X. 2022. Bytetrack: Multi-object tracking by associating every detection box. In *European Conference on Computer Vision*, 1–21. Springer.

Zhao, H.; Jiang, L.; Jia, J.; Torr, P. H.; and Koltun, V. 2021. Point transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16259–16268.

Zheng, J.; Liu, X.; Liu, W.; He, L.; Yan, C.; and Mei, T. 2022. 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*, 20228–20237.

Zhu, Z.; Guo, X.; Yang, T.; Huang, J.; Deng, J.; Huang, G.; Du, D.; Lu, J.; and Zhou, J. 2021. Gait recognition in the wild: A benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14789–14799.