LidarGait: Benchmarking 3D Gait Recognition with Point Clouds

Chuanfu Shen1,2, Fan Chao2,3, Wei Wu2, Rui Wang2, George Q. Huang4, Shiqi Yu2,3,*
1 Department of Industrial and Manufacturing Systems Engineering, The University of Hong Kong
2 Department of Computer Science and Engineering, Southern University of Science and Technology
3 Research Institute of Trustworthy Autonomous System, Southern University of Science and Technology
4 Department of Industrial and Systems Engineering, The Hong Kong Polytechnic University

noahshen@connect.hku.hk, {12131100, 12032501, 12232385}@mail.sustech.edu.cn

Abstract

Video-based gait recognition has achieved impressive results in constrained scenarios. However, visual cameras neglect human 3D structure information, which limits the feasibility of gait recognition in the 3D wild world. Instead of extracting gait features from images, this work explores precise 3D gait features from point clouds and proposes a simple yet efficient 3D gait recognition framework, termed LidarGait. Our proposed approach projects sparse point clouds into depth maps to learn the representations with 3D geometry information, which outperforms existing point-wise and camera-based methods by a significant margin. Due to the lack of point cloud datasets, we build the first large-scale LiDAR-based gait recognition dataset, SUSTech1K, collected by a LiDAR sensor and an RGB camera. The dataset contains 25,239 sequences from 1,050 subjects and covers many variations, including visibility, views, occlusions, clothing, carrying, and scenes. Extensive experiments show that (1) 3D structure information serves as a significant feature for gait recognition. (2) LidarGait outperforms existing point-based and silhouette-based methods by a significant margin, while it also offers stable cross-view results. (3) The LiDAR sensor is superior to the RGB camera for gait recognition in the outdoor environment. The source code and dataset have been made available at https://lidargait.github.io.

1. Introduction

Gait is an essential biometric, which has the unique advantage of human identification at a distance without physical contact. Gait empowers real-world applications such as human retrieval, forensic identification, and serving robots. Recently, great progress has been made to pro-

*Corresponding Author

Figure 1. Illustration of (a) camera-based and (b) LiDAR-based gait recognition. Camera-based gait recognition commonly uses silhouettes to learn shape information from a single view. LiDAR-based gait recognition can use 3D structure, shape, and scale information to identify a subject.
models [23, 27, 56], which is usually imprecise in various challenging conditions of low resolution, poor illumination, untrained posture, etc. Fortunately, 3D sensors provide precise 3D perception like human nature, e.g., recognizing a subject from multiple views as illustrated in Fig. 1b.

Visual ambiguity is the alternative limitation of camera-based approaches. To our knowledge, most existing gait datasets [42, 56, 58] only consider camera-based modalities, and fail to acknowledge the challenges of visual ambiguity caused by poor illumination and complex backgrounds in outdoor environments. These factors can significantly harm the performance of upstream tasks like pedestrian detection and segmentation, which in turn affects the accuracy of the gait system in real-world applications. Thus, obtaining precise 3D information for gait description is highly desirable to eliminate visual ambiguity in RGB images.

The remarkable success of 3D applications [6, 14, 33] motivates us to endow gait recognition with precise 3D structural information and accurate human perception, by utilizing LiDAR sensors in challenging outdoor environments. In addition to improving gait recognition, LiDAR sensors offer potential benefits in many scenarios, including robotics, healthcare, social security, and surveillance. For example, robots equipped with LiDAR-based gait recognition can function as 24 × 7 security guards, enhancing community safety. Vehicles fitted with LiDAR sensors can aid in locating lost orders and children. Furthermore, LiDAR is more privacy-preserving than cameras, making it suitable for sensitive scenarios such as nursing homes and kindergartens. Additionally, LiDAR has the potential to enhance biometric security by protecting against Deepfake attacks compared to cameras.

This paper introduces SUSTech1K, the first large-scale LiDAR-based gait dataset to facilitate 3D gait recognition with point clouds. The dataset is captured outdoors using a Velodyne VLS128 LiDAR sensor and an RGB camera mounted together on a robot. Compared to existing datasets listed in Tab. 1, SUSTech1K offers several distinctive features: (1) **Precision.** The SUSTech1K dataset provides 3D point clouds as gait representations with high precision and density, providing precise and robust 3D structure information for recognition. (2) **Scalability.** The dataset captures 25,239 sequences from 1,050 subjects, providing scalability for statistical evaluation. (3) **Diversity.** The dataset includes diverse and realistic challenges, such as illumination, occlusion, dressing, carrying, and more, along with detailed annotations, enabling the community to study the impact of different factors on gait recognition. (4) **Multimodality.** The dataset captures data streams from LiDAR and camera sensors, opening up opportunities for exploring sensor fusion approaches for robust gait recognition.

Given that 3D point clouds are formatted differently from pixels in images and that point-based gait recognition has received little attention, we investigate four cutting-edge methods [15, 36, 37, 54] from the study of point-based object classification [36]. However, we observed that all the implemented point-based methods performed sub-optimally when compared to methods using camera-based silhouettes. We believe the performance gap is primarily due to the difference in feature granularity of the task. The aforementioned point-based methods are primarily designed for coarse-grained object classification, focusing more on global context modeling. In contrast, gait recognition requires extracting fine-grained local information to achieve high accuracy.

To address this issue, we propose a simple yet effective baseline method named the LidarGait. Specifically, LidarGait first projects 3D point clouds into depth images from the LiDAR range view and then employs convolutional networks to extract gait features with 3D structural information from the projection. This approach contrasts point-wise methods that learn global context from sparse point clouds with limited local connectivity. Using convolutional neural networks on projection, LidarGait can efficiently capture the fine-grained and discriminative gait features from sparse point clouds. Extensive experiments demonstrate that (1) LidarGait is effective in maintaining 3D structural information for gait recognition, and including 3D information can significantly contribute to performance improvement, (2) point-based gait recognition equipped with a LiDAR sensor performs stably well on various challenges, convincingly demonstrating its practical significance.

To summarize, our main contributions are as follows: (1) We carry out one of the first studies of 3D gait recognition with point clouds, bringing precise perception and 3D geometry of humans for better practicality in real-world scenarios. (2) We introduce SUSTech1K, the first large-scale LiDAR-based gait recognition benchmark, which includes a range of annotations covering occlusions, viewpoints, carrying, clothing, and distance. (3) We propose a novel point cloud gait recognition framework, LidarGait, outperforming camera-based methods by a large margin.

2. Related Work

**Gait Recognition.** According to the used representations, gait recognition can be generally divided into 2D and 3D representations-based methods [39].

The majority of 2D representations-based methods study gait characteristics directly from images, termed appearance-based [5, 11, 27, 46] methods, which have made surprising high performance based on silhouettes [16, 25, 28, 29] together with other gait templates [4, 16, 45]. The alternative approaches learn human structure [23, 27, 44] and dynamics [44] as gait representations, but they are heavily constrained by model-based estimation models. 3D representation-based methods are generally extracted by
The SUSTech1K dataset is captured by a mobile robot equipped with a 128-beam LiDAR scanner and a monocular camera, providing synchronized multimodal data. It includes 1,050 identities, 25,239 sequences, 763,416 point-cloud frames, and 3,075,575 RGB images with corresponding silhouettes. The SUSTech1K dataset is a synchronized multimodal dataset, with timestamped frames for each modality of frames. In addition, we protect the privacy of the participants by blurring their faces and obtaining informed consent. To the end, we manually annotate various walking conditions in SUSTech1K.

### Data Collection

The dataset was collected in July 2022 in three scenes on the SUSTech campus using an industrial camera and a LiDAR sensor. The camera captured RGB imagery streams at a resolution of $1,280 \times 980$ and 30 frames per second (FPS), while the point cloud streams were recorded at 10 FPS. We synchronized the LiDAR and camera with the GPS clock and timestamped each frame to enable collaboration between the two modalities for a robust gait recognition system. The current indoor datasets mainly focus on subject-centered gait recognition.

---

**Table 1. Comparison of publicly available datasets for gait recognition.**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Year</th>
<th>Subject #</th>
<th>Seq #</th>
<th>View #</th>
<th>Data Type</th>
<th>3D</th>
<th>Multimodal</th>
<th>Outdoor</th>
</tr>
</thead>
<tbody>
<tr>
<td>CASIA-C [43]</td>
<td>2006</td>
<td>153</td>
<td>1,530</td>
<td>1</td>
<td>Infra-red, Silhouettes</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>KY4D [20]</td>
<td>2010</td>
<td>42</td>
<td>168</td>
<td>16</td>
<td>Silhouettes, RGB, 3D Volumetrics</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>TUM-GAID [17]</td>
<td>2012</td>
<td>305</td>
<td>3,370</td>
<td>1</td>
<td>Audio, Video, Depth</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>SZTAKI-LGA [3]</td>
<td>2016</td>
<td>28</td>
<td>11</td>
<td>1</td>
<td>3D Point Cloud</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>FVG [52]</td>
<td>2019</td>
<td>226</td>
<td>2,856</td>
<td>3</td>
<td>RGB</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>PCG [49]</td>
<td>2020</td>
<td>30</td>
<td>60</td>
<td>1</td>
<td>3D Point Cloud</td>
<td>✓</td>
<td>✓</td>
<td>✗</td>
</tr>
<tr>
<td>GREW [58]</td>
<td>2021</td>
<td>26,345</td>
<td>128,671</td>
<td>882</td>
<td>Silhouettes, 2D/3D Skeleton, Flow</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>Gait3D [56]</td>
<td>2022</td>
<td>4,000</td>
<td>25,309</td>
<td>39</td>
<td>Silhouettes, 2D/3D Skeleton, 3D Mesh</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>SUSTech1K</td>
<td>2023</td>
<td>1,050</td>
<td>25,239</td>
<td>12</td>
<td>RGB, Silhouettes, 3D Point Cloud</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

---

**Figure 2. Data acquisition setup.** Each participant is first instructed to normally walk along four round-trip paths and four one-way paths, then walk again with a random variance along the same paths.

---

**3. The SUSTech1K Benchmark**

The SUSTech1K dataset is captured by a mobile robot equipped with a 128-beam LiDAR scanner and a monocular camera, providing synchronized multimodal data. It includes 1,050 identities, 25,239 sequences, 763,416 point-cloud frames, and 3,075,575 RGB images with corresponding silhouettes. The SUSTech1K dataset is a synchronized multimodal dataset, with timestamped frames for each modality of frames. In addition, we protect the privacy of the participants by blurring their faces and obtaining informed consent. To the end, we manually annotate various walking conditions in SUSTech1K.

**Data Collection.** The dataset was collected in July 2022 in three scenes on the SUSTech campus using an industrial camera and a LiDAR sensor. The camera captured RGB imagery streams at a resolution of $1,280 \times 980$ and 30 frames per second (FPS), while the point cloud streams were recorded at 10 FPS. We synchronized the LiDAR and camera with the GPS clock and timestamped each frame to enable collaboration between the two modalities for a robust gait recognition system. The current indoor datasets mainly focus on subject-centered gait recognition.
with a clean background, while other existing datasets focus on pedestrians from surveillance views. In contrast, SUSTech1K was collected from a robot view. The existing gait datasets such as OUMVLP [42], and CASIA-E [40] set the capture range to around 8 meters. Our experimental setup is inherited from existing gait datasets such as OUMVLP [42], and CASIA-E [40].

SUSTech1K was collected from a robot view. The existing datasets, such as OUMVLP [42], and CASIA-E [40], set the capture range to around 8 meters. Our experimental setup is inherited from existing gait datasets such as OUMVLP [42], and CASIA-E [40].

Variance. In practice, we instructed each subject to walk with random attributes during their second round. The SUSTech1K dataset preserves the variances found in existing datasets, such as Normal, Bag, Clothes Changing, Views and Object Carrying, while also considering other common but challenging variances encountered outdoors, including Occlusion, Illumination, Uniform, and Umbrella. By categorizing these walking sequences into different subsets based on their variances, we can further explore the impact of different variations on the gait recognition performance of the two modalities.

Annotations and Representations. The continuous data streams are first manually segmented into sequences based on the predefined trajectories shown in Fig. 2. Each sequence is then labeled according to the aforementioned variances. Finally, the camera-based and lidar-based sequences are processed separately to obtain gait representations.

For camera-based representations, we first applied human detection [13], tracking [51], and segmentation [32] on the raw RGB imagery streams, to generate camera-based gait representations with RGB images and corresponding silhouettes. In cases where the tracking algorithm produced inaccurate bounding boxes, we manually corrected them. It should be noted that the performance of the segmentation algorithm deteriorates in low-light conditions, resulting in suboptimal performance.

For LiDAR-based representations, obtaining LiDAR-based representations was relatively easy because there was only one subject walking in the experimental area at a time. To protect the privacy of uninvolved passers and to generate clean gait representations in the point cloud format, we only release the area range to $[-5, -12] m$ for the X axis, $[-3, 3] m$ for the Y axis, and $[-2, 3] m$ for the Z axis. Moreover, we applied noise removal [9] and ground removal [21] on each frame to clean the lidar data.

Statistics. Fig. 4a indicates that the two modalities have the same number of sequences, while the RGB modality has three times more frames per sequence than the LiDAR modality. The imagery gait representations provide more details and dense information when the camera-based representations are in the resolution of $128 \times 128$ as shown in Fig. 4b. When we resize the imagery to the resolution of $64 \times 64$, the ratio of pixels vs points is approximately 1:1, allowing for a more direct and fair comparison of the two modalities. In the end, the distribution of attributes in Fig. 4c, shows the diversity of the SUSTech1K dataset.

Evaluation Metrics. To establish a more challenging and realistic setting, the SUSTech1K dataset is evaluated under an open-set setting [35,42], where train and test set splits are without sample overlapping. The evaluation protocol follows the cross-view recognition setting as commonly used in CASIA-B [50] and OUMVLP [42], where probe sets of the same view calculate the similarity to gallery sets of each view. The probe sets are grouped into many subsets according to the attributes to evaluate the impact of attributes, then perform cross-view retrieval task. The prevailing Rank-1 accuracy and Rank-5 are adopted as the evaluation metric.

4. Gait Recognition with Point Clouds

4.1. Problem Setting

In this section, we introduce the problem setting of 3D gait recognition with point clouds. Given a point cloud dataset $P = \{P_i | i = 1, 2, ..., N; j = 1, 2, ..., m_i\}$ with N identities and $m_i$ sequence for each identity $y_i$. Each point cloud sequence $P_i^j \in \mathbb{R}^{T \times N \times C}$ is with $T$ frames and $N$ points for each frame, where $C$ is the number of feature...
channels. Our goal is to learn a network $N_\theta(\cdot)$ that can produce the feature embedding $F_i$ to represent the associated identity $y_i$.

We propose the LidarGait, as shown in Fig. 5, to tackle the 3D gait recognition task, formulated as:

$$F_i = N_\theta(G(P_i^1))$$  

(1)

where the projection function $G$ operates on point clouds and generates depth images from the LiDAR front view. The feature extractor $N_\theta$ is composed of two components. 1) a structural feature encoder $S$ that captures spatially local connectivity from projected front-view depth images. 2) a temporal aggregation network $T$ that models dynamical conjunction along sequential input, which can be formulated as:

$$N_\theta(\cdot) = T(S(G(P_i^1), \cdots, S(G(P_i^{n_i}))))$$  

(2)

LidarGait first receives sequential 3D point clouds and then extracts spatial-temporal representation from projected depth maps. To end, LidarGait is optimized by combining triplet and cross-entropy loss.

4.2. LidarGait

Point-to-Depth Projection. The range-scanned point clouds from a Velodyne VLS128 LiDAR scanner, can be projected and discretized into a 2D point map, using the following projection function [22]:

$$r = \left[ \text{atan2}(y, x)/\Delta \theta \right]$$

$$c = \left[ \text{arcsin}(z/\sqrt{x^2 + y^2 + z^2})/\Delta \phi \right]$$  

(3)

where 3D point $p = (x, y, z)^T$ is mapped to its corresponding 2D pixel coordinates $(r, c)$ in the projected depth image. The $\Delta \theta$ and $\Delta \phi$ represent the average horizontal and vertical angle resolution between consecutive beam emitters. According to the configuration of the LiDAR, the $\Delta \theta$ and $\Delta \phi$ are set to 0.192 and 0.2, respectively. The resulting 2D point map is similar to cylindrical images. Each element in the map at position $(r, c)$ is filled with $d$, where $d = \sqrt{x^2 + y^2}$. In the rare case where multiple points are projected to the same 2D position, the point closest to the observer is kept. If no 3D point is projected onto a particular 2D position, the corresponding element in the point map is filled with 0. Then the depth projection is normalized and converted to RGB images from the 1-channel images.

Structural Representation Learning. LidarGait extracts abstract structural features from sequences of depth images using a convolutional network $S$. The $S$ is a spatial feature encoder that can use any existing silhouette-based backbone. In this work, we use GaitBase [10] as our feature encoder $S$. As opposed to camera-based methods that use silhouettes as input, point-wise methods [36, 37, 54] extract representation directly from point clouds. However, point-wise methods underperform camera-based recognition methods as shown in Tab. 2, despite using the informative 3D structures of pedestrians. We attribute this performance gap to the fact that current point-based models are optimized for global feature modeling, which is suitable for distinguishing objects with large inter-class differences. However, gait recognition requires capturing fine-grained features to distinguish individuals with small inter-class distances and large intra-class distances. To address this challenge, LidarGait utilizes convolutional networks to extract gait representations from projection depths, which are better at capturing such fine-grained features.

Temporal Fusion. To aggregate the features from the variable length of depth sequences, we use Set Pooling [5] as the temporal feature aggregator, which enables the model to capture the final sequence-level gait representation.

MV-LidarGait. In addition to obtaining projected depths from the perspective of the LiDAR sensor, point clouds can also be projected from orthographic views as described in SimpleView [15]. To verify whether the LiDAR range-
view projection is adequate and to explore the effectiveness of other projected views, we extend LidarGait to MV-LidarGait, which projects point sets into two extra orthogonal views, as illustrated in Fig. 6. Each rendered view is independently extracted by a feature encoder and fused in a frame-by-frame manner.

4.3. Training and Inference

The model is trained using a combined loss function that includes the $BA^+$ triplet loss [5] and the cross-entropy loss [31], with weighted hyperparameters $\alpha$ and $\beta$, respectively:

$$L = \alpha L_{tri} + \beta L_{ce}$$  \hspace{1cm} (4)

During inference, the similarity between each probe-gallery pair is measured using the Euclidean distance.

5. Experiments

5.1. Experimental Setup

Evaluation Protocol. All experiments are conducted on the SUSTech1K dataset, which is divided into three splits: a training set with 250 identities and 6,011 sequences, a validation set with 6,025 sequences from 250 unseen identities, and a test set with the remaining 550 identities and 13,203 sequences. The SUSTech1K dataset provides gait sequences from multiple viewpoints, enabling the study of cross-view gait recognition in both camera and LiDAR modalities. The cross-view evaluation protocol [42, 50] in CASIA-B and OUMVLP is adopted for SUSTech1K as well. During the test, the sequences in normal conditions are grouped into gallery sets, and the sequences in variant conditions are taken as probe sets.

Evaluation on Each Condition. To investigate the impact of various realistic factors on gait recognition in the wild, including clothes changing, poor illumination, object carrying, occlusion, and wearing uniforms, we group all probes with the different covariates into multiple subsets for evaluation. For instance, the umbrella subset consists of probes with an umbrella to evaluate the effect of carrying an umbrella, with the gallery set containing all sequences in normal conditions.

5.1.1 Comparative Methods

As detailed in Sec. 4.2, LidarGait utilizes the GaitBase with set pooling to capture 3D gait features on range-view depths. We evaluate LidarGait with the below methods.

Camera-based Methods. To evaluate the performance of the camera-based modality, we implement four cutting-edge methods: GaitSet [5], GaitBase [10], GaitPart [11], and GaitGL [31]. The network parametric setting is identical to the configuration for CASIA-B, which has the equivalent scale of the training set to SUSTech1K.

LiDAR-based Methods. We implement four commonly used approaches in point cloud classification including PointNet [36], PointNet++ [37], PointTransformer [54], and SimpleView [15]. Among them, the first three methods [36, 37, 54] are point-wise models, while SimpleView [15] is a representative projection-based method.

Implementation Details. All the camera-based silhouettes and LiDAR-based depth images are aligned using the method introduced in [19] and then resized in the resolution of $64 \times 64$. For LiDAR modality methods, we use the SGD optimizer with a weight decay of 0.0005 and an initial learning rate of 0.1. The learning rate is reduced by a factor...
Table 2. Evaluation with different attributes on SUSTech1K valid + test set. The bolded and underlined values represent the first and second-best results, respectively.

<table>
<thead>
<tr>
<th>Model</th>
<th>Publication</th>
<th>Modality</th>
<th>Normal</th>
<th>Bag</th>
<th>Clothing</th>
<th>Probe Sequence (Rank-1 acc)</th>
<th>Overall Rank1</th>
<th>Rank5</th>
</tr>
</thead>
<tbody>
<tr>
<td>GaitSet [5]</td>
<td>AAAI2019</td>
<td></td>
<td>69.10</td>
<td>62.85</td>
<td>37.44</td>
<td>65.01 63.08 61.00 67.19 23.04</td>
<td>65.04</td>
<td>84.76</td>
</tr>
<tr>
<td>GaitGL [31]</td>
<td>ICCV2021</td>
<td></td>
<td>67.11</td>
<td>66.16</td>
<td>35.92</td>
<td>63.31 61.58 58.07 66.59 17.88</td>
<td>63.14</td>
<td>82.82</td>
</tr>
<tr>
<td>GaitBase [10]</td>
<td>CVPR2023</td>
<td>Camera</td>
<td>81.46</td>
<td>77.48</td>
<td>49.60</td>
<td>75.77 75.55 76.66 81.40 25.92</td>
<td>76.12</td>
<td>89.39</td>
</tr>
<tr>
<td>PointNet [36]</td>
<td>CVPR2017</td>
<td></td>
<td>43.59</td>
<td>37.27</td>
<td>25.72</td>
<td>28.78 19.85 30.05 44.29 27.35</td>
<td>31.33</td>
<td>39.75</td>
</tr>
<tr>
<td>PointNet++ [37]</td>
<td>NIPS2017</td>
<td>LiDAR</td>
<td>55.90</td>
<td>52.22</td>
<td>41.60</td>
<td>49.60 47.84 45.91 54.16 52.49</td>
<td>50.78</td>
<td>82.18</td>
</tr>
<tr>
<td>PointTransformer [54]</td>
<td>ICCV2021</td>
<td></td>
<td>53.19</td>
<td>48.08</td>
<td>32.05</td>
<td>43.20 39.06 41.75 47.87 47.12</td>
<td>44.37</td>
<td>76.70</td>
</tr>
<tr>
<td>SimpleView [15]</td>
<td>ICML2021</td>
<td></td>
<td>72.33</td>
<td>68.75</td>
<td>57.15</td>
<td>63.26 49.20 62.52 79.72 66.54</td>
<td>64.83</td>
<td>85.77</td>
</tr>
<tr>
<td>LidarGait</td>
<td>Ours LiDAR</td>
<td></td>
<td>91.80</td>
<td>88.64</td>
<td>74.56</td>
<td>89.03 67.50 80.86 94.53 90.41</td>
<td>86.77</td>
<td>96.08</td>
</tr>
</tbody>
</table>

of 0.1 at the 20,000th and 30,000th iterations, and the total number of iterations is set to 40,000. For methods using camera modality, the Adam optimizer is used to prevent the issue of gradient vanishing because of low-quality silhouettes. The triplet and cross-entropy loss weights are set to 1 and 0.1, respectively. The batch size \((p, k, l)\) is set to \((8, 8, 10)\) and \((8, 16, 30)\) for lidar-based methods and camera-based methods, respectively, where \(p\) denotes the number of IDs, \(k\) for the number of sequence of training samples per ID, and \(l\) is the number of frame per sequence. All comparison methods are trained using the same training strategy as LidarGait. The OpenGait [10] codebase is used to conduct all experiments.

5.2. Comparative Results

Following the cross-view evaluation protocol [42, 50], we evaluate all methods on each subset with different conditions, and we report the cross-view accuracy matrix in Fig. 7 for a detailed performance comparison between LiDAR and camera modality. We report the average of the accuracy matrix in Tab. 2, obtaining the following observations: (1) LidarGait shows its superiority to all existing point-based and camera-based methods, which is mainly beneficial by integration with 3D geometry information. (2) LidarGait achieves state-of-the-art results in all conditions except the umbrella subset. It is mainly caused that umbrellas are erased after segmentation on RGB images, while the umbrellas are kept in point sets. (3) The methods using silhouettes make a poor performance at night. Point-based methods provide more promising results, and LidarGait outperforms others by a large margin. (4) All silhouette-based models [5, 10, 11, 31] achieve higher accuracy than point-based models [36, 37, 54] in point cloud classification. This concludes that it is necessary to design point-based models for 3D gait recognition specifically. (5) The models utilizing the order of the frames in sequences, i.e. GaitPart, and GaitGL, obtain lower results. While other set-based methods, i.e. GaitSet, and GaitBase, perform better accuracy. It means that temporal cues may be impacted in the outdoor scenes because of low-quality silhouettes.
tion than Lidar, so the silhouettes from the camera can have more details. (2) Though LiDAR generates point clouds in sparse space, the depth of information makes a magnificent improvement to the accuracy. Integrating depth information can improve rank-1 accuracy from 64.70% to 86.77%, validating the necessity and effectiveness of 3D information for gait recognition.

**Effectiveness of Other Projected View.** The right-side view (RSV) is obtained by positioning a virtual camera orthogonal to the LiDAR range view (RV) on the right-hand side. The Bird-eye’s view (BEV) projects point clouds onto a plane above the point clouds of pedestrians. Based on the results presented in Tab. 3, we have the following observations:

1. LidarGait achieves the best performance on the range view projection when a single viewpoint is used as input.
2. The right-side view also provides a reliable representation, which performs comparably to camera-based methods.
3. Although the accuracy of Bird’s-eye views is only 26.33%, learning gait features from BEV images provides interesting evidence that gait recognition can potentially be achieved at the Bird’s-eye view.
4. MV-LidarGait can be improved (+0.73%) from LidarGait by combining multiple viewpoints, with the improvement mainly coming from the umbrella subset.
5. The integration of BEV does not enhance the performance of MV-LidarGait, indicating that BEV only provides redundant information already contained in the other two viewpoints.

More experiments and exemplar data on SUSTech1K are included in the supplementary material.

### 6. Discussion

**Ethical Discussion.** The SUSTech1K dataset has been reviewed by the Southern University of Science and Technology Institutional Review Board (SUSTech IRB). All the subjects involved in the dataset signed a written consent to agree that their data can be collected, processed, used, and shared for research purposes. The dataset can be distributed only for non-commercial research purposes with the case-by-case dataset access application. The human faces are blurred from RGB images to protect sensitive privacy. The

---

**Table 3. Effectiveness of each projected view.** MV-LidarGait achieves the best performance.

<table>
<thead>
<tr>
<th>Model</th>
<th>Used views</th>
<th>Normal</th>
<th>Umbrella</th>
<th>Overall Rank1</th>
</tr>
</thead>
<tbody>
<tr>
<td>LidarGait</td>
<td>BEV</td>
<td>39.92</td>
<td>12.12</td>
<td>26.33</td>
</tr>
<tr>
<td></td>
<td>RSV</td>
<td>70.61</td>
<td>47.02</td>
<td>62.67</td>
</tr>
<tr>
<td></td>
<td>RV</td>
<td>91.80</td>
<td>67.50</td>
<td>86.77</td>
</tr>
<tr>
<td>MV-LidarGait</td>
<td>RV + RSV</td>
<td>91.29</td>
<td>70.91</td>
<td>87.50</td>
</tr>
<tr>
<td></td>
<td>RV + RSV + BEV</td>
<td>91.72</td>
<td>69.43</td>
<td>87.47</td>
</tr>
</tbody>
</table>

---

**Conclusions and Future work**

In this paper, we introduce the LiDAR sensor to provide reliable anthropometric parameters for the human body, and to perceive pedestrians in unconstrained scenes. First, we proposed a novel multi-view projection network for point cloud gait recognition, named LidarGait, to exploit 3D human geometry from multi-view representations. Moreover, we build the first large-scale multimodal 3D point cloud gait recognition dataset, termed SUSTech1K, to facilitate the research of gait recognition with point cloud data. SUSTech1K contains 25,239 sequences with 1,050 subjects and covers various visibility, views, occlusions, clothing, carry, and scenes. Lastly, our proposed method achieves remarkable results on the SUSTech1K dataset, showing the superiority of LiDAR and the effectiveness of LidarGait.

LidarGait has obtained remarkable results in various scenarios, yet it does not perform well when subjects carry umbrellas. The reason should be that the umbrellas are wrongly included in the point cloud. Better performance can be achieved if the umbrellas are removed from the point clouds. Besides, LidarGait only takes one modality as input currently. SUSTech1K dataset is a multimodal dataset with synchronized RGB images and point clouds. Much better results should be achieved if the two modalities are fused.

**Acknowledgements**

We thank Jingzhe Ma and Junhao Liang for their invaluable assistance in the SUSTech1K collection, and Xinlai Liu for his careful proofreading. We also thank the BDStar company for providing the LiDAR sensor that helps us explore this research. This work was supported in part by the National Natural Science Foundation of China under Grant 61976144, in part by the Stable Support Plan Program of Shenzhen Natural Science Fund under Grant 20200925155017002, in part by the National Key Research and Development Program of China under Grant 2020AAA0140002, in part by the Shenzhen Technology Plan Program under Grant JSGG20201103091002008, and in part by the HKSAR RGC Theme-based Research Scheme under Grant T32-707-22-N.

---

2[https://lidargait.github.io/](https://lidargait.github.io/)


