LidarGait: Benchmarking 3D Gait Recognition with Point Clouds

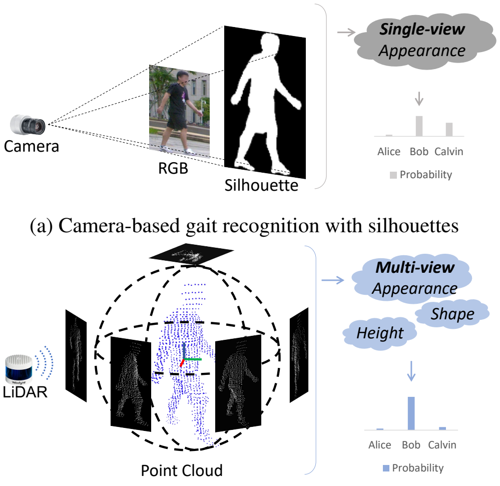
Chuanfu Shen[1,2], Fan Chao[2,3], Wei Wu[2], Rui Wang[2], George Q. Huang[4], Shiqi Yu[2,3]*
[1] Department of Industrial and Manufacturing Systems Engineering, The University of Hong Kong
[2] Department of Computer Science and Engineering, Southern University of Science and Technology
[3] Research Institute 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
gq.huang@polyu.edu.hk, yusq@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 promote gait recognition from in-the-lab setting [19, 42, 50] to in-the-wild scenario [17, 52, 56, 58]. Despite these studies have made significant contributions to recent advances [5, 7, 11, 30, 31, 38, 55], two inherent problems still remain (1) *lack of 3D geometry information*, and (2) *poor feasibility in the real-world scenario*.



(a) Camera-based gait recognition with silhouettes

(b) LiDAR-based gait recognition with point clouds

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.

Existing camera-based methods [18, 53] are counterintuitive to human nature. When recognizing a subject [56, 57], humans consider not only the 2D appearance characteristics, but also 3D geometry structure information like height, shape, and viewpoints. Differently, camera-based gait recognition methods [5, 26, 31] either capture 2D representations from a single viewpoint, as shown in Fig. 1a, or exploit 3D representations from estimated 3D pose/mesh

*Corresponding Author

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 contributes 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, 42, 46] methods, which have made surprising high performance based on silhouettes [16, 25, 26, 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

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

Dataset	Year	Subject #	Seq #	View #	Data Type	3D	Multimodal	Outdoor
CASIA-B [50]	2006	124	13,640	11	RGB, Silhouettes	✗	✗	✗
CASIA-C [43]	2006	153	1,530	1	Infrared, Silhouettes	✗	✗	✓
KY4D [20]	2010	42	168	16	Silhouettes, RGB, 3D Volumetrics	✓	✗	✗
TUM-GAID [17]	2012	305	3,370	1	Audio, Video, Depth	✓	✓	✓
SZTAKI-LGA [3]	2016	28	11	1	3D Point Cloud	✓	✗	✓
OU-MVLP [42]	2018	10,307	288,596	14	Silhouettes	✗	✗	✗
FVG [52]	2019	226	2,856	3	RGB	✗	✗	✓
PCG [49]	2020	30	60	1	3D Point Cloud	✓	✗	✗
GREW [58]	2021	26,345	128,671	882	Silhouettes, 2D/3D Skeleton, Flow	✗	✗	✗
Gait3D [56]	2022	4,000	25,309	39	Silhouettes, 2D/3D Skeleton, 3D Mesh	✓	✗	✓
OUMVLP-Mesh [24]	2022	10,307	288,596	14	3D Mesh	✓	✗	✗
SUSTech1K	2023	**1,050**	**25,239**	**12**	**RGB, Silhouettes, 3D Point Cloud**	✓	✓	✓

sensors [12, 17] or estimation models [25, 27]. The commonly used 3D sensors such as Kinect, provide 3D structured data, but they only facilitate in an indoor and close-distance environment [12]. Meanwhile, multi-cameras reconstruction [2] and 3D estimation models [25, 27, 44, 53, 56] provide considerable 3D geometry, but the performance is far behind the requirements of real-world applications as reported in [58].

Gait Recognition Benchmark. There are three types of publicly available datasets: in-the-lab [19, 42, 50], synthetic [8], and in-the-wild datasets [17, 34, 56, 58]. The in-the-lab datasets [19, 42, 46, 50], represented by CASIA series [43, 46, 50] and OU-ISIR series [19, 42], advance the investigation of the feasibility of gait recognition. The recent synthetic datasets [8] are to overcome the difficulty in data acquisition and annotation of gait, providing more synthetic data with a variety of annotations but introducing cross-domain issue [26] at the same time. The in-the-wild datasets [34, 56, 58] are to promote gait recognition research in the unconstrained environment. The recent works [1, 3, 49] based on LiDAR sensor are closely related to our work, while the main concern is that the existing datasets include at most 30 subjects, which cannot guarantee statistically reliable performance evaluation of LiDAR-based gait recognition. Because of insufficient 3D representations for data-driven gait recognition, as shown in Tab. 1, a dataset with accurate 3D representations is essential.

Point Cloud and 3D Object Classification. LiDAR, which stands for Light Detection and Ranging, projects laser pulses to the targets and then generate point cloud sets. Each point represents a data point in Cartesian coordinates (X, Y, Z). Point cloud data is sparsely distributed, remaining a significant challenge in modeling correlation and geometry. 3D object classification explore projection-based [15, 41, 48], point-wise [36, 37, 54], and graph-wise models [47, 48] to capture discriminative feature on point cloud data for object classification. In this paper, we select many representative models of 3D object classification and compare them with our proposed method to comprehensively study 3D point-based gait recognition.

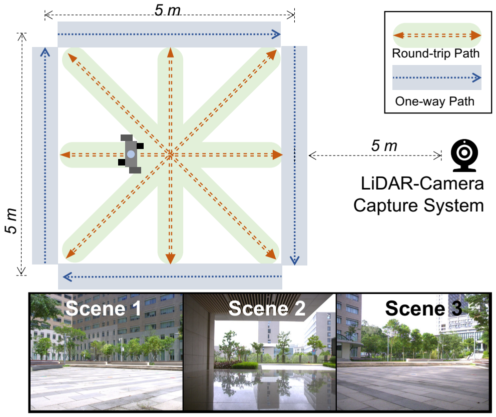
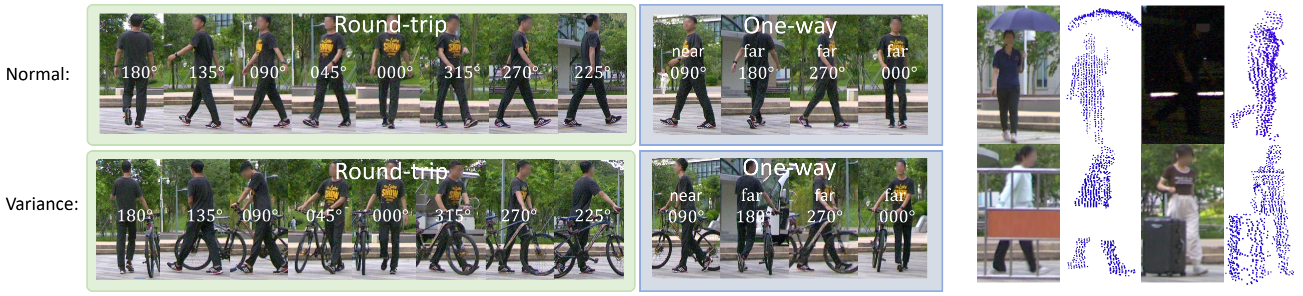


Figure 2. Data acquisition setup. Each participant is first instructed to normally walk along four round-trip paths and four one-way paths, then walk again with a random variance along the same paths.

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



(a) Samples of 12 views in two conditions for a subject.

(b) Diverse attributes in SUSTech1K.

Figure 3. (a) Each participant walks normally (top row), followed by walking with a random variance (bike for this subject) as shown in the bottom row. (b) SUSTech1K collects data in point cloud and RGB modality with diverse realistic variances.

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 at a comparative distance as shown in Fig. 2. Each subject first walked normally along the *one-way paths* and the *round-trip paths*, and then walked again with an extra random attribute, such as carrying any object, as shown in Fig. 3a. Each subject can have a total of 48 gait sequences *(= [4 × 2 (round-trip) + 4 × 1(one-way)] × 2 (twice) × 2 (modality))*.

Variances. 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 $[-5, -12m]$ for the X axis, $[-3m, 3m]$ for the Y axis, and $[-2m, 3m]$ for the Z axis. Moreover, we applied noise removal [9] and ground removal [21] on each frame to clean the lidar data.

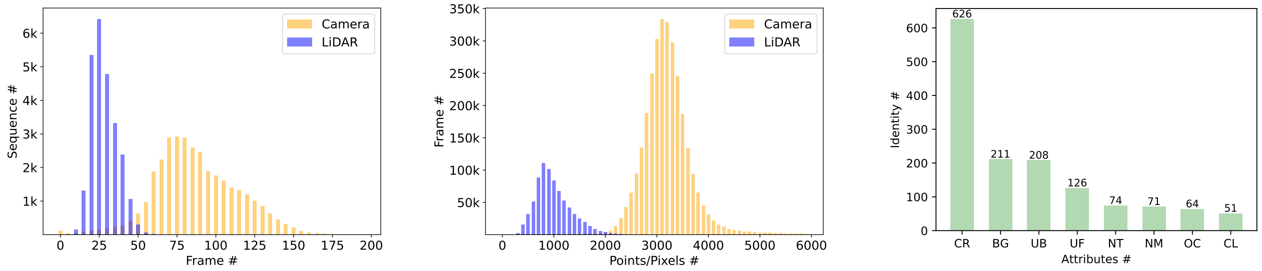
Statistics. Fig. 4a indicates that the two modalities have the same number of sequences, while the RGB modality has three times more frames per sequence than the LiDAR modality. The imagery gait representations provide more details and dense information when the camera-based representations are in the resolution of 128×128 as shown in Fig. 4b. When we resize the imagery to the resolution of 64×64, 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 $\mathcal{P} = \{\mathcal{P}_i^j | i = 1, 2..., N; j = 1, 2, ..., m_i\}$ with N identities and m_i sequence for each identity y_i. Each point cloud sequence $\mathcal{P}_i^j \in \mathbb{R}^{T \times N \times C}$ is with T frames and N points for each frame, where C is the number of feature



(a) sequence # over frame #

(b) frame # over points/pixels #

(c) Attributes distribution

Figure 4. Statistics about SUSTech1K dataset. LiDAR modality and RGB modality are represented in blue and yellow, respectively. It shows that SUSTech1K dataset is scalable, multimodal, and diverse for the study of 3D gait recognition. *CR, BG, UB, UF, NT, NM, OC,* and *CL* denote attributes of Carrying, Bag, Umbrella, Uniform, Night, Occlusion, and Clothing, respectively. Best viewed in color.

channels. Our goal is to learn a network $N_\theta(\cdot)$ that can produce the feature embedding $\mathcal{F}_i^j$ 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:

$$\mathcal{F}_i^j = N_\theta(\mathcal{G}(\mathcal{P}_i^j)) \tag{1}$$

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

$$N_\theta(\cdot) = \mathcal{T}(\mathcal{S}(\mathcal{G}(\mathcal{P}_i^1), \cdots, \mathcal{S}(\mathcal{G}(\mathcal{P}_i^{m_i}))) \tag{2}$$

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]:

$$\begin{aligned} r &= \lfloor \text{atan2}(y, x)/\Delta\theta \rfloor \\ c &= \lfloor \arcsin(z/\sqrt{x^2+y^2+z^2})/\Delta\phi \rfloor \end{aligned} \tag{3}$$

where 3D point $\mathbf{p} = (x, y, z)^\top$ is mapped to its corresponding 2D pixel coordinates (r, c) in the projected depths 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, only 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 $\mathcal{S}$. The $\mathcal{S}$ is a spatial feature encoder that can use any existing silhouette-based backbone. In this work, we use GaitBase [10] as our feature encoder $\mathcal{S}$. 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-

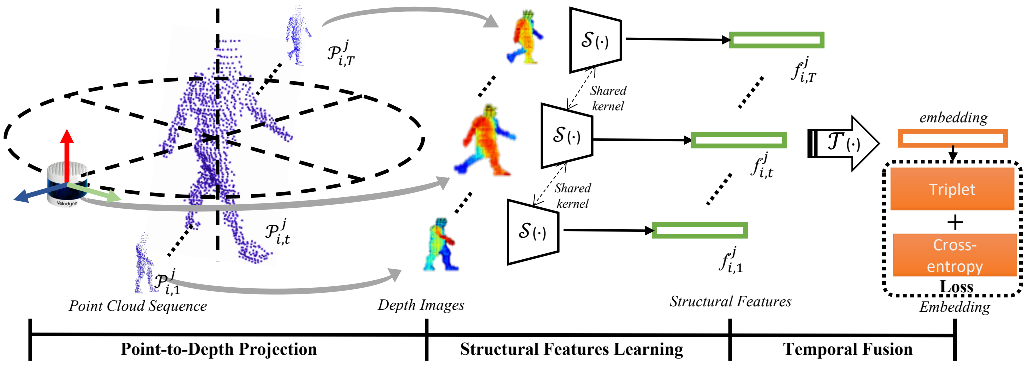


Figure 5. The framework of LidarGait for 3D gait recognition with point clouds. LidarGait receives a sequence of point clouds, extracts representations from range-view projection depths, and aggregates sequential features by set pooling.

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. Traning 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 α and β, respectively:

$$L = \alpha L_{tri} + \beta L_{ce} \tag{4}$$

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.

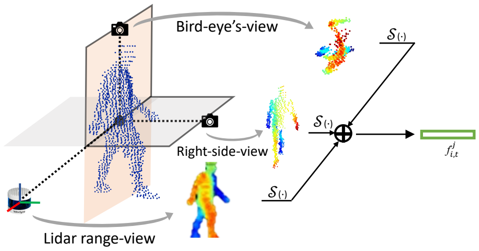


Figure 6. Illustraition of MV-LidarGait, which aggregates extra depths projection from the right-side and bird-eye view.

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 × 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

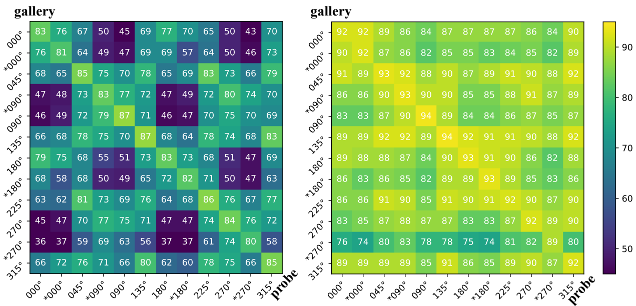
Table 2. Evaluation with different attributes on SUSTech1K *valid* + *test* set. The bolded and underlined values represent the first and second-best results, respectively.

Model	Publication	Modality	Probe Sequence (*Rank-1 acc*)								Overall	
			Normal	Bag	Clothing	Carrying	Umbrella	Uniform	Occlusion	Night	*Rank1*	*Rank5*
GaitSet [5]	AAAI2019	Camera	69.10	68.25	37.44	65.01	63.08	61.00	67.19	23.04	65.04	84.76
GaitPart [11]	CVPR2019		62.20	62.81	33.08	59.53	57.25	54.85	57.20	21.75	59.19	80.79
GaitGL [31]	ICCV2021		67.11	66.16	35.92	63.31	61.58	58.07	66.59	17.88	63.14	82.82
GaitBase [10]	CVPR2023		<u>81.46</u>	<u>77.48</u>	49.60	<u>75.77</u>	**75.55**	<u>76.66</u>	<u>81.40</u>	25.92	<u>76.12</u>	<u>89.39</u>
PointNet [36]	CVPR2017	LiDAR	43.59	37.27	25.72	28.78	19.85	30.05	44.29	27.35	31.33	59.75
PointNet++ [37]	NIPS2017		55.90	52.22	41.60	49.60	47.84	45.91	54.16	52.49	50.78	82.38
PointTransformer [54]	ICCV2021		53.19	48.08	32.05	43.20	39.06	41.75	47.87	47.12	44.37	76.70
SimpleView [15]	ICML2021		72.33	68.75	<u>57.15</u>	63.26	49.20	62.52	79.72	<u>66.54</u>	64.83	85.77
LidarGait	Ours	LiDAR	**91.80**	**88.64**	**74.56**	**89.03**	<u>67.50</u>	**80.86**	**94.53**	**90.41**	**86.77**	**96.08**

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[1].

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.



(a) Camera (b) LiDAR

Figure 7. Cross-view performance comparison between LiDAR and camera for gait recognition. We report rank-1 accuracy (%) on the cross-view protocol. * indicates viewpoint at a longer distance. Best viewed in color and pdf.

Cross-view Gait Recognition. We conduct a detailed comparison of two modalities of cross-view gait recognition in Fig. 7. The identical feature encoder, GaitBase, is utilized for two modalities to make ablative results. We can make the following observations: (1) the distance from subjects to sensors indeed impacts the performance for both two modalities. (2) Camera-based method achieves poor performance when query sets are at views of $0°, 90°, 180°$ (see purple pixel in Fig. 7a). The same phenomenon can be found on CASIA-B [5] and OUMVLP [42]. However, LiDAR-based methods can perform stably in cross-view settings, validating the effectiveness of 3D structure for cross-view gait recognition.

5.3. Ablation Study

Effectiveness of 3D Geometry Information. To evaluate the effectiveness of depth information for gait recognition, we compare four types of data as input: (1) Camera silhouettes: the camera-based silhouettes are obtained by segmentation results of RGB images. (2) LiDAR silhouettes: LiDAR silhouettes are obtained by range-view projection of point cloud sets. (3) LiDAR depth: the depth information is added. From Fig. 8, we can observe that: (1) When depth information is not included, the performance of LiDAR silhouettes is much lower than the accuracy of camera silhouettes. This is because the camera has a much higher resolu-

[1] https://github.com/ShiqiYu/OpenGait

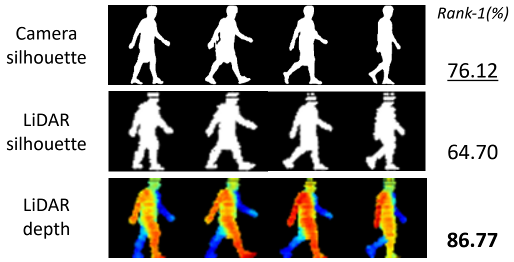


Figure 8. Ablation study on the effectiveness of depth information for performance. Best viewed in color.

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**[2].

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 recorded data can only be used for 20 years since this paper publishes. After this date, all data will be deleted and not allowed to be used.

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

Model	Used views	Normal	Umbrella	Overall *Rank1*
LidarGait	BEV	39.92	12.12	26.33
	RSV	70.61	47.02	62.67
	RV	**91.80**	67.50	86.77
MV-LidarGait	RV + RSV	91.29	**70.91**	**87.50**
	RV + RSV + BEV	91.22	69.43	87.47

7. 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/

References

[1] Jeongho Ahn, Kazuto Nakashima, Koki Yoshino, Yumi Iwashita, and Ryo Kurazume. 2v-gait: Gait recognition using 3d lidar robust to changes in walking direction and measurement distance. In *2022 IEEE/SICE International Symposium on System Integration (SII)*, pages 602–607, 2022. 3

[2] Gunawan Ariyanto and Mark S Nixon. Model-based 3d gait biometrics. In *2011 international joint conference on biometrics (IJCB)*, pages 1–7, 2011. 3

[3] Csaba Benedek, Bence Gálai, Balázs Nagy, and Zsolt Jankó. 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, 2016. 3

[4] A.F. Bobick and J.W. Davis. The recognition of human movement using temporal templates. *IEEE TPAMI*, 23(3):257–267, 2001. 2

[5] Hanqing Chao, Yiwei He, Junping Zhang, and Jianfeng Feng. Gaitset: Regarding gait as a set for cross-view gait recognition. In *AAAI*, pages 8126–8133, 2019. 1, 2, 5, 6, 7

[6] Peishan Cong, Xinge Zhu, Feng Qiao, Yiming Ren, Xidong Peng, Yuenan Hou, Lan Xu, Ruigang Yang, Dinesh Manocha, and Yuexin Ma. Stcrowd: A multimodal dataset for pedestrian perception in crowded scenes. In *CVPR*, pages 19608–19617, 2022. 2

[7] Huanzhang Dou, Pengyi Zhang, and Wei Su. Metagait: Learning to learn an omni sample adaptive representation for gait recognition. In *ECCV*, 2022. 1

[8] Huanzhang Dou, Wenhu Zhang, Pengyi Zhang, Yuhan Zhao, Songyuan Li, Zequn Qin, Fei Wu, Lin Dong, and Xi Li. Versatilegait: A large-scale synthetic gait dataset with finegrainedattributes and complicated scenarios. *arXiv preprint arXiv:2101.01394*, 2021. 3

[9] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, pages 226–231, 1996. 4

[10] Chao Fan, Junhao Liang, Chuanfu Shen, Saihui Hou, Yongzhen Huang, and Shiqi Yu. Opengait: Revisiting gait recognition toward better practicality. *arXiv preprint arXiv:2211.06597*, 2022. 5, 6, 7

[11] Chao Fan, Yunjie Peng, Chunshui Cao, Xu Liu, Saihui Hou, Jiannan Chi, Yongzhen Huang, Qing Li, and Zhiqiang He. Gaitpart: Temporal part-based model for gait recognition. In *CVPR*, pages 14225–14233, 2020. 1, 2, 6, 7

[12] Péter Fankhauser, Michael Bloesch, Diego Rodriguez, Ralf Kaestner, Marco Hutter, and Roland Siegwart. Kinect v2 for mobile robot navigation: Evaluation and modeling. In *2015 International Conference on Advanced Robotics (ICAR)*, pages 388–394, 2015. 3

[13] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. 4

[14] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 2

[15] Ankit Goyal, Hei Law, Bowei Liu, Alejandro Newell, and Jia Deng. Revisiting point cloud shape classification with a simple and effective baseline. In *ICML*, pages 3809–3820, 2021. 2, 3, 5, 6, 7

[16] Jinguang Han and Bir Bhanu. Individual recognition using gait energy image. *IEEE TPAMI*, 28(2):316–322, 2005. 2

[17] Martin Hofmann, Jürgen Geiger, Sebastian Bachmann, Björn Schuller, and Gerhard Rigoll. 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, 2014. 1, 3

[18] Xiaohu Huang, Duowang Zhu, Hao Wang, Xinggang Wang, Bo Yang, Botao He, Wenyu Liu, and Bin Feng. Context-sensitive temporal feature learning for gait recognition. In *ICCV*, pages 12909–12918, 2021. 1

[19] H. Iwama, M. Okumura, Y. Makihara, and Y. Yagi. The ouisir gait database comprising the large population dataset and performance evaluation of gait recognition. *IEEE TIFS*, 7, Issue 5:1511–1521, 2012. 1, 3, 6

[20] Y. Iwashita, R. Baba, K. Ogawara, and R. Kurazume. Person identification from spatio-temporal 3d gait. In *Int. Conf. Emerging Security Technologies (EST)*, 2010. 3

[21] Seungjae Lee, Hyungtae Lim, and Hyun Myung. Patchwork++: Fast and robust ground segmentation solving partial under-segmentation using 3d point cloud. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 13276–13283, 2022. 4

[22] Bo Li, Tianlei Zhang, and Tian Xia. Vehicle detection from 3d lidar using fully convolutional network. *arXiv preprint arXiv:1608.07916*, 2016. 5

[23] Xiang Li, Yasushi Makihara, Chi Xu, and Yasushi Yagi. End-to-end model-based gait recognition using synchronized multi-view pose constraint. In *ICCVW*, pages 4106–4115, 2021. 2

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

[25] Xiang Li, Yasushi Makihara, Chi Xu, Yasushi Yagi, Shiqi Yu, and Mingwu Ren. End-to-end model-based gait recognition. In *ACCV*, 2020. 2, 3

[26] Junhao Liang, Chao Fan, Saihui Hou, Chuanfu Shen, Yongzhen Huang, and Shiqi Yu. Gaitedge: Beyond plain end-to-end gait recognition for better practicality. *arXiv preprint arXiv:2203.03972*, 2022. 1, 2, 3

[27] Rijun Liao, Shiqi Yu, Weizhi An, and Yongzhen Huang. A model-based gait recognition method with body pose and human prior knowledge. *PR*, 98:107069, 2020. 2, 3

[28] Beibei Lin, Chen Liu, Lincheng Li, Robby T Tan, and Xin Yu. Uncertainty-aware gait recognition via learning from dirichlet distribution-based evidence. *arXiv preprint arXiv:2211.08007*, 2022. 2

[29] Beibei Lin, Yu Liu, and Shunli Zhang. Gaitmask: Mask-based model for gait recognition. In *BMVC*, pages 1–12, 2021. 2

[30] Beibei Lin, Shunli Zhang, and Feng Bao. Gait recognition with multiple-temporal-scale 3d convolutional neural net-

work. In *Proceedings of the 28th ACM international conference on multimedia*, pages 3054–3062, 2020. 1

[31] Beibei Lin, Shunli Zhang, and Xin Yu. Gait recognition via effective global-local feature representation and local temporal aggregation. In *ICCV*, pages 14648–14656, 2021. 1, 6, 7

[32] Yi Liu, Lutao Chu, Guowei Chen, Zewu Wu, Zeyu Chen, Baohua Lai, and Yuying Hao. Paddleseg: A high-efficient development toolkit for image segmentation. *arXiv preprint arXiv:2101.06175*, 2021. 4

[33] Jieru Mei, Alex Zihao Zhu, Xinchen Yan, Hang Yan, Siyuan Qiao, Liang-Chieh Chen, and Henrik Kretzschmar. Waymo open dataset: Panoramic video panoptic segmentation. In *ECCV*, pages 53–72, 2022. 2

[34] Zihao Mu, Francisco M Castro, Manuel J Marín-Jiménez, Nicolás Guil, Yan-Ran Li, and Shiqi Yu. Resgait: The real-scene gait dataset. In *2021 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–8, 2021. 3

[35] Mark S Nixon, Tieniu Tan, and Rama Chellappa. *Human identification based on gait*, volume 4. Springer Science & Business Media, 2010. 4

[36] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *CVPR*, pages 652–660, 2017. 2, 3, 5, 6, 7

[37] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *NeurIPS*, 30, 2017. 2, 3, 5, 6, 7

[38] Chuanfu Shen, Beibei Lin, Shunli Zhang, George Q Huang, Shiqi Yu, and Xin Yu. Gait recognition with mask-based regularization. *arXiv preprint arXiv:2203.04038*, 2022. 1

[39] Chuanfu Shen, Shiqi Yu, Jilong Wang, George Q Huang, and Liang Wang. A comprehensive survey on deep gait recognition: algorithms, datasets and challenges. *arXiv preprint arXiv:2206.13732*, 2022. 2

[40] Chunfeng Song, Yongzhen Huang, Weining Wang, and Liang Wang. Casia-e: a large comprehensive dataset for gait recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 4

[41] Hang Su, Subhransu Maji, Evangelos Kalogerakis, and Erik Learned-Miller. Multi-view convolutional neural networks for 3d shape recognition. In *ICCV*, pages 945–953, 2015. 3

[42] Noriko Takemura, Yasushi Makihara, Daigo Muramatsu, Tomio Echigo, and Yasushi Yagi. Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition. *IPSJ Trans. on Computer Vision and Applications*, 10(4):1–14, 2018. 1, 2, 3, 4, 6, 7

[43] Daoliang Tan, Kaiqi Huang, Shiqi Yu, and Tieniu Tan. Efficient night gait recognition based on template matching. In *ICPR*, pages 1000–1003, 2006. 3

[44] Torben Teepe, Ali Khan, Johannes Gilg, Fabian Herzog, Stefan Hörmann, and Gerhard Rigoll. GaitGraph: Graph convolutional network for skeleton-based gait recognition. In *ICIP*, pages 2314–2318, 2021. 2, 3

[45] Chen Wang, Junping Zhang, Jian Pu, Xiaoru Yuan, and Liang Wang. Chrono-gait image: A novel temporal template for gait recognition. In *ECCV*, pages 257–270, 2010. 2

[46] Liang Wang, Tieniu Tan, Huazhong Ning, and Weiming Hu. Silhouette analysis-based gait recognition for human identification. *IEEE TPAMI*, 25(12):1505–1518, 2003. 2, 3

[47] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *Acm Transactions On Graphics (tog)*, 38(5):1–12, 2019. 3

[48] Xin Wei, Ruixuan Yu, and Jian Sun. View-gcn: View-based graph convolutional network for 3d shape analysis. In *CVPR*, pages 1850–1859, 2020. 3

[49] Hiroyuki Yamada, Jeongho Ahn, Oscar Martinez Mozos, Yumi Iwashita, and Ryo Kurazume. Gait-based person identification using 3d lidar and long short-term memory deep networks. *Advanced Robotics*, 34(18):1201–1211, 2020. 3

[50] Shiqi Yu, Daoliang Tan, and Tieniu Tan. A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In *ICPR*, pages 441–444, 2006. 1, 3, 4, 6, 7

[51] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. In *ECCV*, pages 1–21, 2022. 4

[52] Ziyuan Zhang, Luan Tran, Feng Liu, and Xiaoming Liu. On learning disentangled representations for gait recognition. *IEEE TPAMI*, 2020. 1, 3

[53] Ziyuan Zhang, Luan Tran, Xi Yin, Yousef Atoum, Xiaoming Liu, Jian Wan, and Nanxin Wang. Gait recognition via disentangled representation learning. In *CVPR*, pages 4710–4719, 2019. 1, 3

[54] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16259–16268, 2021. 2, 3, 5, 6, 7

[55] Jinkai Zheng, Xinchen Liu, Xiaoyan Gu, Yaoqi Sun, Chuang Gan, Jiyong Zhang, Wu Liu, and Chenggang Yan. Gait recognition in the wild with multi-hop temporal switch. In *ACMMM*, pages 6136–6145, 2022. 1

[56] Jinkai Zheng, Xinchen Liu, Wu Liu, Lingxiao He, Chenggang Yan, and Tao Mei. Gait recognition in the wild with dense 3d representations and a benchmark. In *ICCV*, pages 20228–20237, 2022. 1, 2, 3

[57] Zhedong Zheng, Xiaohan Wang, Nenggan Zheng, and Yi Yang. Parameter-efficient person re-identification in the 3d space. *IEEE TNNLS*, 2022. 1

[58] Zheng Zhu, Xianda Guo, Tian Yang, Junjie Huang, Jiankang Deng, Guan Huang, Dalong Du, Jiwen Lu, and Jie Zhou. Gait recognition in the wild: A benchmark. In *CVPR*, pages 14789–14799, 2021. 1, 2, 3