Gait Recognition in the Wild with Dense 3D Representations and A Benchmark

Jinkai Zheng[1*] Xinchen Liu[2†] Wu Liu[2†] Lingxiao He[2] Chenggang Yan[1] Tao Mei[2]
[1]Hangzhou Dianzi University, Hangzhou, China [2]Explore Academy of JD.com, Beijing, China
{zhengjinkai3, cgyan}@hdu.edu.cn, {liuxinchen1, liuwu1, helingxiao3, tmei}@jd.com

Abstract

*Existing studies for gait recognition are dominated by 2D representations like the silhouette or skeleton of the human body in constrained scenes. However, humans live and walk in the unconstrained 3D space, so projecting the 3D human body onto the 2D plane will discard a lot of crucial information like the viewpoint, shape, and dynamics for gait recognition. Therefore, this paper aims to explore dense 3D representations for gait recognition in the wild, which is a practical yet neglected problem. In particular, we propose a novel framework to explore the 3D Skinned Multi-Person Linear (SMPL) model of the human body for gait recognition, named **SMPLGait**. Our framework has two elaborately-designed branches of which one extracts appearance features from silhouettes, the other learns knowledge of 3D viewpoints and shapes from the 3D SMPL model. In addition, due to the lack of suitable datasets, we build the first large-scale 3D representation-based gait recognition dataset, named **Gait3D**. It contains 4,000 subjects and over 25,000 sequences extracted from 39 cameras in an unconstrained indoor scene. More importantly, it provides 3D SMPL models recovered from video frames which can provide dense 3D information of body shape, viewpoint, and dynamics. Based on Gait3D, we comprehensively compare our method with existing gait recognition approaches, which reflects the superior performance of our framework and the potential of 3D representations for gait recognition in the wild. The code and dataset are available at: https://gait3d.github.io.*

1. Introduction

Visual gait recognition, which aims to identify a target person using her/his walking pattern in a video, has been studied for over two decades [29, 41]. Existing approaches and datasets are dominated by 2D gait representations such as silhouette sequences [54], Gait Energy Images (GEIs) [10], 2D skeletons [61], as shown in Figure 1. However, the human body is a 3D non-rigid object, so the 3D-to-2D projection discards a lot of useful information of shapes, viewpoints, and dynamics while presenting ambiguity for gait recognition. Therefore, this paper is focused on 3D gait recognition that is valuable yet neglected by the community.

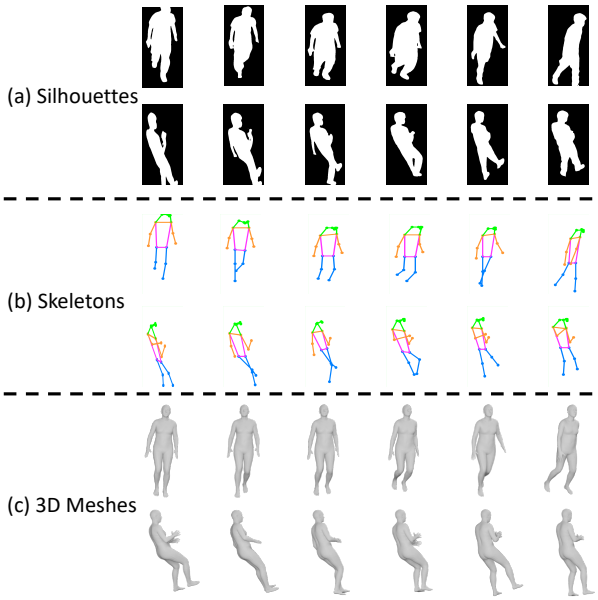


Figure 1. Different gait representations of the same person from two viewpoints. Compared with silhouettes and skeletons, 3D meshes retain the shapes and viewpoints of the human body in the 3D space. (Best viewed in color.)

Recently, deep learning-based methods have dominated the state-of-the-art performance on the widely adopted 2D gait recognition benchmarks like CASIA-B [36] and OU-MVLP [35] by directly learning discriminative features from silhouette sequences [5, 8, 55] or GEIs [44]. Despite the excellent results on the in-the-lab datasets, these methods cannot work well in the wild scenarios which have more diverse 3D viewpoints of cameras and more complex environmental interference factors like occlusions [61]. Although several works exploit 3D cylinders [3] or 3D skeletons [40], these sparse 3D models also

*This work was done when Jinkai Zheng was an intern at Explore Academy of JD.com.
†Corresponding author.

lose helpful information of human bodies like viewpoints and shapes. Fortunately, the development of parameterized human body models like the Skinned Multi-Person Linear (SMPL) model [27] and 3D human mesh recovery approaches [17, 19, 33] makes it possible to estimate precise 3D meshes and viewpoints of human bodies in video frames. The advantages of 3D meshes for gait recognition are two-fold: 1) the 3D mesh can provide not only the pose but also the shape of the human body in the 3D space, which is crucial for learning discriminative features of gait, and 2) the 3D viewpoint can be explored to normalize the orientations of human bodies during cross-view matching.

To this end, we design a novel 3D SMPL model-based Gait recognition framework, i.e., **SMPLGait**, to explore the 3D gait representations for human identification. Our SMPLGait framework has two branches based on deep neural networks. One branch takes the silhouette sequence of a person as the input to learn appearance features like clothing, hairstyle, and belongings. However, due to the extreme viewpoint changes in the wild, the shape of the human body can be distorted, which makes the appearance ambiguous, as shown in Figure 1. To overcome this challenge, we design a 3D Spatial-Transformation Network (3D-STN) as the other branch to learn 3D knowledge of viewpoint and shape from the 3D human mesh. The 3D-STN takes the 3D SMPL model of each frame as the input to learn a spatial transformation matrix. By applying the spatial transformation matrix to the appearance features, these features from different viewpoints are normalized in the latent space. By this means, the gait sequences of the same person will be closer in the feature space.

Nevertheless, there is no suitable dataset that provides 3D meshes of human bodies in the wild. Therefore, to facilitate the research, we build the first large-scale 3D mesh-based gait recognition dataset, named **Gait3D**, from high-resolution videos captured in the wild. Compared to existing datasets listed in Table 1, the Gait3D dataset has the following featured properties: **1)** Gait3D contains 4,000 subjects with over 25,000 sequences captured by 39 cameras in an unconstrained indoor scene which makes it scalable for research and applications. **2)** It provides precise 3D human meshes recovered from video frames which can provide 3D pose and shape of human bodies as well as accurate viewpoint parameters. **3)** It also provides conventional 2D silhouettes and keypoints which can be explored for gait recognition with multi-modal data.

In summary, the contributions of this paper are as follows:

- We make one of the first attempts towards 3D gait recognition in the real-world scenario, which aims to explore dense 3D representations of the human body for gait recognition.
- We propose a novel 3D gait recognition framework based on the SMPL model, named SMPLGait, to explore 3D human meshes for gait recognition.
- We build the first large-scale 3D gait recognition dataset, named Gait3D, which provides the 3D human meshes of gait collected from unconstrained scenarios.

Through comprehensive experiments, we not only evaluate existing 2D silhouettes/skeleton-based approaches but also demonstrate the effectiveness of the proposed SMPLGait method, which reflects the potential of 3D representations for gait recognition. Moreover, the combination of 3D and 2D representations further improves the performance which shows the complementarity of multi-modal representations.

2. Related Work

Gait Recognition. We review the 2D and 3D representations-based gait recognition methods separately.

2D gait recognition methods can be classified into model-based and model-free approaches [41]. Early methods mainly belong to the model-based which defines a structural human body model. Then, gait patterns are modeled by parameters like lengths of limbs, angles of joints, and relative positions of body parts [3, 46]. The model-free methods mainly adopt the silhouettes obtained by background subtraction from video frames [5, 8, 10, 14, 15, 21, 31, 44, 55, 56]. In particular, Han *et al.* proposed to aggregate a sequence of silhouettes into a compact Gait Energy Image (GEI) [10] which was widely used by following methods [31, 44]. Recently, due to the success of deep learning for computer vision tasks [23–25, 48–52], deep Convolutional Neural Networks (CNNs) also dominated the performance of gait recognition. For example, Shiraga *et al.* [31] and Wu *et al.* [44] proposed to learn effective features from GEIs by and significantly outperformed previous methods. The most recent methods started to learn discriminative features directly from the silhouette sequences using larger CNNs or multi-scale structures and achieved the state-of-the-art results [5, 8, 15, 21]. Despite the excellent performance on in-the-lab datasets, e.g., CASIA-B and OU-LP, these methods usually fail in the wild as shown in the experiments on GREW [61] and our Gait3D.

3D representations has also been studied since the early years of gait recognition. For example, Urtasun and Fua [40] proposed an approach to gait analysis that depended on 3D temporal motion models using an articulated skeleton. Zhao *et al.* [57] applied a local optimization algorithm to track 3D motion for gait recognition. Yamauchi *et al.* [47] proposed the first method using 3D pose estimated from RGB frames for walking human recognition. Ariyanto and Nixon [3] built a 3D voxel-based dataset using a complex multi-camera system

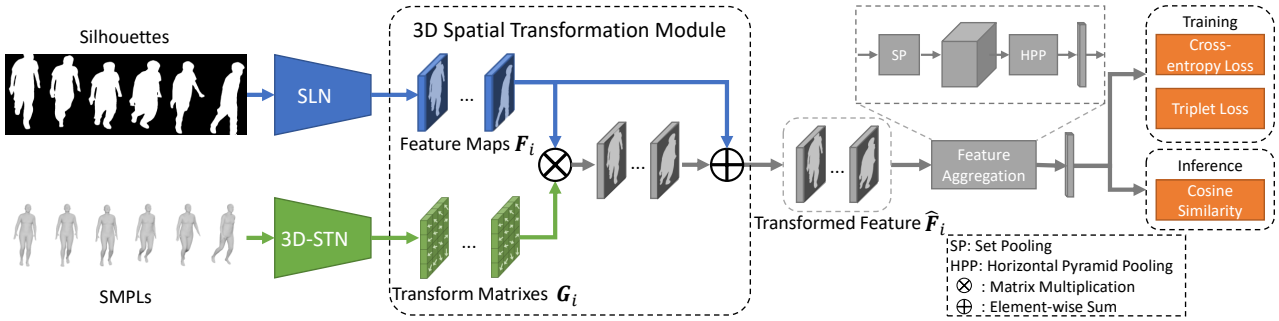


Figure 2. The architecture of the SMPLGait framework for 3D gait recognition in the wild.

and proposed a structural model of articulated cylinders with 3D Degrees of Freedom at each joint to model the human lower legs. However, these methods either discard rich 3D information like viewpoints and shapes or are limited by devices for real-world applications. In summary, to overcome the problem of 2D methods and explore 3D representations for gait recognition in the wild, we aim to explore 3D mesh as a rich representation with the viewpoint and shape of the human body.

Gait Recognition Datasets. Current publicly available gait recognition datasets mainly belong to two series, i.e., the CASIA series [36, 43, 54] and the OU-ISIR series [1, 13, 16, 28, 38, 39, 45] as listed in Table 1. The CASIA series were built in the early research of gait recognition, which facilitated the initial exploration of RGB images and silhouettes for gait representations [10, 36]. Despite its smaller number of subjects, the CASIA-B [54] is still the most widely used dataset for the evaluation of silhouette-based methods. The OU-ISIR series were first built ten years ago and developed comprehensive variants such as walking with different speeds [38], clothing styles [13], and bags [39], subjects of different ages [45], and annotations of 2D pose [1]. Due to their large population, the OU-LP [16] and OU-MVLP [35] also became the most popular datasets for current research. However, the above datasets were collected in constrained scenes like labs [16, 54] or a small defined area in a campus [30,43]. Most recently, researchers started to narrow the gap between in-the-lab research and real-world application. As a contemporaneous study of our work, Zhu *et al.* [61] constructed the GREW dataset from natural videos collected in an open area. However, there is no dataset that provides rich 3D representations for gait recognition in the wild. Therefore, we need to build a new dataset that is collected from complex scenes and with dense 3D meshes for gait recognition in the wild.

3D Human Mesh Recovery. 3D representations has attracted a lot of attention in the computer vision community [26, 59]. The 3D human body can be represented by point clouds [22], voxels [40], parameterized blend shape [2], etc. Among them, the Skinned Multi-Person Linear (SMPL) model [27] is a skinned vertex-based model that can accurately represent a wide variety of body shapes in natural human poses. With the SMPL model, an arbitrary 3D human body can be represented by a linear combination of a group of shape, pose, scale, and viewpoint parameters. Based on the SMPL model, a series of 3D human mesh recovery approaches are developed to estimate accurate 3D shapes, poses, and viewpoints of human bodies from natural images [17, 19, 33, 34]. These methods provide us an opportunity to obtain 3D human meshes from in-the-wild videos for 3D mesh-based gait recognition.

3. The 3D Gait Recognition Method

3.1. Overview

The overall architecture of the proposed 3D SMPL-based Gait Recognition framework, SMPLGait, is shown in Figure 2. There are two branches of the framework. For the first branch, we take the sequence of silhouettes as input which has a rich knowledge of the appearance and use a CNN-based model to extract 2D spatial features from each frame. For the second branch, the SMPLs of the human body are fed into the 3D-Spatial-Transformation Network (3D-STN), which aims to learn the latent transformation matrixes from the 3D viewpoints and shapes. Then the 3D Spatial Transformation Module aligns the 2D appearance features in the latent space using the learned transformation matrixes. Finally, the transformed feature of each frame is aggregated into a sequence-level feature for sequence-to-sequence matching in training or inference. Next, we will introduce the above modules in detail.

3.2. Network Structure

The Silhouette Learning Network (SLN) aims to learn the appearance knowledge of humans from silhouettes that contain 2D spatial information like clothing and hairstyle. The SLN has six convolutional layers which are similar to the backbone of GaitSet [5]. As is shown in Figure 2, the sequences of silhouettes are fed into a CNN. We formulate $X_{sil} = \{\mathbf{x}_i\}_{i=1}^{L}$ as the input sequence, where $\mathbf{x}_i \in \mathbb{R}^{H \times W}$

Dataset	Year	Subject #	Seq #	Cam #	Data Type	Speed	Wild	3D-View
CASIA-A [43]	2003	20	240	3	RGB, Silh.	✗	✗	✗
USF HumanID [30]	2005	122	1,870	2	RGB	✗	✗	✗
CASIA-B [54]	2006	124	13,640	11	RGB, Silh.	✗	✗	✗
CASIA-C [36]	2006	153	1,530	1	Infrared, Silh.	✓	✗	✗
OU-ISIR Speed [38]	2010	34	306	1	Silh.	✓	✗	✗
OU-ISIR-LP [16]	2012	4007	31,368	2	Silh.	✗	✗	✗
OU-LP Bag [39]	2018	62,528	187,584	1	Silh.	✗	✗	✗
OU-MVLP [35]	2018	10,307	288,596	14	Silh.	✗	✗	✗
OU-MVLP Pose [1]	2020	10,307	288,596	14	2D Pose	✗	✗	✗
GREW [61]	2021	26,345	128,671	882	Silh., 2D/3D Pose, Flow	✗	✓	✗
Gait3D	-	**4,000**	**25,309**	**39**	**Silh., 2D/3D Pose, 3D Mesh&SMPL**	✓	✓	✓

Table 1. Comparison of publicly available datasets for gait recognition. Speed, Wild, and 3D-View indicate whether the dataset contains inconstant walking speed, is captured in the wild, and has viewpoint variations in the 3D space, respectively.

is the i-th binary frame, L is the length of the sequence, H and W are the height and width of the silhouette image. For a frame $\mathbf{x}_i$, the process can be formulated as:

$$\mathbf{F}_i = F(\mathbf{x}_i), \tag{1}$$

where $F(\cdot)$ is the CNN-based backbone and $\mathbf{F}_i \in \mathbb{R}^{h\times w}$ is the frame-level feature map for frame $\mathbf{x}_i$ [1].

The 3D Spatial Transformation Network (3D-STN) is proposed to solve viewpoint changes in real 3D scenarios. 3D SMPL parameters related to 3D viewpoints, shapes, and poses are the input of this module. Assuming $Y_{sp} = \{y_i\}_{i=1}^{L}$ is the input SMPLs, where $\mathbf{y}_i \in \mathbb{R}^D$ is the SMPL vector of i-th frame, D is the dimension of the SMPL vector which contains 24×3 dimensions of 3D human body pose, 10 dimensions of 3D body shape, and 3 dimensions of camera scale and translation parameters. The 3D-STN consists of three fully connected (FC) layers with neuron number $= 128 \Rightarrow 256 \Rightarrow h \times w$, where h and w are the height and width of the feature map from the Silhouette Learning Network. Each FC layer is followed by batch normalization and the ReLU activation function. We use dropout for the last two FC layers to eliminate overfitting. The forward process of 3D-STN can be formulated as:

$$\mathbf{g}_i = G(\mathbf{y}_i), \tag{2}$$

where $G(\cdot)$ is the 3D-STN and $\mathbf{g}_i$ is the frame-level transformation vector for frame i.

The 3D Spatial Transformation Module is designed to align the 2D appearance feature map $\mathbf{F}_i \in \mathbb{R}^{h\times w}$ using the transformation vector $\mathbf{g}_i$ in the feature space, as shown in Figure 2. We first reshape the transformation vector $\mathbf{g}_i$ to a matrix $\mathbf{G}_i \in \mathbb{R}^{w\times h}$. Then, for convenience of computation, we expand $\mathbf{F}_i$ and $\mathbf{G}_i$ to square matrixes by zero padding on the short edge. After that, we apply $\mathbf{G}_i$ to $\mathbf{F}_i$ by

$$\widehat{\mathbf{F}}_i = \mathbf{F}_i \cdot (I + \mathbf{G}_i), \tag{3}$$

[1]For convenience of notation, we omit the channel of the feature map.

where I is an identity matrix and $\cdot$ is matrix multiplication. At last, we adopt Set Pooling (SP) and Horizontal Pyramid Pooling (HPP) in GaitSet [5] to aggregate $\widehat{\mathbf{F}}_i$ into the final feature vector for sequence-to-sequence matching. For more details of the SMPLGait framework, please refer to **the supplementary material**.

3.3. Training and Inference

Our two-branch 3D gait recognition framework is trained in an end-to-end manner. The network of our framework is optimized by a loss function with two components:

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

where L_{tri} is the triplet loss, L_{ce} is the cross entropy loss. α and β are the weighting parameters.

During inference, we use the sequences of silhouettes and SMPLs as the inputs of the two branches, respectively. The cosine similarity is used to measure the similarity between a query-gallery pair.

4. The Gait3D Benchmark

To facilitate the research of 3D gait recognition, we present a novel large-scale dataset, named Gait3D, which has several featured properties compared to existing datasets in Table 1. First of all, the Gait3D dataset consists of 4,000 subjects, 25,000+ sequences, and over 3 million bounding boxes captured by cameras of arbitrary 3D viewpoints, which makes it more scalable for training deep CNNs. Moreover, it provides accurate 3D human meshes estimated from video frames, which contains the poses and shapes of human bodies as well as viewpoints in the 3D space. Furthermore, Gait3D also provides 2D silhouettes and 2D/3D keypoints obtained by the state-of-the-art image segmentation and pose estimation methods fine-tuned on our dataset. Therefore, multi-modal data can be explored for gait recognition. In addition, Gait3D is collected in a large supermarket in which people usually walk at irregular speeds and routes, and can be occluded by other people or

objects. The above properties also make Gait3D a scalable but challenging dataset for gait recognition which can be reflected by the evaluation in Section 5.

4.1. Data Collection and Pre-processing

To collect a high-quality in-the-wild dataset for real applications, we collect the seven-day raw videos from 39 cameras mounted in a large supermarket. The scenes of the cameras include the entrance, the goods shelf area, the freezer area, the dining area, the checkout counter, etc. For the videos each day, we randomly sample two segments of continuous two-hour videos. At last, we obtain about 1,090 hours videos with 1,920 $\times$ 1,080 resolution and 25 FPS. Note that, we are authorized by the management of the supermarket to access and process the data for research purposes. In addition, all subjects were noticed that the data is collected only for research purposes. With the videos, we use the open-source FFmpeg [2] to decode the raw videos into frames at 25 FPS to keep the continuity of gait sequences. To guarantee the high quality of the dataset, the annotation process is performed by three main steps as follows.

4.2. Dataset Construction

4.2.1 Person detection and tracking from frames

For each frame extracted from the raw videos, we adopt the CenterNet [60] fine-tuned on our dataset as the person detector since it is an efficient anchor-free object detector [3]. To achieve accurate person tracking in videos, we exploit the Intersection-over-Union (IoU) and person re-identification (ReID) features of bounding boxes in two adjacent frames to measure their similarity. The ReID feature is extracted by an open-source person ReID framework, FastReID [4] [11] pretrained on several public person ReID datasets. When two persons are highly overlapped, the tracking algorithm can easily misjudge them as one person, i.e., ID switching. To solve this problem, we employ human annotators to clean sequences that may contain more than one pedestrian. By this means, we guarantee that each sequence only belongs to one person. Then, we discard the sequences shorter than 25 frames or longer than 500 frames and obtain about 50,000 sequences in total.

4.2.2 Cross-camera sequence matching

With the above sequences, we should cluster the sequences of the same person in all cameras. To achieve effective and efficient cross-camera matching of the same person, we also utilize the person ReID features obtained by FastReID [11]. For each sequence, we first use a pose estimation model, i.e., HRNet [5] [42] fine-tuned on our dataset [6], to select a high-quality frame for cross-camera matching. After that, we utilize FastReID to extract the features of the selected frames of all sequences. Through an unsupervised clustering method, i.e., DBSCAN [7], we roughly obtain 5,336 clusters of sequences. Then, we employ human annotators to filter out the outlier sequences in each group. By discarding the groups containing only one sequence, we finally obtain 4,000 subjects and 25,309 sequences for generating the gait representations.

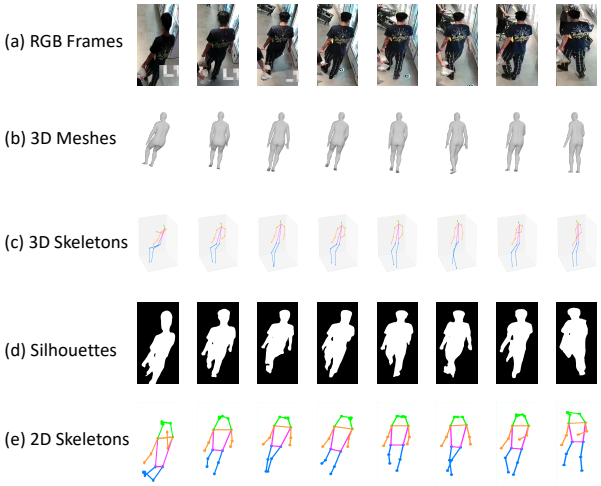


Figure 3. Examples of gait representations in the Gait3D dataset. The sizes are normalized for visualization. (Best viewed in color.)

4.2.3 Generation of gait representations

With the clean sequences of 4,000 IDs, we generate the 3D SMPL parameter, 3D mesh, 3D pose, 2D silhouette, and 2D pose for each frame. For the 3D SMPL, 3D mesh, and 3D pose, we exploit a state-of-the-art 3D human mesh recovery method, ROMP [7] [33], since it can efficiently output these three representations in an end-to-end framework. For the 2D silhouette, we use the semantic segmentation method, HRNet-segmentation [8] [42], to obtain the silhouette of the person in each frame. For the 2D pose, we also utilize the HRNet to estimate the 2D keypoints of the person in each frame. We keep the original resolution and aspect ratio of the frame without resizing or normalization. Some examples of gait representations in our dataset are shown in Figure 3. It is worth noting that we will only release the generated gait representations but not release any RGB frames to protect the privacy of the subjects.

[2] http://ffmpeg.org/ under the GNU LGPL License v2.1.

[3] 4,000 person bounding boxes are labeled for fine-tuning the detector.

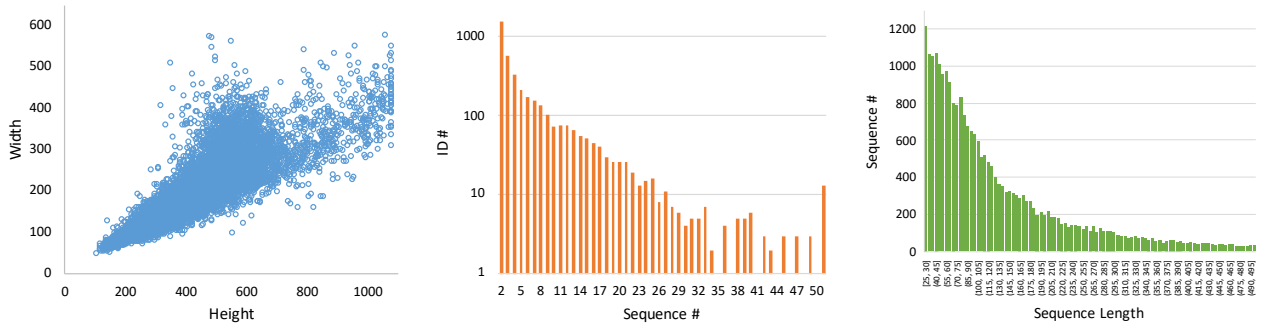
[4] https://github.com/JDAI-CV/fast-reid under the Apache 2.0 license.

[5] https://github.com/HRNet/HRNet-Human-Pose-Estimation under the MIT License.

[6] 4,000 images are labeled to fine-tune the pose estimator.

[7] https://github.com/Arthur151/ROMP under the MIT License.

[8] https://github.com/HRNet/HRNet-Semantic-Segmentation under the MIT license.



(a) Statistics of frame sizes (b) ID # over sequence # (c) Sequence # over sequence lengths

Figure 4. Statistics about the Gait3D dataset.

4.3. Dataset Statistics and Evaluation Protocol

The statistics about the sizes of frames, ID numbers over sequence numbers, and sequence numbers over sequence lengths are shown in Figure 4. From Figure 4 (a), we can find that most frames range from $100 \sim 400 \times 200 \sim 800$ which are larger than person bounding boxes of existing datasets. Figure 4 (b) shows that most IDs have $2 \sim 25$ sequences, which guarantees the high reappearance times of subjects. Figure 4 (c) reflects that most sequences are longer than 50 frames (2 seconds) and the longest sequence has 500 frames, which reflects the complexity of the gait sequences in the unconstrained scenes. The above statistics demonstrate that the Gait3D dataset is scalable but challenging for gait recognition research.

To facilitate the research, we split the 4,000 IDs of the Gait3D dataset into the train/test subsets with 3,000/1,000 IDs, respectively. For the test set, we further randomly select one sequence from each ID to build the query set with 1,000 sequences, while the rest of the sequences become the gallery set with 5,369 sequences. Our evaluation protocol is based on the open-set instance retrieval setting like existing gait recognition datasets [16] and the person ReID task [58]. Given a query sequence, we measure its similarity between all sequences in the gallery set. Then a ranking list of the gallery set is returned by the descending order of the similarities. We report the average Rank-1 and Rank-5 identification rates over all query sequences. We also adopt the mean Average Precision (mAP) and mean Inverse Negative Penalty (mINP) [53] which consider the recall of multiple instances and hard samples.

5. Experiments

In the experiments, we first evaluate several State-Of-The-Art (SOTA) 2D gait recognition methods and our SMPLGait on the Gait3D dataset. Then, we analyze the influence of the frame size, the sequence length, and the scale of training IDs on the performance of gait recognition.

5.1. Evaluation of Existing Methods

Here, we evaluate eight SOTA 2D gait recognition methods including six model-free methods and two model-based methods. We also compare our 3D gait recognition method (SMPLGait) with these methods.

5.1.1 Model-free Approaches

The details of model-free approaches are as follows:

1) GEINet [31] is one of the first methods that adopts a four-layer CNN to learn gait features from GEIs using the cross-entropy loss.

2) GaitSet [5] is the representative method that utilizes a 10-layer CNN to directly learn discriminative gait features from silhouette sequences. The GaitSet is trained by the batch all triplet loss [12].

3) GaitPart [8] adopts the idea of multi-scale feature learning. It horizontally divides a silhouette image into fixed parts to learn discriminative micro-motion features.

4) GLN [14] is an efficient and effective method to learn compact features from gait sequences, which achieves the SOTA performance using only a 256-D feature.

5) GaitGL [21] is also a CNN-based framework to learn both global and local features from gait sequences.

6) CSTL [15] applies multi-scale learning on the temporal dimension of the sequence to learn both long-term and short-term motion for gait recognition.

Implementation Details: During training, we train the above models except GLN with the same configuration. The batch size is $32 \times 4 \times 30$, where 32 denotes the number of IDs, 4 denotes the number of training samples per ID, and 30 is the sequence length. The models are trained for 1,200 epochs with the initial Learning Rate (LR)=1e-3 and the LR is multiplied by 0.1 at the 200-th and 600-th epochs. The optimizer is Adam [18] and the weight decay is set to 5e-4. For GLN, we follow the two-stage training as in [14]. The model trained in the first stage is used as the pre-trained model of the second stage. Both of the two stages are trained with the same configuration of other methods. During testing, we use the cosine similarity to measure the

Input Size (W×H)		88×128				44×64			
Methods	Publication	R-1 (%)	R-5 (%)	mAP (%)	mINP	R-1 (%)	R-5 (%)	mAP (%)	mINP
GEINet [31]	ICB 2016	7.00	16.30	6.05	3.77	5.40	14.20	5.06	3.14
GaitSet [5]	AAAI 2019	42.60	63.10	33.69	19.69	36.70	58.30	30.01	17.30
GaitPart [8]	CVPR 2020	29.90	50.60	23.34	13.15	28.20	47.60	21.58	12.36
GLN [14]	ECCV 2020	42.20	64.50	33.14	19.56	31.40	52.90	24.74	13.58
GaitGL [21]	ICCV 2021	23.50	38.50	16.40	9.20	29.70	48.50	22.29	13.26
CSTL [15]	ICCV 2021	12.20	21.70	6.44	3.28	11.70	19.20	5.59	2.59
PoseGait [20]	PR 2020	0.24	1.08	0.47	0.34	-	-	-	-
GaitGraph [37]	arXiv 2021	6.25	16.23	5.18	2.42	-	-	-	-
SMPLGait w/o 3D	Ours	47.70	67.20	37.62	22.24	42.90	63.90	35.19	20.83
SMPLGait	Ours	**53.20**	**71.00**	**42.43**	**25.97**	**46.30**	**64.50**	**37.16**	**22.23**

Table 2. Comparison of the state-of-the-art gait recognition methods on Gait3D. As the inputs of the model-based methods, i.e., PoseGait and GaitGraph, are unrelated to the frame size, we only report one group of results.

similarity between each pair of query and gallery sequences. For the GaitSet, GaitPart, GLN, and GaitGL models, we adopt the implementations in the open-source OpenGait toolbox [9] since they outperform the original codes.

5.1.2 Model-based Approaches

We compare two representative model-based methods which use 2D or 3D skeletons as the input.

1) PoseGait [20] first exploits OpenPose [4] to extract the 2D keypoints from RGB frames, then uses the method in [6] to estimate the 3D keypoints of human bodies. Based on the 3D skeletons, it defines several parameters such as joint angle, limb length, and joint motion together with the pose features as the gait representation. In our implementation, we train it for 700 epochs with a batch size of 128. The LR is set to 1e-3. The optimizer is Adam [18] and weight decay is equal to 5e-4.

2) GaitGraph [37] is a recent model-based gait recognition method. It models the 2D skeleton as a graph and adopts a Graph Convolution Network, i.e., the ResGCN [32], to learn features by the contrastive loss. We train GaitGraph in two stages. The setting of the first stage is the same as PoseGait, and the model trained in the first stage is used as the pre-trained model of the second stage. In the second stage, we fine-tune it for 250 epochs.

5.1.3 Implementation Details of the SMPLGait

For our SMPLGait, we use the loss in Equ. 4 for training. In 3D-STN, we set the dropout rate to 0.2 for FC layers. The hyper-parameters in Equ. 4 are set as α=1.0 and β=0.1. Other settings are the same as those in Section 5.1.1.

5.1.4 Experimental Results

The results of model-free methods, model-based methods, and our SMPLGait are listed in Table 2.

For model-free methods, we can first observe that the overall performance of the SOTA methods is much worse than their performance on in-the-lab datasets like the CASIA-B [54] and OU-ISIR series [16, 35]. This reflects that there is a huge gap between the in-the-lab research and the in-the-wild application that is much more challenging. Meanwhile, the performance of the SOTA model-free methods varies significantly. For example, the GEI-based method, i.e., GEINet obtains the worst results, which indicates that the GEIs discard too much useful information for gait recognition. Moreover, the methods considering the order of the frames in sequences, i.e., GaitPart, GLN, GaitGL, and CSTL, obtain lower accuracy. It means that the temporal information in the wild scene is hard to learn, because people may stop then continue to walk with varying speeds and routes in unconstrained scenarios. On the contrary, the methods considering frames as an unordered set, i.e., GaitSet, obtain better results.

For model-based methods, we can find that they are greatly worse than model-free methods on the Gait3D dataset. This is because the input of the model-based methods only has a few sparse human body joints, which seriously lacks useful gait information, such as body shape, appearance, and so on. In addition, the walking speed and route are uncertain in real scenarios, which also greatly affects the performance of the model-based methods that aim to model the temporal dynamics of the human body.

Finally, our SMPLGait outperforms other methods by a large margin, which indicates the potential of 3D representations for gait recognition in the wild.

5.1.5 Ablation Study of SMPLGait

We also conduct an ablation study on the key components in SMPLGait by removing the 3D branch (SMPLGait w/o 3D). [10] The results are listed in Table 2. This comparison shows that the integration of 2D and 3D representations can better address the challenges of gait recognition in the wild.

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

[10] It should be noticed that SMPLGait w/o 3D is equal to OpenGait Baseline [9]

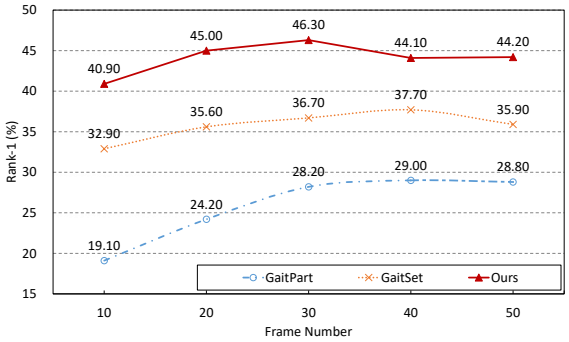


Figure 5. The effect of frame numbers in sequences.

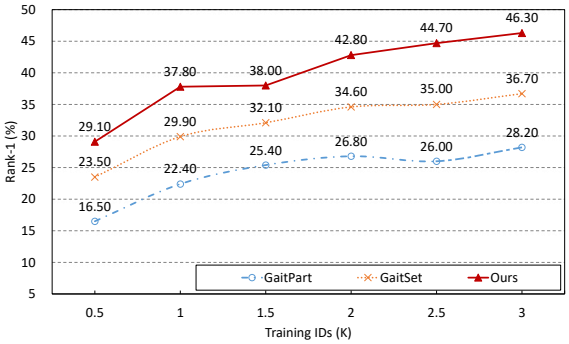


Figure 6. The effect of different training ID numbers.

5.2. More Analysis on the Gait3D Dataset

We choose two SOTA gait recognition methods, i.e., GaitPart and GaitSet, and our SMPLGait (Ours) to analyze the influence of input size, frame number in sequences, and training ID number to the accuracy. All the models are evaluated on the whole Gait3D test set.

Input Size. We explore two input sizes of 88×128 and 44×64 for the compared methods, as shown in Table 2. From the results, we can observe that the performance of almost all methods is improved with larger input size. There is an exception, i.e., GaitGL, which obtain worse accuracy with larger input size. This may be because that GaitGL adopts the 3D CNN as the backbone. When using a larger input size, the 3D CNN learns more misalignment information about frames in physical space, which makes it more difficult to be optimized.

Number of Training Frames We randomly sample $10\sim 50$ frames from original gait sequences during training. The Rank-1 accuracy is illustrated in Figure 5. The results show that as the number of frames increases the performance first increases and then decreases, while the best performance occurs around 30 frames per sequence. This indicates that more frames could not bring higher accuracy. The reason may be that there is a lot of redundant or noisy information caused by uncertain speeds and routes of persons, which will bring ambiguous features for gait recognition.

Scale of Training IDs We fix other settings and use 0.5K $\sim$ 3KIDs with an increment of 0.5K for training. As shown in Figure 6, the performance of the models grows stably with more training IDs. These results reflect the scalability of our Gait3D dataset.

More experiments and exemplar results on Gait3D can be found in **the supplementary material**.

6. Discussion

Ethical Issues. There are two main ethical issues of this paper: 1) privacy, and 2) data bias. For the first issue, we will try our best to protect the privacy of the subjects involved in our dataset. Firstly, we will not release any human cognizable data like original videos, RGB frames, and bounding boxes of persons. Second, the dataset will be distributed only for research purposes via the case-by-case application with a strict license. To eliminate data bias, the genders and ages of subjects are relatively balanced.

Future Work. Despite the proposed baseline method for 3D gait recognition, there are many potential directions for this challenging task. For example, one direction is to study how to design a deep CNN for learning more discriminative features directly from 3D meshes. The second direction is how to learn the temporal information of gait representation, because the walking speed and route in Gait3D are irregular, it is significantly different from the datasets built in the lab. Another interesting direction is how to fuse the multi-modal information like silhouette, 2D/3D skeleton, and 3D mesh for gait recognition in the wild.

More discussions about the limitations and potential negative impact can be found in **the supplementary material**.

7. Conclusion

Gait recognition in the wild faces significant challenges such as extreme viewpoint changes, occlusions of the human body, and complex clutter in the environment. Existing methods using 2D silhouettes or skeletons will fail in the wild because crucial information like 3D viewpoints and shapes of human bodies is discarded. Therefore, this paper proposes a 3D SMPL model-based framework (SMPLGait) which is the first method to explore dense 3D representations for gait recognition in the wild. To facilitate the research, we build the first large-scale 3D gait recognition dataset (Gait3D) from cameras deployed in a large supermarket. It provides diverse gait representations including 3D meshes, 3D SMPLs, 3D poses, 2D silhouettes, and 2D poses for over 25,000 gait sequences of 4,000 subjects. We hope Gait3D can provide researchers with a new perspective of gait recognition.

Acknowledgements. This work was supported in part by the National Key Research and Development Program of China under Grant 2020AAA0103800, in part by the National Nature Science Foundation of China under Grant 61931008 and Grant U21B2024.

References

[1] Weizhi An, Shiqi Yu, Yasushi Makihara, Xinhui Wu, Chi Xu, Yang Yu, Rijun Liao, and Yasushi Yagi. Performance evaluation of model-based gait on multi-view very large population database with pose sequences. *IEEE TBBIS*, 2(4):421–430, 2020. 3, 4

[2] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. SCAPE: shape completion and animation of people. *ACM TOG*, 24(3):408–416, 2005. 3

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

[4] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE TPAMI*, 43(1):172–186, 2021. 7

[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, 3, 4, 6, 7

[6] Ching-Hang Chen and Deva Ramanan. 3d human pose estimation = 2d pose estimation + matching. In *CVPR*, pages 5759–5767, 2017. 7

[7] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, pages 226–231, 1996. 5

[8] 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 14213–14221, 2020. 1, 2, 6, 7

[9] Chao Fan, Chuanfu Shen, Junhao Liang, and Shiqi Yu. OpenGait. https://github.com/ShiqiYu/OpenGait. 7

[10] Ju Han and Bir Bhanu. Individual recognition using gait energy image. *IEEE TPAMI*, 28(2):316–322, 2006. 1, 2, 3

[11] Lingxiao He, Xingyu Liao, Wu Liu, Xinchen Liu, Peng Cheng, and Tao Mei. FastReID: A pytorch toolbox for general instance re-identification. *CoRR*, abs/2006.02631, 2020. 5

[12] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *CoRR*, abs/1703.07737, 2017. 6

[13] Md. Altab Hossain, Yasushi Makihara, Junqiu Wang, and Yasushi Yagi. Clothing-invariant gait identification using part-based clothing categorization and adaptive weight control. *PR*, 43(6):2281–2291, 2010. 3

[14] Saihui Hou, Chunshui Cao, Xu Liu, and Yongzhen Huang. Gait lateral network: Learning discriminative and compact representations for gait recognition. In *ECCV*, pages 382–398, 2020. 2, 6, 7

[15] 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. 2, 6, 7

[16] Haruyuki Iwama, Mayu Okumura, Yasushi Makihara, and Yasushi Yagi. The OU-ISIR gait database comprising the large population dataset and performance evaluation of gait recognition. *IEEE TIFS*, 7(5):1511–1521, 2012. 3, 4, 6, 7

[17] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, pages 7122–7131, 2018. 2, 3

[18] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 6, 7

[19] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. VIBE: video inference for human body pose and shape estimation. In *CVPR*, pages 5252–5262, 2020. 2, 3

[20] Rijun Liao, Shiqi Yu, Weizhi An, and Yongzhen Huang. A model-based gait recognition method with body pose and human prior knowledge. *PR*, 98, 2020. 7

[21] 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. 2, 6, 7

[22] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C. Kot. NTU RGB+D 120: A large-scale benchmark for 3d human activity understanding. *IEEE TPAMI*, 42(10):2684–2701, 2020. 3

[23] Wu Liu, Qian Bao, Yu Sun, and Tao Mei. Recent advances in monocular 2d and 3d human pose estimation: A deep learning perspective. *ACM Computing Surveys*, 2022. 2

[24] Xinchen Liu, Wu Liu, Huadong Ma, and Huiyuan Fu. Large-scale vehicle re-identification in urban surveillance videos. In *ICME*, pages 1–6, 2016. 2

[25] Xinchen Liu, Wu Liu, Tao Mei, and Huadong Ma. PROVID: progressive and multimodal vehicle reidentification for large-scale urban surveillance. *IEEE TMM*, 20(3):645–658, 2018. 2

[26] Xueyi Liu, Xiaomeng Xu, Anyi Rao, Chuang Gan, and Li Yi. Autogpart: Intermediate supervision search for generalizable 3d part segmentation. *CoRR*, abs/2203.06558, 2022. 3

[27] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: a skinned multi-person linear model. *ACM TOG*, 34(6):248:1–248:16, 2015. 2, 3

[28] Yasushi Makihara, Hidetoshi Mannami, and Yasushi Yagi. Gait analysis of gender and age using a large-scale multi-view gait database. In *ACCV*, pages 440–451, 2010. 3

[29] Sourabh A. Niyogi and Edward H. Adelson. Analyzing and recognizing walking figures in XYT. In *CVPR*, pages 469–474, 1994. 1

[30] Sudeep Sarkar, P. Jonathon Phillips, Zongyi Liu, Isidro Robledo Vega, Patrick Grother, and Kevin W. Bowyer. The HumanID gait challenge problem: Data sets, performance, and analysis. *IEEE TPAMI*, 27(2):162–177, 2005. 3, 4

[31] Kohei Shiraga, Yasushi Makihara, Daigo Muramatsu, Tomio Echigo, and Yasushi Yagi. GEINet: View-invariant gait recognition using a convolutional neural network. In *ICB*, pages 1–8, 2016. 2, 6, 7

[32] Yi-Fan Song, Zhang Zhang, Caifeng Shan, and Liang Wang. Stronger, faster and more explainable: A graph convolutional baseline for skeleton-based action recognition. In *ACM MM*, pages 1625–1633, 2020. 7

[33] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Michael J. Black, and Tao Mei. Monocular, one-stage, regression of multiple 3d people. In *ICCV*, pages 11179–11188, 2021. 2, 3, 5

[34] Yu Sun, Wu Liu, Qian Bao, Yili Fu, Tao Mei, and Michael J Black. Putting people in their place: Monocular regression of 3d people in depth. In *CVPR*, 2022. 3

[35] 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 TCVA*, 10:4, 2018. 1, 3, 4, 7

[36] Daoliang Tan, Kaiqi Huang, Shiqi Yu, and Tieniu Tan. Efficient night gait recognition based on template matching. In *ICPR*, pages 1000–1003, 2006. 1, 3, 4

[37] Torben Teepe, Ali Khan, Johannes Gilg, Fabian Herzog, Stefan Hörmann, and Gerhard Rigoll. GaitGraph: Graph convolutional network for skeleton-based gait recognition. *CoRR*, abs/2101.11228, 2021. 7

[38] Akira Tsuji, Yasushi Makihara, and Yasushi Yagi. Silhouette transformation based on walking speed for gait identification. In *CVPR*, pages 717–722, 2010. 3, 4

[39] Md. Zasim Uddin, Trung Ngo Thanh, Yasushi Makihara, Noriko Takemura, Xiang Li, Daigo Muramatsu, and Yasushi Yagi. The OU-ISIR large population gait database with real-life carried object and its performance evaluation. *IPSJ TCVA*, 10:5, 2018. 3, 4

[40] Raquel Urtasun and Pascal Fua. 3d tracking for gait characterization and recognition. In *FGR*, pages 17–22, 2004. 1, 2, 3

[41] Changsheng Wan, Li Wang, and Vir V. Phoha. A survey on gait recognition. *ACM CSUR*, 51(5):89:1–89:35, 2019. 1, 2

[42] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation learning for visual recognition. *IEEE TPAMI*, 43(10):3349–3364, 2021. 5

[43] Liang Wang, Tieniu Tan, Huazhong Ning, and Weiming Hu. Silhouette analysis-based gait recognition for human identification. *IEEE TPAMI*, 25(12):1505–1518, 2003. 3, 4

[44] Zifeng Wu, Yongzhen Huang, Liang Wang, Xiaogang Wang, and Tieniu Tan. A comprehensive study on cross-view gait based human identification with deep cnns. *IEEE TPAMI*, 39(2):209–226, 2017. 1, 2

[45] Chi Xu, Yasushi Makihara, Gakuto Ogi, Xiang Li, Yasushi Yagi, and Jianfeng Lu. The OU-ISIR gait database comprising the large population dataset with age and performance evaluation of age estimation. *IPSJ TCVA*, 9:24, 2017. 3

[46] Chew-Yean Yam, Mark S. Nixon, and John N. Carter. Automated person recognition by walking and running via model-based approaches. *PR*, 37(5):1057–1072, 2004. 2

[47] Koichiro Yamauchi, Bir Bhanu, and Hideo Saito. Recognition of walking humans in 3d: Initial results. In *CVPRW*, pages 45–52, 2009. 2

[48] Chenggang Yan, Biao Gong, Yuxuan Wei, and Yue Gao. Deep multi-view enhancement hashing for image retrieval. *IEEE TPAMI*, 43(4):1445–1451, 2021. 2

[49] Chenggang Yan, Yiming Hao, Liang Li, Jian Yin, Anan Liu, Zhendong Mao, Zhenyu Chen, and Xingyu Gao. Task-adaptive attention for image captioning. *IEEE TCSVT*, 32(1):43–51, 2022. 2

[50] Chenggang Yan, Zhisheng Li, Yongbing Zhang, Yutao Liu, Xiangyang Ji, and Yong-Dong Zhang. Depth image denoising using nuclear norm and learning graph model. *TOMM*, 16(4):122:1–122:17, 2021. 2

[51] Chenggang Yan, Lixuan Meng, Liang Li, Jiehua Zhang, Zhan Wang, Jian Yin, Jiyong Zhang, Yaoqi Sun, and Bolun Zheng. Age-invariant face recognition by multi-feature fusionand decomposition with self-attention. *TOMM*, 18(1s):1–18, 2022. 2

[52] Chenggang Yan, Tong Teng, Yutao Liu, Yongbing Zhang, Haoqian Wang, and Xiangyang Ji. Precise no-reference image quality evaluation based on distortion identification. *TOMM*, 17(3s):1–21, 2021. 2

[53] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven C. H. Hoi. Deep learning for person re-identification: A survey and outlook. *CoRR*, abs/2001.04193, 2020. 6

[54] 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, 7

[55] Shaoxiong Zhang, Yunhong Wang, and Annan Li. Cross-view gait recognition with deep universal linear embeddings. In *CVPR*, pages 9095–9104, 2021. 1, 2

[56] 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. 2

[57] Guoying Zhao, Guoyi Liu, Hua Li, and Matti Pietikäinen. 3d gait recognition using multiple cameras. In *FGR*, pages 529–534, 2006. 2

[58] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, pages 1116–1124, 2015. 6

[59] Xingyi Zhou, Arjun Karpur, Chuang Gan, Linjie Luo, and Qixing Huang. Unsupervised domain adaptation for 3d keypoint estimation via view consistency. In *ECCV (12)*, volume 11216, pages 141–157, 2018. 3

[60] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *CoRR*, abs/1904.07850, 2019. 5

[61] 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 *ICCV*, pages 14789–14799, 2021. 1, 2, 3, 4