The Journal of Supercomputing (2023) 79:12106–12125
https://doi.org/10.1007/s11227-023-05143-0

Gait recognition based on 3D human body reconstruction and multi-granular feature fusion

Chunyun Meng[1] · Xiaobing He[2] · Zhen Tan[1] · Li Luan[3]

Accepted: 23 February 2023 / Published online: 7 March 2023

Abstract

Gait recognition is a crucial video-based biometric approach that allows for the identification of pedestrians from the motion of their walk over a distance without direct contact. Despite significant advances in this field, most existing approaches for gait recognition rely on silhouette sequence extraction, which can result in redundant information when the behavior of pedestrians changes, such as with the addition of coats or bags. To alleviate this, we propose an end-to-end gait recognition method based on 3D human body reconstruction to effectively remove this redundant information and generate compact, discriminative gait representations. Furthermore, to make full use of the spatial characteristics of pedestrians, we propose a multi-granular feature fusion module to model gait representations at multiple granularities. Our method is evaluated on the Outdoor-Gait and CASIA-B datasets and shows improved performance and robustness.

Keywords Gait recognition · 3D reconstruction · Cross-condition · Multi-granular

Chunyun Meng and Xiaobing He have contributed equally.

✉ Zhen Tan
tanzhen208@163.com

Chunyun Meng
cymeng@stu.just.edu.cn

Xiaobing He
2221908017@stmail.ujs.edu.cn

Li Luan
luanli@mail.ustc.edu.cn

[1] School of Economics and Management, Jiangsu University of Science and Technology, Zhenjiang 212100, Jiangsu, China

[2] School of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang 212013, Jiangsu, China

[3] School of Public Affairs, University of Science and Technology of China, Hefei 230026, Anhui, China

1 Introduction

Vision-based biometric technology has made significant advancements in the computer vision community. Popular techniques include fingerprint recognition [1], vein biometrics [2], face identification [3], iris biometrics [4], and gait recognition [5]. Among these, gait recognition is a relatively new method that aims to identify individuals from a distance without any physical contact. This contactless and long-distance recognition approach has many advantages, such as the lack of need for cooperation, difficulty in camouflage, and strong adaptability to different environments. As such, it has great potential for use in medical motion analysis [6], security monitoring, criminal investigations, and other monitoring systems in the future. However, there are still many challenges to be addressed before gait recognition can be fully integrated into real-world applications.

Current gait recognition methods primarily focus on extracting features from the gait silhouette sequence, which can lead to a lack of local information in pedestrian contour segmentation, such as missing legs or feet in certain frames of a video. Additionally, clothing and accessories worn by pedestrians, such as coats and backpacks, can also negatively impact recognition performance. These additional factors not only obscure the pedestrian's walking posture but also add irrelevant information, which can greatly hinder subsequent learning, particularly in cross-condition recognition [7].

In order to deal with the issues of occlusion, clothing, and accessories, some researchers have proposed using human pose estimation networks to generate skeleton sequences for extracting gait features. While methods based on human pose estimation can be robust, they often fail to capture important visual information such as the details of the human body, resulting in poor recognition performance.

To address the issues previously mentioned, we propose an end-to-end gait recognition method based on 3D human body reconstruction. Our method generates a new gait contour sequence using a 3D human body reconstruction method. Usually, 3D view gait descriptor-based techniques [8] require a complex and costly setup of multiple calibrated cameras, limiting their use to controlled environments. However, our proposed method overcomes this limitation by allowing for 3D reconstruction directly from original video frames, eliminating the need for costly camera setups, and expanding applicability to a wider range of environments. In comparison with silhouette sequences, the 3D reconstruction method does not include any redundant information other than the body, which means that previous problems such as clothing and accessories will not affect the analysis. Additionally, the 3D reconstruction allows for the extraction of more informative features that can effectively reflect the pedestrian's gait. A visual comparison of different gait representations is shown in Fig. 1, where it can be observed that the gait information extracted from 3D human body reconstruction is clearer and more complete than that from silhouette sequences. Furthermore, to fully utilize the global and local spatial information of pedestrians, we propose a multi-granular feature fusion module which models temporal–spatial dependencies at multiple levels to achieve better representation ability. Our contributions can be summarized as follows:

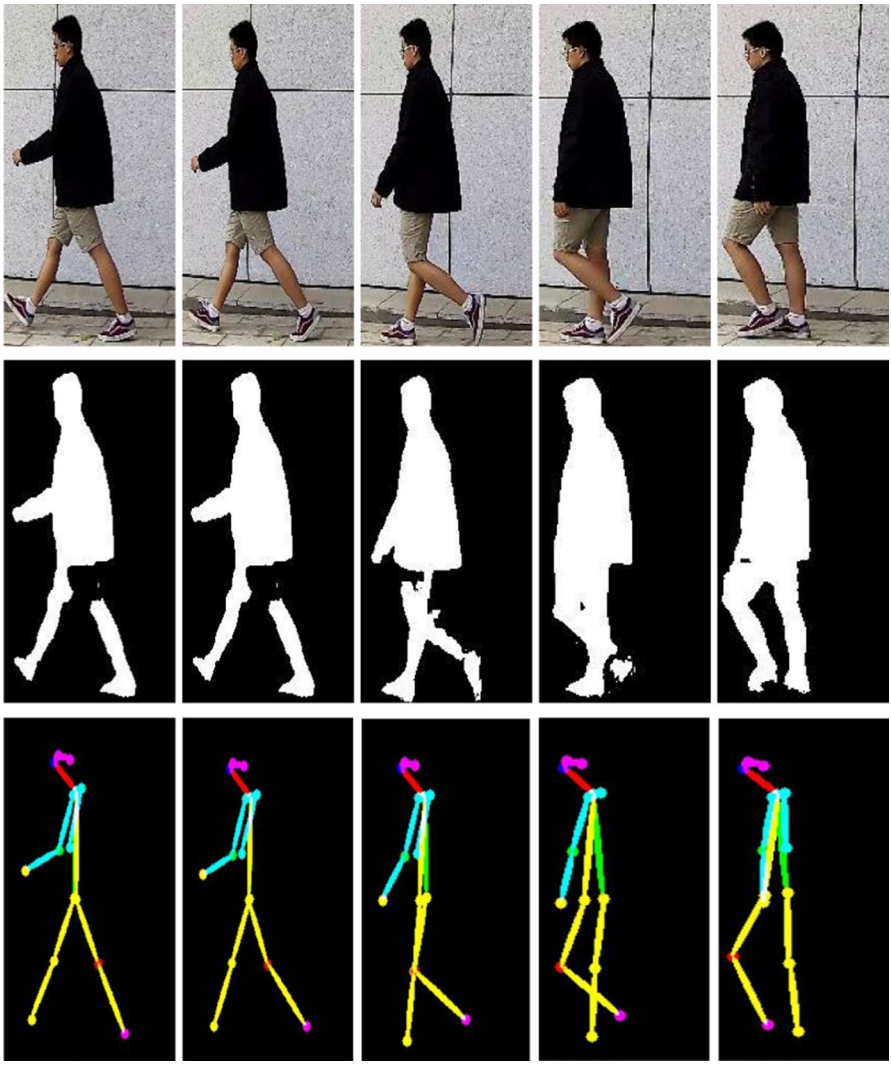

Fig. 1 Comparison of different gait representations. The first row is the original video frames, the second row is the silhouette images, and the third row is the pose sequence

- our proposed method leverages the power of 3D human body reconstruction to overcome the challenges posed by changes in pedestrian appearance and attire, such as coat wearing and bag carrying. Our approach generates a new gait contour sequence that contains information about the pedestrian's body, eliminating the need to consider irrelevant or redundant information. Unlike traditional methods that require a setup of multiple calibrated cameras or preprocessing of video streams, our model can be directly applied to the

original video frames. This greatly simplifies the gait recognition process and enhances its robustness and efficiency.

- To address the issue of underutilizing spatial features in gait recognition methods, we introduce a multi-granular feature fusion module that effectively captures the temporal–spatial information representation of pedestrians from both global and local perspectives. This allows for a more comprehensive understanding of the gait characteristics and helps in enhancing recognition performance.

2 Related works

2.1 Model-based approaches

2.1.1 Traditional gait recognition

The traditional gait recognition techniques mainly focus on utilizing information about the human body structure and the motion patterns of various body parts to identify gait characteristics. This information is then used to generate gait features for recognition purposes. For instance, Lee and Grimson [9] divided the pedestrian gait silhouette into 7 regions, each of which is fitted by an elliptic curve and then calculated the elliptic parameters as gait feature representation. Cunado et al. [10] considered that the leg motions follow the simple harmonic motion and then modeled this rule for gait recognition. In order to analyze the gait motion, Yoo et al. [11] utilized 2D stick shaped to represent the human body model and obtained the angle signals of various parts of the body through linear regression analysis. Yam et al. [12] used the pendulum model to guide the process of motion extraction. Urtasun et al. [13] extended the method of Cunado et al. [10] to 3D space and proposed a 3D human motion model based on principal component analysis (PCA) in order to overcome the influence of occlusion and motion direction changes. Dockstader et al. [14] proposed a hierarchical structure model which used a group of dots and lines to represent the human body and a periodic swing model to describe the gait pattern. Most of these traditional methods rely on specific environments and devices, such as fully controllable multi-camera collaborative environments, making such methods difficult to apply in practice. In contrast, our approach relies only on common cameras, greatly simplifying the constraint mention of recognition scenes.

2.1.2 Method based on RGB video frame

The methods for gait recognition based on RGB images can be separated into two categories: human pose estimation and 3D reconstruction. These techniques have garnered much attention in recent years and offer valuable insight into the field of gait recognition. By using human pose estimation instead of silhouette extraction, the gait recognition method based on human pose sequences represents a departure from traditional methods. Liao et al. [15] proposed a gait recognition method PTSN based on human pose sequences for the first time. It used the open-source pose estimation algorithm to extract human posture information from the original

video sequence. After obtaining the standardized gait pose sequence, it used a pose-based temporal–spatial network to learn gait feature representation. Inspired by the success of GCNs in skeleton-based action recognition, Teepe et al. [16] combined skeleton poses with graph convolution network (GCN) [17] to obtain a modern model-based gait recognition method. The gait recognition methods based on pose estimation ignore the information of human body shape, which reduces the accuracy of gait recognition. To make up for the lack of body shape in human pose-based gait recognition methods, some researchers have started to try to replace human pose sequences with 3D human reconstruction. Li et al. [18] extracted pose and shape features by fitting the SMPL model and subsequently feed the pose and shape features to a recognition network. Several of the above methods do not take into account multiple perspectives, so Khan et al. [19] proposed a view-invariant gait representation for cross-view gait recognition using the temporal–spatial motion characteristics of walking conditions.

2.2 Appearance-based approaches

2.2.1 Gait recognition based on template

The process of constructing gait templates involves subtraction of the background and creation of a human contour through a weighted average of each frame. These templates come in various forms, including Gait Energy Image (GEI) [20], Gait Entropy Image (GEnI) [21], Gait Flow Image (GFI) [22], and Chrono-Gait Image (CGI) [23]. Currently, GEI is considered the simplest and most efficient among these gait template types. Gait recognition methods based on the template can fall into two categories. The first is to extract gait features for discrimination using traditional metric learning methods (e.g., linear discriminant analysis [20], tensor representation discriminant analysis [24], random subspace [25], combined intensity and spatial metric learning [26]), or deep neural network [27–30]. The second is to generate gait representations under different conditions into the same covariate conditions using subspace analysis methods [32–36] or generative adversarial networks (GANs) [38, 39]. The template-based gait recognition method takes a single image after weighted averaging as input and does not make full use of the temporal information of the video, while our method takes video frames as input and learns short-range temporal–spatial features through the motion capture module.

2.2.2 Method based on gait silhouette sequence

The methods based on gait silhouette sequence use the silhouette sequence as the input directly. It is divided into three categories based on the way of extracting temporal information: 3DCNN based [40, 41], LSTM based [42], and set based [43, 44]. The 3DCNN-based methods directly extract the temporal–spatial features of gait sequences through 3D convolution network, but these methods usually have more parameters and are difficult to train. Zhang et al. [42] proposed a new auto-encoder framework to extract gait-related features from the original RGB video and

used three-layer LSTM to model the temporal changes of gait sequence. However, the LSTM-based method is considered to retain the unnecessary constraints of periodic gait. To avoid this problem, GaitSet [43] assumed that the appearance of the silhouette contains its position information and proposed to take the gait as a set to extract temporal–spatial features in the way of temporal pooling, which is simple and effective. Further, based on GaitSet, GaitPart [44] designed a temporal–spatial model for each part of the human body, making full use of the part-level features of pedestrians. The silhouette-based gait recognition method uses silhouette sequences as input. The silhouette sequences not only lose local body information in the process of generation but also contain redundant information such as coats and backpacks, which has a negative impact on gait recognition, while the 3D reconstructed sequences can effectively remove these redundant information.

3 Proposed method

3.1 Overall framework

In this study, we propose an approach for gait recognition where the original video frames of a pedestrian are taken as input and the length of the gait sequence is 30. New gait contour sequences are generated using the Human Mesh Recovery (HMR, 3D human body reconstruction) [45]. The frame-level part feature extractor (FPFE) [44] is then used to extract pedestrian gait features on the gait contour sequences. The multi-granular feature fusion (MGFF) module is employed to model the temporal–spatial representations of pedestrians from multiple granularities based on the generated spatial convolution features. Subsequently, the full connection layer (FC) is utilized to produce column vectors for identifying instances. Finally, the entire network is trained using the triplet loss function. The overall framework of our approach is depicted in Fig. 2.

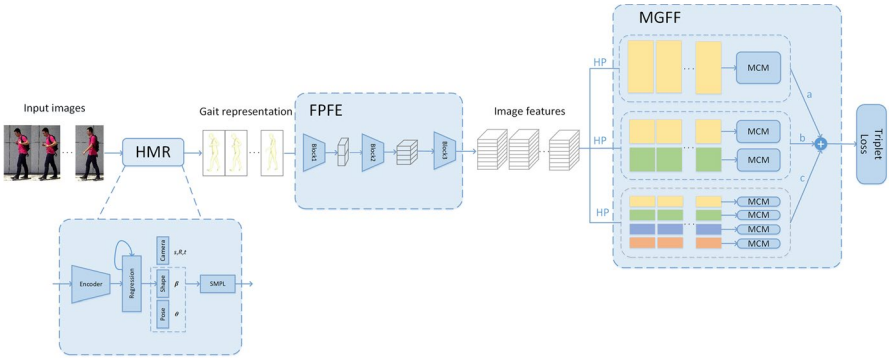

Fig. 2 The framework of our method. s, R, t, β and θ, respectively, represent the camera scaling, the rotation, translation parameters, shape parameters, and attribute parameters. SMPL is a parametric 3D model of human body. Block1, Block2, and Block3 are convolutional blocks of FPFE. HP indicates horizontal pooling. MCM is motion capture module. a, b, and c represent the weight of each granularity

3.2 3D human body reconstruction

Traditional gait recognition methods can be challenged by variations in the input video, such as changes in clothing or carrying objects, which are commonly encountered in real-world scenarios. To tackle this issue, we adopt a 3D human body reconstruction approach to generate a compact and discriminative gait representation, instead of relying on the silhouette feature that has been commonly used but may contain redundant information. 3D human body reconstruction is capable of generating 3D human mesh sequences that incorporate parametric pose and shape features. These sequences are advantageous for gait recognition in cross-state scenes since they do not include redundant information other than the human body, such as clothing and accessories, which is the case with silhouette gait sequences. Compared with simple human pose sequences, 3D human reconstruction produces more refined results that contain both body shape and pose information, resulting in better discrimination for gait recognition. Therefore, 3D human body reconstruction is an effective approach for gait recognition tasks. Our method uses the Human Mesh Recovery module to reconstruct the mesh of the human body from a single RGB image. The HMR is based on the principles of generative adversarial networks and consists of an encoder and a discriminator. The i-th image is fed through the encoder, whose backbone network is a ResNet-50, to extract image features. Then, a parametric regression (iterative 3D regression network) is performed on the features to learn an 85-dimensional vector $\Theta_i = \{s, R, t, \beta, \theta\}$ that includes the camera parameters, such as scaling, rotation, and translation, as well as the shape and attitude parameters of the individual. The shape parameter β describes the height, weight, and body proportions, while the attitude parameter θ describes the joint locations. The learned parameters $\hat{\theta}_i$ and $\hat{\beta}_i$ are then input into the SMPL [46] model, which results in the 3-D joint coordinate of the model. The 3-D joint is then projected onto the image plane using the camera parameters to obtain a predicted 2D image. The SMPL model refers to the Skinned Multi-Person Linear Model, which is a parameterization of the human body.

With the help of the 3D human body reconstruction module (i.e., HMR), we have generated a new 5-dimensional gait representation vector with the size of $N \times S \times C \times H \times W$, where N represents the batch size, S stands for the number of frames, C is the number of channels, and $H \times W$ indicates the resolution of the generated gait feature maps.

3.3 Frame-level part feature extractor

With the aim of enhancing the learning of fine-grained features of frames, we employ the frame-level part feature extractor to extract the local spatial features of each frame. FPFE consists of three blocks, and each block is composed of two focal convolution layers (FConv) that divide the previous feature maps horizontally into n predefined parts, followed by regular convolution operations on each

Table 1 The structure of frame-level part feature extractor. In-C, Out-C, Kernel, Pad, and n are input channels, output channels, the size of kernels, padding, and the number of predefined blocks in FConv, respectively. MaxPool and stride represent the maximum pool operation and the distance of a kernel movement

Frame-level part feature extractor						
Block	Layer	In-C	Out-C	Kernel	Pad	*n*
Block1	FConv1	1	32	5	2	1
	FConv2	32	32	3	1	1
MaxPool, kernel size=2, stride=2						
Block2	FConv3	32	64	3	1	4
	FConv4	64	64	3	1	4
MaxPool, kernel size=2, stride=2						
Block3	FConv5	64	128	3	1	8
	FConv6	128	128	3	1	8

part. After three blocks, the output feature maps are concatenated. The detailed network structure is shown in Table 1.

3.4 Multi-granular feature fusion module

To make the most of the spatial features of pedestrians, we propose the multi-granular feature fusion module to model the multi-granularity features of pedestrians. The MGFF module consists of three branches, each of which ($\mathrm{MGFF}_{(i,\cdot)}$) is responsible for modeling the short-range temporal–spatial representation of a specific granularity using the motion capture module (MCM). The first branch, $\mathrm{MGFF}_{(1,\cdot)}$, extracts global temporal–spatial features, while the second branch, $\mathrm{MGFF}_{(2,\cdot)}$, and third branch, $\mathrm{MGFF}_{(3,\cdot)}$, extract two-part and four-part features, respectively, to focus on finer-grained details. Unlike the most of existing gait recognition methods that only consider either global or local features, our approach models multiple levels of features for improved discriminative performance. Figure 3 shows the specific structure of $\mathrm{MGFF}_{(2,\cdot)}$ as an example.

Let $p_{(i,j)}$ represent the j-th level of the i-th branch in the multi-granular feature fusion module. The part-level generative features are obtained by inputting the vector $p_{(i,j)}$ into $\mathrm{MCM}_{(i,j)}$, as expressed by:

$$v_{(i,j)} = \mathrm{MCM}_{(i,j)}(p_{(i,j)}). \tag{1}$$

In Eq. 1, the motion capture module is designed to learn a more fine-grained gait representation. The MCM is composed of the Micro-motion Template Builder (MTB) module and the temporal pooling (TP) module. The MTB module maps the part-level feature vector $p_{(i,j)}$ to $q_{(i,j)}$, i.e., $q_{(i,j)} = \mathrm{MTB}(p_{(i,j)})$. The TP module then extracts the most discriminative motion feature vector $v_{(i,j)}$, i.e., $v_{(i,j)} = \mathrm{TP}(q_{(i,j)})$.

3.4.1 MTB module

The MTB module includes two similar parts, each with a different convolution kernel size. The first part, ConvNet1d, is a small network composed of two 1-D

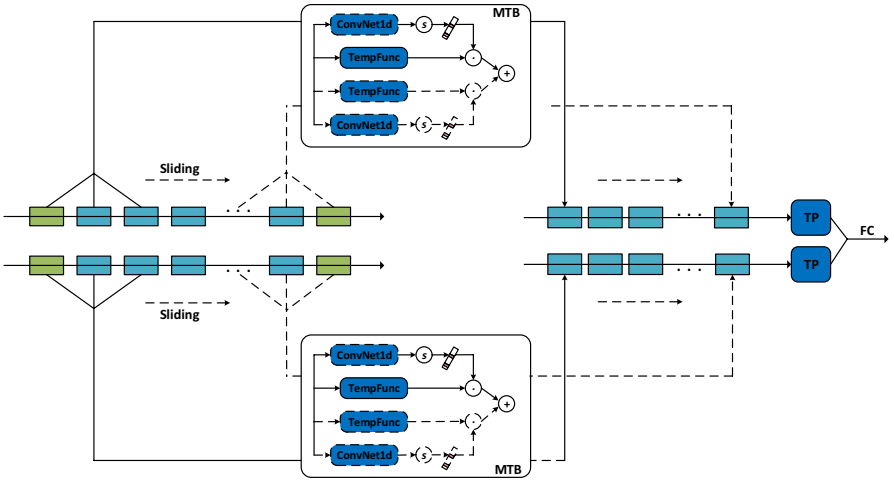

Fig. 3 The structure of $\mathrm{MGFF}_{(2,\cdot)}$. ConvNet1d is a small network composed of two 1-D convolutional layers, Tempfunc is a template function composed of Avgpool1d and Maxpool1d functions, and s is a sigmoid function. TP is a temporary pooling

convolution layers. As shown in Fig. 3, ConvNet1d is utilized to produce a temporary vector $p1_{(i,j)}$, which is depicted as:

$$p1_{(i,j)} = \mathrm{ConvNet1}d(p_{(i,j)}). \tag{2}$$

The second part, Tempfunc, utilizes the concept of a Gait Energy Image to average multiple frames in the sequence. By applying two statistical functions, Tempfunc generates another temporary vector $p2_{(i,j)}$. This can be expressed as:

$$p2_{(i,j)} = \mathrm{Avgpool1}d(p_{(i,j)}) + \mathrm{Maxpool1}d(p_{(i,j)}). \tag{3}$$

Further, to obtain a more discriminative micro-motion representation, the channel attention mechanism is introduced in the MTB module. This mechanism reweights the feature vector at each time, resulting in the final micro-motion representation $q_{(i,j)}$. Mathematically, it can be formulated as:

$$q_{(i,j)} = p2_{(i,j)} \cdot \mathrm{Sigmoid}(p1_{(i,j)}). \tag{4}$$

3.4.2 TP module

After MTB, we get several gait motion representations, from which part-level features can be extracted by TP module. TP module uses $\max(\cdot)$ as the statistical function, i.e.,

$$\mathrm{TP}(q^t_{(i,j)}) = \max(q^1_{(i,j)}, q^2_{(i,j)}, ..., q^t_{(i,j)}), \tag{5}$$

where t is the number of frames.

For obtaining the part-level feature vector $v_{(i)}$, we sum the outputs $v_{(i,\cdot)}$ of each branch $\mathrm{MGFF}_{(i,\cdot)}$ using the following equation:

$$v_{(i)} = \sum_{\cdot=0}^{j} v_{(i,\cdot)}. \tag{6}$$

Finally, by weighting the feature vectors of each branch, we can obtain the final feature vector v:

$$v = av_{(1)} + bv_{(2)} + cv_{(3)}, \tag{7}$$

where a, b, and c are the weights of each branch.

3.5 Loss function

We use the separate Batch All (BA+) triplet loss function to optimize our model, which helps bring samples with the same ID closer in the feature space and separates samples with different IDs further apart. We also utilize the popular triplet loss for video detection tasks. The triplet loss calculates the Euclidean distance between an anchor sample, a positive sample, and a negative sample in the embedding space and aims to make the distance between the anchor and positive samples closer than the distance between the anchor and negative samples. Specifically, given a triplet of image sequences, i.e., anchor sample a, positive sample p, and negative sample n, the triplet loss function can be expressed as:

$$L = [D(f(x_a^i), f(x_p^i)) - D((f(x_a^i), f(x_n^i)) + \beta]_+, \tag{8}$$

where $f(x_a^i)$, $f(x_p^i)$ and $f(x_n^i)$ are the features from anchor samples, positive samples, and negative samples, respectively. $D(,\ )$ denotes the Euclidean distance measure between features, and β is the margin.

4 Experiment

4.1 Datasets and metric

4.1.1 Outdoor-Gait

The Outdoor-Gait [47] dataset is a comprehensive Outdoor-Gait dataset, consisting of 138 individuals and three scenes for each person. Each scene is divided into 3 walking conditions, including 4 normal walking (NM) sequences, 4 walking sequences wearing coat and jacket (CL), and 4 walking sequences with bag (BG). Each walking sequence consists of a single view (90°) of the person walking, and there are $3 * (4 + 4 + 4) = 36$ sequences for each person. During the training process, 69 individuals are used as the training set and the remaining 69 individuals are

used as the test set. The dataset includes both original video frame sequences and gait silhouette sequences.

4.1.2 CASIA-B

The CASIA-B [48] dataset is a large-scale, multi-view gait dataset consisting of 124 individuals. Each individual has three walking conditions including 6 normal walking sequences (NM), 2 walking sequences wearing a coat and jacket (CL), and 2 walking sequences with a bag (BG). Each walking sequence is captured from 11 views ($0°, 18°, 36°, ..., 180°$), spanning from 0° to 180°. In total, each individual has $(6 + 2 + 2) * 11 = 110$ sequences. The first 74 individuals in the database are used for training, and the last 50 individuals are used for testing. The dataset includes both original video frame sequences and gait silhouette sequences.

4.1.3 Rank-1

In our experiments, the effectiveness of the proposed model was evaluated using the Rank-1 recognition accuracy, which measures the ability to correctly identify a sequence in the gallery that has the same ID as the sequence in the Probe. Specifically, the Rank-1 accuracy was calculated by comparing the probe sequence with all sequences in the gallery and determining whether the highest ranked match has the same ID as the probe.

4.2 Implementation details

In this section, we will provide a detailed explanation of the implementation and network structure of our experiments, including the FPFE and MTB modules.

In our experiments, we selected 30 frames for each sequence to be used for training, and the separate Batch All (BA+) triplet loss is used to train the network where the margin β in Eq. 8 was set to 0.2. The batch size for the Outdoor-Gait dataset was set to (4, 8), and the input frame resolution was cropped to 128×88. For the CASIA-B dataset, the batch size was set to (8, 16), and the input frame resolution was cropped to 64×44. We perform 160k iterations for both datasets. In addition, the Adam optimization algorithm was used with a learning rate of 1e-4 and momentum of 0.9. Prior to training, the 3D reconstruction network was pretrained on the MSCOCO-2017 object dataset [49].

The frame-level part feature extractor module is designed to extract meaningful features from gait sequences that represent the unique gait patterns of pedestrians. This module comprises multiple focal convolution network layers and MaxPooling layers, as shown in Table 1. The notations In-C, Out-C, Kernel, and Pad represent the number of input channels, the number of output channels, the size of the kernel, and padding, respectively. The Micro-motion Template Builder module is used to learn the micro-motion representations from the part-level gait features obtained from the FPFE. As seen in Table 2, the MTB module consists of convolution layers and pooling layers. The notations used for the MTB are the same as those used for

Table 2 The structure of Micro-motion Template Builder. *C* and *s* represent the input channel and the squeeze ratio, respectively. '|' is used to divide MTB1 and MTB2

Module	MTB1 MTB2			
Layer	Conv1d-1	Conv1d-2	Avgpool1d	Maxpool1d
In-C	*C*|*C*	*C*/*s*|*C*/*s*	×	×
Out-C	*C*/*s*|*C*/*s*	*C*|*C*	×	×
Kernel	3|3	1|3	3|5	3|5
Pad	1|1	0|1	1|2	1|2

the FPFE. In addition, we use the symbols *C* and *s* to denote the number of channels and the compression ratio between the input and output channels, separated by a '|' symbol.

4.3 Main results

In this experiment, we validated our method on the Outdoor-Gait dataset. It is worth noting that previous gait recognition methods have mostly been based on GEI or silhouette sequence data, as shown in the middle row of Fig. 1. These binary images are generated from the original RGB video frames, meaning that previous works have rarely performed gait recognition directly on the original RGB video frames. Additionally, the reliance on silhouette sequence data as input introduces an extra step of image preprocessing into the gait recognition task, and the recognition accuracy is greatly impacted by the quality of silhouette sequence generation, leading to decreased robustness and increased noise.

The proposed method uses a novel approach based on 3D human body reconstruction and trains on original RGB video data directly. Unlike existing works, our model eliminates the need for GEI or silhouette sequence data, making it more practical and easier to implement in real-world scenarios. The results of cross-condition recognition experiments conducted on the Outdoor-Gait dataset are presented in Table 3. The table compares our method with other gait recognition methods based on GEI or silhouette sequence data and shows that the mean accuracy of our method outperforms these methods in recognizing the same pedestrian under different walking conditions. While our method may show a slight deficiency when the cross-conditions between gallery and probe are the same, i.e., Gallery-NM→Probe-NM, Gallery-BG→Probe-BG, Gallery-CL→Probe-CL, it shows a huge gap over comparative methods in other scenarios. This result is twofold: first, our method avoids the negative effect of redundant information such as coat wearing or bag carrying that seriously impacts the performance of other methods when the conditions between the gallery and probe are different. Second, our method is more feasible for real-world applications as it utilizes original RGB video sequences rather than carefully labeled silhouette data, which may result in a loss of compact and discriminative representation when the conditions between the gallery and probe are the same. Despite this trade-off, our method directly uses original video data and has higher mean accuracy, making it a promising alternative for gait recognition.

Table 3 Experimental results on Outdoor-Gait dataset. NM, CL, and BG are, respectively, normal walking sequences, walking sequences wearing coat and jacket, and walking sequences with bag

Type	Gallery	NM			BG			CL			Mean
	Probe	NM	BG	CL	NM	BG	CL	NM	BG	CL	
Silhouette-based	GEI+PCA [50]	85.0	38.9	29.5	30.7	92.5	20.8	29.2	22.9	86.7	48.5
	GEI-Net [27]	93.2	59.2	55.8	44.2	96.6	27.5	45.9	36.7	93.7	61.4
	GaitNet [47]	96.9	89.1	60.2	92.0	97.1	59.7	58.7	55.4	97.3	78.5
3D Reconstruction-based	Ours	90.0	83.6	80.9	77.5	90.3	72.6	72.2	66.4	88.8	80.2

Table 4 Experimental results on CASIA-B dataset

Type	Gallery	NM		
	Probe	NM	BG	CL
Silhouette-based	CNN-LB [28]	89.9	72.4	54.0
	GaitSet [43]	95.0	87.2	70.4
	GaitPart [44]	96.2	91.5	78.7
RGB-based	PoseGait [15]	60.5	39.6	29.8
	GaitMesh [51]	76.6	42.0	32.8
	GaitGraph [16]	87.7	74.8	66.3
	Ours	88.2	77.8	70.5

Our model was further validated through experiments on the large CASIA-B gait dataset, as seen in Table 4. The results demonstrate that our model produces relatively satisfied results even when raw video frames are used as input. In the GBG-based models, the accuracy of our model is better than others. However, it is important to note that the CASIA-B dataset, being published earlier, contains many poor quality data in its 3D human body reconstruction which negatively impacts recognition accuracy. In contrast, the Outdoor-Gait dataset features higher pixel quality and provides more effective 3D human body reconstruction for gait representation, leading to better results. It should be mentioned that the CASIA-B dataset has 11 different viewing angles for each walking condition, and the recognition accuracy is calculated as the average across these 11 angles.

Our experiments suggest that our method is effective and competitive in gait recognition tasks. As demonstrated in Tables 3 and 4, our proposed approach achieves the highest mean accuracy compared with other silhouette sequence-based models on the high-resolution dataset. Furthermore, when compared with the RGB-based models, our method exhibits higher recognition accuracy. During the process of 3D human body reconstruction, there may be some failed cases. In our experiments, the main reasons for failures are the presence of deviations in body tilt angles and slender limbs, resulting in inaccurate reconstruction of the true human contour, as shown in Fig. 4. There may be two reasons for these situations: firstly, some original RGB images have poor image quality, which affects the precision of human body reconstruction; secondly, the 3D human body reconstruction method used in our experiments has limitations in reconstructing fine-grained details of the human body.

4.4 Ablation study

To demonstrate the effectiveness of our proposed method, we conducted ablation experiments on both the Outdoor-Gait and CASIA-B datasets. Our method was compared against several state-of-the-art methods that rely on GEI and silhouette sequences as inputs, without the use of 3D human body reconstruction. As seen from Tables 5 and 6, the integration of the multi-granular feature fusion module has led to a improvement in the recognition accuracy on both datasets compared to current approaches.

Table 5 Results on Outdoor-Gait dataset

Gallery	NM		
Probe	NM	BG	CL
GEI+PCA [50]	85.0	38.9	29.5
GEI-Net [27]	93.2	59.2	55.8
GaitNet [47]	96.9	89.1	60.2
Ours	97.6	91.0	84.0

Fig. 4 The failure cases of 3D human body reconstruction

Table 6 Experimental results of MGFF on CASIA-B dataset

Gallery NM#1-4		0°−180°											Mean
Probe		0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	
NM#5-6	CNN-LB [28]	82.6	90.3	96.1	94.3	90.1	87.4	89.9	94.0	94.7	91.3	78.5	89.9
	GaitSet [43]	90.8	97.9	99.4	96.9	93.6	91.7	95.0	97.8	98.9	96.8	85.8	95.0
	GaitNet [47]	91.2	92.0	90.5	95.6	86.9	92.6	93.5	96.0	90.9	88.8	89.0	91.6
	GaitPart [44]	94.1	98.6	99.3	98.5	94.0	92.3	95.9	98.4	99.2	97.8	90.4	96.2
	Ours	94.0	99.0	99.6	98.5	95.0	92.5	96.0	98.5	99.6	98.4	91.0	96.6
BG#1-2	CNN-LB [28]	64.2	80.6	82.7	76.9	64.8	63.1	68.0	76.9	82.2	75.4	61.3	72.4
	GaitSet [43]	83.8	91.2	91.8	88.8	83.3	81.0	84.1	90.0	92.2	94.4	79.0	87.2
	GaitNet [47]	83.0	87.8	88.3	93.3	82.6	74.8	89.5	91.0	86.1	81.2	85.6	85.7
	GaitPart [44]	89.1	94.8	96.7	95.1	88.3	84.9	89.0	93.5	96.1	93.8	85.8	91.5
	Ours	88.7	94.4	96.0	94.8	89.0	84.6	89.9	94.9	95.9	94.0	86.1	91.7
CL#1-2	CNN-LB [28]	37.7	57.2	66.6	61.1	55.2	54.6	55.2	59.1	58.9	48.8	39.4	54.0
	GaitSet [43]	61.4	75.4	80.7	77.3	72.1	70.1	71.5	73.5	73.5	68.4	50.0	70.4
	GaitNet [47]	42.1	58.2	65.1	70.7	68.0	70.6	65.3	69.4	51.5	50.1	36.6	58.9
	GaitPart [44]	70.7	85.5	86.9	83.3	77.1	72.5	76.9	82.2	83.8	80.2	66.5	78.7
	Ours	71.9	84.5	89.2	85.1	80.0	74.9	79.2	83.7	84.2	82.6	68.2	80.3

To further demonstrate the efficacy of the multi-granular feature fusion module, we conducted additional experiments on the CASIA-B dataset. We performed ablation tests to evaluate each component of the multi-granular feature fusion module

Table 7 Ablation study on the CASIA-B dataset. In the first column, a, b, and c indicate the weight of different granularity

a	b	c	Gallery	NM		
			Probe	NM	BG	CL
1			Accuracy	95.9	90.6	78.4
	1			96.0	90.0	78.8
		1		96.2	91.5	78.7
1	1	1		96.6	91.7	80.3

separately and compared them to our complete model. As seen in Table 7, the results reveal that the multi-granular feature fusion module consistently delivers improved performance across different walking conditions. To account for the 11 viewing angles in the CASIA-B dataset, the final results were obtained by taking the average recognition accuracy across all 11 angles.

4.5 Visual analysis

In order to show a more intuitive performance of our model effectively, we present a visualization of our results in Fig. 5. The figure is comprised four parts, each representing the original video frames (a), silhouette image sequences (b), pose sequences (c), and image sequences after our 3D human body reconstruction (d). In each row, the first row represents the BG condition, the second row represents the CL condition, and the third row represents the NM condition.

As seen in Fig. 5, the 3D human body reconstruction effectively eliminates the negative impact of extraneous information, such as coats and backpacks, on the performance of gait recognition. Furthermore, it effectively compensates for the lack of local information in the original video frames. Consequently, our proposed method with 3D human body reconstruction performs much better than existing methods that only rely on silhouette image sequences. This leads to more robust and superior results, due to the compact and discriminative gait representation provided by our model.

5 Conclusion

We have designed a novel end-to-end gait recognition method that leverages 3D human body reconstruction to improve recognition performance. By using a HMR module to generate a compact and discriminative gait representation that eliminates the negative effects of redundant information, our method avoids the issues that plague existing methods when dealing with huge changes in video. To further enhance the recognition ability, we introduced a multi-granular feature fusion module that effectively leverages global and local features of pedestrians at multiple granularities. Our method was conducted on two popular gait recognition datasets, the Outdoor-Gait and CASIA-B, and it was shown to outperform similar

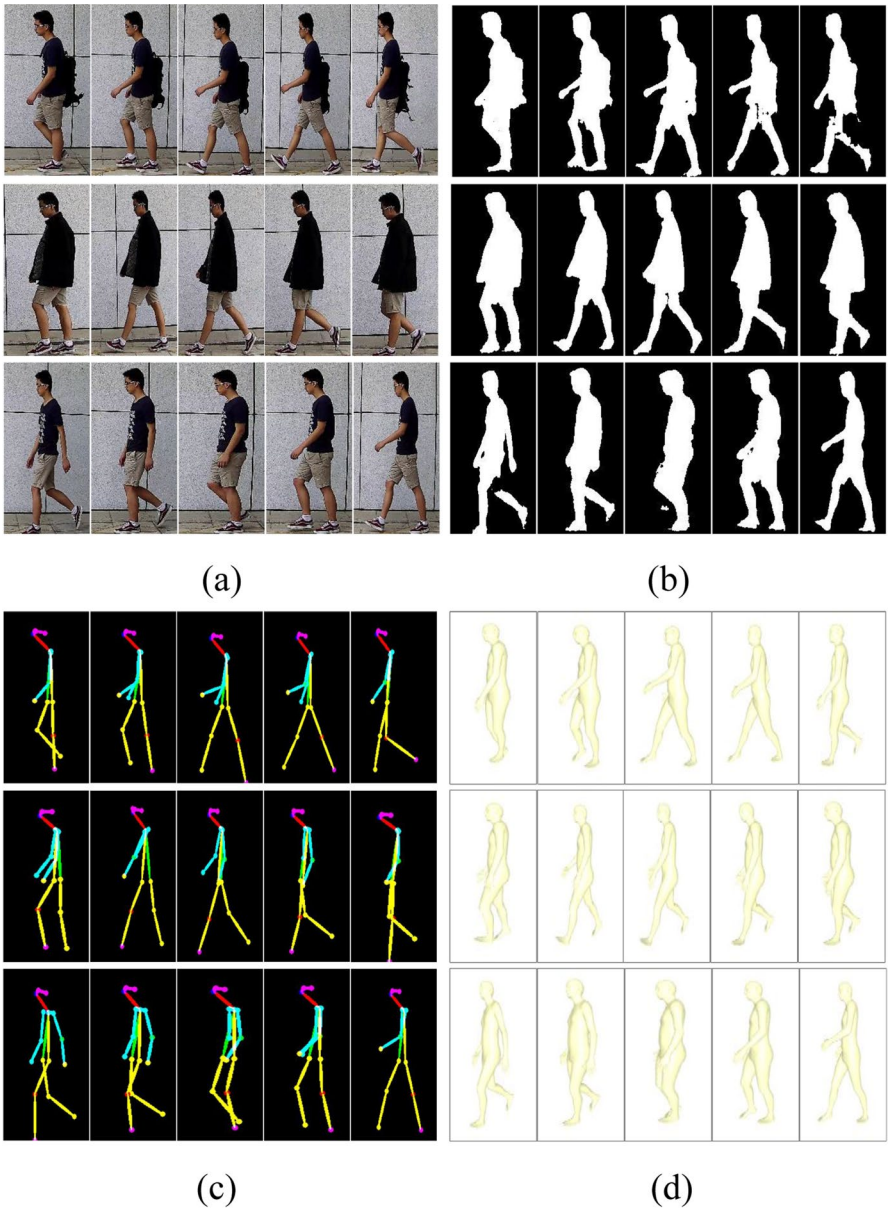

Fig. 5 Visualization results. **a** Original video frames, **b** silhouette image sequences, **c** pose sequences, and **d** image equences after 3D human body reconstruction. For each row, the first row is the BG condition, the second row is the CL condition, and the third row is the NM condition

state-of-the-art methods. Visualization results illustrate that our 3D reconstruction-based model can learn a more discriminative and nonredundant gait representation, greatly contributing to improved gait recognition performance.

Author contributions CM and XH helped in conceptualization, methodology, software, and formal analysis; CM contributed to validation, writing—original draft preparation investigation, and visualization; LL was involved in investigation and writing—review and editing; XH curated the data; ZT helped in supervision, project administration, and funding acquisition. All authors have read and agreed to the published version of the manuscript.

Funding This work was supported by National Social Science Fund of China, under Grant 16AJL008.

Data availability All datasets used are public.

Declarations

Conflict of interest The authors declare no conflict of interest.

Ethical approval Not applicable.

References

1. Lin C, Kumar A (2018) Contactless and partial 3d fingerprint recognition using multi-view deep representation. Pattern Recognit 83:314–327
2. Yang W, Wang S, Hu J, Zheng G, Yang J, Valli C (2019) Securing deep learning based edge finger vein biometrics with binary decision diagram. IEEE Trans Ind Inf 15(7):4244–4253
3. Liu Y, Wei F, Shao J, Sheng L, Yan J, Wang X (2018) Exploring disentangled feature representation beyond face identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2080–2089
4. Ahmed T, Sarma M (2019) Hash-based space partitioning approach to iris biometric data indexing. Expert Syst Appl 134:1–13
5. Singh JP, Jain S, Arora S, Singh UP (2018) Vision-based gait recognition: a survey. IEEE Access 6:70497–70527
6. Sethi D, Bharti S, Prakash C (2022) A comprehensive survey on gait analysis: history, parameters, approaches, pose estimation, and future work. Artif Intell Med 129:102314
7. Wu H, Tian J, Fu Y, Li B, Li X (2020) Condition-aware comparison scheme for gait recognition. IEEE Trans Image Process 30:2734–2744
8. Khan MH, Farid MS, Grzegorzek M (2021) Vision-based approaches towards person identification using gait. Comput Sci Rev 42:100432
9. Lee L, Grimson WEL (2002) Gait analysis for recognition and classification. In: Proceedings of Fifth IEEE International Conference on Automatic Face Gesture Recognition, pp. 155–162. IEEE
10. Cunado D, Nixon MS, Carter JN (1997) Using gait as a biometric, via phase-weighted magnitude spectra. In: International Conference on Audio-and Video-based Biometric Person Authentication, pp. 93–102. Springer
11. Yoo J-H, Nixon MS (2003) Markerless human gait analysis via image sequences. In Proceedings of International Society of Biomechanics 19th Congress, pp. 1–5
12. Yam C, Nixon MS, Carter JN (2004) Automated person recognition by walking and running via model-based approaches. Pattern Recognit 37(5):1057–1072
13. Urtasun R, Fua P (2004) 3d tracking for gait characterization and recognition. In: Sixth IEEE International Conference on Automatic Face and Gesture Recognition, 2004. Proceedings., pp. 17–22. IEEE
14. Dockstader SL, Berg MJ, Tekalp AM (2003) Stochastic kinematic modeling and feature extraction for gait analysis. IEEE Trans Image Process 12(8):962–976
15. Liao R, Cao C, Garcia EB, Yu S, Huang Y (2017) Pose-based temporal-spatial network (ptsn) for gait recognition with carrying and clothing variations. In: Chinese Conference on Biometric Recognition, pp. 474–483. Springer

16. Teepe T, Khan A, Gilg J, Herzog F, Hörmann S, Rigoll G (2021) Gaitgraph: graph convolutional network for skeleton-based gait recognition. In: 2021 IEEE International Conference on Image Processing (ICIP), pp. 2314–2318. IEEE
17. Kipf TN, Welling M (2016) Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907
18. Li X, Makihara Y, Xu C, Yagi Y, Yu S, Ren M (2020) End-to-end model-based gait recognition. In: Proceedings of the Asian Conference on Computer Vision
19. Khan MH, Farid MS, Grzegorzek M (2020) A non-linear view transformations model for cross-view gait recognition. Neurocomputing 402:100–111
20. Han J, Bhanu B (2005) Individual recognition using gait energy image. IEEE Trans Pattern Anal Mach Intell 28(2):316–322
21. Bashir K, Xiang T, Gong S (2009) Gait recognition using gait entropy image. In 3rd International Conference on Imaging for Crime Detection and Prevention, pp. 1–6. IET
22. Lam TH, Cheung KH, Liu JN (2011) Gait flow image: a silhouette-based gait representation for human identification. Pattern Recognit 44(4):973–987
23. Wang C, Zhang J, Wang L, Pu J, Yuan X (2011) Human identification using temporal information preserving gait template. IEEE Trans Pattern Anal Mach Intell 34(11):2164–2176
24. Xu D, Yan S, Tao D, Zhang L, Li X, Zhang H-J (2006) Human gait recognition with matrix representation. IEEE Trans Circuits Syst Video Technol 16(7):896–903
25. Guan Y, Li C-T, Roli F (2014) On reducing the effect of covariate factors in gait recognition: a classifier ensemble method. IEEE Trans Pattern Anal Mach Intell 37(7):1521–1528
26. Makihara Y, Suzuki A, Muramatsu D, Li X, Yagi Y (2017) Joint intensity and spatial metric learning for robust gait recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5705–5715
27. Shiraga K, Makihara Y, Muramatsu D, Echigo T, Yagi Y (2016) Geinet: View-invariant gait recognition using a convolutional neural network. In: 2016 International Conference on Biometrics (ICB), pp. 1–8. IEEE
28. Wu Z, Huang Y, Wang L, Wang X, Tan T (2016) A comprehensive study on cross-view gait based human identification with deep cnns. IEEE Trans Pattern Anal Mach Intell 39(2):209–226
29. Takemura N, Makihara Y, Muramatsu D, Echigo T, Yagi Y (2017) On input/output architectures for convolutional neural network-based cross-view gait recognition. IEEE Trans Circuits Syst Video Technol 29(9):2708–2719
30. Li X, Makihara Y, Xu C, Yagi Y, Ren M (2019) Joint intensity transformer network for gait recognition robust against clothing and carrying status. IEEE Trans Inf Forensics Secur 14(12):3102–3115
31. Kusakunniran W, Wu Q, Zhang J, Li H (2010) Support vector regression for multi-view gait recognition based on local motion feature selection. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 974–981. IEEE
32. Makihara Y, Sagawa R, Mukaigawa Y, Echigo T, Yagi Y (2006) Gait recognition using a view transformation model in the frequency domain. In: European Conference on Computer Vision, pp. 151–163. Springer
33. Tsuji A, Makihara Y, Yagi Y (2010) Silhouette transformation based on walking speed for gait identification. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 717–722. IEEE
34. Muramatsu D, Shiraishi A, Makihara Y, Uddin MZ, Yagi Y (2014) Gait-based person recognition using arbitrary view transformation model. IEEE Trans Image Process 24(1):140–154
35. Mansur A, Makihara Y, Aqmar R, Yagi Y (2014) Gait recognition under speed transition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2521–2528 (2014)
36. Akae N, Mansur A, Makihara Y, Yagi Y (2012) Video from nearly still: an application to low frame-rate gait recognition. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1537–1543. IEEE
37. Yu S, Chen H, Garcia Reyes EB, Poh N (2017) Gaitgan: Invariant gait feature extraction using generative adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 30–37
38. Yu S, Liao R, An W, Chen H, García EB, Huang Y, Poh N (2019) Gaitganv 2: invariant gait feature extraction using generative adversarial networks. Pattern Recognit 87:179–189
39. He Y, Zhang J, Shan H, Wang L (2018) Multi-task gans for view-specific feature learning in gait recognition. IEEE Trans Inf Forensics Secur 14(1):102–113

40. Tran D, Bourdev L, Fergus R, Torresani L, Paluri M (2015) Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4489–4497
41. Wolf T, Babaee M, Rigoll G (2016) Multi-view gait recognition using 3d convolutional neural networks. In: 2016 IEEE International Conference on Image Processing (ICIP), pp. 4165–4169. IEEE
42. Zhang Z, Tran L, Yin X, Atoum Y, Liu X, Wan J, Wang N (2019) Gait recognition via disentangled representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4710–4719
43. Chao H, He Y, Zhang J, Feng J (2019) Gaitset: Regarding gait as a set for cross-view gait recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 8126–8133
44. Fan C, Peng Y, Cao C, Liu X, Hou S, Chi J, Huang Y, Li Q, He Z (2020) Gaitpart: Temporal part-based model for gait recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14225–14233
45. Kanazawa A, Black MJ, Jacobs DW, Malik J (2018) End-to-end recovery of human shape and pose. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7122–7131
46. Loper M, Mahmood N, Romero J, Pons-Moll G, Black MJ (2015) Smpl: a skinned multi-person linear model. ACM Trans Gr (TOG) 34(6):1–16
47. Song C, Huang Y, Huang Y, Jia N, Wang L (2019) Gaitnet: an end-to-end network for gait based human identification. Pattern Recognit 96:106988
48. Yu S, Tan D, Tan T (2006) A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In: 18th International Conference on Pattern Recognition (ICPR'06), vol. 4, pp. 441–444. IEEE
49. Lin T-Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL (2014) Microsoft coco: common objects in context. In: European Conference on Computer Vision, pp. 740–755. Springer
50. Wang L, Tan T, Ning H, Hu W (2003) Silhouette analysis-based gait recognition for human identification. IEEE Trans Pattern Anal Mach Intell 25(12):1505–1518
51. Li X, Makihara Y, Xu C, Yagi Y (2022) Multi-view large population gait database with human meshes and its performance evaluation. IEEE Trans Biom Behav Identity Sci 4(2):234–248