Multi-View Large Population Gait Database With Human Meshes and Its Performance Evaluation

Xiang Li, Yasushi Makihara, Chi Xu, and Yasushi Yagi, *Senior Member, IEEE*

***Abstract*—Existing model-based gait databases provide the 2D poses (i.e., joint locations) extracted by general pose estimators as the human model. However, these 2D poses suffer from information loss and are of relatively low quality. In this paper, we consider a more informative 3D human mesh model with parametric pose and shape features, and propose a multi-view training framework for accurate mesh estimation. Unlike existing methods, which estimate a mesh from a single view and suffer from the ill-posed estimation problem in 3D space, the proposed framework takes asynchronous multi-view gait sequences as input and uses both multi-view and single-view streams to learn consistent and accurate mesh models for both multi-view and single-view sequences. After applying the proposed framework to the existing OU-MVLP database, we establish a large-scale gait database with human meshes (i.e., OUMVLP-Mesh), containing over 10,000 subjects and up to 14 view angles. Experimental results show that the proposed framework estimates human mesh models more accurately than similar methods, providing models of sufficient quality to improve the recognition performance of a baseline model-based gait recognition approach.**

***Index Terms*—Asynchronous multi-view sequences, gait database, gait recognition, three-dimensional human pose/shape estimation.**

I. Introduction

THE WAY in which humans walk contains numerous cues (e.g., static shape and dynamic pose movement) that indicate their unique identities. Compared with other traditional biometrics (e.g., face, iris and fingerprint), this gait information has many advantages, such as availability at a distance, identification without cooperation, and difficulty of deception. Therefore, gait has become an ideal biometric feature for human identity recognition at a distance, with widespread applications in surveillance and forensics [1], [2], [3].

Gait recognition approaches are generally divided into those based on appearance [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20] and those based on a model [21], [22], [23], [24], [25], [26]. The appearance-based approaches have been dominant in recent years because of simple yet effective silhouette-based representations (e.g., gait energy images [4], frequency-domain features [14], and cropped silhouettes). However, these representations are relatively sensitive to many common covariates, such as viewing angles, clothing, and objects that are being carried. To solve this, various metric learning-based methods and deep learning-based networks have been explored. The model-based approaches are mainly based on representations of human pose structure and movement (e.g., skeletons). They are therefore less sensitive to those covariates, but such representations are difficult to extract accurately.

This trend is the same for the existing multi-view gait databases in Table I. Most of them provide silhouettes, which are easy to extract through background subtraction or other segmentation methods. This is because the video sensors employed (e.g., color camera) in generating these databases cannot provide direct model-based representations, and the extra estimation is difficult and expensive. In recent years, there have been significant improvements in human pose estimation from videos or images owing to the success of deep learning, which makes it easy to extract poses from existing gait databases. For example, OUMVLP-Pose uses state-of-the-art pose estimators (i.e., OpenPose [27] and AlphaPose [28]) to estimate the skeletons of the existing large-scale multi-view gait database OU-MVLP [29]. Additionally, some model-based gait recognition approaches [30], [31] have used OpenPose to extract the poses of the well-known CASIA-B [32] and a new CASIA-E database [31], [33]. The CASIA-E database contains 1014 subjects, which is much larger than CASIA-B database. However, it has not been released. The available CASIA-B database only contains 124 subjects, making it insufficient to fully demonstrate the effectiveness of model-based gait recognition approaches that use deep neural networks.

Although OpenPose and AlphaPose provide convenient solutions for pose estimation from existing gait databases, we argue that there are still some limitations. First, these are general pose estimators trained on single-view images that are not specifically designed for pose estimation from multi-view gait video sequences, and thus fail to provide accurate poses for all frames. Second, they mainly use a 2D skeleton as the extracted pose feature, but the 2D skeleton will change as the viewing angle varies. In addition, it loses 1D information. For example, the stride length is generally missing from front-view cases,

Manuscript received November 24, 2021; revised February 24, 2022 and April 29, 2022; accepted May 4, 2022. Date of publication May 12, 2022; date of current version June 21, 2022. This work was supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI under Grant JP19H05692 and Grant JP20H00607. This article was recommended for publication by Associate Editor M. Nixon upon evaluation of the reviewers' comments. *(Corresponding author: Xiang Li.)*

The authors are with the Department of Intelligent Media, SANKEN, Osaka University, Osaka 565-0871, Japan (e-mail: li@am.sanken.osaka-u.ac.jp; makihara@am.sanken.osaka-u.ac.jp; xu@am.sanken.osaka-u.ac.jp; yagi@am.sanken.osaka-u.ac.jp).

Digital Object Identifier 10.1109/TBIOM.2022.3174559

TABLE I
EXISTING MAJOR PUBLICLY AVAILABLE MULTI-VIEW GAIT DATABASES. THE FINAL ROW DESCRIBES OUR DATABASE

Database	Year	Data type	# Subjects	# Views	View range
CMU MoBo [34]	2001	RGB, Silhouettes	25	6	0°-360°
SOTON small [35]	2001	RGB, Silhouettes	12	4	-
USF [36]	2005	RGB, Silhouettes	122	2	-
CASIA-A [37]	2001	RGB, Silhouettes	20	3	0°, 45°, 90°
CASIA-B [32]	2005	RGB, Silhouettes	124	11	0°-180°
AVA [38]	2013	RGB, Silhouettes	20	6	-
WOSG [39]	2013	Short-wave infrared	155	8	-
KY4D [40]	2010	3D volumetric model	42	16	0°-360°
FVG [41]	2019	RGB, Silhouettes	226	3	-45°, 0°, 45°
OU-ISIR LP [42]	2012	Silhouettes	4,016	4	55°-85°
OU-MVLP [29]	2018	Silhouettes	10,307	14	0°-90°, 180°-270°
OUMVLP-Pose [43]	2020	2D skeleton	10,307	14	0°-90°, 180°-270°
OUMVLP-Mesh (Ours)	**2021**	**3D Human Mesh**	**10,307**	**14**	**0°-90°, 180°-270°**

and the body width is usually hidden in side-view cases. Third, they ignore body shape information, which actually contributes to the high recognition accuracy of existing gait recognition works.

Therefore, we consider a more powerful 3D human model than the 2D skeleton. According to our previous studies [44], [45], a 3D human mesh model with parametric pose and shape features (i.e., skinned multi-person linear (SMPL) model [46]) is an effective choice for model-based gait recognition. Compared with the simple sparse 2D skeleton, the SMPL model is more complex and informative, and can encode a full 3D mesh of a human body. Reference [44] describes how the SMPL model is independently estimated from a single-view gait sequence, but the possible multi-view gait sequences are not fully utilized in the training phase, which causes an ill-posed estimation problem in 3D space. To make use of multi-view gait sequences, reference [45] introduced a synchronized multi-view pose constraint to force the poses of multi-view sequences to be similar to each other. However, this constraint alone cannot completely solve the problem. Moreover, both of these methods only infer the SMPL model for single-view gait sequences, and are not applicable for inferring a unified model from multi-view gait sequences.

To obtain a more accurate SMPL model for both multi-view and single-view gait sequences, we propose a multi-view training framework containing multi-view and single-view streams for asynchronous multi-view gait sequences. We further apply the proposed framework to the existing OU-MVLP database, and build the first large-scale gait database with human meshes, named OUMVLP-Mesh.

The contributions of this study can be summarized as follows.

(1) First large-scale multi-view gait database with 3D human meshes[1]*:*

Beyond skeletons, we construct the first multi-view gait database with 3D human meshes (i.e., SMPL model). The SMPL model is complex and informative, making it conducive to promoting the development of model-based gait recognition and boosting the recognition performance. Besides gait recognition, the proposed database can also be used for other gait analyses (e.g., aging progression/regression, training data as genuine gait models for adversarial learning). Built upon the existing OU-MVLP database, OUMVLP-Mesh contains 10,307 subjects with up to 14 viewing angles (from 0° to 90°, 180° to 270° at 15° intervals).

(2) Framework that handles asynchronous multi-view input gait sequences:

To avoid the strict requirement of synchronized multi-view input gait sequences, we introduce a phase sequence estimator for the phase information (gait stance) of input sequences, and synchronize the estimated SMPL models of those sequences for further fusion. As such, our method can handle asynchronous multi-view gait sequences, enabling a wide range of applications.

(3) Multi-view training framework that contains multi-view and single-view streams:

We propose a multi-view training framework containing two streams with shared weights. One is the multi-view stream, in which the estimated view-specific SMPL models from multi-view inputs are fused to form a unified model. The second is the single-view stream, in which the SMPL models for each single-view input are independently estimated. The two streams are constrained to produce similar estimations. As such, the proposed method not only estimates more accurate models through the multi-view stream in the case of multi-view inputs, but also recovers more accurate 3D information through the single-view stream when only single-view inputs are available.

(4) More accurate human models:

Compared with OUMVLP-Pose [43], the proposed OUMVLP-Mesh database has better-quality human models and significantly improves the recognition performance using the same benchmark.

The remainder of this paper is organized as follows. Section II introduces some existing gait databases and 3D human pose and shape estimation approaches. Section III presents the proposed multi-view training framework. Section IV describes the constructed OUMVLP-Mesh

[1] OUMVLP-Mesh is available at http://www.am.sanken.osaka-u.ac.jp/BiometricDB/GaitLPMesh.html.

database. Section V describes the evaluation of the proposed framework and analysis of the OUMVLP-Mesh database. Section VI discusses the performance of the proposed method, and Section VII concludes this paper and discusses ideas for future work.

II. Related Work

A. Gait Databases

Existing major publicly available multi-view gait databases are summarized in Table I. Most of them (i.e., CMU Mobo [34], SOTON small [35], USF [36], CASIA-A [37], CASIA-B [32], AVA [38], WOSG [39], FVG [41], OU-ISIR LP [42], OU-MVLP [29]) contain the silhouettes that are widely used for appearance-based gait recognition approaches. For example, the CMU Mobo database collects 25 subjects walking on a treadmill with six surrounding cameras ranging from 0° to 360°. The SOTON small database contains 12 subjects walking indoors captured under four cameras in the normal side, normal elevated, oblique, and frontal views. The USF database contains 122 subjects walking outdoors in an elliptical path captured under two cameras in the left and right side. The CASIA-A database contains 20 subjects with three views (0°, 45°, 90°). The CASIA-B database contains 124 subjects with large view variations from 0° to 180° with 18° intervals, which is the most widely used database for cross-view gait recognition approaches. The AVA database contains 20 subjects under six view angles. The WOSG database contains 155 subjects walking in an active, outdoor scene captured under eight view variations. Recently, a newly released FVG database focuses more on frontal-view gait sequences with minimal gait information, including 226 subjects walking outdoors near the frontal middle, left, and right view angles. However, these databases cover relatively few subjects, and may therefore not be suitable for evaluating recent deep learning-based approaches in a statistically reliable way. The OU-ISIR LP database contains a relatively larger number (4016) of subjects, but its view variation is limited to only four angles (55°, 65°, 75°, and 85°). The OU-MVLP database is currently the largest gait database, containing 10,307 subjects and 14 view angles ranging from 0° to 90° and 180° to 270° at 15° intervals.

Only two databases (i.e., KY4D [40] and OUMVLP-Pose [43]) provide model-based representations. The KY4D database contains the 3D volumetric models of 42 walking subjects, as reconstructed by the volumetric intersection technique from the gait sequences captured by 16 cameras. Built on the OU-MVLP database, the OUMVLP-Pose database is the first gait database with pose sequences extracted by deep learning-based pose estimators (i.e., OpenPose and AlphaPose). The extracted pose representation is the body skeleton, containing 18 joints in 2D image-based coordinates, namely Nose, Neck, RShoulder, RElbow, RWrist, LShoulder, LElbow, LWrist, RHip, RKnee, RAnkle, LHip, LKnee, LAnkle, REye, LEye, REar, and LEar. However, the KY4D database only covers a small number of subjects, and the 2D skeletons in the OUMVLP-Pose database suffer from information loss and relatively inaccurate estimation by the general pose estimators. This paper describes the use of a more complex and informative 3D mesh model, and proposes a multi-view training framework for more accurate estimation.

B. 3D Human Pose and Shape Estimation

Research on 3D human pose and shape estimation mainly focuses on inferring a parametric model of the human body (e.g., SMPL) from images or videos. Early studies such as SMPLify [47] use optimization-based methods that fit the parameters of SMPL to 2D keypoint detections. More recent works [48], [49], [50] tend to use regression-based methods that regress the model parameters through deep networks. For example, Pavlakos *et al.* [48] used ConvNet to predict 2D pose heat maps and silhouettes from an input RGB image, then designed two individual regression networks to regress the pose and shape parameters of the SMPL respectively. Kanazawa *et al.* [49] proposed an end-to-end human mesh recovery (HMR) network to directly regress the pose and shape parameters of the SMPL from a single RGB image. The two methods are designed for images and fail to produce stable results using videos. Kocabas *et al.* [50] proposed the "Video Inference for Body Pose and Shape Estimation" (VIBE) method to produce accurate and natural motion sequences of SMPL parameters for videos. VIBE uses a temporal encoder and regressor with gated recurrent units (GRUs) [51] to capture sequential human motion, then employs a motion discriminator for adversarial learning to obtain realistic human motion using a large-scale 3D motion-capture dataset named AMASS [52] as the ground-truth motion. All of the aforementioned methods focus on single-view images or videos.

However, single-view images or videos suffer from information loss when estimating a 3D model. For example, if a person walks towards a camera (i.e., captured in the frontal view), the forward–backward motion (e.g., stride length) cannot be observed as clearly as when captured from the side-view. Thus, multi-view images or videos are essential for more accurate 3D human body model estimation. The word "multi-view" refers to the person being captured by synchronized cameras from different view angles; thus, the person has almost the same pose in the multi-view images or videos. There has been relatively little research in this area [53], [54]. Liang and Lin [53] proposed a multi-view multi-stage regression method that takes synchronized multi-view images as input. The estimated pose and shape parameters of the SMPL are iteratively transferred across multiple views, while the estimated camera calibration parameters are transferred across iteration stages. Shin and Halilaj [54] developed a learnable volumetric aggregation approach that fuses information from multi-view images, enabling accurate SMPL models to be reconstructed. Note that these two methods are image-based and require synchronized multi-view images. In contrast, our proposed method is video-based and can deal with asynchronous gait sequences.

C. SMPL Databases

As a strong form of supervision, databases with 3D annotations are essential for 3D human pose and shape estimation.

However, such databases are difficult for humans to annotate because the task is ambiguous. Existing 3D databases (e.g., Human 3.6M [55], MPI-INF-3DHP [56]) use marker-based or marker-less motion capture (MoCap) systems to infer the 3D pose (joint locations) based on an approximate skeletal body structure, commonly treated as the ground-truth 3D pose. Reference [49] describes how the "Motion and Shape Capture" (MoSh) method [57] can be used to fit the ground-truth SMPL parameters from the raw 3D MoCap markers in Human 3.6M. Reference [52] extends MoSh to MoSh++ for more accurate model fitting and establishes a collective database with the SMPL parameters from 15 existing databases. These databases with SMPL are mainly inferred from the MoCap markers using MoSh or MoSh++, a technique that is not applicable for existing multi-view gait databases that have no markers. In contrast, our method has no such limitation and can be applied to gait databases with only multi-view RGB sequences.

III. Multi-View Training Framework

A. Overview

We aim to generate accurate human models (i.e., SMPL model [46]) for existing multi-view gait databases. In our problem setting, we assume that the training set contains asynchronous multi-view RGB gait sequences. Therefore, we make the best possible use of the asynchronous multi-view sequences for accurate model fitting. An overview of the proposed multi-view training framework is shown in Fig. 1. This is a generative adversarial network-based framework that contains a two-stream encoder (i.e., multi-view and single-view stream encoder) for the model fitting and a discriminator for the adversarial learning. In the test case, based on the given sequences, either the multi-view stream encoder or the single-view stream encoder is used for the inference.

The SMPL model factorizes the human body into shape $\boldsymbol{\beta}$ and pose $\boldsymbol{\theta}$ parameters. The shape $\boldsymbol{\beta} \in \mathbb{R}^{10}$ describes the height, weight, and body proportions of individuals, forming the first 10 coefficients in a principal component analysis shape space. The pose $\boldsymbol{\theta} \in \mathbb{R}^{69}$ describes the joint locations, which is the relative 3D rotations of 23 joints in an axis–angle representation. From these parameters, the SMPL model outputs a triangulated mesh $\mathcal{M}(\boldsymbol{\theta}, \boldsymbol{\beta}) \in \mathbb{R}^{6890\times 3}$, which is differentiable with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$. The 3D joint locations X_{3D} can be obtained by linear regression from the mesh vertices. In addition, the global root rotation parameter $\boldsymbol{r} \in \mathbb{R}^3$ in axis–angle representation and the camera parameter $\boldsymbol{k} = [s, \boldsymbol{t}] \in \mathbb{R}^3$ are estimated, where s is the scale and $\boldsymbol{t} = [t_x, t_y]$ is the translation. As such, the 2D projection of the 3D joints is computed as $X_{2D} = s\Pi(RX_{3D}) + \boldsymbol{t}$, where R is the global rotation matrix computed from $\boldsymbol{r}$ and Π is an orthographic projection. All of these parameters form a full SMPL parameter $\Theta = [\boldsymbol{k}, \boldsymbol{r}, \boldsymbol{\theta}, \boldsymbol{\beta}] \in \mathbb{R}^{85}$.

B. Two-Stream Encoder

We design a two-stream encoder to deal with arbitrary input sequences in the test phase. As shown in Fig. 1, one stream is the multi-view stream that fuses the estimated view-specific SMPL parameters from multi-view asynchronous RGB sequences of the same subject to a unified SMPL parameter. The other is the single-view stream, which independently estimates the SMPL parameters for each single-view RGB sequence. The two streams share weights with each other.

1) Multi-View Stream: The multi-view stream contains two modules: a phase sequence estimator and a multi-view union regressor.

Phase sequence estimator: Without imposing multi-view consistency on input multi-view RGB sequences (i.e., asynchronous sequences) of the same subject, we first estimate the sequential phase labels of those sequences through a phase sequence estimator $\mathcal{P}$. These labels are then used for the synchronization task. The estimation is represented as

$$\hat{\boldsymbol{P}} = \mathcal{P}(S), \tag{1}$$

where $S = \{I_1, \ldots, I_n\}$ is an input sequence with n frames and $\hat{\boldsymbol{P}} = \{\hat{\boldsymbol{p}}_1, \ldots, \hat{\boldsymbol{p}}_n\}$ are the corresponding estimated phase labels. The phase label $\hat{\boldsymbol{p}} \in \mathbb{R}^2$ is a 2D point on a unit circle representing the cyclic phase labels [58].

The phase sequence estimator $\mathcal{P}$ is a convolutional neural network (CNN) model as shown in Fig. 2. Initially, there are four convolutional (Conv) layers with a kernel size of 4×4 and a stride of two, and a single fully connected (FC) layer. Each Conv layer is followed by a batch-normalization layer and ReLU activation function. The number of filters is increased from 64 to 512. The FC layer outputs 100-dimensional features. A GRU module is then used to learn the sequential phase information from the latent features, which is a three-layer GRU in which the hidden layer contains 100 nodes. Finally, another FC layer is used to estimate the 2D sequential phase labels $\hat{\boldsymbol{P}}$. A normalization layer is also applied to ensure that $\|\hat{\boldsymbol{p}}_i\|^2 = 1, i \in (1, \ldots, n)$.

Assume that the ground-truth phase labels are $\boldsymbol{P} = \{\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_n\}$. We first compute the estimation loss as follows:

$$L_{\text{p\_mse}} = \frac{1}{n}\sum_{i=1}^{n}\|\hat{\boldsymbol{p}}_i - \boldsymbol{p}_i\|_2^2. \tag{2}$$

Because gait is a continuous movement, the estimated phase labels are assumed to change smoothly and sequentially. We therefore introduce a smoothness loss $L_{\text{p\_smo}}$ as

$$L_{\text{p\_smo}} = \frac{1}{n-1}\sum_{i=1}^{n-1}\|\hat{\boldsymbol{p}}_{i+1} - \hat{\boldsymbol{p}}_i\|_2^2 + \frac{1}{n-2}\sum_{i=2}^{n-1}\|\hat{\boldsymbol{p}}_{i+1} - 2\hat{\boldsymbol{p}}_i + \hat{\boldsymbol{p}}_{i-1}\|_2^2, \tag{3}$$

and a penalty loss $L_{\text{p\_pen}}$ for adjacent frames with disordered phase labels (i.e., reverse evolution of gait stances) as

$$L_{\text{p\_pen}} = \frac{1}{|C|}\sum_{\{i,i+1\}\in C}\|\hat{\boldsymbol{p}}_{i+1} - \hat{\boldsymbol{p}}_i\|_2^2, \tag{4}$$

where C denotes the set of adjacent frame index pairs with disordered phase labels.

Finally, these three loss terms constitute the total loss of the phase sequence estimator, which is represented as

$$L_{\text{phase}} = \lambda_{\text{p\_mse}}L_{\text{p\_mse}} + \lambda_{\text{p\_smo}}L_{\text{p\_smo}} + \lambda_{\text{p\_pen}}L_{\text{p\_pen}}, \tag{5}$$

where $\lambda_{\text{p\_mse}}$, $\lambda_{\text{p\_smo}}$, and $\lambda_{\text{p\_pen}}$ are hyperparameters.

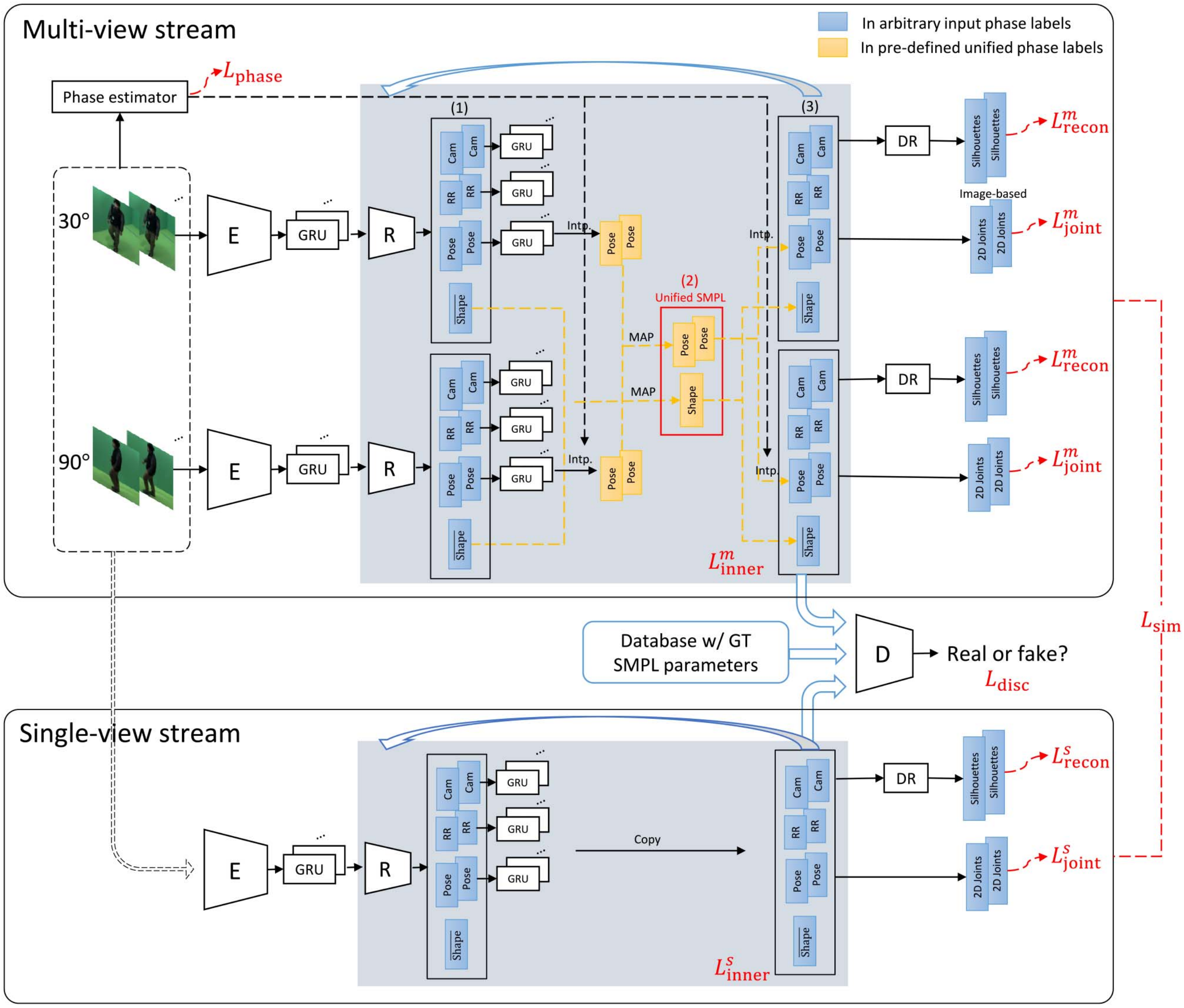


Fig. 1. Proposed multi-view training framework. There is a two-stream encoder (multi-view and single-view streams) and a discriminator. The multi-view stream receives asynchronous multi-view RGB sequences of the same subject as the input, and estimates both view-specific and multi-view unified SMPL parameters through the multi-view union regressor. Here, we take two views as examples. The single-view stream receives each single-view sequence as input and estimates the corresponding SMPL model. The discriminator receives both estimated and ground-truth SMPL models for adversarial learning. There are many losses in the framework, which are jointly trained together.

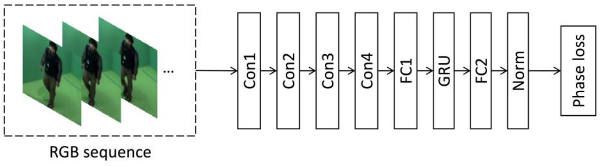


Fig. 2. Network architecture of phase sequence estimator.

Multi-view union regressor: A conventional iterative regressor (e.g., HMR) only regresses the SMPL parameters for each single-view sequence. In contrast, the proposed multi-view union regressor can regress both view-specific and multi-view union SMPL parameters (i.e., unified SMPL) for multi-view sequences, where the union SMPL parameters only contain the pose and shape parameters because the camera and root rotation parameters are view-dependent.

Given asynchronous multi-view RGB sequences of the same subject, the proposed multi-view union regressor first estimates the view-specific SMPL parameters of the each input sequence (see (1) in Fig. 1). We use the same encoder $\mathcal{E}$ (i.e., ResNet-50 [59]) and regressor $\mathcal{R}$ as HMR [49]. Considering the input sequence, we further add a GRU module after the encoder to learn the sequential information for the input sequence, similar to [50]. The regressor $\mathcal{R}$ outputs the SMPL parameters $\hat{\boldsymbol{\Theta}}^m = [(\hat{\boldsymbol{k}}_1^m, \ldots, \hat{\boldsymbol{k}}_n^m), (\hat{\boldsymbol{r}}_1^m, \ldots, \hat{\boldsymbol{r}}_n^m), (\hat{\boldsymbol{\theta}}_1^m, \ldots, \hat{\boldsymbol{\theta}}_n^m), \hat{\boldsymbol{\beta}}^m]$ of the input sequence, where $\hat{\boldsymbol{\beta}}$ is the averaged shape parameter over the sequence. Similarly, we use separate GRU modules to obtain continuous camera $\hat{\boldsymbol{k}}$, root rotation $\hat{\boldsymbol{r}}$, and pose $\hat{\boldsymbol{\theta}}$ parameters.

The estimated sequential phase labels $\hat{\boldsymbol{P}}$ from the phase sequence estimator are then used to interpolate the view-specific pose parameters $\hat{\boldsymbol{\theta}}^m$ in an arbitrary phase sequence (same as the input RGB sequence) to give the new parameters $\hat{\boldsymbol{\theta}}^{m\prime}$ in a pre-defined unified phase sequence for the synchronization. More specifically, we define unified sequential phase labels that are evenly distributed on the phase representation circle. Given the estimated sequential phase labels, we compute the linear weights between the two sequential phase labels for the interpolation. After the interpolation, we obtain the synchronized pose parameters $\hat{\boldsymbol{\theta}}^{m\prime}$ and the shape parameters $\hat{\boldsymbol{\beta}}^{m}$ from multi-view sequences. Through mean average pooling, we obtain the unified pose and shape parameters of the SMPL model for the input subject (see (2) in Fig. 1).

Finally, we apply reverse interpolation to obtain the updated view-specific SMPL parameters $\hat{\boldsymbol{\Theta}}^{m}$ of each single-view sequence for a complete loop (see (3) in Fig. 1). The procedure is repeated several times to ensure a stable estimation.

Following our previous study [44], we propose three loss terms to train the multi-view union regressor. The first is the inner loss L^m_{inner}, which ensures the temporal continuity of the camera, root rotation, and pose parameters. This is defined as the weighted sum of the first- and second-order smoothness terms of each parameter:

$$L^m_{\text{inner}} = \lambda_{\text{cam}} L^m_{\text{cam}} + \lambda_{\text{root}} L^m_{\text{root}} + \lambda_{\text{pose}} L^m_{\text{pose}}, \tag{6}$$

where L^m_{cam} is represented as

$$L^m_{\text{cam}} = \frac{1}{n-1}\sum_{i=1}^{n-1}\left\|\hat{\boldsymbol{k}}^m_{i+1} - \hat{\boldsymbol{k}}^m_{i}\right\|_2^2 + \frac{1}{n-2}\sum_{i=2}^{n-1}\left\|\hat{\boldsymbol{k}}^m_{i+1} - 2\hat{\boldsymbol{k}}^m_{i} + \hat{\boldsymbol{k}}^m_{i-1}\right\|_2^2. \tag{7}$$

The other two terms, L^m_{root} and L^m_{pose}, can be obtained in a similar manner by replacing $\hat{\boldsymbol{k}}^m$ with $\hat{\boldsymbol{r}}^m$ and $\hat{\boldsymbol{\theta}}^m$ in Eq. (7). λ_{cam}, λ_{root}, and λ_{pose} are hyperparameters.

The second loss term is the reconstruction loss L_{recon}, which ensures that the estimated SMPL parameters consider the thickness of various body parts. This term is defined as the mean squared error between the rendered silhouettes and the corresponding ground-truth:

$$L^m_{\text{recon}} = \frac{1}{n}\sum_{i=1}^{n}\left\|\hat{b}^m_i - b_i\right\|_2^2, \tag{8}$$

where the rendered silhouettes $\hat{B}^m = \{\hat{b}^m_1, \ldots, \hat{b}^m_n\}$ are generated through the neural renderer [60] as a differentiable renderer.

The final loss term is the joint loss L_{joint}, which stabilizes the training process and prevents corruption. This is defined as the mean squared error between the estimated 2D joint locations and the corresponding ground-truth based on the image coordinates:

$$L^m_{\text{joint}} = \frac{1}{n}\sum_{i=1}^{n}\left\|\hat{x}^m_i - x_i\right\|_2^2, \tag{9}$$

where the 2D joint locations $\hat{X}^m_{2D} = \{\hat{x}^m_1, \ldots, \hat{x}^m_n\}$ are computed from the SMPL parameters.

Here, we only consider the supervision of 2D joints, because 3D joint estimation is ambiguous for existing single-view approaches (e.g., HMR [49], VIBE [50]), making it unsuitable for use as the pseudo-ground-truth.

In summary, these three loss terms constitute the training loss L_{mv} for multi-view union regressor, which is represented as

$$L_{\text{mv}} = \lambda_{\text{inner}} L^m_{\text{inner}} + \lambda_{\text{recon}} L^m_{\text{recon}} + \lambda_{\text{joint}} L^m_{\text{joint}}, \tag{10}$$

where λ_{inner}, λ_{recon}, and λ_{joint} are hyperparameters.

2) Single-View Stream: The single-view stream is a reduced version of the multi-view stream produced by deleting the phase sequence estimator and the multi-view union module. The SMPL parameters $\hat{\boldsymbol{\Theta}}^s = [(\hat{\boldsymbol{k}}^s_1, \ldots, \hat{\boldsymbol{k}}^s_n), (\hat{\boldsymbol{r}}^s_1, \ldots, \hat{\boldsymbol{r}}^s_n), (\hat{\boldsymbol{\theta}}^s_1, \ldots, \hat{\boldsymbol{\theta}}^s_n), \hat{\boldsymbol{\beta}}^s]$ are independently estimated for each input sequence. We can also generate the rendered silhouettes $\hat{B}^s = \{\hat{b}^s_1, \ldots, \hat{b}^s_n\}$ and the 2D joint locations $\hat{X}^s_{2D} = \{\hat{x}^s_1, \ldots, \hat{x}^s_n\}$.

Regarding the training loss, the single-view stream contains similar losses as the multi-view stream. The first is the inner loss L^s_{inner}, which is defined as

$$L^s_{\text{inner}} = \lambda_{\text{cam}} L^s_{\text{cam}} + \lambda_{\text{root}} L^s_{\text{root}} + \lambda_{\text{pose}} L^s_{\text{pose}} + \lambda_{\text{shape}} L^s_{\text{shape}}. \tag{11}$$

This is slightly different to L^m_{inner}, because it also includes the shape loss L^s_{shape}, which forces the shape parameters $\hat{\boldsymbol{\beta}}^s$ from different sequences of the same subject to be consistent with each other. λ_{cam}, λ_{root}, and λ_{pose} are the same as in Eq. (6).

The second and third loss terms are the same as for the multi-view stream, and are defined as

$$L^s_{\text{recon}} = \frac{1}{n}\sum_{i=1}^{n}\left\|\hat{b}^s_i - b_i\right\|_2^2,$$

$$L^s_{\text{joint}} = \frac{1}{n}\sum_{i=1}^{n}\left\|\hat{x}^s_i - x_i\right\|_2^2. \tag{12}$$

Finally, these three loss terms constitute the training loss L_{sv} for the single-view stream, which is represented as

$$L_{\text{sv}} = \lambda_{\text{inner}} L^s_{\text{inner}} + \lambda_{\text{recon}} L^s_{\text{recon}} + \lambda_{\text{joint}} L^s_{\text{joint}}, \tag{13}$$

where λ_{inner}, λ_{recon}, and λ_{joint} are the same as in Eq. (10).

3) Constraint Between Two Streams: The multi-view stream is capable of capturing the 3D nature of the subject because of the multi-view inputs. Therefore, we impose additional similarity constraints L_{sim} on the estimated SMPL parameters, rendered silhouettes, and 2D joint locations for both the multi-view and single-view streams. As a result, the single-view stream learns to capture more 3D information from the multi-view stream. The similarity loss L_{sim} is defined as

$$\begin{aligned} L_{\text{sim}} = \lambda_{\text{smpl}}\left\|\hat{\boldsymbol{\Theta}}^m - \hat{\boldsymbol{\Theta}}^s\right\|_2^2 + \lambda_{\text{recon}}\left\|\hat{B}^m - \hat{B}^s\right\|_2^2 \\ + \lambda_{\text{joint}}\left\|\hat{X}^m_{2D} - \hat{X}^s_{2D}\right\|_2^2, \end{aligned} \tag{14}$$

where λ_{smpl}, λ_{recon}, and λ_{joint} are hyperparameters; λ_{recon} and λ_{joint} are the same as in Eq. (10).

C. Discriminator

Because we do not have strong 3D supervision (e.g., the ground-truth SMPL parameters), the main supervision is provided by indirect 2D supervision of the silhouette masks and

2D joint locations. Thus, there is a possibility that unrealistic 3D body shapes and poses may minimize the 2D losses because of the ambiguity of unused 1D information. Even in the case of fine-tuning from a well-trained model (e.g., HMR [49] or VIBE [50]) with good initialization, we cannot guarantee that the final model will estimate realistic 3D human bodies, especially for a sequential input with motion. To overcome this issue, we use a discriminator $\mathcal{D}$ to determine whether a sequence of the estimated SMPL parameters is real. More specifically, we consider two sub-discriminators: a frame-level discriminator, responsible for judging the real model of each frame, and a sequence-level discriminator, responsible for judging the real motion of each sequence.

The frame-level discriminator refers to HMR [49] to treat the pose and shape parameters for each single frame independently. Given the estimated pose and shape parameters $\hat{\Phi}_i = [\hat{\boldsymbol{\theta}}_i, \hat{\boldsymbol{\beta}}_i] \in \mathbb{R}^{79}$ of frame i, we train 25 discriminators: one for the shape, 23 for the joints (one for each joint), and a final one for all of the joints together. All pose discriminators receive 9D rotation matrices that have been converted from $\hat{\boldsymbol{\theta}}_i$ via the Rodrigues formula. They share the common feature space of the rotation matrices and only differ in the final classifiers.

The sequence-level discriminator refers to VIBE [50] to treat the pose and shape parameters for each sequence. Given the estimated sequence of pose and shape parameters $\hat{\Phi} = [\hat{\Phi}_1, \ldots, \hat{\Phi}_n]$, we first use a multi-layer GRU model [50] to extract the latent features $h_i = f(\hat{\Phi}_i)$ of the i-th frame. A weighted sum is then used to aggregate the latent features, where the weight of each frame is computed through a self-attention module. Finally, we use a linear layer to classify whether the input $\hat{\Phi}$ constitutes real motion.

For the estimations from both discriminators and the two streams (i.e., $(\hat{\Phi}^m, \hat{\Phi}^m)$ and $(\hat{\Phi}^s, \hat{\Phi}^s)$), we choose the AMASS dataset [52] as the ground-truth SMPL database, and use this for adversarial learning. The adversarial loss function for the two-stream encoder is defined as

$$L_{\text{adv}} = \mathbb{E}\Big[\log\Big(1 - \mathcal{D}\Big(\hat{\Phi}^m, \hat{\Phi}^m\Big)\Big)\Big] + \mathbb{E}\Big[\log\Big(1 - \mathcal{D}\Big(\hat{\Phi}^s, \hat{\Phi}^s\Big)\Big)\Big], \tag{15}$$

and the loss function for the discriminator is defined as

$$\begin{aligned} L_{\text{disc}} = {} & \mathbb{E}\Big[\log\mathcal{D}\Big(\hat{\Phi}^m, \hat{\Phi}^m\Big)\Big] + \mathbb{E}\Big[\log\mathcal{D}\Big(\hat{\Phi}^s, \hat{\Phi}^s\Big)\Big] \\ & + \mathbb{E}\big[\log(1 - \mathcal{D}(\Phi, \Phi))\big], \end{aligned} \tag{16}$$

where Φ and Φ are the ground-truth pose and shape parameters of the single frame and sequence, respectively.

D. Joint Loss Function

Finally, we train the proposed two-stream encoder in an end-to-end manner with a joint loss that combines the aforementioned losses, i.e., Eqs. (5), (10), (13), (14), and (15), which is defined as

$$L_{\text{total}} = L_{\text{phase}} + L_{\text{mv}} + L_{\text{sv}} + L_{\text{sim}} + L_{\text{adv}}. \tag{17}$$

The two losses L_{disc} and L_{total} are iteratively minimized.

E. Training Details and Inference

Input data: The proposed method requires the human-centered cropping RGB sequences and the corresponding silhouettes for training. The sizes of the RGB and silhouette sequences are 224 × 224 and 64 × 64, respectively. See [44] for how to obtain these data. We extract the sequences of 2D joint locations from the cropped RGB sequences using a state-of-the-art pose estimator, i.e., VIBE [50], and use these sequences as the pseudo-ground-truth 2D joint locations. Each sequence contains $n = 25$ continuous frames, because this covers a gait period for most subjects. For sequences of less than 25 frames, we repeatedly select from the frames at the beginning of the sequence.

Training details and parameters: We train the proposed method on 4 Quadro RTX 8000 GPUs using the training set of OU-MVLP containing 5,153 subjects. We first train the phase sequence estimator $\mathcal{P}$ to ensure stable estimated phases. The ground truth labels are computed using the same method as in [58]. Specifically, we first interpolate the original gait cycle to the common gait cycle with a fixed number of frames for all subjects (e.g., evenly choose 20 frames from the original gait cycle for simplicity). Then, we assign cyclic phase labels to a subject and choose it as the standard. We shift the starting frame of the standard and compute the sum of silhouette difference between another subject and this shifted standard for each amount of shift. The two subjects are considered synchronized at the amount of shift with the smallest difference, and have the same phase label. Finally, we reversely interpolate the phase label of the common gait cycle to the original gait cycle. We use the Adam optimizer [61] and set the mini-batch size to 64 randomly selected RGB sequences. The training stage runs for 10 epochs with an initial learning rate of 10^{-4}, and then runs for another five epochs with a reduced learning rate of 10^{-5}. The hyperparameters in Eq. (5) are experimentally set as $\lambda_{\text{p\_mse}} = 1$, $\lambda_{\text{p\_smo}} = 0.01$, and $\lambda_{\text{p\_pen}} = 0.001$.

We use the pre-trained phase sequence estimator $\mathcal{P}$, the encoder $\mathcal{E}$, and the regressor $\mathcal{R}$ of HMR [49] to initialize the corresponding parts of our model; the discriminator uses the default initialization. The GRU modules in the two-stream encoder also have three layers, but different numbers of hidden-layer nodes. Basically, we set the hidden layer to have the same number of nodes as the size of the input feature of the GRU module (e.g., the number of nodes is set to 2048 for the GRU module after the encoder $\mathcal{E}$). To prevent the initial estimation from being destroyed by the pre-trained models, the GRU module is assigned an initial weight of zero and outputs an updated input feature, which is added to the original input feature to give the final output. After initialization, the whole framework is fine-tuned in an end-to-end manner. We use the Adam optimizer [61] for all the models, and set the mini-batch size to 32, i.e., eight subjects and four RGB sequences from different views for each subject. The training runs for 10,000 iterations with an initial learning rate of 10^{-4}, and then for another 10,000 iterations with a reduced learning rate of 10^{-5}. The hyperparameter settings are the same as for the phase sequence estimator in the pre-training process. The remaining hyperparameters are set based on our

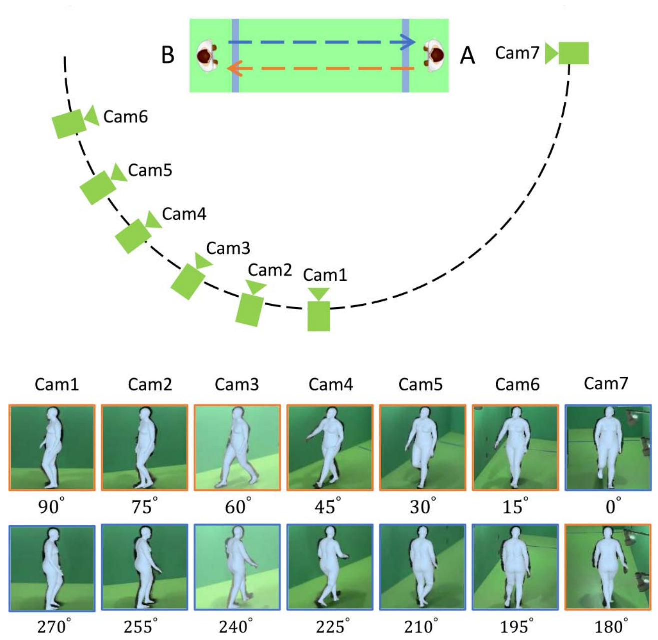


Fig. 3. Camera setup of OU-MVLP and the estimated human mesh models from multiple views.

previous studies [44], [45], i.e., all parameters set to 1, except for $\lambda_{\text{cam}} = 0.1$, $\lambda_{\text{root}} = 0.1$, $\lambda_{\text{pose}} = 0.00001$, and $\lambda_{\text{joint}} = 100$.

Inference: The two-stream design means that the proposed method can infer the SMPL parameters from either multi-view gait sequences or a single-view gait sequence in the test phase. In this study, we use the single-view stream to infer each single-view gait sequence of the entire OU-MVLP, including both training and test sets. This gives the proposed OUMVLP-Mesh database. As such, we can present a fair comparison with existing model-based gait recognition approaches, which suppose that the multi-view sequences are available in the training set while only single-view sequences are available for the probe and gallery in the test set.

If this database was to be used as a kind of pseudo-ground-truth, the best-effort models would be beneficial, i.e., SMPL parameters inferred from all views for the training subjects, although this is not considered in the performance evaluations.

IV. OUMVLP-Mesh Database

Built upon OU-MVLP, the proposed OUMVLP-Mesh database contains 10,307 subjects with up to 14 viewing angles (from 0° to 90°, 180° to 270° at 15° intervals). The raw images are captured at a resolution of 1280×980 pixels at 25 fps under the camera setup shown in Fig. 3. There are seven network cameras placed at 15° intervals along a quarter of a circle. The subjects walk from location A to location B and back to location A, producing 14 view sequences. See [29] for more details about the OU-MVLP database.

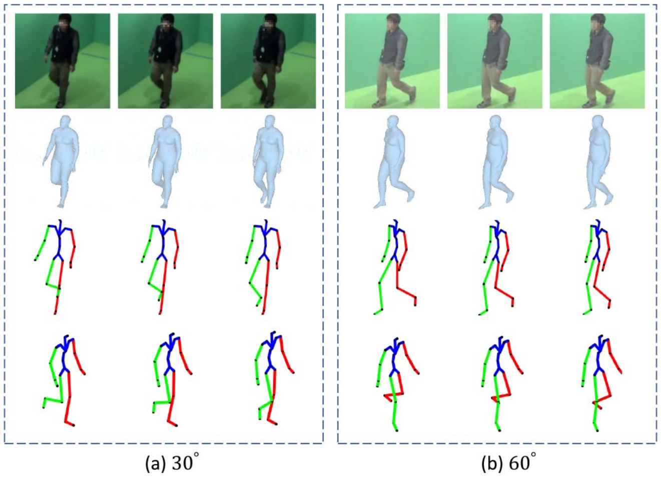


Fig. 4. Examples of the proposed OUMVLP-Mesh database from two different views. The first row is the cropped RGB sequences; the second, third, and fourth rows are the corresponding human mesh models, 2D image-based joint locations, and 3D human-centered joint locations, respectively.

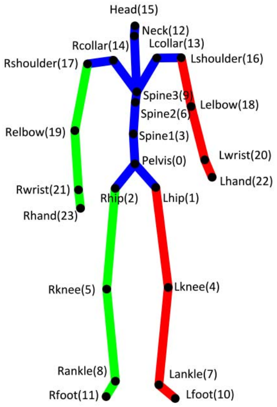


Fig. 5. Human skeleton model with 24 joints.

The human mesh models are estimated from OU-MVLP using the proposed method, as described in Section III. The estimated models from all 14 views are shown in Fig. 3. Some detailed examples are shown in Fig. 4. Besides the 3D meshes, the corresponding skeletons are also provided, and have two different sets of coordinates. The image-based 2D coordinates are the same as those of the poses in OUMVLP-Pose; however, 1D information has been lost and the coordinates change with the viewing angle of the input image. The human-centered 3D coordinates recover the missing information of the input 2D images and are aligned with a common coordinate that is robust to the viewing angle. There are 24 joints in the skeleton shown in Fig. 5, including Pelvis, Lhip, Rhip, Spine1, Lknee, Rknee, Spine2, Lankle, Rankle, Spine3, Lfoot, Rfoot, Neck, Lcollar, Rcollar, Head, Lshoulder, Rshoulder, Lelbow, Relbow, Lwrist, Rwrist, Lhand, and Rhand.

V. Performance Evaluation

We evaluate the proposed method from two aspects. The first is the evaluation of the proposed multi-view training framework for the human mesh estimation. The second is the evaluation of the proposed OUMVLP-Mesh database through cross-view gait recognition approaches.

A. Evaluation Metrics

For the evaluation of human mesh estimation, we mainly focus on the joints and choose two metrics: (1) mean per-joint position error (MPJPE), which is computed as the mean of the

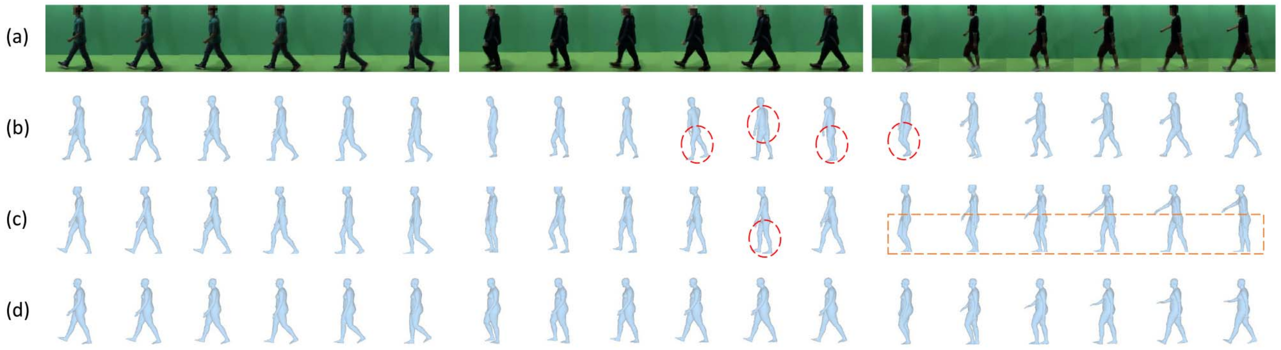


Fig. 6. Examples of estimated human meshes from three different test subjects (side view). (a) shows the continuous input sequences. (b), (c) and (d) show the results of HMR, VIBE, and the proposed method, respectively. The red dotted circle shows a single-frame pose error. The orange dotted rectangle shows consecutive pose errors.

Euclidean distances between the estimated and ground-truth joint locations after root joint alignment (e.g., pelvis joint); (2) Procrustes aligned MPJPE (PA-MPJPE), which is the MPJPE after rigid alignment of the estimation with the ground-truth via Procrustes analysis. For the evaluation of gait recognition, we choose the commonly used rank-1 identification rate (Rank-1) and equal error rate (EER).

TABLE II
MPJPE AND PA-MPJPE [MM] OF ALL METHODS. THE BEST RESULTS ARE SHOWN IN BOLD. THIS CONVENTION IS CONSISTENT THROUGHOUT THE PAPER

Method	18 joints		14 joints	
	MPJPE	PA-MPJPE	MPJPE	PA-MPJPE
HMR [49]	140.3	88.8	114.9	76.7
VIBE [50]	123.8	79.7	102.6	65.2
Ours	**112.6**	**69.7**	**97.2**	**60.8**

B. Evaluation of Human Mesh Estimation

We compare the estimated human meshes of the proposed method with two benchmarks, i.e., HMR and VIBE, in Fig. 6. While both HMR and VIBE have good estimates for the left-most example, there are also wrong estimates for the remaining two examples. We can see that HMR produces more pose errors in a single frame, because it is an image-based human mesh estimation method that does not consider the sequential information between adjacent frames. The video-based VIBE method has the ability to estimate more consecutive poses. However, it may sometimes cause severe consecutive pose errors in the left and right leg positions because of the unclear positional relationship between the two legs in the side-view case (see the row (c) of the right-most example in Fig. 6). In contrast, the proposed method avoids such errors and achieves the best estimation results. This is because the proposed method considers sequential estimation, similar to VIBE, and also uses a multi-view training framework that forces the single-view stream to learn from the multi-view stream. As such, the model is able to estimate accurate left and right leg positions based on the implication from the multi-view sequences in the training, even in the case of difficult scenes (i.e., side view). HMR and VIBE fail to capture the actual body shape for different subjects and show a common shape, while the proposed method learns some characteristics of the actual body shape through the reconstruction loss L_{recon} related to the silhouettes.

We now present a quantitative analysis on a synthesized multi-view gait dataset including the 3D ground-truth poses, because OU-MVLP does not include ground-truth poses. To prepare the synthetic dataset, we first generate 10 human models (five males and five females) using the Makehuman software,[2] and select 10 3D pose sequences (walking pose from 10 subjects) from the CMU MoCap dataset [62]. Using the Blender software,[3] one pose sequence is fitted to one human model and asynchronous walking sequences are generated with 30 frames from four different viewing angles (0°, 30°, 60°, 90°) under a simple green background. The fitted 3D pose sequences provide the corresponding ground-truths. Because the ground-truth joints are not exactly the same as our estimated joints, we only consider the 18 joints that are the same, namely Lhip, Rhip, Lknee, Rknee, Lankle, Rankle, Lfoot, Rfoot, Neck, Head, Lshoulder, Rshoulder, Lelbow, Relbow, Lwrist, Rwrist, Lhand, and Rhand. We also prepare a reduced set of 14 joints by removing the most difficult joints (i.e., toe and finger joints, namely Lfoot, Rfoot, Lhand, and Rhand). From the results presented in Table II, it is clear that the proposed method achieves the smallest estimation errors.

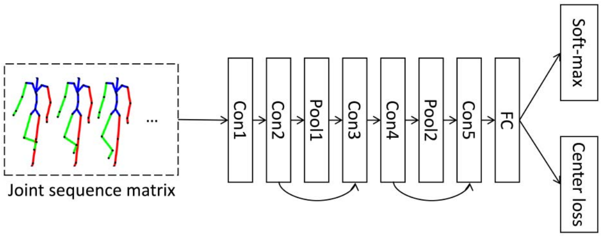


Fig. 7. Network architecture of CNN-Pose.

C. Evaluation of Cross-View Gait Recognition

1) OU-MVLP:

Comparison with different types of joints: We compare the proposed OUMVLP-Mesh with the existing OUMVLP-Pose through a model-based cross-view gait recognition baseline, i.e., CNN-Pose [43]. The detailed network architecture of CNN-Pose is shown in Fig. 7. It takes a joint sequence matrix

[2] https://www.makehumancommunity.org/

[3] https://www.blender.org/

TABLE III
COMPARISON WITH DIFFERENT TYPES OF JOINTS ON OU-MVLP. MEAN RANK-1 RATE AND EER [%] OF ALL 14 VIEW COMBINATIONS USING THE BASELINE METHOD CNN-POSE. '–' INDICATES NOT PROVIDED. '*' INDICATES THE RESULTS IN THE ORIGINAL PAPER

Joint type	Original joints		The same 13 joints	
	Rank-1	EER	Rank-1	EER
OpenPose*	14.76	14.24	-	-
OpenPose	15.73	13.04	12.14	15.44
AlphaPose*	20.42	14.01	-	-
AlphaPose	28.06	11.34	23.63	12.77
Ours-IM2D	38.30	5.50	31.92	6.08
Ours-HC3D	**48.63**	**3.81**	**43.71**	**4.22**

as input, and learns the discriminative features through a backbone CNN. Finally, two losses are used for training. The joint sequence matrix consists of a joint location sequence with N consecutive frames, where each column vector comes from a single frame.

OUMVLP-Pose has two types of joints, as extracted by OpenPose and AlphaPose, and there are 18 2D joints per image. The proposed OUMVLP-Mesh also contains two types of joints: one in the image-based 2D coordinates (IM2D) and the other in the human-centered 3D coordinates (HC3D). There are 24 joints per image. We use these four joint types for the experiments. Considering the different number of joints, we conduct two experiments, one using the original joints and another using the same 13 joints. We re-implemented CNN-Pose and used the same training settings for all types of joints to ensure a fair comparison. More specifically, we followed the original settings in [43] to choose the protocol, number of input frames in a sequence, and mini-batch size. We also used the Adam optimizer with an initial learning rate of 10^{-4}. After 100,000 iterations, we set the learning rate to 10^{-5} and continued for a further 50,000 iterations.

Table III presents the recognition results. Our re-implementation of OpenPose and AlphaPose achieves better performance than reported in the original paper [43]. This may be because we chose better training settings. A comparison using the 2D joints (i.e., OpenPose, AlphaPose, and Ours-IM2D) shows that the proposed method outperforms OpenPose and AlphaPose by a large margin, demonstrating that our approach achieves better pose quality. A comparison using the joints given by the proposed method (i.e., Ours-IM2D and Ours-HC3D) shows that the proposed HC3D joints achieve better performance, indicating that the proposed method can successfully restore accurate HC3D joints that are robust to view changes. The detailed Rank-1 rates and EERs of all view combinations for Ours-HC3D are presented in Tables IV and V. When comparing the performance of different numbers of joints, we find that a higher number of joints tends to produce better performance because more information is available.

Comparison with state-of-the-art methods: Table VI compares Ours-HC3D with some state-of-the-art methods, including four appearance-based methods (i.e., GaitSet [11], GaitPart [63], GLN [64], 3DLocal [65]) and two model-based methods (i.e., ModelGait [44], MvModelGait [45]). Unlike other tables within this paper, which only show the Rank-1 rates without non-enrolled probes, we also present the results with non-enrolled probes. Compared with the state-of-the-art appearance-based methods, Ours-HC3D exhibits worse performance. This is because we use the simple CNN-Pose for feature extraction and the joints have no body shape (or appearance) features. This encourages us to use more informative models (e.g., 3D mesh vertices) that include both shape and pose features and explore more effective networks to boost the performance.

We also note that the proposed method exhibits worse performance than our previous model-based gait recognition studies (i.e., ModelGait and MvModelGait). This is because we mainly focus on model estimation in this paper, while our previous studies are end-to-end frameworks that consider both model estimation and gait recognition. Despite the higher performance, they require more GPU resources and are very time-consuming because of the larger batch sizes required for recognition purposes, making them unsuitable for exploring various recognition networks. Additionally, these model-based approaches mainly rely on single-view sequences for model estimation, so their estimated human models are not as accurate as those given by the proposed method.

2) CASIA-B: Similar to OU-MVLP, we first train the proposed multi-view training framework on the training set of CASIA-B, then generate SMPL models for the whole CASIA-B using the single-view stream. Finally, we compare with an existing work PoseGait [31]. We follow the same protocol as PoseGait, which uses the first 62 subjects for training while the remaining 62 subjects for test. Because PoseGait uses 3D poses, which are first estimated from original 2D poses extracted by OpenPose and then transformed into the human-centered coordinate to reduce the effect of view angles, we also select our HC3D joints for comparison. CNN-Pose is used as the recognition network. The results are shown in Table VII. The proposed method achieves higher performance, which shows the better quality of our HC3D joints. Besides, the generation of our HC3D joints can be done in an end-to-end way, which is more convenient than multi-step estimation of 3D poses in PoseGait.

D. Ablation Study

Multi-view stream: We conduct an ablation study of the proposed method by deleting the multi-view stream while retaining the single-view stream. The results are presented in Table VIII. Without the multi-view stream, the proposed method cannot learn from the multi-view sequences, and so there is a certain performance degradation for both IM2D and HC3D joints. The performance of these two joint types becomes similar, which indicates that the HC3D joints estimated using only single-view stream training are not as accurate as those produced by the proposed two-stream training. We also visualize the estimated joints in Fig. 8. Training using only the single-view stream results in left and right leg position flip errors in the side-view case, similar to VIBE, and the estimated HC3D joints of the two sequences from almost the same phase are inconsistent. In contrast, the

TABLE IV
DETAILED RANK-1 RATES [%] USING OURS-HC3D OF ALL VIEW COMBINATIONS ON OU-MVLP. "P" DENOTES THE PROBE VIEW, "G" DENOTES THE GALLERY VIEW

P \ G	0°	15°	30°	45°	60°	75°	90°	180°	195°	210°	225°	240°	255°	270°	Mean
0°	**71.31**	53.06	40.34	32.23	26.26	21.42	19.17	32.63	32.87	32.18	27.72	21.36	19.80	19.17	32.11
15°	61.54	**82.74**	71.89	59.24	46.86	38.02	30.64	38.69	50.18	50.57	46.17	38.10	33.52	30.63	48.48
30°	51.64	73.11	**82.30**	76.27	63.86	51.97	42.75	36.97	49.18	58.63	57.13	47.93	43.75	40.16	55.40
45°	42.21	61.14	76.87	**84.17**	76.88	64.56	51.43	32.95	46.96	58.96	61.01	55.15	52.50	48.56	58.10
60°	31.10	47.40	62.41	75.98	**84.21**	73.69	57.56	27.37	38.81	50.89	56.86	60.38	56.35	53.98	55.50
75°	27.11	39.13	53.11	65.38	74.66	**82.63**	73.33	24.41	33.35	45.75	53.31	55.33	62.68	64.62	53.91
90°	24.33	33.65	43.54	52.65	60.41	73.98	**81.00**	22.59	29.67	40.93	48.08	52.24	62.16	68.05	49.52
180°	36.93	37.42	34.64	31.52	26.93	22.28	19.80	**79.10**	58.99	50.66	37.94	27.29	23.28	21.10	36.28
195°	35.51	43.77	43.86	40.65	35.20	28.06	24.39	54.42	**82.21**	72.74	57.95	41.31	32.07	27.04	44.23
210°	34.31	45.56	52.60	52.16	46.70	39.49	34.92	46.38	72.75	**85.37**	78.47	59.04	48.07	39.28	52.51
225°	30.11	40.75	49.52	54.37	50.89	46.06	41.00	35.46	59.64	78.31	**85.06**	70.44	59.04	46.85	53.39
240°	22.76	32.40	40.69	48.11	54.78	47.09	43.11	24.99	41.05	57.41	70.08	**80.49**	65.36	51.13	48.53
255°	22.46	30.38	38.04	46.57	50.83	55.82	54.76	22.88	33.16	48.15	59.57	66.37	**79.21**	67.96	48.30
270°	21.92	27.77	34.88	42.68	48.53	56.28	60.49	20.17	26.99	39.22	47.75	51.64	67.38	**78.29**	44.57
Mean	36.67	46.31	51.76	54.43	53.36	50.10	45.31	35.64	46.84	54.98	56.22	51.93	50.37	46.92	48.63

TABLE V
DETAILED EERS [%] USING OURS-HC3D OF ALL VIEW COMBINATIONS ON OU-MVLP. "P" DENOTES THE PROBE VIEW, "G" DENOTES THE GALLERY VIEW

P \ G	0°	15°	30°	45°	60°	75°	90°	180°	195°	210°	225°	240°	255°	270°	Mean
0°	**3.05**	3.92	4.35	4.86	5.88	6.08	6.48	5.48	5.33	5.34	5.68	6.67	6.52	6.29	5.42
15°	3.56	**2.50**	2.85	3.24	3.63	4.35	4.68	4.60	3.90	4.00	4.39	5.06	4.82	4.84	4.03
30°	3.82	2.70	**2.14**	2.45	2.79	3.17	3.54	4.82	3.77	3.28	3.34	3.66	3.71	3.71	3.35
45°	4.57	3.13	2.39	**2.06**	2.44	2.74	3.20	4.98	4.02	3.32	3.05	3.31	3.36	3.22	3.27
60°	5.33	3.78	2.76	2.36	**2.13**	2.45	2.86	5.59	4.56	3.67	3.34	3.16	2.99	3.17	3.44
75°	5.58	4.30	3.24	2.72	2.44	**1.97**	2.16	5.68	4.70	3.60	3.14	3.23	2.57	2.63	3.43
90°	5.68	4.55	3.61	3.11	2.89	2.16	**2.10**	5.67	4.87	3.74	3.14	3.22	2.41	2.30	3.53
180°	4.69	4.67	4.51	4.94	5.36	5.70	5.57	**2.42**	3.39	3.59	4.30	5.00	5.20	5.43	4.63
195°	5.04	3.95	3.96	4.23	4.72	5.07	5.20	3.38	**2.12**	2.72	3.07	4.16	4.46	4.92	4.07
210°	4.82	3.88	3.41	3.35	3.74	3.89	4.06	3.85	2.53	**1.79**	2.25	3.00	3.30	3.57	3.39
225°	5.34	4.27	3.56	3.36	3.58	3.58	3.82	4.50	3.05	2.07	**1.94**	2.60	2.75	3.12	3.40
240°	6.22	5.30	4.20	3.81	3.65	3.81	3.87	5.72	3.87	2.98	2.47	**2.30**	2.49	3.14	3.84
255°	6.14	4.99	4.15	3.61	3.51	3.13	3.12	5.72	4.57	3.44	2.78	2.85	**2.05**	2.57	3.76
270°	6.08	5.03	4.09	3.46	3.37	2.71	2.67	5.89	5.05	3.89	3.20	3.31	2.38	**2.02**	3.80
Mean	4.99	4.07	3.52	3.40	3.58	3.63	3.81	4.88	3.98	3.39	3.29	3.68	3.50	3.64	3.81

TABLE VI
COMPARISON WITH STATE-OF-THE-ART METHODS USING OURS-HC3D ON OU-MVLP. THE RANK-1 RATES AND EERS [%] ARE FOR EACH PROBE VIEW AVERAGED OVER THE 14 GALLERY VIEWS, WHERE THE IDENTICAL VIEW IS EXCLUDED. THE UPPER AND LOWER BLOCKS OF RANK-1 RATES ARE THE RESULTS WITHOUT AND WITH NON-ENROLLED PROBES, RESPECTIVELY. "–" DENOTES NOT PROVIDED. "†" DENOTES APPEARANCE-BASED METHODS. "‡" DENOTES MODEL-BASED METHODS

	Methods	Probe view														
		0°	15°	30°	45°	60°	75°	90°	180°	195°	210°	225°	240°	255°	270°	Mean
Rank-1	GaitSet [11] †	84.7	93.6	96.7	96.7	93.6	95.3	94.2	86.9	92.8	96.0	96.1	93.0	94.5	92.8	93.3
	GaitPart [63] †	-	-	-	-	-	-	-	-	-	-	-	-	-	-	95.1
	GLN [64] †	89.3	95.8	**97.9**	**97.8**	**96.0**	96.7	96.1	90.7	95.3	**97.7**	**97.5**	95.7	96.2	95.3	95.6
	3DLocal [65] †	-	-	-	-	-	-	-	-	-	-	-	-	-	-	**96.5**
	ModelGait [44] ‡	92.8	96.2	96.8	96.3	94.7	96.6	96.6	93.5	95.4	96.3	96.7	96.5	96.5	96.2	95.8
	MvModelGait [45] ‡	**93.5**	**96.5**	97.1	96.9	95.7	**96.8**	**97.1**	**93.7**	**95.6**	96.6	97.0	**97.1**	**97.1**	**97.0**	96.2
	Ours ‡	29.1	45.8	53.3	56.1	53.3	51.7	47.1	33.0	41.3	50.0	51.0	46.1	45.9	42.0	46.1
	GaitSet [11] †	79.5	87.9	89.9	90.2	88.1	88.7	87.8	81.7	86.7	89.0	89.3	87.2	87.8	86.2	87.1
	GaitPart [63] †	82.6	88.9	90.8	91.0	89.7	89.9	89.5	85.2	88.1	90.0	90.1	89.0	89.1	88.2	88.7
	GLN [64] †	83.8	90.0	91.0	91.2	90.3	90.0	89.4	85.3	89.1	90.5	90.6	89.6	89.3	88.5	89.2
	3DLocal [65] †	86.1	**91.2**	**92.6**	**92.9**	**92.2**	**91.3**	**91.1**	86.9	**90.8**	**92.2**	**92.3**	**91.3**	**91.1**	**90.2**	**90.9**
	ModelGait [44] ‡	87.1	89.4	90.9	89.6	88.9	90.0	89.8	87.9	89.2	89.2	89.8	90.2	89.4	89.3	89.3
	MvModelGait [45] ‡	**87.7**	89.7	91.1	90.1	89.8	90.3	90.3	**88.1**	89.4	89.4	90.0	90.8	90.0	89.7	89.7
	Ours ‡	27.2	43.0	49.6	52.5	50.5	48.5	44.2	30.9	38.6	46.3	47.5	43.5	42.9	39.3	43.2
EER	GaitSet [11] †	1.45	0.93	0.76	0.75	0.99	0.79	0.86	2.80	1.61	1.53	2.20	1.83	1.15	1.00	1.33
	ModelGait [44] ‡	0.34	0.34	0.20	0.18	0.31	0.26	0.17	0.28	0.28	0.36	0.34	0.21	0.20	0.20	0.26
	MvModelGait [45] ‡	**0.29**	**0.29**	**0.18**	**0.14**	**0.24**	**0.23**	**0.15**	**0.24**	**0.22**	**0.27**	**0.24**	**0.18**	**0.17**	**0.17**	**0.21**
	Ours ‡	5.61	4.15	3.44	3.36	3.54	3.54	3.64	4.80	4.22	3.51	3.51	3.96	3.89	3.93	3.94

proposed two-stream training produces accurate estimations. These results demonstrate the importance of the multi-view streams for the proposed method.

Training and test view number: We investigate the effect on different number of view angles in the both training and test sets. We set up six cases by increasing the viewing

TABLE VII
COMPARISON WITH POSEGAIT [31] ON CASIA-B. MEAN RANK-1 RATE [%] OF ALL 11 VIEW COMBINATIONS UNDER THREE WALKING CONDITIONS

Methods	Probe sets		
	NM	BG	CL
PoseGait [31]	60.92	39.16	29.71
Ours	**76.64**	**42.01**	**32.81**

TABLE VIII
ABLATION EXPERIMENT. MEAN RANK-1 RATE AND EER [%] OF ALL 14 VIEW COMBINATIONS USING THE BASELINE CNN-POSE METHOD

Multi-view stream	Joint type	24 joints	
		Rank-1	EER
×	Ours-IM2D	32.27	6.68
	Ours-HC3D	31.86	6.34
√	Ours-IM2D	38.30	5.50
	Ours-HC3D	**48.63**	**3.81**

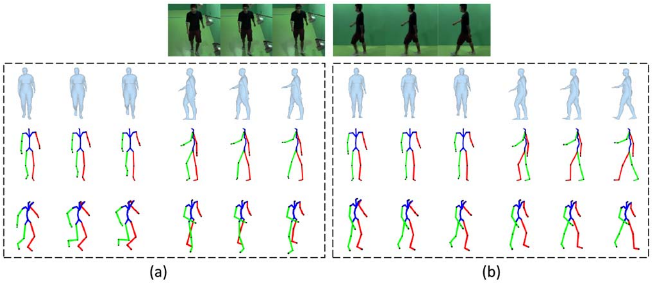


Fig. 8. Ablation study on the multi-view stream. Given two sequences of the same subject in front and side-view cases, (a) shows the estimation results using only single-view stream training, and (b) shows the results using the proposed two-stream training.

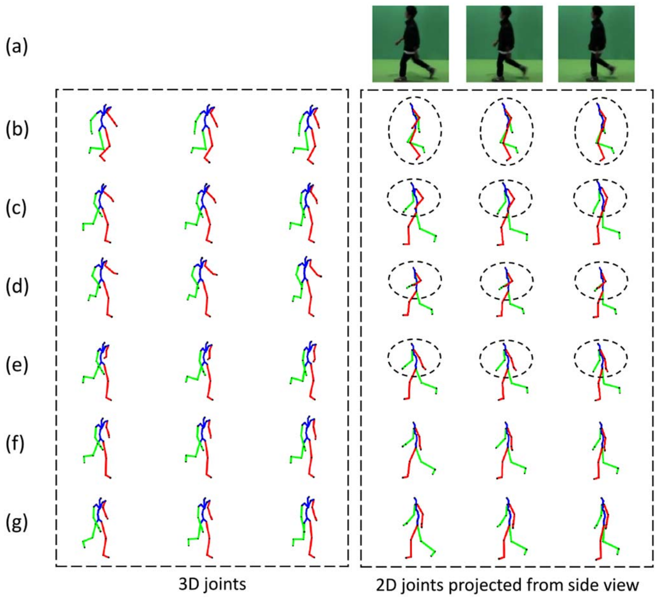


Fig. 9. Ablation study on the training and test view number. (b) to (g) show the estimated joints for six cases from case (1) (one front view) to case (6) (all 14 views); left side shows the unified 3D joints estimated using the multi-view stream; right side shows the corresponding 2D joints projected from side view. (a) shows the ground truth from side view. The black dotted circle shows incorrect poses compared with the ground truth.

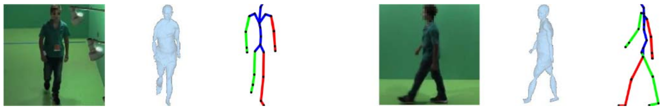


Fig. 10. Ablation study on the discriminator. Models are trained w/o discriminator.

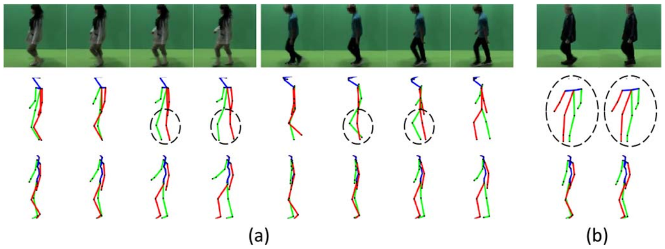


Fig. 11. Some failure cases of the estimated 2D joints produced by OpenPose and AlphaPose. The final row shows the results of the proposed method for comparison. (a) Leg position errors, where the left and right sides are from OpenPose and AlphaPose, respectively; (b) total failure cases of OpenPose. The black dotted circle shows the error frame.

range: (1) 1 view [0°]; (2) 2 views [0°, 15°]; (3) 3 views [0°, 15°, 30°]; (4) 5 views [0°, 15°, 30°, 45°, 60°]; (5) 7 views [0°, 15°, 30°, 45°, 60°, 75°, 90°]; (6) all 14 views. For each case, we use the available views to train the proposed method, and then also combine all available views using the multi-view stream to generate unified 3D joints. We further project the 3D joints to 2D joints from the side view and compare with the ground truth RGB sequences in the side view. The results are shown in Fig. 9. With only 1 view [0°] (see the row (b) of Fig. 9), the proposed method fails to capture the 3D nature of human model and results in totally incorrect poses. With the increasing of view number (viewing range), the proposed method has more information to generate more accurate 3D models.

Discriminator: We also confirm the effect of the discriminator by deleting it from the main framework, and show the visualization of some samples in Fig. 10. Without the discriminator, although the poses are consistent with the input images, the body meshes do not look like real people.

VI. DISCUSSION

A. Advantages Over OpenPose and AlphaPose

Because OpenPose and AlphaPose are general pose estimators trained on pose databases, they may be unsuitable for specific gait databases and could cause many estimation errors. In particular, in the side-view case, the human body is self-occluded and not fully observed. We illustrate some failure cases in Fig. 11. From the results, both OpenPose and AlphaPose produce left and right leg flip errors, and OpenPose may sometimes estimate totally incorrect poses. However, the proposed method gives more accurate estimates, because it is specially designed for gait databases and makes full use of the multi-view sequences in the training stage. Besides the 2D pose, it also provides more informative human models, such as the 3D pose and mesh models.

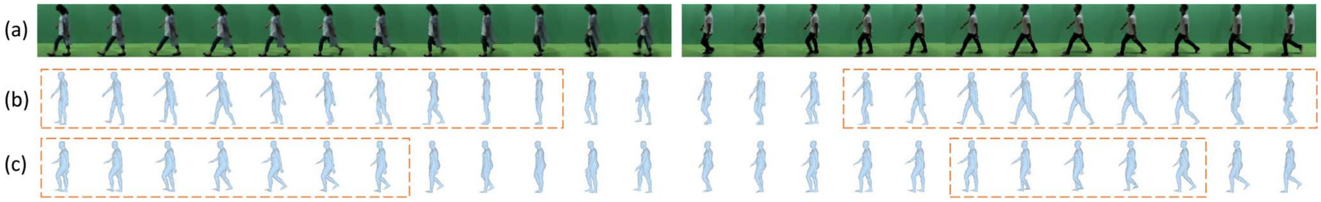


Fig. 12. Failure cases of the proposed method on poses. (a) shows the continuous input sequences. (b) and (c) show the results of VIBE, and the proposed method, respectively. The orange dotted rectangle shows consecutive pose errors.

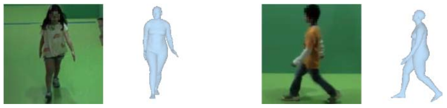


Fig. 13. Failure cases of the proposed method on children's shapes.

B. Use of Multi-View Estimation Stream

Because only single-view sequences exist for the probe and gallery views in the test phase, we only use the single-view estimation stream of the proposed method to infer the proposed OUMVLP-Mesh. However, the multi-view estimation stream would be more useful if multi-view sequences of the same subject were known simultaneously (e.g., multi-view gait recognition), because the estimates from all views could be fused to produce a more accurate unified human model. In addition, the multi-view estimation stream does not require synchronized multi-view sequences, nor does it limit the number of views, making it easier and more extensive to use. We believe our design could inspire other computer vision tasks that deal with multi-view sequences (e.g., multi-view action recognition, where actions include not only the periodic actions but also temporally segmented actions whose starting and ending poses are well defined (e.g., standing up from a sitting position)).

C. Limitations of the Recognition Performance

Because the proposed method is not trained for recognition purposes in an end-to-end manner, it does not outperform the state-of-the-art gait recognition approaches (see Table VI). However, we consider the proposed OUMVLP-Mesh to have the potential to be a valuable database if made publicly available. Thus, it should not be optimized to a specific recognition module to avoid bias, retain greater generality, and enable a wider range of potential applications other than gait recognition databases.

D. Failure Cases

We present some failure cases of the proposed method in Figs. 12 and 13. During training, we use the estimated 2D joints by VIBE as the pseudo ground-truth. Although, the pseudo ground-truth might have some errors (see the row (c) of Fig. 6), the proposed method could fix it through the proposed multi-view training framework (see the row (d) of Fig. 6). However, as shown in Fig. 12, when the estimated 2D joints by VIBE have more errors (i.e., long-term errors), the proposed method could only fix some of them and unavoidably show some errors. For these long-term errors, it's more likely to occur in side view sequences which are the most difficult case to distinguish between left and right legs.

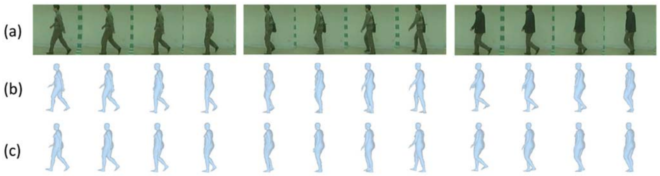


Fig. 14. Examples of estimated human meshes on CASIA-B. (a) shows the continuous input sequences. (b) and (c) show the meshes estimated using the model trained on CASIA-B and OU-MVLP, respectively.

The proposed method also performs poorly on images of children, failing to capture their shape accurately as shown in Fig. 13, which is similar to most 3D human pose and shape estimation methods (e.g., HMR, VIBE). This is because the SMPL model is mainly generated from adults, and does not include children's shapes in its shape space. Despite this, we find that the pose is well estimated.

E. Cross-Dataset Experiment

We make a cross-dataset experiment to investigate the generalization capability of the proposed method. Specifically, we first use the trained model on OU-MVLP to directly estimate the human meshes of CASIA-B. The estimated meshes of CASIA-B are shown in Fig. 14. We find subtle pose differences between the meshes estimated using the model trained on OU-MVLP and those estimated using the model trained on CASIA-B. We further check the quality of the meshes for gait recognition by training CNN-Pose on it using the same protocol as in Section V-C2. The Rank-1 rate for NM, BG, and CL conditions are 64.81%, 38.35%, and 27.01% respectively, with a certain accuracy drop compared with the results trained on the CASIA-B (last row in Table VII). From the results, when trained on the OU-MVLP, the proposed method can handle relatively small domain gap between CAISA-B and OU-MVLP, but with some degradation in recognition quality. However, if faced with more complex and uncontrolled test scenarios, the proposed method may need to be trained on more similar scenarios to improve generalization capability.

VII. Conclusion

We have introduced a multi-view training framework for 3D human mesh model estimation from asynchronous multi-view gait sequences. Using this framework, we generate the OUMVLP-Mesh database based upon an existing multi-view gait database, i.e., OU-MVLP. The proposed OUMVLP-Mesh provides more informative human models (e.g., 3D meshes, 3D/2D joint locations) than current databases. Experimental results show that the proposed framework is able to estimate human mesh models more accurately than the methods compared in this study, and the estimated human mesh models are of sufficient quality to improve the recognition performance of a baseline model-based gait recognition approach. Current model-based gait recognition approaches are still mainly designed for 2D poses. With the proposed 3D pose and the more complex 3D mesh, we believe that more suitable networks are worth exploring to achieve better recognition performance. Additionally, our current method cannot estimate children's shapes accurately because of the limitations of the SMPL model, and so we will attempt to improve it by introducing the children's shape space.

Acknowledgment

The authors thank Stuart Jenkinson, PhD, from Edanz (https://jp.edanz.com/ac) for editing a draft of this manuscript.

References

[1] I. Bouchrika, M. Goffredo, J. Carter, and M. Nixon, "On using gait in forensic biometrics," *J. Forensic Sci.*, vol. 56, no. 4, pp. 882–889, 2011.

[2] H. Iwama, D. Muramatsu, Y. Makihara, and Y. Yagi, "Gait verification system for criminal investigation," *IPSJ Trans. Comput. Vis. Appl.*, vol. 5, pp. 163–175, Oct. 2013.

[3] N. Lynnerup and P. Larsen, "Gait as evidence," *IET Biometr.*, vol. 3, no. 2, pp. 47–54, Jun. 2014.

[4] J. Han and B. Bhanu, "Individual recognition using gait energy image," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 2, pp. 316–322, Feb. 2006.

[5] Y. Guan, C. T. Li, and F. Roli, "On reducing the effect of Covariate factors in gait recognition: A classifier ensemble method," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 7, pp. 1521–1528, Jul. 2015.

[6] Y. Makihara, A. Suzuki, D. Muramatsu, X. Li, and Y. Yagi, "Joint intensity and spatial metric learning for robust gait recognition," in *Proc. CVPR*, Jul. 2017, pp. 5705–5715.

[7] K. Shiraga, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi, "GEINet: View-invariant gait recognition using a convolutional neural network," in *Proc. ICB*, Jun. 2016, pp. 1–8.

[8] Z. Wu, Y. Huang, L. Wang, X. Wang, and T. Tan, "A comprehensive study on cross-view gait based human identification with deep CNNs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 2, pp. 209–226, Feb. 2017.

[9] N. Takemura, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi, "On input/output architectures for convolutional neural network-based cross-view gait recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 9, pp. 2708–2719, Sep. 2019.

[10] K. Zhang, W. Luo, L. Ma, W. Liu, and H. Li, "Learning joint gait representation via quintuplet loss minimization," in *Proc. CVPR*, 2019, pp. 4695–4704.

[11] H. Chao, Y. He, J. Zhang, and J. Feng, "GaitSet: Regarding gait as a set for cross-view gait recognition," in *Proc. AAAI*, 2019, pp. 1–8.

[12] X. Li, Y. Makihara, C. Xu, Y. Yagi, and M. Ren, "Joint intensity transformer network for gait recognition robust against clothing and carrying status," *IEEE Trans. Inf. Forensics Security*, vol. 14, pp. 3102–3115, 2019.

[13] W. Kusakunniran, Q. Wu, J. Zhang, and H. Li, "Support vector regression for multi-view gait recognition based on local motion feature selection," in *Proc. CVPR*, San Francisco, CA, USA, Jun. 2010, pp. 1–8.

[14] Y. Makihara, R. Sagawa, Y. Mukaigawa, T. Echigo, and Y. Yagi, "Gait recognition using a view transformation model in the frequency domain," in *Proc. ECCV*, Graz, Austria, May 2006, pp. 151–163.

[15] Y. Makihara, A. Tsuji, and Y. Yagi, "Silhouette transformation based on walking speed for gait identification," in *Proc. CVPR*, San Francisco, CA, USA, Jun. 2010, pp. 717–722.

[16] D. Muramatsu, A. Shiraishi, Y. Makihara, M. Z. Uddin, and Y. Yagi, "Gait-based person recognition using arbitrary view transformation model," *IEEE Trans. Image Process.*, vol. 24, pp. 140–154, 2015.

[17] A. Mansur, Y. Makihara, R. Aqmar, and Y. Yagi, "Gait recognition under speed transition," in *Proc. CVPR*, Jun. 2014, pp. 2521–2528.

[18] N. Akae, A. Mansur, Y. Makihara, and Y. Yagi, "Video from nearly still: An application to low frame-rate gait recognition," in *Proc. CVPR*, Providence, RI, USA, Jun. 2012, pp. 1537–1543.

[19] S. Yu *et al.*, "GaitGANv2: Invariant gait feature extraction using generative adversarial networks," *Pattern Recognit.*, vol. 87, pp. 179–189, Mar. 2019.

[20] Y. He, J. Zhang, H. Shan, and L. Wang, "Multi-task GANs for view-specific feature learning in gait recognition," *IEEE Trans. Inf. Forensics Security*, vol. 14, pp. 102–113, 2019.

[21] D. Wagg and M. Nixon, "On automated model-based extraction and analysis of gait," in *Proc. 6th IEEE Int. Conf. Automat. Face Gesture Recognit.*, 2004, pp. 11–16.

[22] C. Yam, M. Nixon, and J. Carter, "Automated person recognition by walking and running via model-based approaches," *Pattern Recognit.*, vol. 37, no. 5, pp. 1057–1072, 2004.

[23] A. Bobick and A. Johnson, "Gait recognition using static activity-specific parameters," in *Proc. CVPR*, vol. 1, 2001, pp. 423–430.

[24] D. Cunado, M. Nixon, and J. Carter, "Automatic extraction and description of human gait models for recognition purposes," *Comput. Vis. Image Understand.*, vol. 90, no. 1, pp. 1–41, 2003.

[25] K. Yamauchi, B. Bhanu, and H. Saito, "3D human body modeling using range data," in *Proc. ICPR*, Aug. 2010, pp. 3476–3479.

[26] G. Ariyanto and M. Nixon, "Marionette mass-spring model for 3D gait biometrics," in *Proc. 5th IAPR Int. Conf. Biometr.*, Mar. 2012, pp. 354–359.

[27] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: Realtime multi-person 2D pose estimation using part affinity fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 172–186, Jan. 2021.

[28] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "RMPE: Regional multi-person pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2334–2343.

[29] N. Takemura, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi, "Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition," *IPSJ Trans. Comput. Vis. Appl.*, vol. 10, no. 4, pp. 1–14, 2018.

[30] R. Liao, C. Cao, E. B. Garcia, S. Yu, and Y. Huang, "Pose-based temporal-spatial network (PTSN) for gait recognition with carrying and clothing variations," in *Proc. 12th Chin. Conf. Biometr. Recognit.*, 2017, pp. 474–483.

[31] R. Liao, S. Yu, W. An, and Y. Huang, "A model-based gait recognition method with body pose and human prior knowledge," *Pattern Recognit.*, vol. 98, Feb. 2020, Art. no. 107069.

[32] S. Yu, D. Tan, and T. Tan, "A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition," in *Proc. ICPR*, vol. 4. Hong Kong, China, Aug. 2006, pp. 441–444.

[33] Y. Zhang, Y. Huang, L. Wang, and S. Yu, "A comprehensive study on gait biometrics using a joint CNN-based method," *Pattern Recognit.*, vol. 93, pp. 228–236, Sep. 2019.

[34] R. Gross and J. Shi, "The CMU motion of body (MoBo) database," Robot. Inst., Carnegie Mellon Univ., Pittsburgh, PA, USA, Rep. CMU-RI-TR-01-18, Jun. 2001.

[35] M. Nixon, J. Carter, J. Shutler, and M. Grant, "Experimental plan for automatic gait recognition," Southampton, Rep., 2001.

[36] S. Sarkar, P. J. Phillips, Z. Liu, I. R. Vega, P. G. Ther, and K. W. Bowyer, "The HumanID gait challenge problem: Data sets, performance, and analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 2, pp. 162–177, Feb. 2005.

[37] L. Wang, H. Ning, W. Hu, and T. Tan, "Gait recognition based on procrustes shape analysis," in *Proc. Int. Conf. Image Process.*, vol. 3, 2002, p. 3.

[38] D. López-Fernández, F. J. Madrid-Cuevas, Á. Carmona-Poyato, M. J. Marín-Jiménez, and R. Muñoz-Salinas, "The AVA multi-view dataset for gait recognition," in *Activity Monitoring by Multiple Distributed Sensing*, P. L. Mazzeo, P. Spagnolo, and T. B. Moeslund, Eds. Cham, Switzerland: Springer, 2014, pp. 26–39.

[39] B. DeCann, A. Ross, and J. Dawson, "Investigating gait recognition in the short-wave infrared (SWIR) spectrum: Dataset and challenges," in *Proc. Biometr. Surveillance Technol. Human Activity Identific.*, vol. 8712, 2013, pp. 101–116. [Online]. Available: https://doi.org/10.1117/12.2018145

[40] Y. Iwashita, R. Baba, K. Ogawara, and R. Kurazume, "Person identification from spatio-temporal 3D gait," in *Proc. Int. Conf. Emerg. Security Technol.*, 2010, pp. 30–35.

[41] Z. Zhang *et al.*, "Gait recognition via disentangled representation learning," in *Proc. CVPR*, Long Beach, CA, USA, Jun. 2019, pp. 4705–4714.

[42] H. Iwama, M. Okumura, Y. Makihara, and Y. Yagi, "The OU-ISIR gait database comprising the large population dataset and performance evaluation of gait recognition," *IEEE Trans. Inf. Forensics Security*, vol. 7, pp. 1511–1521, 2012.

[43] W. An *et al.*, "Performance evaluation of model-based gait on multi-view very large population database with pose sequences," *IEEE Trans. Biometrics, Behav., Ident. Sci.*, vol. 2, no. 4, pp. 421–430, Oct. 2020.

[44] X. Li, Y. Makihara, C. Xu, Y. Yagi, S. Yu, and M. Ren, "End-to-end model-based gait recognition," in *Proc. Asian Conf. Comput. Vis. (ACCV)*, Nov. 2020, pp. 3–20.

[45] X. Li, Y. Makihara, C. Xu, and Y. Yagi, "End-to-end model-based gait recognition using synchronized multi-view pose constraint," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV) Workshops*, Oct. 2021, pp. 4106–4115.

[46] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A skinned multi-person linear model," *ACM Trans. Graph.*, vol. 34, no. 6, p. 248, Nov. 2015.

[47] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black, "Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image," in *Computer Vision*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham, Switzerland: Springer Int., 2016, pp. 561–578.

[48] G. Pavlakos, L. Zhu, X. Zhou, and K. Daniilidis, "Learning to estimate 3D human pose and shape from a single color image," in *Proc. CVPR*, Jun. 2018, pp. 459–468.

[49] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik, "End-to-end recovery of human shape and pose," in *Proc. CVPR*, 2018, pp. 7122–7131.

[50] M. Kocabas, N. Athanasiou, and M. J. Black, "VIBE: Video inference for human body pose and shape estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020.

[51] K. Cho, B. V. Merrienboer, C. Gulcehre, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proc. EMNLP*, 2014, pp. 1–11.

[52] N. Mahmood, N. Ghorbani, N. F. Troje, G. Pons-Moll, and M. J. Black, "AMASS: Archive of motion capture as surface shapes," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5441–5450.

[53] J. Liang and M. Lin, "Shape-aware human pose and shape reconstruction using multi-view images," in *Proc. ICCV*, Aug. 2019, pp. 4351–4361.

[54] S. Shin and E. Halilaj, "Multi-view human pose and shape estimation using learnable volumetric aggregation," Nov. 2020, *arXiv:2011.13427*.

[55] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 7, pp. 1325–1339, Jul. 2014.

[56] D. Mehta *et al.*, "Monocular 3D human pose estimation in the wild using improved CNN supervision," in *Proc. 5th Int. Conf. 3D Vis. (3DV)*, 2017, pp. 506–516.

[57] M. Loper, N. Mahmood, and M. J. Black, "MoSh: Motion and shape capture from sparse markers," *ACM Trans. Graph.*, vol. 33, no. 6, p. 220, 2014.

[58] C. Xu, Y. Makihara, X. Li, Y. Yagi, and J. Lu, "Gait recognition from a single image using a phase-aware gait cycle reconstruction network," in *Proc. ECCV*, 2020, pp. 386–403.

[59] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Proc. ECCV*, 2016, pp. 1–15.

[60] H. Kato, Y. Ushiku, and T. Harada, "Neural 3D mesh renderer," in *Proc. CVPR*, 2018, pp. 3907–3916.

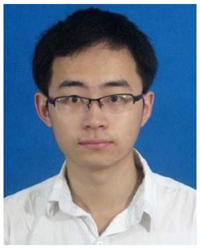
[61] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[62] "NSF EIA-0196217 CMU Graphics Lab Motion Capture Library." [Online]. Available: http://mocap.cs.cmu.edu/ (Accessed: Jul. 8, 2021).

[63] C. Fan *et al.*, "GaitPart: Temporal part-based model for gait recognition," in *Proc. CVPR*, Jun. 2020, pp. 14213–14221.

[64] S. Hou, C. Cao, X. Liu, and Y. Huang, "Gait lateral network: Learning discriminative and compact representations for gait recognition," in *Computer Vision (ECCV)*. Cham, Switzerland: Springer Int., 2020, pp. 382–398.

[65] Z. Huang *et al.*, "3D local convolutional neural networks for gait recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 14920–14929.



Xiang Li received the Ph.D. degree in engineering from the Nanjing University of Science and Technology, China, in 2021. He worked as a Visiting Researcher in 2016 and a Specially Appointed Researcher (part-time) from 2017 to 2020 with SANKEN, Osaka University, Japan, where he is currently a Specially Appointed Researcher (full-time). His research interests are computer vision, image processing, and gait recognition.

Yasushi Makihara received the B.S., M.S., and Ph.D. degrees in engineering from Osaka University in 2001, 2002, and 2005, respectively, where he was appointed as a Specially Appointed Assistant Professor (full-time), an Assistant Professor, and an Associate Professor with The Institute of Scientific and Industrial Research in 2005, 2006, and 2014, respectively, and currently a Professor with the Institute for Advanced Co-Creation Studies. His research interests are computer vision, pattern recognition, and image processing including gait recognition, pedestrian detection, morphing, and temporal super resolution. He has obtained several honors and awards, including the 2nd International Workshop on Biometrics and Forensics in 2014, the IAPR Best Paper Award, the 9th IAPR International Conference on Biometrics in 2016, the Honorable Mention Paper Award, and the Commendation for Science and Technology by the Minister of Education, Culture, Sports, Science and Technology, Prizes for Science and Technology, Research Category in 2014. He has served as an Associate Editor-in-Chief for *IEICE Transactions on Information and Systems*, an Associate Editor for *IPSJ Transactions on Computer Vision and Applications*, the Program Co-Chair for the 4th Asian Conference on Pattern Recognition in 2017, and the Area Chair for ICCV 2019, CVPR 2020, and ECCV 2020. He is a member of IPSJ, IEICE, RSJ, and JSME.

Chi Xu received the Ph.D. degree in engineering from the Nanjing University of Science and Technology, China, in 2021. She worked as a Visiting Researcher in 2016 and a Specially Appointed Researcher (part-time) from 2017 to 2020 with SANKEN, Osaka University, Japan, where she is currently a Specially Appointed Researcher (full-time). Her research interests are gait recognition, machine learning, and image processing.

Yasushi Yagi (Senior Member, IEEE) received the Ph.D. degree from Osaka University in 1991, where he is a Professor with the Institute of Scientific and Industrial Research. In 1985, he joined the Product Development Laboratory, Mitsubishi Electric Corporation, where he worked on robotics and inspections. He became a Research Associate in 1990, a Lecturer in 1993, an Associate Professor in 1996, and a Professor in 2003 with Osaka University, where he was also the Director of the Institute of Scientific and Industrial Research from 2012 to 2015 and the Executive Vice President from 2015 to 2019. His research interests are computer vision, medical engineering, and robotics. He was awarded the ACM VRST2003 Honorable Mention Award, the IEEE ROBIO2006 Finalist of T.J. Tan Best Paper in Robotics, the IEEE ICRA2008 Finalist for Best Vision Paper, the MIRU2008 Nagao Award, and the PSIVT2010 Best Paper Award. International conferences for which he has served as Chair include: FG1998 (Financial Chair), OMINVIS2003 (Organizing chair), ROBIO2006 (Program co-chair), ACCV2007 (Program chair), PSVIT2009 (Financial chair), ICRA2009 (Technical Visit Chair), ACCV2009 (General chair), ACPR2011 (Program co-chair) and ACPR2013 (General chair). He has also served as an Editor for IEEE ICRA Conference Editorial Board from 2007 to 2011. He is the Editorial Member of *International Journal of Computer Vision* and the Editor-in-Chief of *IPSJ Transactions on Computer Vision and Applications*. He is a Fellow of IPSJ and a Member of IEICE and RSJ.