GaitAGE: Gait Age and Gender Estimation Based on an Age- and Gender-specific 3D Human Model

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

Abstract—Gait-based human age and gender estimation has potential applications in visual surveillance, such as searching for specific pedestrian groups and automatically counting customers by different ages/genders. Unlike most existing methods that exploit widely used appearance-based gait features (e.g., gait energy image and silhouettes) or simple model-based gait features (e.g., leg length, stride width/frequency, and head-to-body ratio), we explore a recently popular 3D human mesh model (i.e., skinned multi-person linear model (SMPL)), which is more robust to various covariates (e.g., view angles). Furthermore, instead of the commonly used gender-neutral SMPL model, we propose a simple yet effective method to generate more realistic age- and gender-specific human mesh models by interpolating among male, female, and infant SMPL models using two learned age and gender weights. The age weight controls the proportion of importance between male/female and infant models, which is learned in a data-driven scheme by considering the paired relation between ground-truth ages and age weights. The gender weight controls the proportion of importance between male and female models, which indicates the gender probability. Then, we explore the use of generated realistic mesh models for age and gender estimation. Finally, the human mesh reconstruction and age and gender estimation modules are integrated into a unified end-to-end framework for training and testing. The experimental results on the OU-MVLP and FVG datasets demonstrated that the proposed method achieved both good mesh reconstruction and state-of-the-art age and gender estimation results.

Index Terms—Gait-based age and gender estimation, 3D human mesh reconstruction, age- and gender-specific SMPL.

I. INTRODUCTION

HUMAN age and gender are two important soft biometric traits that are crucial for many practical applications in surveillance scenarios, such as specific pedestrian group search (e.g., finding lost children or the elderly), automatic customer counting scenarios by different age groups or genders for product marketing research, and age/gender-specific recommendations in entertainment scenarios. They can also be used as auxiliary information for popular vision tasks (e.g., person detection, tracking, and re-identification). Generally, the face and gait are two major biometrics used in this topic [1]–[3]. Considering that facial images are often low resolution in surveillance scenarios or even covered by face masks because of the threat of COVID-19, gait is more suitable because it is accessible remotely.

In most previous gait-based age and gender estimation studies, researchers considered age estimation [5], [6] and gender classification [7], [8] separately. And in some recent studies,

X. Li, Y. Makihara, C. Xu and Y. Yagi are with the Department of Intelligent Media, SANKEN, Osaka University, Osaka 567-0047, Japan. Email: {li, makihara, xu, yagi}@am.sanken.osaka-u.ac.jp
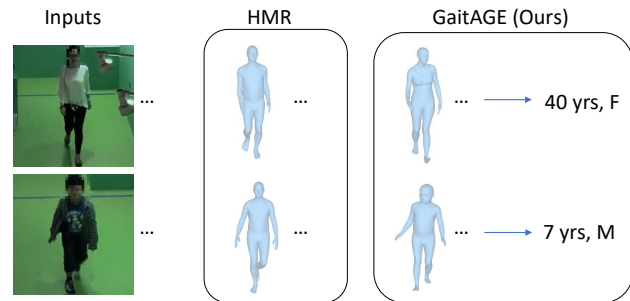
Fig. 1. Given input images, an existing mesh reconstruction method (HMR [4]) reconstructs gender-neutral mesh models while ignoring age and gender information. By contrast, the proposed GaitAGE aims to reconstruct age- and gender-specific mesh models, and further estimate the age and gender from the meshes.

researchers have also considered simultaneous estimation in multi-task pipelines [9]–[11]. The gait features used generally fall into two categories according to the task: appearance-based and model-based features. Appearance-based features mainly contain silhouette-based representations, such as raw silhouette sequences and averaged silhouettes (known as gait energy images (GEIs) [12]), which are more likely to be affected by various covariates (e.g., clothing, carrying, and view angles) and it is also true for gait recognition [13]. While model-based features are more robust to covariates that mainly contain some anthropometric and kinematic representations, such as leg length, stride width/frequency, and head-to-body ratio.

Recently, an advanced human model representation (i.e., skinned multi-person linear model (SMPL) [14]) was proposed. It describes the human body without clothes and carried objects through multiple 3D meshes from a canonical view; hence, it is suitable for gait-based analysis against various covariates. In previous studies [15], [16], researchers validated the effectiveness of SMPL in gait-based identity recognition. Thus, in this paper, we explore the use of SMPL for gait-based age and gender estimation.

However, we note that, like many pure SMPL reconstruction approaches [4], [17], [18], existing approaches [15], [16] for gait recognition only use the default gender-neutral SMPL model while ignoring subjects' gender information. Similarly, a recent study [19] proposed using SMPL for online model-based gait age and gender estimation but also fell into the same problem. In addition, the SMPL model [14] mainly uses adults for training, so it fails to describe children's body shapes and loses the subjects' age information. As shown in Fig. 1, gender-neutral and age-independent SMPL models are

unsuitable for describing the given female and male samples.

There are some studies in which researchers have included age or gender information in the SMPL reconstruction. For gender information, in [20], the authors applied an auxiliary gender classifier to determine whether to use a male or female SMPL model; however, it requires training an additional gender classifier and is somewhat inefficient. For age information, AGORA [21] first generates an extension to linearly blend an adult SMPL body template and infant SMIL template [22] with an optimized weight $\alpha \in [0, 1]$ to approximately describe the shape of children. When α is close to 0, this means that the adult template has a larger weight. When α is close to 1, this means that the infant template has a larger weight. Considering that α is an age-relevant parameter, it can also be used for age estimation.

Inspired by this approach, we consider both age and gender weights (i.e., the age weight α and gender weight w_g), and propose a simple but effective method to generate more realistic age- and gender-specific SMPL models (called SMPL+AG) by interpolating among male, female, and infant SMPL models. Thus, we include both age and gender information in a unified model. Regarding α learning, we rely on the simple assumption that the younger subject tends to have a larger weight for the infant model than the older subject in the interpolation, and thus use a relative relation loss for subject pairs to estimate α in a data-driven manner. Moreover, the proposed SMPL+AG model with age and gender weights is not only responsible for realistic model fitting but also suitable for age and gender estimation in this study. Then, we design age and gender estimation networks to estimate both age and gender information from the reconstructed SMPL+AG models. Finally, we integrate the human mesh reconstruction, and age and gender estimation into an end-to-end training framework.

The contributions are summarized as follows:

(1) SMPL+AG reconstruction

We first generate a novel gait representation (i.e., SMPL+AG model) by interpolating among male, female, and infant SMPL models using learned age and gender weights, which can describe more realistic age- and gender-specific mesh models than existing mesh reconstruction methods, as shown in Fig. 1.

(2) Data-driven learning scheme for the relation between an age and age weight α

Unlike the existing method [23] that defines a pseudo relation between age weight α and four age groups (adults, teenagers, children, and infants), we use a data-driven learning scheme through a pair-based relative relation loss for age weight α estimation.

(3) SMPL+AG model for age and gender estimation

We propose a unified framework to use the reconstructed SMPL+AG model for effective **gait a**ge and **g**ender **e**stimation, which is referred as **GaitAGE**. We also conducted ablation experiments on various sub-features derived from the SMPL+AG model, such as the latent shape feature, pose feature, 3D key joints, and 3D vertices. The experimental results on the OU-MVLP [24] and FVG [25] datasets demonstrated the effectiveness of the proposed method for both mesh reconstruction, and age and gender estimation.

The remainder of this paper is organized as follows: In Section II, we present related work on gait-based age and gender estimation methods, and 3D human pose and shape estimation methods. In Section III, we introduce the proposed end-to-end framework for simultaneous human mesh reconstruction, and age and gender estimation. In Section IV, we present the qualitative and quantitative experimental results. In Section V, we discuss the study. In Section VI, we conclude the paper and consider future work.

II. RELATED WORK

A. Gait-based age and gender estimation

Age estimation. In early studies [26], [27], researchers mainly focused on the classification of several typical age groups (e.g., children, adults, and the elderly), and tended to use model-based features, such as leg length, stride width/frequency, and head-to-body ratio. With the advent of large-scale gait datasets [28] and the popularity of appearance-based features, particularly GEIs, in many recent studies, researchers have applied various traditional machine learning and deep learning techniques to improve the accuracy and precision of age estimation. For example, in some traditional methods, manifold analysis [29], [30] is first used on GEIs to determine the discriminative subspace, then regression methods (e.g., Gaussian process regression [31] and support vector regression [30]) are used to estimate a single age value. Other deep learning methods use various deep convolutional neural networks (e.g., dense convolutional network [32] and deep residual network [6]) on GEIs to learn deep features, which achieve better performance. In addition to a single age value regression, Sakata et al. [5] proposed a label distribution learning framework to estimate the probability distribution of the estimated age, which provided a confidence level for the estimated age.

Gender classification. Compared with age estimation, gender classification is a relatively easy task. In most studies, researchers achieved promising classification accuracy, even using traditional methods. Some of them used model-based features. For example, Huang and Wang [33] fitted the ellipse model to silhouettes and extracted ellipse parameters as features. In most of the remaining studies, researchers used appearance-based features (e.g., GEI or averaged gait image [34]), which avoided additional model fitting. Then they applied some strategies to help to extract more effective features, such as component segmentation guided by human prior knowledge [35], sparse reconstruction-based metric learning [36], and the discrete cosine transform [7]. Finally, they used traditional classifiers (e.g., SVM or XGBoost) to classify the gender.

Multi-task estimation. In some studies, researchers also considered simultaneous age and gender estimation in deep learning-based multi-task pipelines. For example, Marín-Jiménez et al. [37] proposed a multi-task convolutional neural network (CNN) that receives a sequence of optical flow measurements as input and outputs several biometric features, including identity, gender, and age. In [38], the authors proposed a joint CNN-based framework for both gait recognition

and gait-based soft biometric (e.g., age, gender, and body shape) estimation, which receives GEIs and silhouettes as input. Zhang et al. [10] also used GEIs as input, and combined the independent age and gender distribution estimation into a joint distribution estimation. Lau and Chan [11] proposed a tree structure convolutional neural network for age and gender estimation in both single-view and multi-view gait settings. In previous studies, researchers required multiple frames of a gait period; however, Xu et al. [39] investigated a real-time age and gender estimation framework for a single silhouette.

Compared with the aforementioned methods, the proposed method has two main differences: 1) the end-to-end multi-task framework combines both gait feature extraction, and age and gender estimation; and 2) we use a novel model-based gait representation (i.e., SMPL+AG). In [8], although the authors also used the shape feature of the traditional SMPL model to estimate gender, the used SMPL model was age and gender independent.

B. 3D human pose and shape estimation

In most studies on 3D human pose and shape estimation, researchers have aimed at fitting the parameters of SMPL models from monocular images or videos. There are two main paradigms: optimization-based [40] and regression-based [4], [18], [41]. The former attempts to fit the SMPL parameters to the detected 2D key points, whereas the latter is more popular and uses the powerful nonlinear mapping properties of deep neural networks to regress the model parameters. In these paradigms, 3D supervision, such as the ground-truth SMPL parameters or 3D joint annotation, is undoubtedly important for regressing 3D models from 2D images in a single view. However, because of the less accurate 3D ground truth and limited information in a single view, in some studies [17], [42], [43], researchers have attempted to include more information using multi-view inputs for more accurate SMPL parameter inference. However, in all those studies, the researchers only considered the default gender-neutral SMPL model; that is, they ignored the subjects' age and gender information. Although there are some studies [20], [21], [23] in which the researchers considered including age and gender information, they only included one of them (either age or gender) and focused only on model fitting. By contrast, we propose a unified model that merges age and gender simultaneously, and aim at both model fitting, and additional age and gender estimation.

III. METHODOLOGY

A. Overview

Given an RGB sequence of a walking subject, our goal is to extract the 3D mesh model of the subject, and estimate his/her age and gender information from the mesh model. The framework of the proposed method is shown in Fig. 2. For each RGB sequence, we first extract the latent feature through an encoder (E) and GRU unit (G), and feed it into a regressor (R), which learns the SMPL+AG parameters (i.e., SMPL, α, w_g). Then, we use the SMPL+AG parameters to generate the fused mesh model by interpolating among the male, female, and

infant models. We include a discriminator (D) for realistic model fitting. Finally, we apply an age and gender estimator (AG) for age and gender estimation. We jointly train all losses in an end-to-end framework.

B. SMPL+AG: age- and gender-specific SMPL

SMPL+AG is defined as the age- and gender-specific 3D body mesh for a given human body, which is a linearly interpolated mesh model among male, female, and infant SMPL models. It contains the original SMPL parameters $\boldsymbol{\Theta}$ in [4], and additional age weight α and gender weight w_g, that is, $\boldsymbol{\Psi} = [\boldsymbol{\Theta}, \alpha, w_g]$.

Specifically, the SMPL parameter $\boldsymbol{\Theta} = [\boldsymbol{k}, \boldsymbol{r}, \boldsymbol{\theta}, \boldsymbol{\beta}] \in \mathbb{R}^{85}$ consists of several sub-parameters, where $\boldsymbol{k} = [s, t_x, t_y] \in \mathbb{R}^3$ is the camera parameter that represents scale and translation, $\boldsymbol{r} \in \mathbb{R}^3$ is the global root rotation parameter, $\boldsymbol{\theta} \in \mathbb{R}^{69}$ is the pose parameter that represents the relative rotations of 23 joints, and $\boldsymbol{\beta} \in \mathbb{R}^{10}$ is the shape parameter that represents the first 10 coefficients in principal component analysis shape space. From the pose and shape parameters, we can generate a triangulated mesh $\mathcal{M}(\boldsymbol{\theta}, \boldsymbol{\beta}) \in \mathbb{R}^{6890 \times 3}$ with 6,890 vertices and its corresponding 3D key joints $J_{3D} \in \mathbb{R}^{24 \times 3}$. Then, we use the camera and global root rotation parameters to project the 3D key joints to the corresponding 2D key joints $J_{2D} \in \mathbb{R}^{24 \times 2}$ on the 2D image plane in a weak-perspective camera model. The SMPL models, for example, neutral-SMPL, male-SMPL, female-SMPL, and infant-SMPL (i.e., SMIL [22]), differ according to the training dataset (e.g., both male and female, male-only, female-only, and infant-only). As mentioned in Section I, most existing methods ignore age and gender information in SMPL reconstruction. Therefore, we propose a simple but effective method to additionally estimate age weight $\alpha \in [0, 1]$ (e.g., 1 for infants and 0 for adults) and gender weight $w_g \in [0, 1]$ (e.g., 0 for females and 1 for males), and use them for the linear interpolation as follows:

$$\mathcal{M} = w_1 \mathcal{M}_m(\boldsymbol{\theta}, \boldsymbol{\beta}) + w_2 \mathcal{M}_f(\boldsymbol{\theta}, \boldsymbol{\beta}) + w_3 \mathcal{M}_i(\boldsymbol{\theta}, \boldsymbol{\beta}), \quad (1)$$

where $\mathcal{M}_m$, $\mathcal{M}_f$, and $\mathcal{M}_i$ are the meshes for the male-SMPL, female-SMPL, and infant-SMPL models, respectively, generated by the same $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$; and $w_1 = (1 - \alpha)w_g$, $w_2 = (1 - \alpha)(1 - w_g)$, and $w_3 = \alpha$ are their corresponding linear weights, respectively. An example is shown in Fig. 3. We can obtain the key joints J_{3D} and J_{2D} in the same manner.

C. SMPL+AG model reconstruction

In the framework, the encoder, regressor, and discriminator belong to the model reconstruction networks, which are responsible for SMPL+AG model reconstruction. We introduce them as follows:

Encoder. For each input sequence $S = \{I_1, ..., I_n\}$ with n frames, we use ResNet-50 [44] as the encoder to first extract latent features $F = \{f_1, ..., f_n\}, f_i \in \mathbb{R}^{2048}$, similar to many existing human mesh estimation methods (e.g., HMR [4] and VIBE [18]). Considering the continuity of gait motion, we add a gated recurrent unit (GRU) module [45] after the encoder to learn sequential information for the input sequence. It is a three-layer GRU with a hidden dimension size of 2048, which
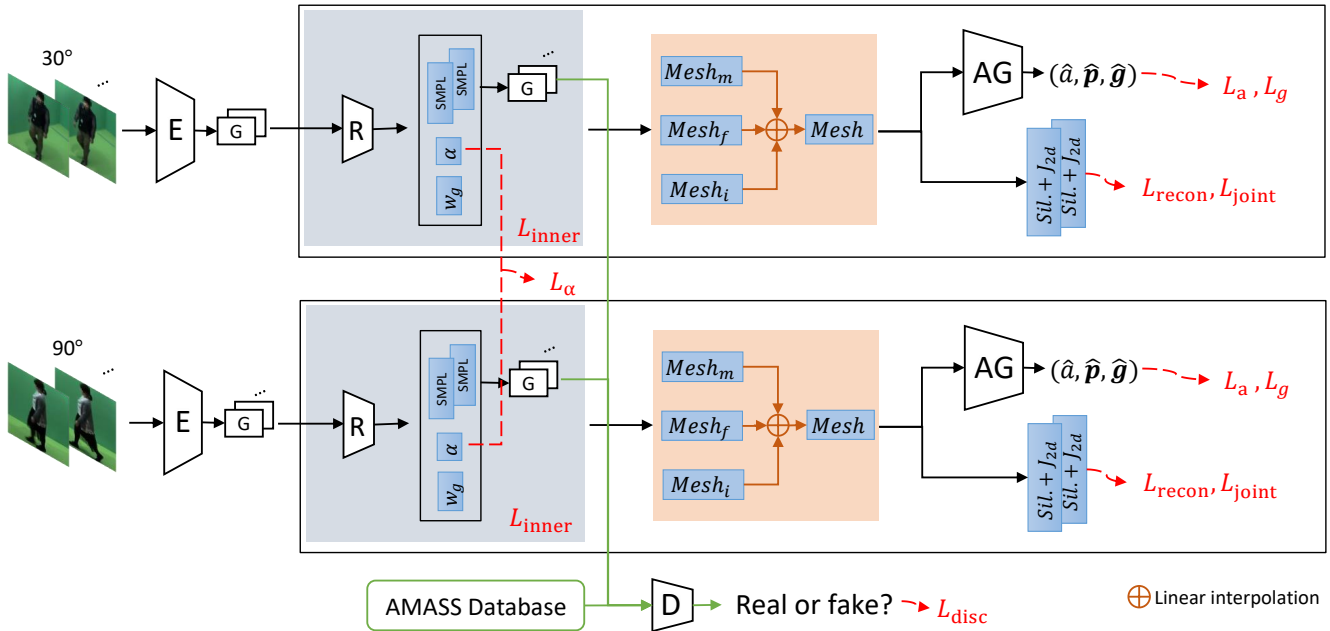
Fig. 2. Framework of the proposed method. Every RGB sequence is considered independently for SMPL+AG model reconstruction (i.e., SMPL, α, w_g), and age and gender estimation. All networks (e.g., the encoder (E), GRU unit (G), regressor (R), discriminator (D), and age and gender estimator (AG)) share weights with each other. Then, for a paired estimated age weight α (a different subject pair as an example in this figure), we use a pair-based relative relation loss L_α to learn α in a data-driven manner.
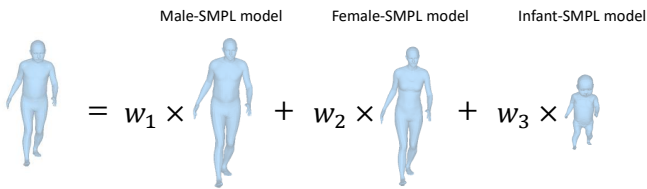

Fig. 3. Example of the proposed SMPL+AG model, where $w_1 = 0.5$, $w_2 = 0$, and $w_3 = 0.5$ ($\alpha = 0.5$ and $w_g = 1$).

is the same size as the input feature. Thus, the output features are represented as $G = \{g(f_1), ..., g(f_n)\}$, where each $g(f_i) \in \mathbb{R}^{2048}$ is learned based on previous frames. Additionally, for stable training, we use a residual design to add the GRU output feature to the input feature as the final feature of the encoder, that is, $F = F + G$.

Regressor. The regressor receives the final encoded features F and regresses the SMPL+AG parameters $\hat{\boldsymbol{\Psi}} = [(\hat{\boldsymbol{k}}_1, ..., \hat{\boldsymbol{k}}_n), (\hat{\boldsymbol{r}}_1, ..., \hat{\boldsymbol{r}}_n), (\hat{\boldsymbol{\theta}}_1, ..., \hat{\boldsymbol{\theta}}_n), \hat{\boldsymbol{\beta}}, \hat{\alpha}, \hat{w}_g]$ for every input sequence, where $\hat{\boldsymbol{\beta}}$ is the averaged shape parameter, and $\hat{\alpha}$ and $\hat{w}_g$ are the averaged age and gender weights, respectively. Additionally, we use separate GRU modules to obtain the continuous camera $\hat{\boldsymbol{k}}$, root rotation $\hat{\boldsymbol{r}}$, and pose $\hat{\boldsymbol{\theta}}$ parameters.

Following [4], we use the regressor in an iterative error feedback loop to recurrently make progressive changes to the current estimate. We initialize it using the mean $\boldsymbol{\Psi}_0$ and regress the residual $\Delta\boldsymbol{\Psi}_t = R([F, \boldsymbol{\Psi}_t])$ in each iteration t. Then, it updates the current estimation: $\boldsymbol{\Psi}_{t+1} = \boldsymbol{\Psi}_t + \Delta\boldsymbol{\Psi}_t$. The difference between our SMPL+AG and the original SMPL parameter regression is the two additional parameter regressions for age and gender weights α, w_g in the regressor.

Finally, we use the estimated SMPL+AG parameters $\hat{\boldsymbol{\Psi}}$ to generate the age- and gender-specific mesh $\hat{M}$ and 2D key joints $\hat{J}_{2D}$, as introduced in Section III-B. We also generate the rendered silhouettes $\hat{Sil}$ of the meshes using the differentiable neural renderer [46].

The regressor loss has three terms, which are represented as

$$L_{\text{reg}} = \lambda_{\text{inner}}L_{\text{inner}} + \lambda_{\text{recon}}L_{\text{recon}} + \lambda_{\text{joint}}L_{\text{joint}}, \quad (2)$$

where λ_{inner}, λ_{recon}, and λ_{joint} are the hyperparameters. $L_{\text{inner}} = \lambda_{\text{cam}}L_{\text{cam}} + \lambda_{\text{root}}L_{\text{root}} + \lambda_{\text{pose}}L_{\text{pose}} + \lambda_{\text{shape}}L_{\text{shape}}$ is the inner loss for the SMPL+AG parameters (i.e., camera, root rotation, pose, and shape parameters). Similar to [15], we define L_{cam}, L_{root}, and L_{pose} as the sum of the first- and second-order smoothness losses for each parameter to ensure temporal continuity within each sequence. We define L_{shape} as the mean squared error (MSE) loss for the shape features from various sequences of the same subject to ensure shape consistency for each subject. Because we do not have direct supervision of the SMPL+AG parameters, we use indirect 2D supervision on the silhouettes and 2D key joints. Thus, we define $L_{\text{recon}} = \|\hat{Sil} - Sil\|_2^2$ and $L_{\text{joint}} = \|\hat{J}_{2D} - J_{2D}\|_2^2$ as the MSE loss between the rendered silhouettes $\hat{Sil}$ or estimated 2D key joints $\hat{J}_{2D}$ and the ground-truth silhouettes Sil or 2D key joints J_{2D}.

Discriminator. To ensure realistic mesh model reconstruction, we apply both frame-level and sequence-level discriminators for adversarial learning. Regarding the frame-level discriminator, we consider the pose and shape parameters $\hat{\Phi}_i = [\hat{\boldsymbol{\theta}}_i, \hat{\boldsymbol{\beta}}_i] \in \mathbb{R}^{79}$ of each frame independently. Regarding the sequence-level discriminator, we consider the pose and shape parameters of all frames $\hat{\boldsymbol{\Phi}} = [\hat{\Phi}_1, ..., \hat{\Phi}_n]$

in a sequence together to learn realistic continuous motion. Regarding the ground-truth SMPL database, we choose the AMASS dataset [47]. The adversarial loss function is then represented as

$$L_{\text{adv}} = \mathbb{E}[\log(1 - \mathcal{D}(\hat{\Phi}, \hat{\mathbf{\Phi}}))], \qquad (3)$$

and the loss function for the discriminator is represented as

$$L_{\text{disc}} = \mathbb{E}[\log \mathcal{D}(\hat{\Phi}, \hat{\mathbf{\Phi}})] + \mathbb{E}[\log(1 - \mathcal{D}(\Phi, \mathbf{\Phi}))], \qquad (4)$$

where $(\Phi, \mathbf{\Phi})$ are the ground truths and $(\hat{\Phi}, \hat{\mathbf{\Phi}})$ are the estimations.

D. Constraint on age weight α and gender weight w_g

The age and gender weights, α and w_g, respectively, control the interpolation between the male-SMPL, female-SMPL, and infant-SMPL models, which is vital for the proposed SMPL+AG model.

Regarding the age weight, in [23], the authors used an ad hoc method that learns α based on a predefined pseudo relation between four age groups (adults, teenagers, children, and infants) and α. However, the pseudo relation might not describe the real relation between α and the ground-truth age well because there is no such ground truth. By contrast, we rely on the simple assumption that a younger subject tends to have a larger age weight for the infant model than an older subject in the interpolation; thus, we use a relative relation loss for subject pairs to estimate α in a data-driven manner. Specifically, we apply the signed quadratic contrastive loss used in [48], which was originally designed for gait relative attribute learning. Given the i-th pair of estimated age weights $(\hat{\alpha}_{1,i}, \hat{\alpha}_{2,i})$ and the corresponding ground-truth age $(a_{1,i}, a_{2,i})$ in a mini-batch, we define three relative relations and their labels y_i as

$$\begin{cases} d_i = 0, (y_i = 0 & if\ a_{1,i} = a_{2,i}|(a_{1,i} \geq T\ \&\ a_{2,i} \geq T)) \\ d_i > 0, (y_i = 1 & if\ a_{1,i} < a_{2,i}\ \&\ a_{1,i} < T) \\ d_i < 0, (y_i = -1 & if\ a_{1,i} > a_{2,i}\ \&\ a_{2,i} < T) \end{cases},$$

$$(5)$$

where $d_i = \hat{\alpha}_{1,i} - \hat{\alpha}_{2,i}$ is the signed L1 distance between the paired age weight and T is an age threshold. This design is based on three guidelines: 1) the subjects in a pair should have similar age weights if they are the same age or both adults (i.e., older than T years old); 2) if the first subject in a pair is younger than the second subject and his/her age is less than T years old, then his/her age weight is larger than that of the other subject (i.e., $d_i > 0$); 3) if the second subject in a pair is younger than the first subject and his/her age is less than T years old, then his/her age weight is larger than that of the other subject (i.e., $d_i < 0$). The age weight relative relation loss is then summarized as

$$L_{\alpha} = \sum_{i=1}^{M} [(1 - |y_i|)d_i^2 + |y_i| \max\{0, (m - y_i d_i)|m - y_i d_i|\}],$$

$$(6)$$

where M is the total number of pairs in the mini-batch and m is a constant margin.

Regarding gender weight w_g, we essentially do not impose any constraint on it because it can be learned through the
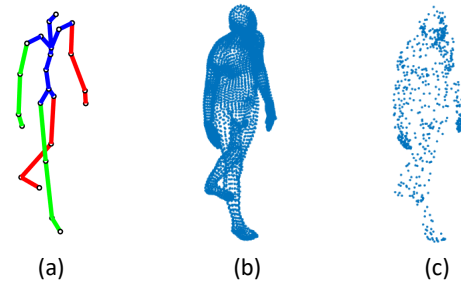

Fig. 4. Example of the sub-features derived from the SMPL+AG model: (a) 3D key joints, (b) 3D vertices, and (c) downsampled 3D vertices.
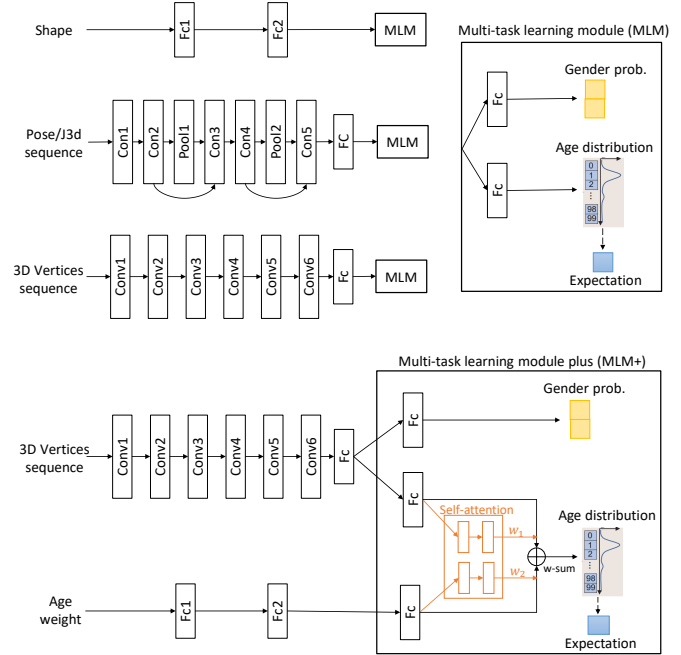

Fig. 5. Various age and gender estimators for various sub-features derived from the SMPL+AG parameters.

proposed age and gender estimator. However, when using shape or pose features for age and gender estimation, it fails to obtain correct estimations. This is because shape and pose features are not generated from w_g; hence, the backpropagated gradients from the age and gender estimator fail to update w_g. Therefore, in this case, we add an additional loss for w_g similar to L_g in Section III-E.

E. Age and gender estimator

After obtaining the SMPL+AG parameters $\hat{\mathbf{\Psi}}$ of an input sequence, we use them for age and gender estimation. Because previous SMPL-based works on gait recognition [15], [16] report that shape features perform better than pose features, i.e., the difference in sub-features of the SMPL model affects the accuracy of individual recognition tasks, we therefore investigate whether the difference in sub-features affects the accuracy of age and gender estimation tasks too.

Specifically, we choose the following five types of sub-features: (1) shape feature; (2) pose feature sequence; (3) 3D key joint sequence; (4) 3D vertices sequence; and (5)

3D vertices sequence and age weight $\hat{\alpha}$. We compute the 3D key joints and 3D vertices by setting the global root rotation parameter to zero, which makes them have a common view-invariant coordinate (also called the human-centered coordinate in [43]). Some sub-features are shown in Fig. 4. For various features, we design their corresponding networks and then feed them into a common multi-task learning module (MLM) for simultaneous age and gender estimation. Note that for each model training, we only choose one of the five features. Fig. 5 shows the detailed network designs.

Specifically, for the shape feature $\hat{\boldsymbol{\beta}} \in \mathbb{R}^{10}$, we feed it into two fully connected (FC) layers with 1024 and 512 nodes.

For the pose feature sequence $[\hat{\boldsymbol{\theta}}_1, ..., \hat{\boldsymbol{\theta}}_n] \in \mathbb{R}^{69 \times n}$ and 3D key joint sequences $[\hat{J}_1, ..., \hat{J}_n] \in \mathbb{R}^{24 \times 3 \times n}$, we first reshape them into 2D matrices of dimension $69 \times n$ and $72 \times n$, and then feed them into the CNN-Pose network [49]. There are five convolutional layers (Conv) with kernel size 3×3, and two max-pooling layers (Pool) with kernel size 2×2 and stride 2. The channel numbers of the Conv layers are 32, 64, 64, 128, and 128. Each Conv layer is followed by a ReLU activation function. There are also two skip connections in the third and fifth Conv layers. The final FC layer has 512 nodes.

For the 3D vertices sequence $[\hat{\mathcal{M}}_1, ..., \hat{\mathcal{M}}_n] \in \mathbb{R}^{6890 \times 3 \times n}$, because the original number of vertices (i.e., 6,890) is too large, we first evenly downsample it 10 times (see (c) in Fig. 4) and then reshape it into a 2D matrix of dimension $2067 \times n$. We apply a new network, considering that there is a larger number of dense points than key joints. There are six Conv layers with kernel size 3×3. The channel numbers of Conv layers are 64, 64, 128, 128, 256, and 256. Each Conv layer is followed by a BatchNorm layer and ReLU activation function. All Conv layers have a stride of 2 in the vertical direction, and the third and sixth Conv layers have a stride of 2 in the horizontal direction. The final FC layer has 512 nodes.

The upper three networks are then followed by an MLM module, which separately learns age and gender embeddings using two independent FC layers with 512 nodes. Finally, it regresses the age label distribution $\hat{\boldsymbol{p}} = [\hat{p}_0, \hat{p}_1, .., \hat{p}_{99}] \in \mathbb{R}^{100}$, where $\hat{p}_i$ is the estimated probability for integer age i, and gender probability $\hat{\boldsymbol{g}} = [\hat{g}, 1 - \hat{g}] \in \mathbb{R}^2$ with the softmax function. The estimated age is the expectation of the estimated age probability distribution $\hat{a} = \sum_{i=0}^{99} i \hat{p}_i$.

Moreover, considering that age weight $\hat{\alpha} \in \mathbb{R}^1$ is age-relevant, we combine it with the 3D vertices sequence to further improve age estimation accuracy. We use two FC layers with 512 nodes to extract the latent feature, which we then feed into an MLM+ module. The module fuses the age embeddings from the 3D vertices sequence and age weight via a self-attention module that learns their respective summed weights.

We assume that the estimated ages, age label distributions, and gender probabilities of the multi-view and single-view streams are $(\hat{a}, \hat{\boldsymbol{p}}, \hat{\boldsymbol{g}})$, and the corresponding ground-truth age, age label distribution, and gender class are $(a, \boldsymbol{p}, g)$, where we set $\boldsymbol{p}$ to a discrete normal distribution $\mathcal{N}(a, 1)$ and $g \in \{0, 1\}$. Following [39], the age loss is defined as

$$L_\text{a} = JS(\hat{\boldsymbol{p}} \| \boldsymbol{p}) + \|\hat{a} - a\|_1 + \lambda_\text{ap} L_\text{ap}, \tag{7}$$

where JS is the Jensen–Shannon divergence loss [50] and L_ap is the age pair similarity loss that forces the estimated ages to be similar for the same subject sequences in the mini-batch. The gender loss is defined as

$$L_\text{g} = BCE(\hat{\boldsymbol{g}}, g) + \lambda_\text{gp} L_\text{gp}, \tag{8}$$

where BCE is the binary cross-entropy loss and L_gp is the gender pair similarity loss that forces the estimated genders to be similar for the same subject sequences in the mini-batch.

The age and gender losses are then combined to an union loss represented as

$$L_\text{ag} = \lambda_\text{a} L_\text{a} + \lambda_\text{g} L_\text{g}, \tag{9}$$

F. Extension to multi-view sequences

In the case in which there are multi-view sequences of the same subject in the training set, we can easily extend the current single-view framework to improve model fitting accuracy by including another multi-view regression stream similar to the approach in our previous study [43]. Specifically, because multi-view sequences of the same subject are usually asynchronous with different starting phases (e.g., start with the double-support or single-support phase), we first apply a phase estimator to estimate the phase labels of each sequence, and synchronize the estimated view-specific SMPL+AG parameters for arbitrary phase labels to the predefined unified phase labels through linear interpolation based on the estimated phase labels. After the synchronization of the multi-view sequences, we can simply apply mean average pooling to obtain the unified SMPL+AG parameters. Note that we exclude the camera and root rotation parameters because they are view-dependent. Finally, we apply reverse interpolation to obtain the updated view-specific SMPL+AG parameters and further estimate the age and gender information from the multi-view stream. We maintain the sharing of network parameters between the single-view and multi-view streams.

The multi-view stream loss function L_mv has similar terms to the single-view stream, including the regressor loss L_reg, adversarial loss L_adv, discriminator loss L_disc, age weight relative relation loss L_α, and age and gender loss L_a, L_g. Additionally, we include the phase estimation loss L_phase, which is defined in [43]. Moreover, we also define the similarity loss L_sim for the estimated SMPL+AG parameters, and age and gender estimation results between the single-view and multi-view streams. This enables the single-view stream to learn 3D information from the multi-view stream.

G. Joint loss functions

We train the entire framework with all the aforementioned losses in an end-to-end manner. To enable a clear understanding, we use superscripts to differentiate the losses from the single-view and multi-view streams, where the single-view stream loss is $L_\text{sv} = L_\text{reg}^s + L_\text{adv}^s + \lambda_\alpha L_\alpha^s + \lambda_\text{ag} L_\text{ag}^s$ and the multi-view stream loss is $L_\text{mv} = L_\text{phase}^m + L_\text{reg}^m + L_\text{adv}^m + \lambda_\alpha L_\alpha^m + \lambda_\text{ag} L_\text{ag}^m$. We combine them into the total loss:

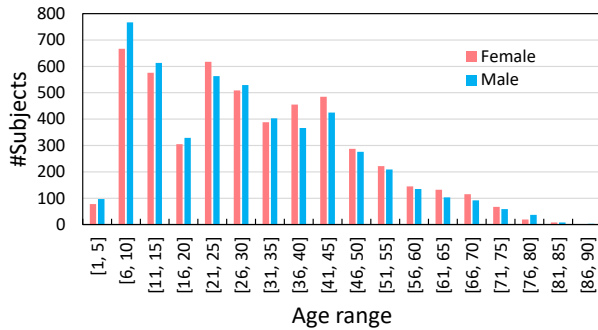$$L_\text{total} = L_\text{sv} + L_\text{mv} + L_\text{sim}, \tag{10}$$

Fig. 6. Subjects's gender and age distribution of OU-MVLP.

where L_{sim} is the similarity loss between the single-view and multi-view streams. Finally, we iteratively minimize L_{total} and $L_{\text{disc}} = L_{\text{disc}}^s + L_{\text{disc}}^m$.

IV. EXPERIMENTS

A. Datasets

We used two datasets, that is, OU-MVLP [24] and FVG [25], to evaluate the proposed method.

OU-MVLP is one of the world's largest gait datasets, with a wide view variation. It contains 10,307 subjects (5,114 males and 5,193 females) with ages ranging from 2 to 87 years old. Fig. 6 shows subjects' gender and age distribution. Each subject was captured from 14 views, ranging from 0°–90° and 180°–270° in 15° intervals, with two sequences ("00" and "01") captured for each view. Following the original protocol, we used 5,153 subjects for training and the other disjoint 5,154 subjects for testing.

FVG is a front-view gait dataset of subjects in real-world outdoor scenes. It contains 226 subjects captured with variations from three near-front views (-45°, 0°, and 45°). There are five variations in the data: normal, walking speed (slow and fast), clothing changes, carrying/wearing change (bag or hat), and clutter background (multiple persons). Because the dataset does not provide age and gender information, we only evaluated gender estimation performance on this dataset using our annotated gender information[1]. As a result, there were 146 males and 80 females. Following the new data split on this dataset, we used 136 subjects (85 males and 51 females) for training and the other disjoint 90 subjects (61 males and 29 females) for testing. Additionally, 12 subjects in the test set were collected twice in the elapsed time. Thus, there was a total of 306 sequences in the test set. We only chose normal sequences (with sequence IDs "01", "02", and "03") for this experiment.

B. Training details and inference

In the training phase, we prepared cropped RGB sequences and the corresponding silhouettes, referring to [15], for OU-MVLP. For FVG, we used PointRend [51] to automatically detect and segment human bodies from raw RGB sequences. Additionally, because the sequence contained some standing

[1] We read the gender labels from raw RGB images.

frames at the beginning and some occluded frames at the end, we only used approximately 60% of the middle frames. We resized the cropped RGB sequences and silhouettes to 224×224 and 64×64, respectively. We extracted the pseudo ground-truth 2D key joints J_{2D} using VIBE [18]. We chose 25 continuous frames for each sequence because this covered a gait period for most subjects. If the sequence had fewer than 25 frames, we repeated the selection from the beginning.

We trained the proposed model on four Quadro RTX 8000 GPUs. We chose Adam as the optimizer. The mini-batch size (P, K) means that there were P subjects, and each subject had K sequences from various views in the mini-batch. We set the initial learning rate to 10^{-4} for all networks, except the age and gender estimator, which we set to 10^{-5}. For OU-MVLP, we set the mini-batch size to $(12, 4)$ and the total iteration to 20,000, and decreased the learning rate by a factor of 10 at 10,000 iterations. For FVG, we set the mini-batch size to $(12, 3)$ and the total iteration to 1,000, and decreased the learning rate by a factor of 10 at 500 iterations. Additionally, there were many weight parameters in the total loss function. For simplicity, we kept the weight hyperparameters unchanged for the same loss functions as in [39], [43] (e.g., $\lambda_a = 1$, $\lambda_g = 2$, $\lambda_{ap} = 1$, $\lambda_{gp} = 0.01$), and experimentally set $\lambda_\alpha = 10$ and $\lambda_{ag} = 0.8$ to achieve a balance between the model reconstruction performance and age and gender estimation accuracy. Regarding the hyperparameters in L_α, we empirically set $T = 18$ because 18 is regarded as the age threshold for adults, and experimentally set $m = 0.4$. Additionally, because the multi-view stream was similar to that in [43], please refer to that paper for more details. For inference, we chose 25 frames in each sequence and only used the single-view stream for model reconstruction, and age and gender estimation. Without being specified, the proposed method uses the 3D vertices + $\hat{\alpha}$ of the SMPL+AG model for age and gender estimation.

C. Evaluation metrics

For the performance evaluation of age estimation, we used two widely used metrics: the mean absolute error (MAE) and cumulative score (CS). Given the estimated age $\hat{a}_i$ and ground-truth age a_i of a test sequence, the MAE is defined as the MAE of these two ages:

$$\text{MAE} = \frac{1}{N_s} \sum_{i=1}^{N_s} |\hat{a}_i - a_i|, \tag{11}$$

where N_s is the total number of test sequences. The CS measures the percentages of test sequences under various absolute error tolerances, for example, the CS(j) is computed as

$$\text{CS(j)} = \frac{N_{e \leq j}}{N_s}, \tag{12}$$

where $N_{e \leq j}$ is the number of test sequences whose MAE is less than j years old. For the performance evaluation of gender classification, we used the correct classification rate (CCR).
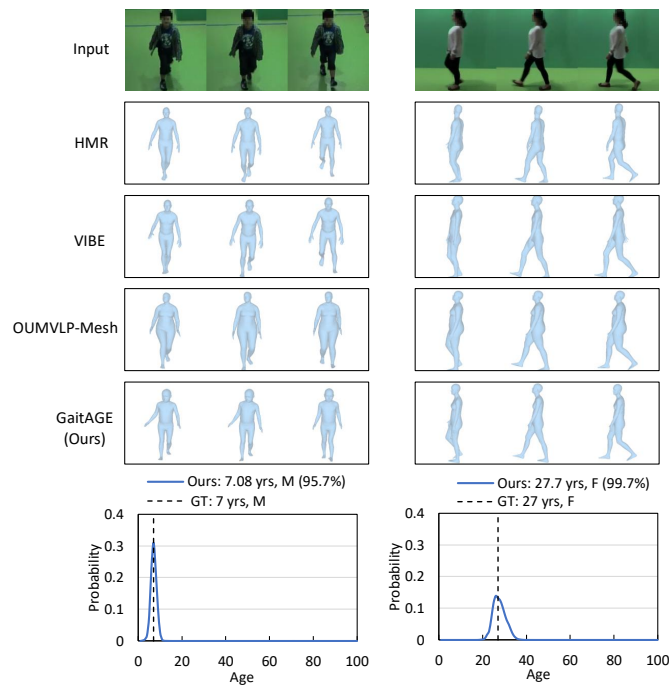
Fig. 7. Visualization of two examples (left) and (right). The rows from the top to bottom are the input sequences, reconstructed mesh models of HMR [4], VIBE [18], OUMVLP-Mesh [43], the proposed method, and the estimated age distribution and expected age of proposed method.

D. Visualization

Figure 7 visualizes two reconstructed mesh examples for the proposed method and three comparison methods (i.e., HMR [4], VIBE [18], and OUMVLP-Mesh [43]). Regarding the proposed method, we chose 3D vertices + $\hat{\alpha}$ as the input of the age and gender estimator by default because it was the best sub-feature. The three comparison methods failed to reconstruct age- and gender-specific mesh models for the selected examples, which is not convincing, particularly for the children. By contrast, the proposed method captured age and gender information from the inputs well and reconstructed reasonable age- and gender-specific mesh models. We also noticed that some mesh models of HMR from the side view had left and right leg flip errors, for example, the middle frame of the right example of HMR. This is because HMR was single-view method and may not have captured the correct position of the legs from very difficult side view scenarios, whereas the proposed method handled this because it had the same multi-view design as that in [43]. Additionally, we presented the estimated age distribution and expected age of the proposed method, where we achieved good age estimation results. The age distribution curve indicates the confidence level in the prediction: the higher the curve, the higher the confidence. From the results, the proposed method was more confident in predicting the ages of the child than the adult. This is because as children grew, their gait regularly changed in physical characteristics (e.g., taller height, smaller head-to-body ratio), which results in obvious gait differences between children of different ages, in other words, gaits are sensitive to aging during childhood, leading to easier age estimation

and higher confidence. However, adults had limited physical variation in gait (e.g., between 20 and 45 years old), in other words, gaits are insensitive to aging during adulthood, thus making age estimation more difficult and less confident. These are consistent with the results of [5].

E. Comparison with state-of-the-art methods

OU-MVLP. We compared the proposed method with state-of-the-art gait-based age and gender estimation methods (i.e., GEINet [5], Upadhyay's method [7], TBResNet [10], Xu's method [39], and Shehata's method [19]); two famous gait recognition networks (i.e., GaitSet [52] and GaitGL [53]); and a series of well-known ResNet backbones (i.e., ResNet-18, ResNet-34, ResNet-50 [44]). We reimplemented GEINet [5], which only estimates an age label distribution using the KL divergence loss function. For methods [7], [10], [19], [39], we used the results in the original papers. Because GaitSet [52] and GaitGL [53] were initially designed for gait recognition, we slightly modified them by adding the proposed MLM module to regress the age and gender labels. In addition to silhouettes, we used RGB sequences as input for evaluation. In this case, we changed the input channels of the first Conv layer of GaitSet and GaitGL to 3. We also added the MLM module to ResNet series models. For various models, we set the batch size so that it was the same as ours and chose the appropriate number of iterations with the aim to achieve the best results. Note that the silhouettes and RGB inputs were the same as those of the proposed method. The results of all methods over 14 views are shown in Table I. From the results, the proposed method achieved the best age and gender estimation accuracy, with an MAE 0.32 lower than that of the second-best method [19] and a CCR 0.29 % higher than that of the second-best ResNet-18. Considering OU-MVLP contains 130,968 test samples from about 5000 subjects and 14 views, our improvements on MAE and CCR are not minor, e.g., CCR improvement of 0.29% mean the number of samples to obtain correct gender estimation results increases by 379. Regarding the CSs at 1, 5, and 10-year error tolerances, the proposed method also achieved the highest percentages.

Table II shows the MAE and CCR for each view angle of the proposed method and two comparison benchmarks (i.e., Shehata's method [19] and ResNet-18, which achieved the second-best age and gender estimation results, respectively). From the results, all methods showed differences in the age and gender estimation for different views. Specifically, we found that the results for the seven front view angles (i.e., 0°, 15°, ..., 90°) were better than those for the seven back view angles (i.e., 180°, 195°, ..., 270°). The reason may be that some motion information, such as hand movements, is often occluded by the human body when viewed from the back. Additionally, the results from the near-side views (e.g., 75° and 90°) were better than those from the near-front view (e.g., 0° and 15°). This may be because the body thickness and gait stride are more visible from side views, which provides more information for age and gender estimation. Compared with the two appearance-based methods, the proposed model-based method obtains the smallest standard deviation under all

TABLE I
COMPARISON OF THE PROPOSED GAITAGE WITH STATE-OF-THE-ART METHODS ON OU-MVLP. MAE [YEAR], CSs [%], AND CCR [%] ARE THE RESULTS FOR ALL 14 VIEWS. "SF" INDICATES SINGLE FRAME INPUT. "*" INDICATES THAT WE SLIGHTLY CHANGED THE NETWORK TO FIT THE 3-CHANNEL INPUT. "N/A" MEANS NOT APPLICABLE. "-" INDICATES THAT DATA WERE NOT PROVIDED IN THE ORIGINAL PAPER. THE BEST RESULTS ARE IN BOLD.

| Methods	Inputs	Metrics				
		MAE↓	CS(1)↑	CS(5)↑	CS(10)↑	CCR↑
GEINet [5]	GEI	7.61	18.39	52.16	72.26	N/A
Upadhyay's [7]	GEI	N/A	N/A	N/A	N/A	95.33
TBResNet [10]	GEI	6.70	20.57	55.87	76.59	96.71
Xu's [39]	SF	8.39	15.84	48.00	68.40	94.27
Xu's [39]	Sil.	6.63	-	-	-	96.04
GaitSet	Sil.	6.55	20.06	56.55	77.15	95.80
GaitGL	Sil.	6.85	19.35	55.32	75.65	95.52
GaitSet*	RGB	5.66	24.52	61.00	81.27	96.60
GaitGL*	RGB	6.03	22.78	58.77	79.34	96.46
ResNet-18	RGB	5.84	21.28	60.92	80.53	97.58
ResNet-36	RGB	6.34	22.89	59.48	77.89	97.11
ResNet-50	RGB	6.77	17.07	55.17	79.34	96.78
Shehata's [19]	RGB	5.44	-	-	-	97.3
GaitAGE (Ours)	RGB	**5.12**	**25.24**	**63.50**	**84.32**	**97.87**

Fig. 8. MAE [year] of the proposed method at the 5-year interval.

Fig. 9. Scatter plots of the estimated age and ground-truth age for the proposed method at 0° and 90°.

14 view angles, indicating the superiority of our model-based method in reducing the viewpoint influence.

Additionally, we also presented some detailed results of the proposed method on different genders, age groups and scatter plots. Table III shows the results under the same gender for the proposed method. We found that the results for males were better than those for females. This is because females have a wide range of appearance variations (e.g., clothes, hairstyle, and shoes) than males, making the age and gender estimation more difficult. This is consistent with the insights in [28]. Fig. 8 and Fig. 9 show the MAE of the proposed method at 5-year intervals[2] and the scatter plots of the estimated age and ground-truth age for the proposed method at 0° and 90°, respectively. From the two figures, we found that for subjects under 15 years old, the proposed method achieved very good age estimation results, with the lowest MAE (1–2 years old), because the gait of children changes significantly as they grow. When subjects become adults, larger MAEs (5–10 years old) inevitably occur because the gait of adults changes slowly and has more variations. When subjects were older, the proposed method achieved the worst age estimation results, with the largest MAE over 10 years, because the gait of the elderly also varies greatly and depends largely on their health status. Healthy older people may have a very young gait. Another important reason is that the number of elderly subjects was very limited as shown in Fig. 6. For example, there were only two subjects in the age group [86, 90]. This inevitably leads to the trained model focusing on the more representative younger age groups rather than the older age groups.

FVG. Because there was no official age and gender informa-

[2]Note that five adults (with subject IDs "03368", "05336", "06578", "08186", and "08262") were incorrectly labeled as age group [1, 5], which greatly increased the MAE; hence, we simply removed them from this evaluation.
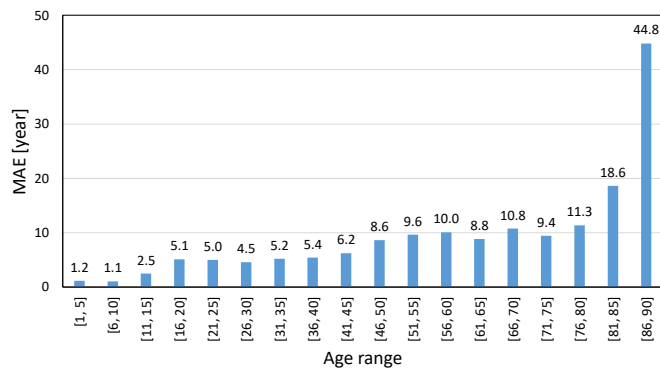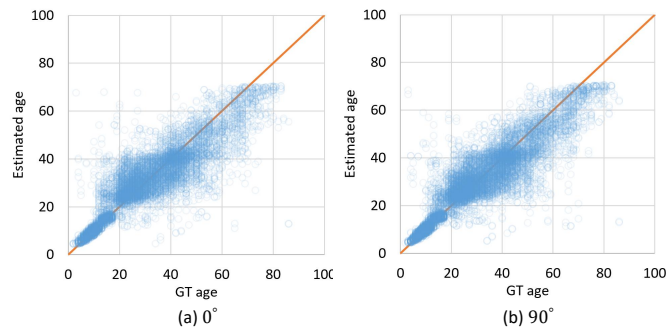
tion, we only evaluated the gender estimation performance on this dataset using our annotated gender information. To fit the gender-only estimation, we removed the age-related parts from the entire framework, for example, the regressed age weight α, infant-SMPL model in the model interpolation, and age label regression in the MLM module. Thus, we only report the results of the proposed method using the 3D vertices sequences trained from scratch. For comparison, we chose the method in [8] because it is also an SMPL-based gender estimation method. However, it only uses the shape feature of SMPL in conjunction with a traditional SVM classifier to estimate the gender of images. Because the protocol was not stated clearly in [8], we attempted to provide most of the setting information and show the comparison results in Table IV. Despite the different protocols, the proposed method achieved a very good gender estimation accuracy of over 95% CCR on the outdoor dataset.

F. Ablation study

We compared the performance of the proposed SMPL+AG model with the original SMPL model in an ablation experiment. For the two mesh models, we used several derived sub-features as input for the age and gender estimator, as mentioned in Section III-E. Particularly, 3D vertices + $\hat{\alpha}$ were only available for the proposed SMPL+AG model because $\hat{\alpha}$ was not in the original SMPL model. The comparison results are shown in Table V. From the results, we observed that for each sub-feature, the proposed SMPL+AG model consistently achieved a lower MAE and higher CCR than the

TABLE II
DETAILED MAE [YEAR] AND CCR [%] OF SHEHATA'S METHOD [19], RESNET-18, AND THE PROPOSED GAITAGE FOR EACH VIEW ANGLE ON OU-MVLP. THE BEST RESULTS ARE IN BOLD.

Methods	Metrics	0°	15°	30°	45°	60°	75°	90°	180°	195°	210°	225°	240°	255°	270°	Mean±std
Shehata's [19]	MAE↓	5.53	5.18	5	5.25	5.20	5.14	5.08	6.18	5.96	5.79	5.71	5.48	5.33	5.29	5.44±0.34
	CCR↑	97.3	**97.8**	97.7	97.7	97.4	97.8	97.7	95.4	97	97.2	97.1	97.3	97.7	97.6	97.3 ±0.60
ResNet-18	MAE↓	6.11	5.87	5.79	5.57	5.39	5.37	5.50	6.59	6.40	6.20	6.09	5.75	5.71	5.61	5.84±0.36
	CCR↑	96.83	97.28	97.70	97.77	98.07	97.68	97.65	96.92	97.71	97.66	97.53	**97.70**	97.69	97.82	97.58±0.33
GaitAGE (Ours)	MAE↓	**5.20**	**5.05**	**4.97**	**4.97**	**4.81**	**4.82**	**4.91**	**5.63**	**5.50**	**5.44**	**5.35**	**5.07**	**5.01**	**5.00**	**5.12±0.25**
	CCR↑	**97.68**	**97.65**	**98.05**	**98.11**	**98.19**	**98.28**	**98.03**	**97.44**	**97.94**	**97.82**	**97.66**	97.59	**97.74**	**97.87**	**97.87±0.24**

TABLE III
MAE [YEAR] AND CCR [%] UNDER THE SAME GENDER FOR THE PROPOSED METHOD ON OU-MVLP.

Gender	Metrics	
	MAE↓	CCR↑
Female	5.39	96.99
Male	4.84	98.79
All	5.12	97.87

TABLE IV
CCR [%] OF THE PROPOSED GAITAGE AND GIULIA'S METHOD [8] ON FVG. THE PROPOSED METHOD IS A SEQUENCE-BASED METHOD, WHEREAS GIULIA'S METHOD [8] IS AN IMAGE-BASED METHOD.

Methods	Feature	Train	Test	CCR↑
Giulia's [8]	shape	5650 images	1130 images	87.38
GaitAGE (Ours)	3D vert.	408 seq.	306 seq.	**95.10**

TABLE V
ABLATION EXPERIMENTS FOR THE PROPOSED SMPL+AG MODEL AND ORIGINAL SMPL MODEL ON OU-MVLP. MAE [YEAR] AND CCR [%] ARE THE RESULTS FOR ALL 14 VIEWS. THE BEST RESULTS ARE IN BOLD.

Mesh model	Sub-features	Metrics	
		MAE↓	CCR↑
SMPL	Shape	5.25	96.27
SMPL+AG	Shape	5.20	97.67
SMPL	Pose	5.27	96.44
SMPL+AG	Pose	5.19	96.64
SMPL	3D key joints	5.25	95.73
SMPL+AG	3D key joints	5.19	97.74
SMPL	3D vertices	5.18	96.10
SMPL+AG	3D vertices	5.14	97.86
SMPL+AG	3D vertices + $\hat{\alpha}$	**5.12**	**97.87**

original SMPL model, which demonstrated that the proposed SMPL+AG model captured more age- and gender-specific information. From the visualization in Fig. 7, the proposed SMPL+AG model also provided more reasonable mesh models for the respective ages and genders. In a comparison of various sub-features, 3D vertices + $\hat{\alpha}$ obtained the best results because it contained more information than the other sub-features. Although the results of the shape, pose sequence, and 3D key joint sequences were slightly worse, they were still good choices, considering the low-dimensional feature space. Compared with SMPL-based gait recognition tasks [15], [16], the difference in sub-features does not significantly affect the accuracy in age and gender estimation tasks.

V. DISCUSSION

A. Learned age and gender weights

We analyzed the learned age and gender weight of the proposed SMPL+AG model, since misestimated age and gender weights may result in unsuitable human models and further affect the subsequent estimations. Thus, we investigated the frequency of the misestimated age and gender weights.

As for the age weight, we visualized the mean and standard deviation of the learned age weight $\hat{\alpha}$ for each single ground-truth age for all test sequences in Fig. 10. As mentioned earlier, the age weights are designed as the linear interpolation weights between adult and infant-SMPL models. That is, when it is close to 0, the estimated model is more like an adult; when it is close to 1, the estimated model is more like an infant. From

the results, even without direct supervision on the age weights during training, we achieved the expected curves thanks to the proposed relative relation loss L_α. More specifically, the estimated $\hat{\alpha}$ gradually decreased before 18 years old, which indicates the reduction of interpolation weight of the infant-SMPL model. This caused the interpolated SMPL model to change from an infant-like model to an adult-like model, which is consistent with the normal changes in physical characteristics (e.g., height, head-to-body ratio) during the growth of children. When the child became an adult (older than 18 years old), the estimated $\hat{\alpha}$ was less than 0.1, which indicates that the infant-SMPL model had very little weight, and thus the interpolation relied more on the adult-SMPL model. The estimated $\hat{\alpha}$ also remained almost constant between the ages of 18 and 60, because the physical characteristics of adults are relatively stable over the years. Finally, the estimated $\hat{\alpha}$ increased slightly as adults got older (over 60 years old). This is because the elderly may have arched backs and knees (i.e., become a little shorter in height and consequently slightly larger head-to-body ratio), resulting in a slight increase in age weights. As for the gender weight, we counted the proportion of correctly estimated samples to the total test sequences and found that nearly 98% of the sequences have the correct gender weights.

Therefore, we think that although the misestimated age and gender weights would affect the subsequent age and gender estimation, it rarely happens. In most cases, the proposed method could obtain accurate age and gender weights. The ablation study in subsection 4.6 also verifies the clear improvement of the proposed method by introducing the age and gender
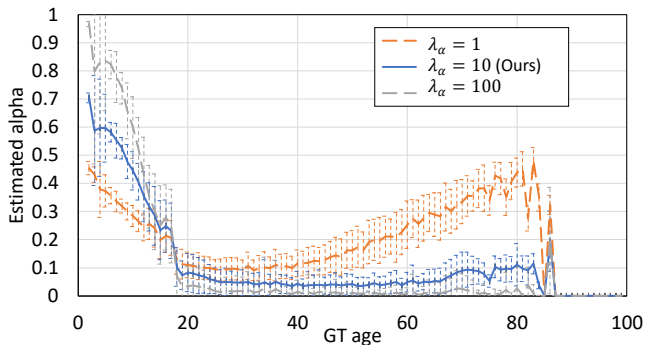
Fig. 10. Mean and standard deviation of learned age weight $\hat{\alpha}$ for each single ground-truth age on all test sequences of OU-MVLP. Blue solid line is the proposed method with $\lambda_\alpha = 10$; orange and gray dotted lines are with $\lambda_\alpha = 1$ and $\lambda_\alpha = 100$, respectively. Best viewed in color.

weights to SMPL model.

B. Sensitivity analysis of hyperparameters

In Table VI, we analyzed the sensitivity of the experimentally set hyperparameters (i.e., weight hyperparameter λ_α and margin m of the age weight relative relation loss, and weight hyperparameter λ_{ag} of the age and gender loss) of the proposed method for all test sequences on OU-MVLP. Specifically, when varying one hyperparameter across a range of values, we kept the other hyperparameters constant. The varied ranges of values were set to $[1, 10, 100]$, $[0.3, 0.4, 0.5]$, $[0.6, 0.8, 1]$ for λ_α, m, and λ_{ag}, respectively. To evaluate the performance of model reconstruction, the reconstruction error, calculated as the mean squared error between the GT silhouettes and rendered silhouettes of all test sequences, is also reported. Regarding different λ_α, we also showed the mean and standard deviation of the learned age weight $\hat{\alpha}$ for each single ground-truth age in Fig. 10. From the results, the smaller λ_α (i.e., 1) didn't meet our expectations mentioned above. While the larger λ_α (i.e., 100) led to poor reconstruction performance and age and gender estimation accuracies. Thus, the moderate λ_α (i.e., 10) was chosen for the proposed method. Regarding m, as it is relatively insensitive to age and gender estimation accuracies, we set it to 0.4 considering the best reconstruction performance. Similarly, we also chose the moderate λ_{ag} (i.e., 0.8) considering the balance between reconstruction performance and age and gender estimation accuracies.

C. Failure cases

In Fig. 11, we showed two failure examples with large age estimation errors. The first young male (17 years old) was overestimated to be approximately 32 years old. This might be because he sometimes had a hunched posture when he walked, which is more likely to occur in older people. The estimated age distribution also had two peaks at approximately 15 and 40 years old, which indicates less confidence in the expected age. The second adult female (46 years old) was underestimated to be approximately 22 years old. This might be because she had a slim body shape, which is more likely to occur in younger

TABLE VI
HYPERPARAMETER SENSITIVITY ANALYSIS FOR THE PROPOSED METHOD ON OU-MVLP. MAE [YEAR] AND CCR [%] ARE THE RESULTS FOR ALL 14 VIEWS.

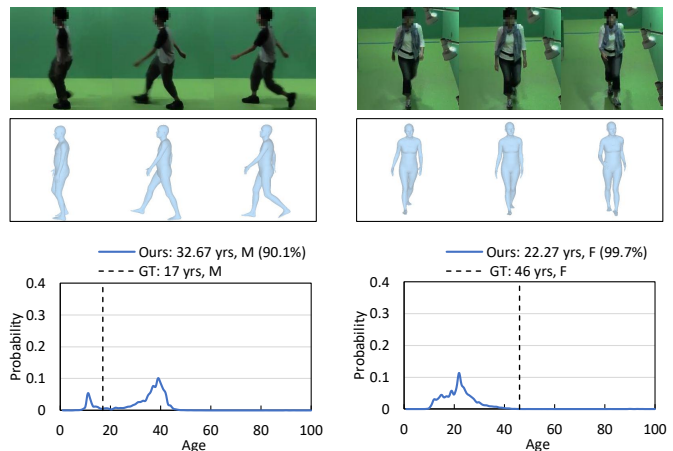| Hyperparameters		Metrics		Reconstruction error↓
		MAE↓	CCR↑	
λ_α	1	5.09	98.07	0.0628
	10	5.12	97.87	0.0597
	100	5.16	97.27	0.0606
m	0.3	5.09	98.00	0.0617
	0.4	5.12	97.87	0.0597
	0.5	5.13	98.01	0.0607
λ_{ag}	0.6	5.09	96.61	0.0592
	0.8	5.12	97.87	0.0597
	1	5.11	97.99	0.0601

Fig. 11. Two examples (left) and (right) with large age estimation errors. The rows from the top to bottom are the input sequences, reconstructed mesh models of the proposed method, and estimated age and gender results.

people. The estimated age distribution was flatter, with a larger standard deviation, which also reduced confidence in the expected age. Therefore, the estimated age distribution was more meaningful than a single estimated age, and provided the confidence level of the estimation results. Additionally, the two examples also hinted at a gap between the chronological age and physiological (gait) age [28]. Older subjects may have been underestimated because of their younger gait pattern (e.g., slim shape and large stride) and vice versa.

We further compared the failure cases between the proposed method and ResNet-18, which are shown in Fig. 12. The first row shows the examples where the proposed method succeeds and ResNet-18 fails, while the second row shows the examples where ResNet-18 succeeds and the proposed method fails. The results show that ResNet-18 directly uses the input RGB information and pays more attention to clothing styles, thus it may overestimate the age of young subjects wearing mature clothing (e.g., dresses or suits) and underestimate the age of older subjects wearing fashion clothing for young people. In contrast, the proposed method is based on the human model and pays more attention on the body shape and motion, thus it may overestimate the age of young subjects with fatter or taller body shapes and underestimate the age of older subjects
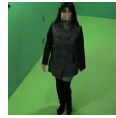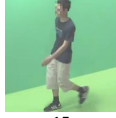
This article has been accepted for publication in IEEE Transactions on Biometrics, Behavior, and Identity Science. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TBIOM.2024.3424841

JOURNAL OF LATEX CLASS FILES, VOL. 14, NO. 8, AUGUST 2015
12

Fig. 12. A comparison of failure cases between the proposed method and ResNet-18. The numbers under each image are the ground truth ages, the estimated ages of the proposed method and ResNet-18. The red numbers mean the estimated ages that are close to the ground truth ages.

with younger gait patterns.

D. Limitations

Since it is often said that model-based methods require relatively high-resolution human images to achieve good model fitting, the proposed method may inevitably suffer from poor data quality (e.g., low resolution human size) in real scenes. Thus, we further analyzed the effect of poor-quality data. More specifically, we first simulated the poor data quality on the test set by reducing the image resolution by a factor of 4 from 224×224 to 56×56, then used the low-resolution data for testing. ResNet-18 was chosen for comparison. To fit the training model, the low-resolution test data were firstly unsampled to 224×224 before inference. The results show that the MAE and CCR of the proposed method dropped from 5.12 to 13.49 and 97.87% to 94.23% respectively, while the MAE and CCR of ResNet-18 dropped from 5.84 to 13.08 and 97.58% to 91.78% respectively. Not only the proposed model-based method but also the appearance-based method directly using RGB suffer significant degradation, which means that the impact of data deficiencies (e.g., poor data quality) is common to both model-based and appearance-based methods. The findings also motive us to improve the robustness against poor-quality data in future work.

E. Inference time

We computed the inference time per sequence of the proposed model-based method and ResNet-18. It turns out that our method takes more time than ResNet-18 (i.e., 53.1 ms vs. 8.1 ms), but it can still run near real-time (i.e., about 19 fps). We think it is acceptable considering advantages against appearance-based methods such as higher accuracy and view invariance.

VI. Conclusion

In this paper, we introduced a gait-based age and gender estimation approach based on reconstructed human meshes. To describe realistic age- and gender-specific mesh models, we proposed the SMPL+AG model, which uses two additional age and gender weights to interpolate among male, female, and infant SMPL models. For the age weight, we proposed a data-driven learning scheme to estimate it by considering the paired relation between ground-truth ages and age weights. We then explored the effectiveness of various sub-features derived from the SMPL+AG model, and designed corresponding networks for age and gender estimation. The experimental results demonstrated that the proposed method achieved both good mesh estimation and state-of-the-art age and gender estimation results. Considering real-world applications, we could improve two aspects of the approach in future work. First, we could improve the quality and stability of mesh reconstruction by introducing more training data under various capture conditions. Second, we could improve age and gender estimation accuracy by introducing other effective networks or other complementary gait-unrelated information, such as hairstyle, clothing type, and shoes.

References

[1] C. B. Ng, Y. H. Tay, and B.-M. Goi, "Recognizing human gender in computer vision: A survey," in *PRICAI 2012: Trends in Artificial Intelligence*, P. Anthony, M. Ishizuka, and D. Lukose, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 335–346.

[2] N. Mansouri, "Automatic age estimation: A survey," *Computación y Sistemas*, vol. 24, 2020.

[3] T. B. Aderinola, T. Connie, T. S. Ong, W.-C. Yau, and A. B. J. Teoh, "Learning age from gait: A survey," *IEEE Access*, vol. 9, pp. 100 352–100 368, 2021.

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

[5] A. Sakata, Y. Makihara, N. Takemura, D. Muramatsu, and Y. Yagi, "How confident are you in your estimate of a human age? uncertainty-aware gait-based age estimation by label distribution learning," in *IEEE International Joint Conference on Biometrics*, Sep. 2020.

[6] S. Zhang, Y. Wang, and A. Li, "Gait-based age estimation with deep convolutional neural network," in *2019 International Conference on Biometrics (ICB)*, 2019, pp. 1–8.

[7] J. Upadhyay and T. Gonsalves, "Robust and lightweight system for gait-based gender classification toward viewing angle variations," *AI*, vol. 3, no. 2, pp. 538–553, 2022.

[8] G. Martinelli, N. Garau, and N. Conci, "Gender recognition from 3d shape parameters," in *Image Analysis and Processing. ICIAP 2022 Workshops*, P. L. Mazzeo, E. Frontoni, S. Sclaroff, and C. Distante, Eds. Cham: Springer International Publishing, 2022, pp. 203–214.

[9] A. Sakata, N. Takemura, and Y. Yagi, "Gait-based age estimation using multi-stage convolutional neural network," *IPSJ Transactions on Computer Vision and Applications*, vol. 11, no. 4, pp. 1–10, 2019.

[10] S. Zhang, Y. Wang, and A. Li, "Gait energy image-based human attribute recognition using two-branch deep convolutional neural network," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, pp. 1–1, 2022.

[11] L. K. Lau and K. Chan, "Tree structure convolutional neural networks for gait-based gender and age classification," *Multimedia Tools and Applications*, vol. 82, pp. 2145–2164, 2023.

[12] J. Han and B. Bhanu, "Individual recognition using gait energy image," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 2, pp. 316– 322, 2006.

[13] A. Parashar, A. Parashar, W. Ding, R. S. Shekhawat, and I. Rida, "Deep learning pipelines for recognition of gait biometrics with covariates: a comprehensive review," *Artificial Intelligence Review*, vol. 56, pp. 8889–8953, 2023.

[14] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A skinned multi-person linear model," *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, vol. 34, no. 6, pp. 248:1–248:16, Oct. 2015.

[15] X. Li, Y. Makihara, C. Xu, Y. Yagi, S. Yu, and M. Ren, "End-to-end model-based gait recognition," in *Proceedings of the Asian Conference on Computer Vision (ACCV)*, November 2020.

[16] X. Li, Y. Makihara, C. Xu, and Y. Yagi, "End-to-end model-based gait recognition using synchronized multi-view pose constraint," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, October 2021, pp. 4106–4115.

[17] J. Liang and M. Lin, "Shape-aware human pose and shape reconstruction using multi-view images," in *ICCV*, 08 2019.

[18] M. Kocabas, N. Athanasiou, and M. J. Black, "Vibe: Video inference for human body pose and shape estimation," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[19] A. Shehata, A. Alsherfawi, L. Gäher, X. Li, Y. Makihara, and Y. Yagi, "Online model-based gait age and gender estimation," in *Proc. of the Int. Joint Conf. on Biometrics (IJCB2023)*, Ljubljana, Slovenia, Sept. 2023, pp. 761–770.

[20] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. Osman, D. Tzionas, and M. J. Black, "Expressive body capture: 3d hands, face, and body from a single image," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 10 975–10 985.

[21] P. Patel, C.-H. P. Huang, J. Tesch, D. T. Hoffmann, S. Tripathi, and M. J. Black, "Agora: Avatars in geography optimized for regression analysis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 13 468–13 478.

[22] N. Hesse, S. Pujades, J. Romero, M. J. Black, C. Bodensteiner, M. Arens, U. G. Hofmann, U. Tacke, M. Hadders-Algra, R. Weinberger *et al.*, "Learning an infant body model from rgb-d data for accurate full body motion analysis," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 792–800.

[23] Y. Sun, W. Liu, Q. Bao, Y. Fu, T. Mei, and M. J. Black, "Putting people in their place: Monocular regression of 3d people in depth," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 13 243–13 252.

[24] 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. on Computer Vision and Applications*, vol. 10, no. 4, pp. 1–14, 2018.

[25] Z. Zhang, L. Tran, X. Yin, Y. Atoum, J. Wan, N. Wang, and X. Liu, "Gait recognition via disentangled representation learning," in *CVPR*, Long Beach, CA, June 2019.

[26] H. Mannami, Y. Makihara, and Y. Yagi, "Gait analysis of gender and age using a large-scale multi-view gait database," in *Proc. of the 10th Asian Conf. on Computer Vision*, Queenstown, New Zealand, Nov. 2010, pp. 975–986.

[27] B. K. Y. Chuen, T. Connie, O. T. Song, and M. Goh, "A preliminary study of gait-based age estimation techniques," in *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, Dec 2015, pp. 800–806.

[28] C. Xu, Y. Makihara, G. Ogi, X. Li, Y. Yagi, and J. Lu, "The ou-isir gait database comprising the large population dataset with age and performance evaluation of age estimation," *IPSJ Transactions on Computer Vision and Applications*, vol. 9, no. 1, p. 24, Dec 2017. [Online]. Available: https://doi.org/10.1186/s41074-017-0035-2

[29] J. Lu and Y. P. Tan, "Ordinary preserving manifold analysis for human age and head pose estimation," *IEEE Transactions on Human-Machine Systems*, vol. 43, no. 2, pp. 249–258, March 2013.

[30] X. Li, Y. Makihara, C. Xu, Y. Yagi, and M. Ren, "Gait-based human age estimation using age group-dependent manifold learning and regression," *Multimedia Tools Appl.*, vol. 77, no. 21, p. 28333–28354, Nov. 2018.

[31] Y. Makihara, M. Okumura, H. Iwama, and Y. Yagi, "Gait-based age estimation using a whole-generation gait database," in *Proc. of the Int. Joint Conf. on Biometrics (IJCB2011)*, Washington D.C., USA, Oct. 2011, pp. 1–6.

[32] A. Sakata, Y. Makihara, N. Takemura, D. Muramatsu, and Y. Yagi, "Gait-based age estimation using a densenet," in *Asian Conference on Computer Vision*. Springer, 2018, pp. 55–63.

[33] G. Huang and Y. Wang, "Gender classification based on fusion of multi-view gait sequences," in *Proc. of the 8th Asian Conf. on Computer Vision*, vol. 1, Nov. 2007, pp. 462–471.

[34] X. Li, S. Maybank, S. Yan, D. Tao, and D. Xu, "Gait components and their application to gender recognition," *Trans. on Systems, Man, and Cybernetics, Part C*, vol. 38, no. 2, pp. 145–155, Mar. 2008.

[35] S. Yu, T. Tan, K. Huang, K. Jia, and X. Wu, "A study on gait-based gender classification," *IEEE Trans. on Image Processing*, vol. 18, no. 8, pp. 1905–1910, Aug. 2009.

[36] J. Lu, G. Wang, and P. Moulin, "Human identity and gender recognition from gait sequences with arbitrary walking directions," *IEEE Transactions on information Forensics and Security*, vol. 9, no. 1, pp. 51–61, 2013.

[37] M. J. Marín-Jiménez, F. M. Castro, N. Guil, F. de la Torre, and R. Medina-Carnicer, "Deep multi-task learning for gait-based biometrics," in *2017 IEEE International Conference on Image Processing (ICIP)*, 2017, pp. 106–110.

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

[39] C. Xu, Y. Makihara, R. Liao, H. Niitsuma, X. Li, Y. Yagi, and J. Lu, "Real-time gait-based age estimation and gender classification from a single image," in *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2021, pp. 3459–3469.

[40] 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 – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 561–578.

[41] G. Pavlakos, L. Zhu, X. Zhou, and K. Daniilidis, "Learning to estimate 3d human pose and shape from a single color image," in *CVPR*, June 2018.

[42] S. Shin and E. Halilaj, "Multi-view human pose and shape estimation using learnable volumetric aggregation," in *arXiv preprint arXiv:2011.13427*, 11 2020.

[43] X. Li, Y. Makihara, C. Xu, and Y. Yagi, "Multi-view large population gait database with human meshes and its performance evaluation," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 4, no. 2, pp. 234–248, 2022.

[44] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *ECCV*, 2016.

[45] 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 *EMNLP*, 2014.

[46] H. Kato, Y. Ushiku, and T. Harada, "Neural 3d mesh renderer," in *CVPR*, 2018.

[47] N. Mahmood, N. Ghorbani, N. F. Troje, G. Pons-Moll, and M. J. Black, "Amass: Archive of motion capture as surface shapes," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[48] Y. Hayashi, A. Shehata, Y. Makihara, D. Muramatsu, and Y. Yagi, "Deep gait relative attribute using a signed quadratic contrastive loss," in *2020 25th International Conference on Pattern Recognition (ICPR)*, 2021, pp. 8484–8491.

[49] W. An, S. Yu, Y. Makihara, X. Wu, C. Xu, Y. Yu, R. Liao, and Y. Yagi, "Performance evaluation of model-based gait on multi-view very large population database with pose sequences," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 2, no. 4, pp. 421–430, 2020.

[50] J. Lin, "Divergence measures based on the shannon entropy," *IEEE Transactions on Information Theory*, vol. 37, no. 1, pp. 145–151, 1991.

[51] A. Kirillov, Y. Wu, K. He, and R. Girshick, "Pointrend: Image segmentation as rendering," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9799–9808.

[52] H. Chao, Y. He, J. Zhang, and J. Feng, "Gaitset: Regarding gait as a set for cross-view gait recognition," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 8126–8133.

[53] B. Lin, S. Zhang, and X. Yu, "Gait recognition via effective global-local feature representation and local temporal aggregation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 14 648–14 656.

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, a Specially Appointed Researcher (part-time) from 2017 to 2020, and a Specially Appointed Researcher (full-time) from 2021 to 2022 with SANKEN, Osaka University, Japan, where he is currently Specially Appointed Assistant Professor. 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, ECCV 2020, ICCV 2021, CVPR 2023, and ICCV 2023. He is a member of IPSJ, IEICE, RSJ, and JSME.

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.

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, a Specially Appointed Researcher (part-time) from 2017 to 2020, and a Specially Appointed Researcher (full-time) from 2021 to 2022 with SANKEN, Osaka University, Japan, where she is currently a Specially Appointed Assistant Professor. Her research interests are gait recognition, machine learning, and image processing.