End-to-end Model-based Gait Recognition

Xiang Li[1,2], Yasushi Makihara[2], Chi Xu[1,2], Yasushi Yagi[2],
Shiqi Yu[3], and Mingwu Ren[1]

[1] Nanjing University of Science and Technology, Nanjing, China
{lixiangmzlx, xuchisherry}@gmail.com, renmingwu@mail.njust.edu.cn
[2] Osaka University, Osaka, Japan
{makihara, yagi}@am.sanken.osaka-u.ac.jp
[3] Southern University of Science and Technology, Shenzhen, China
yusq@sustech.edu.cn

Abstract. Most existing gait recognition approaches adopt a two-step procedure: a preprocessing step to extract silhouettes or skeletons followed by recognition. In this paper, we propose an end-to-end model-based gait recognition method. Specifically, we employ a skinned multi-person linear (SMPL) model for human modeling, and estimate its parameters using a pre-trained human mesh recovery (HMR) network. As the pre-trained HMR is not recognition-oriented, we fine-tune it in an end-to-end gait recognition framework. To cope with differences between gait datasets and those used for pre-training the HMR, we introduce a reconstruction loss between the silhouette masks in the gait datasets and the rendered silhouettes from the estimated SMPL model produced by a differentiable renderer. This enables us to adapt the HMR to the gait dataset without supervision using the ground-truth joint locations. Experimental results with the OU-MVLP and CASIA-B datasets demonstrate the state-of-the-art performance of the proposed method for both gait identification and verification scenarios, a direct consequence of the explicitly disentangled pose and shape features produced by the proposed end-to-end model-based framework.

1 Introduction

A video of a walking person contains abundant information about his/her gait pose sequence (i.e., dynamic parameters) and his/her body shape (i.e., static parameters). Gait recognition aims to identify a walking person using such dynamic and static parameters, and has a number of merits: identification at a distance; identification without subject cooperation; and difficulty in spoofing. These merits make gait recognition one of the most important solutions for person identification from CCTV footage, which has a wide range of applications in surveillance systems, forensics, and criminal investigation [1–3].

Gait recognition approaches are generally either model-based [4–12] or appearance-based [13–31]. The former mainly relies on the parameters of an articulated human model (e.g., the size of a link and a joint angle sequence), which are less sensitive to apparent changes caused by covariates (e.g., view angles,

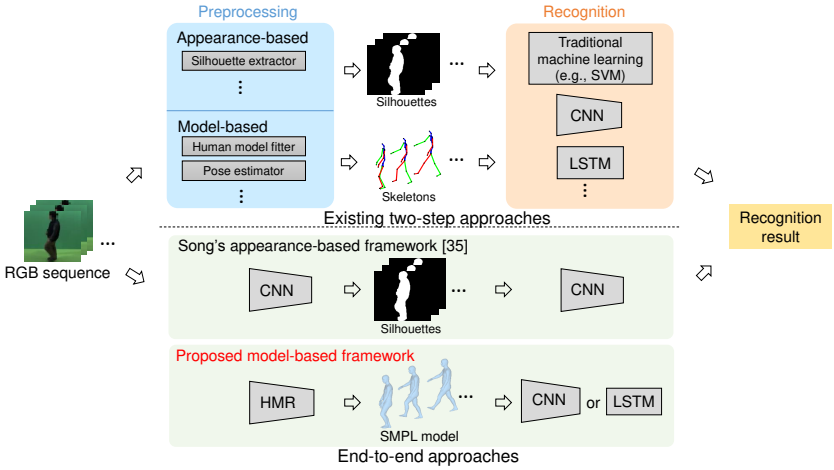

Fig. 1. While existing two-step approaches require a preprocessing step to first extract skeletons or silhouettes from RGB gait sequences before using a recognition network, the proposed method is an end-to-end model-based framework that estimates an SMPL model using a pre-trained HMR network, and subsequently uses the explicit parameterized pose and shape features of the SMPL model for recognition.

carrying status, and clothing), but are generally hard to extract accurately, particularly from low-resolution images. The latter approach mainly relies on human silhouette-based representations (e.g., gait energy images (GEIs) [13], frequency-domain features [25], and chrono-gait images [32]), which are easy to extract, even from low-resolution images, but are relatively sensitive to the aforementioned covariates.

Both approaches typically apply a preprocessing step to extract the pose structure or silhouettes from raw RGB videos with some additional algorithms, as shown in Fig. 1. For example, human body pose estimation methods (e.g., OpenPose [33]) are used to extract the joint locations; traditional background subtraction or deep learning-based semantic segmentation methods (e.g., RefineNet [34]) are used to segment the silhouettes. The preprocessing step is, however, not optimally designed for the latter recognition step. Hence, two-step approaches are not necessarily optimal in terms of recognition.

Several studies have proposed end-to-end appearance-based gait recognition frameworks [35, 36] that make the whole process optimal for recognition tasks, as they directly output the recognition results from the input RGB image sequences. For example, Song et al. [35] proposed a framework combining two convolutional neural networks (CNNs) that handle silhouette segmentation and gait recognition, respectively. Zhang et al. [36] proposed an end-to-end model that directly extracts latent pose and appearance features from masked RGB image sequences via disentangled representation learning (DRL), and obtains recognition results by subsequently feeding the latent pose feature sequence (i.e., motion informa-

tion) to a long short-term memory (LSTM) framework. These methods [35, 36] are appearance-based approaches, and hence still suffer from the aforementioned sensitivity to covariates (e.g., view).

By contrast, the bottleneck of model-based approaches (e.g., difficulty of human model fitting) is being resolved by recent advances in deep learning-based human model fitting [33, 37, 38]. For example, Kanazawa et al. [37] proposed a human mesh recovery (HMR) network that could directly estimate a parametric 3D human model (i.e., a skinned multi-person linear (SMPL) model [39]) to describe the 3D human body and shape from a single RGB image. In this paper, we therefore describe the integration of a human model estimation method [37] with gait recognition, and propose the first end-to-end model-based gait recognition approach.

The contributions of this study can be summarized as follows.

(1) End-to-end model-based gait recognition.

We propose an end-to-end model-based gait recognition method for the first time. More specifically, given an RGB gait sequence, we first extract pose and shape features by fitting the SMPL model and subsequently feed the pose and shape features to a recognition network. The whole network is then trained in an end-to-end manner. Unlike the existing DRL method [36], which provides implicit (or not physically interpretable) pose and appearance features, our method produces explicit (or physically interpretable) pose and shape features as a direct consequence of the model-based approach, which can be regarded as a kind of explicit DRL method for gait recognition.

(2) A human model fine-tuning scheme using a differentiable renderer without pose supervision.

We fine-tune a pre-trained HMR network [37] to make it optimal for gait recognition tasks as well as being adapted to different subject populations in a target gait dataset (i.e., transfer learning). Because gait datasets do not usually provide ground-truth poses (i.e., 2D joint locations), which are needed to train the HMR network [37], we instead use silhouette masks of the target dataset. More specifically, we fine-tune the HMR network using the reconstruction loss between the silhouette masks and rendered silhouettes from an estimated 3D human model with a differentiable renderer, in addition to other losses (e.g., recognition loss).

(3) State-of-the-art performance using our model-based approach.

We achieve state-of-the-art performance on two gait datasets, namely OU-MVLP [20] and CASIA-B [40], demonstrating the robustness of the proposed method against various covariates (e.g., views, clothes, carried objects). Because appearance-based approaches are believed to be better than model-based approaches in the gait recognition research community (indeed, model-based approaches have not outperformed appearance-based approaches for the last decade, to the best of our knowledge), our work is sure to generate some excitement by demonstrating that a model-based approach can outperform state-of-the-art appearance-based approaches.

2 Related Work

2.1 Gait Recognition

Model-based approaches mainly rely on the recognition of the human pose structure and movement. Early traditional approaches [4–9] suffered from a high computational cost and inaccurate human model fitting to a low-resolution image, and hence achieved relatively poor recognition performance. Thanks to recent advances in human pose estimation methods (e.g., Flowing ConvNets [41] and OpenPose [33]), human model estimation is no longer the bottleneck, resulting in an increasing number of model-based approaches [10–12]. These usually perform two-step recognition, in which the pose information (e.g., joint heat maps or exact joint locations) is first estimated, and then LSTM or CNN is applied to further extract discriminant features for recognition. These techniques have enabled great improvements compared with traditional model-based approaches.

Appearance-based approaches mainly rely on the recognition of human silhouettes. Unfortunately, the silhouette-based representations are easily affected by many covariates (e.g., view angles, carrying status, and clothing), which cause larger intra-subject differences than the subtle inter-subject differences. The many attempts to mitigate these negative effects of covariates can generally be separated into discriminative and generative approaches. Discriminative approaches [13–23, 42] mainly focus on extracting discriminative features or subspaces that are invariant to the covariates, whereas generative approaches [24–31] specialize in generating gait features under the same covariate conditions. In particular, the introduction of deep learning techniques has significantly improved the recognition accuracy. For example, some CNN-based discriminative approaches [19–23] apply a Siamese/triplet network to learn similarity metrics for a pair/triplet of input GEIs/silhouettes. Some generative adversarial network-based approaches [30, 31] generate a gait template from other covariate conditions to produce a canonical condition with no covariates (e.g., no carried object) or a target covariate condition (e.g., side view).

Both families, however, require a preprocessing step to extract the joint locations or silhouette-based representations given RGB gait sequences. Although [35] proposed an end-to-end framework for silhouette segmentation and gait recognition, it relies on the extracted silhouettes (i.e., appearance-based representation), and hence still suffers from sensitivity to covariates (e.g., view).

2.2 Disentangled Representation Learning

DRL is generally designed to separate the input data into disjoint meaningful variables, and has been widely studied in many areas (e.g., pose-invariant face recognition [43], shape-guided human image generation [44]). With regard to gait, there have been relatively few studies. Zhang et al. [36] employed an autoencoder-based CNN to directly extract latent pose and appearance features from RGB imagery, and then applied LSTM-based pose features for recognition. Li et al. [45] proposed a disentanglement method for the identity and covariate

features for GEIs. They first used an encoder to extract the latent identity and covariate features, and then used a two-stream decoder to reconstruct GEIs with and without covariates.

The latent features used in [36, 45] are not, however, physically interpretable because of the implicit disentanglement process. Moreover, the features may not be perfectly disentangled when certain assumptions (e.g., appearance consistency within a sequence) are violated. In contrast, our method realizes explicit disentanglement into the physically interpretable 3D joints and shape features of an SMPL model.

2.3 3D Human Pose and Shape Estimation

Most approaches formulate this problem as the estimation of parameters for a parametric 3D human model (i.e., SMPL [39]). For example, Pavlakos et al. [38] used a ConvNet to predict 2D pose heat maps and silhouettes from an RGB image, and then used these to infer the pose and shape parameters of the SMPL through two individual networks. Kanazawa et al. [37] designed an end-to-end HMR network that directly estimates the pose and shape parameters of human meshes from a single RGB image. The HMR network is trained in an adversarial manner to determine whether the estimated parameters are real or not using a large database of 3D human meshes.

As the estimated SMPL model accurately captures the statistics of human 3D pose and shape, it could be used for numerous tasks (e.g., motion imitation or appearance transfer in human image synthesis [46]). In this paper, we explore the usage of the SMPL model for gait recognition. We mainly rely on the existing HMR network [37] to estimate the pose and shape parameters.

3 End-to-end Model-based Gait Recognition

3.1 Overview

An overview of the proposed network is shown in Fig. 2. Given a period of size-normalized and registered RGB gait sequence, the SMPL model containing the pose and shape features is first estimated using a pre-trained HMR network [37]. The estimated shape features are further aggregated by an average pooling layer to produce a common shape for all frames. Thereafter, we render 2D binary silhouettes with a differentiable renderer from the 3D human mesh generated by the common shape, frame-by-frame pose, and camera parameters. The HMR network is then fine-tuned without pose supervision, instead using the reconstruction loss between the silhouette masks in the gait datasets and the rendered silhouettes. In this way, the estimated SMPL model becomes more identity-preserving. Finally, the shape feature and the latent pose features filtered by LSTM/CNN are used for recognition. The whole network is trained in an end-to-end manner with multiple losses, which are responsible for the identity-preserving SMPL model estimation and gait recognition.

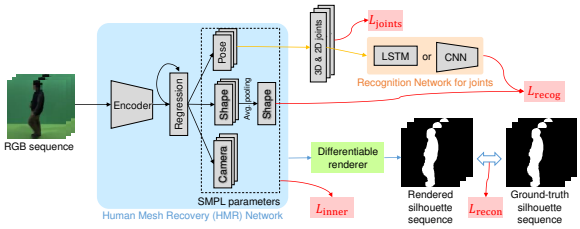

Fig. 2. Overview of the proposed network. For each input RGB sequence, an HMR [37] network first extracts the SMPL parameters, which are further rendered to 2D binary silhouettes for comparison with the silhouette masks in the gait datasets. The shape and pose features of the SMPL are then used for recognition.

3.2 HMR Network [37]

Given an input RGB gait sequence, the HMR network estimates a 3D human body mesh derived from the SMPL model [39], which factors the human body into shape $\boldsymbol{\beta}$ and pose $\boldsymbol{\theta}$ components. The shape $\boldsymbol{\beta} \in \mathbb{R}^{10}$ is defined in a low-dimensional principal component analysis shape space, which describes the height, weight, and body proportions of individuals. The pose $\boldsymbol{\theta} \in \mathbb{R}^{72}$ is the combination of the relative 3D rotations of 23 joints in an axis–angle representation (23×3) with respect to its parent in the kinematic tree and the 3D root orientation, which describes the joint locations of individuals. The SMPL model outputs a triangulated mesh with 6,890 vertices, and is differentiable with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$. Additionally, the weak-perspective camera parameter $\boldsymbol{k} \in \mathbb{R}^3$ is also estimated. As a result, the HMR network first encodes the input image into a 2,048-dimensional feature using ResNet-50 [47], and then predicts the 85-dimensional vector $\boldsymbol{\Theta} = [\boldsymbol{\beta}^T, \boldsymbol{\theta}^T, \boldsymbol{k}^T]^T$ using an iterative 3D regression network.

The HMR network was trained using six datasets with different properties: four datasets [48–51] with 2D joint information (i.e., pose and body part length) from over 4,000 subjects and two datasets [52, 53] with full 3D information, including body part thickness, but from a limited number of subjects (i.e., just five and eight subjects, respectively). Therefore, the subject diversity is good for the pose and the body part length, while it is limited for the body part thickness. Consequently, the estimation accuracy for the body part thickness is not as high as that for the pose and body part length (e.g., joint locations).

3.3 Shape Consistency and Pose/Camera Smoothness

Because the original HMR network is designed for a single image, it ignores the shape consistency and the pose and camera continuities within a sequence. We therefore modify the HMR network by introducing the inner loss L_{inner} of the estimated parameters $\boldsymbol{\Theta}$ (i.e., combination of shape consistency and pose/camera smoothness) as additional constraints, as described below.

Suppose that N gait sequences $\{S_i\}(i = 1, \ldots, N)$ are given and the i-th sequence is composed of T_i frames. The HMR network first estimates the pa-

rameters $\boldsymbol{\Theta}_i^j = [{\boldsymbol{\beta}_i^j}^T, {\boldsymbol{\theta}_i^j}^T, {\boldsymbol{k}_i^j}^T]^T$ for the j-th frame of the i-th sequence frame-by-frame. To make the shape $\boldsymbol{\beta}_i^j$ consistent within the input sequence S_i, we aggregate each frame in an average pooling layer to produce a common shape $\bar{\boldsymbol{\beta}}_i = \frac{1}{T_i}\sum_{j=1}^{T_i}\boldsymbol{\beta}_i^j$. This common shape $\bar{\boldsymbol{\beta}}_i$ then replaces each frame's shape $\boldsymbol{\beta}_i^j$ in the final estimated SMPL model.

Furthermore, suppose that N gait sequences are composed of C subjects, and denote the set of sequence indices for the c-th subject as $\mathcal{S}_c$. We then introduce the shape-based inner loss L_{shape} to make the common shapes $\bar{\boldsymbol{\beta}}_i$ within the same subject closer to each other:

$$L_{\text{shape}} = \frac{1}{C}\sum_{c=1}^{C}\frac{1}{|\mathcal{S}_c|(|\mathcal{S}_c|-1)}\sum_{i\in\mathcal{S}_c}\sum_{l\in\mathcal{S}_c-\{i\}}\|\bar{\boldsymbol{\beta}}_i - \bar{\boldsymbol{\beta}}_l\|_2^2. \tag{1}$$

As for the pose $\boldsymbol{\theta}_i^j$ and camera $\boldsymbol{k}_i^j$, they are considered to change smoothly within a gait sequence S_i. We therefore introduce the pose-based inner loss L_{pose} and camera-based inner loss L_{cam} as

$$\begin{aligned}
L_{\text{pose}} &= \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{T_i-1}\sum_{j=1}^{T_i-1}\|\boldsymbol{\theta}_i^{j+1}-\boldsymbol{\theta}_i^j\|_2^2 + \frac{1}{T_i-2}\sum_{j=2}^{T_i-1}\|\boldsymbol{\theta}_i^{j+1}-2\boldsymbol{\theta}_i^j+\boldsymbol{\theta}_i^{j-1}\|_2^2\right),\\
L_{\text{cam}} &= \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{T_i-1}\sum_{j=1}^{T_i-1}\|\boldsymbol{k}_i^{j+1}-\boldsymbol{k}_i^j\|_2^2 + \frac{1}{T_i-2}\sum_{j=2}^{T_i-1}\|\boldsymbol{k}_i^{j+1}-2\boldsymbol{k}_i^j+\boldsymbol{k}_i^{j-1}\|_2^2\right),
\end{aligned} \tag{2}$$

where the first and second terms of each expression are first- and second-order smoothness terms, respectively.

The final inner loss L_{inner} is a weighted summation of the abovementioned three losses, and can be written as

$$L_{\text{inner}} = \lambda_{\text{shape}}L_{\text{shape}} + \lambda_{\text{pose}}L_{\text{pose}} + \lambda_{\text{cam}}L_{\text{cam}}, \tag{3}$$

where λ_{shape}, λ_{pose}, and λ_{cam} are the weight parameters.

3.4 Transfer Learning using a Differentiable Renderer

As the pre-trained HMR network is insufficient in terms of subject diversity, particularly for the body part thickness (see subsection 3.2), we fine-tune the HMR network to improve the coverage. Because publicly available gait datasets never provide the ground-truth SMPL parameters $\{\boldsymbol{\theta}, \boldsymbol{\beta}\}$, we instead exploit the silhouette masks provided by the gait datasets, which adequately reflect the body part thickness. More specifically, we render 2D binary silhouettes from the 3D human mesh obtained from the HMR network, and fine-tune the HMR network so that the rendered silhouettes match the silhouette masks in the gait datasets. In this way, we can adapt the HMR network to the subject population in the gait datasets through transfer learning.

For the purpose of the 3D-to-2D projection, we introduce a neural renderer [54] as a differentiable renderer, which can be integrated with the whole framework through end-to-end training. Given the shape, pose, and camera parameters $\boldsymbol{\Theta}_i^j$ of the j-th frame of the i-th sequence, as estimated by the HMR network, we render the binary silhouette with the differential render as $\hat{\boldsymbol{b}}(\boldsymbol{\Theta}_i^j)$. We then compute the reconstruction loss between the silhouette masks $\{\boldsymbol{b}_i^j\}$ and the rendered silhouettes as

$$L_{\text{recont}} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{T_i}\sum_{j=1}^{T_i}\|\boldsymbol{b}_i^j - \hat{\boldsymbol{b}}(\boldsymbol{\Theta}_i^j)\|_2^2. \tag{4}$$

3.5 Constraints on Joint Locations

If we fine-tune the HMR network using only the silhouette masks, we may produce corrupted poses from overfitting to the silhouette masks because the poses (or joint locations) are ambiguous in a textureless silhouette mask. For example, the silhouette masks may have some extra parts with loose clothes or carried objects, which are never considered in SMPL models of real humans, and unlikely joint locations may be estimated by overfitting the rendered silhouettes to the silhouette masks with these extra parts. However, thanks to the good coverage of the subject diversity for the pose and body part length, the pre-trained HMR network can estimate the joint locations accurately, even for external datasets (i.e., the gait datasets considered in this paper).

Therefore, we introduce the joint loss L_{joints} to constrain the 3D and 2D joint locations. The parameters $\boldsymbol{\Theta}_i^j$ and $\tilde{\boldsymbol{\Theta}}_i^j$ at the j-th frame of the i-th sequence given by the fine-tuned and pre-trained HMR networks, respectively, are transformed to the 3D and 2D joint locations using the mapping functions $\boldsymbol{x}_{\text{3D}}(\boldsymbol{\Theta})$ and $\boldsymbol{x}_{\text{2D}}(\boldsymbol{\Theta})$, respectively. The joint loss L_{joints} is defined as

$$L_{\text{joints}} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{T_i}\sum_{j=1}^{T_i}\sum_{\text{dim}\in\{\text{2D,3D}\}}\|\boldsymbol{x}_{\text{dim}}(\boldsymbol{\Theta}_i^j) - \boldsymbol{x}_{\text{dim}}(\tilde{\boldsymbol{\Theta}}_i^j)\|_2^2. \tag{5}$$

3.6 Disentangled Shape and Pose Features for Gait Recognition

The explicitly disentangled shape and pose features given by the HMR network are separately fed into recognition networks. For simplicity, let us drop the sequence index in this subsection.

Regarding the shape feature, we directly use the averaged shape feature $\boldsymbol{f}_{\text{shape}} = \bar{\boldsymbol{\beta}} \in \mathbb{R}^{10}$ of a given sequence for recognition.

For the pose feature, we first concatenate the 3D and 2D joint locations for a sequence with T frames as $P = \{\boldsymbol{p}^j | \boldsymbol{p}^j \in \mathbb{R}^{120}, j = 1, \ldots, T\}$,[4] and feed them into an LSTM/CNN for spatiotemporal feature extraction. We use a three-layer LSTM with 256 hidden units in each cell following [36]. Once we obtain an

[4] Five dimensions of 23 joints + one root joint sum up to $5 \times (23 + 1) = 120$.

LSTM output sequence $H = \{\boldsymbol{h}^j | \boldsymbol{h}^j \in \mathbb{R}^{256}, j = 1, \ldots, T\}$, we take the average $\boldsymbol{f}_{\text{pose}}^{\text{LSTM}} = (1/T)\sum_{j=1}^{T} \boldsymbol{h}^j \in \mathbb{R}^{256}$ as the pose feature for recognition.

We also use a CNN on the pose features, for which [12] mentioned that CNNs could outperform the LSTM. First, a 2D matrix is formed from the concatenated 3D and 2D joint locations of T frames as $[\boldsymbol{p}^1, \ldots, \boldsymbol{p}^T] \in \mathbb{R}^{120\times T}$. We employ a relatively simple CNN architecture consisting of three convolutional layers and one fully connected layer. Each convolutional layer is followed by a batch normalization layer and ReLU activation function. We set the kernel size to 3×3, and use 64, 128, and 256 channels in the respective layers. The feature size is reduced in the row direction by setting the vertical stride to 2, as there is a limited frame number in the column direction. Finally, we use the 52-dimensional output from the last fully connected layer as the pose features for recognition, $\boldsymbol{f}_{\text{pose}}^{\text{CNN}} \in \mathbb{R}^{52}$. Each feature $\boldsymbol{f} \in \{\boldsymbol{f}_{\text{shape}}, \boldsymbol{f}_{\text{pose}}^{\text{LSTM}}, \boldsymbol{f}_{\text{pose}}^{\text{CNN}}\}$ is taken separately for training and testing.

In the training stage, following [20, 45], we use different loss functions ($L_{\text{recog}} \in \{L_{\text{trip}}, L_{\text{cont}}\}$) for different recognition tasks (gait identification and verification). For gait identification, we assume that a mini-batch of sequences contains N_{trip} triplets $\{\boldsymbol{f}_Q^i, \boldsymbol{f}_G^i, \boldsymbol{f}_I^i\}(i = 1, \ldots, N_{\text{trip}})$, where a subject of $\boldsymbol{f}_G^i$ is the same as that of $\boldsymbol{f}_Q^i$ whereas a subject of $\boldsymbol{f}_I^i$ is different from that of $\boldsymbol{f}_Q^i$. The triplet loss [55] is defined as

$$L_{\text{trip}} = \frac{1}{N_{\text{trip}}} \sum_{i=1}^{N_{\text{trip}}} \max(m - \|\boldsymbol{f}_Q^i - \boldsymbol{f}_I^i\|_2^2 + \|\boldsymbol{f}_Q^i - \boldsymbol{f}_G^i\|_2^2, 0), \tag{6}$$

where m is the margin.

For gait verification, we assume that a mini-batch of sequences contains N_{pair} pairs $\{\boldsymbol{f}_P^i, \boldsymbol{f}_G^i\}(i = 1, \ldots, N_{\text{pair}})$, and the i-th pair has a binary label y_i (if $y_i = 1$, the pair is from the same subject sequence; if $y_i = 0$, the pair is from different subject sequences). The normalized contrastive loss [56] for the same and different subject pairs is defined as

$$L_{\text{cont}} = \frac{1}{N_s} \sum_{i=1}^{N_{\text{pair}}} y_i \|\boldsymbol{f}_P^i - \boldsymbol{f}_G^i\|_2^2 + \frac{1}{N_d} \sum_{i=1}^{N_{\text{pair}}} (1 - y_i) \max(m - \|\boldsymbol{f}_P^i - \boldsymbol{f}_G^i\|_2^2, 0), \tag{7}$$

where N_s, N_d are the number of same and different subject pairs in this batch.

3.7 Joint Loss Function

We train the whole network in an end-to-end manner with a joint loss function. This is a weighted summation of the aforementioned losses, and is defined as

$$L_{\text{total}} = \lambda_{\text{inner}} L_{\text{inner}} + \lambda_{\text{recont}} L_{\text{recont}} + \lambda_{\text{joints}} L_{\text{joints}} + \lambda_{\text{recog}} L_{\text{recog}}, \tag{8}$$

where λ_{inner}, λ_{recont}, and λ_{joints} are the weight parameters.

In a test case, we use the trained model to extract the feature $\boldsymbol{f}$ for the input RGB sequences, then compute the L2 distance between $\boldsymbol{f}$ for two sequences as a dissimilarity score for matching.

4 Experiments

4.1 Datasets

We evaluated our method on two datasets: OU-MVLP [20] and CASIA-B [40].

OU-MVLP is currently the largest database with various view variations. There are 10,307 subjects with 14 view angles (0°, 15°, ..., 90°; 180°, 195°, ..., 270°). Each subject has two sequences (#00, #01). Following the protocol of [20], 5,153 subjects were selected for the training set and the remaining 5,154 subjects were placed in the test set. Sequence #00 and sequence #01 were assigned as a probe and a gallery, respectively. Probes that did not have corresponding galleries (i.e., non-enrolled probes) were excluded from the test set.

CASIA-B is one of the most frequently used gait databases. There are 124 subjects with 11 view angles (0°, 18°, ..., 180°). Each subject has 10 sequences per view. Six are in the normal walking condition (NM), two are carrying a bag (BG), and the remaining two are wearing a coat (CL). We applied the protocol of using the first 74 subjects for training, with the remaining 50 subjects used for testing. In the test cases, NM #1–4 were assigned as galleries, and the other six were divided into three probes: NM #5–6, BG #1–2, and CL #1–2.

4.2 Implementation Details

We simply obtained the size-normalized and registered human silhouettes from the raw silhouettes provided by the datasets based on the region center and height, and cropped the raw RGB sequences using the corresponding locations. As for other real-scene videos, some state-of-the-art human detection and segmentation methods (e.g., Mask R-CNN [57] and RefineNet [34]) could be applied to obtain the required RGB sequences and silhouettes for training and testing.

The cropped RGB sequences were then scaled to 224×224 to fit the pre-trained HMR [37] model, and the corresponding silhouette masks were resized to a smaller 64×64 size to reduce the memory requirements. The frame number T in a period was set to 15. Thus, we generally selected 15 frames at equal intervals from a period in sequences, while omitting sequences with fewer than 15 frames.

For both datasets, other layers besides the HMR (which is initialized with a pre-trained model provided by [37]) used the default initialization. Adam [58] was chosen as the optimizer. The learning rate was initially set to 10^{-4}, and then decreased to 10^{-5} after a certain number of iterations M. For OU-MVLP, we set $M = 30$K and the total number of iterations to 60K. For CASIA-B, we basically set $M = 10$K and the total number of iterations to 15K; sometimes, we decreased the learning rate earlier or applied early stopping to prevent overfitting to the small training set. For each training iteration, we randomly selected eight subjects and eight sequences per subject to generate all possible triplets or pairs in a mini-batch. The margin m in Eqs. (6) and (7) was set to 0.2. The hyper-parameters of the inner loss and the joint loss functions in Eqs. (3) and (8) were experimentally set to $\lambda_{\text{shape}} = 1$, $\lambda_{\text{pose}} = 0.1$, $\lambda_{\text{cam}} = 0.00001$, $\lambda_{\text{inner}} = 1$, $\lambda_{\text{recont}} = 1$, $\lambda_{\text{joints}} = 100$, and $\lambda_{\text{recog}} = 1$.

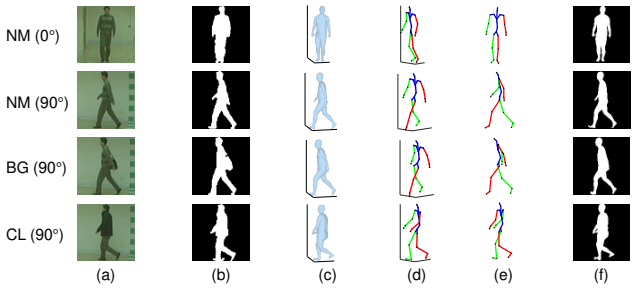

Fig. 3. Visualization of estimation results under different views and walking conditions on CASIA-B: (a) RGB inputs; (b) silhouette masks; (c) estimated SMPL models; (d) estimated 3D joint locations; (e) estimated 2D joint locations; (f) rendered silhouettes.

We evaluated the recognition performance in terms of the Rank-1 identification rate (denoted as Rank-1) and the equal error rate (denoted as EER) for the identification and verification tasks, respectively.

4.3 Feature Visualization

Some of the estimation results given by the proposed method are visualized in Fig. 3. This figure shows that the proposed method can estimate reasonable SMPL models under view, clothing, and carrying status variations, i.e., it obtains a similar body thickness between frontal and side views, and successfully excludes carried bags or clothing in the estimated SMPL models to some extent (e.g., the third row of Fig. 3). This is beneficial for robust gait recognition against changes in view, clothing, and carried objects. In addition, the pose (or joint locations) is also reasonably estimated, although some of the joints are not realistic (i.e., both left and right arms stay forward); this will be examined in future research.

4.4 Comparison using OU-MVLP

The benchmarks included both appearance-based approaches [60, 25, 18–20, 22, 61][5] and a model-based approach [11]. The deep models were trained using all 14 views. As mentioned in subsection 3.4, we trained and tested three features given by the proposed method (i.e., "pose_LSTM", "pose_CNN", and "shape") separately. Following [20], we selected four typical views (i.e., 0°, 30°, 60°, 90°) and present the test results averaged over each angular difference in Table 1. We also present more comprehensive results for some benchmarks in Table 2.

The results indicate that the proposed method using the shape feature outperforms state-of-the-art appearance- and model-based approaches in terms of both identification and verification tasks. As the view variation increases from 0° to 90°, the accuracy of the existing methods decreases rapidly, whereas our

[5] While the original GaitSet paper [22] reported results including the non-enrolled probes, the results here exclude the non-enrolled probes to ensure a fair comparison.

Table 1. Rank-1 rates and EERs of comparison methods on OU-MVLP based on the angular differences. "Mean" is the average result over all 16 probe and gallery view combinations. PTSN-O and PTSN-α denote PTSN with OpenPose [33] and with AlphaPose [59] as pose extractors, respectively. The first and second blocks are appearance-based and model-based approaches, respectively. Ours (ensemble) is an ensemble of "shape" and "pose_CNN". Bold and italic bold indicate the best and the second-best accuracies, respectively. This convention is consistent throughout this paper.

Methods	Rank-1 [%]					EER [%]				
	Angular difference					Angular difference				
	0°	30°	60°	90°	Mean	0°	30°	60°	90°	Mean
Direct Match	77.4	2.4	0.2	0.0	20.3	6.5	25.2	41.4	46.2	27.2
LDA [60]	81.6	10.1	0.8	0.1	24.4	6.2	22.7	35.7	40.1	24.0
VTM [25]	77.4	2.7	0.6	0.2	20.5	6.5	26.8	34.2	38.5	25.0
GEINet [18]	85.7	40.3	13.8	5.4	40.7	2.4	5.9	12.7	17.2	8.1
LB [19]	89.9	42.2	15.2	4.5	42.6	1.0	3.3	6.7	9.3	4.3
Takemura's [20]	89.5	55.0	30.0	17.3	52.7	1.0	2.0	3.4	4.2	2.4
PSTN [61]	93.9	69.2	41.9	25.9	63.1	0.6	1.5	2.8	3.7	1.9
GaitSet [22]	99.1	96.4	86.9	79.8	92.6	0.25	0.42	0.68	0.92	0.51
PTSN-O [11]	48.1	18.4	6.6	2.1	18.8	8.8	11.9	15.4	20.5	13.1
PTSN-α [11]	59.7	31.0	13.2	4.7	30.4	6.6	8.7	12.3	17.8	10.2
Ours (shape)	**99.6**	***97.7***	***95.0***	**91.4**	***96.7***	**0.12**	**0.19**	***0.23***	***0.29***	***0.19***
Ours (pose_LSTM)	88.1	61.9	47.3	35.1	61.5	0.36	0.55	0.79	1.06	0.62
Ours (pose_CNN)	98.3	91.9	81.9	74.5	88.8	***0.19***	***0.30***	0.41	0.58	0.33
Ours (ensemble)	***99.5***	**98.2**	**95.2**	***91.3***	**96.9**	**0.12**	**0.19**	**0.14**	**0.28**	**0.17**

Table 2. Rank-1 rates and EERs averaged over the 14 gallery views, where the identical view is excluded. Ours is an ensemble. GaitPart [42] only provides the mean Rank-1 rate.

	Methods	Probe view														
		0°	15°	30°	45°	60°	75°	90°	180°	195°	210°	225°	240°	255°	270°	Mean
Rank -1 [%]	PTSN-O	6.4	11.0	15.4	18.8	17.6	15.1	8.8	5.2	10.6	10.5	17.3	14.6	11.6	7.7	12.2
	PTSN-α	11.8	19.0	23.9	26.5	24.9	20.6	14.7	6.1	11.6	14.2	22.1	21.3	17.9	14.3	17.8
	GaitSet	***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	-	-	-	-	-	-	-	-	-	-	-	-	-	-	***95.1***
	Ours	**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**
EER [%]	PTSN-O	16.0	13.3	13.0	11.2	11.6	12.8	17.1	17.6	14.9	18.9	12.3	13.2	14.9	18.2	14.6
	PTSN-α	15.1	12.5	11.9	11.1	11.2	12.5	14.8	22.2	17.8	21.3	11.8	11.9	13.0	14.7	14.4
	GaitSet	***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***
	Ours	**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**

method mitigates the degradation. This demonstrates the robustness of our method against view variations.

However, our method performs worse with the pose feature than with the shape feature. This is partly because the extraction accuracy of the pose feature is sometimes degraded by self-occlusions when only a single-view sequence is used (e.g., both arms stay forward in Fig. 3). We also find that "pose_CNN" performs better than "pose_LSTM," which is consistent with the conclusions of the recent pose-based gait recognition study [12]. Despite the abovementioned difficulty, our method using the pose feature (CNN) outperforms most of the benchmarks and is even comparable to GaitSet [22], the state-of-the-art appearance-based method (i.e., slightly worse Rank-1 rate, but better EER).

Moreover, we constructed an ensemble of "shape" and "pose_CNN" by averaging their dissimilarity scores, and this produced a further improvement.

Table 3. Rank-1 rates [%] of comparison methods on CASIA-B using the first 74 subjects for training. The mean result over all 10 gallery views for each probe view is given, where the identical view is excluded.

Probe	Methods	Probe view											
		0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	Mean
NM #5–6	ViDP [62]	-	-	-	64.2	-	60.4	-	65.0	-	-	-	-
	CNN ensemble [19]	88.7	95.1	98.2	96.4	94.1	91.5	93.9	97.5	98.4	95.8	85.6	94.1
	Takemura's [20]	83.2	91.2	95.8	93.4	91.2	87.8	89.4	93.6	96.0	95.8	81.6	90.8
	PSTN [61]	87.0	93.8	96.2	94.4	92.2	91.8	92.0	95.0	96.0	96.4	84.8	92.7
	Song's GaitNet [35]	75.6	91.3	91.2	92.9	92.5	91.0	91.8	93.8	92.9	94.1	81.9	89.9
	Zhang's GaitNet [36]	91.2	92.0	90.5	95.6	86.9	92.6	93.5	96.0	90.9	88.8	89.0	91.6
	GaitSet [22]	90.8	***97.9***	**99.4**	96.9	93.6	91.7	95.0	***97.8***	***98.9***	96.8	85.8	95.0
	GaitPart [42]	94.1	**98.6**	***99.3***	**98.5**	94.0	92.3	95.9	**98.4**	**99.2**	***97.8***	90.4	96.2
	PoseGait [12]	55.3	69.6	73.9	75.0	68.0	68.2	71.1	72.9	76.1	70.4	55.4	68.7
	Ours (shape)	**97.1**	97.3	98.4	***98.4***	***97.4***	**98.3**	**97.7**	96.2	96.9	97.1	***97.5***	***97.5***
	Ours (pose_LSTM)	65.1	59.5	67.2	67.9	66.2	68.1	72.0	66.0	65.2	65.4	64.0	66.1
	Ours (pose_CNN)	87.1	88.3	93.8	95.4	92.1	92.8	90.5	90.7	88.5	92.4	91.7	91.2
	Ours (ensemble)	***96.9***	97.1	98.5	***98.4***	**97.7**	***98.2***	***97.6***	97.6	98.0	**98.4**	**98.6**	**97.9**
BG #1–2	LB [19]	64.2	80.6	82.7	76.9	64.8	63.1	68.0	76.9	82.2	75.4	61.3	72.4
	Zhang's GaitNet [36]	83.0	87.8	88.3	93.3	82.6	74.8	89.5	91.0	86.1	81.2	85.6	85.7
	GaitSet [22]	83.8	91.2	91.8	88.8	83.3	81.0	84.1	90.0	***92.2***	**94.4**	79.0	87.2
	GaitPart [42]	89.1	**94.8**	**96.7**	**95.1**	88.3	**94.9**	89.0	***93.5***	**96.1**	***93.8***	***85.8***	***91.5***
	PoseGait [12]	35.3	47.2	52.4	46.9	45.5	43.9	46.1	48.1	49.4	43.6	31.1	44.5
	Ours (shape)	***92.0***	91.7	92.2	93.0	***92.7***	91.6	***92.8***	92.3	88.4	86.5	83.4	90.6
	Ours (pose_LSTM)	53.9	50.5	52.4	51.9	49.1	50.6	47.1	49.4	47.0	44.2	45.7	49.3
	Ours (pose_CNN)	86.8	81.2	84.6	86.8	84.9	83.0	83.9	82.8	82.1	84.0	83.2	83.9
	Ours (ensemble)	**94.8**	***92.9***	***93.8***	***94.5***	**93.1**	***92.6***	**94.0**	**94.5**	89.7	93.6	**90.4**	**93.1**
CL #1–2	LB [19]	37.7	57.2	66.6	61.1	55.2	54.6	55.2	59.1	58.9	48.8	39.4	54.0
	Zhang's GaitNet [36]	42.1	58.2	65.1	70.7	68.0	70.6	65.3	69.4	51.5	50.1	36.6	58.9
	GaitSet [22]	61.4	75.4	80.7	77.3	72.1	70.1	71.5	73.5	***73.5***	68.4	50.0	70.4
	GaitPart [42]	70.7	**85.5**	**86.9**	**83.3**	77.1	72.5	**76.9**	**82.2**	**83.8**	**80.2**	66.5	**78.7**
	PoseGait [12]	24.3	29.7	41.3	38.8	38.2	38.5	41.6	44.9	42.2	33.4	22.5	36.0
	Ours (shape)	***72.1***	74.1	77.2	79.0	***77.3***	***76.7***	75.2	76.0	70.1	72.8	**74.8**	75.1
	Ours (pose_LSTM)	40.4	42.5	41.7	38.9	34.9	34.9	37.5	36.1	34.8	33.5	32.0	37.0
	Ours (pose_CNN)	63.0	62.4	66.3	65.2	61.9	58.2	58.3	59.1	56.8	55.4	55.6	60.2
	Ours (ensemble)	**78.2**	***81.0***	***82.1***	***82.8***	**80.3**	**76.9**	***75.5***	***77.4***	72.3	***73.5***	***74.2***	***77.6***

4.5 Comparison using CASIA-B

Table 3 compares the results from our method with those given by other benchmarks. We only report the Rank-1 rates, because few previous studies evaluated the verification accuracy. For three different probe sets, we trained three separate models for our method. Our method (ensemble) is comparable with the latest state-of-the-art GaitPart [42], i.e., 1.7% and 1.6% better for "NM" and "BG," but 1.1% worse for "CL", and achieves the best or the second-best accuracy in most cases. Additionally, compared with the two-step model-based approach with pose information [12], our pose-based method (pose_CNN) exhibits significantly better performance. This is because our method trains the human model fitting along with recognition tasks in an end-to-end manner, whereas the method of [12] extracts poses with the pre-trained OpenPose [33], which is not optimized for recognition.

4.6 Ablation Study of Loss Functions

We conducted ablation studies using our method with the shape feature (i.e., the best individual feature), and analyze the effect of the proposed loss func-

Table 4. Ablation study of our method (shape) on CASIA-B under NM probe setting.

Pre-trained HMR	Fine-tuned HMR	Fine-tuned loss			L_{recog}	Rank-1 [%]
		L_{inner}	L_{joints}	L_{recon}		
√						4.1
	√		√	√	√	94.3
	√	√		√	√	***94.8***
	√	√	√		√	93.6
	√	√	√	√	√	**97.5**

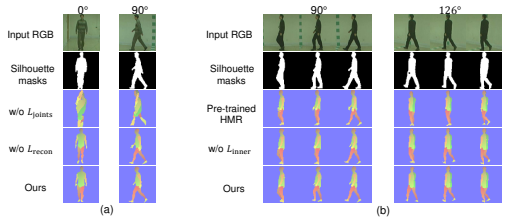

Fig. 4. Rendered semantic silhouettes comparison: (a) effect of L_{joints} and L_{recon}; (b) effect of L_{inner}.

tions on the recognition performance in Table 4. The pre-trained HMR without recognition loss L_{recog} yields the worst accuracy (the first row), whereas introducing the recognition loss L_{recog} significantly improves the accuracy (the second to the fifth rows). When we separately turn off the inner loss L_{inner}, joint loss L_{joints}, and reconstruction loss L_{recon} (see the second to the fourth rows), the accuracy decreases by 3–4%, indicating the contribution of each loss term. Moreover, we show a qualitative comparison of the rendered semantic silhouettes in Fig. 4. In line with our expectations, L_{joints} avoids the pose corruption problem and L_{recon} enables wider body shapes to be learned from the silhouette masks; L_{inner} helps learn continuous poses, which can fix the error estimations of the pre-trained HMR. All the results indicate that our method can accurately estimate the SMPL models from RGB sequences, which is beneficial for future gait recognition studies.

5 Conclusion

We have proposed an end-to-end model-based gait recognition approach that models the human gait via an SMPL model and provides explicitly disentangled shape and pose representations. We evaluated the recognition performance of shape and pose features for different recognition tasks. The experimental results on two datasets show that our method outperforms a number of state-of-the-art approaches. Because current recognition networks for pose features ignore the structure information between joints (e.g., limb length), more suitable networks will be investigated in the future.

Acknowledgement. This work was supported by JSPS KAKENHI Grant No. JP18H04115, JP19H05692, and JP20H00607, and the National Natural Science Foundation of China (Grant No. 61727802).

References

1. Bouchrika, I., Goffredo, M., Carter, J., Nixon, M.: On using gait in forensic biometrics. Journal of Forensic Sciences **56** (2011) 882–889
2. Iwama, H., Muramatsu, D., Makihara, Y., Yagi, Y.: Gait verification system for criminal investigation. IPSJ Transactions on Computer Vision and Applications **5** (2013) 163–175
3. Lynnerup, N., Larsen, P.: Gait as evidence. IET Biometrics **3** (2014) 47–54
4. Wagg, D., Nixon, M.: On automated model-based extraction and analysis of gait. In: Proc. of the 6th IEEE Int. Conf. on Automatic Face and Gesture Recognition. (2004) 11–16
5. Yam, C., Nixon, M., Carter, J.: Automated person recognition by walking and running via model-based approaches. Pattern Recognition **37** (2004) 1057–1072
6. Bobick, A., Johnson, A.: Gait recognition using static activity-specific parameters. In: CVPR. Volume 1. (2001) 423–430
7. Cunado, D., Nixon, M., Carter, J.: Automatic extraction and description of human gait models for recognition purposes. Computer Vision and Image Understanding **90** (2003) 1–41
8. Yamauchi, K., Bhanu, B., Saito, H.: 3d human body modeling using range data. In: ICPR. (2010) 3476–3479
9. Ariyanto, G., Nixon, M.: Marionette mass-spring model for 3d gait biometrics. In: Proc. of the 5th IAPR International Conference on Biometrics. (2012) 354–359
10. Feng, Y., Li, Y., Luo, J.: Learning effective gait features using lstm. In: ICPR. (2016) 325–330
11. Liao, R., Cao, C., Garcia, E.B., Yu, S., Huang, Y.: Pose-based temporal-spatial network (ptsn) for gait recognition with carrying and clothing variations. In: Proceedings of the 12th Chinese Conference on Biometric Recognition. (2017) 474–483
12. Liao, R., Yu, S., An, W., Huang, Y.: A model-based gait recognition method with body pose and human prior knowledge. Pattern Recognition **98** (2020) 107069
13. Han, J., Bhanu, B.: Individual recognition using gait energy image. IEEE Transactions on Pattern Analysis and Machine Intelligence **28** (2006) 316– 322
14. Xu, D., Yan, S., Tao, D., Zhang, L., Li, X., jiang Zhang, H.: Human gait recognition with matrix representation. IEEE Trans. Circuits Syst. Video Technol **16** (2006) 896–903
15. Lu, J., Tan, Y.P.: Uncorrelated discriminant simplex analysis for view-invariant gait signal computing. Pattern Recognition Letters **31** (2010) 382–393
16. Guan, Y., Li, C.T., Roli, F.: On reducing the effect of covariate factors in gait recognition: A classifier ensemble method. IEEE Transactions on Pattern Analysis and Machine Intelligence **37** (2015) 1521–1528
17. Makihara, Y., Suzuki, A., Muramatsu, D., Li, X., Yagi, Y.: Joint intensity and spatial metric learning for robust gait recognition. In: CVPR. (2017) 5705–5715
18. Shiraga, K., Makihara, Y., Muramatsu, D., Echigo, T., Yagi, Y.: Geinet: View-invariant gait recognition using a convolutional neural network. In: ICB. (2016)
19. Wu, Z., Huang, Y., Wang, L., Wang, X., Tan, T.: A comprehensive study on cross-view gait based human identification with deep cnns. IEEE Transactions on Pattern Analysis and Machine Intelligence **39** (2017) 209–226
20. Takemura, N., Makihara, Y., Muramatsu, D., Echigo, T., Yagi, Y.: On input/output architectures for convolutional neural network-based cross-view gait recognition. IEEE Transactions on Circuits and Systems for Video Technology **29** (2019) 2708–2719

21. Zhang, K., Luo, W., Ma, L., Liu, W., Li, H.: Learning joint gait representation via quintuplet loss minimization. In: CVPR. (2019)
22. Chao, H., He, Y., Zhang, J., Feng, J.: Gaitset: Regarding gait as a set for cross-view gait recognition. In: AAAI. (2019)
23. Li, X., Makihara, Y., Xu, C., Yagi, Y., Ren, M.: Joint intensity transformer network for gait recognition robust against clothing and carrying status. IEEE Transactions on Information Forensics and Security (2019) 1–1
24. Kusakunniran, W., Wu, Q., Zhang, J., Li, H.: Support vector regression for multi-view gait recognition based on local motion feature selection. In: CVPR, San Francisco, CA, USA (2010) 1–8
25. Makihara, Y., Sagawa, R., Mukaigawa, Y., Echigo, T., Yagi, Y.: Gait recognition using a view transformation model in the frequency domain. In: ECCV, Graz, Austria (2006) 151–163
26. Makihara, Y., Tsuji, A., Yagi, Y.: Silhouette transformation based on walking speed for gait identification. In: CVPR, San Francisco, CA, USA (2010)
27. Muramatsu, D., Shiraishi, A., Makihara, Y., Uddin, M., Yagi, Y.: Gait-based person recognition using arbitrary view transformation model. IEEE Trans. on Image Processing **24** (2015) 140–154
28. Mansur, A., Makihara, Y., Aqmar, R., Yagi, Y.: Gait recognition under speed transition. In: CVPR. (2014) 2521–2528
29. Akae, N., Mansur, A., Makihara, Y., Yagi, Y.: Video from nearly still: an application to low frame-rate gait recognition. In: CVPR, Providence, RI, USA (2012) 1537–1543
30. Yu, S., Liao, R., An, W., Chen, H., Garcia, E.B., Huang, Y., Poh, N.: Gaitganv2: Invariant gait feature extraction using generative adversarial networks. Pattern Recognition **87** (2019) 179 – 189
31. He, Y., Zhang, J., Shan, H., Wang, L.: Multi-task gans for view-specific feature learning in gait recognition. IEEE Transactions on Information Forensics and Security **14** (2019) 102–113
32. Wang, C., Zhang, J., Wang, L., Pu, J., Yuan, X.: Human identification using temporal information preserving gait template. IEEE Trans. on Pattern Analysis and Machine Intelligence **34** (2012) 2164 –2176
33. Cao, Z., Hidalgo, G., Simon, T., Wei, S.E., Sheikh, Y.: OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. In: arXiv preprint arXiv:1812.08008. (2018)
34. Lin, G., Milan, A., Shen, C., Reid, I.D.: Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. CVPR (2017) 5168–5177
35. Song, C., Huang, Y., Huang, Y., Jia, N., Wang, L.: Gaitnet: An end-to-end network for gait based human identification. Pattern Recognition **96** (2019) 106988
36. Zhang, Z., Tran, L., Yin, X., Atoum, Y., Wan, J., Wang, N., Liu, X.: Gait recognition via disentangled representation learning. In: CVPR, Long Beach, CA (2019)
37. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: CVPR. (2018) 7122–7131
38. Pavlakos, G., Zhu, L., Zhou, X., Daniilidis, K.: Learning to estimate 3d human pose and shape from a single color image. In: CVPR. (2018)
39. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. ACM Trans. Graphics (Proc. SIGGRAPH Asia) **34** (2015) 248:1–248:16
40. Yu, S., Tan, D., Tan, T.: A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In: ICPR. Volume 4., Hong Kong, China (2006) 441–444

41. Pfister, T., Charles, J., Zisserman, A.: Flowing convnets for human pose estimation in videos. In: ICCV. (2015)
42. Fan, C., Peng, Y., Cao, C., Liu, X., Hou, S., Chi, J., Huang, Y., Li, Q., He, Z.: Gaitpart: Temporal part-based model for gait recognition. In: CVPR. (2020)
43. Tran, L., Yin, X., Liu, X.: Disentangled representation learning gan for pose-invariant face recognition. In: CVPR. (2017)
44. Esser, P., Sutter, E., Ommer, B.: A variational u-net for conditional appearance and shape generation. In: CVPR. (2018)
45. Li, X., Makihara, Y., Xu, C., Yagi, Y., Ren, M.: Gait recognition via semi-supervised disentangled representation learning to identity and covariate features. In: CVPR. (2020)
46. Liu, W., Zhixin Piao, Min Jie, W.L.L.M., Gao, S.: Liquid warping gan: A unified framework for human motion imitation, appearance transfer and novel view synthesis. In: ICCV. (2019)
47. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: ECCV. (2016)
48. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV. (2014)
49. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2d human pose estimation: New benchmark and state of the art analysis. In: CVPR. (2014)
50. Johnson, S., Everingham, M.: Learning effective human pose estimation from inaccurate annotation. In: CVPR. (2011)
51. Johnson, S., Everingham, M.: Clustered pose and nonlinear appearance models for human pose estimation. In: BMVC. (2010)
52. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. IEEE Transactions on Pattern Analysis and Machine Intelligence **36** (2014) 1325–1339
53. Mehta, D., Rhodin, H., Casas, D., Fua, P., Sotnychenko, O., Xu, W., Theobalt, C.: Monocular 3d human pose estimation in the wild using improved cnn supervision. In: Fifth International Conference on 3D Vision (3DV). (2017)
54. Hiroharu Kato, Y.U., Harada, T.: Neural 3d mesh renderer. In: CVPR. (2018)
55. Wang, J., Song, Y., Leung, T., Rosenberg, C., Wang, J., Philbin, J., Chen, B., Wu, Y.: Learning fine-grained image similarity with deep ranking. In: CVPR. (2014)
56. Hadsell, R., Chopra, S., LeCun, Y.: Dimensionality reduction by learning an invariant mapping. In: CVPR. Volume 2. (2006) 1735–1742
57. He, K., Gkioxari, G., Dollár, P., Girshick, R.B.: Mask R-CNN. In: ICCV. (2017)
58. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint (2014)
59. Fang, H.S., Xie, S., Tai, Y.W., Lu, C.: Rmpe: Regional multi-person pose estimation. In: The IEEE International Conference on Computer Vision (ICCV). (2017)
60. OTSU, N.: Optimal linear and nonlinear solutions for least-square discriminant feature extraction. In: ICPR. (1982) 557–560
61. Xu, C., Makihara, Y., Li, X., Yagi, Y., Lu, J.: Cross-view gait recognition using pairwise spatial transformer networks. IEEE Transactions on Circuits and Systems for Video Technology (2020) 1–1
62. Hu, M., Wang, Y., Zhang, Z., Little, J.J., Huang, D.: View-invariant discriminative projection for multi-view gait-based human identification. IEEE Transactions on Information Forensics and Security **8** (2013) 2034–2045