Occlusion-Aware Human Mesh Model-Based Gait Recognition

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

***Abstract*— Partial occlusion of the human body caused by obstacles or a limited camera field of view often occurs in surveillance videos, which affects the performance of gait recognition in practice. Existing methods for gait recognition against occlusion require a bounding box or the height of a full human body as a prerequisite, which is unobserved in occlusion scenarios. In this paper, we propose an occlusion-aware model-based gait recognition method that works directly on gait videos under occlusion without the above-mentioned prerequisite. Specifically, given a gait sequence that only contains non-occluded body parts in the images, we directly fit a skinned multi-person linear (SMPL)-based human mesh model to the input images without any pre-normalization or registration of the human body. We further use the pose and shape features extracted from the estimated SMPL model for recognition purposes, and use the extracted camera parameters in the occlusion attenuation module to reduce intra-subject variation in human model fitting caused by occlusion pattern differences. Experiments on occlusion samples simulated from the OU-MVLP dataset demonstrated the effectiveness of the proposed method, which outperformed state-of-the-art gait recognition methods by about 15% rank-1 identification rate and 2% equal error rate in the identification and verification scenarios, respectively.**

***Index Terms*— Partial occlusion, gait recognition, human mesh model.**

I. Introduction

GAIT recognition is a popular biometric that recognizes people from their unique gait features, including the body shape and walking posture characteristics. The gait has distinct advantages over other biometrics (e.g., DNA, fingerprint, and face), such as long-distance capture without subject cooperation and applicability to low-resolution images. Therefore, gait recognition is considered to have great potential in applications that use CCTV footage, such as surveillance, forensics, and criminal investigation [1], [2], [3].

In practical applications, gait recognition is also subject to several challenging factors, including walking speed [4], [5], observation view [6], [7], the carried object [8], [9], and occlusion [10], [11]. Among them, occlusion is a common covariate in captured gait videos, in which part of the walking person is temporarily or continuously occluded by obstacles, or caused by a limited camera field of view (see Fig. 1 for examples). The occlusion of the body leads to a lack of body shape, pose, and motion information, which greatly affects the performance of gait recognition.

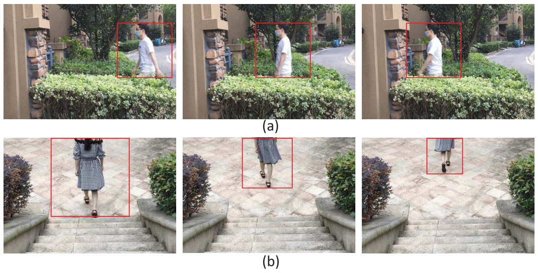

Fig. 1. Examples of occlusion in real life. (a) The lower part of the subject is continuously occluded by the flower bed during walking. (b) Because of the limited camera field of view, most parts of the subject can be observed when she is close to the camera, and as she walks away, the occlusion gradually increases.

Previous studies on gait recognition against occlusion are mainly divided into two categories: reconstruction-based approaches and reconstruction-free approaches. Reconstruction-based approaches first reconstruct non-occluded silhouette images [11], [12] or gait features (e.g., gait energy image (GEI) [13]) of the entire body [10] from the given gait sequence under occlusion. Reconstruction-free approaches directly extract gait features from occluded images without regenerating full-body images [14], [15], or apply matching only in the same visible (i.e., non-occluded) regions of a matching pair [16], [17].

Most existing methods work on cropped images that are size-normalized and registered based on a full human body; that is, they use a full-body bounding box as a prerequisite for cropping despite the full body being unobserved in occlusion scenarios (see Fig. 2(a) and (b)).

Unlike the above-mentioned appearance-based approaches to occlusion handling with a prerequisite, model-based approaches (e.g., ModelGait [18]) have the potential to handle occlusion without a prerequisite in a more natural manner. This is because we can not only obtain the body shape and pose parameters but also locate a full body position (i.e., a bounding box for a full body) as a result of human model fitting, even from a partially occluded image. In fact, a human mesh model (e.g., skinned multi-person linear

Manuscript received 4 August 2022; revised 6 December 2022; accepted 26 December 2022. Date of publication 11 January 2023; date of current version 27 January 2023. This work was supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI under Grant JP20H00607. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Emanuele Maiorana. *(Corresponding author: Chi Xu.)*

The authors are with the Institute of Scientific and Industrial Research, Osaka University, Osaka 567-0047, Japan (e-mail: xu@am.sanken.osaka-u.ac.jp; makihara@am.sanken.osaka-u.ac.jp; li@am.sanken.osaka-u.ac.jp; yagi@am.sanken.osaka-u.ac.jp).

This article has supplementary downloadable material available at https://doi.org/10.1109/TIFS.2023.3236181, provided by the authors.

Digital Object Identifier 10.1109/TIFS.2023.3236181

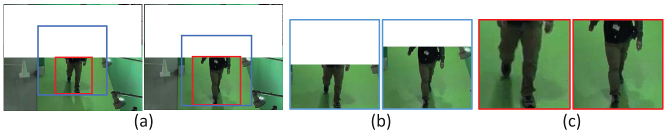

Fig. 2. Different bounding boxes used by existing works and our method. (a) Bounding boxes shown in original RGB with simulated occlusion (blue: existing works, red: our method). Existing works use a full-body bounding box including occluded body parts for cropping by assuming a known human height, whereas our method uses a bounding box containing only visible body parts. (b) Cropped images used by existing works. Human scales and body centers in the input sequence are normalized by full-body bounding boxes. (c) Inputs of our method. Body scales and centers may vary in a sequence due to cropping and resizing.

(SMPL) [19]) has been successfully estimated from a single image with partial occlusion [20], [21], [22].

Therefore, we propose a model-based gait recognition method that combines human mesh (i.e., SMPL model) estimation and gait recognition to address the occlusion issue. The contributions of this work are three-fold.

A. Model-Based Gait Recognition Method Against Occlusion Without a Prerequisite

We propose the first end-to-end model-based method that acts on occluded RGB gait videos. Unlike existing methods for occlusion-handling appearance-based gait recognition, our model-based method directly uses a bounding box of only visible body parts without the above-mentioned prerequisite (see Fig. 2(a) and (c)), simulating pedestrian detection results in real occluded scenes. Additionally, existing approaches to occlusion-handling human model fitting work on a single input image, whereas we estimate a temporally consistent and continuous sequence of human model parameters, which are beneficial for subsequent recognition modules. We train the entire framework in an end-to-end manner to achieve a trade-off between model estimation and recognition accuracy.

B. Occlusion-Aware Framework to Reduce the Effect of Occlusion on Model Fitting

Rather than simply retraining ModelGait [18] on occlusion data, we also incorporate a gated recurrent unit (GRU) module [23] for temporal information capture, and an occlusion attenuation module to reduce intra-subject variation in the model estimation. Specifically, because the difference between occlusion patterns in the input sequence may lead to different human model estimations, even for the same subject, we introduce an occlusion attenuation module to alleviate the dependence of model estimation on occlusion by considering the model parameters that imply the input occlusion patterns. Thus, the obtained human models become more similar for the same subject, which is beneficial to the matching task.

C. State-of-the-Art Performance on Occlusion Data

We evaluated the proposed method using OU-MVLP [24], which is the world's largest gait dataset with wide view variations. We prepared various types of occlusion to simulate occlusion scenarios that often occur in real life. Compared with existing state-of-the-art gait recognition methods, the proposed method achieved superior performance in both identification and verification scenarios.

II. Related Work

A. Gait Recognition Against Occlusion

1) Reconstruction-Free Approaches: Some typical reconstruction-free approaches directly apply various machine learning techniques to the occluded gait sequence to extract gait features that are relatively insensitive to occlusion [14], [15], [25], [26], such as a statistical analysis-based weighted averaging method [15]. A few researchers have attempted to apply matching to the same visible regions in a pair of samples by dividing the body into several parts [16], [17], [27], [28], [29]. In multi-gait recognition [30], body parts affected by inter-subject occlusion are excluded using an automatic tracking and segmentation method, and features then directly extracted from the obtained single-gait images, without special treatment for a relatively small occlusion.

However, these methods may not work well for a large occlusion, particularly when the same visible regions of a matching pair are very small (e.g., the upper body of a probe is occluded, whereas the lower body of a gallery is occluded).

2) Reconstruction-Based Approaches: Reconstruction-based approaches first reconstruct silhouettes or gait features without occlusion before the feature learning and matching process [10], [11], [12], [31]. For example, Muramatsu et al. [10] reconstructed an entire frequency domain feature (FDF) [6] directly from a partially occluded FDF using a subspace-based method. An approach based on a conditional generative adversarial network was proposed in [11], which combines silhouette reconstruction and gait recognition in a unified convolutional neural network (CNN) framework.

However, these methods require the prerequisite of a full-body bounding box to ensure a size-normalized and body center-registered silhouette sequence under occlusion, which increases the difficulty of applying them to real-world scenes.

B. Robust Gait Recognition Against Various Covariates

In addition to occlusion, there are other challenging covariates that may affect gait recognition performance, such as walking speed [4], [5], [32], [33], clothing and carrying [8], [9], [34], [35], [36], [37], [38], and view angles [6], [7], [39], [40], [41], [42], [43]. A variety of approaches have been proposed by designing gait representations combined with metric learning or deep learning techniques. For example, a metric learning technique called random subspace method [4] and a gait representation called single-support GEI [5] were designed for speed variations; body part-based templates [35] and a generative adversarial network-based method [38] were proposed for clothing and carrying factors; a CNN-based method named GEINet [39] was proposed for view variations, and later, various CNN structures with a pair of input GEIs were investigated in [7] and [41].

Rather than focusing on a specific covariate, some recent works address more general gait recognition scenarios [44], [45], [46], [47], [48], [49], [50], [51], [52], [53]. With the development of deep learning, most approaches work directly on silhouettes or RGB images, which significantly improves the performance compared to GEI. GaitSet [44] ignored the order information in silhouette sequences and treated the input as a set, achieving landmark recognition accuracy.

GaitPart [47] further improved performance by considering body division and micro-motions (i.e., short-range temporal features) contained in each body part. In [49], global and local features are combined in conjunction with 3D CNNs. To eliminate the effects of color and texture in the input RGBs, [45], [46] extracted pose features via disentangled representation learning, and [53] synthesized silhouettes that mask bodies with trainable edges.

Besides the aforementioned appearance-based methods, model-based methods [18], [54], [55], [56], [57], [58], [59], [60] have also shown impressive results recently. After obtaining body skeletons from RGB images using pose estimation works (e.g., OpenPose [61]), CNN models [54], [55] and graph convolutional networks [57], [58], [59] were employed to learn pose features for recognition. The first gait database with pose sequences, OUMVLP-Pose, was proposed in [56], promoting model-based gait recognition research. To exploit both shape and pose features for recognition, Li et al. [18] proposed an end-to-end model-based method by incorporating the human mesh recovery (HMR) framework [62], which estimates the SMPL human models [19] for the subsequent recognition network. The multi-view training framework in [60] further improved the human model estimation accuracy, and a database with human meshes, OUMVLP-Mesh, was constructed accordingly.

Compared to appearance-based methods, model-based methods have the potential to be more robust to occlusions without the full-body bounding box prerequisite thanks to the model fitting framework.

C. 3D Human Shape and Pose Estimation From an Occluded Image

In several recent studies on 3D human shape and pose estimation, the researchers focused on the estimation of full-body keypoints or a full human body mesh (e.g., SMPL model [19]) from an image under occlusion, where the body was partially occluded by an object [20], [22], [63] or the camera field of view [21]. For example, Cheng et al. [63] proposed an occlusion-aware framework for 3D pose estimation from a video, which combines estimation of keypoint confidence heatmaps with constraints on optical-flow consistency, suppressing unreliable estimation of occluded keypoints. In [20], the object-occluded human body was represented as a partial UV map, and the SMPL model estimation task was therefore converted to a UV map inpainting problem. In [21], Rockwell et al. handled partial images from a consumer video, where only part of a person was visible, which made human model estimation more difficult. Using a simple self-training framework, combined with cropping operations and a confident sample selection scheme, the HMR model [62] was successfully adapted to unlabeled partial images (i.e., without the ground-truth SMPL parameters or joint positions) to reconstruct the human mesh (hereafter, we refer to this method as partial HMR).

Compared with pure keypoints, the SMPL human model contains more information such as body shape, which is more beneficial to the gait recognition task. Additionally, although the partial HMR performed well for SMPL estimation on a single occluded image, it is not necessarily suitable for direct application to continuous video frames because temporal information is ignored.

III. Occlusion-Aware Model-Based Gait Recognition

A. Overview

An overview of the proposed method is shown in Fig. 3. Given a gait video with occlusion, we crop the unoccluded body parts using a square bounding box for each frame (see Fig. 1), and then resize the cropped images to a unified image size while maintaining the aspect ratio. We use a sequence encoder to estimate the 3D human mesh (i.e., the body shape and pose parameters), and global rotation and camera parameters for each cropped input image. Thereafter, we alleviate the intra-subject variation of the estimated SMPL model parameters induced by various occlusion patterns using an occlusion attenuation module. Finally, we feed the 3D joint locations obtained from the pose parameters and shape parameters averaged over frames into a recognition module.

B. SMPL Model

We briefly introduce the SMPL model [19] used in our work. We parameterize the SMPL model using the shape $\boldsymbol{\beta} \in \mathbb{R}^{10}$ (e.g., by expressing the body height, weight, and body proportions) and pose $\boldsymbol{\theta} \in \mathbb{R}^{69}$ parameters (i.e., the relative 3D rotation of 23 joints in the axis-angle representation). A triangulated mesh with 6,980 vertices can be output from the SMPL using a differentiable function. We compute the 3D joint locations from the vertices using linear regression. We can further obtain the 2D projection using the 3D global rotation $\boldsymbol{r} \in \mathbb{R}^3$ with a weak-perspective camera model, which is composed of parameters $\boldsymbol{\sigma} = [s, \boldsymbol{t}] \in \mathbb{R}^3$, where s and $\boldsymbol{t} \in \mathbb{R}^2$ are the scale and translation parameters, respectively.

C. Sequence Encoder

We estimate the SMPL parameters from each input frame using a sequence encoder, which is composed of a feature extractor, GRU module [23], and regressor. Unlike the partial HMR [21] and ModelGait [18], our method includes a GRU module to capture temporal information [64] when inferring SMPL model parameters. Particularly, we use bidirectional GRU (BiGRU) to learn latent features from both past and future frames because it temporally constrains the body shape and pose parameters better when occlusion changes within a sequence.

Given a sequence $S_i (i = 1, \ldots, N)$, where N is the number of sequences used for training, and the sequence is composed of T_i frames $\{I_i^1, \ldots, I_i^{T_i}\}$, we first extract a 2,048D feature from each frame using ResNet-50 [65]. Then we feed the extracted feature into a BiGRU with a hidden size of 1,024, which outputs the updated feature based on both past and future frames. To better merge the output feature learned from two directions, we use an additional fully connected (FC) layer with 2,048 neurons after the BiGRU. Finally, we regress the SMPL parameters using a 3D regressor with iterative feedback [62]. Finally, the estimated parameters

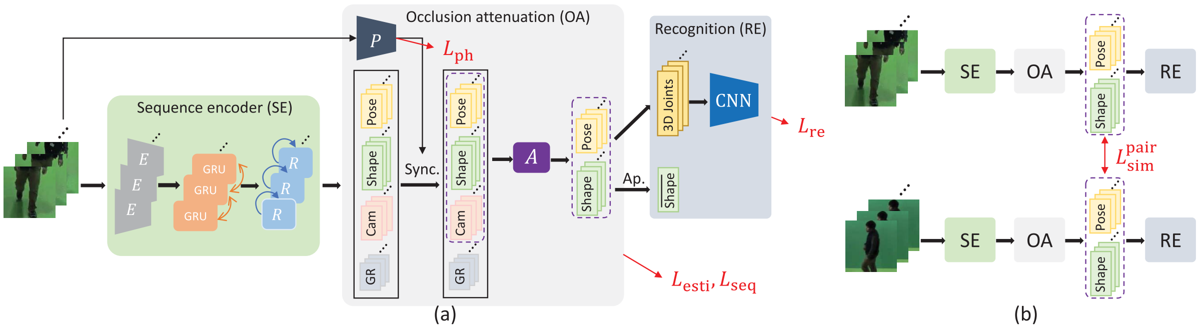

Fig. 3. Overview of the proposed method. (a) Given an RGB sequence, a sequence encoder first estimates an initial SMPL model, and then an occlusion attenuation module mitigates the dependence of the estimated body parameters on the input occlusion. The attenuated body parameters are further fed into the recognition module. E, R, P, and A denote the feature extractor, regressor, phase estimation network, and attenuation transformation, respectively. GR, Sync., and Ap. represent global rotation, phase synchronization, and average pooling, respectively. (b) To supervise the occlusion dependency attenuation, we further define a paired similarity loss $L_{\text{sim}}^{\text{pair}}$ to make the attenuated body parameters more consistent between the same subject input pair.

$\hat{\boldsymbol{\Theta}}_i^j = [\hat{\boldsymbol{\beta}}_i^j, \hat{\boldsymbol{\theta}}_i^j, \hat{\boldsymbol{r}}_i^j, \hat{\boldsymbol{\sigma}}_i^j] \in \mathbb{R}^{85}$ for the j-th frame of the i-the input sequence $I_i^j (j = 1, \ldots, T_i)$ can be written as

$$\hat{\boldsymbol{\Theta}}_i^j = R(G(E(I_i^j))), \tag{1}$$

where E, G, and R represent the feature extractor, GRU module, and regressor, respectively.

D. Occlusion Attenuation Module

We incorporate an occlusion attenuation module to mitigate the intra-subject variation of the estimated SMPL model parameters induced by differences between the input occlusion patterns (e.g., upper-body and lower-body occlusion). Because intra-subject pose variation should be computed in the same phase (i.e., gait stance), we first estimate the phase sequence of an input gait video and then synchronize the phases among sequences.

1) Phase Estimation: We use a CNN network P to estimate the phase label sequence from the input RGB images, which is denoted by

$$\hat{\boldsymbol{P}}_i = P(S_i), \tag{2}$$

where $\hat{\boldsymbol{P}}_i = \{\hat{\boldsymbol{p}}_i^1, \ldots, \hat{\boldsymbol{p}}_i^{T_i}\}$, and $\hat{\boldsymbol{p}}_i^j \in \mathbb{R}^2$ denotes the estimated phase for the j-th frame of the input sequence S_i, and is expressed as a sine and cosine function-based cyclic 2D vector, similar to that used in [66].

To obtain the estimated phase, we first feed each input frame into four convolutional layers, where each layer has a subsequent batch normalization layer [67] and ReLU activation function [68]. The size of each convolution kernel is 4 × 4, the stride is two, and the number of channels is increased from 16 to 128. Then we use an FC layer to extract a 50D feature. Similar to the methodology in Sec. III-C, we further apply a BiGRU and another FC layer to learn temporal information from past and future frames. We feed the obtained 100D feature into the final FC layer to regress the 2D phase label, which is followed by a normalization layer to maintain $\|\hat{\boldsymbol{p}}_i^j\|_2 = 1$.

The estimated phase is supervised by its ground-truth phase $\boldsymbol{p}_i^j$ using an estimation loss, which is defined as

$$L_{\text{esti}}^{\text{phase}} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{T_i}\sum_{j=1}^{T_i}\|\hat{\boldsymbol{p}}_i^j - \boldsymbol{p}_i^j\|_2^2. \tag{3}$$

To maintain the temporal continuity of the estimated phase label sequence, we define a smoothness loss as

$$L_{\text{smoo}}^{\text{phase}} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{T_i - 1}\sum_{j=1}^{T_i-1}\|\hat{\boldsymbol{p}}_i^{j+1} - \hat{\boldsymbol{p}}_i^j\|_2^2 \right. \\ \left. + \frac{1}{T_i - 2}\sum_{j=2}^{T_i-1}\|\hat{\boldsymbol{p}}_i^{j+1} - 2\hat{\boldsymbol{p}}_i^j + \hat{\boldsymbol{p}}_i^{j-1}\|_2^2\right). \tag{4}$$

Additionally, we penalize disordered phase labels between adjacent frames (i.e., reverse evolution of gait stances), which is formulated as

$$L_{\text{penal}}^{\text{phase}} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{|\mathcal{P}_i|}\sum_{(j,j+1)\in\mathcal{P}_i}\|\hat{\boldsymbol{p}}_i^{j+1} - \hat{\boldsymbol{p}}_i^j\|_2^2, \tag{5}$$

where $\mathcal{P}_i$ denotes the set of adjacent frame index pairs with disordered phase labels for the i-th input sequence.

We compute the entire loss function for the phase estimation network as follows:

$$L_{\text{ph}} = w_{\text{esti}}^{\text{phase}} L_{\text{esti}}^{\text{phase}} + w_{\text{smoo}}^{\text{phase}} L_{\text{smoo}}^{\text{phase}} + w_{\text{penal}}^{\text{phase}} L_{\text{penal}}^{\text{phase}}, \tag{6}$$

where $w_{\text{esti}}^{\text{phase}}$, $w_{\text{smoo}}^{\text{phase}}$, and $w_{\text{penal}}^{\text{phase}}$ are the weight parameters for the above three losses.

2) Phase Synchronization: Then we synchronize the initial SMPL parameters output by the sequence encoder using linear interpolation based on the estimated phase label sequence. Specifically, we first define a canonical gait period with T frames, where the phase evolves uniformly in these frames. Then we compute the interpolation weights between a canonical phase and its two neighboring estimated phases, and finally interpolate the initial SMPL parameter sequence $\hat{\boldsymbol{\Theta}}_i = \{\hat{\boldsymbol{\Theta}}_i^1, \ldots, \hat{\boldsymbol{\Theta}}_i^{T_i}\}$ into the synchronized SMPL sequence $\hat{\boldsymbol{\Theta}}'_i = \{\hat{\boldsymbol{\Theta}}'^1_i, \ldots, \hat{\boldsymbol{\Theta}}'^T_i\}$, where each $\hat{\boldsymbol{\Theta}}'^j_i (j = 1, \ldots, T)$ corresponds to a canonical phase label.

3) Occlusion Dependency Attenuation: Next, we attempt to reduce the intra-subject variation of the estimated SMPL model caused by the occlusion pattern variations by transforming the phase-synchronized SMPL body parameters. The SMPL parameter vector for the j-th $(j = 1, \ldots, T)$ canonical phase is denoted by $\hat{\boldsymbol{\Theta}}'^j_i = [\hat{\boldsymbol{\beta}}'^j_i, \hat{\boldsymbol{\theta}}'^j_i, \hat{\boldsymbol{r}}'^j_i, \hat{\boldsymbol{\sigma}}'^j_i] \in \mathbb{R}^{85}$, where

$\hat{\boldsymbol{\beta}}'^{j}_{i}$, $\hat{\boldsymbol{\theta}}'^{j}_{i}$, $\hat{\boldsymbol{r}}'^{j}_{i}$, and $\hat{\boldsymbol{\sigma}}'^{j}_{i}$ are the interpolated shape, pose, global rotation, and camera parameters, respectively. In the SMPL model fitting process, the camera parameters used for 2D projection, that is, the scale and translation parameters, reflect the occluded body region. For example, the larger the occluded region, the larger the scale value; different occlusion positions (i.e., top and bottom occlusion) also result in different translation parameters. Therefore, we introduce the camera parameters as a cue for the occlusion patterns, and transform the shape and pose parameters $\hat{\boldsymbol{\Phi}}'^{j}_{i} = [\hat{\boldsymbol{\beta}}'^{j}_{i}, \hat{\boldsymbol{\theta}}'^{j}_{i}] \in \mathbb{R}^{79}$ to more occlusion-independent parameters as follows:

$$\tilde{\boldsymbol{\Phi}}^{j}_{i} = \hat{\boldsymbol{\Phi}}'^{j}_{i} + A([\hat{\boldsymbol{\Phi}}'^{j}_{i}, \hat{\boldsymbol{\sigma}}'^{j}_{i}]), \tag{7}$$

where A denotes the attenuation transformation, $\tilde{\boldsymbol{\Phi}}^{j}_{i} = [\tilde{\boldsymbol{\beta}}^{j}_{i}, \tilde{\boldsymbol{\theta}}^{j}_{i}] \in \mathbb{R}^{79}$, and $\tilde{\boldsymbol{\beta}}^{j}_{i} \in \mathbb{R}^{10}$ and $\tilde{\boldsymbol{\theta}}^{j}_{i} \in \mathbb{R}^{69}$ are the shape and pose parameters after transformation, respectively.

We implement the transformation via an FC layer, which learns the updates for the body parameters. To ensure the transformation successfully mitigates the intra-subject variation (i.e., occlusion dependence) of the SMPL model, we define a paired similarity loss as

$$L^{\text{pair}}_{\text{sim}} = \frac{1}{M}\sum_{m=1}^{M}\frac{1}{T|\mathcal{S}_m|(|\mathcal{S}_m|-1)}\sum_{S_i\in\mathcal{S}_m}\sum_{\substack{S_k\in\mathcal{S}_m\\ k\neq i}}\sum_{j=1}^{T}\|\tilde{\boldsymbol{\Phi}}^{j}_{i} - \tilde{\boldsymbol{\Phi}}^{j}_{k}\|^2_2, \tag{8}$$

where $\mathcal{S}_m (m = 1, \ldots, M)$ denotes the sequence set of the m-th subject, and M is the number of training subjects. We finally obtain a sequence of model parameters in the canonical phases: $\{\tilde{\boldsymbol{\Theta}}^{1}_{i}, \ldots \tilde{\boldsymbol{\Theta}}^{T}_{i}\}$, where $\tilde{\boldsymbol{\Theta}}^{j}_{i} = [\tilde{\boldsymbol{\beta}}^{j}_{i}, \tilde{\boldsymbol{\theta}}^{j}_{i}, \hat{\boldsymbol{r}}'^{j}_{i}, \hat{\boldsymbol{\sigma}}'^{j}_{i}] \in \mathbb{R}^{85}$.

E. SMPL Supervision

1) Sequential Constraint: Because the shape parameters should be temporally consistent within a sequence, we first apply average pooling to the shape parameters to obtain the unified shape of the i-th sequence: $\bar{\tilde{\boldsymbol{\beta}}}_{i} = \frac{1}{T}\sum_{j=1}^{T}\tilde{\boldsymbol{\beta}}^{j}_{i} \in \mathbb{R}^{10}$; hence, we replace the estimated parameters with $\tilde{\boldsymbol{\Theta}}^{j}_{i} = [\bar{\tilde{\boldsymbol{\beta}}}_{i}, \tilde{\boldsymbol{\theta}}^{j}_{i}, \hat{\boldsymbol{r}}'^{j}_{i}, \hat{\boldsymbol{\sigma}}'^{j}_{i}]$. To maintain the sequential property of the obtained model parameters, we further define a sequence loss L_{seq} similar to that in [18], which ensures the consistency of the shape parameters between same-subject sequences, in addition to the temporal continuity of the pose, global rotation, and camera parameters.

2) Supervision With Ground Truth: Ground-truth SMPL parameters are not provided in publicly available gait datasets. By contrast, considering that occluded gait images are generated by cropping the corresponding full-body images (i.e., images without occlusion) in the training stage, we may use the SMPL model parameters obtained by applying state-of-the-art model-based gait recognition [18] to full-body images as a pseudo ground-truth for supervision. We can also compute the pseudo ground-truth camera parameters for the occluded image by converting the parameters for the full-body image because the spatial location of the cropped area (i.e., input image) in the full-body image is known during the training phase.

Similar to the methodology in Sec. III-D, we first interpolate the ground-truth pose, global rotation, and camera parameters to the canonical phases, and compute the SMPL estimation loss as follows:

$$L^{\text{SMPL}}_{\text{esti}} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{T}\sum_{j=1}^{T}\|\tilde{\boldsymbol{\Theta}}^{j}_{i} - \boldsymbol{\Theta}^{j}_{i}\|^2_2, \tag{9}$$

where $\boldsymbol{\Theta}^{j}_{i}$ is the pseudo ground-truth SMPL parameters after phase synchronization.

To constrain the 3D joint locations computed from the estimated SMPL mesh vertices, we also define a joint estimation loss $L^{\text{joint}}_{\text{esti}}$ similar to Eq. (9).

We combine $L^{\text{SMPL}}_{\text{esti}}$ and $L^{\text{joint}}_{\text{esti}}$ as follows:

$$L_{\text{esti}} = w^{\text{SMPL}}_{\text{esti}}L^{\text{SMPL}}_{\text{esti}} + w^{\text{joint}}_{\text{esti}}L^{\text{joint}}_{\text{esti}}, \tag{10}$$

where $w^{\text{SMPL}}_{\text{esti}}$ and $w^{\text{joint}}_{\text{esti}}$ are the weight parameters.

F. Recognition Module

We then use the estimated shape and pose parameters for recognition. We directly exploit the unified shape $\bar{\tilde{\boldsymbol{\beta}}}_{i} \in \mathbb{R}^{10}$ as the shape feature of the i-th input sequence: $\boldsymbol{f}^{i}_{\text{shape}} = \bar{\tilde{\boldsymbol{\beta}}}_{i}$. Considering the insights mentioned in [55] that CNN is more suitable than long short-term memory (LSTM) for temporal data learning in gait recognition, and the superior performance of CNN compared to LSTM shown in [18], we also employ CNN for pose feature extration in the proposed method. Specifically, we input the 3D joint locations obtained from the estimated SMPL mesh into the CNN used in [18] to extract discriminative pose features $\boldsymbol{f}^{i}_{\text{pose}} = cnn(\tilde{X}_i) \in \mathbb{R}^{52}$, where $\tilde{X}_i = \{\tilde{\boldsymbol{x}}^{1}_{i}, \ldots, \tilde{\boldsymbol{x}}^{T}_{i}\}$ is the sequence of joint locations.

We finally use the shape/pose features to compute the triplet loss to optimize recognition performance, which is defined as

$$L_{\text{re}} = \frac{1}{N_{\text{trip}}}\sum_{n=1}^{N_{\text{trip}}}\max(\text{margin} + d^{\text{gen}}_{n} - d^{\text{imp}}_{n}, 0)^2, \tag{11}$$

where N_{trip} is the number of triplets in a mini-batch, and d^{gen}_{n} and d^{imp}_{n} are the dissimilarities of the genuine pair and imposter pair for the n-th triplet, respectively.

G. Unified Loss Function

To ensure a trade-off between model estimation and recognition accuracy, we use a unified loss to optimize the entire framework in an end-to-end manner, which is defined as

$$L_{\text{uni}} = w_{\text{ph}}L_{\text{ph}} + w^{\text{pair}}_{\text{sim}}L^{\text{pair}}_{\text{sim}} + w_{\text{seq}}L_{\text{seq}} + w_{\text{esti}}L_{\text{esti}} + w_{\text{re}}L_{\text{re}}, \tag{12}$$

where w_{ph}, $w^{\text{pair}}_{\text{sim}}$, w_{seq}, w_{esti}, and w_{re} are the weights for each term.

IV. Experiments

A. Data Preparation

There is no publicly available gait dataset that focuses on occlusion variations. Although occlusion is also considered in the GREW dataset [69], RGB data is not provided. In addition, it is unsuitable for systematic evaluation w.r.t. occlusion types, patterns, ratios, and views (i.e., labels are not provided). Therefore, similar to other gait recognition works that artificially simulate occlusion samples (e.g., [10], [11], [16]), we used the OU-MVLP dataset [24] to simulate several types of occlusion patterns that may often appear in real life. OU-MVLP is the world's largest gait dataset with a wide view variation, and contains 10,307 subjects, each captured from 14 views (0°–90° and 180°–270° in 15° intervals). Informed consent was obtained from each subject. We chose four typical views, 0°, 30°, 60°, and 90°, for both the training and test evaluation. Following [24], we used 5,153 subjects for training and the other disjoint 5,154 subjects for testing.

We mainly focused on occlusion in the vertical direction, which makes it impossible to fully observe the height of the human body. We considered two occlusion scenarios: fixed occlusion ratio and changing occlusion ratio in a sequence (i.e., the proportion of occluded body height changed). The fixed occlusion ratio simulates a scenario such as a subject in the side view being occluded by a relatively long obstacle (e.g., flower bed), or the front view being occluded by an object moving with the subject (e.g., a large suitcase). The changing occlusion ratio simulates a scenario such as a subject in the front view being occluded because of the limited camera field of view, or the side view being occluded by an obstacle at a certain angle to the walking direction (e.g., a billboard).

For simplicity, we set a rectangular occlusion region on the entire body, and used a square bounding box to crop the remaining unoccluded part as the input image (similar to Fig. 1), simulating the pedestrian detection results (e.g., YOLOv5 [70]) for occluded images in real scenes (i.e., cropped images for similar real and artificial occlusion effects are similar). We prepared four occlusion patterns accordingly: fixed occlusion ratio at the top (FT) and bottom (FB), and changing occlusion ratio at the top (CT) and bottom (CB) (see Fig. 4). For each occlusion pattern, we generated samples with three occlusion ratios. Specifically, for FT and FB, 20%, 40%, and 60% of the height was occluded in each frame; for CT, the occlusion ratio in a gait period gradually changed from 60% to 20%, 40% to 0% (no occlusion in the last frame), and 20% to 0% in the first half period and no occlusion in the second half period (denoted by 60%, 40%, and 20%, for simplicity), whereas the occlusion ratio for CB was the opposite.[1] In the training phase, we used all samples in different views, occlusion patterns, and ratios to train a unified model.

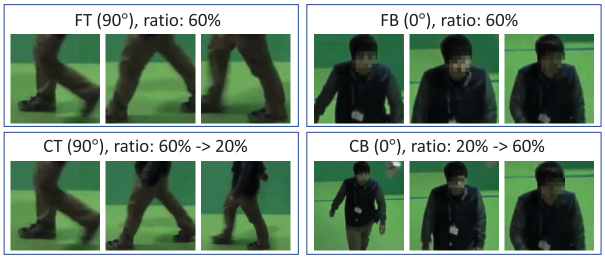

Fig. 4. Examples of occlusion patterns. Frames in FT and CT were simulated using a sequence from 90°, and the occlusion ratio was 60% and a variation from 60% to 20%, respectively. Frames in FB and CB were simulated using a sequence from 0°, and the occlusion ratio was set similar to the values used in FT and CT, respectively.

B. Implementation Details

We resized the cropped unoccluded regions to 224 × 224, and used 25 consecutive frames in a sequence as input,[2] which covers approximately one gait cycle for most subjects in OU-MVLP. We first trained the phase estimation network to obtain stable phase estimation results, and then included other parts to jointly train the entire framework in an end-to-end manner. We initialized the feature extractor E and regressor R using a pre-trained partial HMR [21]. We zero-initialized the GRU module G and the attenuation layer A to learn the updates of their respective input features. We initialized the recognition module with default parameters. We trained the network using the Adam optimizer [71], and set the batch size to 8 × 8, which represents eight subjects with eight samples per subject chosen as a mini-batch. We set the learning rate to 10^{-4} for the first 60K iterations, and decreased it by 0.1 for the last 70K iterations. We set the weight parameters in the loss functions to 1, except for $w_{\text{smoo}}^{\text{phase}} = 0.01$ and $w_{\text{penal}}^{\text{phase}} = 0.001$ in Eq. (6), and $w_{\text{sim}}^{\text{pair}} = 0.001$ and $w_{\text{esti}} = 100$ in Eq. (12). We set the margin in Eq. (11) to 0.2. Following [18], we separately trained and tested the shape and pose features for recognition.

We used the rank-1 identification rate and equal error rate (EER) [72] to evaluate recognition performance in identification and verification scenarios, respectively.

C. Visualization of SMPL Estimation

We chose samples with different occlusion patterns and views simulated based on sequences from the same test subject to visualize the SMPL models estimated by the proposed method. For comparison, we also show the SMPL estimated by the pre-trained ModelGait [18] in Fig. 5. Because ModelGait was originally designed for full-body images, the SMPL models estimated from the occluded images contained large errors (e.g., temporally discontinuous walking poses in Fig. 5(c)). By contrast, the body shape and pose estimated by the proposed method fit the input images well, which demonstrates its effectiveness in fitting the SMPL model to occlusion data. Although the estimation errors remained (e.g., stride in the double-support phase was smaller than the ground-truth), the

[1] Because the camera position for OU-MVLP was set relatively high (5 m), the occlusion ratio for CT and CB changed in opposite directions. For a fair comparison, we assumed that the occlusion ratio changed consistently in different views if the obstacles in different views had different angles from the viewing direction (e.g., in the front view, the angle between the obstacle and the viewing direction was perpendicular, and in the side view, the angle was almost parallel).

[2] We repeated frames from the beginning when the sequence was less than 25 frames.

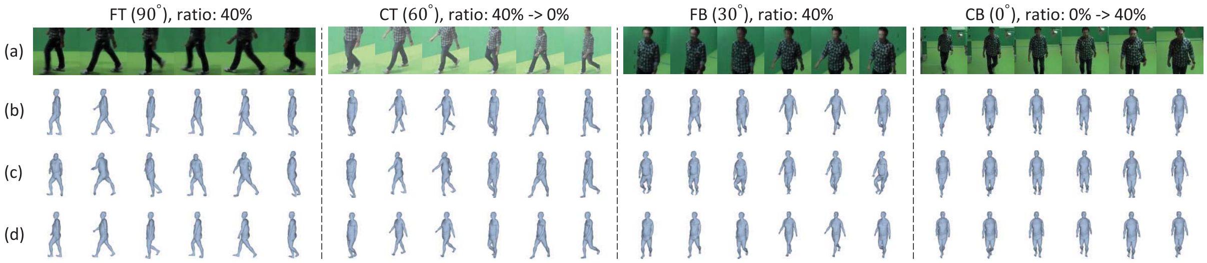

Fig. 5. Examples of the SMPL models estimated by the proposed method and ModelGait [18]. Four input samples with different occlusion patterns and different views were simulated based on sequences captured from the same test subject. The occlusion ratio in the four input samples was 40%. (a) Input sequence (interval of four frames). (b) Ground-truth SMPL of the corresponding full-body images. (c) SMPL estimated by ModelGait [18]. The SMPL models were projected using the camera parameters of the corresponding full-body images. (d) SMPL estimated by the proposed method. To keep the phases consistent with the input images in the visualization, the final obtained SMPL parameters were inversely interpolated into the original phases. See the supplementary material for more visualization examples.

TABLE I

Rank-1 Identification Rate [%] (Denoted by Rank-1) and EER [%] for Each Comparison Method in the Case of the Same Occlusion Pattern. We Report the Mean Results of Each Occlusion Pattern by Taking the Average of All 16 Combinations of the Four Occlusion Ratios (i.e., 0%, 20%, 40%, and 60%) in the Probe and Gallery. Note That for Each Occlusion Ratio Combination, We Averaged the Result Over All 16 Probe and Gallery View Pairs (i.e., View Pairs From 0°, 30°, 60°, and 90°). Bold and Bold Italic Indicate the Best and Second-Best Results

Methods	Rank-1 [%]					EER [%]				
	Occlusion pattern					Occlusion pattern				
	FT	CT	FB	CB	Mean	FT	CT	FB	CB	Mean
ModelGait (only test)	15.5	35.7	13.9	30.7	24.0	22.80	8.75	26.48	11.27	17.33
ModelGait (train&test)	22.8	43.2	27.5	49.7	35.8	6.11	3.75	6.07	3.53	4.86
GaitSet [44]	29.2	47.4	28.0	46.0	37.6	5.57	3.16	6.44	3.48	4.66
GaitGL [49]	36.6	52.7	***37.7***	***55.0***	45.5	5.18	3.17	6.10	3.30	4.44
Ours-shape	***44.1***	***69.0***	25.9	53.3	***48.1***	4.37	**0.83**	6.79	***1.93***	***3.48***
Ours-pose	36.2	52.8	29.2	46.6	41.2	***4.17***	2.49	***5.21***	3.10	3.75
Ours-fusion	**54.5**	**78.9**	**39.9**	**67.6**	**60.2**	**2.87**	**0.83**	**4.21**	**1.52**	**2.36**
ModelGait [18] (upper bound)	96.9					0.17				

shapes and poses between different occlusion patterns were more similar than ModelGait. This illustrates that the proposed method reduced the influence of occlusion on the model fitting results, to some extent.

D. Comparison With General Gait Recognition Methods

Unlike video-based person Re-ID, we do not use the color and texture information as appearance-based gait features; hence, video-based person Re-ID methods are beyond the scope of comparison, just as other gait recognition works using RGB inputs have done (e.g., GaitNet [45], [46], ModelGait [18]). However, no existing gait recognition method has addressed occlusion without a prerequisite. Therefore, we compared our method with two state-of-the-art appearance-based gait recognition methods, GaitSet [44] and GaitGL [49], and a state-of-the-art model-based method, ModelGait [18]. We retrained GaitSet, GaitGL, and ModelGait using the same training data as our method for comparison. We also provided the testing results directly using the pretrained ModelGait. Because we trained the proposed method separately using the shape and pose features, we also applied score-level fusion similar to that used in [18]. As a reference, we showed a type of upper bound accuracy, that is, the recognition results on the full-body images provided by ModelGait in [18].

1) Same Occlusion Pattern: The results in the same occlusion pattern case (i.e., the occlusion patterns in the probe and gallery were the same) are shown in Table I. Although ModelGait [18] achieved prominent recognition performance on full-body images, it did not perform well on occluded images, which is consistent with the qualitative comparison in Sec. IV-C. The proposed method also achieved significantly better performance than GaitSet [44] and GaitGL [49], which demonstrates the superiority of the proposed occlusion-aware framework in handling occlusion without a prerequisite. The shape features performed better than the pose features, except for the FB pattern, and their fusion achieved the best results for both identification and verification scenarios, outperforming the benchmarks by about 15% for the rank-1 rate and 2% for the EER, on average.

Among the four occlusion patterns, the performance of patterns with a changing occlusion ratio (i.e., CT and CB) was relatively higher than that of the fixed occlusion ratio (i.e., FT and FB). This is understandable because the sequence in CT and CB patterns contained frames with smaller occlusion ratios than the corresponding frames in the FT and FB patterns. Most methods performed worse on the bottom occlusion than the top occlusion, particularly for the FB pattern. This demonstrates that both the shape and pose features of the lower body were more important for gait recognition.

Table II shows the results for the CT occlusion pattern using the proposed method with the fusion scheme. Essentially, the results under the same view (i.e., 0° view difference) were much better than those under different views. As the occlusion ratio difference increased, accuracy gradually decreased, particularly for the largest occlusion ratio difference (i.e., 60%),

TABLE II

Mean Rank-1 Rate (Before Slash) and EER (After Slash) [%] of the Proposed Method (Fusion) for the CT Occlusion Pattern. The Results Are Computed for Each Occlusion Ratio Difference and Each View Difference Between the Probe and Gallery. R and V Denote the Occlusion Ratio Difference and the View Difference, Respectively

V / R	0°	30°	60°	90°	Mean
0%	96.7 / 0.43	90.2 / 0.54	80.7 / 0.78	70.6 / 1.06	87.0 / 0.64
20%	93.4 / 0.53	85.3 / 0.65	75.5 / 0.90	65.8 / 1.21	82.4 / 0.75
40%	86.2 / 0.69	76.4 / 0.85	66.3 / 1.11	57.0 / 1.45	73.9 / 0.95
60%	75.2 / 0.92	63.5 / 1.12	54.1 / 1.42	45.3 / 1.83	61.8 / 1.23
Mean	90.2 / 0.59	81.6 / 0.73	71.8 / 0.99	62.2 / 1.31	78.9 / 0.83

TABLE III

Rank-1 Rate (Before Slash) and EER (After Slash) [%] of the Comparison Methods in the Cross-Occlusion Pattern Case. We Considered Three Specific Probe and Gallery Occlusion Ratio Pairs for Evaluation. We Obtained the Mean Result of Each Occlusion Ratio Pair by Averaging All 16 Combinations of the Four Occlusion Patterns and All 16 View Combinations. Probe and Gallery Are Denoted by P and G, Respectively

Occlusion ratio pair	GaitSet [44]	GaitGL [49]	Ours-fusion
P: 20%, G: 40%	35.8 / 4.38	43.0 / 4.38	**63.0 / 1.91**
P: 40%, G: 40%	28.3 / 5.80	32.9 / 6.13	**54.0 / 2.61**
P: 60%, G: 40%	14.3 / 9.45	18.8 / 9.63	**27.1 / 6.11**
Mean	26.2 / 6.55	31.6 / 6.71	**48.0 / 3.54**

TABLE IV

Mean Rank-1 Rate (Before Slash) and EER (After Slash) [%] of the Proposed Method (Fusion) for Each Combination of the Four Occlusion Patterns. The Probe and Gallery Both Had an Occlusion Ratio of 40%. Each Result Was the Average of 16 Probe and Gallery View Pairs

G / P	FT	CT	FB	CB
FT	84.0 / 0.80	80.2 / 0.96	13.5 / 6.65	51.1 / 2.02
CT	79.2 / 0.96	87.1 / 0.68	23.1 / 4.83	69.9 / 1.21
FB	10.3 / 6.58	18.4 / 4.73	62.6 / 2.29	40.0 / 2.85
CB	48.2 / 2.01	68.2 / 1.22	45.5 / 2.97	82.4 / 0.97

where the loss of a large amount of the individual body shape and pose information caused by considerable occlusion highly affected the SMPL model fitting and recognition performance. Additionally, the view difference also had a great impact on performance. This is because we trained a unified model that considered multiple occlusion patterns, occlusion ratios, and views, which significantly increased the training difficulty.

2) Cross-Occlusion Pattern: We then compared the proposed method (fusion), GaitSet, and GaitGL in the cross-occlusion pattern case (i.e., the occlusion patterns in the probe and gallery were different). As shown in Table III, the proposed method clearly outperformed other methods for all three occlusion ratio pairs. We report the results of the proposed method for each combination of the four occlusion patterns when the probe and gallery had the same occlusion ratio (i.e., 40%) in Table IV. Specifically, the performance was better when the occlusion patterns in the probe and gallery were the same. For the different pattern case, the recognition results between similar occlusion positions (e.g., FT vs. CT) were relatively better. Additionally, the performance of the occlusion pattern pair with a changing occlusion ratio (i.e., CT vs. CB) was also higher than a fixed occlusion ratio.

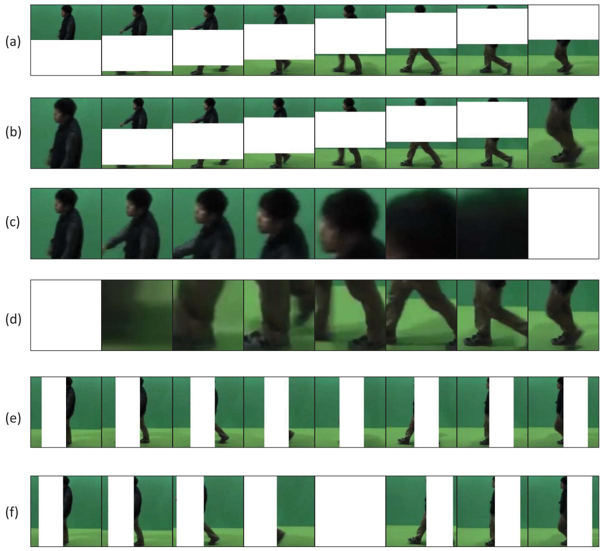

Fig. 6. Illustration of different bounding box settings for experiments in Sec. IV-E. Here, an occlusion degree of 50% defined in [11] was taken as an example. (a) Full-body bounding box used in [11] for RDBT pattern. (b) Visible-part bounding box with a single body fragment used by our method for RDBT pattern. Person tracking was assumed to work well, which results in visible upper and lower body parts still considered to be in the same bounding box. The first/last frame contains only visible upper/lower body fragment. (c) Visible-part bounding box with upper-body fragment used by our method for RDBT pattern. (d) Visible-part bounding box with lower-body fragment used by our method for RDBT pattern. For (c) and (d), we assumed the visible upper and lower body parts were tracked with two different bounding boxes, and used each fragment separately for recognition. (e) Full-body bounding box used in [11] for RDLR pattern. (f) Visible-part bounding box with a single body fragment used by our method for RDLR pattern. Visible left and right parts were assumed to be successfully tracked as belonging to the same person, with a temporally complete sequence. All-white images indicate the body fragment was fully occluded.

E. Comparison With Other Occlusion-Handling Gait Recognition Work

Since other works on gait recognition against occlusion require prerequisites of full-body bounding boxes, we compared with a CNN-based occlusion-handling method [11] by following their settings as much as possible. Specifically, in the case of vertical occlusion pattern, namely relative dynamic occlusion gradually moves from the bottom to top (RDBT), we used a visible-part bounding box while assuming that person tracking worked perfectly under occlusion. In this case, as shown in Fig. 6(b), the visible upper and lower body parts were still considered to belong to the same person, which resulted in only a single body fragment; hence, the differences from the full-body bounding box setting in Fig. 6(a) are just the first and last frames.

In the case of horizontal occlusion pattern, that is, relative dynamic occlusion gradually moves from the left to right (RDLR), we also considered a similar setting to the RDBT pattern for our method, i.e., visible left and right parts were considered to be the same person being successfully tracked, resulting in a temporally complete sequence. Compared to the full-body bounding box used in [11] (see Fig. 6(e)), in our setting (see Fig. 6(f)), the human scale and body center computed based only on the visible parts changed in a sequence. Moreover, the body may be fully occluded in several middle frames of a sequence (e.g., 5 or 6 middle frames may be completely occluded with a defined occlusion degree of 50%[3]).

Additionally, we also considered a more challenging setting of the visible-part bounding box for the RDBT pattern in Fig. 6(c) and (d), i.e., the visible upper and lower body parts were tracked into independent sequences, which splitted the human region into two fragments, each of which was used for recognition separately. In fact, this occlusion setting is even more difficult than that in Sec. IV-D, because with the original setting of 50% occlusion degree in [11], the actual occlusion ratio in each upper/lower-body fragment changed between 50% and 100%. Unlike the vertical occlusion pattern, the setting of visible-part bounding box with only left or right body fragments (i.e., assuming the visible left and right parts were tracked into two independent sequences) would result in each independent sequence containing less than half a gait cycle of the original sequence (e.g., about 10 frames per left/right fragment sequence under 50% occlusion degree). Since the phase synchronization used in the currently proposed method may fail if less than half a gait cycle is available, we did not consider this setting for the horizontal occlusion pattern.

Following the protocol in [11], 3,000 subjects from OU-MVLP were used for training and another disjoint 3,000 subjects were used for testing, while both the training and testing sets contained only samples from the side view. Three occlusion degrees were considered, i.e., 30%, 40%, and 50%. Since there were much fewer training samples compared to the experiments in Sec. IV-D, we reduced the number of training iterations to 20K.

The comparison results are shown in Table V. Under similar bounding box settings for the RDBT pattern (i.e., only the first and last frames are different), the proposed method achieved significantly better performance than the previous work [11]. The proposed method also outperformed even under a more challenging setting, where only the visible upper/lower fragment was used each time. This illustrates the superiority of the proposed model-based method in vertical occlusion handling for gait recognition. Furthermore, using lower-body fragment again obtained better results than upper-body fragment, which is consistent with the findings in Sec. IV-D, demonstrating the importance of the lower body for gait recognition.

In addition, the proposed method also clearly outperformed [11] for the horizontal occlusion pattern. Even if several middle frames were fully occluded in this case, the proposed method still worked thanks to the phase synchronization process (i.e., interpolation of missing phases from other visible frames), as well as the modules (i.e., GRU) and constraints for maintaining the temporal continuity and consistency of the estimated models within a sequence. Therefore, the proposed method can also handle horizontal occlusions if effective tracking results are obtained in the pre-processing.

TABLE V

RANK-1 RATE, RANK-5 RATE, AND EER [%] OF EACH COMPARISON METHOD AND SETTING. SINGLE, UPPER, AND LOWER FRAG. REPRESENT A VISIBLE-PART BOX WITH A SINGLE-BODY, UPPER-BODY, AND LOWER-BODY FRAGMENT, RESPECTIVELY. THE DIGITS AFTER THE OCCLUSION PATTERNS ARE THE OCCLUSION DEGREES AS DEFINED IN [11]

(a) RDBT occlusion pattern

Methods	RDBT_30 vs. RDBT_30			RDBT_50 vs. RDBT_50		
	Rank-1	Rank-5	EER	Rank-1	Rank-5	EER
Uddin's [11]	82.5	89.0	5.9	77.3	86.3	6.2
Ours (single frag.)	**99.3**	**99.8**	**0.1**	**98.7**	**99.9**	**0.1**
Ours (upper frag.)	90.4	97.9	0.6	83.3	95.6	1.0
Ours (lower frag.)	96.5	99.5	0.3	89.8	98.1	0.6

(b) RDLR occlusion pattern

Methods	RDLR_30 vs. RDLR_30			RDLR_50 vs. RDLR_50		
	Rank-1	Rank-5	EER	Rank-1	Rank-5	EER
Uddin's [11]	82.4	88.6	6.0	75.9	85.4	6.6
Ours (single frag.)	**98.7**	**100**	**0.1**	**93.6**	**99.6**	**0.3**

F. Ablation Study

We analyze the effects of individual components in Table VI. More specifically, we consider the GRU module in the sequence encoder, and the phase synchronizer and attenuation module in the occlusion attenuation framework. Because the attenuation transformation A cannot exist independently of the phase synchronization process, we did not use it when turning off the phase synchronizer. Because phase synchronization mainly affects the pose features, we used the pose features for the ablation experiments. Furthermore, we explored the impact of model initialization for SMPL regression, i.e., replacing the initialization of the feature extractor E and regressor R from pre-trained partial HMR [21] to standard HMR [62].

Based on the results, the entire proposed method performed better than the ablative methods, which demonstrates that all components contributed to the proposed method. For example, if the occlusion attenuation module was excluded, the average rank-1 identification rate decreased from 41.2% to 32.5%, which indicates the effectiveness of the module for occlusion handling. On the other hand, although changing the initialization for SMPL regression slightly degraded the performance, the effects of initialization was smaller compared to other proposed network components.

V. DISCUSSION

A. Validating Color Effects in Model Fitting

Different from video-based occluded person Re-ID methods (e.g., [73], [74], [75]) that directly encode color and texture as discriminative appearance features, gait recognition excludes color and texture features that are subject to clothes change [45]. Therefore, to validate the invariance to color and texture changes of the human model fitting in the proposed framework, we compared the estimated SMPL models between the original input and the corresponding color-retouched images. Specifically, we manually blurred the

[3] The input image size used in [11] is 128 × 88, and the size of occluded region is 128 × 44 (i.e., 50% of the image width is occluded). Because our input image size is 224 × 224, to make the comparison as fair as possible, we kept the same occluded areas relative to the human body in our inputs. Although the occlusion visually appeared to be less than 50% of the entire image width, for most frames the occluded body part was greater than 50% horizontally.

TABLE VI

MEAN RANK-1 RATE [%] OF THE PROPOSED METHOD WITH THE POSE FEATURES FOR ABLATION EXPERIMENTS IN THE SAME OCCLUSION PATTERN CASE. THE RESULT FOR EACH PATTERN WAS AVERAGED OVER 16 COMBINATIONS OF THE FOUR OCCLUSION RATIOS AND 16 VIEW COMBINATIONS. E, R, P, AND A DENOTE THE FEATURE EXTRACTOR, REGRESSOR, PHASE ESTIMATION NETWORK, AND ATTENUATION TRANSFORMATION, RESPECTIVELY

Sequence encoder		Occlusion attenuation		Occlusion pattern				
E & R init.	GRU	P+sync.	A	FT	CT	FB	CB	Mean
Partial HMR [21]	×	√	√	***34.7***	51.4	***27.3***	42.9	39.1
Partial HMR [21]	√	×	×	25.9	41.5	23.6	39.2	32.5
Partial HMR [21]	√	√	×	34.0	50.3	26.0	43.0	38.3
HMR [62]	√	√	√	***34.7***	***52.1***	***27.3***	***45.0***	***39.8***
Partial HMR [21]	√	√	√	**36.2**	**52.8**	**29.2**	**46.6**	**41.2**

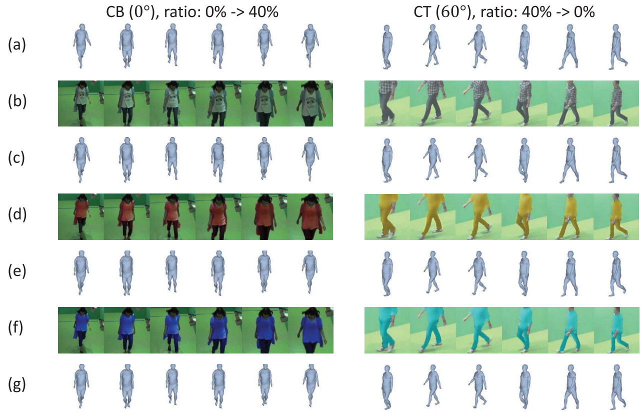

Fig. 7. Examples of the SMPL models estimated from color-retouched images. (a) The ground-truth SMPL of the original input images. (b) Original input images. (c) SMPL estimated from (b). (d) Color-retouched images of (b). (e) SMPL estimated from (d). (f) Another set of color-retouched images of (b). (g) SMPL estimated from (f). The estimated SMPL models are similar between (c), (e), and (g).

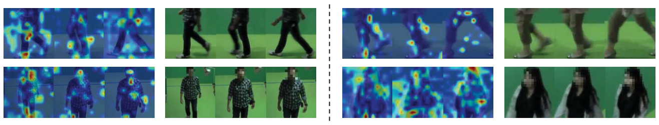

Fig. 8. Grad-CAM visualizations on input images and the corresponding raw input images. The top two examples are top occlusions, and the bottom two are bottom occlusions.

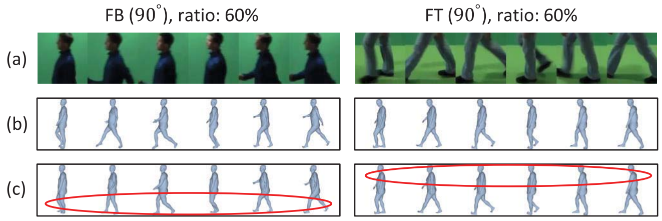

Fig. 9. Two failure examples of the proposed method. Left: input samples with the FB occlusion pattern; right: input samples with the FT occlusion pattern. The occlusion ratio in the two input samples was 60%. (a) Frames in the input sequence. (b) The ground-truth SMPL of the corresponding full-body images. (c) SMPL estimated by the proposed method. Erroneous parts are highlighted by red circles.

original clothes pattern, and painted the clothes with another color using image editing software. We experimented with several colors different from the background to enable the modeling fitting framework to distinguish between human and background regions. On the other hand, model estimation may fail if the clothes are of the same color as the background, which is a challenging case not only for model-based methods, but also for appearance-based tasks (e.g., segmentation step in it).

As shown in Fig. 7, the body shape and pose estimated from the color-retouched images are similar to the original images, demonstrating the robustness of the human model fitting to color and texture changes. Although estimation differences (e.g., slight smaller stride for the color-retouched images) still exist to some extent, this is because the color was retouched by manual painting, which may also inevitably change the background color around the body contour and further affect the model estimation. By contrast, the estimation results are almost identical between two different color-retouched images (Figs. 7(e) and (g)). This illustrates that changes in color and texture do not affect the model fitting framework, and thus, the gait features extracted from the estimated human model do not contain color and texture information, which is different from the person Re-ID works.

B. Analysis of Learned Features

To analyze where the proposed method learns useful features for model fitting, we used a visualization method, Gradient-weighted Class Activation Mapping (Grad-CAM) [76], to show important regions in input images in Fig. 8. Based on the heat maps, for top occlusion, our model focuses more on stride and joints (e.g., knees, ankles, crotch, and visible elbows) to predict possible upper body poses (e.g., arm swings and back bends), and may use the shape of legs (e.g., fat or thin, leg length) to estimate the upper body shape (e.g., fat or thin, height). Similarly, for bottom occlusion, the shape and some joints of the upper body (e.g., shoulders, elbows, torso shape, and visible thighs) may be more useful for predicting the lower body shape and pose. In addition, faces are also often in focus when bottom occlusion occurs. This is to help estimate the full-body model including the head part, which may be oriented differently than the torso (e.g., looking to the other side while walking forward).

C. Limitations

Figure 9 shows some typical failure examples of the proposed method, which were chosen from the challenging occlusion cases, that is, fixed occlusion pattern with a large occlusion ratio (i.e., 60%). Compared with the ground-truth, the estimated SMPL models contained some errors. For example, for the FB pattern, the estimated stride lengths in the double-support phases were smaller than those shown in the ground-truth; for the FT pattern, the estimated upper body was relatively straight, while the ground-truth upper body was bent forward. This illustrates that it is still difficult to well estimate a full human body model just from relatively small non-occluded body parts, whereas subtle individual pose and shape characteristics are important cues for gait recognition. Considering that we used direct shape parameters and pose features extracted by a simple CNN for recognition in the proposed method, one possible way to improve the recognition performance is to incorporate more effective feature extraction that takes into account possible estimation errors in occluded body regions (e.g., attention mechanism), aiming to improve the occlusion invariance of the extracted shape and pose features. Additionally, as mentioned in Sec. IV-D, the unified model involving multiple occlusion patterns, occlusion ratios, and views greatly increased the training difficulty. Considering that the results between patterns with similar occlusion positions (i.e., top/bottom) were better than the results between the top and bottom patterns, one possible solution is to train a model for top and bottom occlusion separately, combined with simple preprocessing of occlusion position detection.

In gait biomechanics studies [77], [78], it has been demonstrated that there is a flexible neuronal coupling between upper and lower limb muscles during human walking; however, it remains difficult to precisely define such neuronal connections. Therefore, it becomes more difficult for human model estimation from occluded gait video, which estimates full-body models based only on captured upper-/lower-body images rather than precise muscle activation data (e.g., electromyographic responses) used in biomechanics works. On the other hand, this also implies a potential future direction for incorporating gait biomechanics to improve performance.

Currently, 25 consecutive frames are taken as input to the proposed network. If the input 25 frame-sequence is a few frames (e.g., 5 or 6) less than a full gait cycle, the proposed method still works thanks to the phase synchronization process, where missing phases can be interpolated based on existing frames in the input, as well as the GRU module and the temporal continuity and consistency constraints in the supervision. On the other hand, if the input is less than half a gait cycle, phase interpolation may not work well due to large temporal gap between starting and ending frames, which is a limitation of the currently proposed method. In the future, to mitigate this problem, we may consider incorporating the idea of reconstructing a full gait cycle from limited frames, which is often done in low frame-rate gait recognition [79], [80] and single-image gait recognition [66].

VI. Conclusion

In this paper, we proposed an occlusion-aware model-based gait recognition method to handle occlusion without a prerequisite. Given an occluded gait sequence, we estimate the SMPL models directly from the input images by incorporating the occlusion attenuation module, and further use the models to extract the shape and pose features for the recognition task. Experiments on simulated occlusion illustrated the effectiveness of the proposed method.

In the future, a more effective feature extraction module is worth investigating to gain more robustness against occlusion. Additionally, while we focused on artificially simulated occlusion samples in this study, we will conduct experiments with more realistic scenes after collecting sufficient data. Considering real occlusion may come from various obstacles of various shapes and colors, we can combine pedestrian detection with object detection works as preprocessing, which helps locate occluded regions within human bounding boxes to mitigate the impact of the complex real-world scenes on the proposed method including human model estimation (e.g., painting the detected occlusion region with a regular shape and a color distinct from the body).

References

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

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

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

[4] Y. Guan and C.-T. Li, "A robust speed-invariant gait recognition system for walker and runner identification," in *Proc. Int. Conf. Biometrics (ICB)*, Jun. 2013, pp. 1–8.

[5] C. Xu, Y. Makihara, X. Li, Y. Yagi, and J. Lu, "Speed invariance vs. stability: Cross-speed gait recognition using single-support gait energy image," in *Proc. 13th Asian Conf. Comput. Vis. (ACCV)*, Taipei, Taiwan, Nov. 2016, pp. 52–67.

[6] Y. Makihara, R. Sagawa, Y. Mukaigawa, T. Echigo, and Y. Yagi, "Gait recognition using a view transformation model in the frequency domain," in *Proc. 9th Eur. Conf. Comput. Vis.*, Graz, Austria, May 2006, pp. 151–163.

[7] Z. Wu, Y. Huang, L. Wang, X. Wang, and T. Tan, "A comprehensive study on cross-view gait based human identification with deep CNNs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 2, pp. 209–226, Feb. 2016.

[8] Y. Makihara, A. Suzuki, D. Muramatsu, X. Li, and Y. Yagi, "Joint intensity and spatial metric learning for robust gait recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6786–6796.

[9] X. Li, Y. Makihara, C. Xu, Y. Yagi, and M. Ren, "Gait recognition via semi-supervised disentangled representation learning to identity and covariate features," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13309–13319.

[10] D. Muramatsu, Y. Makihara, and Y. Yagi, "Gait regeneration for recognition," in *Proc. Int. Conf. Biometrics (ICB)*, May 2015, pp. 169–176.

[11] M. Z. Uddin, D. Muramatsu, N. Takemura, M. A. R. Ahad, and Y. Yagi, "Spatio-temporal silhouette sequence reconstruction for gait recognition against occlusion," *IPSJ Trans. Comput. Vis. Appl.*, vol. 11, no. 1, pp. 1–18, Dec. 2019.

[12] A. Roy, S. Sural, J. Mukherjee, and G. Rigoll, "Occlusion detection and gait silhouette reconstruction from degraded scenes," *Signal, Image Video Process.*, vol. 5, no. 4, pp. 415–430, Nov. 2011.

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

[14] G. Zhao, L. Cui, and H. Li, "Gait recognition using fractal scale," *Pattern Anal. Appl.*, vol. 10, no. 3, pp. 235–246, Jul. 2007.

[15] J. Ortells, R. A. Mollineda, B. Mederos, and R. Martín-Félez, "Gait recognition from corrupted silhouettes: A robust statistical approach," *Mach. Vis. Appl.*, vol. 28, nos. 1–2, pp. 15–33, Feb. 2017.

[16] P. Nangtin, P. Kumhom, and K. Chamnongthai, "Gait identification with partial occlusion using six modules and consideration of occluded module exclusion," *J. Vis. Commun. Image Represent.*, vol. 36, pp. 107–121, Apr. 2016, doi: 10.1016/j.jvcir.2016.01.008.

[17] Y. Iwashita, K. Uchino, and R. Kurazume, "Gait-based person identification robust to changes in appearance," *Sensors*, vol. 13, no. 6, pp. 7884–7901, 2013. [Online]. Available: https://www.mdpi.com/1424-8220/13/6/7884

[18] X. Li, Y. Makihara, C. Xu, Y. Yagi, S. Yu, and M. Ren, "End-to-end model-based gait recognition," in *Proc. Asian Conf. Comput. Vis. (ACCV)*, Nov. 2020, pp. 1–17.

[19] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A skinned multi-person linear model," *ACM Trans. Graph.*, vol. 34, no. 6, pp. 1–16, Nov. 2015, doi: 10.1145/2816795.2818013.

[20] T. Zhang, B. Huang, and Y. Wang, "Object-occluded human shape and pose estimation from a single color image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7376–7385.

[21] C. Rockwell and D. F. Fouhey, "Full-body awareness from partial observations," in *Computer Vision—ECCV*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham, Switzerland: Springer, 2020, pp. 522–539.

[22] K. Lin, L. Wang, and Z. Liu, "End-to-end human pose and mesh reconstruction with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 1954–1963.

[23] K. Cho et al., "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1–15.

[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. Comput. Vis. Appl.*, vol. 10, no. 1, pp. 1–14, 2018.

[25] S. Yu, D. Tan, K. Huang, and T. Tan, "Reducing the effect of noise on human contour in gait recognition," in *Advances in Biometrics*, S.-W. Lee and S. Z. Li, Eds. Berlin, Germany: Springer, 2007, pp. 338–346.

[26] C. Chen, J. Liang, H. Zhao, H. Hu, and J. Tian, "Frame difference energy image for gait recognition with incomplete silhouettes," *Pattern Recognit. Lett.*, vol. 30, no. 11, pp. 977–984, Aug. 2009. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167865509000920

[27] N. V. Boulgouris and Z. X. Chi, "Human gait recognition based on matching of body components," *Pattern Recognit.*, vol. 40, no. 6, pp. 1763–1770, 2007. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320306004857

[28] T. Whytock, A. Belyaev, and N. M. Robertson, "On covariate factor detection and removal for robust gait recognition," *Mach. Vis. Appl.*, vol. 26, no. 5, pp. 661–674, Jul. 2015, doi: 10.1007/s00138-015-0681-2.

[29] S. H. Shaikh, K. Saeed, and N. Chaki, "Gait recognition using partial silhouette-based approach," in *Proc. Int. Conf. Signal Process. Integr. Netw. (SPIN)*, Feb. 2014, pp. 101–106.

[30] X. Chen, J. Weng, W. Lu, and J. Xu, "Multi-gait recognition based on attribute discovery," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 7, pp. 1697–1710, Jul. 2018.

[31] M. Hofmann, D. Wolf, and G. Rigoll, "Identification and reconstruction of complete gait cycles for person identification in crowded scenes," in *Proc. Int. Conf. Comput. Vis. Theory Appl. (VISAPP)*, Jan. 2011, pp. 594–597.

[32] A. Tsuji, Y. Makihara, and Y. Yagi, "Silhouette transformation based on walking speed for gait identification," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Francisco, CA, USA, Jun. 2010, pp. 717–722.

[33] W. Kusakunniran, Q. Wu, J. Zhang, and H. Li, "Gait recognition across various walking speeds using higher order shape configuration based on a differential composition model," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 42, no. 6, pp. 1654–1668, Dec. 2012.

[34] K. Bashir, T. Xiang, and S. Gong, "Gait recognition using gait entropy image," in *Proc. Int. Conf. Crime Detection Prevention*, Dec. 2009, pp. 1–6.

[35] M. A. Hossain, Y. Makihara, J. Wang, and Y. Yagi, "Clothing-invariant gait identification using part-based clothing categorization and adaptive weight control," *Pattern Recognit.*, vol. 43, no. 6, pp. 2281–2291, Jun. 2010.

[36] X. Li, Y. Makihara, C. Xu, D. Muramatsu, Y. Yagi, and M. Ren, "Gait energy response function for clothing-invariant gait recognition," in *Proc. 13th Asian Conf. Comput. Vis. (ACCV)*, Taipei, Taiwan, Nov. 2016, pp. 257–272.

[37] X. Li, Y. Makihara, C. Xu, Y. Yagi, and M. Ren, "Joint intensity transformer network for gait recognition robust against clothing and carrying status," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 12, pp. 3102–3115, Dec. 2019.

[38] X. Li, Y. Makihara, C. Xu, Y. Yagi, and M. Ren, "Gait recognition invariant to carried objects using alpha blending generative adversarial networks," *Pattern Recognit.*, vol. 105, Sep. 2020, Art. no. 107376. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320320301795

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

[40] S. Yu, H. Chen, E. B. G. Reyes, and N. Poh, "GaitGAN: Invariant gait feature extraction using generative adversarial networks," in *Proc. CVPR Workshops*, Jul. 2017, pp. 532–539.

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

[42] Y. He, J. Zhang, H. Shan, and L. Wang, "Multi-task GANs for view-specific feature learning in gait recognition," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 1, pp. 102–113, Jan. 2019.

[43] C. Xu, Y. Makihara, X. Li, Y. Yagi, and J. Lu, "Cross-view gait recognition using pairwise spatial transformer networks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 1, pp. 260–274, Jan. 2021.

[44] H. Chao, Y. He, J. Zhang, and J. Feng, "Gaitset: Regarding gait as a set for cross-view gait recognition," in *Proc. 33th AAAI Conf. Artif. Intell. (AAAI)*, 2019, pp. 8126–8133.

[45] Z. Zhang et al., "Gait recognition via disentangled representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4705–4714.

[46] Z. Zhang, L. Tran, F. Liu, and X. Liu, "On learning disentangled representations for gait recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 345–360, Jan. 2022.

[47] C. Fan et al., "GaitPart: Temporal part-based model for gait recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 14213–14221.

[48] S. Hou, C. Cao, X. Liu, and Y. Huang, "Gait lateral network: Learning discriminative and compact representations for gait recognition," in *Computer Vision—ECCV*. Berlin, Germany: Springer-Verlag, Aug. 2020, pp. 382–398, doi: 10.1007/978-3-030-58545-7_22.

[49] B. Lin, S. Zhang, and X. Yu, "Gait recognition via effective global-local feature representation and local temporal aggregation," in *Proc. ICCV*, Oct. 2021, pp. 14648–14656.

[50] X. Huang et al., "Context-sensitive temporal feature learning for gait recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 12889–12898.

[51] Z. Huang et al., "3D local convolutional neural networks for gait recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 14900–14909.

[52] T. Chai, A. Li, S. Zhang, Z. Li, and Y. Wang, "Lagrange motion analysis and view embeddings for improved gait recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 20217–20226.

[53] J. Liang, C. Fan, S. Hou, C. Shen, Y. Huang, and S. Yu, "Gaitedge: Beyond plain end-to-end gait recognition for better practicality," in *Computer Vision—ECCV*. Berlin, Germany: Springer-Verlag, Oct. 2022, pp. 375–390, doi: 10.1007/978-3-031-20065-6_22.

[54] R. Liao, C. Cao, E. B. Garcia, S. Yu, and Y. Huang, "Pose-based temporal-spatial network (PTSN) for gait recognition with carrying and clothing variations," in *Biometric Recognition*, J. Zhou, Y. Wang, Z. Sun, Y. Xu, L. Shen, J. Feng, S. Shan, Y. Qiao, Z. Guo, and S. Yu, Eds. Cham, Switzerland: Springer, 2017, pp. 474–483.

[55] R. Liao, S. Yu, W. An, and Y. Huang, "A model-based gait recognition method with body pose and human prior knowledge," *Pattern Recognit.*, vol. 98, p. 107069, Feb. 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S003132031930370X

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

[57] T. Teepe, A. Khan, J. Gilg, F. Herzog, S. Hörmann, and G. Rigoll, "Gaitgraph: Graph convolutional network for skeleton-based gait recognition," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2021, pp. 2314–2318.

[58] T. Teepe, J. Gilg, F. Herzog, S. Hörmann, and G. Rigoll, "Towards a deeper understanding of skeleton-based gait recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops*, Jun. 2022, pp. 1569–1577.

[59] L. Wang, J. Chen, and Y. Liu, "Frame-level refinement networks for skeleton-based gait recognition," *Comput. Vis. Image Understand.*, vol. 222, Sep. 2022, Art. no. 103500. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1077314222000972

[60] X. Li, Y. Makihara, C. Xu, and Y. Yagi, "Multi-view large population gait database with human meshes and its performance evaluation," *IEEE Trans. Biometrics, Behav., Identity Sci.*, vol. 4, no. 2, pp. 234–248, Apr. 2022.

[61] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh, "OpenPose: Realtime multi-person 2D pose estimation using part affinity fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, pp. 172–186, 2021.

[62] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik, "End-to-end recovery of human shape and pose," in *Proc. Comput. Vis. Pattern Regognition (CVPR)*, 2018, pp. 1–10.

[63] Y. Cheng, B. Yang, B. Wang, Y. Wending, and R. Tan, "Occlusion-aware networks for 3D human pose estimation in video," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 723–732.

[64] M. Kocabas, N. Athanasiou, and M. J. Black, "VIBE: Video inference for human body pose and shape estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5253–5263.

[65] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Computer Vision—ECCV*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham, Switzerland: Springer, 2016, pp. 630–645.

[66] C. Xu, Y. Makihara, X. Li, Y. Yagi, and J. Lu, "Gait recognition from a single image using a phase-aware gait cycle reconstruction network," in *Proc. 16th Eur. Conf. Comput. Vis. (ECCV)*, Aug. 2020, pp. 386–403.

[67] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015, *arXiv:1502.03167*.

[68] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. 27th Int. Conf. Int. Conf. Mach. Learn.* Madison, WI, USA: Omnipress, 2010, pp. 807–814. [Online]. Available: http://dl.acm.org/citation.cfm?id=3104322.3104425

[69] Z. Zhu et al., "Gait recognition in the wild: A benchmark," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 14789–14799.

[70] G. Jocher. (2020). *YOLOV5*. [Online]. Available: https://github.com/ultralytics/yolov5

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

[72] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The FERET evaluation methodology for face-recognition algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 10, pp. 1090–1104, Oct. 2000.

[73] P. Chen et al., "Occlude them all: Occlusion-aware attention network for occluded person Re-ID," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 11833–11842.

[74] C. Yan, G. Pang, J. Jiao, X. Bai, X. Feng, and C. Shen, "Occluded person re-identification with single-scale global representations," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 11875–11884.

[75] S. Gao, J. Wang, H. Lu, and Z. Liu, "Pose-guided visible part matching for occluded person ReID," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11744–11752.

[76] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

[77] V. Dietz, K. Fouad, and C. M. Bastiaanse, "Neuronal coordination of arm and leg movements during human locomotion," *Eur. J. Neurosci.*, vol. 14, no. 11, pp. 1906–1914, Dec. 2001. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1046/j.0953-816x.2001.01813.x

[78] J. L. Stephenson, S. J. D. Serres, and A. Lamontagne, "The effect of arm movements on the lower limb during gait after a stroke," *Gait Posture*, vol. 31, no. 1, pp. 109–115, 2010.

[79] Y. Makihara, A. Mori, and Y. Yagi, "Temporal super resolution from a single quasi-periodic image sequence based on phase registration," in *Proc. 10th Asian Conf. Comput. Vis.*, Queenstown, New Zealand, Nov. 2010, pp. 107–120.

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

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.

Yasushi Makihara received the B.S., M.S., and Ph.D. degrees in engineering from Osaka University in 2001, 2002, and 2005, respectively. 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, Osaka University, in 2005, 2006, and 2014, respectively. He is 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 is a member of IPSJ, IEICE, RSJ, and JSME. He has obtained several honors and awards, including the Second International Workshop on Biometrics and Forensics in 2014, the IAPR Best Paper Award, the Ninth 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 the 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 Fourth 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.

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 a Specially Appointed Assistant Professor. His research interests are computer vision, image processing, and gait recognition.

Yasushi Yagi (Senior Member, IEEE) received the Ph.D. degree from Osaka University in 1991. He is currently 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 is a fellow of IPSJ and a member of IEICE and RSJ. 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. His international conferences for which he has served as the Chair include: FG1998 (the Financial Chair), OMINVIS2003 (the Organizing Chair), ROBIO2006 (the Program Co-Chair), ACCV2007 (the Program Chair), PSVIT2009 (the Financial Chair), ICRA2009 (the Technical Visit Chair), ACCV2009 (the General Chair), ACPR2011 (the Program Co-Chair), and ACPR2013 (the 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*.