Occluded Gait Recognition via Silhouette Registration Guided by Automated Occlusion Degree Estimation

Chi Xu Shogo Tsuji Yasushi Makihara Xiang Li Yasushi Yagi
Osaka University, Osaka, Japan
{xu, tsuji, makihara, li, yagi}@am.sanken.osaka-u.ac.jp

Abstract

Gait recognition tasks often face significant difficulties caused by partial occlusions of the human body. To address this challenge, we propose a silhouette registration method based on flexible estimation of the spatial scale associated with the occluding elements. Existing appearance-based methods require prior knowledge about the spatial scale of the human body in relation to the input image, or a bounding box that includes the actual full body. In our method, the region corresponding to the silhouette of the body is estimated directly from visible body parts within the image. This estimate is then used to normalize and register the human body by adapting it to the scale of the occlusions. To reduce the occlusion difference between elements of a matching pair, which may lead to substantial intra-subject variation when the difference is large, we use a pairwise mask to extract common visible regions for subsequent feature learning and matching. Experiments on the synthetic occluded OU-MVLP dataset demonstrate the effectiveness of the proposed method, which successfully improves recognition performance when matching pairs present occlusion differences. We discuss specific characteristics of the proposed silhouette registration and pairwise masking methods with the aid of detailed quantitative and qualitative evaluations, in the hope of providing useful insights for future research on this topic.

1. Introduction

Human gait is a behavioral biometric characteristic associated with human walking. Even without subject cooperation, it can be authenticated from low-resolution images captured from distant vantage points [39]. As a consequence, gait recognition has become increasingly popular in diverse applications such as surveillance, criminal investigation, and forensics [2, 21, 33]. While readily available from camera footage, gait recognition also presents important challenges: factors such as viewing angle [34, 53], clothing [35, 27], walking speed [13, 52], and occlusions [48, 51] may greatly affect recognition performance in real-world applications.

Partial occlusion of the human body is often seen in captured gait videos. Occlusion may be caused by obstacles such as trees and buildings, or by disappearance of body parts as they move out of view. To reduce the impact of partial occlusion on feature extraction, most existing occluded gait recognition methods seek to reconstruct silhouettes of the full body from occluded images [41, 36, 48, 10], or to extract relatively occlusion-invariant features through metric learning techniques [23, 38, 40]. However, these methods rely on easily accessible knowledge of how human height relates to the input images, or bounding boxes for the full body. They use this information to normalize and register body sizes and full-body centers within silhouette sequences before carrying out subsequent steps such as feature extraction and matching, however this information is typically unavailable in real-world occlusion scenarios. For example, the normalized and registered silhouettes shown in Fig. 1(a, I) are generated by first localizing the top of the head, which is not visible in typical occlusion scenarios. As shown in Fig. 1(a, II), if the bounding box only contains the visible region, as is often the case for pedestrian detection on occluded gait videos, the human size and body center position may vary greatly with the degree of occlusion, greatly degrading gait recognition.

The above issues also apply to recent state-of-the-art appearance-based gait recognition methods, such as GaitGL [31]: recognition accuracy drops severely without silhouette registration, especially when the degree of occlusion varies greatly between probe and gallery (see green data point corresponding to 60% occlusion in Fig. 1(b)). When silhouettes are pre-registered with full-body bounding boxes, recognition accuracy is greatly improved by the normalization of body size and position throughout the matching sequences (see red line in Fig. 1(b)).

To address the above challenges, we tackle the occlusion problem in gait recognition by applying body registration to partially occluded silhouettes prior to feature extraction. Instead of relying on full-body bounding boxes, which are often unavailable in real-world situations, we only consider

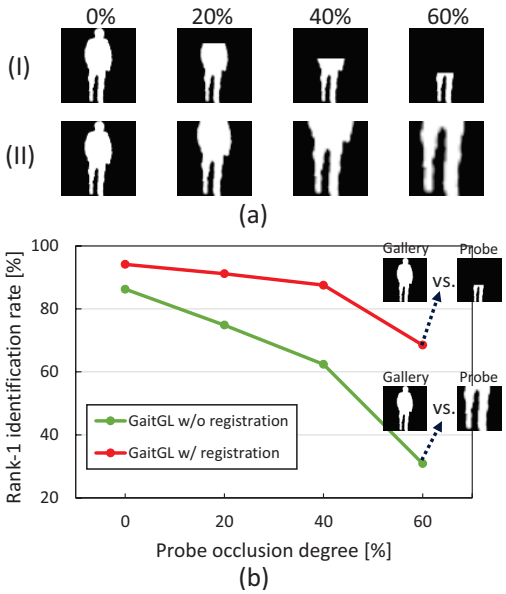

Figure 1: (a) Occluded silhouettes obtained with and without registration with full-body bounding boxes. Top digit labels indicate the degree of occlusion (as % of body height) for (I) silhouettes registered using full-body bounding boxes and (II) silhouettes of visible parts without full-body registration. Human size and body center position vary with degree of occlusion. (b) Rank-1 identification rates of GaitGL [31] applied to the synthetic occluded OU-MVLP dataset [45], with (red) and without (green) silhouette registration. Both probes and galleries were in front view (similar to (a)). Galleries consisted of samples without occlusion (0% occlusion), while probes varied between 0% and 60% occlusion. Silhouette registration significantly improves robustness against occlusion, especially when occlusion is strong (60%).

the bounding boxes associated with visible parts (similar to Fig. 1(a, II)) and use them to estimate the degree of occlusion in the form of an occlusion ratio (ratio between occluded parts and whole body). Our study makes two important contributions, which we discuss separately below.

1. Silhouette registration under occlusion without prior knowledge of full-body bounding boxes

Different from existing appearance-based methods, which require prior knowledge about body size within images, we directly input silhouettes for visible body parts only, which is closer to real application scenarios for captured occluded gait videos. After registering the silhouette based on the estimated occlusion ratio, we use a pairwise mask to pick corresponding visible regions between elements of a matching pair, further improving performance for pairs with different degrees of occlusion.

2. Performance analysis and discussion supported by both quantitative and qualitative evaluations

We evaluated performance on the synthetic occluded OU-MVLP dataset [45], which contains a large number of samples covering a wide range of views and degrees of synthetic occlusion. Compared with state-of-the-art methods, our proposed approach substantially improves recognition performance for matching pairs with occlusion ratio variations, especially in the case of large variations. We also conducted detailed quantitative and qualitative evaluations to analyze the impact of the proposed silhouette registration and pairwise masking methods, alongside a discussion of possible future directions for this research effort.

2. Related Work

2.1. Occluded gait recognition

Existing approaches for occluded gait recognition rely primarily on appearance-based features, and roughly fall into two categories: reconstruction-based and reconstruction-free approaches. Reconstruction-based methods involve reconstruction of full-body silhouettes [16, 41, 48, 10] or gait feature templates [36] from occluded samples, and use the reconstructed images for subsequent feature learning. Compared to approaches based on traditional machine learning [41, 36], recent generative network-based methods [48, 10] yield better reconstruction and recognition performance. Reconstruction-free methods [54, 58, 9, 40] use metric learning techniques to directly extract relatively occlusion-invariant features from occluded silhouettes, without the need for full explicit reconstruction as performed by reconstruction-based methods. To reduce intra-subject variations between matching pairs, methods based on body partition were proposed for matching similar non-occluded regions [3, 23, 50, 38].

A major drawback of the above methods is that they all assume knowledge of the full-body bounding box even under partial occlusions, to prevent performance degradation caused by variations in body size and position. The model-based approach OA-ModelGait [51] represents an exception: this method directly fits a skinned multi-person linear (SMPL) [32] model to the input RGB image containing only visible body parts. Although OA-ModelGait achieved state-of-the-art recognition performance under occlusion, its accuracy degraded with increasing degree of occlusion as a consequence of difficulties with model fitting, especially for cases involving severe occlusion.

2.2. General gait recognition

Instead of using gait feature templates such as gait energy image (GEI) [14], recent gait recognition methods rely on deep networks to directly operate on silhouettes [7, 12, 17, 31, 18, 19, 5, 11] for more effective gait recognition that is robust to various covariates. For example, GaitSet [7] achieved excellent performance by regarding

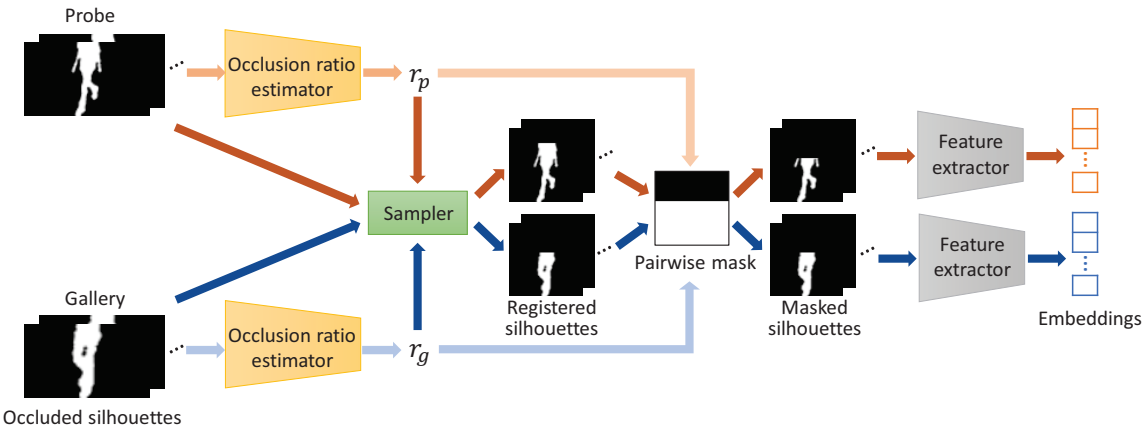

Figure 2: Overview of the proposed method, which involves an occlusion ratio estimator, a sampler, a pairwise mask, and a feature extractor.

the input sequence as a set, leading to further research on set-based methods [17, 8]. GaitGL [31] used 3D CNNs to aggregate temporal features, and combined these networks with an ensemble framework to exploit both global and local gait features. Besides silhouettes, RGB images were also explored as input to the networks, and used for feature extraction after excluding useless color information [57, 56, 29].

Model-based methods are also developing rapidly. Methods based on CNNs [30, 1] or graph convolutional networks [47, 46, 49] often utilize skeleton joints obtained from pose estimation methods (such as OpenPose [4]) for pose feature learning. SMPL-based methods [28, 26] estimate SMPL models from input RGBs, and perform recognition using both shape and pose features from the estimated models.

The appearance-based methods mentioned above require body size and position registration for all sequences, regardless of whether their input consists of silhouettes or RGB images. Model-based methods can handle diverse inputs more flexibly, however model estimation errors for unregistered inputs may increase intra-subject variability.

2.3. Silhouette registration for gait recognition

Silhouette registration was introduced during the early stages of research on appearance-based gait recognition [42]. After segmenting the foreground region into a silhouette, this region is scaled to a fixed size and its center is shifted horizontally to keep it consistent across all frames, based on the notion that scaling and shifting provide size invariance and compensate for placements errors of bounding boxes. When generating commonly used gait feature representations, such as GEI [14] and frequency-domain features [34], the above normalization and registration procedures are regarded as necessary preprocessing steps for raw silhouettes. For these reasons, appearance-based gait studies typically start with normalized and registered silhouettes that are available from widely used gait databases, such as CASIA-B [55] and OULP [22]. A recent end-to-end gait recognition network [29], which took RGB images as input, also included a differentiable size normalization module for handling synthesized silhouettes from the inputs.

Based on the above considerations, we propose that registration and normalization are also important for silhouettes under occlusion, and we introduce methods for performing these procedures while retaining applicability to realistic occlusion scenarios.

3. Proposed method

3.1. Overview

Figure 2 summarizes our proposed method. As a starting point, we focus on upper body occlusion, where the upper body is partly occluded by obstacles or goes out of view of the camera.

Given an occluded silhouette, represented as a bounding box containing only visible regions, its degree of occlusion is quantified by an occlusion ratio estimator. The registered silhouette is then obtained using a differentiable sampler, which transforms the occluded silhouette into normalized body size and center position based on the estimated occlusion ratio. To reduce differences in the extent of occlusion between a pair of probe and gallery samples, each pair of registered silhouettes is pairwise masked to retain the common visible region. The resulting masked silhouettes are fed to a gait recognition network that extracts features for matching.

3.2. Occlusion ratio estimator

Occlusion ratio is a continuous label denoting the proportion of occluded parts relative to body height. Given input occluded silhouette $I_{i,j}(i = 1, \ldots, N, j = 1, \ldots, M)$,

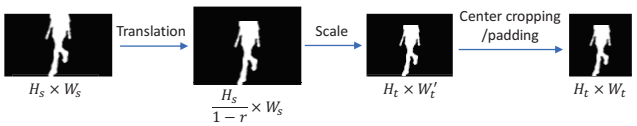

Figure 3: Image registration operated by the sampler. Input image size is fixed at $H_s \times W_s$, and changes to $H_t \times W_t'$ after affine transformation, where W_t' is subject to occlusion ratio r. The final registered image is obtained by center cropping or padding, with a fixed size of $H_t \times W_t$.

where N is the number of sequences in a mini-batch and M is the number of frames in a sequence, the occlusion ratio estimator E outputs the occlusion ratio as

$$\hat{r}_{i,j} = E(I_{i,j}). \tag{1}$$

Estimator E consists of three convolutional layers and two fully connected layers. Each convolutional filter takes size 5×5, and the number of filters is increased from 32 to 128. Each convolutional layer is followed by a ReLU activation function [37], a batch normalization layer [20], and a max pooling operation with kernel size 2×2. We then use a fully connected layer with 128 output neurons, a ReLU activation function [37], and another fully connected layer to regress the one-dimensional occlusion ratio.

Assuming that a given video sequence only contains a relatively short fragment of walking action (e.g., one gait cycle), and that occlusion remains relatively stable for the duration of the sequence, we average estimates across frames to obtain an occlusion ratio for the entire sequence, thus mitigating potentially large frame-level estimation errors. This procedure can be formalized as

$$\bar{\hat{r}}_i = \frac{1}{M} \sum_{j=1}^{M} \hat{r}_{i,j}, \tag{2}$$

where $\bar{\hat{r}}_i$ is the estimated occlusion ratio of the i-th sequence.

The estimator operates under the supervision of a mean squared error (MSE) loss, defined as

$$L_{\text{MSE}} = \frac{1}{N} \sum_{i=1}^{N} (\bar{\hat{r}}_i - r_i)^2, \tag{3}$$

where r_i is the ground truth occlusion ratio of the i-th sequence.

3.3. Silhouette registration

Following the above procedures, the occluded silhouette is registered according to the estimated occlusion ratio: body size and center position are normalized, and the occluded region is replaced by black pixels (see Fig. 3). To achieve differentiable registration, we used a sampler derived from the spatial transformer network [24] that registers images via affine transformation.

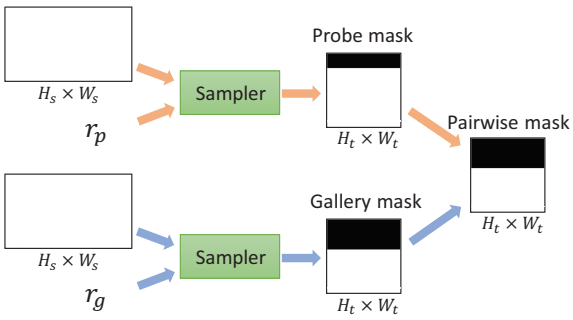

Figure 4: Generation of the pairwise mask. All-white images indicate the template mask with all-white pixels. The transformation in Fig. 3 is also applied to the sampler here.

Given input silhouette $I \in \mathbb{R}^{H_s \times W_s}$, where H_s and W_s are image height and width and r is occlusion ratio, we first applied a downward translation to reveal the occluded body parts while maintaining their aspect ratio. This operation is implemented by translation matrix A_{trans}, defined as

$$A_{\text{trans}} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & \frac{rH_s}{1-r} \\ 0 & 0 & 1 \end{bmatrix}. \tag{4}$$

The upper black region in Fig. 3 represents the occluded body part. After translation, image height increases to $\frac{H_s}{1-r}$.

The translated image is then scaled to a fixed image height H_t, where the scaling matrix A_{scale} is defined as

$$A_{\text{scale}} = \begin{bmatrix} \frac{(1-r)H_t}{H_s} & 0 & 0 \\ 0 & \frac{(1-r)H_t}{H_s} & 0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{5}$$

The affine transformation applied to the input image can therefore be computed as the product of the above translation and scaling matrices:

$$A_{\text{aff}} = A_{\text{scale}} A_{\text{trans}} = \begin{bmatrix} \frac{(1-r)H_t}{H_s} & 0 & 0 \\ 0 & \frac{(1-r)H_t}{H_s} & rH_t \\ 0 & 0 & 1 \end{bmatrix}, \tag{6}$$

where A_{aff} is the affine transformation matrix.

In practical implementations, instead of applying a forward transformation from the source image to the target image, the reverse operation is typically applied to transform the target image into the source image, so as to facilitate the calculation of pixel correspondence [6]. This operation is implemented by the inverse affine transformation matrix, defined as

$$A_{\text{aff}}^{-1} = \begin{bmatrix} \frac{H_s}{(1-r)H_t} & 0 & 0 \\ 0 & \frac{H_s}{(1-r)H_t} & -\frac{rH_s}{1-r} \\ 0 & 0 & 1 \end{bmatrix}. \tag{7}$$

Because the width W_t' of the resulting transformed image is subject to input occlusion ratio r, we generate a registered image of fixed width W_t by simultaneously cropping or padding both sides of the image. This operation is performed to retain the alignment between body center and horizontal image center (see Fig. 3).

3.4. Pairwise masking

Although body size and position are normalized after registration, there may be residual variations within the compensated occluded region between elements of a matching pair. These variations are caused by differences in occlusion ratio, and may lead to intra-subject variation in the subsequently extracted gait features. To mitigate these effects, we applied a pairwise mask to exclude maximally occluded regions, while retaining common visible parts for each matching pair.

Similar to silhouette registration, we use the sampler to retain differentiability of the masking operation. Given an all-white template of size $H_s \times W_s$, and occlusion ratios r_p and r_g for probe and gallery, we first generated two individual masks of size $H_t \times W_t$ using the transformation described in Sec. 3.3 (see Fig. 4). We then obtained a pairwise mask by performing the element-wise product between individual mask images, corresponding to the mask with larger occlusion ratio.

3.5. Feature extraction

After pairwise masking each frame of the two sequences from a matching pair, we employ a state-of-the-art silhouette-based gait recognition network (GaitGL [31]) as final feature extractor. Each masked probe and gallery sequence is individually fed to GaitGL, to output gait features to be used for matching during the testing phase. During the training phase, feature learning is supervised using batch-all triplet loss [15] and cross-entropy loss, similar to the original work [31].

4. Experiments

4.1. Dataset

We simulated occluded gait samples using OU-MVLP [45], a large-scale wide-view gait dataset. In this dataset, gait sequences of 10,307 subjects were captured across 14 views in the ranges 0°–90° and 180°–270° with an interval of 15°. For fair comparison with existing works (e.g., [51]), we selected the same four views: 0°, 30°, 60°, and 90°. We considered four degrees of occlusion for each sequence: 0% (no occlusion), 20%, 40%, and 60% (see Fig. 5 for examples). Following the official protocol [45], we used 5,153 subjects for training and 5,154 disjoint subjects for testing. For each subject, the sequence labeled "00" was used as probe, and "01" was used as gallery.

4.2. Implementation details

To fully preserve visible body parts after transformation and cropping using the sampler, we set the input image to a relatively large dimension ($H_s \times W_s = 64 \times 110$). We set the registered image size to $H_t \times W_t = 64 \times 64$, the number of frames per input sequence to 30, and the mini-batch size to 8×16 (8 subjects with 16 sequences per subject). We used Adam optimizer [25] for network training, and set the initial learning rate to 10^{-4}. Because the estimation of occlusion ratio has a large impact on silhouette registration and feature extraction, we first trained the occlusion ratio estimator for 30 epochs while reducing the learning rate by 0.1 at 20 epochs. We then trained the feature extractor using the ground truth occlusion ratio, and followed the settings of the original GaitGL algorithm [31]: we reduced the learning rate by 0.1 at 150K and 200K iterations, for a total of 210K iterations. We used samples from all four views and four degrees of occlusion to train a single model.

To improve the robustness of the feature extractor to errors in occlusion ratio estimation, we implemented data augmentation by simulating a maximum $\pm 2\%$ estimation error during training. Taking an input with a ground truth occlusion ratio of 20% as an example, we prepared five augmented registration samples corresponding to estimates ranging from 18% to 22%. Because the operation of the pairwise mask is based on the maximum occlusion ratio within a mini-batch, we randomly selected a ratio as the maximum for each mini-batch, rather than simply performing random sampling. We adopted this procedure to balance the amount of training data across different occlusion ratios.

4.3. Accuracy of occlusion ratio estimation

Because silhouette registration relies on occlusion ratio estimates, we first evaluate the accuracy of ratio estimation for the test set using the mean absolute error (MAE), defined as $\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N} |\hat{r}_i - r_i|$, where N is the number of test sequences, and $\hat{r}_i$ and r_i are the estimated and ground truth occlusion ratios of the i-th test sequence, respectively.

Table 1 shows that errors in occlusion ratio estimation are more pronounced for larger occlusions (40% and 60%). However, the mean estimation error is at worst only 0.14%, and overall MAE across all test samples is 0.07%. With a target image size of 64 pixels following registration, Eq. (6) indicates that estimation errors below 1.56% do not produce appreciable effects on silhouette registration. As shown in Fig. 5, the registered silhouettes based on the estimated occlusion ratios match the corresponding ground truth images. After registration, different body scales and center positions across different inputs are successfully aligned (see Fig. 5(a)), demonstrating that the occlusion ratio estimator provides reliable estimates for subsequent silhouette registration.

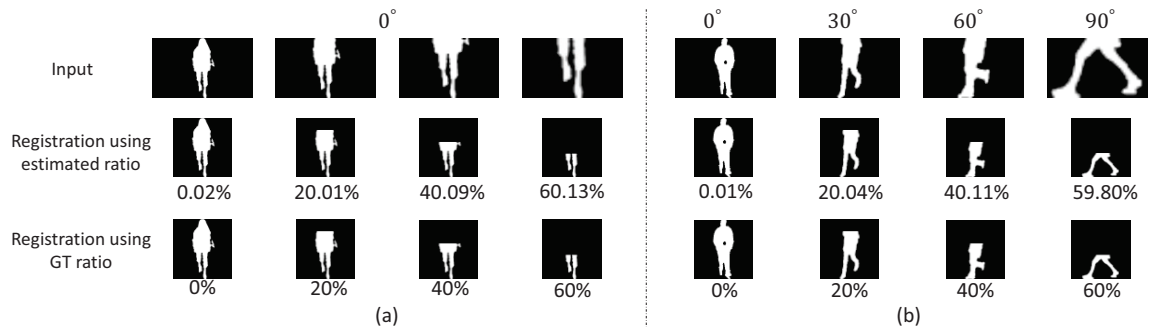

Figure 5: Examples of silhouette registration. (a) Samples from the same subject at 0°. (b) Samples from different subjects at 0°, 30°, 60°, and 90° (different columns from left to right). Top row shows input frames, middle and bottom rows show registered silhouettes based on estimated and ground truth occlusion ratios, respectively. Because estimation errors were small, middle and bottom rows are virtually identical.

Table 1: MAEs [%] of test samples computed for each combination of ground truth occlusion ratios (R) and views (V).

R \ V	0°	30°	60°	90°
0%	0.02	0.02	0.02	0.02
20%	0.05	0.05	0.05	0.06
40%	0.14	0.12	0.11	0.10
60%	0.11	0.10	0.09	0.09

4.4. Comparison with state-of-the-art algorithms

We compared the recognition performance achieved by our algorithm with results obtained using state-of-the-art appearance-based gait recognition methods. More specifically, we focused on GaitSet [7] and GaitGL [31], where GaitGL represents the backbone of our proposed method (without silhouette registration). We also included OA-ModelGait [51] for comparison. We evaluated recognition performance using the rank-1 identification rate and equal error rate (EER) for identification and verification tasks, respectively.

Table 2 shows overall comparison results. The proposed method significantly outperformed appearance-based methods, demonstrating the effectiveness of the silhouette registration procedure for handling occlusions. Compared with OA-ModelGait, the proposed method achieved better rank-1 rate but somewhat worse EER. With relation to this apparently inconsistent result, it should be noted that inconsistent performance trends often occur between identification and verification tasks in gait recognition, as a consequence of different computational criteria [44].

To inspect our results in greater detail, we report rank-1 rates and EERs of the proposed method against OA-ModelGait. These metrics were computed for each occlusion ratio difference and view difference between probe and gallery sets. As shown in Table 3, OA-ModelGait achieved better performance when matching pairs present equal or smaller occlusion differences, however our proposed method produced superior performance for large occlusion ratio differences, especially 60%. EER values for large ratio differences under small view variations (0° and 30°) were reduced in the verification task too, demonstrating that the proposed silhouette registration method successfully improved recognition performance for large occlusion variations.

Table 2: Mean rank-1 rate and EER [%] for each method. Results were averaged over all 4×4 combinations of occlusion ratios across probe and gallery matching pairs, and for each occlusion ratio pair, we took the average of all 4×4 view combinations across probe and gallery pairs. Bold and italic bold denote best and second-best results, respectively.

Methods	Rank-1	EER
GaitSet	54.0	2.75
GaitGL	56.9	2.74
OA-ModelGait	***71.6***	**1.01**
Proposed	**73.6**	***1.45***

Our method is primarily designed to target issues connected with occlusion, rather than addressing view variations. Because OA-ModelGait reconstructs 3D human models from input RGBs, which are more informative than 2D silhouettes, it presents intrinsic advantages over appearance-based methods when dealing with cross-view matching [51]. We discuss this issue in more detail in Sec. 5.

4.5. Ablation study

In this section, we validate the effectiveness of each component of the proposed method, except for the inevitable gait feature extractor component. We focus on the occlusion ratio estimator, the sampler-based registration operation, and pairwise masking, which constitute the entire proposed silhouette registration process. The first row in Table 4 lists baseline results obtained without silhouette registration, which are equivalent to those associated with GaitGL in Table 2. The second row details the impact of pairwise masking. Because the registration operation is essential to the proposed method, we cannot selectively exclude it for analysis. The third row presents results obtained using the proposed method. Because the registration operation depends on occlusion ratio estimation, we report the results obtained using ground truth

Table 3: Mean rank-1 rate [%] and EER [%] of the proposed method versus OA-ModelGait. Results were computed for each occlusion ratio difference and each view difference between probe and gallery sets. For example, the 40% occlusion ratio difference (Rd) included probe and gallery ratio pairs of 0% versus 40%, 20% versus 60%, 40% versus 0%, and 60% versus 20%, similar to the view difference (Vd) calculation.

(a) Rank-1 rates [%] of the proposed method

Rd \ Vd	0°	30°	60°	90°	Mean
0%	91.5	81.1	70.1	59.9	78.3
20%	**89.9**	78.3	66.7	55.8	75.5
40%	**87.7**	**74.3**	**62.5**	50.6	**71.7**
60%	**82.5**	**64.2**	**51.9**	**37.9**	**62.4**
Mean	**88.8**	**76.2**	64.7	53.3	**73.6**

(b) Rank-1 rates [%] of OA-ModelGait

Rd \ Vd	0°	30°	60°	90°	Mean
0%	**98.0**	**90.0**	**81.4**	**70.9**	**87.5**
20%	86.8	**78.9**	**71.5**	**62.8**	**77.0**
40%	74.2	65.2	58.4	**51.2**	64.0
60%	46.7	39.0	34.5	29.9	38.7
Mean	81.4	73.3	**66.1**	**57.8**	71.6

(c) EERs [%] of the proposed method

Rd \ Vd	0°	30°	60°	90°	Mean
0%	0.69	1.14	1.64	1.95	1.25
20%	0.77	1.26	1.79	2.11	1.37
40%	**0.86**	1.40	1.99	2.36	1.53
60%	**1.07**	**1.75**	2.50	2.99	**1.92**
Mean	0.81	1.32	1.89	2.24	1.45

(d) EERs [%] of OA-ModelGait

Rd \ Vd	0°	30°	60°	90°	Mean
0%	**0.24**	**0.42**	**0.64**	**0.93**	**0.49**
20%	**0.60**	**0.76**	**0.94**	**1.15**	**0.81**
40%	1.00	**1.18**	**1.36**	**1.62**	**1.24**
60%	1.95	2.16	**2.39**	**2.62**	2.22
Mean	**0.78**	**0.95**	**1.15**	**1.40**	**1.01**

occlusion ratio labels in the last row, which represents an upper bound for the proposed method.

From the above results, it is evident that the proposed method works best when all components are included. Overall performance when using ground truth labels is equivalent to the performance obtained using estimates, which demonstrates that the impact of ratio estimation errors on silhouette registration is sufficiently small, consistent with the results in Sec. 4.3.

To visualize the effects of pairwise masking on subsequent feature learning, we used Gradient-weighted Class Activation Mapping [43] (Grad-CAM) to highlight relevant regions from which features were learned. Surprisingly, the resulting visualizations in Fig. 6 show that, while the proposed method learned features from the entire visible body

Table 4: Mean rank-1 rate [%] and EER [%] for ablation experiments. Results were averaged over all 16 occlusion ratio combinations and 16 view pairs. GT denotes ground truth.

Occlusion ratio estimator	Registration operation	Pairwise masking	Rank-1	EER
×	×	×	56.9	2.74
Estimated	√	×	70.0	1.59
Estimated	√	√	**73.6**	**1.45**
GT	√	√	**73.6**	**1.45**

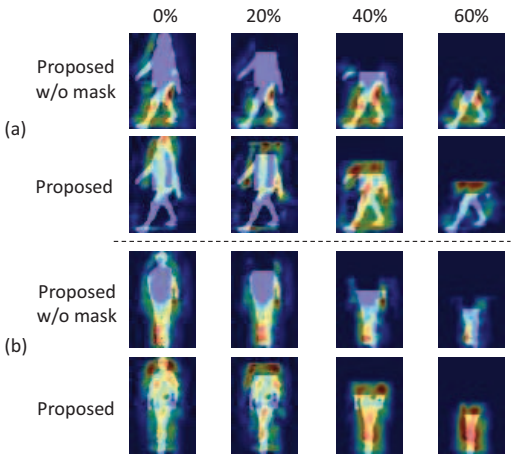

Figure 6: Feature visualization using Grad-CAM [43] for (a) samples from 60° and (b) from 0°. For both (a) and (b), first and second rows shows heat maps learned with the proposed method without and with pairwise masking, respectively. Occlusion ratio (%) increases from left to right.

parts, ablation without masking focused mainly on the lower legs, even though the upper body is largely unaffected when occlusion ratio is small. Taking into account that we only experimented with upper body occlusions, it appears that the absence of pairwise masking prompts the network to apply an implicit mask to extract features only from common visible regions across all samples (the lower 40%).

In realistic scenarios, occlusions may impact different body parts, making it difficult to obtain common visible regions for all data. For example, the upper body may be occluded in a given matching pair, while the lower body may be occluded in another pair. Therefore, although excluding pairwise masking from the proposed method did not severely reduce performance in our experiments (Table 4), this result may not extend to multiple occlusion types given their impact on feature learning. Under these more challenging conditions, we hypothesize that our proposed method may still be able to extract reasonable features by fully exploiting the visible region for each matching pair.

5. Limitations and open issues

In this study we only focus on upper body occlusion, but we hope to incorporate other types of body part occlusion

(alone or in combination) in future work. Under those conditions, an occlusion ratio estimator with a one-dimensional ratio output is inadequate. To address this limitation, we plan to extend the regression output to multiple dimensions, where each dimension represents a ratio label for a specific region (such as upper, lower, left, and right), and the sampler operation is adjusted accordingly based on the location of the occlusions.

While the proposed method achieved significant performance improvements for matching pairs with large occlusion ratio differences under small view variations (Sec. 4.4), the performance difference with respect to OA-ModelGait is relatively small in the case of large view variations (90°) as a consequence of the inherent advantages in cross-view gait recognition associated with 3D human models. Although OA-ModelGait suffered from severe model estimation errors at large occlusions, it nevertheless achieved impressive recognition performance when matching pairs present equally large occlusion ratios (40% versus 40% and 60 % versus 60%; see Table 5). In light of these results, we plan to incorporate OA-ModelGait into the proposed registration method. In this approach, input RGB pairs would be initially registered to the same human scales and visible regions, before applying model fitting.

Unlike binary (black versus white) silhouettes, which present consistently sharp boundaries between foreground and background, RGB images with various backgrounds in which occluded regions are simply replaced by a fixed color at registration may contain boundary gaps, which may in turn affect model fitting. This issue may be addressed by incorporating instance segmentation, which may mitigate boundary differences by excluding the original background.

Another limitation of our study is the representation of occlusion strength using a continuous ratio label, which is easy to estimate for simulated rectangular occlusions. Real scenes, however, contain complex occlusions with irregular shapes (such as regions occluded by trees or shrubs), which are difficult to characterize using simple ratios. Because large-scale occluded gait datasets are lacking, existing occluded gait recognition studies mainly experimented with synthetic occlusion data [51, 10, 48, 36, 9]. The applicability of these results to realistic occlusion scenarios remains an open issue that needs to be validated and investigated in future studies. Furthermore, in the case of matching pairs with different occlusion locations (such as left versus right occlusions), relevant features may differ significantly as a consequence of limited common visible regions. These scenarios represent another challenging problem that will require further investigation in future work.

Table 5: Rank-1 rate [%] and EER [%] for the proposed method versus OA-ModelGait when probe (P) and gallery (G) present the same occlusion ratio. Different rows show results for each pair of the same occlusion ratio, and different columns show mean values computed for pairs with corresponding view difference.

(a) Rank-1 rates [%] for the proposed method

P&G occlusion ratio	P&G view difference			
	0°	30°	60°	90°
0% vs. 0%	96.4	90.8	81.1	73.6
20% vs. 20%	94.7	86.5	75.4	66.5
40% vs. 40%	92.9	83.7	72.5	62.7
60% vs. 60%	82.0	63.6	51.4	36.9

(b) Rank-1 rates [%] for OA-ModelGait

P&G occlusion ratio	P&G view difference			
	0°	30°	60°	90°
0% vs. 0%	**99.0**	**96.2**	**90.6**	**82.5**
20% vs. 20%	**98.6**	**94.6**	**87.5**	**77.8**
40% vs. 40%	**98.5**	**93.3**	**85.2**	**75.0**
60% vs. 60%	**95.8**	**76.1**	**62.3**	**48.4**

(c) EERs [%] of the proposed method

P&G occlusion ratio	P&G view difference			
	0°	30°	60°	90°
0% vs. 0%	0.43	0.75	1.12	1.28
20% vs. 20%	0.57	0.96	1.39	1.66
40% vs. 40%	0.65	1.05	1.50	1.77
60% vs. 60%	1.12	1.78	2.56	3.08

(d) EERs [%] of OA-ModelGait

P&G occlusion ratio	P&G view difference			
	0°	30°	60°	90°
0% vs. 0%	**0.15**	**0.22**	**0.34**	**0.49**
20% vs. 20%	**0.20**	**0.29**	**0.43**	**0.62**
40% vs. 40%	**0.20**	**0.33**	**0.50**	**0.75**
60% vs. 60%	**0.40**	**0.84**	**1.29**	**1.86**

6. Conclusion

In this study, we propose an occluded gait recognition method based on silhouette registration, which directly processes bounding boxes restricted only to visible regions of input silhouettes. In our approach, an occlusion ratio estimator, a sampler, a pairwise mask, and a feature extractor are sequentially applied to obtain gait features from automatically registered silhouettes. Our experiments with the simulated occluded OU-MVLP database demonstrate substantial performance improvements obtained with the proposed method for large occlusion variations. We discuss future challenges and directions for research on this topic using evidence from quantitative and qualitative evaluations. In this context, we propose the incorporation of model-based methods after properly handling boundary issues. In future work, we plan to extend the repertoire of occlusion types used for evaluation, including the collection of realistic occlusion data to facilitate future research on real-world application scenarios.

Acknowledgments. This work was partly supported by JST Moonshot R&D Grant Number JPMJMS2215, and JSPS KAKENHI Grant Number JP19H05692 and Grant JP20H00607.

References

[1] Weizhi An, Shiqi Yu, Yasushi Makihara, Xinhui Wu, Chi Xu, Yang Yu, Rijun Liao, and Yasushi Yagi. Performance evaluation of model-based gait on multi-view very large population database with pose sequences. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2(4):421–430, 2020.

[2] I. Bouchrika, M. Goffredo, J. Carter, and M. Nixon. On using gait in forensic biometrics. *Journal of Forensic Sciences*, 56(4):882–889, 2011.

[3] Nikolaos V. Boulgouris and Zhiwei X. Chi. Human gait recognition based on matching of body components. *Pattern Recognition*, 40(6):1763–1770, 2007.

[4] 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 Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[5] Tianrui Chai, Annan Li, Shaoxiong Zhang, Zilong Li, and Yunhong Wang. Lagrange motion analysis and view embeddings for improved gait recognition. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20217–20226, 2022.

[6] J. Chaki and N. Dey. *A Beginner's Guide to Image Preprocessing Techniques*. Intelligent Signal Processing and Data Analysis. CRC Press, 2018.

[7] Hanqing Chao, Yiwei He, Junping Zhang, and Jianfeng Feng. Gaitset: Regarding gait as a set for cross-view gait recognition. In *Proc. of the 33th AAAI Conference on Artificial Intelligence (AAAI 2019)*, 2019.

[8] Hanqing Chao, Kun Wang, Yiwei He, Junping Zhang, and Jianfeng Feng. Gaitset: Cross-view gait recognition through utilizing gait as a deep set. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3467–3478, 2022.

[9] Changhong Chen, Jimin Liang, Heng Zhao, Haihong Hu, and Jie Tian. Frame difference energy image for gait recognition with incomplete silhouettes. *Pattern Recognition Letters*, 30(11):977–984, 2009.

[10] Dhritimaan Das, Ayush Agarwal, and Pratik Chattopadhyay. Gait recognition from occluded sequences in surveillance sites. In Leonid Karlinsky, Tomer Michaeli, and Ko Nishino, editors, *Computer Vision – ECCV 2022 Workshops*, pages 703–719, Cham, 2023. Springer Nature Switzerland.

[11] Huanzhang Dou, Pengyi Zhang, Wei Su, Yunlong Yu, and Xi Li. Metagait: Learning to learn an omni sample adaptive representation for gait recognition. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part V*, page 357–374, Berlin, Heidelberg, 2022. Springer-Verlag.

[12] Chao Fan, Yunjie Peng, Chunshui Cao, Xu Liu, Saihui Hou, Jiannan Chi, Yongzhen Huang, Qing Li, and Zhiqiang He. Gaitpart: Temporal part-based model for gait recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14213–14221, 2020.

[13] Yu Guan and Chang-Tsun Li. A robust speed-invariant gait recognition system for walker and runner identification. In *Proc. of the 6th IAPR International Conference on Biometrics*, pages 1–8, 2013.

[14] J. Han and B. Bhanu. Individual recognition using gait energy image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(2):316– 322, 2006.

[15] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *CoRR*, abs/1703.07737, 2017.

[16] Martin Hofmann, Daniel Wolf, and Gerhard Rigoll. Identification and reconstruction of complete gait cycles for person identification in crowded scenes. In *VISAPP 2011 - Proceedings of the International Conference on Computer Vision Theory and Application*, pages 594–597, 01 2011.

[17] Saihui Hou, Chunshui Cao, Xu Liu, and Yongzhen Huang. Gait lateral network: Learning discriminative and compact representations for gait recognition. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX*, page 382–398, Berlin, Heidelberg, 2020. Springer-Verlag.

[18] Xiaohu Huang, Duowang Zhu, Hao Wang, Xinggang Wang, Bo Yang, Botao He, Wenyu Liu, and Bin Feng. Context-sensitive temporal feature learning for gait recognition. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12889–12898, 2021.

[19] Zhen Huang, Dixiu Xue, Xu Shen, Xinmei Tian, Houqiang Li, Jianqiang Huang, and Xian-Sheng Hua. 3d local convolutional neural networks for gait recognition. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14900–14909, 2021.

[20] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *CoRR*, abs/1502.03167, 2015.

[21] H Iwama, D. Muramatsu, Y. Makihara, and Y. Yagi. Gait verification system for criminal investigation. *IPSJ Transactions on Computer Vision and Applications*, 5:163–175, Oct. 2013.

[22] H. Iwama, M. Okumura, Y. Makihara, and Y. Yagi. The ou-isir gait database comprising the large population dataset and performance evaluation of gait recognition. *IEEE Transactions on Information Forensics and Security*, 7(5):1511–1521, Oct. 2012.

[23] Yumi Iwashita, Koji Uchino, and Ryo Kurazume. Gait-based person identification robust to changes in appearance. *Sensors*, 13(6):7884–7901, 2013.

[24] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and koray kavukcuoglu. Spatial transformer networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 2017–2025. Curran Associates, Inc., 2015.

[25] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint*, arXiv: 1412.6980 (2014), 2014.

[26] Xiang Li, Yasushi Makihara, Chi Xu, and Yasushi Yagi. Multi-view large population gait database with human meshes and its performance evaluation. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 4(2):234–248, 2022.

[27] Xiang Li, Yasushi Makihara, Chi Xu, Yasushi Yagi, and Mingwu Ren. Gait recognition via semi-supervised disentangled representation learning to identity and covariate features. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[28] Xiang Li, Yasushi Makihara, Chi Xu, Yasushi Yagi, Shiqi Yu, and Mingwu Ren. End-to-end model-based gait recognition. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, November 2020.

[29] Junhao Liang, Chao Fan, Saihui Hou, Chuanfu Shen, Yongzhen Huang, and Shiqi Yu. Gaitedge: Beyond plain end-to-end gait recognition for better practicality. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part V*, page 375–390, Berlin, Heidelberg, 2022. Springer-Verlag.

[30] Rijun Liao, Shiqi Yu, Weizhi An, and Yongzhen Huang. A model-based gait recognition method with body pose and human prior knowledge. *Pattern Recognition*, 98:107069, 2020.

[31] Beibei Lin, Shunli Zhang, and Xin Yu. Gait recognition via effective global-local feature representation and local temporal aggregation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14648–14656, October 2021.

[32] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Smpl: A skinned multi-person linear model. *ACM Trans. Graph.*, 34(6), Oct. 2015.

[33] Niels Lynnerup and Peter Kastmand Larsen. Gait as evidence. *IET Biometrics*, 3(2):47–54, 6 2014.

[34] Y. Makihara, R. Sagawa, Y. Mukaigawa, T. Echigo, and Y. Yagi. Gait recognition using a view transformation model in the frequency domain. In *Proc. of the 9th European Conference on Computer Vision*, pages 151–163, Graz, Austria, May 2006.

[35] Y. Makihara, A. Suzuki, D. Muramatsu, X. Li, and Y. Yagi. Joint intensity and spatial metric learning for robust gait recognition. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6786–6796, July 2017.

[36] Daigo Muramatsu, Yasushi Makihara, and Yasushi Yagi. Gait regeneration for recognition. In *2015 International Conference on Biometrics (ICB)*, pages 169–176, 2015.

[37] Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, pages 807–814, USA, 2010. Omnipress.

[38] Prasit Nangtin, Pinit Kumhom, and Kosin Chamnongthai. Gait identification with partial occlusion using six modules and consideration of occluded module exclusion. *J. Vis. Comun. Image Represent.*, 36(C):107–121, Apr. 2016.

[39] Mark S. Nixon, Tieniu N. Tan, and Rama Chellappa. *Human Identification Based on Gait*. Int. Series on Biometrics. Springer-Verlag, Dec. 2005.

[40] Javier Ortells, Ramón Alberto Mollineda, Boris Mederos, and Raúl Martín-Félez. Gait recognition from corrupted silhouettes: a robust statistical approach. *Machine Vision and Applications*, 28:15–33, 2017.

[41] Aditi Roy, Shamik Sural, Jayanta Mukherjee, and Gerhard Rigoll. Occlusion detection and gait silhouette reconstruction from degraded scenes. *Signal, Image and Video Processing*, 5:415–430, 2011.

[42] S. Sarkar, J.P. Phillips, Z. Liu, I.R. Vega, P. Gro ther, and K.W. Bowyer. The humanid gait challenge problem: Data sets, performance, and analysis. *IEEE Transactions of Pattern Analysis and Machine Intelligence*, 27(2):162–177, 2005.

[43] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017.

[44] K. Shiraga, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi. Geinet: View-invariant gait recognition using a convolutional neural network. In *2016 International Conference on Biometrics (ICB)*, pages 1–8, 2016.

[45] Noriko Takemura, Yasushi Makihara, Daigo Muramatsu, Tomio Echigo, and Yasushi Yagi. Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition. *IPSJ Trans. on Computer Vision and Applications*, 10(4):1–14, 2018.

[46] Torben Teepe, Johannes Gilg, Fabian Herzog, Stefan Hörmann, and Gerhard Rigoll. Towards a deeper understanding of skeleton-based gait recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1569–1577, June 2022.

[47] Torben Teepe, Ali Khan, Johannes Gilg, Fabian Herzog, Stefan Hörmann, and Gerhard Rigoll. Gaitgraph: Graph convolutional network for skeleton-based gait recognition. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 2314–2318, 2021.

[48] Md. Zasim Uddin, Daigo Muramatsu, Noriko Takemura, Md Atiqur Rahman Ahad, and Yasushi Yagi. Spatio-temporal silhouette sequence reconstruction for gait recognition against occlusion. *IPSJ Transactions on Computer Vision and Applications*, 11:1–18, 2019.

[49] Likai Wang, Jinyan Chen, and Yuxin Liu. Frame-level refinement networks for skeleton-based gait recognition. *Computer Vision and Image Understanding*, 222:103500, 2022.

[50] Tenika Whytock, Alexander Belyaev, and Neil M. Robertson. On covariate factor detection and removal for robust gait recognition. *Mach. Vision Appl.*, 26(5):661–674, July 2015.

[51] Chi Xu, Yasushi Makihara, Xiang Li, and Yasushi Yagi. Occlusion-aware human mesh model-based gait recognition. *IEEE Transactions on Information Forensics and Security*, 18:1309–1321, 2023.

[52] 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. of the 13th Asian Conf. on Computer Vision (ACCV 2016)*, pages 52–67, Taipei, Taiwan, Nov. 2016.

[53] S. Yu, H. Chen, E. B. G. Reyes, and N. Poh. Gaitgan: Invariant gait feature extraction using generative adversarial networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 532–539, July 2017.

[54] Shiqi Yu, Daoliang Tan, Kaiqi Huang, and Tieniu Tan. Reducing the effect of noise on human contour in gait recognition. In Seong-Whan Lee and Stan Z. Li, editors, *Advances in Biometrics*, pages 338–346, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg.

[55] S. Yu, D. Tan, and T. Tan. A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In *Proc. of the 18th Int. Conf. on Pattern Recognition*, volume 4, pages 441–444, Hong Kong, China, Aug. 2006.

[56] Ziyuan Zhang, Luan Tran, Feng Liu, and Xiaoming Liu. On learning disentangled representations for gait recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):345–360, 2022.

[57] Ziyuan Zhang, Luan Tran, Xi Yin, Yousef Atoum, Xiaoming Liu, Jian Wan, and Nanxin Wang. Gait recognition via disentangled representation learning. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4705–4714, 2019.

[58] Guoying Zhao, Li Cui, and Hua Li. Gait recognition using fractal scale. *Pattern Analysis and Applications*, 10:235–246, 2007.