Is Multi-Person Gait Recognition Feasible under Mutual Occlusion? A Human Model Regression-based Approach

Supplementary Material

We provide more details regarding the architecture of the proposed framework, the computation of pseudo-GT SMPL parameters and location descriptors, and the implementation of comparison methods in the supplementary material.

7. Architecture details of the proposed framework

STE. The architecture of our STE is based on SwinV2-small [41] with a 16×16 window size. The model takes 256×256 sized images as input and first divides them into 4×4 patches. It then processes the image through four stages, with the number of blocks in each stage set to $[2, 2, 18, 2]$ and the number of attention heads per stage set to $[3, 6, 12, 24]$. The output feature map has a resolution of 8×8 patches, each with a feature dimension of 768.

MPR. The architecture of our MPR is based on a standard transformer decoder architecture with multi-head self-attention as [17] and an MLP. It consists of 6 layers, each with multi-head self-attention, multi-head cross-attention, and feed-forward blocks, with layer normalization. It has a 2048 hidden dimension, 8 heads for self- and cross-attention, and a hidden dimension of 1024 in the feed-forward MLP block. It processes on a learnable 2048-dimensional SMPL query token as input, which cross-attends to the latent features extracted from STE. Lastly, a single-layer MLP on the output token regresses the SMPL parameters.

Location Estimator. The Location Estimator takes the latent features extracted from the STE as input and first computes the mean over the 8×8 patches, resulting in a 768-dimensional feature vector. Two separate MLPs, each consisting of a single fully connected layer, are then used to estimate the 3-dimensional location descriptor for each person.

CNN architecture for joint-based recognition. We employed a lightweight CNN-based architecture for joint-based feature extraction. The network takes the temporally concatenated joint matrix from 3D joints in a sequence as input, then processes it through three residual blocks, each consisting of 3×3 convolutional layers followed by batch normalization and ReLU activation. The stride along the column dimension is set to 2 to progressively reduce spatial resolution, while the feature channels are increased from 64 to 128 and then to 256. Following the residual blocks, an adaptive average pooling layer and a fully connected layer are applied to project the representation into a 64-dimensional feature vector. Finally, the output feature serves as the gait feature embedding for further recognition.

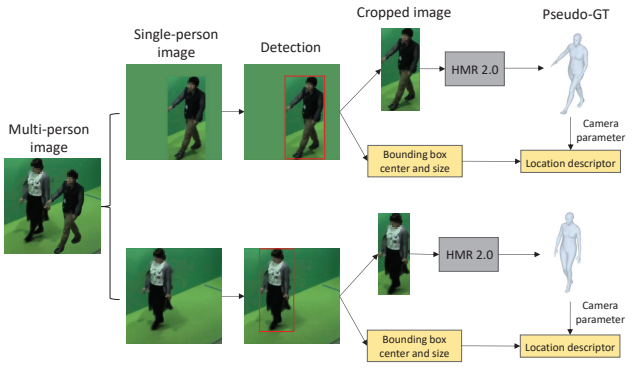

Figure 7. Pipeline for computing the pseudo-GT SMPL parameters and location descriptors for multi-person images.

8. Computation of pseudo-GT

As mentioned in Section 4.1, to generate reliable pseudo-GT for training supervision, including location descriptors and SMPL parameters, we preserved corresponding single-person images aligned with the positions of each subject in the composited multi-person images, ensuring complete and unobstructed body information. Fig. 7 shows the whole pipeline of pseudo-GT computation. Given the single-person gait images, we first used the SOTA detection method VitDet [38] to obtain the bounding box of each person, then applied HMR 2.0 [17] to estimate the SMPL parameters of the detected person. Using the bounding box center and size, and the camera parameters of SMPL, we calculated the location descriptor according to the Equation 6 in Sec. 3.5. Following [17], the focal length f is set to 5,000 by default. This process enabled us to compute the pseudo-GT SMPL parameters and location descriptors for each individual in the multi-person gait images, ensuring accurate supervision during model training.

9. Implementation details of comparison methods

To evaluate performance, we chose GaitBase [13], DeepGaitV2 [14], and ModelGait [35] as comparison methods. For GaitBase and DeepGaitV2, we used the official open-source repository OpenGait. Given the smaller number of training subjects in our dataset (i.e., 1,000), we adopt training configurations similar to the default settings used for

CASIA-B [65]. For ModelGait, considering the existence of pseudo-GT SMPL parameters, we added the same loss ($\mathcal{L}_{\mathrm{SMPL}}$) as the proposed method, which provides a more accurate supervision.