Occluded Gait Recognition with Mixture of Experts: An Action Detection Perspective

Panjian Huang[1,3⋆], Yunjie Peng[2,4⋆], Saihui Hou[1,3†], Chunshui Cao[3], Xu Liu[3], Zhiqiang He[2,4], and Yongzhen Huang[1,3†]

[1] School of Artificial Intelligence, Beijing Normal University
[2] School of Computer Science and Technology, Beihang University
[3] WATRIX.AI
[4] AI Lab, Lenovo Research

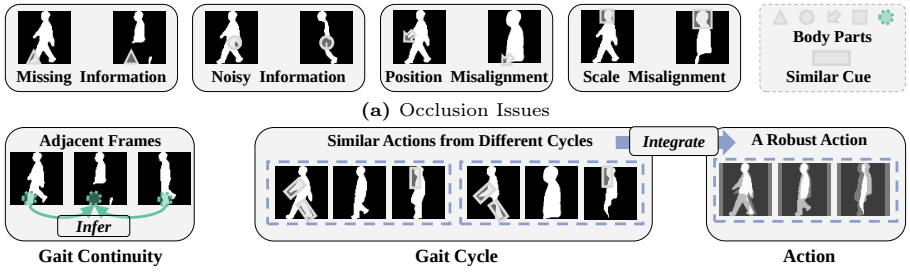
Abstract. Extensive occlusions in real-world scenarios pose challenges to gait recognition due to missing and noisy information, as well as body misalignment in position and scale. We argue that rich dynamic contextual information within a gait sequence inherently possesses occlusion-solving traits: 1) Adjacent frames with gait continuity allow holistic body regions to infer occluded body regions; 2) Gait cycles allow information integration between holistic actions and occluded actions. Therefore, we introduce an action detection perspective where a gait sequence is regarded as a composition of actions. To detect accurate actions under complex occlusion scenarios, we propose an Action Detection Based Mixture of Experts (GaitMoE), consisting of Mixture of Temporal Experts (MTE) and Mixture of Action Experts (MAE). MTE adaptively constructs action anchors by temporal experts and MAE adaptively constructs action proposals from action anchors by action experts. Especially, action detection as a proxy task with gait recognition is an end-to-end joint training only with ID labels. In addition, due to the lack of a unified occluded benchmark, we construct a pioneering Occluded Gait database (OccGait), containing rich occlusion scenarios and annotations of occlusion types. Extensive experiments on OccGait, OccCASIA-B, Gait3D and GREW demonstrate the superior performance of GaitMoE. OccGait is available at https://github.com/BNU-IVC/OccGait.

Keywords: Occluded Gait Recognition · Dynamic Contextual Information · Action Detection · Mixture of Experts

1 Introduction

Gait recognition has attracted increasing attention and gained broad applications in crime prevention, forensic identification, and social security [44] due to its ability to accurately identify walking patterns of pedestrians from a distance in complex surveillance scenarios, *e.g.*, viewing angles, cloth-changing and

⋆ Equal contribution
† Corresponding author



(a) Occlusion Issues

(b) Dynamic Contextual Information (Ours)

Fig. 1: (a) The main four occlusion issues in gait recognition. (b) The dynamic contextual information within a gait sequence can infer and integrate occlusion information.

illumination conditions [40]. Applying upstream tasks such as tracking, segmentation, size normalization, and alignment to preprocess raw videos, the obtained gait representations (*e.g.*, silhouettes or skeletons) make existing methods [3, 4, 10, 12, 23–25, 31, 32, 34, 52] achieve accurate identification. However, current studies overlook occlusions largely existing in practical scenarios, *e.g.*, occluded by carrying, obstacles, the crowd, or moving out of camera view. As shown in Fig. 1(a), body regions occluded by obstacles or the crowd lead to missing and noisy information, while partial visual body regions with size normalization cause position and scale misalignment, which significantly degrading fine-grained feature matching [56,58]. Direct solutions, *e.g.*, simply discarding or persisting occluded frames, do not adequately address occlusion issues since partially visual body regions in occluded frames may still contain key discriminative regions while allowing them to persist will cause erroneous feature extraction and matching. Therefore, occlusion issues have become one of the biggest bottlenecks in gait recognition.

To address occlusion issues in gait recognition, we rethink the dynamic contextual information within a gait sequence: **(i) Gait Continuity.** Adjacent frames with continual motion enable body regions in holistic frames to infer the same body regions in occluded frames. As shown in Fig. 1(b), for the current frame missing lower body information, the preceding and subsequent holistic frames can still infer approximate motion in the occluded regions. **(ii) Gait Cycle.** As shown in Fig. 1(b), we regard the current misaligned frame and adjacent frames as an action. Combined with the gait cycle, *i.e.*, a gait sequence is formed by the repetition of a series of actions, to discover actions with similar cues, holistic and occluded actions can be integrated into a robust action.

Driven by the above analysis, we introduce a new perspective for occluded gait recognition, "Action Detection". The paradigm of action detection aims to predict the action boundaries and categories from an untrimmed video, where predefined consecutive frames as action anchors represent potential actions and further action anchors with high actionness scores generate action proposals [33, 43,59]. Specific to gait recognition, we associate action anchors to represent adjacent frames with gait continuity and further consider the gait cycle to construct

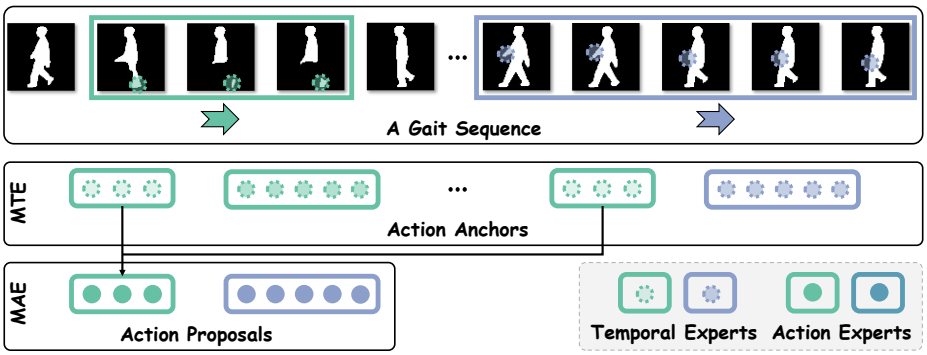


Fig. 2: Action composition. Each temporal expert focuses on one body region with individual temporal size, constructing action anchors. Each action expert integrates similar action anchors from different gait cycles, constructing one action proposal.

action proposals from similar action anchors. Therefore, a gait sequence can be regarded as a composition of actions.

Considering a single model that struggles to capture holistic and diverse actions under complex occlusion scenarios, *e.g.*, the uncertainty of occluded body regions and duration, we propose an Action Detection Based Mixture of Experts (GaitMoE). Mixture of Experts (MoE) follows a divide-and-conquer philosophy, breaking down complex problems into simple sub-problems. Each sub-problem is handled by a dedicated expert, contributing collectively to solve the overall complexity. As shown in Fig. 2, each temporal expert focuses on the corresponding body region and temporal granularity, sliding at the entire gait sequence to construct action anchors. Subsequently, each action expert integrates similar action anchors with one action type, constructing action proposals. Finally, instead of localization and classification in action detection, action proposals are collectively as discriminative features for identification.

Additionally, the absence of publicly available gait databases with quantifiable occlusion metrics poses an extreme challenge for occluded gait recognition. To this end, we establish a pioneering Occluded Gait database (OccGait) with two characteristics: **(i) Diverse Occlusion Scenarios.** Each subject has 4 different types of occlusion situations, including None of Occlusion, Carrying Occlusion, Crowd Occlusion, and Static Occlusion. **(ii) Explicit Occlusion Types.** OccGait provides explicit occlusion types for each gait sequence, which enables to qualify and quantify occlusion issues.

Our main contributions can be summarized as follows:

- To address occlusion challenges, we introduce an action detection perspective where an Action Detection Based Mixture of Experts (GaitMoE) structures a gait sequence as a composition of action.
- To qualify and quantify occlusion issues, we build a novel Occluded Gait recognition benchmark (OccGait), including diverse occlusion scenarios and explicit annotations of occlusion types.

– To evaluate effectiveness and robustness, extensive experimental results on OccGait, OccCASIA-B, Gait3D, and GREW demonstrate that our method significantly outperforms other state-of-the-art methods.

2 Related Work

2.1 Gait Recognition

Gait Recognition is mainly categorized into appearance-based and model-based approaches. Appearance-based methods [3, 4, 10, 12, 23–25, 31, 32, 34, 52] usually uses templates of compressing a sequence of gait silhouettes (*e.g.*, Gait Energy Image), set of frames and sequence of frames as inputs, extracting fine-grained features (*e.g.*, spatial-temporal and part-level representations). Model-based methods [14,47,48,61–63] explicitly model human body structure, *e.g.*, 2D or 3D skeletons and meshes. Additionally, some researches [2, 30, 42] take other data types as inputs, such as RGB frames, optical flow and point clouds. However, most of these methods usually neglect the fact that real-world scenarios introduce a significant amount of occlusion.

2.2 Occluded Gait Recognition

We introduce occluded gait recognition from two aspects: **(i) DataBases.** For synthesis-based databases, Chen *et al.* [6] simulate occluded scenarios based on CMU Mobo [15] through adding horizontal or vertical black bars. Uddin *et al.* [50] synthesize relative static and dynamic occlusions based on OUMVLP [45] by a background rectangle mask in a fixed position or gradually changed position. Delgado-Escano *et al.* [9] generate crowd occlusions based on CASIA-B [60] and TUM-GAID [20] by augmenting persons in raw videos. Xu *et al.* [56, 58] synthesize occluded scenarios by simulating cropping and size-normalized silhouettes. For real-world databases, Hofmann *et al.* [21] collect TUM-IITKGP, including static occlusions (*e.g.*, standing people, a backpack, gown and hands in pocket), dynamic occlusions (*e.g.*, two walking people). Chattopadhyay *et al.* [5] construct a frontal and occluded gait database by estimating Kinect depth. Li *et al.* [29] present an OG RGB+D dataset captured by Azure Kinect DK sensor, containing occlusions with carrying, clothing and the crowd. However, these databases have some limitations, *e.g.*, the single occlusion scenario, small occlusion regions, or not publicly available yet. **(ii) Architectures.** For reconstruction-based approaches, Xu *et al.* [58] re-normalize and register silhouettes from learned holistic information (*e.g.*, scales) before the following feature extraction and matching process. Xu *et al.* [56] estimate SMPL with pose and shape features from RGB occluded videos. Uddin *et al.* [50] reconstruct a sequence from the occluded sequence by a conditional deep generative adversarial network. Peng *et al.* [37] register and recover occluded silhouettes with a self-supervised alignment module and temporal recovery transformer. Guo *et al.* [16] propose a Physics-Augmented Autoencoder (PAA) that generates physically intermediate representations through a graph-convolution-based encoder and a

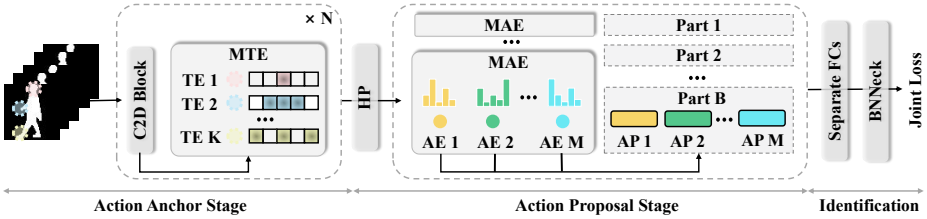


Fig. 3: The overview of GaitMoE. Best viewed in colors, the input sequence is firstly extracted to frame-level features by C2D Block (*e.g.*, 2D CNNs or a residual block), and Mixture of Temporal Experts (MTE) dispatch temporal experts (TE) with different temporal bounding boxes to the corresponding channel segments, forming action anchors. After horizontal partitioning (HP), Mixture of Action Experts (MAE) is independent for each part-level feature. For each channel segment, one action expert (AE) integrates similar action anchors along the temporal dimension by weighted sum operation, forming one action proposal. Finally, The concatenated action proposals as part features for identification.

physics-based decoder, which enhances the ability to reconstruct input skeleton sequences even when partially occluded. For reconstruction-free approaches, Gupta *et al.* [17] propose a occlusion-aware module by synthetic occlusions to detect occlusion type information to guide gait recognition training. Zhu *et al.* [64] use SMPLify-X that provides the body shape feature decoupled from its pose and strong prior, which enables to generate the complete shape even with mild occlusions.

2.3 Mixture of Experts

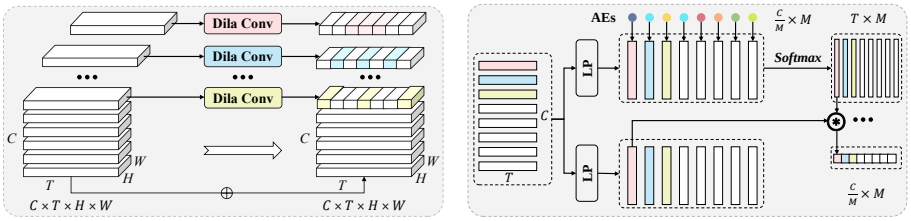
Mixture of Experts (MoE) is a sparsely-activated architecture where a router network output weights for aggregating multiple experts [41]. The philosophy of divide and conquer allows MoE to reduce computational cost and increase model capacity, and it has been widely extended to Vision Transformer [28,38,55]. Each expert in MoE is dispatched with one sub-data (*e.g.*, image patches, data from one domain) and maintains specialization, which is named "Expert". This work extends MoE to detect fine-grained actions in occluded gait recognition and makes each expert concentrate on one representative action.

3 Methodology

GaitMoE mainly consists of Mixture of Temporal Experts (MTE) and Mixture of Action Experts (MAE). We give a brief overview of the full process in Fig. 3.

3.1 Mixture of Temporal Experts

Although a gait sequence has filtered texture information and performed coarse-aligned registration, the complex occlusions in real scenarios cause the uncer-



(a) MTE **(b)** MAE

Fig. 4: (a) Mixture of Temporal Experts. Dila Conv (DC) represents Dilated Convolution, predefining action anchors with different dilated ratios. (b) Mixture of Action Experts. LP denotes the linear projection. Similar action anchors adaptively integrate into action proposals.

tainty of occluded body regions and duration. To this end, we adopt multi-scale mechanisms in temporal and channel dimensions to extract fine-grained features.

Action Anchors. As we know, a gait sequence possesses continuity with significant mutual information between each frame and adjacent frames. For example, when we observe a person lifting their leg, it is likely to be followed by a swinging leg. Some existing approaches employ temporal modeling to capture such relationships, *e.g.*, 3D CNNs and LSTMs. However, uncertain and complex occlusions interfere with the dynamic information. To alleviate these issues, Fig. 4(a) shows that MTE predefines various sizes of temporal experts, which are dilated convolutions with different dilated ratios for corresponding channel segments. At each temporal position, each temporal expert independently constructs action anchors from adjacent temporal positions, which is why we name it "Temporal Expert". Considering adjacent occluded frames may introduce noise information to current holistic frames, MTE preserves partial channel segments for adaptively selecting clean regions. Let $\mathcal{X} \in \mathbb{R}^{\mathcal{C}\times\mathcal{T}\times\mathcal{H}\times\mathcal{W}}$ denote frame-level features extracted from silhouettes by C2D Block, where $\mathcal{C}, \mathcal{T}, \mathcal{H}, \mathcal{W}$ represent channel, consecutive $\mathcal{T}$ frames, height and width dimensions. The process of MTE is formulated as follows:

$$\mathcal{Y} = Concat(DC_i(\mathcal{X}_i), i = 1, 2, \cdots, \mathcal{K}, \mathcal{X}_{[\mathcal{S}-\mathcal{K}+1,\mathcal{S}]}) \tag{1}$$

where $\mathcal{X}_i \in \mathbb{R}^{\frac{\mathcal{C}}{\mathcal{S}}\times\mathcal{T}\times\mathcal{H}\times\mathcal{W}}$, $\mathcal{Y} \in \mathbb{R}^{\mathcal{C}\times\mathcal{T}\times\mathcal{H}\times\mathcal{W}}$, $\mathcal{S}$ is the number of channel segments, $\mathcal{K}$ is the number of temporal experts, DC is dilated convolution, and i is the segment index. In our work, we set $\mathcal{S} = 8$, $\mathcal{K} = 4$, and DC_i is 3D CNN with kernel size $(3, 1, 1)$, stride $(1, 1, 1)$, padding $(i, 0, 0)$, and dilated ratio i. In addition, residual learning is embedded within MTE for easing training.

3.2 Mixture of Action Experts

Although action anchors contain rich dynamic information, they may have a large amount of redundant and invalid actions. Therefore, we adopt prototype-based architecture to adaptively select and integrate discriminative and similar

action anchors as action proposals. Since a gait sequence can be regarded as the composition of actions, when the sequence is occluded resulting in many invalid actions, the action prototype needs to detect the most discriminative action type and aggregate this action type from occluded and holistic actions. In addition, GaitMoE adopts horizontal pooling (HP) for fine-grained part features $\mathcal{P} \in \mathbb{R}^{\mathcal{C} \times \mathcal{T}}$, and MAE is independent for each part. Here, we omit the part index for simplicity.

Action Proposals. To capture discriminative actions only with ID labels, MAE shown in Fig. 4(b) predefines a set of learnable action prototypes where an action prototype adaptively learns a type of action for recognition, that is why we name it "Action Expert". For fine-grained action extraction, we dispatch each action expert to the corresponding channel segment (*e.g.*, action anchors), and action experts "watch" contextual action anchors in the entire temporal dimension for filtering action anchors with occlusions, and select and integrate similar action anchors as action proposals. Different to MoEs [13, 27, 36, 41] where the routers generally adopt Top-K selection and sparsely memorize information, recent MoE works has shown remarkable performance with a fixed hash router [39], or convolutional experts [7]. In this work, we introduce a soft selection for balancing the training. Let $\mathcal{A} \in \mathbb{R}^{\frac{\mathcal{C}}{\mathcal{M}} \times \mathcal{M}}$ represent $\mathcal{M}$ action experts with $\frac{\mathcal{C}}{\mathcal{M}}$ dimension, and we obtain action queries $\mathcal{Q}$ by identify mapping on $\mathcal{A}$, action keys $\mathcal{K}$ and values $\mathcal{V}$ by different linear projections on $\mathcal{P}$. Then, we dispatch each action query to the corresponding channel segment of $\mathcal{K}$ to evaluate action anchors by scores where the higher the score, the more reliable the action anchor, and vice versa. To make one action expert concentrate on one most representative action, we use the softmax function to calculate the scores within the corresponding channel segment of $\mathcal{K}$ and weighted sum operation with the corresponding channel segment of $\mathcal{V}$ as an action proposal. The formulation is as follows:

$$\mathcal{Q}_i = \mathcal{A}_i, \quad \mathcal{K}_i = \mathcal{P}_i \mathcal{W}^{\mathcal{K}}, \quad \mathcal{V}_i = \mathcal{P}_i \mathcal{W}^{\mathcal{V}} \tag{2}$$

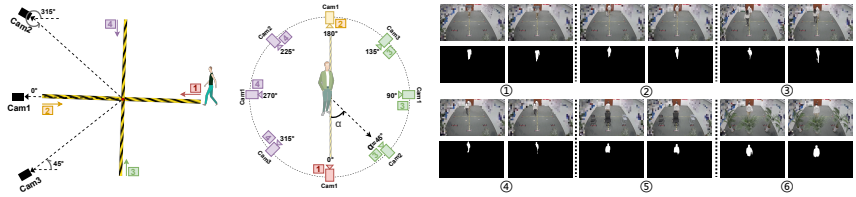
$$\mathcal{F}_i = \sum_{j=1}^{\mathcal{T}} \mathcal{O}_{i,j} \otimes \mathcal{V}_{i,j}, \quad \mathcal{O}_{i,j} = \frac{exp(\mathcal{Q}_{i,j} \mathcal{K}_{i,j}^T)}{\sum_{j=1}^{\mathcal{T}} exp(\mathcal{Q}_{i,j} \mathcal{K}_{i,j}^T)} \tag{3}$$

where i is the channel segment index, $i \in 1, 2, \cdots, \mathcal{M}$, j is the action anchor index along the temporal dimension of the i segment, $j \in 1, 2, \cdots, \mathcal{T}$, $\mathcal{W}^{\mathcal{K}}, \mathcal{W}^{\mathcal{V}} \in \mathbb{R}^{\frac{\mathcal{C}}{\mathcal{M}} \times \frac{\mathcal{C}}{\mathcal{M}}}$, $\mathcal{P}_i \in \mathbb{R}^{\mathcal{T} \times \frac{\mathcal{C}}{\mathcal{M}}}$. $\mathcal{Q}_{i,j}, \mathcal{K}_{i,j}, \mathcal{V}_{i,j} \in \mathbb{R}^{1 \times \frac{\mathcal{C}}{\mathcal{M}}}$. Finally, we obtain $\mathcal{F}$ as one part feature by concatenating action proposals $[\mathcal{F}_1, \mathcal{F}_2, \cdots, \mathcal{F}_{\mathcal{M}}]$ along the channel dimension. $\mathcal{F}$ is fed into Separate FCs and BNNeck for identification.

3.3 Joint Loss

GaitMoE is an end-to-end joint learning framework only with ID labels, introducing action detection as a proxy task with gait recognition. The joint loss includes two types: Triplet Loss [19] $\mathcal{L}_{tp}$ and Cross Entropy Loss $\mathcal{L}_{ce}$, which constrains each part independently. This formulation is as follows:

$$\mathcal{L} = \mathcal{L}_{tp} + \beta \mathcal{L}_{ce} \tag{4}$$



(a) Layout and camera views **(b)** Occlusions

Fig. 5: (a) The layout and camera views during collection. (b) The 3 types of occlusion scenarios where the number ①② are for Carrying Occlusion, ③④ are for Crowd Occlusion and ⑤⑥ are for Static Occlusion, and None Occlusion is omitted for simplicity.

where the hyper-parameter β is for balancing the two terms.

4 The OccGait Benchmark

Due to the absence of a comprehensive gait database for quantitatively and qualitatively analyzing the impact of various occlusion types, we collect the Occluded Gait Database (OccGait), including 101 subjects with 4 types of occluded scenarios, 8 camera views, and over 80k sequences. *It is worth noting that we hope the OccGait can serve as a starting point to promote robust gait recognition for practical applications, similar to the transition from the indoor databases,* e.g.*, CASIA-X Series [44, 46, 53, 60] and OU-X Series [1, 22, 26, 35, 45, 49, 51, 57] to the wild databases,* e.g.*, Gait3D [62] and GREW [65].*

4.1 Data Collection and Pre-processing

The OccGait is collected in an indoor gait recognition laboratory. During OccGait collection, we obtain authorization from all subjects who are informed for academic data collection in advance. Privacy is also the highest priority in our research. As the left in Fig. 5(a), we place 3 cameras (Cam1 of 0°, Cam2 of 45°, Cam3 of 315°) with 1920 × 1080 resolution in the square area. During the data collection process, the subjects follow this walk route (*i.e.*, 1-2-3-4). We filter out data with overlapping camera views caused by the combination of 3 cameras and walking directions, and obtain gait sequences with 8 camera views on the right in Fig. 5(a). To qualify and quantify realistic and complex occluded scenarios, we set 4 types of occluded situations with diverse occlusions shown in Fig. 5(b), which are None of Occlusion (*i.e.*, Normal Walking as NM), Carrying Occlusion (CA, ①②), Crowd Occlusion (CR, ③④), and Static Occlusion (ST, ⑤⑥).

All subjects walk the route four times for NM and two times for CA, CR and ST, respectively, which denotes NM01, NM02, NM03, NM04, CA01, CA02, CR01, CR02, ST01 and ST02.

None of Occlusion. To simulate walking status in real scenarios, subjects in their clothing, walk with their walking speed. As shown in Fig. 5(a), there are no obstacles in this situation, and the full body of each subject is fully visible.

Carrying Occlusion. As shown in Fig. 5(b)(①②), we establish two common carrying scenarios for daily life: umbrellas and luggage. People with an umbrella on a rainy day partially obstruct the upper of the body, while luggage occludes both the torso and the lower of the body.
Crowd Occlusion. Gait recognition often requires retrieval in crowded scenes, and human body occlusion can lead to significant interference, such as occlusion of body edges and incorrect dynamic information, as shown in Fig. 5(b)(③④). We design two types of crowd occlusion, the subject walking with another person in different directions (opposite and parallel), which results in partial occlusion of gait sequences at certain moments and complete occlusion of gait sequences at all times.
Static Occlusion. Gait recognition will be deployed in a wide range of scenarios, such as in squares, malls, and other locations with numerous static obstructions. Fig. 5(b)(⑤) demonstrates our placement of plants and chairs in the walking route. Fig. 5(b)(⑥) shows that the complex and irregular static obstructions hinder the lower body region.
Data Pre-processing. We adopt MaskFormer [8], a segmentation algorithm pre-trained on large datasets (including numerous occluded scenarios), to extract silhouettes from the original RGB data as input.

4.2 Evaluation Protocol

OccGait is divided into training and testing sets. The 51 individuals with odd numbers as the training set while the remaining 50 with even numbers as the test set. Our gait evaluation follows the protocols of the previous gait evaluation [60]. Given a query sequence, we measure its distance to each sequence in the gallery, retrieving the subject with the closest distance from the gallery. To quantify occluded gait analysis, we use the NM01 and NM02 of each subject in the testing set as the gallery, evaluating their Rank 1 performance under different occlusions and viewing angles.

5 Experiments

5.1 Datasets

We first conduct extensive qualitative and quantitative occlusion analyses on OccGait and OccCASIA-B [37]. Subsequently, we further validate the generalizability and practicality of our method on Gait3D [62] and GREW [65].
OccGait is for real-scenario occlusion evaluation built by this work. The details have been discussed in Section 4.
OccCASIA-B is a synthetic occluded gait database [37] from CASIA-B and has similar basic statistics, containing 124 subjects, 3 different walking situations, *etc.*, Walking in Normal (NM), Walking with a Bag (BG) and Walking with Different Clothes (CL), 11 camera views from uniform interval of 18° in $[0°, 180°]$. To simulate occlusion situations, OccCASIA-B sets 4 types of occlusions: None Occlusion (NO), Crowd Occlusion (CO), Static Occlusion (SO) and

Table 1: The Rank-1 accuracy (%) on OccGait for different probe views excluding the identical-view cases. For evaluation, the sequences of NM01 and NM02 for each subject are taken as the gallery. The benchmark adopts None Occlusion (NO), Carrying Occlusion (CA), Crowd Occlusion (CR) and Static Occlusion (ST).

	Method	Probe View								Average
		0°	45°	90°	135°	180°	225°	270°	315°	
NM	GaitSet [4]	65.7	91.7	89.1	90.7	66.7	91.9	89.4	92.1	84.7
	GaitPart [12]	62.9	92.0	88.9	89.6	59.6	89.9	87.3	90.7	82.6
	GaitGL [32]	73.9	94.3	92.6	93.4	68.1	93.1	91.7	92.6	87.5
	STOR [37]	73.7	94.6	92.0	93.3	73.3	94.1	92.6	92.6	88.3
	GaitBase [11]	68.4	91.6	88.7	91.1	74.9	93.6	88.1	91.7	86.0
	GaitMoE(ours)	**81.0**	**96.0**	**94.0**	**95.1**	**81.3**	**94.7**	**94.1**	**94.7**	**91.4**
CA	GaitSet [4]	50.1	74.3	79.7	79.1	47.4	73.6	74.0	76.0	69.3
	GaitPart [12]	42.0	69.9	74.9	78.6	37.6	66.4	66.0	63.7	62.4
	GaitGL [32]	48.6	79.7	84.6	87.9	38.7	76.7	78.1	70.4	70.6
	STOR [37]	55.9	83.6	83.6	86.3	54.4	81.7	83.1	81.7	76.3
	GaitBase [11]	58.1	82.0	84.3	85.6	54.3	79.9	80.1	79.3	75.4
	GaitMoE(ours)	**68.3**	**87.7**	**88.9**	**89.6**	**61.7**	**86.4**	**86.6**	**87.3**	**82.1**
CR	GaitSet [4]	58.3	84.7	80.1	77.4	52.3	77.9	77.6	85.6	74.2
	GaitPart [12]	48.1	81.9	76.4	69.6	39.0	67.6	68.1	79.3	66.3
	GaitGL [32]	47.4	89.0	81.3	77.1	41.0	75.0	77.1	87.6	71.9
	STOR [37]	55.6	88.4	86.0	81.7	52.0	81.7	83.0	88.4	77.1
	GaitBase [11]	62.4	86.9	83.1	82.1	**60.0**	**85.4**	80.7	87.6	78.5
	GaitMoE(ours)	**63.1**	**90.6**	**86.6**	**84.4**	57.6	81.6	**84.9**	**90.3**	**79.9**
ST	GaitSet [4]	54.1	86.3	86.7	82.3	54.1	86.7	86.4	83.6	74.2
	GaitPart [12]	44.6	83.7	85.7	77.3	39.1	83.4	84.1	77.4	71.9
	GaitGL [32]	36.7	87.7	90.7	80.0	37.6	86.6	90.4	82.3	74.0
	STOR [37]	50.3	90.0	91.3	87.4	54.9	90.1	91.3	89.1	80.6
	GaitBase [11]	57.7	89.9	87.4	85.9	59.9	88.9	87.4	85.4	80.3
	GaitMoE(ours)	**62.9**	**93.3**	**92.1**	**89.4**	**63.6**	**91.1**	**93.6**	**91.4**	**84.7**

Detect Occlusion (DO), which denotes a walking person without occlusions, occluded by another one in a crowded area, occluded by static occlusions, *e.g.*, benches, bicycles and fire hydrants, and losing body regions in the up, down, left, or right direction. The OccCASIA-B takes the first 74 subjects as the training set where each gait sequence with 0.6 of occlusion probability generates one of the 4 types of occluded scenarios. The remaining 50 subjects are used for occlusion evaluation. For each occlusion benchmark, except for the first 4 NM sequences used as the holistic gallery set, the remaining sequences are generated with the corresponding occlusion scenario.

Gait3D samples two segments of continuous two-hour video clips from each of seven-day raw videos in a supermarket, including complex covariates (*e.g.*, occlusions, view angles) for practical gait recognition. It contains 3000 subjects with 25309 sequences, taking 2000 subjects as the training dataset and 1000 subjects as the testing dataset.

GREW is a large-scale wild gait database, containing 26345 subjects with 128671 sequences captured by 882 cameras. It provides 4 types of silhouettes, optical flow, and 2D/3D human poses. The benchmark takes 20000 subjects as the training dataset and 6000 subjects as the testing dataset, and each subject provides two sequences for the gallery set and two sequences for the probe set.

Table 2: The Rank-1 accuracy (%) on OccCAISA-B across different views, excluding the identical-view cases. The NO, CO, SO, and DO denote the testing sets of Non-Occlusion, Crowd Occlusion, Static Occlusion, and Detection Occlusion accordingly. Based on the walking condition, probe sequences are grouped into Normal Walking (NM), Carrying Bags (BG), and Cloth-changing Condition (CL).

Methods	NO				CO				SO				DO			
	NM	BG	CL	Mean	NM	BG	CL	Mean	NM	BG	CL	Mean	NM	BG	CL	Mean
GaitSet [4]	92.4	83.0	65.2	80.2	80.7	69.6	50.4	66.9	86.0	76.6	57.8	73.5	85.7	72.7	52.7	70.4
GaitPart [12]	92.3	84.9	68.9	82.0	80.2	70.6	53.3	68.1	85.0	77.3	60.5	74.3	80.7	68.4	52.4	67.2
GaitGL [32]	94.5	89.2	75.3	86.4	84.4	74.8	57.7	72.3	87.4	81.8	66.9	78.7	86.2	76.2	61.5	74.6
STOR [37]	95.9	90.9	77.1	88.0	88.8	80.7	64.3	77.9	91.2	85.6	69.9	82.3	93.2	86.3	70.1	83.2
GaitBase [11]	94.4	88.5	68.7	83.9	87.6	78.0	57.5	74.4	90.1	82.7	62.0	78.3	89.9	80.4	58.5	76.3
GaitMoE(ours)	**96.2**	**91.5**	**80.7**	**89.5**	**90.0**	**83.3**	**68.3**	**80.5**	**91.9**	**86.0**	**73.9**	**83.9**	**93.4**	**87.3**	**75.3**	**85.3**

Table 3: Comparisons on Gait3D and GREW.

Method	Venue	Gait3D		GREW	
		Rank-1	mAP	Rank-1	Rank-5
GaitSet [4]	AAAI19	36.7	30.0	46.3	63.6
GaitPart [12]	CVPR20	28.2	47.6	44.0	60.7
GaitGL [32]	ICCV21	29.7	22.3	47.3	63.6
SMPLGait [62]	CVPR22	46.3	37.2	-	-
MTSGait [62]	MM22	48.7	37.6	55.3	71.3
GaitBase [11]	CVPR23	64.6	-	60.1	-
DANet [34]	CVPR23	48.0	-	-	-
GaitGCI [10]	CVPR23	50.3	39.5	68.5	80.8
DyGait [54]	ICCV23	66.3	56.4	71.4	83.2
HSTL [52]	ICCV23	61.3	55.5	62.7	76.6
GaitMoE-T(ours)	-	**71.3**	**62.5**	**74.4**	**84.9**
GaitMoE-B(ours)	-	**73.7**	**66.2**	**79.6**	**89.1**

5.2 Implementation Details

We provide details about the training process. **Inputs.** All datasets are resized to 64×44. In addition, we employ spatial alignment module [37] as pre-processing to re-align input silhouettes for OccGait and OccCASIA-B. We adopt batch size [P, K] and the number of iterations, $[8, 16], 40K$ for OccGait, OccCASIA-B, $[32, 4], 60K$ for Gait3D, and $[32, 4], 180K$ for GREW. We sample 30 frames of each gait sequence in the training stage and all frames are used for inference. **Network.** For OccGait and OccCASIA-B, we stack three 2D convolution blocks as our Baseline with the number of channels (64, 128, 256). Each 2D convolution block is followed by an MTE. After Horizontal Pooling with the part parameter of 64, we set individual MAE for each part as in Fig. 4(b) and $\mathcal{M}$ is set to 16. For Gait3D and GREW, we replace our Baseline with GaitBase-like architecture (4 residual blocks or 10 residual blocks) [11,18], setting channels to (64, 128, 256, 512). More Details are shown in *Supplementary Materials*. **Optimization.** GaitMoE is an end-to-end joint training framework only with ID labels. We use the optimizer of SGD with an initial learning rate of 0.1, which is decreasing by a factor of 0.1 per $[10K, 20K, 30K]$, $[10K, 20K, 30K]$, $[20K, 40K, 50K]$, $[80K, 120K, 150K]$ for OccGait, OccCASIA-B, Gait3D and GREW, respectively.

Table 4: Impact of MTE and MAE on OccGait, OccCASIA-B and Gait3D.

		OccGait				OccCASIA-B				Gait3D	
MTE	MAE	NM	CA	CR	ST	NO	CO	SO	DO	Rank-1	mAP
		87.7	69.0	75.1	77.2	85.8	75.9	79.6	80.0	59.2	48.6
✓		89.7	78.3	77.3	81.2	88.2	76.9	82.1	83.6	66.2	56.6
	✓	89.4	78.4	77.4	81.8	87.6	78.3	81.7	82.5	67.2	57.3
✓	✓	**91.4**	**82.1**	**79.9**	**84.7**	**89.5**	**80.5**	**83.9**	**85.3**	**71.3**	**62.5**

Table 5: The number of Action Experts in MAE on OccGait.

MTE	MAE	NM	CA	CR	ST	Mean
✓	1	89.5	78.3	78.0	82.1	82.0
✓	8	90.6	80.7	79.1	83.3	83.4
✓	16	**91.4**	**82.1**	**79.9**	**84.7**	**84.5**
✓	32	90.8	81.2	79.9	84.6	84.1

For joint loss, we set $\beta = 0.1$ for OccGait, OccCASIA-B, and $\beta = 1.0$ for Gait3D and GREW. All the models are trained on NVIDIA 8×3090 GPUs.

5.3 Main Results

To evaluate the effectiveness of GaitMoE, we first compare with other state-of-the-art methods on OccGait and OccCASIA-B with controlled occlusions, and then further make comparisons on Gait3D and GREW with complex covariates (*e.g.*, uncertain occlusions).

Comparison on OccGait. To qualitatively and quantitatively validate GaitMoE, we build the real-scenario gait database, OccGait, containing a wide range of occlusions and explicit annotations. Meanwhile, spatial and scale misalignment may occur in all occlusion scenarios. Tab. 1 shows that GaitMoE outperforms other state-of-the-art methods under all types of occlusions. Notably, CR causes interference from other pedestrians to overshadow the main subject information at certain angles, causing the model to overly focus on the obstructer, not the target. However, the Average results still show SoTA in the main manuscript, which validates the occlusion-solving traits.

Comparison on OccCASIA-B. Tab. 2 shows the overall results where set-based and temporal-based gait recognition methods (*i.e.*, GaitSet, GaitPart, GaitGL, GaitBase) suffer from severe degradation under the occlusion scenario, comparing to their performances in their paper on holistic CASIA-B. Although STOR with silhouette registration alleviates the misalignment issue, the complex walking patterns under occlusions make the feature extraction difficult. Our proposed method outperforms all methods by a large margin, especially in the extremely challenging cloth-changing (CL) condition, *e.g.*, exceeds GaitBase by 12.0% in **NO**, and 16.8% in **DO**. The experimental results demonstrate that GaitMoE effectively filters invalid occluded actions and extracts robust actions.

Comparisons on Gait3D and GREW. We have validated the effectiveness of our method on the controlled occlusion environment (*i.e.*, OccCASIA-B and OccGait). The large-scale wild gait databases also provide complex and uncertain

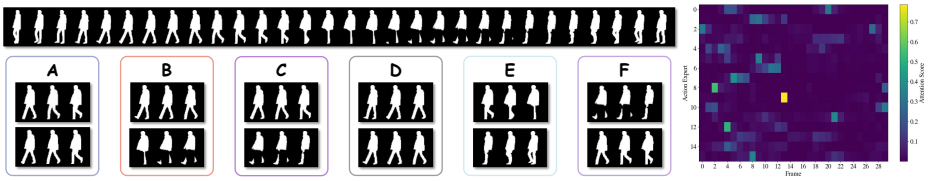


Fig. 6: The visualization of action composition. Here are an occluded gait sequence, action proposals and the selection map of action anchors. Alphabet denotes the action proposal. Each rectangle denotes one action proposal, and each row within one action proposal denotes an action anchor.

occlusion scenarios. As shown in Tab. 3, GaitMoE-T (4 residual blocks) and GaitMoE-B (10 residual blocks) also achieve the highest performance among state-of-the-art methods, which further proves the generalizability and practicality of our method.

5.4 Ablation Study

In this section, we mainly qualitatively and quantitatively validate the MTE and MAE in OccGait, OccCASIA-B and Gait3D. Besides, we provide the visualization of action detection on OccCASIA-B.

The Effectiveness of MTE and MAE. Tab. 4 shows that each of MTE and MAE can extract better dynamic information. Especially, the combination of MTE and MAE *i.e.*, GaitMoE, achieves a significant increase over Baseline, which proves that our proposed method discovers the representative actions by first coarse action extraction (*i.e.*, action anchors), and then fine-grained action extraction (*i.e.*, action proposals). Besides, we select (8, 4) for ($\mathcal{S}$, $\mathcal{K}$) on temporal experts since a smaller $\mathcal{S}$ or larger $\mathcal{K}$ degrades original information, and a larger $\mathcal{S}$ or smaller $\mathcal{K}$ restricts dynamic information. More Details are shown in *Supplementary Materials*. We also quantify the impact on the number of action experts. Tab. 5 illustrates that more experts may result in redundant actions, while fewer experts may not be able to capture the diverse actions. We select 16 as the parameter of GaitMoE.

The Visualization of Action Composition. To better understand and interpret the action detection in gait recognition, we visualize the key module MAE of action detection in Fig. 6, we select part index 60, action anchor with 2 dilated ratios and 6 action proposals as the example. MTE enables to infer the occluded region information by consecutive frames. For an example in E, the right occluded frame can hallucinate the missing region through the middle frame. MAE enables to select and integrate the most discriminative similar action anchors from different gait cycles. For an example in F, although missing information occurs, MAE combines multiple similar action anchors to further confirm this action type (*i.e.*, swinging legs), integrating occluded information.

Trade-off between Accuracy and Efficiency. As Fig. 7 shows, we compare all parameters of these models. For FLOPs, models input a 30-frame gait se-

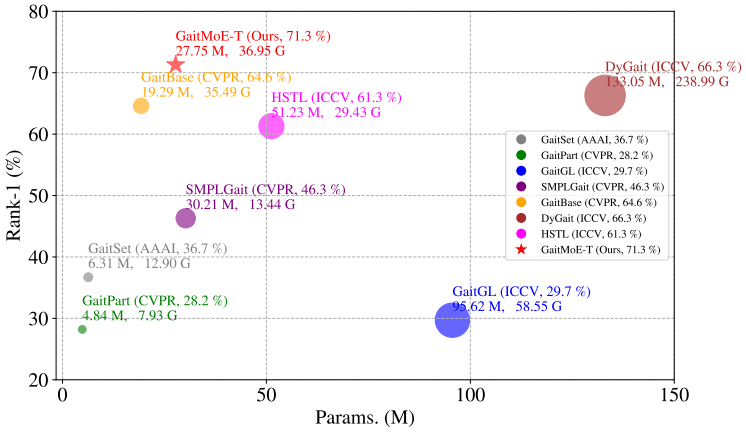


Fig. 7: The model comparisons on accuracy and efficiency. Rank-1 (%), Param. (M) and FLOPs. (G) on Gait3D.

quence without Separate FCs and BNNeck for significant comparisons. GaitMoE makes a trade-off and achieves SoTA without substantially increasing computational cost. In contrast, DyGait demands significant computation due to 3D convolutions and GaitBase offers better efficiency with 2D convolutions but lower accuracy.

5.5 Conclusion and Limitations

In this paper, we introduce an action detection perspective where a gait sequence is regarded as a composition of actions, allowing holistic body regions to infer occluded body regions and information integration between holistic and occluded actions. To detect accurate actions under complex occlusion scenarios, we propose an Action Detection Based Mixture of Experts (GaitMoE) to leverage dynamic contextual information *i.e.*, gait continuity and gait cycle, to construct action anchors and action proposals. To obtain qualitative and quantitative occlusion analysis, we propose a novel Occluded Gait Recognition benchmark (OccGait) as a pioneering database with a wide range of occlusion scenarios and explicit annotations. Extensive experimental results have demonstrated that GaitMoE effectively captures accurate and robust actions for occluded gait recognition. In addition. we provide some limitations where the selection map of action anchors shown in Fig. 6 shows that some of the experts in GaitMoE present redundancy and similarity. We will explore the optimization and design for expert selection in the future.

Acknowledgment

This work is jointly supported by National Natural Science Foundation of China (62276025, 62206022), Beijing Municipal Science & Technology Commission (Z23

1100007423015) and Shenzhen Technology Plan Program (KQTD2017033109321 7368).

References

1. An, W., Yu, S., Makihara, Y., Wu, X., Xu, C., Yu, Y., Liao, R., Yagi, Y.: 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. Bashir, K., Xiang, T., Gong, S., Mary, Q., et al.: Gait representation using flow fields. In: BMVC. pp. 1–11 (2009)
3. Chai, T., Li, A., Zhang, S., Li, Z., Wang, Y.: Lagrange motion analysis and view embeddings for improved gait recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20249–20258 (2022)
4. Chao, H., He, Y., Zhang, J., Feng, J.: Gaitset: Regarding gait as a set for cross-view gait recognition. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 8126–8133 (2019)
5. Chattopadhyay, P., Sural, S., Mukherjee, J.: Frontal gait recognition from occluded scenes. Pattern Recognition Letters **63**, 9–15 (2015)
6. Chen, C., Liang, J., Zhao, H., Hu, H., Tian, J.: Frame difference energy image for gait recognition with incomplete silhouettes. Pattern Recognition Letters **30**(11), 977–984 (2009)
7. Chen, X., Li, H., Li, M., Pan, J.: Learning a sparse transformer network for effective image deraining. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5896–5905 (2023)
8. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1290–1299 (2022)
9. Delgado-Escano, R., Castro, F.M., R. Cózar, J., Marin-Jimenez, M.J., Guil, N.: Mupeg—the multiple person gait framework. Sensors **20**(5), 1358 (2020)
10. Dou, H., Zhang, P., Su, W., Yu, Y., Lin, Y., Li, X.: Gaitgci: Generative counterfactual intervention for gait recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5578–5588 (2023)
11. Fan, C., Liang, J., Shen, C., Hou, S., Huang, Y., Yu, S.: Opengait: Revisiting gait recognition towards better practicality. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9707–9716 (2023)
12. Fan, C., Peng, Y., Cao, C., Liu, X., Hou, S., Chi, J., Huang, Y., Li, Q., He, Z.: Gaitpart: Temporal part-based model for gait recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14225–14233 (2020)
13. Fedus, W., Zoph, B., Shazeer, N.: Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. The Journal of Machine Learning Research **23**(1), 5232–5270 (2022)
14. Fu, Y., Meng, S., Hou, S., Hu, X., Huang, Y.: Gpgait: Generalized pose-based gait recognition. arXiv preprint arXiv:2303.05234 (2023)
15. Gross, R.: The cmu motion of body (mobo) database. Carnegie Mellon University, The Robotics Institute (2001)

16. Guo, H., Ji, Q.: Physics-augmented autoencoder for 3d skeleton-based gait recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19627–19638 (2023)
17. Gupta, A., Chellappa, R.: You can run but not hide: Improving gait recognition with intrinsic occlusion type awareness. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 5893–5902 (2024)
18. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
19. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737 (2017)
20. Hofmann, M., Geiger, J., Bachmann, S., Schuller, B., Rigoll, G.: The tum gait from audio, image and depth (gaid) database: Multimodal recognition of subjects and traits. Journal of Visual Communication and Image Representation **25**(1), 195–206 (2014)
21. Hofmann, M., Wolf, D., Rigoll, G.: Identification and reconstruction of complete gait cycles for person identification in crowded scenes. In: Proc. Intern. Conf. on Computer Vision Theory and Applications (VISAPP), Algarve, Portugal (2011)
22. Hossain, M.A., Makihara, Y., Wang, J., Yagi, Y.: Clothing-invariant gait identification using part-based clothing categorization and adaptive weight control. Pattern Recognition **43**(6), 2281–2291 (2010)
23. Hou, S., Cao, C., Liu, X., Huang, Y.: Gait lateral network: Learning discriminative and compact representations for gait recognition. In: European conference on computer vision. pp. 382–398. Springer (2020)
24. Hou, S., Liu, X., Cao, C., Huang, Y.: Set residual network for silhouette-based gait recognition. IEEE Transactions on Biometrics, Behavior, and Identity Science **3**(3), 384–393 (2021)
25. Huang, X., Zhu, D., Wang, H., Wang, X., Yang, B., He, B., Liu, W., Feng, B.: Context-sensitive temporal feature learning for gait recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12909–12918 (2021)
26. Iwama, H., Okumura, M., Makihara, Y., Yagi, Y.: 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 (2012)
27. Lepikhin, D., Lee, H., Xu, Y., Chen, D., Firat, O., Huang, Y., Krikun, M., Shazeer, N., Chen, Z.: Gshard: Scaling giant models with conditional computation and automatic sharding. arXiv preprint arXiv:2006.16668 (2020)
28. Li, B., Yang, J., Ren, J., Wang, Y., Liu, Z.: Sparse fusion mixture-of-experts are domain generalizable learners. arXiv e-prints pp. arXiv–2206 (2022)
29. Li, N., Zhao, X.: A multi-modal dataset for gait recognition under occlusion. Applied Intelligence **53**(2), 1517–1534 (2023)
30. Liang, J., Fan, C., Hou, S., Shen, C., Huang, Y., Yu, S.: Gaitedge: Beyond plain end-to-end gait recognition for better practicality. arXiv preprint arXiv:2203.03972 (2022)
31. Lin, B., Zhang, S., Bao, F.: Gait recognition with multiple-temporal-scale 3d convolutional neural network. In: Proceedings of the 28th ACM international conference on multimedia. pp. 3054–3062 (2020)
32. Lin, B., Zhang, S., Yu, X.: Gait recognition via effective global-local feature representation and local temporal aggregation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14648–14656 (2021)

33. Lin, C., Xu, C., Luo, D., Wang, Y., Tai, Y., Wang, C., Li, J., Huang, F., Fu, Y.: Learning salient boundary feature for anchor-free temporal action localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3320–3329 (2021)
34. Ma, K., Fu, Y., Zheng, D., Cao, C., Hu, X., Huang, Y.: Dynamic aggregated network for gait recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22076–22085 (2023)
35. Makihara, Y., Mannami, H., Yagi, Y.: Gait analysis of gender and age using a large-scale multi-view gait database. In: Computer Vision–ACCV 2010: 10th Asian Conference on Computer Vision, Queenstown, New Zealand, November 8-12, 2010, Revised Selected Papers, Part II 10. pp. 440–451. Springer (2011)
36. Mustafa, B., Riquelme, C., Puigcerver, J., Jenatton, R., Houlsby, N.: Multimodal contrastive learning with limoe: the language-image mixture of experts. Advances in Neural Information Processing Systems **35**, 9564–9576 (2022)
37. Peng, Y., Cao, C., He, Z.: Occluded gait recognition. In: 2023 International Joint Conference on Neural Networks (IJCNN). pp. 1–8. IEEE (2023)
38. Riquelme, C., Puigcerver, J., Mustafa, B., Neumann, M., Jenatton, R., Susano Pinto, A., Keysers, D., Houlsby, N.: Scaling vision with sparse mixture of experts. Advances in Neural Information Processing Systems **34**, 8583–8595 (2021)
39. Roller, S., Sukhbaatar, S., Weston, J., et al.: Hash layers for large sparse models. Advances in Neural Information Processing Systems **34**, 17555–17566 (2021)
40. Sepas-Moghaddam, A., Etemad, A.: Deep gait recognition: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence (2022)
41. Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., Dean, J.: Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. arXiv preprint arXiv:1701.06538 (2017)
42. Shen, C., Fan, C., Wu, W., Wang, R., Huang, G.Q., Yu, S.: Lidargait: Benchmarking 3d gait recognition with point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1054–1063 (2023)
43. Shi, D., Zhong, Y., Cao, Q., Ma, L., Li, J., Tao, D.: Tridet: Temporal action detection with relative boundary modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18857–18866 (2023)
44. Song, C., Huang, Y., Wang, W., Wang, L.: Casia-e: a large comprehensive dataset for gait recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(3), 2801–2815 (2022)
45. Takemura, N., Makihara, Y., Muramatsu, D., Echigo, T., Yagi, Y.: Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition. IPSJ transactions on Computer Vision and Applications **10**, 1–14 (2018)
46. Tan, D., Huang, K., Yu, S., Tan, T.: Efficient night gait recognition based on template matching. In: 18th international conference on pattern recognition (ICPR'06). vol. 3, pp. 1000–1003. IEEE (2006)
47. Teepe, T., Gilg, J., Herzog, F., Hörmann, S., Rigoll, G.: Towards a deeper understanding of skeleton-based gait recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1569–1577 (2022)
48. Teepe, T., Khan, A., Gilg, J., Herzog, F., Hörmann, S., Rigoll, G.: Gaitgraph: Graph convolutional network for skeleton-based gait recognition. In: 2021 IEEE International Conference on Image Processing (ICIP). pp. 2314–2318. IEEE (2021)
49. Tsuji, A., Makihara, Y., Yagi, Y.: Silhouette transformation based on walking speed for gait identification. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 717–722. IEEE (2010)

50. Uddin, M.Z., Muramatsu, D., Takemura, N., Ahad, M.A.R., Yagi, Y.: Spatio-temporal silhouette sequence reconstruction for gait recognition against occlusion. IPSJ Transactions on Computer Vision and Applications **11**(1), 1–18 (2019)
51. Uddin, M.Z., Ngo, T.T., Makihara, Y., Takemura, N., Li, X., Muramatsu, D., Yagi, Y.: The ou-isir large population gait database with real-life carried object and its performance evaluation. IPSJ Transactions on Computer Vision and Applications **10**(1), 1–11 (2018)
52. Wang, L., Liu, B., Liang, F., Wang, B.: Hierarchical spatio-temporal representation learning for gait recognition. arXiv preprint arXiv:2307.09856 (2023)
53. Wang, L., Tan, T., Ning, H., Hu, W.: Silhouette analysis-based gait recognition for human identification. IEEE transactions on pattern analysis and machine intelligence **25**(12), 1505–1518 (2003)
54. Wang, M., Guo, X., Lin, B., Yang, T., Zhu, Z., Li, L., Zhang, S., Yu, X.: Dygait: Exploiting dynamic representations for high-performance gait recognition. arXiv preprint arXiv:2303.14953 (2023)
55. Wang, W., Bao, H., Dong, L., Bjorck, J., Peng, Z., Liu, Q., Aggarwal, K., Mohammed, O.K., Singhal, S., Som, S., et al.: Image as a foreign language: Beit pretraining for all vision and vision-language tasks. arXiv preprint arXiv:2208.10442 (2022)
56. Xu, C., Makihara, Y., Li, X., Yagi, Y.: Occlusion-aware human mesh model-based gait recognition. IEEE transactions on information forensics and security **18**, 1309–1321 (2023)
57. Xu, C., Makihara, Y., Ogi, G., Li, X., Yagi, Y., Lu, J.: The ou-isir gait database comprising the large population dataset with age and performance evaluation of age estimation. IPSJ Transactions on Computer Vision and Applications **9**(1), 1–14 (2017)
58. Xu, C., Tsuji, S., Makihara, Y., Li, X., Yagi, Y.: Occluded gait recognition via silhouette registration guided by automated occlusion degree estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3199–3209 (2023)
59. Yang, L., Peng, H., Zhang, D., Fu, J., Han, J.: Revisiting anchor mechanisms for temporal action localization. IEEE Transactions on Image Processing **29**, 8535–8548 (2020)
60. Yu, S., Tan, D., Tan, T.: A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In: 18th International Conference on Pattern Recognition (ICPR'06). vol. 4, pp. 441–444. IEEE (2006)
61. Zhang, C., Chen, X.P., Han, G.Q., Liu, X.J.: Spatial transformer network on skeleton-based gait recognition. Expert Systems p. e13244 (2023)
62. Zheng, J., Liu, X., Liu, W., He, L., Yan, C., Mei, T.: Gait recognition in the wild with dense 3d representations and a benchmark. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20228–20237 (2022)
63. Zhu, H., Zheng, W., Zheng, Z., Nevatia, R.: Gaitref: Gait recognition with refined sequential skeletons. arXiv preprint arXiv:2304.07916 (2023)
64. Zhu, H., Zheng, Z., Nevatia, R.: Gait recognition using 3-d human body shape inference. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 909–918 (2023)
65. Zhu, Z., Guo, X., Yang, T., Huang, J., Deng, J., Huang, G., Du, D., Lu, J., Zhou, J.: Gait recognition in the wild: A benchmark. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 14789–14799 (2021)