GPGait: Generalized Pose-based Gait Recognition

Yang Fu[1]*, Shibei Meng[1]*, Saihui Hou[1,2]†, Xuecai Hu[1], Yongzhen Huang[1,2]†
[1] School of Artificial Intelligence, Beijing Normal University [2] WATRIX.AI
{yangfu, mengshibei}@mail.bnu.edu.cn, {housaihui, huxc1208, huangyongzhen}@bnu.edu.cn

Abstract

Recent works on pose-based gait recognition have demonstrated the potential of using such simple information to achieve results comparable to silhouette-based methods. However, the generalization ability of pose-based methods on different datasets is undesirably inferior to that of silhouette-based ones, which has received little attention but hinders the application of these methods in real-world scenarios. To improve the generalization ability of pose-based methods across datasets, we propose a ***G****eneralized* ***P****ose-based* ***Gait*** *recognition (****GPGait****) framework. First, a Human-Oriented Transformation (HOT) and a series of Human-Oriented Descriptors (HOD) are proposed to obtain a unified pose representation with discriminative multi-features. Then, given the slight variations in the unified representation after HOT and HOD, it becomes crucial for the network to extract local-global relationships between the keypoints. To this end, a Part-Aware Graph Convolutional Network (PAGCN) is proposed to enable efficient graph partition and local-global spatial feature extraction. Experiments on four public gait recognition datasets, CASIA-B, OUMVLP-Pose, Gait3D and GREW, show that our model demonstrates better and more stable cross-domain capabilities compared to existing skeleton-based methods, achieving comparable recognition results to silhouette-based ones. Code is available at* https://github.com/BNU-IVC/FastPoseGait.

1. Introduction

Gait recognition is an essential task in the human identification field. Compared with existing biometric identification methods, such as face, fingerprint, and iris recognition, it can capture long-distance gait features without the cooperation of subjects. Existing studies of gait recognition can mainly be divided into two streams, appearance-based [4–6, 11, 12, 17, 39] and model-based methods [19, 21, 23, 31, 32, 38]. Specifically, the appearance-based methods try to directly learn gait features from the silhouette sequences and have been the dominant approach for a long time. And the model-based methods try to explicitly estimate human body structures (*e.g.*, keypoints or 3D mesh) for gait recognition. Despite that the performance is inferior to the appearance-based methods at the moment, the model-based methods have their own advantage of being robust to carrying and clothing, which is appealing to practical applications and thus deserves continuous attention.

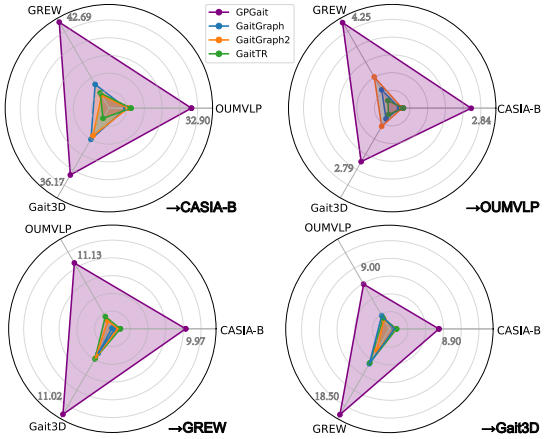

Figure 1. GPGait achieves state-of-the-art generalization performance across four popular gait datasets. In each subgraph, three vertices correspond to the source domain. Arrows (→) point to the target domain. Triangles in different colors represent the generalization ability of different models under cross-domain settings.

Model-based methods mostly take human poses (*i.e.*, keypoints) as the input which encodes visual clues for body structure and proportion in an explicit way. Benefiting from the rapid development in pose estimation [2, 7, 16, 29] and graph-based models (GCN [28, 36] and Transformer [24]), pose-based methods have achieved fairly surprising results in some cases [31, 32, 38], *e.g.*, GaitTR [38] introduces Spatial Transformer to establish overall spatial relationships between keypoints. The model achieves much higher accuracy than previous pose-based methods [31, 32] on CASIA-B [37], even surpassing the accuracy of appearance-based methods [4, 6, 11, 20] in clothes-changing conditions.

*Equal contribution
†Corresponding Author

However, a vital problem is ignored in these research, *i.e.*, generalization ability. Through a preliminary study as shown in Fig.1, we find that the performance of these methods tends to drastically degrade when testing gait sequences from unseen environments, limiting the application in realistic scenarios. Our analysis suggests that the decline in cross-domain performance can be linked to various factors: (1) scale variations due to the distance to cameras, (2) tilt and horizontal views due to the deployment of cameras, (3) offsets within the camera coordinate system. All these factors can cause intense changes in data distributions, resulting in a dramatic decline in performance when there are variations in cameras and environments.

To promote the research on model-based gait recognition, we aim to design a framework for **G**eneralized **P**ose-based **Gait** Recognition (GPGait) that can effectively improve the generalization ability of pose-based methods. Particularly, we try to solve the problem from two perspectives: *a human-oriented input that is comparable across different cameras*, and *a part-aware model that extracts fine-grained body features for recognition*.

In terms of input format, we propose a Human-Oriented Transformation (HOT) and a series of Human-Oriented Descriptors (HOD) to obtain a unified and enriched representation. Specifically, HOT consists of three steps, namely affine transform, body rescale, and body alignment, through which the original skeleton sequences captured in the camera coordinate system are transformed into unified representations in the human-oriented coordinate system. Then, to enrich the input, we carefully design a module named Human-Oriented Descriptors (HOD) to generate individual-invariant features of bone and angle to explicitly reflect the body proportion and structure.

Regarding the modeling, we argue that the fine-grained learning for different human parts is the key to extracting discriminative gait features and improving the generalization ability. Although we can obtain a unified representation with HOT and HOD, the uniform gait expression exhibits less variation over time compared to the original pose sequence. Therefore, it is imperative to capture the local-global relationships between the keypoints, where local features can capture the slight changes in pose and the global ones can represent the entire human structure. Inspired by the recent progress in the field of gait recognition [3, 4, 6, 11, 20] and domain generalization [5], we design a Part-Aware Graph Convolutional Network (PAGCN) which can efficiently implement graph partition and local-global relationship construction through mask operations on the adjacency matrix.

To summarize, we make the following three major contributions:

- In the HOT module, a series of human-oriented operations are proposed to facilitate a uniform input that overcomes problems caused by various environmental covariances. The input is further enriched in the HOD module to explicitly reflect the skeleton structure and movement.

- We present PAGCN to achieve efficient graph partition and local-global feature relation extraction under the unified pose representation. With different part-specific masks and a well-designed network structure, the method can not only capture fine-grained features and distinct local relations but also reduce the amount of calculations and the number of parameters required.

- Extensive experiments demonstrate that the proposed GPGait framework achieves state-of-the-art generalization results in all scenarios (indoor and outdoor) under cross-domain settings. Especially, the result of the cross-domain test on GREW→CASIA-B outperforms previous methods by a large margin of 34.69%.

2. Related Work

2.1. Gait Recognition

Gait recognition methods can be classified into two main categories: appearance-based [4, 6, 9, 11, 12, 17] and model-based methods [14, 15, 18, 19, 31, 32, 35] depending on the type of data input. Appearance-based methods usually rely on silhouette sequences [4, 6, 11, 12, 17] or their transformation, such as Gait Energy Image (GEI) [9] and Chrono-Gait Image (CGI) [34]. Model-based methods, on the other hand, represent the human body as mesh [14, 15, 35] or a set of keypoints [18, 19, 31, 32].

Silhouette-based methods GaitSet [4] aggregates temporal information in silhouette sequences using a statistical function to adapt to different frame rates. GaitPart [6] uses Focal Convolutional Layer to extract fine-grained local features and Micro-motion Template Builder with different window sizes to extract local temporal information. GaitGL [20] uses 3D Convolution to extract global and local spatiotemporal information. LagrangeGait [3] adds a local motion extractor and a viewpoint branch based on GaitGL to get more discriminative local temporal information.

Pose-based methods PoseGait [19] employs 3D pose information to generate multi-feature vectors and uses a CNN to extract the gait information in both spatial and temporal dimensions. GaitGraph [32] and GaitGraph2 [31] adopt Graph Convolutional Network for gait recognition, treating keypoints as nodes and limbs as edges to form a topology graph. GaitTR [38] and GaitMixer [23] use the self-attention [33] to explore long-range spatial correlations, and temporal convolution with a large kernel size to extract long temporal information.

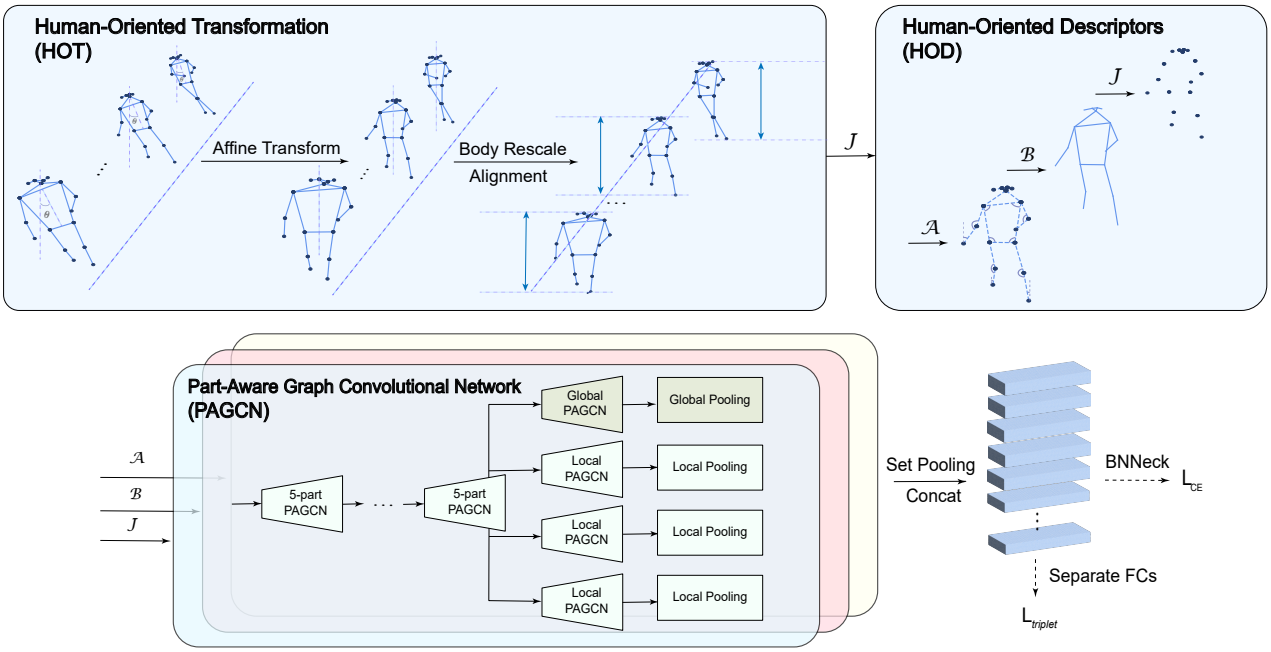

Figure 2. The framework of GPGait. The original pose sequence is first transformed into a unified representation by Human-Oriented Transformation. Then, angle, bone and joint features generated by Human-Oriented Descriptors are learned dependently through a multi-branch network named PAGCN. The output features are concatenated in part dimension and learned by a fully connected layer separately. Finally, a widely used BNNeck [22] is adopted to adjust feature space. Triplet loss [10] and cross-entropy loss are utilized to supervise the whole training process.

2.2. Generalization in Gait Recognition

The generalization ability of the model is a practically significant concern. From GaitSet [4] to LagrangeGait [3], the size of all the silhouette data is aligned based on methods in [30]. Each silhouette is put in the central position to generate a uniform representation, which is more stable and robust than the original input. Cross-domain experiments are done by [40] with the backbone of GaitSet between popular public datasets and Gait3D to prove the importance of in-the-wild datasets. GaitEdge [17] evaluates their method on cross-dataset settings to demonstrate the irrelevance to RGB texture and color. Self-supervised learning has also been employed by [5] to improve the ability to generalize in unseen domains. In skeleton-based methods, PoseGait [19] and CNN-Pose [1] use the human neck to align the skeletons and spine length to rescale the body. However, the spine length is sensitive to occlusion which introduces noise and even errors. Rashmi *et al*. [25] uses the dataset-independent statistics to rescale the skeleton, which still lacks the ability to generalize for the scale changes of skeletons across datasets.

3. Method

In this section, we first present the pipeline of GPGait. Then we display a Human-Oriented Transformation (HOT) and a series of Human-Oriented Descriptors (HOD) which generate a unified and enriched pose representation, followed by the description of the Part-Aware Graph Convolutional Network (PAGCN) that is specially designed for efficient fine-grained feature extraction.

3.1. Pipeline

As shown in Fig.2, the pose sequences are first fed into Human-Oriented Transformation (HOT) to get a unified pose representation. Then a series of Human-Oriented Descriptors (HOD) are obtained to generate discriminative and domain-invariant features. Due to the different distributions of joints, angles and bones, we perform different feature extractions with a parameter-independent multi-branch architecture. The multi-features are learned through Part-Aware Graph Convolutional Network (PAGCN) which can enable efficient human graph partitioning and local-global relationship building. The pooling operations on semantic body parts are utilized at the end of the network to get the final embedding for recognition. We use the separate triplet and cross-entropy losses to supervise the training process.

3.2. Human-Oriented Transformation

In this part, as Fig.2 shows, we transform the original data from various cameras to a stable and unified pose representation. Specifically, HOT consists of three phases of affine transform, body rescale and alignment to eliminate

the environment covariance like viewpoints, distance away from the camera, offset noises, *et al*.

First, an affine transform is used to overcome the problem of slant skeletons resulting from different camera views. Formally, we suppose the original 2D frame set of one subject sequence is $P = \{p \in \mathbb{R}^{T_{in} \times V_{in} \times C_{in}}\}$, where T_{in}, V_{in}, C_{in} denote the number of input frames, joints, and coordinates, respectively. We regard the neck $\mathbf{p_{neck}}$ as the median position of the right shoulder and left shoulder. Similarly, the location of hip joint $\mathbf{p_{hip}}$ is the average of the right hip and the left hip. The spine is considered as the line between the neck and hip. We take the spine as the axis, and the neck as a center to transform the inclined skeleton perpendicular to the ground. The rotation angle θ is calculated as:

$$\theta = \arctan\left(\frac{p_{neck}^{c_x} - p_{hip}^{c_x}}{p_{neck}^{c_y} - p_{hip}^{c_y}}\right), \tag{1}$$

where c_x is the coordinate of x-axis and c_y is the coordinate of y-axis. Affine transform is applied exclusively to sequences with serious slant problems when the angle θ is larger than the threshold ϕ. Otherwise, the original pose P is directly adapted. We assume the concatenation of P and $\mathbf{p_{neck}}$ in C dimension is $D = \{\mathbf{d_i} | i = 1, 2, ..., V_{in}\}$. The process can be formulated as:

$$P_a = \begin{cases} M_a D, & \theta \geqslant \phi \\ P, & \text{otherwise} \end{cases} \tag{2}$$

$$M_a = \begin{bmatrix} \cos\theta & -\sin\theta & 1-\cos\theta & \sin\theta \\ \sin\theta & \cos\theta & -\sin\theta & 1-\cos\theta \end{bmatrix}, \tag{3}$$

$$\mathbf{d_i} = \begin{bmatrix} p_i^{c_x} & p_i^{c_y} & p_{neck}^{c_x} & p_{neck}^{c_y} \end{bmatrix}^\top, \tag{4}$$

where P_a is the output of the affine transform, M_a is the affine transform matrix, p_i^k denotes the i-th joint at the k-th coordinate, and $k \in \{c_x, c_y\}$. After the series of operations, the slanted human spine in all the sequences is perpendicular to the ground.

Second, considering the scale of the skeleton in each frame is variational, we propose a simple yet effective body-rescale (in Eq.5) method to achieve a uniform height of skeletons by dividing the difference between the maximum and minimum keypoint values along the vertical axis.

$$\mathbf{p_i'} = \mathbf{pa_i} * \frac{h_{unif}}{\max(\mathbf{pa^{c_y}}) - \min(\mathbf{pa^{c_y}})}, \tag{5}$$

where $\mathbf{pa}$ is taken from P_a and h_{unif} is a factor to control the uniform body height.

Third, an alignment operation (in Eq.6) is performed on the human joints, in which the camera coordinate is converted to a human-oriented coordinate:

$$\mathbf{j_i} = \mathbf{p_i'} - \mathbf{p'_{neck}}. \tag{6}$$

The unified pose sequence we get in this module is $\mathcal{J} = \{\mathbf{j_i} | i = 1, 2, ..., V_{in}\}$.

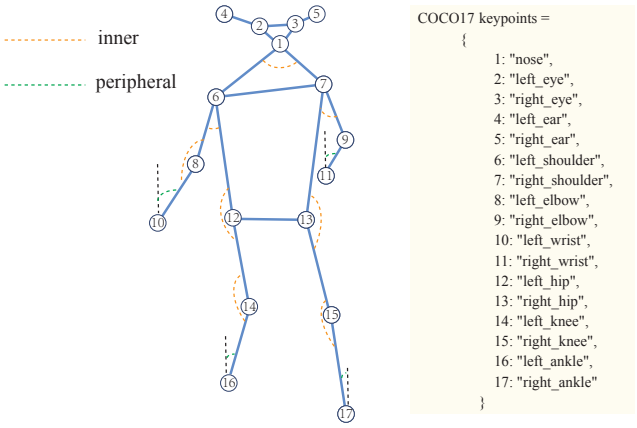

Figure 3. The detailed description of COCO2017 format with 17 keypoints. The figure also illustrates the sketch of inner and peripheral angles calculated in Human-Oriented Descriptors.

3.3. Human-Oriented Descriptors

Recent works [19,31] have noticed that multi-input characterization can lead to improved recognition performance. However, these works tend to underestimate the potential of the second-order information related to bones and angles. Specifically, the input of bone and angle features usually have different distributions from that of joint features, which can result in a totally different relation-learning process in GCN. The direct input of joints or the earlier fusion leads to distinct degradation in performance. So we explicitly add information of bone and angle to learn discriminative gait signatures.

For the bone features (in Eq.7), we consider the human bone as vectors, i.e., $\mathcal{B} = \{\mathbf{b_i} | i = 1, 2, ..., V_{in}\}$.

$$\mathbf{b_i} = \mathbf{j_i} - \mathbf{j_{adj(i)}}, \tag{7}$$

where $adj(i)$ denotes the adjacency joint of i-th joint. For angles of the skeletons, unlike previous works [28,31] using angles between bones and the horizontal or vertical line, we design a human-oriented angle-calculating method. Specifically, we use inner angles and peripheral angles shown in Fig.3, i.e., $\mathcal{A} = \{a_i | i = 1, 2, ..., V_{in}\}$.

$$a_i = \begin{cases} \arccos\left(\dfrac{s_l^2(i) + s_r^2(i) - s_{opp}^2(i)}{2 * s_l(i) * s_r(i)}\right), & \text{inner} \\ \arctan\left(\dfrac{j_i^{c_x} - j_{adj(i)}^{c_x}}{j_i^{c_y} - j_{adj(i)}^{c_y}}\right), & \text{peripheral} \end{cases} \tag{8}$$

where the $s_l(i), s_r(i), s_{opp}(i)$ denotes the length of adjacent sides and the opposite side of the i-th joint. $adj(i)$ denotes the adjacency joints of peripheral skeleton joints and j_i^k denotes i-th joint at k-th coordinate. Specifically, taking the ⑭ joint in Fig.3 as an example, $[s_l, s_r]$ denotes the length of [⑭-⑫, ⑭-⑯], while s_{opp} denotes the length of ⑫-⑯.

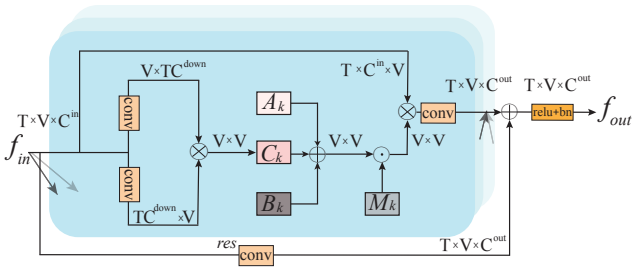

Figure 4. The detailed structure of Part-Aware Graph Convolutional Network (PAGCN) block.

3.4. Part-Aware Graph Convolutional Network

PAGCN Block Although we get a unified representation to eliminate the camera covariance across datasets after HOT and HOD, there are actually not so many changes in body scale and offset when the subject is moving. So, it is important to capture the local variation and the overall structural information for the network. However, the traditional methods using graph convolutional network [31, 32] or spatial transformer [38] only construct the global relations between the keypoints, in which the slight move of some keypoints around the aligned center would be ignored. According to this, we propose a Part-Aware Graph Convolutional Network (PAGCN) to get local-global representations by using different partitioned masks. Our PAGCN block can be formulated in Eq.9 and described in Fig.4. The partitioned masks following the structure of the human body are shown in Fig.5.

$$f_{out} = \sum_{k}^{K_v} W_k(f_{in}(A_k + B_k + C_k) \odot M_k), \qquad (9)$$

where A_k denotes the predefined adjacency matrix of natural human structure, B_k denotes the parameterized adjacency matrix which can be updated in an end-to-end learning manner, C_k denotes the self-attention adjacency matrix which is used to construct global connections of joints for each sequence, the K_v denotes the number of graph subset. $\odot$ stands for element-wise product. Especially, the value of M_k depends on the different body partition strategies G, as shown in:

$$M_k(i_1, i_2) = \begin{cases} 1, & p_{i_1}, p_{i_2} \in g, g \subset G \\ 0, & \text{otherwise} \end{cases} \qquad (10)$$

where $G = \{g_1, g_2, .., g_n\}$, n means the number of different body partition strategies shown in Fig.5. And the global PAGCN block is with $M_k(i_1, i_2) = 1$. Different from the previous work [26], PAGCN allows for explicit graph partition and unleashes the deep GCN's expressive power in the training phase through mask operations, which is verified in the experimental study.

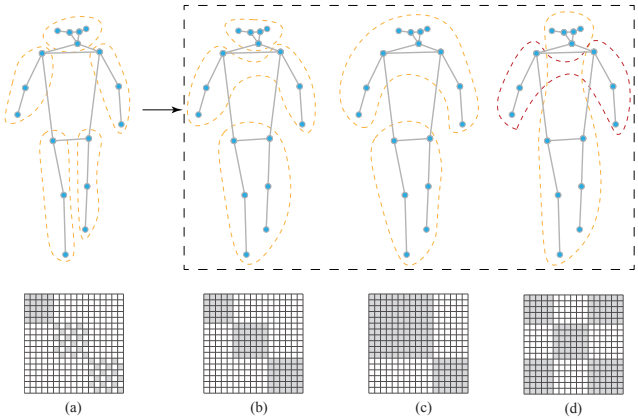

Figure 5. The partition masks following the human body structure. The larger body parts in (b), (c) and (d) are the combinations of 5 small parts in (a). The small black boxes in the adjacency matrices of keypoints refer to 1 and the white ones refer to 0.

PAGCN Backbone The design of the backbone is mainly based on the following considerations: **1)** One part cannot include too few keypoints because each of them contains limited information. **2)** In order to make each part as distinguishable as possible, we try to avoid including the information of keypoints outside of the part during the process of learning the relationships . **3)** The backbone should be light and efficient. **4)** To adapt to diverse human feature distributions and enhance performance, multiple descriptor features should be applied using different parameters. Motivated by these principles, the overall backbone is composed of three parameter-independent branches. And in each branch, we stack 5-part PAGCN blocks with shared parameters in the first few layers to construct relations between keypoints in small parts as shown in Fig.2. In the last few layers, we use larger-part PAGCN blocks to learn the relationships inside the major body parts and ensure better learning of part-specific relationships by avoiding parameter sharing across different partition masks. The whole architecture of the network can not only learn the fine-grained body part features but also prioritizes significant consideration of the second-order joint information.

In previous works, using split feature maps is considered effective for silhouette-based methods in gait recognition tasks. Instead of pyramid mapping [4, 8], we use semantic body feature mapping for the final recognition since every keypoint has its own semantic information and each body part has specific abilities for identification. We perform the pooling operation on the predefined body parts and on the whole body features to get the final embedding. We assume that $f_M \in \mathbb{R}^{T \times V \times C}$ is the output of a last-layer PAGCN block, C is the output channels, T is the number of frames in a sequence and V is the number of keypoints. The feature mapping in spatial dimension can be formulated as:

$$f_{vp}^n = \mathrm{ap}_n(f_M) + \mathrm{mp}_n(f_M), \qquad (11)$$

where $\mathrm{ap}(\cdot)$ refers to the statistical function *average pooling* and $\mathrm{mp}(\cdot)$ refers to *max pooling* along the V dimension. n is the n-th body part. The output $f_{vp}^{n} \in \mathbb{R}^{T\times 1\times C}$. Then, a set pooling function is used to aggregate the temporal information:

$$f_{tp}^{n} = \mathrm{sp}(f_{vp}^{n}), \tag{12}$$

where $\mathrm{sp}(\cdot)$ is the *max pooling* function along the T dimension and $f_{tp}^{n} \in \mathbb{R}^{1\times 1\times C}$. The final representation for recognition is to concatenate the output of each part from three branches including *joint*, *angle* and *bone* branch:

$$F_{branch} = \mathrm{cat}(f_{tp}^{1}, ..., f_{tp}^{n}), \tag{13}$$

$$F_{final} = \mathrm{cat}(F_{joint}, F_{angle}, F_{bone}), \tag{14}$$

3.5. Optimization

During training, the triplet loss [10] $L_{triplet}$ and the cross-entropy loss L_{CE} are calculated for each part independently and the average of all parts is used to supervise the learning process of the model:

$$L_{joint} = L_{triplet} + \gamma L_{CE}, \tag{15}$$

where the hyper-parameter γ is used to balance the weights of two losses.

4. Experiments

4.1. Datasets and Implementation Details

To validate the effectiveness of GPGait, we conduct experiments under indoor (CASIA-B and OUMVLP-Pose) and outdoor (Gait3D and GREW) scenarios, respectively.

CASIA-B [37] contains 124 subjects and each subject is required to walk under three conditions. For each condition, 11 sequences from multiple viewpoints range from 0° to 180°. During the testing stage, sequences of four normal conditions are regarded as the gallery, and the rest sequences are regarded as the probe. HRNet [29] is used to extract pose data from the RGB videos following the CASIA-B release agreement.

OUMVLP-Pose [1] is based on the large gait recognition dataset OUMVLP [30] which contains 10,307 subjects, and the sequences of each subject are collected from 14 views (0°, 15°, ..., 90°; 180°, 195°, ..., 270°). There are 2 sequences (#00-01) under each view. 5,153 subjects are split for training and 5,154 subjects are split for testing. During the test, sequences with index #01 are regarded as the gallery and those with index #00 are used as the probe.

Gait3D [40] is a wild dataset containing 4,000 subjects with over 25,000 sequences. The data is collected in a supermarket from 39 cameras. 3,000 subjects are used for training and the rest for testing. At the testing stage, one sequence of each subject is selected to build the probe, and the rest sequences become the gallery.

GREW [41] is also a wild dataset that includes almost 3,500 hours from 882 cameras. GREW dataset is split into a training set, a validation set and a testing set which contains 20,000, 345 and 6,000 subjects respectively. During testing, two sequences of each subject are regarded as the probe and the other two sequences are regarded as the gallery.

Implementation Details In all experiments, h_{unif} is set to a fixed number of 225. The threshold ϕ in Eq.2 is set to 0.1 radians. We design the network capacity referring to the baselines [31, 32, 38] and do not deliberately tune it. Specifically, for CASIA-B and Gait3D, the number of 5-part blocks is 3 and the vector sizes of each block are (64, 64, 128). While for OUMVLP and GREW, the number of 5-part blocks is 4 and the vector sizes of each block are (64, 128, 128, 128, 128). During training, the length of pose sequences is fixed to 30 for OUMVLP and 60 for others in an unordered selecting manner. For data augmentations, we apply left-right flipping of the skeleton with a probability of 0.01 and gaussian noises are added to each keypoint with a probability of 0.3. The optimizer is adam [13] with a one-cycle learning rate schedule [27] of three phases, where initial, maximum, and final learning rates are set to 1e-5, 1e-3, and 1e-8. We adjust the batch size and the number of iterations to fit different dataset scales. **1)** On CASIA-B, we train the model for 40k iterations with a batch size of (4, 32). **2)** On Gait3D, we train the model for 60k iterations with a batch size of (32, 4). **3)** On OUMVLP and GREW, the model is trained for 150k iterations with a batch size of (32, 16), (32, 8), respectively. During the test stage, all the frames of a sequence are fed into the network.

4.2. Performance Comparison

Evaluation for Cross-Domain Settings For a comprehensive comparison, we conduct cross-domain experiments on previous pose-based methods to evaluate the ability of generalization, *i.e.*, train the model on the source dataset (Source) and test on other datasets (Target). As shown in Tab.1, our method outperforms the existing pose-based methods over all of the source-target dataset pairs, suggesting our model's excellent generalization ability. In particular, compared to the second-best methods trained on the outdoor datasets, we achieve state-of-the-art results with considerable margins of 23.86%, 34.69% Rank-1 on Gait3D→CASIA-B and GREW→CASIA-B respectively. We have also surpassed the second-best result trained on indoor datasets by a significant margin of 9.35%, 10.07% Rank-1 on CASIA-B→GREW and OUMVLP-Pose→GREW respectively.

Furthermore, the cross-domain results of GPGait are

Table 1. Rank-1 accuracy (%) on four popular datasets: cross-domain and single-domain performance of GPGait and recent state-of-the-art pose-based methods, excluding identical-view case of indoor datasets. [1]

Source Dataset	Method	Target Dataset						
		CASIA-B				OUMVLP-Pose	GREW	Gait3D
		NM	BG	CL	Mean			
CASIA-B	GaitGraph [32]	86.37	76.5	65.24	76.04	0.07	0.45	0.90
	GaitGraph2 [31]	80.29	71.40	63.80	71.83	0.07	0.48	1.10
	GaitTR [38]	**94.72**	**89.29**	**86.65**	**90.22**	0.07	0.62	1.10
	GPGait(ours)	93.60	80.15	69.29	81.01	**2.84**	**9.97**	**8.90**
OUMVLP-Pose	GaitGraph [32]	4.85	4.84	3.90	4.53	4.24	0.67	1.50
	GaitGraph2 [31]	8.83	7.62	5.13	7.19	**70.68**	0.85	1.40
	GaitTR [38]	10.10	8.26	5.17	7.84	39.77	1.06	2.60
	GPGait(ours)	**44.36**	**31.97**	**22.35**	**32.90**	59.11	**11.13**	**9.00**
GREW	GaitGraph [32]	10.54	7.73	5.73	8.00	0.17	10.18	4.40
	GaitGraph2 [31]	8.85	7.18	5.13	7.05	0.22	34.78	8.30
	GaitTR [38]	7.60	6.36	6.40	6.79	0.06	48.58	7.30
	GPGait(ours)	**57.87**	**45.98**	**24.23**	**42.69**	**4.25**	**57.04**	**18.50**
Gait3D	GaitGraph [32]	16.47	12.18	8.29	12.31	0.27	3.14	8.60
	GaitGraph2 [31]	12.32	9.93	5.43	9.23	0.09	2.39	11.20
	GaitTR [38]	4.50	3.90	3.96	4.12	0.06	4.38	7.20
	GPGait(ours)	**48.83**	**40.26**	**19.43**	**36.17**	**2.79**	**11.02**	**22.40**

found to even outperform or approach the source-domain results of recent methods. For instance, the cross-domain result (GREW→Gait3D) of GPGait outperforms the best source-domain results (Gait3D→Gait3D) of other methods with a margin of 7.3% in Rank-1 accuracy. These superior performances collectively demonstrate the generalization capabilities of our proposed method.

Comparison on Source Domain The comparison in Tab.1 indicates that GPGait achieves nearly comparable or even better performance than recent pose-based methods in most cases. Especially for Gait3D, GPGait outperforms previous works by a considerable margin of 11.2%. Also, compared with other methods that failed to produce a stable result on specific datasets, GPGait achieves a relatively stable performance without any distinct performance degradation on single-domain settings across all four datasets. A detailed analysis of the performance on the source domain is included in Sec.5.

[1](a) We take great efforts to build a unified framework for pose-based gait recognition named FastPoseGait (https://github.com//BNU-IVC/FastPoseGait) and re-run these experiments for a more fair comparison by sticking to the original implementations as much as possible. The results are a little different but comparable to those in the initial submission as well as those in the corresponding papers.

(b) In the literature, there are two versions of pose-based CASIA-B estimated by HRNet [29] and SimCC [16] respectively. Given that Gait3D and GREW are generated by HRNet, we finally use the HRNet version of CASIA-B for a unified experimental setting. We will also provide some results for the SimCC version of CASIA-B in our codebase.

(c) For OUMVLP-Pose, the sequences are generated by AlphaPose [7] consisting of 18 keypoints for each frame. In our experiments, we transform the keypoints into the COCO2017 format with 17 keypoints for the cross-domain evaluation.

4.3. Ablation Study

To verify the effectiveness of the components in GPGait, a series of ablation studies are conducted on CASIA-B and Gait3D to show the source-domain and cross-domain performance of indoor and outdoor evaluation systematically.

Analysis of Human-Oriented Transformation In this section, we first analyze the effectiveness of HOT compared with other normalization methods. Then we insert HOT into different backbones to verify generalization ability after normalizing the pose data.

As shown in Tab.2, compared with previous methods, the results of GPGait demonstrate the best generalization ability. **a)** Compared with spine-unit normalization [1, 19], HOT outperforms it by a large margin in the source domain as well. On the one hand, it proves the effectiveness of employing such a simple method to achieve remarkable results. On the other hand, the spine as a body part is not a good solution for some special situations like occlusion which introduces abnormal calculations. **b)** Compared with the dataset-independent normalization method [25], HOT demonstrates excellent generalization results. This is because the dataset statistics preserve the relative height of the human body in one dataset. However, the relative information is lost when the model is applied to another sequence recorded by a different camera of different datasets, which makes it less applicable in the real world.

In addition, to demonstrate the reusability of HOT on different backbones, we apply it to previous methods which are shown in Tab.3. **a)** Although HOT reduces the performance of the previous pose-based methods on CASIA-B,

Table 2. Analysis of different pose representations, in which we control the structure of the network.

Method	CASIA-B→Gait3D		Gait3D→CASIA-B	
	Source	Target	Source	Target
HOT(ours)	81.01	**8.90**	22.40	**36.17**
Spine-Unit [1, 19]	74.53	5.50	14.50	15.74
Dataset-Independent [25]	87.03	1.50	9.90	11.54

Table 3. Analysis of HOT, in which we compare the performance of HOT on different backbones.

Method(w/wo)	CASIA-B→Gait3D		Gait3D→CASIA-B	
	Source	Target	Source	Target
GaitGraph	40.91	1.96	11.00	29.24
GaitGraph2	45.66	3.20	12.20	24.49
GaitTR	64.37	2.40	8.10	21.74
GPGait	81.01	**8.90**	22.40	**36.17**
GaitGraph	76.04	0.90	8.60	12.31
GaitGraph2	71.83	1.10	11.20	9.23
GaitTR	90.22	1.10	7.20	4.12
GPGait	86.15	**2.70**	16.00	**20.30**

Table 4. Analysis of Human-Oriented Descriptors, in which we keep the network architecture consistent.

Setting			CASIA-B→Gait3D		Gait3D→CASIA-B	
Joint	Bone	Angle	Source	Target	Source	Target
✓			77.10	7.20	17.90	31.12
	✓		76.02	7.20	18.30	31.78
		✓	43.78	2.90	5.30	14.18
✓	✓		80.08	7.70	19.20	32.46
	✓	✓	79.40	7.70	18.30	32.74
✓		✓	77.42	7.00	17.00	34.46
✓	✓	✓	**81.01**	**8.90**	**22.40**	**36.17**

its generalization ability remains relatively steady across different backbones. Also, all the methods with HOT on Gait3D improve by a large margin. This demonstrates that HOT can be effectively adapted to other works, enabling them to achieve better and more stable results. **b)** For settings without HOT, GPGait still performs relatively stable and outperforms most previous works, which further demonstrates the superiority of PAGCN.

Analysis of Human-Oriented Descriptors In Tab.4, the results show the impact of using different combinations of generated features in Human-Oriented Descriptors. It can be seen that using only a single modality as input can not yield satisfactory results. When combining the input types, corresponding improvements can be achieved in both the source domain and the target domain. Furthermore, the multi-features of joint, bone and angle can significantly boost the performance. The benefit of multi-features generated by HOD can explicitly describe discriminant information of the human body that includes human keypoints, human body structure, and gait movement.

Impact of Multi-Branch in PAGCN This section aims to explore the contribution of multi-branch architecture to learn the features generated by HOD. In single-branch settings, we concatenate the three types of features in the channel dimension and put them into a one-branch network as the operations in GaitTR [38]. The multi-branch settings consist of parameter-independent branches to extract three types of features. In Tab.5, the use of multi-branch networks can lead to better performance, showing its effectiveness in learning specific expressions. This is mainly because each branch can focus on extracting information from different types of features separately. In contrast, a single-branch network merges the information from different data distributions, making the learning process challenging.

Table 5. Analysis of Multi-branch, in which we control the backbone of PAGCN.

Setting	CASIA-B→Gait3D		Gait3D→CASIA-B	
	Source	Target	Source	Target
Single-Branch	75.54	7.10	18.90	33.04
Multi-Branch	**81.01**	**8.90**	**22.40**	**36.17**

Table 6. Analysis of Partition, in which we control the whole network structure.

Partition	CASIA-B→Gait3D		Gait3D→CASIA-B	
	Source	Target	Source	Target
w	**81.01**	**8.90**	**22.40**	**36.17**
w/o	76.47	7.10	20.40	35.15

Impact of Partition in PAGCN A completely identical network without multiplying the M_k in Eq.9 is designed to verify the effectiveness of partition masks, where the number of parameters is completely the same as well. In Tab.6, the remarkable performance of the network leveraging masks proves that it is beneficial to extract the discriminative body features at both local and global levels with the help of partition on the adjacency matrices.

5. Discussion

Source-Domain Results It can be seen that GPGait does not achieve state-of-the-art but comparable source-domain results on some datasets. The main reason is that HOT eliminates certain relative information in one dataset. For example, the relative information of body height is a discriminative factor to improve single-domain performance. But it does not exist in real-world scenarios due to the various heights of camera viewpoints. HOT rescales all the skeletons to a uniform height, which can force the network more concerned about gait-relevant features and further enhance the practicality of skeleton-based methods. Therefore, it is reasonable for the performance degradation on some datasets, and we believe future research that incorporates advanced feature extraction methods under the human-oriented representation can improve the performance of pose-based methods on both source-domain and

cross-domain settings.

Compared with Silhouette-based Methods Compared to recent silhouette-based methods, the performance of GPGait is still limited. An evident explanation is that the poses used in GPGait lose body shape compared with the silhouettes. But pose-based methods have their own strengths like the robustness to wearings and explicit modeling for proportions and relations of body parts, which deserves further and continuous exploration.

Prospect Pose as an important modality for human representation contributes a lot to gait recognition as well. However, the lack of ability to generalize limits the application and further development of skeleton-based gait recognition. Our GPGait framework takes both the input and method into account, proposing a viable solution to address the generalization problem of pose-based methods. We are expecting the unified representation of skeletons can be utilized in later works for a fair and applicable future of pose-based methods. And PAGCN offers a promising approach for establishing part relations and extracting fine-grained gait information, which can be readily adapted for future research. Overall, more advanced pose-based methods are expected to further narrow the gap between the lab and the real world and promote the development of gait recognition.

6. Conclusion

In this paper, we present a generalized pose-based framework (GPGait), which transforms the arbitrary human pose into a unified representation and make full use of human pose characteristics to extract multi-features in Human-Oriented Transformation and Human-Oriented Descriptors. Part-Aware Graph Convolutional Network allows efficient partitions of the human graph and the effective learning of local-global relations. Experiments on four benchmarks (including indoor and outdoor scenarios) have indicated that GPGait achieves the highest accuracy on cross-domain settings and the most stable performance on single-domain settings, which also demonstrates the great potential of pose-based gait recognition.

Acknowledgement This work is jointly supported by National Natural Science Foundation of China (62276025, 62206022), Fundamental Research Funds for the Central Universities (2021NTST31) and Shenzhen Technology Plan Program (KQTD20170331093217368).

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] 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.

[3] Tianrui Chai, Annan Li, Shaoxiong Zhang, Zilong Li, and Yunhong Wang. Lagrange motion analysis and view embeddings for improved gait recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20249–20258, 2022.

[4] Hanqing Chao, Yiwei He, Junping Zhang, and Jianfeng Feng. Gaitset: Regarding gait as a set for cross-view gait recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8126–8133, 2019.

[5] Chao Fan, Saihui Hou, Jilong Wang, Yongzhen Huang, and Shiqi Yu. Learning gait representation from massive unlabelled walking videos: A benchmark. *arXiv preprint arXiv:2206.13964*, 2022.

[6] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14225–14233, 2020.

[7] Hao-Shu Fang, Jiefeng Li, Hongyang Tang, Chao Xu, Haoyi Zhu, Yuliang Xiu, Yong-Lu Li, and Cewu Lu. Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[8] Yang Fu, Yunchao Wei, Yuqian Zhou, Honghui Shi, Gao Huang, Xinchao Wang, Zhiqiang Yao, and Thomas Huang. Horizontal pyramid matching for person re-identification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8295–8302, 2019.

[9] Jinguang Han and Bir Bhanu. Individual recognition using gait energy image. *IEEE transactions on pattern analysis and machine intelligence*, 28(2):316–322, 2005.

[10] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

[11] Saihui Hou, Chunshui Cao, Xu Liu, and Yongzhen Huang. Gait lateral network: Learning discriminative and compact representations for gait recognition. In *European conference on computer vision*, pages 382–398, 2020.

[12] Saihui Hou, Xu Liu, Chunshui Cao, and Yongzhen Huang. Gait quality aware network: toward the interpretability of silhouette-based gait recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

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

[14] 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.

[15] 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*, 2020.

[16] Yanjie Li, Sen Yang, Peidong Liu, Shoukui Zhang, Yunxiao Wang, Zhicheng Wang, Wankou Yang, and Shu-Tao Xia. Simcc: A simple coordinate classification perspective for human pose estimation. In *European Conference on Computer Vision*, pages 89–106, 2022.

[17] 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 *European Conference on Computer Vision*, pages 375–390, 2022.

[18] Rijun Liao, Chunshui Cao, Edel B Garcia, Shiqi Yu, and Yongzhen Huang. Pose-based temporal-spatial network (ptsn) for gait recognition with carrying and clothing variations. In *Chinese Conference on Biometric Recognition*, pages 474–483, 2017.

[19] 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.

[20] 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*, pages 14648–14656, 2021.

[21] Xiaokai Liu, Zhaoyang You, Yuxiang He, Sheng Bi, and Jie Wang. Symmetry-driven hyper feature gcn for skeleton-based gait recognition. *Pattern Recognition*, 125:108520, 2022.

[22] Hao Luo, Wei Jiang, Youzhi Gu, Fuxu Liu, Xingyu Liao, Shenqi Lai, and Jianyang Gu. A strong baseline and batch normalization neck for deep person re-identification. *IEEE Transactions on Multimedia*, 22(10):2597–2609, 2019.

[23] Ekkasit Pinyoanuntapong, Ayman Ali, Pu Wang, Minwoo Lee, and Chen Chen. Gaitmixer: skeleton-based gait representation learning via wide-spectrum multi-axial mixer. *arXiv preprint arXiv:2210.15491*, 2022.

[24] Chiara Plizzari, Marco Cannici, and Matteo Matteucci. Spatial temporal transformer network for skeleton-based action recognition. In *International Conference on Pattern Recognition*, pages 694–701, 2021.

[25] M Rashmi and Ram Mohana Reddy Guddeti. Human identification system using 3d skeleton-based gait features and lstm model. *Journal of Visual Communication and Image Representation*, 82:103416, 2022.

[26] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12026–12035, 2019.

[27] Leslie N Smith and Nicholay Topin. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial intelligence and machine learning for multi-domain operations applications*, volume 11006, pages 369–386, 2019.

[28] Yi-Fan Song, Zhang Zhang, Caifeng Shan, and Liang Wang. Stronger, faster and more explainable: A graph convolutional baseline for skeleton-based action recognition. In *proceedings of the 28th ACM international conference on multimedia*, pages 1625–1633, 2020.

[29] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5686–5696, 2019.

[30] 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 transactions on Computer Vision and Applications*, 10:1–14, 2018.

[31] 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*, pages 1569–1577, 2022.

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

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[34] Chen Wang, Junping Zhang, Liang Wang, Jian Pu, and Xiaoru Yuan. Human identification using temporal information preserving gait template. *IEEE transactions on pattern analysis and machine intelligence*, 34(11):2164–2176, 2011.

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

[36] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[37] Shiqi Yu, Daoliang Tan, and Tieniu Tan. A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In *International conference on pattern recognition*, volume 4, pages 441–444, 2006.

[38] Cun Zhang, Xing-Peng Chen, Guo-Qiang Han, and Xiang-Jie Liu. Spatial transformer network on skeleton-based gait recognition. *arXiv preprint arXiv:2204.03873*, 2022.

[39] Yuqi Zhang, Yongzhen Huang, Liang Wang, and Shiqi Yu. A comprehensive study on gait biometrics using a joint cnn-based method. *Pattern Recognition*, 93:228–236, 2019.

[40] Jinkai Zheng, Xinchen Liu, Wu Liu, Lingxiao He, Chenggang Yan, and Tao Mei. 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*, pages 20228–20237, 2022.

[41] Zheng Zhu, Xianda Guo, Tian Yang, Junjie Huang, Jiankang Deng, Guan Huang, Dalong Du, Jiwen Lu, and Jie Zhou. Gait recognition in the wild: A benchmark. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 14789–14799, 2021.