Beyond Sparse Keypoints: Dense Pose Modeling for Robust Gait Recognition

Wenpeng Lang
Beijing Normal University
Beijing, China
wenpenglang@mail.bnu.edu.cn

Saihui Hou
Beijing Normal University
Beijing, China
housaihui@bnu.edu.cn

Yongzhen Huang*
Beijing Normal University
WATRIX.AI
Beijing, China
huangyongzhen@bnu.edu.cn

Abstract

Gait recognition has emerged as a promising biometric technology due to its ability to operate at a distance without subject cooperation. While pose-based methods offer advantages over appearance-based approaches in robustness and interpretability, their performance has been limited by the sparse keypoint representations of current pose estimation frameworks. We identify two critical limitations: (1) incomplete motion representation due to insufficient keypoints for dynamic body parts, and (2) lack of shape information from minimal skeleton points. This paper presents DPGait, a novel framework that addresses these challenges through innovations in both upstream processing and downstream modeling. First, we enhance pose estimation by extending the standard COCO keypoint format with additional motion-sensitive points and shape-descriptive keypoints inspired by human mesh estimation. Second, we propose a divide-and-conquer modeling strategy that processes dense keypoints through group convolution with cross-group attention, coupled with multi-granularity supervision for improved training. Our comprehensive experiments demonstrate state-of-the-art performance in pose-based gait recognition, achieving 85.8% rank-1 accuracy on SUSTech1K—surpassing leading silhouette-based methods for the first time. The results validate that dense pose representation combined with our novel modeling approach significantly advances the field of gait recognition.

CCS Concepts

• **Computing methodologies** → **Biometrics**.

Keywords

Gait Recognition; Heatmap Representation; Skeleton and Surface Points; Dense Pose Gait Modeling

ACM Reference Format:
Wenpeng Lang, Saihui Hou, and Yongzhen Huang. 2025. Beyond Sparse Keypoints: Dense Pose Modeling for Robust Gait Recognition. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3746027.3755685

*Corresponding author

MM '25, Dublin, Ireland.

ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3755685

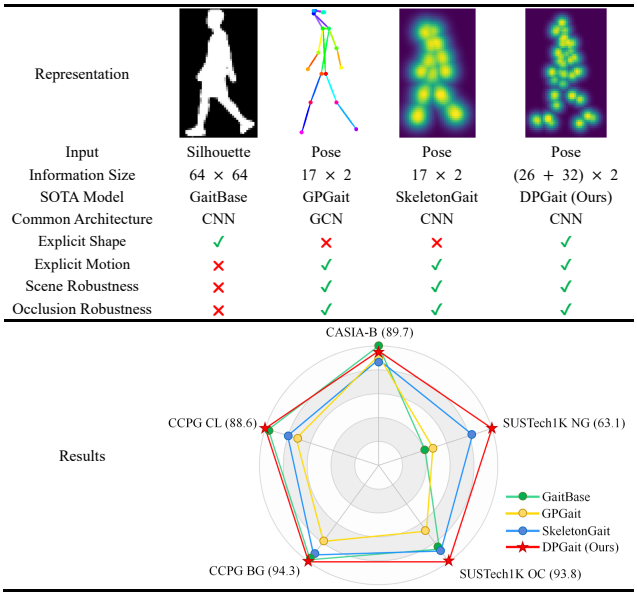

Figure 1: Motivation of the proposed DPGait approach. Top: Comparison of of dense pose keypoints over other representations. Bottom: DPGait demonstrates outstanding performance and strong robustness in complex environments.

1 Introduction

Human gait is a distinctive biometric characteristic that has emerged as an important research area in identity recognition. Unlike other biometric modalities such as facial features, fingerprints, or iris patterns, gait can be captured at a distance without requiring subject cooperation. Furthermore, gait recognition maintains robust identification accuracy even when individuals undergo significant changes in walking posture, speed, or clothing, as it analyzes subtle patterns in walking dynamics. These advantages, combined with gait's inherent anti-spoofing properties and the ubiquity of surveillance infrastructure, make it particularly valuable for security monitoring and criminal investigation applications [31, 32, 35].

Current gait recognition approaches fall into two primary categories: appearance-based and model-based methods. Appearance-based techniques, which have dominated the field for years, typically process binary silhouette sequences as input [5, 8, 9, 11,

17, 25, 29, 43]. However, the increasing prevalence of outdoor datasets [1, 21, 22, 33, 42, 51] have revealed limitations of them, as silhouette sequences often contain corrupted frames due to segmentation errors. In contrast, model-based approaches that utilize human pose keypoints [23, 33, 44] have gained prominence with recent advances in pose estimation technology [12, 19, 20, 24, 36].

Pose-based gait recognition offers three key advantages over silhouette-based methods:

(1) Keypoints are less affected by changes in appearance, such as clothing, carried items, or occlusions.
(2) Pose keypoints are typically stored in sparse coordinates, reducing storage overhead compared to high-dimensional pixel-based silhouettes.
(3) Keypoints provide clear semantic information about the human body, facilitating module design and enhancing interpret ability.

Despite these benefits, current pose-based methods relying on the COCO keypoint format [26] face two critical limitations:

(1) **Incomplete Motion Representation**: The sparse COCO format inadequately captures body dynamics. For example, representing the entire foot with a single keypoint fails to encode crucial motion patterns like arch movement during walking [26].
(2) **Limited Shape Information**: Existing pose estimation models [19, 24, 36, 37] output only a minimal set of skeleton points, insufficient for representing individual shape variations.

We argue that these limitations explain why current pose-based methods underperform compared to silhouette-based approaches. This work addresses these challenges through two main aspects of the gait recognition pipeline.

Upstream Processing: Recognizing that upstream processing quality determines downstream performance potential, we enhance pose estimation by extending the sparse COCO format. Our approach incorporates additional keypoints to better capture motion dynamics, particularly in the head and feet regions known to be crucial for identification [22]. We also introduce shape-descriptive keypoints inspired by human mesh estimation techniques. Using this enriched keypoint format, we collect training data to develop a novel pose estimation model that better represents both motion and shape characteristics.

Downstream Modeling: Building on our dense keypoint representation, we propose a novel modeling framework employing a **divide-and-conquer** strategy. Specifically, we encode keypoint coordinates as multichannel heatmaps, separating them into motion and shape groups. Group convolution processes each keypoint type independently, while channel-wise attention enables cross-group information exchange. We further enhance training through multi-granularity supervision, applying identity classification to both fused features and individual keypoint-group features. This approach improves model training without increasing inference complexity. The resulting solution achieves state-of-the-art performance among pose-based methods, even competitive with recent silhouette-based models.

In summary, our main contributions are:

- We conduct the first systematic study of the challenges caused by sparse keypoint formats in pose-based gait recognition. By innovating the upstream pose estimation task, we obtain dense human skeleton and surface points that are more effective for gait recognition.
- We propose a divide-and-conquer strategy for modeling, leveraging dense keypoints and introducing an efficient approach for keypoint modeling and supervision.
- Extensive experiments demonstrate that our approach achieves state-of-the-art results in pose-based gait recognition. Notably, our method achieves a rank-1 accuracy of 85.8% on the SUSTech1K dataset, surpassing the leading silhouette-based models.

2 Related Work

2.1 Gait Recognition

Contemporary gait recognition approaches can be broadly categorized into two paradigms based on input representation: appearance-based and model-based methods. Appearance-based techniques primarily process silhouette sequences, while model-based approaches utilize structural human pose representations encoded through skeleton keypoints.

Silhouette-based Methods. As the most extensively studied modality, silhouette-based approaches have evolved through kinds of spatial-temporal modeling techniques. While early works are mostly based on silhouette templates such as GEI [34], GaitSet [5], a set-based paradigm that treats silhouette sequences as unordered collections, employing statistical pooling for temporal aggregation. Subsequent work by GaitPart [11] introduced fine-grained part-level feature extraction with multi-scale temporal modeling. The field advanced further with GaitGL [25], which established a unified spatiotemporal framework through 3D convolutions for simultaneous global-local feature learning. More recently, GaitBase [9] and DeepGaitV2 [8] demonstrated that simplified network architectures could achieve competitive performance while improving computational efficiency.

Pose-based Methods. Skeleton keypoint representations have emerged as an alternative modality, with several notable architectural innovations. GaitGraph [39] and its successor GaitGraph2 [38] formulated the problem as graph learning, employing Graph Convolutional Networks (GCNs) to model structural relationships between keypoints. The transformer architecture was introduced to gait recognition by GaitTR [50], leveraging self-attention mechanisms for long-range spatiotemporal modeling. Recent advances include GPGait [14], which enhanced cross-domain generalization through part-aware graph convolutions, and PAA [15], which incorporated biomechanical priors via physics-augmented autoencoders. SkeletonGait [10] bridged the gap between modalities by encoding keypoints as heatmaps, enabling the transfer of Convolutional Neural Networks (CNNs) from silhouette-based approaches.

2.2 Human Pose and Mesh Estimation

The fields of human pose estimation and mesh estimation provide fundamental representations for modeling human motion and shape characteristics, respectively.

Human Pose Estimation. Modern pose estimation techniques fall into two methodological categories: regression-based approaches that directly predict keypoint coordinates [3, 37], and heatmap-based methods that model keypoint distributions through spatial likelihood maps [19, 24, 36]. The field has progressed from basic joint estimation [18, 26] to comprehensive full-body localization encompassing hands and feet [12, 20].

Human Mesh Estimation. Mesh estimation techniques have evolved from early SMPL-based approaches [27] to sophisticated surface reconstruction methods. Initial works like Pose2Mesh [6] and DecoMR [49] demonstrated mesh recovery from monocular inputs using keypoints or IUV maps. Subsequent advances such as THUNDR [48] and VirtualMarkers [30] achieved higher fidelity through dense surface point prediction. This progress has been enabled by the development of large-scale datasets, with data collection methodologies transitioning from motion capture systems [27] to synthetic data generation pipelines [4, 41, 46].

3 Our Approach

Our approach introduces innovations in both upstream representation and downstream modeling. Section 3.1 presents our enhanced pose representation, while Section 3.2 details the corresponding recognition architecture.

3.1 Upstream: Skeleton and Surface Points

3.1.1 Enhance Motion Representation. **Human Skeleton Points** in pose estimation typically mark biomechanically significant joints that characterize limb movement dynamics (e.g., joint rotation centers and limb endpoints). Since CNN-Pose [2] introduced the first gait-oriented skeleton dataset, most pose-based gait recognition systems have adopted the COCO format [26], which defines a 17-keypoint human topology. Current gait datasets typically generate these keypoints using pose estimators like HRNet [36], SimCC [24], or AlphaPose [13]. Established methods including GaitGraph, GaitGraph2, GaitTR, and GPGait build upon this format while developing advanced feature extraction strategies.

However, our analysis identifies significant information loss in the COCO format. While it captures major torso joints and facial landmarks, it omits critical regions like the head and feet - areas proven essential for gait analysis in CCPG dataset studies [22]. We address this limitation through a two-stage enhancement:

(1) **Format Upgrade**: We replace COCO with more comprehensive Halpe-Body format as illustrated in Figure 2, providing richer motion information for downstream recognition.
(2) **Estimation Improvement**: We employ RTMPose [19] for higher precision keypoint detection, overcoming estimation errors prevalent in existing gait datasets.

Notably, increased keypoint density does not always improve recognition performance. As Figure 2 demonstrates, formats like COCO-WholeBody (133 points) concentrate excessive detail on finger joints - features both difficult to estimate reliably in low-resolution surveillance footage and minimally relevant to gait dynamics. We will provide quantitative analysis of this trade-off in Section 4.3.1.

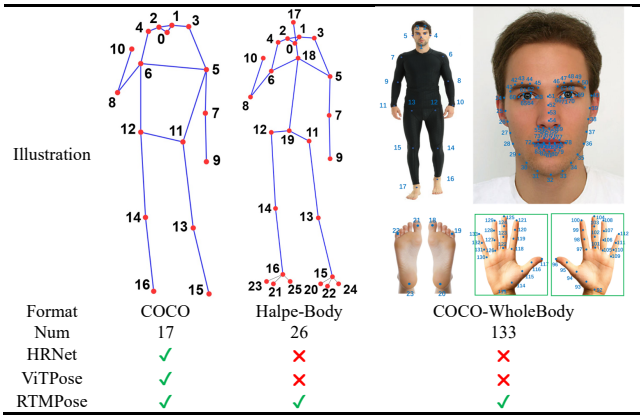

Figure 2: Three popular keypoint formats and typical pose estimation models.

3.1.2 Mitigate Shape Representation. To complement skeleton keypoints that primarily capture motion dynamics, we propose **Human Surface Points** - a novel set of anatomical landmarks specifically designed to encode body shape characteristics for gait recognition. This represents the first systematic integration of explicit shape representation in pose-based gait analysis.

Our surface point selection methodology builds upon virtual marker techniques from mesh estimation [30], employing archetypal analysis to automatically identify optimal surface landmarks. This data-driven approach provides two fundamental advantages:

(1) **Optimal Shape Encoding**: The optimization process automatically identifies the minimal set of surface points that best preserves the complete shape information, avoiding the information loss inherent in manual landmark selection.
(2) **Anatomically Valid Positions**: All selected points are constrained to lie precisely on or within a proximity to the actual mesh surface, ensuring they maintain realistic spatial relationships and biomechanical validity throughout motion sequences.

Formally, the point selection is formulated as a constrained optimization problem. Let $B \in \mathbb{R}^{M\times K}$ be a selection matrix mapping M mesh vertices to K surface points, and $X = [x_1, ..., x_N] \in \mathbb{R}^{3\times N\times M}$ represent a mesh sequence with N frames. We jointly optimize B with a reconstruction matrix $A \in \mathbb{R}^{K\times M}$ to minimize:

$$\min_{A,B} ||X - XBA||^2 \tag{1}$$

The values in B are constrained to approach 0 or 1, enforcing valid point selection from the mesh. After optimization, we apply three practical constraints for gait recognition: (1) Symmetry enforcement to maintain consistency with common data augmentation strategies. (2) Spatial separation from skeleton keypoints to ensure complementary information. (3) Resolution robustness by avoiding overly dense clusters (e.g., hands). Finally, we establish $K = 32$ as a reasonable surface point count, between shape representation fidelity and computational efficiency for real-world gait analysis scenarios.

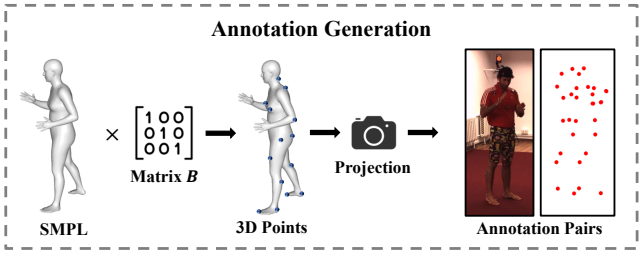

Figure 3: The pipeline for obtaining human surface points.

3.1.3 Data Preparation and Training Strategy. For skeleton point annotation, seven public datasets (AI Challenger, MS COCO, CrowdPose, MPII, sub-JHMDB, Halpe, and PoseTrack18) are integrated to cover diverse scenarios including single-person poses, complex multi-person interactions, and temporal motion sequences. This comprehensive collection ensures robust pose estimation across various real-world conditions.

Particularly, surface point annotation involves a specialized pipeline combining 3D motion capture data (Human3.6M, Surreal, SynBody, and GTA-Human) with camera projection. As shown in Figure 3, we first extract 3D surface points from mesh models, then project them to 2D image coordinates to create annotation pairs that maintain consistency with conventional skeleton point formats. This approach bridges the gap between 3D shape representation and 2D pose estimation requirements.

Our training framework builds on the RTMPose architecture, modified with parallel prediction heads for both skeleton points (17 keypoints) and surface points (32 keypoints). The model is initialized with pretrained weights for skeleton point estimation, with the surface point branch trained using our projected annotations. This dual-branch design allows simultaneous estimation of motion dynamics and shape characteristics. The trained models and annotation tools will be released to support reproducible research.

3.2 Downstream: Dense Pose Gait Modeling

3.2.1 Overview. Our Dense Pose Gait modeling (DPGait) framework, illustrated in Figure 4, processes enhanced human pose representations through a multi-stage neural architecture. The pipeline begins by transforming both skeleton and surface points into structured feature representations using Semantic-Guided Heatmap Partition (SGHP), generating normalized heatmaps of two types that are concatenated along the channel dimension. The backbone network then processes these combined features through three specialized components: Depth-Wise Deformable Convolution (DWDConv) extracts channel-specific spatial patterns, Separate Point Convolution (SPConv) models spatiotemporal relationships within each point type, and Part-Guided Channel Recalibration (PGCR) discovers implicit relationships between two types of points focusing on motion and shape through attention mechanisms. The entire system is optimized end-to-end using Multi-Granularity Supervision (MGS), which enforces discriminative feature learning at multiple levels from each type of keypoint.

3.2.2 Semantic-Guided Heatmap Partition (SGHP). Given a sequence of 2D keypoints $X \in \mathbb{R}^{V\times T\times 2}$, where V, T represent the number of points and frames, we first transform them into heatmap representations. Following the heatmap generation approach of SkeletonGait [10], we obtain a heatmap sequence $S \in \mathbb{R}^{V\times T\times H\times W}$, where H and W denote the spatial dimensions of each heatmap. Particularly, to maintain consistency with our surface point processing and avoid manual definition of skeleton connections, we deliberately omit the generation of limb heatmaps for skeleton points.

Recognizing that compressing dense point heatmaps into a single channel may introduce noise interference, we propose a semantic-guided multichannel partitioning strategy to optimize the trade-off between storage efficiency and information preservation. Drawing inspiration from the LandmarkGait [45] parsing approach, we generate a refined heatmap sequence $\tilde{S} \in \mathbb{R}^{4\times T\times H\times W}$ through:

$$\tilde{S} = \max\left(S; \{S_i \in V_{head}\}; \{S_i \in V_{upper}\}; \{S_i \in V_{lower}\}\right) \quad (2)$$

where V_{head}, V_{upper}, and V_{lower} represent three semantically distinct partitions of human body points.

Unlike conventional multi-modal approaches that employ separate feature extractors with cross-branch communication, our method treats skeleton and surface points as complementary representations of human motion and shape within a unified convolutional architecture. This integrated approach capitalizes on their inherent similarity while preserving their distinct semantic contributions to gait recognition. Finally we get the input $\hat{S} \in \mathbb{R}^{8\times T\times H\times W}$;

$$\hat{S} = [\tilde{S}_{skeleton}; \tilde{S}_{surface}] \quad (3)$$

3.2.3 Depth-wise Deformable Convolution (DWDConv). Based on the understanding that each point within a heatmap represents different body structures, we expect the model to separately focus on and learn independent features from each channel during early feature learning stages. At the same time, since different body parts exhibit varying positions and motion intensities during walking, the model needs to adaptively focus on relevant spatial regions [7]. To meet these requirements, we introduce Depth-wise Deformable Convolution (DWDConv). Specifically, the DWDConv first splits the input heatmap, and for each location p_0^c on the output feature map y of heatmap channel c, the operation is defined as:

$$y(p_0^c) = \sum_{p_n^c \in \mathcal{R}} w(p_n^c) \cdot s(p_0^c + p_n^c + \Delta p_n^c) \quad (4)$$

where $\mathcal{R}$ denotes the regular grid sampling locations, w denotes the convolution weights, Δp_n^c represents the learned spatial offsets that enable adaptive receptive field and s uses bilinear interpolation in fractional sampling. As shown in Figure 5, after processing each heatmap channel independently, we concatenate them along the channel dimension. To further increase feature dimensionality, we employ K parallel DWDConv operations to map the original heatmaps into a higher-dimensional feature space:

$$y(p_0) = [y(p_0^0); ...; y(p_0^7)] \in \mathbb{R}^8 \quad (5)$$

$$Y(p_0) = [DWDConv_0(y(p_0)); ...; DWDConv_K(y(p_0))] \in \mathbb{R}^C \quad (6)$$

where $C = 8 \times K$. This approach effectively captures both channel-specific patterns and their spatial relationships while maintaining computational efficiency.

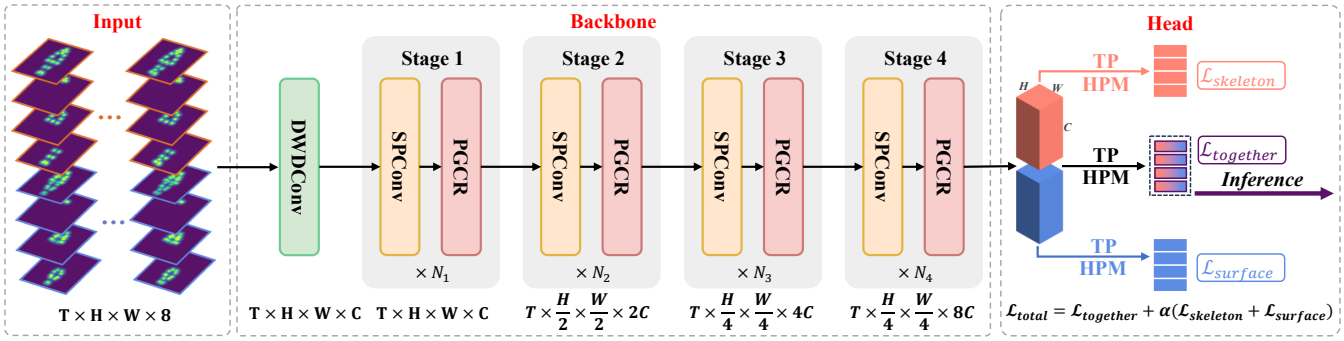

Figure 4: Overview of the DPGait pipeline. In the input phase, skeleton and surface points are converted into multichannel heatmaps via Semantic-guided Heatmap Partition (SGHP). In the backbone, the input is first mapped to shallow feature space using Depth-wise Deformable Convolution (DWDConv). Then, spatiotemporal modeling is conducted across four stages, each implementing the "divide-and-conquer" mechanism through Separate Point Convolution (SPConv) and Part-Guided Channel Recalibration(PGCR). In the Head, model training is driven by Multi-Granularity Supervision and TP/HPM represent Temporal Pooling/Horizontal Pyramid Mapping, which are common operations in GaitBase [9].

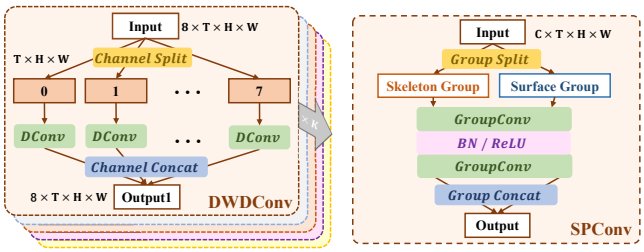

Figure 5: Illustration of DWDConv (Left) and SPConv (Right). Left: DWDConv learns initial distribution through deformable convolution kernels. Right: SPConv learns spatiotemporal features through grouped convolutions.

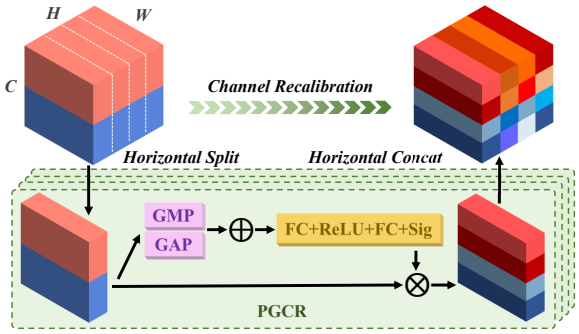

Figure 6: Part-Guided Channel Recalibration (PGCR) mechanism. Feature blocks are horizontally partitioned into multiple bins, with channel-wise reweighting performed through squeeze-and-excitation operations.

3.2.4 Separate Point Convolution (SPConv). In our framework, the structured feature distribution is preserved as human surface points and skeleton points maintain separate information flows during input and shallow modeling. To enhance spatiotemporal feature learning while preserving motion-shape separability in deep feature space, we propose Separate Point Convolution (SPConv) with a grouped architecture. As shown in Figure 5, SPConv employs two internal grouped convolutions that, when stacked across four stages in high-dimensional space, demonstrate two key advantages: (1) Efficient spatiotemporal modeling with parameter count reduced to 1/2 compared to standard convolution, and (2) Parallel processing within a single-branch architecture that maintains information separability through uniform convolution operations while ensuring precise feature alignment.

3.2.5 Part-Guided Channel Recalibration (PGCR). While dense point inputs enable simultaneous modeling of motion and shape characteristics, they also introduce noise from occlusions and geometric distortions that may degrade recognition performance. Our Part-guided Channel Recalibration (PGCR) addresses this by enhancing discriminative body regions through localized feature refinement. Building on local partition techniques [11, 16], PGCR first segments global features into localized bins, then performs differential enhancement via channel-wise attention. Each bin undergoes parallel global average and max pooling before two fully-connected layers with ReLU and Sigmoid activations learn channel-wise recalibration weights. Crucially, PGCR maintains independent weights per bin to accommodate varying motion patterns across body parts. Formally, given the feature sequence $F = \{F_1, F_2, ..., F_N\}$, PGCR processes each horizontally sliced bin F_{ij} ($i \in N$, $j \in J$) where N denotes the temporal length of the feature and J denotes the number of bins through:

$$X_{ij} = GAP(F_{ij}) + GMP(F_{ij}) \tag{7}$$

$$Y_{ij} = \sigma(W_j^2 \delta(W_j^1 X_{ij})) \tag{8}$$

$$\tilde{F_{ij}} = Y_{ij} \otimes F_{ij} \tag{9}$$

where GAP/GMP perform spatial squeezing, W_j^1/W_j^2 are FC layer weights, δ/σ denote ReLU/Sigmoid, and $\otimes$ is channel-wise multiplication. The recalibrated bins are then combined into a final feature sequence $\tilde{F}$.

Table 1: Training configurations across different datasets.

DataSet	Batch Size	Milestones	Total Steps
CASIA-B	(8, 16)	(20k, 40k, 50k)	60k
CCPG	(8, 16)	(20k, 30k, 40k)	60k
SUSTech1K	(8, 8)	(20k, 30k, 40k)	50k

3.2.6 Multi-Granularity Supervision (MGS). Our supervision strategy operates at multiple levels: primary optimization integrates all features through $L_{together}$, while auxiliary losses $L_{skeleton}$ and $L_{surface}$ maintain discriminability in separate point spaces enabled by our network design. The composite loss combines three terms through:

$$L_{total} = L_{together} + \alpha(L_{skeleton} + L_{surface}) \tag{10}$$

where α balances the loss terms and each combines triplet and cross-entropy losses assisted by BNNeck [28] for effective feature learning.

4 Experiment

4.1 Datasets

Since the upstream acquisition part of this work requires obtaining human skeleton points and surface points from RGB images, we conducted tests on three popular gait datasets: CASIA-B [47], CCPG [22], and SUSTech1K [33]. We strictly follow the official evaluation protocols in our experiments.

4.2 Implementation Details

In terms of data preparation, the input of all experiments is estimated by the RTMPose, and we ensure that the frames of surface points and skeleton points are strictly aligned. The generated heatmap size is 64×64. Due to the increase in the number of points in the dense pose, it is noteworthy that we set the gaussian variance of the heatmap $\sigma_{gau} = 4$ in our method. In terms of model training, we use the same data augmentation as SkeletonGait. During training, 30 frames of ordered sequences are randomly selected as input, while all frames are used during testing. The optimizer is SGD with an initial learning rate of 0.1 and a weight decay of 0.0005. As shown in Table 1, for different datasets we adjust batch size and milestones, where the learning rate is multiplied by 0.1. In terms of hyperparameter settings, the mapping channel $C = 64$, the number of stacks in the four stages is $N_1 = N_2 = N_3 = N_4 = 1$, the horizontal bin number $J = 16$ in the PGCR, and the loss weight $\alpha = 0.2$ to adjust the auxiliary loss in MGS.

4.3 Performance Comparison

4.3.1 Compare with Different Pose Estimation Methods. We first analyze the impact of different keypoint formats and estimation models by taking SkeletonGait and CCPG as benchmarks. As shown in Table 2, we reach the following conclusions: (1) When using the same input format COCO, compared to other estimation models, RTMPose significantly improves the performance of SkeletonGait by its superior architectural design and more strategies for cross-domain training. (2) Comparing the three keypoint formats supported by RTMPose, we notice that the Halpe format performs well under all conditions. We think that the additional foot keypoints in Halpe compared to COCO can more effectively capture the foot dynamics during walking, while the added body keypoints enhance the modeling ability of tilt movements. These features make Halpe a more suitable skeleton point format for gait recognition. (3) Although the COCO-WholeBody format provides more dense annotations of full-body key points, due to factors such as occlusion or low resolution in gait datasets, the estimation results of facial and hand keypoints are less stable. This phenomenon shows that there is a certain deviation between the existing upstream pose estimation and the downstream gait recognition.

Table 2: Performance comparison on different pose estimation format and models.

Model	Format	Num	CCPG		
			CL	UP	DN
			R-1(%)		
HRNet	COCO	17	52.4	65.4	72.8
ViTPose	COCO	17	70.7	82.4	84.2
RTMPose	COCO	17	76.2	85.6	85.0
RTMPose	Halpe-Body	26	**85.0**	**92.0**	**90.4**
RTMPose	COCO-WholeBody	133	77.7	85.4	87.3

4.3.2 Compare with Pose-based Methods. Our proposed DPGait introduces a dense point representation. Notably, we think that the human skeleton points and surface points well balance the motion and shape information required for fine-grained recognition. Therefore, as shown in Table 3 and Table 4, the overall Rank-1 in CCPG and SUSTech1K has been significantly improved compared to the second-best method. For example, the accuracy of Rank-1 in CCPG-CL / SUSTech1K-CL increases by 12.4% / 19.3%, respectively, compared to the second-best method. For CASIA-B, it also alleviates the problem that the heatmap method is less fitted than the point coordinate method, with the mean accuracy approaching GaitTR.

4.3.3 Compare with Silhouette-based Methods. As shown in Table 3 and Table 4, the rapid upstream development of pose estimation (e.g., RTMPose) has opened up new potential for pose-based gait recognition. For instance, previous works such as GPGait and SkeletonGait have already surpassed the silhouette method in some metrics, which strengthens confidence in future research of pose-based gait recognition. Our proposed DPGait further narrows the gap between silhouette-based gait recognition and pose-based gait recognition. For example, compared to GaitBase, the metrics on the three datasets have been largely surpassed. Meanwhile, we would like to emphasize that the dense pose method demonstrates stronger scene robustness and occlusion robustness in complex scenarios. In the SUSTech1K dataset, silhouettes are often affected by segmentation sensitivity and failed to get high results in OC, CL, and NG. However, learning the prior of the human body can often provide more clues. Ultimately, it also achieves an overall Rank-1 accuracy of 85.8%, which surpasses the best silhouette-based model for the first time.

Table 3: Performance comparison on CASIA-B and CCPG. Best methods of silhouette are in underline, and that of pose are in bold. The results are reported in the Rank-1 (R-1) accuracy.

Input	Method	Source	CASIA-B				CCPG				
			NM	BG	CL	Mean	CL	UP	DN	BG	Mean
			R-1				R-1				
Silhouette	GaitSet [5]	AAAI2019	95.8	90.0	75.4	86.8	77.5	85.0	82.9	87.5	83.2
	GaitPart [11]	CVPR2020	96.1	90.7	78.7	88.5	79.2	85.3	86.5	88.0	84.8
	GaitBase [9]	CVPR2023	97.6	94.0	77.4	89.7	88.5	92.7	93.4	93.2	92.0
Pose	GaitGraph [39]	ICIP2021	88.2	79.9	68.9	79.0	28.9	34.0	39.7	28.9	32.9
	GaitGraph2 [38]	CVPRW2022	85.2	73.7	64.4	74.4	21.1	25.6	29.1	25.2	25.3
	GaitTR [50]	ES2023	95.3	88.2	**86.6**	**90.0**	43.4	53.0	42.8	19.8	39.8
	GPGait [14]	ICCV2023	94.3	82.7	70.5	82.5	64.1	72.9	78.4	74.3	72.4
	SkeletonGait [10]	AAAI2024	92.2	79.5	66.8	79.5	76.2	85.6	85.0	89.3	84.0
	DPGait (Ours)	-	**97.2**	**90.3**	78.2	88.2	**88.6**	**94.0**	**93.9**	**94.3**	**92.7**

Table 4: Performance comparison on SUSTech1K. Best methods of silhouette are in underline, and that of pose are in bold. The results are reported in the Rank-1 (R-1) and Rank-5 (R-5) accuracy.

Input	Method	Source	Probe Sequence (R-1)								Overall	
			NM	BG	CL	CA	UM	UN	OC	NG	R-1	R-5
Silhouette	GaitSet [5]	AAAI2019	69.1	68.2	37.4	65.0	63.1	61.0	67.2	23.0	65.0	84.8
	GaitPart [11]	CVPR2020	62.2	62.8	33.1	59.5	57.2	54.8	57.2	21.7	59.2	80.8
	GaitGL [25]	ICCV2021	67.1	66.2	35.9	63.3	61.6	58.1	66.6	17.9	63.1	82.8
	GaitBase [9]	CVPR2023	81.5	77.5	49.6	75.8	75.5	76.7	81.4	25.9	76.1	89.4
Pose	GaitGraph [39]	ICIP2021	20.9	22.2	8.6	21.2	15.2	21.0	35.5	17.6	21.2	45.5
	GaitGraph2 [38]	CVPRW2022	26.0	26.4	11.8	24.6	19.4	23.1	32.9	19.3	24.8	49.3
	GaitTR [50]	ES2023	39.1	36.3	28.5	35.0	26.5	38.5	50.2	25.9	35.8	61.5
	GPGait [14]	ICCV2023	51.8	50.0	38.6	47.8	42.3	50.3	64.0	30.2	49.2	72.4
	SkeletonGait [10]	AAAI2024	74.5	70.0	45.0	67.6	65.0	72.7	83.3	52.3	69.2	87.4
	DPGait (Ours)	-	**90.3**	**85.7**	**64.3**	**85.3**	**83.7**	**88.8**	**93.8**	**63.1**	**85.8**	**95.3**

Table 5: Performance comparison on cross-domain evaluations between CASIA-B, CCPG and SUSTech1K. Best methods are in bold. The results are reported in the Rank-1 accuracy.

Model	Train on CCPG					Train on SUSTech1K					Train on CASIA-B			
	CASIA-B			SUSTech1K		CASIA-B			CCPG		CCPG		SUSTech1K	
	NM	BG	CL	OC	Overall	NM	BG	CL	CL	BG	CL	BG	OC	Overall
GaitGraph [39]	5.5	4.8	4.2	1.5	1.6	5.2	4.5	3.8	1.2	3.0	2.3	2.9	2.0	1.9
GaitGraph2 [38]	7.9	6.6	5.7	1.3	1.2	10.6	9.0	6.4	2.6	5.2	1.8	1.2	1.2	1.5
GaitTR [50]	5.2	4.8	4.3	1.3	1.1	6.7	6.5	5.9	1.7	2.2	2.6	1.9	1.3	1.5
GPGait [14]	41.7	33.4	21.7	7.7	4.4	57.4	44.8	23.9	**9.8**	21.4	**12.6**	20.1	9.5	5.3
SkeletonGait [10]	46.0	36.7	31.3	14.5	8.9	58.7	45.1	26.8	6.4	21.0	8.9	16.8	21.2	12.1
DPGait (Ours)	**56.1**	**46.9**	**33.8**	**18.9**	**14.3**	**71.5**	**59.6**	**28.3**	7.9	**28.7**	10.7	**21.7**	**29.1**	**20.7**

4.3.4 Cross-domain Evaluation. Concerned that the introduction of surface points on the human body might lead to loss of generalization ability of pose-based gait recognition, as shown in Table 5, we conduct cross-domain evaluation, that is, training on source dataset and evaluating on target dataset. From the results, it can be seen that the cross-domain ability of our method has been improved in most cases, further indicating that shape information is more important for gait recognition under some conditions.

4.4 Ablation Study

Our method achieves significant improvements in the upstream acquisition. Meanwhile, to verify the simple yet effective model, we conduct an ablation study of each module in SUSTech1K as shown

Figure 7: The visualization of skeleton and dense points, which are marked by connected lines and isolated points, respectively.

Table 6: Ablation study of each module on SUSTech1K.

Base	SGHP	DWDConv	SPConv	PGCR	MGS	SUSTech1K	
						R-1	R-5
✓					✓	80.2	91.6
✓	✓				✓	80.9	93.2
✓	✓	✓			✓	81.5	93.4
✓	✓	✓	✓		✓	82.7	94.0
✓	✓	✓	✓	✓		85.0	94.9
✓	✓	✓	✓	✓	✓	85.8	95.3

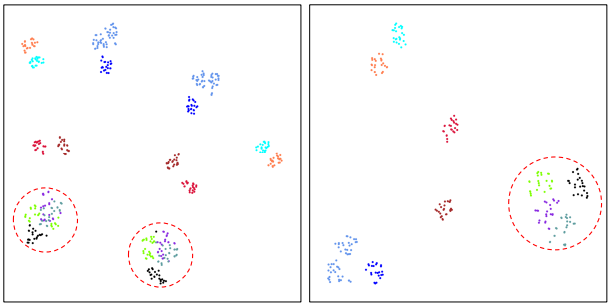

Figure 8: The t-SNE visualization of feature distributions in SkeletonGait and DPGait. Different colors represent different identities selected from SUSTech1K.

in Table 6. "Base" can be regarded as SkeletonGait with Halpe-body format human skeleton surface points as input. SGHP reduces information loss through multi-heatmaps, which helps the model quickly lock onto the candidate identities. DWDConv and SPConv can effectively extract the information on the initial distribution and feature space of the two types of points, respectively. Finally, PGCR enhances the key channel information of the local region to improve accuracy. It is worth mentioning that, throughout all other module ablation studies, we consistently employ MGS to supervise the inputs of both point types that have proven helpful in most cases. To specifically validate its effectiveness, we remove this loss function for isolated validation in the last two rows.

4.5 Visualization

Figure 7 provides the visualization of the skeleton and surface points estimated by our trained model. Moreover, to validate the robust discriminative capability of our method in learning skeleton and surface point information, we use t-SNE [40] technique to visualize feature distributions of 10 randomly selected identities from SUSTech1K. As illustrated in Figure 8, comparative analysis reveals that our method achieves enhanced inter-identity differentiation, particularly in the annotated regions within red dashed circles.

5 Conclusion

In this work, we systematically address the critical issue of motion and shape information deficiency in pose-based gait recognition, proposing the DPGait framework used for dense pose modeling. Regarding upstream acquisition, we employ an advanced pose estimation model to generate more accurate human skeleton keypoints and extend the model's output to human surface points using human mesh data, achieving dual representation of motion and shape. For the downstream modeling, we propose a "divide-and-conquer" spatiotemporal modeling mechanism combined with multichannel heatmaps to effectively extract fine-grained human features. Experiments show that DPGait achieves the best performance among pose-based methods and competitive results with silhouette-based methods. These findings validate the potential of keypoint representation in complex scenarios and lay the foundation for next-generation robust gait recognition paradigms.

Acknowledgments

This work is jointly supported by National Natural Science Foundation of China (62276025, 62206022, 62476027) and the Fundamental Research Funds for the Central Universities (2253200026).

References

[1] 2021. Gait Recognition in the Wild: A Benchmark. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. 14769–14779. doi:10.1109/ICCV48922.2021.01452

[2] Weizhi An, Shiqi Yu, Yasushi Makihara, Xinhui Wu, Chi Xu, Yang Yu, Rijun Liao, and Yasushi Yagi. 2020. 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 (2020), 421–430. doi:10.1109/TBIOM.2020.3008862

[3] Qingyuan Cai, Xuecai Hu, Saihui Hou, Li Yao, and Yongzhen Huang. 2024. Disentangled diffusion-based 3d human pose estimation with hierarchical spatial and temporal denoiser. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 38. 882–890.

[4] Zhongang Cai, Mingyuan Zhang, Jiawei Ren, Chen Wei, Daxuan Ren, Zhengyu Lin, Haiyu Zhao, Lei Yang, Chen Change Loy, and Ziwei Liu. 2024. Playing for 3D Human Recovery. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024), 1–12. doi:10.1109/TPAMI.2024.3450537

[5] Hanqing Chao, Yiwei He, Junping Zhang, and Jianfeng Feng. 2019. GaitSet: regarding gait as a set for cross-view gait recognition. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence* (Honolulu, Hawaii, USA) *(AAAI'19/IAAI'19/EAAI'19)*. AAAI Press, Article 996, 8 pages. doi:10.1609/aaai.v33i01.33018126

[6] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. 2020. Pose2Mesh: Graph Convolutional Network for 3D Human Pose and Mesh Recovery from a 2D Human Pose. In *European Conference on Computer Vision (ECCV)*.

[7] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. 2017. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*. 764–773.

[8] Chao Fan, Saihui Hou, Yongzhen Huang, and Shiqi Yu. 2024. Exploring Deep Models for Practical Gait Recognition. arXiv:2303.03301 [cs.CV] https://arxiv.org/abs/2303.03301

[9] Chao Fan, Junhao Liang, Chuanfu Shen, Saihui Hou, Yongzhen Huang, and Shiqi Yu. 2023. OpenGait: Revisiting Gait Recognition Toward Better Practicality. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 9707–9716. doi:10.1109/CVPR52729.2023.00936

[10] Chao Fan, Jingzhe Ma, Dongyang Jin, Chuanfu Shen, and Shiqi Yu. 2024. SkeletonGait: gait recognition using skeleton maps. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence (AAAI'24/IAAI'24/EAAI'24)*. AAAI Press, Article 185, 8 pages. doi:10.1609/aaai.v38i2.27933

[11] Chao Fan, Yunjie Peng, Chunshui Cao, Xu Liu, Saihui Hou, Jiannan Chi, Yongzhen Huang, Qing Li, and Zhiqiang He. 2020. GaitPart: Temporal Part-Based Model for Gait Recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 14213–14221. doi:10.1109/CVPR42600.2020.01423

[12] Hao-Shu Fang, Jiefeng Li, Hongyang Tang, Chao Xu, Haoyi Zhu, Yuliang Xiu, Yong-Lu Li, and Cewu Lu. 2022. AlphaPose: Whole-Body Regional Multi-Person Pose Estimation and Tracking in Real-Time. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022).

[13] Hao-Shu Fang, Jiefeng Li, Hongyang Tang, Chao Xu, Haoyi Zhu, Yuliang Xiu, Yong-Lu Li, and Cewu Lu. 2023. AlphaPose: Whole-Body Regional Multi-Person Pose Estimation and Tracking in Real-Time. *IEEE Trans. Pattern Anal. Mach. Intell.* 45, 6 (June 2023), 7157–7173. doi:10.1109/TPAMI.2022.3222784

[14] Yang Fu, Shibei Meng, Saihui Hou, Xuecai Hu, and Yongzhen Huang. 2023. GPGait: Generalized Pose-based Gait Recognition. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. 19538–19547. doi:10.1109/ICCV51070.2023.01795

[15] Hongji Guo and Qiang Ji. 2023. Physics-Augmented Autoencoder for 3D Skeleton-Based Gait Recognition. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. 19570–19581. doi:10.1109/ICCV51070.2023.01798

[16] Saihui Hou, Xu Liu, Chunshui Cao, and Yongzhen Huang. 2023. Gait Quality Aware Network: Toward the Interpretability of Silhouette-Based Gait Recognition. *IEEE Transactions on Neural Networks and Learning Systems* 34, 11 (2023), 8978–8988. doi:10.1109/TNNLS.2022.3154723

[17] Zhen Huang, Dixiu Xue, Xu Shen, Xinmei Tian, Houqiang Li, Jianqiang Huang, and Xian-Sheng Hua. 2021. 3D Local Convolutional Neural Networks for Gait Recognition. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. 14900–14909. doi:10.1109/ICCV48922.2021.01465

[18] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. 2014. Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36, 7 (jul 2014), 1325–1339.

[19] Tao Jiang, Peng Lu, Li Zhang, Ningsheng Ma, Rui Han, Chengqi Lyu, Yining Li, and Kai Chen. 2023. RTMPose: Real-Time Multi-Person Pose Estimation based on MMPose. doi:10.48550/ARXIV.2303.07399

[20] Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. 2020. Whole-Body Human Pose Estimation in the Wild. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

[21] Wu Liu Lingxiao He Chenggang Yan Tao Mei Jinkai Zheng, Xinchen Liu. 2022. Gait Recognition in the Wild with Dense 3D Representations and A Benchmark. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[22] Weijia Li, Saihui Hou, Chunjie Zhang, Chunshui Cao, Xu Liu, Yongzhen Huang, and Yao Zhao. 2023. An In-Depth Exploration of Person Re-Identification and Gait Recognition in Cloth-Changing Conditions. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 13824–13833. doi:10.1109/CVPR52729.2023.01328

[23] Xiang Li, Yasushi Makihara, Chi Xu, and Yasushi Yagi. 2022. Multi-View Large Population Gait Database With Human Meshes and Its Performance Evaluation. *IEEE Transactions on Biometrics, Behavior, and Identity Science* 4, 2 (2022), 234–248. doi:10.1109/TBIOM.2022.3174559

[24] Yanjie Li, Sen Yang, Peidong Liu, Shoukui Zhang, Yunxiao Wang, Zhicheng Wang, Wankou Yang, and Shu-Tao Xia. 2022. SimCC: A Simple Coordinate Classification Perspective for Human Pose Estimation. In *Computer Vision – ECCV 2022*, Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (Eds.). Springer Nature Switzerland, Cham, 89–106.

[25] Beibei Lin, Shunli Zhang, and Xin Yu. 2021. Gait Recognition via Effective Global-Local Feature Representation and Local Temporal Aggregation. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. 14628–14636. doi:10.1109/ICCV48922.2021.01438

[26] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Computer Vision – ECCV 2014*, David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (Eds.). Springer International Publishing, Cham, 740–755.

[27] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. 2015. SMPL: A Skinned Multi-Person Linear Model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)* 34, 6 (Oct. 2015), 248:1–248:16.

[28] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. 2019. Bag of tricks and a strong baseline for deep person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*. 0–0.

[29] Kang Ma, Ying Fu, Chunshui Cao, Saihui Hou, Yongzhen Huang, and Dezhi Zheng. 2024. Learning Visual Prompt for Gait Recognition. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 593–603. doi:10.1109/CVPR52733.2024.00063

[30] Xiaoxuan Ma, Jiajun Su, Chunyu Wang, Wentao Zhu, and Yizhou Wang. 2023. 3D Human Mesh Estimation From Virtual Markers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 534–543.

[31] Yasushi Makihara, Mark S. Nixon, and Yasushi Yagi. 2021. *Gait Recognition: Databases, Representations, and Applications*. Springer International Publishing, Cham, 487–499. doi:10.1007/978-3-030-63416-2_883

[32] Alireza Sepas-Moghaddam and Ali Etemad. 2023. Deep Gait Recognition: A Survey . *IEEE Transactions on Pattern Analysis & Machine Intelligence* 45, 01 (Jan. 2023), 264–284. doi:10.1109/TPAMI.2022.3151865

[33] Chuanfu Shen, Chao Fan, Wei Wu, Rui Wang, George Q. Huang, and Shiqi Yu. 2023. LidarGait: Benchmarking 3D Gait Recognition With Point Clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 1054–1063.

[34] Kohei Shiraga, Yasushi Makihara, Daigo Muramatsu, Tomio Echigo, and Yasushi Yagi. 2016. GEINet: View-invariant gait recognition using a convolutional neural network. In *2016 International Conference on Biometrics (ICB)*. 1–8. doi:10.1109/ICB.2016.7550060

[35] M. Sivarathinabala, S. Abirami, and R. Baskaran. 2017. *A Study on Security and Surveillance System Using Gait Recognition*. Springer International Publishing, Cham, 227–252. doi:10.1007/978-3-319-44790-2_11

[36] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. 2019. Deep High-Resolution Representation Learning for Human Pose Estimation. In *CVPR*.

[37] Xiao Sun, Bin Xiao, Shuang Liang, and Yichen Wei. 2017. Integral human pose regression. *arXiv preprint arXiv:1711.08229* (2017).

[38] Torben Teepe, Johannes Gilg, Fabian Herzog, Stefan Hörmann, and Gerhard Rigoll. 2022. Towards a Deeper Understanding of Skeleton-based Gait Recognition. doi:10.48550/arXiv.2204.07855

[39] Torben Teepe, Ali Khan, Johannes Gilg, Fabian Herzog, Stefan Hörmann, and Gerhard Rigoll. 2021. Gaitgraph: Graph Convolutional Network for Skeleton-Based Gait Recognition. In *2021 IEEE International Conference on Image Processing (ICIP)*. 2314–2318. doi:10.1109/ICIP42928.2021.9506717

[40] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research* 9, 86 (2008), 2579–2605. http://jmlr.org/papers/v9/vandermaaten08a.html

[41] Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. 2017. Learning from Synthetic Humans. In *CVPR*.

[42] Chenye Wang, Saihui Hou, Aoqi Li, Qingyuan Cai, and Yongzhen Huang. 2025. RA-GAR: A Richly Annotated Benchmark for Gait Attribute Recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 7591–7599.

[43] Lei Wang, Bo Liu, Fangfang Liang, and Bincheng Wang. 2023. Hierarchical Spatio-Temporal Representation Learning for Gait Recognition. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. doi:10.1109/ICCV51070.2023.01799

[44] Yanxiang Wang, Xian Zhang, Yiran Shen, Bowen Du, Guangrong Zhao, Lizhen Cui, and Hongkai Wen. 2022. Event-Stream Representation for Human Gaits Identification Using Deep Neural Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 7 (2022), 3436–3449. doi:10.1109/TPAMI.2021.3054886

[45] Zengbin Wang, Saihui Hou, Man Zhang, Xu Liu, Chunshui Cao, Yongzhen Huang, and Shibiao Xu. 2023. LandmarkGait: Intrinsic Human Parsing for Gait Recognition. In *Proceedings of the 31st ACM International Conference on Multimedia* (Ottawa ON, Canada) *(MM '23)*. Association for Computing Machinery, New York, NY, USA, 2305–2314. doi:10.1145/3581783.3611840

[46] Zhitao Yang, Zhongang Cai, Haiyi Mei, Shuai Liu, Zhaoxi Chen, Weiye Xiao, Yukun Wei, Zhongfei Qing, Chen Wei, Bo Dai, Wayne Wu, Chen Qian, Dahua Lin, Ziwei Liu, and Lei Yang. 2023. SynBody: Synthetic Dataset with Layered Human Models for 3D Human Perception and Modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 20282–20292.

[47] Shiqi Yu, Daoliang Tan, and Tieniu Tan. 2006. 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. 441–444. doi:10.1109/ICPR.2006.67

[48] Mihai Zanfir, Andrei Zanfir, Eduard Gabriel Bazavan, William T. Freeman, Rahul Sukthankar, and Cristian Sminchisescu. 2021. THUNDR: Transformer-Based 3D Human Reconstruction With Markers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 12971–12980.

[49] Wang Zeng, Wanli Ouyang, Ping Luo, Wentao Liu, and Xiaogang Wang. 2020. 3D Human Mesh Regression with Dense Correspondence. In *CVPR*.

[50] Cun Zhang, Xing-Peng Chen, Guo-Qiang Han, and Xiang-Jie Liu. 2023. Spatial transformer network on skeleton-based gait recognition. *Expert Systems* 40 (01 2023). doi:10.1111/exsy.13244

[51] Shinan Zou, Chao Fan, Jianbo Xiong, Chuanfu Shen, Shiqi Yu, and Jin Tang. 2024. Cross-Covariate Gait Recognition: A Benchmark. *Proceedings of the AAAI Conference on Artificial Intelligence* 38, 7 (Mar. 2024), 7855–7863.