Cut out the Middleman: Revisiting Pose-based Gait Recognition

Yang Fu[1], Saihui Hou[1,2(✉)], Shibei Meng[1], Xuecai Hu[1(✉)], Chunshui Cao[2], Xu Liu[2], and Yongzhen Huang[1,2]

[1] School of Artificial Intelligence, Beijing Normal University
[2] WATRIX.AI

Abstract. Recent pose-based gait recognition methods, which utilize human skeletons as the model input, have demonstrated significant potential in handling variations in clothing and occlusions. However, methods relying on such skeleton to encode pose are constrained mainly by two problems: (1) poor performance caused by the shape loss, and (2) lack of generalizability. Addressing these limitations, we revisit pose-based gait recognition and develop **GaitHeat**, a heatmap-based framework that largely enhances performance and robustness by utilizing a new modality to encode pose rather than keypoint coordinates. We make our efforts from two aspects, the pipeline and the extraction of multi-channel heatmap features. Specifically, the process of resizing and centering is performed in the RGB space to largely preserve the integrity of heatmap information. To boost the generalization across various datasets further, we propose a pose-guided heatmap alignment module to eliminate the influence of gait-irrelevant covariates. Furthermore, a global-local network incorporating an efficient fusion branch is designed to improve the extraction of semantic information. Compared to skeleton-based methods, GaitHeat exhibits superior performance in learning gait features and demonstrates effective generalization across different datasets. Experiments on three datasets reveal that our proposed method achieves state-of-the-art results for pose-based gait recognition, comparable to that of silhouette-based approaches. All the source code is available at https://github.com/BNU-IVC/FastPoseGait.

Keywords: Gait Recognition · Heatmap Representation · Generalization Ability

1 Introduction

Gait, an essential biometric characteristic, has garnered considerable attention for its application in identification tasks. Unlike other biometrics, gait is inherently difficult to disguise and can be captured at a long distance. Leveraging these advantages, gait recognition has been increasingly deployed in security applications [24, 27, 30], including suspect tracking and identity verification.

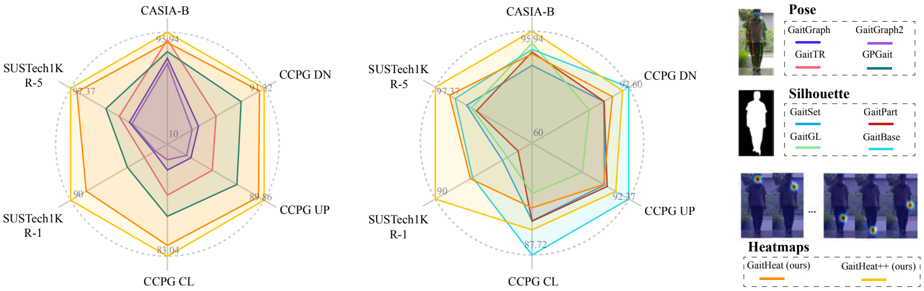

Fig. 1: GaitHeat achieves state-of-the-art performance for pose-based gait recognition and comparable results to silhouette-based approaches. Best viewed in color.

Recent progress in gait recognition has primarily utilized two types of input: silhouettes and skeletons. The silhouettes reserve the human shape and Convolution Neural Network (CNN) is mainly utilized to extract spatial-temporal patterns [1,4,6,13,14,18,21,29]. Contrastively, the skeletons encode the human pose and Graph Convolution Network (GCN) is usually employed to discover the individualized features [7,20,33,34,39]. Although the silhouette-based research holds state-of-the-art on most existing benchmarks, it has great potential to perform gait recognition on human pose which is theoretically robust to the carrying and clothing covariates.

However, the current pose-based gait recognition with skeletons has considerable drawbacks. (1) **Poor Performance.** Despite the recent improvement [7, 19,33,34,39], the performance of skeleton-based methods is frustratingly inferior to the silhouette-based ones on most benchmarks [1,4,6,21]. The probable reason is that the shape information is almost inevitably lost in the skeletons. (2) **Lack of Generalization Ability.** The skeleton-based methods often fail to generalize the different domains [7,33,34,39], which is largely caused by the dependency on the accurate keypoint positions.

Keeping these limitations in mind, we revisit the entire pipeline of pose-based gait recognition starting from RGB frames and find that the current pose-based research takes it for granted that the skeletons are used as the intermediate representations between pose estimation and gait recognition. However, we argue that *it is not definitely necessary to take the skeletons to encode the pose from the perspective of gait recognition.* Based on this point of view, we further observe that most state-of-the-art methods for pose estimation [31,37] regress the skeletons according to the heatmaps that encode the probability distribution of each keypoint separately. A natural idea comes to us: *can we adopt the heatmaps to perform gait recognition to skip the skeletons?* Intuitively, the heatmaps have two important advantages to exactly deal with the drawbacks of skeletons mentioned above. (1) **Encoding the pose and partially reserving the shape information.** (2) **Enhanced robustness to the errors of keypoint predictions.** For clarity, we call the heatmap-based strategy *Cut out the Middleman* where the skeletons are omitted in the entire pipeline.

However, performing pose-based gait recognition with heatmaps also faces crucial challenges. (1) *Pretreatment Sensitivity.* Analogous to silhouettes, the heatmaps of various sizes need to be pretreated into a fixed resolution for the convenience of feature extraction. However, the distribution of the edge area encoding the shape in multi-channel heatmaps is more complex than the corresponding binary silhouette, and the resize operation [5] is more likely to corrupt the individualized details. Furthermore, due to changes in the view and human movement, the positions of human bodies in the heatmap are not strictly aligned, which reduces the generalization ability of the model. (2) *Integration Confusion.* Instead of a single channel for a silhouette or a graph for a skeleton, there are a couple of heatmaps estimated from a frame to encode the human pose, as shown in Fig. 1. How to effectively integrate them and extract discriminative features that sufficiently utilize human part semantic information, is challenging and remains an open question.

In this work, we make a pioneering attempt to conduct pose-based gait recognition with heatmaps and provide a simple yet effective framework to tackle the above challenges. Specifically, for the first challenge, we innovate the entire pipeline and perform the pretreatment in RGB space ahead of pose estimation to ensure that the generated heatmaps can be more easily adopted as the input for recognition without further processing. Our key intuition lies in that the RGB frames are less sensitive to the resize operations and the alignment can largely be achieved implicitly by using the same detection and pose estimation models. This simple yet insightful improvement can remarkably benefit downstream recognition. To tackle more challenging scenarios, such as body rotation and bias, we introduce a Pose-Guided Heatmap Alignment module to further eliminate the dataset covariance and improve the generalization ability of our network. For the second challenge, we investigate the partial fusion strategies to integrate multi-channel headmaps of a frame, and effectively incorporate them in a global-local framework.

To summarize, the main contributions of this work can be boiled down to three aspects:

- We present a new perspective on pose-based gait recognition and propose to adopt the heatmaps, which are generated from the upstream task, instead of the skeletons as the intermediate representations.
- We point out the key challenges to performing gait recognition based on heatmaps and provide a simple yet non-trivial solution based on the insights of the entire pipeline and a comprehensive study.
- Extensive experiments demonstrate the potential of pose-based gait recognition with heatmaps. For example, on SUSTech1K [28], our approach outperforms the recent silhouette-based and skeleton-based baselines by a large margin, *e.g.*, $\uparrow$ 13.88% *vs.* GaitBase [4] and $\uparrow$ 47.55% *vs.* GPGait [7] in terms of rank-1 accuracy.

2 Related Work

2.1 Gait Recognition

Appearance-based Methods. Appearance-based methods are the prevalent choice for extracting gait patterns, focusing on spatial-temporal correlations and fine-grained features. For instance, GaitSet [1] treats gait silhouette sequences as an unordered set, employing a set pooling mechanism to integrate temporal information. GaitPart [6] incorporates a micro-motion capture module for modeling temporal dependencies. GaitGL [21] advances this approach with a 3D convolution-based feature extractor that captures both global and local spatial-temporal details. DyGait [35] introduces a Dynamic Augmentation Module, enhancing the spatial-temporal representation of the body's dynamic regions. DroneGait [16] contributes to the field with a novel dataset obtained from different vertical angles, utilizing distillation to refine gait recognition from high perspectives. GaitParsing [36] marks a significant leap forward by leveraging semantic parsing to boost gait recognition accuracy. GaitBase [4] emerges as a new baseline model, distinguished by its uncomplicated yet robust design.

Model-based Methods. PoseGait [20] leverages 3D human body keypoints as the representation for feature extraction, utilizing human pose and prior knowledge. These features are then processed by a CNN to extract gait information. GaitGraph [34] and its more advanced version GaitGraph2 [33] adopt GCN for gait recognition, regarding the human body as a graph. GaitTR [39] and Gait-Mixer [26] bring innovations by introducing self-attention mechanisms, which allow for broader spatial relationships, and by using temporal convolution with enlarged kernels to capture extended temporal patterns. GPGait [7] introduces Human-Oriented Transformation alongside a Part-Aware Graph Convolutional Network, enhancing generalization across various datasets. Additionally, PAA [9] presents a physics-augmented autoencoder for 3D skeleton-based gait recognition, achieving notable advancements in performance.

2.2 Heatmap Representation

Heatmap is widely used in pose estimation [31, 37] as an intermediate representation since it can preserve the spatial structure of input image compared to regression-based methods. Liu *et al.* [22] pioneer the use of heatmaps as input for action recognition. PoTion [2] aggregates heatmaps across the temporal dimension into a 2D input using color encodings. It utilizes a shallow CNN to extract features that complement traditional appearance and motion streams. PoseConv3D [3] advances this technique by introducing 3D volume heatmaps and employing 3D convolutional neural networks to capture spatial and temporal information, showcasing superior performance across various action recognition datasets. This idea has been extended to gait recognition by Liao *et al.* [19] and Fan *et al.* [5]. Liao *et al.* explore the byproduct of pose estimation as the input, while Fan *et al.* regenerate a heatmap from the pose coordinates as the input feature. Both works lack a specific network design to fully leverage the semantic information of the new modality. In addition, the former method has not demonstrated competitive results on gait datasets compared with skeleton-based methods [7,33,34,39], while the latter suffers from the redundant post-processing and shape loss.

3 Methods

Overview As illustrated in Fig. 2(b), RGB images undergo human detection for human bounding boxes. The images, cropped to focus on the human area according to these boxes, are subsequently fed into a pose estimation algorithm to generate the corresponding heatmaps. Instead of decoding these heatmaps into discrete keypoints in Fig. 2(a), we utilize the human-box-centered and informative heatmaps as the input of our downstream model. As shown in Fig. 3, Pose-Guided Heatmap Alignment is utilized to reduce the effects of covariance and improve generalization across various datasets. Following this, a parameter-sharing global-local network is employed to thoroughly exploit the semantic spatial context and positional cues at both global and local levels. Moreover, we propose a multi-stage feature fusion branch to achieve a compact embedding, facilitating rapid retrieval. To integrate the independent partial features in the local branches, the Max Response (MR) is introduced. Subsequently, Set Pooling [1] and Horizontal Pyramid Mapping (HPM) [1] are employed to extract part features, followed by the application of BNNeck [23] to refine the feature space. The entire process is supervised using both Triplet loss [12] and Cross-entropy loss.

3.1 Revisiting Pose-based Gait Recognition

Heatmaps *vs.* Skeletons As Fig. 2(a) shows, the previous pose-based method involves an extra step of transforming heatmaps into a skeleton. In this process, discrete 2D skeletons in the original images are derived by decoding, based on

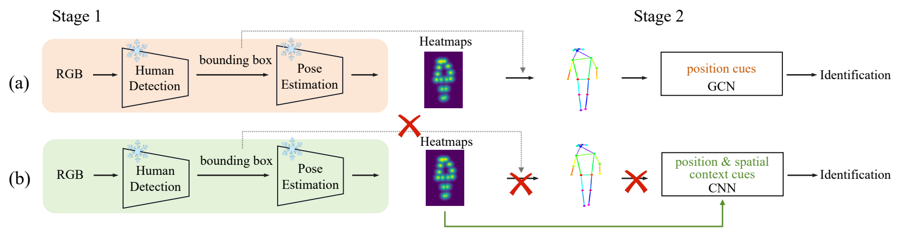

Fig. 2: A comparison of existing pose-based methods with ours at a framework level. (a) Existing methods: shape-loss and estimation error-sensitive. (b) Ours: shape-retention and estimation error-insensitive. We leverage intermediate visual presentations from the upstream task in *Stage1* for further gait feature extraction in *Stage2*.

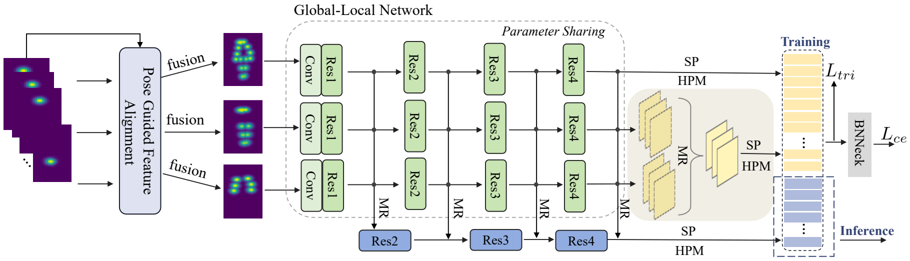

Fig. 3: An overview of GaitHeat. 'Res' denotes residual block [10], 'MR' stands for the *max response* operation, 'SP' indicates set pooling [1], and 'HPM' refers to horizontal pyramid mapping [1]. We first align the heatmaps with a parameter-free Pose-Guided Heatmap Alignment module and then fuse them by a human partition strategy. Finally, we extract informative features with a parameter-sharing global-local network and design a multi-stage feature fusion branch for efficient storage and retrieval. More details about **Pose-Guided Heatmap Alignment** can be found in Fig. 4.

the positions of the maximum values in the heatmap and the locations of the bounding boxes. As shown in Fig. 2(b), we bypass the decoding step and obtain the heatmaps of size $V \times H \times W$ from the upstream, where V represents the number of joints, H and W denote the height and width of a heatmap, respectively. Compared to discrete 2D skeletons, the advantages of 2D heatmaps can be summarized in two aspects: (1) Informative Shapes: previous pose-based gait recognition encounters the issue of human shape loss, which leads to the inferior performance of pose-based methods. The 2D heatmap we proposed mitigates this issue by providing informative shape information. As demonstrated in Fig. 2, the heatmap displays keypoint peaks within a silhouette-like shade, offering a richer representation of body shape. (2) Enhanced Robustness: Heatmaps represent the probability distribution of each keypoint in a continuous space, in contrast to discrete point representation. This feature enables a model to overcome uncertainties and variations, which is valuable in scenarios with imprecise or obscured keypoints. Even at night, heatmaps can still offer partial locational cues with informative shape information, whereas discrete keypoints might fail or yield unreliable results.

3.2 Heatmap-based Gait Recognition

In this section, we tackle two major challenges in pose-based gait recognition using heatmaps: sensitivity to pretreatment and confusion in integration. To address these issues and demonstrate the viability and effectiveness of heatmap-based gait recognition, we introduce a simple yet effective solution. This involves innovating the heatmap generation process, introducing an alignment module to mitigate the influence of covariance, and investigating various strategies for the effective integration and extraction of multi-channel heatmaps.

Challenge 1: Pretreatment Sensitivity

Pretreatment in RGB Space. The previous methods for gait recognition typically involve pretreatment steps to achieve a generalized representation. For instance, GPGait [7] employs Human-Oriented Transformation to center and rescale the skeleton, while GaitSet [1] utilizes normalization methods [32] to center and resize silhouettes.

However, the distribution of values in multi-channel heatmaps is more complex than that of binary silhouettes, making the resize operation more prone to the loss of individualized details. In contrast, our approach performs pretreatment in the RGB space before pose estimation, ensuring that the resulting heatmaps can be used for recognition without resizing processing. The key idea is that RGB frames are less sensitive to resizing, and centering can be implicitly achieved using consistent detection and pose estimation models.

Specifically, as shown in Fig. 2(b), our heatmap generation process involves three steps. (1) Detection: The human detection algorithm determines the corresponding bounding box for a human using top-left and bottom-right coordinates. (2) Resize in RGB space: Padding is applied to the detected box with a fixed ratio,*e.g.,*×1.25, ensuring that the predicted heatmap remains within the boundary. To centralize the human area in RGB space and avoid complex post-processing, the cropped image based on the detection box is resized to a fixed dimension of $3 \times H^{'} \times W^{'}$. For the area of the resized detection box that extends beyond the RGB image, we apply zero-padding on the image to preserve the human ratio. This step ensures the body ratio remains unchanged and results in a consistently sized image. (3) Pose estimation: The resized image is then inputted into pose estimation, leading to the generation of heatmaps sized $V \times H \times W$. These heatmaps feature a fixed size and a human-centered property, thus eliminating the need for complex and lossy post-processing in [5].

Pose-Guided Heatmap Alignment. Although generating heatmaps in RGB space has to some extent centralized the human area, dealing with significant human tilt and bias, as caused by camera perspectives and human movement, remains a challenge. To tackle the issue, we proposed a Pose-Guided Heatmap Alignment module (PGHA) to mitigate the effects of covariance, which involves a rotation transform to correct tilt and a translation transform to adjust for movement bias. The proposed module utilizes a few keypoints from the heatmaps and guides the

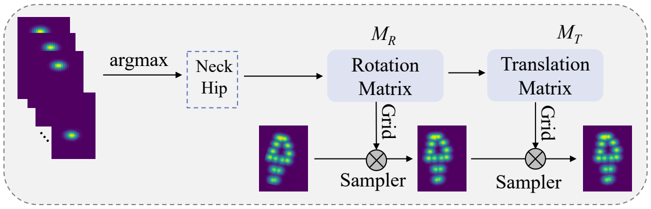

Fig. 4: The Pose-Guided Heatmap Alignment Module in GaitHeat.

alignment without relying on additional parameters. As illustrated in Fig. 4, the PGHA consists of two main steps: (1) *Generating the transform matrix M*, *i.e.*, rotation matrix and translation matrix. (2) *Generating grid and sampling*, where the grid establishes a pixel-wise location mapping between the transformed and original images. During the sampling phase, bilinear interpolation is utilized to precisely determine the pixel values for the transformed image based on the corresponding mapping positions of the original image.

Specifically, for the given multi-channel heatmaps X, we employ argmax to locate maximum response positions P_{neck} and P_{hip} within the corresponding heatmaps channels. Considering the line connecting the neck and hip as analogous to the human spine, which is presumed to align vertically in most scenarios, we assess the rotation angle θ.

$$\theta = \arctan(P^x_{neck} - P^x_{hip}, P^y_{neck} - P^y_{hip}). \tag{1}$$

We utilize the rotation matrix M_R to rotate the heatmaps around the midpoint P_{mid} between the neck and hip, thereby eliminating the effects of inclination covariance.

$$M_R = \begin{bmatrix} \cos(-\theta) & -\sin(-\theta) & (1-\cos(-\theta)) \times P^x_{mid} + \sin(-\theta) \times P^y_{mid} \\ \sin(-\theta) & \cos(-\theta) & -\sin(-\theta) \times P^x_{mid} + (1-\cos(-\theta)) \times P^y_{mid} \end{bmatrix}. \tag{2}$$

Then we produce the translation matrix M_T, aimed at reducing the influence of positional deviations due to the human movement. In detail, we calculate the bias between rotated P_{neck} and the predefined fixed position P_{align} across both the x-axis and y-axis. The biases identified are then used to formulate M_T:

$$\begin{aligned} [x_{bias}, y_{bias}]^\top &= [P^x_{align}, P^y_{align}]^\top - M_R[P^x_{neck}, P^y_{neck}, 1]^\top, \\ M_T &= \begin{bmatrix} 1 & 0 & -x_{bias} \\ 0 & 1 & -y_{bias} \end{bmatrix}. \end{aligned} \tag{3}$$

After obtaining the transformation matrices, we employ the same approach as in the STN [15] to generate a grid and sampling corresponding value using the transformation matrixes, *i.e* M_R and M_T. For instance, given a location (x^t, y^t) in the transformed image, the corresponding location (x^s, y^s) in the input image is determined by the transformation matrix M.

$$\begin{pmatrix} x^s \\ y^s \end{pmatrix} = M \begin{pmatrix} x^t \\ y^t \\ 1 \end{pmatrix}. \tag{4}$$

The value at (x^t, y^t) in the transformed image is then sampled from the input image in (x^s, y^s) using bilinear interpolation for precise value acquisition, as (x^s, y^s) might not be integer.

Challenge 2: Integration Confusion

Multi-Channel Heatmap Integration. For a set of T heatmaps $X_{in} \in \mathbb{R}^{V\times T\times H\times W}$ aligned by PGHA module, the global features are derived by fusing the entire body's heatmaps through Max operation along the V dimension. This process aggregates the information across all channels to capture a holistic representation of the body's pose. For the local feature, a variety of partitioning strategies, denoted as $U = \{u_1, ...u_K\}$, are utilized to partially fuse the heatmaps. These strategies are designed to selectively integrate semantic information from different regions of human body, enabling the model to focus on specific aspects of the pose that are most discriminative for identification tasks. The operations are formalized as follows:

$$X_g = Max\{X_{in}\}, X_l = concatenate(Max\{X_{in}^{u_1}\}, ..., Max\{X_{in}^{u_K}\}), \tag{5}$$

where $X_g \in \mathbb{R}^{T\times H\times W}$ represents the global heatmaps, while $X_l \in \mathbb{R}^{K\times T\times H\times W}$ denotes the local heatmaps. The *concatenate* means concatenating the local features from K kinds of partitioning strategies in the batch dimension for parameter sharing.

A Global-Local Framework. Gait recognition, being a fine-grained task, requires the model to not only discover the overall structure but also needs the model to focus on the specific human body regions to extract discriminative identification features. With this in mind, we propose utilizing a global-local network to meet these demands and fully explore the semantic information of heatmaps. In this framework, the global branch processes information from the entire human body, whereas the local one is dedicated to enhancing the details from specific body areas. To further accelerate the model's learning process and improve the interaction between the global context with local detailed information, as illustrated in Fig. 3, these heatmaps are fed into a parameter-sharing network, denoted as F_{share} to obtain the feature maps f_g and f_l, respectively:

$$f_g = F_{share}(X_g), f_l = F_{share}(X_l), \tag{6}$$

where $f_g \in \mathbb{R}^{D\times T\times \frac{H}{4}\times \frac{W}{4}}$ and $f_l \in \mathbb{R}^{K\times D\times T\times \frac{H}{4}\times \frac{W}{4}}$ with D denoting the channel dimension. To enhance the interaction between these parts, we use a Max Response (MR) operation. It merges partial representations into a global-like representation by maximizing the response of each part, thereby improving the complementarity of the features across different body partitions.

$$\begin{aligned} f_{agg} &= MR(f_{local}) \\ &= \max\{f_{local_1}, ...f_{local_K}\}, \end{aligned} \tag{7}$$

where aggregated local feature $f_{agg} \in \mathbb{R}^{D\times T\times \frac{H}{4}\times \frac{W}{4}}$. After MR, set pooling separately merges the local feature f_{agg} and global feature f_g in the temporal dimension. Subsequently, the local and global features independently undergo horizontal pyramid mapping HPM [8] and BNNeck [23] for loss optimization.

A Multi-Stage Fusion Branch. Given the high-dimensional feature representation resulting from multi-branches feature extraction, which could require extensive storage and decelerate the retrieval process, a multi-stage fusion branch is introduced. This approach employs the Max Response operation to fuse features from different receptive fields at various stages, and then extracts gait features from these fused features using a parameter-independent network, denoted as F_{fusion}. Taking the fusion branch at the i-th layer, denoted as f^i_{fusion}, as an example, we use MR to fuse multiple features from the global-local branches, and then add the fusion features produced by the current layer F^i_{fusion}.

$$f^i_{fusion} = MR(f^i_{local}, f^i_{global}) + F^i_{fusion}(f^{i-1}_{fusion}). \tag{8}$$

Training and Testing. During the training phase, we deploy both triplet loss [12] and cross-entropy loss to simultaneously train the global, local and fusion branches. The loss function is defined as:

$$L = (L^{global}_{tri} + L^{global}_{ce}) + (L^{local}_{tri} + L^{local}_{ce}) + (L^{fusion}_{tri} + L^{fusion}_{ce}). \tag{9}$$

In this formula, $L^{global}_{tri}, L^{local}_{tri}, L^{fusion}_{tri}$ denote the triplet losses, $L^{global}_{ce}, L^{local}_{ce}, L^{fusion}_{ce}$ represent the cross-entropy losses. *In the testing phase, we only take the fusion embedding for inference.*

4 Experiments and Results

4.1 Settings

Datasets For comprehensive comparisons, we utilize three well-known gait datasets: CASIA-B [38], CCPG [17], and SUSTech1K [28]. We strictly follow the official evaluation protocols in our experiments. More details about the datasets can be found in the supplementary materials.

Implementation Details

Data Pre-processing. Following the revisited pipeline, we are able to generate heatmaps of a fixed size, 64×48, from the top-down pose estimation HRNet [31] for three datasets. To evaluate the impact of our method on the latest pose estimation algorithms, we propose GaitHeat++, which maintains the same architecture but differs by employing heatmaps from ViTPose [37]. For the data cleaning, we adopt the average score as a metric to assess the frame quality and set a threshold of 0.4 to eliminate frames of low quality.

Gait Recognition. During training, we randomly select 30 frames from the unordered gait sequence set, while for testing, we utilize all available frames. The backbone in GaitHeat is ResNet9 [11] adapted from the GaitBase [4]. Furthermore, we reproduce four pose-based gait recognition algorithms based on FastPoseGait [25]. When θ described in Eq. (1) is larger than the threshold γ, we adopt the operation of PGHA. γ is 5° for CASIA-B and SUSTech1K, and 20° for CCPG, based on the statistics of specific datasets. To further mitigate the risk of overfitting, we introduce a probability of 0.2 to perform PGHA when θ is smaller than γ.

4.2 Performance Comparsion

Compared with Skeleton-based Methods GaitHeat exhibits significant performance superiority over existing pose-based methods. Specifically, as shown in Tab. 1 and Tab. 2, compared with the second-best result, it delivers an 18.84% increase in rank-1 accuracy for CCPG in CL, and 34.49% for SUSTech1K. Moreover, GaitHeat demonstrates exceptional robustness in uncontrolled environments, as illustrated in Fig. 5. In such scenarios, traditional methods based on silhouettes and keypoints sometimes fail to accurately capture the human pose. However, heatmaps, as employed by GaitHeat, effectively encode the human pose even in challenging low-light conditions. Specifically, within the SUSTech1K dataset under nighttime conditions, while earlier pose-based methods reach their highest rank-1 accuracies of 31.8%, GaitHeat surpasses these best-performing methods by a significant margin of 29.65%, underlining its effectiveness and robustness in accurately capturing human gait patterns across diverse and challenging conditions.

Compard with Other State-of-the-Arts Methods The comparison method mentioned above utilizes the pose estimation outcomes from HRNet [31]. To further explore the potential of the heatmap-based method, we employ high-quality prediction results from ViTPose [37] as input, naming this enhanced approach GaitHeat++. As shown in Tab. 1 and 2, GaitHeat++ significantly enhances accuracies beyond the original versions, achieving results comparable to those of silhouette-based methods. For instance, compared with its vanilla version, GaitHeat++ improves the rank-1 accuracy by 7.77% in CASIA-B, 9.45% of CL in CCPG, and 13.06% in SUSTech1K. Furthermore, in comparison with the silhouette-based method of GaitGL, GaitHeat++ secures a 4.11% improvement in rank-1 accuracy on CASIA-B. Remarkably, even when compared to

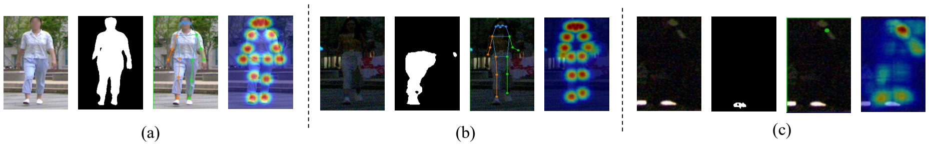

Fig. 5: The visualization of some examples in SUSTech1K. From left to right: the original RGB image, silhouette, keypoints, and heatmap. In case (a), the three representations accurately depict daytime scene features. In case (b), segmentation fails in the evening. In case (c), under darkness, the heatmap provides more positional cues and shape information.

Table 1: Performance of state-of-the-art pose-based and silhouette-based methods on CASIA-B [38] and CCPG [17]. The bold values represent the best result of pose-based methods.

Method	Input	CASIA-B				CCPG							
		NM	BG	CL	Mean	CL		UP		DN		BG	
		R-1 (%)				R-1 & mAP(%)							
GaitSet (AAAI'19) [1]	Sils	95.00	87.20	70.40	84.20	77.80	46.58	83.28	60.94	82.38	61.85	85.84	64.18
GaitPart (CVPR'20) [6]		96.20	91.50	78.70	88.80	75.83	44.10	82.67	60.15	83.75	59.93	86.43	63.38
GaitGL (ICCV'21) [21]		97.40	94.50	83.60	91.83	73.81	35.52	80.59	49.04	79.83	48.14	83.36	52.58
GaitBase (CVPR'23) [4]		97.60	94.00	77.40	89.67	87.72	58.56	92.37	72.93	92.60	73.10	93.17	76.58
GaitGraph (ICIP'21) [34]	Pose	86.37	76.50	65.24	76.04	20.50	11.56	30.74	20.13	39.11	23.73	31.14	19.68
GaitGraph2 (CVPRW'22) [33]		80.29	71.40	63.80	71.83	15.54	5.64	21.06	8.90	26.21	10.26	20.14	8.89
GaitTR (ES'23) [39]		94.81	87.65	88.07	90.18	40.92	18.65	46.62	26.29	47.75	27.16	42.41	24.15
GPGait (ICCV'23) [7]		93.59	80.15	69.30	81.01	54.75	25.78	65.60	38.44	71.06	41.04	65.36	37.83
GaitHeat (Ours)		98.22	92.14	74.15	88.17	73.59	41.06	83.19	58.04	86.47	60.27	90.53	66.36
GaitHeat++ (Ours)		**99.60**	**97.88**	**90.35**	**95.94**	**83.04**	**54.74**	**89.86**	**71.43**	**91.32**	**72.40**	**93.09**	**77.21**

Table 2: Evaluation with different attributes on SUSTech1K [28]. The bold values represent the best result of pose-based methods. NM, BG, CL, CA UM, UN, OC and NG are abbreviations of Normal, Bag, Clothing, Carrying, Umbrella, Uniform, Occlusion and Night.

Method	Input	Probe Sequence (Rank-1 acc)								Overall	
		NM	BG	CL	CA	UM	UN	OC	NG	R-1	R-5
GaitSet (AAAI'19) [1]	Sils	69.10	68.25	37.44	65.01	63.08	61.00	67.19	23.04	65.04	84.76
GaitPart (CVPR'20) [6]		62.20	62.81	33.08	59.53	57.25	54.85	57.20	21.75	59.19	80.79
GaitGL (ICCV'21) [21]		67.11	66.16	35.92	63.31	61.58	58.07	66.59	17.88	63.14	82.82
GaitBase (CVPR'23) [4]		81.46	77.48	49.60	75.77	75.55	76.66	81.40	25.92	76.12	89.39
LidarGait (CVPR'23) [28]	LiDAR	91.80	88.64	74.56	89.03	67.50	80.86	94.53	90.41	86.77	96.08
GaitGraph (ICIP'21) [34]	Pose	22.80	20.72	8.69	19.36	14.36	22.22	31.45	18.38	19.96	43.49
GaitGraph2 (CVPRW'22) [33]		25.98	21.95	7.77	22.03	17.65	22.28	28.46	20.81	22.11	46.27
GaitTR (ES'23) [39]		31.39	33.31	18.58	31.30	27.97	36.83	38.44	21.94	31.71	57.21
GPGait (ICCV'23) [7]		43.96	40.98	24.28	41.42	38.34	47.00	57.99	31.80	42.45	65.41
GaitHeat (Ours)		81.47	77.21	44.32	77.60	71.57	78.21	87.79	61.45	76.94	91.63
GaitHeat++ (Ours)		**93.46**	**91.08**	**76.01**	**90.41**	**84.74**	**88.18**	**96.71**	**73.01**	**90.00**	**97.37**

the best-performing method on SUSTech1K [28], which utilizes Lidar modality, GaitHeat++ still demonstrates superior performance with a significant improvement margin of 3.23%. This impressive achievement highlights the robustness of our method across various environments, including challenging conditions such as low-light situations at night.

Cross-domain Evaluation To evaluate the generalization capabilities of pose-based methods, we conduct cross-dataset tests, *i.e.*, training on source datasets and evaluating on target datasets. As presented in Tab. 3, GaitHeat demonstrates strong generalization when evaluated on CASIA-B and SUSTech1K. However, its performance on the CCPG testing set is less impressive, with a considerable result in the BG setting but weaker performance in clothing conditions (CL, UP, DN). We attribute this discrepancy to differences in dataset distribution. The CASIA-B and SUSTech1K feature fewer cloth-changing sequences compared to CCPG, which challenges GaitHeat's ability to filter out gait-irrelevant features in the presence of significant clothing variation. These analyses indicate that the distribution of datasets can influence the generaliza-

Table 3: Cross-Domain evaluation on three popular datasets of recent state-of-the-art pose-based methods. CA, CC and SU are abbreviations of CASIA-B, CCPG and SUSTech1K. The bold values represent the best result.

(a) Test on CASIA-B

Method	Train Set→Test Set							
	CC→CA				SU→CA			
	NM	BG	CL	Mean	NM	BG	CL	Mean
GaitGraph	4.93	4.81	3.62	4.45	3.81	3.65	3.11	3.52
GaitGraph2	7.67	6.49	5.47	6.54	13.65	10.68	6.35	10.23
GaitTR	4.37	4.30	4.37	4.35	5.38	4.78	4.74	4.97
GPGait	40.80	33.09	19.15	31.01	56.36	44.41	22.71	41.16
GaitHeat	54.18	46.30	32.98	44.49	**77.59**	63.57	23.16	54.77
GaitHeat++	**58.85**	**50.45**	**35.14**	**48.15**	74.78	**66.37**	**26.73**	**55.96**

(b) Test on SUSTech1K

Method	Train Set→Test Set			
	CC→SU		CA→SU	
	Rank1	Rank5	Rank1	Rank5
GaitGraph	1.03	3.72	1.1	3.88
GaitGraph2	0.82	3.12	0.81	2.95
GaitTR	0.61	2.23	0.84	3.03
GPGait	2.48	6.94	3.48	8.55
GaitHeat	11.45	25.00	10.51	22.41
GaitHeat++	**20.40**	**38.99**	**15.56**	**31.53**

(c) Test on CCPG

Method	Train Set→Test Set							
	CA→CC				SU→CC			
	CL	UP	DN	BG	CL	UP	DN	BG
GaitGraph	1.63	1.47	1.45	1.88	0.95	1.73	2.89	1.96
GaitGraph2	0.86	0.52	1.28	1.54	1.42	2.43	4.85	9.04
GaitTR	2.66	2.43	1.96	3.07	1.03	1.73	1.62	1.71
GPGait	**11.85**	**16.46**	**19.75**	21.50	**9.49**	**13.61**	**21.45**	21.16
GaitHeat	5.63	9.79	18.30	**21.59**	5.41	11.90	19.06	**27.47**
GaitHeat++	8.16	13.69	17.19	20.48	2.36	4.42	7.15	13.06

tion ability of the model. With data from a more diverse distribution, the model can achieve stronger generalization ability.

4.3 Ablation Study

Impact of Heatmap Type. The heatmap representation in GaitHeat offers a more detailed shape depiction of the human body and avoids the loss of discriminative information caused by the complex normalization process, compared to the regenerated heatmap utilized in SkeletonGait [5]. For a fair comparison, we adopt the same approach as SkeletonGait to generate a heatmap from keypoint coordinates. As shown in Tab. 4, the original heatmap substantially surpasses the regenerated method in performance. This can be attributed to the ability of the original heatmap to preserve essential pose and shape information from the upstream task, whereas the process of converting keypoints back into a heatmap leads to irreversible loss of information. Additionally, utilizing the heatmap from upstream reduces the dependency on complex transformations between heatmaps and keypoints, significantly improving efficiency in practical scenarios. By retaining shape information from earlier stages and minimizing information loss through extensive post-process operations, GaitHeat achieves a 12.92% increase in rank-1 accuracy over the regenerated approach on CASIA-B, 17.6% on CCPG in the CL setting, underscoring the effectiveness and efficiency of utilizing heatmap representations in gait task.

Impact of Pose-Guided Heatmap Alignment. The Pose-Guided Heatmap Alignment (PGHA) module plays a crucial role in aligning heatmaps to enhance the

Table 4: Effect of different heatmap types.

Method	CASIA-B				CCPG							
	NM	BG	CL	Mean	CL		UP		DN		BG	
	R-1 (%)				R-1 & mAP(%)							
SkeletonGait	91.47	76.59	57.69	75.25	55.99	27.05	65.86	40.24	74.21	45.22	70.90	44.29
ours	**98.22**	**92.14**	**74.15**	**88.17**	**73.59**	**41.06**	**83.19**	**58.04**	**86.47**	**60.27**	**90.53**	**66.36**

Table 5: Effect of Pose-Guided Heatmap Alignment(PGHA). The arrows (→) point from the source domain to the target domain.

Setting	CASIA-B→CASIA-B				CCPG→CASIA-B				CCPG→CCPG				CASIA-B→CCPG			
	NM	BG	CL	Mean	NM	BG	CL	Mean	CL	UP	DN	BG	CL	UP	DN	BG
w/o PGFA	98.36	**93.04**	**77.43**	**89.61**	**54.46**	**46.82**	31.69	44.32	70.63	82.50	84.26	89.16	3.09	7.54	12.00	14.59
w PGFA	**98.22**	92.14	74.15	88.17	54.18	46.30	**32.98**	**44.49**	**73.59**	**83.19**	**86.47**	**90.53**	**5.63**	**9.79**	**18.30**	**21.59**

generalization of the model across various datasets, while also ensuring comparable performance in the source domain. As demonstrated in Tab. 5, PGHA is highly effective in mitigating discrepancies across different dataset domains, addressing issues like variations in human body orientation and spatial displacements. For example, employing PGHA leads to a promising performance improvement of 7% on CASIA-B→CCPG in BG setting. This showcases PGHA's capacity to bridge the gap between distinct datasets, enhancing the generalization abilities and effectiveness in cross-domain setting.

5 Conclusion and Limitations

Conclusion Our work introduces a new pipeline that bypasses the step of decoding heatmap into coordinates found in previous pose-based gait recognition approaches, retaining the informative and human-centered representation from pose estimation. Our method not only fully utilizes information from upstream but also reduces dependency on complex post-normalization. We also propose a simple baseline to demonstrate its effectiveness. Our model significantly outperforms previous pose-based gait recognition methods in terms of performance, robustness, and generalization across various datasets in most cases.

Limitations and Future Work While we have set up a simple baseline to evaluate the effectiveness of heatmaps from our revised pipeline, GaitHeat still encounters some limitations that can be further explored in future work. a) The heatmap representation requires more computation and storage compared to coordinates. b) The semantic information in the heatmap is only preliminarily explored. c) The objectives of upstream tasks and GaitHeat are different, where the former focuses on body structure and the latter on gait patterns. Effectively combining these two tasks (*e.g.*, training in an end-to-end manner) can encourage the former to also concentrate on gait features, rather than solely on body structure.

Acknowledgements

This work is jointly supported by National Natural Science Foundation of China (62276025, 62206022), Beijing Municipal Science & Technology Commission (Z23 1100007423015) and Shenzhen Technology Plan Program (KQTD2017033109321 7368).

References

1. Chao, H., He, Y., Zhang, J., Feng, J.: Gaitset: Regarding gait as a set for cross-view gait recognition. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 8126–8133 (2019)
2. Choutas, V., Weinzaepfel, P., Revaud, J., Schmid, C.: Potion: Pose motion representation for action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7024–7033 (2018)
3. Duan, H., Zhao, Y., Chen, K., Lin, D., Dai, B.: Revisiting skeleton-based action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2969–2978 (2022)
4. Fan, C., Liang, J., Shen, C., Hou, S., Huang, Y., Yu, S.: Opengait: Revisiting gait recognition towards better practicality. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9707–9716 (2023)
5. Fan, C., Ma, J., Jin, D., Shen, C., Yu, S.: Skeletongait: Gait recognition using skeleton maps. In: Proceedings of the AAAI Conference on Artificial Intelligence (2024)
6. Fan, C., Peng, Y., Cao, C., Liu, X., Hou, S., Chi, J., Huang, Y., Li, Q., He, Z.: Gaitpart: Temporal part-based model for gait recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14225–14233 (2020)
7. Fu, Y., Meng, S., Hou, S., Hu, X., Huang, Y.: Gpgait: Generalized pose-based gait recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19595–19604 (2023)
8. Fu, Y., Wei, Y., Zhou, Y., Shi, H., Huang, G., Wang, X., Yao, Z., Huang, T.: Horizontal pyramid matching for person re-identification. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 8295–8302 (2019)
9. Guo, H., Ji, Q.: Physics-augmented autoencoder for 3d skeleton-based gait recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19627–19638 (2023)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
11. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14. pp. 630–645. Springer (2016)
12. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person reidentification. arXiv preprint arXiv:1703.07737 (2017)
13. Hou, S., Cao, C., Liu, X., Huang, Y.: Gait lateral network: Learning discriminative and compact representations for gait recognition. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX. pp. 382–398. Springer (2020)

14. Hou, S., Liu, X., Cao, C., Huang, Y.: Gait quality aware network: toward the interpretability of silhouette-based gait recognition. IEEE Transactions on Neural Networks and Learning Systems (2022)
15. Jaderberg, M., Simonyan, K., Zisserman, A., et al.: Spatial transformer networks. Advances in neural information processing systems **28** (2015)
16. Li, A., Hou, S., Cai, Q., Fu, Y., Huang, Y.: Gait recognition with drones: A benchmark. IEEE Transactions on Multimedia (2023)
17. Li, W., Hou, S., Zhang, C., Cao, C., Liu, X., Huang, Y., Zhao, Y.: An in-depth exploration of person re-identification and gait recognition in cloth-changing conditions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13824–13833 (2023)
18. Liang, J., Fan, C., Hou, S., Shen, C., Huang, Y., Yu, S.: Gaitedge: Beyond plain end-to-end gait recognition for better practicality. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part V. pp. 375–390. Springer (2022)
19. Liao, R., Li, Z., Bhattacharyya, S.S., York, G.: Posemapgait: A model-based gait recognition method with pose estimation maps and graph convolutional networks. Neurocomputing **501**, 514–528 (2022)
20. Liao, R., Yu, S., An, W., Huang, Y.: A model-based gait recognition method with body pose and human prior knowledge. Pattern Recognition **98**, 107069 (2020)
21. Lin, B., Zhang, S., Yu, X.: Gait recognition via effective global-local feature representation and local temporal aggregation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14648–14656 (2021)
22. Liu, M., Yuan, J.: Recognizing human actions as the evolution of pose estimation maps. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1159–1168 (2018)
23. Luo, H., Gu, Y., Liao, X., Lai, S., Jiang, W.: 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. pp. 0–0 (2019)
24. Makihara, Y., Nixon, M.S., Yagi, Y.: Gait recognition: Databases, representations, and applications. Computer Vision: A Reference Guide pp. 1–13 (2020)
25. Meng, S., Fu, Y., Hou, S., Cao, C., Liu, X., Huang, Y.: Fastposegait: A toolbox and benchmark for efficient pose-based gait recognition. arXiv preprint arXiv:2309.00794 (2023)
26. Pinyoanuntapong, E., Ali, A., Wang, P., Lee, M., Chen, C.: Gaitmixer: skeleton-based gait representation learning via wide-spectrum multi-axial mixer. arXiv preprint arXiv:2210.15491 (2022)
27. Sepas-Moghaddam, A., Etemad, A.: Deep gait recognition: A survey. IEEE transactions on pattern analysis and machine intelligence **45**(1), 264–284 (2022)
28. Shen, C., Fan, C., Wu, W., Wang, R., Huang, G.Q., Yu, S.: Lidargait: Benchmarking 3d gait recognition with point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1054–1063 (2023)
29. Shiraga, K., Makihara, Y., Muramatsu, D., Echigo, T., Yagi, Y.: Geinet: View-invariant gait recognition using a convolutional neural network. In: 2016 international conference on biometrics (ICB). pp. 1–8. IEEE (2016)
30. Sivarathinabala, M., Abirami, S., Baskaran, R.: A study on security and surveillance system using gait recognition. Intelligent techniques in signal processing for multimedia security pp. 227–252 (2017)
31. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: 2019 IEEE/CVF Conference on Computer Vision

and Pattern Recognition (CVPR). pp. 5686–5696 (2019). https://doi.org/10.1109/CVPR.2019.00584

32. Takemura, N., Makihara, Y., Muramatsu, D., Echigo, T., Yagi, Y.: Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition. IPSJ transactions on Computer Vision and Applications **10**, 1–14 (2018)
33. Teepe, T., Gilg, J., Herzog, F., Hörmann, S., Rigoll, G.: Towards a deeper understanding of skeleton-based gait recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1569–1577 (2022)
34. Teepe, T., Khan, A., Gilg, J., Herzog, F., Hörmann, S., Rigoll, G.: Gaitgraph: Graph convolutional network for skeleton-based gait recognition. In: 2021 IEEE International Conference on Image Processing (ICIP). pp. 2314–2318. IEEE (2021)
35. Wang, M., Guo, X., Lin, B., Yang, T., Zhu, Z., Li, L., Zhang, S., Yu, X.: Dygait: Exploiting dynamic representations for high-performance gait recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13424–13433 (2023)
36. Wang, Z., Hou, S., Zhang, M., Liu, X., Cao, C., Huang, Y.: Gaitparsing: Human semantic parsing for gait recognition. IEEE Transactions on Multimedia (2023)
37. Xu, Y., Zhang, J., Zhang, Q., Tao, D.: Vitpose: Simple vision transformer baselines for human pose estimation. Advances in Neural Information Processing Systems **35**, 38571–38584 (2022)
38. Yu, S., Tan, D., Tan, T.: A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In: 18th international conference on pattern recognition (ICPR'06). vol. 4, pp. 441–444. IEEE (2006)
39. Zhang, C., Chen, X.P., Han, G.Q., Liu, X.J.: Spatial transformer network on skeleton-based gait recognition. Expert Systems p. e13244 (2023)