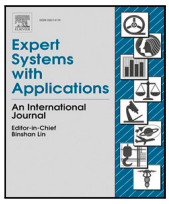
Expert Systems With Applications 275 (2025) 126946

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa



Empirical study of human pose representations for gait recognition

Nicolás Cubero [a,*], Francisco M. Castro [a], Julián R. Cózar [a], Nicolás Guil [a], Manuel J. Marín-Jiménez [b]

[a] Department of Computer Architecture, University of Málaga, Spain
[b] Department of Computing and Numerical Analysis, University of Córdoba, Spain

ARTICLE INFO

Keywords:
Gait recognition
Human pose
Biometrics
Deep learning

ABSTRACT

Gait recognition has gained attention for its ability to identify individuals from afar. Current state-of-the-art approaches predominantly utilize visual information, such as silhouettes, or a combination of visual data and basic body pose information, including skeleton joint coordinates. However, the role of human pose in gait recognition is still underexplored, often leading to poorer results compared to visual approaches. In this work, we propose a novel hierarchical limb-based representation that enhances the depiction of body pose and can be applied to various pose descriptors. Our representation consists of three hierarchical levels: full body, body limbs (arms and legs), and middle limbs (forearms, lower arms, thighs, and shins). This structure enriches the gait description of the overall pose by incorporating the specific movements of each limb. Particularly, we investigate the application of our hierarchical arrangement using two different rich pose descriptors: heatmaps derived from 2D body skeletons and a dense representation obtained from pixel-wise estimation of body pose (*i.e* DensePose). Furthermore, we introduce the PoseGaitGL family of models to better leverage the features derived from our pose representations. By employing our hierarchical pose representations, the proposed model achieves state-of-the-art results in pose-based gait recognition. Thus, the hierarchical heatmap-based and hierarchical DensePose representations attain Rank-1 accuracy of 82.2% and 92.0%, respectively, on the cross-view setup of CASIA-B, and 99.3% and 99.8%, respectively, on TUM-GAID, establishing a new benchmark for pose-based methods. Source code is available at https://github.com/Nico-Cubero/PoseGaitGL.

1. Introduction

The goal of gait-based people identification, also known as *gait recognition*, is to identify individuals by analyzing the way they walk. As a biometric descriptor, gait offers several advantages, including its uniqueness for each person and its non-invasive nature: gait identification can be performed from a distance and does not require the subject to actively cooperate with any capturing device, unlike other biometrics such as fingerprint or iris recognition methods (Singh et al., 2018). In recent years, significant efforts have been directed towards improving gait recognition for people identification, resulting in an increasing number of publications (Sepas-Moghaddam & Etemad, 2023) and expanding applications in various fields such as surveillance (Muramatsu et al., 2013), forensic investigation (Bouchrika et al., 2011), and medical research (Haussler et al., 2024).

Various modalities have been proposed in the literature to capture gait characteristics. These include silhouettes (Fan et al., 2020; Lin et al., 2021; Ma et al., 2023), optical flow (Castro et al., 2024; Castro, Marín-Jiménez, Guil, de la Blanca, 2017), inertial sensors (Delgado-Escaño et al., 2019), gray images (Delgado-Escaño et al., 2021), and even the use of multiple modalities simultaneously (Castro et al., 2020). However, most studies mainly utilize the binary silhouette as a primary descriptor (Fan et al., 2020; Lin et al., 2021; Ma et al., 2023). The silhouette represents changes in the appearance of the human body during walking, providing a rich description of gait characteristics. On the downside, the silhouette includes additional information related to appearance such as body contours, clothing, and other items that are not directly related to gait. Therefore, models relying on such descriptors may be adversely affected by variations in body shape, which can impact their overall performance. The same occurs with other descriptors based on appearance, such as RGB or optical flow.

As an alternative to using silhouettes and other shape-based descriptors, other approaches utilize pose as a gait descriptor. The pose represents the positions of a specific set of joints, without unnecessary shape information, as illustrated in Fig. 1. Typically, gait recognition

* Corresponding author.
E-mail addresses: ncubero@uma.es (N. Cubero), fcastro@uma.es (F.M. Castro), julian@uma.es (J.R. Cózar), nguil@uma.es (N. Guil), mjmarin@uco.es (M.J. Marín-Jiménez).

https://doi.org/10.1016/j.eswa.2025.126946
Received 24 September 2024; Received in revised form 2 February 2025; Accepted 16 February 2025
Available online 28 February 2025

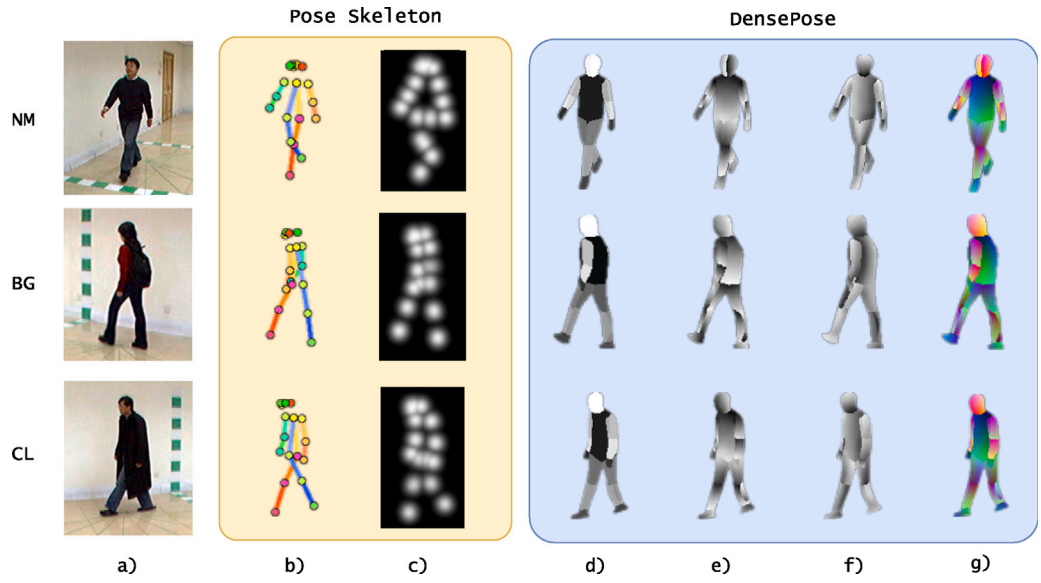


Fig. 1. Pose comparison. Pose descriptors obtained by the different pose models considered in this work. NM: Normal Walking, BG: Walking with Bag, CL: Walking with coats. (a) Raw RGB image on CASIA-B, (b) skeleton model, (c) heatmap image, (d) I body segmentation image (e) U gradient image, (f) V gradient image (g) I + U + V images aggregated on each channel color.

methods that rely on human pose extract the 2D or 3D coordinates of body joints (An et al., 2018; Liao et al., 2017). These methods require specialized architectures to manage the correlations between the motions of different joints to predict the identities of subjects. Nevertheless, these pose-based approaches generally perform worse than those that use visual descriptors such as silhouettes (Fan et al., 2020; Lin et al., 2021; Ma et al., 2023). In our opinion, the information provided by pose-based methods, *i.e.* a set of 2D or 3D coordinates, is quite limited. Typically, these methods capture data from up to 17 joints, each with two or three components, resulting in fewer than 100 values per sample. In contrast, a standard method that uses a silhouette of size 64 × 64 can capture 4096 values per sample. Other approaches suggest utilizing more complex pose descriptors, such as pose heatmaps (Feng et al., 2016; Liao et al., 2022) or DensePose images (Schwarz et al., 2022). On the one hand, heatmaps provide confidence maps for estimating each joint, thereby enhancing the information about the coordinates of the joints. This representation offers greater robustness in scenarios where occlusions occur. However, although models based on heatmap descriptors generally outperform those that rely solely on coordinates (Feng et al., 2016; Liao et al., 2022), they still do not match the performance of models based on appearance descriptors. On the other hand, DensePose (Güler et al., 2018) represents the human body using UV coordinates as a 3D mesh, along with segmentation of body parts. The information provided by this descriptor is significantly richer than that of heatmaps. Conversely, existing methods (Schwarz et al., 2022) do not effectively leverage the information derived from this descriptor, and its performance is still low. Fig. 1c–g illustrate some examples of heatmaps and DensePose.

Recent approaches (Cui & Kang, 2023; Fan et al., 2024; Li et al., 2022, 2020; Peng et al., 2023) have attempted to employ a multimodal setup that combines pose information with silhouettes to enhance model performance. Nonetheless, this multimodal strategy reintroduces body shape information, which diminishes the advantages of using human pose alone.

In this work, we aim to more effectively utilize the information from the gait cycle provided by pose descriptors. We propose a novel hierarchical pose representation for gait recognition that organizes the human body into a three-level structure. This structure consists of the full-body with all its limbs at one level, the main four limbs (the left and right arms, as well as the left and right legs) at second level, and a third level formed by subdividing each limb into two smaller parts, resulting in a total of eight smaller sublimbs. This allows gait recognition approaches to concentrate on the specific movements of each individual limb while also considering the overall gait movement represented at the full body level. Motivated by the higher robustness of pose heatmaps and DensePose descriptors (Güler et al., 2018) we study the application of our hierarchical representation on both these pose descriptors. Moreover, we introduce the PoseGaitGL architecture that leverages the advanced capabilities for visual analysis achieved by leading visual-based state-of-the-art approaches (Chao et al., 2019; Fan et al., 2020; Lin et al., 2021) and is optimized to effectively manage our derived hierarchical heatmap-based, and DensePose-based representations.

The main contributions of this paper are five-fold: *(i)* A hierarchical pose representation that can be applied to any pose descriptor, enriching the gait description and improving the performance of gait recognition methods. *(ii)* An empirical evaluation of our hierarchical representation utilizing two pose descriptors: heatmaps and DensePose. *(iii)* A PoseGaitGL architecture designed to effectively extract gait features from our hierarchical pose representations. *(iv)* A comprehensive experimental study comparing various gait recognition approaches using our hierarchical pose representation. *(v)* State-of-the-art results on the CASIA-B (Yu et al., 2006) and TUM-GAID (Hofmann et al., 2014) datasets in the context of gait recognition based on pose.

The remainder of this paper is structured as follows: we present the related work in Section 2, followed by a description of the human pose representations considered in Section 3. We report and discuss the experimental results in Section 4. We conclude the paper in Section 5, including some future lines of research.

2. Related work

Numerous methods have been proposed in the field of gait recognition in recent decades due to their potential applications in video surveillance, social security, crime prevention, and forensic identification (Sepas-Moghaddam & Etemad, 2023). The most common type of data used in the literature comprises silhouettes or features derived from them. For instance, the Gait Energy Image (GEI) (Han & Bhanu, 2005; Wu et al., 2017) is a straightforward descriptor that condenses the entire sequence of silhouettes into a single frame that captures all the spatio-temporal information. However, GEI is sensitive to variations encountered in real-world scenarios, prompting the development of

new approaches that better exploit the temporal information. GaitSet (Chao et al., 2019), for example, leverages a random stack of silhouettes, and treats each frame independently to extract features that are ultimately combined to form a final descriptor through Horizontal Pyramid Pooling (HPP).

GaitPart (Fan et al., 2020) builds on this trend but introduces a novel part-based model that extracts features from horizontal segments of intermediate convolutional activations. In contrast, GLN (Hou et al., 2020) constructs larger descriptors to enhance the model's discrimination capability by concatenating intermediate convolutional activations, with a compression module added at the end to reduce dimensionality. GaitGL (Lin et al., 2021) adopts the concept of split convolutions from GaitPart and applies it to 3D convolutions, along with a simplified version of HPP. Castro et al. (2024) improves the HPP (Chao et al., 2019), with an Attention HPP that learns to leverage each horizontal split with a different weight, and applies it to optical flow image sequences. Recently, DANet (Ma et al., 2023) has improved the analysis of motion patterns from local regions using a dynamic attention mechanism. While silhouettes effectively convey rich descriptive features of human body shapes and movement patterns throughout the gait cycle, they are sensitive to various factors, such as carrying objects, clothing, lighting conditions, and dynamic backgrounds.

Other studies have explored alternative sensors (Zhao & Zhou, 2017), like floor sensors (Nakajima et al., 2000), accelerometers (Delgado-Escaño et al., 2019, 2020), and wave sensors (Meng et al., 2019). There are also efforts to utilize different visual modalities, including RGB images (Zhang et al., 2019), optical flow (Castro et al., 2024; Castro, Marín-Jiménez, Guil, López-Tapia et al., 2017), depth maps (Castro et al., 2020), or combinations of these modalities (Marín-Jiménez et al., 2021). A noteworthy approach proposed in Delgado-Escaño et al. (2021) employs a teacher–student methodology, in which a student model is trained to mimic the behavior of a teacher model using grayscale images so that optical flow computation becomes unnecessary. This allows the student model to extract visual and motion features from a single input.

A logical progression from these singular modality approaches is to employ multiple modalities simultaneously. This approach leverages data obtained from diverse sources or sensors. For example, Kumar et al. (2018) utilize data from multiple inertial sensors to create a 3D skeleton representation complemented by video images. Castro et al. (2020) design a CNN model that jointly uses optical flow, depth, and grayscale images to enhance the model's overall accuracy. UGaitNet (Marín-Jiménez et al., 2021) proposes a multimodal framework robust against missing data, enabling the model to maintain performance even when an input modality is absent due to external factors.

Alternatively to visual approaches, taking advantages of the robustness of the pose to the gait covariates encountered in real-world scenarios, many approaches adopt the pose as point to create models invariant to variations in shape and viewpoints (Shen et al., 2022). Liao et al. (2017) extracted 2D joints from the human body and fed them into an LSTM and CNN model for gait recognition. In An et al. (2018), the authors improved upon the previous proposal by extracting 3D joints instead of 2D joints. In Liao et al. (2020), Liao et al. computed multiple temporal–spatial features from joint positions, joint angles, motion, and limb length based on a 3D human pose model. Teepe et al. (2021) extracted 2D joints and utilized a Graph Convolutional-derived model to further exploit the spatial information derived from joint positions and limb adjacency. However, the performance of their model remained low. Finally, Liao et al. (2022) extracted features from both pose heatmaps and skeleton graph images, which were constructed by coloring the pose joints and the limbs connecting them.

Despite the advantages of pose-based models over visual-based approaches, pose models generally exhibit lower performance compared to silhouette-based models. To address this issue, other proposals have combined pose information with shape-based descriptors. Li et al. (2022) propose a multimodal approach that integrates pose heatmaps with silhouettes. Their model incorporates Transformer blocks to mitigate the limitations associated with pose data. Cui and Kang (2023) computes a combined descriptor from features extracted from the silhouette and the pose at the spatial level, and secondly, they fuse it across the temporal dimension. Peng et al. (2023) introduced a multi-scale gait graph network to extract features from the pose skeleton and aggregate them with silhouette features obtained from GaitPart (Fan et al., 2020) or GaitGL (Lin et al., 2021). In Li et al. (2020), the authors used a 3D pose model inspired by the Human Mesh Recovery Model (HMR) (Kanazawa et al., 2018) in combination with silhouettes, which were fed into an ensemble of CNN and LSTM models. They also proposed end-to-end training using raw RGB frames, with the HMR included to fine-tune the poses and silhouettes derived from the model on the target dataset. BigGait (Ye et al., 2024) aims to teach the model to generate the optimal image descriptor from the original RGB. Specifically, DINOv2 (Oquab et al., 2023) is used to extract inner features from the original RGB frames, and the model learns to denoise these features with the silhouette maps. Fan et al. (2024) constructed a heatmap skeleton map based on a Gaussian approximation of the pose and combined it with the silhouette. Another approach was presented in Schwarz et al. (2022), where the authors used the average of DensePose UV mapping images (Güler et al., 2018) as input along a temporal sequence of frames.

In this work, we propose a novel hierarchical representation of gait based on body pose, which complements the overall description of body motion with specific motion patterns for each individual limb. This approach can be applied to any pose descriptor to enhance gait recognition. We evaluate our novel representation using two gait descriptors: heatmaps and DensePose. Our experimental results demonstrate the advantages of our proposed pose representation approach across various state-of-the-art models for gait recognition, clearly showcasing the improvements achieved with our method.

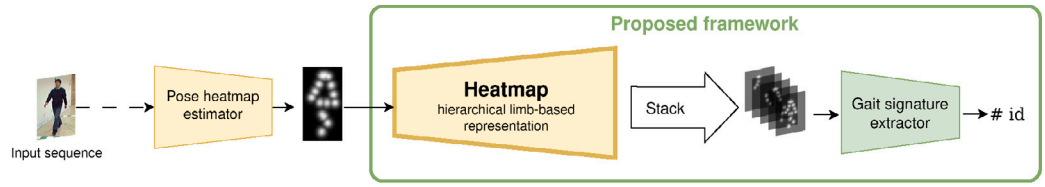
3. Pose representations & gait architecture

In this work, we explore human pose representations that differ from the traditional approaches commonly used in previous pose-based research. Specifically, we propose a hierarchical limb-based representation that considers various splits of body parts. Our novel representation is implemented in this study based on two different pose descriptors: (i) features based on heatmaps produced by a keypoint-based pose extraction model (see Fig. 1c) and (ii) dense features extracted from DensePose (Güler et al., 2018) (Fig. 1d–g). Additionally, to enhance feature extraction from our pose descriptors, we introduce a new gait recognition model architecture, the *PoseGaitGL* architecture, adapted from existing state-of-the-art models.

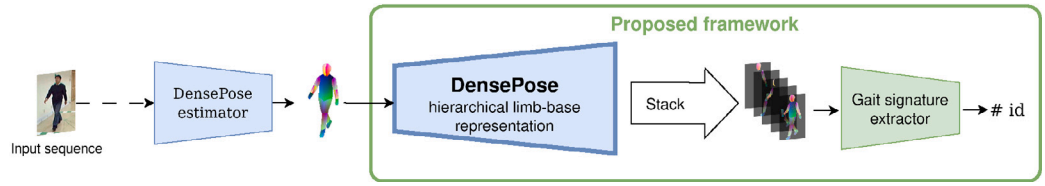
This way, starting from RGB frames, we compute the pose features (heatmaps or DensePose) using an out-of-the-box model for each pose feature. Once these pose features are obtained, we compute the hierarchical pose representation, which is then input into our proposed PoseGaitGL model. This model extracts a gait signature that is used to identify the subjects, as illustrated in Fig. 2. Note that in this figure, two similar processing pipelines are depicted: the pipeline in Fig. 2(a) illustrates the processing using heatmap features, while Fig. 2(b) demonstrates the use of DensePose features. Next, we will focus on the main contributions of the proposed pipelines.

3.1. Hierarchical limb-based representation

To obtain a richer representation of poses, we propose a novel hierarchical limb-based representation composed of three levels of hierarchy: *Human body*, the first level containing the full body representation; *Limbs*, the second level containing four isolated representations of the main body limbs (left and right arms, and left and right legs); and *Middle limbs*, the third level containing eight representations of the sub-limbs from the previous level (the left and right upper-arms

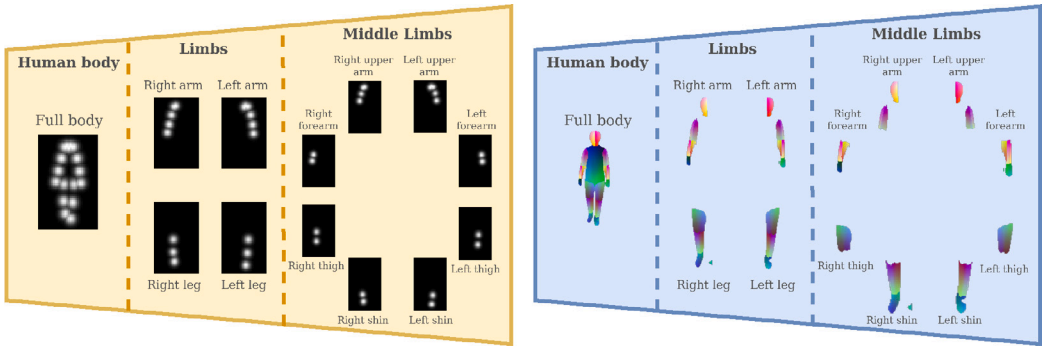


(a) Pipeline using heatmap descriptor.



(b) Pipeline using DensePose descriptor.

Fig. 2. General pipeline of our approach. It is illustrated the pipeline of our approach using either the heatmap or the DensePose descriptors. Note that only one input frame is shown for clarity in the display. (Best viewed in digital format).



(a) Representation for heatmaps.

(b) Representation for DensePose.

Fig. 3. Hierarchical limb-based representation for pose. Each pose heatmap 3(a) or DensePose pose 3(b) frame is rearranged into a 3-level hierarchical limb-based representation together with the full body.

and forearms, and the left and right thighs and shins). This structure allows the model to correlate the movements of different limbs during the walking patterns of subjects, extracting distinct key features from each limb, while the full-body level ensures maintaining a global vision of the human body's movements.

In this work, we explore the application of our pose representation to two robust pose features that provide richer information than traditional pose skeleton key points: heatmaps and DensePose. However, it is important to note that our representation can be applied to any pose descriptor or any other descriptor that supports a limb-based or part-based segmentation of the human body.

3.1.1. *Heatmap hierarchical limb-based representation*

A *heatmap* is a feature map generated by a keypoint-based pose extractor network before calculating the output coordinates for each body joint. These maps comprise one channel for each body joint and represent the probability distribution of the corresponding joint location. A higher value indicates greater confidence from the network in detecting the joint, while a lower value may suggest that the joint is not visible or that the network is less certain about its estimation. This approach provides richer information than simple 2D or 3D coordinates, as each joint is represented as a probability distribution across a region of the image. This allows the model to disregard minor variations in the positions of body joints caused by noise or occlusions. Conversely, joints that are occluded will have lower probabilities, enabling the model to focus less on these less confident joints and effectively handle occlusions. Fig. 3(a) illustrates an example of our heatmap hierarchical representation composed of the full-body representation (first hierarchical level), the four images representing different body limbs (second hierarchical level), and the eight images representing the middle limbs (third hierarchical level).

3.1.2. *DensePose hierarchical limb-based representation*

The DensePose feature is derived from the DensePose estimator model (Güler et al., 2018), which creates a dense 3D surface-based representation of the human body using three images: *(i)* a body Part Segmentation Image (I), that segments the human body into 24 different parts, each colored in varying shades of gray. A texture planar gradient is applied to each body part, indicating the horizontal and vertical relative coordinates of each point with respect to the origin of that body part. *(ii)* a mapping image U indicating the horizontal gradient coordinates of this ; and *(iii)* a mapping image U indicating the vertical gradient coordinates. Several examples of this representation can be seen in Fig. 1(d–g), which showcases the three images, I, U, and V, along with an RGB representation of this 3D mesh. The RGB image is created by stacking the three images as channels to provide a colored representation.

For our representation, we use the DensePose model to extract the V, and I images from each video frame. We have decided to discard the U image from the computed descriptors, as it does not provide any improvement. We then rearrange each of these images into our hierarchical limb-based representation based on the body part segmentation from image I, as depicted in Fig. 3(b).

3.2. *PoseGaitGL architecture*

We propose a model based on GaitGL (Lin et al., 2021), which achieves top results for silhouette data. Nevertheless, our hierarchical

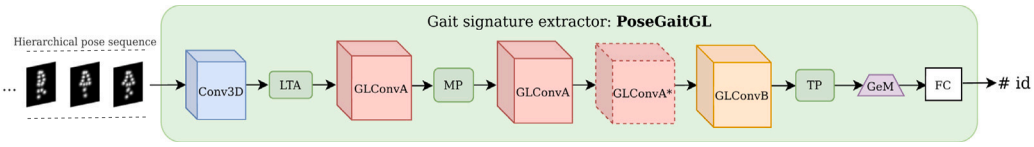


Fig. 4. PoseGaitGL: Gait signature extractor. Our signature extractor accepts as input a sequence of hierarchical pose heatmaps or DensePose. The abbreviations used in our model are as follows: LTA stands for Local Temporal Aggregation, MP indicates Max Pooling, TP refers to Temporal Pooling, GeM represents Generalized-Mean pooling, and FC means Fully Connected layer. Additionally, the GLConvA block has been duplicated as GLConvA* to enhance the model's learning capability. More details can be found in the main text. Best viewed in digital format.

representation can be used with any visual modality model, including optical flow or silhouettes. By leveraging a well-known gait recognition model, we can aim to demonstrate the improvements offered by our hierarchical limb-based pose representation. To adapt GaitGL (Lin et al., 2021) to our pose representation, we implement some modifications to the model architecture. Since our representation has more channels than silhouettes, which only have a single channel, we resize the first 'Conv3D' block to accommodate the new input shape. Additionally, to enhance the model's learning capacity due to the increased number of input channels, we add extra trainable blocks to the architecture. To prevent drastic changes that could lead to unpredictable behavior, we duplicate the 'GLConvA1' block. We refer to our newly designed models as PoseGaitGL-HM for the *pose heatmaps* modality and PoseGaitGL-DP for the *DensePose* modality. A diagram of our final model is presented in Fig. 4, where the newly added block is illustrated with a dotted block.

4. Experiments and results

In this section, we present the experimental results of our approach. We start by introducing the datasets (see Section 4.1) and provide some implementation details (see Section 4.2). Next, we compare our hierarchical limb-based representation with other state-of-the-art approaches (see Section 4.3) and conduct a thorough ablation study (see Section 4.4). Following that, we present inference metrics for the model (see Section 4.5) and introduce a visual analysis of the quality of the transformed features generated by the models (see Section 4.6). Finally, we discuss the strengths and weaknesses of the pose estimation models (see Section 4.7).

4.1. Datasets

We conduct our experimental study using the CASIA-B dataset (Yu et al., 2006) and the 'TUM Gait from Audio, Image and Depth' (TUM-GAID) dataset (Hofmann et al., 2014).

In the CASIA-B dataset, 124 subjects walk in an indoor environment while being recorded from 11 different viewpoints (from 0° to 180°, in steps of 18°). The video resolution is 320 × 240 pixels with a frame rate of 24 fps. Three walking conditions are considered: normal walking (*NM*), carrying a bag (*BG*), and wearing a coat (*CL*). For our experiments, we follow the *Large-Sample Training* (LT) experimental protocol defined in Wu et al. (2017) for cross-view gait recognition. The sequences from the first 74 subjects across all walking conditions and viewpoints are used for training. For the remaining subjects, the first 4 *NM* sequences are used as a gallery set, while the rest of the walking conditions and types serve as the probe set.

In the TUM-GAID dataset, 305 subjects perform two walking trajectories in an indoor environment. The sequences are captured from a single viewpoint (90°) using a Microsoft Kinect sensor, with a resolution of 640 × 480 pixels at 30 fps. Three scenarios are considered: normal walking (N), carrying a backpack (B), and wearing shoes (S). Additionally, for 32 of the 305 subjects, another case is recorded, referred to as elapsed time (TN-TB-TS), where subjects wear spring-like clothing. In this study, we adhere to the experimental setup defined in Marín-Jiménez et al. (2021). Consequently, the training set consists of 150 subjects with non-elapsed-time sequences (N, B, S), while the test set includes the remaining 155 subjects. Two test setups are examined: one for non-elapsed-time sequences (N, B, and S) and another for elapsed-time sequences (TS, TB, and TS).

It is important to note that other popular datasets, such as OU-MVLP (Takemura et al., 2018) and GREW (Zhu et al., 2021), have not released their original RGB video sequences. As a result, the human pose estimators used in this study cannot be applied. While GREW does provide precomputed body joint coordinates, the analysis would be incomplete since DensePose could not be utilized, and the uncertainty of the provided skeleton coordinates must be set to a constant value.

Table 1
Training hyperparameters. Description of the hyperparameters used to train our PoseGaitGL model.

Hyperparameter description	
# iterations	$80k$
Batch size (P subjects x K samples)	P: 8, K: 8
Optimizer	lr: 10^{-4} (10^{-5} after iter. $70k$)
Regularization	L2 (Weight decay: $5 \cdot 10^{-4}$)
# of filters per conv. block	32, 64, 128, 128, 128
GeM pooling	p initial value: 6.5
Triplet loss margin	0.2
Cross-entropy - Label smooth	0.1

4.2. Implementation details

Our approach is based on the GaitGL model (Lin et al., 2021), and we follow the implementation setup recommended by its authors. We scale and crop our input data to ensure that the subject is always centered in the frame, resulting in an input shape of 64 × 44. For experiments that utilize pose heatmaps, we extract heatmaps using ViTPose (Xu et al., 2022). We also conduct ablation experiments with other pose estimation methods, including AlphaPose (Fang et al., 2017) and HRNet (Sun et al., 2019), which are discussed in Section 4.4.1.

In the experiments involving DensePose, we use images obtained from DensePose (Güler et al., 2018) as input data, specifically the I-V pose images. It is important to note that image 'I' is represented using 25 gray tones (24 for body parts and 1 for the background). Therefore, we scale its values to cover the full grayscale range of $[0, 255]$.

During training, we use input samples consisting of 30 frames to reduce memory usage. Conversely, at test time, we evaluate the model's accuracy using all available video frames. A summary of the training hyperparameters can be found in Table 1.

In our experiments, all models were developed and evaluated using the OpenGait library (Fan et al., 2022) with PyTorch 2.4. To evaluate system performance, we use the standard Rank-1 (R1) accuracy, which is the percentage of correctly classified videos:

$$R1 = \frac{num.\ \text{correct}\ videos}{num.\ \text{total}\ videos} \tag{1}$$

4.3. Comparison to the state of the art

In Table 2, we compare our results with the state-of-the-art on the CASIA-B dataset. Our models, PoseGaitGL-HM and PoseGaitGL-DP, outperform all pose-based models and show results comparable

Table 2
State-of-the-art comparison on CASIA-B. Comparison with other pose-based and shape-based models. The best mean results are marked in bold.

Data	Model	Walking condition			Mean
		NM	BG	CL	
Pose	PoseGait (Liao et al., 2020)	68.7	44.5	39.9	49.7
	End-to-end Pose LSTM (Li et al., 2020)	66.1	49.3	37.0	50.8
	PoseMapGait (Liao et al., 2022)	79.3	61.1	48.1	62.8
	TransGait (Li et al., 2022) (Pose + STM)	84.5	71.2	54.4	70.0
	GaitGraph (Teepe et al., 2021)	87.7	74.8	66.3	76.3
	End-to-end Pose CNN (Li et al., 2020)	91.2	83.9	60.2	78.4
	MSGG (Peng et al., 2023)	93.0	78.1	68.3	79.8
	GaitPart-HM (ours)	92.7	81.2	69.3	81.1
	PoseGaitGL-HM (ours)	93.3	81.8	71.5	**82.2**
Shape	GaitSet (Chao et al., 2019)	95.0	87.2	70.4	84.2
	End-to-end shape model (Li et al., 2020)	97.5	90.6	75.1	87.7
	GaitPart (Fan et al., 2020)	96.2	91.5	78.7	88.8
	End-to-end ensemble (Li et al., 2020)	97.9	93.1	77.6	89.5
	TransGait (Li et al., 2022) (Sil + STM)	97.3	92.8	80.6	90.2
	GaitGL (Lin et al., 2021)	97.4	94.5	83.6	91.8
	TransGait (Li et al., 2022) (*Multimodal*)	98.1	94.9	85.8	92.9
	DANet (Ma et al., 2023)	98.0	95.9	89.9	94.6
	AttenGait (Castro et al., 2024)	98.8	97.7	91.0	95.8
	MMGaitFormer (Cui & Kang, 2023) (*Multimodal*)	98.4	96.0	94.8	96.4
	BiFusion (Peng et al., 2023) (*Multimodal*)	98.6	97.0	94.0	**96.5**
	GaitPart-DP (ours)	96.9	90.0	81.1	89.3
	PoseGaitGL-DP (ours)	97.3	92.8	86.0	92.0

Table 3
State-of-the-art comparison on TUM-GAID. Mean Rank-1 accuracy (%) on TUM-GAID under all walking conditions. The best mean results are marked in bold, while the second best mean results are marked in italics.

Model	Non-elapsed-time			Elapsed-time			Mean
	N	*B*	*S*	*TN*	*TB*	*TS*	
CNN-NN128 (Castro, Marín-Jiménez, Guil, de la Blanca, 2017)	99.7	98.1	95.8	62.5	56.3	59.4	93.9
3D-CNN-7NN-All (Castro et al., 2020)	100	99.4	99.4	75.0	62.5	62.5	96.5
Student-Conv3D-B (Delgado-Escaño et al., 2021)	99.4	97.4	96.4	100	100	100	97.0
PoseGaitGL-HM (ours)	99.7	99.0	99.0	100	100	100	*99.3*
PoseGaitGL-DP (ours)	99.7	100	99.7	100	100	100	**99.8**

to shape-based models. Focusing on the rows for 'Pose-based models', our PoseGaitGL-HM model demonstrates an improvement of 3.8% over the method by Li et al. (2020) and a 2.4% improvement over the MSGG network from Peng et al. (2023). The first method is end-to-end trained using RGB images, which include appearance information, to derive more refined features but allowing the model to learn from that appearance information implicitly. In contrast, our model achieves superior results by relying solely on pose information without any data regarding human body shape. The latter mentioned method proposes a multi-scale graph-based network that leverages information from skeleton data; however, its pose descriptor and model pipeline do not perform as well as our approach. Focusing on the rows related to shape-based models, our proposed PoseGaitGL-DP model outperforms most silhouette-based approaches, with the exception of DANet (Ma et al., 2023). Additionally, it achieves results comparable to those of complex visual descriptor approaches (Castro et al., 2024). In our opinion, these outcomes are very promising, as we achieve results that are closely aligned with the top shape-based approaches, only using a pose representation modality that focuses on describing the spatial distribution of body parts while minimizing the influence of the overall body shape. We would like to note that multimodal approaches (Cui & Kang, 2023; Li et al., 2022; Peng et al., 2023) have been included in this comparison for information purposes as a comparison with methods using a single modality, as ours, would be unfair. We also evaluated the performance of our hierarchical representation based on GaitPart (Fan et al., 2020): GaitPart-HM for the model using heatmaps, and GaitPart-DP for the model using DensePose frames. In both cases, GaitPart, with our hierarchical representation, underperforms compared to our PoseGaitGL model for both pose descriptors.

We also conducted an evaluation of our proposed approach using the TUM-GAID dataset (see Table 3). Both models, based on their respective representations, demonstrate a significant improvement compared to previous state-of-the-art models. Specifically, PoseGaitGL-HM achieves perfect accuracy in the elapsed-time experiment and a very high accuracy in the non-elapsed time test. Similarly, PoseGaitGL-DP reaches perfect accuracy in the elapsed-time test and almost perfect performance in the non-elapsed time test, scoring 99.8%.

4.4. Ablation study

In this section, we provide a comprehensive ablation analysis of our hierarchical representation and the proposed model, primarily focusing on the CASIA-B benchmark, which presents a greater and more significative benchmark. We conduct four experiments: the first three utilize heatmap-based features, while the fourth employs DensePose features. Specifically, we explore alternative models for estimating pose heatmaps and compare their performance with ViTPose. Additionally, we evaluate different body part representations that differ from the hierarchical limb-based representation and examine various modifications of the chosen model. Lastly, we analyze how different grouping strategies for the images affect the DensePose features.

4.4.1. Heatmap extraction models

In our experiments, we extracted pose heatmaps using the ViTPose model (Xu et al., 2022), which currently achieves the highest performance in the MS COCO (Lin et al., 2014) Keypoint Detection validation benchmark. Additionally, we evaluated our model against other popular pose estimation methods, namely AlphaPose (Fang et al., 2017) and HRNet (Sun et al., 2019). Table 4 presents the performance

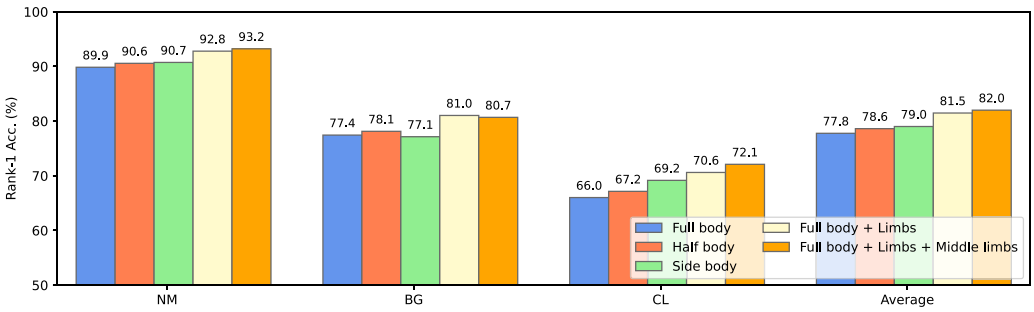


Fig. 5. Heatmap representations comparison. Mean Rank-1 accuracy (%) on CASIA-B under all walking conditions for all the studied pose heatmap representations, excluding identical-view case. Note that accuracy range displayed is cropped to 50-100%. (Best viewed in digital format).

Table 4
Comparison of three pose estimation models on CASIA-B. Empirical comparison of the recognition performance of the PoseGaitGL-HM model for the heatmaps extracted with ViTPose, AlphaPose, and HRNet.

Pose model	Walking condition			Mean
	NM	BG	CL	
ViTPose	93.3	81.8	71.5	**82.2**
AlphaPose	91.4	76.9	67.9	78.7
HRNet	92.2	78.8	70.4	80.5

of PoseGaitGL-HM using ViTPose, AlphaPose, and HRNet on the CASIA-B dataset. The results show that AlphaPose and HRNet achieved mean accuracies that are lower by over 3.5% and 1.7%, respectively, compared to the accuracy obtained by ViTPose. This indicates that ViTPose provides superior pose estimation for gait recognition models.

4.4.2. Body part representations

Apart from the hierarchical limb-based representation described in Section 3.1, we explored additional body part representations. Specifically, we considered two different paradigms for splitting the human body: a body-half-based split and a hierarchical limb-based split.

Body-half-based split. This approach divides the human body into two halves. We examine a horizontal split into upper and lower limbs (half body) and a vertical split into left and right sides (side body). In the first case, the movement of the legs and arms is decoupled, while the second split helps eliminate occlusions between body parts on different sides. In both scenarios, the output is an image with two channels, one representing each body half.

Hierarchical limb-based split. This approach, explained in Section 3.1 and illustrated in Figs. 3(a) and 3(b), involves dividing the human body based on its limbs and considering multiple levels of subdivision. In this ablation study, we assess the performance gains achieved by aggregating the different sublevels of this hierarchy into the overall representation. Specifically, we evaluate the following representations: the 'Full Body' representation (which includes only the first level of hierarchy), 'Full Body + Limbs' (the first and second levels), and 'Full Body + Limbs + Middle limbs' (which includes the third level of hierarchy).

The empirical comparison of these representations is illustrated in Fig. 5. In these experiments, we utilized the original GaitGL architecture (Lin et al., 2021) to evaluate the impact of different representations purely. Focusing on the average accuracy, both 'body-half-based' splits show improvements over the 'Full-body' representation, indicating a better understanding of joint movement by decoupling limbs into distinct groups. Generally, the 'Side-body' representation outperforms the 'Half-body' representation. This suggests that occluded joints in the right arm and leg, particularly when viewed from a 90° angle, significantly affect the model's performance. The use of the 'Side-body' representation mitigates this issue. Furthermore, the hierarchical limb-based splits outperform both the 'Body-half-based' splits and the 'Full-body' representation. The 'Full body + Limbs + Middle limbs' representation, featuring three levels of hierarchy, achieves the highest mean accuracy among all tested representations, improving upon the 'Full-body' case by 4.2%. Overall, it is evident that mean accuracy increases with each additional level of hierarchy— 'Full body + Limbs' improves upon 'Full body', and 'Full body + Limbs + Middle limbs' enhances performance over 'Full body + Limbs'. However, the performance gain from adding the third level of hierarchy, 'Middle limbs', is comparatively less significant.

4.4.3. Model ablation

Since the original GaitGL (Lin et al., 2021) architecture was designed for silhouettes, we evaluate various modifications to adapt it to our optimal 'Hierarchical limb-based' representation.

Architecture size. Our representation consists of three hierarchical levels composed of several channels representing each level (*i.e.*eight for middle limbs, plus four for the limbs, and one for the full body), compared to the single channel used for silhouettes. Therefore, we investigate whether the model is sufficiently complex to process the new information represented in these additional hierarchical levels. To avoid drastic changes to the architecture, we decided to duplicate the core blocks of the model to enhance its learning capability. Specifically, we experimented with the following blocks: 'initial Conv3D', 'GLConvA0', 'GLConvA1', and 'GLConvB2'.

Multi-Branch Architecture. Given that our input data consists of various types of information, we explore a multi-branch architecture where each channel is processed through a separate branch. This approach enables the model to learn different filters for each type of data. In our case, the branches include the 'initial Conv3D' and 'LTA' blocks, which are then combined using an aggregation operation. We examine two distinct operations for this combination: averaging and taking the maximum.

Fig. 6 presents an experimental comparison of the evaluated modifications to the architecture using our preferred representation ('Hierarchical limb-based') as input. Across all scenarios, the model with the duplicated 'GLConvA1' layer, detailed in Fig. 4, consistently outperforms other modifications. When comparing the average accuracy of our modified architecture to the original GaitGL, we observe an improvement of 1.1% in accuracy for the BG scenario. When focusing on the models utilizing parallel branches (represented by the last two bars in Fig. 6), it is evident that these configurations do not effectively enhance the original architecture, regardless of the aggregation function used.

4.4.4. DensePose representation

The following set of experiments focuses on the use of DensePose as the input data (refer to Section 3.1.2). As explained in Section 3.1.2, DensePose provides three different representations of the subject. In this study, we examine the benefits of each representation and their various combinations in relation to the model's final accuracy.

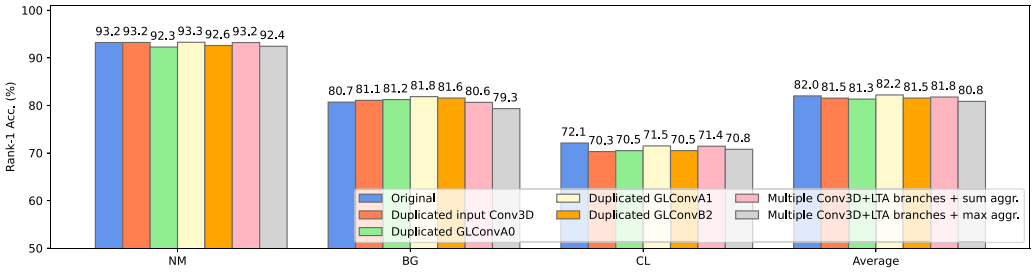


Fig. 6. Model ablation. Mean Rank-1 accuracy (%) on CASIA-B under all walking conditions, excluding identical-view cases, for the model architecture variations studied in the ablation on pose heatmaps. (Best viewed in digital format).

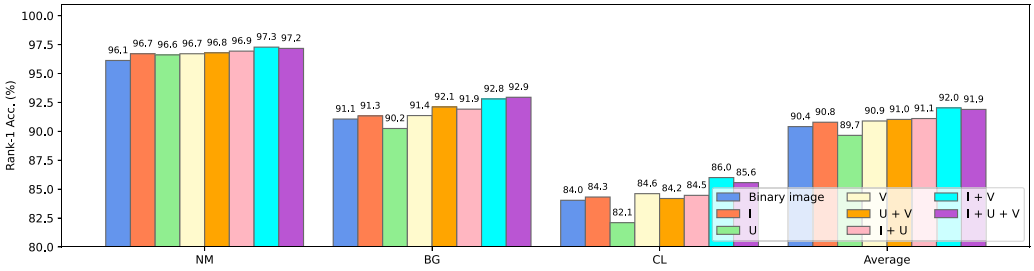


Fig. 7. DensePose representation: I, U, and V channels ablation. Mean Rank-1 accuracy (%) on CASIA-B, under all walking conditions excluding identical-view case. Note that the displayed accuracy range is cropped to 80–100% (Best viewed in digital format).

Fig. 7 summarizes the experiments evaluating all combinations of images: I, U, and V provided by DensePose. Additionally, we include a case using binary silhouettes obtained from binarized DensePose maps, which is equivalent to not utilizing any pose information. While the I and V images demonstrate an improvement compared to the binary silhouette case, the U image performs worse than the binary silhouette. Combining the U image with the I and V images (*i.e.* $I+U$ and $U+V$) does not yield better results than the performances achieved with the individual I and V images. In contrast, the combination of I and V results in a notable improvement of 1.6% over the binary silhouette, whereas adding the U image leads to a decrease in performance.

4.5. Model complexity

In this section, we evaluate the complexity of the two proposed models by analyzing the number of parameters each requires, their memory usage, and their computational complexity. Additionally, we evaluate the inference time per sample.

The PoseGaitGL-HM model consists of 3.99 million parameters, requires 19.72 MB of memory, and operates at 20.67 GFLOPs. In contrast, the PoseGaitGL-DP model has 4.00 million parameters, occupies 16.01 MB of memory, and performs at 20.67 GFLOPs.

For inference time per sample, we conducted experiments on a computer equipped with 32 cores running at 2 GHz, 256 GB of RAM, and an NVIDIA Titan Xp GPU. The recorded inference times were 20.25 ms for PoseGaitGL-HM and 21.22 ms for PoseGaitGL-DP.

4.6. Visual analysis of gait feature transformation

We further evaluate the robustness of the gait signatures extracted by our proposed models using UMAP (see Fig. 8 for projection). In this analysis, we visualize the gait signatures for the first 10 subjects from the CASIA-B test set, specifically subjects #075 to #084. These features are derived from the PoseGaitGL-HM and PoseGaitGL-DP models.

As shown in the figure, both models manage to effectively cluster the gait signatures of the same users and separate them from the signatures of other subjects. This effectiveness holds regardless of the walking condition or the view of the input sequence.

4.7. Discussion on the use of the pose

In contrast to silhouettes that are influenced by changes in body shape—such as those caused by clothing or carried objects—body pose algorithms are currently better able to resist these distractions, as shown in Fig. 9.

However, while one of the main limitations of silhouette-based models is the computation of the silhouettes themselves, the primary limitation of pose-based models lies in the estimation of body pose. Fig. 10 illustrates instances where pose estimation algorithms have produced incorrect results. In these cases, low light conditions and clothing with low contrast and dark colors pose significant challenges for the methods being evaluated.

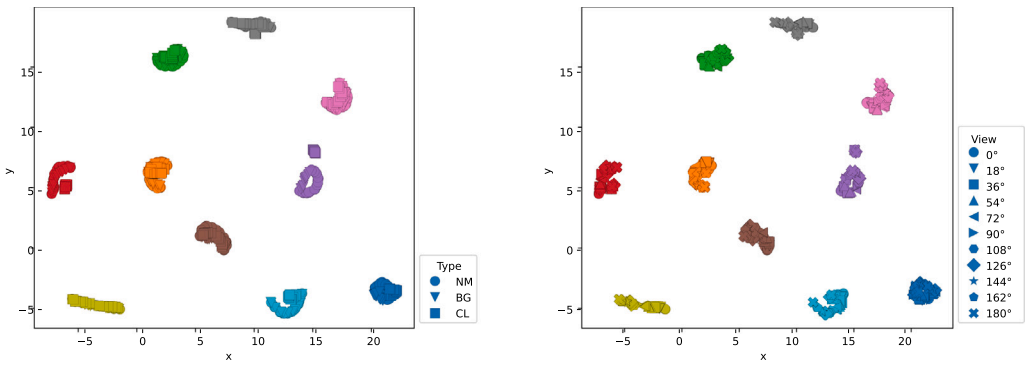
5. Conclusion

This paper presents an experimental study on a novel hierarchical pose-based representation for gait recognition. This representation consists of three hierarchical levels: full body, body limbs, and middle limbs, allowing for better exploitation of key information derived from body poses. We implemented this hierarchical structure using two different pose descriptors: body pose heatmaps and DensePoses.

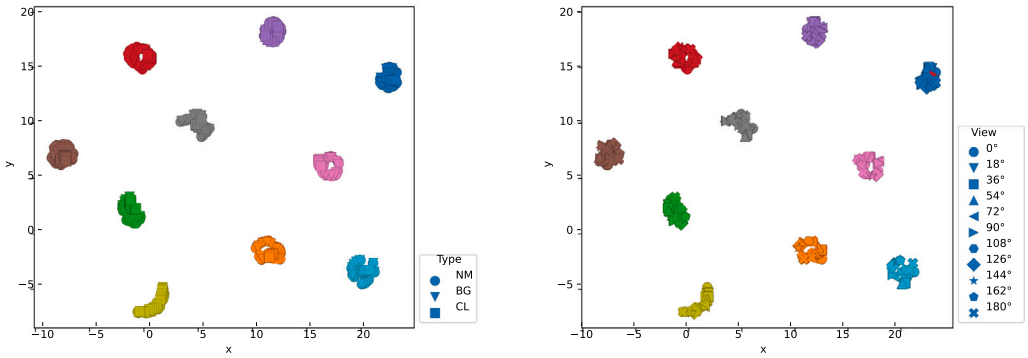
In our study, we employed a model derived from GaitGL; however, our representation is not limited to this model and can be utilized by any image-based gait recognition model.

The experimental results from the CASIA-B and TUM-GAID datasets demonstrate that, compared to existing pose-based methods: (a) our hierarchical pose-based representation improves upon traditional pose descriptors, and (b) the DensePose representation provides richer information than the sparse representation obtained solely from body heatmaps or just silhouettes.

For future work, we plan to further explore more sophisticated pose representations derived from the heatmaps or DensePoses investigated in this study. Additionally, we aim to investigate other paradigms of human pose representation, such as 3D skeletons or body meshes, which could enhance existing methods. Furthermore, this research was



(a) PoseGaitGL-HM.



(b) PoseGaitGL-DP.

Fig. 8. UMAP projection of gait signatures for subjects #075-#084. Each mark represents a signature in the UMAP space, with colors denoting the same subject. The left figures mark the different types of walking condition using symbols (circles, triangles, squares), while the right figures mark the associated views for each gait signature (Best viewed in digital format).

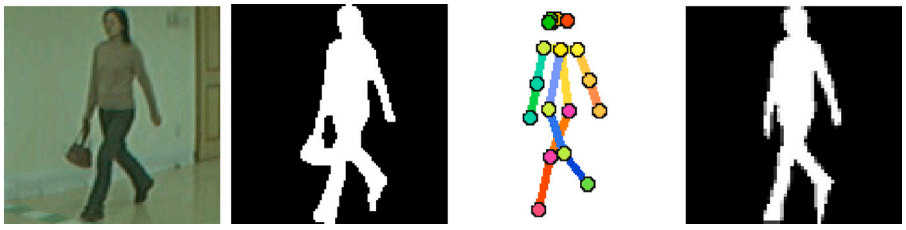


Fig. 9. Person segmentation. Pose estimation models are very robust to carrying objects. From left to right: original RGB image; silhouette included in the original dataset; pose skeleton estimated by ViTPose; and, binary silhouette generated with DensePose.

Fig. 10. Pose estimation errors committed by the ViTPose and DensePose models. The second image illustrates the pose skeleton estimated by ViTPose for the subject depicted in the first image. The fourth image presents the inaccurately segmented image I (*i.e.* missing an arm and featuring a smaller head) as estimated by DensePose for the subject shown in the third image.

limited to datasets providing RGB images for computing our pose representations. These datasets represent a constrained scenario recorded in indoor environments, which do not capture the complexities of real-world situations. In the future, we intend to extend our research towards more realistic scenarios and adapt our models and pose representations to accommodate the input data provided by those datasets. All gait recognition methods that rely on descriptors derived from RGB data (such as pose descriptors, silhouettes, and optical flow) require prior person detection and cropping to effectively compute these descriptors. Generally, standard person detection approaches achieve higher accuracy in most environments, although they can produce errors in noisy or cluttered scenarios. Therefore, it is essential to first evaluate the performance of the pose detection algorithms in the target environment to ensure high-quality gait recognition results.

CRediT authorship contribution statement

Nicolás Cubero: Methodology, Software, Validation, Investigation, Writing – original draft, Writing – review & editing. **Francisco M. Castro:** Methodology, Validation, Investigation, Writing – original draft, Writing – review & editing. **Julián R. Cózar:** Methodology, Writing – original draft, Writing – review & editing. **Nicolás Guil:** Conceptualization, Methodology, Validation, Writing – original draft, Writing – review & editing. **Manuel J. Marín-Jiménez:** Conceptualization, Methodology, Validation, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work has been supported by the Junta de Andalucía of Spain (P20_00430, including European Union funds), the Ministry of Science and Innovation of Spain (PID2019-105396RB-I00 and TED2021-129151B-I00), and the grant "University of Malaga PhD Scholarship Program".

Data availability

The authors do not have permission to share data.

References

An, W., Liao, R., Yu, S., Huang, Y., & Yuen, P. C. (2018). Improving gait recognition with 3D pose estimation. In *Chinese conference on biometric recognition* (pp. 137–147). Springer.

Bouchrika, I., Goffredo, M., Carter, J. N., & Nixon, M. S. (2011). On using gait in forensic biometrics. *Journal of Forensic Sciences*, *56*, URL: https://api.semanticscholar.org/CorpusID:14357171.

Castro, F. M., Delgado-Escaño, R., Hernández-García, R., Marín-Jiménez, M. J., & Guil, N. (2024). AttenGait: Gait recognition with attention and rich modalities. *Pattern Recognition*, Article 110171.

Castro, F. M., Marín-Jiménez, M. J., Guil, N., & de la Blanca, N. P. (2017). Automatic learning of gait signatures for people identification. *vol. 10306*, In *IWANN* (pp. 257–270).

Castro, F. M., Marín-Jiménez, M. J., Guil, N., & de la Blanca, N. P. (2020). Multimodal feature fusion for CNN-based gait recognition: an empirical comparison. *Neural Computing and Applications*, 1–21.

Castro, F. M., Marín-Jiménez, M. J., Guil, N., López-Tapia, S., & de la Blanca, N. P. (2017). Evaluation of CNN architectures for gait recognition based on optical flow maps. In *BIOSIG* (pp. 251–258).

Chao, H., He, Y., Zhang, J., & Feng, J. (2019). GaitSet: Regarding gait as a set for cross-view gait recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*, 8126–8133. http://dx.doi.org/10.1609/aaai.v33i01.33018126.

Cui, Y., & Kang, Y. (2023). Multi-modal gait recognition via effective spatial-temporal feature fusion. In *CVPR* (pp. 17949–17957).

Delgado-Escaño, R., Castro, F. M., Cózar, J. R., Marín-Jiménez, M. J., & Guil, N. (2019). An end-to-end multi-task and fusion CNN for inertial-based gait recognition. *IEEE Access*, *7*, 1897–1908.

Delgado-Escaño, R., Castro, F. M., Cózar, J. R., Marín-Jiménez, M. J., Guil, N., & Casilari, E. (2020). A cross-dataset deep learning-based classifier for people fall detection and identification. *Computer Methods and Programs in Biomedicine*, *184*, Article 105265.

Delgado-Escaño, R., Castro, F. M., Guil, N., & Marín-Jiménez, M. J. (2021). GaitCopy: Disentangling appearance for gait recognition by signature copy. *IEEE Access*, *9*, 164339–164347.

Fan, C., Ma, J., Jin, D., Shen, C., & Yu, S. (2024). SkeletonGait: Gait recognition using skeleton maps. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 1662–1669).

Fan, C., Peng, Y., Cao, C., Liu, X., Hou, S., Chi, J., Huang, Y., Li, Q., & He, Z. (2020). GaitPart: Temporal part-based model for gait recognition. In *CVPR* (pp. 14225–14233).

Fan, C., Shen, C., & Liang, J. (2022). OpenGait. URL: https://github.com/ShiqiYu/OpenGait.

Fang, H.-S., Xie, S., Tai, Y.-W., & Lu, C. (2017). RMPE: Regional multi-person pose estimation. In *ICCV*.

Feng, Y., Li, Y., & Luo, J. (2016). Learning effective gait features using LSTM. In *Proc. ICPR* (pp. 325–330). IEEE.

Güler, R. A., Neverova, N., & Kokkinos, I. (2018). DensePose: Dense human pose estimation in the wild. In *CVPR*.

Han, J., & Bhanu, B. (2005). Individual recognition using gait energy image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(2), 316–322.

Haussler, A. M., Tueth, L. E., May, D. S., Earhart, G. M., & Mazzoni, P. (2024). Refinement of an algorithm to detect and predict freezing of gait in parkinson disease using wearable sensors. *Sensors*, *25*(1), 124.

Hofmann, M., Geiger, J., Bachmann, S., Schuller, B., & Rigoll, G. (2014). The TUM Gait from Audio, Image and Depth (GAID) database: Multimodal recognition of subjects and traits. *Journal of Visual Communication and Image Representation*, *25*(1), 195–206, Visual Understanding and Applications with RGB-D Cameras.

Hou, S., Cao, C., Liu, X., & Huang, Y. (2020). Gait lateral network: Learning discriminative and compact representations for gait recognition. In *European conference on computer vision* (pp. 382–398). Springer.

Kanazawa, A., Black, M. J., Jacobs, D. W., & Malik, J. (2018). End-to-end recovery of human shape and pose. In *Computer vision and pattern recognition*.

Kumar, P., Mukherjee, S., Saini, R., Kaushik, P., Roy, P. P., & Dogra, D. P. (2018). Multimodal gait recognition with inertial sensor data and video using evolutionary algorithm. *IEEE Transactions on Fuzzy Systems*, *27*(5), 956–965.

Li, G., Guo, L., Zhang, R., Qian, J., & Gao, S. (2022). TransGait: Multimodal-based gait recognition with set transformer. *Applied Intelligence: The International Journal of Artificial Intelligence, Neural Networks, and Complex Problem-Solving Technologies*, 1–13.

Li, X., Makihara, Y., Xu, C., Yagi, Y., Yu, S., & Ren, M. (2020). End-to-end model-based gait recognition. In *CVPR*.

Liao, R., Cao, C., Garcia, E. B., Yu, S., & Huang, Y. (2017). Pose-based temporal-spatial network (PTSN) for gait recognition with carrying and clothing variations. In *Chinese conference on biometric recognition* (pp. 474–483). Springer.

Liao, R., Li, Z., Bhattacharyya, S. S., & York, G. (2022). PoseMapGait: A model-based gait recognition method with pose estimation maps and graph convolutional networks. *Neurocomputing*, *501*, 514–528.

Liao, R., Yu, S., An, W., & Huang, Y. (2020). A model-based gait recognition method with body pose and human prior knowledge. *Pattern Recognition*.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. In D. Fleet, T. Pajdla, B. Schiele, & T. Tuytelaars (Eds.), *Computer vision – ECCV 2014* (pp. 740–755). Cham: Springer International Publishing.

Lin, B., Zhang, S., & Yu, X. (2021). Gait recognition via effective global-local feature representation and local temporal aggregation. In *ICCV* (pp. 14648–14656).

Ma, K., Fu, Y., Zheng, D., Cao, C., Hu, X., & Huang, Y. (2023). Dynamic aggregated network for gait recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 22076–22085).

Marín-Jiménez, M. J., Castro, F. M., Delgado-Escaño, R., Kalogeiton, V., & Guil, N. (2021). UGaitNet: Multimodal gait recognition with missing input modalities. *IEEE Transactions on Information Forensics and Security*, *16*, 5452–5462.

Meng, Z., Fu, S., Yan, J., Liang, H., Zhou, A., Zhu, S., Ma, H., Liu, J., & Yang, N. (2019). Gait recognition for co-existing multiple people using millimeter wave sensing. In *Proceedings of the AAAI conference on artificial intelligence*.

Muramatsu, D., Makihara, Y., Iwama, H., Tanoue, T., & Yagi, Y. (2013). Gait verification system for criminal investigation. (pp. 747–748). http://dx.doi.org/10.1109/ACPR.2013.195.

Nakajima, K., Mizukami, Y., Tanaka, K., & Tamura, T. (2000). Footprint-based personal recognition. *IEEE Transactions on Biomedical Engineering*, *47*(11), 1534–1537.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H. V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Howes, R., Huang, P.-Y., Xu, H., Sharma, V., Li, S.-W., Galuba, W., Rabbat, M., Assran, M., Ballas, N., ... Bojanowski, P. (2023). DINOv2: Learning robust visual features without supervision.

Peng, Y., Ma, K., Zhang, Y., & He, Z. (2023). Learning rich features for gait recognition by integrating skeletons and silhouettes. *Multimedia Tools and Applications*, *83*, 1–22. http://dx.doi.org/10.1007/s11042-023-15483-x.

Schwarz, P., Scharinger, J., & Hofer, P. (2022). Gait recognition with DensePose energy images. In G. Rozinaj, & R. Vargic (Eds.), *Systems, signals and image processing* (pp. 65–70).

Sepas-Moghaddam, A., & Etemad, A. (2023). Deep gait recognition: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(1), 264–284. http://dx.doi.org/10.1109/TPAMI.2022.3151865.

Shen, C., Yu, S., Wang, J., Huang, G. Q., & Wang, L. (2022). A comprehensive survey on deep gait recognition: Algorithms, datasets and challenges. URL: https://arxiv.org/abs/2206.13732.

Singh, J., Jain, S., Arora, S., & Singh, D. U. (2018). Vision-based gait recognition: A survey. *IEEE Access*, *PP*, http://dx.doi.org/10.1109/ACCESS.2018.2879896.

Sun, K., Xiao, B., Liu, D., & Wang, J. (2019). Deep high-resolution representation learning for human pose estimation. In *CVPR*.

Takemura, N., Makihara, Y., Muramatsu, D., Echigo, T., & Yagi, Y. (2018). Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition. *IPSJ Transactions on Computer Vision and Applications*, *10*(1), 4.

Teepe, T., Khan, A., Gilg, J., Herzog, F., Hormann, S., & Rigoll, G. (2021). GaitGraph: Graph convolutional network for skeleton-based gait recognition. In *2021 IEEE international conference on image processing*. IEEE, http://dx.doi.org/10.1109/icip42928.2021.9506717.

Wu, Z., Huang, Y., Wang, L., Wang, X., & Tan, T. (2017). A comprehensive study on cross-view gait based human identification with deep CNNs. *IEEE PAMI*, *39*(2), 209–226.

Xu, Y., Zhang, J., Zhang, Q., & Tao, D. (2022). ViTPose: Simple vision transformer baselines for human pose estimation. In *Advances in neural information processing systems*.

Ye, D., Fan, C., Ma, J., Liu, X., & Yu, S. (2024). BigGait: Learning gait representation you want by large vision models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 200–210).

Yu, S., Tan, D., & Tan, T. (2006). A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. *vol. 4*, In *Proc. ICPR* (pp. 441–444).

Zhang, Z., Tran, L., Yin, X., Atoum, Y., Liu, X., Wan, J., & Wang, N. (2019). Gait recognition via disentangled representation learning. In *CVPR* (pp. 4710–4719).

Zhao, Y., & Zhou, S. (2017). Wearable device-based gait recognition using angle embedded gait dynamic images and a convolutional neural network. *Sensors*, *17*(3), 478.

Zhu, Z., Guo, X., Yang, T., Huang, J., Deng, J., Huang, G., Du, D., Lu, J., & Zhou, J. (2021). Gait recognition in the wild: A benchmark. In *ICCV* (pp. 14789–14799).