Multimodal Human Pose feature fusion for Gait recognition *

Nicolás Cubero[1 [0000−0002−4851−9181]], Francisco M. Castro[1 [0000−0002−7340−4976]], Julián R. Cózar[1 [0000−0002−8993−1949]], Nicolás Guil[1 [0000−0003−3431−6516]], and Manuel J. Marín-Jiménez[2 [0000−0001−9294−6714]]

[1] University of Malaga, Bulevar Louis Pasteur. 35, 29071 Malaga, Spain
[2] University of Cordoba, Rabanales campus, 14014 Córdoba, Spain

Abstract. *Gait recognition* allows identifying people at a distance based on the way they walk (i.e. gait) in a non-invasive approach. Most of the approaches published in the last decades are dominated by the use of silhouettes or other appearance-based modalities to describe the Gait cycle. In an attempt to exclude the appearance data, many works have been published that address the use of the human pose as a modality to describe the walking movement. However, as the pose contains less information when used as a single modality, the performance achieved by the models is generally poorer. To overcome such limitations, we propose a multimodal setup that combines multiple pose representation models. To this end, we evaluate multiple fusion strategies to aggregate the features derived from each pose modality at every model stage. Moreover, we introduce a weighted sum with trainable weights that can adaptively learn the optimal balance among pose modalities. Our experimental results show that (a) our fusion strategies can effectively combine different pose modalities by improving their baseline performance; and, (b) by using only human pose, our approach outperforms most of the silhouette-based state-of-the-art approaches. Concretely, we obtain 92.8 % mean Top-1 accuracy in CASIA-B.

Keywords: Gait recognition · human pose · surveillance · biometrics · deep learning · multimodal fusion.

1 Introduction

Gait-based people identification, or simply *gait recognition* aims at recognizing people by their manner of walking. Unlike other biometrical features, such as iris or fingerprints, gait recognition can be performed at a distance without the subject cooperation. Hence, it owns very potential applications in social security or medical research [19], among others, and many works have been published in this area during the last decades.

* Supported by the Junta de Andalucía of Spain (P18-FR-3130, UMA20-FEDERJA-059 and PAIDI P20_00430) and the Ministry of Education of Spain (PID2019-105396RB-I00 and TED2021-129151B-I00). Including European funds, FEDER.

Despite multiple modalities that have been proposed to describe the gait motion, silhouettes are still the most studied modality in literature [3, 5, 14]. Silhouette holds a binary representation of the human body shape. A sequence of silhouettes reflects the body limbs' movements. However, the silhouette also contains information about the human shape and body contours that is unrelated to the motion of the limbs. In this sense, models may be biased by that appearance-based information and their performance could be penalized.

To remove that appearance-based information, many authors propose to use the human pose as an alternative modality. Human pose describes the positions of the body limbs at every instant and removes any other unnecessary shape information, so it is more robust to that appearance-related bias. Typically, pose-based gait recognition approaches exploit the 2D or 3D coordinates of the joints from the human body [11, 1] and extract features from the correlation between the motion of different body joints to predict the identity of the subjects. However, those approaches perform worse than those based on visual descriptors like silhouettes [3, 5, 14]. This is caused by the less information received from the human pose, *i.e.* a set of 2D/3D coordinates versus a typically 64×64 silhouette image. To overcome such limitations, recent approaches [9, 10] propose a multimodal setup that combines the pose information with silhouettes. However, although this multimodal setup reaches a substantial improvement, it brings up again the body shape information, limiting the benefits of the human pose.

In this work, we propose a combination of multiple pose representations in a multimodal setup to overcome the lack of information in every single representation, and hence, to avoid the use of shape descriptors. Moreover, instead of pose coordinates, we use two different representations: *(i)* a set of pose heatmaps images that are extracted from a human pose estimator [23] and; *(ii)* a dense pose representation extracted from DensePose [18] model. These two representations contain richer information than a solely set of coordinates and allow us to build a multimodal model that combines both representations and extracts more valuable features through different fusion strategies.

Therefore, our main contributions are: *(i)* a multimodal setup that exploits information from different pose representation modalities achieving state-of-the-art results on CASIA-B; and, *(ii)* a thorough experimental study comparing different fusion strategies to better combine the information from both pose representations.

The rest of this paper is organized as follows: Sec. 2 presents previous works. Sec. 3 describes our fusion strategies and Sec. 4 contains the experimental results on CASIA-B. Finally, Sec. 5 concludes the work.

2 Related work

Recent gait recognition approaches have been mostly dominated by silhouettes or derived descriptors. GaitSet [3] uses a random stack of silhouettes where each frame is handled independently to extract and combine features through a Horizontal Pyramid Pooling (HPP). GaitPart [5] includes a novel part-based

model that extracts features from horizontal splits of intermediate convolutional activations. GLN [7] introduces concatenation at intermediate convolutional activations and a compression module that is attached at the end of the model to reduce feature dimensionality. GaitGL [14] applies split convolutions within the convolutional pipeline together with a simplified version of the HPP proposed in [3]. As an alternative to silhouettes, other works use descriptors extracted from alternative sensors like accelerometer [4], floor-sensors [17], wave-sensors [16], or visual modalities [2, 15] However, since all those descriptors are based on the human shape, like the silhouette, they are affected by changes in the body shape, illumination, etc.

In an attempt to remove the human shape, many other proposals use human pose descriptors [20] as input modality. Liao *et al.* [11] extract 2D joints from the human body and fed an LSTM and CNN model for gait recognition. In [1], the previous idea was improved by extracting 3D joints instead of 2D joints. In [13], Liao *et al.* compute multiple temporal-spatial features from the joint positions, the joint angles and motion, and limb length from the 3D human pose model. Teepe *et al.* [22] proposes a Graph Convolutional-based model to further exploit the spatial information originated from the 2D joints and their adjacency. Finally, Liao *et al.* [12] extract features from both pose heatmaps and skeleton graph images with a colored joint and limbs skeleton. Although pose-based models have some benefits with respect to visual-based approaches, their performance is lower in comparison with silhouette-based models.

Finally, many works propose multimodal models that exploit pose in combination with other modalities to improve the performance of pose-based models. Li *et al.* [9] propose a multimodal approach that combines pose heatmaps with silhouettes through a set of Transformers blocks that jointly process patches from both modalities. In [10], authors use a 3D pose model inspired by the Human Mesh Recovery Model (HMR), [8] combined with silhouettes that are fed into an ensemble of CNN and LSTM models.

Nevertheless, all these methods, despise proposing genuine strategies for modalities fusion, most of them bring up again the silhouette or another shape-derived, so the models learn the bias inherent to the shape covariates. In contrast, we propose a combination of different pose representation methods, without shape information.

3 Methodology

In this work, we propose multiple fusion strategies to combine and aggregate features extracted from pose (i.e. heatmaps and dense pose images). We start by describing in detail both studied pose representations. Then we describe the key elements of our fusions strategies.

3.1 Pose representations

Hierarchical heatmap representation. *Heatmap* is a feature map generated by a keypoint-based pose extractor network before computing the output coor-

(a) Pose Heatmaps. (b) DensePose.

Fig. 1: **Pose representations.** Both pose heatmaps and dense pose representations are studied as modalities. For 1a, from left to right, each of the joint group channels used as a pose heatmap representation: Right leg and hip, Left leg and hip, Right arm and eye, Left arm and the whole body at last channels. For 1b, images I+V are displayed. Best viewed in digital format.

dinates of each body joint. These maps hold a channel per body joint indicating the probability distribution of the target joint. Thus, a higher value indicates more confident detection while a low value may indicate that the joint is not visible or its estimation is low confident.

Therefore, heatmaps codify richer information than single 2D/3D coordinates and allow the model to be more robust to low-confident joint locations due to noise or occlusions. We regroup the joint heatmaps into the following hierarchical schema, which is composed of five channels: The first four channels contain different parts of the body like the left/right arm and the left/right leg while the last channel contains an image with all the joints of the full body. In this way, this representation allows us to better isolate the movement of each arm or leg from each other while keeping an overall description of the motion of the whole body, in addition, to remove the memory requirements. Heatmap aggrupation is depicted in figure 1a

Dense Pose. DensePose [18] is a dense 3D mesh representation of the human body. This mesh indicates information about body segmentation and the 3D location of each body part.

More concretely, the mesh is codified into three channels: *(i)* a body part segmentation map image I, which splits the human body into 24 segments, where each segment is colored in a different shade of gray – for each body part, a texture planar gradient is used to indicate the horizontal and vertical relative coordinates of each point concerning the origin of each body part; *(ii)* a mapping image U with the horizontal gradient coordinates; and, *(iii)* a mapping image V with the vertical gradient.

3.2 Fusion strategies

We evaluate multiple fusion strategies to aggregate the information from both pose hierarchical heatmaps and the DensePose channels. We first implement strategies for fusing at an *early* stage by aggregating the output layers at a

certain depth through an aggregation function. Secondly, we also explore *late* fusion on the final predictions obtained from each modality.

Early fusion Model architecture is split into two parallel branches, one per modality, until the fusion stage. At the fusion stage, the gait features obtained from each branch are aggregated into one feature map that is fed to the rest of the model layers to obtain the final prediction. We consider different aggregation strategies among the gait features:

- **Concatenation**: Features produced by the last layer of each parallel branch are concatenated along the filter dimension. After fusion, the input filters of the immediate after layer hold duplicated input filters to accommodate the duplicated size while maintaining the number of output filters.
- **Sum**: The features output by each branch are summed into a single feature map. Hence, we sum the output features produced by the last layer on each branch.
- **Weighted sum**: Features are aggregated through a weighted sum. Weights are learned as the rest of the model parameters during the training process. Thus, the model finds the optimal balance among both modalities.
  Let n the number of modalities and β_i the trainable parameter associated with the i-th modality, the weight w_i associated with the i-th modality is computed through softmax function. β_i is divided by a factor $\sqrt{n}$ so to stabilize values as follows:

$$w_i = \text{softmax}(\frac{\beta_i}{\sqrt{n}}) \tag{1}$$

  This fusion also adds a layer normalization operator followed by a residual connection to the original feature values. Early fusion with Weighted sum is illustrated in figure 2.

Late fusion Fusion is performed at inference time on the final prediction output by every single model trained with each modality. At test stage, we compute the final embedding returned by each model for every sequence within gallery and probe sets. Then, softmax probability scores are computed for every sequence, based on the pair distance among the embeddings in the gallery, and the embeddings in the probe. Softmax scores measures, for every sequence, the affinity among the target subject and the subjects in the gallery. Fusion of the softmax probabilities is carried out by means of the following strategies:

- **Product**: Let P_i the set of softmax scores vectors output from the m_i modality. The final softmax score vectors S_{prod} can be computed as follows:

$$S_{prod}(v = c) = \prod_{i=1}^{n} P_i(m_i = c) \tag{2}$$

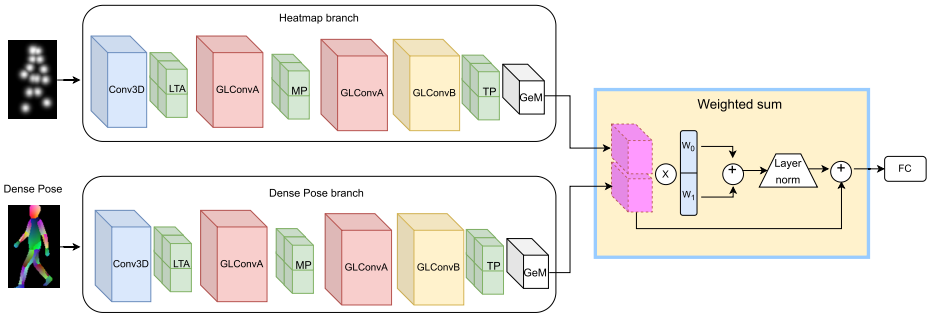

Fig. 2: **Scheme of Early fusion with weighted sum.** Our model has two parallel branches, one is fed with pose heatmaps and the other one with dense pose. Note that fusion is illustrated over the GaitGL [14] architecture on GeM layer. However, we tested other locations as explained in Sec 4.4. GLConv refers to Global-Local Convolution, LTA to Local Temporal Aggregation, MP to Max Pooling, TP to Temporal Pooling; GeM to Generalized-Mean pooling and FC to Fully-Connected layer (Best viewed in digital format).

where $S_{prod}(v = c)$ refers to the probability of assigning the score of video v to the subject c and $P_i(m_i = c)$ is the probability of assigning the identity of subject c to that subject in the modality m_i.

– **Weighted sum**: Final softmax scores S_{ws} are computed from the softmax vector scores P_i output from each modality m_i by a weighted sum as follows:

$$S_{ws}(v = c) = \sum_{i=1}^{n} \lambda_i P_i(m_i = c) \tag{3}$$

Where λ_i is the weight assigned to each modality m_i, subject to $\forall i = 1, .., n,\ \lambda_i > 0$ and $\sum_{i=1}^{n} \lambda_i = 1$.
Considering $n = 2$ modalities, we grid values for λ_1 from 0.1 to 0.9 with steps of 0.1 for one modality, and assign $\lambda_2 = 1 - \lambda_1$ to the other modality.

4 Experiments and results

In this section, we report the experimental results of our fusion approach. Firstly, we describe the datasets and metrics considered to evaluate our models' performance. Then, we report the implementation details of our models. Finally, we report our experimental results and the comparison against the *state of the art*.

4.1 Datasets

We carry out our experimental study on CASIA-B dataset [24]. Note that other popular datasets like OU-MVLP [21] or GREW [25] have not released the original

RGB video sequences, so it is not possible to apply the pose estimators on them to extract the pose heatmaps or the dense poses.

CASIA-B collects 124 subjects walking in an indoor environment while they are recorded from 11 different viewpoints (*i.e.* from 0° to 180° in steps of 18°). Video resolution is 320×240 pixels and fps is 24. For every subject, three walking conditions are considered: normal walking (*NM*), carrying a bag (*BG*) and wearing a coat (*CL*). We follow the *Large-Sample Training* (LT) experimental protocol followed too by [14].

The sequences from the first 74 subjects of all the walking conditions and viewpoints are used for training. For the remaining subjects, the first four *NM* sequences are used as gallery set while the rest of the walking conditions and types are used as probe set.
As evaluation metrics, we use the standard Rank-1 (R1) accuracy to measure the accuracy of our models, *i.e.* the percentage of correctly classified videos: $R1 = \#correct/\#total$.

4.2 Implementation details

As GaitGL [14] is the current state-of-the-art model in gait recognition using silhouettes we employ it with the two proposed modalities: pose heatmaps and dense pose. Notice that the model is trained from scratch in all our experiments.

Table 1 summarises the training hyperparameters. For training, input samples contain 30 frames to reduce memory requirements, while at test time, we use all video frames to evaluate model accuracy.

Regarding pose image preprocessing, our input data is scaled and cropped so that the subject is always located in the middle of the frame, resulting in an input shape of 64×44. Pose heatmaps are obtained through ViTPose [23], while I-V images are obtained from DensePose [18]. Since image I is represented with 25 gray tones (24 body parts + background), we scale its values to the complete gray scale range ($[0, 255]$). We performed preliminary ablation experiments with every single I, U, and V image and concluded that image U does not provide valuable data. Hence, we only use I+V images. Figure 1b shows the I+V channels.

Table 1: **Training hyperparameters.** Description of the hyperparameters used to train GaitGL [14].

GaitGL [14] train hyperparameters	
# 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
GeM pooling	p initial value: 6.5
Triplet loss margin	0.2

For early fusion with the weighted sum, weights β_i for every modality are initialized to 1, and we allow the model to find the optimal values during the

training process. All the models are developed using OpenGait [6] and PyTorch v1.12.1.

4.3 Baseline results

Firstly, we train and evaluate the base GaitGL model with each individual pose heatmaps and dense pose modalities. Hence, we obtain the baseline accuracy that can be obtained per each individual modality.

Table 2 reports the accuracy obtained by the base GaitGL model trained on every single modality. It can be observed that Dense Pose representation achieves higher accuracy as it manages richer information than Pose Heatmaps representation.

Table 2: **Baseline accuracy**. Top-1 accuracy (%) obtained per each single modality at test stage. Mean accuracy per each walking condition (NM, BG and CL) is reported along with the overall accuracy. The best result is highlighted in bold.

Modality	Walking condition			Mean
	NM	*BG*	*CL*	
Hierarchical pose heatmaps	92.7	80.0	70.3	81.0
Dense Pose	96.7	91.6	83.2	**90.5**

4.4 Study of early fusion strategies

In this section, we report a thorough study of the proposed early fusion strategies over each stage of the GaitGL architecture. Thus, we have tested the proposed aggregation strategies at several locations of the model: Conv3D, LTA, first GLConvA layer (called GLConvA0), second GLConvA layer (called GLConvA1) and GLConvB (called GLConvB), TP, GeM and FC.

The mean global top-1 accuracy obtained by each early fusion method over all the fusion stages is summarised in Figure 3.

It can be observed that fusion strategies based on both concatenation (blue bars) and sum (red bars) obtain worse results than the baseline result achieved by the single Dense Pose modality, indicating that a more complex fusion strategy is necessary.

Thus, we also tested fusion through the weighted sum, where the contribution of each modality must be learned. The results of this fusion strategy (green bars) show an important improvement with respect to previous fusion schemes. It can improve in 2.3% the best result achieved by the Dense Pose baseline when applied at the GeM module's output. This best model holds 4.543 M of parameters, and the average inference time per sample on an NVIDIA Titan Xp GPU is 28 ms.

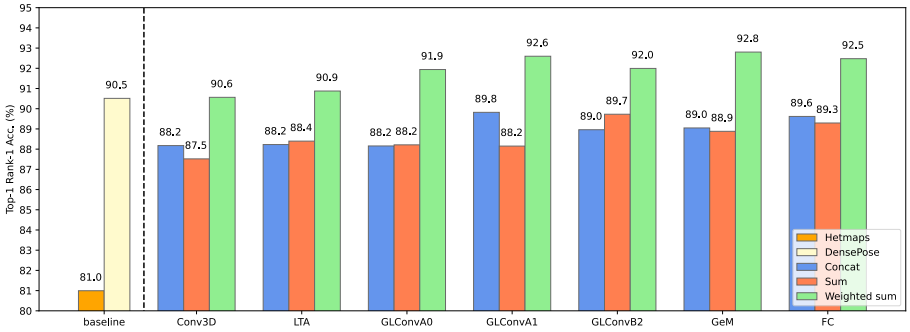

Fig. 3: **Results with Early fusion**. Mean Top-1 for all the early fusion strategies in all the fusion locations. Note that the accuracy scale is cropped to 80-95 %. Best viewed in digital format.

4.5 Study of late fusion strategies

The following tests focus on evaluating our proposed strategies for late fusion: product and weighted sum (w-sum). Thus, Table 3 collects both the mean top-1 accuracy (%) for every walking condition and the global mean for each late fusion strategy.

Table 3: **Results with Late fusion**, for both Product and Weighted sum strategies. Mean Top-1 accuracy, in percentage, for every walking condition is reported along with the overall mean for all the walking conditions. Note that λ_{hm} refers to the weight associated with heatmaps and λ_{dp} the weight associated with densepose.

Fusion	Weights		Walking condition			Mean
			NM	*BG*	*CL*	
Product			97.5	91.5	83.4	90.8
W-sum	λ_{hm}	λ_{dp}				
	0.1	0.9	97.1	92.2	84.0	91.1
	0.2	0.8	97.4	**92.7**	84.8	91.6
	0.3	0.7	**97.6**	**92.7**	**84.9**	**91.7**
	0.4	0.6	**97.6**	92.3	84.4	91.4
	0.5	0.5	97.5	91.5	83.4	90.8
	0.6	0.4	97.1	90.3	82.0	89.8
	0.7	0.3	96.5	88.6	80.0	88.4
	0.8	0.2	95.7	86.5	77.4	86.5
	0.9	0.1	94.6	83.6	74.1	84.1

It shows that both product and w-sum improve the baseline results. For w-sum, when dense pose modality is weighted under 0.5, the performance gets lower

Table 4: **State-of-the-art comparison on CASIA-B.** Comparison with other pose-based and shape-based models. Mean Top-1 accuracy, in percentage, for every walking condition is reported along with the overall mean for all the walking conditions. Best other results per each data type are highlighted in italic-bold.

Data	Model	Walking condition			Mean
		NM	*BG*	*CL*	
Pose	TransGait [9] (Pose + STM)	84.5	71.2	54.4	70.0
	PoseMapGait [12]	79.3	61.1	48.1	62.8
	PoseGait [13]	68.7	44.5	39.9	49.7
	GaitGraph [22]	87.7	74.8	66.3	76.3
	End-to-end Pose LSTM [10]	66.1	49.3	37.0	50.8
	End-to-end Pose CNN [10]	91.2	83.9	60.2	***78.4***
Shape	GaitSet [3]	95.0	87.2	70.4	84.2
	GaitPart [5]	96.2	91.5	78.7	88.8
	GaitGL [14]	97.4	94.5	83.6	91.8
	TransGait [9] (Sil + STM)	97.3	92.8	80.6	90.2
	TransGait [9] (Multimodal)	98.1	94.9	85.8	***92.9***
	End-to-end shape model [10]	97.5	90.6	75.1	87.7
	End-to-end ensemble [10]	97.9	93.1	77.6	89.5
Ours	PoseFusionGaitGL (tw-sum after GeM)	98.4	93.4	86.6	**92.8**

than the baseline accuracy. For most of the late fusion strategies, performance gets higher than early fusion by concatenation and sum methods. Finally, the weighted sum with $\lambda = 0.3$ for the pose heatmaps features and $\lambda = 0.7$ for the dense pose features achieves the best performance but does not improve the best results with the best early fusion strategy.

4.6 Comparison to the state of the art using pose

Table 4 compares our best approach with the state-of-the-art pose-based and shape-based models. Our proposal outperforms all the pose-based methods, including 'End-to-end Pose' [10], which has been trained end-to-end with the original RGB frames. Our model obtains the best accuracy in the 'walking with clothes' (CL) condition and improves over the performance of most of the silhouette-based approaches, such as GaitGL [14], or GaitPart [5], and reaches very competitive results with multimodal TransGait [9] which has been trained using both pose and silhouette modalities.

These results show the capability of our fusion strategies to optimally aggregate features from multiple pose representation modalities using little or no shape data.

5 Conclusions

In this work, we presented an experimental study of multiple fusion pipelines for a multimodal framework for gait recognition that exploits various human

pose representations. Concretely, we consider two pose representations: (a) pose heatmaps rearranged in a hierarchical decomposition of the human limbs and (b) dense pose.

As fusion strategies, we proposed, on the one hand, early fusion on the output descriptors produced by different layers of the model through concatenation and sum. In addition, we introduced a weighted sum, where the weights are learned during the training process and allow the model to leverage both modalities optimally. And, in the other hand, we proposed multiple late fusions on the final softmax probabilities output by each branch.

We evaluated our pose fusion approaches on the base GaitGL model for silhouette-based gait recognition. We maintained the original GaitGL architecture except for the parallel branches and fusion. We also included a comparison against the baseline performance achieved by every single modality.

Our experimental results show that: (a) Concatenation and sum early fusion methods do not allow one modality to enrich the features of the other modality, so results get poor. By contrast, weighted sum allows the model to learn to combine the features produced by each modality in a more optimal way. (b) Late fusion generally obtains good results too, and improves the baseline performance. And, (c) pose fusion obtains higher results than pose-based models and most shape-based models. Our approach reaches comparable performance to other multimodal fusions proposals based on silhouettes.

In future work, we plan to extend our study to alternative pose representations models that provide new perspectives on gait motion. In addition, we consider studying more elaborated fusion strategies. In our study, all the proposed fusion strategies treat equally the whole modality feature without taking into account useful information that is derived from local regions.

References

1. An, W., Liao, R., Yu, S., Huang, Y., Yuen, P.C.: Improving gait recognition with 3d pose estimation. In: Chinese Conference on Biometric Recognition. pp. 137–147. Springer (2018)
2. Castro, F.M., Marín-Jiménez, M.J., Guil, N., de la Blanca, N.P.: Multimodal feature fusion for CNN-based gait recognition: an empirical comparison. Neural Computing and Applications pp. 1–21 (2020)
3. Chao, H., He, Y., Zhang, J., Feng, J.: Gaitset: Regarding gait as a set for cross-view gait recognition. In: Procs. AAAI Conference on Artificial Intelligence (2019)
4. Delgado-Escaño, R., Castro, F.M., Cózar, J.R., Marín-Jiménez, M.J., Guil, N., Casilari, E.: A cross-dataset deep learning-based classifier for people fall detection and identification. Computer methods and programs in biomedicine **184**, 105265 (2020)
5. 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: CVPR. pp. 14225–14233 (2020)
6. Fan, C., Shen, C., Liang, J.: OpenGait (2022), https://github.com/ShiqiYu/OpenGait

7. Hou, S., Cao, C., Liu, X., Huang, Y.: Gait lateral network: Learning discriminative and compact representations for gait recognition. In: European Conference on Computer Vision. pp. 382–398. Springer (2020)
8. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: Computer Vision and Pattern Recognition (CVPR) (2018)
9. Li, G., Guo, L., Zhang, R., Qian, J., Gao, S.: Transgait: Multimodal-based gait recognition with set transformer. Applied Intelligence pp. 1–13 (2022)
10. Li, X., Makihara, Y., Xu, C., Yagi, Y., Yu, S., Ren, M.: End-to-end model-based gait recognition. In: CVPR (2020)
11. Liao, R., Cao, C., Garcia, E.B., Yu, S., Huang, Y.: Pose-based temporal-spatial network (ptsn) for gait recognition with carrying and clothing variations. In: Chinese Conference on Biometric Recognition. pp. 474–483. Springer (2017)
12. 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)
13. Liao, R., Yu, S., An, W., Huang, Y.: A model-based gait recognition method with body pose and human prior knowledge. Pattern Recognition (2020)
14. Lin, B., Zhang, S., Yu, X.: Gait recognition via effective global-local feature representation and local temporal aggregation. In: iccv. pp. 14648–14656 (2021)
15. Marín-Jiménez, M.J., Castro, F.M., Delgado-Escaño, R., Kalogeiton, V., Guil, N.: Ugaitnet: Multimodal gait recognition with missing input modalities. IEEE Transactions on Information Forensics and Security **16**, 5452–5462 (2021)
16. Meng, Z., Fu, S., Yan, J., Liang, H., Zhou, A., Zhu, S., Ma, H., Liu, J., Yang, N.: Gait recognition for co-existing multiple people using millimeter wave sensing. In: Proceedings of the AAAI Conference on Artificial Intelligence (2019)
17. Nakajima, K., Mizukami, Y., Tanaka, K., Tamura, T.: Footprint-based personal recognition. IEEE Trans. Biomedical Engineering **47**(11), 1534–1537 (2000)
18. Riza Alp Güler, Natalia Neverova, I.K.: DensePose: Dense human pose estimation in the wild. In: CVPR (2018)
19. Sepas-Moghaddam, A., Etemad, A.: Deep gait recognition: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence (2022)
20. Shen, C., Yu, S., Wang, J., Huang, G.Q., Wang, L.: A comprehensive survey on deep gait recognition: Algorithms, datasets and challenges (2022), https://arxiv.org/abs/2206.13732
21. 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 Trans. Computer Vision and Applications **10**(1), 4 (2018)
22. Teepe, T., Khan, A., Gilg, J., Herzog, F., Hormann, S., Rigoll, G.: Gaitgraph: Graph convolutional network for skeleton-based gait recognition. In: 2021 IEEE International Conference on Image Processing (ICIP). IEEE (sep 2021)
23. Xu, Y., Zhang, J., Zhang, Q., Tao, D.: ViTPose: Simple vision transformer baselines for human pose estimation. In: NeurIPS (2022)
24. Yu, S., Tan, D., Tan, T.: A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In: Proc. ICPR. vol. 4, pp. 441–444 (2006)
25. Zhu, Z., Guo, X., Yang, T., Huang, J., Deng, J., Huang, G., Du, D., Lu, J., Zhou, J.: Gait recognition in the wild: A benchmark. In: ICCV. pp. 14789–14799 (2021)