Towards a Deeper Understanding of Skeleton-based Gait Recognition

Torben Teepe* Johannes Gilg Fabian Herzog Stefan Hörmann Gerhard Rigoll
Technical University of Munich
*t.teepe@tum.de

Abstract

Gait recognition is a promising biometric with unique properties for identifying individuals from a long distance by their walking patterns. In recent years, most gait recognition methods used the person's silhouette to extract the gait features. However, silhouette images can lose fine-grained spatial information, suffer from (self) occlusion, and be challenging to obtain in real-world scenarios. Furthermore, these silhouettes also contain other visual clues that are not actual gait features and can be used for identification, but also to fool the system. Model-based methods do not suffer from these problems and are able to represent the temporal motion of body joints, which are actual gait features. The advances in human pose estimation started a new era for model-based gait recognition with skeleton-based gait recognition. In this work, we propose an approach based on Graph Convolutional Networks (GCNs) that combines higher-order inputs, and residual networks to an efficient architecture for gait recognition. Extensive experiments on the two popular gait datasets, CASIA-B and OUMVLP-Pose, show a massive improvement (3×) of the state-of-the-art (SotA) on the largest gait dataset OUMVLP-Pose and strong temporal modeling capabilities. Finally, we visualize our method to understand skeleton-based gait recognition better and to show that we model real gait features.

1. Introduction

Gait is a soft biometric with huge, unique advantages compared to hard biometrics like face, iris, or fingerprint. Human gait can be described as the way a person walks, or more formal, the movement pattern of the limbs during motion. Gait patterns can be observed at a distance, without a person's compliance, and are hard to disguise. This is a considerable advantage compared to hard biometrics, which requires the user to interact with a sensor. For applications like surveillance and forensic identification, gait offers vast potential; however, it also entails risks for privacy and misuse.

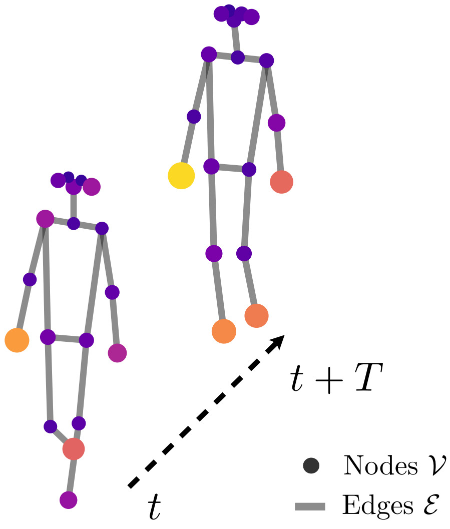

Figure 1. Example of two poses represented as graph at different time steps.

There are also challenges when using gait as a biometric. Gait is sensitive to carried items, worn cloth, and surface type. A different type of footwear like sneakers compared to boots or heels may considerably change a person's gait. The biggest challenge in gait recognition is obtaining unique features invariant to these influences.

Most approaches [2, 4, 5, 7, 8, 13, 21, 27] use silhouettes to extract the gait features from a video sequence. The silhouette is commonly obtained by background subtraction. Depending on the method used for background subtraction, this may cause undesired artifacts on the contour, as shown in the two center frames in Fig. 2. Background subtraction may be reliable in a lab setting but becomes a complex problem in a real-world scenario with cluttered and rapidly changing environments. While most approaches [2, 4, 5, 7, 8, 13, 27] do not consider how to obtain the silhouettes, one approach [21] trains a separate Convolutional Neural Network (CNN) for background subtraction in a real-world scenario. A significant drawback of the silhouette representation is the sensitivity to clothing and carried items. In Fig. 2 the persons bag is clearly visible in the silhouette. Furthermore, the silhouette reveals many appearance clues, like hairstyle and clothes. Hence, these approaches are more comparable to re-identification tasks.

Recent deep learning based pose estimation allow to generate keypoints robust against occlusion, cluttered, and changing backgrounds [23]. These advantages make an approximation to gait features with appearance-based methods obsolete and enable a new generation of model-based gait recognition [14,15,20]. Building upon keypoints brings gait recognition back to an early description of gait [11]:

> A few bright spots describing the motions of the main joints [...] evoke a compelling impression of human walking.

Skeleton-based approaches offer a cleaner gait representation, only capturing the spatial posture and the temporal movement. Thus we can bring back *actual* gait with a focus on motion recognition instead of visual recognition.

Current model-based approaches [14, 15, 20] still lack performance compared to appearance-based methods. We introduce GCNs to process the skeletons described by the keypoints and bridge the gap to appearance-based methods even further.

Our contributions can be summarized as follows:

1. We propose a multi-branch graph-based interpretation of gait together with a GCN architecture that can efficiently learn features on this graph.
2. We provide a deeper understanding of gait with extensive ablation and visualization of our features.
3. Our empirical experiments show SotA results by a huge margin on the largest model-based gait dataset OUMVLP-Pose.

The code and models are publicly available[1].

2. Related Work

Although skeleton-based gait recognition is a relatively new research area, silhouette-based approaches have a long history. This section will give an overview of these two areas of gait recognition and other skeleton-based human understanding that inspired this work.

2.1. Gait Recognition

In recent years, appearance-based approaches using a silhouette representation as input dominated gait recognition. Model-based approaches only played a minor role, but lately, new model-based approaches have emerged using pose estimation and the human skeleton as the gait feature representation. Silhouette-based methods still set the SotA for gait recognition, but recently skeleton-based approaches have started to challenge this lead.

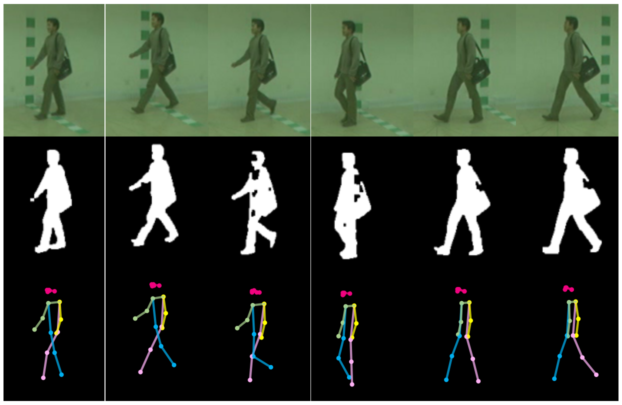

Figure 2. Comparison a sequence of RGB images and the respective gait representation in silhouette image and skeleton, from top to bottom. Images are from the CASIA-B [30] dataset.

2.1.1 Silhouette-based

Silhouette-based methods relied on a binary human image extracted from the original image [13]. Background subtraction can obtain these silhouettes for static scenes, but for dynamic and changing settings, this task becomes more complicated [21].

Silhouette approaches are distinctive for their temporal modeling and group into single-image, sequence-based, and set-based approaches. Early approaches summarized a gait cycle into a single image, i.e., Gait Energy Image (GEI) [2, 8]. These representations lose most of the temporal information but allow for easier processing. Sequence-based approaches focus on each input separately. For modeling the temporal information 3D-CNN [27] or Long Short-Term Memorys (LSTMs) [7] are used. These approaches can comprehend more spatial information and more temporal information at higher computational costs. The set-based approach [4, 5] with shuffled inputs models no temporal information, thus has less computational complexity.

2.1.2 Skeleton-based

Model-based approaches were not very popular in recent years due to the high computing complexity. Robuster and more lightweight pose estimators [3,6] enabled a comeback for model-based approaches. The keypoints of the human body allow modeling a skeleton representation of a person. The first approach to propose the use of pose estimation and utilize pose keypoints is Pose-Based Temporal-Spatial Network (PTSN) [14]. *PTSN* proposed a two-path network architecture with an CNN for spatial modeling, and an LSTM for temporal modeling. The authors show superior results on CASIA-B [30] compared to silhouette-based approaches in special walking conditions in the same-view setting.

Later, the same group proposed *PoseGait* [15], which

[1] https://github.com/tteepe/GaitGraph2

uses 3D pose estimation with handcrafted features. The approach uses the 3D keypoints in euclidean space to calculate the joint angle, bone length, and joint motion. Using these handcrafted features, a CNN then learns high-level spatio-temporal features. This approach is evaluated in a cross-view setting on CASIA-B [30] and shows competitive results to silhouette-based approaches.

We build on the work GaitGraph [25], which introduced the first approach with GCNs for skeleton-based gait recognition. With an adapted GCN architecture for action recognition and integrated spatial and temporal modeling, GaitGraph gave a considerable performance increase for model-based gait recognition.

A recent approach [26] combines the advantages of silhouette-based and skeleton-based approaches in a two-branch GCN and CNN network architecture and achieve even higher recognition rates. Showing that skeleton-based approaches preserve information not captured in silhouette-based methods.

2.2. Skeleton-based Action Recognition

While skeleton-based gait recognition was only recently becoming popular, other areas of human understanding have already employed skeleton features for a few years. First and foremost, the area of skeleton-based action recognition pioneered most of the current graph-based spatio-temporal GCN architectures, including the ones used in this paper [22, 29].

Yan *et al*. [29] introduced GCNs to action recognition with the Spatial-Temporal Graph Convolutional Network (ST-GCN) architecture, in which skeleton data is represented as a graph with natural skeleton connections. ST-GCN interleaves spatial graph convolutions along with temporal convolutions for spatial-temporal modelling. Many more GCN architectures and improvements on the original ST-GCN design have since been proposed. A notable example is the Two-Stream Adaptive Graph Convolutional Network (2s-AGCN) [17] that introduced a adaptive adjacency matrix and pre-computed bone information as second-order input. SGN [31] provides a deeper analysis of these second-order features and proposes a data pre-processing step that adds pre-computed velocity and bone information to the raw keypoint input. Another architecture used in this paper is ResGCN [22], which added multiple residual connections in the ST-GCN blocks and a bottleneck for feature dimension reduction. ResGCN also adopts the higher-order input with a multi-branch input structure.

3. Skeleton-based Gait Recognition

3.1. Notation

We describe the human skeleton as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{v_1, \ldots, v_N\}$ is the set of N nodes representing joints, and $\mathcal{E}$ is the set of edges representing bones captured by an adjacency matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$ with $\mathbf{A}_{i,j} = 1$ if an edge connects from v_i to v_j and $\mathbf{A}_{i,j} = 0$ otherwise. $\mathbf{A}$ is symmetric since $\mathcal{G}$ is undirected. Every node consists of three channels $v_n = (x_n, y_n, c_n)$, with the estimated x, y coordinate and the keypoint confidence c.

For gait recognition, we use a sequence of these graphs. Thus we add a temporal dimension T. The sequence is then defined as the tensor $\mathbf{X} = \{v_{t,n} \in \mathbb{R}^3 \mid t, n \in \mathbb{N}_0, t < T, n < N\}$ and $\mathbf{X} \in \mathbb{R}^{T \times N \times C}$.

Overall the gait sequence can be described structurally by the adjacency matrix $\mathbf{A}$ and the feature tensor $\mathbf{X}$.

3.2. Graph Convolutions

An essential building block of our network architecture are graph convolutions. On skeleton inputs, defined by features $\mathbf{X}$ and graph structure $\mathbf{A}$, the layer-wise update rule of graph convolutions can be applied to features at time t as:

$$\mathbf{X}_t^{(l+1)} = \sigma\left(\tilde{\mathbf{D}}^{-\frac{1}{2}}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-\frac{1}{2}}\mathbf{X}_t^{(l)}\Theta^{(l)}\right), \quad (1)$$

where $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}$ is the skeleton graph with added self-loops to keep identity features. $\tilde{\mathbf{D}}$ is the diagonal degree matrix of $\tilde{\mathbf{A}}$, and $\sigma(\cdot)$ is an activation function. The term $\tilde{\mathbf{D}}^{-\frac{1}{2}}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-\frac{1}{2}}\mathbf{X}_t^{(l)}$ can be intuitively interpreted as a spatial feature aggregation from the messages passed by the direct neighbors. The adjacency matrix $\mathbf{A}$ is obtained using the spatial partition presented in [29].

3.3. Pose Estimation

The skeletons are extracted from the RGB images of the dataset.v Keypoint estimation aims to detect the locations of N keypoints (e.g., shoulder, hip, or knee) from an image $\mathbf{I} \in \mathbb{R}^{W \times H \times 3}$. A common method is the top-down-approach [23], which predicts N heatmaps $\{\mathbf{H}_1, \mathbf{H}_2, \ldots, \mathbf{H}_N\}$ of size $W' \times H'$, where the heatmap $\mathbf{H}_n$ indicates the location of the n-th keypoint. The location of the maximum of these heatmaps $\mathbf{H}_n$ yields the location of the keypoint v_n that define the edges $\mathcal{E}$.

Official keypoints [1] for the OUMVLP dataset keypoints are provided. The authors rely real-time pose estimators to allow real-time applications of the approach. In our work, we decided to extract the 2D keypoints on CASIA-B using HRNet [23] pre-trained on the COCO dataset [16]. It is an offline approach but yields higher keypoint accuracy than real-time approaches. Tab. 1 shows a comparison of the two datasets and the used keypoints. The COCO pose annotations consist of 17 keypoints. We define the bones or edges $\mathcal{E}$ as shown in Fig. 1.

3.4. Network Architecture

For our task, we adapted the ResGCN [22] architecture designed initially for action recognition. Blocks of this architecture are based on the ST-GCN block. and contain a sequential execution of a spatial graph convolutions and a temporal 2D convolutions. The ResGCN approach introduces a bottleneck structure, based on ResNet by He *et al.* [9]. The bottleneck adds two 1×1 convolutional layers before and after a convolution layer to reduce the number of feature channels with a reduction rate r. ResGCN adds this bottleneck for every spatial and temporal block, thus reducing the number of parameters.

Another modification to the original ST-GCN architecture is the addition of residual connections. Song *et al.* [22] propose residual connections that connect the features before and after every spatial and temporal block.

A ResGCN bottleneck block is shown in Fig. 3 and an overview of the overall structure in Fig. 4.

3.5. Multi-Branch Input

As proposed by works in skeleton-based gait recognition [22, 31], we pre-compute features of our skeleton information. In this work, we use three features: 1) joint positions, 2) motion velocities, and 3) bone features.

The first branch contains the joint positions. We also add the relative position of each joint to the center of the pose c (*i.e.* nose or neck):

$$\mathcal{R} = \{v_{t,n} - v_{t,c} \mid t, n \in \mathbb{N}_0, t < T, n < N\}.$$

The second branch uses the motion velocities $\mathcal{F}$ as input. We calculate the difference to the same joint in the two next frames for each joint:

$$\mathcal{F} = \{v_{t+i,n} - v_{t,n} \mid t, n \in \mathbb{N}_0, i \in \{1,2\}, t < T-2, n < N\}.$$

Finally, the input of the last branch is the bone length $\mathcal{L}$ and bone angles $\mathcal{A}$. For the bone length, we subtract the coordinate of every joint n with every connected joint n_{adj}:

$$\mathcal{L} = \{v_{t,n} - v_{t,n_adj} \mid t, n \in \mathbb{N}_0, t < T, n < N\}.$$

Finally, the angle of each bone is:

$$\mathcal{A} = \left\{ \arccos\left(\frac{v_{t,n} - v_{t,n_adj}}{\sqrt{\sum v_{t,n}^2}} \right) \mid t, n \in \mathbb{N}_0, t < T, n < N \right\}.$$

4. Experiments

4.1. Datasets

In recent years, the focus for gait recognition was on silhouette-based approaches. Hence, most datasets provided only silhouettes. One of the widely used datasets, CASIA-B [30], provides RGB images on which we can run the pose estimation. Recently, an extension with the pose data OUMVLP-Pose [1] was published for the largest public gait database OU-MVLP [24]. Tab. 1 shows a comparison of the two datasets. These two popular gait datasets provide a good comparison to other methods.

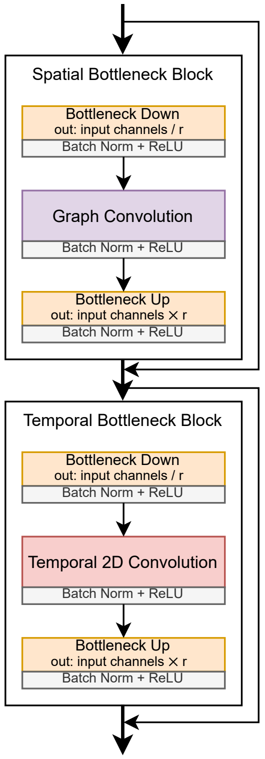

Figure 3. Structure of ResGCN bottleneck block including the residual connections.

CASIA-B [30] is popular multi-view gait dataset with 124 subjects. The dataset contains 3 walking conditions recorded in 11 views (0°, 18°, . . . , 180°). The walking conditions are normal (NM) (6 sequences per subject), walking with a bag (BG) (2 sequences per subject), and wearing a jacket or a coat (CL) (2 sequences per subject). In total each subject contains $11 \times (6 + 2 + 2) = 110$ sequences.

CASIA-B has no official partition of training and test set, but several experiment protocols exist [32]. In this work, we use the popular protocol proposed in [28] for a fair comparison. Furthermore, we use the commonly called large-sample training (LT) partition. The train set contains the first 74 subjects for this partition, and the remaining 50 subjects build the test set. In the test set, the gallery comprises four sequences of the NM condition (NM #1-4), and the remaining six sequences are divided into three probe subsets, i.e., NM subsets containing NM #5-6, BG subsets containing BG #1-2 and CL subsets containing CL #1-2.
We extract the poses from CASIA-B with the pre-trained HRNet [23] pose estimator.

OUMVLP-Pose [1] is based on the multi-view large-scale gait dataset, OUMVLP [24]. OUMVLP is currently the largest gait dataset and contains 10,307 subjects captured by seven cameras in a round-trip walking course, re-

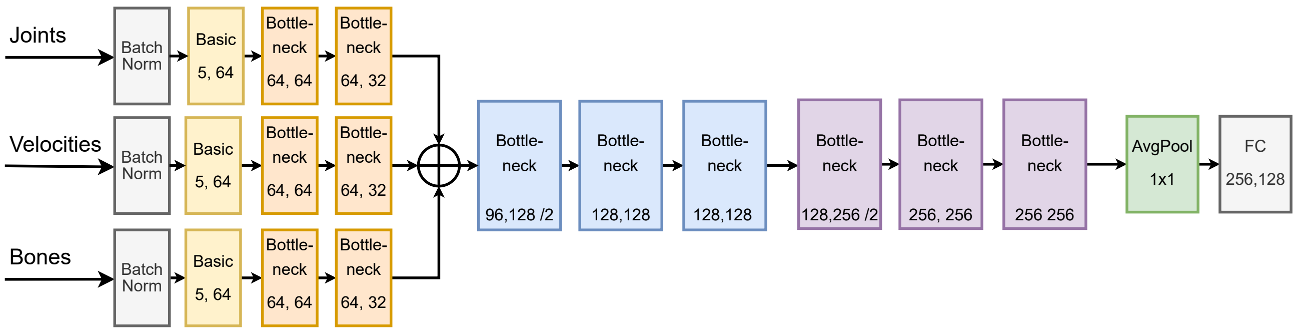

Figure 4. Overview of the multi-branch ResGCN architecture.

	CASIA-B	OUMVLP-Pose	
#IDs	124	10,307	
#Sequences per ID	2 × 3	2	
Keypoint Estimatior	HRNet	OpenPose	AlphaPose
Keypoint Accuracies	75.8 mAP	64.2 mAP	71.0 mAP
#Keypoints	17	18	

Table 1. Comparison of the two skeleton gait datasets. Keypoint accuracy of the pose estimator is reported on the COCO test-dev dataset.

sulting in effectively 14 views in a 15° interval (0°, 15°, . . . , 90°; 180°, 195°, . . . , 270°). Every sequence contains from 18 to 35 frames, and on average, 25 frames. The dataset is split into 5,153 subjects for training and 5,154 subjects for testing. For testing, sequences with index #01 assemble the gallery, while the other sequences are used for the probe set.

OUMVLP-Pose keypoints are extracted from the (unreleased) RGB frames. Two datasets were created using different pre-trained pose estimators, OpenPose [3], and AlphaPose [6], but containing the same frames and subjects.

4.2. Implementation Details

For the training setup, we use an Adam optimizer with a *1-cycle* learning rate schedule [19] with a maximum learning rate of 0.005. The embedding size is 128, the loss function's temperature is 0.01, and the batch size is 768. After 80% of the maximum epochs, we use Stochastic Weight Averaging (SWA) [10]. All the experiments are conducted on a single NVIDIA 3090 GPU with PyTorch. Due to the different properties of the size of the datasets, we employ different model sizes and training strategies.

For **CASIA-B** we use the *ResGCN-N21-R8* architecture with 350 K parameters and a sequence length of $T = 60$ and train for 200 epochs.

For **OUMVLP-Pose** we us a sequence length of $T = 30$ and train for 750 epochs. The network setup is *ResGCN-N51-R4* with 765 K parameters.

Loss As the loss function, we use supervised contrastive (*SupCon*) loss as proposed by [12]. Compared to traditional contrastive losses such as triplet loss or N-pairs loss, this current batch contrastive loss considers all positive and negative samples in the batch. The small size of the skeleton data allows us to run big batch sizes; thus, each batch should contain a positive pair. If an element has only negative pairs or no pairs, it will be ignored.

Augmentation Our network design relies on a constant input sequence length. Thus, we first add mirror padded frames for sequences shorter than the desired sequence length in the temporal dimension. Afterward, we randomly pick a subsequence of the desired length. Additionally, we flip the images from left to right and flip every left joint with the right and vice versa. We add various uniform noises to the estimated keypoints and their confidence to account for pose estimator inaccuracies.

Evaluation & Test Time Augmentation (TTA) At test time, the distance between gallery and probe is defined as the cosine similarity of the corresponding feature vectors. The same sequence padding augmentation from train time is applied, but we pick a sub-sequence of the required length from the sequence center. We also use two additional samples: a left/right flipped sample and a time inverted sample. The resulting three embeddings are concatenated for the later distance calculation.

4.3. Comparison with State-of-the-Art

First, we compare the results of model-based approaches on **CASIA-B** in Tab. 2. Currently, we can only compare two approaches. The other skeleton-based approach PTSN [14] did not publish the results with the same evaluation protocol (excluding the same view). We can still compare to a follow-up paper by the same group PoseGait [15].
All approaches are more robust in the askew angles than in the strict side view, front view, or back view angles (0°, 90°, 180°). Another commonality is that all approaches suffer in performance with different walking conditions. While the drop is more substantial for PoseGait [15], the GCN-based

Gallery NM#1-4		0°-180°											mean
Probe		0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	
NM#5-6	PoseGait [15]	55.3	69.6	73.9	75.0	68.0	68.2	71.1	72.9	76.1	70.4	55.4	68.7
	GaitGraph [25]	85.3	88.5	91.0	92.5	87.2	86.5	88.4	89.2	87.9	85.9	81.9	87.7
	GaitGraph2	78.5	82.9	85.8	85.6	83.1	81.5	84.3	83.2	84.2	81.6	71.8	82.0
BG#1-2	PoseGait [15]	35.3	47.2	52.4	46.9	45.5	43.9	46.1	48.1	49.4	43.6	31.1	44.5
	GaitGraph [25]	75.8	76.7	75.9	76.1	71.4	73.9	78.0	74.7	75.4	75.4	69.2	74.8
	GaitGraph2	69.9	75.9	78.1	79.3	71.4	71.7	74.3	76.2	73.2	73.4	61.7	73.2
CL#1-2	PoseGait [15]	24.3	29.7	41.3	38.8	38.2	38.5	41.6	44.9	42.2	33.4	22.5	36.0
	GaitGraph [25]	69.6	66.1	68.8	67.2	64.5	62.0	69.5	65.6	65.7	66.1	64.3	66.3
	GaitGraph2	57.1	61.1	68.9	66	67.8	65.4	68.1	67.2	63.7	63.6	50.4	63.6

Table 2. Averaged Rank-1 accuracies in percent on CASIA-B per probe angle excluding identical-view cases compared with other model-based methods.

Probe	Gallery (0°- 270°)			
	OpenPose		AlphaPose	
	CNN-Pose	**GaitGraph2**	CNN-Pose	**GaitGraph2**
0°	8.2	32.9	14.3	54.3
15°	13.9	47.7	22.3	68.4
30°	18.1	53.9	27.2	76.1
45°	22.4	56.8	30.0	76.8
60°	21.3	53.9	28.4	71.5
75°	18.2	54.7	23.4	75.0
90°	10.9	45.4	17.2	70.1
180°	7.3	29.0	7.9	52.2
195°	13.5	35.7	13.6	60.6
210°	12.0	34.3	15.6	57.8
225°	20.5	44.3	25.0	73.2
240°	17.3	46.2	24.1	67.8
255°	13.7	46.4	20.2	70.8
270°	9.4	38.4	16.5	65.3
mean	14.8	**44.3**	20.4	**67.1**

Table 3. Averaged rank-1 accuracies on OUMVLP-Pose.

approaches can handle these conditions better.

While our approach builds upon GaitGraph, it can only match the results closely. Due to the different properties of the two datasets, our approach was more optimized towards the larger, more significant dataset OUMVLP. The main difference in our approach is the multi-branch input; this makes the training much faster and much more stable compared to GaitGraph.

Secondly, the results on **OUMVLP-Pose** are in Tab. 3. For this recently released dataset, we can only compare to the baseline set by the authors CNN-Pose [1]. The results show that our GCN-based methods outperform the CNN-based baseline considerably. We can improve the performance by ∼3× on both keypoint types.

4.4. Comparison with Appearance-based Methods

Appearance-based methods still archive the best results in gait recognition. Nevertheless, the new skeleton-based approaches help model-based approaches to close this gap. Tab. 4 shows this comparison.

One hope for skeleton-based gait recognition is to be more invariant to the changes in the walking condition presented in CASIA-B. For now, the results show that this is not the case. Appearance-based methods can handle these conditions better, and their performance drop is less substantial. A reason could be that the pose estimators are also not as accurate on people wearing coats and bags.

Our approach is the first one to evaluate the two most popular gait datasets and give the first complete comparison to appearance-based methods. Skeleton-based approaches made a big step towards the SotA appearance-based methods; however, there is still a significant gap.

It is also notably that the two-branch approach [26] that combines both paradigms can improve the results in all walking conditions. It shows that both features are complementary to archive overall SotA performance. Especially people wearing a coat are much better recognized with both input modalities.

4.5. Ablation Studies

First, we look at the components of our approach in Tab. 5. As a baseline, we use the original ST-GCN architecture. Following, we analyze how much the different components of our approach contribute to the overall performance.

Temporal Modeling The network's ability to model temporal information is investigated by training and testing sorted or shuffled sequences. Tab. 6 shows three configurations. For this ablation study, we can not use the multi-branch inputs since these contain pre-computed temporal

		CASIA-B			OU-MVLP				
Type	Method	NM	BG	CL	0°	30°	60°	90°	mean
appearance-based	GEINet [18]	-	-	-	11.4	41.5	39.5	38.9	35.8
	GaitNet [21]	91.6	85.7	58.9	-	-	-	-	-
	GaitSet [4]	95.0	87.2	70.4	79.5	89.9	88.1	87.8	87.1
	GaitPart [5]	**96.2**	**91.5**	**78.7**	**82.6**	**90.8**	**89.7**	**89.5**	**88.7**
model-based	PoseGait [15]	68.7	44.5	36.0	-	-	-	-	-
	CNN-Pose [1]	-	-	-	12.3	29.3	30.5	18.1	20.4
	GaitGraph [25]	**87.7**	**74.8**	**66.3**	-	-	-	-	-
	GaitGraph2	82.0	73.2	63.6	**54.3**	**76.1**	**71.5**	**70.1**	**67.1**
combined	Two-Branch [26]	**97.7**	**93.8**	**92.7**	-	-	-	-	-

Table 4. Averaged Rank-1 accuracies in percent comparison with both appearance-based and model-based methods. Results for CASIA-B are in the cross-view setup. Results for OU-MVLP in the model-based categories use OUMVLP-Pose AlphaPose keypoints.

Model	Params	Acc
ST-GCN (Baseline)	3.3 M	60.7
ResGCN-N51	645 K	63.7
+ Multi Input	765 K	66.5
+ Augmentation	"	66.6
+ TTA	"	67.1

Table 5. Ablation study of the model components. All experiments are conducted on OUMVLP-Pose and AlphaPose keypoints.

			CASIA-B			OU-MVLP
	Train	Test	NM	BG	CL	AlphaPose
a	Shuffle	Sort	34.7	27.0	19.0	34.1
b	Sort	Sort	72.8	60.1	44.6	63.1
c	Sort	Shuffle	34.3	27.9	18.5	14.5

Table 6. Spatio-temporal Study. Control Condition: shuffle/sort the input sequence at train/test phase. Results are rank-1 accuracies averaged in percent. CASIA-B results are in the cross-view setup. All experiments are without the multi-input pre-computation and TTA.

information and would harm the validity of the ablation.

Our approach shows a good ability to model temporal features. The performance drops substantially when trained with sorted sequences and tested with shuffled sequences (row c). These results further support our claim of bringing back actual temporal features to gait recognition. Tab 6 also illustrates the spatial modeling abilities in row a. Despite the missing temporal and appearance information, the network can still learn appearance-invariant features of the person's underlying physic.

Pose Estimator The OUMVLP-Pose dataset has keypoints extracted by different pose estimators. The two follow a different approach for keypoint estimation, with AlphaPose being a top-down approach and OpenPose being a bottom-up approach. The performance measured in the detector's mean average precision (mAP) also varies. AlphaPose archives a 71.0 mAP and OpenPose a 64.2 mAP on the COCO test-dev dataset.

These two keypoints allow us to analyze how the gait recognition algorithms scale on differently performing pose estimators. Tab. 3 shows the results on OUMVLP-Pose with the two keypoint estimators. We archive almost the same performance gain on both keypoint types compared to CNN-Pose. This scaling performance indicates that skeleton-based methods can scale with the pose estimators' performance, potentially allowing better performance in the future with emerging improved pose estimators.

4.6. Gait Recognition Analysis

A better understanding of what the network learns from the skeleton data is indicated by the discriminate features of gait. Hence we want to study which joints have the highest activation at different times of the gait cycle. Using the activation map technique [33] we calculate the activation of each joint per time frame as shown in Fig. 5.

The Figure shows that the network looks mainly on the outer limbs. The keypoints on the hands and feet have the highest activation over the gait cycle, and the shoulders and face keypoints (except ears) have the lowest activation. The limb with forward motion has a higher activation than fixed parts in the temporal context. For example in sequence a), the right leg swing from frame 5 to frame 20 shows increased foot activation the more it moves. These observations conclude that the arm-swing and the leg-swing are the most discriminate features for our network. The network also takes hints from the hip and head movement. For the head, the ear keypoints are most relevant, presumably be-

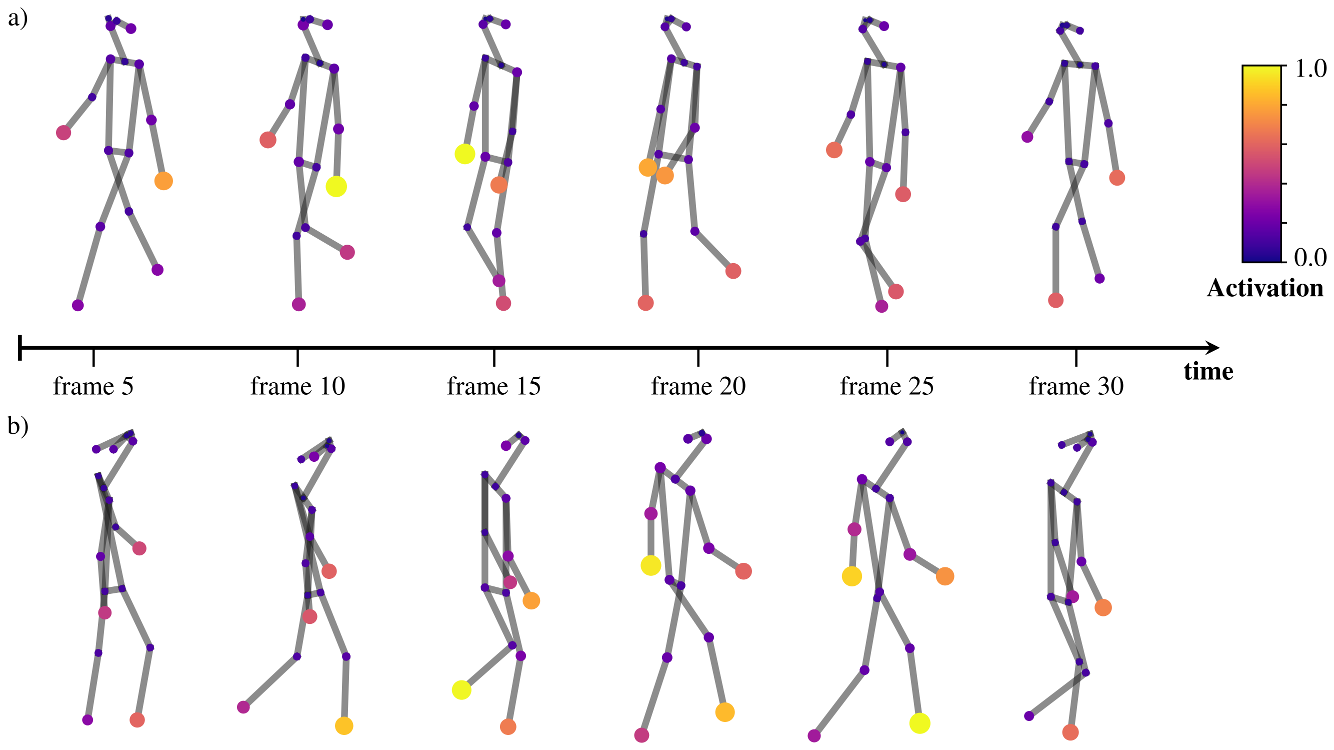

Figure 5. Activated joints of our proposed method on an example gait cycle from the OUMVLP-Pose test set. Higher activated joints are shown in a brighter color and in bigger scale. The view angle for a) is 60°and b) is 255°.

cause they are most outward and the most reliable keypoints detected from the back view. This observation is highly aligned with an early definition of gait "the motion of the living body [is] represented by a few bright spots describing the motions of the main joints" [11] and the intuitive understanding of gait, which is best captured by the outer limbs.

5. Conclusion

In this paper, we present a improved approach for skeleton-based gait recognition and introduce a multi-branch architecture for gait recognition. The architecture uses pre-computed temporal information divided into three branches: joints, motion, and bones. We archive impressive performance gains on OUMVLP-Pose, the largest skeleton-based gait dataset, with an improve of 3× over to the baseline. Compared to previous graph-based approaches, our training is much faster and much more stable. We are the first approach to publish results on both popular gait datasets and set the baseline for future gait recognition research.

In our ablation, we analyze the learned gait representation of the network and show our strong temporal modeling. Furthermore, our visualization of joint activations indicate the strong focus on moving body parts and the outermost joints. Indicating to focus on these joints for further research. Together these two observations confirm that our model captures *real* gait features, instead of performing a visual re-identification. Combined with the advantages of robust pose estimation, this allows for broader spectrum of applications of gait recognition. Our ablation indicates that gait recognition will improve with more accurate pose estimators. This closes the gap for bringing gait recognition from the lab to the real-world.

References

[1] Weizhi An, Shiqi Yu, Yasushi Makihara, Xinhui Wu, Chi Xu, Yang Yu, Rijun Liao, and Yasushi Yagi. Performance evaluation of model-based gait on multi-view very large population database with pose sequences. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2(4):421–430, 2020. 3, 4, 6, 7

[2] Maryam Babaee, Yue Zhu, Okan Köpüklü, Stefan Hörmann, and Gerhard Rigoll. Gait energy image restoration using generative adversarial networks. In *ICIP*, pages 2596–2600, 2019. 1, 2

[3] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. OpenPose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE TPAMI*, 2019. 2, 5

[4] Hanqing Chao, Yiwei He, Junping Zhang, and Jianfeng Feng. GaitSet: Regarding gait as a set for cross-view gait

recognition. In *AAAI*, volume 33, pages 8126–8133, 2019. 1, 2, 7

[5] Chao Fan, Yunjie Peng, Chunshui Cao, Xu Liu, Saihui Hou, Jiannan Chi, Yongzhen Huang, Qing Li, and Zhiqiang He. GaitPart: Temporal part-based model for gait recognition. In *CVPR*. IEEE, June 2020. 1, 2, 7

[6] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. Rmpe: Regional multi-person pose estimation. In *ICCV*, Oct 2017. 2, 5

[7] Yang Feng, Yuncheng Li, and Jiebo Luo. Learning effective gait features using lstm. In *ICPR*, pages 325–330, 2016. 1, 2

[8] Jinguang Han and Bir Bhanu. Individual recognition using gait energy image. *PAMI*, 28(2):316–322, 2005. 1, 2

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 4

[10] Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*, 2018. 5

[11] Gunnar Johansson. Visual perception of biological motion and a model for its analysis. *Perception & psychophysics*, 14(2):201–211, 1973. 2, 8

[12] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673, 2020. 5

[13] Liang Wang, Tieniu Tan, Huazhong Ning, and Weiming Hu. Silhouette analysis-based gait recognition for human identification. *PAMI*, 25(12):1505–1518, 2003. 1, 2

[14] Rijun Liao, Chunshui Cao, Edel B. Garcia, Shiqi Yu, and Yongzhen Huang. Pose-based temporal-spatial network (PTSN) for gait recognition with carrying and clothing variations. In Jie Zhou, Yunhong Wang, Zhenan Sun, Yong Xu, Linlin Shen, Jianjiang Feng, Shiguang Shan, Yu Qiao, Zhenhua Guo, and Shiqi Yu, editors, *Biometric Recognition*, pages 474–483, Cham, 2017. Springer International Publishing. 2, 5

[15] Rijun Liao, Shiqi Yu, Weizhi An, and Yongzhen Huang. A model-based gait recognition method with body pose and human prior knowledge. *PR*, 98:107069, 2020. 2, 5, 6, 7

[16] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014. 3

[17] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *CVPR*, pages 12026–12035, 2019. 3

[18] Kohei Shiraga, Yasushi Makihara, Daigo Muramatsu, Tomio Echigo, and Yasushi Yagi. Geinet: View-invariant gait recognition using a convolutional neural network. In *2016 international conference on biometrics (ICB)*, pages 1–8. IEEE, 2016. 7

[19] Leslie N Smith and Nicholay Topin. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial intelligence and machine learning for multi-domain operations applications*, volume 11006, page 1100612. International Society for Optics and Photonics, 2019. 5

[20] Anna Sokolova and Anton Konushin. Pose-based deep gait recognition. *IET Biometrics*, 8(2):134–143, 2018. 2

[21] Chunfeng Song, Yongzhen Huang, Yan Huang, Ning Jia, and Liang Wang. GaitNet: An end-to-end network for gait based human identification. *PR*, 96:106988, 2019. 1, 2, 7

[22] Yi-Fan Song, Zhang Zhang, Caifeng Shan, and Liang Wang. Stronger, faster and more explainable: A graph convolutional baseline for skeleton-based action recognition. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1625–1633, 2020. 3, 4

[23] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*. IEEE, June 2019. 2, 3, 4

[24] Noriko Takemura, Yasushi Makihara, Daigo Muramatsu, Tomio Echigo, and Yasushi Yagi. Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition. *IPSJ Transactions on Computer Vision and Applications*, 10(1):4, Feb 2018. 4

[25] Torben Teepe, Ali Khan, Johannes Gilg, Fabian Herzog, Stefan Hörmann, and Gerhard Rigoll. GaitGraph: Graph convolutional network for skeleton-based gait recognition. In *ICIP*, pages 2314–2318, 2021. 3, 6, 7

[26] Likai Wang and Jinyan Chen. A two-branch neural network for gait recognition. *arXiv preprint arXiv:2202.10645*, 2022. 3, 6, 7

[27] Thomas Wolf, Mohammadreza Babaee, and Gerhard Rigoll. Multi-view gait recognition using 3d convolutional neural networks. In *ICIP*, pages 4165–4169. IEEE, 2016. 1, 2

[28] Zifeng Wu, Yongzhen Huang, Liang Wang, Xiaogang Wang, and Tieniu Tan. A comprehensive study on cross-view gait based human identification with deep cnns. *IEEE TPAMI*, 39(2):209–226, 2017. 4

[29] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI*, 2018. 3

[30] Shiqi Yu, Daoliang Tan, and Tieniu Tan. A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In *ICPR*, volume 4, pages 441–444, 2006. 2, 3, 4

[31] Pengfei Zhang, Cuiling Lan, Wenjun Zeng, Junliang Xing, Jianru Xue, and Nanning Zheng. Semantics-guided neural networks for efficient skeleton-based human action recognition. In *CVPR*, pages 1112–1121, 2020. 3, 4

[32] Ziyuan Zhang, Luan Tran, Xi Yin, Yousef Atoum, Xiaoming Liu, Jian Wan, and Nanxin Wang. Gait recognition via disentangled representation learning. In *CVPR*, June 2019. 4

[33] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, pages 2921–2929, 2016. 7