GaitZero: Temporal Self-similarity for Unsupervised Gait Recognition

Ayush Gupta
Birla Institute of Technology and Science
Pilani, India

Alexander Matasa
University of Central Florida

Shruti Vyas
University of Central Florida

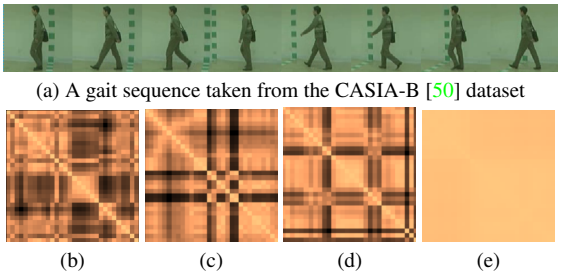
Yogesh S Rawat
University of Central Florida

Abstract

*In this work we focus on vision-based gait recognition. Most of the existing works utilize silhouettes or pose which either requires additional processing or specialized sensors. Moreover, these methods use a supervised approach where subject identities are required. We propose **GaitZero**, a novel **unsupervised** framework for learning gait signature from **RGB videos** which does not require **subject identities** across gait instances. Learning a gait signature directly from RGB videos without the use of labels presents two main challenges, 1) how to learn a meaningful gait signature in presence of appearance covariates, and 2) how to learn a discriminative signature without any labels. We propose to utilize **temporal self-similarity** to extract gait patterns from a video. Temporal self-similarity focuses on the **evolution of gait** and helps in ignoring appearance biases in RGB videos. GaitZero is trained using a **self-contrastive** loss formulation to learn a discriminative gait signature. The proposed **self-contrastive** objective utilizes a **negative sample** with similar appearance which further mitigates the effect of covariates. We demonstrate the effectiveness of the proposed approach on two different benchmark datasets, FVG and CASIA-B. GaitZero achieves ∼**80%** accuracy on FVG dataset without the use of subject identification which is comparable to recent **supervised** methods.*

1. Introduction

Gait represents a person's walking pattern and is one of the many identifying characteristics of humans. It has a wide range of applications in social security [6], authentication [29, 33], and tracking [3]. A subject's gait information can be captured using various wearable sensors, such as pressure sensors [36] and accelerometers [15, 28], however, use of such sensors require subjects' consent and cooperation which limits its applications. Vision-based gait analysis overcomes this limitation due to its non-invasive nature where the gait sequence can be captured from a distance.



(a) A gait sequence taken from the CASIA-B [50] dataset

(b) (c) (d) (e)

Figure 1. Visualization of temporal self-similarity matrices (TSM) inferred by *GaitZero*. Figure 1b: full body, 1c arms region, 1d: legs, and 1e: head. We observe different repetitive patterns for arms and legs with three peaks. The head region shows no pattern due to minimal movement during walking.

There have been several advances in vision-based gait recognition using deep networks with encouraging performance [12, 42]. However, most of these methods utilize a *supervised* approach where subject identification is required for each gait instance. These existing methods can be mainly categorized into two different types based on the modality used; *pose-based* [2, 10, 32, 37], and *silhouette-based* [19, 25, 44]. The use of silhouettes and pose modalities is a positive step towards protection of the privacy of individuals, but it comes with several limitations and challenges. First, converting a gait sequence to a silhouette or pose requires *additional preprocessing*; second, the performance of any method based on these modalities will be limited by the effectiveness of the preprocessing step; and finally, although pose modality can be captured using *specialized sensors*, such as Kinect [51], these sensors can be erroneous in uncontrolled environments and a requirement of such sensors limits the practical application of these approaches. Moreover, *supervised* training of deep networks requires a large amount of annotations which can be very challenging and expensive to collect.

In this work we aim at addressing these limitations and

explore the use of *RGB video* for gait recognition in an *unsupervised* setting. We propose ***GaitZero***, which can learn to extract a meaningful gait signature directly from RGB frames without any subject identification on gait instances. The use of *RGB videos* for learning gait signature in an *unsupervised* approach brings several challenges. First, it can be difficult to learn a discriminative gait signature due to lack of annotations on gait instances. Second, the use of RGB frames adds a lot of covariates such as appearance, clothing, background, etc., which are not important for learning a gait signature. Moreover, lack of annotations makes it harder to ignore such covariates.

GaitZero is our attempt to address some of these challenges and learn an effective gait signature without subject identities. We utilize *temporal self-similarity* to learn the evolution of gait and capture how the body posture changes with time. The use of self-similarity helps in ignoring covariate features as it learns from frame similarities instead of visual features. Contrastive learning has recently shown impressive performance in learning discriminative features without using any labels [9]. Inspired by this, we propose a *self-contrastive* objective for GaitZero which helps in learning a *discriminative* gait signature. We utilize a sample with randomly shuffled frames from a gait sequence as additional negative example which enables GaitZero to ignore appearance biases and focus more on the temporal aspect of gait.

GaitZero is trained end-to-end with the help of self-contrastive loss without any subject labels. We evaluate our approach on two different benchmark datasets including FVG [52] and CASIA-B [50]. We also demonstrate the *robustness* of GaitZero against domain shifts across datasets, where we found that a *single trained model* can perform well on both CASIA-B and FVG datasets. GaitZero is very effective for *knowledge transfer* and it is found that it easily generalizes to a target domain with very *limited unlabelled samples*. We make the following contributions in this work,

- We propose *GaitZero*, an *unsupervised* method for gait recognition using *RGB videos*. This is the first work addressing this problem to the best of our knowledge.
- We propose a novel *TSM Pyramid* architecture which effectively captures evolution of gait to learn a meaningful gait signature.
- We introduce a *self-contrastive* loss for learning a discriminative signature and which enforces the model to ignore *appearance covariates* in RGB videos.

2. Related work

The two main modalities used for vision-based gait recognition are silhouettes [19, 24, 25, 38, 49] and pose [7, 20, 23, 27, 31, 48]. Most existing works in gait recognition typically rely on sequences of silhouettes for learning a gait signature [8, 12, 19, 24, 25, 38, 45, 45, 49]. The temporal nature of gait demands the injection of time-relevant feature information. for which 3D convolutions have been used as shown in [19, 24, 25, 45], and individual frame-level features have been aggregated as shown in [8, 12, 38, 49]. LSTMs [8, 38], transformers [49] and 1D convolutions [12] have proven to be effective for such aggregation. Further, works like [8, 12, 13] show that splitting silhouette frames into small sections can improve signature extraction. Silhouettes ensure that the subjects' privacy is retained. In addition, all the biases and covariates present are easily ignored while learning the gait signature. However, obtaining silhouettes from RGB videos requires additional processing which is not always desirable.

In several works, pose has also been utilized to learn a gait signature [7, 10, 11, 20, 21, 23, 27, 31, 35, 48]. Most of these works assume the availability of accurate pose, while [20, 21] adopt end to end approaches with pose features as their bottleneck. However, these methods also face the existing issue of feature obfuscation created by the scenic covariates. Relying on accurate pose-estimation algorithms introduces another set of challenges while using pose for gait recognition, introducing additional computations without significant gains in performance.

Focusing on real-world application, [52] has attempted to learn gait signatures utilizing RGB videos. Like prior works, this method also extracts frame-level features to develop a signature. We argue that using frame level features is not effective as motion information is not taken into account - which is very crucial for a gait signature.

In this work, we utilize a raw RGB video stream to learn a gait signature which does not require any processing to obtain silhouettes as an additional step. Moreover, all the existing works assume the availability of subject identities. Efforts have been made to learn gait signatures via semi-supervised learning but they require a pre-training step in addition to labels for fine tuning [22, 26, 31]. In contrast, we propose an unsupervised method, which does not require any finetuning using labels. Furthermore, the lack of pre-processing allows for our architecture to be directly applied to many surveillance applications.

3. Proposed approach

Given a video V_i representing a gait sequence for any subject S_i, our goal is to extract a gait signature γ_i such that it is similar to the signature extracted from other gait sequences of the same subject S_i. We randomly sample n contiguous frames to get a segment $V_i^s = \{v_1, v_2, ..., v_n\}$. GaitZero takes these sequential frames as input and provides a gait signature $\gamma = \mathcal{F}(V)$ as output where $\mathcal{F}$ represents the GaitZero model. It consists of three main components, 1) *visual encoder* ($\mathcal{F}_v$), 2) *temporal self-similarity module* ($\mathcal{F}_s$), and 3) *gait evolution encoder* ($\mathcal{F}_e$).

The visual encoder $\mathcal{F}_v$ takes a video V with n

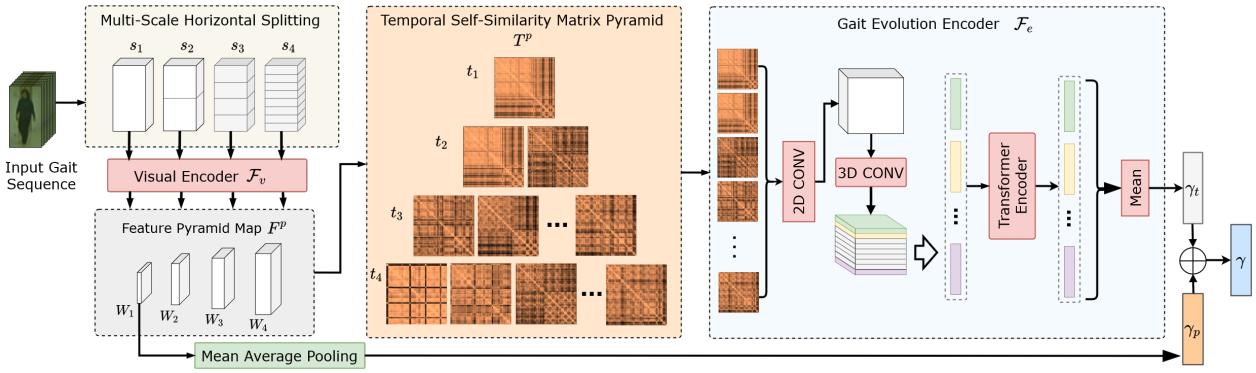


Figure 2. Overview of GaitZero. It takes RGB gait sequence as input and provide a gait signature γ. First the visual encode $\mathcal{F}_v$ extracts frame-wise embeddings for the input video. These embeddings are extracted at multiple scales to build a feature pyramid. This feature pyramid is used further to construct a TSM pyramid and extract temporal gait signature γ_t with the help of gait evolution encoder $\mathcal{F}_e$. The temporal gait signature is aggregated with the posture gait signature γ_p to obtain the final gait signature γ.

frames as input and generates per-frame embeddings $e = [e_1, e_2, ..., e_n]$. These per-frame embeddings are used by the temporal self-similarity module $\mathcal{F}_s$ to obtain a self-similarity matrix M^s. This self-similarity matrix M^s is produced at multiple scales with varying number of horizontal segments in the input video frames. The temporal self-similarity module $\mathcal{F}_s$ finally provides a TSM Pyramid T^p which consists of self-similarity matrices M^s for all the segments at varying scales. The TSM Pyramid T^p is fed to the gait evolution encoder $\mathcal{F}_e$ which outputs the embeddings γ_t encoding the temporal aspect of gait. In addition, the per-frame embeddings e are aggregated together to get posture encodings γ_p and both these embeddings, γ_t and γ_p are combined together to obtain a gait signature γ. An overview of the complete architecture is shown in Figure 2.

3.1. Visual encoder

We extract independent embeddings for each frame in the input video which allows us to compute the temporal self-similarity for gait evolution. The visual encoder $\mathcal{F}_v$ comprises of a ResNet18 [18] model which provides 2D conv features for each frame v_i of the input video V. The output of the visual encoder is a sequence of frame-wise embeddings $e = [e_1, e_2, ..., e_n]$ which is sent to the temporal self-similarity module. To obtain the embeddings for visual posture γ_p for the gait signature, we perform mean-pooling, augmenting the temporal information present in the TSM Pyramid. The mean-pooling of frame-wise features is inspired by Gait Energy Images, GEIs [16], which have proven to be effective in gait recognition.

3.2. Temporal self-similarity

Temporal self-similarity matrices (TSMs) have shown to be a promising for vision based tasks with repetitive patterns [43]. Also, self-similarity has been found effective in prior classical methods for gait analysis [4]. Moreover, TSMs are easily interpretable by humans, which can give important insights about the input sequence. Due to its repetitive nature, we propose to model gait using TSMs. Given a sequence of frame-wise embeddings, a TSM can be constructed by computing the similarity of the embedding of a frame with the embeddings of all other frames.

We use the frame-wise latent embeddings e to construct a self-similarity matrix by M^s by computing all pairwise similarities. First, we send the frame-wise visual embeddings e through a 3D convolution layer to supplement the network with temporal information before constructing the TSM Pyramid. This helps in capturing local temporal information along with frame-wise visual features. The temporal context captures short-term motions [43, 47] and allows the network to differentiate between visually similar frames with different motion. For example, a leg might be moving up or down while walking in a gait sequence. The embeddings after 3D convolution are passed to a max-pooling layer to reduce the dimensionality and obtain frame-wise sequence of embeddings $\hat{e} = [\hat{e}_1, \hat{e}_2, ..., \hat{e}_n]$.

Similarity between two different frames v_i and v_j is computed as $M^s_{i,j} = f(\hat{e}_i, \hat{e}_j)$ using a similarity function $f(.)$. We use euclidean distance to compute the similarity between two frames, $f(i, j) = -||i - j||$.

Feature pyramid map Different body parts follow a unique motion pattern when a person walks. These patterns play an important role independently in defining a gait signature along with the whole body movement. This aspect has been found to be very effective in prior-works for person re-identification [14] and even gait recognition [12]. Motivated by this, we propose to utilize the visual embeddings at different scales where we divide the input video frames into horizontal segments. We take the input video sequence

$V = \{v_1, v_2, ..., v_n\}$ with n frames and split each frame v_i horizontally at C different scales. For each scale $c \in [1, C]$, we obtain 2^{c-1} different segments and this results in a set of horizontal slices L for each frame where,

$$L = \{(s_{i,j}|j \in \{1...2^{i-1}\})|i \in \{1...C\}\}.$$

Here $s_{i,j}$ is the j^{th} slice at the i^{th} scale. Each slice $s_{i,j}$ is part of the video V and consists of n frames. These slices are passed through the visual encoder $\mathcal{F}_v$ which is shared across all the slices and provides a feature pyramid map,

$$F^p = \{(w_{i,j}|j \in \{1...2^{i-1}\})|i \in \{1...C\}\}.$$

where $w_{i,j}$ is feature vector of the j^{th} slice at the i^{th} scale.

TSM pyramid Using the feature pyramid map F^p, we construct a TSM pyramid to encode the evolution of pose at different scales. Each scale W_c in F^p can be given by

$$W_c = \{w_{c,1}, ..., w_{c,2^{c-1}}\},$$

where each $w_{c,j}$ consists of n embeddings, one for each frame. The embeddings $w_{c,j}$ in the feature pyramid map are fed to the temporal self-similarity module $\mathcal{F}_s$ which is shared between all embeddings across all the scales. Computing self-similarity on the frames for these embeddings gives 2^{c-1} self-similarity matrices

$$t_c = \{t_{c,1}, ..., t_{c,2^{c-1}}\},$$

where t_c is a set containing 2^{c-1} number of $n \times n$ self-similarity matrices, at the c^{th} scale. The complete TSM Pyramid $T^p = \{t_1, ..., t_C\}$ is computed by stacking t_c's for all the C scales.

3.3. Gait evolution encoder

The TSM pyramid T^p is then used to obtain temporal embeddings γ_t for the gait signature. The gait evolution encoder $\mathcal{F}_e$ takes the self-similarity matrix M^s and learns a temporal embedding for each matrix in the TSM Pyramid T^p. The encoder $\mathcal{F}_e$ is shared across all the matrices in the pyramid and the obtained embeddings are concatenated to generate the temporal embedding γ_t.

The self-similarity matrix M^s consists of $n \times n$ similarity values between each pair of frames in the input video. The self-similarity matrix is first passed through a 2D convolution layer to transform the similarity values to latent features. A feature vector for each frame is obtained by combining all latent features in one row of M^s, representing the similarity of that frame with all other frames. The features for all n frames are converted to a sequence and passed through a transformer encoder [41]. This architecture is inspired by [43] where a transformer model was found effective for encoding self-similarity matrices. The output sequence is then averaged over all positions to get the final embedding of the input matrix M^s.

3.4. Contrastive objective

We rely on contrastive learning [9] to train the proposed model. In a traditional contrastive objective, N examples are randomly sampled, and a loss is computed on pairs of augmented examples with a total of $2N$ data points. For each positive pair, all other $2(N-1)$ samples are treated as negative examples. In this traditional formulation the negative examples always come from different samples in the mini-batch. We propose a *self-contrastive* objective where we also sample negative examples from the positive pairs.

We propose random shuffling of frames for sampling a negative example for any video. The random shuffling will have a different sequence of frames but it will still have similar appearance in all the frames. However, the gait evolution pattern will be different. Treating this as a negative example will enforce the network to ignore static appearance features for learning a gait signature.

We introduce two such negative samples, resulting in a total of $4N$ data points in any mini-batch with N examples. The self-contrastive loss for a pair of positive examples (i, j) is defined as,

$$l_{i,j} = -\log \frac{\exp(\text{sim}(\gamma_i, \gamma_j)/\tau)}{\sum_{k=1}^{4N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(\gamma_i, \gamma_k)/\tau)}, \quad (1)$$

where γ_i and γ_j represent gait embeddings for a positive pair, γ_i and γ_k represent a negative pair, $\mathbb{1}_{[k \neq i]} \in \{0, 1\}$ is an indicator with value 1 iff $k \neq i$, τ is a temperature parameter, sim is the cosine similarity between a pair of embeddings. The final loss is computed only for the positive pairs, both (i, j) and (j, i), ignoring the additional negative examples as there will be no positive pair for such samples.

4. Experiments

Datasets We perform our experiments on two different benchmark datasets, Frontal View Gait (FVG) [52] and CASIA-B [50]. **FVG** is a recently collected RGB gait dataset [52]. It comprises of 226 subjects, each with 4 different walking conditions and 3 different viewpoints. The dataset is captured keeping frontal view angles in mind and the videos are captured from $0°$and $45°$on both the sides. We follow the evaluation protocol from [52] and use the first 136 subjects for training and the remaining 90 subjects for evaluation. The normal walking video from the frontal view is used as the gallery, and the remaining videos are used as probes. **CASIA-B** is one of the most popular datasets available for gait recognition [50]. Traditionally, this dataset has been used in the form of pre-processed silhouettes, but we use the original RGB videos. It comprises of 124 subjects, where the first 74 subjects are used for training and the remaining 50 subjects for evaluation. We follow the existing protocol where the first four normal walking sequences are used as galleries and the remaining as probes.

Supervision mode	Variation	WS		CB		CL		CBG		ALL	
	TDR@FAR	1%	5%	1%	5%	1%	5%	1%	5%	1%	5%
Supervised	PE-LSTM [52]	79.3	87.3	59.1	78.6	55.4	67.5	61.6	72.2	65.4	74.1
	GEI [17]	9.4	19.5	6.1	12.5	5.7	13.2	6.3	16.7	5.8	16.1
	GEINet [34]	15.5	35.2	11.8	24.7	6.5	16.7	17.3	35.2	13.0	29.2
	DCNN [1]	11.0	23.6	5.7	12.7	7.0	15.9	8.1	20.9	7.9	19.0
	LB [46]	53.4	73.1	23.1	50.3	23.2	38.5	56.1	74.3	40.7	61.6
	GaitNet [52]	91.8	96.6	74.2	85.1	56.8	72.0	92.3	97.0	81.2	87.8
Unsupervised	R2+1D [39] Contrastive [9]	72.3	85.8	69.7	78.8	30.0	45.5	70.0	84.0	61.2	75.0
	R2+1D [39] Triplet [40]	74.6	86.1	69.7	75.8	24.9	44.7	77.2	88.6	63.9	76.2
	ResNet [18] Contrastive [9]	73.3	89.2	75.8	84.9	37.2	53.6	73.0	83.2	65.4	78.6
	ResNet [18] Triplet [40]	81.6	92.1	78.8	87.9	36.8	48.6	80.2	90.3	70.9	81.1
	GaitZero (ours)	**87.1**	**95.1**	**97.0**	**100.0**	**53.6**	**69.2**	**83.4**	**93.7**	**78.1**	**87.6**

Table 1. A comparison of performance on the FVG [52] dataset, showing supervised methods and unsupervised baselines along with GaitZero. Metrics reported here are True Detection Rate at 1% and 5% False Acceptance Rate. The best unsupervised scores are written in bold, and the best supervised scores are underlined. It can be seen that GaitZero is able to achieve a comparable performance with supervised approaches without the use of any labels.

Probes	Methods	Probe view											Mean
		0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	
NM	R2+1D [39] Contrastive [9]	66.7	66.8	58.7	72.0	66.4	68.0	68.5	71.7	56.8	65.4	56.2	68.3
	R2+1D [39] Triplet [40]	59.2	63.0	42.6	74.1	55.0	66.4	60.3	66.4	58.5	63.1	44.7	63.0
	ResNet [18] Contrastive [9]	57.6	61.1	62.3	75.4	62.1	69.4	68.8	69.3	65.4	66.8	62.1	68.6
	ResNet [18] Triplet [40]	**74.3**	54.6	57.3	**78.6**	65.8	72.6	62.0	74.3	63.0	67.3	62.6	69.6
	GaitZero (ours)	69.2	**68.6**	**70.6**	69.8	**78.4**	**76.4**	**75.8**	**78.4**	**69.5**	**70.6**	**74.3**	**72.9**
BG	R2+1D [39] Contrastive [9]	53.9	55.5	45.8	58.8	51.7	53.1	48.5	56.9	31.7	48.2	34.5	52.2
	R2+1D [39] Triplet [40]	49.6	53.0	38.4	63.6	39.6	57.2	43.9	58.0	41.5	48.3	26.4	51.4
	ResNet [18] Contrastive [9]	54.4	56.7	60.0	66.4	45.0	61.6	45.1	58.5	52.7	49.8	42.0	57.0
	ResNet [18] Triplet [40]	**67.7**	49.9	52.4	**69.8**	51.4	57.2	43.4	63.0	46.5	54.7	50.3	58.4
	GaitZero (ours)	64.8	**62.3**	**65.7**	65.5	**70.3**	**69.2**	**66.9**	**68.4**	**59.1**	**62.8**	**69.3**	**65.8**
CL	R2+1D [39] Contrastive [9]	9.8	12.0	10.7	9.5	9.9	11.2	8.8	10.7	7.7	10.5	9.0	10.3
	R2+1D [39] Triplet [40]	8.9	9.3	8.6	12.6	10.0	12.3	8.1	10.8	7.0	8.6	5.6	9.7
	ResNet [18] Contrastive [9]	13.7	13.3	12.4	13.2	9.8	11.7	9.9	14.0	9.3	11.8	8.9	12.0
	ResNet [18] Triplet [40]	12.6	9.2	10.6	10.7	8.1	12.6	9.3	11.6	**9.7**	8.0	7.7	10.2
	GaitZero (ours)	**19.1**	**20.4**	**18.5**	**14.0**	**18.3**	**13.2**	**12.2**	**14.3**	9.3	**14.1**	**15.1**	**15.3**

Table 2. A comparison of performance on the CASIA-B [50] dataset with unsupervised baselines along with GaitZero, excluding identical view cases. Average retrieval accuracy for each probe angle is reported for three different walking conditions.

4.1. Implementation details

We use PyTorch [30] to implement our approach and train our model with Adam optimizer with a batch size of 16 and a learning rate of 1e-4. We use a resolution of 64x32 with n = 32 frames in a video. To construct the feature pyramid map, we split the input video horizontally at C = 4 scales. The ResNet18 model used in the visual encoder is initialized using pretrained weights from ImageNet. We extract the visual features from the third block of ResNet18 followed by average pooling. For our gait evolution encoder transformer, we use 8 heads with 512 feedforward dimensions. The transformer outputs a 256-size embedding for each TSM in the pyramid. We performed person detection using YOLO-v4 [5] on CASIA-B videos to extract the cropped gait sequences. For FVG, we use the preprocessed cropped videos which are provided with the dataset.

Augmentations for contrastive learning We utilize the following set of augmentations to obtain positive pairs for contrastive loss; 1) walking speed augmentation using varying sampling rate to take into account variations in walking speed, 2) gamma augmentation for lighting variations, 3) random horizontal flipping, 4) random cropping of frames, and 5) random starting phase of gait by selecting a random starting frame. The positive pairs for contrastive loss are obtained by randomly applying these augmentations on a given instance of gait sequence. A gait instance can be captured from multiple viewpoints and therefore the positive pairs can be from same or different viewpoints. This also helps in learning a view-invariant gait signature.

4.2. Inference and evaluation metrics

Since we use n frames from a sequence to compute the gait signature, we split the video into segments of n frames each during inference. The mean of the embeddings of each of these segments is considered as the gait signature for the

Variation	WS	CB	CL	CBG	All
R2+1D [39] Contrastive [9]	68.3	57.6	24.9	65.4	56.6
R2+1D [39] Triplet [40]	67.7	60.6	22.8	73.0	58.4
ResNet18 [18] Contrastive [9]	75.0	72.8	37.2	73.5	65.5
ResNet18 [18] Triplet [40]	77.9	69.7	34.6	76.0	67.2
GaitZero (ours)	**89.4**	**97**	**50.6**	**85.2**	**79.5**

Table 3. Accuracy of GaitZero and other unsupervised baselines on the FVG [52] dataset.

subject. The signature of the probes is compared with all the galleries, and the closest match gives the predicted ID of the subject. We use the standard evaluation metrics and protocols for both CASIA-B [50] and FVG [52]. For CASIA-B, we report rank-1 retrieval accuracy and for FVG, true positive rate@1% and 5% false positive rate is used. In addition, we also report rank-1 retrieval accuracy for FVG.

4.3. Baselines

This is the first work focusing on unsupervised gait recognition using RGB videos to best of our knowledge. Therefore, we developed some baseline approaches using existing methods to validate the effectiveness of GaitZero. We developed four different baselines with the help of two different backbones and two unsupervised loss functions. As a backbone, we consider R(2+1)D [39] and ResNet18 [18] architectures. R(2+1)D is one of the best models for learning spatio-temporal features and ResNet18 is one of the best model for extracting spatial features using 2D convolutions. We use R(2+1)D model to directly extract the gait signature using a given input video. In the case of ResNet18, we extract features for each frame and then average them to obtain a gait signature. We utilize two different objective functions to train these models, triplet margin loss [40] and contrastive loss [9].

4.4. Results

FVG dataset The FVG dataset has five different protocols, Walking speed (WS), Carrying Bag (CB), Clothing (CL), Cluttered Background (CBG) and 'All'. The performance of GaitZero on these protocols is shown in Table 1. The clothing protocol is found to be the toughest for our model, since the gait can change significantly with a person wearing a jacket. However, we find that the model performs well on all other conditions, as well as the 'all' protocol.

CASIA-B dataset The CASIA-B dataset consists of three different conditions, Normal Walking (NM), Baggage (BG) and Clothing (CL). There are two different protocols to create the gallery set for a probe, one excluding identical view cases are the other including identical view cases. The corresponding evaluation scores are shown in Table 2 and Table 4 respectively. We observe that GaitZero performs reasonably well on Normal Walking and Baggage conditions. Table 2 also shows the effect of different probe angles on GaitZero. In general, best performance is achieved when probe angles are near to 90°since complete gait information can be captured from such viewpoints, as opposed to extreme angles. Further, including identical viewpoints in the gallery set gives an expected improvement in performance as shown in Table 4, because the gait sequence is inherently similar and the model does not have to filter out view information in the final gait signature.

Method	Probe condition		
	NM	BG	CL
R2+1D [39] Contrastive [9]	68.3	52.2	10.3
R2+1D [39] Triplet [40]	63.0	51.4	9.7
ResNet [18] Contrastive [9]	68.6	57.0	12.0
ResNet [18] Triplet [40]	69.6	58.4	10.2
GaitZero (ours)	**75.3**	**68.6**	**16.2**

Table 4. Performance comparison on the CASIA-B [50] dataset, showing unsupervised baselines along with GaitZero, including identical view cases. Mean scores for all probe views are shown.

Comparisons The comparison of GaitZero with existing approaches and baselines for FVG is shown in Table 1 and Table 3. GaitZero performs better than these baselines on all protocols, even achieving a perfect score in the Carrying Bag condition at 5% False Positive Rate. We also observe that ResNet backbones are performing better than R2+1D baselines, but as shown in Sec 4.6, they are focusing more on the appearance features. We also compare GaitZero with the exisiting *supervised* approaches on FVG. It can be observed that GaitZero achieves a comparable performance with supervised methods, even without the knowledge of the subjects' identity, which is very impressive.

The comparison of GaitZero with the baselines on CASIA-B is shown in Table 2. CASIA-B is a more challenging dataset due to large variations in viewpoints. GaitZero achieves a notable increase over the unsupervised baselines, with the increase being most significant for the Baggage protocol. Since most existing approaches on CASIA-B use either silhouettes or pose modalities, we do not compare with the other supervised approaches. The results for these supervised approaches are provided in the supplementary material for reference.

4.5. Ablation study

Effect of TSM pyramid To study the effect of the TSM pyramid on GaitZero, we remove the pyramid and perform two experiments: 1) using a single TSM instead of the pyramid, and 2) slicing the input video uniformly to get C slices, instead of building a pyramid using multiple scales. The performance reduces in both these experiments, giving 75%

Description	Signature		Loss		Accuracy
	γ_t	γ_p	SC	TC	
γ_p + SC		✓	✓		67.0
γ_t + SC	✓		✓		62.2
γ_t + γ_p + TC	✓	✓		✓	77.4
GaitZero (γ_t + γ_p+ SC)	✓	✓	✓		79.5

Table 5. Effect of different components on GaitZero. SC denotes proposed self-contrastive loss, TC is the traditional contrastive loss [9]. γ_t and γ_p are the temporal and pose signatures.

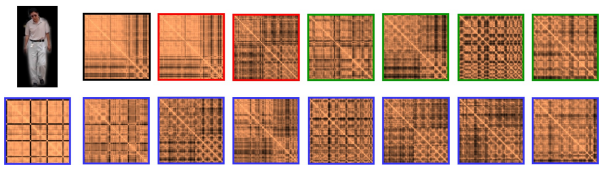


Figure 3. TSMs for an extended video sequence. Same border denotes same scale. The difference between local and global gait evolution can be clearly seen.

accuracy in the single TSM case and 76% in the C uniform slices case. This demonstrates that the TSM pyramid is helpful in extracting a better signature from the input as it enables the model to focus on multiple spatial scales at once, capturing both global and local gait evolution.

Effect of self-contrastive learning In this experiment, we analyse the effect of introducing additional negative samples during training. To do this, we compare the performance of GaitZero trained using the proposed self-contrastive loss formulation with one trained with the traditional contrastive loss [9]. The results are summarized in Table 5. It can be seen that self-contrastive learning indeed helps the model. Treating shuffled frames as additional negatives can be seen as a form of data augmentation, generating more data for the model to learn from. Additionally, it also helps in ignoring covariates like clothing and face which are irrelevant to gait.

Effect of Pose Signature In this experiment, we try to analyze the importance of γ_p, the pose signature. To this extent, we discard the pose encoder from the architecture, and just use γ_t as the overall gait signature. We observe a significant drop in performance, as shown in Table 5. This shows that the pose signature γ_p supplements the temporal signature γ_t to obtain a better discriminative gait signature.

Effect of Temporal signature Next, we observe the effect of removing γ_t on the performance of GaitZero. Without γ_t, the model will only focus on the appearance aspect of gait. Static information has historically proven to be useful as GEIs [16] in gait recognition. However, as mentioned in Table 5, removing the temporal signature γ_t drastically reduces performance, which shows that capturing temporal information of gait helps in getting a better gait signature.

4.6. Discussion and analysis

Analysing TSM Self-similarity matrices are interpretable features in our model. The correlation between the different body parts of the subject and the corresponding self-similarity matrices is illustrated in Fig. 1. Some sampled frames from the video are shown in Fig 1a. The TSM for the horizontal slice at the arms' level is shown in Fig 1c. In the video, only the left arm is visible, and the TSM of the arms shows a distinct boundary at that point, indicating the start of a new phase in the gait cycle. Similarly, Fig 1d indicates three such boundaries, and the leg movement indeed reaches its peak three times. The TSM for the complete frame without slicing is illustrated in Fig. 1b, showing the repetition pattern of the body as a whole. Interestingly, Fig 1e almost shows no variation, as the head displays minimal movement in the gait cycle.

To further visualize the features learnt by the visual encoder $\mathcal{F}_v$, we take an entire video sequence from FVG and compute TSMs for a long video. The resulting pyramid is illustrated in Fig. 3. Interesting patterns can be observed in these TSMs, with distinct boundaries between different phases of gait cycles. The stark contrast between TSMs of different scales also shows the difference between global gait evolution and local pose changes.

Generalization to silhouettes We also conduct an experiment to check the generalizability of our model to silhouette inputs. We find encouraging results, with the model achieving 67.2% (Excluding identical view cases) accuracy on the Normal Walking condition on CASIA-B silhouettes, which shows that TSMs can be used even for silhouette inputs. Detailed results for this experiment are provided in the supplementary material.

Robustness Some existing approaches for gait recognition need different architectures for different datasets [8,12] which limits their generalizability practical scenarios. We solve this problem by combining both FVG and CASIA-B datasets and training a new model on this mixed dataset. It was observed that a single model, with the same weights, can perform well on both the datasets, achieving 75.2% on FVG's 'all' protocol and 72.3% on CASIA-B's Normal Walking condition. This demonstrates the method's robustness to domain shifts across datasets.

We analyze GaitZero for transfer learning, where we used CASIA-B for pre-training as it is much larger in size as compared to FVG. Then, we fine tune the model using limited training data from FVG. It is important to note that

Portion of FVG data	Pretrained	Scratch
0%	67.8	-
5%	77.8	64
10%	78.6	69.4
15%	79.0	73.2
20%	**82.1**	74.3

Table 6. Performance of GaitZero when pre-trained on CASIA-B and fine tuned on a portion of FVG, compared with training from scratch. The metric reported here is TPR@1% FPR. GaitZero is able to surpass the best supervised performance on FVG [52] without using any annotations.

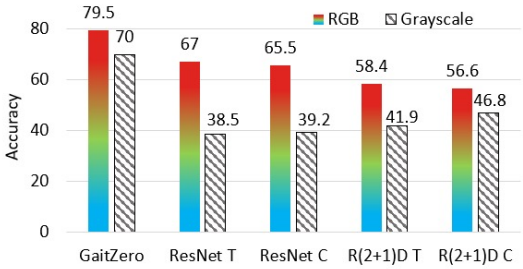


Figure 4. Comparison of GaitZero with different baselines on FVG, evaluated on RGB and Grayscale data. T stands for Triplet loss [40], and C stands for Contrastive loss [9]. It can be inferred that the ResNet baselines are mostly focusing on appearance and color features, while GaitZero is looking at the gait of the subject.

no annotations are used during both pre-training and fine-tuning. As shown in Table 6, GaitZero is able to outperform current supervised methods using this approach, *without using any labels*. This demonstrates the generalization capability of GaitZero across datasets.

Covariate analysis For a deep neural network working on RGB data, it can be difficult to ignore color and appearance biases. In this experiment, we analyze the effects these covariates can have on the network. To test the effect of facial features on GaitZero, we evaluate the model using testing sequences with the face of the subjects removed. As shown in Fig. 5, removing the face does not affect performance. Next, to test the effect of the color of clothing, we evaluate our model and the baselines on grayscale videos. The results are shown in Fig 4. While the ResNet baselines are mostly focusing on the appearance and color features to obtain a high accuracy, GaitZero is indeed extracting the gait signature from RGB videos.

Occlusion We conduct experiments to analyze the effect of occlusion on our method. The results are shown in Fig 5. The input video is cut into four uniform slices, and various slices are discarded while testing. We observe that GaitZero is able to perform well even if some body parts are missing

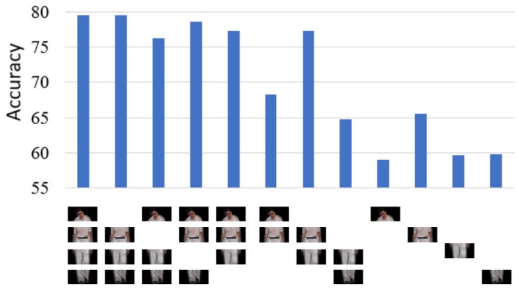


Figure 5. Effect of occlusion on GaitZero during inference. The horizontal axis shows which slices of the input data were used by the model. It can be observed that removing any one horizontal slice does not have a major effect, and the head does not have a significant contribution in the gait signature.

in the input gait sequence, meaning that no single horizontal segment is a critical factor, indicating the method's robustness to occlusion.

4.7. Limitations

Although GaitZero can provide very good performance without the use of annotations, there are some limitations of GaitZero. We observe that the method performs reasonably well in normal conditions and is comparable to supervised approaches, but it struggles with varying clothing conditions. The lack of subject labels is the main reason for this, because while training GaitZero has no knowledge of cross-condition embeddings belonging to the same subject.

4.8. Ethics

Gait recognition is a very important problem in computer vision research. An improper use of this technology can raise privacy issues in the society. In this work we are only using datasets which are officially authorized for research purpose. These datasets are collected, released and used with the consent of participating subjects. Moreover, the proposed method does not require subject identification which further minimizes the privacy concerns.

5. Conclusion

In this work we propose *GaitZero*, an *unsupervised* approach for learning gait signature from *RGB videos*. We utilize *temporal self-similarity* and propose a *TSM pyramid* to extract effective gait features. TSM pyramid models *temporal gait evolution* and operates at different *scales* on the input video. GaitZero is trained using a novel *self-contrastive* objective which helps in learning a discriminative feature and ignores *appearance covariates*. We evaluated GaitZero on two different public benchmark datasets, including CASIA-B and FVG. We observe that GaitZero can achieve *promising performance* which is close to *supervised* state-of-the-art methods on the FVG dataset. This

could be an interesting research direction for gait recognition as it also addresses the issue of *privacy* which is present when we have to annotate a dataset.

References

[1] Munif Alotaibi and Ausif Mahmood. Improved gait recognition based on specialized deep convolutional neural network. *Comput. Vis. Image Underst.*, 164:103–110, 2017. 5

[2] Virginia Ortiz Andersson and Ricardo Matsumura Araujo. Person identification using anthropometric and gait data from kinect sensor. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015. 1

[3] Maryam Babaee, Gerhard Rigoll, and Mohammadreza Babaee. Joint tracking and gait recognition of multiple people in video. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 2592–2596. IEEE, 2017. 1

[4] Chiraz Benabdelkader, Ross Cutler, and Larry Davis. Gait recognition using image self-similarity. *EURASIP Journal on Advances in Signal Processing*, 2004, 04 2004. 3

[5] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020. 5

[6] Ruud M Bolle, Jonathan H Connell, Sharath Pankanti, Nalini K Ratha, and Andrew W Senior. *Guide to biometrics*. Springer Science & Business Media, 2013. 1

[7] Ning Cai, Shiling Feng, Qing Gui, Lei Zhao, Huadong Pan, Jun Yin, and Bin Lin. Hybrid silhouette-skeleton body representation for gait recognition. In *2021 13th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*, pages 216–220, 2021. 2

[8] Hanqing Chao, Yiwei He, Junping Zhang, and Jianfeng Feng. Gaitset: Regarding gait as a set for cross-view gait recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8126–8133, 2019. 2, 7

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2, 4, 5, 6, 7, 8

[10] Adrian Cosma and Ion Emilian Radoi. Wildgait: Learning gait representations from raw surveillance streams. *ArXiv*, abs/2105.05528, 2021. 1, 2

[11] Vitor C de Lima, Victor HC Melo, and William R Schwartz. Simple and efficient pose-based gait recognition method for challenging environments. *Pattern Analysis and Applications*, 24(2):497–507, 2021. 2

[12] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14225–14233, 2020. 1, 2, 3, 7

[13] Yang Fu, Yunchao Wei, Yuqian Zhou, Honghui Shi, Gao Huang, Xinchao Wang, Zhiqiang Yao, and Thomas Huang. Horizontal pyramid matching for person re-identification. *AAAI*, 2019. 2

[14] Yang Fu, Yunchao Wei, Yuqian Zhou, Honghui Shi, Gao Huang, Xinchao Wang, Zhiqiang Yao, and Thomas Huang. Horizontal pyramid matching for person re-identification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8295–8302, 2019. 3

[15] Davrondzhon Gafurov and Einar Snekkenes. Gait recognition using wearable motion recording sensors. *EURASIP Journal on Advances in Signal Processing*, 2009:1–16, 2009. 1

[16] Ju Han and Bir Bhanu. "individual recognition using gait energy image". *IEEE transactions on pattern analysis and machine intelligence*, 28:316–22, 03 2006. 3, 7

[17] Ju Han and Bir Bhanu. "individual recognition using gait energy image". *IEEE transactions on pattern analysis and machine intelligence*, 28:316–22, 03 2006. 5

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3, 5, 6

[19] Zhen Huang, Dixiu Xue, Xu Shen, Xinmei Tian, Houqiang Li, Jianqiang Huang, and Xian-Sheng Hua. 3d local convolutional neural networks for gait recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14920–14929, October 2021. 1, 2

[20] Xiang Li, Yasushi Makihara, Chi Xu, and Yasushi Yagi. End-to-end model-based gait recognition using synchronized multi-view pose constraint. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 4106–4115, October 2021. 2

[21] Xiang Li, Yasushi Makihara, Chi Xu, Yasushi Yagi, Shiqi Yu, and Mingwu Ren. End-to-end model-based gait recognition. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, November 2020. 2

[22] Yanan Li, Yilong Yin, Lili Liu, Shaohua Pang, and Qiuhong Yu. Semi-supervised gait recognition based on self-training. In *2012 IEEE Ninth International Conference on Advanced Video and Signal-Based Surveillance*, pages 288–293, 2012. 2

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

[24] Beibei Lin, Shunli Zhang, and Feng Bao. Gait recognition with multiple-temporal-scale 3d convolutional neural network. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 3054–3062, New York, NY, USA, 2020. Association for Computing Machinery. 2

[25] Beibei Lin, Shunli Zhang, and Xin Yu. Gait recognition via effective global-local feature representation and local temporal aggregation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14648–14656, October 2021. 1, 2

[26] Yiqun Liu, Yi Zeng, Jian Pu, Hongming Shan, Peiyang He, and Junping Zhang. Selfgait: A spatiotemporal representation learning method for self-supervised gait recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021, Toronto, ON, Canada, June 6-11, 2021*, pages 2570–2574. IEEE, 2021. 2

[27] Mengge Mao and Yonghong Song. Gait recognition based on 3d skeleton data and graph convolutional network. In *2020 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–8, 2020. 2

[28] Claudia Nickel, Mohammad O Derawi, Patrick Bours, and Christoph Busch. Scenario test of accelerometer-based biometric gait recognition. In *2011 Third International Workshop on Security and Communication Networks (IWSCN)*, pages 15–21. IEEE, 2011. 1

[29] Javier Ortega-Garcia, Josef Bigun, Douglas Reynolds, and Joaquin Gonzalez-Rodriguez. Authentication gets personal with biometrics. *IEEE signal processing magazine*, 21(2):50–62, 2004. 1

[30] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. 5

[31] Haocong Rao, Siqi Wang, Xiping Hu, Mingkui Tan, Huang Da, Jun Cheng, and Bin Hu. Self-supervised gait encoding with locality-aware attention for person re-identification. In Christian Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 898–905. International Joint Conferences on Artificial Intelligence Organization, 7 2020. Main track. 2

[32] Haocong Rao, Siqi Wang, Xiping Hu, Mingkui Tan, Yi Guo, Jun Cheng, Xinwang Liu, and Bin Hu. A self-supervised gait encoding approach with locality-awareness for 3d skeleton based person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 1

[33] Muhammad Sharif, Mudassar Raza, Jamal Hussain Shah, Mussarat Yasmin, and Steven Lawrence Fernandes. An overview of biometrics methods. *Handbook of Multimedia Information Security: Techniques and Applications*, pages 15–35, 2019. 1

[34] 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, 2016. 5

[35] Anna Sokolova and Anton Konushin. Pose-based deep gait recognition. *IET Biometrics*, 8(2):134–143, 2019. 2

[36] Naoto Takayanagi, Motoki Sudo, Yukari Yamashiro, Sangyoon Lee, Yoshiyuki Kobayashi, Yoshifumi Niki, and Hiroyuki Shimada. Relationship between daily and in-laboratory gait speed among healthy community-dwelling older adults. *Scientific reports*, 9(1):1–6, 2019. 1

[37] Rawesak Tanawongsuwan and Aaron Bobick. Gait recognition from time-normalized joint-angle trajectories in the walking plane. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 2, pages II–II. IEEE, 2001. 1

[38] Suibing Tong, Yuzhuo Fu, Xinwei Yue, and Hefei Ling. Multi-view gait recognition based on a spatial-temporal deep neural network. *IEEE Access*, 6:57583–57596, 2018. 2

[39] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018. 5, 6

[40] Daniel Ponsa Vassileios Balntas, Edgar Riba and Krystian Mikolajczyk. Learning local feature descriptors with triplets and shallow convolutional neural networks. In Edwin R. Hancock Richard C. Wilson and William A. P. Smith, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 119.1–119.11. BMVA Press, September 2016. 5, 6, 8

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 4

[42] Changsheng Wan, Li Wang, and Vir V Phoha. A survey on gait recognition. *ACM Computing Surveys (CSUR)*, 51(5):1–35, 2018. 1

[43] Bastian Wandt and Bodo Rosenhahn. Repnet: Weakly supervised training of an adversarial reprojection network for 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7782–7791, 2019. 3, 4

[44] Liang Wang, Tieniu Tan, Huazhong Ning, and Weiming Hu. Silhouette analysis-based gait recognition for human identification. *IEEE transactions on pattern analysis and machine intelligence*, 25(12):1505–1518, 2003. 1

[45] Thomas Wolf, Mohammadreza Babaee, and Gerhard Rigoll. Multi-view gait recognition using 3d convolutional neural networks. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 4165–4169. IEEE, 2016. 2

[46] 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 Transactions on Pattern Analysis and Machine Intelligence*, 39(2):209–226, 2017. 5

[47] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European conference on computer vision (ECCV)*, pages 305–321, 2018. 3

[48] Ke Xu, Xinghao Jiang, and Tanfeng Sun. Gait identification based on human skeleton with pairwise graph convolutional network. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2021. 2

[49] Lingxiang Yao, Worapan Kusakunniran, Qiang Wu, Jingsong Xu, and Jian Zhang. Collaborative feature learning for gait recognition under cloth changes. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1, 2021. 2

[50] Shiqi Yu, Daoliang Tan, and Tieniu Tan. A framework for evaluating the effect of view angle, clothing and carrying

condition on gait recognition. In *18th International Conference on Pattern Recognition (ICPR'06)*, volume 4, pages 441–444. IEEE, 2006. 1, 2, 4, 5, 6

[51] Zhengyou Zhang. Microsoft kinect sensor and its effect. *IEEE MultiMedia*, 19(2):4–10, Apr. 2012. 1

[52] Ziyuan Zhang, Luan Tran, Xi Yin, Yousef Atoum, Xiaoming Liu, Jian Wan, and Nanxin Wang. Gait recognition via disentangled representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4710–4719, 2019. 2, 4, 5, 6, 8