2024 IEEE International Joint Conference on Biometrics (IJCB) | 979-8-3503-6413-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/IJCB62174.2024.10744527

Distillation-guided Representation Learning for Unconstrained Gait Recognition

Yuxiang Guo[1], Siyuan Huang[1], Ram Prabhakar[1],
Chun Pong Lau[2], Rama Chellappa[1], Cheng Peng[1]
[1]Johns Hopkins University, [2]City University of Hong Kong
{yguo87, shuan124, rprabha3, rchella4, cpeng26}@jhu.edu, cplau27@cityu.edu.hk

Abstract

*Gait recognition holds the promise of robustly identifying subjects based on walking patterns instead of appearance information. While previous approaches have performed well for curated indoor data, they tend to underperform in unconstrained situations, e.g. in outdoor, long distance scenes, etc. We propose a framework, termed **GA**it **DE**tection and **R**ecognition (**GADER**), for human authentication in challenging outdoor scenarios. Specifically, GADER leverages a Double Helical Signature to detect segments that contain human movement and builds discriminative features through a novel gait recognition method, where only frames containing gait information are used. To further enhance robustness, GADER encodes viewpoint information in its architecture, and distills representation from an auxiliary RGB recognition model, which enables GADER to learn from silhouette and RGB data at training time. At test time, GADER only infers from the silhouette modality. We evaluate our method on multiple State-of-The-Arts(SoTA) gait baselines and demonstrate consistent improvements on indoor and outdoor datasets, especially with a significant 25.2% improvement on unconstrained, remote gait data.*

1. Introduction

Unconstrained biometric identification, in outdoor and far-away situations, has been a longstanding challenge [64, 63, 47, 48]. RGB-based face and body recognition systems focus on learning *spatially* discriminative features; however, real-world effects like challenging view angles, low face resolution, changing appearances (e.g., clothes and glasses), and long distance turbulence can significantly distort biometric information. Consequently, RGB-based recognition systems tend to perform inconsistently in remote unconstrained scenarios [32, 33, 53].

Gait analysis provides an alternative modality for human recognition by focusing on learning discriminative features in the *temporal* domain. As such, it can be more robust to challenging, unconstrained situations, especially at range,

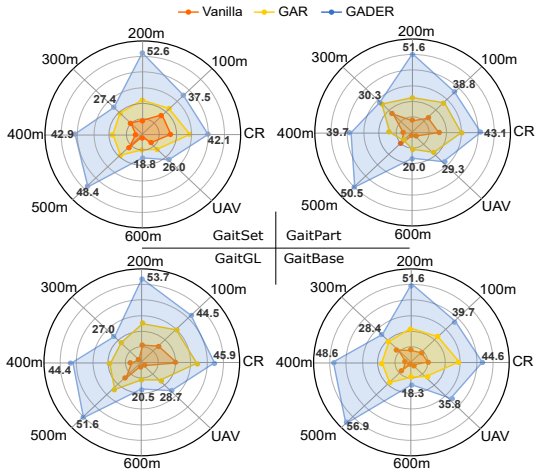


Figure 1. For each baseline, GADER raises the Rank-1 accuracy on BRIAR significantly, especially over 25% at 500m. Close Range(CR), Unmanned Aerial Vehicle(UAV).

and has been applied in many applications such as human authentication [4], health [13] and crime analysis [21], etc.

The field of gait recognition was initially developed [46, 22, 11, 35, 1, 47] using traditional methods, such as template matching [5, 38, 61] and model-based methods [3, 58, 6, 31], but limited by variations in scale and viewing angle and sensitivity to video quality, respectively. Deep learning (DL)-based approaches [9, 2, 37] have made significant advances in image and video-based recognition tasks compared to traditional methods. They are able to generate robust identity embeddings by directly processing the complex temporal information present in gait sequences. This enables effective recognition under the variabilities mentioned above, making the DL-based methods widely preferred.

While DL-based gait recognition performs well for indoor scenes, it often fails to achieve good performance in unconstrained/outdoor scenarios. In this work, we seek to apply gait recognition to unconstrained situations with maximal automation. The recently collected BRIAR [12] dataset contains standing, structured walking, and random

walking sequences, which mimic the real-world challenges in gait recognition. Existing gait recognition methods assume that the subject is always walking with periodic movement and that there are no standing sequences [59, 50]. By making such assumptions, these methods tend to learn suboptimal representations, sometimes only achieving 24% on close range recognition [37]. We also realize that standing segments inherently contain little temporal information, so it is computationally intensive to apply 3D convolutions widely used in gait recognition models. Moreover, the distinctive temporal patterns of standing and gait sequences raise a problem in generating a cohesive feature space for the same identity using one model. This highlights the need for an approach that separates frames that do not contain human motion in order to make gait recognition features more robust.

In the common application scenario of gait recognition - i.e. as a component of end-to-end video recognition, there exists a plethora of additional information at training time that is rarely considered. For instance, RGB images are required for generating human silhouettes. These RGB images contain rich information that can contribute to building robust features not captured by silhouettes alone. Intuitively, the feature space learned by a body recognition model can be used to enhance gait recognition. Another example is viewpoint information. Many gait recognition methods employ the practice of *size normalization* [9], where the original masks are cropped and resized to the same resolution regardless of the subject's distance from the camera, leading to information loss. Particularly, such a resizing ratio over frames can implicitly offer important viewpoint information, which is useful for generating effective embeddings [8]. Unfortunately, this cue is lost in the resizing operation. Based on our experiments, we show that viewpoint information helps to build a robust representation, especially in unconstrained situations.

In this paper, we aim to push the performance of unconstrained gait analysis through a framework named GAit Detection and Recognition system in the wild (GADER). To address the problem of mixing the moving and standing segments in sequences, we introduce a novel *gait detection* module (Fig. 4) that detects the walking and non-walking parts in a sequence, so that gait recognition can just exploit the frames that contain human movement. Instead of using a 3D volume to capture the movement, our gait detector uses the Double Helical Signature [45, 42], a 2D pattern, with a lightweight classification model to segment the walking portion of the input sequence. Thus, we do not provide the entire sequence to the model. Instead, we split the given video into multiple windows of varying lengths to get predictions, followed by Non-Maximum Suppression, to localize the movement duration. This provides a relatively pure gait sequence for gait recognition models, yielding a robust representation and making it suitable for real-world scenarios.

For the gait recognition module, namely GAR, along with the size normalized silhouettes capturing the temporal and body shape information, we further introduce a *cross-modality feature distillation* step, where we guide the intermediate gait features to be more expressive by making them close to features generated from RGB frames. This enables the gait features to maintain their robustness to appearance changes, while also benefiting from the discriminative power of RGB features. As the augmented gait features can be obtained from silhouettes and do not require the RGB frames during inference, we also gain computational efficiency. Additionally, we embed the resizing ratio from the original frame as an attention signal. This *ratio attention* helps to preserve the viewpoint information that is beneficial for robust identity representation.

In summary, GADER makes three contributions:

- We introduce a light-weight gait detector to automatically detect frames that contain human movements so that gait recognition and person ReID can cooperate efficiently.
- We propose a novel gait recognition training strategy, which leverages the color space and size information during training; specifically, knowledge distillation on RGB features is used to enhance silhouette features' capacity.
- We conduct a series of evaluations, i.e. rank retrieval and verification on CASIA-B, Gait3D and BRIAR datasets, showing consistent improvement in applying SoTA gait recognition backbones.

2. Related Work

2.1. Gait Representations

Gait, encompassing both spatial and temporal information, offers various avenues for representation, primarily classified into *appearance-based* and *human model-based*. Preceding the advent of deep learning, many appearance-based representations [46, 11, 61] sought to compress temporal information into a single frame. Han and Bhanu [22] introduced the Gait Energy Image (GEI) template as an average of aligned and normalized silhouette frames. Other popular appearance-based gait representations include Frame Difference Energy Image [11] and Active Energy Image [61]. On the other hand, model-based methods represent the whole human body using well-defined models to represent gait. The methods vary by the different techniques used for modeling the human body, such as hidden Markov models [28, 39], stride length and walking tempo [3], and Velocity Hough Transform [41].

2.2. Gait Recognition

Early *appearance-based* methods employed global gait representation such as silhouettes [9], RGB [20, 62, 30] and

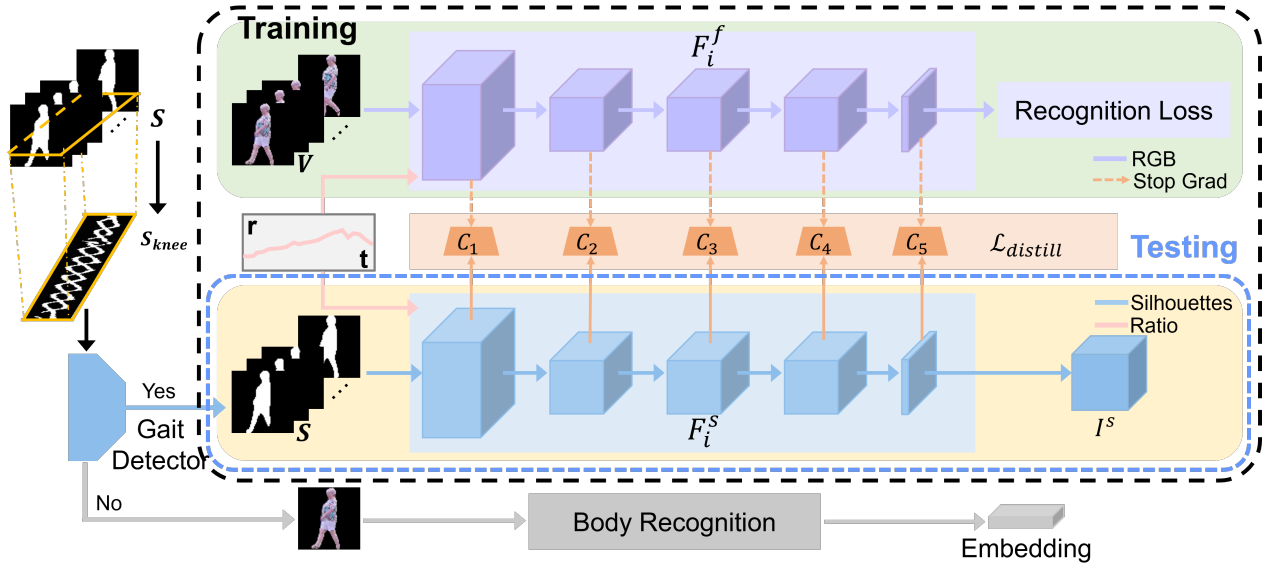


Figure 2. Overview of end-to-end pipeline. GADER consists of two parts: gait detection GAR. The gait detector utilizes gait representation to filter out segments without gait information or incomplete body which will be processed by a body recognition algorithm, and only frames with human movement are fed to GAR. GAR leverages ratio attention and RGB feature space to extract a more robust silhouette feature for recognition.

GEI [49, 55, 25] as input to CNNs. The advent of deep learning technology propelled silhouettes to the forefront, primarily owing to their simplicity, privacy, and discriminative capabilities. Dou *et al.* introduced GaitMPL [14], a progressive method that learns from simple to hard samples. Recently, [40] proposed a dynamic aggregated network (DANet) to represent contextual relationships by encoding pixel features into magnitude and phase. Compared to appearance-based methods, *model-based* approaches [17] often yield smaller input sizes extracted by lightweight models, resulting in reduced computational overhead. GaitGraph [52] and GaitGraph2 [51] represent pose keypoints as a graph and extract features through Graph Convolutional Network (GCN). GaitTR [60] and GaitMixer [44] employ transformer architecture to capture global temporal and spatial relationships. 3D representations extracted using depth sensors [24, 43, 10] and RGB images [36, 35, 7, 29, 18] also have shown promising results.

Recognition in the Wild: Developing an accurate gait representation in the wild is a long-term goal and has been actively researched over the past decade. Towards that aim, Zhu *et al.* curated a large scale gait dataset called GREW [64]. It is a natural video dataset consisting of 128K sequences of 26K identities captured over 882 cameras. Similarly, Zheng *et al.* [63] collected Gait3D dataset that contains silhouettes, 2D/3D keypoints, and 3D meshes for 3D gait recognition. The authors [63] observed that state-of-the-art gait recognition methods do not yield similar superior performance for GREW and Gait3D as they do for indoor datasets like CASIA-B.

3. Method

In this work, we focus on silhouette-based gait recognition, which relies on the binary masks of the subjects in a video. Formally, we denote a video as a 4D tensor, i.e. $\boldsymbol{V} \in \mathbb{R}^{T\times H\times W\times 3}$, where T, H, W, are the frame index, height and width. For each frame t, the subject silhouette $\boldsymbol{s}_t \in \{0,1\}^{H\times W}$ is obtained from an off-the-shelf segmentation model, e.g. [54]. Gait recognition takes $\boldsymbol{S} = [\boldsymbol{s}_t]_{t=1}^{T}$ as the input, and obtains corresponding features $\boldsymbol{f} = F_\theta(\boldsymbol{S})$ from a feature extractor F_θ. Triplet loss [23] is used to constrain the training process of F_θ with respect to ground-truth labels $y \in \{1, 2, \ldots, |\mathcal{Y}|\}$ for each video in the training sets, where $|\mathcal{Y}|$ is the cardinality of the label set. After F_θ is trained, gallery gait silhouettes $\boldsymbol{S}^g$ and probe gait silhouettes $\boldsymbol{S}^p$ are passed through F_θ to obtain the gait feature $\boldsymbol{f}^g$ and $\boldsymbol{f}^p$. To recognize the probe identity, a similarity metric $\mathcal{D}$, such as Euclidean distance or cosine similarity, is used to get $\mathcal{D}(\boldsymbol{f}^g, \boldsymbol{f}^p)$, where the gallery subject g and the probe p are decided to be the same person if they are the close enough in feature space.

3.1. Gait Detector

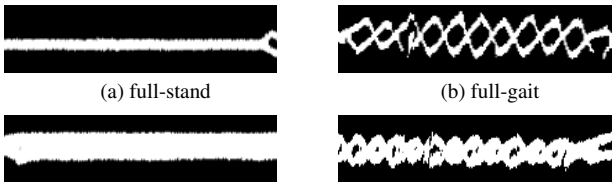
Previous gait recognition methods [9, 16, 37, 34, 8] directly use the silhouettes $\boldsymbol{S}$ as the input, implicitly under the assumption that the video sequence captures *the entire body* and *continuous movement*. While these assumptions are likely to be effective for curated datasets, such as CASIA-

B [59], they are often ineffective for unconstrained videos and will lead to suboptimal performance [26]. To this end, we propose a gait detector to assess if the video sequence contains gait movement with the complete body, analogous to the role played by face detection in face recognition, retaining only those frames that contain human movement for subsequent processing.

3.1.1 Double Helical Signature

An ideal representation of a gait detector should be able to discriminate moving subjects from stationary ones and a partial body from a full body. Since the legs' motion is nearly periodic and contributes significantly to gait recognition, we use the Double Helical Signature (DHS) [45]. DHS is a classic gait representation that captures the movement of the knee to describe gait movement as a function of time. As our input $\boldsymbol{S}$ is normalized to the same size, we can deduce the knee height $\mathcal{H}_{\text{knee}}$ to be approximately a quarter of the overall frame height. By taking a slice from the silhouette sequence, i.e., $\boldsymbol{S}_{\text{knee}}(x,t) = \boldsymbol{S}(x, \mathcal{H}_{\text{knee}}, t), \boldsymbol{S}_{\text{knee}} \in \mathbb{R}^{W \times T}$, we obtain a DHS pattern that indicates human movement.

As shown in Fig. 3, the DHS pattern is discriminating. For the standing case, the DHS pattern shows a constant straight line since there is no movement at the knees. When the subject is walking, a periodic pattern is obtained, known as a type of "Frieze pattern" [42]. Given an incomplete body, DHS appears rather different as $\mathcal{H}_{\text{knee}}$ does not correspond to the knee position anymore; consequently, DHS becomes thicker. Therefore, DHS emerges as a compact and distinctive representation encapsulating temporal movement into a 2D image.



(a) full-stand (b) full-gait (c) part-stand (d) part-gait

Figure 3. Four cases of DHS(a-d) are shown using two variables: full/part - indicates whether the body is complete, and stand/gait - shows whether gait information is present.

3.1.2 Light-weighted Classification

Capitalizing on the distinctive characteristics of DHS, we employ a simple yet effective network to extract segments from $\boldsymbol{S}$ that contain gait information and a complete body. During the training step, we randomly select the start point and duration from the entire DHS and get R_{knee}, which is similar to the sampling strategy commonly applied in gait recognition. Each segment has the same height as DHS's.

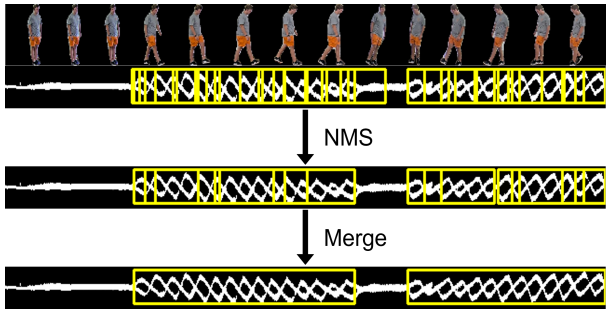


Figure 4. The gait detection process. The gait detector processes split DHS segments to obtain prediction followed by Non-Maximum Suppression (NMS) and concatenation to pinpoint frames where gait information is present in a sequence. The yellow bounding boxes indicate the frames that contain gait.

The fragment is then processed by a five-layer Convolutional Neural Network $\mathcal{M}_\phi$ to get the feature. To handle the varying window lengths, we employ a temporal pooling module in the form of a max-pooling layer to generate a window-length invariant embedding. Subsequently, a four-class multi-layer perceptron (MLP) is applied to obtain the prediction, using the cross entropy loss.

$$P = \text{MLP}(\mathcal{P}_{Max}^{1\times 1\times t}(\mathcal{M}_\phi(R_{\text{knee}}))), \tag{1}$$

where $\mathcal{P}_{Max}^{1\times 1\times t}$ represents a temporal pooling module with size $(1 \times 1 \times t)$, and t is the window's width.

Especially in an unconstrained walking sequence, non-ideal fragments, including turning and standing within a short period, limit detection performance. To precisely identify the segments that contain gait information as well as a complete boy, we first split the entire DHS sequence into multiple windows of varying durations R_{knee}^n, where n represents the number of windows. Each window goes through the well-trained gait classification model and gets a corresponding prediction. We only keep the complete body gait predictions and reduce the predictions by Non-Maximum Suppression (NMS). Finally, we check the reduced windows' inner distance and merge them if the distance is smaller than a predefined threshold. The ratio of the detected movement length to the entire DHS sequence serves as an indicator for determining whether the sequence should be further utilized in the gait recognition module. The process is illustrated in Fig. 4. Thus our gait detector automatically identifies the corresponding movement status of each clip.

3.2. GAit Recognition (GAR)

With the help of the gait detector, we can obtain relatively robust gait sequences for gait recognition. In the recognition stage, we use GAR, which incorporates prior knowledge such as the RGB feature space and the resizing information. The architecture is shown in Fig. 2.

3.2.1 Ratio Attention

Viewpoint information has been shown to improve gait recognition performances [8]; however, such information is difficult to obtain in unconstrained situations. In this work, we leverage resizing ratios to dynamically describe the changing views. The *resizing ratio* is defined as the change in height from the original bounding box to the normalized silhouette height, which is usually 64. Intuitively, the resizing ratios are similar if two videos are recorded from the same viewpoint, and these ratios naturally encode viewpoint information.

To effectively incorporate ratios into the gait feature extraction module, we apply it as an attention mechanism and fuse it into the network. Empirically, attention is an effective way to employ view point information as shown in Sec. 4.3. The ratios are embedded using 1D convolution F_{1DConv}, followed by a sigmoid function $Sigmoid(*)$, i.e.

$$r = Sigmoid(F_{1DConv}(\frac{\boldsymbol{S}_{raw}}{\boldsymbol{S}})), \tag{2}$$

where $\boldsymbol{S}_{raw}$ and $\boldsymbol{S}$ are the body bounding boxes in the original and normalized silhouette, and $r \in \mathbb{R}^{1\times1\times T}$. Notably, we utilize the ratio as a pattern, rather than a discrete representation [8], to characterize the viewing point, effectively and dynamically adapting to the viewpoint changes in unconstrained scenarios.

3.2.2 Cross Modality Distillation

Previous works [16, 37, 34, 56] have shown that silhouette-based recognition is robust to appearance changes such as different clothing and low image quality. On the flip side, segmenting RGB frames into silhouettes leads to loss of useful information for human identification, e.g. the rich content [57, 20]. While gait recognition systems do not have access to color space information at *test time*, they can benefit by learning from the feature space learned by an RGB-based recognition system *during training* to enhance the ability to separate identities. To this end, GAR introduces an auxiliary branch to extract features from RGB. We denote $F^f_{3DConv}, F^s_{3DConv}$ as the first 3D convolution layers to the RGB and silhouette feature extraction backbones. Combining with ratio attention r, GAR first obtains weight features, i.e.

$$\boldsymbol{F}^f_1 = r * F^f_{3DConv}(\boldsymbol{V}), \boldsymbol{F}^s_1 = r * F^s_{3DConv}(\boldsymbol{S}). \tag{3}$$

Subsequently, we apply the rest of backbones to the processed features,

$$\begin{aligned} \boldsymbol{F}^f_{i+1} = F^f_i(\boldsymbol{F}^f_i), \boldsymbol{F}^s_{i+1} = F^s_i(\boldsymbol{F}^s_i), \\ I^f = F^f_N(\boldsymbol{F}^f_N), I^s = F^s_N(\boldsymbol{F}^s_N), \end{aligned} \tag{4}$$

in which I^f and I^s are identification embedding, and $i \in \{1, 2, ..., N\}$ is the convolution block index.

3.3. Loss functions

A cross-modal distillation loss is employed within GAR, which promotes the representation power of gait features based on the learned RGB feature space. Since both modalities have their specific advantages, directly forcing their features to be close leads to an averaged representation that does not benefit from the specificity of each modality. An additional convolutional layer C_i is introduced for the i-th intermediate silhouette feature $\boldsymbol{F}^s_i$, such that the transformed features are constrained to be similar to $\boldsymbol{F}^f_i$. The loss can be described as:

$$\mathcal{L}_{distill} = \frac{1}{N}\sum_i (\mathcal{D}(\varnothing(\boldsymbol{F}^f_i), C_i(\boldsymbol{F}^s_i))), \tag{5}$$

where stop gradient ($\varnothing$) operation is used such that the RGB branch is not affected by the silhouette features in this process. Note that C_i is only used at training time.

Triplet loss [23] $\mathcal{L}^f_{tri}$, $\mathcal{L}^s_{tri}$ is applied to maximize the distance of representations from different subjects and minimize the ones from the same identity for both modalities.

Overall, the training loss is

$$\mathcal{L}_{train} = \lambda_f\mathcal{L}^f_{tri} + \lambda_s\mathcal{L}^s_{tri} + \lambda_{distill}\mathcal{L}_{distill}, \tag{6}$$

where $\lambda_{f,s,distill} = 0.425, 0.425, 0.15$ are the loss hyperparameters used during training.

4. Experiments

4.1. Datasets and Metrics

In this work, we focus on applying gait recognition to unconstrained scenarios with comparatively sparse data curation, exemplified by **BRIAR** [12]. BRIAR consists of 776 and 856 subjects used as training and test sets, respectively. In the test set, there are 493 distractors and 363 target subjects. Compared to other unconstrained datasets, BRIAR data includes different walking status and clothes settings, and incomplete body shapes into consideration, making it very challenging. Similar to previous methods for gait recognition, we also evaluate the proposed approach on **CASIA-B** [59], a controlled, indoor dataset with continuous motion. CASIA-B consists of 124 subjects with three walking conditions which are normal walking (NM), walking with a bag (BG) and walking in a coat or jacket (CL). To further demonstrate the effectiveness of the proposed approach in the unconstrained case, we also test GADER on the **Gait3D** [63] dataset. Gait3D consists of 25,309 sequences recorded by 39 cameras on 4,000 subjects inside a large supermarket.

Considering CASIA-B and Gait3D are pure gait datasets, we assume that all sequences in these datasets fully contain gait, so we only evaluate **GAR**. For the BRIAR

	Rank Retrieval									Verification			
Probe	CR	100m	200m	300m	400m	500m	600m	UAV	Mean	$1e^{-4}$	$1e^{-3}$	$1e^{-2}$	$1e^{-1}$
GaitSet [9]	21.3	20.8	13.4	14.5	8.4	15.4	6.2	11.7	17.6	5.9	15.0	33.0	62.2
↪ w / GAR	31.6	26.1	24.9	22.7	21.7	22.3	14.1	17.1	26.8 ↑ 9.2	9.0	21.2	41.2	67.5
↪ w / GADER	42.1	37.5	52.6	27.4	42.9	48.4	18.8	26.0	35.9 ↑ 18.3	9.3	22.5	45.7	75.6
GaitPart [16]	20.2	17.7	12.4	21.1	10.2	13.8	5.6	6.2	15.7	4.0	11.0	25.9	60.7
↪ w / GAR	32.3	30.1	25.8	29.1	17.8	11.3	14.7	21.5	27.3 ↑ 11.6	10.3	21.0	40.9	74.3
↪ w / GADER	43.1	38.8	51.6	30.3	39.7	50.5	20.0	29.3	36.3 ↑ 20.6	10.4	22.4	45.5	79.7
GaitGL [37]	24.6	18.1	15.5	7.5	11.0	18.1	6.2	6.5	17.3	4.9	12.9	33.1	65.9
↪ w / GAR	35.8	32.3	27.8	21.6	23.0	27.1	15.2	19.8	28.3 ↑ 11.0	7.1	16.4	33.6	60.4
↪ w / GADER	45.9	44.5	53.7	27.0	44.4	51.6	20.5	28.7	37.8 ↑ 20.5	8.2	19.6	41.1	71.3
GaitBase [15]	14.8	13.3	12.3	15.5	8.1	12.2	4.5	6.0	11.6	3.0	7.6	20.9	54.4
↪ w / GAR	24.0	22.6	23.7	23.0	19.1	22.3	14.1	17.1	26.8 ↑ 15.2	5.5	13.1	31.6	65.7
↪ w / GADER	44.6	39.7	51.6	28.4	48.6	56.9	18.3	35.8	36.8 ↑ 25.2	7.2	17.6	39.9	74.5

Table 1. Rank-1 accuracy (%) and verification (TPR(%)@FPR =$1e^{-4}$, $1e^{-3}$, $1e^{-2}$, $1e^{-1}$) on BRIAR. CR is Close Range(<100m); Unmanned Aerial Vehicle (UAV).

dataset, we use the proposed gait detector to extract the segments containing gait information with a complete body and feed them to the GAR model; the remaining frames are processed by a SoTA ReID method, SemReID [27], the resulting system is named as **GADER**.

Evaluation Metric For CASIA-B and BRIAR, *verification* and *rank retrieval* are used to evaluate recognition performance. As for Gait3D, the evaluation follows the open-set instance retrieval setting and calculates the average Rank-1, 5, and 10 accuracies, mean Average Precision (mAP), and mean Inverse Negative Penalty (mINP) over queries.

4.2. Quantitative Evaluation

Evaluation on BRIAR [12] To demonstrate the superior performance of the proposed methods in the wild, we evaluate on BRIAR data with GaitSet [9], GaitPart [16], GaitGL [37] and GaitBase [15] serving as backbones. The Rank-1 accuracies are shown in Table 1. With the help of distilled knowledge from RGB data, our gait recognition models gain improvements compared to their vanilla version, by 9.2%, 11.6%, 11.0% and 15.2%, respectively. GAR leads to better performance than previous works, indicating that the referred discriminative feature space from distillation contributes to a better gait feature. When the gait detector as well as SemReID are integrated, GaitGL with GADER doubles the vanilla's accuracy from 17.3% to 37.8%, showing that it is beneficial to apply the gait detector to eliminate segments without human movements.

When it comes to *verification*, the TAR (%) results are shown in Table 1. For the case of GaitBase, compared to the original, our method improves by 10.7%, achieving 31.6% when FAR=$1e^{-2}$. When the gait detector is included, we observe a big improvement. The recognition system's verification results increase from 20.9% to 39.9% FAR@$1e^{-2}$indicating that gait recognition will perform well if the segments that contain gait information with a complete body are used as input.

Evaluation on CASIA-B [59] To show the effectiveness of the proposed method in indoor scenarios, we present the performance variations across four backbones by incorporating the resizing ratio and RGB feature space. The Rank-1 accuracies are shown in Table 2. Compared to the original backbones, the models with prior knowledge consistently demonstrates an increase of 3.3% for GaitSet on CL. It is crucial to highlight that all improvements with GAR are accomplished without additional parameters involved but with an extra 1D convolution to extract ratio information in the test phase.

We also evaluate the *verification* performance. The TAR(%) results are shown in Table 3. Across the four backbones, GAR contributes to enhancements of 2.7%, 3.4%, 2.0% and 0.7%, attaining verification rates of 78.2%, 81.2%, 81.7% and 80.8%, respectively. It is noteworthy that although the recognition results in NM and BG approach saturation, there remains room for improvement in the verification task.

Evaluation on Gait3D [63] To evaluate our model on a public outdoor dataset, we also did cross-domain evaluation on Gait3D. The results are in Table 4. We see that our proposed method achieves higher performance in all criteria. Especially, Rank-1 increases 2.0% to 23.5%. Since cross-domain evaluation is a challenging task, the results are lower than single-domain ones. The model trained on the BRIAR dataset exhibits superior performance compared to others, indicating that the model learned from the BRIAR dataset generalize well.

Gait Detector Evaluation We trained a gait detector using the DHSs generated from the BRIAR training set, and

Gallery NM#1-4		$0^\circ - 180^\circ$											
Probe		0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	Mean
CL#1-2	GaitSet [9]	61.4	75.4	80.7	77.3	72.1	70.1	71.5	73.5	73.5	68.4	50.0	70.4
	↪ w / GAR	67.5	82.3	84.1	79.2	70.8	68.8	71.4	75.2	77.7	75.9	58.1	73.7 ↑ 3.3
	GaitPart [16]	70.7	85.5	86.9	83.3	77.1	72.5	76.9	82.2	83.8	80.2	66.5	78.7
	↪ w / GAR	78.0	85.7	89.4	84.4	77.7	71.6	77.0	79.7	84.3	81.3	68.2	79.7 ↑ 1.0
	GaitBase [15]	-	-	-	-	-	-	-	-	-	-	-	77.4
	↪ w / GAR	69.3	80.1	82.5	82.1	81.2	78.2	77.8	80.0	83.3	80.2	65.3	78.2 ↑ 0.8
	GaitGL [37]	76.6	90.0	90.3	87.1	84.5	79.0	84.1	87.0	87.3	84.4	69.5	83.6
	↪ w / GAR	80.8	**90.5**	**92.2**	91.0	84.7	**79.7**	**84.8**	89.6	89.7	**87.1**	**72.8**	85.7 ↑ 2.1

Table 2. Rank-1 accuracy (%) on CASIA-B excluding identical-view case for CL#1-2.

Models	NM	BG	CL	Mean
GaitSet [9]	89.7	77.6	59.3	75.5
↪w / GAR	90.8	81.5	62.4	78.2 ↑ 2.7
GaitPart [16]	90.5	78.7	64.2	77.8
↪w / GAR	91.5	83.0	69.3	81.2 ↑ 3.4
GaitBase [15]	91.1	82.8	65.3	79.7
↪w / GAR	**92.3**	**85.2**	67.7	81.7 ↑ 2.0
GaitGL [37]	91.1	82.7	66.6	80.1
↪w / GAR	91.5	83.1	67.9	80.8 ↑ 0.7

Table 3. Verification (TPR(%)@FPR=$1e^{-2}$) on CASIA-B.

Source	Methods	Rank-1	Rank-5	Rank-10	mAP	mINP
CASIA-B	GaitSet [9]	6.9	14.6	-	4.5	-
	GaitGL [37]	8.8	15.7	18.8	5.5	3.1
OU-MVLP	GaitSet [9]	6.1	12.4	-	4.4	-
	GaitGL [37]	16.4	25.8	31.2	13.1	7.3
GREW	GaitSet [9]	16.5	31.1	-	11.7	-
	GaitGL [37]	18.3	31.9	39.2	13.1	7.3
BRIAR	GaitGL [37]	21.5	36.5	42.5	15.2	8.2
	↪w / GAR	23.5	37.4	43.4	15.9	8.2

Table 4. Cross domain evaluation on Gait3D with detector.

it reaches 91.9%, 86.1% and 88.5% accuracy on BRIAR, CASIA-B and Gait3D respectively. For BRIAR, the false positive, true positive, false negative and true negative for the gait detection module are 4.5%, 53.1%, 3.4% and 38.8%, respectively. The high accuracies show that the gait detector is robust to different domains. If the whole DHS is fed to the detector to get the prediction, the detection accuracy would be 82.7%, 85.3% and 87.1%, respectively. The drop demonstrates the necessity of taking the split signature as an input. These results are under the assumption that the clips in curated datasets do not contain segments without gait information or incomplete body.

4.3. Ablation Study

To show the impact of each part in our design, we conduct a series of ablation experiments.

Ratio and RGB help better silhouette embedding. In the gait recognition model, we evaluate the *ratio attention* and *cross modality distillation* by employing GaitGL as the backbone. From Table 5, when we apply cross modality distillation, the recognition accuracy on CL reaches 85.0% and its verification result increases by 0.5% using the exact same model as GaitGL. As for ratio attention, it improves verification and Rank-1 accuracy under all conditions. Compared to the baseline, our proposed GAR gains a remarkable improvement on verification and Rank-1 in CL from 80.1% and 83.3% to 80.8% and 85.7% respectively. These improvements are also shown in the BRIAR dataset, which means the view angle cue from ratio and RGB's feature space help build a representative embedding.

RGB modality is sensitive to appearance change. In Table 6, we evaluate the framework only using the RGB modality with GaitGL as a feature extraction model, i.e. GaitGL$_{RGB}$. We observe that it has lower performance than silhouette-based one, i.e. GaitGL, decreasing from 17.3% to 15.6%. Considering GaitGL is a model focusing on temporal rather than spatial information, we further experiment on CAL [19], which achieves SoTA in public clothes-changing ReID datasets. But it only reaches 17.6%, still lower than gait-based methods. So, as mentioned in [48], even though RGB is an acceptable modality for gait recognition, it is not widely applied due to its sensitivity to appearance change. However, its feature has unique gait information that augments the silhouette feature.

Mixing gait and non-walking adversely affect feature aggregation. To demonstrate the effectiveness of the gait detector, we first train with all BRIAR training set data, i.e. GaitGL w/ stand, including standing, random walking, and structure walking. From Table 6, we see that when standing sequences are included in the training process, the gait recognition performance drops, which means that static sequences corrupt the gait embedding construction. In Table 1, we show the improved performance obtained by applying GADER–GAR on segments that contain gait information with a complete body while employing SemReID [27] on remaining data. GADER's superior performance shows that it is able to process segments with gait information with a complete body rather well. And this improvement is achieved with little cost, a lightweight classi-

Ratio	Cross	R_{NM}	R_{BG}	R_{CL}	$Veri$
✗	✗	97.2	94.1	83.3	80.1
✓	✗	97.4	94.4	85.1	80.4
✗	✓	97.3	**94.5**	85.0	80.6
✓	✓	**97.5**	**94.5**	**85.7**	**80.8**

(a) CASIA-B: recognition accuracy of three probes (R_{NM}, R_{BG}, R_{CL}) and average verification ($Veri$).

Ratio	Cross	Rank-1	$Veri$
✗	✗	17.3	33.1
✓	✗	26.3	33.4
✗	✓	25.2	33.3
✓	✓	**28.3**	**33.6**

(b) BRIAR: mean rank-1 accuracy(Rank-1) and verification ($Veri$).

Table 5. Ablation studies on ratio attention and cross-modality distillation. The results are shown using recognition accuracy and verification (TPR(%)@FPR=$1e^{-2}$) on CASIA-B and BRIAR.

Modality	Models	Rank-1	Rank-1 w/detect
RGB	CAL[19]	17.6	18.3
	GaitGL_{RGB}	15.6	14.8
Silhouette	GaitGL w/stand	24.2	26.4
	GaitGL[37]	17.3	29.8
	↪w / GAR	28.3	**42.5**

Table 6. Comparison among different models tested on full set and detected qualified gait set through gait detection on BRIAR.

fication network, since the DHS is extracted from the input provided to gait recognition. What is more, the well-trained model is robust to domain gaps among different datasets, so it can be directly applied.

4.4. Failure Cases

Segments without gait appear even in curated datasets. We recognize the presence of incomplete body segments or sequences without discernible movement in the BRIAR dataset, simulating real-life gait recognition challenges. But with closer examination, we found this issue also exists in curated datsets. We show some examples from GREW [64] and Gait3D [63] in Fig. 5. These examples emphasize the need for the gait detector even for curated datasets

Figure 5. Some examples showing the absence of human movement in videos appearing in the curated dataset. We record the dataset name and corresponding subject id of the sequence.

Analysis During test time, a fixed window length is used, which can result in *missing* gait sequences or non-gait segments involved, as demonstrated in Fig. 6 first row, but the majority of the segments containing gait information are still caught. Furthermore, since we only have annotations for the entire DHS sequences rather than individual moments, it is challenging to train a robust model to always make correct predictions especially when the appearance is misleading, as shown in Fig. 6 second row. Despite these challenges, our gait detector can accurately identify frames that contain gait signatures.

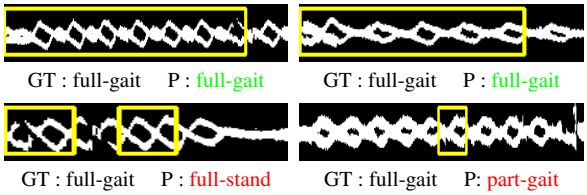


Figure 6. Some gait detector failure cases.

5. Conclusion

In this paper, we present a novel gait detection and recognition approach to address challenging unconstrained conditions. First, we introduce a gait detector to identify frames that contain gait with a complete body. With the help of a gait detector, gait recognition and person ReID can cooperate complementarily to achieve higher performance. Secondly, the gait recognition pipeline utilizes both RGB and silhouette modality to learn robust representations. Notably, we fill the viewpoint information leakage with a simple yet effective ratio attention signal. Additionally, we enhance the silhouette modality embedding through feature distillation from the RGB modality. Such a design helps to leverage the well-learned feature space of RGB modality with the robustness of silhouettes and does not require RGB data at test time. Through extensive experiments, we show that our proposed method improves the performance on Gait3D in cross-domain evaluation and achieves SoTA performance in the standard CASIA-B and the challenging BRIAR dataset.

6. Acknowledgement

This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via [2022-21102100005]. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The US. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

[1] W. An, R. Liao, S. Yu, Y. Huang, and P. C. Yuen. Improving gait recognition with 3d pose estimation. In *Chinese Conference on Biometric Recognition*, pages 137–147. Springer, 2018.

[2] F. Battistone and A. Petrosino. Tglstm: A time based graph deep learning approach to gait recognition. *Pattern Recognition Letters*, 126:132–138, 2019.

[3] C. Ben Abdelkader, R. Cutler, and L. Davis. Stride and cadence as a biometric in automatic person identification and verification. In *Proceedings of Fifth IEEE international conference on automatic face gesture recognition*, pages 372–377. IEEE, 2002.

[4] C. Benedek, B. Gálai, B. Nagy, and Z. Jankó. Lidar-based gait analysis and activity recognition in a 4d surveillance system. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(1):101–113, 2016.

[5] A. F. Bobick and J. W. Davis. The recognition of human movement using temporal templates. *IEEE Transactions on pattern analysis and machine intelligence*, 23(3):257–267, 2001.

[6] N. V. Boulgouris and Z. X. Chi. Human gait recognition based on matching of body components. *Pattern recognition*, 40(6):1763–1770, 2007.

[7] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017.

[8] T. Chai, A. Li, S. Zhang, Z. Li, and Y. Wang. Lagrange motion analysis and view embeddings for improved gait recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20249–20258, 2022.

[9] H. Chao, Y. He, J. Zhang, and J. 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.

[10] P. Chattopadhyay, S. Sural, and J. Mukherjee. Frontal gait recognition from occluded scenes. *Pattern Recognition Letters*, 63:9–15, 2015.

[11] C. Chen, J. Liang, H. Zhao, H. Hu, and J. Tian. Frame difference energy image for gait recognition with incomplete silhouettes. *Pattern Recognition Letters*, 30(11):977–984, 2009.

[12] D. Cornett III, J. Brogan, N. Barber, D. Aykac, S. Baird, N. Burchfield, C. Dukes, A. Duncan, R. Ferrell, J. Goddard, et al. Expanding accurate person recognition to new altitudes and ranges: The BRIAR dataset. *arXiv preprint arXiv:2211.01917*, 2022.

[13] S. Del Din, M. Elshehabi, B. Galna, M. A. Hobert, E. Warmerdam, U. Suenkel, K. Brockmann, F. Metzger, C. Hansen, D. Berg, et al. Gait analysis with wearables predicts conversion to parkinson disease. *Annals of neurology*, 86(3):357–367, 2019.

[14] H. Dou, P. Zhang, Y. Zhao, L. Dong, Z. Qin, and X. Li. Gaitmpl: Gait recognition with memory-augmented progressive learning. *IEEE Transactions on Image Processing*, 2022.

[15] C. Fan, J. Liang, C. Shen, S. Hou, Y. Huang, and S. Yu. Opengait: Revisiting gait recognition towards better practicality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9707–9716, 2023.

[16] C. Fan, Y. Peng, C. Cao, X. Liu, S. Hou, J. Chi, Y. Huang, Q. Li, and Z. 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.

[17] Y. Fu, S. Meng, S. Hou, X. Hu, and Y. Huang. Gpgait: Generalized pose-based gait recognition. *arXiv preprint arXiv:2303.05234*, 2023.

[18] G. Georgakis, R. Li, S. Karanam, T. Chen, J. Košecká, and Z. Wu. Hierarchical kinematic human mesh recovery. In *European Conference on Computer Vision*, pages 768–784. Springer, 2020.

[19] X. Gu, H. Chang, B. Ma, S. Bai, S. Shan, and X. Chen. Clothes-changing person re-identification with rgb modality only. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1060–1069, June 2022.

[20] Y. Guo, C. Peng, C. P. Lau, and R. Chellappa. Multi-modal human authentication using silhouettes, gait and rgb. *arXiv preprint arXiv:2210.04050*, 2022.

[21] A. Hadid, M. Ghahramani, V. Kellokumpu, M. Pietikäinen, J. Bustard, and M. Nixon. Can gait biometrics be spoofed? In *Proceedings of the 21st international conference on pattern recognition (ICPR2012)*, pages 3280–3283. IEEE, 2012.

[22] J. Han and B. Bhanu. Individual recognition using gait energy image. *IEEE transactions on pattern analysis and machine intelligence*, 28(2):316–322, 2005.

[23] E. Hoffer and N. Ailon. Deep metric learning using triplet network. In Y. Bengio and Y. LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*, 2015.

[24] M. Hofmann, J. Geiger, S. Bachmann, B. Schuller, and G. Rigoll. The tum gait from audio, image and depth (gaid) database: Multimodal recognition of subjects and traits. *Journal of Visual Communication and Image Representation*, 25(1):195–206, 2014.

[25] E. Hossain and G. Chetty. Multimodal feature learning for gait biometric based human identity recognition. In *International Conference on Neural Information Processing*, pages 721–728. Springer, 2013.

[26] S. Hou, X. Liu, C. Cao, and Y. Huang. Gait quality aware network: Toward the interpretability of silhouette-based gait recognition. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–11, 2022.

[27] S. Huang, Y. Zhou, R. P. Kathirvel, R. Chellappa, and C. P. Lau. Self-supervised learning of whole and component-based semantic representations for person re-identification. *arXiv preprint arXiv:2311.17074*, 2023.

[28] A. Kale, A. Sundaresan, A. Rajagopalan, N. P. Cuntoor, A. K. Roy-Chowdhury, V. Kruger, and R. Chellappa. Identification of humans using gait. *IEEE Transactions on image processing*, 13(9):1163–1173, 2004.

[29] N. Kolotouros, G. Pavlakos, D. Jayaraman, and K. Daniilidis. Probabilistic modeling for human mesh recovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11605–11614, 2021.

[30] X. Li, Y. Makihara, C. Xu, Y. Yagi, S. Yu, and M. Ren. End-to-end model-based gait recognition. In *Proceedings of the Asian conference on computer vision*, 2020.

[31] X. Li, S. J. Maybank, S. Yan, D. Tao, and D. Xu. Gait components and their application to gender recognition. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 38(2):145–155, 2008.

[32] Y. Li, Y. Li, and N. Vasconcelos. Resound: Towards action recognition without representation bias. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 513–528, 2018.

[33] Y. Li and N. Vasconcelos. Repair: Removing representation bias by dataset resampling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9572–9581, 2019.

[34] J. Liang, C. Fan, S. Hou, C. Shen, Y. Huang, and S. Yu. Gaitedge: Beyond plain end-to-end gait recognition for better practicality. *arXiv preprint arXiv:2203.03972*, 2022.

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

[36] R. Liao, S. Yu, W. An, and Y. Huang. A model-based gait recognition method with body pose and human prior knowledge. *Pattern Recognition*, 98:107069, 2020.

[37] B. Lin, S. Zhang, and X. Yu. Gait recognition via effective global-local feature representation and local temporal aggregation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14648–14656, 2021.

[38] J. Liu and N. Zheng. Gait history image: a novel temporal template for gait recognition. In *2007 IEEE international conference on multimedia and expo*, pages 663–666. IEEE, 2007.

[39] Z. Liu and S. Sarkar. Improved gait recognition by gait dynamics normalization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(6):863–876, 2006.

[40] K. Ma, Y. Fu, D. Zheng, C. Cao, X. Hu, and Y. Huang. Dynamic aggregated network for gait recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22076–22085, 2023.

[41] J. M. Nash, J. N. Carter, and M. S. Nixon. Dynamic feature extraction via the velocity hough transform. *Pattern Recognition Letters*, 18(10):1035–1047, 1997.

[42] S. A. Niyogi and E. H. Adelson. Analyzing gait with spatiotemporal surfaces. In *Proceedings of 1994 IEEE Workshop on Motion of Non-rigid and Articulated Objects*, pages 64–69. IEEE, 1994.

[43] J. F. Nunes, P. M. Moreira, and J. M. R. Tavares. Benchmark rgb-d gait datasets: A systematic review. In *ECCOMAS Thematic Conference on Computational Vision and Medical Image Processing*, pages 366–372. Springer, 2019.

[44] E. Pinyoanuntapong, A. Ali, P. Wang, M. Lee, and C. Chen. Gaitmixer: skeleton-based gait representation learning via wide-spectrum multi-axial mixer. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[45] Y. Ran, Q. Zheng, R. Chellappa, and T. M. Strat. Applications of a simple characterization of human gait in surveillance. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 40(4):1009–1020, 2010.

[46] S. Sarkar, P. J. Phillips, Z. Liu, I. R. Vega, P. Grother, and K. W. Bowyer. The humanid gait challenge problem: Data sets, performance, and analysis. *IEEE transactions on pattern analysis and machine intelligence*, 27(2):162–177, 2005.

[47] A. Sepas-Moghaddam and A. Etemad. Deep gait recognition: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):264–284, 2022.

[48] C. Shen, S. Yu, J. Wang, G. Q. Huang, and L. Wang. A comprehensive survey on deep gait recognition: algorithms, datasets and challenges. *arXiv preprint arXiv:2206.13732*, 2022.

[49] K. Shiraga, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi. Geinet: View-invariant gait recognition using a convolutional neural network. In *2016 international conference on biometrics (ICB)*, pages 1–8. IEEE, 2016.

[50] N. Takemura, Y. Makihara, D. Muramatsu, T. Echigo, and Y. 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):1–14, 2018.

[51] T. Teepe, J. Gilg, F. Herzog, S. Hörmann, and G. Rigoll. Towards a deeper understanding of skeleton-based gait recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1569–1577, 2022.

[52] T. Teepe, A. Khan, J. Gilg, F. Herzog, S. Hörmann, and G. Rigoll. Gaitgraph: Graph convolutional network for skeleton-based gait recognition. In *2021 IEEE international conference on image processing (ICIP)*, pages 2314–2318. IEEE, 2021.

[53] P. Weinzaepfel and G. Rogez. Mimetics: Towards understanding human actions out of context. *International Journal of Computer Vision*, 129(5):1675–1690, 2021.

[54] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019.

[55] Z. Wu, Y. Huang, L. Wang, X. Wang, and T. 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, 2016.

[56] J. Xiao, L. Jing, L. Zhang, J. He, Q. She, Z. Zhou, A. Yuille, and Y. Li. Learning from temporal gradient for semi-supervised action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3252–3262, 2022.

[57] D. Ye, C. Fan, J. Ma, X. Liu, and S. Yu. Biggait: Learning gait representation you want by large vision models. *arXiv preprint arXiv:2402.19122*, 2024.

[58] J.-H. Yoo, D. Hwang, K.-Y. Moon, and M. S. Nixon. Automated human recognition by gait using neural network. In *2008 First Workshops on Image Processing Theory, Tools and Applications*, pages 1–6. IEEE, 2008.

[59] S. Yu, D. Tan, and T. 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.

[60] C. Zhang, X.-P. Chen, G.-Q. Han, and X.-J. Liu. Spatial transformer network on skeleton-based gait recognition. *Expert Systems*, 40(6):e13244, 2023.

[61] E. Zhang, Y. Zhao, and W. Xiong. Active energy image plus 2dlpp for gait recognition. *Signal Processing*, 90(7):2295–2302, 2010.

[62] Z. Zhang, L. Tran, F. Liu, and X. Liu. On learning disentangled representations for gait recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):345–360, 2020.

[63] J. Zheng, X. Liu, W. Liu, L. He, C. Yan, and T. Mei. Gait recognition in the wild with dense 3d representations and a benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20228–20237, 2022.

[64] Z. Zhu, X. Guo, T. Yang, J. Huang, J. Deng, G. Huang, D. Du, J. Lu, and J. Zhou. Gait recognition in the wild: A benchmark. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 14789–14799, 2021.