Long range gait matching using 3D body fitting with gait-specific motion constraints

Mauricio Pamplona Segundo, Cole Hill, Sudeep Sarkar
Computer Science and Engineering, University of South Florida, Tampa, Florida, USA
{mauriciop,coleh,sarkar}@usf.edu

Abstract

Drawing upon the many works in estimating the 3D human body shape and motion in images and video, some have recently proposed using 3D human models for gait recognition to overcome viewpoint variation. However, the problem is the fit quality, particularly in the motion aspects. While the overall 3D shape aspects look good, the limb configurations over the video only capture walking in some cases. To address this problem, we build on the recent trend of fitting a 3D deformable body model – the SMPL model – to gait videos using deep neural networks to obtain disentangled shape and pose representations for each frame. This work is the first to use adversarial training for gait recognition, and it helps us to enforce motion consistency in the network output. To this end, a subset of walking activity instances from the AMASS mocap dataset serves as the natural motion distribution. We benchmark our solution to the state-of-the-art on the well-known USF HumanID and CASIA-B datasets in terms of variations concerning viewpoint, clothing, carrying condition, walking surface, and time. We are either the best or close to the best-reported performance on these datasets. We demonstrate the quality of the 3D fitted models for gait recognition on the newly constructed IARPA BRIAR dataset of IRB consented 375 subjects with videos taken at 100m, 200m, 400m, and 500m. We are among the first to report gait recognition estimates at long range.

1. Introduction

Gait or walking style is a biometric trait involving many aspects of the human body, including the skeleton and musculature. In this work, we consider the problem of matching two videos and verifying if they are from the same person based on gait as captured in the video, taken from long ranges. Unlike most biometrics, such as fingerprint and iris, gait can be used in long-range recognition, making it valuable for many applications, such as securing sensitive locations, remotely identifying poachers of endangered species, and locating lost mentally impaired individuals or children. However, scientific studies have yet to benchmark the ability to perform gait matching at long ranges, where atmospheric effects can result in poor-quality videos.

Gait recognition based on 2D silhouettes is the most prevalent in the literature. These methods work well in controlled conditions (e.g., indoor environments) but drop in performance when silhouettes cannot be reliably extracted, for instance, in outdoor environments, from moving cameras, or at long range with atmospheric turbulence. We build upon recent works [17, 1] that fit a deformable 3D body model. To be precise, we fit the *Skinned Multi-Person Linear* (SMPL) model [24] to a video sequence and use the model parameters for gait recognition. The *Video Inference for Human Body Pose and Shape Estimation* (VIBE) approach [15] is an approach that can fit the SMPL model to a video. The VIBE-fitted SMPL model delivers good body shape, but not pose; the motion coherence over a gait sequence is not guaranteed.

We modify the VIBE's training procedure [15] to utilize an adversarial learning approach to enforce motion consistency in the network output and train it end-to-end for biometric discrimination. However, instead of using a general motion discriminator, as was done in VIBE, we propose using a gait-specific one. Its task is to predict whether a body pose sequence was created by the 3D fitting network or comes from an accurate gait distribution. This distribution consists of a subset of walking activity instances from the *Archive of Motion Capture As Surface Shapes* (AMASS) [25] *motion capture* (mocap) dataset.

This work contributes to the literature with the following: (i) We present the first gait-specific, adversarially regularized, 3D model fitting approach that preserves discriminability among humans based on body shape and poses. (ii) We are among the first to report long-range gait recognition estimates on the newly constructed *IARPA Biometric Recognition and Identification at Altitude and Range* (BRIAR) dataset of 375 IRB-consented subjects with videos taken at 100m, 200m, 400m, and 500m.

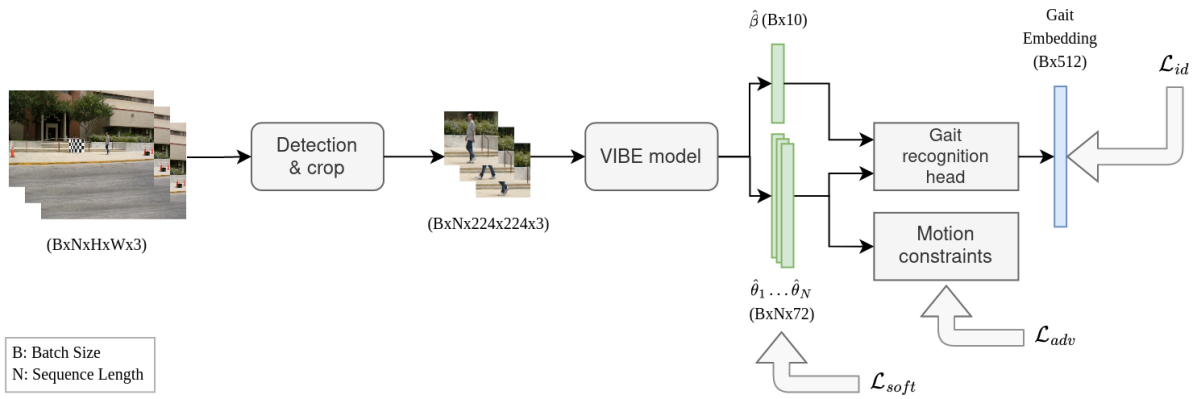


Figure 1. Overview of *GaitVIBE* and novel training losses for (i) adversarial training for gait motion and (ii) a soft-loss to preserve identity.

2. Related works

Nixon *et al*. [26] give an in-depth review of pre-deep learning gait recognition. We focus on deep learning era.

Appearance-based approaches extract a discriminative representation directly from sequences of body silhouette images in good-quality videos, static backgrounds, and well-segmented human shapes. Hossaine and Chetty [13] showed one of the first appearance-based works with deep learning using a *Restricted Boltzmann Machine* (RBM) but with limitations due to the lack of spatial bias in RBMs. Shiraga *et al*. [32] proposed one of the first works in gait recognition using *Convolutional Neural Networks* (CNNs), whose kernel-based structure inherently preserves local spatial relationships. This method used the very old *Gait Energy Image* (GEI) representation, concomitantly proposed by Liu and Sarkar [22] and Han and Bhanu [11]. The representation compresses the gait sequence into a single frame and loses all temporal information. Instead of relying on a handcrafted input, Wu *et al*. [40] fed randomly picked raw silhouettes directly to a CNN and combined the obtained feature vectors to produce a gait representation. Wolf *et al*. [38] presented a 3D convolutional approach shortly after that. They also took the silhouettes directly as input and relied on 3D kernels that can exploit spatio-temporal information. Song *et al*. [33] showed the first end-to-end CNN-based gait recognition pipeline with GaitNet, which extracts silhouettes and discriminative features from RGB imagery with a single network. Chao *et al*. [3] introduced the set-based approach, performing recognition on sets of frames rather than a sequence. Fan *et al*. [9] showed that we could improve performance by using horizontal segments of the silhouette. Although these works achieve high accuracy in controlled datasets, their performance considerably drops with unconstrained scenarios.

Model-based approaches first fit a human body model (skeleton or mesh) to the input image frames and then derive features for recognition from the obtained model parameters. Due to the difficulties in extracting accurate 3D information from imagery, model-based approaches have been limited to state-of-the-art human body pose and shape models and networks that estimate their parameters from images. First, there was progress in accurate pose estimation. Cao *et al*. [2] enabled Liao *et al*. [18] to create the first model-based network, which used both a CNN and a *Long Short-Term Memory* (LSTM) network to create features for gait recognition. Teepe *et al*. [36] further improved skeleton pose-based gait recognition through the use of the *Graph Convolutional Network* (GCN) and an enhanced pose estimation with HRNet [34]. The release of the SMPL by Looper *et al*. [24] provided a realistic model for both human shape and pose. Li *et al*. [17] published an end-to-end gait recognition system using SMPL features as input. Zheng *et al*. [42] introduced a fusion of appearance and model-based approaches, extracting appearance-based features from silhouettes and normalizing them for pose and view utilizing SMPL parameters. Bansal *et al*. [1] integrated SMPL features with the *Vision Transformer* (ViT) [8] to fuse appearance and model information for recognition.

Our proposed approach is an end-to-end model as in Li *et al*. [17]. The major difference is our use of gait-specific adversarial training to enhance motion realism with fully disentangled pose and shape representations, that are biometrically discriminative. We do not rely on body skeletons that blend these two pieces of information. There are other differences, such as their use of the older *Human Mesh Recovery* (HMR) [14] model versus our use of the newer VIBE model [15]. And finally, we use a Transformer network [37] to fuse temporal pose information for recognition.

3. Solution Architecture

Our gait recognition architecture, a gait-specific VIBE (*GaitVIBE*), is depicted in Figure 1. The recognition process can be divided into three stages: (1) person tracking, (2) 3D body reconstruction, and (3) gait description for matching. We provide details about each of these stages.

3.1. Person tracking

We need to locate the person in each frame to use a video for gait recognition. We start by detecting people in each video frame using the *Faster Region-based CNN* (Faster R-CNN) detector [28] available in the Detectron2 library [39]. The outcome of this method is a list of bounding boxes, each enclosing one person in the input frame. Given the controlled nature of the video datasets used in this work (at most one person per video), we keep only the largest detection per frame. Although we chose a state-of-the-art detection approach, different factors impacted its results for this task: noisy bounding boxes, false detections, missed detections, and multiple people in some frames of some videos. To address those issues, we create three different time series of center coordinates and size values from all selected detections in a video. Next, we fit third-degree polynomials to a 150-frame sliding window over these time series with a step size of 50 frames to smooth out the tracks. Moreover, we take the average of the smoothed values for frames that appear in multiple windows. Finally, we crop the minimum enclosing square around the person's bounding box for the entire track and resize the resulting images to 224×224 pixels to form the normalized video.

3.2. 3D body reconstruction

We use the VIBE architecture [15] unchanged to regress parameters from the SMPL model [24] and thus obtain a 3D body representation for each frame of a normalized video.

The SMPL model is a 3D body prior that encodes shape and pose in a disentangled manner. The shape is represented by a 10-dimensional array β containing the weights of the ten most significant principal components obtained from thousands of full-body 3D scans. These weights are used to deform a canonical body mesh into any person's physique. The pose is represented by a 72-dimensional array θ, which contains three rotation angles for 23 body joints plus a global orientation. The SMPL model receives a pair (β, θ) and produces a mesh with 6890 vertices and 13776 faces. It is worth saying that shape and pose are fully disentangled, so one does not affect the other.

VIBE is a general 3D body fitting approach that retrieves a pair $(\hat{\beta}_i, \hat{\theta}_i)$ for the i-th video frame. It is composed of a ResNet-50 backbone [12] that extracts relevant visual features from each frame, a 2-layer *Gated Recurrent Unit* (GRU) [6] that takes care of temporal consistency, and an iterative *Multilayer Perceptron* (MLP) [29] that regresses $\hat{\beta}_i$ and $\hat{\theta}_i$ from the temporally adjusted features. During training, VIBE relies on body shape and pose supervision (*e.g.*, body joint re-projection error, SMPL regression error) and on adversarial motion learning. The latter confronts real examples from a dataset of mocap sequences (AMASS [25]) with the pose sequences created by VIBE to learn how to reproduce realistic motion for general activities.

We average the estimated shape parameter over the entire sequence to provide a single $\hat{\beta}$. Figure 3 renders some of the obtained SMPL fittings from normalized videos using the proposed approach.

3.3. Gait description for matching

The shape $\hat{\beta}$ along with the sequence $\hat{\theta}_1 \dots \hat{\theta}_N$ of pose parameters for a video of N frames serve as input to the gait recognition head, which maps them into a 512-dimensional embedding that allows us to discriminate between individuals. Its architecture is shown in Figure 2.

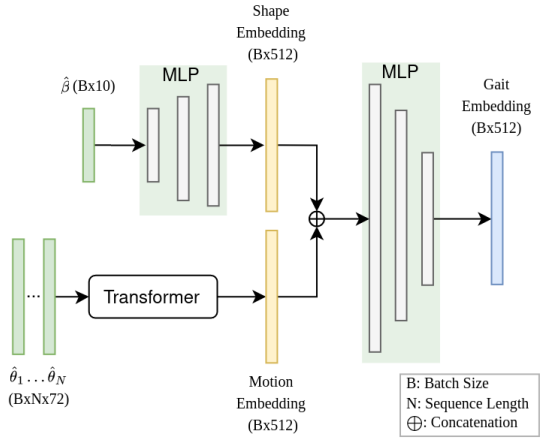


Figure 2. Gait recognition head in our GaitVIBE model.

The $\hat{\beta}$ parameter is projected from 10 to 512 dimensions using a 2-layer MLP with 2048 hidden units. Each layer of the MLP consists of a linear layer followed by ReLU activation and batch normalization. We see this network branch as a decoder that extracts the most discerning body measurements from the SMPL body representation to form a shape embedding. Motion is processed in parallel, starting by converting joint orientation angles in $\hat{\theta}_i$ to 3×3 rotation matrices ($\mathbb{R}^{69} \mapsto \mathbb{R}^{207}$). This is done to avoid abrupt value transitions that occur when using Euler angles (*e.g.*, the angles $0°$ and $359°$ are spatially close but numerically distant) and impact the network performance. The set of rotation matrices for each pose is then projected into 512 dimensions using a linear layer. Following the standard state-of-the-art practice, we add a 512-dimensional trainable token to the beginning of the pose sequence to gather information from the entire video. We feed the updated sequence to a 6-block Transformer encoder [37], each block with eight attention heads and 2048 hidden units, and take the output 512-dimensional embedding corresponding to the added token as the motion embedding. To produce the final embedding, we concatenate the disentangled shape and motion embeddings and pass them through another 2-layer MLP with 2048 hidden units to produce a single 512-dimensional gait embedding. The output of the last layer goes through

an L2 normalization so that those embeddings are unit vectors. For matching, we compare any two gait embeddings using the cosine similarity.

4. Training details

Our model is trained end-to-end using the following combination of loss functions:

$$\mathcal{L} = \mathcal{L}_{id} + \lambda_{soft}\mathcal{L}_{soft} + \lambda_{adv}\mathcal{L}_{adv} \quad (1)$$

with λ_{soft} and λ_{adv} being scaling factors to balance the three loss terms (λ_{soft} was empirically set to 0.06, and λ_{adv} to 0.5). Details about each of these loss functions are given in the following sections.

4.1. Soft reconstruction loss

A soft supervision is applied to the VIBE backbone by using shape and pose estimates from the original VIBE model as a pseudo ground truth:

$$\mathcal{L}_{soft} = ||\beta' - \hat{\beta}||_2 + \lambda_{pose}\sum_{i=1}^{N}||\theta'_i - \hat{\theta}_i||_2 \quad (2)$$

with $(\hat{\beta}, \hat{\theta}_1 \dots \hat{\theta}_N)$ being the network output, $(\beta', \theta'_1 \dots \theta'_N)$ the pseudo ground truth, and λ_{pose} a scaling factor to balance the importance of shape and pose (λ_{pose} was empirically set to 1000). This loss ensures that the model retains valid SMPL representations of shape and pose, as the gait datasets used in our experiments (*USF HumanID Gait Dataset* (USF HumanID), *CASIA Gait Dataset B* (CASIA-B), and BRIAR) do not have any kind of annotation other than the person identity.

Li *et al*. [17] also used pseudo ground truth for training but limited to pose estimates computed with HMR [14].

4.2. Identity loss

To create discrimination representations for gait videos, we apply the Triplet Semi-Hard Loss [31] to the final gait embeddings:

$$\mathcal{L}_{id} = \sum_{a,p,n\in T}\max(\|x^a - x^p\|_2^2 - \|x^a - x^n\|_2^2 + \alpha, 0) \quad (3)$$

where x^a, x^p, and x^n form a triplet of embeddings (x^a and x^p belong to the same subject, and x^n belongs to a different subject), α is the margin term (α was empirically set to 1.0), and T is the set with all semi-hard triplets in a training batch [31].

4.3. Gait-specific adversarial learning

VIBE [15] uses an adversarial loss to produce realistic motion sequences for general activities. However, in this work, we are only interested in walking activities for gait recognition purposes. Thus, we carried out gait-specific adversarial learning by restricting our real motion distribution to a subset of 811 walking instances from the 11000+ mocap sequences in the AMASS dataset [25]. With that, we use the following loss term to enforce gait-specific motion patterns in the output of our network:

$$\mathcal{L}_{adv} = \mathbb{E}_{\hat{\theta}_1 \dots \hat{\theta}_N \sim p_G}\left[(D(\hat{\theta}_1 \dots \hat{\theta}_N) - 1)^2\right] \quad (4)$$

with p_G being the distribution of pose sequences created by our network and D being a motion discriminator with the same architecture used in VIBE (a 2-layer GRU followed by a 2-layer MLP-based self-attention mechanism and a linear layer for classification). The discriminator itself is trained using the following loss term:

$$\begin{aligned}\mathcal{L}_{disc} = \mathbb{E}_{\theta_1 \dots \theta_N \sim p_R}\left[(D(\theta_1 \dots \theta_N) - 1)^2\right] + \\ \mathbb{E}_{\hat{\theta}_1 \dots \hat{\theta}_N \sim p_G}\left[D(\hat{\theta}_1 \dots \hat{\theta}_N)^2\right]\end{aligned} \quad (5)$$

with p_R being the distribution of real pose sequences.

Li *et al*. [16] also used adversarial learning to enhance their reconstruction results. However, as in VIBE, they use the entire AMASS dataset as the source of real motion examples, which results in a discriminator that is not gait-specific. They also use a shape discriminator, which is not needed in our case thanks to the soft reconstruction loss (Section 4.1). Besides, as the SMPL shape parameters are actually weights of a PCA model, if necessary, we could regularize their values in more practical ways (*e.g.*, Mahalanobis distance).

4.4. Other information

The model is implemented in Python 3.8 using Pytorch[27]. We train the 3D reconstruction and gait description modules in an end-to-end fashion. First, we initialize the 3D reconstruction module with the publicly available VIBE weights and randomly initialize all other network parameters (e.g., gait head and discriminator). Next, we optimize the entire architecture using the Adam algorithm for up to 100 epochs with a batch size of 24 60-frame video segments and a learning rate of 5×10^{-5}, except for the discriminator, which is updated at every training iteration of the main network with a learning rate of 10^{-4}.

5. Datasets and experiments

We performed our experiments using the CASIA-B [41], USF HumanID [30], and BRIAR[1] datasets.

The CASIA-B dataset is from 124 subjects, each with ten sequences split into three covariates: 6 normal (NM), 2 carrying a bag (BG), and 2 wearing a coat (CL). They recorded each sequence from 11 view angles in a controlled

[1] https://www.iarpa.gov/research-programs/briar

indoor setting with a simple background and even lighting. We followed the standard evaluation protocol from previous works. We used the first 74 subjects for training and the rest for testing. At test time, sequences NM #1-4 compose the gallery, and the remaining ones form three sets of probes: NM #5-6, BG #1-2, CL #1-2.

The USF HumanID dataset contains gait sequences from 126 (IRB consented) subjects outdoors with variation in the surface, time, shoe, briefcase, clothing, and view (1884 sequences in all). The same subjects are used for both training and testing the network. The evaluation protocol consists of 12 experiments (Exp. A-L) pairing videos from a fixed gallery set to a probe set with a specific composition of covariates. Only 1080 out of the 1884 videos are used in these experiments. We used the gallery set plus all unused videos whose combination of covariates does not appear in any probe set for training (a total of 677 videos).

The BRIAR dataset is being put together by Oak Ridge National labs within the IARPA BRIAR program for public release. It is divided into training and testing sets. The training set consists of 158 subjects with an average of 82 videos per subject. Most subjects contain enrollment videos from ten different viewpoints captured indoors and probe videos captured outdoors at multiple distances (close range, 100m, 200m, 400m, and 500m) with different viewpoints, clothing, and atmospheric turbulence conditions. Some of these conditions are exemplified in Figure 3. The testing set contains 375 subjects with enrollment videos that form the gallery. Its evaluation protocol provides 1599 whole-body video segments with walking activities as probes, which include 44 different identities. No individual is in both training and testing sets.

To our knowledge, the BRIAR dataset is also the first dataset to provide camera distance as a covariate for gait recognition. Long-range gait recognition introduces a unique challenge not present in datasets currently used in the literature. The distance between the camera and the subject causes atmospheric distortion, impairing recognition performance. Also, unlike other large gait recognition datasets (OU-MVLP [35], Gait3D [43], GREW [44]), the BRIAR dataset provides raw video data. The others distribute just sequences of silhouette masks, so we cannot use them. Our approach works directly on the video texture.

5.1. CASIA-B results

For CASIA-B, we report the average Rank-1 identification rates for each probe subset (NM, BG, and CL) when using different views for gallery and probe. Table 1 compares CASIA-B results reported in the literature for different gait recognition methods. *GaitVIBE* stands for the results of the proposed model without any motion constraints, while *GaitVIBE + Adv* includes the adversarial learning strategy. As CASIA-B is a controlled dataset, silhouette extraction tends to be more precise. As a consequence, methods that rely on silhouettes achieve higher accuracy. Still, *GaitVIBE* and *GaitVIBE + Adv* obtain competitive results for the NM and BG subsets, showing their potential to recognize individuals from different views. The CL subset, however, reveals one of the main limitations of SMPL-based approaches: shape fitting is negatively affected by clothing variations. In this case, we can observe that GaitGraph [36] obtains higher performance among non-silhouette-based methods, as it only uses body skeleton information that is more robust to changes in clothes.

	Probe	NM	BG	CL
With silhouettes	GaitNet [33]‡	91.6	85.7	58.9
	GaitSet [3]‡	95.0	87.2	70.4
	GaitPart [9]‡	96.2	91.5	<u>78.7</u>
	GaitGL [20]‡	<u>97.4</u>	**94.5**	**83.6**
	Li *et al*. [17]†‡	**97.9**	<u>93.1</u>	77.6
Without silhouettes	PoseGait [19]‡	63.8	45.5	32.0
	GaitGraph [36]‡	87.7	74.8	**66.3**
	GaitVIBE (our)	<u>93.7</u>	<u>88.0</u>	<u>60.1</u>
	GaitVIBE + Adv (our)	**94.9**	**90.0**	49.5

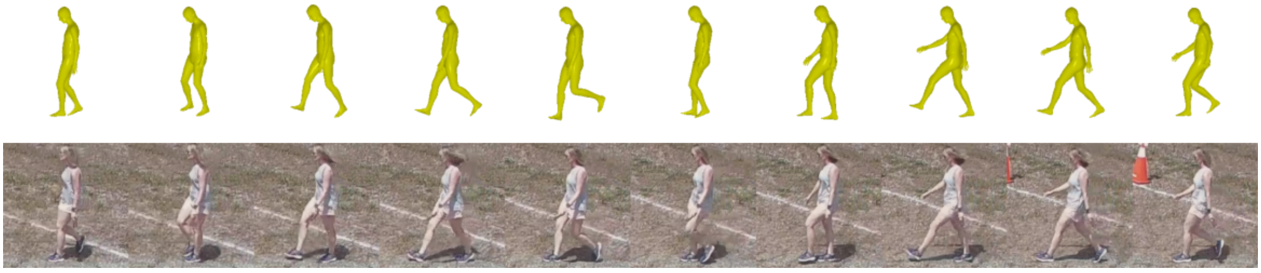
Table 1. Comparison of gait recognition approaches on CASIA-B. Best two Rank-1 results for each category are indicated in bold and underlined characters. †Works that use silhouettes for training only. ‡Results as reported in the literature.

5.2. USF HumanID results

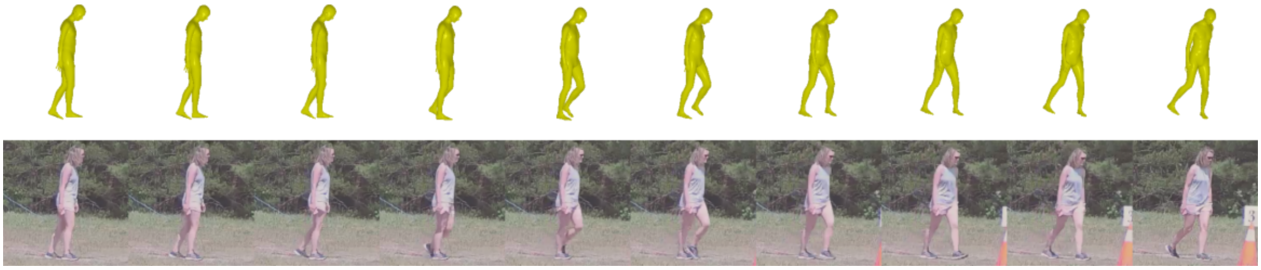
For USF HumanID, we report the Rank-1 identification rate for each probe experiment (Exp. A-L). To provide a point of reference, we include the best USF HumanID performance reported in the literature [7, 10] and the results of three state-of-the-art approaches, GaitGraph [36], GaitPart [9], and GaitSet [3]. Since their publications did not report results on USF HumanID, we used their implementations from the OpenGait repository[2] to train these approaches using our USF HumanID setup. GaitPart and GaitSet require body silhouette images, so we use a background matting method [21] that segments foreground objects to extract those. In our experiments, silhouettes extracted via background matting were better suited for gait recognition purposes than the ones extracted via body segmentation (HTC [4], FCN [23], DeepLabV3 [5]).

Table 2 shows the Rank-1 identification rates of different approaches on USF HumanID. It shows that the proposed model outperforms prior works in all experiments, including the challenging experiments with changes in surface (D-G) and acquisition time (K-L), showing the advantage of using model-based approaches in unconstrained scenarios. Although *GaitVIBE* and *GaitVIBE + Adv* achieve nearly the same result in Table 2, we observed in unreported ex-

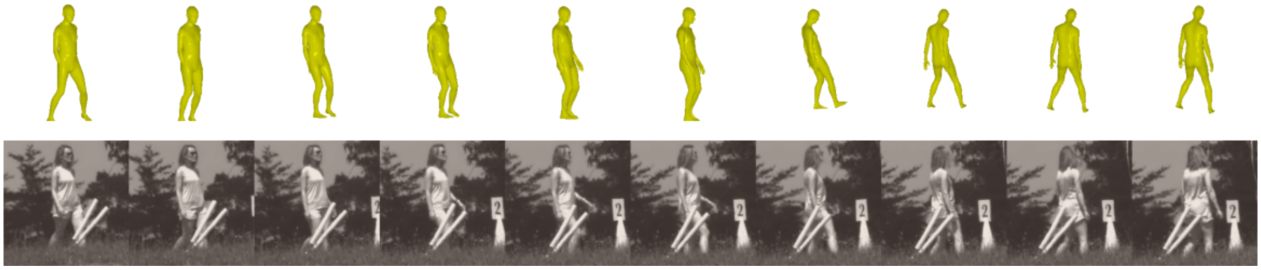
[2] https://github.com/ShiqiYu/OpenGait



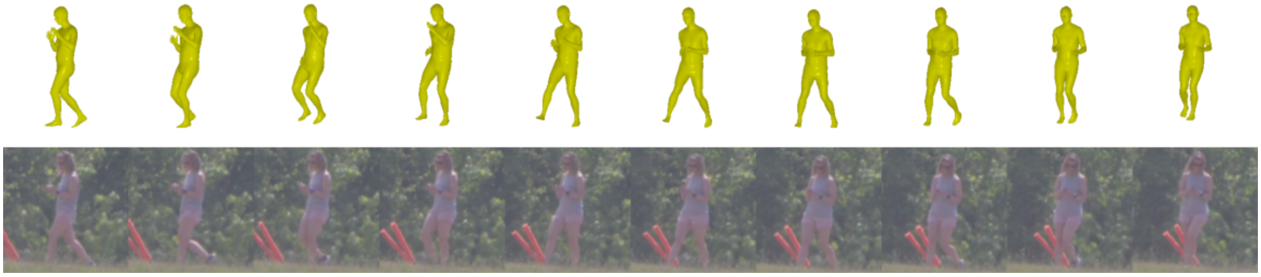
(a) Close range with high elevation



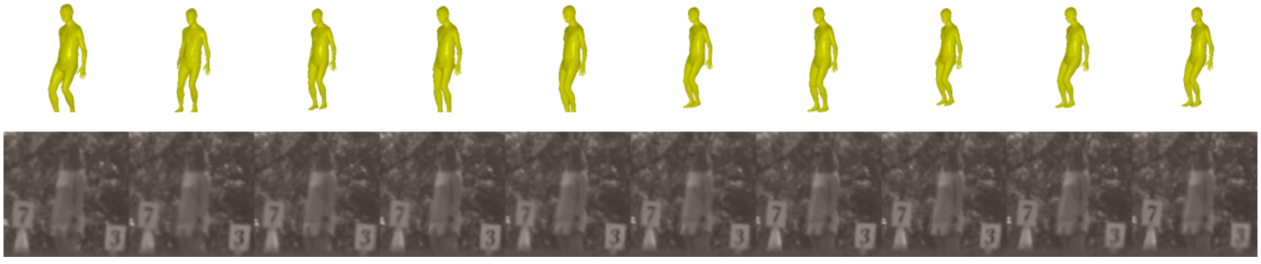
(b) 100m



(c) 200m with body occlusion caused by external object and feet occlusion caused by ground vegetation



(d) 400m with feet occlusion caused by ground vegetation



(e) 500m with lower legs occlusion caused by ground vegetation

Figure 3. Input sequences from the BRIAR dataset at different distances and conditions (viewpoint, sensor, occlusion) with their respective 3D fitting results.

	Probe	A	B	C	D	E	F	G	H	I	J	K	L
With silhouettes	Deng *et al.* [7]‡	97	98	96	92	88	81	82	95	92	82	76	73
	Guan *et al.* [10]‡	**100**	95	94	73	73	55	64	97	99	94	42	42
	GaitPart [9]†	95.90	81.50	81.50	85.60	77.60	84.70	74.10	89.80	86.20	89.00	84.80	90.90
	GaitSet [3]†	**100**	94.40	94.40	93.20	93.10	93.20	89.70	98.30	93.10	97.50	81.80	90.90
Without silhouettes	GaitGraph [36]†	35.20	74.10	13.00	28.80	19.00	23.70	12.10	63.60	50.00	28.00	63.60	24.20
	GaitVIBE (our)	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**97**	**97**
	GaitVIBE + Adv (our)	**100**	**100**	**100**	**100**	**100**	**100**	**100**	99.2	**100**	**100**	**97**	**97**

Table 2. Comparison of gait recognition approaches on USF HumanID. Best two Rank-1 results for each category are indicated in bold and underlined characters. †Results computed using OpenGait. ‡Results as reported in the literature.

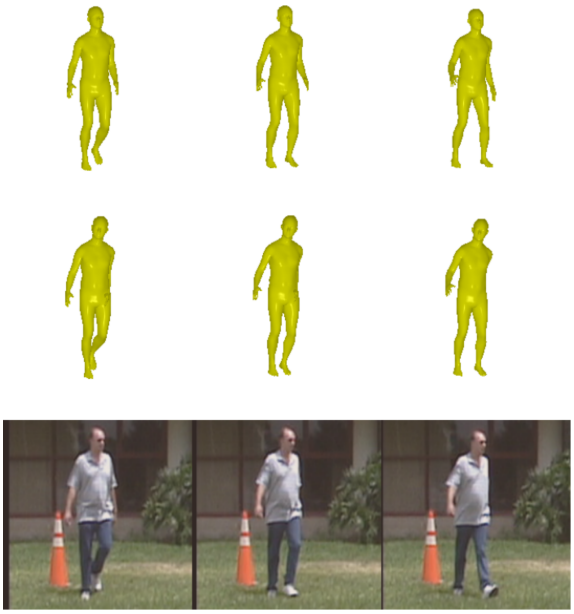


Figure 4. 3D body fitting results for a USF HumanID video (bottom) using VIBE (top) and *GaitVIBE + Adv* (center). Observe the improvements in the position of arms and legs obtained by *GaitVIBE + Adv*, which recovers the walking characteristics of the sequence. Meanwhile, the original VIBE output better approximates a standing position.

periments that *GaitVIBE + Adv* has advantages when training for less time or when parts of the VIBE model are kept frozen (*e.g.*, the ResNet-50 backbone), which are strong indicators of transfer learning capabilities.

Both GaitSet [3] and GaitPart [9] show a considerable drop in performance when evaluated under outdoor conditions. This indicates that silhouette-based models may lend too much weight to details with low repeatability, such as shoe shape or shadow artifacts, which are not a problem in controlled datasets such as CASIA-B. Our whole body representation and disentangled pose force the model to be more robust to these changes. These results corroborate the claim that appearance-based approaches suffer in more realistic scenarios, such as outdoor environments, due to the lack of proper silhouettes. Although the results for GaitSet and GaitPart were computed using OpenGait, this implementation's performance for CASIA-B is on par or better than the original results. Finally, GaitGraph [36] performs poorly on USF HumanID and demonstrates the negative impact of relying only on skeletons for recognition since its descriptors lose important discriminative information. Figure 4 illustrates how our *GaitVIBE + Adv* approach can improve the 3D reconstruction of a walking video thanks to the adversarial loss. Even though we do not have 3D ground truth for the videos used for training, we can still enhance VIBE's output by imposing gait-specific motion constraints.

5.3. BRIAR results

The BRIAR dataset has indoor gallery (see Figure 5) and outdoor probes (see Figure 3), which combined with all other dataset covariates (including long-range acquisition, atmospheric turbulence, clothing, and occlusion) generates a much more challenging evaluation environment for gait recognition. For this dataset, we report verification results in the form of True Positive Rates (TPR) at 1% False Acceptance Rate (FAR) and identification results for Rank-K ($K = \{1, 5, 10\}$) for *GaitVIBE + Adv* and GaitSet [3]. Note that for many probes, silhouette detection failed and/or the 3D body estimation failed. This is evidence of the difficult nature of the data. For this reason, we only use the 1493 probes that could be processed by at least one of the chosen methods. We set the cosine similarity between probe videos that were not correctly processed and all gallery identities to -1 (lowest possible value).

We first evaluate the impact of the number of gallery identities on recognition performance. As presented in Table 3, Rank-1 rates are three times higher when using only the 44 probe identities in the gallery compared to the full 375-subject gallery. The scale of this reduced gallery experiment is similar to CASIA-B's test set. Still, the performance gap is immense even if we consider CASIA-B's worst-performing subset (CL), which highlights the complexity of the BRIAR dataset. Also, compared to GaitSet, we obtain better verification and identification rates, which reaffirms the advantage of model-based approaches in unconstrained scenarios.

But the main innovation on BRIAR is the inclusion of the distance covariate, so we report in Table 4 the results for

	Gallery size	44	100	200	300	375
GaitSet [3]	**TPR**	7.6	7.2	7.5	7.7	7.8
	Rank-1	15.3	11.6	5.8	4.6	4.2
	Rank-5	41.1	30.9	19.8	16.1	13.8
	Rank-10	51.4	43.4	30.2	23.8	21.7
GaitVIBE+Adv	**TPR**	**13.5**	**16.7**	**16.9**	**18.1**	**18.1**
	Rank-1	**23.0**	**15.7**	**9.4**	**7.4**	**6.8**
	Rank-5	**58.1**	**47.4**	**32.2**	**27.0**	**23.0**
	Rank-10	**72.1**	**58.5**	**45.6**	**38.1**	**33.9**

Table 3. Gait recognition results on BRIAR with different gallery sizes for all 1493 probe videos. Best results for each category are indicated in bold.

Distance (m)	**Close**	**100**	**200**	**400**	**500**
TPR	20.1	23.9	4.1	15.8	17.8
Rank-1	8.8	8.7	1.0	4.7	8.6
Rank-5	26.9	31.2	7.7	19.7	21.6
Rank-10	40.5	43.8	15.9	29.0	28.6
# samples	457	356	195	279	185

Table 4. Gait recognition results on BRIAR at different distances with a gallery size of 375 subjects using *GaitVIBE + Adv*.

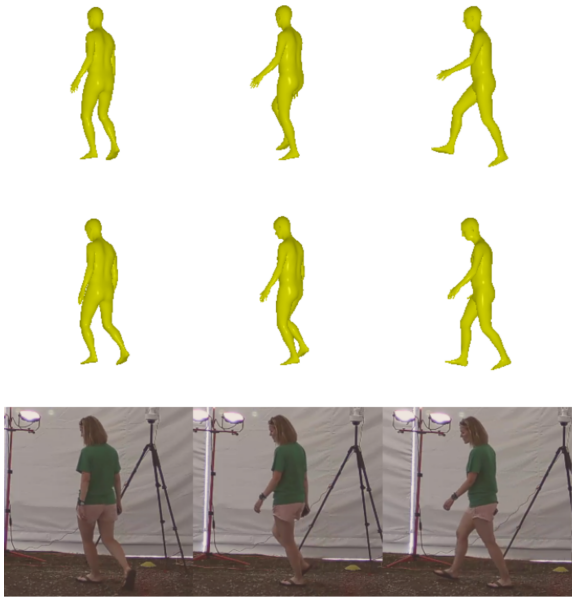


Figure 5. 3D body fitting results for a BRIAR video (bottom) using VIBE (top) and GaitVIBE + Adv (center). *GaitVIBE + Adv* enhances the realism of the walking sequence of the sequence, with both arms and legs adjusted to create a more fluid motion.

the different camera distances available using *GaitVIBE + Adv*. Examples of the different distances and the 3D fitting results obtained by the proposed approach are shown in Figure 3. Close-range videos may be closer to the captured individuals but also include high elevation to diverge from the enrollment conditions (cameras at ground level). For this reason, videos captured 100 meters away from the subjects are the ones with the closest resemblance to the gallery and, consequently, are the ones with the highest recognition performance in terms of verification and identification. Overall, we see a steady drop in performance as conditions move away from ideal, except for videos from the camera at 200 meters. In this specific case, a challenging combination of covariates (sensor color, leg and body occlusions) severely impacted the results, making it the worst-performing subset.

6. Discussion and conclusion

This paper presents a gait recognition method utilizing 3D body fitting with motion constraints via gait-specific adversarial training. We provide qualitative results that show the improvement in gait motion naturalness obtained thanks to the adversarial training. Quantitative results show that, although the proposed approach does not outperform state-of-the-art appearance-based approaches in indoor datasets like CASIA-B, it is superior to them in outdoor datasets like the USF HumanID and BRIAR. This happens because the extracted silhouettes' quality directly affects appearance-based approaches, and too many factors deteriorate them in unconstrained environments. When talking about long-range recognition, silhouette extraction becomes even harder due to additional artifacts such as bad camera focus, atmospheric turbulence, and the absence of a background model. Although model-based approaches are not necessarily a solution to those issues, they rely on prior knowledge that assists in smoothing out noisy information while maintaining meaningful features for recognition. Our results on BRIAR show that its long-range evaluation is far more challenging than other publicly available video datasets, even though recognition at a distance is constantly mentioned as an inherent advantage of gait over other biometrics. To boost recognition performance, future works must address some of the introduced challenges, such as different enrollment and probe conditions, different sensing technologies, and atmospheric turbulence caused by long-range acquisition. In this context, this paper shows that model-based approaches can serve as a better starting point when silhouettes cannot be reliably extracted.

Acknowledgements

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 2022-21102100003. 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 U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

[1] Vaibhav Bansal, Gian Luca Foresti, and Niki Martinel. Cloth-changing person re-identification with self-attention. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 602–610, 2022.

[2] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser 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.

[3] 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.

[4] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. Hybrid task cascade for instance segmentation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4969–4978, 2019.

[5] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2018.

[6] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, 2014.

[7] Muqing Deng, Cong Wang, Fengjiang Cheng, and Wei Zeng. Fusion of spatial-temporal and kinematic features for gait recognition with deterministic learning. *Pattern Recognition*, 67:186–200, 2017.

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[9] 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.

[10] Yu Guan, Chang-Tsun Li, and Fabio Roli. On reducing the effect of covariate factors in gait recognition: A classifier ensemble method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(7):1521–1528, 2015.

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

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

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

[14] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Computer Vision and Pattern Recognition (CVPR)*, 2018.

[15] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5253–5263, 2020.

[16] Xiang Li, Yasushi Makihara, Chi Xu, and Yasushi Yagi. Multi-view large population gait database with human meshes and its performance evaluation. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2022.

[17] 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*, 2020.

[18] Rijun Liao, Chunshui Cao, Edel B Garcia, Shiqi Yu, and Yongzhen Huang. Pose-based temporal-spatial network (ptsn) for gait recognition with carrying and clothing variations. In *Chinese conference on biometric recognition*, pages 474–483. Springer, 2017.

[19] 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.

[20] 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*, pages 14648–14656, 2021.

[21] Shanchuan Lin, Andrey Ryabtsev, Soumyadip Sengupta, Brian Curless, Steve Seitz, and Ira Kemelmacher-Shlizerman. Real-time high-resolution background matting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8758–8767, 2021.

[22] Z. Liu and S. Sarkar. Simplest representation yet for gait recognition: Averaged silhouette. In *International Conference on Pattern Recognition*, volume 1, pages 211–214, 2004.

[23] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440, 2015.

[24] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015.

[25] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of*

the IEEE/CVF international conference on computer vision, pages 5442–5451, 2019.

[26] Mark S Nixon, Tieniu Tan, and Rama Chellappa. *Human identification based on gait*, volume 4. Springer Science & Business Media, 2010.

[27] 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 *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[28] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

[29] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. *Learning Internal Representations by Error Propagation*, pages 318–362. MIT Press, Cambridge, MA, USA, 1986.

[30] Sudeep Sarkar, P Jonathon Phillips, Zongyi Liu, Isidro Robledo Vega, Patrick Grother, and Kevin 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.

[31] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, 2015.

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

[33] Chunfeng Song, Yongzhen Huang, Yan Huang, Ning Jia, and Liang Wang. Gaitnet: An end-to-end network for gait based human identification. *Pattern recognition*, 96:106988, 2019.

[34] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5693–5703, 2019.

[35] Noriko Takemura, Yasushi Makihara, Daigo Muramatsu, Tomio Echigom, and Yasushi Yagi. Gait recognition in the wild: A benchmark. *IPSJ Transactions on Computer Vision and Applications*, 10(4), 2018.

[36] Torben Teepe, Ali Khan, Johannes Gilg, Fabian Herzog, Stefan Hörmann, and Gerhard Rigoll. Gaitgraph: Graph convolutional network for skeleton-based gait recognition. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 2314–2318. IEEE, 2021.

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, 2017.

[38] 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.

[39] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. `https://github.com/facebookresearch/detectron2`, 2019.

[40] Zifeng Wu, Yongzhen Huang, and Liang Wang. Learning representative deep features for image set analysis. *IEEE Transactions on Multimedia*, 17(11):1960–1968, 2015.

[41] 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.

[42] Jinkai Zheng, Xinchen Liu, Wu Liu, Lingxiao He, Chenggang Yan, and Tao 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.

[43] Jinkai Zheng, Xinchen Liu, Wu Liu, Lingxiao He, Chenggang Yan, and Tao Mei. Gait recognition in the wild with dense 3d representations and a benchmark. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[44] Zheng Zhu, Xianda Guo, Tian Yang, Junjie Huang, Jiankang Deng, Guan Huang, Dalong Du, Jiwen Lu, and Jie Zhou. Gait recognition in the wild: A benchmark. In *IEEE International Conference on Computer Vision (ICCV)*, 2021.