Gait Recognition via Disentangled Representation Learning

Ziyuan Zhang, Luan Tran, Xi Yin, Yousef Atoum, Xiaoming Liu
Michigan State University
{zhang835, tranluan, yinxi1, atoumyou, liuxm}@msu.edu

Jian Wan, Nanxin Wang
Ford Research and Innovation Center
{jwan1, nwang1}@ford.com

Abstract

Gait, the walking pattern of individuals, is one of the most important biometrics modalities. Most of the existing gait recognition methods take silhouettes or articulated body models as the gait features. These methods suffer from degraded recognition performance when handling confounding variables, such as clothing, carrying and view angle. To remedy this issue, we propose a novel AutoEncoder framework to explicitly disentangle pose and appearance features from RGB imagery and the LSTM-based integration of pose features over time produces the gait feature. In addition, we collect a Frontal-View Gait (FVG) dataset to focus on gait recognition from frontal-view walking, which is a challenging problem since it contains minimal gait cues compared to other views. FVG also includes other important variations, e.g., walking speed, carrying, and clothing. With extensive experiments on CASIA-B, USF and FVG datasets, our method demonstrates superior performance to the-state-of-the-arts quantitatively, the ability of feature disentanglement qualitatively, and promising computational efficiency.

1. Introduction

Biometrics measures people's unique physical and behavioral characteristics to recognize the identity of an individual. Gait [35], the walking pattern of an individual, is one of the biometrics modalities, e.g., face, fingerprint, and iris. Gait recognition has the advantage that it can operate at a distance without user cooperation. Also, it is difficult to camouflage. Due to these advantages, gait recognition is applicable to many applications such as person identification, criminal investigation, and healthcare.

As other recognition problems in vision, the core of gait recognition lies in extracting *gait-related features* from the video frames of a walking person, where the prior approaches are categorized into two types: appearance-based and model-based methods. The appearance-based methods such as Gait
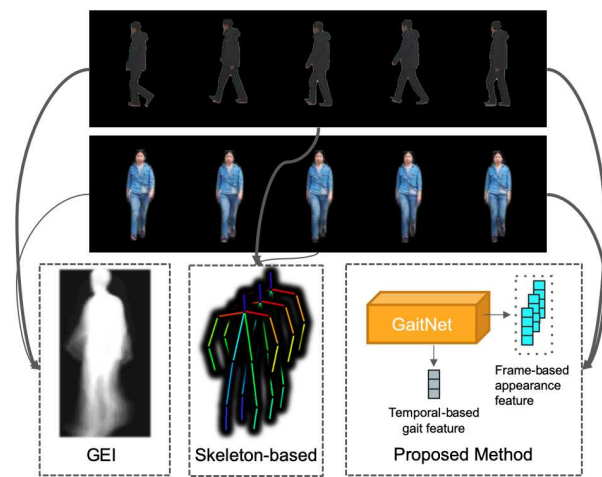
Figure 1: We propose to a novel CNN-based model, termed Gait-Net, to automatically learn the disentangled gait feature from a walking video, as opposed to handcrafted GEI, or skeleton-based features. While many conventional gait databases study side-view imagery, we collect a new gait database where both gallery and probe are captured in frontal-views.

Energy Image (GEI) [20] take the averaged silhouette image as the gait feature. While having a low computational cost and can handle low-resolution imagery, it can be sensitive to variations such as clothes change, carrying, view angles and walking speed [37, 5, 46, 6, 24, 1]. The model-based method first performs pose estimation and takes articulated body skeleton as the gait feature. It shows more robustness to those variations but at a price of a higher computational cost and dependency on pose estimation accuracy [17, 2].

It is understandable that the challenge in designing a gait feature is the necessity of being invariant to the appearance variation due to clothing, viewing angle, carrying, etc. Therefore, our desire is to *disentangle* the gait feature from the visual appearance of the walking person. For both appearance-based or model-based methods, such disentan-

Table 1: Comparison of existing gait databases and our collected FVG database.

Dataset	#Subjects	#Videos	Environment	Resolution	Format	Variations
CASIA-B	124	13,640	Indoor	320×240	RGB	View, Clothing, Carrying
USF	122	1,870	Outdoor	720×480	RGB	View, Ground Surface, Shoes, Carrying, Time
OU-ISIR-LP	4,007	–	Indoor	640×480	Silhouette	View
OU-ISIR-LP-Bag	62,528	–	Indoor	1,280×980	Silhouette	Carrying
FVG (our)	226	2,856	Outdoor	1,920×1,080	RGB	View, Walking Speed, Carrying, Clothing, Background, Time

glement is achieved by manually handcrafting the GEI or body skeleton, since neither has color information. However, we argue that these manual disentanglements may lose certain or create redundant gait information. E.g., GEI learns the average contours over time, but not the dynamic of how body parts move. For body skeleton, under carrying condition, certain body joints such as hands may have fixed positions, and hence are redundant information to gait.

To remedy the issues in handcrafted features, as shown in Fig. 1, this paper aims to automatically disentangle the pose/gait features from appearance features, and use the former for gait recognition. This disentanglement is realized by designing an autoencoder-based CNN, GaitNet, with novel loss functions. For each video frame, the encoder estimates two latent representations, pose feature (i.e., frame-based gait feature) and appearance feature, by employing two loss functions: 1) cross reconstruction loss enforces that the appearance feature of one frame, fused with the pose feature of another frame, can be decoded to the latter frame; 2) gait similarity loss forces a sequence of pose features extracted from a video sequence, of the same subject to be similar even under different conditions. Finally, the pose features of a sequence are fed into a multi-layer LSTM with our designed incremental identity loss to generate the sequence-based gait feature, where two of which can use the cosine distance as the video-to-video similarity metric.

Furthermore, most prior work [20, 46, 33, 12, 2, 7, 13] often choose the walking video of the side view, which has the richest gait information, as the gallery sequence. However, practically other view angles, such as the frontal view, can be very common when pedestrians toward or away from the surveillance camera. Also, the prior work [40, 10, 11, 34] that focuses on frontal view are often based on RGB-D videos, which have richer depth information than RGB videos. Therefore, to encourage gait recognition from the frontal-view RGB videos that generally has the minimal amount of gait information, we collect a high-definition (HD,1080p) frontal-view gait database with a wide range of variations. It has three frontal-view angles where the subject walks from left $45°$, $0°$, and right $45°$ off the optical axes of the camera. For each of three angles, different variants are explicitly captured including walking speed, clothing, carrying, clutter background, etc.

The contributions of this work are the following:

1) We propose an autoencoder-based network, GaitNet, with novel loss functions to explicitly disentangle the pose features from visual appearance and use multi-layer LSTM to obtain aggregated gait feature.

2) We introduce a frontal-view gait database, named FVG, including various variations of viewing angles, walking speeds, carrying, clothing changes, background and time gaps. This is the first HD gait database, with a nearly doubled number of subjects than prior RGB gait databases.

3) Our proposed method outperforms state of the art on three benchmarks, CASIA-B, USF, and FVG datasets.

2. Related Work

Gait Representation. Most prior works are based on two types of gait representations. In appearance-based methods, gait energy image (GEI) [20] or gait entropy image (GEnI) [5] are defined by extracting silhouette masks. Specifically, GEI uses an averaged silhouette image as the gait representation for a video. These methods are popular in the gait recognition community for their simplicity and effectiveness. However, they often suffer from sizeable intra-subject appearance changes due to covariates such as clothing, carrying, views, and walking speed. On the other hand, model-based methods [17] fit articulated body models to images and extract kinematic features such as 2D body joints. While they are robust to some covariates such as clothing and speed, they require a relatively higher image resolution for reliable pose estimation and higher computational costs.

In contrast, our approach learns gait information from raw RGB video frames which contain the richer information, thus with higher potential of extracting discriminative gait features. The most relevant work to ours is [12], which learns gait features from RGB images via Conditional Random Field. Compared to [12], our CNN-based approach has the advantage of being able to leverage a large amount of training data and learning more discriminative representation from data with multiple covariates. This is demonstrated by our extensive comparison with [12] in Sec. 5.2.1.

Gait Databases. There are many classic gait databases such as SOTON Large dataset [39], USF [37], CASIA-B [23], OU-ISIR [32], TUM GAID [23] and etc. We compare our FVG database with the most widely used ones in Tab. 1. CASIA-B is a large multi-view gait database with three variations: view angle, clothing, and carrying. Each subject is captured from 11 views under three conditions: normal walking (NM), walking in coats (CL) and walking while carrying bags (BG). For each view, 6, 2 and 2 videos are recorded from normal, coats and bags conditions. USF database has
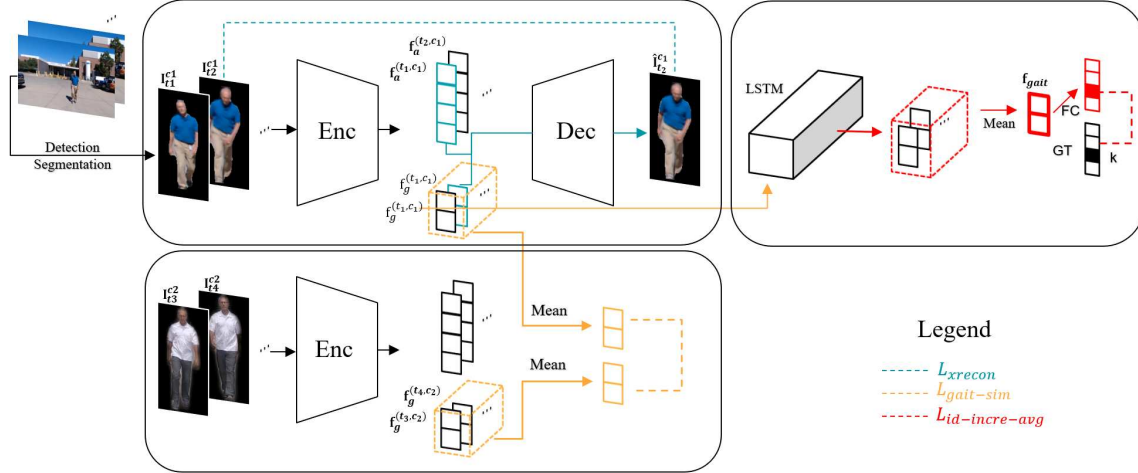
Figure 2: Overall architecture of our proposed approach, with three novel loss functions.

122 subjects with five variations, totaling 32 conditions for each subject. It contains two view angles (left and right), two ground surface (grass and concrete), shoes change, carrying condition and time. While OU-ISIR-LP and OU-ISIR-LP-Bag are large datasets, we can not leverage them as only the silhouette is publicly released.

Unlike those databases, our FVG database focuses on the frontal view, with 3 different near frontal-view angles towards the camera, and other variations including walking speed, carrying, clothing, cluttered background and time.

Disentanglement Learning. Besides model-based approaches [43, 42, 31] representing data with semantic latent vectors; data-driven disentangled representation learning approaches are gaining popularity in computer vision community. DrNet [14] disentangles content and pose vectors with a two-encoders architecture, which removes content information in the pose vector by generative adversarial training. The work of [3] segments foreground masks of body parts by 2D pose joints via U-Net [36] and then transforms body parts to desired motion with adversarial training. Similarly, [15] utilizes U-net and Variational Auto Encoder (VAE) to disentangle an image into appearance and shape. DR-GAN [44, 45] achieves state-of-the-art performances on pose-invariant face recognition by explicitly disentangling pose variation with a multi-task GAN [19].

Different from [14, 3, 15], our method has only one encoder to disentangle the appearance and gait information, through the design of novel loss functions without the need for adversarial training. Unlike DR-GAN [45], our method does not require adversarial training, which makes training more accessible. Further, pose labels are used in DR-GAN training so as to disentangle identity feature from the pose. However, to disentangle gait and appearance feature from the RGB information, there is no gait nor appearance *label* to be utilized for our method, since the type of walking pattern

or clothes cannot be defined as discrete classes.

3. Proposed Approach

Let us start with a simple example. Assuming there are three videos, where videos 1 and 2 capture subject A wearing t-shirt and long down coat respectively, and in video 3 subject B wears the same long down coat as in video 2. The objective is to design an algorithm, from which the gait features of video 1 and 2 are the same, while those of video 2 and 3 are different. Clearly, this is a challenging objective, as the long down coat can easily dominate the feature extraction, which would make videos 2 and 3 to be more similar than videos 1 and 2 in the latent space of gait features. Indeed the core challenge, as well as the objective, of gait recognition is to extract gait features that are discriminative among subjects, but invariant to different confounding factors, such as viewing angles, walking speeds and appearance.

Our approach to achieve this objective is via feature disentanglement - separating the gait feature from appearance information for a given walking video. As shown in Fig. 2, the input to our model is a video frame, with background removed using any off-the-shelf pedestrian detection and segmentation method [21, 9, 8]. An encoder-decoder network, with carefully designed loss functions, is used to disentangle the appearance and pose features for each video frame. Then, a multi-layer LSTM explores the temporal dynamics of pose features and aggregates them to a sequence-based gait feature for the identification purpose. In this section, we first present the feature disentanglement, followed by temporal aggregation, and finally implementation details.

3.1. Appearance and Pose Feature Disentanglement

For the majority of gait recognition datasets, there is a limited appearance variation within each subject. Hence, appearance could be a discriminate cue for identification during training as many subjects can be easily distinguished

by their clothes. Unfortunately, any networks or feature extractors relying on appearance will not generalize well on the test set or in practice, due to potentially diverse clothing or appearance between two videos of the same subject.

This limitation on training sets also prevents us from learning good feature extractors if solely relying on identification objective. Hence we propose to learn to disentangle the gait feature from the visual appearance in an unsupervised manner. Since a video is composed of frames, disentanglement should be conducted on the frame level first. Because there is no dynamic information within a video frame, we aim to disentangle the pose feature from the visual appearance for a frame. The dynamics of pose features over a sequence will contribute to the gait feature. In other words, we view the pose feature as the manifestation of video-based gait feature at a specific frame.

To this end, we propose to use an encoder-decoder network architecture with carefully designed loss functions to disentangle the pose feature from appearance feature. The encoder, $\mathcal{E}$, encodes a feature representation of each frame, $\mathbf{I}$, and explicitly splits it into two parts, namely appearance $\mathbf{f}_a$ and pose $\mathbf{f}_g$ features:

$$\mathbf{f}_a, \mathbf{f}_g = \mathcal{E}(\mathbf{I}). \tag{1}$$

These two features are expected to fully describe the original input image. As they can be decoded back to the original input through a decoder $\mathcal{D}$:

$$\widetilde{\mathbf{I}} = \mathcal{D}(\mathbf{f}_a, \mathbf{f}_g). \tag{2}$$

We now define the various loss functions defined for learning the encoder, $\mathcal{E}$, and decoder $\mathcal{D}$.

Cross Reconstruction Loss. The reconstructed $\widetilde{\mathbf{I}}$ should be close to the original input $\mathbf{I}$. However, enforcing self-reconstruction loss as in typical auto-encoder can't ensure the appearance $\mathbf{f}_a$ learning appearance information across the video and $\mathbf{f}_g$ representing pose information in each frame. Hence we propose the cross reconstruction loss, using an appearance feature $\mathbf{f}_a^{t_1}$ of one frame and pose feature $\mathbf{f}_g^{t_2}$ of another one to reconstruct the latter frame:

$$\mathcal{L}_{\text{xrecon}} = \left\| \mathcal{D}(\mathbf{f}_a^{t_1}, \mathbf{f}_g^{t_2}) - \mathbf{I}_{t_2} \right\|_2^2, \tag{3}$$

where $\mathbf{I}_t$ is the video frame at the time step t.

The cross reconstruction loss, on one hand, can play a role as the self-reconstruction loss to make sure the two features are sufficiently representative to reconstruct video frames. On the other hand, as we can pair a pose feature of a current frame to the appearance feature of *any* frame in the same video to reconstruct the same target, it enforces the appearance features to be similar across all frames.

Gait Similarity Loss. The cross reconstruction loss prevents the appearance feature $\mathbf{f}_a$ to be over-represented, containing pose variation that changes between frames. However, appearance information may still be leaked into pose

feature $\mathbf{f}_g$. In an extreme case, $\mathbf{f}_a$ is a constant vector while $\mathbf{f}_g$ encodes all the information of a video frame. To make $\mathbf{f}_g$ "cleaner", we leverage multiple videos of the same subject. Extra videos can introduce the change in appearance. Given two videos of the same subject with length n_1, n_2 in two different conditions c_1, c_2. Ideally, c_1, c_2 should contain difference in the person's appearance, i.e., cloth changes. While appearance changes, the gait information should be consistent between two videos. Since it's almost impossible to enforce similarity on $\mathbf{f}_g$ between video frames as it requires precise frame-level alignment; we enforce the similarity between two videos' averaged pose features:

$$\mathcal{L}_{\text{gait-sim}} = \left\| \frac{1}{n_1} \sum_{t=1}^{n_1} \mathbf{f}_g^{(t,c_1)} - \frac{1}{n_2} \sum_{t=1}^{n_2} \mathbf{f}_g^{(t,c_2)} \right\|_2^2. \tag{4}$$

3.2. Gait Feature Learning via Aggregation

Even when we can disentangle appearance and pose information for each video frame, the current feature $\mathbf{f}_g$ only contains the walking pose of the person in a specific instance, which can share similarity with another specific instance of a very different person. Here, we are looking for discriminative characteristics in a person walking pattern. Therefore, modeling its temporal change is critical. This is where temporal modeling architectures like the recurrent neural network or long short-term memory (LSTM) work best.

Specifically, in this work, we utilize a multi-layer LSTM structure to explore spatial (e.g., the shape of a person) and mainly, temporal (e.g., how the trajectory of subjects' body parts changes over time) information on pose features. As shown in Fig. 2, pose features extracted from one video sequence are feed into a 3-layer LSTM. The output of the LSTM is connected to a classifier C, in this case, a linear classifier is used, to classify the subject's identity.

Let $\mathbf{h}^t$ be the output of the LSTM at time step t, which is accumulative after feeding t pose features $\mathbf{f}_g$ into it:

$$\mathbf{h}^t = \text{LSTM}(\mathbf{f}_g^1, \mathbf{f}_g^2, ..., \mathbf{f}_g^t). \tag{5}$$

Now we define the loss function for LSTM. A trivial option for identification is to add the classification loss on top of the LSTM output of the final time step:

$$\mathcal{L}_{\text{id-single}} = -\log(C_k(\mathbf{h}^n)), \tag{6}$$

which is the negative log likelihood that the classifier C correctly identifies the final output $\mathbf{h}^n$ as its identity label k.

Identification with Averaged Feature. By the nature of LSTM, the output $\mathbf{h}^t$ is greatly affected by its last input $\mathbf{f}_g^t$. Hence the LSTM output, $\mathbf{h}^t$, can be varied across time steps. With a desire to obtain a gait feature that can be robust to the stopping instance of a walking cycle, we propose to use the

averaged LSTM output as our gait feature for identification:

$$\mathbf{f}_{\text{gait}}^t = \frac{1}{t} \sum_{s=1}^{t} \mathbf{h}^s. \tag{7}$$

The identification loss can be rewritten as:

$$\mathcal{L}_{\text{id-avg}} = -\log(C_k(\mathbf{f}_{\text{gait}}^n))$$
$$= -\log\left(C_k\left(\frac{1}{n}\sum_{s=1}^{n}\mathbf{h}^s\right)\right). \tag{8}$$

Incremental Identity Loss. LSTM is expected to learn that the longer the video sequence, the more walking information it processes then the more confident it identifies the subject. Instead of minimizing the loss on the final time step, we propose to use all the intermediate outputs of every time step weighted by w_t:

$$\mathcal{L}_{\text{id-inc-avg}} = \frac{1}{n}\sum_{t=1}^{n} -w_t \log\left(C_k\left(\frac{1}{t}\sum_{s=1}^{t}\mathbf{h}^s\right)\right). \tag{9}$$

To this end, the overall training loss function is:

$$\mathcal{L} = \mathcal{L}_{\text{id-inc-avg}} + \lambda_r \mathcal{L}_{\text{xrecon}} + \lambda_s \mathcal{L}_{\text{gait-sim}}. \tag{10}$$

The entire system, encoder-decoder, and LSTM are jointly trained. Updating $\mathcal{E}$ to optimize $\mathcal{L}_{\text{id-inc-avg}}$ also helps to further generate pose feature that has identity information and on which LSTM is able to explore temporal dynamics. At the test time, the output $\mathbf{f}_{\text{gait}}^t$ of LSTM is the gait feature of the video and used as the identity feature representation for matching. The cosine similarity score is used as the metric.

3.3. Implementation Details

Segmentation and Detection. Our network receives video frames with the person of interest segmented. The foreground mask is obtained from the state-of-the-art detection, Mask R-CNN [21]. Instead of using a zero-one mask by hard thresholding, we keep the soft mask returned by the network, where each pixel indicates the probability of being a person. This is partially due to the difficulty in choosing a threshold. Also, it prevents the loss in information due to the mask estimation error. We use a bounding box with a fixed ratio of width: height $= 1:2$ with the absolute height and center location given by the Mask R-CNN network. Input is obtained by pixel-wise multiplication between the mask and RGB values which is then resized to 32×64.

Network hyperparameter. Our encoder-decoder network is a typical CNN. Encoder consisting of 4 stride-2 convolution layers following by Batch Normalization and Leaky ReLU activation. The decoder structure is an inverse of the encoder, built from transposed convolution, Batch Normalization and Leaky ReLU layers. The final layer has a
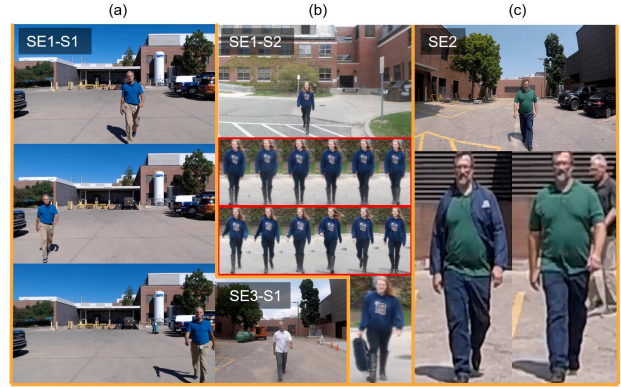
Figure 3: Examples of FVG Dataset. (a) Samples of the near frontal middle, left and right walking view angles in session 1 (SE1) of the first subject (S1). SE3-S1 is the same subject in session 3. (b) Samples of slow and fast walking speed for another subject in session 1. Frames in top red boxes are slow and in the bottom red box are fast walking. Carrying bag sample is shown below. (c) samples of changing clothes and with cluttered background from one subject in session 2.

Sigmoid activation to bring the value into $[0, 1]$ range as the input. The classification part is a stacked 3-layer LSTM [18], which has 256 hidden units in each of cells.

Adam optimizer [27] is used with the learning rate of 0.0001, and the momentum of 0.9. For each batch, we use video frames from 32 different clips. Since video lengths are varied, a random crop of 20-frame sequence is applied; all shorter videos are discarded. For Eqn. 9, we set $w_t = t^2$ while other options such as $w_t = 1$ also yield similar performance. The λ_r and λ_s (Eqn. 10) are set to 0.1 and 0.005 in all experiments.

4. Front-View Gait Database

Collection. To facilitate the research of gait recognition from frontal-view angles, we collect the Front-View Gait (FVG) database in a course of two years 2017 and 2018. During the capturing, we place the camera (Logitech C920 Pro Webcam or GoPro Hero 5) on a tripod at the height of 1.5 meter. We ask each of 226 subjects to walk toward the camera 12 times starting from around 16 meters, which results in 12 videos per subject. The videos are captured at $1,080 \times 1,920$ resolution with the average length of 10 seconds. The height of human in the video ranges from 101 to 909 pixels. These 12 walks have the combination of three angles toward the camera ($-45°$, $0°$, $45°$ off the optical axes of the camera), and four variations.

FVG is collected in three sessions. In session 1, in 2017, videos from 147 subjects are collected with four variations (normal walking, slow walking, fast walking, and carrying status). In session 2, in 2018, videos from additional 79 subjects are collected. Variations are normal, slow or fast walking speed, clothes or shoes change, and twilight or clus-
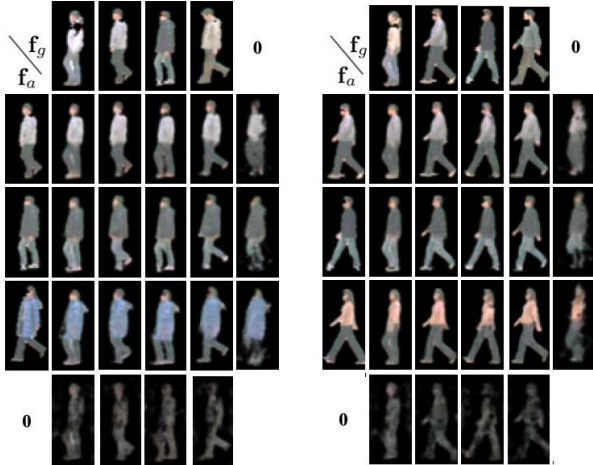
Figure 4: Synthesized frames on CASIA-B by decoding the various combination of $\mathbf{f}_a$ and $\mathbf{f}_g$. Left and right parts are two examples. For each example, $\mathbf{f}_a$ is extracted from images in the first column and $\mathbf{f}_g$ is extracted from images in the first row. $\mathbf{0}$ vector has the same dimension as $\mathbf{f}_g$ or $\mathbf{f}_a$, accordingly.

tered background. Finally in session 3, we collect repeated 12 subjects in year 2018 for extreme challenging test with the same setup as section 1. The purpose is to test how time gaps affect gait, along with changes in cloth/shoes or walking speed. Fig. 3 shows exemplar images from FVG.

Protocols. Different from prior gait databases, subjects in FVG are walking toward the camera, which creates a great challenge on exploiting gait information as the difference in consecutive frames can be much smaller than side-view walking. We focus our evaluation on variations that are challenging, e.g., different appearance, carrying a bag, or are not presented in other databases, e.g., cluttered background, along with view angles.

To benchmark research on FVG, we define 5 evaluation protocols, among which there are two commonalities: 1) the first 136 and rest 90 subjects are used for training and testing respectively; 2) the video 2, the normal frontal-view walking, is used as the gallery. The 5 protocols differ in their specific probe data, which cover the variations of Walking Speed (WS), Carrying Bag (CB), Changing Clothes (CL), Cluttered Background (CBG), and all variations (All). At the top part of Fig. 6, we list the detailed probe set for all 5 protocols. E.g., for the WS protocol, the probes are video $4-9$ in session 1 and video $4-6$ in session 2.

5. Experiments

Databases. We evaluate the proposed approach on three gait databases, CASIA-B [47], USF [37] and FVG. As mentioned in Sec. 2, CASIA-B, and USF are the most widely used gait databases, making the comparison with prior work easier. We compare our method with [46, 12, 29, 30] on these two databases, by following the respective experimental protocols of the baselines. These are either the most
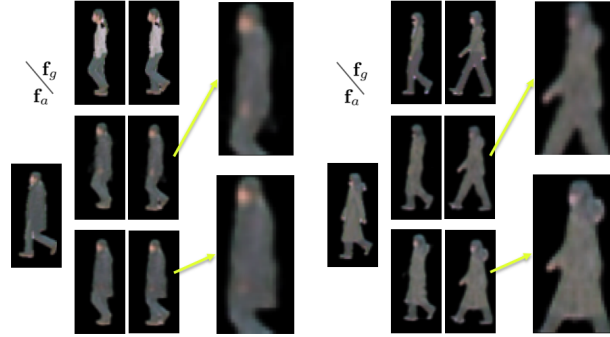
Figure 5: Synthesized frames on CASIA-B by decoding $\mathbf{f}_a$ and $\mathbf{f}_g$ from different variations (NM vs. CL). Left and right parts are two examples. For each example, $\mathbf{f}_a$ is extracted from the most left column image (CL) and $\mathbf{f}_g$ is extracted from the top row images (NM). Top row synthesized images are generated with model trained without $\mathcal{L}_{\text{gait-sym}}$ loss, bottom row is with the loss. To show the differences, details in generated images are magnified.

Table 2: Ablation study on our disentanglement loss and classification loss. By removing or replacing with other loss functions, Rank-1 recognition rate on cross NM and CL condition degrades.

Disentanglement Loss	Classification Loss	Rank 1
-	$\mathcal{L}_{\text{id-inc-avg}}$	56.0
$\mathcal{L}_{\text{xrecon}}$	$\mathcal{L}_{\text{id-inc-avg}}$	60.2
$\mathcal{L}_{\text{xrecon}} + \mathcal{L}_{\text{gait-sim}}$	$\mathcal{L}_{\text{id-inc-avg}}$	85.6
$\mathcal{L}_{\text{xrecon}} + \mathcal{L}_{\text{gait-sim}}$	$\mathcal{L}_{\text{id-avg}}$	62.6
$\mathcal{L}_{\text{xrecon}} + \mathcal{L}_{\text{gait-sim}}$	$\mathcal{L}_{\text{id-single}}$	26.0
$\mathcal{L}_{\text{xrecon}} + \mathcal{L}_{\text{gait-sim}}$	$\mathcal{L}_{\text{id-ae}}$ [41]	71.2

recent and state-of-art work or classic gait recognition methods. The OU-ISIR database [32] is not evaluated, and related methods [33] are not compared since our work consumes RGB video input, but OU-ISIR only releases silhouettes.

5.1. Ablation Study

Feature Visualization. To aid on understanding our features, we randomly pair $\mathbf{f}_a$, $\mathbf{f}_g$ features from different images and visualize the resultant paired feature by feeding it into our learned decoder $\mathcal{D}$. As shown in Fig. 4, each result is generated by paring the appearance $\mathbf{f}_a$ in the first column, and the pose $\mathbf{f}_g$ in the first row. The synthesized images show that indeed $\mathbf{f}_a$ contributes all the appearance information, e.g., cloth, color, texture, contour, as they are consistent across each row. Meanwhile, $\mathbf{f}_g$ contributes all the pose information, e.g., position of hand and feet, which share similarity across columns. We also visualize features $\mathbf{f}_a$, $\mathbf{f}_g$ *individually* by forcing the other feature to be a zero vector $\mathbf{0}$. Without $\mathbf{f}_g$, the reconstructed image still shares appearance similarity with $\mathbf{f}_a$ input but does not show a clear walking pose. Meanwhile, when removing $\mathbf{f}_a$, the reconstructed image still mimics the pose of $\mathbf{f}_g$'s input.

Disentanglement with Gait Similarity Loss. With the cross reconstruction loss, the appearance feature $\mathbf{f}_a$ can be

Table 3: Recognition accuracy cross views under NM on CASIA-B dataset. One single GaitNet module is trained for all the view angles.

Methods	0°	18°	36°	54°	72°	108°	126°	144°	162°	180°	Average
CPM [12]	13	14	17	27	62	65	22	20	15	10	24.1
GEI-SVR [29]	16	22	35	63	95	95	65	38	20	13	42.0
CMCC [28]	18	24	41	66	96	95	68	41	21	13	43.9
ViDP [26]	8	12	45	80	**100**	**100**	81	50	15	8	45.4
STIP+NN [30]	–	–	–	–	84.0	86.4	–	–	–	–	–
LB [46]	18	36	67.5	93	99.5	99.5	92	66	36	18	56.9
L-CRF [12]	38	**75**	68	93	98	99	**93**	67	76	39	67.8
GaitNet (our)	**68**	74	**88**	91	99	98	84	**75**	76	**65**	**81.8**

Table 4: Comparison on CASIA-B with cross view and conditions. Three models are trained for NM-NM, NM-BG, NM-CL. Average accuracies are calculated excluding probe view angles.

Gallery NM #1-4	0°-180°					36°-144°			
Probe NM #5-6	0°	54°	90°	126°	Mean	54°	90°	126°	Mean
CCA [4]	–	–	–	–	–	66.0	66.0	67.0	66.3
ViDP [26]	–	64.2	60.4	65.0	–	87.0	87.7	89.3	88.0
LB [46]	82.6	94.3	87.4	94.0	89.6	98.0	98.0	99.2	98.4
GaitNet (our)	**91.2**	**95.6**	**92.6**	**96.0**	**93.9**	**99.1**	**99.0**	**99.2**	**99.1**
Probe BG #1-2	0°	54°	90°	126°	Mean	54°	90°	126°	Mean
LB-subGEI [46]	64.2	76.9	63.1	76.9	70.3	89.2	84.3	91.0	88.2
GaitNet (our)	**83.0**	**86.6**	**74.8**	**85.8**	**82.6**	**90.0**	**85.6**	**92.7**	**89.4**
Probe CL #1-2	0°	54°	90°	126°	Mean	54°	90°	126°	Mean
LB-subGEI [46]	37.7	61.1	54.6	59.1	53.1	77.3	74.5	74.5	75.4
GaitNet (our)	**42.1**	**70.7**	**70.6**	**69.4**	**63.2**	**80.0**	**81.2**	**79.4**	**80.2**

enforced to represent static information that shares across the video. However, as discussed, the feature $\mathbf{f}_g$ can be spoiled or even encode the whole video frame. Here we show the need for the gait similarity loss $\mathcal{L}_{\text{gait-sym}}$ on the feature disentanglement. Fig. 5 shows the cross visualization of two different models learned with and without $\mathcal{L}_{\text{gait-sym}}$. Without $\mathcal{L}_{\text{gait-sym}}$ the decoded image shares some appearance characteristic, e.g., cloth style, contour, with $\mathbf{f}_g$. Meanwhile with $\mathcal{L}_{\text{gait-sym}}$, appearance better matches with $\mathbf{f}_a$.

Joints Location as Pose Feature. In literature, there is a large amount of effort in human pose estimation [17]. Aggregating joint locations over time could be a good candidate for gait features. Here we compare our framework with a baseline, named PE-LSTM, using pose estimation results as the input to the same LSTM as ours. Using state-of-the-art pose estimator [16], we extract 14 joints' locations and feed to the LSTM. This network achieves the recognition accuracy of 65.4% TDR at 1% FAR on the ALL protocol of FVG dataset, where our method outperforms it with 81.2%. This result demonstrates that our pose feature $\mathbf{f}_g$ does explore more discriminate feature than the joints' locations alone.

Loss Function's Impact on Performance. As the system consists of multiple loss functions, here we analyze the effect of each loss function on the final recognition performance. Tab. 2 reports the recognition accuracy of different variants of our framework on CASIA-B dataset under NM and CL. We first explore the effects of different disentanglement losses. Using $\mathcal{L}_{\text{id-inc-avg}}$ as the classification loss, we train different variants of our framework: a baseline without any disentanglement losses, a model with $\mathcal{L}_{\text{xrecon}}$, and

our full model with both $\mathcal{L}_{\text{xrecon}}$ and $\mathcal{L}_{\text{gait-sim}}$. The baseline achieves the accuracy of 56.0%. Adding the $\mathcal{L}_{\text{xrecon}}$ slightly improves the performance to 60.2%. By combining with $\mathcal{L}_{\text{gait-sim}}$, our model significantly improves the performance to 85.6%. Between $\mathcal{L}_{\text{xrecon}}$ and $\mathcal{L}_{\text{gait-sim}}$, the gait similarity loss plays a more critical role as $\mathcal{L}_{\text{xrecon}}$ is mainly designed to constrain the appearance feature $\mathbf{f}_a$, which does not directly involve identification.

Using the combination, $\mathcal{L}_{\text{xrecon}}$ and $\mathcal{L}_{\text{gait-sim}}$, we benchmark different options for classification loss as presented in Sec. 3.1, as well as the autoencoder loss by Srivastava et al. [41]. The model using the conventional identity loss on the final LSTM output $\mathcal{L}_{\text{id-single}}$ achieves the rank-1 accuracy of 26.0%. Using the average output of LSTM as identity feature, $\mathcal{L}_{\text{id-average}}$, shows to improve the performance to 62.6%. The autoencoder loss [41] achieves a good performance, 71.2%. However, it is still far from our proposed incremental identity loss $\mathcal{L}_{\text{id-inc-avg}}$'s performance.

5.2. Evaluation on Benchmark Datasets
5.2.1 CASIA-B
Since various experimental protocols have been defined on CASIA-B, for a fair comparison, we strictly follow the respective protocols in the baseline methods. Following [46], Protocol 1 uses the first 74 subjects for training and rest 50 for testing, regarding variations of NM (normal), BG (carrying bag) and CL (wearing a coat) with crossing view angles of 0°, 54°, 90°, and 126°. Three models are trained for comparison in Tab. 4. For the detailed protocol, please refer to [46]. Here we mainly compare our performance to Wu et al. [46], along with other methods [26]. Under multiple view angles and cross three variations, our method (GaitNet) achieves the best performance on all comparisons.

Recently, Chen et al. [12] propose new protocols to unify the training and testing where only one single model is being trained for each protocol. Protocol 2 focuses on walking direction variations, where all videos used are in NM. The training set includes videos of first 24 subjects in all view angles. The rest 100 subjects are for testing. The gallery is made of four videos at 90° view for each subject. Videos from remaining view angles are the probe. The rank 1 recognition accuracy are reported in Tab. 3. Our GaitNet achieves the best average accuracy of 81.8% across ten view angles, with significant improvement on extreme views. E.g., at

Table 5: Comparison with [12] and [46] under different walking conditions on CASIA-B by accuracies. One single GaitNet model is trained with all gallery and probe views and the two conditions.

Probe	Gallery	GaitNet (our)		L-CRF [12]		LB [46]		RLTDA [25]	
		BG	CL	BG	CL	BG	CL	BG	CL
54	36	91.6	**87.0**	**93.8**	59.8	92.7	49.7	80.8	69.4
54	72	90.0	**90.0**	**91.2**	72.5	90.4	62.0	71.5	57.8
90	72	**95.6**	**94.2**	94.4	88.5	93.3	78.3	75.3	63.2
90	108	87.4	**86.5**	**89.2**	85.7	88.9	75.6	76.5	72.1
126	108	90.1	**89.8**	**92.5**	68.8	93.3	58.1	66.5	64.6
126	144	**93.8**	**91.2**	88.1	62.5	86.0	51.4	72.3	64.2
Mean		91.4	89.8	91.5	73.0	90.8	62.5	73.8	65.2

Table 6: Definition of FVG protocols and performance comparison. Under each of the 5 protocols, the first/second columns indicate the indexes of videos used in gallery/probe.

	Index of Gallery & Probe videos									
Session 1	2	4-9	2	10-12	–	–	–	–	2	1,3-12
Session 2	2	4-6	–	–	2	7-9	2	10-12	2	1,3-12
Session 3	–	–	–	–	–	–	–	–	–	1-12
Variation	WS		CB		CL		CBG		All	
TDR@FAR	1%	5%	1%	5%	1%	5%	1%	5%	1%	5%
PE-LSTM	79.3	87.3	59.1	78.6	55.4	67.5	61.6	72.2	65.4	74.1
GEI [20]	9.4	19.5	6.1	12.5	5.7	13.2	6.3	16.7	5.8	16.1
GEINet [38]	15.5	35.2	11.8	24.7	6.5	16.7	17.3	35.2	13.0	29.2
DCNN [1]	11.0	23.6	5.7	12.7	7.0	15.9	8.1	20.9	7.9	19.0
LB [46]	53.4	73.1	23.1	50.3	23.2	38.5	56.1	74.3	40.7	61.6
GaitNet (our)	91.8	96.6	74.2	85.1	56.8	72.0	92.3	97.0	81.2	87.8

view angles of $0°$, and $180°$, the improvement margins are 30% and 26% respectively. This shows that GaitNet learns a better view-invariant gait feature than other methods.

Protocol 3 focuses on appearance variations. Training sets have videos under BG and CL. There are 34 subjects in total with $54°$ to $144°$ view angles. Different test sets are made with the different combination of view angles of the gallery and probe as well as the appearance condition (BG or CL). The results are presented in Tab. 5. We have comparable performance with the state-of-art method L-CRF [12] on BG subset while significantly improving the performance on CL subset. Note that due to the challenge of CL protocol, there is a significant performance gap between BG and CL for all methods except ours, which is yet another evidence that our gait feature has strong invariance to all major gait variations.

Across all evaluation protocols, GaitNet consistently outperforms state of the art. This shows the superior of GaitNet on learning a robust representation under different variations. It is contributed to our ability to disentangle pose/gait information from other static variations.

5.2.2 USF

The original protocol of USF [37] does not define a training set, which is not applicable to our method, as well as [46], that require data to train the models. Hence following the experiment setting in [46], we randomly partition the dataset into the non-overlapping training and test sets, each with half of the subjects. We test on Probe A, defined in [46], where the probe is different from the gallery by the viewpoint. We achieve the identification accuracy of $99.5 \pm 0.2\%$, which

Table 7: Runtime (ms per frame) comparison on FVG dataset.

Methods	Pre-processing	Inference	Total
PE-LSTM	22.4	0.1	22.5
GEINet [38]	0.5	1.5	2.0
DCNN [1]	0.5	1.7	2.2
LB [46]	0.5	1.3	1.8
GaitNet (our)	0.5	1.0	1.5

is better than the reported $96.7 \pm 0.5\%$ of LB network [46], and $94.7 \pm 2.2\%$ of multi-task GAN [22].

5.2.3 FVG

Given that FVG is a newly collected database and no reported performance from prior work, we make the efforts to implement 4 classic or state-of-the-art methods on gait recognition [20, 38, 1, 46]. For each of 4 methods and our GaitNet, one model is trained with the 136-subject training set and tested on all 5 protocols.

As shown in Tab. 6, our method shows state-of-the-art performance compared with other methods, including the recent CNN-based methods. Among 5 protocols, CL is the most challenging variation as in CASIA-B. Comparing with all different methods GEI based methods suffer from frontal view due to the lack of walking information.

5.3. Runtime Speed

System efficiency is an essential metric for many vision systems including gait recognition. We calculate the efficiency while each of the 5 methods processing one video of USF dataset on the same desktop with GeForce GTX 1080 Ti GPU. As shown in Tab. 7, our method is significantly faster than the pose estimation method because of 1) efficiency of Mask R-CNN; 2) an accurate, yet slow, version of AlphaPose [16] is required for gait recognition.

6. Conclusions

This paper presents an autoencoder-based method termed GaitNet that can disentangle appearance and gait feature representation from raw RGB frames, and utilize a multi-layer LSTM structure to further explore temporal information to generate a gait representation for each video sequence. We compare our method extensively with the state of the arts on CASIA-B, USF, and our collected FVG datasets. The superior results show the generalization and promising of the proposed feature disentanglement approach. We hope that in the future, this disentanglement approach is a viable option for other vision problems where motion dynamics needs to be extracted while being invariant to confounding factors, e.g., expression recognition with invariance to facial appearance, activity recognition with invariance to clothing.

References

[1] Munif Alotaibi and Ausif Mahmood. Improved Gait recognition based on specialized deep convolutional neural networks. *Computer Vision and Image Understanding (CVIU)*, 164:103–110, 2017. 1, 8

[2] Gunawan Ariyanto and Mark S Nixon. Marionette mass-spring model for 3D gait biometrics. In *International Conference on Biometrics (ICB)*, 2012. 1, 2

[3] Guha Balakrishnan, Amy Zhao, Adrian V Dalca, Fredo Durand, and John Guttag. Synthesizing Images of Humans in Unseen Poses. In *Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[4] Khalid Bashir, Tao Xiang, and Shaogang Gong. Cross-View Gait Recognition Using Correlation Strength. In *British Machine Vision Conference (BMVC)*, 2010. 7

[5] Khalid Bashir, Tao Xiang, and Shaogang Gong. Gait Recognition Using Gait Entropy Image. In *International Conference on Imaging for Crime Detection and Prevention (ICDP)*, 2010. 1, 2

[6] Khalid Bashir, Tao Xiang, and Shaogang Gong. Gait recognition without subject cooperation. *Pattern Recognition Letters*, 31(13):2052–2060, 2010. 1

[7] Aaron F Bobick and Amos Y Johnson. Gait Recognition Using Static, Activity-Specific Parameters. In *Computer Vision and Pattern Recognition (CVPR)*, 2001. 2

[8] Garrick Brazil and Xiaoming Liu. Pedestrian Detection with Autoregressive Network Phases. In *Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[9] Garrick Brazil, Xi Yin, and Xiaoming Liu. Illuminating Pedestrians via Simultaneous Detection and Segmentation. In *International Conference on Computer Vision (ICCV)*, 2017. 3

[10] Pratik Chattopadhyay, Aditi Roy, Shamik Sural, and Jayanta Mukhopadhyay. Pose Depth Volume extraction from RGB-D streams for frontal gait recognition. *Journal of Visual Communication and Image Representation*, 25(1):53–63, 2014. 2

[11] Pratik Chattopadhyay, Shamik Sural, and Jayanta Mukherjee. Frontal Gait Recognition From Incomplete Sequences Using RGB-D Camera. *IEEE Transactions on Information Forensics and Security*, 9(11):1843–1856, 2014. 2

[12] Xin Chen, Jian Weng, Wei Lu, and Jiaming Xu. Multi-Gait Recognition Based on Attribute Discovery. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 40(7):1697–1710, 2018. 2, 6, 7, 8

[13] David Cunado, Mark S Nixon, and John N Carter. Automatic extraction and description of human gait models for recognition purposes. *Computer Vision and Image Understanding*, 90(1):1–41, 2003. 2

[14] Emily L Denton et al. Unsupervised Learning of Disentangled Representations from Video. In *Neural Information Processing Systems (NeurIPS)*, 2017. 3

[15] Patrick Esser, Ekaterina Sutter, and Björn Ommer. A Variational U-Net for Conditional Appearance and Shape Generation. In *Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[16] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. RMPE: Regional Multi-Person Pose Estimation. In *International Conference on Computer Vision (ICCV)*, 2017. 7, 8

[17] Yang Feng, Yuncheng Li, and Jiebo Luo. Learning Effective Gait Features Using LSTM. In *International Conference on Pattern Recognition (ICPR)*, 2016. 1, 2, 7

[18] Felix A Gers, Jürgen Schmidhuber, and Fred Cummins. Learning to forget: continual prediction with LSTM. *Neural Computation*, 1999. 5

[19] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. In *Neural Information Processing Systems (NeurIPS)*, 2014. 3

[20] Ju Han and Bir Bhanu. Individual Recognition Using Gait Energy Image. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 28(2):316–322, 2006. 1, 2, 8

[21] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *International Conference on Computer Vision (ICCV)*, 2017. 3, 5

[22] Yiwei He, Junping Zhang, Hongming Shan, and Liang Wang. Multi-Task GANs for View-Specific Feature Learning in Gait Recognition. *IEEE Transactions on Information Forensics and Security*, 14(1):102–113, 2019. 8

[23] Martin Hofmann, Jürgen Geiger, Sebastian Bachmann, Björn Schuller, and Gerhard 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. 2

[24] Md Altab Hossain, Yasushi Makihara, Junqiu Wang, and Yasushi Yagi. Clothing-invariant gait identification using part-based clothing categorization and adaptive weight control. *Pattern Recognition*, 43(6):2281–2291, 2010. 1

[25] Haifeng Hu. Enhanced Gabor Feature Based Classification Using a Regularized Locally Tensor Discriminant Model for Multiview Gait Recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 23(7):1274–1286, 2013. 8

[26] Maodi Hu, Yunhong Wang, Zhaoxiang Zhang, James J Little, and Di Huang. View-Invariant Discriminative Projection for Multi-View Gait-Based Human Identification. *IEEE Transactions on Information Forensics and Security*, 8(12):2034–2045, 2013. 7

[27] Diederik P Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[28] Worapan Kusakunniran. Recognizing Gaits on Spatio-Temporal Feature Domain. *IEEE Transactions on Information Forensics and Security*, 9(9):1416–1423, 2014. 7

[29] Worapan Kusakunniran, Qiang Wu, Jian Zhang, and Hongdong Li. Support Vector Regression for Multi-View Gait Recognition based on Local Motion Feature Selection. In *Computer Vision and Pattern Recognition (CVPR)*, 2010. 6, 7

[30] Worapan Kusakunniran, Qiang Wu, Jian Zhang, Hongdong Li, and Liang Wang. Recognizing Gaits Across Views Through Correlated Motion Co-Clustering. *IEEE Transactions on Image Processing*, 23(2):696–709, 2014. 6, 7

[31] Feng Liu, Dan Zeng, Qijun Zhao, and Xiaoming Liu. Disentangling Features in 3D Face Shapes for Joint Face Reconstruction and Recognition. In *Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[32] Yasushi Makihara, Hidetoshi Mannami, Akira Tsuji, Md Altab Hossain, Kazushige Sugiura, Atsushi Mori, and Yasushi Yagi. The OU-ISIR Gait Database Comprising the Treadmill Dataset. *IPSJ Transactions on Computer Vision and Applications*, 4:53–62, 2012. 2, 6

[33] Yasushi Makihara, Atsuyuki Suzuki, Daigo Muramatsu, Xiang Li, and Yasushi Yagi. Joint Intensity and Spatial Metric Learning for Robust Gait Recognition. In *Computer Vision and Pattern Recognition (CVPR)*, 2017. 2, 6

[34] Athira M Nambiar, Paulo Lobato Correia, and Luís Ducla Soares. Frontal Gait Recognition Combining 2D and 3D Data. In *ACM Workshop on Multimedia and Security*, 2012. 2

[35] Mark S Nixon, Tieniu Tan, and Rama Chellappa. *Human Identification Based on Gait*. 2010. 1

[36] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015. 3

[37] Sudeep Sarkar, P Jonathon Phillips, Zongyi Liu, Isidro Robledo Vega, Patrick Grother, and Kevin W Bowyer. The Human ID Gait Challenge Problem: Data Sets, Performance, and Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 27(2):162–177, 2005. 1, 2, 6, 8

[38] Kohei Shiraga, Yasushi Makihara, Daigo Muramatsu, Tomio Echigo, and Yasushi Yagi. GEINet: View-Invariant Gait Recognition Using a Convolutional Neural Network. In *International Conference on Biometrics (ICB)*, 2016. 8

[39] Jamie D Shutler, Michael G Grant, Mark S Nixon, and John N Carter. On a Large Sequence-Based Human Gait Database. In *Applications and Science in Soft Computing*. 2004. 2

[40] Sabesan Sivapalan, Daniel Chen, Simon Denman, Sridha Sridharan, and Clinton Fookes. Gait Energy Volumes and Frontal Gait Recognition using Depth Images. In *International Joint Conference on Biometrics (IJCB)*, 2011. 2

[41] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov. Unsupervised Learning of Video Representations using LSTMs. In *International Conference on Machine Learning (ICML)*, 2015. 6, 7

[42] Luan Tran, Feng Liu, and Xiaoming Liu. Towards High-fidelity Nonlinear 3D Face Morphable Model. In *Computer Vision and Pattern Recognition (CVPR)*, June 2019. 3

[43] Luan Tran and Xiaoming Liu. Nonlinear 3D Face Morphable Model. In *Computer Vision and Pattern Recognition (CVPR)*, June 2018. 3

[44] Luan Tran, Xi Yin, and Xiaoming Liu. Disentangled Representation Learning GAN for Pose-Invariant Face Recognition. In *Computer Vision and Pattern Recognition (CVPR)*, 2017. 3

[45] Luan Tran, Xi Yin, and Xiaoming Liu. Representation Learning by Rotating Your Faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2018. 3

[46] Zifeng Wu, Yongzhen Huang, Liang Wang, Xiaogang Wang, and Tieniu Tan. A Comprehensive Study on Cross-View Gait Based Human Identification with Deep CNN. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 39(2):209–226, 2017. 1, 2, 6, 7, 8

[47] Shiqi Yu, Daoliang Tan, and Tieniu Tan. A Framework for Evaluating the Effect of View Angle, Clothing and Carrying Condition on Gait Recognition. In *International Conference on Pattern Recognition (ICPR)*, 2006. 6