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.

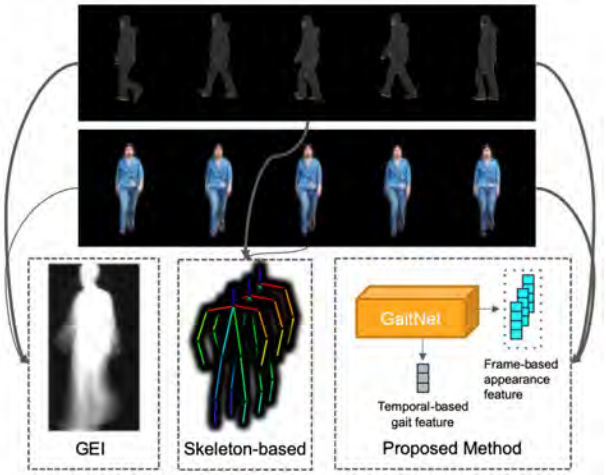

Figure 1: We propose a novel CNN-based model, termed GaitNet, to automatically learn the disentangled gait feature from a walking video, as opposed to handcrafted GEI, or skeleton-based features. While many conventional gait databases study side-view imagery, we collect a new gait database where both gallery and probe are captured in frontal-views.

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 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

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 (ours)	226	2,856	Outdoor	1,920×1,080	RGB	View, Walking Speed, Carrying, Clothing, Background, Time

appearance-based or model-based methods, such disentanglement 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 arts 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

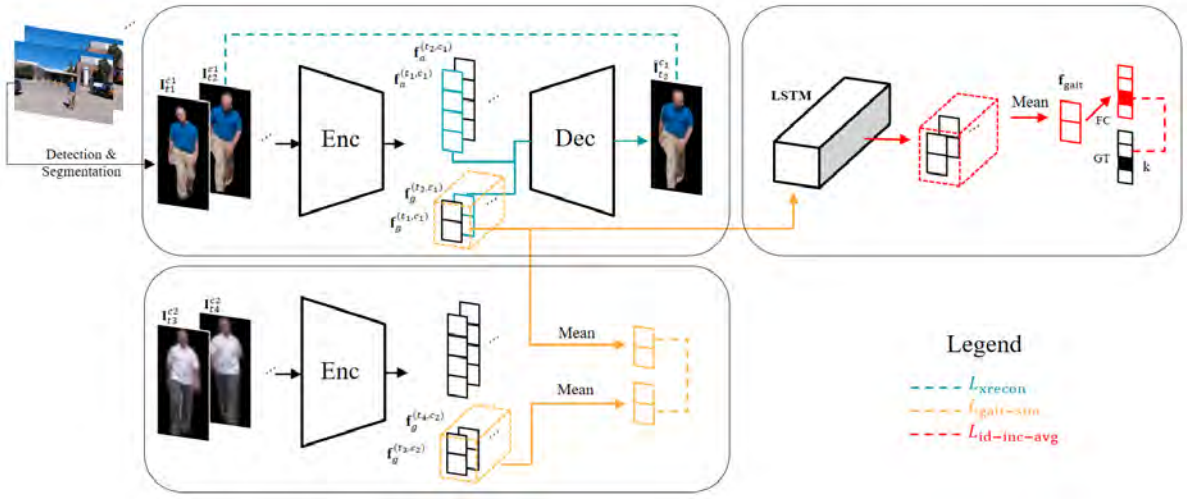

Figure 2: Overall architecture of our proposed approach, with three novel loss functions.

from normal, coats and bags conditions. USF database has 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:

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

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 instance segmentation, 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 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.

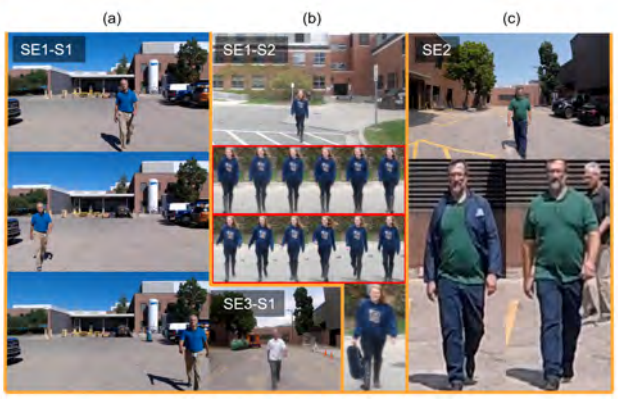

Figure 3: Examples of FVG Dataset. (a) Samples of the near frontal middle, left and right walking view angles in session 1 (SE1) of the first subject (S1). SE3-S1 is the same subject in session 3. (b) Samples of slow and fast walking speed for another subject in session 1. Frames in top red boxes are 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.

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-

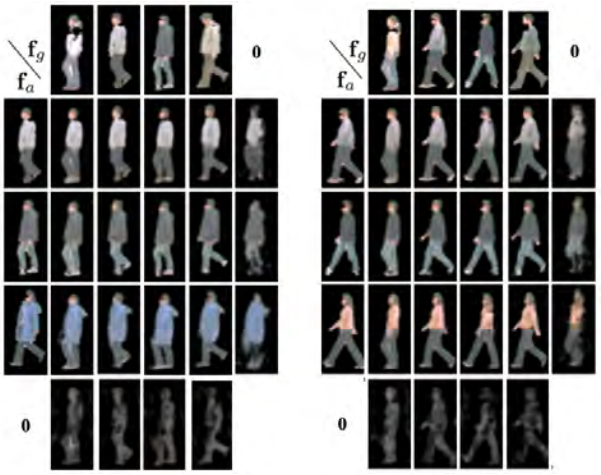

Figure 4: Synthesized frames on CASIA-B by decoding the various combination of $\mathbf{f}_a$ and $\mathbf{f}_g$. Left and right parts are two examples. For each example, $\mathbf{f}_a$ is extracted from images in the first column and $\mathbf{f}_g$ is extracted from images in the first row. $\mathbf{0}$ 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 recent and state-of-the-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.

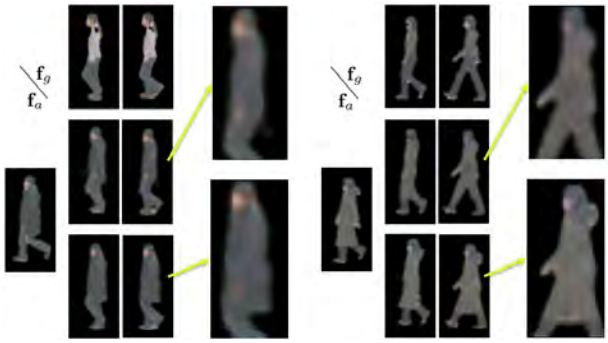

Figure 5: Synthesized frames on CASIA-B by decoding $\mathbf{f}_a$ and $\mathbf{f}_g$ from different variations (NM vs. CL). Left and right parts are two examples. For each example, $\mathbf{f}_a$ is extracted from the most left column image (CL) and $\mathbf{f}_g$ is extracted from the top row images (NM). Top row synthesized images are generated with model trained without $\mathcal{L}_{\text{gait-sim}}$ 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

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 (ours)	**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 (ours)	**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 (ours)	**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 (ours)	**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 $\mathcal{L}_{\text{gait-sim}}$ on the feature disentanglement. Fig. 5 shows the cross visualization of two different models learned with and without $\mathcal{L}_{\text{gait-sim}}$. Without $\mathcal{L}_{\text{gait-sim}}$ the decoded image shares some appearance characteristic, *e.g.*, cloth style, contour, with $\mathbf{f}_g$. Meanwhile with $\mathcal{L}_{\text{gait-sim}}$, 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 (ours)		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 (ours)	**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-the-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 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].

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 (ours)	0.5	1.0	1.5

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.

Acknowledgement

This work was supported with funds from the Ford-MSU Alliance program.

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

CVPR2019 Paper Translation

姓名：周学荣

学号：2017302373

班号：10011704

通过解缠表示学习进行步态识别

Ziyuan Zhang, Luan Tran, Xi Yin, Yousef Atoum, Xiaoming Liu
密歇根州立大学
{zhang835, tranluan, yinxi1, atoumyou, liuxm}@msu.edu

Jian Wan, Nanxin Wang
福特研究与发展中心
{jwan1, nwang1}@ford.com

摘要

步态，个体的行走模式，是最重要的生物识别方式之一。大多数现存的步态识别方法使用轮廓或者 () 身体模型作为步态特征。当受到衣服、携带物和视角等混杂变量的影响时，这些方法的识别性能会降低。为了解决这些问题，我们提出了一种新颖的 AutoEncoder 框架以明确地从 RGB 图像中区分姿势和外观特征，且随着时间的推移，基于 LTSM 的姿势特征整合产生步态特征。另外，我们收集了正面视角步态（FVG）数据集以专注于从正面观察步行识别步态，这是一个具有挑战性的问题，因为与其他视角相比，它含有最少的步态提示。FVG 还包括其他重要的变量，如步行速度、携带物和衣服。通过对 CASIA-B，USF 和 FVG 数据集的大量实验，我们的方法在数量上表现出优于现有技术的性能，定性地区分特征的能力，以及有前途的计算效率。

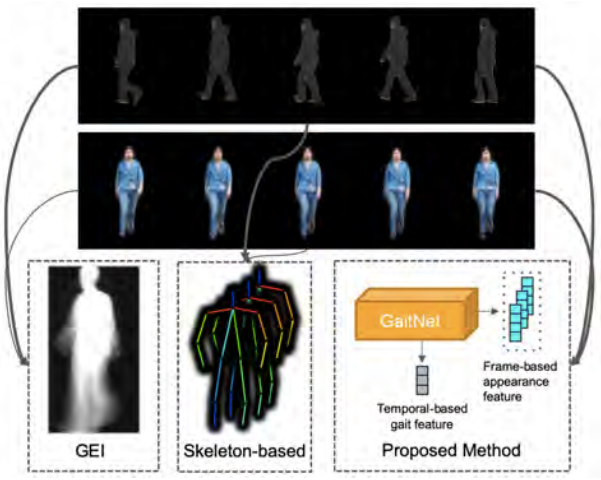

图 1: 我们提出了一种新的基于 CNN 的模型，称为 GaitNet，用于从步行视频中自动学习解缠的步态特征，而不是手工制作的 GEI 或基于骨架的特征。虽然许多传统数据库研究侧视图像，但我们收集了一个新的步态数据库，其中图库和探头都是在正面视图中捕获的。

1. 介绍

生物识别技术测量人们独特的身体和行为特征以识别个体的身份。步态 [35]，个体的行走模式，是生物识别的方式之一，类似的还有面部，指纹和虹膜。步态识别的优点是它可以在没有用户合作的情况下远距离操作，同时很难伪装。由于这些优点，步态识别适用于许多应用，例如人员识别，刑事调查和医疗保健。

和其他的视觉识别问题类似，步态识别的核心在于从步行者的视频帧中提取步态相关的特征，在先前的方法中被分为两种类型：基于外观的方法和基于模型的方法。基于外观的方法，如步态能量图像 (GEI) [20] 将平均轮廓图像作为步态特征。有较低的计算成本且可以处理低分辨率的图像，但它对例如衣服的更换、携带物、视角和步行速度 [37, 5, 46, 6, 24, 1] 之类的变化敏感。基于模型的方法首先执行姿势估计且将关节式人体骨架作为步态特征。它对于这些变化有更强的鲁棒性，但代价是更高的计算成本和对姿态估计精度的依赖 [17, 2]。

不受衣服、视角、携带物等外观变化而改变的必要性是设计步态特征的主要挑战是可以理解的。因此，我们希望将步态特征和行走者的视觉外观区分开来。对于基于外观或基于模型的方法，通过手动手工加工 GEI 或者身体骨架来实现这种区分，因为它们都没有颜色信息。然而，我们认为这些手动区分的方式会丢失某些或创建冗余的步态信息。例如，GEI 随着时间学习平均轮廓，而不是身体部位如何移动的动态。对于身体骨

表 1: 比较现有数据库和我们收集的 FVG 数据库。

数据集	个体数	视频数	环境	分辨率	格式	变量
CASIA-B	124	13,640	户内	320×240	RGB	视角, 穿着, 携带物
USF	122	1,870	户外	720×480	RGB	视角, 地面, 鞋子, 携带物, 时间
OU-ISIR-LP	4,007	–	户内	640×480	Silhouette	视角
OU-ISIR-LP-Bag	62,528	–	户内	1,280×980	Silhouette	携带物
FVG (我们的)	226	2,856	户外	1,920×1,080	RGB	视角, 行走速度, 携带物, 穿着, 背景, 时间

骼，在有携带物的状态下，诸如手的某些身体关节可能具有固定的位置，产生步态的冗余信息。

为了解决手工制作功能上的问题，如图 1所示，本文旨在自动将姿势/步态特征与外观特征进行区分，同时将前者用于步态识别。通过设计具有新颖损失函数的，基于自动编码器 CNN 的 GaitNet 来实现这种区分。对于每个视频帧，编码器通过使用两个损失函数来估计两个潜在的表示，姿势特征（即基于帧的步态特征）和外观特征：1）交叉重建损失强制执行一帧的外观特征，与另一帧的姿势特征融合，可以解码到后一帧；2）步态相似性损失迫使从视频序列中提取的一系列姿势特征即使在不同条件下也是相似的。最后，将一序列的姿势特征馈送到具有我们设计的增量同一性损失的多层 LSTM 中以生成基于序列的步态特征，其中两个可以使用余弦距离作为视频-视频相似性度量。

此外，大多数先前的研究 [20, 46, 33, 12, 2, 7, 13] 经常选择具有最丰富步态信息的侧视步行视频作为图库序列。然而，实际上当行人朝向或远离监控摄像机时，其他视角如正视图可能非常常见。此外，着重于正面视图的先前的研究 [40, 10, 11, 34] 通常基于 RGB-D 视频，其具有比 RGB 视频更丰富的深度信息。因此，为了促进通常具有最少步态信息量的正面 RGB 视频的步态识别，我们收集了具有各种变化的高清（HD,1080p）正面视图步态数据库。它有 3 个正面视角，主体从离摄像机的光轴左 45°，0°，右 45° 的视角。对于三个角度中的每一个，明确地捕获了不同的变体，包括步行速度，衣服，携带物，杂乱背景等。

这项工作的主要贡献如下：

1）我们提出了一个基于自动编码器的网络 GaitNet，它具有新颖的损失函数，可以明确地将姿势特征和视觉外观区分开来，并使用多层 LSTM 来获得聚合步态特征。

2）我们引入了一个名为 FVG 的正面视图步态数据库，包括视角，步行速度，携带物，衣服变化，背景和时间间隙等各种变化。这是第一个高清步态数据库，与以前的 RGB 步态数据库相比，它的早期主题数量增加了一倍。

3）我们提出的方法在 CASIA-B，USF 和 FVG 数据集这三个基准测试中优于现有技术水平。

2. 相关研究

步态表示. 大多数先前的工作基于两种类型的步态表示。在基于外观的方法中，通过提取轮廓掩膜来定义步态能量图像 (GEI）)[20] 或步态熵图像 (GEnI) [5]。具体而言，GEI 使用平均轮廓图像作为视频的步态表示。这些方法因其简单性和有效性而在步态识别社区中很受欢迎。然而，它们经常受到诸如衣服，携带物，观察视角和步行速度之类的协变量的影响，受试者会产生相当大的外观变化。另一方面，基于模型的方法 [17] 将关节模型拟合到图像中并提取运动特征，如 2D 身体关节。虽然它们对诸如衣服和速度的一些协变量具有鲁棒性，但是它们需要相对较高的图像分辨率以用于可靠的姿势估计，以及更高的计算成本。

相比之下，我们的方法从包含更丰富信息的原始 RGB 视频帧学习步态信息，因此具有提取辨别步态特征的更高潜力。与我们最相关的工作是 [12]，它通过条件随机场从 RGB 图像中学习步态特征。与 [12] 相比，我们的基于 CNN 的学习方法具有以下优点：能够利用大量训练数据并从具有多个协变量的数据中学习更多的判别性表示。在 5.2.1中我们与 [12] 进行了广泛的比较，并证明了这一点。

步态数据库. 现在有大量的步态数据库，如 SOTON 大数据集 [39]，USF [37]，OU-ISIR [23]，TUM GAID [32] 等。我们在表 1中将 FVG 数据库与最广泛的数据库进行了比较。CASIA-B 是一个大型多视图步态数据库，

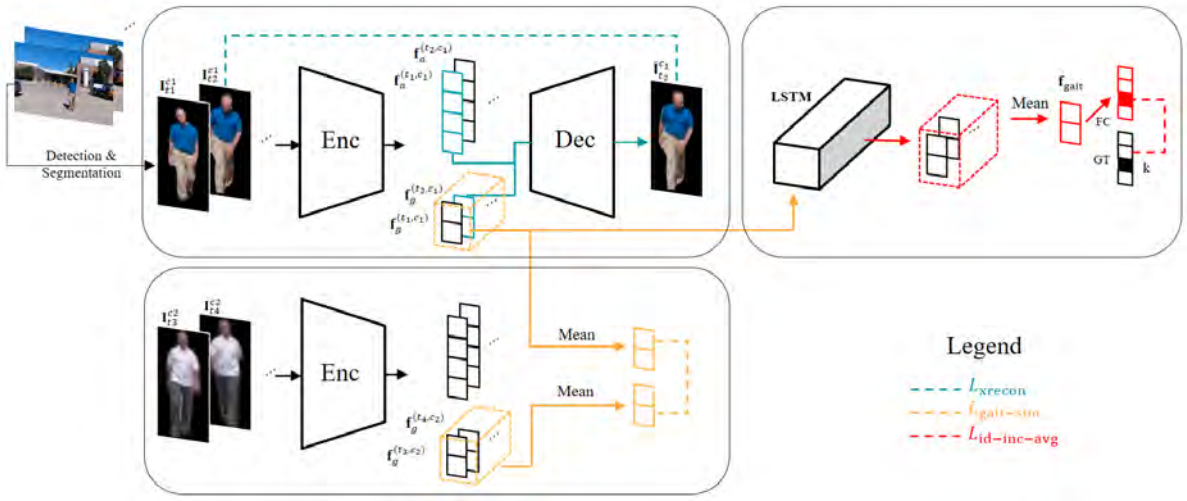

图 2: 我们提出的方法的总体结构，具有三个新的损失函数。

有三种变量：视角，衣服和携带物。在三个条件下从 11 个视角种捕获每一个受试者：正常步行（NM），穿着大衣（CL）和携带行李（BG）行走。对于每个视图，从正常、外套和行李的状态下记录 6、2、2 个视频。USF 数据库有 122 个条目，5 个变量，每个条目有 32 种状态。它包含两个视角（左和右），两种地面（草和混凝土），鞋子的变化，携带物状态和时间变化虽然 OU-ISIR-LP 和 OU-ISIR-Bag 是大型数据集，但我们无法利用它们，因为它们只公开发布了轮廓。

与那些数据库不同，我们的 FVG 数据库侧重于正面视图，具有朝向摄像机的 3 个不同的近前视角，以及其他变化，包括步行速度，携带物，衣服，杂乱的背景和时间。

解缠学习. 除了用语义潜在向量表示数据的基于模型的方法 [43, 42, 31]；数据驱动的解缠学习方法在计算机视觉社区种越来越受欢迎。DrNet [14] 使用双编码器架构区分内容和姿势向量，通过生成对抗训练去除姿势向量中的内容信息。 [3] 的工作借助 U-Net [36] 通过 2D 姿势关节对身体部位的前景蒙版进行分割，然后通过对抗训练将身体部位转换为所需的动作。同样，[15] 利用 U-Net 和变分自动编码器 (VAE) 将图像结构为外观和形状。DR-GAN [44, 45] 通过用多任务 GAN [19] 明确地解析姿势变化达到了姿势不变人脸识别的最先进表现。

与 [14, 3, 15] 不同，我们的方法只需一个编码器来区分外观和步态信息，通过设计新的损失函数而无需对抗训练。与 DR-GAN [45] 不同，我们的方法不需要对抗训练，这使得训练更容易获得。此外，姿势标签用于 DR-GAN 训练以便从姿势中解析身份特征。然而，为了从 RGB 信息中区分步态和外观特征，没有步态和外观标签可用于我们的方法，因为步态模式或衣服的类型不能定义为离散类。

3. 方法提出

让我们从一个简单的例子开始。假设有三个视频，其中视频 1 和 2 分别捕获主体 A 穿着 T 恤和长羽绒服，视频 3 中，主体 B 穿着和视频 2 中相同的长羽绒服。目标是通过视频设计一种算法，视频 1 和 2 的步态特征是相同的，而视频 2 和 3 的步态特征是不同的。很明显，这是一个具有挑战性的目标，因为长羽绒服很容易主导特征的提取，这将使视频 2 和 3 在步态特则会那个的潜在空间中比视频 1 和 2 更相似。实际上，步态识别的核心挑战和目标是提取受试者之间具有辨别力的步态特征，但对于不同的混淆因素例如视角，步行速度和外观是不变的。

我们实现此目标的方法是通过特征解析-分离将步态特征与给定的步行视频的外观信息分开。如图 2所示，我们输入的是一个视频帧，使用任意现成的行人检测和分割方法去除背景 [21, 9, 8]。编码器-解码器网络具有精心设计的损失函数，用于解开每个视频帧的外观和姿势特征。然后，用多层 LSTM 探索姿势特征的时间动态，并将它们聚合为基于序列的步态特征以用于识别目的。在本节中，我们首先介绍特征解析，然后是时间聚合，最后为实现的细节。

3.1. 3.1 外观和姿势特征解缠

对于大多数步态识别数据集，每个受试者的外观变化有限。因此，外观可以是训练期间识别的辨别线索，因为许多对象可以通过他们的衣服容易地区分。不幸的是，由于可能在同一主体的两个视频之间存在各种各样的服装或外观，因此依赖于外观的任何网络或特征提取器将不会在测试集上或实践中很好地归纳。

如果完全依赖于识别目标，这种对训练集的限制也会妨碍我们学习出好的特征提取器。因此，我们建议以无监督学习的方式将步态特征与视觉外观区分开来。由于视频由帧组成，所以应该首先从帧级上进行解析。由于视频帧内没有动态信息，我们的目标是将姿势特征与帧的视觉外观区分开来。姿势特征在序列上的动态变化将有助于步态特征。换句话说，我们将姿势特征视为特定帧处基于视频的步态特征的表现。

为此，我们提议使用具有精心设计的损失函数的编码器-解码器网络架构，以将姿势特征和外观特征分开。编码器 $\mathcal{E}$ 编码的每个帧的特征表示为 $\mathbf{I}$，并明确地分成两个部分，即外观 $\mathbf{f}_a$ 和姿势 $\mathbf{f}_g$ 特征：

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

预计这两个特征将完全描述原始输入图像。因为它们可以通过解码器 $\mathcal{D}$ 解码为原始输入：

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

我们现在为学习编码器 $\mathcal{E}$ 和解码器 $\mathcal{D}$ 定义各种损失函数。

交叉重建损失. 重建的 $\widetilde{\mathbf{I}}$ 应接近原始输入 $\mathbf{I}$。然而，在典型的自动编码器中强制执行自重构损失不能确保外观 $\mathbf{f}_a$ 和表示每帧中姿势信息的 $\mathbf{f}_g$ 在整个视频中学习信息。因此，我们提出交叉重建损失，使用一帧的外观特征 $\mathbf{f}_a^{t_1}$ 和另一帧的姿势特征 $\mathbf{f}_g^{t_2}$ 来重建后一帧：

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

其中 $\mathbf{I}_t$ 是时间点 t 的视频帧。

一方面，交叉重建损失可以起到自重构损失的作用，以确保两个特征足以负责重构视频帧。另一方面，由于我们可以将当前帧的姿势特征与同一视频中的任何帧的外观特征配对以重建相同的目标，因此它强制所有帧上的外观特征相似。

步态相似性损失. 交叉重建损失防止外观特征 $\mathbf{f}_a$ 被过度表示，包含在帧之间改变的姿势变化。但是，外观信息仍可能会泄露成姿势特征 $\mathbf{f}_g$。在极端情况下，$\mathbf{f}_a$ 是常数矢量，而 $\mathbf{f}_g$ 编码视频帧的所有信息。为了使 $\mathbf{f}_g$ “更清晰 “，我们利用同一主体的多个视频。额外的视频导致了外观的变化。给定两个具有相同主体、长度为 n_1 和 n_2，处于不同状态 c_1 和 c_2 的两个视频。理想地，c_1 和 c_2 应该包含人的外观差异，即布料的变化。外观变化时，两个视频之间的步态信息应该保持一致。由于几乎不可能在视频帧之间强制实现 $\mathbf{f}_g$ 的相似性，因为它需要精确的帧级对齐；我们强制执行两个视频的平均姿势特征之间的相似性：

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

3.2. 通过聚合进行步态特征学习

即使我们可以分开每个视频帧的外观和姿势信息，当前特征 $\mathbf{f}_g$ 仅包含特定实例中的人的行走姿势，其可以与非常不同的人的另一特定实例分享相似性。这里，我们正在寻找一个人行走模式的辨别特征。因此，对其时间变化进行建模至关重要。这就是回归神经网络或长短期记忆 (LSTM) 等时间建模体系结构最佳的地方。

具体来说，在该工作中，我们利用多层 LSTM 结构来探索空间 (例如人的形状)、并且主要是时间 (例如对象的身体部位的轨迹如何随时间变化) 上的关于姿势特征的信息。如图2所示，从一个视频序列中提取的姿势特征被馈送到 3 层 LSTM 中。LSTM 的输出连接到分类器 C，在这种情况下，使用线性分类器对主体的身份进行分类。

设 $\mathbf{h}^t$ 为时间步长 t 的 LSTM 的输出，在将姿势特征 $\mathbf{f}_g$ 输入其中之后累积 t：

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

现在我们定义 LSTM 的损失函数。识别的一个简单的选项是在最终时间步长的 LSTM 输出之上添加分类损失：

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

这是分类器 C 正确识别最终输出 $\mathbf{h}^n$ 作为其身份标签 k 的负对数似然性。

使用平均特征进行识别. 根据 LSTM 的性质，输出 $\mathbf{h}^t$ 受其最后输入的 $\mathbf{f}_g^t$ 的影响很大。因此，LSTM 输出的

$\mathbf{h}^t$ 可以跨时间步长变化。由于希望获得对步行周期的停止实例具有鲁棒性的步态特征，我们建议使用平均 LSTM 输出作为我们的步态特征进行识别：

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

识别损失可以改写为：

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

增量身份损失. 预计 LSTM 将了解到的视频序列越长，其处理的步行信息越多，则识别该主体的确信度越大。我们建议使用 w_t 加权的每个时间步长的所有中间输出，而不是最小化最后时间步长的损失：

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

到此，整体训练损失函数是：

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

整个系统，编码器-解码器和 LSTM 是联合训练的。更新 $\mathcal{E}$ 以优化 $\mathcal{L}_{\text{id-inc-avg}}$ 还有助于进一步生成具有身份信息的姿势特征，并且 LSTM 能在其上探索时间动态。在测试时，LSTM 的输出 $\mathbf{f}_{\text{gait}}^t$ 是视频的步态特征，并用作匹配的身份特征表现。余弦相似度得分用作度量。

3.3. 实施细节

分割和检测. 我们的网络接收感兴趣的人的分段视频帧。前景掩模是从最先进的实例分割 Mask R-CNN [21] 获得的。我们不是通过硬阈值使用零一掩码，而是保留网络返回的软掩码，其中每个像素表示成为一个人的概率。这部分因为难以选择阈值。而且，它防止了由于掩膜估计错误导致的信息丢失。我们使用一个固定比率为宽：高 $= 1:2$ 的边界框，其中绝对高度和中心位置由 Mask R-CNN 网络给出。通过掩模和 RGB 值之间的逐像素乘法获得输入，然后将其大小调整为 32×64。

网络超参数. 我们的编码器-解码器网络是典型的 CNN。编码器由 4 个步幅-2 循环层组成，后面是批量标准化和 Leaky ReLU 激活。解码器结构是编码器的反转，由转置卷积构建，批量标准化和 Leaky ReLU 层构建。最后一层由一个 Sigmoid 函数激活，将值作为输入带入 $[0, 1]$ 范围。分类部分是堆叠的 3 层 LSTM [18]，每个单元中有 256 个隐藏单元。

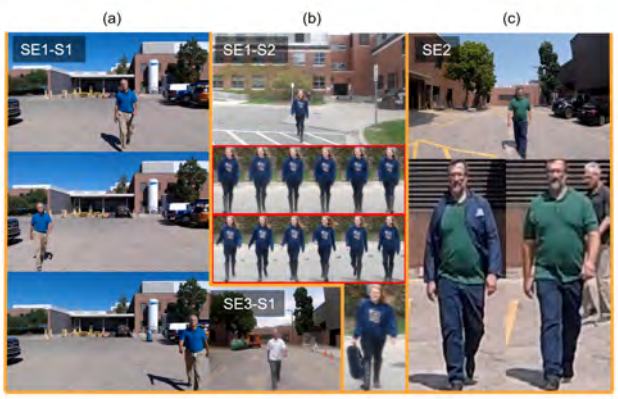

图 3: FVG 数据集的实例。(a) 第一对象 ($S1$) 的会话 1($SE1$) 中的近前，中，后行走视角的样本。SE3-S1 是在会话 3 中相同的受试者。(b) 第 1 阶段中另一名受试者的慢速和快速步行样本。顶部红色框内的帧较慢，底部红色框内的帧较快。携带物如下所示。(c) 会话 2 中来自一名受试者的衣服更换和杂乱背景的样本。

Adam 优化器 [27] 的学习率为 0.0001，动量为 0.9。对于每个批次，我们使用来自 32 个不同剪辑的视频帧。由于视频长度是变化的，因此应用 20 帧序列的随机剪裁；所有较短的视频都被丢弃。对于等式 9，我们设置 $w_t = t^2$，而其他选项如 $w_t = 1$ 也能产生类似的性能。在所有实验中，λ_r 和 λ_s（等式 10）设定为 0.1 和 0.005。

4. 前视步态数据库

采集. 为了便于从正面视角研究步态识别，我们在 2017 年和 2018 年的两年中收集了前视步态（FVG）数据库。在捕捉期间，我们放置了相机（Logitech C920 Pro 网络摄像头或 GoPro Hero5）在 1.5 米高的三脚架上。我们要求 226 个受试者中的每一位从大约 16 米开始 12 次走向相机，结果是每个受试者的 12 个视频。这些视频的分辨率为 $1,080 \times 1,920$，平均长度为 10 秒。视频中人的高度范围为 101 到 909 像素。这 12 个步行视频拥有相对于相机的三个角度的组合（偏离相机的光轴 $-45°, 0°, 45°$），以及 4 种变化。

FVG 分三次收集。在 2017 年的第一个季度，收集了来自 147 名受试者的视频，其中包括四种变化（正常

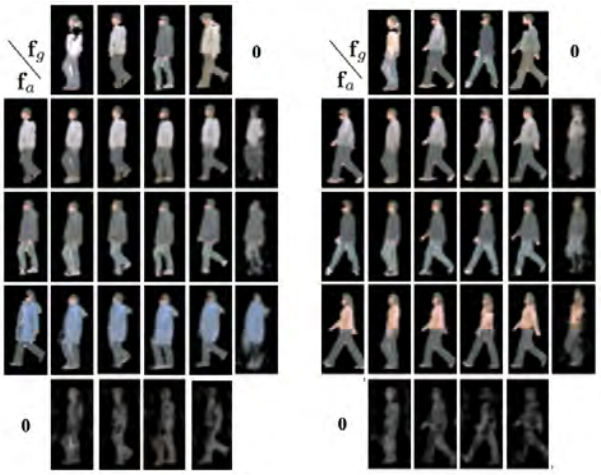

图 4: CASIA-B 上的合成帧通过解码 $\mathbf{f}_a$ 和 $\mathbf{f}_g$ 形成各种组合。左右两部分是两个例子。对于每个示例，$\mathbf{f}_a$ 和 $\mathbf{f}_g$ 分别从第一列和第一行中的图像提取。因此，$\mathbf{0}$ 向量具有和 $\mathbf{f}_g$ 或 $\mathbf{f}_a$ 相同的维度。

步行，慢走，快走和携带物状态）。在 2018 年的第二个季度，79 个其他受试者的视频被收集。变量是正常、缓慢或快速的步行速度，衣服或鞋子的变化，以及黄昏或群集的背景。最后在第三季度，我们在 2018 年收集重复的 12 个个体进行极端性测试，其设置与第一部分相同。目的是测试时间间隔如何影响步态，以及衣服/鞋子或步行速度的变化。图 3显示了来自 FVG 的实例性图像。

协议. 与先前的步态数据库不同，FVG 中的受试者正朝着相机行走，这对于利用步态信息产生了巨大的挑战，因为连续帧中的差异可能比侧视行走要小得多。我们将评估集中在具有挑战性的变化上，例如携带包的不同外观，或者不在其他数据库中呈现的变量如杂乱的背景，以及视角。

为了对 FVG 进行基准研究，我们定义了 5 个评估协议，其中有两个共性：1) 前 136 个和其余 90 个主题分别用于训练和测试；2) 视频 2，正常的正面视图行走，用作图库。这 5 种方案的特定探针数据不同，它们涵盖了步行速度（WS），携带包（CB），衣服更换（CL），杂乱背景（CBG）和所有变化（全部）的变化。在图 6的顶部，我们列出了 5 种协议的详细探针集。例如，对于 WS 协议，探测器在季度 1 中是视频 $4-9$，在季度 2 中是视频 $4-6$。

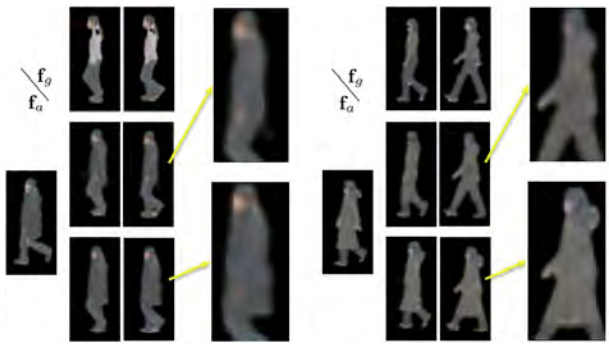

图 5: CASIA-B 上的合成帧通过解码来自不同变量 (NM 对 CL) 的 $\mathbf{f}_a$ 和 $\mathbf{f}_g$ 生成。左右两部分是两个例子。对每个例子，$\mathbf{f}_a$ 从大多数左列图像 (CL) 提取，$\mathbf{f}_g$ 从大多数顶行图像 (NM) 提取。生成的顶行合成图像没有 $\mathcal{L}_{\text{gait-sim}}$ 损失，底行是损失。为了显示差异，生成图像中的细节被放大。

表 2: 消融研究我们的解缠损失和分类损失。通过移除或替换其他损失函数，跨越 NM 和 CL 条件的 rank-1 识别率降低。

解缠损失	分类损失	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

5. 实验

数据库 我们评估了三种步态数据库的建议方法，CASIA-B [47]，USF [37] 和 FVG。如第 2节所述，CASIA-B 和 USF 是使用最广泛的步态数据库，使得较以前的工作比较容易。我们通过遵循基线的相应实验方案，在这两个数据库上比较我们的方法与 [46, 12, 29, 30] 。这些是最新的和最先进的工作或经典的步态识别方法。OU-ISIR 数据库 [32] 未被评估，相关方法 [33] 未进行比较，因为我们的工作使用 RGB 视频输入，但 OU-ISIR 仅发布轮廓。

5.1. 消融研究

特征可视化. 为了帮助理解我们的特征，我们随机对来自不同图像的 $\mathbf{f}_a$,$\mathbf{f}_g$ 特征进行配对，并通过将其馈送到我们的学习解码器 $\mathcal{D}$ 种来可视化所得到的配对特征。

表 3: 在 CASIA-B 数据集的 NM 状态下识别准确性的交叉视图。针对所有视角训练一个 GaitNet 模块。

方法	0°	18°	36°	54°	72°	108°	126°	144°	162°	180°	平均
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 (我们的)	**68**	74	**88**	91	99	98	84	**75**	76	**65**	**81.8**

表 4: CASIA-B 交叉视图和状态的比较。针对训练 NM-NM，NM-BG，NM-CL 三种模型。计算平均精度，不包括探针视角。

Gallery NM #1-4	0°-180°					36°-144°			
Probe NM #5-6	0°	54°	90°	126°	平均	54°	90°	126°	平均
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 (ours)	**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°	平均
LB-subGEI [46]	64.2	76.9	63.1	76.9	70.3	89.2	84.3	91.0	88.2
GaitNet (ours)	**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°	平均	54°	90°	126°	平均
LB-subGEI [46]	37.7	61.1	54.6	59.1	53.1	77.3	74.5	74.5	75.4
GaitNet (ours)	**42.1**	**70.7**	**70.6**	**69.4**	**63.2**	**80.0**	**81.2**	**79.4**	**80.2**

如图 4所示，通过对第一列种的外观 $\mathbf{f}_a$ 和第一行种的姿势 $\mathbf{f}_g$ 来进行配对生成每个结果。合成图像显示 $\mathbf{f}_a$ 确实体现了所有外观信息，例如布料、颜色、纹理、轮廓，因为它们在每行上是一致的。同时，$\mathbf{f}_g$ 贡献所有姿势信息，例如手和脚的位置，其在列之间共享相似性。我们还通过强制另一个特征为零的向量 $\mathbf{0}$ 来单独可视化特征 $\mathbf{f}_a$，$\mathbf{f}_g$。重建的图像仍然与输入共享外观相似性 $\mathbf{f}_a$，但是没有显示清晰的行走姿势 $\mathbf{f}_g$。同时，当去除 $\mathbf{f}_a$ 时，重建的图像仍模仿 $\mathbf{f}_g$ 输入的姿势。

具有步态相似性损失的解缠. 通过交叉重建损失，可以强制外观特征 $\mathbf{f}_a$ 表现视频中共享的静态信息。然而，如所讨论的，特征 $\mathbf{f}_g$ 可以被破坏甚至编码整个视频帧。在这里，我们显示在特征解开时需要的步态相似性损失 $\mathcal{L}_{\text{gait-sim}}$。图 5显示出了使用和不使用 $\mathcal{L}_{\text{gait-sim}}$ 学习的两个不同模型的交叉可视化。没有 $\mathcal{L}_{\text{gait-sim}}$ 的情况下，解码图像与 $\mathbf{f}_g$ 共享一些外观特征，例如布样式，轮廓。与有 $\mathcal{L}_{\text{gait-sim}}$ 的情况相比，外观更好匹配 $\mathbf{f}_a$。

作为姿势特征的关节位置. 文献中有许多对人体姿势的估计而作出的努力 [17]。随着时间的推移聚合的关节位置可能是表示步态特征的良好候选者。在这里，我们将我们的框架和名为 PE-LSTM 的基线进行比较，使用姿势估计结果作为与我们相同的 LSTM 的输入。使用最先进的姿势估计器 [16]，我们提取 14 个关节的位置并提供给 LSTM。该网络在 FVG 数据集的 ALL 协议上实现了在 1%FAR 下的 65.4%TDR 的识别准确度，而我们的方法以 81.2% 的结果占优。这个结果表明我们的姿势特征 $\mathbf{f}_g$ 确实比单独的关节位置探索到更多的辨别特征。

损失函数对性能的影响. 由于系统由多个损失函数组成，因此我们在此分析每个损失函数对最终识别性能的影响。表格 2报告了我们的框架的不同变体在 NM 和 CL 的状态下对 CASIA-B 数据集的识别准确性。我们首先探讨不同解缠损失的影响。使用 $\mathcal{L}_{\text{id-inc-avg}}$ 作为分类损失，我们的训练我们的框架的不同变体：没有任何解缠损失的基线，具有 $\mathcal{L}_{\text{xrecon}}$ 的模型，以及具有 $\mathcal{L}_{\text{xrecon}}$ 和 $\mathcal{L}_{\text{gait-sim}}$ 的完整模型。基线的准确率为 56.0%。添加 $\mathcal{L}_{\text{xrecon}}$ 可以将性能略微提高至 60.2%。通过与 $\mathcal{L}_{\text{gait-sim}}$ 相结合，我们的模型性能显著提高到了 85.6%。在 $\mathcal{L}_{\text{xrecon}}$ 和 $\mathcal{L}_{\text{gait-sim}}$ 之间，步态相似性损失起着更为关键的作用，因为 $\mathcal{L}_{\text{xrecon}}$ 主要用于约束外观特征 $\mathbf{f}_a$，其不直接涉及识别。

使用 $\mathcal{L}_{\text{xrecon}}$ 和 $\mathcal{L}_{\text{gait-sim}}$ 组合，我们对分类损失的不同选项进行了基准测试，如第 3.1节中所述，以及 Srivastava 等人的自动编码器的损失 [41]。在最终 LSTM 输出 $\mathcal{L}_{\text{id-single}}$ 上使用传统同一性损失的模型实现了 26.0% 的 rank-1 准确度。使用 LSTM 的作为身份特征的平均输出 $\mathcal{L}_{\text{id-average}}$，显示出性能提高到 62.6%。自动编码器损失 [41] 实现了良好的性能，71.2%。然而，它仍然不及我们提出的增量身份损失 $\mathcal{L}_{\text{id-inc-avg}}$ 的表现。

表 5: 在 CASIA-B 下与 [12] 和 [46] 在不同步行状态下比较准确性。一个单一的 GaitNet 模型以图库和探索视图两个条件训练。

Probe	Gallery	GaitNet (ours)		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
平均		91.4	**89.8**	**91.5**	73.0	90.8	62.5	73.8	65.2

5.2. 5.2 对基准数据集的评估

5.2.1 CASIA-B

由于 CASIA-B 上已经定义了各种实验方案，为了公平比较，我们严格遵循基线方法中的相应方案。在 [46] 后，方案 1 使用前 74 个受试者进行训练，余下 50 个进行测试关于 NM（正常），BG（携带包），CL（穿着大衣）的变化，其交叉视角为 0°，54°，90° 和 126°。在表格 4中是训练的三种模型的比较，有关详细的协议请参阅 [46]。这里我们主要比较我们和吴等人的模型的表现 [46]，以及其他方法 [26]。在多视角和三种变化下，我们的方法（GaitNet）在所有比较中都达到了最佳性能。

最近，陈等人 [12] 提出了新的协议来统一训练和测试，其中只有一个模型按所有协议训练。协议 2 侧重于步行方向的变化，其中所有视频都处于 NM 状态。训练集包括所有视角中的前 24 个受试者的视频。其余 100 名受试者进行测试。该图库由每个主体位于 90° 视角的四个视频组成。来自其余视角的视频是检测器。表格 3中报告了 rank-1 的识别准确度。我们的 GaitNet 在十个视角内达到了 81.8% 的最佳平均精度，并且在极端视图方面有着显著改善。例如，在 0° 和 180° 的视角下，改善幅度为 30% 和 26%。这表明 GaitNet 学会了比其他方法更好的视图不变步态功能。

协议 3 侧重于外观变化。训练集有在 BG 和 CL 状态下的视频。总共有 34 个受试者，拍摄视角从 54° 到 144°。使用图库中不同的角度、探头和不同的外观条件（BG 或 CL）制作成不同组合的测试组。结果显示在表格 5中。我们在 BG 子集上具有与最先进的方法 L-CRF [12] 相当的性能，同时显著提高了 CL 子集的性能。注意由于 CL 协议的挑战，除了我们以外的所有方法，BG 和 CL 之间都存在显著的性能差距，这是我们的步态特征对所有主要步态变化具有强不变性的另一个证据。

表 6: FVG 协议的定义和性能比较。在 5 个状态中每个状态下，第一/第二列表示在 gallery/probe 中使用的视频的索引。

Gallery &Probe 视频索引										
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
变量	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 (我们的)	**91.8**	**96.6**	**74.2**	**85.1**	**56.8**	**72.0**	**92.3**	**97.0**	**81.2**	**87.8**

在所有的评估协议中，GaitNet 实种优于现有技术水平。这显示了 GaitNet 在不同变化下学习能力稳健的优越性。它有助于我们将姿势/步态信息和其他静态变化区分开来。

5.2.2 USF

USF [37] 的原始协议没有定义训练集，这不适用于我们的方法， [46] 同理，它们需要数据来训练模型。因此，按照 [46] 中的实验设置，我们将数据集随机分成不重叠的训练和测试集，分别包含了一半的受试者。我们在 [46] 中定义的探针 A 上进行测试，其中探针与视点不同。我们的识别准确率达到 $99.5 \pm 0.2\%$，优于报告的 LB 网络的 $96.7 \pm 0.5\%$ [46] 和多任务 GAN 的 $94.7 \pm 2.2\%$ [22]。

5.2.3 FVG

鉴于 FVG 是一个新收集的数据库，并且没有报告先前工作的表现，我们努力实施 4 种经典或最先进的步态识别方法 [20, 38, 1, 46]。对于 4 种方法和我们的 GaitNet 中的每一种方法，使用 136 个受试者训练集训练一个模型并且在 5 个协议上进行测试。

如表格. 6所示，我们的方法表现出相比其他方法的最先进性能，包括最近的基于 CNN 的方法。在 5 种协议中，CL 是 CASIA-B 中最具有挑战性的变量。与所有不同的方法相比，基于 GEI 的方法由于受到正面视图的影响缺乏步行信息。

5.3. 运行速度

系统效率是包括步态识别在内的许多视觉系统的基本指标。我们使用 GeForce GTX 1080 Ti GPU 在同

表 7: FVG 数据集的运行时间 (每毫秒每帧)

方法	预处理	推算	总共
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 (我们的)	0.5	1.0	1.5

样的桌面上分别使用 5 种方法处理一个 USF 数据集的视频以计算效率。结果如表格 7所示，我们的方法明显快于姿势估计方法，原因是 1）Mask R-CNN 的效率;2) 步态识别需要准确但是缓慢的 AlphaPose [16] 版本。

6. 结论

本文提出了一种基于自动编码器的方法，称为 GaitNet，可以分开原始 RGB 帧的外观和步态特征表示，并利用多层 LSTM 结构进一步探索时间信息，为每个视频序列生成步态表示。我们将我们的方法和 CASIA-B，USF 和我们收集的 FVG 数据集的现状进行了广泛的比较。优越的结果表命了所提出的特征解缠方法的普遍性和有前途。我们希望在将来，这种解缠方法对于其他视觉问题也是可行的选择，其中需要提取运动动力学同时对混杂因素不变，例如具有面部外观不变性的表情识别，具有衣服不变性的活动识别。

鸣谢

这项研究得到了福特-密歇根州立大学联盟计划的资金支持。

参考文献

[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.

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

[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.

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

[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.

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

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

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

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

[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.

[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.

[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.

[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.

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

[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.

[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.

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

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

[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.

[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.

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

[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.

[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.

[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.

[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.

[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.

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

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

[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.

[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.

[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.

[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.

[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.

[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.

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

[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.

[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.

[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.

[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.

[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.

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

[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.

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

[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.

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

[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.

[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.