On Learning Disentangled Representations for Gait Recognition

Ziyuan Zhang, Luan Tran, Feng Liu, *Member, IEEE*, and Xiaoming Liu, *Senior Member, IEEE*

Abstract—Gait, the walking pattern of individuals, is one of the important biometrics modalities. Most of the existing gait recognition methods take silhouettes or articulated body models as gait features. These methods suffer from degraded recognition performance when handling confounding variables, such as clothing, carrying and viewing angle. To remedy this issue, we propose a novel AutoEncoder framework, GaitNet, to explicitly disentangle appearance, canonical and pose features from RGB imagery. The LSTM integrates pose features over time as a dynamic gait feature while canonical features are averaged as a static gait feature. Both of them are utilized as classification features. 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 SOTA quantitatively, the ability of feature disentanglement qualitatively, and promising computational efficiency. We further compare our GaitNet with state-of-the-art face recognition to demonstrate the advantages of gait biometrics identification under certain scenarios, e.g., long distance/lower resolutions, cross viewing angles. Source code is available at http://cvlab.cse.msu.edu/project-gaitnet.html.

Index Terms—Gait recognition, deep convolutional neural networks, disentangled representation learning, auto-encoder, LSTM, canonical representation, face recognition.

1 Introduction

Biometrics measures people's unique physical and behavioral characteristics to recognize the identity of an individual. Gait [1], the walking pattern of an individual, is one of biometrics modalities besides face, fingerprint, iris, *etc*. Gait recognition has the advantage that it can operate at a distance without users' cooperation. It has been applied to many applications such as person identification, criminal investigation, and healthcare [2].

As other recognition problems, gait data can usually be captured by five types of sensors [3], *i.e.*, RGB camera, RGB-D camera [4], [5], accelerometer [6], floor sensor [7], and continuous-wave radar [8]. Among them, RGB camera is not only the most popular one due to the low sensor cost, but also the most challenging one since RGB pixels might not be effective in capturing the motion cues. This work studies gait recognition from RGB cameras.

The core of gait recognition lies in extracting *gait features* from the video frames of a walking person, where the prior work can be categorized into two types: appearance-based and model-based methods. The appearance-based methods, *e.g.*, Gait Energy Image (GEI) [9], take the averaged silhouette image as the gait feature. While having a low computational cost and being able to handle low-resolution imagery, it can be sensitive to variations such as cloth change, carrying, viewing angles and walking speed [10]–[17]. The model-based methods use the articulated body skeleton from pose estimation as the gait feature. They show more robustness to aforementioned variations but at a price of a higher computational cost and dependency on pose estimation accuracy [18]–[20].

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 non-gait-related appearance of the walking person. For both appearance-based or model-based methods, such disentanglement is achieved by manually handcrafting the GEI-like [9], [11] or body skeleton-like [18]–[20] features, since neither has color or texture information. However, we argue that these manual disentanglements may be sensitive to changes in walking condition. In other words, they can lose certain or create redundant gait information. E.g., GEI-like features have distinct silhouettes for the same subject wearing different clothes. For skeleton-like features, when carrying accessories (*e.g.*, bags, umbrella), certain body joints such as hands may have fixed positions, and hence are redundant information to gait.

To remedy the aforementioned issues in handcrafted features, as shown in Fig. 1 (a), this paper proposes a novel approach to learn gait representations from the RGB video directly. Specifically, we aim to automatically disentangle dynamic pose features (trajectory of gait) from pose-irrelevant features. To further distill identity information from pose-irrelevant features, we disentangle the pose-irrelevant features into appearance (*i.e.,* clothing) and canonical features. Here, the canonical feature refers to a standard and unique representation of human body, such as body ratio, width and limb lengths, *etc*. The pose features and canonical features are discriminative in identity and are used for gait recognition. Fig. 1 (b) visualizes the three disentangled features.

This disentanglement is realized by designing an autoencoder-based Convolutional Neural Network (CNN), GaitNet, with novel loss functions. For each video frame, the encoder estimates three latent representations: pose, canonical and appearance features, by employing three loss functions: 1) cross reconstruction loss enforces that the canonical and appearance features of one frame, fused with the pose feature of another frame, can be decoded to the latter frame; 2) pose similarity loss forces a sequence of pose features extracted from a video sequence, of the same

• *Ziyuan Zhang, Luan Tran, Feng Liu, and Xiaoming Liu are with the Department of Computer Science and Engineering, Michigan State University. E-mail: {zhang835, tranluan, liufeng6}@msu.edu, liuxm@cse.msu.edu*

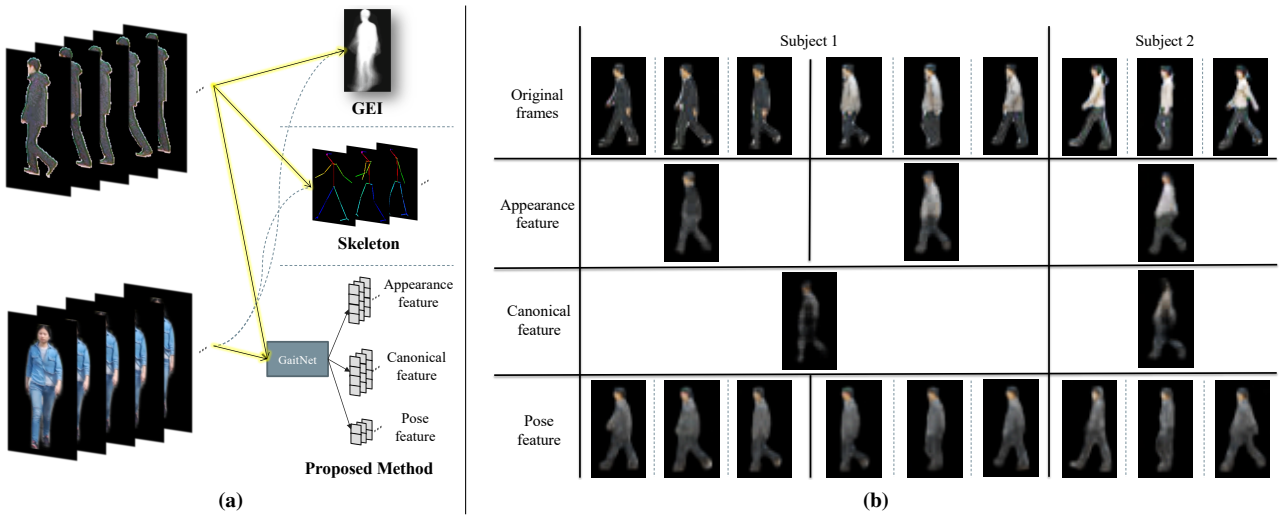

Fig. 1: (a) While conventional gait databases capture side-view imagery, we collect a new gait database (FVG) with focus on more challenging frontal views. We propose a novel CNN-based model, termed GaitNet, to directly learn the disentangled appearance, canonical and pose features from walking videos, as opposed to handcrafted GEI or skeleton features. (b) Given 2 videos of Subject 1 and 1 video of Subject 2, feature visualizations by our decoder in Fig. 3 show that, the appearance feature is video-specific capturing clothing information; the canonical feature is subject-specific capturing the overall body shape at a standard pose; the pose feature is frame-specific capturing body poses at individual frames.

subject to be similar even under different conditions; 3) canonical consistency loss favors consistent canonical features among videos of the same subject 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 dynamic gait feature. The average of canonical features results in the sequence-based static gait feature. Given two gait videos, the cosine distances between their respective dynamic and static gait features are computed and their summation is the final video-to-video gait similarity metric.

In addition, most prior work [9], [11], [15], [18], [21]–[26] choose the walking video of the side view, which has the richest gait information, as the gallery sequence. However, in practices other viewing angles, such as the frontal view, can be very common when pedestrians walk toward or away from the surveillance camera. Also, the prior work [27]–[30] that focuses on frontal view are often based on RGB-D videos, which have additional depth information than RGB. Therefore, to encourage gait recognition from frontal-view RGB videos that generally has the minimal amount of gait information, we collect a high-definition (HD, 1080p) Frontal-View Gait database, named FVG, 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, multiple people, *etc*.

A preliminary version of this work was published in the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2019 [31]. We extend the work from three aspects. 1) Instead of disentangling features in two components: pose and pose-irrelevant [31], we further decouple the pose-irrelevant features into discriminative canonical feature and appearance feature. By devising an effective canonical consistency loss, the canonical feature helps to improve gait recognition accuracy. 2) We conduct more insightful ablation studies to analyze the relationship between our disentanglement losses and features, gait recognition over time, and contributions of dynamic and static gait features. 3) We perform side-by-side comparison between gait recognition and the state-of-the-art (SOTA) face recognition on the same dataset.

In summary, this paper makes the following contributions:

◇ Our proposed GaitNet directly learns disentangled representations from RGB videos, which is in sharp contrast to the conventional appearance-based or model-based methods.

◇ We introduce a Frontal-View Gait database, including various variations of viewing angles, walking speeds, carrying, clothing changes, backgrounds and time gaps. This is the first HD gait database, with nearly twice the number of subjects compared to existing RGB gait databases.

◇ Our proposed method outperforms the state of the arts on three benchmarks, CASIA-B, USF, and FVG datasets.

◇ We demonstrate the strength of gait recognition over face recognition in the task of person recognition from surveillance-quality videos.

2 Related Work

Gait Representation. Most prior works are based on two types of gait representations. In appearance-based methods, gait energy image (GEI) [9] or gait entropy image (GEnI) [11] 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 [19], [20] 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 representation directly from raw RGB video frames which contain richer information, thus with higher potential of extracting more discriminative gait features. The most relevant work to ours is [35], which learns gait features from RGB images via Conditional Random Field. Compared to [35], our

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

Dataset	#Subjects	#Videos	Environment	FPS	Resolution	Format	Variations
CASIA-B [32]	124	13,640	Indoor	25	320×240	RGB	View, Clothing, Carrying
USF [10]	122	1,870	Outdoor	30	720×480	RGB	View, Ground Surface, Shoes, Carrying, Time
OU-ISIR-LP [33]	4,007	-	Indoor	-	640×480	Silhouette	View
OU-ISIR-LP-Bag [34]	62,528	-	Indoor	-	1,280×980	Silhouette	Carrying
FVG (ours)	226	2,856	Outdoor	15	1,920×1,080	RGB	View, Walking Speed, Carrying, Clothing, Multiple people, Time

proposed approach learns two complimentary features: dynamic gait, and static gait features, and has the advantage of being able to leverage a large amount of training data and learning more discriminative representation from data with multiple covariates. In addition, some recent works [12], [16], [23], [24], [36] use CNN to learn more discriminative features from GEI. However, the source of the learning, GEI, already loses dynamic information since a random shuffle of video frames results in the *identical* GEI feature. In contrast, the proposed GaitNet learns features from RGB imagery instead, which allows the network to explore richer information for representation learning. This is demonstrated by our comparison with [12], [35] in Sec. 5.2.1 and Sec. 5.2.3.

Gait Databases. There are many classic gait databases such as SOTON Large dataset [37], USF [10], CASIA-B [32], OU-ISIR [34], and TUM GAID [38]. We compare our FVG database with the widely used ones in Tab. 1. CASIA-B is a large multi-view gait database with three variations: viewing 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 captured in NM, CL and BG conditions, respectively. USF database has 122 subjects with five variations, totaling 32 conditions per subject. It contains two viewing angles (left and right), two ground surfaces (grass and concrete), shoe change, carrying condition and time. While OU-ISIR-LP and OU-ISIR-LP-Bag are large databases, only silhouettes are publicly released in both of them. In contrast, our FVG focuses on the frontal view, with 3 different near frontal-view angles toward the camera, and other variations including walking speed, carrying, clothing, multiple people and time.

Disentanglement Learning. Besides model-based approaches representing data with semantic latent vectors [39]–[42], data-driven disentangled representation learning approaches are gaining popularity in the computer vision community. DrNet [43] 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 [44] segments foreground masks of body parts by 2D pose joints via U-Net [45] and then transforms body parts to desired motion with adversarial training. Similarly, [46] utilizes U-net and Variational Auto Encoder (VAE) [47] to disentangle an image into appearance and shape. DR-GAN [48], [49] achieves SOTA performances on pose-invariant face recognition by explicitly disentangling pose variation with a multi-task GAN [50]. Different from [43], [44], [46], our method has only one encoder to disentangle the three latent features, through the design of novel loss functions without the need for adversarial training. Further, pose labels are used in DR-GAN training so as to disentangle identity feature from the pose. However, to disentangle pose and appearance features from RGB, there is no pose nor appearance *label* to be utilized for our method, since it is nontrivial to define the types of walking pattern or clothes as discrete classes.

TABLE 2: Symbols and notations.

Symbol	Dim.	Notation
s	scalar	Index of subject
c	scalar	Condition
t	scalar	Time step in a video
n	scalar	Number of frames in a video
$\mathbf{X}^c$	matrices	Gait video under condition c
$\mathbf{x}^{c,t}$	matrix	Frame t of video $\mathbf{X}^c$
$\hat{\mathbf{x}}$	matrix	Reconstructed frame via $\mathcal{D}$
$\mathcal{E}$	-	Encoder network
$\mathcal{D}$	-	Decoder network
C^{sg}	-	Classifier for $\mathbf{f}_c$
C^{dg}	-	Classifier for $\mathbf{f}_{\text{dyn-gait}}$
$\mathbf{f}_p$	64×1	Pose feature
$\mathbf{f}_c$	128×1	Canonical feature
$\mathbf{f}_a$	128×1	Appearance feature
$\mathbf{f}_{\text{dyn-gait}}$	256×1	Dynamic gait feature
$\mathbf{f}_{\text{sta-gait}}$	128×1	Static gait feature
$\mathbf{h}^t$	128×1	The output of LSTM at step t
$\mathcal{L}_{\text{xrecon}}$	-	Reconstruction loss
$\mathcal{L}_{\text{pose-sim}}$	-	Pose similarity loss
$\mathcal{L}_{\text{cano-sim}}$	-	Canonical similarity loss
$\mathcal{L}_{\text{id-inc-avg}}$	-	Incremental identity loss

Gait vs. Face recognition. Both gait and face are popular biometrics modalities, especially in covert identification-at-a-distance applications. Hence, it is valuable to understand the pros and cons of each modality if the SOTA gait recognition and face recognition algorithms are deployed. Along this direction, most of the prior works focus on the fusion of both modalities and evaluate on relatively small datasets [51]–[53]. In contrast, we conduct comprehensive evaluations using SOTA face and gait recognition algorithms, across various conditions of CASIA-B and FVG databases. Further, the performances are measured along the video duration to explore the impact of person-to-camera distances.

3 Proposed Approach

3.1 Overview

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 extracted feature, which would make video 2 and 3 to be more similar than 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

(a) The same subject **(b)** Different subjects

Fig. 2: If we may ignore the differences in color/texture of clothing and the body pose, there are inherent body characteristics that are different across subjects (b), and invariant within the same subject (a). These include overall body shape, arm length, torso vs. leg ratio, *etc*. We define *canonical feature* to specifically describe these characteristics.

angles, walking speeds and changing clothes. Table 2 summarizes the symbol and notation used in this paper.

Our approach to achieve this objective is feature disentanglement. In our preliminary work [31], we disentangle features into two components: pose and "appearance" features. However, further research discovered that the "appearance" feature still contains certain discriminative information, which can be useful for identity classification. For instance, as in Fig. 2, imagining if we would ignore the body pose, *e.g.*, position of arms and legs, and clothing information, *e.g.*, color and texture of clothes, we may still tell apart different subjects by their *inherent body characteristics*, which can include categories of overall body shape (*e.g.*, rectangle, triangle, inverted triangle, and hourglass [54]), arm length, torso vs. leg ratio [55], *etc*. In other words, even when different people wearing exactly the same clothing and standing still, these characteristics are still subject dependent. In the meantime, for the same subject under various conditions, these characteristics are relatively constant. In this work, we term the feature describing these characteristics as the *canonical feature*. Hence, given a walking video $\mathbf{X}^c$ under condition c, our framework disentangle the encoded feature into three components: the pose feature $\mathbf{f}_p$, the appearance feature $\mathbf{f}_a$ and the canonical feature $\mathbf{f}_c$. We also term the concatenation of $\mathbf{f}_a$ and $\mathbf{f}_c$ as the pose-irrelevant feature, which is conceptually equivalent to the "appearance" feature in [31]. The pose feature describes the positions of body parts, and their dynamic over time is essentially the core element of gait; the canonical feature defines the unique characteristics of individual body; and the appearance feature describes the subject's clothing.

The above feature disentanglement can be naturally implemented as an encoder-decoder network. Specifically, as depicted in Fig. 3, the input to our GaitNet is a video sequence, with background removed using any off-the-shelf pedestrian detection and segmentation method [56]–[58]. With carefully designed loss functions, an encoder is learned to disentangle the pose, canonical and appearance features for each video frame. Then, a multi-layer LSTM explores the temporal dynamics of pose features and aggregates them to a sequence-based dynamic gait feature. In the meantime, the average of all the canonical features is defined as the static gait feature. Measuring distances of both dynamic and static features between the gallery and probe walking videos provides the final matching score. In this section, we first present the feature disentanglement, followed by temporal aggregation, model inference and finally implementation details.

TABLE 3: The properties of three disentangled features in terms of its constancy across frames and conditions, and discriminativeness. These properties are the basis for us to design loss functions for feature disentanglement.

	Constant Across Frames	Constant Across Conditions	Discriminative
$\mathbf{f}_a$	Yes	No	No
$\mathbf{f}_c$	Yes	Yes	Yes
$\mathbf{f}_p$	No	Yes	Yes for $\mathbf{f}_p$ over t

3.2 Feature Disentanglement

For the majority of gait datasets, there is limited intra-subject appearance variation. Hence, appearance could be a discriminative cue for identification during training as many subjects can be easily distinguished by their clothes. Unfortunately, any 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 ideal feature extractors if solely relying on identification objective. Hence we propose to learn to disentangle the canonical and pose feature from the visual appearance. Since a video is composed of frames, disentanglement should be conducted at the frame level first.

Before presenting the details of how we conduct disentanglement, let us first understand the various properties of three types of features, as summarized in Tab. 3. These properties are crucial in guiding us to define effective loss functions for disentanglement. The appearance feature mainly describes the clothing information of the subject. Hence it is constant within a video sequence, but often different across different conditions. Of course it is not discriminative among individuals. The canonical feature is subject-specific, and is therefore constant across both video frames, and conditions. The pose feature is obviously different across video frames, but is assumed to be constant across conditions. Since the pose feature is the manifestation of video-based gait information at a specific frame, the pose feature itself might not be discriminative. However, the dynamics of pose features over time will constitute the dynamic gait feature, which is discriminative among individuals.

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

$$\mathbf{f}_a, \mathbf{f}_c, \mathbf{f}_p = \mathcal{E}(\mathbf{x}). \tag{1}$$

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

$$\hat{\mathbf{x}} = \mathcal{D}(\mathbf{f}_a, \mathbf{f}_c, \mathbf{f}_p). \tag{2}$$

We now define the various loss functions to jointly learn the encoder $\mathcal{E}$ and decoder $\mathcal{D}$.

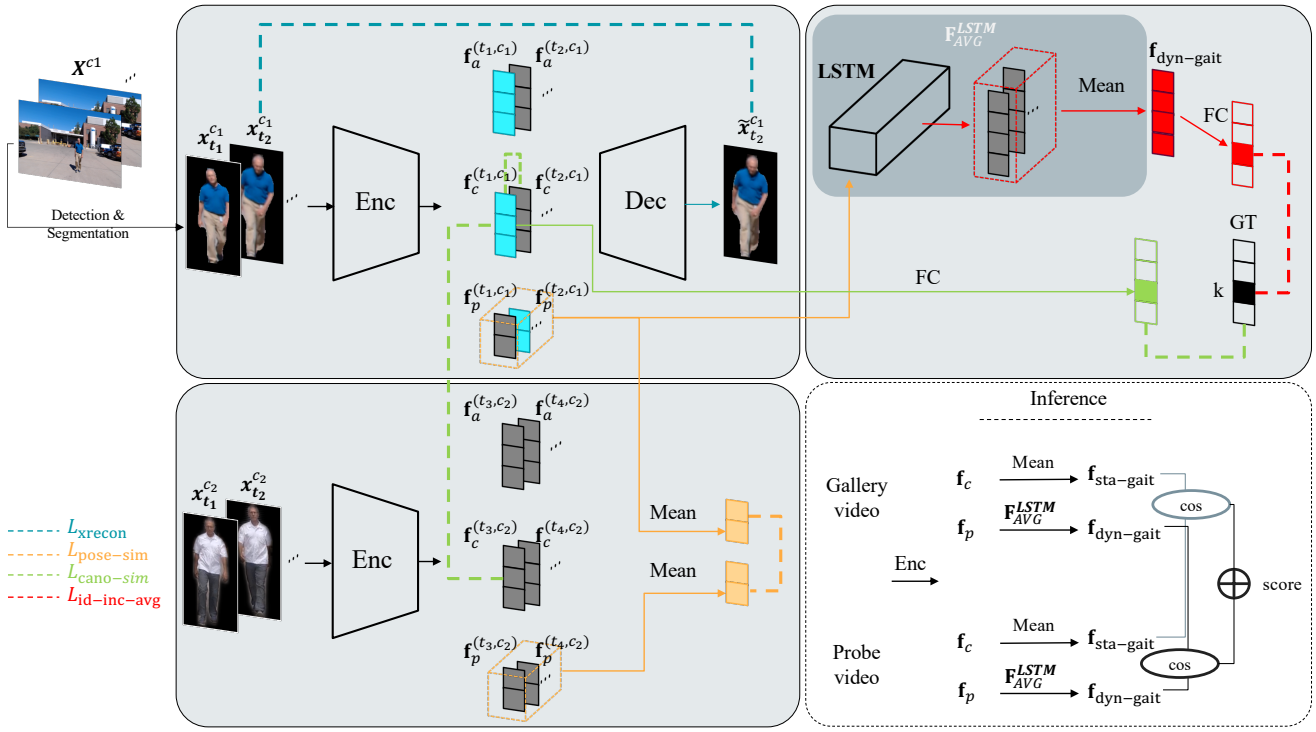

Fig. 3: The overall architecture of proposed GaitNet. The bottom right block indicates the inference process, while the remaining illustrates the training process with the four color-coded loss functions. All the symbols can be refereed in Tab. 2. The blue dash line, which indicate the cross reconstruction loss, enforces both the canonical and appearance features to be similar across all frames within a video. The yellow dash line, which represents the pose similarity loss, encourages $\mathbf{f}_p$ containing only the pose information. The green dash line is the canonical consistency loss, enforces the encoder to extract the unique body characteristics of each subject into $\mathbf{f}_c$.

Cross Reconstruction Loss. The reconstructed image $\hat{\mathbf{x}}$ should be close to the original input $\mathbf{x}$. However, enforcing self-reconstruction loss as in typical auto-encoder cannot ensure the meaningful disentanglement as in our design. Hence, we propose the cross reconstruction loss, using the appearance feature $\mathbf{f}_a^{t_1}$ and canonical feature $\mathbf{f}_c^{t_1}$ of frame t_1 and the pose feature $\mathbf{f}_p^{t_2}$ of frame t_2 to reconstruct the latter frame:

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

The cross reconstruction loss, on one hand, can act as the self-reconstruction loss to make sure the three features are sufficiently representative to reconstruct a video frame. On the other hand, as we can pair a pose feature of a current frame with the canonical and appearance features of *any* frame in the same video to reconstruct the same target, it enforces both the canonical and appearance features to be similar across all frames within a video. Indeed, according to Tab. 3, between the pose-irrelevant feature, $\mathbf{f}_a \& \mathbf{f}_c$, and the pose feature $\mathbf{f}_p$, the main distinct property is that the former is constant across frames while the latter is not. This is the basis for designing our cross reconstruction loss.

Pose Similarity Loss. The cross reconstruction loss is able to prevent the pose-irrelevant feature, $\mathbf{f}_a \& \mathbf{f}_c$, to be contaminated by the pose information that changes across frames. If not, *i.e.*, $\mathbf{f}_a$ or $\mathbf{f}_c$ contains some pose information, $\mathcal{D}(\mathbf{f}_a^{t_1}, \mathbf{f}_c^{t_1}, \mathbf{f}_p^{t_2})$ and $\mathbf{x}^{t_2}$ would have different poses. However, clothing/texture and body information may still be leaked into the pose feature $\mathbf{f}_p$. In the extreme case, $\mathbf{f}_c$ and $\mathbf{f}_a$ could be constant vectors while $\mathbf{f}_p$ encodes all the information of a video frame.

To encourage $\mathbf{f}_p$ including *only* the pose information, we leverage multiple videos of the same subject. Given two videos of the same subject with length n_1, n_2 in two different conditions c_1, c_2, they contain difference in the person's appearance, *i.e.*, cloth changes. Despite appearance changes, the gait information is assumed to be constant between two videos. Since it's almost impossible to enforce similarity on $\mathbf{f}_p$ between video frames as it requires precise frame-level alignment, we minimize the similarity between two videos' averaged pose features:

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

According to Tab. 3, the pose feature is constant across conditions, which is the basis of our pose similarity loss.

Canonical Similarity Loss. The canonical feature describes the subject's body characteristics, which is unique over all video frames. To be specific, for two videos of the same subject k in two different conditions c_1, c_2, the canonical feature is constant across both frames and conditions, as illustrated in Tab. 3. Tab. 3 also states that the canonical feature is discriminative across subjects. Hence, to enforce the two constancy and the discriminativeness, we define

the canonical consistency loss as follows:

$$\begin{aligned}\mathcal{L}_{\text{cano-sim}} =& \frac{2}{n_1(n_1-1)}\sum_{i\neq j}\left\|\mathbf{f}_c^{(t_i,c_1)}-\mathbf{f}_c^{(t_j,c_1)}\right\|_2^2 \\ &+\frac{1}{n_1}\sum_i\left\|\mathbf{f}_c^{(t_i,c_1)}-\mathbf{f}_c^{(t_i,c_2)}\right\|_2^2 \\ &+\frac{1}{n_1}\sum_i -\log(C_k^{sg}(\mathbf{f}_c^{(t_i,c1)}))),\end{aligned} \tag{5}$$

where the three terms measure the consistency across frames in a single video, consistency across different videos of the same subject, and identity classification using a classifier C^{sg} which is a fully connection layer with ReLU activation, respectively.

3.3 Gait Feature Learning and Aggregation

Even when we can disentangle pose, canonical and appearance information for each video frame, the $\mathbf{f}_p$ and $\mathbf{f}_c$ have to be aggregated over time, since 1) gait recognition is conducted between two videos instead of two images; 2) not all the $\mathbf{f}_c$ from every single frame is guaranteed to have same canonical information; 3) the current feature $\mathbf{f}_p$ only represents the walking pose of the person at a specific instance, which can share similarity with another instance of a different individual. Here, we are looking for discriminative characteristics in a person's walking pattern. Therefore, modeling its aggregation for $\mathbf{f}_c$ and temporal change for $\mathbf{f}_p$ is critical.

3.3.1 Static Gait Feature via Canonical Feature Aggregation

After learning $\mathbf{f}_c$ for every single frame as defined in Eqn. 5, we explore the best representation of $\mathbf{f}_c$ features across all frames of a video sequence. Since $\mathbf{f}_c$ is assumed to be constant over time, we compute the averaged $\mathbf{f}_c$ features as a way to aggregate the canonical features over time. Given that $\mathbf{f}_c$ describes the body characteristics as if we freeze the gait, we call the aggregated $\mathbf{f}_c$ as the static gait feature $\mathbf{f}_{\text{sta-gait}}$.

$$\mathbf{f}_{\text{sta-gait}} = \frac{1}{n}\sum_{t=1}^{n}\mathbf{f}_c^t. \tag{6}$$

3.3.2 Dynamic Gait Feature via Pose Feature Aggregation

For temporal modeling of poses, 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 temporal information of pose features, *e.g.*, how the trajectory of subjects' body parts changes over time. As shown in Fig. 3, pose features extracted from one video sequence are fed into a 3-layer LSTM. The output of the LSTM is connected to a classifier C^{dg}, 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 $\mathbf{f}_p$ into it:

$$\mathbf{h}^t = \text{LSTM}(\mathbf{f}_p^1, \mathbf{f}_p^2, ..., \mathbf{f}_p^t). \tag{7}$$

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:

$$\mathcal{L}_{\text{id-single}} = -\log(C_k^{dg}(\mathbf{h}^n)), \tag{8}$$

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

Identification with Averaged Feature. By the nature of LSTM, the output $\mathbf{h}^t$ can be greatly affected by its last input $\mathbf{f}_p^t$. Hence the LSTM output, $\mathbf{h}^t$, could be unstable across time steps. With a desire to obtain a gait feature that is robust to the final instance of a walking cycle, we choose to use the averaged LSTM output as our gait feature for identification:

$$\mathbf{f}_{\text{dyn-gait}}^t = \frac{1}{t}\sum_{s=1}^{t}\mathbf{h}^s. \tag{9}$$

The identification loss can be rewritten as:

$$\begin{aligned}\mathcal{L}_{\text{id-avg}} &= -\log(C_k^{dg}(\mathbf{f}_{\text{dyn-gait}}^n)) \\ &= -\log\left(C_k^{dg}\left(\frac{1}{n}\sum_{s=1}^{n}\mathbf{h}^s\right)\right).\end{aligned} \tag{10}$$

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

$$\mathcal{L}_{\text{id-inc-avg}} = \frac{1}{\sum_{t=1}^{n} w_t}\sum_{t=1}^{n} -w_t\log\left(C_k^{dg}\left(\frac{1}{t}\sum_{s=1}^{t}\mathbf{h}^s\right)\right), \tag{11}$$

where we set $w_t = t^2$ and other options such as $w_t = 1$ also yield similar performance. In the experiments, we will ablate the impact of three options in classification loss: $\mathcal{L}_{\text{id-single}}$, $\mathcal{L}_{\text{id-avg}}$, and $\mathcal{L}_{\text{id-inc-avg}}$. To this end, the overall loss function is:

$$\mathcal{L} = \mathcal{L}_{\text{id-inc-avg}} + \lambda_r\mathcal{L}_{\text{xrecon}} + \lambda_d\mathcal{L}_{\text{pose-sim}} + \lambda_s\mathcal{L}_{\text{cano-sim}}. \tag{12}$$

The entire system, including 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 from which LSTM is able to explore temporal dynamics.

3.4 Model Inference

Since GaitNet takes one video sequence as input and outputs $\mathbf{f}_{\text{dyn-gait}}$ and $\mathbf{f}_{\text{sta-gait}}$ as shown in Fig. 3, one single score is needed to measure the similarity between the gallery and probe videos for either gait authentication or identification. During testing, both $\mathbf{f}_{\text{sta-gait}}$ and $\mathbf{f}_{\text{dyn-gait}}$ are used as the identity features for score calculation. We use the cosine similarity scores, normalized to the range of $[0, 1]$ via min-max. The static and dynamic scores are finally fused by a weighted sum rule:

$$\text{Score} = (1-\alpha)\frac{\mathbf{f}_{\text{sta-gait}}^g\cdot\mathbf{f}_{\text{sta-gait}}^p}{\left\|\mathbf{f}_{\text{sta-gait}}^g\right\|\left\|\mathbf{f}_{\text{sta-gait}}^p\right\|} + \alpha\frac{\mathbf{f}_{\text{dyn-gait}}^g\cdot\mathbf{f}_{\text{sta-gait}}^p}{\left\|\mathbf{f}_{\text{dyn-gait}}^g\right\|\left\|\mathbf{f}_{\text{dyn-gait}}^p\right\|}, \tag{13}$$

where g and p represent gallery and probe, respectively.

3.5 Implementation Details

Detection and Segmentation. Our GaitNet receives video frames with the person of interest segmented. The foreground mask is obtained from the SOTA instance segmentation algorithm, Mask R-CNN [56]. Instead of using a zero-one mask by hard thresholding, we maintain 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 an appropriate threshold suitable

TABLE 4: The architecture of $\mathcal{E}$ and $\mathcal{D}$ networks. Note the layer with ()* is removed for experiments with small training sets, *i.e.*, all ablation studies in Sec. 5.1, to prevent overfitting.

$\mathcal{E}$			$\mathcal{D}$		
Layers	Filters/Stride	Output Size	Layers	Filters/Stride	Output Size
Conv1	3x3/1	64x32x64	FC	-	4x2x512
MaxPool1	3x3/2	32x16x64	FCConv1	3x3/2	8x4x256
Conv2	3x3/1	32x16x256	FCConv2	3x3/2	16x8x128
MaxPool2	3x3/2	16x8x256	FCConv3	3x3/2	32x16x64
Conv3	3x3/2	16x8x512	FCConv4	3x3/2	32x16x3
(Conv4	3x3/2	16x8x512)*			
MaxPool3	3x3/2	4x2x512			
FC	-	320			

for multiple databases. Also, it remedies 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. The input of GaitNet is obtained by pixel-wise multiplication between the mask and the $[0, 1]$-normalized RGB values, and then resizing to 32×64 pixels. This applies to all the experiments on CASIA-B, USF and FVG datasets in Sec. 5.

Network Structure and Hyperparameter. Our encoder-decoder network is a typical CNN, illustrated in Tab. 4. Different from our preliminary work [31], we replace stride-2 convolution layers with stride-1 convolution layers and max pooling layers, since we find the latter is able to achieve the similar results with less hyper-parameter searching for different training scenarios. Each convolution layer is followed by Batch Normalization and Leaky ReLU activation. The decoder structure, similar to [59], is built from transposed 2D convolution, Batch Normalization and Leaky ReLU layers. The final layer is a Sigmoid activation which can output the value into $[0, 1]$ range as the input. All the transposed convolutions are with stride of 2 to upsample images and all the Leaky ReLU are with slope of 0.2. The classification part is a stacked 3-layer LSTM [60], which has 256 hidden units in each cell. We empirically show that 3-layer LSTM achieves higher performance than 1, 2, or 4-layer. The length of $\mathbf{f}_a$, $\mathbf{f}_c$ and $\mathbf{f}_p$ is 128, 128 and 64 respectively, as shown in Tab. 2.

The Adam optimizer [61] is initialized with the learning rate of 0.0001, and the momentum of 0.9. To prevent over-fitting, the weights decay of 0.001 is applied to all the experiments, and the learning rate decays by multiplying 0.9 in every 500 iterations. For each batch, we use video frames from 16 or 32 different clips depending on different experiment protocols. Since video lengths are varied, a random crop of 20-frame sequence is applied during training; all shorter videos are discarded. The λ_r, λ_s and λ_d in Eqn. 12 are all set to 1 in all experiments.

4 Front-View Gait (FVG) Database

Collection. To facilitate the research of gait recognition from frontal-view angles, we collect the Front-View Gait (FVG) database in 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 1.50 meters. We require each of 226 subjects to walk toward the camera 12 times starting from around 16 meters away from the camera, which results in 12 videos per subject. The videos are captured at $1,080 \times 1,920$ resolution with 15 FPS and the average length of 10 seconds. The height of body in the video ranges from 101 to 909 pixels, and the height of faces ranges from 17 to 467 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. As detailed in Tab. 5, FVG is collected in three sessions with five variations: normal, walking speed (slow and fast), clothing changes, carrying/wearing change (bag or hat), and clutter background (multiple persons). The five variations are well balanced in three sessions. Fig. 4 shows exemplar images from FVG.

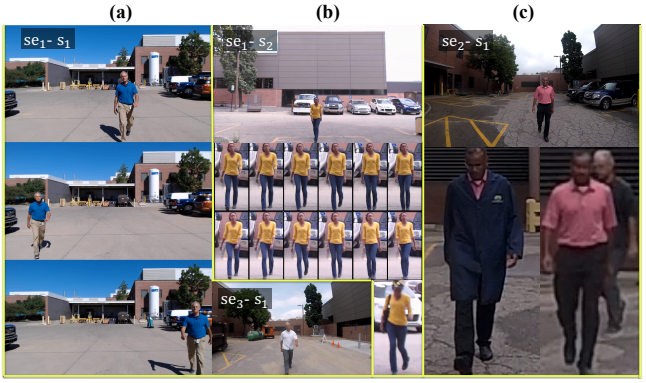

Fig. 4: Examples of FVG Dataset. (a) Samples of the near frontal middle, left and right walking viewing angles in Session 1 (se_1) of the first subject (s_1). se_3-s_1 is the same subject in Session 3. (b) Samples of slow and fast walking speed for another subject in Session 1. Frames in the second row are normal and in the third row are fast walking. Carrying bag and wearing hat sample is shown below. (c) Samples of changing clothes and with multiple people background from one subject in Session 2.

TABLE 5: The FVG database. The last 5 rows show the specific variations that are captured by each of 12 videos per subject.

Collection Year	2017			2018					
Session	1			2			3		
Number of Subjects	147			79			12		
Viewing Angle (°)	-45	0	45	-45	0	45	-45	0	45
Normal	1	2	3	1	2	3	1	2	3
Fast / Slow Walking	4/7	5/8	6/9	4	5	6	4	5	6
Carrying Bag / Hat	10	11	12	-	-	-	-	-	-
Change Clothes	-	-	-	7	8	9	7	8	9
Multiple Person	-	-	-	10	11	12	10	11	12

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 visual difference in consecutive frames is normally much smaller than side-view walking. We focus our evaluation on variations that are challenging, *e.g.*, different clothes, carrying a bag while wearing a hat, or are not presented in prior databases, *e.g.*, multi-person. To benchmark research on FVG, we define 5 evaluation protocols, among which there are two commonalities: 1) the first 136 and remaining 90 subjects are used for training and testing respectively; 2) the video 2, the normal frontal-view walking, is always used as the gallery. The 5 protocols differ in their respective probe data, which cover the variations of Walking Speed (WS), Carrying Bag while Wearing a Hat (BGHT), Changing Clothes (CL), Multiple Persons (MP), and all variations (ALL). At the top part of Tab. 5, we list the detailed probe sets for all 5 protocols. For instance, for the WS protocol, the probes are video $4-9$ in Session 1 and video $4-6$ in Session 2. In all protocols, the performance metrics are the True Accept Rate (TAR) at 1% and 5% False Alarm Rate (FAR).

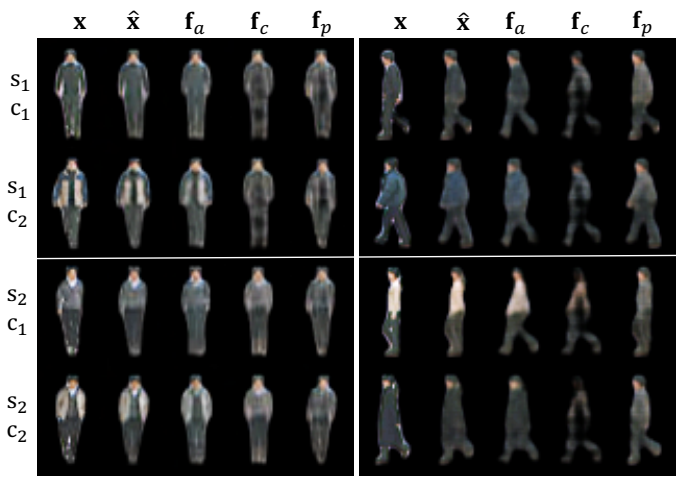

Fig. 5: Synthesis by decoding three features individually, $\mathbf{f}_a$, $\mathbf{f}_c$ and $\mathbf{f}_p$, and their concatenation. Left and right parts are two learnt models on frontal and side views of CASIA-B. The top two rows are two frames of the same subject under different conditions (NM vs. CL) and the bottom two are another subject. The reconstructed frames $\hat{\mathbf{x}}$ closely match the original input. $\mathbf{f}_c$ shows consistent body shape for the same subject while different for different subjects. $\mathbf{f}_a$ recovers the appearance of clothes, at the pose specified by $\mathbf{f}_c$. The body pose of $\mathbf{f}_p$ matches with the input frame.

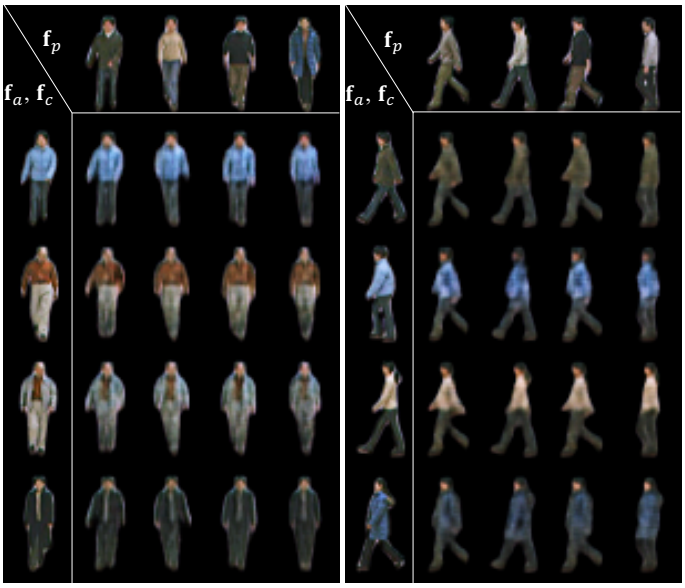

Fig. 6: Synthesis by decoding pairs of pose features $\mathbf{f}_p$ and pose-irrelevant features, $\{\mathbf{f}_a, \mathbf{f}_c\}$. Left and right parts are examples of frontal and side views of CASIA-B. In either part, each of 4×4 synthetic images is $\mathcal{D}(\mathbf{f}_a^l, \mathbf{f}_c^l, \mathbf{f}_p^t)$, where $\{\mathbf{f}_a^l, \mathbf{f}_c^l\}$ is extracted from images in the first column and $\mathbf{f}_p^t$ is from the top row. The synthetic images resemble the appearance of the first column and the pose of the top row.

5 Experimental Results

We evaluate the proposed approach on three gait databases, CASIA-B [32], USF [10] and FVG. As mentioned in Sec. 2, CASIA-B and USF are the most widely used gait databases, which helps us to make the comprehensive comparison with prior works. We compare our method with [12], [35], [62], [63] on these two databases, by following the respective experimental protocols of the baselines. These are either the most recent and SOTA work, or classic gait recognition methods. The OU-ISIR database [34] is not evaluated, and related results [26] are not compared since our work consumes RGB video input, but OU-ISIR only releases silhouettes. Finally, we also conduct experiments to compare our gait recognition with the state-of-the-art face recognition method ArcFace [64] on the CASIA-B and FVG datasets.

5.1 Ablation Study

5.1.1 Feature Visualization Through Synthesis

While our decoder is only useful in training, but not model inference, it can enable us to visualize the disentangled features as a synthetic image, by feeding either the feature itself, or their random concatenation, to our learned decoder $\mathcal{D}$. This synthesis helps to gain more understanding of the feature disentanglement.

Visualization of Features in One Frame. Our decoder requires the concatenation of three vectors for synthesis. Hence, to visualize each individual feature, we concatenate it with two vectors of zeros and then feed to decoder. In Fig. 5, we show the disentanglement visualization of 4 subjects (two frontal and two side views), each under the NM and CL conditions. First of all, the canonical feature discovers a *standard* body pose that is consistent across both subjects, which is more visible in the side view. Under such a standard body pose, the canonical feature then depicts the unique body shape, which is consistent within a subject but different between subjects. The appearance feature faithfully recovers the color and texture of clothing, at the standard body pose specified by the canonical feature. The pose feature captures the walking pose of the input frame. Finally, combining all three features can closely reconstruct the original input. This shows that our disentanglement not only preserves all information of the input, but also fulfills all the desired properties described in Tab. 3.

Visualization of Features in Two Frames. As shown in Fig. 6, each result is generated by pairing the pose-irrelevant feature $\{\mathbf{f}_a, \mathbf{f}_c\}$ in the first column, and the pose feature $\mathbf{f}_p$ in the first row. The synthesized images show that indeed pose-irrelevant feature contributes all the appearance and body information, *e.g.*, cloth, body width, as they are consistent across each row. Meanwhile, $\mathbf{f}_p$ contributes all the pose information, *e.g.*, positions of hand and feet, which share similarity across columns. Despite that concatenating vectors from different subjects may create samples outside the input distribution of $\mathcal{D}$, the visual quality of synthetic images shows that $\mathcal{D}$ is versatile to these new samples.

5.1.2 Feature Visualization Through t-SNE

To gain more insight into the frame-level features $\mathbf{f}_a$, $\mathbf{f}_c$, $\mathbf{f}_p$ and sequence-level LSTM feature aggregation, we apply t-SNE [65] to these features to visualize their distribution in a 2D space. With the learnt models in Sec. 5.1.1, we randomly select two videos under NM and CL conditions for each of 5 subjects.

Fig. 7 (a,b) visualizes the $\mathbf{f}_a$ and $\mathbf{f}_c$ features. Obviously, for the appearance feature $\mathbf{f}_a$, the margins between intra-class and inter-class distances are unpromising, which shows that $\mathbf{f}_a$ has limited discrimination power. In contrast, the canonical feature $\mathbf{f}_c$ has both the compact intra-class variations and separable inter-class differences – useful for identity classification. In addition, we visualize the $\mathbf{f}_p$ from $\mathcal{E}$ and its corresponding $\mathbf{f}_{\text{dyn-gait}}$ at each time step in Fig. 7 (c-d). As defined in Eqn. 4, we enforce the averaged $\mathbf{f}_p$ of the same subject to be consistent under different conditions. Since Eqn. 4 only minimizes the intra-class distance, it cannot guarantee the discrimination among subjects. However,

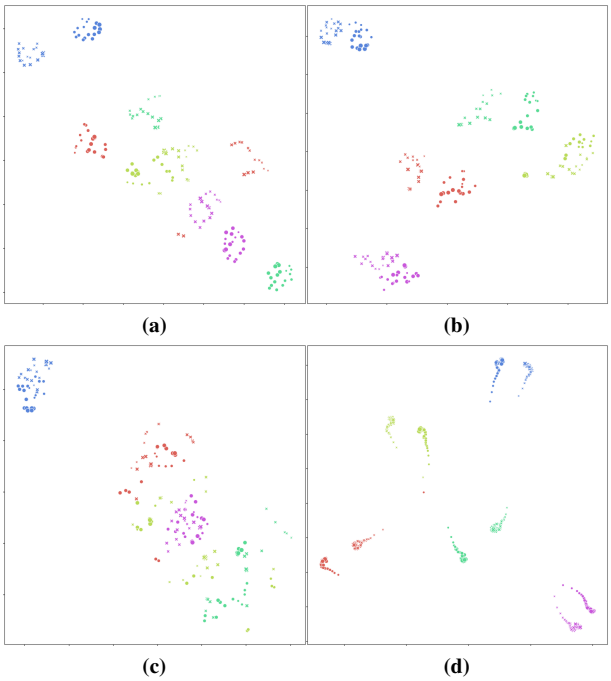

Fig. 7: The t-SNE visualization of (a) appearance features $\mathbf{f}_a$, (b) canonical features $\mathbf{f}_c$, (c) pose features $\mathbf{f}_p$, and (d) dynamic gait features $\mathbf{f}_{\text{dyn-gait}}$. We select 5 subjects each with two videos of NM vs. CL conditions. Each point represents a single frame, whose color is for subject ID, shape of 'dot' and 'cross' is NM and CL respectively, and size is frame index. We see that $\mathbf{f}_c$ and $\mathbf{f}_{\text{dyn-gait}}$ are far more discriminative than $\mathbf{f}_a$ and $\mathbf{f}_p$.

after aggregation by the LSTM network, distances of points at longer time duration for inter-class are substantially enlarged.

5.1.3 Loss Function's Impact on Performance

Disentanglement with Pose Similarity Loss. With the cross reconstruction loss, the appearance feature $\mathbf{f}_a$ and canonical feature $\mathbf{f}_c$ can be enforced to represent static information that shares across the video. However, as discussed, $\mathbf{f}_p$ could be contaminated by the appearance information or even encode the entire video frame. Here we show the benefit of the pose similarity loss $\mathcal{L}_{\text{pose-sim}}$ to feature disentanglement. Fig. 8 shows the cross visualization of two different models learned with and without $\mathcal{L}_{\text{pose-sim}}$. Without $\mathcal{L}_{\text{pose-sim}}$ the decoded image shares some appearance and body characteristic, *e.g.*, cloth style, contour, with $\mathbf{f}_p$. Meanwhile, with $\mathcal{L}_{\text{pose-sim}}$, appearance better matches with $\mathbf{f}_a$ and $\mathbf{f}_c$.

Loss Function's Impact on Recognition Performance. As there are various options in designing our framework, we ablate their effect on the final recognition performance from three perspectives: the disentanglement loss, the classification loss, and the classification feature. Tab. 6 reports the Rank-1 recognition accuracy of different variants of our framework on CASIA-B under NM vs. CL and lateral view. The model is trained with all videos of the first 74 subjects and tested on the remaining 50 subjects.

We first explore the effects of different disentanglement losses applied to $\mathbf{f}_{\text{dyn-gait}}$ and use $\mathbf{f}_{\text{dyn-gait}}$ only for classification. 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 model with both $\mathcal{L}_{\text{xrecon}}$ and $\mathcal{L}_{\text{pose-sim}}$. The

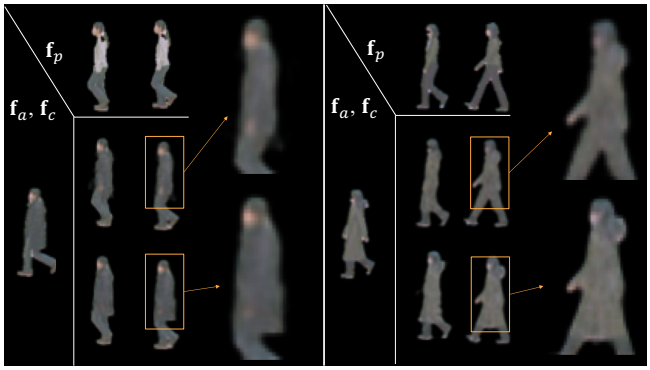

Fig. 8: Synthesis on CASIA-B by decoding pose-irrelevant feature $\{\mathbf{f}_a, \mathbf{f}_c\}$ and pose feature $\mathbf{f}_p$ from videos under NM vs. CL conditions. Left and right parts are two examples. For each example, $\{\mathbf{f}_a, \mathbf{f}_c\}$ is extracted from the first column (CL) and $\mathbf{f}_p$ is from the top row (NM). Top row synthetic images are generated from model trained *without* $\mathcal{L}_{\text{pose-sim}}$ loss, bottom row is *with* the loss. To show the difference, details in synthetic images are magnified.

TABLE 6: Ablation study on various options of the disentanglement loss, classification loss, and classification features. A GaitNet model is trained on NM and CL conditions of lateral view with the first 74 subjects of CASIA-B and tested on remaining subjects.

Disentanglement Loss	Classification Loss	Classification Feature	Rank-1
-	$\mathcal{L}_{\text{id-inc-avg}}$	$\mathbf{f}_{\text{dyn-gait}}$	56.0
$\mathcal{L}_{\text{xrecon}}$	$\mathcal{L}_{\text{id-inc-avg}}$	$\mathbf{f}_{\text{dyn-gait}}$	60.2
$\mathcal{L}_{\text{xrecon}} + \mathcal{L}_{\text{pose-sim}}$	$\mathcal{L}_{\text{id-inc-avg}}$	$\mathbf{f}_{\text{dyn-gait}}$	85.6
$\mathcal{L}_{\text{xrecon}} + \mathcal{L}_{\text{pose-sim}} + \mathcal{L}_{\text{cano-sim}}$	$\mathcal{L}_{\text{id-single}}$	$\mathbf{f}_{\text{dyn-gait}}$ & $\mathbf{f}_{\text{sta-gait}}$	72.5
$\mathcal{L}_{\text{xrecon}} + \mathcal{L}_{\text{pose-sim}} + \mathcal{L}_{\text{cano-sim}}$	$\mathcal{L}_{\text{id-ae}}$ [66]	$\mathbf{f}_{\text{dyn-gait}}$ & $\mathbf{f}_{\text{sta-gait}}$	76.5
$\mathcal{L}_{\text{xrecon}} + \mathcal{L}_{\text{pose-sim}} + \mathcal{L}_{\text{cano-sim}}$	$\mathcal{L}_{\text{id-avg}}$	$\mathbf{f}_{\text{dyn-gait}}$ & $\mathbf{f}_{\text{sta-gait}}$	82.6
$\mathcal{L}_{\text{xrecon}} + \mathcal{L}_{\text{pose-sim}} + \mathcal{L}_{\text{cano-sim}}$	$\mathcal{L}_{\text{id-inc-avg}}$	$\mathbf{f}_a$	33.4
$\mathcal{L}_{\text{xrecon}} + \mathcal{L}_{\text{pose-sim}} + \mathcal{L}_{\text{cano-sim}}$	$\mathcal{L}_{\text{id-inc-avg}}$	$\mathbf{f}_{\text{sta-gait}}$	76.3
$\mathcal{L}_{\text{xrecon}} + \mathcal{L}_{\text{pose-sim}} + \mathcal{L}_{\text{cano-sim}}$	$\mathcal{L}_{\text{id-inc-avg}}$	$\mathbf{f}_{\text{dyn-gait}}$	85.9
$\mathcal{L}_{\text{xrecon}} + \mathcal{L}_{\text{pose-sim}} + \mathcal{L}_{\text{cano-sim}}$	$\mathcal{L}_{\text{id-inc-avg}}$	$\mathbf{f}_{\text{dyn-gait}}$ & $\mathbf{f}_{\text{sta-gait}}$	**92.1**

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

We also compare the effects of different classification losses applied to $\mathbf{f}_{\text{dyn-gait}}$. Even though the classification loss only affects $\mathbf{f}_{\text{dyn-gait}}$, we report the performance with both $\mathbf{f}_{\text{dyn-gait}}$ and $\mathbf{f}_{\text{sta-gait}}$ for a direct comparison with our full model in the last row. With the disentanglement loss of $\mathcal{L}_{\text{xrecon}}$, $\mathcal{L}_{\text{pose-sim}}$ and $\mathcal{L}_{\text{cano-sim}}$, we benchmark different options of the classification loss as presented in Sec. 3.2, as well as the autoencoder loss by Srivastava *et al.* [66]. The model using the conventional identity loss on the final LSTM output $\mathcal{L}_{\text{id-single}}$ achieves the rank-1 accuracy of 72.5%. Using the average output of LSTM as the identity feature, $\mathcal{L}_{\text{id-avg}}$ improves the accuracy to 82.6%. The autoencoder loss [66] achieves a good performance of 76.5%. However, it is still far from our proposed incremental identity loss $\mathcal{L}_{\text{id-inc-avg}}$'s performance at 92.1%. Fig. 9 further visualizes the $\mathbf{f}_{\text{dyn-gait}}$ over time, for two models learnt with $L_{\text{id-single}}$ and $\mathcal{L}_{\text{id-inc-avg}}$ loss respectively. Clearly, even with less than 10 frames, the model with $\mathcal{L}_{\text{id-inc-avg}}$ shows more discriminativeness, which also increases rapidly as time progresses.

Finally, we compare different features in computing the final classification score. The performance is based on the model with full disentanglement losses and $\mathcal{L}_{\text{id-inc-avg}}$ as the classification loss.

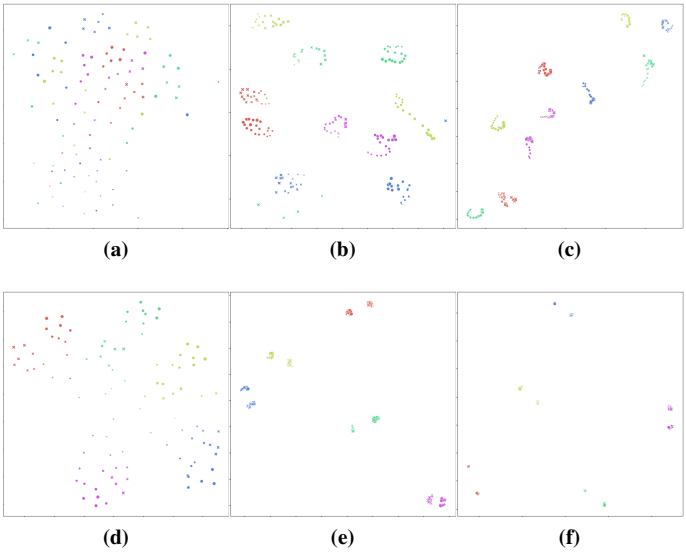

Fig. 9: The t-SNE visualization of $\mathbf{f}_{\text{dyn-gait}}$ from 5 subjects, each with 2 videos (NM vs. CL). The symbols are defined the same as Fig. 7. The top and bottom rows are two models learnt with $L_{\text{id-single}}$ and $\mathcal{L}_{\text{id-inc-avg}}$ loss respectively. From left to tight, the points are $\mathbf{f}_{\text{dyn-gait}}$ of the first 10 frames, 10-30 frames, and 30-60 frames. Learning with $\mathcal{L}_{\text{id-inc-avg}}$ leads to more discriminative dynamic features for the entire duration.

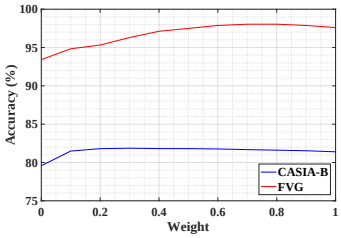

Fig. 10: Recognition by fusing $\mathbf{f}_{\text{dyn-gait}}$ and $\mathbf{f}_{\text{sta-gait}}$ scores with different weights as defined in Eqn. 13. Rank-1 accuracy and TAR@1% FAR is calculated for CASIA-B and FVG, respectively.

When $\mathbf{f}_a$ is utilized in cosine distance calculation, the rank-1 accuracy is merely 33.4%, while $\mathbf{f}_{\text{sta-gait}}$ and $\mathbf{f}_{\text{dyn-gait}}$ achieve 76.3% and 85.9% respectively. The results prove the learnt $\mathbf{f}_c$ and $\mathbf{f}_p$ are effective for classification while $\mathbf{f}_a$ has limited discriminative power. Also, by combining both $\mathbf{f}_{\text{sta-gait}}$ and $\mathbf{f}_{\text{dyn-gait}}$ features, the recognition performance can be further improved to 92.1%. We believe that such performance gain is owing to the complementary discriminative information offered by $\mathbf{f}_{\text{sta-gait}}$ w.r.t. $\mathbf{f}_{\text{dyn-gait}}$.

5.1.4 Dynamic vs. Static Gait Features

Since $\mathbf{f}_{\text{dyn-gait}}$ and $\mathbf{f}_{\text{sta-gait}}$ are complementary in classification, it is interesting to understand their relative contributions, especially in the various scenarios of gait recognition. This amounts to exploring a global weight α for the GaitNet on various training data, where α ranges from 0 to 1. There are three protocols on CASIA-B and hence three GaitNet models are trained respectively. We calculate the weighted score of all three models on the training data of protocol 1, since it is the most comprehensive and representative protocol covering all the viewing angles and conditions. The same experiment is conducted on "ALL" protocol of the FVG dataset.

As shown in Fig. 10, GaitNet has the best average performance on CASIA-B when α is around 0.2, while on FVG α is around

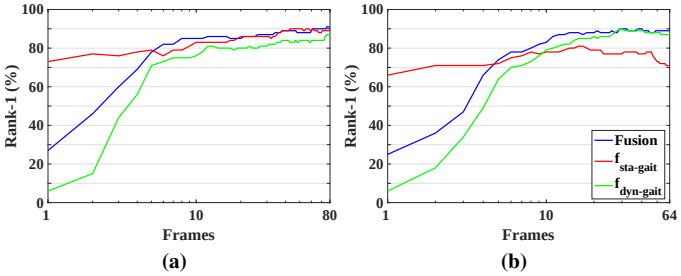

Fig. 11: Recognition performance at different video lengths. We use different feature scores ($\mathbf{f}_{\text{sta-gait}}$, $\mathbf{f}_{\text{dyn-gait}}$, and their fusion) on NM-CL, BG conditions of CAISA-B. (a) is on frontal-frontal view and (b) is on side-side views.

0.75. According to Eqn. 13, $\mathbf{f}_{\text{sta-gait}}$ has relatively more classification contributions on CASIA-B. One potential reason is that it is more challenging to match dynamic walking poses under large range of viewing angles. In comparison, FVG favors $\mathbf{f}_{\text{dyn-gait}}$. Since FVG is an all-frontal-walking dataset containing varying distances or resolutions, dynamic gait is relatively easier to learn with the fixed view, while $\mathbf{f}_{\text{sta-gait}}$ might be sensitive to resolution changes.

Nevertheless, note that in the two extreme cases, where only $\mathbf{f}_{\text{sta-gait}}$ or $\mathbf{f}_{\text{dyn-gait}}$ is used, there is relatively small performance gap between them. This means that either feature is effective in classification. Considering this observation and the balance between databases, we choose to set α=0.5, which will be used in all subsequent experiments.

5.1.5 Gait Recognition Over Time

One interesting question to study is that, how many video frames are needed to achieve reliable gait recognition. To answer this question, we compare the performance with different feature scores ($\mathbf{f}_{\text{sta-gait}}$, $\mathbf{f}_{\text{dyn-gait}}$ and their fusion) for identification, with different video lengths. As shown in Fig. 11, both dynamic and static features achieve stable performance starting from about 10 frames, after which the gain in performance is relatively small. At 15 FPS, a clip of 10 frames is equivalent to merely 0.7 seconds of walking. Further, the static gait feature has notable good performance even with a single video frame. This impressive result shows the strength of our GaitNet in processing very short clips. Finally, for most of the frames in this duration, the fusion outperforms both the static and dynamic gait feature alone.

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 [12], Protocol 1 uses the first 74 subjects for training and remaining 50 for testing, regarding variations of NM (normal), BG (carrying bag) and CL (wearing a coat) with crossing viewing angles of 0° to 180°. Three models are trained for comparison in Tab. 7. For the detailed protocol, please refer to [12]. Here we mainly compare our work to Wu *et al.* [12], along with other methods [36], [67]. We denote our preliminary work [31] as GaitNet-pre and this work as GaitNet. Under multiple viewing angles and across three variations, GaitNet achieves the best performance compared to all SOTA methods and GaitNet-pre since $\mathbf{f}_c$ can distill more discriminative information under various viewing angles and conditions.

TABLE 7: 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 viewing angles.

Gallery NM #1-4	0°-180° (exclude identical viewing angle)											
Probe NM #5-6	0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	Mean
ViDP [67]	-	-	-	64.2	-	60.4	-	65.0	-	-	-	-
LB [12]	82.6	90.3	96.1	94.3	90.1	87.4	89.9	94.0	94.7	91.3	78.5	89.9
3D MT network [12]	87.1	93.2	**97.0**	94.6	90.2	88.3	91.1	93.8	**96.5**	**96.0**	85.7	92.1
J-CNN [36]	87.2	93.2	96.3	**95.9**	**91.6**	86.5	89.8	93.8	95.1	93.0	80.8	91.2
GaitNet-pre [31]	91.2	92.0	90.5	95.6	86.9	92.6	93.5	**96.0**	90.9	88.8	89	91.6
GaitNet	**93.1**	92.6	90.8	92.4	87.6	**95.1**	**94.2**	95.8	92.6	90.4	**90.2**	**92.3**
Probe BG #1-2	0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	Mean
LB-subGEI [12]	64.2	80.6	82.7	76.9	64.8	63.1	68.0	76.9	82.2	75.4	61.3	72.4
J-CNN [36]	73.1	78.1	83.1	81.6	71.6	65.5	71.0	80.7	79.1	78.6	68.0	75.0
GaitNet-pre [31]	83.0	87.8	88.3	93.3	82.6	74.8	89.5	91.0	86.1	81.2	85.6	85.7
GaitNet	**88.8**	**88.7**	**88.7**	**94.3**	**85.4**	**92.7**	**91.1**	**92.6**	**84.9**	**84.4**	**86.7**	**88.9**
Probe CL #1-2	0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	Mean
LB-subGEI [12]	37.7	57.2	66.6	61.1	55.2	54.6	55.2	59.1	**58.9**	48.8	39.4	53.98
J-CNN [36]	46.1	58.4	64.4	64.2	55.5	50.5	54.7	55.8	53.3	**51.3**	39.9	54.01
GaitNet-pre [31]	42.1	58.2	65.1	70.7	68.0	70.6	65.3	**69.4**	51.5	50.1	36.6	58.9
GaitNet	**50.1**	**60.7**	**72.4**	**72.1**	**74.6**	**78.4**	**70.3**	68.2	53.5	44.1	**40.8**	**62.3**

TABLE 8: Recognition accuracy cross views under NM on CASIA-B dataset. One single GaitNet model is trained for all the viewing angles.

Methods	0°	18°	36°	54°	72°	108°	126°	144°	162°	180°	Average
CPM [35]	13	14	17	27	62	65	22	20	15	10	24.1
GEI-SVR [62]	16	22	35	63	95	95	65	38	20	13	42.0
CMCC [68]	18	24	41	66	96	95	68	41	21	13	43.9
ViDP [67]	8	12	45	80	**100**	**100**	81	50	15	8	45.4
STIP+NN [63]	-	-	-	-	84.0	86.4	-	-	-	-	-
LB [12]	18	36	67.5	93	99.5	99.5	92	66	36	18	56.9
L-CRF [35]	38	75	68	93	98	99	93	67	76	39	67.8
GaitNet-pre [31]	68	74	88	91	99	98	84	75	76	65	81.8
GaitNet	**82**	**83**	86	91	93	98	**92**	**90**	**79**	**79**	**87.3**

Recently, Chen *et al.* [12] propose new protocols to unify the training and testing where only one single model is trained for each protocol. Protocol 2 focuses on walking direction variations, where all videos used are in the NM subset. The training set includes videos of first 24 subjects in all viewing angles. The rest 100 subjects are for testing. The gallery is made of four videos at 90° view for each subject. The first two videos from remaining viewing angles are the probe. The Rank-1 recognition accuracies are reported in Tab. 8. GaitNet achieves the best average accuracy of 87.3% across 10 viewing angles, with significant improvement on extreme views compared to our preliminary work [31]. For example, at viewing angles of 0°, and 180°, the improvement margins are both 14%. This shows that more discriminative gait information, such as a canonical body shape information, under different views are learned in $\mathbf{f}_c$, which contributes to the final recognition accuracy.

Protocol 3 focuses on appearance variations. Training sets have videos under BG and CL. There are 34 subjects in total with 54° to 144° viewing angles. Different test sets are made with the different combination of viewing angles of the gallery and probe as well as the appearance condition (BG or CL). The results are presented in Tab. 9. Our preliminary work has comparable performance as the SOTA method L-CRF [35] on BG subset while significantly outperforming on CL subset. The proposed GaitNet outperforms on both subsets. 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.

TABLE 9: Comparison with [35] and [12] 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	GaitNet-pre [31]	JUCNet [23]	L-CRF [35]	LB [12]	RLTDA [69]
Subset				BG			
54	36	93.5	91.6	91.8	**93.8**	92.7	80.8
54	72	**94.1**	90.0	93.9	91.2	90.4	71.5
90	72	**98.6**	95.6	95.9	94.4	93.3	75.3
90	108	**99.3**	87.4	95.9	89.2	88.9	76.5
126	108	**99.5**	90.1	93.9	92.5	93.3	66.5
126	144	**90.0**	93.8	87.8	88.1	86.0	72.3
Mean		**95.8**	91.4	93.2	91.5	90.8	73.8
Subset				CL			
54	36	**97.5**	87.0	-	59.8	49.7	69.4
54	72	**98.6**	90.0	-	72.5	62.0	57.8
90	72	**99.3**	94.2	-	88.5	78.3	63.2
90	108	**99.6**	86.5	-	85.7	75.6	72.1
126	108	**98.3**	89.8	-	68.8	58.1	64.6
126	144	86.6	**91.2**	-	62.5	51.4	64.2
Mean		**96.7**	89.8	-	73.0	62.5	65.2

Across all evaluation protocols, GaitNet consistently outperforms the 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 appearance variations. Comparing with our preliminary work, the canonical feature $\mathbf{f}_c$ contains discriminative power which can further improve the recognition performance.

TABLE 10: 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.

Protocol	WS		BGHT		CL		MP		ALL	
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	-	-	-	-	-	-	-	-	-	$1-12$
TAR@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 [9]	9.4	19.5	6.1	12.5	5.7	13.2	6.3	16.7	5.8	16.1
GEINet [15]	15.5	35.2	11.8	24.7	6.5	16.7	17.3	35.2	13.0	29.2
DCNN [14]	11.0	23.6	5.7	12.7	7.0	15.9	8.1	20.9	7.9	19.0
LB [12]	53.4	73.1	23.1	50.3	23.2	38.5	56.1	74.3	40.7	61.6
GaitNet-pre [31]	91.8	96.6	74.2	85.1	56.8	72.0	92.3	97.0	81.2	87.8
GaitNet	**96.2**	**97.5**	**92.3**	**96.4**	**70.4**	**87.5**	**92.5**	**96.0**	**91.9**	**96.3**

5.2.2 USF

The original protocol of USF [10] and the methods [70]–[73] does not define a training set, which is not applicable to our method, as well as [12], that require data to train the models. Hence following the experiment setting in [12], which randomly partitions the dataset into the non-overlapping training and test sets, each with half of the subjects. We test on Probe A, defined in [12], where the probe is different from the gallery by the viewpoint. We achieve the identification accuracy of $99.7 \pm 0.2\%$, which is better than $99.5 \pm 0.2\%$ of our preliminary work GaitNet-pre [31], the reported $96.7 \pm 0.5\%$ of LB network [12], and $94.7 \pm 2.2\%$ of multi-task GAN [74].

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 SOTA methods on gait recognition [9], [12], [14], [15]. Furthermore, given the large amount of effort in human pose estimation [20], aggregating joint locations over time can be a good candidate for gait features. Therefore we define another baseline, named PE-LSTM, using pose estimation results as the input to the same LSTM and classification loss as ours. Using SOTA 2D pose estimation [75], we extract 14 joints' locations, feed to the 3-layer-LSTM, and train with our proposed LSTM incremental loss. For each of 5 baselines and our GaitNet, one model is trained with the 136-subject training set and tested on all 5 protocols.

As shown in Tab. 10, our method shows state-of-the-art performance compared with baselines, 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. Again, thanks to the discriminative canonical feature $\mathbf{f}_c$, GaitNet achieves better recognition accuracies than GaitNet-pre. Also, the superior performance of our GaitNet over PE-LSTM demonstrates that our feature $\mathbf{f}_p$ and $\mathbf{f}_c$ do explore more discriminate information than the joints' locations alone.

5.3 Comparison to Face Recognition

Face recognition aims to identify subjects by extracting discriminative identity features, or representation, from face images. Due to the vigorous development in the past few years, face recognition system is one of the most studied and deployed systems in the vision community, even superior to human on some tasks [76].

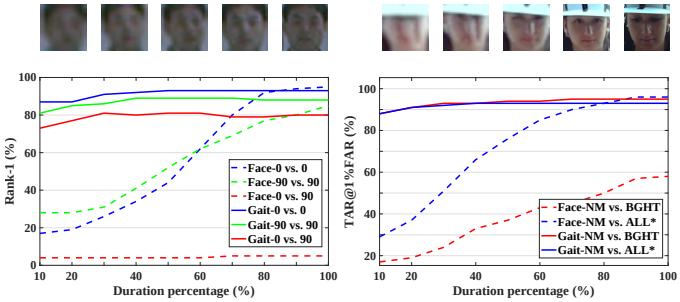

Fig. 12: Comparison of gait and face recognition on CASIA-B and FVG. Classification accuracy scores along with video duration percentage are calculated. (a) In CASIA-B, both gait and face recognition are performed in three scenarios: frontal-frontal (0° vs. 0°), side-side (90° vs. 90°) and frontal-side (0° vs. 90°). (b) In FVG, both recognitions use NM vs. BGHT and NM vs. ALL* protocols. Detected face examples are shown on the top of each frontal and side view plots under various video duration percentage.

However, the challenge is particularly prominent in the video surveillance scenario, where low-resolution and/or non-frontal faces are acquired at a distance. While gait, as a behavioral biometric compared to face, might have more advantages in those scenarios since the dynamic information can be more resistant even at a lower resolution and different viewing angles. Especially for GaitNet, $\mathbf{f}_{\text{sta-gait}}$ and $\mathbf{f}_{\text{dyn-gait}}$ can have complementary contributions in changing distances, resolutions and viewing angles. Therefore, to explore the advantages and disadvantages of gait recognition and face recognition in surveillance scenario, we compare our GaitNet with the most recent SOTA face recognition method, ArcFace [64], on the CASIA-B and FVG databases.

Specifically, for face recognition, we first employ SOTA face detection algorithm RetinaFace [77] to detect face and ArcFace to extract features for each frame of gallery and probe videos. Then the features over all frames of a video are aggregated by average pooling, an effective scheme used in prior video-based face recognition work [78]. We measure the similarity of features by their cosine distance. To keep consistency with above gait recognition experiments, both face and gait report TAR at 1% FAR for FVG and Rank-1 score for CASIA-B. To evaluate effects of time, we use the entire sequence as gallery and partial (*e.g.*, 10%) sequence as probe on 10 points on the time axis ranging from 10% to 100%.

Fig. 13: Examples in CASIA-B and FVG where the SOTA face recognizer ArcFace fails. The first row is the image of probe set; the second row is the recognized wrong person in gallery; and the third row shows the genuine gallery. The first three columns are three scenarios of CASIA-B and the last two columns are two protocols of FVG.

5.3.1 Gait vs. Face Recognition on CASIA-B

In this experiment, we select the videos of the NM as gallery and both CL and BG are probes. We compare gait and face recognition in three scenarios: frontal-frontal, side-side and side-frontal viewing angles. Fig. 12 shows the Rank-1 scores over the time duration. As the video begins, GaitNet is significantly superior to face in all scenarios since our $\mathbf{f}_{\text{sta-gait}}$ can capture discriminative information such as body shape in low-resolution images, as mentioned in Sec. 5.1.5, while faces are of too low resolution to perform meaningful recognition. As time progresses, GaitNet is stable to the resolution change and view variations, with increasing accuracy. In comparison, face recognition always has lower accuracies throughout the entire duration, except the frontal-frontal view face recognition slightly outperforms gait in the last 20% of the duration, which is expected as this is toward the ideal scenario for face recognition to shine. Unfortunately, for side-side or side-frontal views, face recognition continues to struggle even at the end of the duration.

5.3.2 Gait vs. Face Recognition on FVG

We further compare GaitNet with ArcFace on FVG with NM-BGHT and NM-ALL* protocols. Note that the videos of NM-BGHT contain variations in carrying bags and wearing hat. The videos of ALL*, different from ALL in Tab. 10, including all the variations in FVG except carrying and wearing hat variations (refer to Tab. 5 for details). As shown in Fig. 12, on the BGHT protocol, gait outperforms face in the entire duration, since wearing hat dramatically affects face recognition but not gait recognition. For ALL* protocol, face outperforms gait in the last 20% duration because by then low resolution is not an issue and FVG has frontal-view faces.

Figure 13 shows some examples in CASIB-B and FVG, which are incorrectly recognized by face recognition. We also show some images (video frames) for which our GaitNet fails to recognize in Fig. 14. The low resolution and illumination conditions in these videos are the main reasons for failure. Note that while video-based alignment [79], [80] or super-resolution approaches [81] might help to enhance the image quality, their impact to recognition is beyond the scope of this work.

Fig. 14: Failure cases of GaitNet on CASIB-B and FVG due to blurry and illumination conditions. The rows and columns are defined the same as Fig. 13.

TABLE 11: Runtime (ms per frame) comparison on FVG dataset.

Methods	Pre-processing	Inference	Total
PE-LSTM	224.4	0.1	224.5
GEINet [15]	89.5	1.5	91.0
DCNN [14]	89.5	1.7	91.2
LB [12]	89.5	1.3	90.8
GaitNet (ours)	89.5	1.0	90.5

5.4 Runtime Speed

System efficiency is an essential metric for many vision systems including gait recognition. We calculate the efficiency while each of the 5 gait recognition methods processing one video of FVG dataset on the same desktop with GeForce GTX 1080 Ti GPU. All the coding are implemented in PyTorch Framework of Python programming language. Parallel computing of batch processing is enabled for GPU on all the inference models, where batch size is number of samples in the probe. Alphapose and Mask-R-CNN takes batch size of 1 as input in inference. As shown in Tab. 11, our method is faster than the pose estimation method because of 1) an accurate, yet slow, version of AlphaPose [75] is required for model-based gait recognition method; 2) only low-resolution input of 32×64 pixels is needed for GaitNet. Further, our method has similar efficiency as the recent CNN-based gait recognition methods.

6 Conclusion

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 leverage 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 promise 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.

ACKNOWLEDGMENTS

This work was partially sponsored by the Ford-MSU Alliance program, and the Army Research Office under Grant Number W911NF-18-1-0330. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

REFERENCES

[1] M. S. Nixon, T. Tan, and R. Chellappa, *Human Identification Based on Gait*. Springer Science & Business Media, 2010.

[2] M. S. Seyfioğlu, S. Z. Gürbüz, A. M. Özbayoğlu, and M. Yüksel, "Deep learning of micro-doppler features for aided and unaided gait recognition," in *2017 IEEE Radar Conference (RadarConf)*. IEEE, 2017, pp. 1125–1130.

[3] C. Wan, L. Wang, and V. V. Phoha, "A survey on gait recognition," *ACM Computing Surveys (CSUR)*, vol. 51, no. 5, pp. 89:1–89:35, 2018.

[4] Y. Wang, B. Du, Y. Shen, K. Wu, G. Zhao, J. Sun, and H. Wen, "EV-Gait: Event-based robust gait recognition using dynamic vision sensors," in *Computer Vision and Pattern Recognition (CVPR)*, 2019.

[5] Q. Zou, L. Ni, Q. Wang, Q. Li, and S. Wang, "Robust gait recognition by integrating inertial and RGBD sensors," *IEEE Transactions on Cybernetics*, vol. 48, no. 4, pp. 1136–1150, 2017.

[6] Y. Zhang, G. Pan, K. Jia, M. Lu, Y. Wang, and Z. Wu, "Accelerometer-based gait recognition by sparse representation of signature points with clusters," *IEEE Transactions on Cybernetics*, vol. 45, no. 9, pp. 1864–1875, 2014.

[7] L. Middleton, A. A. Buss, A. Bazin, and M. S. Nixon, "A floor sensor system for gait recognition," in *Workshop on Automatic Identification Advanced Technologies (AutoID)*, 2005.

[8] W. Wang, A. X. Liu, and M. Shahzad, "Gait recognition using WiFi signals," in *Pervasive and Ubiquitous Computing (UbiComp)*, 2016.

[9] J. Han and B. Bhanu, "Individual Recognition Using Gait Energy Image," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 28, no. 2, pp. 316–322, 2005.

[10] S. Sarkar, P. J. Phillips, Z. Liu, I. R. Vega, P. Grother, and K. W. Bowyer, "The Human ID Gait Challenge Problem: Data Sets, Performance, and Analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 27, no. 2, pp. 162–177, 2005.

[11] K. Bashir, T. Xiang, and S. Gong, "Gait Recognition Using Gait Entropy Image," in *International Conference on Imaging for Crime Detection and Prevention (ICDP)*, 2010.

[12] Z. Wu, Y. Huang, L. Wang, X. Wang, and T. Tan, "A Comprehensive Study on Cross-View Gait Based Human Identification with Deep CNN," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 39, no. 2, pp. 209–226, 2016.

[13] M. A. Hossain, Y. Makihara, J. Wang, and Y. Yagi, "Clothing-invariant gait identification using part-based clothing categorization and adaptive weight control," *Pattern Recognition*, vol. 43, no. 6, pp. 2281–2291, 2010.

[14] M. Alotaibi and A. Mahmood, "Improved Gait recognition based on specialized deep convolutional neural networks," *Computer Vision and Image Understanding (CVIU)*, vol. 164, pp. 103–110, 2017.

[15] K. Shiraga, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi, "GEINet: View-Invariant Gait Recognition Using a Convolutional Neural Network," in *International Conference on Biometrics (ICB)*, 2016.

[16] S. Yu, R. Liao, W. An, H. Chen, E. B. G. Reyes, Y. Huang, and N. Poh, "GaitGanv2: Invariant gait feature extraction using generative adversarial networks," *Pattern Recognition*, vol. 87, pp. 179–189, 2019.

[17] T. T. Verlekar, P. L. Correia, and L. D. Soares, "View-invariant gait recognition system using a gait energy image decomposition method," *IET Biometrics*, vol. 6, no. 4, pp. 299–306, 2017.

[18] G. Ariyanto and M. S. Nixon, "Marionette mass-spring model for 3D gait biometrics," in *International Conference on Biometrics (ICB)*, 2012.

[19] S. Choi, J. Kim, W. Kim, and C. Kim, "Skeleton-based gait recognition via robust Frame-level matching," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 10, pp. 2577–2592, 2019.

[20] Y. Feng, Y. Li, and J. Luo, "Learning Effective Gait Features Using LSTM," in *International Conference on Pattern Recognition (ICPR)*, 2016.

[21] A. F. Bobick and A. Y. Johnson, "Gait Recognition Using Static, Activity-Specific Parameters," in *Computer Vision and Pattern Recognition (CVPR)*, 2001.

[22] D. Cunado, M. S. Nixon, and J. N. Carter, "Automatic extraction and description of human gait models for recognition purposes," *Computer Vision and Image Understanding (CVIU)*, vol. 90, no. 1, pp. 1–41, 2003.

[23] K. Zhang, W. Luo, L. Ma, W. Liu, and H. Li, "Learning joint gait representation via quintuplet loss minimization," in *Computer Vision and Pattern Recognition (CVPR)*, 2019.

[24] Y. Makihara, D. Adachi, C. Xu, and Y. Yagi, "Gait recognition by deformable registration," in *Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2018.

[25] D. Tao, X. Li, X. Wu, and S. J. Maybank, "General tensor discriminant analysis and gabor features for gait recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 29, no. 10, pp. 1700–1715, 2007.

[26] Y. Makihara, A. Suzuki, D. Muramatsu, X. Li, and Y. Yagi, "Joint Intensity and Spatial Metric Learning for Robust Gait Recognition," in *Computer Vision and Pattern Recognition (CVPR)*, 2017.

[27] S. Sivapalan, D. Chen, S. Denman, S. Sridharan, and C. Fookes, "Gait Energy Volumes and Frontal Gait Recognition using Depth Images," in *International Joint Conference on Biometrics (IJCB)*, 2011.

[28] P. Chattopadhyay, A. Roy, S. Sural, and J. Mukhopadhyay, "Pose Depth Volume extraction from RGB-D streams for frontal gait recognition," *Journal of Visual Communication and Image Representation*, vol. 25, no. 1, pp. 53–63, 2014.

[29] P. Chattopadhyay, S. Sural, and J. Mukherjee, "Frontal Gait Recognition From Incomplete Sequences Using RGB-D Camera," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 11, pp. 1843–1856, 2014.

[30] A. M. Nambiar, P. L. Correia, and L. D. Soares, "Frontal Gait Recognition Combining 2D and 3D Data," in *ACM Workshop on Multimedia and Security*, 2012.

[31] Z. Zhang, L. Tran, X. Yin, Y. Atoum, X. Liu, J. Wan, and N. Wang, "Gait Recognition via Disentangled Representation Learning," in *Computer Vision and Pattern Recognition (CVPR)*, 2019.

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

[33] H. Iwama, M. Okumura, Y. Makihara, and Y. Yagi, "The OU-ISIR Gait Database Comprising the Large Population Dataset and Performance Evaluation of Gait Recognition," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 5, pp. 1511–1521, 2012.

[34] Y. Makihara, H. Mannami, A. Tsuji, M. A. Hossain, K. Sugiura, A. Mori, and Y. Yagi, "The OU-ISIR Gait Database Comprising the Treadmill Dataset," *IPSJ Transactions on Computer Vision and Applications*, vol. 4, pp. 53–62, 2012.

[35] X. Chen, J. Weng, W. Lu, and J. Xu, "Multi-Gait Recognition Based on Attribute Discovery," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 40, no. 7, pp. 1697–1710, 2017.

[36] Y. Zhang, Y. Huang, L. Wang, and S. Yu, "A comprehensive study on gait biometrics using a joint CNN-based method," *Pattern Recognition*, vol. 93, pp. 228–236, 2019.

[37] J. D. Shutler, M. G. Grant, M. S. Nixon, and J. N. Carter, "On a Large Sequence-Based Human Gait Database," in *Applications and Science in Soft Computing*, 2004.

[38] M. Hofmann, J. Geiger, S. Bachmann, B. Schuller, and G. Rigoll, "The TUM Gait from Audio, Image and Depth (GAID) database: Multimodal recognition of subjects and traits," *Journal of Visual Communication and Image Representation*, vol. 25, no. 1, pp. 195–206, 2014.

[39] L. Tran and X. Liu, "Nonlinear 3D Face Morphable Model," in *Computer Vision and Pattern Recognition (CVPR)*, 2018.

[40] ——, "On learning 3D face morphable model from in-the-wild images," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2019, doi: 10.1109/tpami.2019.2927975.

[41] L. Tran, F. Liu, and X. Liu, "Towards High-fidelity Nonlinear 3D Face Morphable Model," in *Computer Vision and Pattern Recognition (CVPR)*, 2019.

[42] F. Liu, D. Zeng, Q. Zhao, and X. Liu, "Disentangling Features in 3D Face Shapes for Joint Face Reconstruction and Recognition," in *Computer Vision and Pattern Recognition (CVPR)*, 2018.

[43] E. Denton and B. Vighnesh, "Unsupervised Learning of Disentangled Representations from Video," in *Neural Information Processing Systems (NeurIPS)*, 2017.

[44] G. Balakrishnan, A. Zhao, A. V. Dalca, F. Durand, and J. Guttag, "Synthesizing Images of Humans in Unseen Poses," in *Computer Vision and Pattern Recognition (CVPR)*, 2018.
[45] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015.
[46] P. Esser, E. Sutter, and B. Ommer, "A Variational U-Net for Conditional Appearance and Shape Generation," in *Computer Vision and Pattern Recognition (CVPR)*, 2018.
[47] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.
[48] L. Tran, X. Yin, and X. Liu, "Disentangled Representation Learning GAN for Pose-Invariant Face Recognition," in *Computer Vision and Pattern Recognition (CVPR)*, 2017.
[49] ——, "Representation Learning by Rotating Your Faces," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2018, doi: 10.1109/tpami.2018.2868350.
[50] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Nets," in *Neural Information Processing Systems (NeurIPS)*, 2014.
[51] G. Shakhnarovich, L. Lee, and T. Darrell, "Integrated face and gait recognition from multiple views," in *Computer Vision and Pattern Recognition (CVPR)*, 2001.
[52] A. Kale, A. K. RoyChowdhury, and R. Chellappa, "Fusion of gait and face for human identification," in *Acoustics, Speech, and Signal Processing (ICASSP)*, 2004.
[53] X. Zhou and B. Bhanu, "Integrating face and gait for human recognition at a distance in video," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 37, no. 5, pp. 1119–1137, 2007.
[54] L. J. Connell, P. V. Ulrich, E. L. Brannon, M. Alexander, and A. B. Presley, "Body shape assessment scale: Instrument development foranalyzing female figures," *Clothing and Textiles Research Journal (CTRJ)*, vol. 24, no. 2, pp. 80–95, 2006.
[55] K. Cheung, S. Baker, and T. Kanade, "Shape-from-silhouette of articulated objects and its use for human body kinematics estimation and motion capture," in *Computer Vision and Pattern Recognition (CVPR)*, 2003.
[56] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *International Conference on Computer Vision (ICCV)*, 2017.
[57] G. Brazil, X. Yin, and X. Liu, "Illuminating Pedestrians via Simultaneous Detection and Segmentation," in *International Conference on Computer Vision (ICCV)*, 2017.
[58] G. Brazil and X. Liu, "Pedestrian Detection with Autoregressive Network Phases," in *Computer Vision and Pattern Recognition (CVPR)*, 2019.
[59] A. Radford, L. Metz, and S. Chintala, "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks," in *International Conference on Learning Representations (ICLR)*, 2016.
[60] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: continual prediction with LSTM," *Neural Computation*, vol. 12, no. 10, pp. 2451–2471, 2000.
[61] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *arXiv preprint arXiv:1412.6980*, 2014.
[62] W. Kusakunniran, Q. Wu, J. Zhang, and H. Li, "Support Vector Regression for Multi-View Gait Recognition based on Local Motion Feature Selection," in *Computer Vision and Pattern Recognition (CVPR)*, 2010.
[63] W. Kusakunniran, Q. Wu, J. Zhang, H. Li, and L. Wang, "Recognizing Gaits Across Views Through Correlated Motion Co-Clustering," *IEEE Transactions on Image Processing*, vol. 23, no. 2, pp. 696–709, 2013.
[64] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Computer Vision and Pattern Recognition (CVPR)*, 2019.
[65] L. v. d. Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.
[66] N. Srivastava, E. Mansimov, and R. Salakhudinov, "Unsupervised Learning of Video Representations using LSTMs," in *International Conference on Machine Learning (ICML)*, 2015.
[67] M. Hu, Y. Wang, Z. Zhang, J. J. Little, and D. Huang, "View-Invariant Discriminative Projection for Multi-View Gait-Based Human Identification," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 12, pp. 2034–2045, 2013.
[68] W. Kusakunniran, "Recognizing Gaits on Spatio-Temporal Feature Domain," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 9, pp. 1416–1423, 2014.
[69] H. 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*, vol. 23, no. 7, pp. 1274–1286, 2013.
[70] W. Chen, J. Zhang, L. Wang, J. Pu, and X. Yuan, "Human identification using temporal information preserving gait template," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 34, no. 11, pp. 2164–2176, 2011.
[71] D. Xu, S. Yan, D. Tao, S. Lin, and H.-J. Zhang, "Marginal fisher analysis and its variants for human gait recognition and content-based image retrieval," *IEEE Transactions on Image processing*, vol. 16, no. 11, pp. 2811–2821, 2007.
[72] Y. Guan, C.-T. Li, and F. Roli, "On reducing the effect of covariate factors in gait recognition: a classifier ensemble method," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 37, no. 7, pp. 1521–1528, 2014.
[73] H. Aggarwal and D. K. Vishwakarma, "Covariate conscious approach for gait recognition based upon Zernike moment invariants," *IEEE Transactions on Cognitive and Developmental Systems (TCDS)*, vol. 10, no. 2, pp. 397–407, 2017.
[74] Y. He, J. Zhang, H. Shan, and L. Wang, "Multi-Task GANs for View-Specific Feature Learning in Gait Recognition," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 1, pp. 102–113, 2018.
[75] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "RMPE: Regional Multi-Person Pose Estimation," in *International Conference on Computer Vision (ICCV)*, 2017.
[76] X. Yin, *Representation Learning and Image Synthesis for Deep Face Recognition*. Michigan State University, 2018.
[77] J. Deng, J. Guo, Z. Yuxiang, J. Yu, I. Kotsia, and S. Zafeiriou, "RetinaFace: Single-stage dense face localisation in the wild," in *arXiv preprint arXiv:1905.00641*, 2019.
[78] S. Gong, Y. Shi, and A. K. Jain, "Low quality video face recognition: Multi-mode aggregation recurrent network (MARN)," in *International Conference on Computer Vision Workshops (ICCVW)*, 2019.
[79] X. Liu, "Video-based face model fitting using adaptive active appearance model," *Image and Vision Computing*, vol. 28, no. 7, pp. 1162–1172, 2010.
[80] Y. Tai, Y. Liang, X. Liu, L. Duan, J. Li, C. Wang, F. Huang, and Y. Chen, "Towards highly accurate and stable face alignment for high-resolution videos," in *AAAI Conference on Artificial Intelligence (AAAI)*, vol. 33, 2019, pp. 8893–8900.
[81] Y. Chen, Y. Tai, X. Liu, C. Shen, and J. Yang, "FSRNet: End-to-end learning face super-resolution with facial priors," in *Computer Vision and Pattern Recognition (CVPR)*, 2018.

Ziyuan Zhang is now pursuing his B.S. in Computer Science from Michigan State University. His research areas of interest include deep learning and computer vision.

Luan Tran received his B.S. in Computer Science from Michigan State University with High Honors in 2015. He is now pursuing his Ph.D. also at Michigan State University in the area of deep learning and computer vision. His research areas of interest include deep learning and computer vision, in particular, face modeling and face recognition.

Feng Liu is currently a post-doc researcher in the Computer Vision Lab at Michigan State University. He received the Ph.D. degree in Computer Science from Sichuan University, China in 2018. His main research interests focus on computer vision and pattern recognition, specifically for 3D modeling, 2D and 3D face recognition.

Xiaoming Liu is a Professor at the Department of Computer Science and Engineering of Michigan State University. He received the Ph.D. degree in Electrical and Computer Engineering from Carnegie Mellon University in 2004. Before joining MSU in Fall 2012, he was a research scientist at General Electric (GE) Global Research. His research interests include computer vision, machine learning, and biometrics. As a co-author, he is a recipient of Best Industry Related Paper Award runner-up at ICPR 2014, Best Student Paper Award at WACV 2012 and 2014, and Best Poster Award at BMVC 2015. He has been the Area Chair for numerous conferences, including CVPR, ECCV, ICCV, NeurIPS, and ICLR. He is the Program Chair of WACV 2018, BTAS 2018, AVSS 2022, and General Chair of FG 2023. He is an Associate Editor of Pattern Recognition Letters, Pattern Recognition, and IEEE Transactions on Image Processing. He is an IAPR fellow. He has authored more than 150 scientific publications, and has filed 22 U.S. patents.