Improving Disentangled Representation Learning for Gait Recognition using Group Supervision

Lingxiang Yao, Worapan Kusakunniran,* *Senior Member, IEEE*, Peng Zhang, *Senior Member, IEEE*, Qiang Wu, *Senior Member, IEEE*, Jian Zhang, *Senior Member, IEEE*
*Corresponding Author: worapan.kun@mahidol.edu

***Abstract*—For decades, gait has been gathering extensive interest due to the advantage that it can be measured from a distance without physical contact. However, for image/video-based gait recognition, its performance can be remarkably influenced by exterior factors, such as viewing angles and clothing changes. Thus, in this paper, a group-supervised disentangled representation learning network is proposed for gait recognition to extract features invariant to these factors. First, sequences are explicitly disentangled into pose, gait, appearance, and view features through a generic encoder-decoder framework. To ensure feature adaptability and independency, a disentanglement swap module is specifically adopted during our encoder-decoder process through a series of swap operations based on the feature attributes. Following the feature disentanglement, a disentanglement aggregation module is also specially proposed for pose, gait, and appearance features to enhance their effectiveness. Finally, the enhanced three features are concatenated together for gait recognition. Relevant experiments certify that compared with other disentangled representation learning-based gait recognition methods, our proposed method enables a more excellent recognition result, despite fewer gait frames being utilized.**

***Index Terms*—Gait Recognition, Deep Learning, Disentangled Representation Learning**

I. Introduction

GAIT has many remarkable advantages over other kinds of biometric authentication [45]. First, each person presents his/her walking patterns in a sufficiently unique manner; thus, it is difficult to disguise other people's gaits. Second, gait works well in an unconstrained condition, given that it can be measured from a distance without proximal sensing or physical contact. Given these advantages, recognition using gait is more attractive than other biometrics for surveillance applications. In Denmark and the UK, gait analysis has already been utilized to collect evidence for convicting criminals [46], [48].

Over decades, many different methods have been developed for gait recognition [34]. However, most of them can only obtain a prominent recognition result when their probe/gallery gaits are in a similar condition [31]. It becomes more challenging as people's gaits are influenced by other factors and the used probe/gallery gaits become less similar. Examples of factors that can have this adverse influence on gait recognition are: bearing loads [36], [49], [56], clothing variations [1], [6], [13], walking modes [19], [29], [44], *etc*. There also exist lots of other factors that are connected with the external environments. Examples of these factors are: view changes [20], [25], [47], background, illumination, road surface materials and smoothness, *etc*.

The core of gait recognition is to extract gait-related features for each walking sequence, and the first challenge of extracting these features is to guarantee their invariableness to the adverse influences mentioned above, such as clothing, carrying, viewing angles, *etc*. [59]. Therefore, in this paper, a new method is raised for gait recognition by disentangling such invariant gait-related features from the appearances of each person. Motivated by the achievements of disentangled representation learning (DRL) in the computer vision community [26], [33], [38], [59], the proposed method intends to disentangle pose, gait, appearance, and view features for each person from their segmented binary silhouette sequences. Given that gait is closely related to human bodies and their movements, in our method, the final output invariant gait-related features are hybridized by three parts, namely, pose features, gait features, and appearance features.

More specifically, in our method, the feature disentanglement is learned using an encoder-decoder framework with a group of input sequences and a specifically raised disentanglement swap module. Motivated by [9], sequences sharing the same attribute values are randomly sampled and formed into the input groups. The encoder encodes each silhouette frame and explicitly splits its encoded feature representations into some meaningful parts. Meanwhile, since these split parts fully represent their encoded frame, with the decoder they can be decoded back to prototype. Additionally, given that sequences sharing the same attribute values are inclined to capture similar features for their shared attributes [9], it is rational for our disentanglement swap module to swap their corresponding split feature parts. For example, if there are two sequences caught under the same viewing angle, then their disentangled view features should be similar and thus naturally can be swapped. Meanwhile, even if the swapped view features are used, they can still be decoded back to their original inputs. By this means, we can enforce the disentanglement of the same attribute values to be similar, thereby achieving the consistency of gait features for the same person in different conditions.

In this paper, a disentanglement aggregation module is also specially developed to enhance the feature effectiveness. Basically, this enhancement can be separated into two different stages. First, it attempts to boost the discrimination capabilities of our disentangled gait and appearance features. Motivated by [37], our feature enhancement is conducted in a self-supervised manner across successive feature channels and among different sequence frames. Second, it aims to map the frame-based pose, gait, and appearance features into sequence-based features. For

pose features, it is significant to extract their temporal changes, since the disentangled pose features of a frame can only denote the pose of a specific instance, which may be similar to another instance of another different person [58]. For the enhanced gait and appearance features, although they can be relatively steady within each sequence, this mapping operation can help enforce the feature consistency, thereby increasing their effectiveness.

We summarize our contributions as follows.

- In this paper, a group-supervised DRL method is raised to tackle the gait recognition problem for the first time. First, input sequences are explicitly disentangled into pose, gait, appearance, and view features. After disentanglement, the combination of pose, gait, and appearance features is used for the final gait recognition.
- To enforce the feature efficiency and independency, in this paper, a disentanglement swap module is specially utilized during our encoder-decoder process. Moreover, to enhance the feature reliability and effectiveness, a disentanglement aggregation module is also specially utilized in this paper.
- Experiments using relevant datasets certify that our proposed method outperforms other DRL gait recognition methods. Moreover, these experiments also verify that our proposed method enables a remarkable performance with only a few gait frames being sampled.

The rest of this paper is organized as follows. Related work is reviewed in Section II. The proposed methods are presented in Section III. Experiment results are shown in Section IV, and conclusions are given in Section V.

II. Related Work

A. Gait Representation

For decades, a number of different representations have been raised for gait recognition. Generally, these gait representations can be categorized into two types: appearance-based or model-based representations [49], [58], [59].

Appearance-based representations are mostly extracted from human silhouettes, and then different subjects can be recognized by measurements that represent human shapes/movements [31]. One of the most widely used appearance-based representations is gait energy image (GEI), which is a silhouette template averaged over a full gait cycle [10]. Motion silhouette image (MSI) is another representation similar to GEI, in which each pixel is denoted as a descriptor of its movements in the temporal domain across all silhouettes that are part of a single gait cycle [24]. Appearance-based representations are widely utilized in gait recognition for their efficiency and simplicity. However, given their connection to human silhouettes, these representations are vulnerable to appearance changes caused by covariates, *e.g.*, clothing, *etc*.

Model-based representations are generally grasped based on the dynamics knowledge of human bodies. A gait model, made up of information about different body parts and how each part keeps relative with others, is first required, and then representations are extracted from this fit gait model. A major strength of these representations is that these gait models ensure that only image data corresponding to allowable human shapes and movements can be adopted for feature extraction, which reduces the effects of noise [31]. However, although these representations indicate a better robustness to appearance changes, their performance is highly dependent on the quality of gait models, and a relatively high resolution input image is always required for reliable pose estimation and gait model construction [58], [59].

B. Deep Learning-Based Gait Recognition

Many deep learning-based gait recognition methods have been recently proposed. According to their used input information, these methods can be basically divided into two categories, *i.e.*, template-based or sequence-based methods [4].

For template-based methods, a pre-process of extracting gait templates from sequential images or videos is first needed. One of the most widely used templates is the above-mentioned GEI [10]. Once gait templates are obtained, different deep learning-based networks can be adopted to extract the representations of gait, enhancing the characterization capabilities [4]. [35] raised GEINet using GEI as input. [53] fine-tuned the Siamese Neural Network for feature extraction. In [56], a View Transformation Generative Adversarial Network (VT-GAN) was developed for GEI to achieve transformation across two arbitrary views using a single generic model. In [55], an Identity-preserved Variation Normalizing Generative Adversarial Network (VN-GAN) was proposed to extract purely identity-related representations from GEI. For template-based methods, a major disadvantage is that they can lose the individual information of each frame, because in most cases, gait templates are attained by frame stacking and averaging. In addition, since only one or two gait templates can be learned from each sequence, it may also lead to the problem of insufficient input training data and over-fitting.

Sequence-based methods directly use successive gait frames as input. In [41], a 3D-CNN network was proposed to generate features in multiple views. In [8], heat maps were first captured as features of each frame, and then LSTM was used to translate the features of each frame into a feature of each entire sequence. In [2], features were first grasped for skeleton key-points and then attached to skeleton edges. LSTM was utilized to jointly model structured data and temporal information by finding long short-term dependencies from graph structure. Moreover, a 3D-CNN module was proposed in [27] to grasp spatial-temporal features from small and large temporal scales. Also, in [28], another 3D-CNN module was developed to gather global and local features in a principle manner. A major advantage of these methods is that they enable handling individual information of each frame. Meanwhile, more temporal information can be captured since specialized structures are adopted [4]. However, for these methods, large amounts of computation resources are required, which may limit their usage in real-world applications [4].

Lately, multiple 2D-CNN networks have also been proposed to approach gait recognition in the sequence-based manner. These networks assume that the appearance of a silhouette contains the position information; thus, the order information of a sequence is not required in gait recognition [4]. For example, in [4], each frame was first independently processed

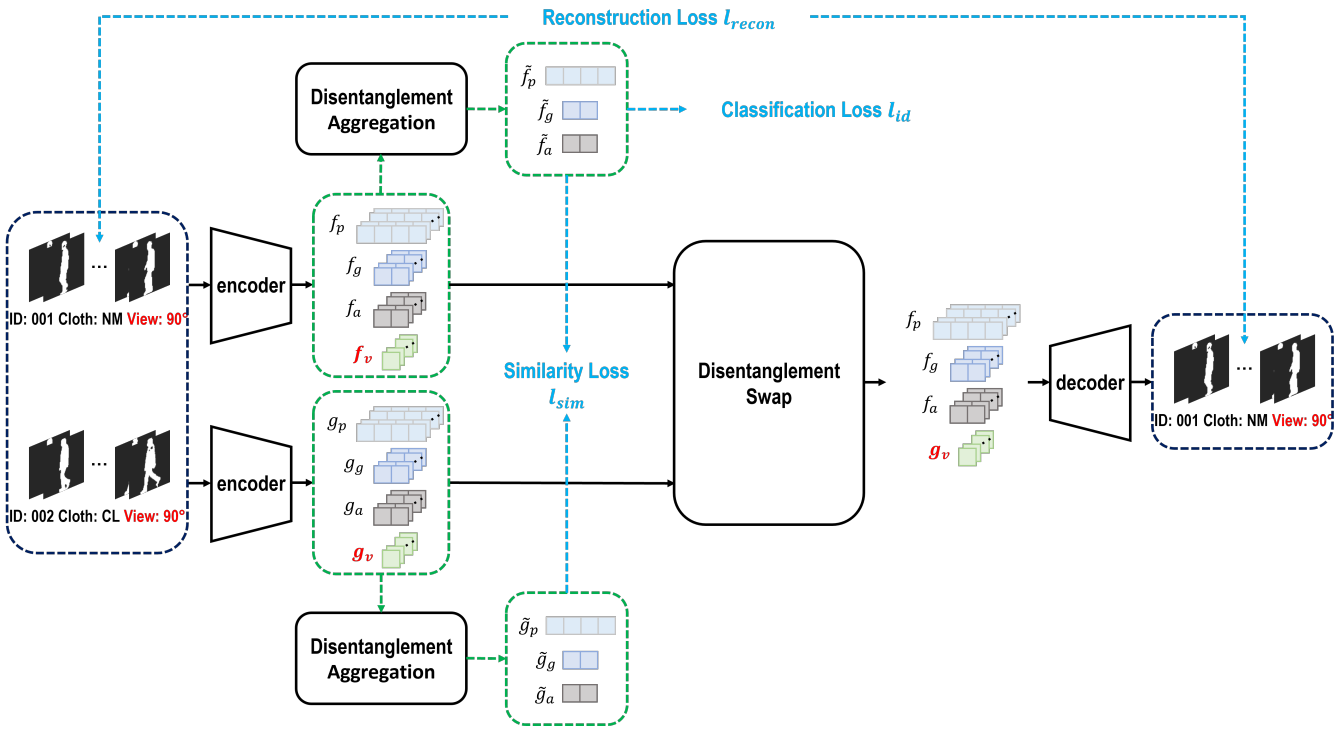


Fig. 1: Framework of the proposed network. To be more intuitive, in each group only two sequences are shown.

via a series of 2D convolution units with shared weight. After that, a global pooling module was used at the top to translate the information of each frame into a feature of each entire sequence. Compared with most gait recognition methods, a more excellent result has been obtained by [4] under its assumption.

C. Disentangled Representation Learning

Recently, data-driven DRL methods are garnering popularity in the computer vision community [26], [33], [38], [59]. DRL aims to learn features by decoupling the underlying structure of data into disjoint meaningful components [26]. To some extent, DRL helps interpret deep models and indicate which hidden features are actually learned in the model training processes [26].

In [59], an AutoEncoder structure was proposed to explicitly disentangle pose and appearance features from RGB frames. In [58], a further improved AutoEncoder structure was developed. Except for pose and appearance features, canonical features that provided the basic and unique representations of human bodies were also disentangled. In [26], GEI was first disentangled into identity and covariate features, and then these features were used to simultaneously recreate the input GEI and its canonical version with no covariates in a semi-supervised way. In [3], a covariate feature control gate was utilized to make up for the feature loss by introducing additional semantic labels. An attention module was also adopted to lead the identity and covariate parts to give attention to multiple spatial regions. Moreover, in [49], skeleton silhouette images were created and then disentangled into pose and canonical features. A view invariant layer was also utilized to reduce the influence of view changes.

In contrast to [3], [26], our proposed DRL network directly learns gait representations from randomly sampled silhouettes. Therefore, our learned representations should be categorized as silhouette-based representations, and the proposed method can also be deemed as a sequence-based method. The most relevant work to ours is [58], [59], and there also exist great differences across these three methods. First, [58], [59] learn features from successive RGB frames, while our method grasps features from randomly sampled binary silhouettes. In addition, [58], [59] leave the disentangled appearance features out of the final hybridized gait features, while in our method, these appearance features are also adopted after reducing their cloth-changing influence with our proposed attention module. Third, while [58], [59] use loss functions to keep the canonical consistency and pose similarity, in our method, the reliability and effectiveness of our disentangled features are preserved through swap operations in the decoder reconstruction process. Overall, compared with the above-mentioned DRL methods, our proposed method enables a more prominent performance on relevant datasets.

III. Proposed Methods

The main challenge of image/video-based gait recognition is to explore gait-related features from walking sequences, which are discriminative for each person and invariant to confounding factors, *e.g.*, viewing angles and clothing variations [59]. In our paper, we approach this challenge via feature disentanglement.

As shown in Fig. 1, the input to our DRL model is a group of sequences, and in each sequence, randomly sampled

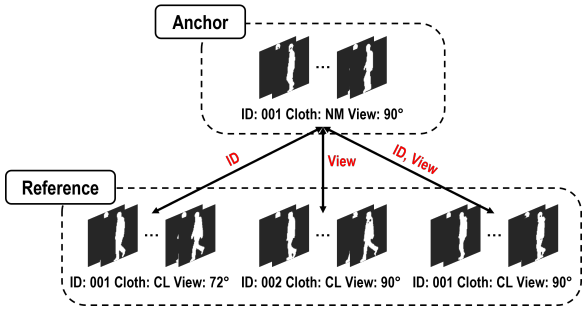


Fig. 2: Sample of a group input. Each edge links two sequences sharing at least one same attribute.

silhouettes are utilized. An encoder-decoder framework with a specifically proposed disentanglement swap module is used to generate the disentangled pose, gait, appearance, and view features for each sequence. Similar to canonical features [58], in our method, the disentangled gait features also reveal the unique and immanent information of gait, *e.g.*, stride. A disentanglement aggregation module is also utilized in our method to enhance the robustness of these disentangled gait and appearance features, followed by mapping these frame-based pose, gait, and appearance features into sequence-based features for identification purposes.

A. Group-Supervised Feature Disentanglement

1) Group-Supervised Learning:

Basically, each gait dataset consists of a variety of sequences $D = \{\chi_1, \chi_2, \cdots, \chi_n\}$, and each sequence is associated with a set of m attributes $D_a = \left\{(a^1_{\chi_i}, a^2_{\chi_i}, \cdots, a^m_{\chi_i})\right\}^n_{i=1}$. Thus, each attribute value can be a element of an attribute set $a^j \in A^j$. For example, in CASIA Gait Dataset B [60], A^1 can denote subject IDs $A^1 = \{001, 002, \cdots, 124\}$, A^2 can denote dressing modes $A^2 = \{\mathrm{NM}, \mathrm{BG}, \mathrm{CL}\}$, and then A^3 can refer to viewing angles $A^3 = \{0°, 18°, 36°, \cdots, 180°\}$.

As Fig. 1 indicates, the input to our DRL model is a group of sequences, and within each group, four sequences are included. Among these sequences, one sequence is deemed as the anchor sequence, and then the other sequences are deemed as the reference sequences. Between the anchor sequence and its each reference sequence, they share at least one same attribute. Fig. 2 presents a group input from CASIA Gait Dataset B [60].

2) Feature Disentanglement:

A universal encoder-decoder framework contains an encoder $\mathcal{E}$ and a decoder $\mathcal{D}$. This encoder encodes each frame and splits the encoded feature representation into some independent parts explicitly. Since these split parts can fully describe the encoded frame, they can be decoded back to the encoded frame with the decoder. Our network is developed based on the basic encoder-decoder framework, and each encoded feature representation is carefully split based on the anatomic studies of human gait and the attribute sets of gait datasets.

Taking CASIA Gait Dataset B [60] for instance, the encoded feature representations are split into pose, gait, appearance, and view features. Since silhouettes are randomly sampled in each sequence, pose features vary from one sequence to another. Gait features can remain relatively stable, since they depict the instinct gait information of each person. Appearance features are much more vulnerable to clothing changes, and such changes can generally have different effects on different human body parts [13]. View features are very sensitive to viewing angles, but distinct from appearance features, this influence is global. Thus, in our model, the feature disentanglement is divided into two stages.

Assuming each input sequence $\tilde{\chi}, \tilde{\chi} \subset \chi_i, i = 1, 2, \cdots, n$ is composed of N randomly sampled silhouettes, and the feature representations of each silhouette encoded by $\mathcal{E}$ can be denoted as a set $f = \left\{f^1, f^2, \cdots, f^N\right\}$. In our model, inspired by [40], for the $k-$th silhouette, the encoded feature representations f^k are first decomposed into view features f^k_v and view-disrelated features f^k_{dv} in an orthogonal manner,

$$f^k = f^k_v \cdot f^k_{dv} \tag{1}$$

where $f^k_v = \left\|f^k\right\|_2$ and $f^k_{dv} = \frac{f^k}{\|f^k\|_2}$, with $\|\|_2$ denoting the L_2 norm. After that, the view-disrelated features f^k_{dv} are explicitly split into three parts, namely, pose features f^k_p, gait features f^k_g, and appearance features f^k_a.

$$f^k_{dv} = \left[f^k_p, f^k_g, f^k_a\right] \tag{2}$$

Correspondingly, our latent representations used for decoder reconstruction are also developed in two stages. Pose, gait, and appearance features are first concatenated, and then multiplied with view features.

3) Disentanglement Swap:

Motivated by [9], in which two samples sharing the same attribute value (*e.g.*, both under the viewing angle of 90°) have identical latent values for this shared attribute (*e.g.*, view) although other attribute values (*e.g.*, pose, appearance) may vary, in our model, a disentanglement swap module is proposed to help enforce the feature consistency for each group through the shared semantic attributes. A more detailed legend is illustrated in Fig. 3 for the group sequences demonstrated in Fig. 2.

Given that sequences sharing the same attribute tend to share similar latent features for this shared attribute, in our model, the disentanglement swap module aims to swap attributes between sequences by swapping each corresponding split parts between each feature representations. Based on the shared attribute, our swap operations can be divided into three types.

(a) no-swap-op It does not contain any swap operations, and it is only fit for the anchor sequences. Its main purpose is to prove that our split feature parts can be assembled and decoded back to their original sequences (see Fig. 3(a)).

(b) one-swap-op It simply contains one swap operation, and it is used for the cases when only subject IDs or viewing angles are the same between each anchor and the reference sequences. Specifically, for these cases, only the split parts that share the same attribute values can be swapped. As Fig. 3(b-2) indicates, given that only the viewing angles are the same, only their corresponding view features can be swapped. Fig. 3(b-1) offers another swap sample when only the subject IDs are kept the same. It is worth noting that the swaps of pose features are avoided in our model, since the silhouettes of each sequence

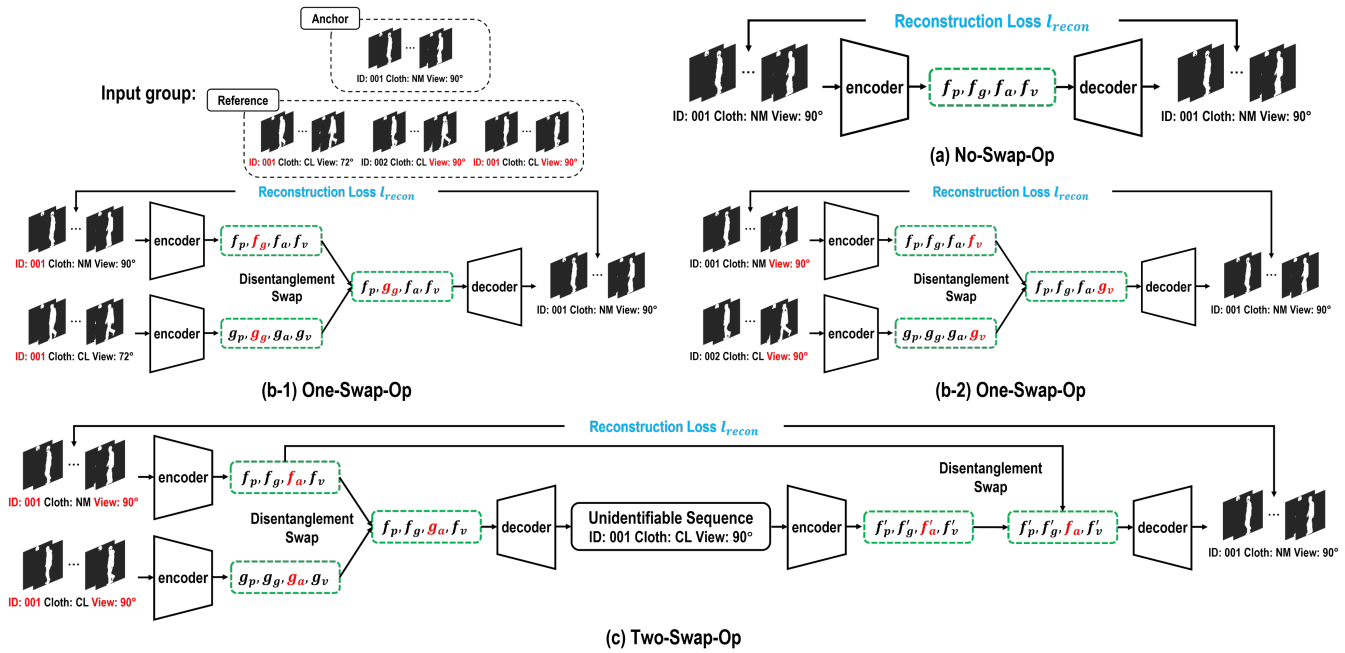


Fig. 3: Samples of disentanglement swap. (a): no-swap-op. (b-1, b-2): one-swap-op. (c): two-swap-op.

are randomly sampled; thus, the pose features are essentially different in spite of their attribute distributions.

(c) two-swap-op It consists of two swap operations, and it is fit for the cases when only dressing modes are varying between each anchor and their reference sequences. As Fig. 3(c) shows, the corresponding appearance features f_a, g_a are first swapped. After that, an unidentifiable sequence is generated based on the swapped features g_a. Following, this unidentifiable sequence is also encoded, and its corresponding appearance features f'_a are swapped using the former f_a. If, after the two swap operations, we are able to recover our original sequence, it implies that this attribute swap of dressing modes does not affect the split latent information from other attributes [9]. Distinct from subject IDs and viewing angles, dressing modes are an attribute that cannot be quantized, and usually we can only define a general concept of its values. For example, even if two persons are both dressed in long coats (CL), they still might be different from each other due to the coat length, thickness, *etc*. Therefore, in our model, a different swap operation is proposed for dressing modes.

4) Reconstruction Loss:

After each swap operation, the reconstructed sequence ought to be similar to its anchor sequence. In our model, an MSE loss is calculated after each reconstruction to enforce this similarity, and the reconstruction loss l_{recon} shown in Fig. 1 illustrates the sum of the four reconstruction losses shown in Fig. 3.

On the one hand, this reconstruction loss can ensure that our disentangled four features are fairly representative for each gait sequence. On the other hand, combined with our attribute swap operations, it also guarantees the availability and independency for our disentangled features.

B. Feature Learning and Aggregation

Motivated by [4], [37], [39], in our model, a disentanglement aggregation module is also specifically proposed for generating the finally used gait-related features. Basically, this module has two purposes: the first is to raise the discrimination capabilities of our disentangled gait and appearance features; the second is to transform frame-based pose, gait, and appearance features into sequence-based features. Eventually, these three sequence-based features are concatenated together for gait recognition.

1) Pose Feature Aggregation:

Learning the temporal changes of pose features is important, because for each frame, the disentangled pose features can only characterize the walking pose of a specific instance, which may share similarity with another instance of another person [58].

Given that the appearance of a silhouette reveals the position information [4], in our model, the temporal information of pose features is directly learned through a max-pooling operation.

$$\tilde{f}_p = \mathrm{maxpool}(f_p) \tag{3}$$

It is worth noticing that LSTM [8] has been discarded in this module for two main reasons. First, the output of LSTM can be varied over time steps, since it is significantly influenced by the last input [8]. Moreover, LSTM works better with longer input; however, its computation resources will also be increased. Here, we choose the max-pooling operation mainly because it is easy to implement but more stable and efficient with fewer resources being required.

2) Gait Feature Aggregation:

In our model, the disentangled gait features mainly represent the static and interior information of each human body; thus, for each person, their disentangled gait features are assumed

to stay unchanged over time, which also reveals that it is not necessary for our model to extract their temporal changes. Specifically, in order to generate more robust finally used gait-related features, our gait feature aggregation is divided into two stages.

In the first stage, motivated by [37], the feature effectiveness is improved by using the correlation within features and across frames. It is reasonable to use this underlying correlation, since gait can be deemed as a dynamically coupled pendulum model, and all human body parts are joined in a regular manner. In our model, two statistical functions are first assembled among each channel and their neighbors. After that, in order to extract more robust representations, the channel-wise attention is introduced to re-weight the disentangled gait features.

$$\begin{aligned} f_g^{a_1} &= [\mathrm{Avgpool1d}(f_g), \mathrm{Maxpool1d}(f_g)] \\ f_g^{logits_1} &= \mathrm{Conv1dNet}(f_g^{a_1}) \\ f_g^{ch} &= f_g \times \mathrm{Sigmoid}(f_g^{logits_1}) \end{aligned} \tag{4}$$

Further, similar to [37] communicating across different tokens, our model also adopts a communicating mechanism to enhance the feature robustness from different input frames. It is rational for our model to adopt the correlation of different frames, since within each sequence, our disentangled gait features can always be kept stable, thereby reducing the occurrences of outliers.

$$\begin{aligned} f_g^{a_2} &= \left[\mathrm{avepool}(f_g^{ch}), \mathrm{maxpool}(f_g^{ch})\right] \\ f_g^{logits_2} &= \mathrm{Conv1dNet}(f_g^{a_2}) \\ f_g^{fr} &= f_g^{ch} \times \mathrm{Sigmoid}(f_g^{logits_2}) \end{aligned} \tag{5}$$

Specifically, in this stage, our channel-wise operation aims at communication across different channels for each frame, while our frame-wise operation allows communication over different frames for each channel. In total, the two operations are bonded for each sequence to enable interaction across space and time.

In the second stage, a max-pooling operation is also adopted to map the frame-based improved gait features into a feature of the full sequence, similar to our aforementioned pose features.

$$\tilde{f}_g = \mathrm{maxpool}(f_g^{fr}) \tag{6}$$

Lastly, it is also worth noting that although our proposed gait feature aggregation module is similar to the networks proposed by [7], [37], significant differences still exist among these three methods. In [7], the attention mechanism is only adopted within short-range frames to capture the micro-motion patterns, while in our module, the attention mechanism is not only used among different frames but also utilized across feature channels. Also, our DRL network is proposed based on the assumption that the appearance of a silhouette contains its position information [4]; thus, in our network, all motion features are implicitly extracted. Moreover, different from [37], which repeatedly employs MLPs across spatial patches and feature channels, our module is built based on convolutions and attention mechanism.

3) Appearance Feature Aggregation:

In our model, the disentangled appearance features generally refer to the shape information of human bodies. Different from gait features that can keep stable all the time, these features can be significantly influenced by clothing variations. Hence, in this model, we aim to reduce this adverse cloth-changing influence.

Stimulated by [39] using a shift code to synthesize images in a certain direction, in our model, shift codes are also adopted to reduce the influence caused by clothing changes. Similar to the above-mentioned gait features, our feature effectiveness is also boosted within each features and among different frames. First, a method similar to gait feature aggregation is adopted to grasp the most salient parts f_a^{ch} for features. However, different from f_g^{ch}, which present the inherent gait information of each person, f_a^{ch} give more attention to the exterior dressing styles, and they should be excluded from our disentangled appearance features. Thus, our improved appearance features can be formulated as,

$$\tilde{f}_a^{ch} = f_a - \alpha_1 * f_a^{ch} \tag{7}$$

where α_1 is a learnable shift code.

The following feature improvement among frames is similar as we mentioned above with α_2 as another shift code. Also, a max-pooling operation is utilized to translate the frame-based appearance features into a sequence-based feature.

4) Similarity Loss and Classification Loss:

Our aggregated gait features basically extract the unique and inherent information of human gait; hence, for each person, their aggregated gait features are assumed to have similarities across different sequences. Similarly, for sequences within each group sharing the same IDs and views, after weakening their clothing effects in a self-supervised manner, our aggregated appearance features are also supposed to stay similar. Hence, to enforce the feature efficiency and consistency, a similarity loss l_{sim} is used in our model based on the L1 loss.

Furthermore, to enhance the effectiveness of our finally used gait-related features, a Batch All (BA_+) triplet loss [12] is also used in our model to serve as the classification loss l_{id}.

To this end, the overall training loss function is,

$$l = l_{id} + \lambda_s * l_{sim} + \lambda_r * l_{recon}. \tag{8}$$

IV. Experiments

In this section, we will verify the robustness of our proposed method on two broadly used gait datasets: CASIA Gait Dataset B [60] and OU-ISIR Treadmill Gait Dataset B [30].

Specifically, the training and testing details are first shown in Section IV-A. Then, a comparison of our proposed method and some other gait recognition methods has been organized on the aforementioned two datasets in Section IV-B and IV-C. Finally, ablation experiments are revealed in Section IV-D. Comparison results have certified the robustness of our proposed method on disentangling robust representations for gait recognition.

TABLE I: Averaged rank-1 accuracies (%) on CASIA-B following the protocols in [4], excluding identical-view cases.

Gallery NM #1-4 / Probe		Input Modality	0°-180°											
			0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	Mean
NM #5-6	LB [42]	Silhouettes	82.6	90.3	96.1	94.3	90.1	87.4	89.9	94.0	94.7	91.3	78.5	89.9
	Joint CNN [57]	GEI+Silhouettes	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 [59]	RGB frames	91.2	92.0	90.5	95.6	86.9	92.6	93.5	96.0	90.9	88.8	89.0	91.6
	GaitNet [58]	RGB frames	93.1	92.6	90.8	92.4	87.6	95.1	94.2	95.8	92.6	90.4	90.2	92.3
	SSDGait [49]	Skeleton silhouettes	85.7	82.1	82.6	83.5	85.8	89.7	88.1	84.0	84.7	85.7	87.1	85.4
	Proposed	Silhouettes	87.9	95.2	97.0	95.1	90.5	88.0	90.9	94.8	96.5	93.7	82.7	92.0
	Proposed	RGB frames	**97.0**	**98.0**	**98.2**	**99.1**	**99.4**	**97.3**	**95.8**	**98.5**	**98.0**	**97.8**	**97.8**	**97.9**
BG #1-2	LB [42]	Silhouettes	64.2	80.6	82.7	76.9	64.8	63.1	68.0	76.9	82.2	75.4	61.3	72.4
	Joint CNN [57]	GEI+Silhouettes	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 [59]	RGB frames	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 [58]	RGB frames	88.8	88.7	88.7	94.3	85.4	92.7	91.1	92.6	84.9	84.4	86.7	88.9
	SSDGait [49]	Skeleton silhouettes	75.2	77.6	76.9	78.2	81.1	81.1	80.9	79.3	76.5	74.3	70.2	77.4
	Proposed	Silhouettes	77.9	88.8	91.8	90.1	84.4	79.7	83.5	89.3	92.2	89.5	77.5	85.9
	Proposed	RGB frames	**96.3**	**97.8**	**97.8**	**97.8**	**96.9**	**93.4**	**92.4**	**97.6**	**95.8**	**95.0**	**89.8**	**95.5**
CL #1-2	LB [42]	Silhouettes	37.7	57.2	66.6	61.1	55.2	54.6	55.2	59.1	58.9	48.8	39.4	54.0
	Joint CNN [57]	GEI+Silhouettes	46.1	58.4	64.4	64.2	55.5	50.5	54.7	55.8	53.3	51.3	39.9	54.0
	GaitNet-pre [59]	RGB frames	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 [58]	RGB frames	50.1	60.7	72.4	72.1	74.6	**78.4**	70.3	68.2	53.5	44.1	40.8	62.3
	SSDGait [49]	Skeleton silhouettes	**64.7**	68.6	69.3	72.0	**76.6**	75.9	**78.5**	73.6	68.1	**70.2**	**64.5**	71.1
	Proposed	Silhouettes	60.9	**75.6**	**81.0**	**78.1**	72.6	67.8	73.0	**77.1**	**76.8**	70.0	53.3	**71.5**
	Proposed	RGB frames	67.3	66.9	59.3	57.2	48.8	36.6	32.6	35.7	36.6	36.1	31.8	46.2

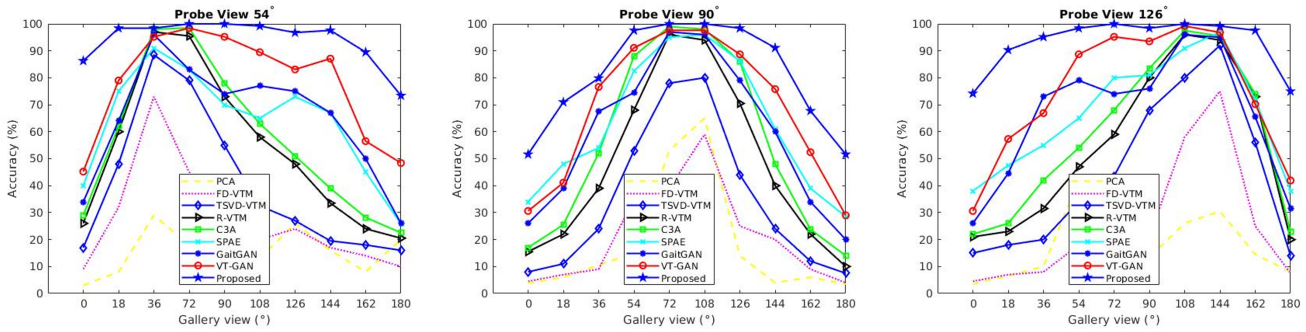


Fig. 4: Comparisons of the NM subset under the probe viewing angles of 54°, 90°, and 126° following the protocols in [52]. (Methods: PCA [10], FD-VTM [32], TSVD-VTM [21], R-VTM [20], C3A [43], SPAE [51], GaitGAN [50], VT-GAN [56])

A. Training and Testing Details

Stimulated by [11], our proposed network has an unbalanced encoder-decoder framework. Our encoder $\mathcal{E}$ is very similar to GaitSet [4], consisting of three convolution units followed by a Batch Normalization layer [16] and another Leaky ReLU layer. The decoder $\mathcal{D}$ has a basic structure as GaitNet [58], [59], built from three successive series of stride-2 transposed convolution, Batch Normalization [16], and Leaky ReLU layers. A Sigmoid activation is also utilized at its top to bring values back into the $[0, 1]$ range as our input sequences. Moreover, our disentangled pose, gait, appearance, and view features are all empirically set as 512, 256, 256, and 1 dimensional vectors, respectively.

In our training phase, each input is a group of four silhouette sequences, and each sequence is built by 10 randomly sampled silhouette frames in a size of 64×64. Each time, a batch in a size of 8×8 is randomly sampled for our training, which indicates that in each batch, the number of persons and the number of groups each person has in this batch are both 8. Adam [17] is adopted as our optimizer, and its learning rate is set as 0.0001. For l_{id}, the margin in BA_+ triplet loss [12] is set as 0.2. Furthermore, the λ_s and λ_r in Eq. 8 are set as 1 in all experiments.

In the testing phase, the batch size is set as 1, and the input is the entire silhouette sequence rather than a group of randomly sampled silhouette frames.

B. Comparison on CASIA Gait Dataset B

As the most widely used gait dataset, CASIA Gait Dataset B captures gait videos from 124 persons under 11 viewing angles ($0°, 18°, 36°, \ldots, 180°$) [60]. For a person under each viewing angle, 10 gait videos are captured, 6 in normal dressings (NM), 2 with a bag (BG), and 2 with a long coat (CL). In addition, in this dataset, the segmented gait silhouettes are also directly offered.

Many different experiment protocols have been proposed for CASIA Gait Dataset B [59], [60]. To present a fair comparison, we strictly abide by the protocols of each baseline method, and our comparison can be divided into three parts.

TABLE II: Rank-1 accuracy (%) across views under NM on CASIA-B following the protocols in [42].

Gallery NM #1-4 / Probe NM #5-6		Input Modality	90°										
			0°	18°	36°	54°	72°	108°	126°	144°	162°	180°	Mean
Methods	CPM [5]	RGB frames	13	14	17	27	62	65	22	20	15	10	24.1
	GEI-SVR [22]	Silhouettes	16	22	35	63	95	95	65	38	20	13	42.0
	CMCC [23]	Silhouettes	18	24	41	66	96	95	68	41	21	13	43.9
	ViDP [15]	Silhouettes	8	12	45	80	100	100	81	50	15	8	45.4
	STIP+NN [18]	RGB frames	-	-	-	-	84.0	86.4	-	-	-	-	-
	LB [42]	Silhouettes	18	36	67.5	93	99.5	99.5	92	66	36	18	56.9
	L-CRF [5]	RGB frames	38	75	68	93	98	99	93	67	76	39	67.8
	GaitNet-pre [59]	RGB frames	68	74	88	91	99	98	84	75	76	65	81.8
	GaitNet [58]	RGB frames	82	83	86	91	**93**	**98**	**92**	**90**	79	79	87.3
	Proposed	Silhouettes	65.9	77.2	86.9	87.6	79.5	78.8	85.0	88.1	82.0	65.9	79.2
	Proposed	RGB frames	**86.8**	**88.5**	**89.8**	**91.7**	92.3	88.8	91.7	88.6	**88.6**	**85.2**	**89.1**

TABLE III: Rank-1 accuracy (%) across views under BG and CL on CASIA-B following the protocols in [5].

Probe, Gallery (θ_p, θ_g) / Subset		Input Modality	54°, 36°		54°, 72°		90°, 72°		90°, 108°		126°, 108°		126°, 144°		Mean	
			BG	CL	BG	CL	BG	CL	BG	CL	BG	CL	BG	CL	BG	CL
Methods	RLTDA [14]	Silhouettes	80.8	69.4	71.5	57.8	75.3	63.2	76.5	72.1	66.5	64.6	72.3	64.2	73.8	65.2
	LB [42]	Silhouettes	92.7	49.7	90.4	62.0	93.3	78.3	88.9	75.6	93.3	58.1	86.0	51.4	90.8	62.5
	L-CRF [5]	RGB frames	93.8	59.8	91.2	72.5	94.4	88.5	89.2	85.7	92.5	68.8	88.1	62.5	91.5	73.0
	JUCNet [54]	Silhouettes	91.8	-	93.9	-	95.9	-	95.9	-	93.9	-	87.8	-	93.2	-
	GaitNet-pre [59]	RGB frames	91.6	87.0	90.0	90.0	95.6	94.2	87.4	86.5	90.1	89.8	93.8	**91.2**	91.4	89.8
	GaitNet [58]	RGB frames	93.5	**97.5**	94.1	**98.6**	98.6	**99.3**	**99.3**	**99.6**	**99.5**	**98.3**	90.0	86.6	95.8	**96.7**
	ICDNet [26]	Silhouettes	-	-	-	-	90.0	76.7	87.8	66.7	-	-	-	-	88.9	71.7
	Proposed	Silhouettes	**96.6**	90.5	**98.9**	88.8	**99.4**	96.1	98.9	98.3	98.3	94.4	**98.9**	87.2	**98.5**	95.3

In our first comparison, we follow the protocols proposed by [4]. The first 74 persons are utilized for training, and the remaining 50 persons are utilized for testing. In the testing set, the first 4 NM videos (NM #1-4) are taken as gallery, and the remaining 6 videos are separated into 3 probe subsets, *i.e.*, NM #5-6, BG #1-2, and CL #1-2 [4]. Table. I shows the comparison results of our proposed method and some other DRL gait recognition methods. Results given in this table are averaged on the 11 gallery views with all identical views excluded. Also, in Table. I, two results are given for our proposed method, and the main difference is their input modalities. We can see that our proposed method can attain the best result in the NM-NM and NM-BG cases with RGB frames being used. A comparable result has also been attained in these two cases when we handle silhouettes as input. Compared with silhouettes, RGB frames can afford richer information, thereby providing a higher possibility of extracting more discriminative gait-related features [58]. However, in the NM-CL case, the highest recognition accuracy is attained with silhouettes being utilized. One potential reason could be that in the NM-CL case, silhouettes are more relevant with the human shapes rather than the texture appearances that RGB frames are more concerned with. Also, it can be seen from Table. I that our method has suffered severe performance degradation when probe views are around the side view of $90°$. One reason is that gait features are different between the frontal views of $0°$-$90°$ and the back views of $90°$-$180°$. Another reason is that in our method, no view transformation is adopted, which can be improved before the final recognition as [49] did.

In our second comparison, we follow the protocols proposed by [52]. The first 62 persons are taken for training, and the next 62 persons are utilized for testing. Its gallery and probe sets are divided in the same manner as [4]. Due to limited space, a brief comparison is given in Fig. 4 for the NM subset from the probe viewing angles of $54°$, $90°$, and $126°$. From this figure, we can see that our proposed method has presented a more remarkable performance than most gait recognition methods, especially for the cases where probe and gallery view gaps are large.

In our third comparison, we abide by the protocols proposed by [42], which focus on walking direction variations [58], [59]. Therefore, in this comparison, only videos in the NM subset are used. The first 24 persons are sampled for training, and the remaining 100 persons are utilized for testing. In the testing set, the first 4 NM videos (NM #1-4) under the $90°$ viewing angle are used as gallery, and the other 2 NM videos (NM #5-6) under the remaining 10 viewing angles are tackled as probe. The comparison results of the proposed method with some other gait recognition methods are reported in Table. II. Our proposed method has attained the best average accuracy of 89.1% across 10 viewing angles, with significant improvement compared to [58], [59] using the same RGB frames as input. Moreover, compared with methods using silhouettes as input, our proposed method has also achieved the highest average accuracy, outperforming [42] by over 20%.

For our final comparison, we observe the protocols proposed in [5], which give more attention to appearance variations [58], [59]. The first 34 persons are utilized for training, and the other persons are chosen for testing. In the testing set, the NM videos are sampled as gallery, and the other videos are divided into the BG and CL probe subsets. 6 probe/gallery view pairs are tested between the normal viewing angles from

TABLE IV: Rank-1 accuracy (%) on OU-ISIR Treadmill-B.

Probe Set	Proposed	[46]	[1]	[6]
Type 0	**100.0**	99.7	94.0	**100.0**
Type 4	**100.0**	**100.0**	94.1	98.5
Type 8	**100.0**	**100.0**	94.5	94.1
Type A	**100.0**	**100.0**	91.6	91.2
Type C	**100.0**	**100.0**	94.5	94.1
Type E	**100.0**	**100.0**	91.5	91.2
Type G	**100.0**	99.8	89.1	98.5
Type I	**100.0**	**100.0**	98.5	98.5
Type J	**100.0**	**100.0**	91.5	91.2
Type L	**100.0**	**100.0**	90.0	**100.0**
Type N	**100.0**	**100.0**	85.5	**100.0**
Type P	**100.0**	**100.0**	91.1	**100.0**
Type R	**100.0**	**100.0**	86.2	88.2
Type T	**100.0**	**100.0**	95.0	94.1
Type U	**100.0**	**100.0**	95.5	94.1
Type V	**100.0**	**100.0**	91.6	91.2
Type X	**100.0**	**100.0**	90.1	**100.0**
Type Z	**100.0**	**100.0**	87.2	98.5

36° to 144°, and each time, one model is trained and tested for one probe/gallery view pair (θ_p, θ_g) [5]. Table. III offers the comparison results for our proposed method and some other gait recognition methods. We can see that the proposed method has attained the highest mean accuracy for the BG subset, outperforming GaitNet-pre [59] by 7.1% and exceeding GaitNet [58] by 2.7%. Additionally, for the CL subset, it also illustrates a performance comparable to GaitNet-pre [59] and GaitNet [58] with silhouettes being utilized.

For all evaluation protocols, our proposed method enables consistent achievement of a prominent gait recognition performance. In addition, in many cases, our proposed method has outperformed the state-of-the-art methods. To sum up, these comparisons have certified the effectiveness of our proposed method on extracting a robust disentangled representation for gait recognition under different variations.

C. Comparison on OU-ISIR Treadmill Gait Dataset B

To our knowledge, OU-ISIR Treadmill Gait Dataset B offers the maximum number of clothing variations [6]. Specifically, it gathers gait sequences from 68 subjects under 32 combinations of clothing types [1], [30], and each sequence is recorded twice on the same day. Given that it reveals a comprehensive analysis of the influence of clothing variations [1], it is recommended to verify the robustness of our proposed method on this dataset.

For this comparison, our training set is built by one sequence of each subject under each clothing combination, thereby, 2,176 sequences are sampled. 32 testing sets are formed by the left sequence of each subject based on their clothing combinations. It is worth mentioning that since view changes are not contained in OU-ISIR Treadmill Gait Dataset B, in our method view features are ignored, with only pose, gait, and appearance features being disentangled from each input sequence.

Table. IV offers the brief comparison results of our proposed method and some other gait recognition methods, and it can be seen that in all probe sets, our proposed method has attained the state-of-the-art recognition performance. A stronger robustness against clothing variations has been presented for our proposed method in this comparison. Once a clothing combination exists in the training set, it will surely be recognized in the evaluation process. To sum up, this comparison has verified the robustness of our disentangled gait-related features.

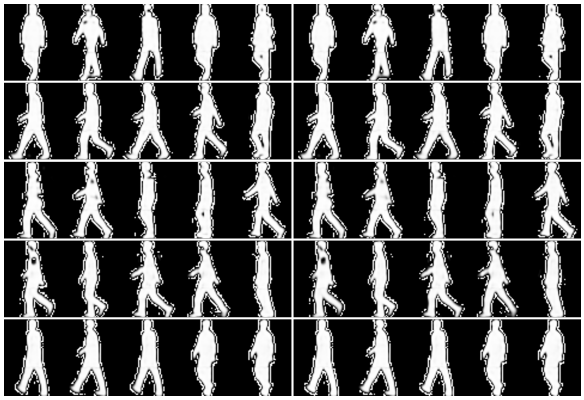


Fig. 5: Samples of the reconstructed frames with and without swapping view features.

D. Ablation Experiments on CASIA Gait Dataset B

In this part, we probe into the factors that can significantly influence the effectiveness of our disentangled features.

1) Influence of input frame numbers:

The first four lines of Table. V present the influence of different input frame numbers. We can see that our mean accuracy of the CL subset first rises with the increase of the frame number, and then it will stabilize at its best performance when more than 10 frames are sampled each time. In our method, each input frame is first independently disentangled with shared parameters, and a disentanglement aggregation module is followed to transform our frame-based disentangled features f_p, f_a, and f_g into three sequence-based features. With more frames being sampled, the robustness of these sequence-based features will also be raised. However, due to the limitation of the frame-based features, this improvement is not sustainable, and it will stabilize at its peak.

Also, compared with GaitSet [4] where 10 frames are used as its input (in this case the accuracy of GaitSet [4] is 65.1%), our proposed method has attained a more remarkable performance, which verifies the superiority of our proposed method.

2) Influence of input frame resolutions:

Experiments show that the resolution of the input frames has a huge impact on the disentangled features. As Table. V shows, there is a prominent increase if we improve the input resolution from 64×64 to 128×128. This is because a large frame enables more local details for gait recognition.

3) Influence of disentanglement swap:

The 7-th and 8-th lines of Table. V evaluate the effectiveness of our proposed disentanglement swap module. It can be found that a performance improvement has been attained through this disentanglement swap module, which verifies that it is efficient for our disentanglement swap module to enforce features of the same attributes to be similar and independent from others.

TABLE V: Studies of disentangled features on CASIA-B following the protocols in [4], excluding identical-view cases.

Row No.	Input Frame Number	Input Frame Resolution	Disentanglement Swap	Disentanglement Aggregation	Disentanglement Schemes	Disentanglement Dimensions (f_p, f_g, f_a, f_v)	Mean Accuracy(%) of the CL subset
1	5	64×64	$\surd$	$\surd$	(f_p, f_g, f_a, f_v)	$(512, 256, 256, 1)$	61.3
2	10						71.5
3	15						73.1
4	20						72.1
5	10	64×64	$\surd$	$\surd$	(f_p, f_g, f_a, f_v)	$(512, 256, 256, 1)$	71.5
6		128×128					73.6
7	10	64×64	$\times$*1	$\surd$	(f_p, f_g, f_a, f_v)	$(512, 256, 256, 1)$	69.6
8			$\surd$				71.5
9	10	64×64	$\surd$	$\times$*2	(f_p, f_g, f_a, f_v)	$(512, 256, 256, 1)$	66.5
10				$\surd$ w/o shift codes			71.1
11				$\surd$ w/ shift codes			71.5
12	10	64×64	$\surd$	$\surd$	(f_p, f_a)	$(512, 256, 256, 1)$	54.7
13					(f_p, f_g, f_a)		53.7
14					$(f_p, f_g, f_a, f_v^{dp})$*3		71.5
15					$(f_p, f_g, f_a, f_v^{sp})$*4		63.5
16	10	64×64	$\surd$	$\surd$	(f_p, f_g, f_a, f_v)	$(512, 256, 256, 1)$	71.5
17						$(256, 512, 256, 1)$	65.2
18						$(256, 256, 512, 1)$	66.2

*1 The final reconstruction loss l_{recon} only involves the reconstruction loss from the no-swap-op.
*2 The similarity loss l_{sim} is not involved in this feature disentanglement process.
*3 View features f_v^{dp} are first orthogonally decomposed from our encoded feature representations.
*4 View features f_v^{sp} are simply split from our encoded feature representations without decomposition.

Moreover, Fig. 5 shows some reconstructed sequences of the same person with and without view feature swapping. It can be seen that these reconstructed frames are similar with or without feature swapping, which strongly supports the disentanglement swap module we proposed in this paper.

4) Influence of disentanglement aggregation:

The robustness of the proposed disentanglement aggregation module has also been validated. A prominent increase has been attained in Table. V across the 9-th, 10-th, and 11-th lines using our proposed module with shift codes. As we stated above, this disentanglement aggregation module not only maps our frame-based disentangled features into sequence-based features, but more significantly also enhances their discrimination capabilities among feature channels and sequence frames. Furthermore, the similarity loss is followed to enforce our aggregated features to stay more consistent, which further enhances their efficiency.

5) Influence of disentanglement schemes:

In Table. V, the 12-th, 13-th, 14-th, and 15-th lines indicate a comparison of four different disentanglement schemes. We can see that it is more effective to first orthogonally decompose our encoded representations into view features and view-disrelated features, and then explicitly split these view-disrelated features into pose, gait, and appearance features. Different from bearing loads and clothing variations, view changes tend to have a huge influence on whole human bodies. Thus, it is rational for us to decompose view features from other disentangled features.

6) Influence of disentangled feature dimensions:

The last three lines of Table. V validate the influence of different feature dimensions. The best performance is obtained when the dimensions of our disentangled features f_p, f_g, f_a, f_v are set as 512, 256, 256, and 1, respectively. To some extent, it also shows the importance of pose in gait recognition.

V. Conclusion

Gait recognition is remarkably influenced by factors, such as viewing angles and clothing changes. Therefore, in this paper, a group-supervised DRL method is proposed for gait recognition to grasp features invariant to these factors. First, each sequence is explicitly disentangled into pose, gait, appearance, and view features through an encoder-decoder framework. To ensure the feature adaptability and independency, a disentanglement swap module is specially utilized during our encoder-decoder process with a series of swap operations based on the feature attributes. Moreover, to enhance the feature practicality and effectiveness, a disentanglement aggregation module is also specially utilized after our feature disentanglement. Finally, the aggregated pose, gait, and appearance features are concatenated together for gait recognition. Experiments using relevant datasets have verified that this proposed method can achieve a more prominent result than other DRL gait recognition methods.

References

[1] R. Anusha and C. Jaidhar, "Clothing invariant human gait recognition using modified local optimal oriented pattern binary descriptor," *Multimedia Tools and Applications*, vol. 79, pp. 2873–2896, 2019.

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

[3] T. Chai, X. Mei, A. Li, and Y. Wang, "Semantically-guided disentangled representation for robust gait recognition," *2021 IEEE International Conference on Multimedia and Expo*, 2021.

[4] H. Chao, Y. He, J. Zhang, and J. Feng, "Gaitset: Regarding gait as a set for cross-view gait recognition," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 8126–8133.

[5] X. Chen, J. Weng, W. Lu, and J. Xu, "Multi-gait recognition based on attribute discovery," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, pp. 1697–1710, 2018.

[6] M. Deng and C. Wang, "Gait recognition under different clothing conditions via deterministic learning," *IEEE/CAA Journal of Automatica Sinica*, pp. 1–10, 2018.

[7] C. Fan, Y. Peng, C. Cao, X. Liu, S. Hou, J. Chi, Y. Huang, Q. Li, and Z. He, "Gaitpart: Temporal part-based model for gait recognition," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14 213–14 221, 2020.

[8] Y. Feng, Y. Li, and J. Luo, "Learning effective gait features using lstm," *2016 23rd International Conference on Pattern Recognition*, pp. 325–330, 2016.

[9] Y. Ge, S. Abu-El-Haija, G. Xin, and L. Itti, "Zero-shot synthesis with group-supervised learning," *ArXiv*, vol. abs/2009.06586, 2021.

[10] J. Han and B. Bhanu, "Individual recognition using gait energy image," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, pp. 316–322, 2006.

[11] K. He, X. Chen, S. Xie, Y. Li, P. Doll'ar, and R. B. Girshick, "Masked autoencoders are scalable vision learners," *ArXiv*, vol. abs/2111.06377, 2021.

[12] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *ArXiv*, vol. abs/1703.07737, 2017.

[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, pp. 2281–2291, 2010.

[14] 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, pp. 1274–1286, 2013.

[15] M. Hu, Y. Wang, Z. Zhang, 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, pp. 2034–2045, 2013.

[16] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*. PMLR, 2015, pp. 448–456.

[17] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2015.

[18] W. Kusakunniran, "Recognizing gaits on spatio-temporal feature domain," *IEEE Transactions on Information Forensics and Security*, vol. 9, pp. 1416–1423, 2014.

[19] W. Kusakunniran, Q. Wu, J. Zhang, and H. Li, "Gait recognition across various walking speeds using higher order shape configuration based on a differential composition model," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 42, pp. 1654–1668, 2012.

[20] ——, "Gait recognition under various viewing angles based on correlated motion regression," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, pp. 966–980, 2012.

[21] W. Kusakunniran, Q. Wu, H. Li, and J. Zhang, "Multiple views gait recognition using view transformation model based on optimized gait energy image," *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*, pp. 1058–1064, 2009.

[22] W. Kusakunniran, Q. Wu, J. Zhang, and H. Li, "Support vector regression for multi-view gait recognition based on local motion feature selection," *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 974–981, 2010.

[23] 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, pp. 696–709, 2014.

[24] T. H. Lam and R. Lee, "Human identification by using the motion and static characteristic of gait," *18th International Conference on Pattern Recognition*, vol. 3, pp. 996–999, 2006.

[25] S. Li, W. Liu, and H. Ma, "Attentive spatial–temporal summary networks for feature learning in irregular gait recognition," *IEEE Transactions on Multimedia*, vol. 21, pp. 2361–2375, 2019.

[26] X. Li, Y. Makihara, C. Xu, Y. Yagi, and M. Ren, "Gait recognition via semi-supervised disentangled representation learning to identity and covariate features," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13 306–13 316, 2020.

[27] B. Lin, S. Zhang, and F. Bao, "Gait recognition with multiple-temporal-scale 3d convolutional neural network," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020.

[28] B. Lin, S. Zhang, and X. Yu, "Gait recognition via effective global-local feature representation and local temporal aggregation." in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 14 648–14 656.

[29] Y. Makihara, D. Adachi, C. Xu, and Y. Yagi, "Gait recognition by deformable registration," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 674–67 410, 2018.

[30] 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.

[31] Y. Makihara, D. S. Matovski, M. S. Nixon, J. N. Carter, and Y. Yagi, "Gait recognition: Databases, representations, and applications," *Computer Vision: A Reference Guide*, pp. 1–13, 2020.

[32] Y. Makihara, R. Sagawa, Y. Mukaigawa, T. Echigo, and Y. Yagi, "Gait recognition using a view transformation model in the frequency domain," in *European Conference on Computer Vision*, 2006.

[33] X. Peng, X. Yu, K. Sohn, D. N. Metaxas, and M. Chandraker, "Reconstruction-based disentanglement for pose-invariant face recognition," *2017 IEEE International Conference on Computer Vision*, pp. 1632–1641, 2017.

[34] A. Sepas-Moghaddam and A. Etemad, "Deep gait recognition: A survey," *ArXiv*, vol. abs/2102.09546, 2021.

[35] K. Shiraga, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi, "Geinet: View-invariant gait recognition using a convolutional neural network," *2016 International Conference on Biometrics*, pp. 1–8, 2016.

[36] S. Singh and K. K. Biswas, "Biometric gait recognition with carrying and clothing variants," in *International Conference on Pattern Recognition and Machine Intelligence*. Springer, 2009, pp. 446–451.

[37] I. O. Tolstikhin, N. Houlsby, A. Kolesnikov, L. Beyer, X. Zhai, T. Unterthiner, J. Yung, D. Keysers, J. Uszkoreit, M. Lucic, and A. Dosovitskiy, "Mlp-mixer: An all-mlp architecture for vision," *ArXiv*, vol. abs/2105.01601, 2021.

[38] L. Tran, X. Yin, and X. Liu, "Disentangled representation learning gan for pose-invariant face recognition," *2017 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1283–1292, 2017.

[39] A. Voynov and A. Babenko, "Unsupervised discovery of interpretable directions in the gan latent space," in *International Conference on Machine Learning*, 2020, pp. 9786–9796.

[40] Y. Wang, D. Gong, Z. Zhou, X. Ji, H. Wang, Z. Li, W. Liu, and T. Zhang, "Orthogonal deep features decomposition for age-invariant face recognition," in *European Conference on Computer Vision*, 2018.

[41] T. Wolf, M. Babaee, and G. Rigoll, "Multi-view gait recognition using 3d convolutional neural networks," *2016 IEEE International Conference on Image Processing*, pp. 4165–4169, 2016.

[42] Z. Wu, Y. Huang, L. Wang, X. Wang, and T. Tan, "A comprehensive study on cross-view gait based human identification with deep cnns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, pp. 209–226, 2017.

[43] X. Xing, K. jun Wang, T. Yan, and Z. Lv, "Complete canonical correlation analysis with application to multi-view gait recognition," *Pattern Recognition*, vol. 50, pp. 107–117, 2016.

[44] C. Xu, Y. Makihara, X. Li, Y. Yagi, and J. Lu, "Speed invariance vs. stability: Cross-speed gait recognition using single-support gait energy image," in *Asian Conference on Computer Vision*, 2016.

[45] K. Xu, X. Jiang, and T. Sun, "Gait recognition based on local graphical skeleton descriptor with pairwise similarity network," *IEEE Transactions on Multimedia*, 2021.

[46] L. Yao, W. Kusakunniran, Q. Wu, J. Zhang, and J. Xu, "Part-based collaborative spatio-temporal feature learning for cloth-changing gait recognition," *2020 25th International Conference on Pattern Recognition*, pp. 2057–2064, 2021.

[47] L. Yao, W. Kusakunniran, Q. Wu, J. Zhang, Z. Tang, and W. kou Yang, "Robust gait recognition using hybrid descriptors based on skeleton gait energy image," *Pattern Recognition Letters*, 2019.

[48] M. Ye, C. Yang, V. Stanković, L. Stanković, and S. Cheng, "Distinct feature extraction for video-based gait phase classification," *IEEE Transactions on Multimedia*, vol. 22, pp. 1113–1125, 2020.

[49] J. Yoo and K. Park, "Skeleton silhouette based disentangled feature extraction network for invariant gait recognition," *2021 International Conference on Information Networking*, pp. 687–692, 2021.

[50] S. Yu, H. Chen, E. B. G. Reyes, and N. Poh, "Gaitgan: Invariant gait feature extraction using generative adversarial networks," *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 532–539, 2017.

[51] S. Yu, H. Chen, Q. Wang, L. Shen, and Y. Huang, "Invariant feature extraction for gait recognition using only one uniform model," *Neurocomputing*, vol. 239, pp. 81–93, 2017.

[52] 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.

[53] C. Zhang, W. Liu, H. Ma, and H. Fu, "Siamese neural network based gait recognition for human identification," *2016 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 2832–2836, 2016.
[54] K. Zhang, W. Luo, L. Ma, W. Liu, and H. Li, "Learning joint gait representation via quintuplet loss minimization," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4695–4704, 2019.
[55] P. Zhang, Q. Wu, and J. Xu, "Vn-gan: Identity-preserved variation normalizing gan for gait recognition," *2019 International Joint Conference on Neural Networks*, pp. 1–8, 2019.
[56] ——, "Vt-gan: View transformation gan for gait recognition across views," *2019 International Joint Conference on Neural Networks*, pp. 1–8, 2019.
[57] 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.
[58] Z. Zhang, L. Tran, F. Liu, and X. Liu, "On learning disentangled representations for gait recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. PP, 2020.
[59] Z. Zhang, L. Tran, X. Yin, Y. Atoum, X. Liu, J. Wan, and N. Wang, "Gait recognition via disentangled representation learning," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4705–4714, 2019.
[60] S. Zheng, J. Zhang, K. Huang, R. He, and T. Tan, "Robust view transformation model for gait recognition," *2011 18th IEEE International Conference on Image Processing*, pp. 2073–2076, 2011.

Lingxiang Yao received the B.S. degree from the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China. He is currently a Ph.D. student with the School of Electrical and Data Engineering, University of Technology Sydney, NSW, Australia. His research interests include gait recognition, person reidentification and deep learning.

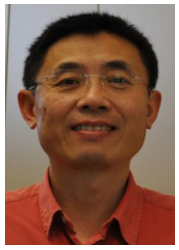
Worapan Kusakunniran received the B.Eng. degree in computer engineering from the University of New South Wales (UNSW), Australia in 2008, and the Ph.D. degree in computer science and engineering from UNSW, in cooperation with the Neville Roach Laboratory, National ICT Australia, Australia in 2013. He is currently a lecturer with the Faculty of Information and Communication Technology, Mahidol University, Thailand.

He is the author of several papers in top international conferences and journals. He served as a program committee member for many international conferences and workshops. Also, he has served as a reviewer for several international conferences and journals, such as ICPR, ICIP, PR, TIP, and TIFS. He was a recipient of the ICPR Best Biometric Student Paper Award in 2010, and also a winner of several national and international innovation contests. His research interests include biometrics, pattern recognition, medical image processing, computer vision, multimedia, and machine learning.

Peng Zhang obtained his PhD degree from University of Technology Sydney in 2020. He is currently working as a lecturer in Shandong University of Science and Technology. He has published more than 20 research papers in major journals and conferences, such as IEEE T-IP, IEEE T-CSVT, PR, WACV, *etc*. His research interests include metric learning, deep learning and their applications on person reidentification, gait recognition, *etc*.

Qiang Wu received the B.Eng. and M.Eng. degrees from the Harbin Institute of Technology, Harbin, China, in 1996 and 1998, respectively, and the Ph.D. degree from the University of Technology Sydney, Australia, in 2004.

He is currently an Associate Professor and a Core Member of the Global Big Data Technologies Centre, University of Technology Sydney. His research interests include computer vision, image processing, pattern recognition, machine learning, and multimedia processing. His research outcomes are applied span over fields such as video security surveillance, biometrics, video data analysis, and humancomputer interaction. His research outcomes have been published in many premier international conferences, including ECCV, CVPR, ICIP, and ICPR, and the major international journals, such as IEEE TIP, IEEE TSMC-B, IEEE TCSVT, IEEE TIFS, PR, PRL, and Signal Processing.



Jian Zhang (SM'04) received the B.S. degree in electronics from East China Normal University, China, the M.S. degree in computer science from Flinders University, Australia, and the Ph.D. degree in electrical engineering from the University of New South Wales (UNSW), Australia. From 2004 to 2011, he was a Principal Researcher and a Project Leader with Data61, Australia, and a Conjoint Associate Professor with the School of Computer Science and Engineering, UNSW. He is currently an Associate Professor with the Global Big Data Technologies Centre, University of Technology Sydney, Australia. He has authored or co-authored over 140 paper publications, book chapters, and six issued U.S. and Chinese patents. His current interests include social multimedia signal processing, large-scale image and video content analytics, retrieval and mining, 3D-based computer vision, and intelligent video surveillance systems.

Dr. Zhang was an Associate Editor of the IEEE Transactions on Circuits and Systems for Video Technology from 2006 to 2015. He has been an Associate Editor of the IEEE Transactions on Multimedia and the EURASIP Journal on Image and Video Processing since 2016.