This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2021.3112564, IEEE Transactions on Circuits and Systems for Video Technology

Collaborative Feature Learning for Gait Recognition under Cloth Changes

Lingxiang Yao, Worapan Kusakunniran, *Senior Member, IEEE*, Qiang Wu, *Senior Member, IEEE*, Jingsong Xu, *Senior Member, IEEE*, Jian Zhang, *Senior Member, IEEE*

***Abstract*—Since gait can be utilized to identify individuals from a far distance without their interaction and coordination, recently many gait recognition methods have been proposed. However, due to a real-world scenario of clothing changes, a degradation occurs for most of these methods. Thus in this paper, a more efficient gait recognition method is proposed to address the problem of clothing variances. First, part-based gait features are formulated from two different perspectives, *i.e.*, the separated body parts that are more robust to clothing changes and the estimated human skeleton key-point regions. It is reasonable to formulate such features for cloth-changing gait recognition, because these two perspectives are both less vulnerable to clothing changes. Given that each feature has its own advantages and disadvantages, a more efficient gait feature is generated in this paper by assembling these two features together. Moreover, since local features are more discriminative than global features, in this paper more attention is focused on the local short-range features. Also, unlike most methods, in our method we treat the estimated key-point features as a set of word embeddings, and a transformer encoder is specifically used to learn the dependence of each correlative key-points. The robustness and effectiveness of our proposed method are certified by experiments on CASIA Gait Dataset B, and it has achieved the state-of-the-art performance on this dataset.**

***Index Terms*—Gait recognition, Cloth changes, Deep Learning**

I. Introduction

DIFFERENT from most biometric features, gait provides a noninvasive manner for identifying a person at a distance without their coordination, which allows gait to be widely used in security surveillance and forensic identification. In Denmark and UK, gait analysis already has been used to collect evidence in criminal cases for convicting suspects [14], [46], [66].

It is feasible to recognize a person by gait, since each person presents his/her walking pattern in a fairly unique manner [64]. However, in real-world applications, the result of image/video-based gait recognition is influenced by several external factors. Among these factors, clothing variations can be deemed as one of the most challenging factors to gait recognition [2], [4]. One reason that the problem of clothing changes is a huge challenge to gait recognition is that people can walk with any casual style of dressing in their daily lives, *e.g.*, shirts, jackets, hoodies, *etc*, which significantly alters their walking appearances and affects the available visual features to be used in the future recognition procedure [40], [76]. Fig. 1 shows 6 samples from CASIA Gait Dataset B in three different dressing styles [75]. It can be found that people's walking appearances are remarkably changed due to the changing dressing styles, especially for the upper bodies. Besides, a performance comparison is also indicated in Table. I for recently proposed gait recognition methods. These methods all have attained an excellent result in the case without clothing changes. However, once clothing change occurs, these methods all suffer performance degradation. For example, although [33] has achieved the best performance in the NM case, its accuracy in the CL case only reaches 81.5%, reduced by over 20%. Thus compared with other factors, clothing changes can be treated as one of the most challenging factors for gait recognition.

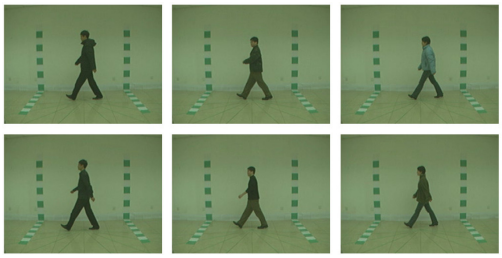

Fig. 1: Samples from CASIA Gait Dataset B.

For decades, different methods have been proposed to tackle the cloth-changing problem for gait recognition. Broadly, these methods can be classified into two categories, *i.e.*, model-based methods and appearance-based methods [2], [4], [47].

For model-based methods, a pre-process of locating skeleton key-points is first required, and gait features are extracted from the estimated human models [74]. Comprised of information of different body parts and connections between correlative parts, these estimated human models represent a primitive expression of each subject, which minimizes the effects on gait of clothing variations. In [11], legs were taken as an interlinked pendulum, and gait features were developed from the frequency variations of each thigh inclination. In [5], gait signatures were developed from human joints through elliptic Fourier descriptors. In [59], gait features were integrated by fusing static and dynamic body biometrics. Static body information was attained by Procrustes Shape Analysis, and dynamic gait information was acquired by tracking each person and recovering the joint-angle trajectories of their lower limbs. For the aforementioned methods, the main

Manuscript received June 11, 2021; revised July 28, 2021 and August 23, 2021; accepted September 10, 2021. *(Corresponding author: Worapan Kusakunniran.)*

Lingxiang Yao, Qiang Wu, Jingsong Xu and Jian Zhang are with the School of Electrical and Data Engineering, University of Technology Sydney, Ultimo, NSW 2007, Australia (e-mail: Lingxiang.Yao@student.uts.edu.au, Qiang.Wu@uts.edu.au, Jingsong.Xu@uts.edu.au, Jian.Zhang@uts.edu.au).

Worapan Kusakunniran is with the Faculty of Information and Communication Technology, Mahidol University, Salaya, Nakhon Pathom 73170, Thailand (email: worapan.kun@mahidol.edu).

Copyright © 2021 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending a request to pubs-permissions@ieee.org.

1051-8215 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Northeastern University. Downloaded on September 24,2021 at 01:33:35 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2021.3112564, IEEE Transactions on Circuits and Systems for Video Technology

TABLE I: Averaged rank-1 accuracies (%) on CASIA-B under the LT setting, excluding identical-view cases.

Method			Performance			
Reference	Year	Venue	NM	BG	CL	Average
[39]	2016	Int. J. Biom.	90.8	45.9	45.3	60.7
[66]	2017	IEEE T-PAMI	94.1	72.4	54.0	73.5
[74]	2019	CVPR	93.9	82.6	63.2	79.9
[53]	2019	IET Biom.	94.5	78.6	51.6	74.9
[7]	2019	AAAI	95.0	87.2	70.4	84.2
[73]	2020	IEEE T-PAMI	92.3	88.9	62.3	81.2
[25]	2020	IEEE Access	95.1	87.9	74.0	85.7
[16]	2020	CVPR	96.2	91.5	78.7	88.8
[22]	2020	ECCV	96.8	**94.0**	77.5	89.4
[33]	2020	ACCV	**97.9**	93.1	77.6	89.5
[34]	2020	MM	96.7	93.0	**81.5**	**90.4**
[52]	2021	ICPR	95.7	90.7	72.4	86.3
[51]	2021	IEEE T-Biom.	95.2	89.7	74.7	86.5

challenge is that the predicted locations of human joints are not always robust on markerless motions [56], [68]. Recently, with the rapid development of deep learning-based technologies, the prediction precision of human joints also has been dramatically improved [6], [17]. Inspired by this, many deep learning-based methods also have been proposed to handle this cloth-changing problem. For example, in [70], SGEI, a novel model-based gait feature, was created based on the detected human skeleton key-points by [6]. However, although the human joints are detected more accurately, SGEI still can only obtain a comparable result to appearance-based features [70]. One reason is that compared with apperance-based methods, only innate model information, *e.g.*, joint locations, can be used in these model-based methods. Another reason is that for most datasets there is little difference in each person's silhouettes, and somewhat human appearances can be treated as the most discriminative features [37], whereas most of them are totally neglected in model-based methods.

For appearance-based methods, a pre-process of extracting a series of human silhouettes is first needed, and gait features are explored from these extracted human silhouettes. For example, GEI, one of the most widely-used gait features in this category, is averaged from an entire gait cycle of human silhouettes [19]. However, given that the gait of each person can be significantly influenced by clothing changes, many methods in this category prefer to explore features from the non/less affected body parts rather than the entire human silhouette. In [2], a heavier weight was assigned to the body parts that were unaffected by clothing changes. In [13], features were hybridized from each silhouette area, and the median widths of lower limb parts and the holistic silhouette. In [38], GEI was first segmented into three different body parts. After removing the co-factored information in each part, a co-factored GEI was formed by linking these segmented parts. For the aforementioned methods, the biggest challenge is how to delimit which parts are more robust to clothing changes and which parts are more vulnerable to these variations [48]. In order to obtain a more robust recognition result, many different methods have been utilized to solve this segmentation problem. For example, in [26], silhouettes were first separated into seven parts based on the anatomical studies of gait. Meanwhile, pixel points were used as boundaries to detect each particular factor. Among the separated seven parts, only the part where its factor was absent could be used for the final classification. Moreover, experiment results in [49] also have proved that compared with classical appearance-based methods, it would be more efficient to extract gait features from the non/less affected body parts.

In this paper, a robust part-based method is proposed for gait recognition to handle the cloth-changing problem. It is rational to use part-based strategies for cloth-changing gait recognition. For each person, in most cases clothing changes can only affect some parts of gait and the remaining parts still can keep steady. As Fig. 1 shows, setting a fitted shirt and a regular pair of pants as the standard dressing style, if someone wears a thick cotton-padded jacket, it only affects the upper body parts and the other parts keep unchanged. Besides, related experiments also certify that part-based features are more robust for cloth-changing gait recognition [26]. Thus, in our method we pay more attention to generating high-quality part-based gait recognition features.

Following the above-mentioned ideas and inspired by recent development of gait recognition and human skeleton detection, in this paper part-based features are attained from two different perspectives. First, given that gait is entangled with appearance and appearance-based methods also have achieved a prominent result for cloth-changing gait recognition by using certain part-based strategies [32], [37], [74], in our method features are first generated from the body parts which are non/less influenced by clothing changes. Specifically, based on the anatomical studies of gait, human silhouettes are first separated into three different parts, *i.e.*, the head, the crus, and the left human body parts. As Fig. 1 indicates, the upper bodies are much likely to be affected by clothing changes compared with the other two parts. Hence, in our method we mainly focus on generating features from the head and crus parts. Besides, considering that local features are more helpful than global features in gait recognition [16], [66], in this paper a temporal pooling function is specifically used to extract the local short-range features instead of the global long-range features. Second, given that skeleton key-points illustrate more robustness to clothing variations, it is advisable to extract some part-based semantic features from each key-point region. Thus, in this paper a CNN backbone and a key-point prediction network are first used to extract these local features. After that, we treat these local features as word embeddings and propose a transformer encoder to jointly extract the semantic dependence between each correlative key-points.

As stated above, in this paper we address the cloth-changing problem by extracting robust part-based features from different perspectives, either appearance-based or skeleton-based. Given that each feature has its own limitations, in this paper these two features are concatenated to equip complementary functions in fields where the other is lacking. For features from the skeleton key-points, although more information has been retained in our features, the topology information of correlative key-points has been totally lacking. Considering that features from the divided silhouettes can offer the information from the non/less affected body parts, to some extent they will remedy the bereft topology information. For features from the segmented silhouettes, there is a possibility of suffering from inaccurate body segmentation, especially the skeleton key-points. However, such a loss can be complemented by adding features from the predicted key-point regions. Given the above, it is rational for us to concatenate the

1051-8215 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Northeastern University. Downloaded on September 24,2021 at 01:33:35 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2021.3112564, IEEE Transactions on Circuits and Systems for Video Technology

extracted features together for achieving a more efficient cloth-changing gait recognition. Furthermore, related experiments on CASIA Gait Dataset B will verify that our method outperforms other gait recognition methods for the cloth-changing problem.

In addition, main contributions of this paper are summarized as follows.

- In this paper, a new method is proposed to settle the cloth-changing challenge for gait recognition. First, features are developed from two different perspectives, either from the body parts which are less vulnerable to clothing variations or from the predicted skeleton key-point regions. Through their combination, a more robust gait feature is hybridized in this paper for cloth-changing gait recognition.
- A novel deep network is proposed in this paper for feature extraction. Two different subnetworks are included in this network, aiming to extract gait features from two different perspectives respectively.
- Since local features are more efficient than global features in gait recognition, in our method a new temporal pooling function is specifically proposed to extract the local short-range features. Moreover, different from most methods, in this paper we see all extracted semantic key-point features as word embeddings, and a transformer encoder is utilized to explore the dependence of correlative key-points.
- The proposed method can obtain the state-of-the-art result for cloth-changing gait recognition on the most frequently used gait dataset, CASIA Gait Dataset B.

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 Recognition Using Bag-of-Words Method

Inspired by [29], [57], a potential part-based feature learning method is proposed in this paper for gait recognition to address the cloth-changing problem.

The method that represents an object through a bag of visual words is broadly used in computer vision programs [35]. Many gait recognition methods also obtain a remarkable performance by using this technology. In [42], the walking sequence of each person was first encoded by a list of code words, thus the entire sequence could be represented by a feature vector implying the existence of each code word. In [29], considering the variations of both spatial and temporal directions, a bag-of-words method was developed based on the space-time interest points (STIPs). Descriptors were generated on a 3D patch in a neighborhood of each STIP, and a bag-of-words model was used to generate gait features from each descriptor set. In [1], another descriptor was produced for each STIP by extending LBP-TOP. A hierarchical K-means algorithm was utilized to map these descriptors into a set of code words. In [3], a bag-of-words method was proposed for Kinect to analyze and determine the severity of human gait. Moreover, in [45], a bag-of-words feature learning method was proposed to analyze the movements of patients with pathology. First, each time series were divided into subsequences using an overlapping sliding window. Similar patterns were identified as a visual word, and then a vocabulary was generated using these identified words. Thus, word features could be computed based on the similarity between the subsequences and the vocabulary. To sum up, it is reasonable to consider local features as a series of visual words in gait recognition.

Stimulated by the aforementioned bag-of-words methods, in this paper we also handle the extracted local features as a series of code words. Motivated by [57], a CNN backbone and a key-point detection model are first used to extract the local features of each skeleton key-point region. After that, a fully-connected layer is used to map these local features into word embeddings. Motivated by [15], a transformer encoder is specially proposed in our paper to model the dependence of correlative key-points. There are significant differences between the above-mentioned bag-of-words methods and our proposed method. First, distinct from the aforementioned methods whose word dictionaries are generated by extracted features, in our method the dictionary is predefined based on human skeleton key-points, which implies that in our method the local features are endued with meanings even since they are extracted. Second, for the above-mentioned methods, a global description is usually assembled by counting the occurrences of each visual word [35], while for our method the description is formulated by a transformer encoder, and this description will be more focused on portraying the dependence of each correlative key-points.

B. Transformers and Self-Attention

Due to their effectiveness across a range of fields like natural language processing, computer vision, *etc*, transformer models have garnered immense interest recently [43], [50], [55], [60].

Transformer models are multi-layer architectures, formed by arranging transformer blocks on top of another [54]. Generally, transformer blocks are comprised of a multi-head self-attention mechanism, a position-wise feed-forward network, a few layer normalization layers, and residual connectors. First, each input passes through an embedding layer to convert its one-hot token into a d dimensional embedding. This new embedding contains position encoding, and delivers to a multi-headed self-attention module. The input and output are connected through a residual connector and a layer normalization layer. After that, this novel output passes on to a two-layer feed-forward network, in which the input and output are also connected in a residual fashion by a layer normalization layer. For transformer models, the crucial characteristic is the multi-headed self-attention mechanism. To some degree, this self-attention mechanism can be deemed as a graph-like inductive bias, aiming to combine all tokens through a relevance-based pooling operation [54], [55].

According to the multi-headed self-attention mechanism, for computer vision applications to images would require that each pixel should connect to all other pixels [15], which is infeasible considering its computation cost. Thus, several approximations have been proposed for introducing transformers into computer vision applications. For example, [41] limited the self-attention mechanism to focus on local neighborhood of each pixel rather than the whole image. In [10], sparse factorization was utilized for the attention matrix in order to be more practical to images. In [15], images were first divided into a series of patches. After

1051-8215 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Northeastern University. Downloaded on September 24,2021 at 01:33:35 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2021.3112564, IEEE Transactions on Circuits and Systems for Video Technology

embedding each patch and appending positional embeddings, a sequence of vectors was formed and then delivers to a standard transformer encoder. Recently, a lot of networks also have been proposed for computer vision by assembling CNN and the self-attention mechanism [24], [30], [65]. This combination mainly focuses on two manners, either exploring more feature maps or further handling the output of CNN [15]. Furthermore, lately in some networks convolutional layers are fully replaced by these multi-headed self-attention blocks, and a distinct improvement has been verified in most datasets [44], [58].

In addition, transformers also have been introduced into gait recognition. In [31], a unified joint intensity transformer model was designed. For each input pair of GEIs, a sample-dependent joint intensity metric was first generated. Then, a joint intensity transformer module was used to obtain the spatial dissimilarity from the joint intensity metric. In [67], another pairwise spatial transformer network was founded. Each input pair of GEIs was first registered into a mediate subspace using a pairwise spatial transformer, followed by a recognition network to calculate the dissimilarity score for the registered pairs.

Inspired by [15], a transformer encoder is specially proposed in this paper to learn the dependence for correlative key-points. Different from [31], [67] using transformers to learn the spatial dissimilarity or to transform features into a shared space, in our method the transformer is used to model the relationship across correlative key-points. More details of this transformer encoder can be found in Section III-C.

C. Horizontal Pyramid Mapping

It is a common practice in gait recognition to divide features into different parts, because different body parts cause different influence on human gaits [32], [69].

Horizontal Pyramid Matching (HPM) method was proposed in [18] to capture various partial information of a given person. In [18], after capturing feature maps through a CNN backbone, a horizontal pyramid pooling was leveraged to segment feature maps into various local and global spatial bins. For each spatial bin, an average-pooling operation and a max-pooling operation were both utilized to extract discriminative information of each person part, followed by a convolutional layer to reduce feature dimension. In [7], HPM was further progressed into Horizontal Pyramid Mapping (HPM), where the initial convolutional layer was replaced by independent fully connect layers.

III. Proposed Methods

A. Overview

The core of image/video-based gait recognition is to explore gait-related features from the walking sequences [74]. Within a gait dataset, the walking sequence of each target can be tackled as a permutation of n frames, $\chi = \{x_i | i = 1, 2, 3, ..., n\}$. Thus, the gait-related features of each sequence can be denoted as,

$$f = H(G(F(\chi))) \tag{1}$$

where F means to extract gait-related features from each frame independently with a convolutional network. G denotes to map the features of each frame into a feature of the whole sequence, and H means to project the sequence feature into another more discriminative subspace to enhance its differentiation.

In our proposed method, gait features are extracted from two different perspectives, either from each estimated skeleton key-point region or from the segmented body parts that are non/less affected by clothing changes. Moreover, considering the strong complementarity of these two perspectives, a more robust part-based gait feature is integrated in our method by concatenating these two extracted gait features. Explicitly, assuming $\tilde{\chi}$ means a successive clip captured from χ, the concatenated gait feature integrated in our method can be formulated as,

$$f = H(G(F_{bp}(\tilde{\chi}_{bp}))) \bigoplus G(F_{kp}(\tilde{\chi}_{kp})) \tag{2}$$

where $\tilde{\chi}_{bp}$ represents the non/less affected body parts separated from the silhouettes of $\tilde{\chi}$, $\tilde{\chi}_{kp}$ represents the estimated skeleton key-point regions of $\tilde{\chi}$, F_{bp} and F_{kp} represent two subnetworks proposed in our paper to extract the gait-related features of $\tilde{\chi}_{bp}$ and $\tilde{\chi}_{kp}$, and $\bigoplus$ represents the concatenation of features. Eq. 2 is an enhanced version of Eq. 1, specifically proposed to tackle the cloth-changing gait recognition problem.

Fig. 2 presents the flowchart of our proposed method. On the one hand, after obtaining the silhouettes of each sequence, they are separated into the affected and non/less affected parts based on the anatomy information. Besides, a subnetwork is specially proposed in our paper to generate the gait-related features from the separated non/less affected body parts. On the other hand, a well-trained skeleton detection backbone is used in our method to locate the position of each skeleton key-point. A subnetwork is also specially utilized in this paper to explore the gait-related features from the located skeleton key-point regions. It is worth noting that our proposed network is an end-to-end architecture. The operations of human segmentation and skeleton estimation can be regarded as a preparatory process, and it is reasonable to assemble these operations into each proposed subnetwork.

B. Extracting Features from the Non/less Affected Body Parts

1) Segmenting the affected and non/less affected body parts: It is a widely-used operation in gait recognition to approach the cloth-changing problem by applying features from the non/less affected body parts. For these methods, the most decisive point is how to precisely segment human bodies into the affected and non/less affected parts. Basically, human bodies are segmented based on the anatomy information [12]. For a person height H, his/her body can be separated around some semantic positions, *e.g.*, knees $(0.285H)$, pelvis $(0.48H)$, waist $(0.535H)$ and neck $(0.87H)$ [12]. However, as Fig.1 illustrates, the upper bodies of each person are greatly affected by clothing variances. Thus, in this paper we mainly focus on two parts, the head part covering between the top of one's head and one's neck, and the crus part covering between one's knees and one's feet. It is worth noting that the body parts segmented in this paper are a little bit wider than they ought to be. We extend the limitation for two reasons. First, the discrimination power of a wider body part is stronger, but its robustness to clothing changes will become weaker, and *vice versa* for a narrower body part [2]. Second, apparel design is not always completely consistent with the human anatomical structures. Thus, in this paper when we segment human bodies,

1051-8215 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Northeastern University. Downloaded on September 24,2021 at 01:33:35 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2021.3112564, IEEE Transactions on Circuits and Systems for Video Technology

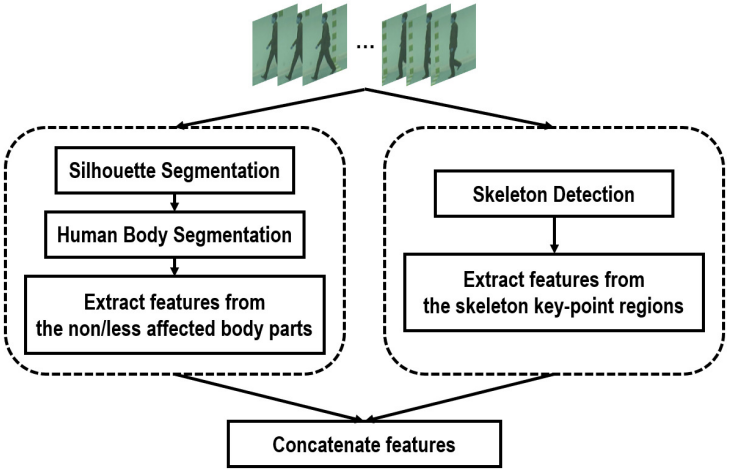

Fig. 2: Overview structure of the proposed network.

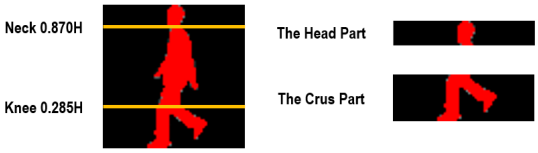

Fig. 3: Segmentation of human bodies.

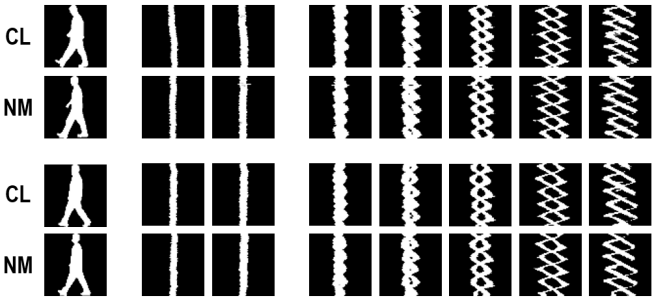

Fig. 4: Snapshots from the $T-W$ view of two persons in different dressing styles. Samples in the 2rd and 3st columns are obtained from the head part and samples in the 4-8th columns are obtained from the crus part.

a more flexible strategy is adopted by moderately extending the segmentation limitation. Fig. 3 presents a segmentation sample of the head and crus parts by using this strategy.

Fig. 4 shows some snapshots of the segmented head and crus parts from the $T-W$ view. Different from the common $H-W$ view, the $T-W$ view captures the displacement of a horizontal section across a time period, and the captured snapshots can be deemed as an aggregation of trajectories for points on the same horizontal line. As Fig. 4 shows, although these two people are dressed in different clothes, for each person there is no obvious difference between their captured $T-W$ snapshots. One major reason is that these snapshots are generated from their non/less affected head and crus parts, which also proves that it is logical to moderately extend the limitation when separating the human bodies. Meanwhile, it also certifies that the dynamic features of the non/less affected body parts have greater robustness against clothing changes, which motivates us to include some temporal cues when extracting the gait-related features. Moreover, Fig. 4 also shows that for each person there exists a prominent margin between their head and crus patterns. Therefore, in our method these two segmented parts should be processed respectively.

2) Extracting features from the non/less affected body parts: Following the above-mentioned ideas and inspired by the rapid development of gait recognition [16], [32], [66], in our method a subnetwork is specifically proposed to extract the gait-related features from the non/less affected body parts. The architecture of this subnetwork is indicated in Fig. 5. It can be seen that this subnetwork consists mainly of two branches, which are applied to handle the non/less affected head and crus parts respectively. These two branches share the same structure, both having three convolutional units. In each convolutional unit, two continuous convolutional operations are adopted. After each convolutional unit, there is a pooling module specially designed to extract the local micro-motion patterns within each short-range frames. At the top of our subnetwork, a global max-pooling function (G in Eq. 1 and Eq. 2) and a HPM module (H in Eq. 1 and Eq. 2) are arranged. More specifically, the global max-pooling function is proposed to integrate the short-range features into a long-range feature of the entire sequence, and the HPM module is adopted to map the features of each sequence into a more robust feature space to enhance the discrimination abilities. The final features used for identification are formed by assembling the features of the non/less affected head and crus parts.

Motivated by [7], [16], [66], a pooling module is specifically utilized in this paper to explore the local micro-motion patterns from the short-range frames. For gait, the frames with a similar visual appearance are more likely to appear at regular intervals, which highlights that the long-range dependencies, *e.g.*, longer than a complete gait cycle, can be redundant and ineffective for gait recognition [16]. Therefore, compared with the long-range dependencies, the local short-range features, *i.e.*, micro-motion patterns, are more efficient and discriminative for periodic gait. Moreover, experiments in [7], [66] also indicate that for a CNN model, pixels in feature maps of a deep layer are more involved with global coarse-grained cues while pixels in feature maps of

1051-8215 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Northeastern University. Downloaded on September 24,2021 at 01:33:35 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2021.3112564, IEEE Transactions on Circuits and Systems for Video Technology

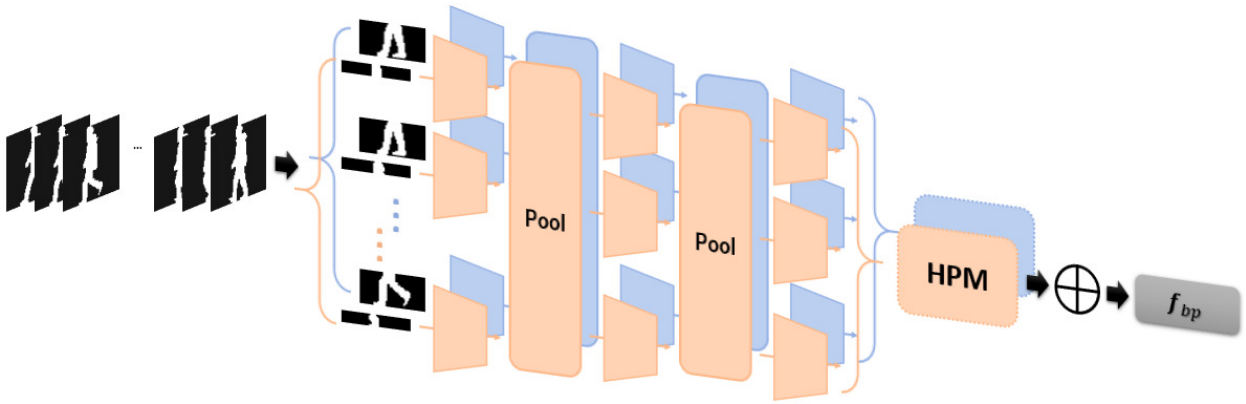

Fig. 5: The network proposed for extracting features from the non/less affected body parts.

a shallow layer are more involved with local fine-grained cues. Thus, in this paper we focus on extracting the local short-range micro-motion patterns from the shallow convolutional units.

Enlightened by GEI which generates a robust spatiotemporal descriptor by averaging frames over a complete gait cycle [19], in our method the local micro-motion patterns are generated by utilizing a global max-pooling function. Specifically, assuming $F(f_i, r) = \{f_k | k = i - r, \ldots, i + r\}$ denotes the feature maps extracted from the i-th and its r-neighbor frames, so the micro-motion patterns of the i-th frame can be generated as,

$$mp_i = maxpool(F(f_i, r)) \tag{3}$$

Different from GEI directly generating features over a whole gait cycle, in our method the gait-related features are generated in a progressive manner. As Fig. 5 illustrates, two max-pooling functions are contained in the proposed subnetwork. The working mechanism of the two max-pooling functions is similar to a sliding-window detector. The first pooling function is proposed to transform the frame-level features contained in each window into a micro-level feature. The second pooling function aims to integrate our obtained micro-level features into a more efficient micro-motion feature of a window in a larger size. Moreover, it is worth noting that there exists a great difference between [16] and our proposed method. In [16] its micro-motion features are directly generated from the global part-level features at the top, while in our method the micro-motion features are generated in a progressive way from our local feature maps of shallow units. Compared with [16], more discriminative local motion features can be maintained in our proposed method. Besides, in order to fuse multi-scale information, two different window sizes, *i.e.*, 3 and 5, are utilized in our method.

Moreover, HPM is also used in our paper to map the features of each body part into a more effective feature space. HPM has different scales, thus it can guide the deep networks to focus on features with different sizes, so as to gather the local and global information [7]. It is rational to apply HPM in gait recognition, because it is a common practice for gait recognition to segment features into different strips. In most gait recognition networks, features are split along the horizontal direction. However in our paper, since human bodies already have been separated into the affected and non/less affected parts based on the human height, we prefer to split the extracted features in the vertical direction. If HPM has S scales, then features will be split into 2^{s-1} strips in the vertical direction, and the feature f of the strip z_s will be formulated as,

$$f = maxpool(z_s) + avgpool(z_s) \tag{4}$$

where $maxpool$ and $avgpool$ represent the global max-pooling and global average pooling functions, respectively.

C. Extracting Features from the Skeleton Key-point Regions

It is rational for gait recognition to tackle the cloth-changing problem by using the skeleton key-point generated features. As stated above, made up of information about each body part and relationships between correlative parts, these features are more robust to approach the clothing variations than the appearance-based features [59], [70]. Following the above-mentioned ideas and inspired by the development of human pose estimation [6], [9], [17], a subnetwork is proposed in this paper to extract gait-related features from the human skeleton key-point regions.

Fig. 6 indicates the framework of our subnetwork. It consists of three modules, *i.e.*, semantic feature extracting module, key-point dependence learning module, and final feature generating module. The three modules are jointly trained in an end-to-end way, and a competent gait-related feature can be finally learned for cloth-changing gait recognition.

1) Semantic Feature Extracting Module: The main target of this module is to extract semantic features from each estimated skeleton key-point region. Motivated by recent development of gait recognition and human pose estimation [6], [9], [16], [17], for this module, a CNN backbone is used to learn feature maps, and a human key-point estimation model is used to predict key-points at the same time. Regarding the predicted key-point heat maps as the feature filter, then we can get the semantic features of each corresponding key-point.

Specifically, for each frame I, we can obtain its feature maps f_f and its key-point heat maps f_{kp} through the CNN backbone and the human key-point estimation model. Following, through an outer product ($\bigotimes$) and a fully-connected layer ($g()$), we can attain a group of semantic features from each key-point region. The whole process can be represented by Eq. 5, and K denotes the pre-defined key-point number. It is worth noting that f_{kp} in this paper is attained by normalization with a softmax function,

1051-8215 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Northeastern University. Downloaded on September 24,2021 at 01:33:35 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2021.3112564, IEEE Transactions on Circuits and Systems for Video Technology

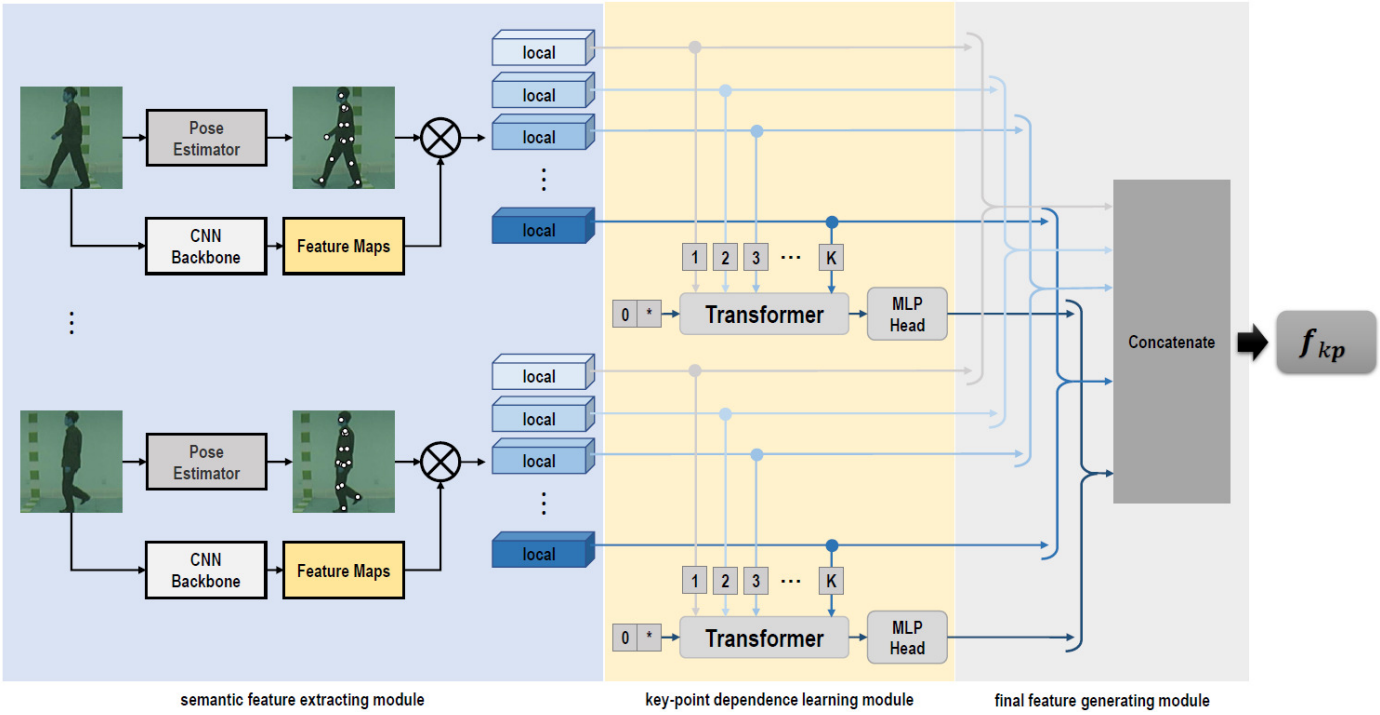

Fig. 6: The network proposed for extracting features from the skeleton key-point regions.

which can prevent from noise and outliers and has been proved more effective in relevant experiments [57].

$$V^S = \{v_k^S\}_{k=1}^K = g(f_f \bigotimes f_{kp}) \tag{5}$$

Moreover, it is also worth noting that in this module only the parameters of the CNN backbone can be learned in our training stage. For human key-point estimation, a pretrained model will be directly used, and no further optimization will be involved.

2) Key-point Dependence Learning Module: The main goal of this module is to learn the high-order dependence among the correlative key-points, which is motivated by two causes. First, for gait recognition, the model-based features are usually made up of two factors, the information about each body part and the connections between the correlative parts. Also, although more information has been included in our earlier obtained key-point features, they may still suffer from the problems of information loss compared with appearance-based features. Hence, in order to offset the lost information and the connections between each correlative key-points, a module is specifically proposed in this paper to explore the dependence of the correlative key-points.

Motivated by [15], [29], in our method we treat the extracted semantic key-point features as a set of word embeddings, and a transformer encoder is used to explore the dependence of these embeddings. Basically, this encoder is comprised of alternative multi-head self-attention and MLP blocks. For each correlative key-points, the corresponding embeddings are associated as an effective input sentence for the transformer encoder. Besides, a learnable embedding is also added to serve as the output results of the transformer encoder. Moreover, position embeddings are also used to retain the semantic information for each key-point. Thus, for a frame I, after attaining its semantic features of each key-point $V^S = \{v_k^S\}_{k=1}^K$ via Eq. 5, its correlative dependence $V^D = \{v_k^D\}_{k=1}^R$ can be formulated as,

$$V^D = f_D(V^S) \tag{6}$$

where f_D denotes our proposed key-point dependence learning module, and R denotes the number of connections between the correlative semantic key-points.

More specifically, for each $v_k^D, k \in \{1, 2, \cdots, R\}$, assuming the semantic features of its n correlative key-points can be represented as $\{v_{k_1}^S, v_{k_2}^S, \cdots, v_{k_n}^S\}, v_{k_i}^S \in V^S, i \in \{1, 2, \cdots, n\}$, thus v_k^D can be formulated as,

$$\begin{aligned} z_0 &= [v^*; v_{k_1}^S; v_{k_2}^S; \cdots; v_{k_n}^S] + v_{pos} \\ z_1 &= \mathrm{MSA}(\mathrm{LN}(z_0)) + z_0 \\ z_2 &= \mathrm{MLP}(\mathrm{LN}(z_1)) + z_1 \\ v_k^D &= \mathrm{LN}(z_2) \end{aligned} \tag{7}$$

where v^* means the learnable embedding to serve as the output results. v_{pos} represents the position embeddings to preserve the key-point semantic knowledge. LN denotes LayerNorm layers. MSA means the multi-head self-attention used in transformers, and MLP represents the involved MLP blocks.

Different from most methods seeing the semantic features as nodes and adopting graph convolutional networks to model the high-order relationship knowledge [57], in our method we take the semantic key-point features as word embeddings and adopt a transformer encoder to learn the high-order dependence from the correlative key-points. It is advisable to adopt a transformer encoder to model such dependence. Transformers, in particular self-attention-based structures, have dominated a wide range of tasks in natural language processing (NLP) [55]. Recently, they also have achieved a prominent result in many computer vision

1051-8215 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Northeastern University. Downloaded on September 24,2021 at 01:33:35 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2021.3112564, IEEE Transactions on Circuits and Systems for Video Technology

tasks [15], [36], [62], [65]. Attention mechanisms have become an integral part of dependence modeling without regard to their positions of the input or output sequences [27], [55]. Moreover, self-attention mechanisms establish this dependence further by relating different positions for each single sequence. Compared with graph convolutional methods learning the relation through the nearest nodes, in our method the dependence is modeled by directly connecting the correlative key-points, which makes the learned dependence to be more concise and efficient.

In addition, considering that the upper bodies of each person can be significantly influenced by clothing changes, thus in this module we more focus on the detected lower body parts. Given the dynamics studies of gait, three different dependence will be learned in our module, one for the left leg, one for the right leg, and the other for the entire lower body part.

3) Final Feature Generating Module: The primary target of this module is to integrate the final gait-related features for gait recognition. A global max-pooling function is first used to map the features of each frame into a feature of the whole sequence. After that, a concatenation operation is adopted to hybridize all features together, and the concatenated features will be utilized as our final gait-related features for gait recognition.

Specifically, for a frame clip $\tilde{\chi} = \{x_i | i = 1, 2, 3, ..., m\}$, we can produce the semantic key-point features for each frame V_i^S via Eq. 5 and extract the correlative dependence for each frame V_i^D via Eq. 6. Thus, the procedures of generating the final gait-related features $f_{kp} = \{v_k\}_{k=1}^{K+R}$ can be formulated as,

$$f_{kp} = maxpool(V_i^S) \bigoplus maxpool(V_i^D) \tag{8}$$

It is suitable to concatenate the extracted semantic key-point features and their contained correlative dependence as our gait-related features, since model-based features are basically made up of two types of information, *i.e.*, the information of different body parts and the connections between correlative parts. Also, it is worth noting that the global max-pooling function not only wraps the features of each frame into a feature of the sequence, but also reduces the influence of incorrect key-point prediction. Although the detection precision of human key-points has been remarkably improved, they still suffer from the occluded cases. However in our method, rather than use a single frame as input, we feed our proposed network with a clip of successive frames. Considering that occlusion can only exist in several frames and each key-point feature is produced from a predicted region, our pooling function can restrain the impacts of incorrect key-point detection and maintain the most discriminative components for gait recognition.

D. Assembling Features from Two Different Perspectives

As stated above, in our method two different part-based gait-related features are extracted for gait recognition to address the cloth-changing problem. Considering the great complementary of these two features, a more efficient gait feature is hybridized in our paper by concatenating these two features together.

Specifically, let f_{bp} represent the features generated from the non/less body parts and let f_{kp} represent the features generated from the skeleton key-point regions, then our final attained gait features f can be formulated as,

$$f = f_{bp} \bigoplus f_{kp} \tag{9}$$

It is advisable to unite the extracted two gait-related features together as our final attained gait features, because each feature has its own advantages and disadvantages, and the combination will provide complementary functions in fields where the other is lacking. For features from the skeleton key-point regions, the topology knowledge of correlative key-points has been entirely lacking, although their semantic dependence has been specially described in our method. On the other hand, given that features from the non/less affected body parts can provide useful spatial clews for gait recognition, to some extent they can make up for the missing topology knowledge. Meanwhile, for features from the non/less affected body parts, although in this paper we have reasonably extended the segmentation limitation, they may still suffer from deviant body segmentation, especially at some key-point positions. However, such a loss can be remedied by using features from the skeleton key-point regions. Overall, given the high complementary of these two features, it is advisable for us to combine these two features together for a more robust cloth-changing gait recognition. In addition, relevant experiments on CASIA Gait Dataset B also verify that our method outperforms other gait recognition methods for the cloth-changing problem.

In addition, the triplet loss is used in this paper to distinguish differences among the concatenated gait features.

TABLE II: Comparison on CASIA-B under the same normal viewing angle by accuracies (%).

Probe Set	ours	[69]	[7]	[4]	[13]
36°(nm)	**100.0**	**100.0**	**100.0**	90.5	89.5
36°(cl)	**100.0**	**100.0**	**100.0**	90.9	91.1
54°(nm)	**100.0**	**100.0**	**100.0**	91.1	88.2
54°(cl)	**100.0**	**100.0**	**100.0**	93.2	91.9
72°(nm)	**100.0**	**100.0**	**100.0**	94.7	88.7
72°(cl)	**100.0**	**100.0**	**100.0**	96.5	89.5
90°(nm)	**100.0**	**100.0**	**100.0**	93.5	87.1
90°(cl)	**100.0**	99.2	**100.0**	95.1	88.7
108°(nm)	**100.0**	**100.0**	**100.0**	92.7	-
108°(cl)	**100.0**	99.2	**100.0**	94.1	-
126°(nm)	**100.0**	**100.0**	**100.0**	91.1	-
126°(cl)	**100.0**	**100.0**	**100.0**	91.5	-
144°(nm)	**100.0**	99.7	**100.0**	92.2	-
144°(cl)	**100.0**	**100.0**	**100.0**	93.5	-

IV. Experiments

In this section, we will evaluate the efficiency of our method on the most widely-used gait database, CASIA Gait Dataset B. A comparison is first made on this dataset between our method and some other state-of-the-art gait recognition methods. After that, ablation experiments are conducted. Relevant experiments will indicate that the proposed method outperforms other state-of-the-art methods for cloth-changing gait recognition.

A. Training and Testing Details

As the most broadly used gait database, CASIA Gait Dataset B [75] contains videos for 124 subjects from 11 viewing angles

1051-8215 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Northeastern University. Downloaded on September 24,2021 at 01:33:35 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2021.3112564, IEEE Transactions on Circuits and Systems for Video Technology

TABLE III: Comparison on CASIA-B under different walking conditions by accuracies (%).

(Probe, Gallery)	ours	PartGait [69]	GaitNet [74]	L-CRF [8]	LB [66]	RLTDA [23]	CPM [63]	NN [72]
(54°, 36°)	**95.5**	93.7	87.0	59.8	49.7	69.4	19.4	16.4
(54°, 72°)	**97.8**	94.1	90.0	72.5	62.0	57.8	22.2	11.2
(90°, 72°)	**98.9**	98.8	94.2	88.5	78.3	63.2	48.5	23.5
(90°, 108°)	**98.9**	98.7	86.5	85.7	75.6	72.1	55.9	25.9
(126°, 108°)	**96.7**	94.9	89.8	68.8	58.1	64.6	26.7	16.7
(126°, 144°)	**97.8**	93.5	91.2	62.5	51.4	64.2	29.2	19.2
Mean	**97.6**	95.6	89.8	73.0	62.5	65.2	33.7	18.9

TABLE IV: Averaged rank-1 accuracies (%) on CASIA-B under three different settings, excluding identical-view cases.

Gallery NM#1-4		Modality	0°-180°											
Probe CL#1-2			0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	Mean
ST(24)	GaitSet [7]	Silhouettes	29.4	43.1	49.5	48.7	42.3	40.3	44.9	47.4	43.0	35.7	25.6	40.9
	PartGait [69] (30f)	Silhouettes	34.0	47.1	51.0	54.0	52.9	48.9	49.8	50.3	48.2	41.4	30.5	46.2
	PartGait [69] (64f)	Silhouettes	38.1	52.3	57.9	59.1	56.2	51.3	53.8	56.6	56.3	48.0	31.2	51.0
	Ours	RGB frames	**62.0**	**69.5**	**72.7**	**71.3**	**62.5**	**55.9**	**61.9**	**64.8**	**60.7**	**58.8**	**48.1**	**62.5**
MT(62)	AE [71]	Silhouettes	18.7	21.0	25.0	25.1	25.0	26.3	28.7	30.0	23.6	23.4	19.0	24.2
	MGAN [20]	Silhouettes	23.1	34.5	36.3	33.3	32.9	32.7	34.2	37.6	33.7	26.7	21.0	31.5
	GaitSet [7]	Silhouettes	52.0	66.0	72.8	69.3	63.1	61.2	63.5	66.5	67.5	60.0	45.9	62.5
	PartGait [69] (30f)	Silhouettes	59.2	74.7	77.4	74.5	69.5	66.3	69.8	74.4	73.6	69.2	52.5	69.2
	PartGait [69] (64f)	Silhouettes	61.8	77.6	83.1	80.4	74.3	70.5	75.7	80.8	81.1	74.9	54.9	73.4
	Ours	RGB frames	**85.7**	**91.2**	**92.3**	**89.7**	**86.8**	**83.7**	**87.2**	**90.2**	**89.8**	**88.3**	**78.8**	**87.6**
LT(74)	CNN-LB [66]	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
	GaitSet [7]	Silhouettes	61.4	75.4	80.7	77.3	72.1	70.1	71.5	73.5	73.5	68.4	50.0	70.4
	[25]	Silhouettes	64.7	79.4	84.1	80.4	73.7	72.3	75.0	78.5	77.9	71.2	57.0	74.0
	GLN [22]	Silhouettes	70.6	82.4	85.2	82.7	79.2	76.4	76.2	78.9	77.9	78.7	64.3	77.5
	GaitPart [16]	Silhouettes	70.7	85.5	86.9	83.3	77.1	72.5	76.9	82.2	83.8	80.2	66.5	78.7
	[52]	Silhouettes	63.4	77.3	80.1	79.4	72.4	69.8	71.2	73.8	75.5	71.7	62.0	72.4
	[51]	Silhouettes	65.8	80.7	82.5	81.1	72.7	71.5	74.3	74.6	78.7	75.8	64.4	74.7
	PartGait [69] (30f)	Silhouettes	64.2	80.9	83.0	79.5	74.3	69.1	74.8	78.5	81.0	77.0	60.3	74.8
	PartGait [69] (64f)	Silhouettes	71.8	86.6	87.7	83.2	78.3	75.4	81.0	85.2	84.9	82.0	64.1	80.0
	GaitNet [74]	RGB frames	42.1	-	-	70.7	-	70.6	-	69.4	-	-	-	63.2
	[33]	RGB frames	78.2	81.0	82.1	82.8	80.3	76.9	75.5	77.4	72.3	73.5	74.2	77.6
	Ours	RGB frames	**90.4**	**94.2**	**93.9**	**92.3**	**88.0**	**87.2**	**88.8**	**91.7**	**90.9**	**92.2**	**87.3**	**90.6**

(0°, 18°, 36°, . . . , 180°). For each subject, ten video sequences are contained, six normal sequences (NM), two sequences with a long coat (CL), and two other sequences carrying a bag (BG). Some frames sampled from this dataset are presented in Fig. 1. We measure the robustness of our proposed method on CASIA Gait Dataset B, mainly because this dataset is the only publicly available dataset that offers the original gait videos for skeleton key-point estimation.

In all experiments, our input is a clip of successive frames in size of 64×64, and each clip length is set at 30. Meanwhile, for CASIA Gait Dataset B, frames are aligned by the same method in [7]. In our training phase, a batch in size of 8×8 is randomly sampled from the training dataset, which illustrates that in each batch the number of subjects and the number of clips every one has are both set at 8. For the subnetwork creating features from the non/less affected body parts, the convolutional channels are set at 32, 64 and 128 respectively, and the scales S of HPM are set at 5. For the subnetwork creating features from the skeleton key-point regions, a network similar to GaitSet [7] is treated as our backbone, and a well-trained HRNet model [9] is also used for human skeleton detection. Adam is chosen as the optimizer, and the learning rate is set ai $1e{-}4$ [28]. Besides, the margin of BA_+ triplet loss is set at 0.2 [21]. However in the testing stage, our batch size is set at 1.

B. Comparison on CASIA Gait Dataset B

Our experiments contain three different parts.

In the first part, for each viewing angle, a general training set of 496 sequences is aggregated by the first three NM sequences and the first CL sequence of each subject. Two different testing sets are utilized in our testing phase. The first testing set is built by the left three NM sequences of each subject, and the second testing set is formed by their left CL sequences. Table. II shows the comparison results of our proposed method and some other gait recognition methods at the viewing angles of 36°, 54°, . . . , 144°. This experiment is intended for evaluating the robustness of our proposed method, thus only the common viewing angles are utilized in this experiment. It can be seen that our proposed method outperforms the other methods with an evident margin, achieving the accuracy of 100% for most cases. Thus, based on this experiment, we can conclude that compared with other gait recognition methods, the proposed method is more feasible and efficient when handling the cloth-changing problem.

In the second part, both viewing angles and clothing changes are taken into consideration. As Table. III reveals, in this part 6 probe/gallery view pairs are first combined from the 7 common viewing angles. For each probe/gallery view pair, a training set is built by sequences of the first 34 subjects at the probe/gallery viewing angles (θ_p, θ_g). In the testing phase, a probe set is built

1051-8215 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Northeastern University. Downloaded on September 24,2021 at 01:33:35 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2021.3112564, IEEE Transactions on Circuits and Systems for Video Technology

TABLE V: Averaged rank-1 accuracies (%) on CASIA-B using setting LT, excluding identical-view cases.

LT Setting		Modality	Probe Views											
			0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	Mean
Gallery NM#1-4 Probe BG#1-2	CNN-LB [66]	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
	GaitSet [7]	Silhouettes	83.8	91.2	91.8	88.8	83.3	81.0	84.1	90.0	92.2	94.4	79.0	87.2
	[25]	Silhouettes	84.3	91.2	93.4	91.8	86.1	80.3	84.4	90.0	93.7	90.8	80.1	87.9
	GLN [22]	Silhouettes	91.1	97.7	97.8	95.2	92.5	91.2	92.4	96.0	97.5	95.9	88.1	94.0
	GaitPart [16]	Silhouettes	89.1	94.8	96.7	95.1	88.3	94.9	89.0	93.5	96.1	93.8	85.8	91.5
	[52]	Silhouettes	87.3	93.7	94.8	93.1	88.1	84.5	88.8	93.5	96.3	93.3	83.9	90.7
	[51]	Silhouettes	86.0	93.3	95.1	92.1	88.0	82.3	87.0	94.2	95.9	90.7	82.4	89.7
	PartGait [69] (64f)	Silhouettes	73.0	84.9	86.8	82.7	80.0	75.5	81.3	85.8	86.7	84.1	70.9	81.1
	GaitNet [74]	RGB frames	83.0	-	-	86.6	-	74.8	-	85.8	-	-	-	82.6
	[33]	RGB frames	94.8	92.9	93.8	94.5	93.1	92.6	94.0	94.5	89.7	93.6	90.4	93.1
	Ours	RGB frames	**96.0**	**96.2**	**97.3**	**96.1**	**93.9**	**91.8**	**93.0**	**95.7**	**96.6**	**97.3**	**94.0**	**95.3**
Gallery CL#1-2 Probe NM#5-6	GaitSet [7]	Silhouettes	61.7	69.1	72.7	75.6	72.7	69.8	69.8	73.4	71.8	68.4	55.0	69.1
	PartGait [69] (64f)	Silhouettes	71.7	79.8	86.5	85.1	78.5	73.6	79.4	87.5	86.5	79.3	65.1	79.4
	Ours	RGB frames	**86.2**	**90.5**	**92.8**	**92.9**	**89.7**	**89.8**	**91.7**	**93.9**	**94.2**	**92.9**	**87.3**	**91.1**
Gallery CL#1-2 Probe BG#1-2	GaitSet [7]	Silhouettes	57.4	63.3	68.3	69.6	63.0	58.7	62.5	68.7	69.9	62.0	51.9	63.2
	PartGait [69] (64f)	Silhouettes	65.5	78.1	80.5	77.7	73.5	68.0	74.7	79.7	80.8	72.9	62.1	73.9
	Ours	RGB frames	**84.4**	**86.9**	**88.8**	**85.9**	**84.2**	**82.0**	**84.9**	**89.7**	**89.7**	**90.5**	**85.2**	**86.6**

by the CL sequences of the left subjects at the viewing angle of θ_p, and a gallery set is made up of the left NM sequences at the angle of θ_g. Table. III gives the results of this proposed method and some other gait recognition methods. Compared with other methods, the proposed method achieves the highest accuracy in this view-changing and cloth-changing environment. The mean accuracy of our proposed method is 97.6%, outperforming [69] by 2.0%. Although our method is not intended for handling the view-changing challenge, it still has a better robustness against viewing variations. Thus, compared with other gait recognition methods, the proposed method has wider practical prospects in real-world applications.

In the last part, our proposed method is compared with state-of-the-art deep learning-based gait recognition methods. Three normal experiment conditions, *i.e.*, small-sample training (ST), medium-sample training (MT), and large-sample training (LT), are adopted in this experiment [7]. In ST, a lightweight training set is produced by sequences of the first 24 subjects. In MT, the training set is produced by sequences of the first 62 subjects. In LT, the training set is constructed with sequences of the first 74 subjects. For all three conditions, the testing sets are integrated by sequences of the remaining subjects. The two CL sequences are used as the probe, and the first four NM sequences are used as the gallery. Table. IV shows the comparative results between our proposed method and other state-of-the-art gait recognition methods. The results indicated in Table. IV are averaged on the 11 gallery views, and identical views are excluded in this table. Besides, two results are given for PartGait [69], and their major difference is that their input frame numbers are different. It can be seen from Table. IV that our proposed method has presented the state-of-the-art performance in all conditions. For example, the mean accuracy of our proposed method in LT is 90.6%, and it outperforms GaitPart [16] by 11.9%. Besides, although more than double the silhouette frames are input in PartGait [69], our proposed method still exceeds PartGait [69] by almost 10%. In order to illustrate the robustness of this proposed method, more comparisons are given in Table. V in the LT setting. To sum up, these experiments have certified the robustness of our proposed method when approaching cloth-changing gait recognition.

In all, experiments on CASIA Gait Dataset B have indicated that the proposed method is efficient and flexible when tackling the cloth-changing problem in gait recognition. Compared with other methods, our proposed method has achieved the state-of-the-art performance on this dataset. A better robustness against clothing changes has been verified by these experiments for the proposed cloth-changing gait recognition method.

C. Ablation Experiments of Module Effectiveness

In this part, ablation experiments are intended to evaluate the effectiveness of each module used in our proposed network.

1) Features from the Non/less Affected Body Parts: The first three lines of Table. VI indicate the effectiveness of integrating features from the non/less affected head and crus parts.

The first and the third lines of Table. VI certify the efficiency of our proposed local micro-motion pooling modules. It can be seen that our mean accuracy will be increased by 2.7% through these two pooling modules. Also, compared with PartGait [69] and GaitPart [16], although no key-point features are imported, our proposed method still achieves a comparable performance. Different from GaitPart [16] integrating micro-motion patterns from the top layer, in our method the micro-motion patterns are extracted from the shallow units, thus more local micro-motion cues can be maintained in the proposed method. Different from PartGait [69] extracting temporal features from the $T-W$ view respectively, in our method the temporal information is directly formulated from the $H-W$ view. Based on these comparisons, it can be found that our proposed modules are more effective at extracting micro-motion features for gait recognition.

Also, the second and the third lines of Table. VI illustrate the impacts of our proposed HPM and the normally used HPM [7]. In our HPM, since human bodies are already separated into the affected and non/less affected parts based on the human height, the extracted features are split in the vertical direction, while in the normally used HPM [7] these features are segmented in the horizontal direction. We can see that the mean accuracy will be increased to 97.8% using our proposed HPM. To sum up, these

1051-8215 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Northeastern University. Downloaded on September 24,2021 at 01:33:35 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2021.3112564, IEEE Transactions on Circuits and Systems for Video Technology

TABLE VI: Accuracies (%) of different modules on CASIA-B using setting LT, excluding identical-view cases.

f_{bp}		f_{kp}		MEAN ACCURACY
Local Micro-motion Pooling	HPM	Only Key-Point Features	Only Key-Point Dependence	
	√			77.1
√				78.2
√	√			79.8
		√		72.1
			√	50.8
		√	√	74.4
√	√	√	√	90.6

comparative experiments have certified the effectiveness of our proposed improved HPM.

Finally, it also can be found from this table that even without key-point features, the proposed method already has achieved a more splendid result than PartGait [69] and GaitPart [16] when the same input frame number is utilized.

2) Features from the Skeleton Key-point Regions: Table. VI also presents the effectiveness of solely using features from the estimated skeleton key-point regions, simply using dependence of correlative key-points, and their concatenations.

It can be seen that compared with either of them solely being used, an improvement has been shown for their concatenations. However, compared with features of the non/less affected body parts, these key-point concatenations still cannot work well.

In our method, in order to preserve more information of each detected skeleton key-point, semantic features are formed from the predicted key-point regions. Besides, semantic dependence of each correlative key-points is also formulated to enhance the characterization capabilities of our obtained key-point features. Thus, a better result can be attained through their combination. However, compared with features of the non/less affected body parts, some discriminative information is still lost in these key-point features, *e.g.*, the lengths and angles of thighs and calves. Meanwhile, these topology information has a significant part in gait recognition [5], [11]. Hence, it is rational to combine these two features for a more robust cloth-changing gait recognition.

Besides, in order to present the effectiveness of our proposed key-point dependence learning module, a simple comparison is also arranged in this part between our proposed method and the widely-used graph convolutional networks [57], [61]. In [57], a high-order relation module is specifically proposed to grasp the high-order relation information between correlative key-points. Hence, in this part a comparison network is raised by replacing our key-point dependence learning module with this high-order relation module. The same experiment settings are adopted, but its mean accuracy only reaches 72.4%, lower than the proposed method by 2.0%. In [57], the key-point relation is grasped from the nearest points. In our method, the dependence is grasped by jointing all correlative key-points, thus this learned dependence can be more concise and effective. To sum up, we can conclude that it is advisable to adopt our proposed key-point dependence learning module to explore the high-order dependence between correlative key-points for gait recognition.

3) Assembling Features: The effectiveness of concatenating these two aforementioned features is also verified in Table. VI. As Table. VI reveals, the mean accuracy of directly adopting features of the non/less affected body parts reaches 79.8%, and the mean accuracy of directly utilizing features of the predicted skeleton key-point regions is 74.4%. However, their assembled accuracy increases to 90.6%, higher than either of them.

It is a common practice for gait recognition to unite different features together for a more remarkable recognition result [59], [70]. Each feature has its own strengths and deficiencies, thus a proper feature combination can offer complementary functions in fields where the others are missing. In our method, these two features are extracted from two different perspectives, and each of them offers different information of gait. Given the diversity and the complementarity of these two features, a more efficient gait feature can be developed in our method for cloth-changing gait recognition through their combination.

More specifically, two different perspectives are emphasized in these two features. Features from the predicted skeleton key-point regions more focus on generating information about each specific skeleton key-point and their implicit relationships, and features from the non/less affected body parts are more capable of capturing the shape information about each non/less affected body area. Thus, through their combination, a healthy tendency of incorporating points into planes can be created for these two features. Meanwhile, the combination also can complement the deficiencies that the other feature is suffering. On the one hand, features from the non/less affected body parts will help remedy the topology information that features of the predicted skeleton key-point regions are missing, such as the lengths of thighs and calves. On the other hand, features from the estimated skeleton key-point regions to some extent may help reduce the influence of inaccurate body segmentation which features of the non/less body parts can be suffering, especially some skeleton key-point positions. Thus, given the difference and the complementary of these two features, it is advisable to combine these two features together for a more efficient cloth-changing gait recognition.

D. Studies of Features from the Non/less Affected Body Parts

In this part, ablation experiments are intended to evaluate the efficiency of features produced from the non/less affected body parts. First, attention is attached to our separated different body parts, which shows that it is advisable to assemble the head and crus parts together for cloth-changing gait recognition. Second, a comparison experiment is designed to evaluate the robustness of our proposed local micro-motion patterns.

The first five lines of Table. VII show the comparison results of different body parts and their combinations. It is evident that

1051-8215 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Northeastern University. Downloaded on September 24,2021 at 01:33:35 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2021.3112564, IEEE Transactions on Circuits and Systems for Video Technology

TABLE VII: Accuracies (%) of different body parts.

Body Parts		NM	BG	CL
Local	Head	44.3	39.4	24.7
	Upper Body	79.6	45.8	21.2
	Crus	79.5	72.2	74.3
	Head+Upper+Crus	95.2	87.4	74.4
	Head+Crus	86.3	80.0	**79.8**
Global [16]	Head+Crus	84.3	77.7	75.6

TABLE VIII: Complexity Analysis.

Model	Params(M)	GFLOPs(G)
GaitSet [7]	0.563	13.039
PartGait [69]	1.687	26.197
GaitNet [74]	101.338	54.892
subnetwork for f_{bp}	1.125	8.194
subnetwork for f_{kp}	150.798	57.026
the entire network	151.923	65.220

different from the head and crus parts that are non/less affected by clothing variances, the upper body parts can be dramatically affected. The accuracy is declined by over 50% when the target subject changes his/her dressing styles from the normal clothes to a long coat. Given that more stable and efficient gait features can be supplied from the head and crus parts, in our method we mainly focus attention on these two parts. Also, compared with the combination of head and crus, the combination of the entire body parts can merely get a comparable result for the CL case. It also certifies that it is reasonable for us to pay more attention to the head and crus parts when settling the cloth-changing gait recognition problem.

The last two lines of Table. VII compares our proposed local micro-motion patterns and the micro-motion patterns proposed in [16]. In our method, the micro-motion patterns are generated from the shallow units, while in [16] the micro-motion patterns are generated at the top after all convolutional calculation and a HPM module. It is evident that our local micro-motion patterns can indicate a better performance, outperforming [16] by 4.2%. Experiments in [66] have verified that for gait recognition local features are more powerful than global features. In our method, the micro-motion patterns are captured from local feature maps of shallow units, thus more local fine-grained cues are included in our generated micro-motion patterns. In [16], however, these micro-motion patterns are captured at the top, thus more global coarse-grained cues can be supplied by these captured patterns. Therefore, compared with the micro-motion patterns utilized in [16], our proposed local micro-motion patterns are more robust and efficient for cloth-changing gait recognition.

E. Complexity Analysis

In our method, the proposed network is trained and tested on two NVIDIA V100 GPUs. 50,000 iterations are involved in the training phase, and it takes about 25.5 hours to finish our whole training. A brief complexity analysis of this network is given in Table. VIII. We can see that although numerous parameters are adopted in our proposed network, the subnetwork for key-point features holds a bigger proportion. The subnetwork for features of the non/less affected body parts only takes a few parameters, even fewer than [69]. Moreover, for the subnetwork of learning features from key-point regions, the parameters required by the first semantic feature extracting module account for two-thirds. One main reason is that for this module a fully-connected layer is adopted to embed our semantic features of each key-point. In the future work, it can be replaced by resource-saving methods, *e.g.*, the pooling operations in [57].

V. Conclusion

In this paper, a novel method is proposed for gait recognition to tackle the cloth-changing problem. First, part-based features are formulated from two perspectives. Given that in most cases clothing variances can only influence some parts of gait, in this paper features are first generated from the body parts which are non/less affected by clothing changes. Meanwhile, considering that skeleton key-points are more robust to clothing changes, in this paper the second features are generated from the estimated skeleton key-point regions. Moreover, because each feature has its own strengths and weaknesses, in this paper a more efficient feature is hybridized by associating these two features together. This association can provide complementary functions in fields where the other is lacking, which can improve the performance of gait recognition. Since local features are more effective than global features in gait recognition, in this paper a new temporal pooling function is specially proposed to create the local short-range features. In addition, different from most methods, in our method we deem the extracted semantic key-point features as a set of word embeddings, and a transformer encoder is proposed to build the dependence of each correlative key-points. Related experiments on CASIA Gait Dataset B verify that the proposed method outperforms other gait recognition methods for settling the cloth-changing problem.

References

[1] B. Abdolahi and N. Gheissari, "Gait recognition using dynamic texture descriptors," *2012 2nd International eConference on Computer and Knowledge Engineering (ICCKE)*, pp. 6–11, 2012.

[2] M. Altab 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.

[3] B. AmirH.Kargar, A. Mollahosseini, T. Struemph, W. Pace, R. D. Nielsen, and M. Mahoor, "Automatic measurement of physical mobility in get-up-and-go test using kinect sensor," *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 3492–3495, 2014.

[4] 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.

[5] I. Bouchrika and M. Nixon, "Exploratory factor analysis of gait recognition," 2008, pp. 1–6.

[6] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, pp. 172–186, 2021.

[7] H. Chao, Y. He, J. Zhang, and J. Feng, "Gaitset: Regarding gait as a set for cross-view gait recognition," in *AAAI*, 2019.

[8] 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.

[9] B. Cheng, B. Xiao, J. Wang, H. Shi, T. Huang, and L. Zhang, "Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5385–5394, 2020.

1051-8215 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Northeastern University. Downloaded on September 24,2021 at 01:33:35 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2021.3112564, IEEE Transactions on Circuits and Systems for Video Technology

[10] R. Child, S. Gray, A. Radford, and I. Sutskever, "Generating long sequences with sparse transformers," *ArXiv*, vol. abs/1904.10509, 2019.

[11] D. Cunado, M. Nixon, and J. Carter, "Using gait as a biometric, via phase-weighted magnitude spectra," in *Proceedings of 1st Int. Conf. on Audio- and Video-Based Biometric Person Authentication*, 1997.

[12] W. T. Dempster and G. Gaughran, "Properties of body segments based on size and weight," *American Journal of Anatomy*, vol. 120, pp. 33–54, 1967.

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

[14] M. Deng and C. Wang, "Human gait recognition based on deterministic learning and data stream of microsoft kinect," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, pp. 3636–3645, 2019.

[15] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *ArXiv*, vol. abs/2010.11929, 2020.

[16] C. Fan, Y. Peng, C. Cao, X. Liu, S. Hou, J.-N. Chi, Y. Huang, Q. Li, and Z.-Q. 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.

[17] H. Fang, S. Xie, Y.-W. Tai, and C. Lu, "Rmpe: Regional multi-person pose estimation," *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2353–2362, 2017.

[18] Y. Fu, Y. Wei, Y. Zhou, H. Shi, G. Huang, X. Wang, Z. Yao, and T. Huang, "Horizontal pyramid matching for person re-identification," in *AAAI*, 2019.

[19] 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.

[20] 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, pp. 102–113, 2019.

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

[22] S. Hou, C. Cao, X. Liu, and Y. Huang, "Gait lateral network: Learning discriminative and compact representations for gait recognition," in *ECCV*, 2020.

[23] 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.

[24] H. Hu, J. Gu, Z. Zhang, J. Dai, and Y. Wei, "Relation networks for object detection," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3588–3597, 2018.

[25] G. Huang, Z. Lu, C.-M. Pun, and L. Cheng, "Flexible gait recognition based on flow regulation of local features between key frames," *IEEE Access*, vol. 8, pp. 75 381–75 392, 2020.

[26] M. S. Islam, A. Matin, J. Paul, M. Rokanujjaman, and M. Altab Hossain, "A new effective part selection approach for part-based gait recognition," in *16th Int'l Conf. Computer and Information Technology*, March 2014, pp. 181–184.

[27] Y. Kim, C. Denton, L. Hoang, and A. M. Rush, "Structured attention networks," *ArXiv*, vol. abs/1702.00887, 2017.

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

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

[30] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, "Visualbert: A simple and performant baseline for vision and language," *ArXiv*, vol. abs/1908.03557, 2019.

[31] X. Li, Y. Makihara, C. Xu, Y. Yagi, and M. Ren, "Joint intensity transformer network for gait recognition robust against clothing and carrying status," *IEEE Transactions on Information Forensics and Security*, vol. 14, pp. 3102–3115, 2019.

[32] ——, "Gait recognition via semi-supervised disentangled representation learning to identity and covariate features," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13 306–13 316, 2020.

[33] X. Li, Y. Makihara, C. Xu, Y. Yagi, S. Yu, and M. Ren, "End-to-end model-based gait recognition," in *ACCV*, 2020.

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

[35] K. Liu and J. Yang, "Recognition of people reoccurrences using bag-of-features representation and support vector machine," *2009 Chinese Conference on Pattern Recognition*, pp. 1–5, 2009.

[36] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," *ArXiv*, vol. abs/2103.14030, 2021.

[37] S. Lombardi, K. Nishino, Y. Makihara, and Y. Yagi, "Two-point gait: Decoupling gait from body shape," *2013 IEEE International Conference on Computer Vision*, pp. 1041–1048, 2013.

[38] A. Matin, J. Paul, and T. Sayeed, "Segment based co-factor detection and elimination for effective gait recognition," in *2017 IEEE International Conference on Imaging, Vision Pattern Recognition*, 2017, pp. 1–5.

[39] B. Mohamed, S. Mohamed, R. Tlemsani, and L. Mostefai, "Gait recognition based on model-based methods and deep belief networks," *Int. J. Biom.*, vol. 8, pp. 237–253, 2016.

[40] A. Nandy, R. Chakraborty, and P. Chakraborty, "Cloth invariant gait recognition using pooled segmented statistical features," *Neurocomputing*, vol. 191, pp. 117–140, 2016.

[41] N. Parmar, A. Vaswani, J. Uszkoreit, L. Kaiser, N. M. Shazeer, A. Ku, and D. Tran, "Image transformer," *ArXiv*, vol. abs/1802.05751, 2018.

[42] J. Qin, T. Luo, W. Shao, R. Chung, and K. Chow, "A bag-of-gait model for gait recognition," 2012.

[43] C. Raffel, N. M. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *ArXiv*, vol. abs/1910.10683, 2020.

[44] P. Ramachandran, N. Parmar, A. Vaswani, I. Bello, A. Levskaya, and J. Shlens, "Stand-alone self-attention in vision models," in *NeurIPS*, 2019.

[45] E. Rastegari and H. Ali, "A bag-of-words feature engineering approach for assessing health conditions using accelerometer data," *Smart Health*, vol. 16, p. 100116, 2020.

[46] I. Rida, N. Al-Máadeed, and S. Al-Maadeed, "Robust gait recognition: a comprehensive survey," *IET Biom.*, vol. 8, pp. 14–28, 2019.

[47] I. Rida, S. Al-Maadeed, and A. Bouridane, "Unsupervised feature selection method for improved human gait recognition," *2015 23rd European Signal Processing Conference (EUSIPCO)*, pp. 1128–1132, 2015.

[48] I. Rida, X. Jiang, and G. L. Marcialis, "Human body part selection by group lasso of motion for model-free gait recognition," *IEEE Signal Processing Letters*, vol. 23, pp. 154–158, 2016.

[49] I. Rida, N. A. Maadeed, G. L. Marcialis, A. Bouridane, R. Hérault, and G. Gasso, "Improved model-free gait recognition based on human body part," 2017.

[50] A. Roy, M. Saffar, A. Vaswani, and D. Grangier, "Efficient content-based sparse attention with routing transformers," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 53–68, 2021.

[51] A. Sepas-Moghaddam and A. Etemad, "View-invariant gait recognition with attentive recurrent learning of partial representations," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 3, pp. 124–137, 2021.

[52] A. Sepas-Moghaddam, S. Ghorbani, N. Troje, and A. Etemad, "Gait recognition using multi-scale partial representation transformation with capsules," *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 8045–8052, 2021.

[53] A. Sokolova and A. Konushin, "Pose-based deep gait recognition," *ArXiv*, vol. abs/1710.06512, 2019.

[54] Y. Tay, M. Dehghani, D. Bahri, and D. Metzler, "Efficient transformers: A survey," *ArXiv*, vol. abs/2009.06732, 2020.

[55] A. Vaswani, N. M. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *ArXiv*, vol. abs/1706.03762, 2017.

[56] C. Wang, 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*, vol. 34, pp. 2164–2176, 2012.

[57] G. Wang, S. Yang, H. Liu, Z. Wang, Y. Yang, S. Wang, G. Yu, E. Zhou, and J. Sun, "High-order information matters: Learning relation and topology for occluded person re-identification," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6448–6457, 2020.

[58] H. Wang, Y. Zhu, B. Green, H. Adam, A. Yuille, and L.-C. Chen, "Axial-deeplab: Stand-alone axial-attention for panoptic segmentation," in *ECCV*, 2020.

1051-8215 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Northeastern University. Downloaded on September 24,2021 at 01:33:35 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2021.3112564, IEEE Transactions on Circuits and Systems for Video Technology

[59] L. Wang, H. Ning, T. Tan, and W. Hu, "Fusion of static and dynamic body biometrics for gait recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 2, pp. 149–158, 2004.

[60] S. Wang, B. Z. Li, M. Khabsa, H. Fang, and H. Ma, "Linformer: Self-attention with linear complexity," *ArXiv*, vol. abs/2006.04768, 2020.

[61] Y. Wang, X. Zhang, Y. Shen, B. Du, G. Zhao, L. C. C. Lizhen, and H. Wen, "Event-stream representation for human gaits identification using deep neural networks." *IEEE transactions on pattern analysis and machine intelligence*, vol. PP, 2021.

[62] Y. Wang, Z. Xu, X. Wang, C. Shen, B. Cheng, H. Shen, and H. Xia, "End-to-end video instance segmentation with transformers," *ArXiv*, vol. abs/2011.14503, 2020.

[63] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4724–4732, 2016.

[64] D. A. Winter, *Biomechanics and motor control of human movement*. John Wiley & Sons, 2009.

[65] B. Wu, C. Xu, X. Dai, A. Wan, P. Zhang, M. Tomizuka, K. Keutzer, and P. Vajda, "Visual transformers: Token-based image representation and processing for computer vision," *ArXiv*, vol. abs/2006.03677, 2020.

[66] 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, no. 2, pp. 209–226, 2017.

[67] C. Xu, Y. Makihara, X. Li, Y. Yagi, and J. Lu, "Cross-view gait recognition using pairwise spatial transformer networks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, pp. 260–274, 2021.

[68] C. Yam and M. Nixon, "Model-based gait recognition," 2009, pp. 633–639.

[69] 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 (ICPR)*, pp. 2057–2064, 2021.

[70] 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.

[71] 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.

[72] S. Yu, D. Tan, and T. Tan, "A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition," *18th International Conference on Pattern Recognition (ICPR'06)*, vol. 4, pp. 441–444, 2006.

[73] 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.

[74] Z. Zhang, L. Tran, X. Yin, Y. Atoum, X. Liu, J. Wan, and N. Wang, "Gait recognition via disentangled representation learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4710–4719.

[75] S. Zheng, J. Zhang, K. Huang, R. He, and T. Tan, "Robust view transformation model for gait recognition," in *2011 18th IEEE International Conference on Image Processing*, 2011, pp. 2073–2076.

[76] Y. Zhou, Y. Huang, Q. Hu, and L. Wang, "Kernel-based semantic hashing for gait retrieval," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, pp. 2742–2752, 2018.

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 re-identification 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.

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.

Jingsong Xu received the B.S. degree in computer science and the Ph.D. degree in pattern recognition from the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China, in 2007 and 2014, respectively. He is currently a Research Fellow with the Global Big Data Technologies Center, University of Technology Sydney, Ultimo, NSW, Australia. His research interests include computer vision, pattern recognition, and machine learning.

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.

1051-8215 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Northeastern University. Downloaded on September 24,2021 at 01:33:35 UTC from IEEE Xplore. Restrictions apply.