Flexible Gait Recognition Based on Flow Regulation of Local Features Between Key Frames

GUOHENG HUANG1, (Member, IEEE), ZENG LU1, (Graduate Student Member, IEEE), CHI-MAN PUN2, (Senior Member, IEEE), AND LIANGLUN CHENG1, (Senior Member, IEEE)

1School of Computers, Guangdong University of Technology, Guangzhou 510006, China
2Department of Computer and Information Science, University of Macau, Macau 999078, China
Corresponding author: Zeng Lu (zenglu@mail2.gdut.edu.cn)

This work was supported in part by the National Nature Science Foundation of China under Grant 61702111, in part by the National Natural Science Foundation of Guangdong Joint Fund under Grant U1801263 and Grant U1701262, in part by the Guangdong Provincial Key Laboratory of Cyber-Physical System under Grant 2016B030301008, in part by the Guangdong Research and Development plan projects in key areas under Grant 2018B010109007, and in part by the “Blue Fire Plan” (Huizhou) Industry-University-Research Joint Innovation Fund 2017 Project of the Ministry of Education under Grant CXZJHZ201730.

ABSTRACT The information contained in gait frames is different, and the contribution of different frames to recognition tasks is also different. However, each frame has the same degree of attention in the input layer, this prevents the network from focusing on keyframes. Therefore, we propose a keyframe extraction module via information weighting, make network can pay more attention to the high contribution frame at the input layer, and the extraction of the distinctive features is improved. Moreover, the range of motion in different parts of the human body is different, the temporal and spatial correlation of local feature between silhouettes is different. Based on the discovery, we propose a Local Features Flow Regulation module to calculate the correlation coefficient of the local features of each silhouette, and the regulation coefficient is generated by the correlation coefficient. The regulation coefficient is applied to regulate the flow of local features, this enables the network to capture areas with more spatial and temporal features. Through the extraction of frame-level features and the interaction of local features between frames, the network can extract the most discriminative features from global to local flexibly. During training, each horizontal part is trained separately. The training can adjust the regulation coefficients, and the network is more flexible and expressive. Our model has a good performance on cross-viewing and complex environments of CASIA-B dataset. In the case of normal environment and complex environment (pedestrian with backpacks and in coats), the rank-1 of the proposed model is 95.1%, 87.9% and 74.0% respectively, higher than state-of-the-art.

INDEX TERMS Gait recognition, key frame extraction, local features flow regulation, attention mechanism.

I. INTRODUCTION
Gait is a non-cooperative biometric feature, and the gait can be used to identify the identity of a pedestrian within a few tens of meters. This biometric is difficult to disguise and does not require subject cooperation when tested. However, due to changes in perspectives, changes in pedestrian clothing and environmental covariance, gait recognition still faces many problems.

In the early stages, traditional gait recognition methods are built on the template [1]–[6]. This kind of methods generate gait templates based on mathematical geometry or other principles, and a template is utilized to extract features from the gait sequence to produce an image for gait recognition. The template-based method extracts the features of the gait directly and eliminates some covariance to improve recognition accuracy. But it is difficult for the gait template in fitting, and the template is not flexible enough. Therefore, performance is significantly reduced in the case of cross-view recognition. In contrast, the silhouette-based gait recognition methods do not assume certain model of the
human subject, but analyze the spatio-temporal shape and motion features of the silhouette images which are extracted from the sequences [7]. In addition, Bayesian probability and time-frequency signals of gait are applied by researchers to verify gait [8], [9]. These are two types of very effective representations of gait. However, neither the time-frequency representation nor the probability can be easily integrated into the neural network. In contrast, the Euclidean distance or cosine similarity is easily calculated by matrix calculation in the neural network, which is convenient for training in the network.

Gait recognition method based on view transform is proposed to convert gait sequences from different views to the same view [10], or project sequences of gaits at different views onto a plane independent of the views [11], or use the correlation between gait sequences from different views [12]. Finally, the identification match is exploited in the case of the same view. The problem of cross-view recognition is solved to some extent by the view transform model. However, the view transform model does not pay much attention to the time and fine-grained spatial information of the gait sequences. Therefore, it causes the model not to adapt to complex environments.

In recent years, deep learning has been extensively used in various fields of computer vision. Convolutional neural network is particularly excellent in terms of image feature extraction. Non-linear mapping of multiple convolutional layers can adaptively extract features independent of views [13]–[15], the extracted features have strong expressiveness and generalization. The deep-learning method achieve a 95% average cross-view average rank-1 accuracy under normal walking conditions. The Gait Energy Image (GEIs) [16] and the gait silhouette sequences are used as input by many deep learning models to extract features directly from a backbone network. This input makes each frame equally important to the network, without considering the importance of each frame, and does not pay attention to the spatiotemporal correlation locally on the gait silhouette sequences. Long and Short-Term Memory networks (LSTM) and 3 Dimensional Convolutional Neural Networks (3D-CNN) are utilized to extract temporal features of gait sequences [14], [17]. However, it is low-efficient in computation and difficult to train for these networks. Moreover, current gait recognition methods based on deep learning mainly extract global features from each image in the sequence. These network structures have been unable to capture the correlation and spatiotemporal information of local features between frames. The information in the gait sequence is not fully utilized.

Through the observation of human walking, we notice that the change range of each horizontal part of human body is different. This shows that the gait sequence of each horizontal part of the human body contains different spatiotemporal information. Some parts provide a lot of spatiotemporal information, while some parts provide little spatiotemporal information. For example, when walking, the feet swing the most, while the head is usually still. Therefore, the spatiotemporal information on feet will be more helpful for gait recognition.

The current gait recognition methods do not consider the screening of input data. And, there is also a lack of methods for analyzing the correlation between local features of the gait contour frame and the extraction of local features of the gait contour. In this paper, based on the above motivation and current research status of gait recognition, we propose a flexible and simple network structure. with three contributions:

1) In order to increase the contribution of key frames to feature extraction, the gait silhouette sequence in the network is readjusted by using the attention mechanism based on the information it contains.
2) For the gait silhouette of each frame, the gait feature is made with a uniform horizontal division. After obtaining the correlation between the local features of the frames and the features of the corresponding parts of other frames, the correlation coefficients are used to regulate the flow of fine-grained local features in each frame. The features of reorganization will be more flexible and have better expressiveness. The model will be more stable after a lot of sample training.
3) The proposed method has achieved state of the art on CASIA-B dataset. The rank-1 of the proposed model is 95.1%, 87.9% and 74.0% on NM, BG and CL respectively.

The proposed network can select key gait frames through each frame of information, and can further capture the correlation and spatiotemporal information of local features between frames. Finally extract more discriminative features.

The remaining parts of this paper are organized as follows: a review of related theories applied in the proposed method is presented in Section 2. in Section 3, we elaborated on the principles of each module in our network. In Section 4, we do a lot of exploratory experiments, comparative experiments and ablation experiments to verify the effectiveness, stability and generalization ability of the network. In Section 5, we summarized the work of the paper.

II. RELATED WORK
The traditional methods of gait recognition usually meet with the problem of underfitting, which cannot fully lay out the key information in the gait sequences. The traditional method cannot be applied to the complex scene. In this paper, we mainly refer to the method based on deep learning.

A. EXTRACTION OF KEY INFORMATION IN SEQUENCE DATA
Video person re-identification and gait recognition are similar, it is also used to identify people from image sequences. Recently, the mainstream method of video person re-identification is to extract a feature vector from each frame of a video by convolutional neural network [18]. Then these features are sent to Recurrent Neural Network (RNN) in order to extract temporal features. Finally, network make a max
pooling or mean pooling feature vectors for all frames to represent the feature vector of the video sequence. In 2018, Song et al. proposed the region-based quality estimation network [20] to assign high weights to images without pedestrians being obscured. This work selected the images containing more useful information and improves the accuracy of identification. It is proved that the selection of key information is helpful for identification.

The attention mechanism is a very helpful mechanism for deep learning. In 2017, Hu et al. proposed Squeeze-and-Excitation Networks (SENet), which improved the accuracy by modeling the correlation between feature channels and strengthening key feature maps [19]. Yang et al. proposed hierarchical attention networks for the document class [21]. This work demonstrated the effectiveness of attention mechanism in long text classification. Inspired by their work, we propose attention-based key frame extraction based on information weighting in gait sequences, that simply and efficiently extract the key frames with this attention mechanism.

B. SPATIOTEMPORAL INFORMATION EXTRACTION BETWEEN SEQUENCE FRAMES

In 2006, Han et al. proposed the use of Gait Energy Images (GEIs) [16] to recognize human identities. The representation of the average silhouette of gait sequences is used as input to various network models. The gait energy image retains some of the temporal motion features of gait sequences, it can save a lot of computing resources that input GEIs as images into the network. However, it should lose a lot of spatio-temporal information between frames. GEIs are used as input to many gait recognition networks based on deep learning.

Since 2016, convolutional neural networks had been widely used in gait recognition. Zhang et al. proposed a representation of gait characteristics based on convolutional networks named Deep Gait [22]. Shiraga et al. proposed GEINet which used GEIs as input to convolutional networks to extract features [13]. Wu et al. extensively studied the gait recognition method across views [14]. Deep convolutional network-based method took advantage of the highly nonlinear mapping of convolutional layers and plenty of trainable parameters. It was possible to learn the stability features that overcome the view changes and environmental impact. The gait recognition network based on deep learning showed good performance in the case of multiple views and environments, which had gone beyond the traditional gait recognition algorithm.

Gait sequences and GEIs [16] are used as inputs to many deep learning-based gait recognition networks. There is fewer spatial-temporal information contained in GEIs. In the method of using the gait sequence as the input of the network. The gait silhouette sequences need to be input into the LSTM [14] or 3D-CNN [17] in order. It took a lot of computation time and it is difficult to train the network. Moreover, these methods only extracted the spatio-temporal information of the global features of the images between frames, but do not captured the spatio-temporal information of the local features.

In 2019, Chao et al. proposed GaitSet [15]. The authors assume that each frame contains its ordinal information in the sequential gait sequence. GaitSet do not take a gait sequence as input. GaitSet is invoked as a network input in the form of a Set, which simplify the network structure, speed up network computing, and facilitate the embedding of new modules. Images are randomly extract from the gait sequence to form a Set and input into the network. Backbone of GaitSet only consist of convolutional layers. The Set is specially processed when it is input into the convolutional layer and when it is output from the convolutional layer, the specific operation is reshaping of the tensor of the Set. Inspired by GaitSet, our model also uses the Set as input. GaitSet also proposed a method of aggregating all gait frames into a single image. This method is named as Set Pooling.

C. METHODS OF HUMAN BODY FEATURE DIVISION

Region of Interest (RoI) is generally used in object detection, and was often used in Re-id to divide global features to obtain local features [23], [24]. However, these methods generate a large number of unnecessary proposals, which affect the speed of the calculations.

In 2018, Sun et al. proposed a method of evenly dividing the person global features in the horizontal direction and then separately training every part [25]. This method is simple, just horizontally divide the pedestrian’s global feature map to extract local features. The method has achieved good results. In our method, this kind of feature division method is adopted, and the division method is also used in the training part of our network. This causes the feature blocks at the end of the network to be in contact with the feature blocks in the middle of the network.

III. METHOD

In this Section, we describe the main modules we have proposed in the global network. The Set of gait silhouettes is a four-dimensional tensor. The number of Set frames in the second dimension of the tensor. This tensor is input into the network for forward propagation. In the process of forward propagation, the feature map Set is input to the key frame extraction module to retrieve key frames. The feature map Set is subsequently sent to the proposed Local Feature Flow Regulation module for flow regulation. Then, the feature map Set is subjected to aggregation operation after multiple down samplings in the convolution. The final feature is used to calculate multiple loss functions with joint training. The architecture of the global network is shown as FIGURE 1.

A. KEY FRAME EXTRACTION BASED ON INFORMATION WEIGHTING

If the gait silhouettes are randomly composed of a Set of frames images as the input data to the network. The extraction of data is not targeted, resulting in some unidentifiable and highly similar silhouettes being extracted into the Set.
FIGURE 1. The structure of global network: The overall structure of the network is simple. The network is mainly composed of a convolutional layer and a pooling layer. There are two main modules in the network, key frame extraction module and local features flow regulation module.

FIGURE 2. The structure of key frame extraction: the orange cube represents the feature map for each frame in the Set. The purple rectangle represents the fully connected layer. The red circle indicates that the key score corresponding to each frame is calculated. This is detrimental to the extraction of temporal and spatial features, and it will affect the training effect of the network and the recognition effect of the final model. To solve these problems, we have proposed a key frame extraction module based on information weighting. The feature map is extracted for each frame of the image in the network to ensure that the network is end-to-end. Using the frame-level attention mechanism, the key score for each frame is predicted and weighted onto the original feature map.

The feature map Set \( F \) is first reshaped and then input into the key frame extraction module. \( F \) is performed to reduce the dimensionality of the convolutional layer and output \( F^* \) after the max pooling layer. After being expanded, the feature map for each frame is input into the same fully connected layer to calculate the critical score \( S \).

\[
\begin{align*}
F^* &= \text{maxpool}(W_{k,c}F) \\
S_i &= W_{k,fc1}(W_{k,fc1}F_i^* + b_{k,fc1}) + b_{k,fc2} \\
S_i &= \frac{\exp(s_i)}{\sum_{j=1}^{\text{framenum}} \exp(s_j)}
\end{align*}
\]

where \( W_{k,c}, W_{k,fc1} \) and \( W_{k,fc2} \) are the parameters of the convolutional layer, the first fully connected layer, and the second fully connected layer, respectively. \( b_{k,fc1} \) and \( b_{k,fc2} \) are the bias parameters of the two fully connected layers. \( s_i \) is the key score for each frame, \( i \in \{1, 2, 3, ..., \text{framenum}\} \). \( s_i \) is executed a probabilistic operation, getting \( S = [s_1, s_2, ..., s_{\text{framenum}}] \). To make each key score in \( S \) be multiplied by each frame feature map in \( F \), need to expand the tensor and change the shape of \( S \),

\[
F_{\text{key}} = \text{reshape}(S) \cdot F
\]

B. LOCAL FEATURES FLOW REGULATION

The purpose of the key frame extraction module is to extract features at the frame level. The purpose of the Local Features Flow Regulation module is to regulate the local features between frames. Further, based on the frame level, more subtle, the module is a local level of feature regulation.

We note that the correlation between the corresponding local features of the frame can be utilized to calculate the correlation coefficient to regulate the local features of each frame feature. The correlation between frames is measured by the projection between corresponding local features. \( \text{Pr}_a \) is a self-projection of a local feature vector \( a \) on this vector \( a \). \( \text{Pr}_b \) is a projection of a local feature vector \( a \) on corresponding local feature vector \( b \) of other frames. If the Euclidean distance between the two projections is large, the correlation between \( b \) and \( a \) is small. On the contrary, it shows that the correlation between these two vectors is high. The information contained in the randomly extracted \( N \) images is limited, and it is necessary to extract the most discriminative features from the local parts of each frame. We propose the Local Features Flow Regulation, which uses the correlation between the local features of the frames to regulate the flow of local features in each frame, and then extracts the most discriminative features in each frame.

C. CORRELATION MATRIX CORRESPONDING TO LOCAL FEATURES BETWEEN FRAMES

The feature maps of all frames are horizontally divided into \( \text{stripenum \, stripes} \), each of which has a correlation matrix. In the process of calculating the correlation matrix, the tensor shape of the features input into the module has 4 dimensions,
The above equations are essentially to obtain cosine similarity of two identical feature vectors. For instance, we use $a_i^m$ to indicate $f_{m,j}^i$. $x_m^i$ is the unit vector. Therefore, $x_m^i \cdot x_n^i = \|x_m^i\| \|x_n^i\| \cos \theta = \cos \theta = \sum_{j=1}^{\text{featurenum}} f_{m,j}^i$.

$A$ represents the self-projection vector, then use $A$ to calculate the self-projection matrix $\overline{A}$. The transposes of $A$ and $A$ are added together to produce a self-projection matrix. The transposes of $A$ and $A$ is represented by the Equation 8 and 9, and $\overline{A}$ is the self-projection matrix. For the convenience of description, in the Equation 10, we use $a_m^i$ instead of $\sum_{j=1}^{\text{featurenum}} f_{m,j}^i$, $m$ indicates $m$-th frame.

\[
A = \left[\begin{array}{c}
(a_1^1, a_2^1, \ldots, a_{\text{framenum}}^1)\\
(a_1^2, a_2^2, \ldots, a_{\text{framenum}}^2)\\
\vdots\\
(a_1^n, a_2^n, \ldots, a_{\text{framenum}}^n)
\end{array}\right]
\]

(8)

\[
A^T = \left[\begin{array}{c}
(a_1^1, a_2^1, \ldots, a_{\text{framenum}}^1)\\
(a_1^2, a_2^2, \ldots, a_{\text{framenum}}^2)\\
\vdots\\
(a_1^n, a_2^n, \ldots, a_{\text{framenum}}^n)
\end{array}\right]
\]

(9)

\[
\overline{A} = A + A^T
\]

(10)

Second, we calculate the mutual projection matrix of the same stripe between frames, and use the following equations to calculate. The multiplication of 2 in equation (11) is explained by the addition of two matrices in equation (10).

In essence, it is to get cosine similarity of two different feature vectors. For instance, $x_m^i$ and $x_n^i$ represent the feature vector of the $i$-th stripe on $m$-th frame and $n$-th frame respectively, they are all unit vectors. $x_m^i \cdot x_n^i = \|x_m^i\| \|x_n^i\| \cos \theta = \cos \theta$, $\cos \theta$ is the mutual-projection of $x_m^i$ and $x_n^i$.

For ease of description, we assume that there are 3 frames, $\overline{\alpha}$, $\overline{\beta}$ and $\overline{\chi}$ are used to indicate the feature vectors of the $i$-th stripe of these 3 frames, respectively. Assume that the length of the feature vectors is 3. The specific calculation process is described in the example of equation (11). $\overline{AB}$ represents the mutual-projection matrix of the $i$-th stripe between frames. $\overline{AB}$ is generated by multiplying the $\overline{f}$ with the transposed matrix of $\overline{f}$.

\[
\overline{AB} = 2 \cdot \overline{f} \times \overline{f}^T
\]

\[
e.g. \quad 2 \cdot \begin{pmatrix}
\alpha_1 & \alpha_2 & \alpha_3 \\
\beta_1 & \beta_2 & \beta_3 \\
\chi_1 & \chi_2 & \chi_3
\end{pmatrix}
\times
\begin{pmatrix}
\alpha_1 & \beta_1 & \chi_1 \\
\alpha_2 & \beta_2 & \chi_2 \\
\alpha_3 & \beta_3 & \chi_3
\end{pmatrix}
\]

\[
= \begin{pmatrix}
2 \overline{\alpha} \cdot \overline{\alpha} & 2 \overline{\alpha} \cdot \overline{\beta} & 2 \overline{\alpha} \cdot \overline{\chi} \\
2 \overline{\beta} \cdot \overline{\alpha} & 2 \overline{\beta} \cdot \overline{\beta} & 2 \overline{\beta} \cdot \overline{\chi} \\
2 \overline{\chi} \cdot \overline{\alpha} & 2 \overline{\chi} \cdot \overline{\beta} & 2 \overline{\chi} \cdot \overline{\chi}
\end{pmatrix}
\]

(11)

Finally, subtract the two matrices to get the correlation matrix of the $i$-th stripe. The self-projection matrix $\overline{A} \overline{A}$ is subtracted from the mutual-projection matrix $\overline{A} \overline{B}$ to generate a correlation matrix $M$. The element in $M$ represents the self-projection of the feature vector of the $i$-th stripe in a frame minus the mutual projection of the feature vector of the $i$-th stripe in this frame and the feature vector of

\[
\overline{AB} - \overline{A} \overline{A}
\]

\[
= \begin{pmatrix}
A_1 - M_1 & A_2 - M_2 & A_3 - M_3 \\
\vdots & \vdots & \vdots
\end{pmatrix}
\]

\[
= \begin{pmatrix}
\overline{f}_1 \cdot f_1^1 & \overline{f}_1 \cdot f_1^2 & \overline{f}_1 \cdot f_1^3 \\
\vdots & \vdots & \vdots
\end{pmatrix}
\]

\[
= \begin{pmatrix}
\overline{f}_1 \cdot f_1^1 & \overline{f}_1 \cdot f_1^2 & \overline{f}_1 \cdot f_1^3 \\
\vdots & \vdots & \vdots
\end{pmatrix}
\]
the i-th stripe in other frames, which is the difference between the self-projection and the mutual projection. Self-projection and mutual projection are measured by cosine similarity.

\[ M_i = \overline{A}A - \overline{A}B \]  

(12)

The value of the self-projection is greater than or equal to the value of the mutual projection. Therefore, the value in \( M_i \) is greater than or equal to 0. In the actual calculation, correlation matrices of all the stripes are calculated simultaneously and are denoted as \( M \).

D. FLOW REGULATION COEFFICIENT DETERMINATION

The flow regulation coefficients of the local features are calculated using the correlation matrix \( M \). The purpose of flow regulation is to extract features of the stripes with the most frequent changes between frames and enhance the features of these stripes, thereby improving the extraction efficiency of spatial and temporal information. The goal is to extract the most discriminative feature of gait set.

For the convenience of the description, we still explain the calculation process of the regulation coefficient on the i-th stripe. The main principle of calculating the flow regulation coefficient is to weaken the features on the stripe with similarity correlation and enhance the features on the stripe with weak similarity in all frames. The relevant factors are measured using three different types of projections, they are the maximum projection difference \( \text{max}_D_i \), the minimum projection difference \( \text{min}_D_i \) and the mean projection difference \( \text{mean}_D_i \). \( \text{max}_D_i \), \( \text{min}_D_i \) and \( \text{mean}_D_i \) are all calculated by \( M \). \( \text{max}_D_i \) is the largest value of \( M \), \( \text{min}_D_i \) is the smallest value of \( M \), \( \text{mean}_D_i \) is the mean value of all elements in the lower middle triangular matrix of \( M \).

The calculation of the regulation coefficient is not simply determined by the three projection differences. \( w_i \) is set to the adaptive parameter of the i-th stripe so that the calculation of the regulation coefficient has adaptability. The calculation process of the regulation coefficient is as follows.

\[ c_i = \exp \left( -\frac{w_i^2}{\text{max}_D_i + \text{min}_D_i + \text{mean}_D_i} \right) \]  

(13)

where \( w_i \) is a numerator and used to adjust the generation of the regulation coefficient \( c_i \). When the denominator is too large or too small, \( w_i \) will not make \( c_i \) too small or too large. \( w_i \) increase network adaptability and avoid gradient explosion and gradient disappearance. In the above Equation 13, \( c_i \) increases as \( \text{max}_D_i \), \( \text{min}_D_i \) and \( \text{mean}_D_i \) increases. When \( \text{max}_D_i \), \( \text{min}_D_i \) and \( \text{mean}_D_i \) are large, it means that the features on the i-th stripe between each frame have great differences. Pedestrians have more motion information on the i-th stripe. In this way, the feature of the i-th stripe will be enhanced by \( c_i \).

After the above calculation, the flow regulation coefficients of all stripe are obtained, \( c = [c_1, c_2, \ldots, c_{\text{stripenum}}] \). \( h_f \) is the number of rows of the feature map of each frame in \( f \).

The feature map of each frame in \( f \) is divided into \( \text{stripenum} \) stripes, each stripe corresponds to a control coefficient, each stripe contains \( h_f/\text{stripenum} \) rows. In order to make the control coefficient one-to-one with each stripe in \( f \) rows of each frame, \( c_i \) is performed to perform the copy operation while ensuring that the order of each regulation coefficient is unchanged. \( c_i \) is copied \( h_f/\text{stripenum} \) times.

\[ c = [c_1, \ldots, c_1, c_2, \ldots, c_2, c_{\text{stripenum}}, \ldots, c_{\text{stripenum}}] \]  

(14)

where \( c \) is executed some necessary shape changes. Then multiply \( c \) and \( f \) to control the feature flow of each stripe in \( f \).

\[ f_{\text{flow}} = \text{reshape}(c) \cdot f \]  

(15)

E. ASYMMETRIC CORRESPONDENCE OF LOCAL FEATURES

After the Set Pooling module, the feature Set \( g \) is generated and will be horizontally divided into \( \text{partnum} \) regions. This way of dividing features is the same as the way of dividing features in the Local Features Flow Regulation module, it is only different in the number of divisions. The horizontal division method makes a local correspondence between the two position features, and this correspondence can be adjusted by changing \( \text{stripenum} \) and \( \text{partnum} \). Due to the flexibility of this correspondence, it is named as asymmetric correspondence. It has been shown in FIGURE 4.

When \( \text{partnum} \) is less than \( \text{stripenum} \), a part in \( g \) corresponds to multiple stripes in \( f_{\text{flow}} \). When \( \text{partnum} \) is greater than \( \text{stripenum} \), multiple parts in \( g \) correspond to the same stripe in \( f_{\text{flow}} \). When \( \text{partnum} \) is equal to \( \text{stripenum} \), a part in \( g \) corresponds to a stripe in \( f_{\text{flow}} \). This relationship is not an integer number of integers in most cases.

In the training part of the network, a joint training method of classification and metric is adopted. Train each part separately. There are \( \text{partnum} \) parts in \( g \), and a classifier will be trained separately for each part. The training of metric is the same as classification, each part has independent parameters. The asymmetric correspondence between \( f_{\text{flow}} \) and \( g \) makes each part to communicate with each other during training. This fuzzy correspondence improves the flexibility.
and stability of the network, as well as the ability to express discriminative features.

F. JOINT TRAINING OF MULTIPLE LOSS FUNCTIONS

The network uses joint training of classification loss and metric loss in the training process to make full use of the only category labels in the dataset. Joint training allows the network model to correctly classify the gaits while expanding the distance between classes and reducing the distance within the class, improving the stability and robustness of the model.

Refer to the setting of the loss function in Beyond Part Models [25] and GaitSet [15], we calculate separate metric loss [31], [32] for each part. Then, add the loss of each part to form total loss.

\[
\text{totalloss} = \sum_{i=1}^{\text{partnum}} \text{claloss}(g_i) + \sum_{i=1}^{\text{partnum}} \text{tripleloss}(g_i) \tag{16}
\]

IV. EXPERIMENT

A. DATASETS AND TRAINING DETAILS

We conduct the experiments on the CASIA-B dataset [26]. The CASIA-B dataset contains 124 subjects. Each subject contains 6 kinds of conditions of normal walking (NM), 2 kinds of conditions of walking with a bag (BG), 2 kinds of conditions of walking with coat (CL). Persons walking with a bag will have some interference with gait recognition. Persons wearing coats can play a good camouflage effect and can be used to cover large gait postures. There are 11 different views (0°, 18°, ..., 180°) for each walking condition, the spacing between two adjacent views is 18°. There are 110 gait sequences in each subject. CASIA-B contains various views and includes three different walking conditions, which are very challenging and not available in many other gait datasets.

The input to the network is a set of aligned silhouettes. The alignment here is built on the method proposed by Takemura et al. [27]. A patch of size 64 × 44 is randomly sampled from the training set, P represents P different subjects and N represents that each subject contains N gait sequences. In the experiment, P is set to 6, K is set to 12, and there are 72 gait sequences in a batch. 30 silhouettes are randomly extracted from each sequence to constitute each set. A batch contains 72 sets. FIGURE 5 shows the gait data of three different conditions.

Because this dataset is not officially divided into training sets and test sets, we use the popular divisions in the current papers. Mainly referring to the division method in GaitSet [15], we conducted small sample training (ST), medium sample training (MT) and large sample training (LT). In ST, the first 24 subjects (marked in 001-024) are used for training and the remaining 100 subjects were left for testing. In MT, the first 62 subjects were used for training and the remaining 62 subjects were left for testing. In LT, the first 74 subjects are used for training and the remaining 50 subjects are left for testing. In the test sets of all three settings, the first 4 sequences of the NM condition (NM #1-4) are kept in the gallery, and the rest 6 sequences are divided into 3 probe subsets, i.e. NM subsets containing NM #5-6, BG subsets containing BG #1-2 and CL subsets containing CL #1-2.

In testing, a sample q in each probe set Q is input into the network to produce a final representation \( F_q \). All samples in Gallery G are also subjected to the same operation to get any \( F_g \). Finally, using the Euclidean distance to compare the distance between \( F_q \) and each \( F_g \). Get rank-n as experimental results. The learning rate \( lr \) is set to 0.0001, and for small sample training (ST), we train 60,000 iterations. For medium sample training (MT), we trained 70,000 iterations. For large sample training (LT), we trained 80,000 iterations. All experimental results were averaged over 11 views and excluded from the same view. All experiments are performed on a GTX 1080Ti graphics card.

B. REGULATION OF BLOCK CORRESPONDENCE

We execute a large sample (LT) experiment on the CASIA-B dataset to explore the regulation of block correspondence. At the middle and end of the network, \( f_{f_{\text{flow}}} \) and \( g \) are respectively subjected to horizontal division operations. At the middle and end of the network, \( f_{f_{\text{flow}}} \) and \( g \) are respectively subjected to horizontal division operations. This operation makes them locally correspond to each other, and this correspondence can be controlled by changing \( \text{stripenum} \) and \( \text{partnum} \). The correspondence between partitions can be divided into three cases. The first is the case where \( \text{stripenum} \) is smaller than \( \text{partnum} \), the second is the case where \( \text{stripenum} \) is equal to \( \text{partnum} \), and the third is the case where \( \text{stripenum} \) is larger than \( \text{partnum} \). In FIGURE 6, the three different correspondences are shown.

In the experiment, we fixed \( \text{partnum} \) to 6 and only changed \( \text{stripenum} \) to achieve the purpose of adjusting the correspondence.
In order to comprehensively cover the above three cases in the experiment, $\text{stripenum} \in \{2, 4, 6, 8, 16, 32\}$. For each $\text{stripenum}$, we conducted comparative experiments under three different conditions (NM, BG and CL). Through the experiments, we found that different $\text{stripenum}$ will get different recognition accuracy rank-1.

In FIGURE 7, the dotted line diagram clearly describes the change of rank-1 with $\text{stripenum}$ in three different conditions. The accuracy curves of NM, BR and CL are at three different levels from high to low, but the trend of their changes is synchronous. Under the two conditions of NM and BG, the variation is small, while in the CL condition, the variation is larger. The rank-1 shows a trend of increasing first and then decreasing with the increase of $\text{stripenum}$, and peaking at the position set to 8. TABLE 1 shows the specific rank-1 of different $\text{stripenum}$ under three conditions.

In the case where $\text{stripenum}$ is less than $\text{partnum}$. One stripe contains multiple parts. In the process of back propagation of the single training of each $p$ at the end of the network, the coefficients and related parameters in the Local features flow regulation module is adjusted. But one stripe only has one flow regulation coefficient, and multiple parts training is used to control the regulation coefficients in one stripe. This makes it hard for adjacent stripes to be well connected, making the trained model not effective.

In the case where $\text{stripenum}$ is equal to $\text{partnum}$. The one-to-one correspondence between stripe and part makes the network inflexible.

The experiment proves that rank-1 is the best when $\text{stripenum}$ is larger than $\text{partnum}$. Especially when $\text{stripenum}$
TABLE 2. Contrast experiment results of rank-1 for small sample training (ST) on the CASIA-B dataset.

| Gallery NM#1-4 | Probe | 0° | 18° | 36° | 54° | 72° | 90° | 108° | 126° | 144° | 162° | 180° | mean |
|----------------|-------|----|-----|-----|-----|-----|-----|------|------|------|------|------|------|
| ST (24)        |       |    |     |     |     |     |     |      |      |      |      |      |      |
| NM#5-6 ViDP[11]| -     | -  | -   | 59.1| -   | 50.2| -   | 57.5 | -    | -    | -    | -    | -    |
| CMCC[28]       | 46.3  | -  | -   | 52.4| -   | 48.3| -   | 56.9 | -    | -    | -    | -    | -    |
| CNN-LB[14]    | 54.8  | -  | -   | 77.8| -   | 64.9| -   | 76.1 | -    | -    | -    | -    | -    |
| GaitSet[15]    | 64.6  | 83.3| 90.4| 86.5| 80.2| 75.5| 80.3| 86.0 | 87.1 | 81.4 | 59.6 | 79.5 |
| Ours           | 67.4  | 81.6| 88.8| 87.0| 80.7| 74.9| 79.2| 86.7 | 88.2 | 82.0 | 66.7 | 80.2 |
| BG#1-2 GaitSet[15]| 55.8 | 70.5| 76.9| 75.5| 69.7| 63.4| 68.0| 75.8 | 76.2 | 70.7 | 52.5 | 68.6 |
| Ours           | 57.8  | 70.6| 77.1| 76.2| 70.1| 64.3| 68.7| 76.0 | 75.4 | 70.3 | 54.6 | 69.2 |
| CL#1-2 GaitSet[15]| 29.4 | 43.1| 49.5| 48.7| 42.3| 40.3| 44.9| 47.4 | 43.0 | 35.7 | 25.6 | 40.9 |
| Ours           | 33.4  | 47.1| 53.1| 48.8| 46.1| 41.2| 47.4| 47.7 | 47.1 | 39.0 | 29.3 | 43.6 |

TABLE 3. Contrast experiment results of rank-1 for medium sample training (MT) on the CASIA-B dataset.

| Gallery NM#1-4 | Probe | 0° | 18° | 36° | 54° | 72° | 90° | 108° | 126° | 144° | 162° | 180° | mean |
|----------------|-------|----|-----|-----|-----|-----|-----|------|------|------|------|------|------|
| MT (62)        |       |    |     |     |     |     |     |      |      |      |      |      |      |
| NM#5-6 AE[29]  | 49.3  | 61.5| 64.4| 63.6| 63.7| 58.1| 59.9| 66.5 | 64.8 | 56.9 | 44.0 | 59.3 |
| MGAN[30]       | 54.9  | 65.9| 72.1| 74.8| 71.1| 65.7| 70.0| 75.6 | 76.2 | 68.6 | 53.8 | 68.1 |
| GaitSet[15]    | 86.8  | 95.2| 98.0| 94.5| 91.5| 89.1| 91.1| 95.0 | 97.4 | 93.7 | 80.2 | 92.0 |
| Ours           | 86.7  | 95.4| 97.8| 96.3| 91.6| 87.0| 91.4| 96.8 | 95.9 | 93.3 | 82.5 | 92.2 |
| BG#1-2 AE[29]  | 29.8  | 37.7| 39.2| 40.5| 43.8| 37.5| 43.0| 42.7 | 36.3 | 30.6 | 28.5 | 37.2 |
| MGAN[30]       | 48.5  | 58.5| 59.7| 58.0| 53.7| 49.8| 54.0| 61.3 | 59.5 | 55.9 | 43.1 | 54.7 |
| GaitSet[15]    | 79.9  | 89.8| 91.2| 86.7| 81.6| 76.7| 81.0| 88.2 | 90.3 | 88.5 | 73.0 | 84.3 |
| Ours           | 80.1  | 89.9| 91.3| 87.8| 84.0| 75.8| 81.1| 88.6 | 90.7 | 85.5 | 73.7 | 84.4 |
| CL#1-2 AE[29]  | 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[30]       | 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[15]    | 52.0  | 66.0| 72.8| 69.3| 63.1| 61.2| 63.5| 66.5 | 67.5 | 60.0 | 45.9 | 62.5 |
| Ours           | 58.3  | 71.1| 76.8| 71.5| 64.5| 58.9| 64.0| 68.5 | 68.8 | 59.5 | 49.1 | 64.6 |

is set to 8, one part in g will contain two stripes in f_flow. In this correspondence, by single training of each part, one part can affect two adjacent stripes to make the two stripes better connect to adjust the weight of the feature and the flow regulation coefficient. Setting stripenum to 8 does not make one part contain too many stripes, thus avoiding over-containment. In the case where stripe is set to 16 or 32, it will cause the same part to affect too many stripes, and result in a decrease in the recognition accuracy of the model.

These three graphs in FIGURE 8 clearly describe the change of rank-1 with stripenum under three conditions. The recognition accuracy when stripenum is larger than partnum is generally higher than the recognition accuracy when stripenum is less than partnum.

The experiments in this session fully verify that the local feature asymmetric correspondence between the middle and the end of the network plays a positive role in the adjustment of the network parameters. Setting the appropriate stripenum and partnum can significantly improve the recognition accuracy, and especially improve the recognition accuracy in the most challenging CL condition. The experiment also verified the validity of the local features flow regulation module. The local features flow regulation module and the network finally adjusted the individual training of P to form a perfect combination. The local features flow regulation module modules and the single training of each part are adjusted to each other to form a perfect combination.

C. CONTRAST EXPERIMENT

This section mainly lists the rank-1 precision comparison results between our model and other state-of-the-art models on the CASIA-B dataset, stripenum is set to 8. To fully verify the stability of the model, we conduct comparative experiments on small samples (ST), medium samples (MT) and large samples (LT). The main results are shown in the table below.

In TABLE 2, although the small sample training only uses 24 subjects, our model has achieved very good results. The effect of the model in NM and BG walking conditions is very good, rank-1 of our model is slightly higher than the highest rank-1 of the current gait recognition model. The CL condition is a walking condition in which a person wears a coat, and wearing a coat can have a good camouflage effect and can be used to cover a large gait posture. The rank-1 of our model is significantly higher than that of the current gait recognition model under the complex condition of CL. Our rank-1 is higher than the highest rank-1 of the current gait recognition model in all views.

Contrast experimental results of medium sample training (MT) and large sample training (LT) are listed in TABLE 3 and TABLE 4, respectively. As the training samples increase, the rank-1 of our model is getting higher and higher. Compared with the current gait model, there is a steady improvement in rank-1. Especially under CL complex walking conditions, our rank-1 is higher than the highest.
TABLE 4. Contrast experiment results of rank-1 for large sample training (LT) on the CASIA-B dataset.

| Gallery NM#1-4 | Probe | 0° | 18° | 36° | 54° | 72° | 90° | 108° | 126° | 144° | 162° | 180° | mean |
|----------------|-------|----|-----|-----|-----|-----|-----|------|------|------|------|------|------|
|                | CNN-3D[14] | 87.1 | 93.2 | 97.0 | 94.6 | 90.2 | 88.3 | 91.1 | 93.8 | 96.5 | 96.0 | 85.7 | 92.1 |
|                | CNN-Ensemble[14] | 88.7 | 95.1 | 98.2 | 96.4 | 94.1 | 91.5 | 93.9 | 97.5 | 98.4 | 95.8 | 85.6 | 94.1 |
|                | GaitSe[15] | 90.8 | 97.9 | 99.4 | 96.9 | 93.6 | 91.7 | 95.0 | 97.8 | 98.9 | 96.8 | 85.8 | 95.0 |
|                | Ours     | 91.1 | 97.9 | 99.6 | 97.3 | 94.3 | 91.9 | 94.9 | 98.1 | 98.8 | 96.2 | 86.6 | 95.1 |
| LT (74)        | BG#1-2 | CNN-LB[14] | 64.2 | 80.6 | 82.7 | 76.9 | 64.8 | 63.1 | 68.0 | 76.9 | 82.2 | 75.4 | 61.3 |
|                | GaitSe[15] | 83.8 | 91.2 | 91.8 | 88.8 | 83.3 | 81.0 | 84.1 | 90.0 | 92.2 | 94.4 | 79.0 | 87.2 |
|                | Ours     | 84.3 | 91.2 | 93.4 | 91.8 | 86.1 | 80.3 | 84.4 | 90.9 | 93.7 | 90.8 | 80.1 | 87.9 |
|                | CL#1-2  | CNN-LB[14] | 37.7 | 57.2 | 66.6 | 61.1 | 55.2 | 54.6 | 55.2 | 59.1 | 58.9 | 48.8 | 39.4 |
|                | GaitSe[15] | 61.4 | 75.4 | 80.7 | 77.3 | 72.1 | 70.1 | 71.5 | 73.5 | 73.5 | 68.4 | 50.0 | 70.4 |
|                | Ours     | 64.7 | 79.4 | 84.1 | 80.4 | 73.7 | 72.3 | 75.0 | 78.5 | 77.9 | 71.2 | 57.0 | 74.0 |

TABLE 5. Rank-1 of single module and module combinations.

| Method       | Condition | NM  | BG  | CL  |
|--------------|-----------|-----|-----|-----|
| Single Backbone | 93.115    | 85.234 | 69.679 |
| KFS          | 94.382    | 86.764 | 71.418 |
| LFFR         | 94.306    | 87.028 | 72.895 |
| KFS+LFFR     | 95.135    | 87.91  | 74.018 |

Single Backbone indicates that a backbone network is used and it does not add any modules. LFFR indicates a backbone that only adds the local features flow regulation module. KFS+LFFR indicates a network with two modules.

Combining KFS and LFFR can get better results, especially in CL walking conditions, the recognition accuracy is greatly improved. Experiments demonstrate that the combination of KFS and LFFR modules can further improve the recognition accuracy of the network.

V. CONCLUSION

In this paper, we propose a key frame extraction based on information weighting module and a Local Feature Flow Regulation module. The frame weighting module enables the network to increase the attention to key frames at the input end and improve the extraction efficiency of high key features. The Local Feature Flow Regulation module can control the local channel size in each frame and flexibly adjust the traffic size of the local features, so that the network can extract the temporal and spatial features of the gait silhouette locally and increase the network expression capability. The Local Feature Flow Regulation and the individual training of each part adjust each other. Such adjustment increases the flexibility of the network. Network flexibility improves the expressive ability of the network and the accuracy of gait recognition. Compared to other networks, our network pays more attention to the analysis of local features and relationships between them in space and time, thus mining more valuable information to recognize identity. Through a lot of experiments, we have proved the effectiveness of the Key frame extraction module and the Local Features Flow Regulation module. With proper stripenum and partnum selected, our algorithm can make a big breakthrough on CASIA-B dataset. Under the three walking conditions on the CASIA-B dataset, the recognition accuracy of our network is improved compared with the current state-of-the-art. Especially in the most complex walking conditions CL, rank-1 of our network can achieve 74%.

The proposed algorithm module can also be embedded into all networks that process video streams. It is used to extract the correlation and spatiotemporal information of local features between frames. In this paper, we use a fixed division method for global features. This kind of division method is simple. But because it is a hard division, this will result in
less precise alignment. The method has its limitations. In the future we will research an adaptive global feature division to improve our algorithm module to improve the flexibility of the network.

REFERENCES

[1] G. Guan, T. Yang, and W. Liu, “Gait recognition with skeleton information by using ensemble learning,” in Proc. 10th Int. Congr. Image Signal Process., Biomed. Eng. Informat. (CISP-BMEI), Oct. 2017, pp. 1–7.

[2] A. Kale, A. N. Rajagopalan, N. Cuntoor, and V. Kruger, “Gait-based recognition of humans using continuous HMMs,” in Proc. 5th IEEE Int. Conf. Autom. Face Gesture Recognit., May 2002, pp. 336–341.

[3] J. Little and J. Boyd, “Recognizing people by their gait: The shape of motion,” Videer, J. Comput. Vis. Res., vol. 1, no. 2, pp. 1–32, 1998.

[4] H. Murase and R. Sakai, “Moving object recognition in eigenspace representation: Gait analysis and lip reading,” Pattern Recognit. Lett., vol. 17, no. 2, pp. 155–162, Feb. 1996.

[5] S. A. Niyogi and E. H. Adelson, “Analyzing and recognizing walking figures in XVT,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), vol. 94, Jun. 1994, pp. 469–474.

[6] G. Shakhnarovich, L. Lee, and T. Darrell, “Integrated face and gait recognition from multiple views,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR), vol. 1, Dec. 2001, p. 1.

[7] L. Zhang, L. Zhang, D. Tao, and B. Du, “A sparse and discriminative tensor to vector projection for human gait feature representation,” Signal Process., vol. 106, pp. 245–252, Jan. 2015.

[8] B. Lyonnet, C. Ioana, and M. G. Amin, “Human gait classification using microDoppler time-frequency signal representations,” in Proc. IEEE Radar Conf., May 2010, pp. 915–919.

[9] A. I. Bazin and M. S. Nixon, “Gait verification using probabilistic methods,” in Proc. 7th IEEE Workshops Appl. Comput. Vis. (WACV/MOTION), vol. 1, Jan. 2005, pp. 60–65.

[10] Y. Maikawa, K. Tadokoro, T. Sato, K. Sugiyama, and T. Yagi, “Gait recognition using a view transformation model in the frequency domain,” in Proc. Eur. Conf. Comput. Vis. Springer, 2006, pp. 151–163.

[11] M. Hu, Y. Wang, Z. Zhang, J. J. Little, and D. Huang, “View-invariant discriminative projection for multi-view gait-based human identification,” IEEE Trans. Inf. Forensics Security, vol. 8, no. 12, pp. 2034–2045, Dec. 2013.

[12] W. Kusakunniran, Q. Wu, J. Zhang, H. Li, and L. Wang, “Recognizing gait across views through correlated motion co-clustering,” IEEE Trans. Image Process., vol. 23, no. 2, pp. 696–709, Feb. 2014.

[13] K. Shiraga, Y. Maikawa, D. Muramatsu, T. Echigo, and Y. Yagi, “GEINet: View-invariant gait recognition using a convolutional neural network,” in Proc. Int. Conf. Biometrics (ICB), Jan. 2016, pp. 1–8.

[14] 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 Trans. Pattern Anal. Mach. Intell., vol. 39, no. 2, pp. 209–226, Feb. 2017.

[15] H. Chao, Y. He, J. Zhang, and J. Feng, “GaitSet: Regarding gait as a set for cross-view gait recognition,” in Proc. AAAI Conf. Artif. Intell., vol. 33, Jul. 2019, pp. 8126–8133.

[16] J. Han and B. Bhanu, “Individual recognition using gait energy image,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 28, no. 2, pp. 316–322, Feb. 2006.

[17] R. Liao, C. Cao, E. B. Garcia, S. Yu, and Y. Huang, “Pose-invariant and perspective spatial network (PTSN) for gait recognition with carrying and clothing variations,” in Proc. Chin. Conf. Biometric Recognit. Springer, 2017, pp. 474–483.

[18] H. Liu, Z. Jie, K. Jayashree, M. Qi, J. Jiang, S. Yan, and J. Feng, “Video-based person re-identification with accumulative motion context,” IEEE Trans. Circuits Syst. Video Technol., vol. 28, no. 10, pp. 2788–2802, Oct. 2018.

[19] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, “Squeeze-and-excitation networks,” IEEE Trans. Pattern Anal. Mach. Intell., Jun. 2019, pp. 7132–7141.

[20] G. Song, B. Leng, Y. Liu, C. Hetang, and S. Cai, “Region-based quality estimation network for large-scale person re-identification,” in Proc. 32nd AAAI Conf. Artif. Intell., 2018.

[21] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, “Hierarchical attention networks for document classification,” in Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol., 2016, pp. 1480–1489.

[22] C. Li, X. Min, S. Sun, W. Lin, and Z. Tang, “DeepGait: A learning deep convolutional representation for view-invariant gait recognition using joint Bayesian,” Appl. Sci., vol. 7, no. 3, p. 210, 2017.

[23] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, and X. Tang, “Spindle net: Person re-identification with human body region guided feature decomposition and fusion,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 1077–1085.

[24] L. Zheng, Y. Huang, H. Lu, and Y. Yang, “Pose-invariant embedding for deep person re-identification,” IEEE Trans. Image Process., vol. 28, no. 9, pp. 4500–4509, Sep. 2019.

[25] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, “Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline),” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2018, pp. 480–496.

[26] S. Yu, D. Tan, and T. Tan, “A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition,” in Proc. 18th Int. Conf. Pattern Recognit. (ICPR), vol. 4, 2006, pp. 441–444.

[27] N. Takemura, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi, “Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition,” IPSJ Trans. Comput. Vis., vol. 10, no. 1, p. 4, Dec. 2018.

[28] 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, May 2017.

[29] Y. He, J. Zhang, H. Shan, and L. Wang, “Multi-task GANs for view-specific feature learning in gait recognition,” IEEE Trans. Inf. Forensics Security, vol. 14, no. 1, pp. 102–113, Jan. 2019.

[30] J. Gao and R. Nevatia, “Revisiting temporal modeling for video-based person ReID,” 2018, arXiv:1805.02104. [Online]. Available: http://arxiv.org/abs/1805.02104

[31] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, “Person re-identification by multi-channel parts-based CNN with improved triplet loss function,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 1335–1344.

[32] Q. Xiao, H. Luo, and C. Zhang, “Margin sample mining loss: A deep learning based method for person re-identification,” 2017, arXiv:1710.00478. [Online]. Available: http://arxiv.org/abs/1710.00478

GOUHENG HUANG (Member, IEEE) received the B.Sc. degree in mathematics and applied mathematics and the M.Eng. degree in computer science degrees from South China Normal University, in 2008 and 2012, respectively, and the Ph.D. degree in software engineering from Macau University, in 2017. He is currently an Assistant Professor of computer science with the Guangdong University of Technology. As a key member of the Guangdong Key Laboratory of Cyber-Physical System, he has published many research articles. His research interests include computer vision, pattern recognition, and artificial intelligence. He is currently a CCF member. He has hosted and undertaken a number of national and provincial-level scientific research projects, including the Natural Science Foundation of China and National Key Research and Development Plan.

ZENG LU (Graduate Student Member, IEEE) received the bachelor’s degree in engineering from Yichun University. He is currently pursuing the master’s degree in computer science with the Guangdong University of Technology. His main research fields are computer vision and person recognition.
CHI-MAN PUN (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees in software engineering from the University of Macau, in 1995 and 1998, respectively, and the Ph.D. degree in computer science and engineering from The Chinese University of Hong Kong, in 2002. He is currently an Associate Professor and the Head of the Department of Computer and Information Science, University of Macau. He has investigated several funded research projects and authored/coauthored more than 100 refereed scientific papers in international journals, books, and conference proceedings. His research interests include digital image processing, multimedia forensics and watermarking, pattern recognition, and computer vision. He is a professional member of the ACM. He has also served as the Editorial Member/Referee for many international journals, such as the IEEE Transactions on Pattern Analysis and Machine Intelligence, IEEE Transactions on Image Processing, and IEEE Transactions on Information Forensics and Security.

LIANGLUN CHENG (Senior Member, IEEE) received the B.E. and M.S. degrees in automation from the Huazhong University of Science and Technology, Wuhan, China, in 1988 and 1992, respectively, and the Ph.D. degree in automation from the Chinese Academy of Sciences, in 1999. He is currently a Professor with the Guangdong University of Technology, a Computer Dean of the Guangdong University of Technology, a Doctoral Tutor, an excellent Teacher of Nanyue, and a national-level training target for the Thousand-Ten Thousand project of cross-century talents in Guangdong Province. He is the Executive Director of the Robotics Professional Committee of China Automation Association, a member of China Computer Federation, and the Vice Chairman of Guangdong Automation Association. His main research interests include knowledge graph, knowledge automation, and information physics fusion systems.