The Journal of Supercomputing
https://doi.org/10.1007/s11227-024-06089-7

# Gaitdlf: global and local fusion for skeleton-based gait recognition in the wild

Siwei Wei[1,2] · Weijie Liu[1] · Feifei Wei[3] · Chunzhi Wang[1] · Neal N. Xiong[4]

Accepted: 18 March 2024

## Abstract

A new trend in long-range biometrics, gait recognition, is finding application in a number of different fields including video surveillance. Recently, with the increase in robustness of the pose estimator and the presence of various unpredictable factors in realistic gait recognition, skeleton-based methods with higher robustness have emerged to better meet the challenging gait recognition needs. However, existing approaches primarily focus on extracting global skeletal features, neglecting the intricate motion information of local body parts and overlooking inter-limb relationships. Our solution to these challenges is the dynamic local fusion network (GaitDLF), a novel gait neural network for complex environments that includes a detail-aware stream in addition to the previous direct extraction of global skeleton features, which provides an enhanced representation of gait features. To extract discriminative local motion information, we introduce predefined body part assignments for each joint in the skeletal structure. By segmenting and mapping the overall skeleton based on these limb site divisions, limb-level motion features can be obtained. In addition, we will dynamically fuse the motion features from different limbs and enhance the motion feature representation of each limb by global context information and local context information of the limb-level motion features. The ability to extract gait features between individuals can be improved by aggregating local motion features from different body parts. Based on experiments on CASIA-B, Gait3D, and GREW, we show that our model extracts more comprehensive gait features than the state-of-the-art skeleton-based method, demonstrating that our method is better suited to detecting gait in complex environments in the wild than the appearance-based method.

**Keywords** Gait recognition · Computer vision · Pattern recognition · Deep learning

Extended author information available on the last page of the article

Published online: 01 May 2024

## 1 Introduction

Gait recognition [1, 2] is a biometric technology that identifies human gait patterns based on posture features. Compared to other biometric measures such as facial recognition, iris scanning, and fingerprint identification, gait recognition method offers distinct advantages, being widely applied in domains like video surveillance and criminal investigation. This technology constitutes a facet of computer vision research, striving to identify the subtle and unique variations within human gait patterns that can differentiate one individual from another [3–5].

The most used approach in gait recognition is a network that takes silhouettes as input. Such appearance-based methods are highly sensitive to factors like clothing, carried items, cluttered backgrounds, and occlusions prevalent in complex environments [6, 7]. Thus, the main challenge in gait recognition lies in extracting robust features unaffected by these influences [8, 9].

Recent research has demonstrated excellent recognition accuracy using silhouette images as inputs for appearance-based gait recognition models [10–16]. GaitSet [17] treats gait sequences as sets, an efficient approach later adopted by subsequent research. GaitPart [18] emphasizes dependence between local motion details and time. GaitGL [19] addresses the lack of fine-grained details in spatial global gait representations. 3DLocal [20] extracts limb features through adaptive-scaled 3D local operations. These methods emphasize not only global feature extraction but also detailed limb motion features.

Moreover, with advancements in pose estimation networks, the accuracy of extracting skeletal sequences from video footage has improved consistently. Skeletal-based gait recognition methods promise to have higher robustness in complex real-world Environments. In recent years, due to the improved performance of attitude estimators [21, 22], such model-based methods have garnered increasing attention [23–26]. PoseGait [27] utilizes 3D human posture and prior knowledge to mitigate clothing variations, while GaitGraph [28] introduces graph convolutional networks to learn 2D skeletal gait representations. BiFusion [29] integrates skeletons and outlines to capture rich spatiotemporal gait features.

However, model-based methods currently lag behind appearance-based approaches in terms of performance. This is due to the fact that these methods primarily focus on the global features of skeletal sequences, extracting features from the entire body skeleton and subsequently subjecting them to joint-level pooling, inadvertently overlooking limb-specific motion information and inter-limb motion relationships.

As shown in Fig. 1, during human locomotion, distinct limb sites may exhibit unique motion patterns—for instance, the swing of arms and the stride of legs. These different movement patterns of localized limb sites provide complementary information for the gait recognition process. Notably, within a complete gait cycle, highly mobile areas like arms and legs often carry richer and more varied motion information than other regions.

Therefore, modeling localized limb sites becomes crucial in order to fully exploit the subtle differences in gait patterns and synergies between limb sites.

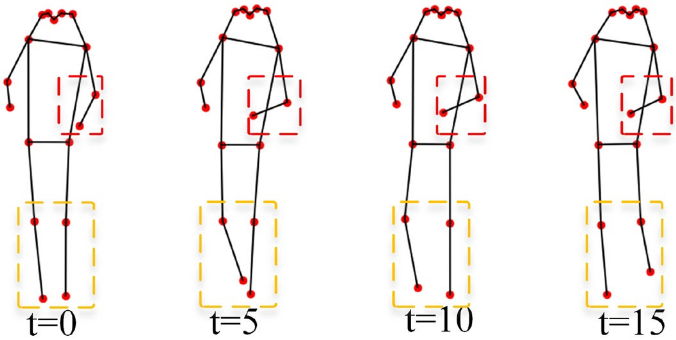

**Fig. 1** Skeleton sequence. Each limb part has a unique movement pattern, and the hands and legs have the greatest variation in movement

In the indoor dataset CASIA-B, however, the silhouette-based gait recognition network is able to achieve very superior performance. However, the performance on wild datasets such as GREW and Gait3D shows that the performance of the gait recognition network still needs to be optimized. In order to enable the gait recognition network to accurately recognize different people in a complex environment full of unpredictable factors, our GiatDLF chooses to use the skeleton data obtained from the pose estimator as the input to the network, since skeleton-based methods are not highly sensitive to factors such as clothing, carried objects, cluttered backgrounds, and occlusions that are prevalent in complex environments. And by focusing on and capturing distinctive motion features of different limb sites, we can achieve more accurate identification and differentiation of gait patterns among individuals. The final conclusion also shows that the fusion of global and local features can enhance the performance and robustness of gait recognition.

In order to address the above challenges of existing methods, Skeleton Local Mapping (SLM) is introduced in this study. This method comprehensively models each limb site of the human skeleton. As shown in Fig. 2, it divides the entire skeletal sequence into several partitions based on predefined limb sites. Then a maximum pooling operation is performed on the joint motion information contained in each partition. In addition, the features obtained from each limb segment are individually mapped to the corresponding feature space through a fully connected layer. The aim is to extract the motion features of each limb segment to produce limb-level motion features. This process transforms limb-specific features into a structured and informative representation. Compared to the traditional approach of direct global joint pooling, this study aims to enhance gait feature extraction in complex environments, enabling the network to extract finer-grained motion features.

To facilitate information exchange and integration among different limb sites, we propose dynamic feature fusion (DFF). The method models different local limb motion features and achieves a dynamic mechanism for adaptive selection of relevant limb parts and features by calculating the attention weights obtained by aggregating the global and local contextual information of different limb-level motion features. This intricate framework effectively eliminates unnecessary redundant

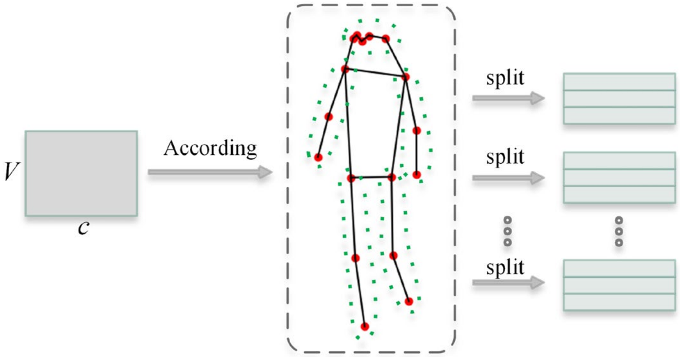

**Fig. 2** Dashed box shows the pre-defined limb parts in the skeleton. $V$ denotes the joint dimension and $C$ denotes the feature dimension. The features of each joint are divided into five parts based on the predefined left hand, right hand, left leg, right leg, and head

information while retaining the salient and distinctive attributes of the data. Our approach aids in comprehending human motion from both global and local perspectives. It adeptly tackles challenges faced by prior methods, enhancing the capability of gait feature extraction.

The primary contributions of this study are summarized as follows:

- We propose a novel gait recognition network called GaitDLF, which takes into account local motion information on top of the original modeling of global skeleton movements in order to extract key gait features, and the results show that this method is more suitable for field environments with complex conditions than existing appearance-based networks.
- In GaitDLF, we perform comprehensive modeling of each limb site to generate limb-level motion features for each limb site. Subsequently, the global and local context information is obtained from different limb-level motion features to obtain the attention matrix, which realizes the dynamic fusion of different limb-level motion features.
- We conduct a comprehensive evaluation on the CASIA-B, Gait3D, and GREW gait datasets. Experimental results on GaitDLF demonstrate its superiority among skeleton-based methods, underscoring the effectiveness of emphasizing local limb motion information in model-based gait recognition. It also demonstrates that the skeleton-based approach is more appropriate than the previous appearance-based approach when facing complex environments in the wild.

## 2 Related work

There are two types of gait recognition methods, which can be broadly categorized into two types based on the data depending on how the data are input: appearance based and model based.

### 2.1 Appearance-based gait recognition

Appearance-based gait recognition methods focus on directly feeding appearance video sequences into the network for gait recognition. This method also performs very well in low resolution conditions, so more and more researchers are studying this method [30, 31]. The research of appearance-based methods has mainly focused on time series modeling and spatial feature extraction due to the rapid development in the fields of video understanding and deep learning.

Appearance-based gait recognition methods capitalize on the processing of video sequences to identify individuals based on their gait. Researchers have paid considerable attention to these methods, as they proved robust under low-resolution conditions [30, 31]. Advances in video analysis and deep learning have spurred focused investigations into spatial feature extraction and temporal modeling within this domain.

Among other things, GaitSet [17] believes that given a collection of gait sequences of one cycle, it is possible to discriminate between different people. They processed each frame of the gait sequence using statistical functions and viewed it as a set of frames. This simple but effective method is known as one of the important contributions to the research of gait recognition this year. GaitPart [18] considers the dependency between local information of input contours and short-range time, and it designs a micro-action pattern builder, which investigates the relationship between the local detail information of the input contours and the short-range time and combines it with different window sizes to extract the complex local spatial-temporal features. In contrast, GaitGL [19] achieves global and local feature extraction by using multiple convolutional layers. The approach is able to capture subtle local motion information, which overcomes the limitations of ignoring detailed information based on global representations and the difficulty of effectively capturing the relationship between neighboring regions based on local region descriptors. CSTL [32] takes inspiration from the observation that humans can discriminate between gaits by adapting their focus to different time scales. It proposes a context-sensitive temporal feature learning (CSTL) network to assess the importance of features by modeling relationships between multi-scale features. By enhancing scales that are more important and suppressing scales that are less important, the network can adapt. Another technique to solve the misalignment problem caused by temporal operations (e.g., temporal convolution) is SSFL. This method extracts the most discriminatory parts of a sequence in order to reassemble the frame of salient spatial features. Additionally, 3DLocal [20] asserts that current methods of providing localized parts do not produce accurate results because they average feature maps. They propose to solve this problem by supporting 3D localization operations that extract body parts in sequences with adaptive spatio-temporal scales, positions and lengths. In this way, a 3D local neighborhood with specific scales, positions, frequencies, and lengths can be used to learn the spatio-temporal patterns of body parts. Meanwhile, GaitEdge [33] argues that end-to-end approaches are inevitably affected by gait-related noise (i.e., low-level texture and color information). An end-to-end framework is presented which effectively masks gait-independent information and unlocks the potential for end-to-end training, synthesizing pedestrian segmentation network output and then feeding it to a recognition network. In order to limit the amount of information available to the recognition network, the synthesized silhouette is comprised of trainable edges and a fixed interior. In addition, a large-scale self-supervised gait recognition benchmark using contrast learning is also proposed by GaitSSB [34], as well as the collection of a large-scale unlabeled gait data set, GaitLU-1 M, consisting of 1.02 million walking sequences. A conceptually simple model of gait recognition, the GaitSSB, is proposed to provide a robust empirical basis for the model.

It is worth noting that in recent years, appearance-based approaches have typically achieved higher recognition accuracies compared to model-based approaches.

### 2.2 Model-based gait recognition

Skeleton-based methods are among the most widely used for model-based gait recognition. In traditional skeleton-based gait recognition methodologies, a great deal of effort is put into extracting discriminative parameters from raw skeleton data to determine the gait and then applying those to the gait recognition algorithms [27, 35, 36]. For example, several spatiotemporal and dynamic parameters for gait recognition were computed by Deng et al. [37]. However, these methods rely on manually designed features, making the whole process complex and undesirable.

To address real-world application demands, researchers collected the GREW [38] and Gait3D [39] datasets in outdoor environments in 2021 and 2022, respectively, marking a shift from indoor to outdoor research. Model-based approaches to gait recognition have gained new vigor with the gradual shift in research direction from indoor to outdoor and the improved robustness of lightweight pose estimators [21, 22]. Model-based methods enable gait recognition to perform more robustly in complex environments in the face of unpredictable factors such as noise interference, mixed backgrounds, unconstrained walking paths, and occlusions.

Specifically, the pose-based spatio-temporal network PTSN [23] uses CNN for spatial modeling and LSTM for temporal modeling for spatio-temporal gait feature extraction. It is the first to propose a gait recognition method using pose estimation and utilizing pose keypoints. PoseGait [27] computes joint angles, bone lengths, and joint motions by means of 3D keypoints in Euclidean space. The method utilizes 3D human pose and a priority knowledge of the human body and overcomes the problem of clothing variations by using hand-crafted 3D pose estimation features. These hand-crafted features are then input into a convolutional neural network to learn advanced spatio-temporal features. The method was evaluated in the cross-view setting of CASIA-B [36], with highly competitive results compared to appearance-based methods.

In order to represent gait more accurately, GaitGraph [28] uses RGB images directly to determine robust skeleton postures. Based on their arguments, these methods misrepresent fine-grained spatial information while silhouette images retain several recognizable visual cues as well as gait features as they lose the fine-grained spatial information. Therefore, they proposed GaitGraph to achieve a number of modern gait recognition methods by combining skeleton poses with graph convolutional networks (GCNs) to create a method based on modern model-based gait recognition. Gaitgraph2 [40] believes that silhouette images contain a wide range of visual cues that aren't actually gait features but can be used for recognizing gait patterns, but they can also provide context for deceiving the system in the process. Hence, they propose an algorithm for gait recognition that combines higher-order inputs with residual networks that use graph convolutional networks (GCNs) to produce an architecture that is efficient and effective for the recognition of gait and its variations. HMRGait [41] propose a model-based approach for end-to-end gait recognition. The study utilizes a skinned multi-person linear (SMPL) model in order to model humans and parameters of this model can then be estimated through the use of a pre-trained human mesh recovery (HMR) network. The reconstruction loss that is introduced between the contour masks of the gait dataset and the contours of the renders of the estimated SMPL model that is generated by the differentiable renderer allow them to compensate for the discrepancy between the gait dataset and the dataset used for pre-training HMR. A study conducted by SMPLGait [29] has found that most existing gait recognition methods learn features from either contours or skeletons, but that the combination of these two data sources may offer more than just one advantage. In the past, multimodal gait recognition methods have mainly utilized the skeleton as an aid in extracting local features, instead of taking advantage of the inherent discriminative properties of skeleton data. In concert with bimodal fusion (BiFusion) techniques, they propose that to learn rich recognition features from gait patterns, the skeleton should be mined for discriminative gait patterns, which can then be combined with contour representations.

We found that many studies in the field of appearance-based approaches emphasize modeling local limb motion details and extracting local motion features through corresponding modules. Drawing inspiration from these appearance-based approaches, our proposed skeleton-based gait recognition network focuses on modeling the relationships between different limb motions to achieve richer feature representations.

## 3 GaitDLF

In this section, we provide a comprehensive elucidation of the feature extraction process within GaitDLF, along with the functions of each pivotal component.

### 3.1 Preliminaries

#### 3.1.1 Human pose estimation

Human pose estimation, also known as joint detection, is used to detect the positions of $V$ joint points (such as wrists, ankles, etc.) in each frame of an image of size $W \times H \times 3$. In gait recognition, datasets are obtained through the original video sequences and pre-trained human pose estimation to obtain human skeleton graph sequences.

### 3.1.2 Notation

A human skeleton graph sequence is presented in the form of a matrix, denoted as $X_i \in \mathbb{R}^{C\times T\times V}$, where $i$ represents that this is the $i$-th sequence, $V$ represents the number of joints included in the human skeleton graph, $T$ the number of frames in a sequence, and $C$ the features of each joint, which include 2D coordinate information and confidence score.

## 3.2 Pipline

The whole process of GaitDLF is shown in the diagram Fig. 3. To be more specific, considering $X_i$ as the representation of the input skeleton sequence, the entire process can be mathematically formalized as follows:

$$Z_i = F\left(X_i\right), \tag{1}$$

$$Y_i = \text{concat}\left(F_{\text{L}}\left(Z_i\right), F_{\text{G}}\left(Z_i\right)\right), \tag{2}$$

where $F(\cdot)$ denotes the Feature Extractor responsible for extracting key motion patterns and features $Z_i$ from the skeleton sequence. $F_{\text{G}}(\cdot)$ represents GAS, $F_{\text{L}}(\cdot)$ represents LDS, $concat(\cdot)$ signifies the channel-wise concatenation operation, and $Y_i$ denotes the final gait features used for loss computation.

In comparison with earlier research, GaitDLF not only emphasizes global motion representation but also underscores local motion representation, thus achieving superior recognition capability. The details of the Feature Extractor are discussed in Sect. 3.3, while the Local Detail Stream is elaborated in Sect. 3.4 and the Global Awareness Stream is detailed in Sect. 3.5.

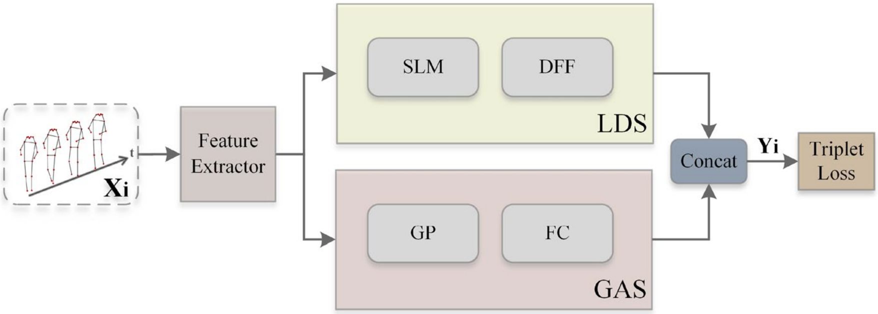

**Fig. 3** Overall flow of GaitDLF. The skeleton sequence $X_i$ is fed into the feature extractor to extract spatio-temporal features and then spliced after passing through the Global Awareness Stream (GAS) and Local Detail Stream (LDS), and finally, the loss is computed. The detailed structure of Skeleton local mapping (SLM) and Dynamic Feature Fusion (DFF) is shown in Figs. 6 and 7

### 3.3 Feature extractor

In our GaitDLF, the first step is to pass the input skeleton sequence through a feature extractor to extract temporal and spatial features.

Our approach mainly addresses how to enable the network to focus on the unique information of different limbs and how to adaptively fuse limb-level features, for which we refer to the structure proposed by GaitGraph2. As shown in Fig. 4, this extractor computes the input skeleton sequence into three features, position, velocity, and skeleton. Then these three features through three branches of ST-GCNs, respectively. To reduce computational complexity, channel splicing is performed in the middle of the network to fuse the features. Subsequently, the fused features continue to extract spatio-temporal features through multi-layer ST-GCNs. The ST-GCN blocks utilize the ability of graph convolutional networks (GCNs) to model skeleton data in the spatial domain to capture the connectivity between human joints through the graph structure. Each ST-GCN block consists of a graph convolutional layer for learning feature representations on the skeleton graph and a temporal convolutional layer for capturing the evolution of movements over time. This combination of modeling in the temporal and spatial domains allows our network to effectively understand and characterize the complexity of human actions. With this in-depth feature extraction, the ST-GCN block significantly improves the accuracy and efficiency of gait recognition [40]. Among them, the improved ST-GCN adds a bottleneck and residuals to the original ST-GCN [42, 43] as shown in Fig. 5.

With this extraction method, we can effectively capture the basic motion patterns and features in the skeleton sequence, thus laying the foundation for the subsequent gait recognition task. It should be emphasized that this feature extractor has the same architecture and the same feature computation method as GaitGraph2 [40].

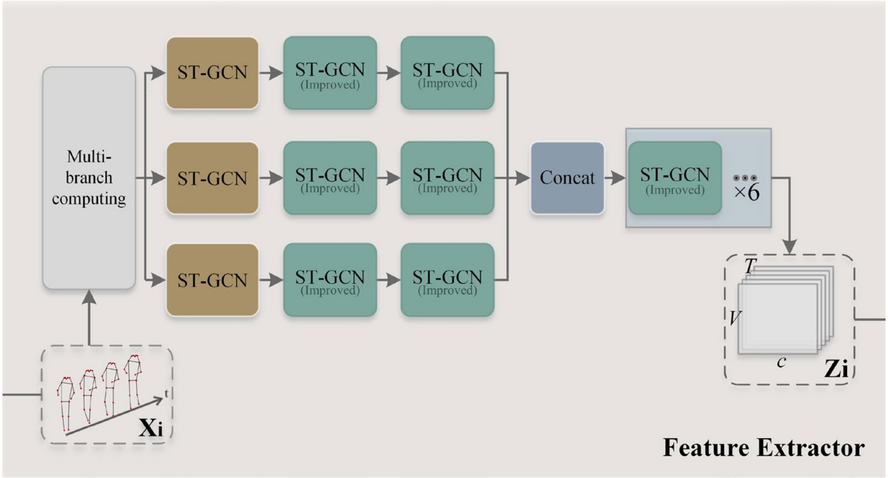

**Fig. 4** Proposed architecture of feature extractor in GaitDLF. It consists of a three-branch computation, some ST-GCNs, and some improved ST-GCNs [40]. The extracted high-level feature $Z_i$ will be fed into LDS and GAS

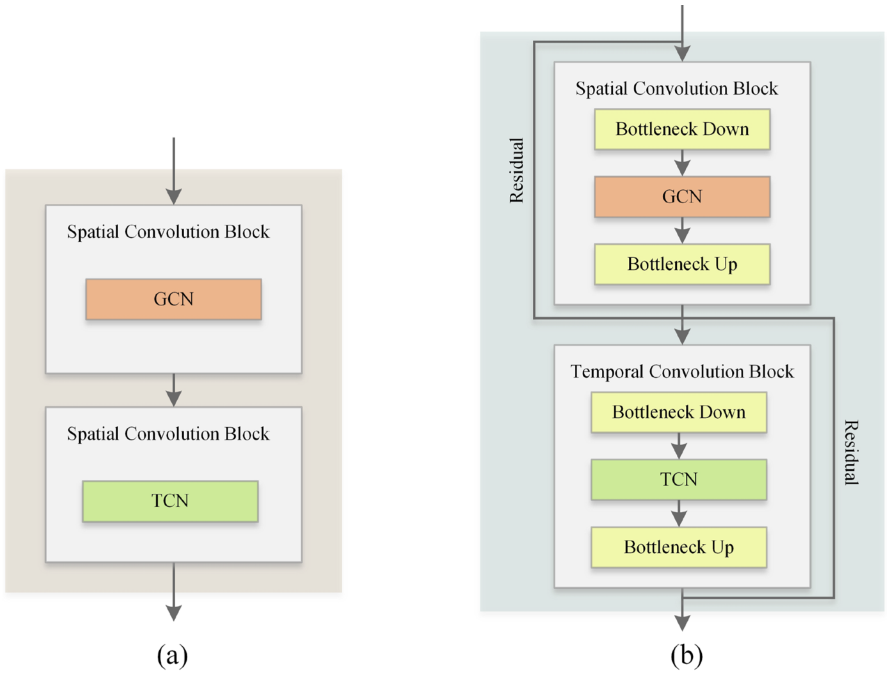
Spatial Convolution Block
GCN
Spatial Convolution Block
TCN
(a)
Residual
Spatial Convolution Block
Bottleneck Down
GCN
Bottleneck Up
Temporal Convolution Block
Bottleneck Down
TCN
Bottleneck Up
Residual
(b)

**Fig. 5** **a** Original ST-GCN consists of a GCN, which is used to learn the feature representation on the skeleton map, and a TCN, which is used to capture the evolution of motion over time. **b** Improved ST-GCN. Based on the original ST-GCN, the number of parameters is reduced using Bottleneck to lighten the network, and residuals are added to enhance the network learning capability

Subsequently, the extracted features are separately fed into the Global Awareness Stream (GAS) and Local Detail Stream (LDS). This facilitates the extraction of both global and local joint motion information. While GAS focuses on encompassing the overall skeletal motion representation, LDS intricately captures local limb motion information, thus enhancing the network's recognition capacity.

### 3.4 Local detail stream

Prior methods commonly employed graph convolutional networks and temporal convolutional networks (TCN) to extract spatial and temporal features from skeleton sequences, culminating in significant recognition accuracy. However, they pooled the extracted high-level features directly across the joint dimensions to obtain the final gait representation. This approach is simple and effective, but since most people's gait movements are very similar and more accurate recognition often requires subtle limb movements, we further model the high-level features based on different limbs, and the proposed Local Detail Stream achieves more accurate recognition accuracy.

The key of LDS is to effectively partition the high-level features $Z_i$ into different limb parts and map them separately to obtain limb-level motion features that carry information unique to different limb parts. This dynamic fusion allows the network to adaptively focus on more significant information and extract finer features. LDS consists of the following two modules.

### 3.4.1 Skeleton local mapping

To achieve this, we first employ the Skeleton local mapping (SLM) process. For a detailed schematic of the SLM process, refer to Fig. 6. In SLM, pooling is first performed along the temporal dimension and then divided into different limb part motion features based on the limb part to which different joints belong. Subsequently, maximum pooling is performed on multiple joints within each limb part to obtain limb-level motion features so that more fine-grained features can be obtained.

We then map the features of each limb segment to a discriminative feature space using separate fully-connected layers, highlighting the motion information of the limb segment and obtaining a more fine-grained representation. The use of a separate fully connected layer allows the model to learn the most appropriate representation parameters for each limb, which helps later modules capture unique information about that limb as it performs its movements, and the fully connected layer can also deepen the network here to enhance the model's representational capabilities. It is worth noting that we believe that the use of average pooling in the global perceptual flow yields information about the global skeleton, whereas max pooling is used here to prevent some locally significant information from being averaged out by the global one. The following mathematical formalization can be used to describe this process:

$$f_{\mathrm{p}} = \delta\left(\mathrm{FCs}\left(\mathrm{Div}\left(\mathrm{TP}\left(Z_i\right)\right)\right)\right), \tag{3}$$

where $\delta(\cdot)$ represents the activation function, $\mathrm{Div}(\cdot)$ denotes the process of dividing into localized limb features, $\mathrm{FCs}(\cdot)$ signifies the fully connected layers for each limb

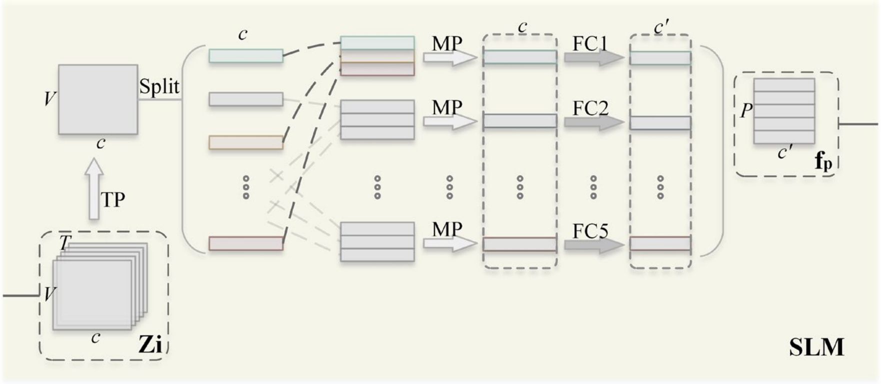

**Fig. 6** Proposed architecture for Skeleton Local Mapping (SLM) in LDS. First, temporal pooling (TP) is performed on *Zi*, then each joint feature is segmented according to the predefined limb parts of the Fig. 2, then pooling (MP) is performed on each limb part, and finally, the mapping is performed using the fully connected layer (FC) in order to obtain limb-level motion features $f_{\mathrm{p}}$

part, TP(·) denotes maximum pooling in the time dimension, and $f_p$ denotes limb-level features.

### 3.4.2 Dynamic feature fusion

In this section, we explain the proposed Dynamic Feature Fusion (DFF) process. See Fig. 7 for a detailed schematic.

The process utilizes the limb-level features to generate global contextual aggregation features and local contextual aggregation features and then uses a sigmoid function to compute the weights on each limb part as well as on each channel, with the aim of establishing adaptive fusion weights at each limb part and channel, thus dynamically fusing different features from different limb parts.

The process generates global context features and local context features using limb-level features and finally sums the global and local context features. The combined global and local context features are obtained, providing a comprehensive feature representation. The sigmoid function is then used to map this information to the 0–1 range to obtain adaptive weights for each limb site and each channel. The limb-level feature points are multiplied by the adaptive weights and then passed through maximum pooling, thus dynamically fusing different features from different limb-level features. To mitigate gradient vanishing, we also add residual connectivity. This then allows the network to effectively learn the important features and ignore the unwanted information. It is worth noting that max-pooling is also used here, with the aim of selecting the max values for different limb parts, retaining the salient information from different limb parts, and using it in the calculation of the attention weights.

Specifically, to fuse information from five limb parts while maintaining a lightweight approach, we opt for Pointwise Convolution (PWConv) as the channel

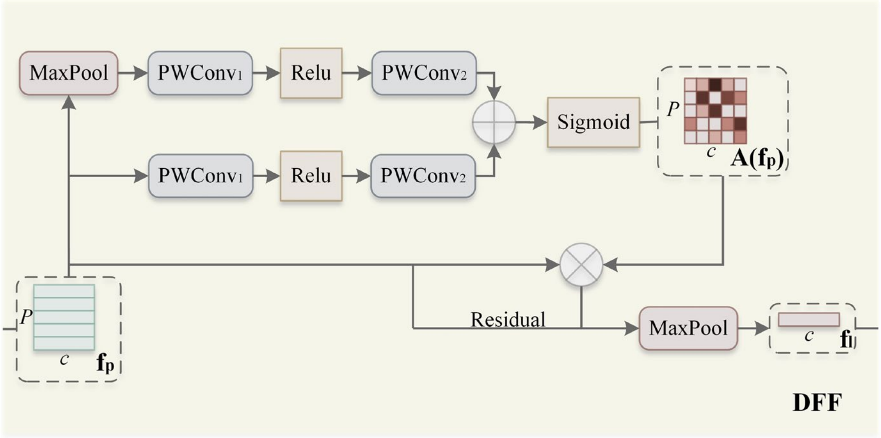

**Fig. 7** Proposed architecture of Dynamic Feature Fusion (DFF) in LDS. Attention weights $A$ are computed after global and local contextual aggregation of the limb-level features $f_{\mathrm{p}}$, and the output $f_{\mathrm{l}}$ is obtained by dot-multiplying with the input and adding residual

context aggregator. PWConv solely leverages point-wise channel interactions at each spatial position. The computation of local context feature $L(f_p)$ and global context feature $G(f_{\mathrm{p}})$ is as follows:

$$L(f_{\mathrm{p}}) = \mathrm{PWConv}_2\big(\delta\big(\mathrm{PWConv}_1(f_{\mathrm{p}})\big)\big), \tag{4}$$

$$G(f_{\mathrm{p}}) = \mathrm{PWConv}_2\big(\delta\big(\mathrm{PWConv}_1\big(MP(f_{\mathrm{p}})\big)\big)\big), \tag{5}$$

The kernel sizes of $\mathrm{PWConv}_1(\cdot)$ and $\mathrm{PWConv}_2(\cdot)$ are $\frac{C}{r} \times C \times 1 \times 1$ and $C \times \frac{C}{r} \times 1 \times 1$. $C$ denotes the feature dimension; $r$ is the channel shrinkage rate, for dimensionality reduction and to make the network more lightweight. It is worth noting that $L(f_{\mathrm{p}})$ retains the same shape as the input features, thereby preserving and emphasizing subtle details in low-level features. Given the global context feature $G(f_{\mathrm{p}})$ and local context feature $L(f_{\mathrm{p}})$, the attention matrix $A$ can be formalized as:

$$A = \sigma(L(f_{\mathrm{p}}) \oplus G(f_{\mathrm{p}})), \tag{6}$$

where $\sigma(\cdot)$ denotes the sigmoid function, $L(f_{\mathrm{p}})$ and $G(f_{\mathrm{p}})$ denote the local context feature and global context feature, and $\oplus$ denotes the matrix addition.

With this strategy, we are able to focus our attention on key body part information and reduce redundant data, thus improving the network's ability to capture and recognize local motion patterns. Dynamic feature fusion (DFF) flexibly adjusts the weights of different parts to achieve accurate information fusion. The local motion features $(f_l)$ can be formalized as:

$$f_{\mathrm{l}} = F_{\mathrm{L}}(Z_i) = MP(f_{\mathrm{p}} \otimes A), \tag{7}$$

where MP$(\cdot)$ represents max-pooling, $\otimes$ denotes the dot product, $f_{\mathrm{l}}$ denotes the locally fused limb part detail features, $A$ represents attention matrix, and $F_{\mathrm{L}}(\cdot)$ represents LDS.

### 3.5 Global awareness stream

Beyond the Local Detail Stream, perceiving the global skeletal structure is also paramount. Effectively utilizing both the global skeletal framework and local limb information can offer rich features for gait recognition. The GAS module in this paper aims to extract global motion features from skeleton sequences, thus giving the network the ability to recognize global motion. Through global average pooling, we reduce the dimensions of the high-level features from output $Z_i$, yielding an overall representation of global motion. Average pooling computes the average of multiple features, thus yielding information about the global skeleton that complements the local information extracted from the local detail stream. Subsequently, we employ a fully connected layer to map these features into the discriminant space to enhance the discriminative and unique nature of the features. Through GAS, we capture overall motion patterns in the skeleton sequences, providing information support from a global perspective for recognition tasks. This process can be expressed as:

$$f_{\mathrm{p}} = F_{\mathrm{G}}(Z_i) = \delta(\mathrm{FC}(\mathrm{GP}(Z_i))), \tag{8}$$

where $\delta(\cdot)$ denotes the activation function, $\mathrm{FC}(\cdot)$ signifies fully connected layers, $\mathrm{GP}(\cdot)$ represents global average pooling, $f_{\mathrm{g}}$ denotes global perception features, and $F_{\mathrm{G}}(\cdot)$ represents GAS.

# 4 Experiments

The following gait datasets are used to evaluate GaitDLF: CASIA-B [36], Gait3D [39], and GREW [38]. The first part of this paper presents three benchmark gait datasets and their corresponding training configurations. Using identical experimental conditions, we compare GaitDLF with state-of-the-art gait methods. Finally, we will conduct ablation experiments to identify how each component of GaitDLF affects performance.

## 4.1 Datasets

A trio of excellent gait datasets are employed, including CASIA-B [36], which is widely used indoors, Gait3D [39], which is famous for its diversity in real-life settings, and GREW [38], which is commonly used in wild environments. As shown in Table 1, the relevant statistics pertaining to the number of sequences and identities are provided. Subsequent sections present a comprehensive outline of the data collection procedures for each dataset, underscoring the notable distinctions between indoor and outdoor datasets.

### 4.1.1 CASIA-B [36]

Is a widely used gait recognition dataset with 124 subjects. Each subject had 10 different walking styles, including 6 normal walking styles, 2 backpack walking styles, and 2 walking styles with different clothes. Viewing angles ranged from 0° to 180° at 18° intervals, and a total of 11 cameras were used. The CASIA-B dataset therefore includes 13,640 gait sequences. We adopted the popular division scheme of existing studies [44], since CASIA-B has never been officially divided into training and test sets. Specifically, training data were collected from 74 subjects (001-074), while test data were collected from 50 subjects (075-124). Four gait sequences were used in the test set in NM condition. The rest of the gait sequences served as probe

**Table 1** Count of labels (Lab) and sequences (Seq) contained in the CASIA-B, GREW, and Gait3D datasets

| Dataset | Year | Train set | | Test set | |
|---|---|---|---|---|---|
| | | Lab | Seq | Lab | Seq |
| CASIA-B [36] | 2006 | 74 | 8140 | 50 | 5500 |
| GREW [38] | 2021 | 20,000 | 102,887 | 6000 | 24,000 |
| Gait3D [39] | 2022 | 3000 | 18,940 | 1000 | 6369 |

sequences. To extract the pose information in the CASIA-B dataset, we used the pretrained HRNet pose estimation model [45] that was pre-trained previously.

#### 4.1.2 Gait3D [39]

Is a large-scale live gait recognition dataset that was captured using 39 cameras in a supermarket, and the dataset includes 4000 different objects and more than 25,000 gait sequences. In the training phase, we use data from 3000 of these objects for training. During the testing phase, all of the remaining 1000 samples will be analyzed, with a maximum of one sequence from each sample being randomly selected as the probe sequence and the remaining 1000 samples serving as the gallery.

#### 4.1.3 GREW [38]

Is a dataset derived from real-world video streams, which are composed of hundreds of cameras and thousands of hours of streams in open systems, as captured by natural cameras. Including 26K identifiers and 128K sequences, the GREW dataset contains a variety of attributes, useful view variations, and more natural challenge factors throughout that make it a great dataset for training, validation, and test purposes. Test objects are composed of two sequences, one of which is considered a probe and the other is considered a gallery, which is composed of two sequences from each test object.

In all our experiments, we follow the official protocols or the most popular protocols for training, testing, and gallery/probe set segmentation in order to achieve the most relevant results possible, using the main evaluation metrics such as Rank-1, Rank-5, Rank-10, and Rank-20 in order to evaluate the effectiveness of our experiments.

### 4.2 Implementation details

#### 4.2.1 Parameter configuration

We use a triplet loss function whose threshold is set to 0.2. The embedding dimension of the last fully connected layer is specified to be 128, and bn layers are added after each activation function. For training, the batch size is configured as ($N$, $S$), indicating that N labels are selected and S sequences are selected for each label. Specifically, for the CASIA-B dataset, the training batch is set to (10, 6), while for the GREW and Gait3D datasets, it is set to (48, 12).

#### 4.2.2 Optimizer

SGD [46] optimizer is used during the optimization process. As a result of these hyperparameters, the initial learning rate is set at 0.1, momentum is set at 0.9, and weight decay is set at 0.005. In order to reduce the learning rate for the GREW dataset by ten, steps of 30K, 60K, 90K, and 120K were taken, which resulted in 150K training iterations. For the Gait3D dataset, there are 32K iterations in total for the training process at a step size of 30K, which results in a reduction in learning rate by a factor of 10 during the training process. The learning rate of the CASIA-B dataset was reduced by a factor of 10 at step sizes of 20K and 40K, and a total of 42K iterations were required for this dataset

#### 4.2.3 Data augmentation

To mitigate the influence of errors arising from pose predictions of task joints in the original RGB videos, a data augmentation strategy is implemented. The coordinates of each joint are augmented with Gaussian noise with a variance of 0.25, which significantly improves the network's ability to cope with errors of this magnitude

During the training phase, a sequence of 30 consecutive frames is arbitrarily selected as input. In the testing phase, all frames within a sequence are input into the network for evaluation. All experimental procedures are conducted using the PyTorch framework on a single NVIDIA GeForce GTX 3080 GPU. Table 2 displays the main hyper-parameters of our experiments.

### 4.3 Comparison with state of the art

The aim of this study is to compare GaitDLF with state-of-the-art skeleton-based methods for gait recognition. A systematic and comprehensive experiment was conducted on three different datasets to validate GaitDLF's effectiveness. These datasets are CASIA-B, Gait3D, and GREW. Our first comparison is between GaitDLF and representative skeleton-based methods, including PoseGait, GaitGraph, and GaitGraph2. This choice was made because of the consistency between these three methods and our own evaluation protocol.

As shown in Table 3 in the CASIA-B indoor dataset, the performance of GaitDLF is only 84.86% in the NM condition, which is not as good as GaitGraph's 87.7%, and only 70.74% in the BG condition, which is also not as good as GaitGraph's 74.8%. But comparing these several skeleton-based methods, our method was able to obtain higher results in the coat-loaded (CL) condition, which suggests that methods focusing on localized limb movements can show higher robustness when the overall skeleton is more occluded by clothing. However, the CASIA-B indoor dataset was collected in the laboratory in order to bridge the gap between laboratory studies and real-world applications. We need methods that are more applicable to wild environments.

**Table 2** Hyperparameters used in the experiments

| Dataset | Batch size | Steps | Frame | Multistep scheduler |
|---|---|---|---|---|
| CASIC-B [36] | (10, 6) | 42k | 40 | (40k) |
| Gait3D [39] | (48, 12) | 32k | 30 | (30k) |
| GREW [38] | (48, 12) | 150k | 30 | (30k, 60k, 90k, 120k) |

**Table 3** Averaged Rank-1 accuracies in percent on CASIA-B per probe angle compared with other model-based methods

| Type | Methods | View | | | | | | | | | | | Mean |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 0° | 18° | 36° | 54° | 72° | 90° | 108° | 126° | 144° | 162° | 180° | |
| NM | PoseGait [27] | 55.3 | 69.6 | 73.9 | 75.0 | 68.0 | 68.2 | 71.1 | 72.9 | 76.1 | 70.4 | 55.4 | 68.7 |
| | GaitGraph [28] | 85.3 | 88.5 | 91.0 | 92.5 | 87.2 | 86.5 | 88.4 | 89.2 | 87.9 | 85.9 | 81.9 | **87.7** |
| | GaitGraph2 [40] | 78.5 | 82.9 | 85.8 | 85.6 | 83.1 | 81.5 | 84.3 | 83.2 | 84.2 | 81.6 | 71.8 | 82.0 |
| | **GaitDLF** | 80.1 | 87.1 | 87.7 | 89.2 | 84.4 | 84.4 | 84.2 | 85.2 | 85.2 | 85.4 | 80.6 | 84.86 |
| BG | PoseGait [27] | 35.3 | 47.2 | 52.4 | 46.9 | 45.5 | 43.9 | 46.1 | 48.1 | 49.4 | 43.6 | 31.1 | 44.5 |
| | GaitGraph [28] | 75.8 | 76.7 | 75.9 | 76.1 | 71.4 | 73.9 | 78.0 | 74.7 | 75.4 | 75.4 | 69.2 | **74.8** |
| | GaitGraph2 [40] | 69.9 | 75.9 | 78.1 | 79.3 | 71.4 | 71.7 | 74.3 | 76.2 | 73.2 | 73.4 | 61.7 | 73.2 |
| | **GaitDLF** | 68.2 | 71.0 | 74.2 | 75.8 | 69.0 | 71.2 | 71.5 | 71.2 | 72.0 | 70.5 | 63.5 | 70.74 |
| CL | PoseGait [27] | 24.3 | 29.7 | 41.3 | 38.8 | 38.2 | 38.5 | 41.6 | 44.9 | 42.2 | 33.4 | 22.5 | 36.0 |
| | GaitGraph [28] | 69.6 | 66.1 | 68.8 | 67.2 | 64.5 | 62.0 | 69.5 | 65.6 | 65.7 | 66.1 | 64.3 | 66.3 |
| | GaitGraph2 [40] | 57.1 | 61.1 | 68.9 | 66.0 | 67.8 | 65.4 | 68.1 | 67.2 | 63.7 | 63.6 | 50.4 | 63.6 |
| | **GaitDLF** | 67.0 | 68.1 | 69.1 | 68.9 | 64.4 | 68.0 | 69.2 | 71.3 | 69.3 | 70.1 | 62.1 | **67.95** |

We highlight the best results by bold markup, and the name of our proposed method

On both the GREW and Gait3D datasets, our method achieves the highest recognition accuracy and has the best recognition rate compared to the best methods on both datasets, as shown in Table 4. Importantly, on the GREW dataset, our accuracy is an impressive 30% higher than that of GaitGraph2. This result highlights the ability of our method to efficiently learn complex and unique gait features from large-scale datasets, which makes it more suitable for real-world gait recognition tasks. It further shows that compared to other methods that pool over the whole skeleton to obtain global features, our proposed modeling and dynamically adaptive fusion of local limbs is clearly more advantageous.

Given the unique characteristics of these three datasets, it is possible to see that our method performs well primarily on larger, more influential datasets (e.g., Gait3D and the GREW dataset in the wild). For the smaller CASIA-B dataset, our method does not achieve the best results and performs second only to GaitGraph. This is

**Table 4** Averaged Rank-1 accuracies in percent on GREW and Gait3D compared with other model-based methods

| Methods | Publication | GREW | | | | Gait3D |
|---|---|---|---|---|---|---|
| | | Rank-1 | Rank-5 | Rank-10 | Rank-20 | |
| PoseGait [27] | PR 2020 | 0.23 | 1.05 | 2.23 | 4.28 | 0.24 |
| GaitGraph [28] | ICIP 2021 | 1.31 | 3.46 | 5.08 | 7.51 | 6.25 |
| GaitGraph2 [40] | CVPRW 2022 | 33.54 | 49.45 | 56.28 | 61.92 | 11.1 |
| **GaitDLF** | Ours | **68.77** | **82.25** | **87.25** | **90.3** | **27.2** |

We highlight the best results by bold markup, and the name of our proposed method because GaitDLF possesses a more complex design than other skeleton-based networks to cope with the challenges of complex scenarios, and thus, on the small dataset, this may have led to overfitting, making GaitDLF difficult to achieve the best performance. However, on large datasets, GaitDLF achieves very high performance.

### 4.4 Comparison with appearance-based methods

Since appearance-based methods have been predominantly used in gait recognition and the silhouette images used include gait information and other recognizable visual cues (e.g., clothing, handbags), the skeleton-based methods only use dynamic skeletal sequences of the subject to extract gait features. Skeletal-based methods, on the other hand, use only the dynamic skeletal sequence of the subject to extract gait features. The GaitDLF is then compared with various state-of-the-art appearance-based methods, such as GaitSet, GaitPart, GaitGL, and GaitBase, in order to assess its performance.

As shown in Table 5, the appearance-based state-of-the-art method outperforms the accuracy of our GaitDLF on the CASIA-B dataset. However, it is worth noting that the CASIA-B dataset was collected under the 2006 constraints and contains only 124 individuals and approximately 13K video sequences, as well as artificially formulated tilt angles, and occlusions. It is shown that in indoor environments, the appearance-based approach can exhibit the highest gait recognition accuracy due to the lack of many complicating factors.

It can be concluded from Tables 3 and 5 that the performance of all the methods is susceptible to the change in walking conditions, based on the performance of skeleton-based and silhouette-based methods in the indoor dataset. Moreover, the recognition performance of the appearance-based methods exhibits large fluctuations when the tilt angle changes, and it can be expected that such fluctuations will be even larger in complex environments. Therefore, in wild environments, skeleton-based methods, due to their greater robustness, are bound to be more advantageous.

One of the most notable features of the GREW dataset is that it is available in 2021 in a highly complex and free environment with 26K individuals and approximately 128K video sequences. The dataset is completely unconstrained, has a large number of undefined variations, and has diverse and useful perspectives. GREW also includes a variety of challenging factors such as complex backgrounds, occlusion, carrying objects, and wearing jewelry. As shown in Table 6, GaitDLF achieves a Rank-1 accuracy of 68.77%, and the Rank-1 accuracy of the state-of-the-art silhouette-based method is 60.1%. This indicates that GaitDLF outperforms the existing state-of-the-art appearance-based methods in an unconstrained wild environment.

Skeletal-based gait recognition can prioritize pose, angle, and directly relevant gait information, whereas appearance-based networks tend to emphasize subjective features such as color and texture. Therefore, for such a large dataset collected in unconstrained and complex environments, a skeleton-based approach that extracts gait features only from skeleton sequences is more appropriate than an appearance-based approach.

**Table 5** Averaged Rank-1 accuracies in percent on CASIA-B per probe type compared with other appearance-based methods

| Type | Methods | View | | | | | | | | | | | Mean |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 0° | 18° | 36° | 54° | 72° | 90° | 108° | 126° | 144° | 162° | 180° | |
| NM | GaitSet [17] | 90.8 | 97.9 | 99.4 | 96.9 | 93.6 | 91.7 | 95.0 | 97.8 | 98.9 | 96.8 | 85.8 | 95.0 |
| | GaitPart [18] | 94.1 | 98.6 | 99.3 | 98.5 | 94.0 | 92.3 | 95.9 | 98.4 | 99.2 | 97.8 | 90.4 | 96.1 |
| | GaitGL [19] | 96.0 | 98.3 | 99.0 | 97.9 | 96.9 | 95.4 | 97.0 | 98.9 | 99.3 | 98.8 | 94.0 | 97.4 |
| | GaitBase [47] | 93.9 | 98.8 | 99.6 | 98.1 | 94.0 | 91.6 | 94.9 | 98.4 | 99.3 | 98.5 | 91.8 | **97.6** |
| | **GaitDLF** | 80.1 | 87.1 | 87.7 | 89.2 | 84.4 | 84.4 | 84.2 | 85.2 | 85.2 | 85.4 | 80.6 | 84.86 |
| BG | GaitSet [17] | 83.8 | 91.2 | 91.8 | 88.8 | 83.3 | 81.0 | 84.1 | 30.0 | 92.2 | 94.4 | 79.0 | 87.2 |
| | GaitPart [18] | 89.1 | 94.8 | 96.7 | 95.1 | 88.3 | 94.9 | 89.0 | 93.5 | 96.1 | 93.8 | 85.8 | 90.7 |
| | GaitGL [19] | 92.6 | 96.6 | 96.8 | 95.5 | 93.5 | 89.3 | 92.2 | 96.5 | 98.2 | 96.9 | 91.5 | **94.5** |
| | GaitBase [47] | 91.9 | 95.5 | 96.8 | 94.7 | 90.9 | 88.9 | 91.7 | 94.9 | 96.2 | 95.5 | 86.3 | 94.0 |
| | **GaitDLF** | 68.2 | 71.0 | 74.2 | 75.8 | 69.0 | 71.2 | 71.5 | 71.2 | 72.0 | 70.5 | 63.5 | 70.74 |
| CL | GaitSet [17] | 61.4 | 75.4 | 80.7 | 77.3 | 72.1 | 70.1 | 71.5 | 73.5 | 73.5 | 68.4 | 50.0 | 70.4 |
| | GaitPart [18] | 70.7 | 85.5 | 86.9 | 83.3 | 77.1 | 72.5 | 76.9 | 82.2 | 83.8 | 80.2 | 66.5 | 78.7 |
| | GaitGL [19] | 76.6 | 90.0 | 90.3 | 87.1 | 84.5 | 79.0 | 84.1 | 87.0 | 87.3 | 84.4 | 69.5 | **83.8** |
| | GaitBase [47] | 60.2 | 77.6 | 82.8 | 78.7 | 74.8 | 72.2 | 76.1 | 78.2 | 76.8 | 72.0 | 56.9 | 77.4 |
| | **GaitDLF** | 67.0 | 68.1 | 69.1 | 68.9 | 64.4 | 68.0 | 69.2 | 71.3 | 69.3 | 70.1 | 62.1 | 67.95 |

We highlight the best mean results by marking them in bold, and the name of our proposed method

**Table 6** Averaged Rank-1, Rank-5, Rank-10, and Rank-20 accuracies in percent on GREW and averaged Rank-1 accuracies in percent on Gait3D compared with other appearance-based methods

| Methods | Publication | GREW | | | | Gait3D |
|---|---|---|---|---|---|---|
| | | Rank-1 | Rank-5 | Rank-10 | Rank-20 | |
| GaitSet [17] | AAAI 2019 | 46.28 | 63.58 | 70.26 | 76.82 | 36.7 |
| GaitPart [18] | CVPR 2020 | 44.01 | 60.68 | 67.25 | 73.47 | 28.2 |
| GaitGL [19] | ICCV 2021 | 47.28 | 63.56 | 69.32 | 74.18 | 29.7 |
| GaitBase [47] | CVPR2023 | 60.1 | 75.4 | 80.38 | 84.16 | **64.6** |
| **GaitDLF** | Ours | **68.77** | **82.25** | **87.25** | **90.3** | 27.2 |

We highlight the best results by bold markup, and the name of our proposed method

The Gait3D dataset contains only 4K individuals and 25K video sequences, which are collected in supermarkets, and the amount of data is much smaller than that of the GREW dataset, as shown in Table 6, the Rank-1 accuracy of GaitDLF is only 27.2%, and that of the state-of-the-art silhouette-based method is 64.4%. It shows that the silhouette-based and skeleton-based methods have their own advantages and usage scenarios.

The above results show that for gait recognition under strict constraints, good results can be achieved using an appearance-based approach, but for unconstrained wild environments with a large number of samples, a skeleton-based approach would be more appropriate. We believe this is caused by the fact that the skeleton sequence contains much less information than the silhouette sequence. Skeleton sequences simply include only the coordinates of each joint and the confidence level, whereas silhouette maps contain a large number of visual cues. However, large datasets have more data volume and diversity, which can help skeleton-based models better capture different variations and characteristics of gait. Therefore, on smaller datasets, the diversity of data may not be sufficient to support the skeleton based model to learn a wider range of gait variations, leading to performance degradation.

In conclusion, a larger capacity and more diverse dataset can help the skeleton-based model better capture a variety of gait variations and features. Compared with appearance-based methods, model-based methods focus only on human gait information and do not focus on rich appearance information, although it performs poorly on small datasets, in datasets with many unpredictable influences, and due to the large amount of data, skeleton-based methods are precisely able to ignore very sensitive appearance features and learn gait features. Therefore, based on this fact, this approach can be more effective in realizing its potential when applying gait recognition tasks in datasets of complexity and large data volumes or even in the real world.

### 4.5 Ablation studies

Our objective in this section is to determine the effectiveness of the modules we have designed by performing an extensive ablation analysis of each module in GaitDLF for each module. Using Gaitgraph2 as a baseline, this study is able to improve the effectiveness of the method by modifying its loss function, data augmentation, and test time augmentation (TTA) strategies in order to improve its performance. During the testing of Gaitgraph2, the left-right flip sample and the time reversal sample are used as two supplementary samples, and the three embedded data obtained are connected to facilitate the subsequent distance calculation. And the distance metric between its gallery samples and probe samples uses the cosine similarity function. On the other hand, our approach avoids the use of supplementary samples during the testing process and instead computes the distance between gallery samples and probe samples by using the Euclidean distance between the respective feature vectors of gallery samples and probe samples. Additionally, auxiliary samples are not utilized, and all frames are input to the network for testing. It is shown in the table that in comparison with the original Gaitgraph2, the adjusted baseline achieves a recognition accuracy of approximately 58%, which shows an increase of approximately 25% when compared to the original Gaitgraph2.

#### 4.5.1 Evaluation of the SLM block

Inspired by the appearance-based approach, we introduce the SLM module, which aims to capture unique motion information from different limb parts to utilize richer gait characteristics. As shown in Table 7, the Rank-1 metric is only 58.2% when only GAS is left in the network, a result that is 10.5% lower than the performance of GaitDLF. This result validates the ability of the SLM module to model different limb parts and emphasizes that focusing on the complex details of localized limb parts is not only effective for appearance-based approaches, but also for model-based approaches.

### 4.5.2 Evaluation of the DFF block

After feature extraction through the SLM module, limb-level motion features corresponding to the five limb parts are obtained. Several simple but effective methods can be used to fuse these five parts, but to further enhance the fusion of these five limb part motion features, we propose an attention-based DFF block. The goal is to dynamically adjust the attention weights of the five limb part motion features to facilitate dynamic fusion to preserve essential information and discard redundant details.

It is worth noting that both global and local channel context information are computed based on the five limb segments during the generation of attention weights. As shown in Table 7, the accuracy of the DFF block when removed from the LDS is 66.1%, a result that is 2.6% lower than the performance of GaitDLF. This result further confirms the efficacy of the proposed DFF block in fusing localized limb segments.

### 4.5.3 Evaluation of the GAS

We have verified the effectiveness of the two modules, of LDS, and we intend to remove the original GAS, according to Table 7, the performance of the network with only LDS is 68.1%, and this result is 0.67% lower than the performance of GaitDLF. It shows that GAS and LDS together can make the network achieve better performance.

**Table 7** Rank-1 accuracy of the network on the GREW dataset, with each of the proposed modules being removed in turn to assess the validity of each module

| Methods | GAS | LDS | | Rank-1 | Rank-5 | Rank-10 | Rank-20 |
|---|---|---|---|---|---|---|---|
| | | SLM | DFF | | | | |
| GaitGraph2 | | | | 33.54 | 49.45 | 56.28 | 61.92 |
| GaitBase | | | | 60.1 | 75.4 | 80.38 | 84.16 |
| our | ✓ | ✓ | ✓ | **68.77** | **82.77** | **87.25** | **90.3** |
| | ✓ | ✓ | | 66.1 | 81.8 | 86.6 | 90 |
| | ✓ | | | 58.2 | 76.2 | 82.1 | 86.2 |
| | | ✓ | ✓ | 68.1 | 82.3 | 86.6 | 89.8 |

We highlight the best results by bold markups

## 4.6 More experiments

GaitDLF will be the subject of more experiments in this section. Besides testing the two modules above, we will also evaluate the role these two modules play in improving the feature extraction capability of GaitDLF by comparing the baseline model without the integrated SLM and DFF modules with the full GaitDLF model. Moreover, we will analyze GaitDLF's complexity and compare it with currently available skeleton-based gait recognition methods.

### 4.6.1 Feature visualization

For the experiments on the GREW dataset, we randomly selected 80 different labels, each containing 12 samples of skeleton sequences, for a total of 960 samples.These samples were input into Baseline and our GaitDLF, respectively, and the respective output gait feature distributions were subsequently visualized and analyzed using t-SNE [48]. As can be seen from the results in Fig. 8, in the comparison of Baseline and GaitDLF feature distributions, we note that Baseline shows some ability in distinguishing between differently labeled samples, even though the feature distributions of these samples are relatively concentrated in the feature space. This concentrated distribution may imply that the features of differently labeled samples are spatially closer together, and may pose a challenge to the network's ability to distinguish between these labels. In contrast, GaitDLF shows slightly better performance in feature extraction, being able to distinguish between samples with different labels with some degree of clarity, as evidenced by the fact that the distribution of samples with different labels is slightly more spread out in the feature space. This also reflects the advantage of GaitDLF in extracting gait features to some extent and also shows that the two modules we developed are effective from the perspective of feature visualization.

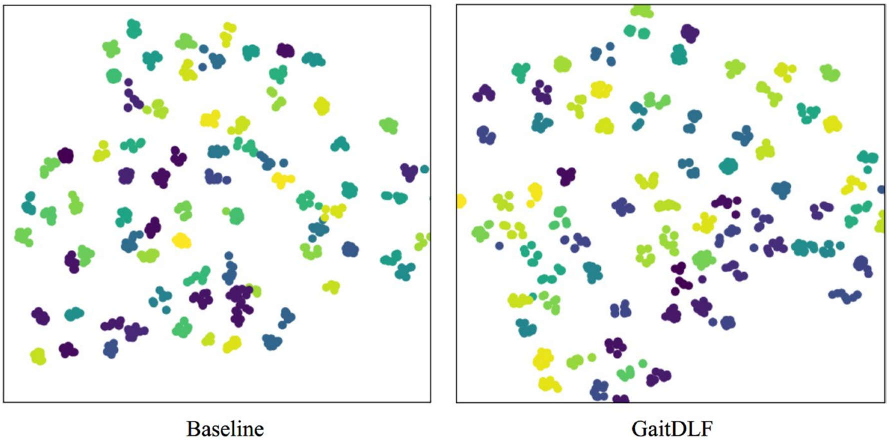

**Fig. 8** Results of visual analysis of a sample of 80 labels selected from the GREW dataset using t-SNE. Same color indicates the same label; different colors indicate different labels

### 4.7 Complexity analysis

We perform a complexity analysis of GaitDLF with the state-of-the-art skeleton-based methods, as shown in Table 8, the parameter of GaitDLF is 0.94 M, which is just 0.18 M more than the 0.76 M of GaitGraph2, while their FLOPs are comparable, indicating that GaitDLF optimizes the computational burden while maintaining a similar model design while maintaining a similar computational burden. More importantly, GaitDLF shows excellent performance on two important outdoor datasets, Gait3D and GREW, with the top-ranked accuracy reaching 27.2% on Gait3D and 68.77% on GREW, which clearly outperforms the other comparison methods. In contrast, Gaitgraph, despite being more lightweight in terms of model size, with a parameter count of only 0.35 M, does not perform as well in terms of actual performance.

## 5 Conclusion

In this paper, we introduce a network that is more adapted to the field environment and focuses on both global and localized limb motion, called the dynamic local fusion network (GaitDLF). Unlike existing skeleton-based approaches, our GaitDLF not only focuses on the motion information of the overall skeleton, but is also able to extract a more discriminative representation of gait features from the local limb motion information. As GaitDLF extracts features at different scales from both local detail streams and global-aware streams, it is able to provide information that is more comprehensive in terms of representation of features.

In addition, a dynamic feature fusion mechanism is used to dynamically integrate the details of different limb parts and retain more critical motion features. Several studies on the CASIA-B, the GREW, and the Gait3D datasets have shown that GaitDLF has a superior discriminative power to other skeleton-based methods, while still meeting the requirements for gait recognition in complex environments. GaitDLF also shows superior discriminative power to skeleton-based methods on the CASIA-B, GREW, and Gait3D datasets. Our experiments reveal the robustness of the skeleton-based approach in datasets with complex environmental factors. Skeleton-based gait recognition methods are better suited for complex and large data sets than appearance-based methods.

**Table 8** Comparison of performance and computational complexity of skeleton-based gait recognition methods on Gait3D and GREW outdoor datasets

| Methods | Gait3D<br>Rank-1 | GREW<br>Rank-1 | params | FLOPs |
|---|---|---|---|---|
| Gaitgraph | 6.25 | 1.31 | 0.35M | 0.0753G |
| Gaitgraph2 | 11.1 | 33.54 | 0.76M | 0.1936G |
| GaitDLF | **27.2** | **68.77** | **0.94M** | **0.1939G** |

We highlight the best results by bold markups

Based on these results, GaitDLF has the potential to be applied in the medical field, especially in the prediction of Parkinson's disease. Parkinson's disease is a neurological disorder that typically affects a patient's gait and motor skills. By using GaitDLF or similar gait recognition technology in field scenarios, physicians and researchers can characterize the final gait of potential patients in the community to identify the disease. This early detection may help with early intervention and treatment, thereby improving the quality of life for patients. In the field of public security and safety, GaitDLF can be used to recognize and track specific pedestrians. This can achieve high accuracy for monitoring pedestrians in cameras, even in complex environments such as crowded locations or city streets.

**Author contributions** SW contributed to the conceptualization, methodology, writing—original draft and writing—review and editing; WL involved in the software, validation, and resources; FW contributed to the supervision, funding acquisition, and resources; CW contributed to the supervision and resources; NX contributed to writing—review and editing, supervision. All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

**Funding** This work is funded by the National Natural Science Foundation of China under Grant No. 61772180 and the Hubei Provincial Natural Science Foundation No. 2023BCB041.

## Declarations

**Conflict of interest** All authors declare that they have no confict of interest.

**Ethical approval** This article does not contain any studies with human participants or animals performed by any of the authors.

## References

1. Ding W, Abdel-Basset M, Hawash H, Moustafa N (2022) Interval type-2 fuzzy temporal convolutional autoencoder for gait-based human identification and authentication. Inf Sci 597:144–165
2. Zhang Z, Wei S, Xi L, Wang C (2024) Gaitmgl: multi-scale temporal dimension and global-local feature fusion for gait recognition. Electronics 13(2):257
3. Wan R, Xiong N, Hu Q, Wang H, Shang J (2019) Similarity-aware data aggregation using fuzzy c-means approach for wireless sensor networks. EURASIP J Wirel Commun Netw 2019:1–11
4. Shen X, Yi B, Liu H, Zhang W, Zhang Z, Liu S, Xiong N (2019) Deep variational matrix factorization with knowledge embedding for recommendation system. IEEE Trans Knowl Data Eng 33(5):1906–1918
5. Shen Y, Fang Z, Gao Y, Xiong N, Zhong C, Tang X (2019) Coronary arteries segmentation based on 3d fcn with attention gate and level set function. IEEE Access 7:42826–42835

6. Zhang W, Zhu S, Tang J, Xiong N (2018) A novel trust management scheme based on dempster-shafer evidence theory for malicious nodes detection in wireless sensor networks. J Supercomput 74:1779–1801
7. Wang Y, Fang W, Ding Y, Xiong N (2021) Computation offloading optimization for uav-assisted mobile edge computing: a deep deterministic policy gradient approach. Wireless Netw 27(4):2991–3006
8. Wang J, Jin C, Tang Q, Xiong NN, Srivastava G (2020) Intelligent ubiquitous network accessibility for wireless-powered mec in uav-assisted b5g. IEEE Trans Netw Sci Eng 8(4):2801–2813
9. Huang S, Zeng Z, Ota K, Dong M, Wang T, Xiong NN (2020) An intelligent collaboration trust interconnections system for mobile information control in ubiquitous 5g networks. IEEE Trans Netw Sci Eng 8(1):347–365
10. Babaee M, Zhu Y, Köpüklü O, Hörmann S, Rigoll G (2019) Gait energy image restoration using generative adversarial networks. In: 2019 IEEE International Conference on Image Processing (ICIP), IEEE, pp 2596–2600
11. Feng Y, Li Y, Luo J (2016). Learning effective gait features using lstm. In: 2016 23rd International Conference on Pattern Recognition (ICPR), IEEE, 325–330
12. Han J, Bhanu B (2005) Individual recognition using gait energy image. IEEE Trans Pattern Anal Mach Intell 28(2):316–322
13. Wang L, Tan T, Ning H, Hu W (2003) Silhouette analysis-based gait recognition for human identification. IEEE Trans Pattern Anal Mach Intell 25(12):1505–1518
14. Song C, Huang Y, Huang Y, Jia N, Wang L (2019) Gaitnet: an end-to-end network for gait based human identification. Pattern Recognit 96:106988
15. Wolf T, Babaee M, Rigoll G (2016) Multi-view gait recognition using 3d convolutional neural networks. In: 2016 IEEE International Conference on Image Processing (ICIP), IEEE, pp 4165–4169
16. Zhao A, Li J, Ahmed M (2020) Spidernet: a spiderweb graph neural network for multi-view gait recognition. Knowl Based Syst 206:106273
17. Chao H, He Y, Zhang J, Feng J (2019) Gaitset: regarding gait as a set for cross-view gait recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol 33. pp 8126–8133
18. Fan C, Peng Y, Cao C, Liu X, Hou S, Chi J, Huang Y, Li Q, He Z (2020) Gaitpart: temporal part-based model for gait recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 14225–14233
19. Lin B, Zhang S, Yu X (2021) Gait recognition via effective global-local feature representation and local temporal aggregation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 14648–14656
20. Huang Z, Xue D, Shen X, Tian X, Li H, Huang J, Hua X.-S (2021) 3d local convolutional neural networks for gait recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 14920–14929
21. Fang H.-S, Xie S, Tai Y.-W, Lu C (2017) Rmpe: Regional multi-person pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision, pp 2334–2343
22. Cao Z, Simon T, Wei S-E, Sheikh Y (2017) Realtime multi-person 2d pose estimation using part affinity fields. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 7291–7299
23. Liao R, Cao C, Garcia E.B, Yu S, Huang Y (2017) Pose-based temporal-spatial network (ptsn) for gait recognition with carrying and clothing variations. In: Biometric Recognition: 12th Chinese Conference, CCBR 2017, Shenzhen, China, October 28-29, 2017, Proceedings vol 12. Springer, pp 474–483
24. Sokolova A, Konushin A (2019) Pose-based deep gait recognition. IET Biom 8(2):134–143
25. Xu C, Makihara Y, Li X, Yagi Y (2023) Occlusion-aware human mesh model-based gait recognition. IEEE Trans Inform Forens Secur 18:1309–1321
26. Li X, Makihara Y, Xu C, Yagi Y (2022) Multi-view large population gait database with human meshes and its performance evaluation. IEEE Trans Biom Behav Identity Sci 4(2):234–248
27. Liao R, Yu S, An W, Huang Y (2020) A model-based gait recognition method with body pose and human prior knowledge. Pattern Recognit 98:107069
28. Teepe T, Khan A, Gilg J, Herzog F, Hörmann S, Rigoll G (2021) Gaitgraph: graph convolutional network for skeleton-based gait recognition. In: 2021 IEEE International Conference on Image Processing (ICIP), IEEE, 2314–2318
29. Peng Y, Ma K, Zhang Y, He Z (2023) Learning rich features for gait recognition by integrating skeletons and silhouettes. Multim Tools Appl 83(3):7273–7294

30. Xiong N, Han W, Vandenberg A (2012) Green cloud computing schemes based on networks: a survey. IET Commun 6(18):3294–3300
31. Zeng Y, Xiong N, Park JH, Zheng G (2010) An emergency-adaptive routing scheme for wireless sensor networks for building fire hazard monitoring. Sensors 10(6):6128–6148
32. Huang X, Zhu D, Wang H, Wang X, Yang B, He B, Liu W, Feng B (2021) Context-sensitive temporal feature learning for gait recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 12909–12918
33. Liang J, Fan C, Hou S, Shen C, Huang Y, Yu S (2022). Gaitedge: beyond plain end-to-end gait recognition for better practicality. In: European Conference on Computer Vision, Springer, pp 375–390
34. Fan C, Hou S, Wang J, Huang Y, Yu S (2022) Learning gait representation from massive unlabelled walking videos: a benchmark. arXiv:2206.13964
35. An W, Yu S, Makihara Y, Wu X, Xu C, Yu Y, Liao R, Yagi Y (2020) Performance evaluation of model-based gait on multi-view very large population database with pose sequences. IEEE Trans Biom Behav Identit Sci 2(4):421–430
36. Yu S, Tan D, Tan T (2006) A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In: 18th International Conference on Pattern Recognition (ICPR'06), vol 4. IEEE, pp 441–444
37. Deng M, Wang C (2018) Human gait recognition based on deterministic learning and data stream of microsoft kinect. IEEE Trans Circuits Syst Video Technol 29(12):3636–3645
38. Zhu Z, Guo X, Yang T, Huang J, Deng J, Huang G, Du D, Lu J, Zhou, J (2021) Gait recognition in the wild: a benchmark. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 14789–14799
39. Zheng J, Liu X, Liu W, He L, Yan C, Mei T (2022) Gait recognition in the wild with dense 3d representations and a benchmark. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 20228–20237
40. Teepe T, Gilg J, Herzog F, Hörmann S, Rigoll G (2022) Towards a deeper understanding of skeleton-based gait recognition. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE, pp 1568–1576
41. Li X, Makihara Y, Xu C, Yagi Y, Yu S, Ren M (2020) End-to-end model-based gait recognition. In: Proceedings of the Asian Conference on Computer Vision
42. Song Y.-F, Zhang Z, Shan C, Wang L (2020) Stronger, faster and more explainable: A graph convolutional baseline for skeleton-based action recognition. In: Proceedings of the 28th ACM International Conference on Multimedia, pp 1625–1633
43. Yan S, Xiong Y, Lin D (2018) Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence, p 32
44. Wu Z, Huang Y, Wang L, Wang X, Tan T (2016) A comprehensive study on cross-view gait based human identification with deep CNNS. IEEE Trans Pattern Anal Mach Intell 39(2):209–226
45. Sun K, Xiao B, Liu D, Wang J (2019) Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 5693–5703
46. Robbins H, Monro S (1951) A stochastic approximation method. Ann Math Stat 22:400–407
47. Fan C, Liang J, Shen C, Hou S, Huang Y, Yu S (2023) Opengait: revisiting gait recognition towards better practicality. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 9707–9716
48. Maaten L, Hinton G (2008) Visualizing data using t-sne. J Mach Learn Res 9(11):2579–2605

## Authors and Affiliations

**Siwei Wei[1,2] · Weijie Liu[1] · Feifei Wei[3] · Chunzhi Wang[1] · Neal N. Xiong[4]**

✉ Feifei Wei
waosfengw@whut.edu.cn

Weijie Liu
102211199@hbut.edu.cn

Chunzhi Wang
chunzhiwang@hbut.edu.cn

Neal N. Xiong
xiongnaixue@gmail.com

[1] School of Computer Science, Hubei University of Technology, Wuhan 430000, China

[2] CCCC Second Highway Consultants Co. Ltd., Wuhan 430056, China

[3] School of Information Management, Hubei University of Economics, Wuhan 430070, China

[4] Northeastern State University, 611 N. Grand Ave, Tulsa, OK 74464, USA