Multi-view gait recognition with joint local multi-scale and global contextual spatio-temporal features

Wenzhe Zhai, Haomiao Li, Chaoqun Zheng, Xianglei Xing*

***Abstract*—Existing gait recognition methods are capable of extracting rich spatial gait information but often overlook fine-grained temporal features within local regions and temporal contextual information across different sub-regions. Considering gait recognition as a fine-grained recognition task and each individual exhibits uniqueness in their movements across different temporal sequences, we propose a local multi-scale and global contextual spatio-temporal (LMGCS) network for gait recognition. It divides the whole gait sequence into sub-sequences with multiple spatio resolutions and extracts multi-scale temporal features. We extract the temporal context information of different sub-sequences with the transformer, and all sub-sequences are fused to form global features. Furthermore, the loss function that combines the triplet loss function and cross-entropy loss function is utilized to prompt the proposed model to fulfill the gait recognition. The proposed method achieved state-of-the-art results on two popular public datasets. It achieved rank-1 accuracy of 98.0%, 95.4%, and 85.0% on the three walk states of the CASIA-B dataset and 90.9% on the OU-MVLP dataset.**

Keywords*—Gait recognition, Fine-grained recognition, Multi-scale feature, Temporal context information.

I. Introduction

GAIT recognition has garnered significant interest from researchers and engineers in computer vision due to its promising applications in human identity verification [1]–[3]. Although gait recognition has attracted the interest of many researchers [4]–[6], it is still a challenging task because gait features are greatly influenced by clothing, environment, and viewing angles. The differences between the gait sequences of different individuals are very subtle.

To alleviate the difficulty of cross-view recognition, the methods from the generative perspective attempt to convert gait sequences of different angles to the same angle or to use a generative model to expand the gait dataset [7]–[9]. For example, Ben *et al.* [10] proposed a coupled patch alignment algorithm that effectively matches a pair of gaits across different views. Huang *et al.* [3] proposed a hierarchical feature aggregation strategy for discriminative feature extraction. These methods aim to extract more discriminative gait feature representations from different viewing angles. Previous methods for gait feature extraction usually consist of spatial feature extraction and temporal feature extraction. The key factor in determining the recognition effect is how to extract effective spatial and temporal features simultaneously.

Currently, many methods have been proposed to extract spatial features [11]–[14]. They consider that different body parts have different motion patterns and divide the gait sequence into different sub-regions in space. Each sub-region is sent to a different network to extract its spatial features based on different motion patterns. For example, Lin *et al.* [14] utilized the global and local spatial features simultaneously and fused them to obtain more robust feature representations. This method focuses solely on spatial features and does not incorporate temporal information, which limits its ability to capture the dynamic characteristics of the data.

To mitigate the limitations mentioned above, some other methods focus on feature extraction in the temporal dimension [13], [15]–[18]. Huang *et al.* [1] proposed a hierarchical feature aggregation strategy for discriminative feature extraction. They used multi-layer 3D convolution or LSTM to extract the temporal features of gait, among which treat the time series as a whole and extract global temporal information [13], [15], [16]. Meanwhile, some methods [17], [18] focus on capturing fine-grained local temporal information in gait by extracting temporal features at multiple time scales and fusing these multi-scale features to obtain more discriminative representations. However, these methods either only focus on global temporal features or only focus on local temporal features. We believe that as gait is a fine-grained movement pattern, extracting gait features only from a global perspective would ignore local detailed temporal information that is more representative of identity. Similarly, focusing only on local temporal features will ignore the global context information of different regions, which is important for discrimination.

By observing the entire gait movement, we consider that people tend to focus on specific local frames, where they pay closer attention to the fine-grained gait features in each segment and derive identity information from consecutive frames. Meanwhile, more discriminative gait features can be obtained by integrating the contextual information from different local segments. According to the voting results of 7 volunteers, people usually focus on local segments where the foot reaches its maximum height off the ground, where the foot just touches the ground, or where the arm is at its highest point, *etc.* They then integrate the contextual information from these segments

The work is supported by the National Natural Science Foundation of China No. 62076078 and the CAAI-Huawei MindSpore Open Fund No. CAAIXSJLJJ-2020-033A. (Corresponding author: Xianglei Xing)

Wenzhe Zhai, Haomiao Li, Caoqun Zheng, and Xianglei Xing are with the College of Intelligent Systems Science and Engineering, Harbin Engineering University, Harbin, 150001, China. e-mail: wenzhezhai@163.com, lhm@hrbeu.edu.cn, chaoqunzheng@hrbeu.edu.cn, xingxl@hrbeu.edu.cn

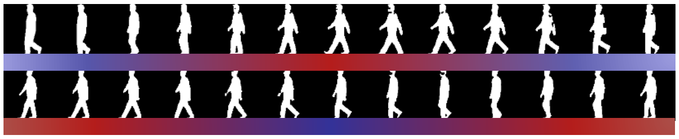

Fig. 1. The gait silhouette images extracted from a 90° view of pedestrians walking normally in the CASIA-B dataset. The color red signifies a higher concentration of personal identity information within the current gait data, while blue indicates a lower concentration of personal identity data.

to make their final judgment. As shown in Fig. 1, the red bar indicates that the current gait moment contains a higher amount of personal identity information, while the blue bar signifies a lesser amount of such information.

The above observations inspire us to propose a local multi-scale and global contextual spatio-temporal (LMGCS) network to mine both the local multi-scale features and the global spatio-temporal features with contextual information. Firstly, since gait recognition is a fine-grained classification problem, the information crucial for identity recognition often appears in a few key adjacent frames. Therefore, we propose a local multi-resolution feature extractor (LMFE) to mine the fine-grained temporal features. We divide the entire gait sequence into many short sub-sequences. Within each sub-sequence, we construct a special structure to focus on the local region and extract fine-grained identity information. A multiple scale scheme is utilized to extract rich local features. After each branch extracted its own fine-grained features, we designed a multi-branch feature fusion (MFF) module to combine the features from different resolutions, which enriches the feature diversity and the representability of different branches. Then, we emphasize the importance of the contextual information for different sub-sequences. Since different persons' walking habits are different, the gait contextual relationship should be unique for each person during the walking process, and learning this gait contextual relationship is beneficial for the task of gait recognition. Previous research usually used max-pooling to extract global temporal features [12], [19], which ignored the contextual relationship between sub-sequences. Unlike previous methods, we propose a global self-attention feature extractor (GSFE) with the transformer structure to learn the contextual information between sub-sequences and employ the learned contextual relationship to adaptively fuse the feature from sub-sequences. Finally, we employ the triplet loss and cross-entropy loss to refine the surgery on the subspace, which enhances both inter-class and intra-class relationships. The main contributions of this paper can be summarized as follows.

1) A local multi-resolution feature extractor (LMFE) is proposed to capture the fine-grained temporal features of each sub-sequence under different resolutions, which is customized for the time series fine-grained classification problem of gait recognition.

2) To enhance the diversity and representability of spatio-temporal features, we propose a multi-branch feature fusion (MFF) module. This module facilitates features from one resolution branch to fuse with features from branches of different resolutions based on the positions of their sub-sequences.

3) A global self-attention feature extractor (GSFE) is introduced to capture the contextual information between sub-sequences globally and employ the learned contextual relationship to extract more discriminative gait features.

The remainder of the paper is structured as follows. In Section II, the related works related to the paper are presented. In Section III, we detail the overall structure and implementation details of the proposed method. Section IV presents the comprehensive experimental analyses and the ablation experiments for each module. The conclusion is presented in Section V.

II. Related Work

This section provides a comprehensive review of recent gait recognition methods. Following the previous works [10], [20], we categorize the methods into two groups, namely spatial feature-based methods and spatio-temporal feature-based methods.

A. Gait Recognition Using Spatial Features

To extract spatial gait features invariant to viewing changes, Wu *et al.* [21] randomly selected raw silhouettes from the gait sequence and accumulated the features extracted by the CNN to compute the set-level representations. The method ignores the motion information from adjacent frames and focuses on capturing the co-occurrences and frequencies of discriminative features. Wu *et al.* [22] accumulated the entire gait sequence into Gait Energy Images (GEIs) and employed CNN to extract features from pairs of GEIs to predict similarity for cross-view gait recognition.

To address the challenges caused by view changes, Xing *et al.* [20] developed the complete canonical correlation analysis model to extract the common spatial features across different viewing angles, which preserves not only the completed correlation information but also the effective discriminant information among the original sets of features. Huang *et al.* [1] introduced a parallel-insight convolution layer, which was integrated with a spatial-temporal dual-attention unit to capture global spatial-temporal information. Moreover, Takemura *et al.* [23] constructed both the Siamese network for the verification task and the triplet network for the identification task. For the large view differences, they have observed that the CNN architectures exhibit insensitivity towards spatial displacement, as the disparity between a matching pair is computed only at the final layer after traversing through convolution and max pooling layers. By contrast, when dealing with subject differences under small view variations, they used CNN architectures to calculate the difference between matching pairs at the input level to enhance their sensitivity towards spatial displacement.

Inspired by the ideas above, the GLN [24] network further leveraged the inherent feature pyramid fusion of deep CNNs to enhance gait representation. Specifically, the silhouette-level and set-level features obtained from various stages are integrated with the lateral connections top-down. To avoid the gait period detection, Gaitset [19] treated the silhouettes in a gait sequence as a disordered set, which is permutation invariant and employed set pooling to fuse the features from the silhouette set. The horizontal pyramid mapping is also employed to learn part representations, which is helpful for gait recognition. Although the representations of both the GEIs and the set pooling consider the information of the whole gait sequence, these approaches primarily emphasize spatial information while neglecting crucial temporal contextual details that contain valuable identification information necessary for analyzing time series data.

B. Gait Recognition Using Spatio-temporal Features

In addition to spatial information, it is important to emphasize on representational temporal information. In early studies, LSTM [13] and 3D convolutional [15] networks are commonly employed to extract temporal features. GaitNet [25] employed the LSTM networks to integrate the pose features disentangled from the appearance features by an autoEncoder framework to capture the temporal dynamics of the whole gait sequence. Wolf *et al.* [15] introduced the 3D convolutions to capture the spatio-temporal features for gait recognition in multiple views.

However, the method overlooks the fine-grained features of local regions. To address this issue, GaitPart [12] utilized a micro-motion capture module to extract fine-grained temporal information within a localized region. Wang *et al.* [26] proposed an end-to-end 3D gait recognition framework named PointGait, which can directly capture informative gait features from point cloud data. Moreover, Huang *et al.* [16] introduced the 3D local operations to extract 3D features of different body parts from the gait sequence with adaptive spatial and temporal parameters, such as scales, locations, and lengths. In addition, Lin *et al.* [17] partitioned the entire gait sequence into sub-sequences and utilized 3D convolution to extract local spatio-temporal features from each sub-sequence. Furthermore, Huang *et al.* [18] extracted temporal features at multiple scales and employed soft attention to aggregate temporal signals to improve spatio-temporal modeling ability.

Extracting gait features solely from a global perspective would overlook local detailed temporal information that is more indicative of identity. Conversely, focusing exclusively on local temporal features would disregard the contextual information of different regions and neglect some valuable discriminative information. The proposed method captures the uniqueness of gait more comprehensively by combining local multi-scale temporal features with global contextual spatio-temporal features to improve recognition accuracy.

III. Method

In this section, we first introduce the framework of the local multi-scale and global contextual spatio-temporal (LMGCS) network. Then, we elaborate on the fundamental components of the proposed method, including the local multi-resolution feature extractor (LMFE), the multi-branch feature fusion (MFF), and the global self-attention feature extractor (GSFE). Finally, we present the loss function of the entire network.

A. Overall framework

The overall framework of the proposed model is illustrated in Fig. 2. It primarily consists of two processes: (1) Spatio-temporal feature extraction process, (2) Local fine-gained global contextual-temporal feature extraction process. The input gait data ($I_j \in R^{h_1 * w_2}, j = 1, 2, \ldots, n$) consists of n frames, where each frame represents a gait silhouette with a height of h and a width of w. To begin with, the input I is fed into the 3D convolutional backbone network $\text{E3D}(\cdot)$ to extract spatio-temporal features, yielding $F = \text{E3D}(I), F \in R^{n*c*h_2*w_2}$. Here, h_2 and w_2 represent the spatial dimensions of feature F in horizontal and vertical directions. c represents the number of channels. We apply the horizontal pyramid mapping to the feature F, which is subsequently partitioned into m local regions in the spatial dimensions. More specifically, the global average pooling $\text{GAP}(\cdot)$ and global maximum pooling $\text{GMP}(\cdot)$ [27] are jointly used to reduce the feature from $h_2 * w_2$ to m dimensions. The formula is as follows,

$$M = \text{GAP}(F) \oplus \text{GMP}(F), \tag{1}$$

where M represents the intermediate feature.

We divide the feature $M \in R^{n*c*m}$ into multiple subsequences in the temporal dimension and extract the fine-grained features by the LMFE module. The LMFE designs multiple branches to mine the fine-grained features from various temporal resolutions. These fine-grained features are subsequently fed into the MFF module, which improves the representative capacity of each branch through information interaction across various resolutions. The GSFE module further captures the global context information from different branches to adaptively complement the local fine-grained features. Details of the LMFE, MFF, and GSFE will be introduced in Sections III-B, III-C, and III-D. The notations are summarized in Table I.

TABLE I. Mathematical Notations.

Notation	Description
I	Input image
$\text{GAP}(\cdot)$	Global average pooling
$\text{GMP}(\cdot)$	Global maximum pooling
$\circ$	2D convolution operation
$\text{Re}(\cdot)$	ReLU activation
$\text{S}(\cdot)$	Softmax function
$\oplus$	Concatenation operator
$\odot$	Multiplication operator

B. Local Multi-resolution Feature Extractor

The structure of the LFME module is illustrated in Fig. 3. The LMFE module prioritizes diverse local motion patterns, which extract the refined temporal features. To extract fine-grained temporal features locally, we partition the entire gait feature sequence into n overlapping subsequences based on scale parameters, where each subsequence is denoted as $M_s^i \in R^{s*c*m}$, with $i \in 1, 2, \ldots, n$. Here, s denotes the number of frames in each subsequence, $*$ indicates the concatenation of tensor dimensions, and i represents the index of the i-th subsequence. Then, the temporal fine-grained features can be extracted from the local subsequence as $T_s^i = \text{Conv1D}(M)$,

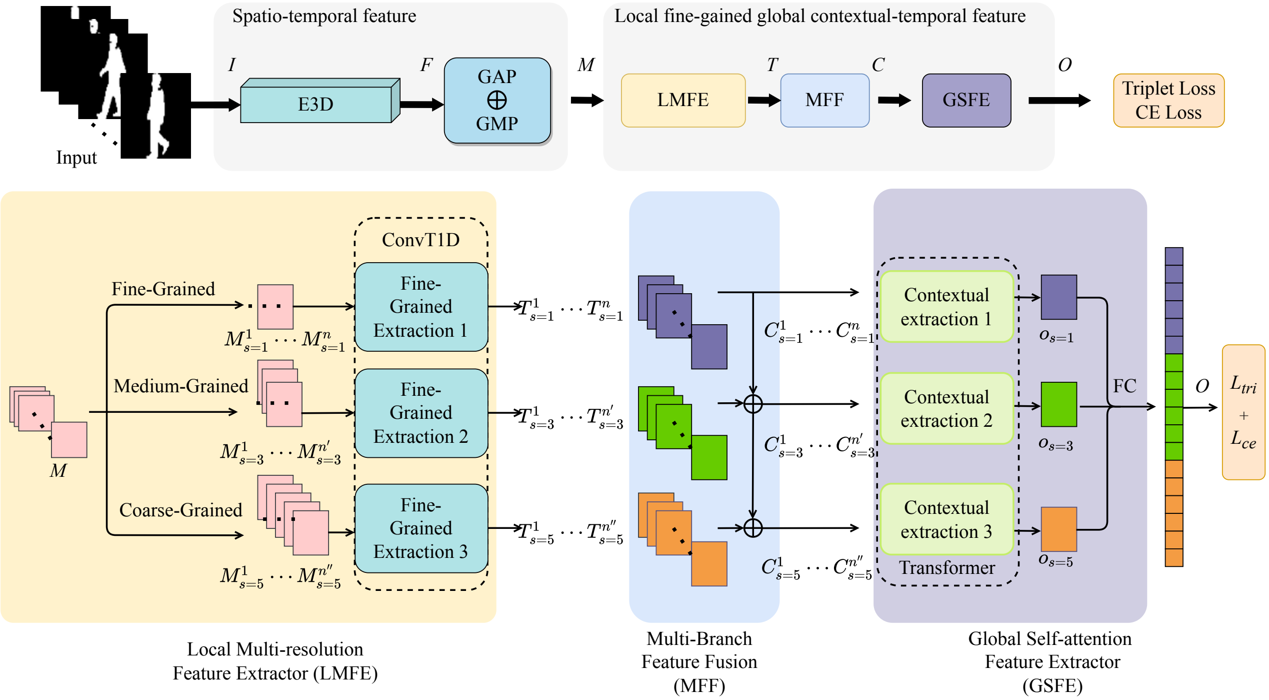
Spatio-temporal feature
Local fine-gained global contextual-temporal feature
Input
I
E3D
F
GAP ⊕ GMP
M
LMFE
T
MFF
C
GSFE
O
Triplet Loss
CE Loss
ConvT1D
Fine-Grained
Medium-Grained
Coarse-Grained
M
Fine-Grained Extraction 1
Fine-Grained Extraction 2
Fine-Grained Extraction 3
$M^1_{s=1} \cdots M^n_{s=1}$
$M^1_{s=3} \cdots M^{n'}_{s=3}$
$M^1_{s=5} \cdots M^{n''}_{s=5}$
$T^1_{s=1} \cdots T^n_{s=1}$
$T^1_{s=3} \cdots T^{n'}_{s=3}$
$T^1_{s=5} \cdots T^{n''}_{s=5}$
$C^1_{s=1} \cdots C^n_{s=1}$
$C^1_{s=3} \cdots C^{n'}_{s=3}$
$C^1_{s=5} \cdots C^{n''}_{s=5}$
Contextual extraction 1
Contextual extraction 2
Contextual extraction 3
Transformer
$o_{s=1}$
$o_{s=3}$
$o_{s=5}$
FC
O
L_{tri} + L_{ce}
Local Multi-resolution Feature Extractor (LMFE)
Multi-Branch Feature Fusion (MFF)
Global Self-attention Feature Extractor (GSFE)

Fig. 2. The overall flowchart of the joint local multi-scale and global contextual spatio-temporal (LMGCS) network. It encompasses the local multi-resolution feature extractor (LMFE), the multi-branch feature fusion (MFF), and the global self-attention feature extractor (GSFE).

where $T^i_s \in \mathrm{R}^{1*c*m}$ and Conv1D$(\cdot)$ describe the feature extraction module. In addition, the module comprises two 1D convolutions with kernel sizes $\{1, 3, 5\}$ and is augmented by the residual block. It is computed as follows,

$$T^i_s = \mathrm{Re}\left(M^i_s \circ k^1 + b^1\right) + \mathrm{Re}\left(\mathrm{Re}\left(M^i_s \circ k^1 + b^1\right) \circ k^2 + b^2\right) \quad (2)$$

where $\circ$ represents the 2D convolution operation. k^1 and k^2 represent the parameters of convolution from the first and second layers. b^1 and b^2 represent the bias parameters of the first and second layers. Re$(\cdot)$ represents the ReLU activation function. We have devised three branches that segment sub-sequences at varying resolutions to improve the extraction of more comprehensive local temporal features.

C. Multi-Branch Feature Fusion

After extracting local temporal features at various resolutions, the mined multi-branch features are forwarded to the MFF module, which facilitates diverse fusion across multiple branches. The MFF module enables each branch to integrate fine-grained features with multi-resolution features from other branches during the merging process while emphasizing its fine-grained information at its respective resolution. This further enriches the feature expression ability of each branch. The following fusion strategy is introduced to integrate features from each corresponding sub-region across three branches. The formula is as follows,

$$C^i_{s'} = \mathrm{H}\left(T^i_{s=1}, T^i_{s=3}, T^i_{s=5}\right), i \in \{1, 2, ..., n\}, \quad (3)$$

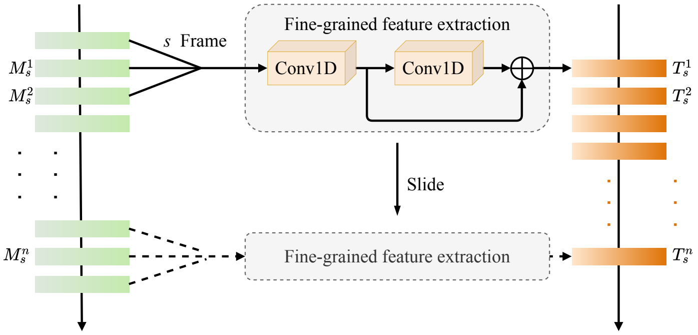

Fig. 3. The detailed structure of local multi-resolution feature extractor (LMFE). It extracts detailed temporal information from each sub-sequence using two 1D convolutions and a residual block to mine the fine-grained temporal features from multi-resolution.

where i represents the i-th subregion and $T^i_s \in \mathrm{R}^{1*c*m}$, $C^i_{s'} \in \mathrm{R}^{1*c*m}$. H$(\cdot)$ represents the multi-branch feature fusion function. The MFF module has two variant structures. 1) Static structure: In the static approach, a straightforward scheme utilizes the summation of branch features as the new branch features. Specifically, we aggregate the fine-grained features into the coarse-grained features, ensuring that the combined output of each branch equals the sum of all original features from branches with a resolution less than or equal to its granularity. As shown in Eq. 4, the first formula denotes that the fused feature from the fine-grained branch remains unchanged and is equivalent to the original fine-grained feature, the second formula employs that the fused medium-grained branch feature is obtained by summing both the original fine-grained and medium-grained branch features, and the third formula represents that the fused coarse-grained branch feature results from combining the original fine-grained, medium-grained, and coarse-grained branch features. This approach is efficient as it does not increase the number of network parameters while simultaneously reducing computational overhead. It is computed as follows,

$$\begin{aligned} C_{s=1}^{i} &= T_{s=1}^{i} \\ C_{s=3}^{i} &= T_{s=1}^{i} \oplus T_{s=3}^{i} \\ C_{s=5}^{i} &= T_{s=1}^{i} \oplus T_{s=3}^{i} \oplus T_{s=5}^{i} \end{aligned} \tag{4}$$

where $\oplus$ denotes the concatenation operator.

2) Attention-based structure: The attention mechanism integrates features from diverse branches, as depicted in Fig. 4. The features from all the branches are concatenated $T^{i} = \left[T_{s=1}^{i}, T_{s=3}^{i}, T_{s=5}^{i}\right] \in \mathrm{R}^{3*c*m}$ and input into the 1D convolution for fusion. The attention weights for each branch are computed by utilizing the correlation between the features of individual branches and the fused features. Finally, the feature of each branch is multiplied by its corresponding attention weight and then combined with the input feature by the residual connection to obtain the updated feature $C^{i} = \left(C_{s=1}^{i}, C_{s=3}^{i}, C_{s=5}^{i}\right) \in \mathrm{R}^{3*c*m}$. It is formulated as follows,

$$C^{i} = \mathrm{S}\left(T^{i} \circ k + b, T^{i}\right) \odot T^{i} + T^{i} \tag{5}$$

where k represents the kernel of the 1D convolution. b denotes the bias. $\mathrm{S}(\cdot)$ represents the softmax function. $\odot$ employs the multiplication of the attention weights.

During the inference stage, the LMFE module divides the input gait sequence into sub-sequences with different resolutions and extracts fine-grained temporal features from each branch. The MFF module then fuses these features across branches, which enriches the feature representation and improves the model's perception of temporal patterns at various granularities. The final global features obtained from the MFF module are fed into the GSFE module to capture global contextual information across the entire gait sequence.

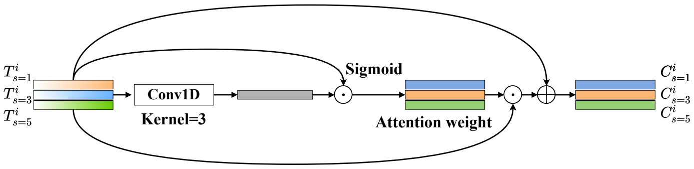

Fig. 4. The variant structure of multi-branch feature fusion (MFF). It combines features from different branches using an attention mechanism to focus on the most relevant information, which enriches the feature diversity and the representability of different branches.

D. Global Self-attention Feature Extractor

The previous studies predominantly employed static fusion methods, such as max(·), mean(·), median(·), or their combination, for global temporal feature fusion. These methods assume equal importance of all local temporal regions in the sequence and fail to adaptively fuse features by considering the contextual relationships among different sub-regions. To tackle this challenge, we employ a transformer architecture, which enables the model to effectively assess the relevance of different subregions within the gait sequence and dynamically modulate their contribution to the feature extraction process. As individuals exhibit unique walking patterns, the temporal characteristics embedded in their global context are also distinct. To capture this uniqueness, we employ a self-attention mechanism to learn the interaction between sub-sequences. Besides, we utilize cross-attention to automatically capture correlations between features from sub-sequences and global temporal features of the entire gait sequence. The GSFE module employs an adaptive fusion mechanism to capture the key regions and important features across the entire time series, which acquire the more expressive feature representation to enhance the performance and accuracy of gait recognition.

Specifically, the spatial feature of each horizontal sub-region $C_{s}^{1}, \ldots, C_{s}^{n}$ obtained in the previous section is inputted into the transformer network to extract global temporal fusion features. The structure of the GSFE module is illustrated in Fig. 5.

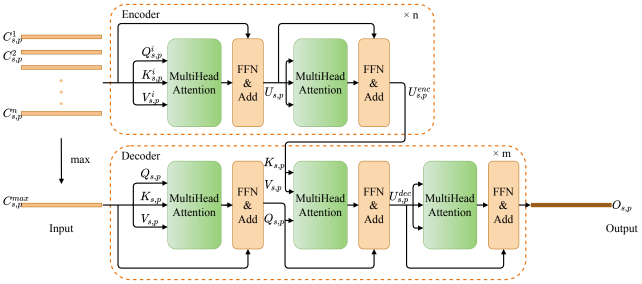

Fig. 5. The detailed structure of global self-attention feature extractor (GSFE). It adopts a transformer structure to extract global contextual information, which learns the contextual information between sub-sequences and employs the learned contextual relationship to adaptively fuse the feature from sub-sequences.

The formula is as follows,

$$\begin{aligned} O_{s,p} &= \operatorname{Transformer}\left(C_{s,p}^{1}, \ldots, C_{s,p}^{n}\right) \\ &= \operatorname{Dec}\left(\operatorname{Enc}\left(C_{s,p}^{1}, \ldots, C_{s,p}^{n}\right), C_{s,p}^{\max}\right), \end{aligned} \tag{6}$$

where the encoder $\operatorname{Enc}(\cdot)$ takes inputs feature $\left\{C_{s,p}^{i} \in \mathrm{R}^{1*c}, i = 1, 2, \ldots, n\right\}$, and $p \in (1, \ldots, \mathrm{m})$ represents the p-th subregion of the horizontal space. The decoder $\operatorname{Dec}(\cdot)$ input $C_{s,p}^{\max} \in \mathrm{R}^{1*c}$ gets $C_{s,p}^{\max} = \max\left(C_{s,p}^{1}, \ldots, C_{s,p}^{n}\right)$ by max-pooling the input $C_{s,p}^{1}, \ldots, C_{s,p}^{n}$ in the time dimension and outputs $O_{s,p} \in \mathrm{R}^{1*c}$. The input features $C_{s,p}^{1}, \ldots, C_{s,p}^{n}$ are sent to the encoder to obtain encoded features. Specifically, $C_{s,p}^{1}, \ldots, C_{s,p}^{n}$ will undergo linear transformation to obtain Q, K, V of the transformer, which is represented as follows,

$$Q_{s,p}^{i} = C_{s,p}^{i} \times W^{Q}; K_{s,p}^{i} = C_{s,p}^{i} \times W^{K}; V_{s,p}^{i} = C_{s,p}^{i} \times W^{V}, \tag{7}$$

where W^{Q}, W^{K}, and W^{V} represent the transformation matrices of Q, K, and V, respectively. The self-attention feature

$u_{s,p}$ is obtained by feeding the self-attention function through $Q_{s,p}, K_{s,p}, V_{s,p}$. The formula is as follows,

$$u_{s,p} = \mathrm{S}\left(Q_{s,p}, K_{s,p}, V_{s,p}\right) = \text{Softmax}\left(\frac{Q_{s,p}\left(K_{s,p}\right)^{T}}{\sqrt{d_k}}\right) V_{s,p}. \tag{8}$$

To enhance the encoding process, we employ a multi-head attention mechanism to capture richer encoding features, similar to the approach in [28]. Then, the self-attention feature $u_{s,p}$ is fed into the feedforward network $\mathrm{FFN}(\cdot)$ to obtain the encoded feature $U_{s,p}$. The formula of the feedforward network is as follows,

$$U_{s,p} = \text{FFN}\left(u_{s,p}\right) = \max\left(0, u_{s,p} \times W_1 + b_1\right) \times W_2 + b_2, \tag{9}$$

where W_1, W_2, b_1, b_2 are the weights and biases of the first and second layers of the feedforward network, respectively. Subsequently, the "Add" operation is employed to implement residual connections and layer normalization within the FFN network. The final encoded feature $U_{s,p}^{enc} \in \mathrm{R}^{n*c}$ as $K_{s,p}$, $V_{s,p}$ is input into the decoder. Similarly, $C_{s,p}^{\max}$ is used as $Q_{s,p}$ and input into the decoder. In the decoder, both $U_{s,p}^{enc}$ and $C_{s,p}^{\max}$ are processed through the self-attention mechanism and feedforward networks, which result in the final decoded features $O_{s,p} = \mathrm{Dec}(U_{s,p}^{enc}, C_{s,p}^{\max})$. The decoded features $O_{s,p}$ of the p-th horizontal spatial region together form the global self-attention temporal context feature $O_s \in \mathrm{R}^{m*c}$ of the s-th branch.

E. Loss Function

To efficiently train the proposed gait recognition model, we utilize the triplet loss function [29], which improves the inter-class distance and reduces the intra-class distance. During training, we perform max-pooling on $C_s^1, ..., C_s^n$ in the temporal dimension to obtain $C_s^{\max} = \max\left(C_s^1, ..., C_s^n\right)$ as the input for the first triplet loss, resulting in L_{tri1}. This component focuses on the local features extracted by LMFE and MFF. It is denoted as follows,

$$\begin{aligned} L_{tri1} = [&\mathrm{D}\left(G\left(C_s^{\max,a_1}\right), G\left(C_s^{\max,b}\right)\right) - \\ &\mathrm{D}\left(G\left(C_s^{\max,a1}\right), G\left(C_s^{\max,a_2}\right)\right) + m]_+, \end{aligned} \tag{10}$$

where a_1 and a_2 represent samples of the same individual, while b denotes samples from a different individual. $C_s^{\max,a1} \in R^{1*c}$ represents the feature vector of sample a_1 at time s, with dimension $1*c$. $C_s^{\max,b} \in R^{1*c}$ represents the feature vector of sample b at time s, with dimension $1*c$. $G(\cdot)$ represents the fully connected layer used for feature mapping. The operation $[\gamma]_+$ is equivalent to $\max(\gamma,\ 0)$. $\mathrm{D}(d_1, d_2)$ represents the Euclidean distance between d_1 and d_2. m is the margin value controlling the minimum required distance between positive and negative pairs.

The component L_{tri2} focuses on the global features extracted by GSFE. It calculates the distance between a positive pair and a negative pair across different sub-sequences. The process takes the output features O_s of the decoder as input. It is formulated as:

$$\begin{aligned} L_{tri2} = [&\mathrm{D}\left(G\left(O_s^{a1}\right), G\left(O_s^{b}\right)\right) - \\ &\mathrm{D}\left(G\left(O_s^{a1}\right), G\left(O_s^{a2}\right)\right) + m]_+, \end{aligned} \tag{11}$$

where O_s is the second triplet loss input gives L_{tri2}. Feature mapping is performed on O_s to change the feature dimension. The total triple loss function L_{com} can be defined as,

$$L_{com} = L_{tri1} + L_{tri2}, \tag{12}$$

where the overall loss function L_{com} is the sum of L_{tri1} and L_{tri2}, which promote both local and global discriminative feature learning.

Furthermore, to refine the classification space, the cross-entropy loss L_{ce} is employed and expressed as follows,

$$L_{ce} = -\sum_{i=1}^{N} y_i \log(p_i), \tag{13}$$

where N is the number of classes, y_i is a binary indicator (0 or 1) if class label i is the correct classification for the observation, and p_i is the predicted probability of the observation being of class i.

To achieve the best performance, triplet loss and cross-entropy loss are used to train our network. It is formulated as follows,

$$L_{all} = L_{com} + L_{ce}, \tag{14}$$

where L_{com} and L_{ce} represent triplet loss and cross-entropy loss respectively. The triplet loss operates from a geometric standpoint, which learns more discriminative feature representations by minimizing the distance between samples of the same class and maximizing the distance between samples of different classes. Besides, the cross-entropy loss is derived from the perspective of maximum likelihood, aiming to optimize the predicted class probability distribution so that it closely matches the true label distribution. By combining these two losses, the model can more effectively extract highly discriminative features, thereby improving its recognition accuracy.

IV. Experiment

This section includes four parts. In the first part, we mainly introduce the two public datasets used in the experiments, called CASIA-B [30] and OU-MVLP [31] datasets. The second part describes the hyperparameter settings of the model. In the third part, we compare the LMGCS with other existing methods on CASIA-B [30] and OU-MVLP [31] datasets. In the fourth part, we conduct the generalization experiments on the Gait3D [32] dataset. In the fifth part, we conduct ablation experiments on the various modules of the LMGCS on the CASIA-B [30] dataset.

A. Dataset

CASIA-B [30] is the most commonly used gait dataset, which includes 124 subjects. Each subject has 10 different sets of data under three conditions: 6 sets of normal walking (NM), 2 sets of walking while carrying a bag (BG), and 2 sets of walking while wearing a coat (CL). Each set contains 11 different angles, ranging from 0° to 180° in increments of 18°. A qualitative illustration is given in Fig. 6. CASIA-B [30] has three general settings: small sample training (ST:

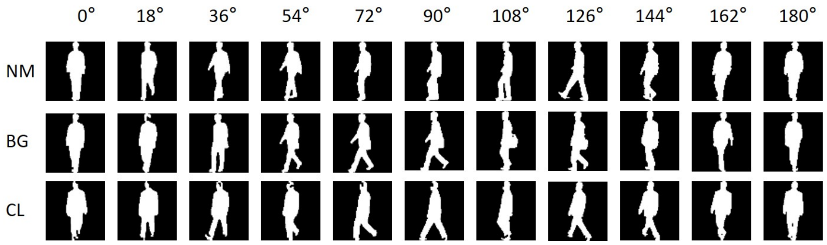

Fig. 6. Examples of cropped gait contours of a subject in the CASIA-B dataset at 0 ° to 180 ° under three different walking states (0 °, 18 °, 36 °, ... , 180 °, 11 evenly spaced angles).

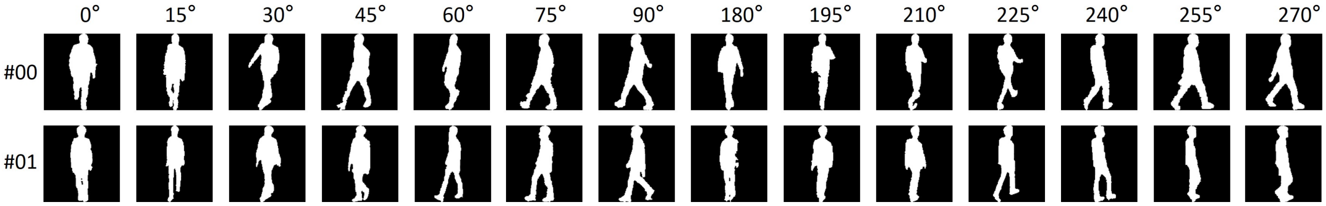

Fig. 7. Example of cropped gait contours for a subject in the OU-MVLP dataset under two walks with 14 different perspectives 0 °, 15 °, ... , 90 ° and 180 °, 195 °, ... , 270 °.

TABLE II. SELECT PARAMETER VALUES FOR THE PROPOSED NETWORK.

Subnetwork	Parameter	Setting
Subnetwork A	Loss Function	Triple Loss + CE Loss
	Optimizer	Adam
	Learning Rate(CASIA-B [30])	Iter<70k:0.0001;Iter>70k:0.00001
	Learning Rate(OU-MVLP [31])	Iter<80k:0.0001;Iter>80k:0.00001
	Margin Value	0.2
	Batch Size(CASIA-B [30])	64
	Batch Size(OU-MVLP [31])	256
	Iter(CASIA-B [30])	80000
	Iter(OU-MVLP [31])	100000
Subnetwork B	Loss Function	Triple Loss + CE Loss
	Optimizer	Adam
	Learning Rate	0.0001
	Margin Value	0.2
	Batch Size(CASIA-B [30])	64
	Batch Size(OU-MVLP [31])	256
	Iter(CASIA-B [30])	30000
	Iter(OU-MVLP [31])	40000

24 subjects for training and 100 for testing), medium sample training (MT: 62 subjects for training and testing, respectively), and large sample training (LT: 74 subjects for training and 50 for testing). The test data is divided into a gallery set and a probe set in each setting. The gallery set includes four NM groups, and the probe set includes the remaining groups.

OU-MVLP [31] is a challenging gait dataset which includes 10,307 subjects. Each subject has 14 views (0°, 15°, ... , 90°, 180°, 195°, ... , 270°), and each view contains two sequences (00 and 01). A qualitative illustration of the OU-MVLP sample is given in Fig. 7. The experimental protocol follows the same procedure as outlined in [11]. All sequences are divided into training and testing sets according to the subjects (5,153 for training and 5,154 for testing). In the testing set, seq01 is the gallery set, and seq00 is the probe set.

B. Implementation Details

1) Training Details: A qualitative illustration is given in Table II. In all the experiments, we use the same processing approach as [18] to align each frame and resize to the size of 64 × 44. We apply Adam as the optimizer [39] with a learning rate of 1e-4 and a momentum of 0.9. In addition, the Leaky ReLU [40] activation function is applied after each convolutional layer. The models are trained on 4 NVIDIA 2080TI GPUs. In a mini-batch, the number of sampled subjects is denoted by p, and the number of sampled sequences per subject is denoted by k. Particularly, (p, k) is set to (8, 8) on CASIA-B [30] and (32, 8) on OU-MVLP [31]. For each input sequence, we sample 30 frames as training data. In CASIA-B [30], we train the subnetwork A for 80K iterations and reduce the learning rate to 1e-5 at 70K iterations first and then train subnetwork B for 30K iterations.

2) Hyper-parameters: (1) The number of channels for four conv layers are set to 32/64, 64/128, 128/256, and 128/256 on CASIA-B [30] and OU-MVLP [31] datasets, respectively. The kernel size is set to 3. (2) For the LMFE, the branch is set to 3, and the value of F is set to 1,3,5 on each branch, respectively. (3) For GSFE, we set the transformer block number of (n, m) to (4,4).

C. Comparison with State-of-the-Art Methods

1) CASIA-B: In Tables III, IV, and V, compared with some experimental methods, the LMGCS model is superior to other competitors. Generally, the models that utilized both spatial and temporal information (*e.g.*, GaitPart [12], GLFE [14], MT3D [17]) outperform the models that rely solely on spatial features (*e.g.*, GaitSet [19]). The modes extracting fine grained temporal features, such as GaitPart [12], GLFE [14], MT3D [17], and the LMGCS, perform better than the GaitSet [19]. Because gait recognition is a fine-grained task, local

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2024.3476384

TABLE III. AVERAGED RANK-1 ACCURACY ON CASIA-B WITH THE SETTING OF ST EXCLUDING IDENTICAL-VIEW CASES. (THE BEST RESULTS ARE MARKED IN **BOLD**).

Gallery NM			$0^o - 180^o$											Mean
Probe			0^o	18^o	36^o	54^o	72^o	90^o	108^o	126^o	144^o	162^o	180^o	
ST	NM	ViDP [33]	-	-	-	59.1	-	50.2	-	57.5	-	-	-	-
		CMCC [34]	46.3	-	-	52.4	-	48.3	-	56.9	-	-	-	-
		CNN-LB [22]	54.8	-	-	77.8	-	64.9	-	76.1	-	-	-	-
		PoseGait [35]	55.3	69.6	73.9	75.0	68.0	68.2	71.1	72.9	76.1	70.4	55.4	68.7
		GaitSet [19]	64.6	83.3	90.4	86.5	80.2	75.5	80.3	86.0	87.1	81.4	59.6	79.5
		GLFE [14]	77.0	87.8	93.9	92.7	83.9	78.7	84.7	91.5	92.5	89.3	74.4	86.0
		LMGCS (Ours)	**77.5**	**88.2**	**94.0**	**92.7**	**83.9**	**79.2**	**85.1**	**91.8**	**92.8**	**90.2**	**75.2**	**86.4**
	BG	PoseGait [35]	35.3	47.2	52.4	46.9	45.5	43.9	46.1	48.1	49.4	43.6	31.1	44.5
		GaitSet [19]	55.8	70.5	76.9	75.5	69.7	63.4	68.0	75.8	76.2	70.7	52.2	68.6
		GLFE [14]	68.1	81.2	87.7	84.9	76.3	70.5	76.1	84.5	87.0	83.6	65.0	78.6
		LMGCS (Ours)	**68.3**	**81.6**	**88.2**	**85.2**	**76.6**	**70.7**	**76.8**	**85.1**	**87.6**	**83.9**	**66.2**	**79.1**
	CL	PoseGait [35]	24.3	29.7	41.3	38.8	38.2	38.5	41.6	44.9	42.2	33.4	22.5	36.0
		GaitSet [19]	29.4	43.1	49.5	48.7	42.3	40.3	44.9	47.4	43.0	35.7	25.6	40.9
		GLFE [14]	46.9	58.7	66.6	65.4	58.3	54.1	59.5	62.7	61.3	57.1	40.6	57.4
		LMGCS (Ours)	**47.2**	**59.3**	**67.3**	**65.9**	**58.6**	**54.3**	**59.8**	**62.8**	**61.3**	**57.9**	**40.8**	**57.7**

TABLE IV. AVERAGED RANK-1 ACCURACY ON CASIA-B WITH THE SETTING OF MT EXCLUDING IDENTICAL-VIEW CASES. (THE BEST RESULTS ARE MARKED IN **BOLD**).

Gallery NM			$0^o - 180^o$											Mean
Probe			0^o	18^o	36^o	54^o	72^o	90^o	108^o	126^o	144^o	162^o	180^o	
MT	NM	AE [36]	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 [37]	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 [19]	86.8	95.2	98.0	94.5	91.5	89.1	91.1	95.0	97.4	93.7	80.2	92.0
		GLFE [14]	93.9	97.6	98.8	97.3	95.2	92.7	95.6	98.1	98.5	96.5	91.2	95.9
		LMGCS (Ours)	**94.2**	**97.9**	**99.2**	**98.0**	**95.8**	**93.2**	**96.0**	**98.3**	**99.0**	**97.0**	**91.6**	**96.4**
	BG	AE [36]	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 [37]	48.5	58.5	59.7	58.0	53.7	49.8	54.0	51.3	59.5	55.9	43.1	54.7
		GaitSet [19]	79.9	89.8	91.2	86.7	81.6	76.7	81.0	88.2	90.3	88.5	73.0	84.3
		GLFE [14]	88.5	95.1	95.9	94.2	91.5	85.4	89.0	95.4	97.4	94.3	86.3	92.1
		LMGCS (Ours)	**88.6**	**95.4**	**96.2**	**94.4**	**92.2**	**85.7**	**89.6**	**95.7**	**97.6**	**94.8**	**86.7**	**92.4**
	CL	AE [36]	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 [37]	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 [19]	52.0	66.0	72.8	69.3	63.1	61.2	63.5	66.5	67.5	60.0	45.9	62.5
		GLFE [14]	70.7	83.2	87.1	84.7	78.2	71.3	78.0	83.7	83.6	77.1	63.1	78.3
		LMGCS (Ours)	**71.2**	**83.6**	**87.9**	**85.2**	**78.6**	**71.6**	**78.5**	**84.1**	**83.7**	**77.5**	**63.4**	**78.7**

fine-grained regions contain more distinctive identity information. Moreover, using different granularity to extract temporal features simultaneously, such as GLFE [14], MT3D [17], and the LMGCS, is better than GaitPart [12], which uses the same granularity to extract temporal features.

As shown in Tables III, IV, and V, the proposed LMGCS achieves best results under different conditions. For example, as shown in Table III, the proposed method achieves superior performance compared to GLFE [14] across several angles in the ST setting. At the 0° angle, the performance of LMGCS improves by 0.5%, and at the 180° angle, the improvement is 0.8%. Additionally, as shown in Table IV, the proposed method consistently improves over GLFE [14] across most angles in the MT setting. It demonstrates the effectiveness of the LMGCS in capturing complex gait features. Besides, compared with the GLFE [14] which utilizes both spatial and temporal information, the proposed LMGCS demonstrated decreases of 0.62%, 0.95%, and 1.67% in NM, BG and CL, as shown in Table V. Furthermore, the LMGCS performs better than GLFE [14], MT3D [17], and GaitPart [12] since the proposed method employs the transformer modules to mine global contextual spatio-temporal features, which is important to the varying temporal context information of different individuals during walking.

2) OU-MVLP: To evaluate the performance of LMGCS, we complete the experiments on the OU-MVLP gait dataset [31], as shown in Table VI. We use the test protocol, which is the same as [12]. The result of the “Mean” scores was the best. Compared to previous methods, the proposed LMGCS method improves the recognition accuracy by approximately 1.7% over GaitSet [19] and 2.5% over GaitPart [12], based on the mean Rank-1 accuracy. The superior performance of the LMGCS approach is attributed to its ability to comprehensively capture the uniqueness of gait by combining local multi-scale temporal features with global contextual spatio-temporal features. As shown in Table VI, the LMGCS performs comparably to 3dLocal [16] at most angles, with differences ranging between 0.1% and 0.9%. This indicates that both methods can achieve comparable performance in many view angles. However, at the extreme angles of 0° and 180°, LMGCS outperforms 3dLocal [16] significantly. Specifically, at 0°, the proposed method achieves an accuracy of 88.4%, whereas 3dLocal [16]

TABLE V. AVERAGED RANK-1 ACCURACY ON CASIA-B WITH THE SETTING OF LT EXCLUDING IDENTICAL-VIEW CASES. (THE BEST RESULTS ARE MARKED IN **BOLD**).

Gallery NM			$0^o - 180^o$											Mean
Probe			0^o	18^o	36^o	54^o	72^o	90^o	108^o	126^o	144^o	162^o	180^o	
LT	NM	CNN-3D [22]	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 [22]	88.7	95.1	98.2	96.4	94.1	91.5	93.9	97.5	98.4	95.8	85.6	94.1
		GaitSet [19]	90.8	97.9	99.4	96.9	93.6	91.7	95.0	97.8	98.9	96.8	85.8	95.0
		ACL [11]	92.0	98.5	100.0	98.9	95.7	91.5	94.5	97.7	98.4	96.7	91.9	96.0
		GEINet [38]	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 [12]	94.1	98.6	99.3	98.5	94.0	92.3	95.9	98.4	99.2	97.8	90.4	96.2
		MT3D [17]	95.7	98.2	99.0	97.5	95.1	93.9	96.1	98.6	99.2	98.2	92.0	96.7
		GLFE [14]	96.0	98.3	99.0	97.9	96.9	95.4	97.0	98.9	99.3	98.8	94.0	97.4
		LMGCS (Ours)	**97.0**	**99.5**	**99.3**	**98.8**	**97.4**	**96.2**	**97.4**	**98.9**	**99.5**	**99.5**	**94.3**	**98.0**
	BG	CNN-LB [22]	64.2	80.6	82.7	76.9	64.8	63.1	68.0	76.9	82.2	75.4	61.3	72.4
		GaitSet [19]	83.8	91.2	91.8	88.8	83.3	81.0	84.1	90.0	92.2	94.4	79.0	87.2
		GEINet [38]	83.8	91.2	91.8	88.8	83.3	81.0	84.1	90.0	92.2	94.4	79.0	87.2
		GaitPart [12]	89.1	94.8	96.7	95.1	88.3	84.9	89.0	93.5	96.1	93.8	85.8	91.5
		MT3D [17]	91.0	95.4	97.5	94.2	92.3	86.9	91.2	95.6	97.3	96.4	86.6	93.0
		GLFE [14]	92.6	96.6	96.8	95.5	93.5	89.3	92.2	96.5	98.2	96.9	**91.5**	94.5
		LMGCS (Ours)	**93.4**	**97.8**	**99.9**	**96.6**	**94.7**	**89.3**	**93.6**	**98.0**	**99.7**	**98.8**	89.0	**95.4**
	CL	MGAN [37]	23.1	34.5	36.3	33.3	32.9	32.7	34.2	37.6	33.7	26.7	21.0	31.5
		CNN-LB [22]	37.7	57.2	66.6	61.1	55.2	54.6	55.2	59.1	58.9	48.8	39.4	54.0
		GaitSet [19]	61.4	75.4	80.7	77.3	72.1	70.1	71.5	73.5	73.5	68.4	50.0	70.4
		GEINet [38]	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 [12]	70.7	85.5	86.9	83.3	77.1	72.5	76.9	82.2	83.8	80.2	66.5	78.7
		MT3D [17]	76.0	87.6	89.8	85.0	81.2	75.7	81.0	84.5	85.4	82.2	68.1	81.5
		GLFE [14]	76.6	90.0	90.3	87.1	84.5	79.0	84.1	87.0	87.3	84.4	69.5	83.6
		LMGCS (Ours)	**79.5**	**91.1**	**93.3**	**88.5**	**84.7**	**79.2**	**84.5**	**88.0**	**88.9**	**85.7**	**71.6**	**85.0**

TABLE VI. AVERAGED RANK-1 ACCURACY ON OU-MVLP EXCLUDING IDENTICAL-VIEW CASES. (THE BEST RESULTS ARE MARKED IN **BOLD**).

NM	Gallery All 14 Views					
	GaitSet [19]	GaitPart [12]	GLFE [14]	CSTL [17]	3dLocal [16]	LMGCS (Ours)
0^o	79.5	82.6	83.8	87.1	86.1	**88.4**
15^o	87.9	88.9	90.0	91.0	91.2	**91.6**
30^o	89.9	90.8	91.0	91.5	**92.6**	91.5
45^o	90.2	91.0	91.2	91.8	**92.9**	91.8
60^o	88.1	89.7	90.3	90.6	**92.2**	91.4
75^o	88.7	89.9	90.0	90.8	**91.3**	91.1
90^o	87.8	89.5	89.4	90.6	**91.1**	90.7
180^o	81.7	85.2	85.3	89.4	86.9	**90.7**
195^o	86.7	88.1	89.1	90.2	90.8	**90.8**
210^o	89.0	90.0	90.5	90.5	**92.2**	91.6
225^o	89.3	90.1	90.6	90.7	**92.3**	91.5
240^o	87.2	89.0	89.6	89.8	**91.3**	91.0
255^o	87.8	89.1	89.3	90.0	**91.1**	90.5
270^o	86.2	88.2	88.5	89.4	**90.2**	90.0
Mean	87.1	88.7	89.2	90.2	**90.9**	**90.9**

only reaches 86.1%, representing an improvement of about 2.67%. At 180°, the LMGCS method scores 90.7%, while 3dLocal [16] reaches 86.9%, leading to a 4.38% improvement. The small performance gap at other angles can be attributed to the fact that 3dLocal [16] effectively captures local spatio-temporal features, leading to comparable performance to the LMGCS at several angles.

However, at extreme angles such as 0° and 180°, 3dLocal [16] struggles to fully extract critical features, as it primarily focuses on local spatio-temporal information and lacks a holistic understanding of the scene. In contrast, the proposed LMGCS extracts local multi-scale features across different temporal resolutions and learns contextual information between different subsequences. It facilitates the effective fusion of features at various resolutions in the LMGCS model, which helps to capture richer global context information. Besides, the variance in accuracy for the proposed LMGCS is around 0.675, while the variance for the 3dLocal [16] method is approximately 3.483. It indicates that the LMGCS has more consistent performance across different viewing angles compared to the 3dLocal [16] method. In conclusion, the proposed method performs better at extreme angles and shows more excellent stability due to its lower variance.

D. Generalization Analysis

To validate the generalization performance of the proposed LMGCS model, we conducted experiments on the Gait3D [32] dataset. The Gait3D dataset encompasses over 25,000 gait sequences from 4,000 participants, which feature diverse information such as different human perspectives and body shapes in the wild. We utilized the 2D Silhouette data to extract the gait feature. Table VII presents a performance comparison of the proposed LMGCS model with other state-of-the-art models based on the R@1, R@5, and mAP metrics. Compared to other competitors, LMGCS demonstrates an advantage across all evaluation metrics. In terms of R@1 and R@5, the LMGCS achieved the best scores of 17.9 and 35.8, respectively. Compared to the CSTL [18] which utilizes spatial and temporal information, the proposed LMGCS achieves a 52.99% improvement in R@1 and an 86.46% improvement in R@5. Furthermore, compared to GaitGraph [41], the proposed method improved the mAP metric by 62.4%. It verifies the generalization ability of the proposed method in complex gait recognition scenarios.

TABLE VII. COMPARISON OF THE STATE-OF-THE-ART GAIT RECOGNITION METHODS ON GAIT3D DATASET.

Methods	R@1	R@5	mAP
GEINet [38]	5.4	14.2	5.1
PoseGait [35]	0.2	1.1	0.5
GaitGraph [41]	6.3	16.2	5.2
CSTL [18]	11.7	19.2	5.6
LMGCS (Ours)	**17.9**	**35.8**	**13.8**

E. Ablation Studies

In Table VIII, the ablation studies provide individual contributions of each key component within the LMGCS network, including the local multi-resolution feature extractor (LMFE), multi-branch feature fusion (MFF), and global self-attention feature extractor (GSFE).

Baseline Configuration (First Row): When all components (LMFE, MFF, and GSFE) are removed, the network achieves Rank-1 accuracy scores of 96.0% (NM), 92.6% (BG), and 81.2% (CL). These results represent the basic configuration of the model without any advanced feature extraction or fusion methods.

Effect of LMFE (Second Row): When LMFE is equipped solely, it achieves accuracy improvements of 1.3% in NM, 1.4% in BG, and 1.5% in CL, which is attributable to the effective extraction of local temporal features. It demonstrates the significance of capturing fine-grained local temporal features in improving the performance.

Effect of MFF and LMFE Combination (Third Row): When MFF is combined with LMFE, the accuracy improves to 84.6% in the CL scenario. This result highlights the importance of integrating multiple temporal features at different granularities to strengthen the model's temporal feature perception.

Effect of GSFE (Fourth Row): When GSFE is combined with LMFE, the Rank-1 accuracy improves the value to 97.8% and 95.3% in the NM, BG scenarios, respectively. It employs the idea that capturing global temporal context is crucial for improving the overall accuracy of the model.

Full Model Configuration (Fifth Row): With the integration of LMFE, MFF, and GSFE, the network achieves the best accuracy in all contexts, with Rank-1 accuracy reaching 98.0% in the NM scenario, 95.4% in the BG scenario, and 85.0% in the CL scenario. It demonstrates the importance of combining local and global temporal features with multi-branch fusion to boost performance.

In summary, the ablation study proves that each component of the LMGCS network contributes to improved accuracy, with the full model configuration achieving the best performance. These results validate the inclusion of these components, as they collectively enhance the model's ability to effectively capture and integrate both local and global temporal features.

TABLE VIII. ABLATION ANALYSIS OF THE KEY COMPONENTS (THE BEST RESULTS ARE MARKED IN **BOLD**).

Methods			Rank-1 Accuracy		
LMFE	MFF	GSFE	NM	BG	CL
✗	✗	✗	96.0	92.6	81.2
✓	✗	✗	97.1	94.0	82.7
✓	✓	✗	97.2	94.9	84.6
✓	✗	✓	97.8	95.3	84.6
✓	✓	✓	**98.0**	**95.4**	**85.0**

In addition, we conducted ablation experiments on individual key modules.

1) Contribution of local multi-resolution feature extractor: Compared with previous models that use a single scale to extract features in the time dimension, we utilized a multi-scale structure to extract fine-grained local temporal features. We conduct different scales in the experiments to explore the effect of multi-scale on the model. Table IX presents the experimental results with LT settings on the CASIA-B [30] dataset. The results show that accuracy steadily improves as the number of branches increases. It indicates that accuracy improvements plateau when the number of branches is set to 3. As shown in the last line, while the experiment indicates that the efficiency of the model is not optimal, the accuracy improves significantly with only a slight performance lag compared to optimal efficiency.

2) Contribution of global self-attention feature extractor: To verify the effectiveness of the GSFE module, we conduct the ablation study as shown in Table X. The experimental results demonstrate that combining GSFE with the baseline method leads to improved accuracy compared to the baseline. It indicates that GSFE is effective as an adaptive collection pooling method. In particular, the average accuracy is further improved when GSFE is combined with GaitSet [19] and MT3D [17]. This indicates that GSFE is universal and can be combined with various base networks to improve accuracy. Additionally, experimental results revealed that the model achieves optimal performance when the number of encoder and decoder layers in GSFE is set to 4.

3) Contribution of multi-branch feature fusion: To explore the fusion pattern for the MFF module, the experimental results are presented in Table XI. The MFF module has two variants, which consist of static structure and attention structure. In the static structure, the fused feature of each branch is the cumulative sum of all preceding branches. The features can be fused in two directions ($T_{s=1} \rightarrow T_{s=5}$ and $T_{s=5} \rightarrow T_{s=1}$). $T_{s=1} \rightarrow T_{s=5}$ indicates the accumulation from fine-grained to coarse-grained features. The transition from $T_{s=1}$ to $T_{s=5}$ indicates the aggregation of features from fine-grained to coarse-grained scales. In the same way, the shift from $T_{s=5}$ to $T_{s=1}$ reflects the accumulation of features from coarse-grained to fine-grained scales. In the attention-based structure, the

TABLE IX. THE IMPORTANCE OF THE DIFFERENT BRANCHES AND THE EXPERIMENT RESULTS WITH THE LT SETTINGS ON THE CASIA-B DATASET.

Multiscale-feature			Rank-1 Accuracy				RTX 3090 Efficiency		
Fine-Grained	Medium-Grained	Coarse-Grained	NM	BG	CL	Mean	Params (M)	FLOPs (G)	Time (ms)
✓			96.0	92.6	81.2	89.9	14.72	2.58	1.41
	✓		95.4	92.2	80.9	89.5	14.72	2.59	5.24
		✓	95.2	92.0	80.6	89.4	14.72	2.59	5.25
✓	✓		97.9	95.2	84.7	92.6	15.77	2.59	5.24
	✓	✓	97.3	94.0	82.8	91.4	15.77	2.59	9.02
✓		✓	97.6	95.0	84.4	92.3	15.77	2.60	9.00
✓	✓	✓	**98.0**	**95.4**	**85.0**	**92.8**	16.82	2.60	9.13

TABLE X. THE EXPERIMENTAL DATA RESULTS OF THE CONTRIBUTION OF GLOBAL SELF-ATTENTION FEATURE EXTRACTOR (THE BEST RESULTS ARE MARKED IN **BOLD**).

Methods	NM	BG	CL
Baseline	96.0	92.6	81.2
Baseline + GSFE(4 × 4)	97.7	94.7	83.4
Baseline + LMFE + GSFE(1 × 1)	97.8	95.0	84.4
Baseline + LMFE + GSFE(4 × 4)	**97.8**	**95.3**	**84.6**
Gaitset [19]	95.0	87.2	70.4
Gaitset [19] + GSFE(4 × 4)	97.0	89.2	74.5
MT3D [17]	96.5	93.4	81.6
MT3D [17] + GSFE(4 × 4)	97.8	94.5	83.0

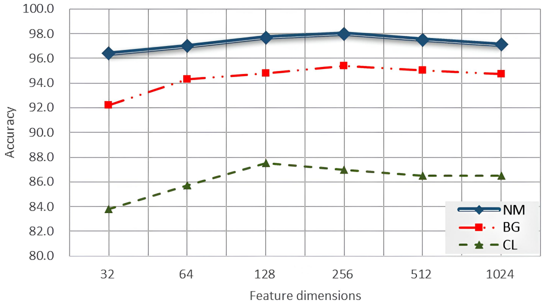

Fig. 8. The relationship between recognition accuracy and feature dimensions. From left to right are the individual results for the NM, BG, and CL scenarios on the CASIA-B dataset.

fused features utilize the attention mechanism and the residual connection to obtain the updated features. In terms of training time and network parameters, the use of static architecture is more efficient than the use of attention-based structure. Especially, the experimental results show that the transition from $T_{s=1}$ to $T_{s=5}$ can achieve higher recognition accuracy. It indicates that gradually accumulating features from fine-grained to coarse-grained levels leads to better performance. Based on the experimental results, we conclude that the fine-grained branch contains more identity information than the medium-grained and coarse-grained branches. Therefore, when the fine-grained branches gradually accumulate to other branches, the gait feature expression ability is enriched and expanded to improve the recognition accuracy of the whole gait recognition network. In the future, we intend to develop more efficient model architectures that reduce parameter count and computational complexity while enhancing accuracy.

4) Selection of feature dimensions: As shown in Table XII and Fig. 8, we explored the relationship between recognition accuracy and feature dimensions by setting the output dimensions to 32, 64, 128, 256, 512, and 1024, respectively. The results indicate that while increasing feature dimensions generally boosts accuracy, excessive dimensions (such as 512 and 1024) slightly reduce performance, especially in the NM and BG scenarios. Specifically, the optimal feature dimension in the NM scenario is 256, with the accuracy reaching 98.0%. Similarly, the BG scenario achieves the best accuracy of 95.4% at 256 dimensions. The CL scenario achieves the best performance at a dimension of 128, with an accuracy rate of 85.5%, but its performance is not satisfactory in the NM and BG scenarios. From this experiment, we conclude that increasing feature dimensions improves recognition accuracy to some degree, but excessive dimensions can decrease testing efficiency and enlarge model size without bringing substantial accuracy improvements. Therefore, setting the feature dimension to 256 strikes an optimal balance between accuracy and efficiency across various scenarios.

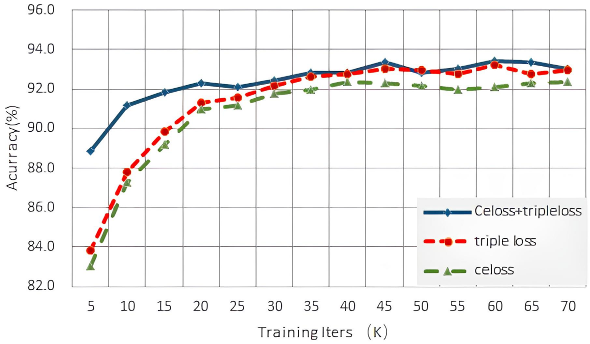

Fig. 9. Comparison of three loss functions.

5) Comparison of different resolution settings: To investigate the impact of different branch resolutions on model accuracy, we address the image-filling issue in each branch and set the resolutions of the branches to 1,5,7, 1,3,5, and 3,5,7, respectively. The experimental results are presented in Table XIII. While the branch settings varied, all other training strategies remained consistent. The results demonstrate that the model achieves optimal accuracy at resolution settings of 1, 3, and 5, which enhances its ability to extract fine-grained temporal features from images.

6) Ablation studies on the loss functions: As indicated in Table XIV, a comparison of the experimental data from the three training strategies shows that combining triplet loss with cross-entropy loss results in higher recognition accuracy. It demonstrates that the combined training loss enhances the performance of gait recognition models, particularly in cases

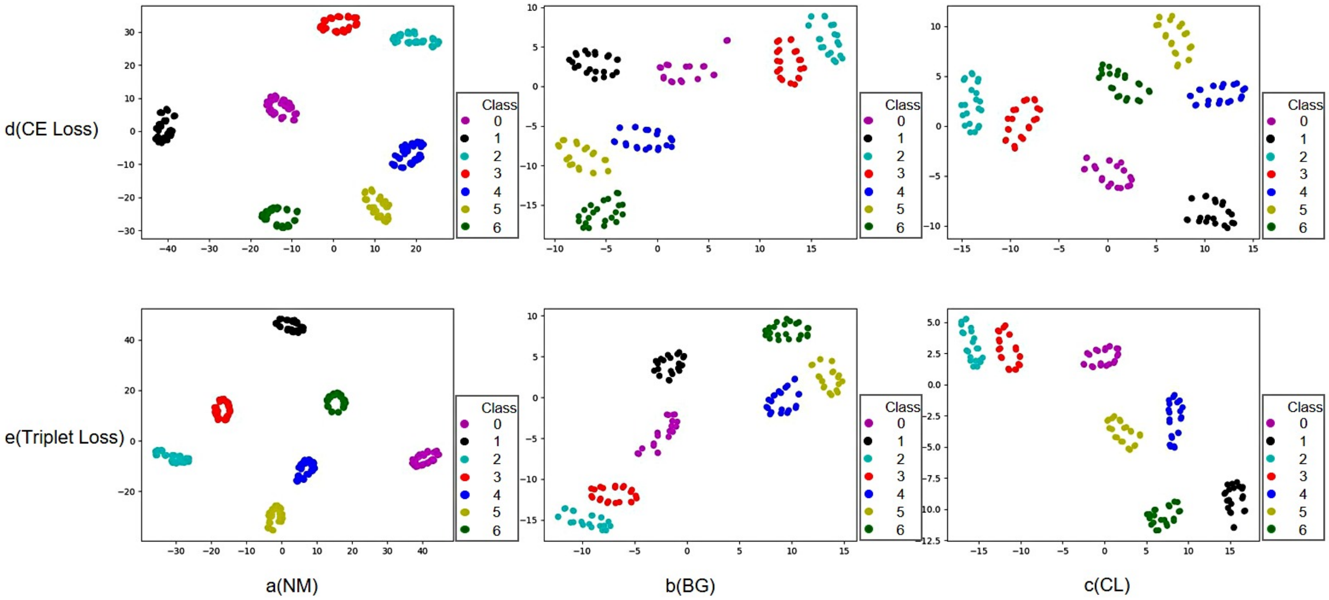

Fig. 10. The t-SNE visualization of the feature spaces using (d) CE Loss and (e) Triplet Loss in (a) NM, (b) BG, and (c) CL scenarios.

TABLE XI. EXPERIMENTAL DATA RESULTS OF MULTI-BRANCH FEATURE FUSION (THE BEST RESULTS ARE MARKED IN **BOLD**).

Fusion Pattern	Inference Time	Model Size	NM	BG	CL
Static structure ($T_{s=1} \to T_{s=5}$)	2m27s	59.0m	98.0	**95.4**	**85.0**
Static structure ($T_{s=5} \to T_{s=1}$)	2m27s	59.0m	97.8	95.1	84.5
Attention-based structure	3m01s	72.0m	**98.5**	95.1	84.4

TABLE XII. THE RELATIONSHIP BETWEEN RECOGNITION ACCURACY AND FEATURE DIMENSIONS (THE BEST RESULTS ARE MARKED IN **BOLD**).

Feature dimensions	NM	BG	CL
32	96.4	92.2	81.8
64	97.0	94.3	83.7
128	97.7	94.8	**85.5**
256	**98.0**	**95.4**	85.0
512	97.5	95.0	84.5
1024	97.1	94.7	84.5

TABLE XIII. THE EXPERIMENTAL RESULTS OF DIFFERENT RESOLUTION SETTINGS.

Fine-Grained	Medium-Grained	Coarse-Grained	NM	BG	CL
1	5	7	93.3	89.8	70.2
1	3	5	**98.0**	**95.4**	**85.0**
3	5	7	91.9	89.2	72.5

where the inter-class distance is large and the intra-class distance is small. In addition, as shown in Fig. 9, the joint use of triplet loss and cross-entropy loss also accelerates training and achieves faster convergence to a local optimum.

To investigate the impact of triplet loss and cross-entropy loss on representation learning, we used the t-SNE method to embed seven groups of high-dimensional features into a two-dimensional space for visualization, as shown in Fig. 10. In detail, the column “a” represents the NM scenario, the column “b” represents the BG scenario, and the column “c” represents the CL scenario. The effects of cross-entropy loss are shown in line “d”, and the effects of triplet loss are illustrated in line “e”. As can be seen in line “d”, the cross-entropy loss does not form distinct clusters for certain samples. By contrast, row “e” indicates that triplet loss has led to denser and more effective clustering, with data points tightly clustered around their centroids. It indicates that individuals are more distinguishable within the gait recognition feature space. The triplet loss plays a more significant role in enhancing recognition accuracy, while the cross-entropy loss is more conducive to faster model training.

TABLE XIV. ABLATION STUDIES ON THE LOSS FUNCTIONS ON CASIA-B DATASET.

Loss Function	NM	BG	CL
CE Loss	97.4	94.3	83.4
Triplet Loss	97.9	95.3	84.8
Triplet Loss + CE Loss	**98.0**	**95.4**	**85.0**

V. CONCLUSION

In this paper, we propose a local multi-scale and global contextual spatio-temporal (LMGCS) network for gait recognition. First, the local multi-resolution feature extractor can capture the fine-grained temporal features. In addition, a multi-branch feature fusion module is employed to improve the spatio-temporal feature diversity. Besides, the global self-attention feature extractor is utilized to extract more discriminative.

Meanwhile, a triplet loss is integrated with the cross-entropy loss to promote the network to accomplish the tasks of gait recognition. Experimental analysis of public datasets reveals the effectiveness of each module in the network. In the future, we will focus on developing more efficient model architectures that minimize parameter count and computational complexity, while aiming to enhance accuracy.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

[1] T. Huang, X. Ben, C. Gong, B. Zhang, R. Yan, and Q. Wu, "Enhanced spatial-temporal salience for cross-view gait recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 10, pp. 6967–6980, 2022.

[2] L. Wang, F. Wang, and T. Yan, "Cross-modal person re-identification combining multi-scale features and confusion learning," *CAAI Transactions on Intelligent Systems*, vol. 19, no. 4, pp. 898–908, 2024.

[3] T. Huang, X. Ben, C. Gong, W. Xu, Q. Wu, and H. Zhou, "Gaitdan: Cross-view gait recognition via adversarial domain adaptation," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[4] X. Ben, P. Zhang, Z. Lai, R. Yan, X. Zhai, and W. Meng, "A general tensor representation framework for cross-view gait recognition," *Pattern Recognition*, vol. 90, pp. 87–98, 2019.

[5] H. Qin, Z. Chen, Q. Guo, Q. J. Wu, and M. Lu, "Rpnet: gait recognition with relationships between each body-parts," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 5, pp. 2990–3000, 2021.

[6] Z. Lyu, Y. Wang, and X. Xing, "A gait representation method based on weighted cca for multi-information fusion," *CAAI Transactions on Intelligent Systems*, vol. 14, no. 3, pp. 449–454, 2019.

[7] X. Xing, R. Gao, T. Han, S.-C. Zhu, and Y. N. Wu, "Deformable generator networks: unsupervised disentanglement of appearance and geometry," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 3, pp. 1162–1179, 2020.

[8] X. Xing, T. Wu, S.-C. Zhu, and Y. N. Wu, "Inducing hierarchical compositional model by sparsifying generator network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 14 296–14 305.

[9] T. Han, X. Xing, and Y. N. Wu, "Learning multi-view generator network for shared representation," in *Proceedings of the International Conference on Pattern Recognition (ICPR)*. IEEE, 2018, pp. 2062–2068.

[10] X. Ben, C. Gong, P. Zhang, X. Jia, Q. Wu, and W. Meng, "Coupled patch alignment for matching cross-view gaits," *IEEE Transactions on Image Processing*, vol. 28, no. 6, pp. 3142–3157, 2019.

[11] Y. Zhang, Y. Huang, S. Yu, and L. Wang, "Cross-view gait recognition by discriminative feature learning," *IEEE Transactions on Image Processing*, vol. 29, pp. 1001–1015, 2019.

[12] C. Fan, Y. Peng, C. Cao, X. Liu, S. Hou, J. Chi, Y. Huang, Q. Li, and Z. He, "Gaitpart: Temporal part-based model for gait recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 14 225–14 233.

[13] A. Sepas-Moghaddam and A. Etemad, "View-invariant gait recognition with attentive recurrent learning of partial representations," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 3, no. 1, pp. 124–137, 2020.

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

[15] T. Wolf, M. Babaee, and G. Rigoll, "Multi-view gait recognition using 3d convolutional neural networks," in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*. IEEE, 2016, pp. 4165–4169.

[16] Z. Huang, D. Xue, X. Shen, X. Tian, H. Li, J. Huang, and X.-S. Hua, "3d local convolutional neural networks for gait recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 14 920–14 929.

[17] B. Lin, S. Zhang, and F. Bao, "Gait recognition with multiple-temporal-scale 3d convolutional neural network," in *Proceedings of the ACM International Conference on Multimedia (ACMMM)*, 2020, pp. 3054–3062.

[18] X. Huang, D. Zhu, H. Wang, X. Wang, B. Yang, B. He, W. Liu, and B. Feng, "Context-sensitive temporal feature learning for gait recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 12 909–12 918.

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

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

[21] Z. Wu, Y. Huang, and L. Wang, "Learning representative deep features for image set analysis," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 1960–1968, 2015.

[22] Z. Wu, Y. Huang, L. Wang, X. Wang, and T. Tan, "A comprehensive study on cross-view gait based human identification with deep cnns," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 2, pp. 209–226, 2016.

[23] N. Takemura, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi, "On input/output architectures for convolutional neural network-based cross-view gait recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 9, pp. 2708–2719, 2017.

[24] S. Hou, C. Cao, X. Liu, and Y. Huang, "Gait lateral network: Learning discriminative and compact representations for gait recognition," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2020, pp. 382–398.

[25] Z. Zhang, L. Tran, F. Liu, and X. Liu, "On learning disentangled representations for gait recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 1, pp. 345–360, 2020.

[26] R. Wang, C. Shen, C. Fan, G. Q. Huang, and S. Yu, "Pointgait: Boosting end-to-end 3d gait recognition with point clouds via spatiotemporal modeling," in *Proceedings of the IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 2023, pp. 1–10.

[27] Y. Fu, Y. Wei, Y. Zhou, H. Shi, G. Huang, X. Wang, Z. Yao, and T. Huang, "Horizontal pyramid matching for person re-identification," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, vol. 33, no. 01, 2019, pp. 8295–8302.

[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[29] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[30] S. Yu, D. Tan, and T. Tan, "A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition," in *Proceedings of the International Conference on Pattern Recognition (ICPR)*, vol. 4. IEEE, 2006, pp. 441–444.

[31] 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 transactions on Computer Vision and Applications*, vol. 10, pp. 1–14, 2018.

[32] J. Zheng, X. Liu, W. Liu, L. He, C. Yan, and T. Mei, "Gait recognition in the wild with dense 3d representations and a benchmark," in *Pro-*

ceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 20 228–20 237.

[33] M. Hu, Y. Wang, Z. Zhang, J. J. Little, and D. Huang, "View-invariant discriminative projection for multi-view gait-based human identification," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 12, pp. 2034–2045, 2013.

[34] W. Kusakunniran, Q. Wu, J. Zhang, H. Li, and L. Wang, "Recognizing gaits across views through correlated motion co-clustering," *IEEE Transactions on Image Processing*, vol. 23, no. 2, pp. 696–709, 2013.

[35] R. Liao, S. Yu, W. An, and Y. Huang, "A model-based gait recognition method with body pose and human prior knowledge," *Pattern Recognition*, vol. 98, p. 107069, 2020.

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

[37] Y. He, J. Zhang, H. Shan, and L. Wang, "Multi-task gans for view-specific feature learning in gait recognition," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 1, pp. 102–113, 2018.

[38] K. Shiraga, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi, "Geinet: View-invariant gait recognition using a convolutional neural network," in *2016 international conference on biometrics (ICB)*. IEEE, 2016, pp. 1–8.

[39] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[40] A. L. Maas, A. Y. Hannun, A. Y. Ng *et al.*, "Rectifier nonlinearities improve neural network acoustic models," in *Proceedings of the International Conference on Machine Learning (ICML) Proc. icml*, vol. 30, no. 1. Atlanta, GA, 2013, p. 3.

[41] T. Teepe, A. Khan, J. Gilg, F. Herzog, S. Hörmann, and G. Rigoll, "Gaitgraph: Graph convolutional network for skeleton-based gait recognition," in *2021 IEEE international conference on image processing (ICIP)*. IEEE, 2021, pp. 2314–2318.

Wenzhe Zhai is pursuing a Ph.D. degree at the College of Intelligent Science Systems and Engineering, Harbin Engineering University, Harbin, China. His research interests include smart city systems, information fusion, crowd analysis, and deep learning.

Haomiao Li Received a B.S. degree in Automation from the School of Control Science and Engineering, Northeast University, Shenyang, China, in 2019. He is currently pursuing a master's degree at Harbin Engineering University, Harbin, China. His current research interests include gait recognition and generative models.

Chaoqun Zheng Received B.S. degree in Automation from the School of Intelligent Science and Engineering, Harbin Engineering University, Harbin, China, in 2023. Her current research interests include gait recognition and generative models.

Xianglei Xing received the M.S. and Ph.D. degrees from the School of Electronic Science and Engineering, Nanjing University, China, in 2006 and 2013, respectively. He is currently a professor with the College of Intelligent System Science and Engineering, Harbin Engineering University. During the years 2017-2019, he was a visiting researcher at UCLA. His research interests include computer vision, statistical modeling, and learning, with a focus on representation learning, deep generative models, sparse and structure learning, and explainable models for computer vision.