Learning Effective Representations from Global and Local Features for Cross-View Gait Recognition

Beibei Lin[1] , Shunli Zhang[1*], Xin Yu[2] , Chuihan Kong[1] and Chenwei Wan[1]

[1]*Beijing Jiaotong University Beijing , China*
[2] *University of Technology Sydney, Australia*

北京交通大學

Introduction

Gait recognition is a typical biometric technology based on the posture of walking. Recently, gait recognition is still challenging because its performance is heavily affected by many complex factors, including clothing, carrying conditions, cross-view, etc.

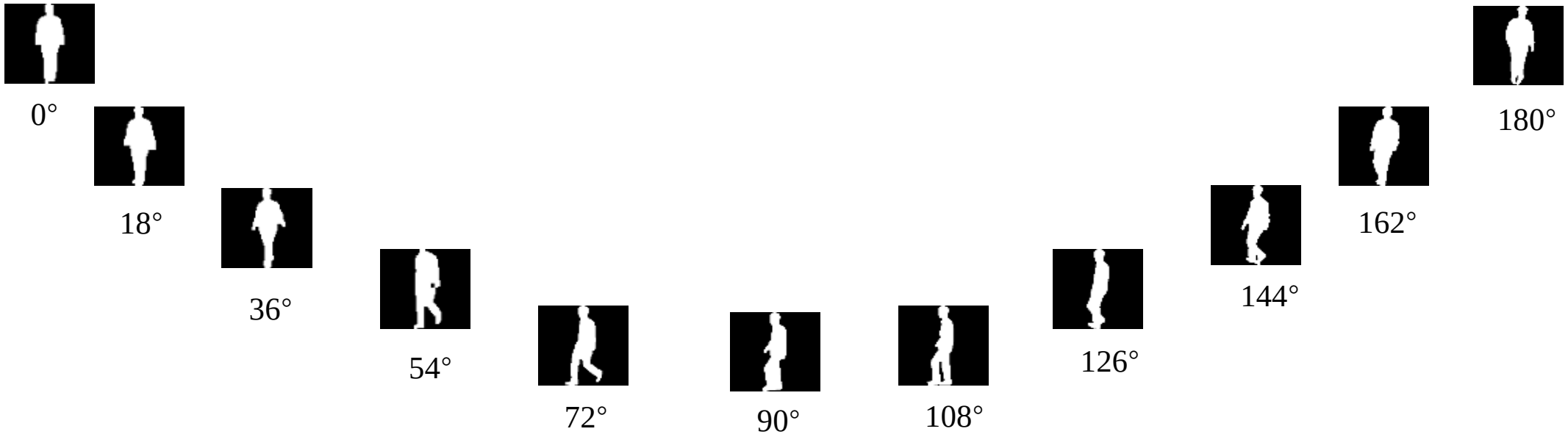

Figure 1 : The human gait with different view angles

Problem 1

Existing deep-learning-based methods can be divided into two different categories, as Fig.2(a)(b) shows.

The first type is called the template-based model.

Disadvantages:
It discards the spatial-temporal information.

The second type is called the sequence-based model.

Disadvantages:
3D CNN requires a fixed length of frames as input

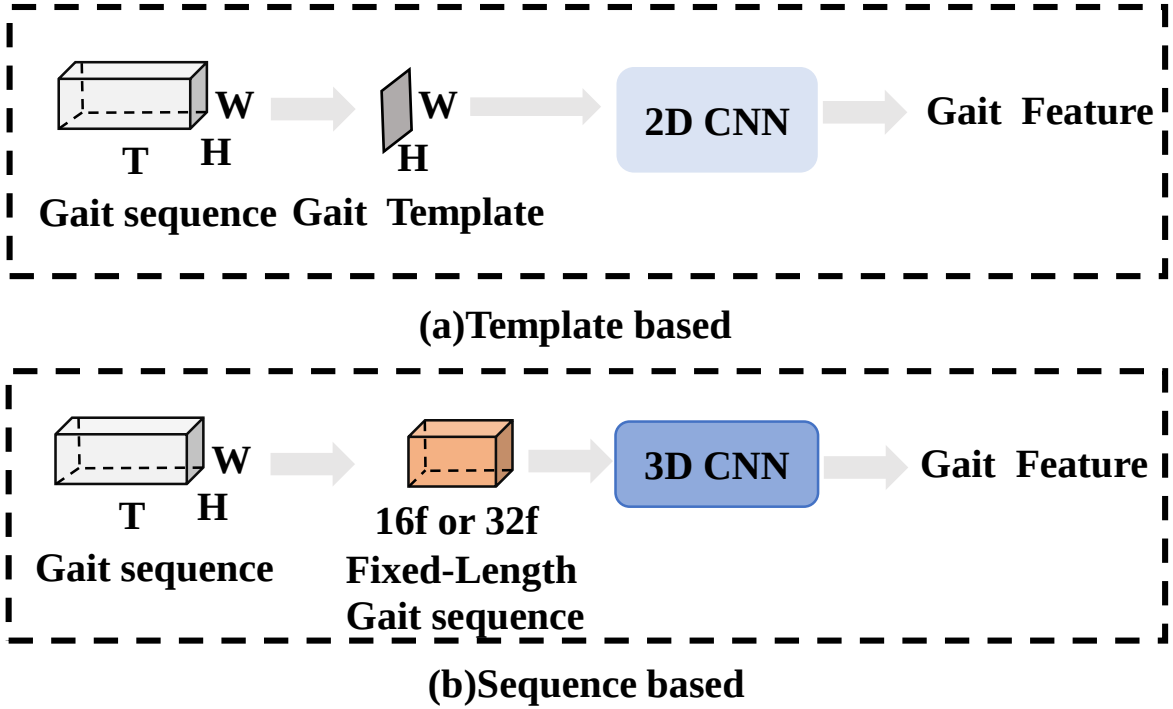

Figure 2 : The input in different levels for gait recognition

Contribution 1

To address the shortcomings in the current gait recognition methods, we propose a novel 3D CNN model with a temporal pooling module, which is shown in Fig.3.The temporal pooling operation is introduced for normalization.
The temporal pooling operation is used to make the input sequences have the same length, which makes the model fit the videos with different lengths.

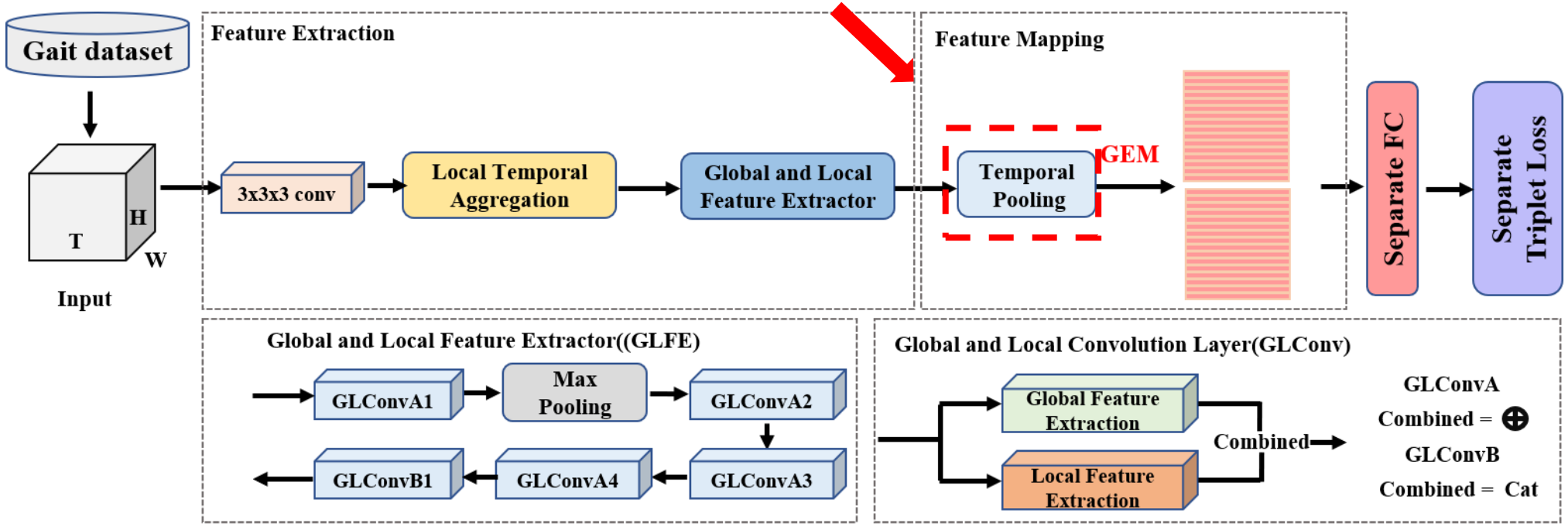

Figure 3 : Overview of the proposed method

Problem 2

Previous works use a specific pattern, “CL-SP-CL-SP-CL,” to extract features, where CL means convolutional layer and SP denotes the spatial pooling layer. However, the spatial information may be lost due to the SP downsampling operation.

Contribution 2

Considering that the temporal information in a gait sequence is periodic, we present the Local Temporal Aggregation operation to replace the first spatial pooling layer, which can integrate temporal information of local clips and maintain more spatial information.

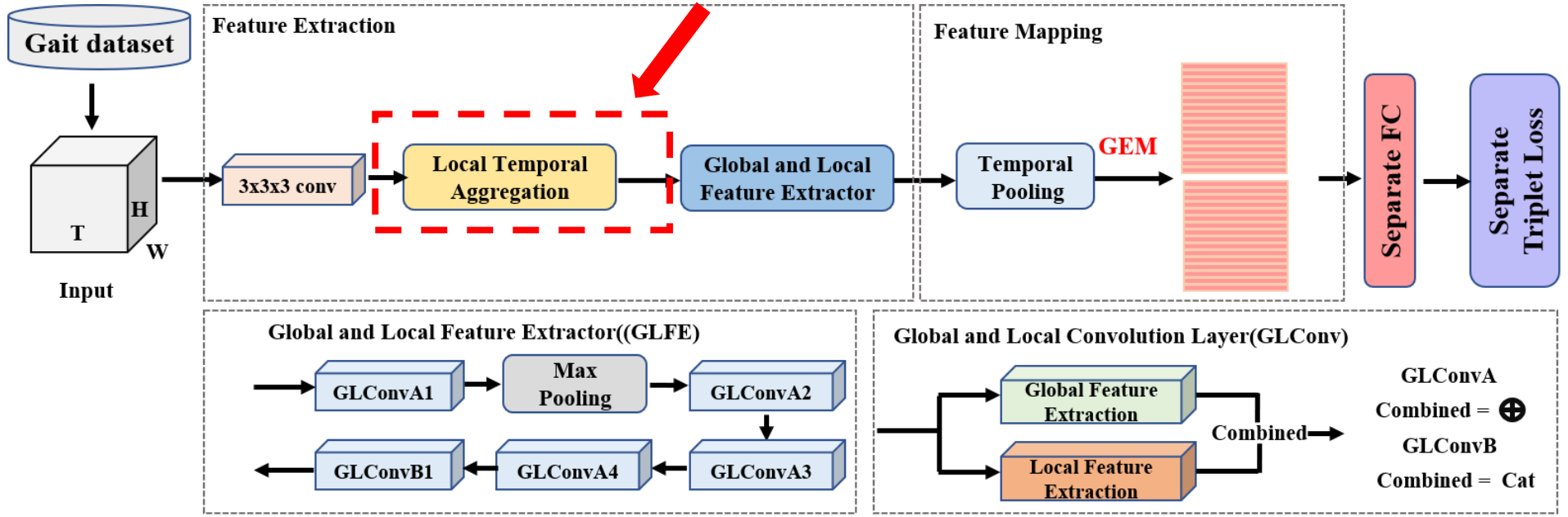

Figure 3 : Overview of the proposed method

Problem3

In general, the feature representations can be divided into two categories: global and local feature-based representation. Global feature-based representation methods extract gait features from whole gait frames. Local feature-based representation methods extract and combine local gait features from local gait parts.

However, the aforementioned methods only utilize either global or local features for representation, thus limiting the recognition performance.

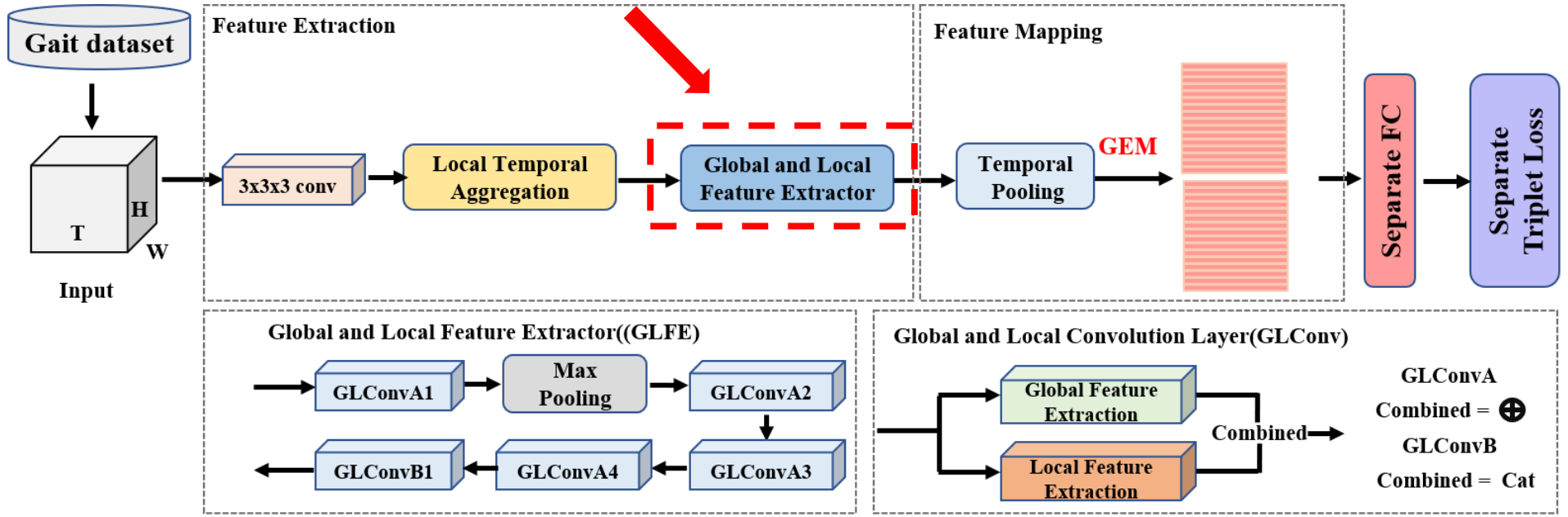

Figure 3 : Overview of the proposed method

Contribution 3

To address the above issues, we propose a novel cross-view gait recognition framework by learning effective representations from global and local features. we propose a novel **Global and Local Feature Extractor** module to extract features, which can take advantage of both global and local information.

The Global and Local Feature Extractor module is implemented by the **Global and Local Convolution Layer**, which contains global and local feature extractors. The global feature extractor can extract the whole gait information, while the local feature extractor is used to extract more details from local feature maps.

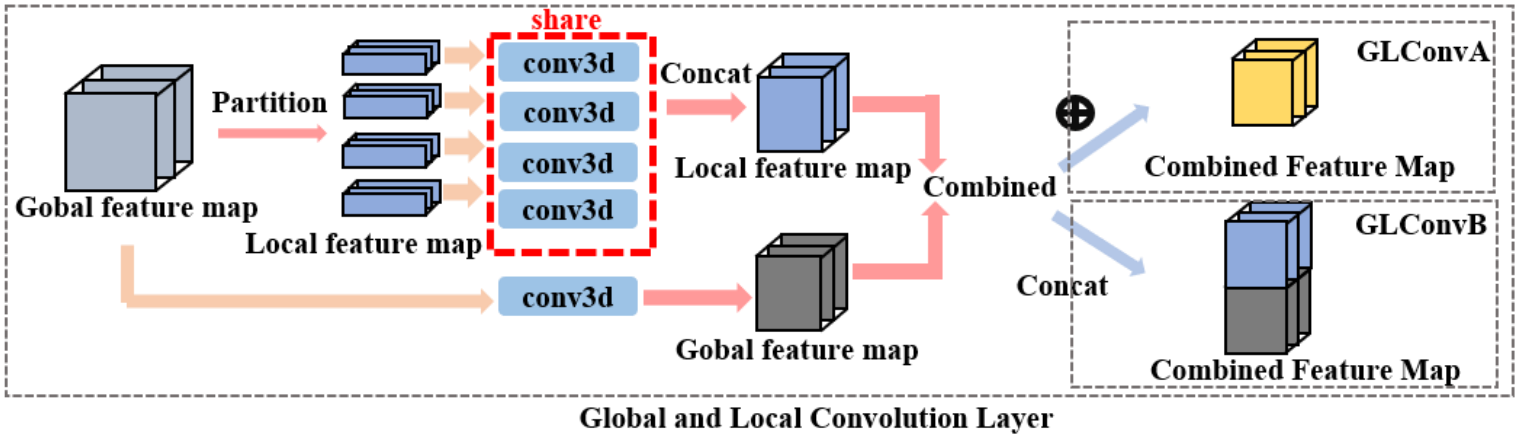

Figure 4: Architecture of Global and Local Convolution Layer

Ablation Study

The proposed recognition framework includes several key modules, e.g. Local Temporal Aggregation, Global and Local Feature Extractor. Hereby, we design different ablation studies to analyze the contribution of each key module. All experiments are conducted with the setting LT. (CASIA-B dataset, 74 subjects are chosen as the training set and the rest 50 subjects are used for test)

The GLFE module is composed of three global and local convolutional layers, i.e. the GLconv layers, each of which includes a global branch and a local branch. To explore the contribution of the global and local branches, we design the ablation study to explore the role of different branches.

*Normal Walking (NM)
*Walking with a Bag (BG)
*Walking in Coats (CL)

Layer Name	In_C	Out_C	Kernel	Global	N-part
First Conv	1	32	(3,3,3)	✓	×
LTA	32	32	(3,1,1)	–	–
GLConvA1	32	64	(3,3,3)	✓	2
Max Pooling, kernel size =(1, 2, 2), stride=(1, 2, 2)					
GLConvA2	64	128	(3,3,3)	✓	2
GLConvB1	128	128	(3,3,3)	✓	2

Table 1: Network parameters of the proposed method on the CASIA-B dataset

GLConvA1		GLConvA2		GLConvB1		NM	BG	CL
Global	N-parts	Global	N-parts	Global	N-parts			
✓	×	✓	×	✓	×	95.7	91.0	80.4
×	2	×	2	×	2	95.7	91.7	82.6
✓	2	✓	2	✓	2	96.4	92.7	83.0
✓	4	✓	4	✓	4	96.4	92.8	82.9
✓	8	✓	8	✓	8	96.3	92.5	82.9

Table 2: Rank-1 accuracy (%) of different combinations for GLFE module on the CASIA-B dataset

Ablation Study

To analyze the contribution of the Local Temporal Aggregation operation, we design the methods with different down-sampling strategies. The experiments are conducted with the LT settings on the CASIA-B dataset.

Existing gait recognition frameworks which always use 2D convolution to extract gait features, in this paper, the proposed framework is built by 3D convolution. Herewith, we design the ablation study to explore the effect of 2D and 3D convolution.

Downsampling		NM	BG	CL
1st pooling layer	2nd pooling layer			
SP	SP	95.5	88.2	78.5
SP	LTA	96.0	91.9	82.6
LTA	SP	**96.4**	**92.7**	**83.0**
LTA	LTA	90.6	84.3	69.6

Table 3: Accuracy(%) of different combinations of down-sampling

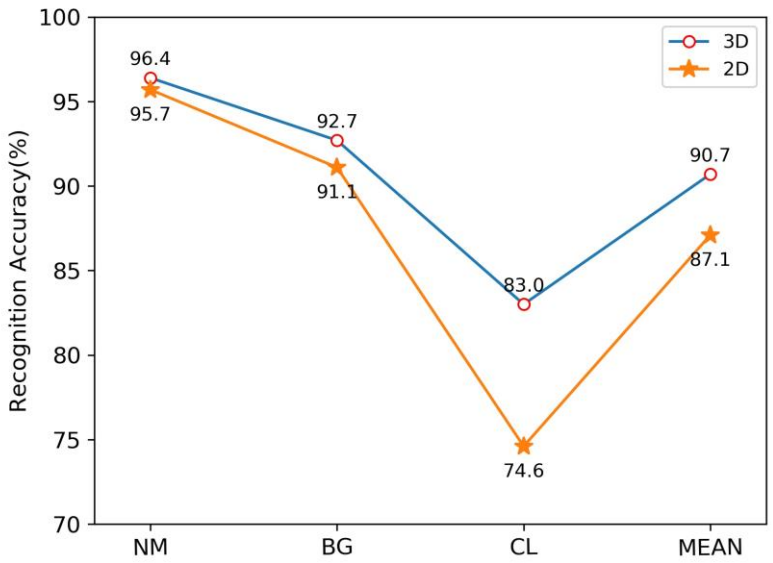

Figure 5 : Accuracy(%) of 2D or 3D Convolution

Experiments

Table 4 indicates that the proposed method achieves the best recognition accuracy on the CASIA-E dataset

Rank	ID	Accuracy
1	**BeibeiLin**	**63.0%**
2	**brl**	**54.1%**
3	**panfengzhang**	**53.4%**
4	**ctsu-ca**	**51.5%**
…..	**….**	**…..**
*****	**jilongwang**	**66.7%**

Table 4 : Rank-1 accuracy (%) on the CASIA-E dataset

Thank You!

北京交通大學