IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 9 August 2024, accepted 10 September 2024, date of publication 18 September 2024,
date of current version 30 September 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3463176

RESEARCH ARTICLE

Landmark-Based Adaptive Graph Convolutional Network for Facial Expression Recognition

DAQI ZHAO, JINGWEN WANG, HAOMING LI, AND DEQIANG WANG, (Senior Member, IEEE)
School of Information Science and Engineering, Shandong University, Qingdao 266237, China
Corresponding author: Deqiang Wang (wdq_sdu@sdu.edu.cn)

This work was supported by the National Key Research and Development Program of China under Grant 2020YFC0833203.

ABSTRACT Facial expression recognition (FER) is an important task in human emotion analysis. Though researchers have made great progress in FER, the semantic information lying in facial landmarks has not been fully exploited. In this paper, we propose a landmark-based adaptive graph convolutional network (LBAGCN) for FER. It takes facial landmarks as input in both phases of model training and inference and thus preserves sensitive privacy information well. Based on the facial landmark sequence extracted from a video clip, we define landmark motion sequence, muscle sequence and muscle motion sequence. Correspondingly, we construct four facial graphs with learnable topologies, named landmark graph, landmark motion graph, muscle graph and muscle motion graph. Each of the facial graphs contains useful semantic information relevant to facial expressions. The proposed LBAGCN consists of four parallel adaptive graph convolutional networks (AGCNs), each of which employs well-designed adaptive graph convolution blocks and multi-branch temporal convolutional network blocks to learn the spatial and temporal features from a facial graph automatically. The class scores of AGCNs are fused to boost the FER accuracy. Extensive experiments have been carried out to validate the proposed LBAGCN on the recognized CK+ dataset and Oulu CASIA dataset. Compared with state-of-the-art methods, the proposed LBAGCN achieves competitive FER accuracy especially on the challenging Oulu CASIA dataset. Testing results on practical platforms confirm that the proposed LBAGCN can be a feasible solution for real-time FER applications.

INDEX TERMS Adaptive graph convolutional network, facial expression recognition, facial graph representation, multi-stream network.

I. INTRODUCTION

In recent years, facial expression recognition (FER) has received increasing attention in the community of computer vision. FER plays an important role in many applications such as affective computing [1], [2] and human-computer interaction [3], [4]. A key challenge in achieving accurate FER is to capture dynamic movement information of facial muscles from videos. Earlier studies such as local binary patterns on three orthogonal planes (LBP-TOP) [5], histograms of oriented 3D spatio-temporal gradients (HOG 3D) [6] and 3D Inception-ResNet (3DIR) [7] mainly focused on extracting local and global features from videos. However, these approaches did not fully exploit the semantic information of facial muscle movements.

In biometric and neurocognitive research field, facial expressions are defined by a set of discrete components, named Action Units (AU), and the relationships between them [8], [9]. Clearly, using a facial graph to represent this kind of relational structure can better retain the underlying semantic information and avoid the interference of redundant information such as glasses and whiskers effectively.

Based on available facial landmarks, some pioneering researchers used facial graph representation to model facial muscle motion and proposed graph convolutional networks (GCNs) to predict facial expressions [10], [11], [12]. In [10], a facial graph was defined by taking 34 selected landmarks as nodes. The nodes were attributed by XY coordinates

The associate editor coordinating the review of this manuscript and approving it for publication was Moussa Ayyash.

of the landmarks and HOG features around the landmarks. The connections between nodes were defined according to psychological relationships, and hop-distance or Euclidean distance was taken as edge attribute. The proposed spatial temporal semantic graph network (STSGN) was composed of STSGN-G and STSGN-A extracting features from the XY coordinates and HOG features respectively. The extracted features were fused to gain better FER accuracy. In [11], the authors constructed a facial graph with predefined topology using 16 selected facial landmarks and 2 derived key points as nodes. Taking a facial image as input, the feature extraction module extracted global and local features using ResNet18 and key point-guided attention. The global and local features were concatenated and fed to the GCN for FER. In [12], guided by 22 selected facial landmarks, 20 regions of interest (ROI) were cropped from a face image and taken as nodes to construct facial graphs. Two GCNs, named Spectral Convolution and double dynamic relationships graph convolutional network (DDRGCN), were designed to extract features from facial graphs for FER. The Spectral Convolution extracted features from a facial graph consisting of all 20 ROIs with predefined topology. The DDRGCN was composed of two parallel sub-GCNs, each of which extracted features from a facial graph consisting of 10 ROIs (left or right face) with learnable topology, and achieved better accuracy. Clearly, DDRGCN neglected connections between ROIs in the left face and ROIs in the right face. Therefore, the dynamic relationships between ROIs are partially learnable indeed if all 20 ROIs are considered. In summary, these methods take HOG features of reginal facial images, facial image or ROIs of facial image as input besides facial landmarks, and hence pose potential risk of leaking sensitive privacy information [13], [14]. Moreover, a few facial landmarks are selected empirically to construct small-scale facial graphs with predefined or partially learnable topologies in these methods. These GCN-based methods are rather lightweight in parameters and computational cost. However, there is still room to explore and exploit the semantic information lying in available facial landmarks such that the FER accuracy can be improved effectively and meanwhile the sensitive privacy information can be preserved well.

In this paper, we propose a landmark-based adaptive graph convolutional network (LBAGCN) for FER. The proposed LBAGCN takes all available facial landmarks as input in both phases of model training and inference. This merit makes it conducive to the preserving of sensitive privacy information [13], [14]. Moreover, it is also suitable for face display sensors in virtual reality [15]. To make better use of available facial landmarks extracted from a video clip, we construct complementary facial graphs with learnable topologies and design adaptive graph convolutional networks (AGCNs) to extract discriminative features from the facial graphs. The constructed facial graphs reflect the dynamics of facial muscle movements well and have good interpretability. Meanwhile, the designed AGCNs can be used to extract spatial and temporal features from facial graphs effectively. Though no input features other than facial landmarks are used, the proposed LBAGCN achieves competitive FER accuracies on recognized datasets. The main contributions of this work are summarized as follows:

1) Based on facial landmarks detected from a video clip and the newly defined virtual facial muscles, we define four complementary sequences, named facial landmark sequence, landmark motion sequence, muscle sequence and muscle motion sequence. These sequences contain rich semantic information on facial expressions and have good interpretability.
2) Based on the defined sequences, four facial graphs with learnable topologies, named landmark graph, landmark motion graph, muscle graph and muscle motion graph, are constructed. These facial graphs are suitable for AGCNs to effectively learn the underlying semantic information and thus gain promising FER accuracy.
3) A LBAGCN scheme with multi-stream AGCN architecture is proposed for FER. The AGCNs are designed according to the defined facial graphs to learn both semantic features and graph topologies. The merit of learnable graph topology enables better learning of semantic information in facial graphs.
4) Experiments have been carried out to demonstrate the effectiveness of the proposed method on two popular datasets, i.e., (CK+) [16] and Oulu CASIA [17]. Compared with state-of-the-art methods, the proposed LBAGCN achieves competitive FER accuracy especially on the more challenging Oulu CASIA dataset.

The remainder of this paper is organized as follows. Section II reviews the relevant works on GCNs and FER. Section III illustrates our proposed LBAGCN. In Section IV, experiment results are presented and discussed. Section V concludes the paper.

II. RELATED WORKS

A. GRAPH CONVOLUTIONAL NETWORKS

To process non-Euclidean data like graphs, GCNs have been developed and studied by researchers broadly [18], [19], [20], [21], [22]. There are mainly two types of GCNs, namely Spectral Convolutional Networks [19], [20], [21], [23] and Spatial Convolutional Networks [19], [21], [23].

Earlier studies focused on spectral domain of graph signals. Defferrard et al. [18] and Bruna et al. [20] constructed spectral filters for feature extraction to realize handwritten digit recognition. Hammond et al. [22] proposed a method combining Wavelet Transform with Graph Fourier Transform to gain better performance. However, a Spectral Convolutional Network relies much on the Laplacian eigen basis, which is closely associated with the graph structure, and thus can be applied to graphs with the same structure only.

Kipf and Welling introduced a first-order approximation to local spectral convolution in [19] and pioneered the study of Spatial Convolutional Networks. Hamilton et al. [21] proposed GraphSAGE to use node feature information, e.g.,

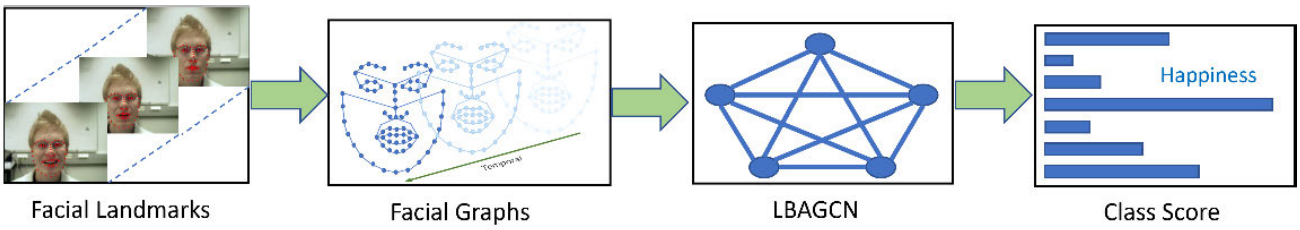


FIGURE 1. Pipeline of the proposed LBAGCN FER scheme. The input 68 facial landmarks (shown as red dots) extracted from videos are taken as input. Facial graphs are then constructed and fed to the following LBAGCN. Finally, the LBAGCN extracts features from the facial graphs and outputs a fused class score.

text attributes, to generate node embeddings for previously unseen data efficiently. Xu et al. [23] presented a theoretical framework to analyze the representational power of Spatial Convolutional Networks. Recently, Spatial Convolutional Networks have been widely applied to various tasks [24], [25], [26].

B. FACIAL EXPRESSION RECOGNITION

Earlier FER methods based on machine learning employed hand-crafted features such as geometry, appearance, and texture. Shan et al. [27] evaluated facial representation empirically based on statistical local features, namely Local Binary Patterns (LBP). Taking motion information in time dimension into consideration, Zhao and Pietikainen [28] introduced the LBP-TOP. Klaser et al. [6] presented a local descriptor based on HOG 3D.

In recent years, researchers have made great progress in deep learning (DL) based FER, where convolutional neural networks (CNNs) and recurrent neural networks (RNNs) are employed to extract features. Jung et al. [29] proposed the deep temporal appearance-geometry network (DTAGN), where a CNN-based deep temporal appearance network (DTAN) was used to extract appearance features from facial image sequences and a multilayer perception-based deep temporal geometry network (DTGN) was employed to extract geometric features from facial landmark sequences. Zhang et al. [30] proposed a PHRNN-MSCNN method for FER. The part-based hierarchical bidirectional recurrent neural network (PHRNN) was used to extract temporal features from landmark sequence, and the multi-signal convolutional neural network (MSCNN) was used to extract spatial features from still frames. Sun et al. [31] proposed a multi-attention shallow and deep model (MSDModel) for FER, where the RNN-based attention shallow model (ASModel) was employed to extract shallow features from facial landmarks and the CNN-based attention deep model (ADModel) was employed to extract deep features from facial images. Ding et al. [32] proposed the FaceNet2ExpNet to address the issue of relatively small facial expression dataset size, where a well-trained face recognition network was used as supervisor to pre-train the convolutional layers of the FER network. Aiming at achieving intensity-invariant FER, Zhao et al. [33] presented peak-piloted deep network (PPDN), which drove the intermediate-layer features of the non-peak expression sample towards those of the peak expression sample.

Along with the development of spectral graph theory, efforts have been made by researchers to implement FER using GCNs. Zhou et al. [10] proposed a two-stream GCN, named STSGN. In STSGN, two sub-networks, STSGN-G and STSGN-A, took the XY coordinates of the facial landmarks and the HOG features extracted from the regional images around the landmarks as input respectively. The final FER was realized by fusing the outputs of STSGN-G and STSGN-A. Liao et al. [11] adopted a key point-guided attention branch and a CNN branch with triplet attention to extract features from input images for their GCN. Moreover, Jin et al. [12] calculated the centers of 20 ROIs related to facial AUs guided by facial landmarks and cropped the ROIs from face image. The cropped ROIs were used to construct facial graphs for FER. Two GCN-based methods, named Spectral Convolution and DDRGCN were proposed to extract features from the constructed facial graphs. Differently from the above mentioned GCN-based methods, our proposed LBAGCN takes spatial location information of the facial landmarks as input only and adopts adaptive graph topology to enable better learning of the underlying semantic information.

III. OUR METHOD

The pipeline of our proposed FER scheme is illustrated in Fig. 1. Given a video clip, Uniform Sampling [34] is usually performed to remove redundancy at frame level. Then, 68 facial landmarks can be obtained frame by frame by using a facial landmark detection algorithm, e.g., the algorithm in [35]. As shown in Fig. 1, the proposed FER scheme takes the facial landmark sequence as input. Based on the facial landmark sequence and its derivatives, four facial graphs, named landmark graph, landmark motion graph, muscle graph and muscle motion graph, are constructed and fed to the following LBAGCN. The LBAGCN consists of four parallel AGC networks (AGCN) as shown in Fig. 2. Each AGCN extracts semantic feature from one facial graph and outputs a class score at its final Softmax layer. The final class score is obtained by calculating a weighted sum of the outputs of all four AGCNs. In what follows, we introduce facial landmark sequence and its derivatives, the facial graphs, the principle

of AGC, the architecture of AGCN and fusion of class scores.

A. FACIAL LANDMARK SEQUENCE AND ITS DERIVATIVES

Given a down-sampled video clip of T frames, the *facial landmark sequence* $X \in R^{T\times N\times 3}$ is defined as

$$X = [\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_T] \tag{1}$$

where, $\mathbf{X}_t \in R^{N\times 3}, t = 1, 2, \ldots, T$ represents the N landmarks detected in frame t and is formulated as

$$\mathbf{X}_t = \begin{bmatrix} x_{t,1} & y_{t,1} & p_{t,1} \\ x_{t,2} & y_{t,2} & p_{t,2} \\ \vdots & \vdots & \vdots \\ x_{t,N} & y_{t,N} & p_{t,N} \end{bmatrix}. \tag{2}$$

Note that the n-th row of $\mathbf{X}_t$ in (2), i.e., $(x_{t,n}, y_{t,n}, p_{t,n})$, represents the attributes of the n-th landmark, where $(x_{t,n}, y_{t,n})$ is the landmark coordinate and $p_{t,n}$ is the confidence probability.

Based on the *facial landmark sequence*, we define three new sequences, named *landmark motion sequence*, *muscle sequence* and *muscle motion sequence*. Detailed definitions are provided in the following.

1) LANDMARK MOTION SEQUENCE

The *landmark motion sequence* is defined to extract facial expression information lying in motion dynamics of facial landmarks. Specifically, the motion attributes of the n-th landmark at frame t are notated by a triplet $(\tilde{x}_{t,n}, \tilde{y}_{t,n}, \tilde{p}_{t,n})$ and computed as

$$\begin{cases} \tilde{x}_{t,n} = x_{t,n} - x_{t+1,n} \\ \tilde{y}_{t,n} = y_{t,n} - y_{t+1,n} \\ \tilde{p}_{t,n} = \left(p_{t,n} + p_{t+1,n}\right) \times 0.5. \end{cases} \tag{3}$$

Based on equation (3), the *landmark motion sequence* $\tilde{X} \in R^{T\times N\times 3}$ is defined as

$$\tilde{X} = \left[\tilde{\mathbf{X}}_1, \tilde{\mathbf{X}}_2, \cdots, \tilde{\mathbf{X}}_T\right] \tag{4}$$

where, $\tilde{\mathbf{X}}_t \in R^{N\times 3}$ represents the landmark motion feature of frame t and is given by

$$\tilde{\mathbf{X}}_t = \begin{bmatrix} \tilde{x}_{t,1} & \tilde{y}_{t,1} & \tilde{p}_{t,1} \\ \tilde{x}_{t,2} & \tilde{y}_{t,2} & \tilde{p}_{t,2} \\ \vdots & \vdots & \vdots \\ \tilde{x}_{t,N} & \tilde{y}_{t,N} & \tilde{p}_{t,N} \end{bmatrix}. \tag{5}$$

2) MUSCLE SEQUENCE

In reality, facial expressions are highly related to the facial muscles and their motions. We define virtual facial muscles based on landmarks detected. The concept of virtual muscle is partly inspired by the skeleton-based action recognition scheme in [24]. Given 68 facial landmarks, we define totally 74 virtual facial muscles, each of which corresponds to a line linking two adjacent landmarks as illustrated in Fig. 3(b). The attributes of the m-th virtual facial muscle, which links two adjacent landmarks i and j, in frame t are notated by a triplet $(\bar{x}_{t,m}, \bar{y}_{t,m}, \bar{p}_{t,m})$ and computed as

$$\begin{cases} \bar{x}_{t,m} = x_{t,i} - x_{t,j} \\ \bar{y}_{t,m} = y_{t,i} - y_{t,j} \\ \bar{p}_{t,m} = \left(p_{t,i} + p_{t,j}\right) \times 0.5. \end{cases} \tag{6}$$

The *muscle sequence* $\bar{X} \in R^{T\times M\times 3}$ is defined as

$$\bar{X} = \left[\bar{\mathbf{X}}_1, \bar{\mathbf{X}}_2, \cdots, \bar{\mathbf{X}}_T\right] \tag{7}$$

where, $\bar{\mathbf{X}}_t \in R^{M\times 3}$ represents the virtual muscle feature of frame t and is given by

$$\bar{\mathbf{X}}_t = \begin{bmatrix} \bar{x}_{t,1} & \bar{y}_{t,1} & \bar{p}_{t,1} \\ \bar{x}_{t,2} & \bar{y}_{t,2} & \bar{p}_{t,2} \\ \vdots & \vdots & \vdots \\ \bar{x}_{t,M} & \bar{y}_{t,M} & \bar{p}_{t,M} \end{bmatrix}. \tag{8}$$

3) MUSCLE MOTION SEQUENCE

The *muscle motion sequence* reflects the motion dynamics of virtual facial muscles. The motion attributes of the m-th virtual facial muscle at frame t are notated by a triplet $(\hat{x}_{t,m}, \hat{y}_{t,m}, \hat{p}_{t,m})$ and computed as

$$\begin{cases} \hat{x}_{t,m} = \bar{x}_{t,m} - \bar{x}_{t+1,m} \\ \hat{y}_{t,m} = \bar{y}_{t,m} - \bar{y}_{t+1,m} \\ \hat{p}_{t,m} = \left(\bar{p}_{t,m} + \bar{p}_{t+1,m}\right) \times 0.5. \end{cases} \tag{9}$$

The *muscle motion sequence* $\hat{X} \in R^{T\times M\times 3}$ is defined as

$$\hat{X} = \left[\hat{\mathbf{X}}_1, \hat{\mathbf{X}}_2, \cdots, \hat{\mathbf{X}}_T\right], \tag{10}$$

where, $\hat{\mathbf{X}}_t \in R^{M\times 3}$ represents the muscle motion feature at frame t and is formulated as

$$\hat{\mathbf{X}}_t = \begin{bmatrix} \hat{x}_{t,1} & \hat{y}_{t,1} & \hat{p}_{t,1} \\ \hat{x}_{t,2} & \hat{y}_{t,2} & \hat{p}_{t,2} \\ \vdots & \vdots & \vdots \\ \hat{x}_{t,M} & \hat{y}_{t,M} & \hat{p}_{t,M} \end{bmatrix}. \tag{11}$$

B. CONSTRUCTION OF FACIAL GRAPHS

Based on the afore defined sequences, we construct four facial graphs, i.e., landmark graph, landmark motion graph, muscle graph and muscle motion graph. The same construction method is used for all these face graphs. Without loss of generality, we introduce the construction of landmark graph in the following.

The topology of the landmark graph consists of two parts: the nodes and the edges. The nodes are facial landmarks detected. The edges represent the relationships between nodes. The mathematical description of the landmark graph is given in **Definition 1**.

Definition 1: Let $G_{LG} = (V_{LG}, E_{LG})$ denote a landmark graph constructed from a facial landmark sequence with T frames and N facial landmarks within each frame. The set of

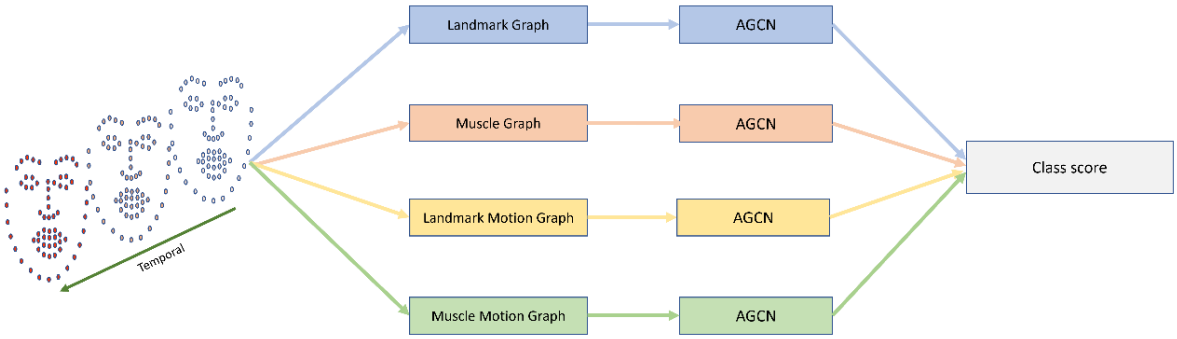
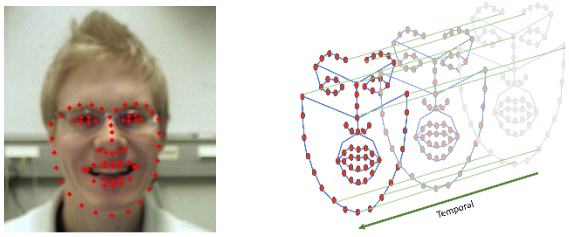


FIGURE 2. Overview of the four-stream AGCN framework adopted by LBAGCN. Based on the raw data of facial landmarks, four facial graphs, named landmark graph, landmark motion graph, muscle graph and muscle motion graph, are constructed and fed to the following AGCNs. Each AGCN extracts features from the corresponding graph and outputs a class score. The final class score is obtained by using a weighted summation method.



(a) facial landmarks (b) facial graph representation

FIGURE 3. Illustration of landmark graph. (a) Example of facial landmarks extracted (red dots indicate facial landmarks). (b) The initial connectivity of landmark graph. The blue lines indicate the initial spatial connectivity of the edges, which is derived from the semantic relationships between the nodes. The green lines indicate the fixed temporal connectivity of the edges (for simplicity, only a portion is shown).

nodes is given by $V_{LG} = \{v_{t,i} | t = 1, \ldots, T; i = 1, \ldots, N\}$, *where* $v_{t,i}$ *denotes the i-th landmark in the t-th frame with attributes* $(x_{t,i}, y_{t,i}, p_{t,i})$. *The set of edges* E_{LG} *consists of two subsets, named spatial subset* E^S_{LG} *and temporal subset* E^T_{LG}. *The spatial subset is given by* $E^S_{LG} = \{s_{t,i,j} | t = 1, \ldots, T;\ i, j = 1, \ldots, N; i \neq j\}$, *where* $s_{t,i,j}$ *represents the connection strength between the i-th landmark and the j-th landmark within the t-th frame. The temporal subset is given by* $E^T_{LG} = \{s_{t_1,t_2,i} | t_1, t_2 = 1, \ldots, T; i = 1, \ldots, N; t_1 \neq t_2\}$, *where* $s_{t_1,t_2,i}$ *represents the connection strength between the i-th landmark in the* t_1*-th frame and that in the* t_2*-th frame.*

For clarity, the landmark graph is illustrated in Fig. 3. In this study, the number of nodes per frame is 68. For the temporal subset E^T_{LG}, the connection strength of an arbitrary edge is set to be $s_{t_1,t_2,i} = 1$. For the spatial subset E^S_{LG}, the connection strength $s_{t,i,j}$ is defined to be independent of t and can be simplified as $s_{i,j}$ by dropping the subscript t. The connection strengths, $\{s_{i,j} | i, j = 1, \ldots, N, i \neq j\}$ are initialized sparsely as shown in Fig. 3(b) and updated iteratively by performing adaptive graph convolution (AGC) during model training. The resultant connection strengths perform better than fixed sparse connections in modeling semantic relationships between facial landmarks.

C. ADAPTIVE GRAPH CONVOLUTION

AGC is at the core of our proposed LBAGCN. As mentioned above, each of the four facial graphs is fed to an AGCN to extract semantic feature and gain a class score. Taking the landmark graph $G_{LG} = (V_{LG}, E_{LG})$ as an example, we provide details on AGC from spatial perspective. An implementation of graph convolution network similar to that in [19] is adopted.

Given the landmark graph $G_{LG} = (V_{LG}, E_{LG})$, an adjacency matrix $\mathbf{A} \in R^{N \times N}$ is defined according to the connection strengths of edges in the spatial subset E^S_{LG} as

$$\mathbf{A} = \begin{bmatrix} 0 & s_{1,2} & \cdots & \cdots & s_{1,N} \\ s_{2,1} & 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 & s_{N-1,N} \\ s_{N,1} & \cdots & \cdots & s_{N,N-1} & 0 \end{bmatrix}. \quad (12)$$

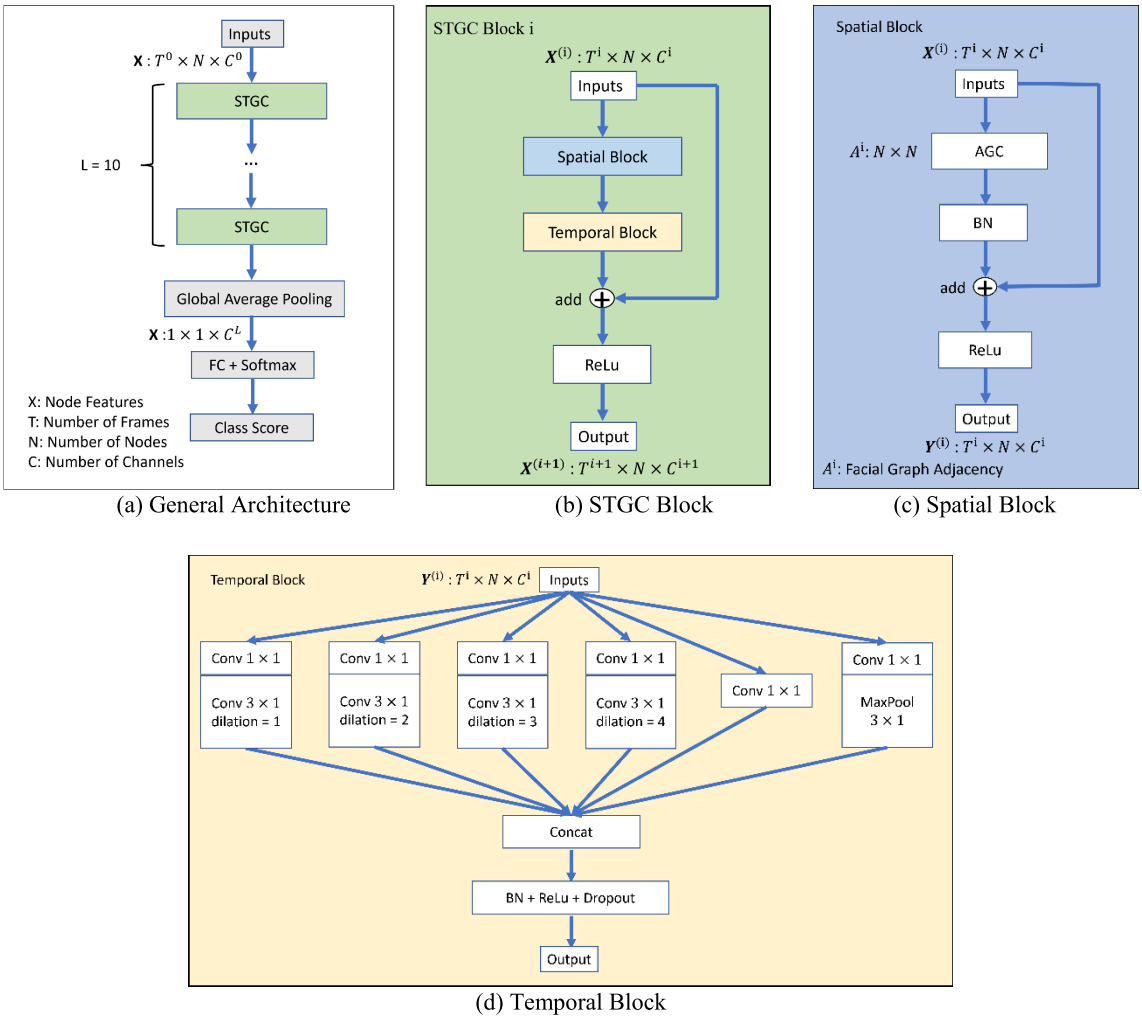
The graph topology of the AGC is defined by the adjacency matrix $\mathbf{A}$ with added self-loops, i.e., $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}$, where $\mathbf{I}$ is identity matrix. The features of nodes in V_{LG}, i.e., the landmark sequence X, are input of the AGC. The learnable weight matrix of the AGC is denoted by $\mathbf{W} \in R^{C \times F}$, where C is the number of input channels, and F denotes the number of output feature maps.

For features of nodes at frame t, i.e., $\mathbf{X}_t$, the layer-wise update rule of the AGC can be given by

$$\mathrm{AGC}(\mathbf{X}_t) = \tilde{\mathbf{D}}^{-\frac{1}{2}} \tilde{\mathbf{A}} \tilde{\mathbf{D}}^{-\frac{1}{2}} \mathbf{X}_t \mathbf{W}, \quad (13)$$

where, $\tilde{\mathbf{D}}$ is the diagonal degree matrix of $\tilde{\mathbf{A}}$. The term $\tilde{\mathbf{D}}^{-\frac{1}{2}} \tilde{\mathbf{A}} \tilde{\mathbf{D}}^{-\frac{1}{2}} \mathbf{X}_t$ can be interpreted as an approximate spatial mean feature aggregation from the neighborhood.

Note that, during the back-propagation, the graph topology $\tilde{\mathbf{A}}$ is also updated to preserve as much semantic information as possible. Thus, the graph topology $\tilde{\mathbf{A}}$ is not only adaptive to the FER task, but also adaptive to the input features of the AGC.



(a) General Architecture (b) STGC Block (c) Spatial Block

(d) Temporal Block

FIGURE 4. Overview of the AGCN architecture. (a) The general architecture of AGCN. (b) The overall design of the Spatial Temporal Graph Convolution (STGC) Block. A STGC Block contains a Spatial Block and a Temporal Block, which are used to extract the spatial and temporal features of the input respectively. (c) Detailed structure of the Spatial Block. The adjacency matrix $\mathbf{A}^{(i)}$ of each Spatial Block exists independently and is updated iteratively as the parameter of Adaptive Graph Convolution (AGC). (d) Detailed structure of the Temporal Block. Dilation convolutions are employed to realize multi-scale temporal feature extraction.

D. ARCHITECTURE OF AGCN

The four parallel AGCNs of the proposed LBAGCN have the same architecture except for their input facial graphs. The architecture and building blocks of an AGCN are shown in Fig. 4. In what follows, we provide detailed information.

1) OVERALL ARCHITECTURE

As shown in Fig. 4(a), an AGCN consists of 10 spatial-temporal graph convolution (STGC) blocks, a global average pooling layer, a fully connected layer and a Softmax classifier. The numbers of channels of the 10 STGC blocks are 64, 64, 64, 64, 128, 128, 128, 256, 256 and 256 accordingly. Stridden temporal convolutions are employed at the fifth and eighth STGC blocks to halve the temporal dimension.

The structure of STGC block is shown in Fig. 4(b). A STGC block consists of a spatial block (SB), a temporal block (TB), a residual connection (RC) and a ReLu activation layer. Let $X^{(i)}$ represent the input of the i-th STGC block, the output can be formulated as

$$X^{(i+1)} = \sigma\left(\mathrm{TB}\left(\mathrm{SB}\left(X^{(i)}\right)\right) + \mathrm{RC}\left(X^{(i)}\right)\right), \tag{14}$$

and

$$\mathrm{RC}\left(X^{(i)}\right) = \begin{cases} X^{(i)}, & i \neq 5, 8 \\ \mathrm{DS}\left(X^{(i)}\right), & i = 5, 8, \end{cases} \tag{15}$$

where, $\sigma(\cdot)$ is the ReLu activation function, and $\mathrm{DS}(\cdot)$ is a down-sampling function to adjust the time dimension of $X^{(i)}$ to be consistent with $\mathrm{TB}\left(\mathrm{SB}\left(X^{(i)}\right)\right)$ at the fifth and eighth STGC blocks.

2) SPATIAL BLOCK

SBs are built based on the above introduced AGC. The detailed design of a SB is shown in Fig. 4(c). A SB consists of an AGC, a RC, a BatchNorm layer and a ReLu activation

layer. For the i-th STGC block, the AGC is defined by the corresponding adjacency matrix $\mathbf{A}^i$. In the phase of model training, $\tilde{\mathbf{A}}^i$ is initialized sparely according to some predefined topology like Fig. 3(b) and updated iteratively using gradient descent to learn a better topology without sparse constraint. Here, the RC is introduced to improve the spatial modelling capability. Given the input $X^{(i)}$, the output of the SB can be expressed as

$$\mathrm{SB}\left(X^{(i)}\right) = \sigma\left(\mathrm{BN}\left(\mathrm{AGC}\left(X^{(i)}\right)\right) + X^{(i)}\right), \tag{16}$$

where, BN(·) is the BatchNorm function.

3) TEMPORAL BLOCK

TBs are designed to extract temporal features effectively. In order to capture multi-scale temporal features, we adopt multi-branch temporal convolutional network (TCN) [36], [37] instead of the conventional single-branch design. Detailed design of a TB is shown in Fig. 4(d). As can be seen, the designed multi-branch TCN consists of six branches including a 1×1 Conv branch, a Max-Pooling branch, and four temporal 1D Conv branches with kernel size 3 and dilations from 1 to 4. The outputs are concatenated, and then a BatchNorm layer, a ReLu activation layer and a dropout layer are employed for further processing. This design is lightweight in both parameters and computational cost. For the i-th TB, given the input $Y^{(i)} \in R^{T^i \times N \times C^i}$, the output can be expressed as

$$\mathrm{TB}\left(Y^{(i)}\right) = \mathrm{Dropout}\left(\sigma\left(\mathrm{BN}\left(\mathrm{TCN}\left(Y^{(i)}\right)\right)\right)\right), \tag{17}$$

where, TCN(·) denotes the function of multi-branch TCN, and Dropout(·) is the dropout function used to prevent overfitting.

E. FUSION OF CLASS SCORES

Given a video clip, the final FER result is obtained by fusing the outputs of the four AGCNs as shown in Fig. 2. Since the contributions of the constructed facial graphs can be different from each other, a weighted summation method is adopted to fuse class scores from all four AGCNs. Specifically, the final class score is given by

$$\mathbf{S} = \sum_{i=1}^{4} \alpha_i \mathbf{S}_i \tag{18}$$

where, $\mathbf{S}_i$ represents the class score vector output at the final Softmax layer of the i-th AGCN, α_i denotes the corresponding weighting factor.

IV. EXPERIMENTS AND DISCUSSIONS

In this section, we evaluate the performance of the proposed LBAGCN on two recognized FER datasets, i.e., Extended Cohn-Kanade (CK+) [16] and Oulu CASIA [17].

A. DATASETS

1) CK+ DATASET

CK+ contains 593 image sequences collected from 123 subjects aged between 18 and 50 years. The image solution is 640×480 pixels. The image sequences vary in duration from 10 frames to 60 frames and incorporate the onset to peak formation of facial expressions. Position annotations of face landmarks are provided for all sequences. In this dataset, seven discrete emotions, i.e., anger, contempt, disgust, fear, happiness, sadness, and surprise, are considered. Totally 327 image sequences have been annotated with emotion labels. In our experiments, the confidence probabilities of face landmarks are set to be 1.0.

2) Oulu CASIA DATASET

Oulu CASIA consists of 480 video sequences of resolution 320×240 pixels collected from 80 subjects aged between 23 and 58 years. Six discrete emotions, i.e., anger, disgust, fear, happiness, sadness and surprise, are taken into consideration. Expressions are captured in strong, weak and dark illumination conditions. All video sequences have been annotated with emotion labels. Each sequence starts at the onset frame and ends at the apex frame of the corresponding expression. However, this dataset does not provide information on facial landmarks. In our experiments, the algorithm in [35] is employed to estimate face landmarks. Oulu CASISA is comparatively more challenging than CK+ in two aspects: image resolution and facial landmark availability.

B. DATA PRE-PROCESSING AND MODEL TRAINING

1) DATA PRE-PROCESSING

Uniform Sampling [34] is adopted to realize down-sampling and data augmentation. For CK+ dataset, an original sequence of facial landmarks is divided uniformly into T segments and one frame is sampled randomly per segment. The sampled facial landmark frames are concatenated again to form an input sequence for the LBAGCN. In this way, a large number of facial landmark sequences can be generated from a video clip. Note that if T is greater than the length of a video clip, we will repeat the video clip to compensate. For Oulu CASIA dataset, the same processing is conducted except that the algorithm in [35] is employed to obtain facial landmarks from the video clips.

2) MODEL TRAINING

Our proposed LBAGCN is implemented by using PyTorch. Cross-entropy is applied as the loss function. Stochastic gradient descent (SGD) optimizer with Nesterov momentum is employed during model training. We set the initial learning rate to 0.1, batch size to 16, and train each model for 100 epochs with the CosineAnnealing LR scheduler. For the optimizer, we set the momentum to 0.9, and weight decay to 5×10^{-4}. The 10-fold cross-validation [38] is adopted on both two datasets. A INVIDIA GeForce RTX 3080Ti GPU is employed for model training and testing.

C. METRICS

In our experiments, number of parameters, computational cost (FLOPs), F1 score and accuracy are taken as metrics to

TABLE 1. Accuracies achieved with different block design on the OULU CASIA dataset.

AGC	TCN	Accuracy (%)
×	×	80.21
×	√	81.04
√	×	84.79
√	√	**87.71**

evaluate the performance of our proposed LBAGCN. Mathematically, the accuracy is defined as

$$\text{Acc} = \frac{1}{10}\sum_{i=1}^{10} \text{Acc}_i, \tag{19}$$

where, Acc_i represents the accuracy obtained with the i-th fold experiment and can be calculated by

$$\text{Acc}_i = \frac{\text{CPL}_i}{\text{GTL}_i}, \tag{20}$$

where, GTL_i and CPL_i are the total number of ground truth labels and that of correct predictions in the i-th fold experiment.

D. ABLATION STUDIES

We conduct ablation studies to evaluate the effectiveness of building blocks of the proposed LBAGCN. Without loss of generality, experiment results on the Oulu CASIA dataset are presented and discussed.

1) EFFECTIVENESS OF AGC AND TCN

As shown in Fig. 4, the AGC in Spatial Block and the multi-branch TCN in Temporal Block are key building blocks of the AGCN. We evaluate the contributions of the AGC and the multi-branch TCN via ablation experiments. For the purpose of comparison, variants of the AGCN are built by removing the AGC and/or the multi-branch TCN. When the AGC is removed, a GCN with fixed graph topology defined by the corresponding adjacency matrix is employed instead. When the multi-branch TCN is removed, a 1D convolution with kernel size 9 is employed instead. In all experiments, T is set to 16 and optimized weights are used to fuse the outputs of the four parallel AGCNs. Experiment results are given in Table 1.

Based on Table 1, we analyze the contributions of the AGC and the multi-branch TCN. In the baseline case that neither the AGC nor the TCN is used, the lowest accuracy of 80.21% is achieved. When the TCN is employed alone, the accuracy is 81.04%. Compared with the baseline, the multi-branch TCN brings an increase of 0.83% in accuracy. It tells that the multi-scale feature extraction ability of the TCN is beneficial to the accuracy. When the AGC is employed alone, the accuracy is 84.79% exhibiting a more significant increase of 4.58%. It reveals that the AGC with learnable graph topology does perform better than a GCN with fixed graph topology in terms of semantic information extraction. Moreover, when both the AGC and the multi-branch TCN are employed, our proposed model achieves the best accuracy of 87.71%, which is 7.50% higher than that of the baseline. These observations confirm the effectiveness of the AGC and the multi-branch TCN.

TABLE 2. Accuracies achieved with different stream combinations on the OULU CASIA dataset.

Landmark	Landmark Motion	Muscle	Muscle Motion	Accuracy (%)
×	√	√	√	83.75
√	√	×	√	84.17
√	×	√	√	85.00
√	√	√	×	85.00
√	√	√	√	**87.71**

TABLE 3. Accuracies achieved with different time segments on the OULU CASIA dataset.

T	Accuracy (%)
8	84.38
16	**87.71**
32	85.00
64	83.33

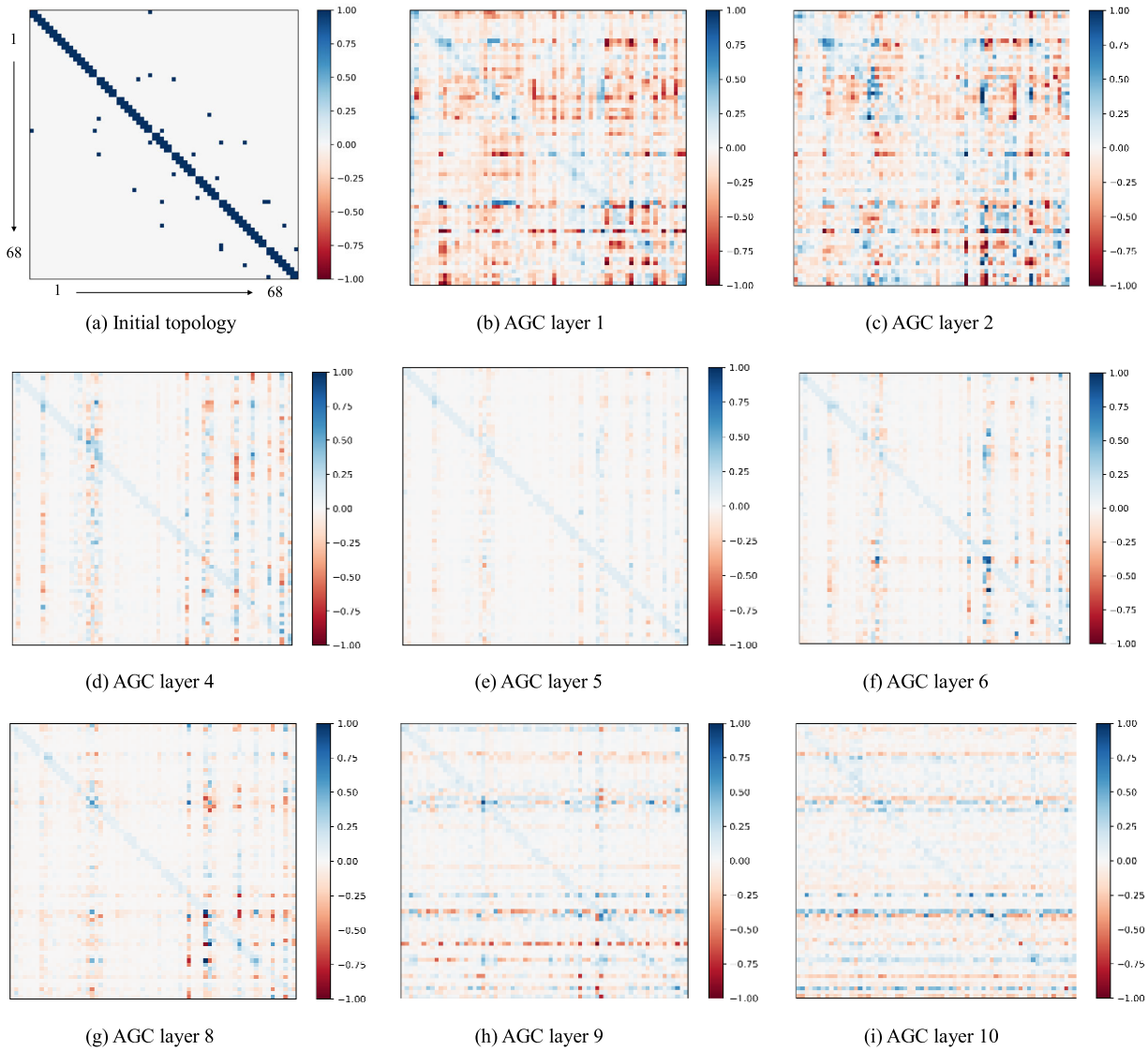
2) EFFECTIVENESS OF MULTI-STREAM DESIGN

As shown in Fig. 2, the proposed LBAGCN has a four-stream architecture. Here, we conduct ablation experiments to explore the contributions of all four streams, i.e., Landmark sequence, Landmark Motion sequence, Muscle sequence and Muscle Motion sequence. Variants of the proposed LBAGCN are built by removing one of the streams. In all experiments, T is set to 16 and optimized weights are used to fuse the outputs of the parallel AGCNs. Experiment results are given in Table 2.

Based on Table 2, we analyze the impact of each stream on the accuracy. The proposed LBAGCN with four streams achieves the best accuracy of 87.71% and is taken as baseline for comparison. Evidently, when each of the streams, namely Landmark, Muscle, Landmark Motion and Muscle Motion, is removed, the corresponding decrease in accuracy is 3.96%, 3.54%, 2.71% and 2.71%. It tells that all four streams contribute much to the accuracy. Moreover, it is worth noting that, though the Muscle, Landmark Motion and Muscle Motion

TABLE 4. Accuracies achieved with different weight distributions on the OULU CASIA dataset.

Weight				Accuracy (%)
Landmark	Muscle	Landmark Motion	Muscle Motion	
1/4	1/4	1/4	1/4	85.83
3/10	3/10	2/10	2/10	85.83
2/6	**2/6**	**1/6**	**1/6**	**87.71**
5/14	5/14	2/14	2/14	86.67
3/10	2/10	3/10	2/10	87.29
2/6	1/6	2/6	1/6	86.67
5/14	2/14	5/14	2/14	87.08



(a) Initial topology (b) AGC layer 1 (c) AGC layer 2

(d) AGC layer 4 (e) AGC layer 5 (f) AGC layer 6

(g) AGC layer 8 (h) AGC layer 9 (i) AGC layer 10

FIGURE 5. **(a) The initial landmark graph topology. (b)-(i) The landmark graph topologies at different AGC layers, i.e., $\tilde{A}^i$, of AGCN learned from the Oulu CASIA dataset.**

streams are derived from the raw Landmark stream, their impacts on the accuracy are considerably significant. This observation confirms that these derivatives do provide useful information to FER tasks.

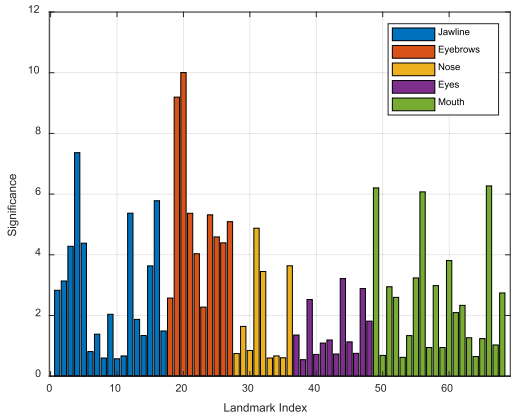
3) ON UNIFORM SAMPLING PARAMETER T

The configurable parameter T plays an important role in our proposed scheme. Both the Uniform Sampling and the LBAGCN depends much on the value of T. Here, we investigate the impact of T on the FER accuracy. Models are built by tuning the value of T. In all experiments, optimized weights are used to fuse the outputs of the parallel AGCNs. Experiment results are given in Table 3.

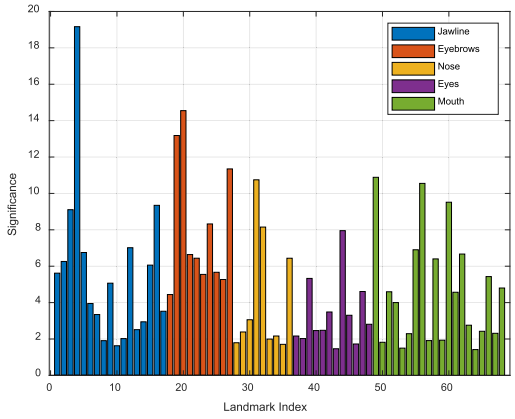
As shown in Table 3, the best accuracy of 87.71% is achieved when T is set to 16. Considerable decreases in accuracy can be observed when T take values of 8, 32 and 64. Intuitively, the reason lies in two folds. For video clips with very long durations, when T is too small, the down-sampled sequences may lose too much of the temporal dynamics of the facial expressions. For video clips with very short durations, when T is too large, the REPEAT operations involved in the Uniform Sampling may introduce crosstalk and thus break the natural dynamics of facial expressions.

4) ON MULTI-STREAM FUSION WEIGHTS

In the proposed LBAGCN, the final class score is obtained by computing a weighted sum of the outputs of the four AGCNs. Setting T to 16, we explore the effect of weight distribution on FER accuracy. Models are built by tuning the weight distribution among AGCNs empirically. Experiment results are given in Table 4.



(a) AGC layer 6



(b) AGC layer 10

FIGURE 6. Significance of landmarks at AGC layer 6 & AGC layer 10. There are totally 68 landmarks belonging to five salient regions of the face, namely Jawline, Nose, Eyebrows, Eyes and Mouth. The salient regions of the face are color-coded for distinction.

As shown in Table 4, the best accuracy of 87.71% is achieved when the weights of Landmark, Muscle, Landmark Motion and Muscle Motion are set to 2/6, 2/6, 1/6 and 1/6 respectively. Other settings of weight distribution exhibit observable losses in accuracy. Intuitively, the weight distribution among Landmark, Muscle, Landmark Motion and Muscle Motion should be determined properly with respect to their contributions to accuracy. From this point, the results in Table 4 is reasonable and compatible with that in Table 2.

E. VISUALIZATION OF LEARNED TOPOLOGIES

According to equation (13), graph topology, i.e., $\tilde{\mathbf{A}}$, plays a key role in AGCN. Taking the Landmark stream as an example, we illustrate the initial graph topology and the graph topologies of different AGC layers of the AGCN learned from the Oulu CASIA dataset. The visualization results are given in Fig. 5. Note that the values close to 0 indicate weak relationships between nodes and vice versa.

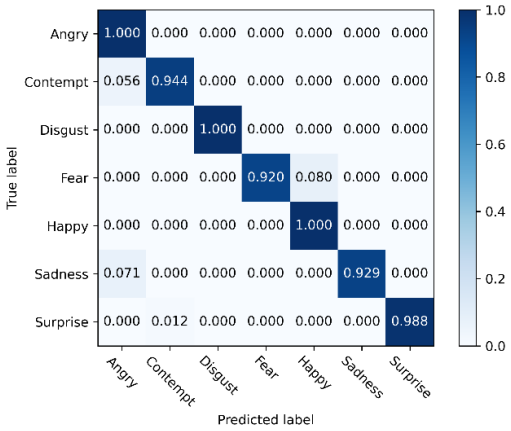
To better understand the AGCN, we interpret the learned topologies in Fig. 5 carefully. As can be observed: (1) The self-loops are preserved well in all the learned topologies of different AGC layers. It indicates that self-loops are useful in preserving identity features during graph convolution. (2) At the first few AGC layers, e.g., AGC layer 1 and AGC layer 2, most of the rows and columns consist of high connection strengths in the learned graph topologies. It implies that, at these layers, each node interacts with as many neighbors as possible to expand the spatial dimension of the feature. (3) At intermediate AGC layers, e.g., AGC layer 4, AGC layer 5 and AGC layer 6, there are a small portion of columns consisting of high connection strengths in the learned graph topologies. It means that the corresponding nodes are assigned higher weights than others by most of the nodes when performing spatial feature aggregation. (4) At the final AGC layers, e.g., AGC layer 9 and AGC layer 10, a small portion of rows consist of high connection strengths in the learned graph topologies. It means that the corresponding nodes assign higher weights to most of their neighbors than others do to gain more discriminative features for the following classifier.

As shown in Fig. 3, the 68 facial landmarks distribute in salient regions including Jawline, Eyebrows, Eyes, Nose and Mouth. The authors of [10], [11] and [12] selected some facial landmarks belonging to Eyebrows, Eyes, Nose and Mouth empirically to construct their facial graphs, and neglected the ones belonging to Jawline. In contrast, we construct facial graphs with learnable topologies using all 68 facial landmarks in LBAGCN. Here, to further understand the learned topologies, we analyze the significance of landmarks (nodes) at typical AGC layers, e.g., AGC layer 6 and AGC layer 10. For the learned topology at AGC layer 6, i.e., $\tilde{\mathbf{A}}^6$, we define the significance of a landmark (node) as the sum of the absolute values of elements of the corresponding column. For the learned topology at AGC layer 10, i.e., $\tilde{\mathbf{A}}^{10}$, we define the significance of a landmark (node) as the sum of the absolute values of elements of the corresponding row. The results are shown in Fig. 6.

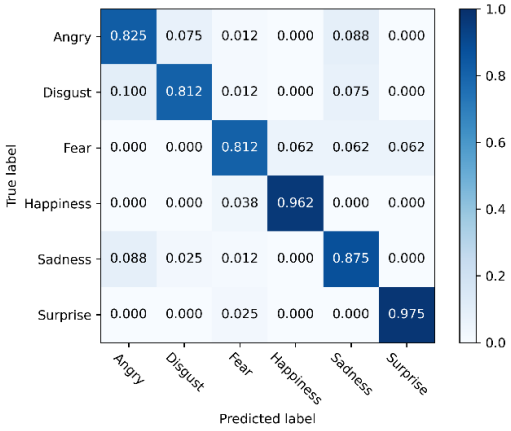
As shown in Fig. 6, in both two AGC layers considered, there exist a few of the landmarks with highest significance in each salient region of the face. It suggests that facial expressions are broadly associated with landmarks in different salient regions of the face. Meanwhile, it can be observed that, in both two AGC layers considered, some landmarks belonging to Jawline can be more significant than those belonging to Nose or Eyes. It reveals that some landmarks belonging to Jawline can be even more valuable for FER than those belonging to Nose or Eyes. In other words, neglecting landmarks belonging to Jawline as done in [10], [11], and [12] may result in loss in terms of FER accuracy. Moreover, if a landmark has high significance in AGC layer 6, it more likely has high

TABLE 5. Parameters, FLOPs, F1 score and Accuracy of the proposed LBAGCN model on CK+ with T = 32 and Oulu CASIA datasets with T = 16.

Dataset	Parameters	FLOPs	F1 Score	Accuracy
CK+	5.48M	4.70G	95.77%	98.17%
Oulu CASIA	5.48M	2.35G	87.67%	87.71%



(a) CK+



(a) Oulu CASIA

FIGURE 7. The confusion matrices of our proposed LBAGCN on CK+ and Oulu CASIA datasets.

significance in AGC layer 10 as well. Intuitively, the facial expressions are closely associated with a relatively small portion of significant landmarks. Thereby, smaller facial graphs can be constructed potentially for resource-limited scenarios by neglecting some landmarks with lowest significance.

F. PERFORMANCE EVALUATION

In this section, we evaluate the proposed LBAGCN on CK+ and Oulu CASIA in terms of number of parameters, FLOPs, F1 score and accuracy. Confusion matrices are provided for misclassification analysis. In addition, the inference times of the models trained on CK+ and Oulu CASIA are tested on typical devices.

1) PERFORMANCE ANALYSIS

The proposed LBAGCN are trained and tested on the CK+ and Oulu CASIA datasets separately. The Uniform Sampling parameter T is set to be 32 and 16 for CK+ and Oulu CASIA respectively. Numerical results are given in Table 5. Confusion matrices on CK+ and Oulu CASIA are shown in Fig. 7.

TABLE 6. Inference time of the proposed LBAGCN model on CK+ and Oulu CASIA.

Device	Inference time	
	CK+	Oulu CASIA
NVIDIA RTX 3080Ti	80.27ms	79.95ms
Desktop computer with Intel(R) Core(TM) i7-7700 @ 3.60GHz and 16GRAM	263.17ms	206.41ms

As can be seen from Table 5, the model trained on CK+ has 5.48M parameters and 4.70GFLOPs, while that trained on Oulu CASIA has 5.48M parameters and 2.35GFLOPs. The F1 score and Accuracy achieved on CK+ are 95.77% and 98.17% respectively. On Oulu CASIA dataset, the F1 score and Accuracy are 87.67% and 87.71% respectively. Evidently, higher F1 score and Accuracy can be achieved on CK+ than on Oulu CASIA. It is because that the facial landmarks provided by CK+ are more precise than those estimated from Oulu CASISA by using the algorithm in [35]. Note that, Oulu CASISA was recorded in more complex illumination conditions (strong, weak and dark) with a comparatively lower resolution of 320 × 240 pixels.

For the case of CK+, one can find from Fig. 7(a) that the model recognizes Angry, Disgust and Happy perfectly. The recognition rates of Surprise, Contempt, Sadness and Fear are 98.80%, 94.40%, 92.90% and 92.00% respectively. Fear and Sadness are most difficult to recognize among all expressions. As can be seen, 8.00% of Fear samples are misclassified into Happy and 7.10% of Sadness samples are misclassified into Angry. Moreover, 5.60% of Contempt samples are misclassified into Angry, while 1.20% of Surprise samples are misclassified into Contempt.

For the case of Oulu CASIA, it can be seen from Fig. 7(b) that high recognition rates of 97.50% and 96.20% can be achieved for Surprise and Happiness. The recognition rates of Sadness, Angry, Disgust and Fear are 87.50%, 82.50%, 81.20% and 81.20% respectively. Disgust, Angry and Sadness are comparatively most difficult to distinguish. As can be seen, 10.00% of Disgust samples and 8.80% of Sadness samples are misclassified into Angry, 8.80% of Angry samples and 7.50% of Disgust samples are misclassed into Sadness, and 7.50% of Angry samples and 2.50% of Sadness samples are misclassified into Disgust. Moreover, Fear is more likely misclassified into Happiness, Sadness and Surprise. As observed, Fear samples are misclassified into Happiness, Sadness and Surprise with the same percentage of 6.20%.

2) INFERENCE TIME

Experiments are conducted to test the inference times of models trained on CK+ and Oulu CASIA. An INVIDIA GeForce RTX 3080Ti and a desktop computer with Intel(R) Core(TM) i7-7700 @ 3.60GHz and 16G RAM are employed. The experiment results are given in Table 6.

TABLE 7. Comparison with state-of-the-art methods.

Method	Architecture	Input	Graph topology	Parameters	FLOPs CK+ / Oulu CASIA	Accuracy (%) CK+	Accuracy (%) Oulu CASIA
DTAN [29]	CNN	Image	--	8.55M	0.13G	91.44	74.38
MSCNN [30]		Image	--	5.25M	0.11G	95.54	77.67
ADModel [31]		Image	--	90.01M	20.46G	94.74	79.17
FN2EN [32]		Image	--	10.60M	3.34G	98.60	**87.71**
PPDN [33]		Image	--	6M	4.95G	97.30	84.59
DTGN [29]	MLP	Landmark	--	0.18M	0.18M	92.35	74.17
PHRNN [30]	BRNN	Landmark	--	8.16M	0.16G	96.36	78.96
ASModel [31]	BiLSTM	Landmark	--	**0.03M**	**0.03M**	89.47	70.83
DTAGN [29]	CNN + MLP	Image + Landmark	--	8.74M	0.13G	97.25	81.46
PHRNN-MSCNN [30]	CNN + BRNN	Image + Landmark	--	13.41M	0.27G	98.50	86.25
MSDModel [31]	CNN + BiLSTM	Image + Landmark	--	90.03M	20.46G	**99.10**	87.33
STSGN-G [10]	GCN	Landmark	Predefined	0.08M	0.04G	90.24	56.82
STSGN-A [10]		HOG	Predefined	0.25M	0.13G	95.71	64.14
STSGN [10]		Landmark + HOG	Predefined	0.33M	0.18G	98.63	87.23
Spectral Convolution [12]		ROI	Predefined	0.11M	1.37M	92.86	67.64
DDRGCN [12]		ROI	Partially Learnable	0.12M	0.78M	94.32	73.28
Our LBAGCN		**Landmark**	**Learnable**	5.48M	4.70G / 2.35G	98.17	**87.71**

As shown in Table 6, the inference times of the models trained on CK+ and Oulu CASIA are 80.27ms and 79.95ms respectively when INVIDIA GeForce RTX 3080Ti is employed. Evidently, more than 12.45 times of inference can be supported by an INVIDIA GeForce RTX 3080Ti per second. When the desktop computer is employed, the inference times of the models trained on CK+ and Oulu CASIA are 263.17ms and 206.41ms respectively. It means that more than 3.79 times of inference can be supported by the desktop computer per second. These observations suggest that the proposed LBAGCN can be a feasible solution for real-time FER applications.

G. COMPARISON WITH STATE-OF-THE-ART METHODS

In this section, we compare the proposed LBAGCN with state-of-the-art methods [10], [12], [29], [30], [31], [32], [33] on CK+ and Oulu CASIA in terms of parameters, computational cost (FLOPs) and accuracy. Numerical results are given in Table 7. For clarity, the architecture, input and graph topology (if applicable) of all methods are provided also.

As can be observed from Table 7, all methods considered perform much better in accuracy on the CK+ dataset than on the Oulu CASIA dataset. The reason lies in two folds: (1) The image resolution of CK+ (640 × 480 pixels) is higher than that of Oulu CASIA (320 × 240 pixels). (2) The precision of facial Landmarks provided by CK+ can be higher than that estimated by the landmark detection algorithm from Oulu CASIA. The advantages of CK+ are beneficial for methods in [29], [30], [31], [32], and [33] to extract better spatial-temporal features and helpful for GCN-based ones [10], [12] to gain better HOGs and ROIs.

Firstly, we compare the proposed LBAGCN with CNN-based methods including DTAN [29], MSCNN [30], ADModel [31], FN2EN [32] and PPDN [33]. This group of baseline methods take facial images as input and thus have the risk of leaking sensitive privacy information. Among them, FN2EN [32] achieves the best accuracies of 98.60% on CK+ and 87.71% on Oulu CASIA with 10.6M parameters and 3.34GFLOPs. Compared with FN2EN [32], the proposed LBAGCN achieves the same accuracy of 88.71% on Oulu CASIA with less parameters (5.48M) and lower computational cost (2.35GFLOPs) as shown in Table 7. On CK+, the proposed LBAGCN shows a gap of 0.43% in accuracy and an increase of 1.36GFLOPs in computational cost compared with FN2EN [32]. Moreover, the proposed LBAGCN outperforms the methods other than FN2EN [32] in accuracy on both CK+ and Oulu CASIA. For instance, though the proposed LBAGCN has much less parameters and lower FLOPs than ADModel [31], it shows accuracy advantages of 3.43% on CK+ and 8.54% on Oulu CASIA. Compared with the lightweight MSCNN [30], which has 5.25M parameters and 0.11GFLOPs, the proposed LBAGCN shows accuracy advantages of 2.63% on CK+ and 10.04% on Oulu CASIA. These observations suggest that the proposed LBAGCN outperforms most of the CNN-based methods considered in terms of accuracy on both CK+ and Oulu CASIA.

Secondly, we compare the proposed LBAGCN with DTGN [29], PHRNN [30] and ASModel [31], which adopt architectures of Multilayer Perception (MLP), Bidirectional RNN (BRNN) and Bidirectional Long Short-Term Memory (BiLSTM) respectively. This group of baseline methods take landmarks as input similarly to LBAGCN and are conductive to the preserving of sensitive privacy information. Among them, PHRNN [30] achieves the best accuracies of 96.36% and 78.96% on CK+ and Oulu CASIA respectively. Compared with PHRNN [30], the proposed LBAGCN has accuracy advantages of 1.82% and 8.75% on CK+ and

Oulu CASIA respectively as shown in Table 7. The proposed LBAGCN has less parameters but higher FLOPs than PHRNN [30]. In comparison with the lightweight DTGN [29] and ASModel [31], the proposed LBAGCN has more parameters and higher FLOPs. Nevertheless, the proposed LBAGCN achieves much higher accuracies on CK+ and Oulu CASIA than DTGN [29] and ASModel [31]. For instance, the accuracy of the proposed LBAGCN is 13.54% higher than that of DTGN [29] and 16.88% higher than that of ASModel [31] on Oulu CASIA. These observations confirm that the proposed LBAGCN outperforms its landmark-based counterparts in terms of accuracy on both CK+ and Oulu CASIA.

Thirdly, we compare the proposed LBAGCN with DTAGN [29], PHRNN-MSCNN [30] and MSDModel [31], which adopt hybrid architectures. This group of baseline methods take both facial images and landmarks as input and thus pose a potential risk of exposing sensitive privacy information. In principle, these methods fuse features extracted from facial images and landmarks to improve the FER accuracy. As shown in Table 7, the proposed LBAGCN outperforms all three baseline methods in accuracy on Oulu CASIA with less parameters. Specifically, the accuracy of the proposed LBAGCN is 6.25%, 1.46% and 0.38% higher than that of DTAGN [29], PHRNN-MSCNN [30] and MSDModel [31] respectively. On CK+, the accuracy of the proposed LBAGCN is higher than that of DTAGN [29] but lower than that of MSDModel [31] and PHRNN-MSCNN [30]. Note that all three baseline methods benefit much from the high-resolution facial images of CK+. It is also observed that the proposed LBAGCN has lower FLOPs than MSDModel [31] but higher FLOPs than DTAGN [29] and PHRNN-MSCNN [30]. These observations reveal that the proposed LBAGCN performs better than the baseline methods in accuracy on the more challenging Oulu CASIA.

Finally, we compare the proposed LBAGCN with its GCN-based counterparts including STSGN-G [10], STSGN-A [10], STSGN [10], Spectral Convolution [12] and DDRGCN [12]. Among this group of baseline methods, STSGN [10] achieves the best accuracies of 98.63% and 87.23% on CK+ and Oulu CASIA respectively. Compared with STSGN [10], the proposed LBAGCN shows a disadvantage of 0.46% in accuracy on CK+ but an advantage of 0.48% on the more challenging Oulu CASIA as shown in Table 7. Note that STSGN [10] benefits from the high-resolution facial images of CK+. Moreover, the proposed LBAGCN achieves significantly higher accuracy than the GCN-based methods other than STSGN [10] especially on Oulu CASIA. As can be seen, the accuracy of the proposed LBAGCN on Oulu CASIA is 30.89%, 23.57%, 20.07% and 14.43% higher than that of STSGN-G [10], STSGN-A [10], Spectral Convolution [12] and DDRGCN [12] respectively. It is also worth noting that the proposed LBAGCN has more parameters and higher FLOPs than its GCN-based counterparts. The reason lies in two folds. On the one hand, the facial graphs of LBAGCN are constructed by using 68 facial landmarks and thus have larger scales than those of its counterparts. Note that the facial graph in [10] is constructed by using 34 landmarks and those in [12] are constructed by using 20 ROIs. On the other hand, the proposed LABGCN has more GCN layers than its counterparts. In fact, the proposed LABGCN has 10 GCN layers, while the STSGN [10] has 3 GCN layers and the DDRGCN [12] has 2 GCN layers.

Based on the above comparisons, several points can be drawn:

(1) The proposed LBAGCN achieves higher accuracy than all the compared methods except for FN2EN [32], PHRNN-MSCNN [30], MSDModel [31] and STSGN [10] on CK+ and achieves the highest accuracy among all methods considered on Oulu CASIA. Note that FN2EN [32], PHRNN-MSCNN [30], MSDModel [31] and STSGN [10] depend much on facial images besides facial landmarks while the proposed LBAGCN takes facial landmarks as input only. It confirms that there do exist rich semantic features in the available facial landmarks and discriminative features can be extracted from facial landmarks effectively by using the proposed LBAGCN.

(2) Among facial landmark-based methods, the proposed LBAGCN achieves significantly higher accuracies than DTGN [29], PHRNN [30], ASModel [31] and STSGN-G [10] on both CK+ and Oulu KASIA. Note that DTGN [29], PHRNN [30], ASModel [31] and STSGN-G [10] employ a few landmarks selected empirically from Eyebrows, Eyes, Nose and Mouth while the proposed LBAGCN make full use of all 68 landmarks. It reveals that semantic features lying in available landmarks can be exploited by the proposed LBAGCN more effectively than its counterparts.

(3) Compared with lightweight methods such as ASModel [31], STSGN-G [10] and DDRGCN [12], the proposed LBAGCN has more parameters and higher FLOPs. It means that the proposed LBAGCN needs to pay some price in terms of memory footprint and computational cost. In practice, it may be worth paying such a price if FER accuracy and preserving of sensitive privacy information are critical factors. By the way, removing some streams from LBAGCN properly is a feasible way to reduce parameters and FLOPs effectively. However, observable loss in accuracy will be paid inevitably. For instance, removing the muscle motion stream from LBAGCN saves 25% of parameters and FLOPs approximately and meanwhile results in an accuracy loss of 2.71% as shown in Table 2.

V. CONCLUSION

In this paper, we propose a novel GCN-based method, named LBAGCN, for FER. Given a facial landmark sequence extracted from a video clip, four complementary facial graphs with learnable topologies are constructed and corresponding AGCNs are designed to extract semantic features from the facial graphs. The class scores of AGCNs are fused by using weighted summation to boost the FER accuracy. Experiment results suggest that the proposed LBAGCN can be used to achieve state-of-the-art FER accuracy especially

on the challenging Oulu CASIA dataset. The LBAGCN takes facial landmarks as input in both phases of model training and inference and thus is conducive to the preserving of sensitive privacy information. Testing results on practical platforms reveal that the proposed LBAGCN can be a feasible solution for real-time FER applications. Although the proposed LBAGCN achieves significantly higher FER accuracy than the existing lightweight FER models, it is more demanding in terms of memory footprint and computational power. In our future study, efforts will be made to reduce the parameters and computational cost of the LBAGCN further.

REFERENCES

[1] W. Li, G. Zeng, J. Zhang, Y. Xu, Y. Xing, R. Zhou, G. Guo, Y. Shen, D. Cao, and F.-Y. Wang, "CogEmoNet: A cognitive-feature-augmented driver emotion recognition model for smart cockpit," *IEEE Trans. Computat. Social Syst.*, vol. 9, no. 3, pp. 667–678, Jun. 2022.

[2] Y. Cimtay, E. Ekmekcioglu, and S. Caglar-Ozhan, "Cross-subject multimodal emotion recognition based on hybrid fusion," *IEEE Access*, vol. 8, pp. 168865–168878, 2020.

[3] L. Chen, M. Wu, M. Zhou, J. She, F. Dong, and K. Hirota, "Information-driven multirobot behavior adaptation to emotional intention in human–robot interaction," *IEEE Trans. Cognit. Develop. Syst.*, vol. 10, no. 3, pp. 647–658, Sep. 2018.

[4] R. Hortensius, F. Hekele, and E. S. Cross, "The perception of emotion in artificial agents," *IEEE Trans. Cognit. Develop. Syst.*, vol. 10, no. 4, pp. 852–864, Dec. 2018.

[5] Y. Wang, H. Yu, B. Stevens, and H. Liu, "Dynamic facial expression recognition using local patch and LBP-TOP," in *Proc. 8th Int. Conf. Human Syst. Interact. (HSI)*, Warsaw, Poland, Jun. 2015, pp. 362–367.

[6] A. Klaeser, M. Marszalek, and C. Schmid, "A spatio-temporal descriptor based on 3D-gradients," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, Leeds, U.K., 2008, pp. 1–11.

[7] B. Hasani and M. H. Mahoor, "Facial expression recognition using enhanced deep 3D convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Honolulu, HI, USA, Jul. 2017, pp. 2278–2288.

[8] E. L. Rosenberg and P. Ekman, *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*. London, U.K.: Oxford Univ. Press, 2020.

[9] D. L. Bimler and G. V. Paramei, "Facial-expression affective attributes and their configural correlates: Components and categories," *Spanish J. Psychol.*, vol. 9, no. 1, pp. 19–31, May 2006.

[10] J. Zhou, X. Zhang, Y. Liu, and X. Lan, "Facial expression recognition using spatial–temporal semantic graph network," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Abu Dhabi, UAE, Oct. 2020, pp. 1961–1965.

[11] L. Liao, Y. Zhu, B. Zheng, X. Jiang, and J. Lin, "FERGCN: Facial expression recognition based on graph convolution network," *Mach. Vis. Appl.*, vol. 33, no. 3, p. 40, Mar. 2022.

[12] X. Jin, Z. Lai, and Z. Jin, "Learning dynamic relationships for facial expression recognition based on graph convolutional network," *IEEE Trans. Image Process.*, vol. 30, pp. 7143–7155, 2021.

[13] K. Chaudhuri, C. Monteleoni, and A. D. Sarwate, "Differentially private empirical risk minimization," *J. Mach. Learn. Res.*, vol. 12, no. 3, pp. 1069–1109, 2011.

[14] M. Jegorova, C. Kaul, C. Mayor, A. Q. O'Neil, A. Weir, R. Murray-Smith, and S. A. Tsaftaris, "Survey: Leakage and privacy at inference time," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 7, pp. 9090–9108, Jul. 2023.

[15] A. Umezawa, Y. Takegawa, K. Suzuki, K. Masai, Y. Sugiura, M. Sugimoto, Y. Tokuda, D. M. Plasencia, S. Subramanian, M. Takahashi, H. Taka, and K. Hirata, "E2-MaskZ: A mask-type display with facial expression identification using embedded photo reflective sensors," in *Proc. Augmented Hum. Int. Conf.*, vol. 33, Mar. 2020, pp. 1–3.

[16] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.-Workshops*, San Francisco, CA, USA, Jun. 2010, pp. 94–101.

[17] G. Zhao, X. Huang, M. Taini, S. Z. Li, and M. Pietikäinen, "Facial expression recognition from near-infrared videos," *Image Vis. Comput.*, vol. 29, no. 9, pp. 607–619, Aug. 2011.

[18] M. Defferrard, X. Bresson, and P. Vandergheynst, "Convolutional neural networks on graphs with fast localized spectral filtering," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 3844–3852.

[19] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," 2016, *arXiv:1609.02907*.

[20] J. Bruna, W. Zaremba, A. Szlam, and Y. LeCun, "Spectral networks and locally connected networks on graphs," 2013, *arXiv:1312.6203*.

[21] W. L. Hamilton, R. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1025–1035.

[22] D. K. Hammond, P. Vandergheynst, and R. Gribonval, "Wavelets on graphs via spectral graph theory," *Appl. Comput. Harmon. Anal.*, vol. 30, no. 2, pp. 129–150, Mar. 2011.

[23] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, "How powerful are graph neural networks?" 2018, *arXiv:1810.00826*.

[24] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proc. AAAI Conf. Artif. Intell.*, New Orleans, LA, USA, 2018, pp. 7444–7452.

[25] B. Yu, H. Yin, and Z. Zhu, "Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting," 2017, *arXiv:1709.04875*.

[26] X. Wang and A. Gupta, "Videos as space-time region graphs," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, vol. 2018, Munich, Germany, 2018, pp. 413–431.

[27] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *Image Vis. Comput.*, vol. 27, no. 6, pp. 803–816, May 2009.

[28] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 915–928, Jun. 2007.

[29] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, "Joint fine-tuning in deep neural networks for facial expression recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Santiago, Chile, Dec. 2015, pp. 2983–2991.

[30] K. Zhang, Y. Huang, Y. Du, and L. Wang, "Facial expression recognition based on deep evolutional spatial–temporal networks," *IEEE Trans. Image Process.*, vol. 26, no. 9, pp. 4193–4203, Sep. 2017.

[31] X. Sun, P. Xia, and F. Ren, "Multi-attention based deep neural network with hybrid features for dynamic sequential facial expression recognition," *Neurocomputing*, vol. 444, pp. 378–389, Jul. 2021.

[32] H. Ding, S. K. Zhou, and R. Chellappa, "FaceNet2ExpNet: Regularizing a deep face recognition net for expression recognition," in *Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Washington, DC, USA, May 2017, pp. 118–126.

[33] X. Zhao, X. Liang, L. Liu, T. Li, Y. Han, N. Vasconcelos, and S. Yan, "Peak-piloted deep network for facial expression recognition," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Amsterdam, The Netherlands. Cham, Switzerland: Springer, Oct. 2016, pp. 425–442.

[34] H. Duan, Y. Zhao, K. Chen, D. Lin, and B. Dai, "Revisiting skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 2959–2968.

[35] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, Jun. 2014, pp. 1867–1874.

[36] Z. Liu, H. Zhang, Z. Chen, Z. Wang, and W. Ouyang, "Disentangling and unifying graph convolutions for skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 140–149.

[37] Y. Chen, Z. Zhang, C. Yuan, B. Li, Y. Deng, and W. Hu, "Channel-wise topology refinement graph convolution for skeleton-based action recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Montreal, QC, Canada, Oct. 2021, pp. 13339–13348.

[38] L. Zhong, Q. Liu, P. Yang, B. Liu, J. Huang, and D. N. Metaxas, "Learning active facial patches for expression analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Providence, RI, USA, Jun. 2012, pp. 2562–2569.

DAQI ZHAO received the B.Sc. degree in communication engineering from Shandong University, Jinan, China, in 2021, where he is currently pursuing the M.Sc. degree with the School of Information Science and Engineering. His research interests include deep learning, facial expression recognition, and graph convolutional networks.

JINGWEN WANG received the B.Sc. degree in electronic information engineering from Shandong University, Jinan, China, in 2021, where she is currently pursuing the M.Sc. degree with the School of Information Science and Engineering. Her research interests include deep learning, neural networks, and behavior detection.

HAOMING LI received the B.Sc. degree in electronic information engineering from Shandong University, Jinan, China, in 2021, where she is currently pursuing the M.Sc. degree with the School of Information Science and Engineering. Her research interests include deep learning, neural networks, and speech emotion recognition.

DEQIANG WANG (Senior Member, IEEE) received the B.S. degree in radio technology and the M.S. degree in signal processing from Shandong University, Jinan, China, in 1990 and 1995, respectively, and the Ph.D. degree in communication and information systems from Beijing University of Posts and Telecommunications, China, in 2005. Since 1995, he has been a Faculty Member with the School of Information Science and Engineering, Shandong University, where he is currently working as a Full Professor. His research interests include deep learning, compressed sensing, wireless communications, and adaptive signal processing.