Geometric Graph Representation With Learnable Graph Structure and Adaptive AU Constraint for Micro-Expression Recognition

Jinsheng Wei, Wei Peng, Guanming Lu, Yante Li, Jingjie Yan, and Guoying Zhao, Fellow, IEEE

Abstract—Micro-expression recognition (MER) holds significance in uncovering hidden emotions. Most works take image sequences as input and cannot effectively explore ME information because subtle ME-related motions are easily submerged in unrelated information. Instead, the facial landmark is a low-dimensional and compact modality, which achieves lower computational cost and potentially concentrates on ME-related movement features. However, the discriminability of facial landmarks for MER is unclear. Thus, this article investigates the contribution of facial landmarks and proposes a novel framework to efficiently recognize MEs with facial landmarks. First, a geometric two-stream graph network is constructed to aggregate the low-order and high-order geometric movement information from facial landmarks to obtain discriminative ME representation. Second, a self-learning fashion is introduced to automatically model the dynamic relationship between landmarks, facial action units and MEs. Furthermore, an adaptive action unit loss is proposed to reasonably build a strong correlation with landmarks, facial action units and MEs. Notably, this work provides a novel idea with much higher efficiency to promote MER, only utilizing graph-based geometric features. The experimental results demonstrate that the proposed method achieves competitive performance with a significantly reduced computational cost. Furthermore, facial landmarks significantly contribute to MER and are worth further study for high-efficient ME analysis.

Index Terms—Micro-expression recognition, facial landmarks, graph network, action units.

Manuscript received 7 November 2022; revised 24 November 2023; accepted 1 December 2023. Date of publication 6 December 2023; date of current version 2 September 2024. This work was partly supported in part by the National Association of public funds Natural Science Research Start up Foundation of Recruiting Talents of Nanjing University of Posts and Telecommunications under Grant NY223030, in part by the Nanjing Science and Technology Innovation Foundation for Overseas Students under Grant NJKCYZZ2023-04, in part by the National Natural Science Foundation of China (NSFC) under Grants 72074038, 61971236, and 62076122, in part by China Scholarship Council under Grant 202008320445, in part by the Research Council of Finland (former Academy of Finland) Academy Professor project EmotionAI under Grants 336116 and 345122, in part by ICT 2023 project TrustFace under Grant 345948, and in part by the University of Oulu & Research Council of Finland Profi 7 under Grant 352788. Recommended for acceptance by R. Poppe. (Corresponding author: Guanming Lu.)

Jinsheng Wei, Guanming Lu, and Jingjie Yan are with the Nanjing University of Posts and Telecommunications (NUPt), Nanjing 210003, China (e-mail: 20220182@njupt.edu.cn; lugm@njupt.edu.cn; yangjingjie@njupt.edu.cn).

Wei Peng is with the Department of Psychiatry and Behavioral Sciences of Stanford Medicine, Stanford University, Stanford, CA 94305 USA (e-mail: wepeng@stanford.edu).

Yante Li and Guoying Zhao are with the Center for Machine Vision and Signal Analysis, University of Oulu, 90014 Oulu, Finland (e-mail: yante.li@oulu.fi; guoying.zhao@oulu.fi).

Digital Object Identifier 10.1109/TAFFC.2023.3340016
of geometric features from facial landmarks to MER. Thus, in this paper, we make the first step to study the discriminability of facial landmarks for MER and employ only facial landmarks as input, for exploring its advantages compared to image-based input. Graph-based methods, as demonstrated in previous studies [22], have proven effective in handling facial landmarks within the non-Euclidean space. Lately, some researchers [8], [21], [23], [24] employed graph-based models to recognize MEs. However, these works adopted a fixed graph structure designed manually according to certain principles (e.g., the conditional probability of different nodes [23], [25]). The fixed adjacency matrix is defined to represent the fixed relationship between nodes. The manual way to set a fixed graph structure cannot comprehensively and effectively model the correlation between landmark-based nodes, which is sub-optimal. To overcome this problem, this paper introduces a learnable adjacency matrix to learn a more reasonable and flexible graph structure based on a pre-defined graph structure.

Besides facial landmarks, Action Unit (AU) [26] encodes the local movement of facial muscles and has been verified to have a strong correlation with MEs. Recently, some works [16], [24] took AU features as input to recognize MEs and used deep learning models to build the relationship between AU features and ME categories. However, these methods need an auxiliary model to extract AU features before inputting AU features into the main model, which increases the model complexity and computational cost. To address this limitation, our research employs the AU labels as a constraint to learn the geometric features related to AUs from facial landmarks. Furthermore, considering that the AU losses in different semantic levels may have different contributions, this paper proposes an Adaptive AU (AAU) loss to automatically learn the constraint intensity of AU loss in different semantic levels. In this way, the multi-scale geometric features are constrained adaptively to have a synchronized pattern with AUs, thereby rationalizing the introduction of facial AU information.

Corresponding to the above problems, the main contributions of this paper can be summarized as follows:

1) This work studies and explores the contribution of facial landmarks for MER. The graph-based models are leveraged to extract discriminative geometric movement features with spatial-temporal information only from facial landmarks. The results demonstrate that facial landmarks are discriminative for MER with a largely reduced computational cost.

2) To comprehensively explore geometric movement features from facial landmark, a Geometric Two-Streams Graph Network (GTS-GN) is proposed to aggregate the low-order and higher-order geometric information from facial landmarks.

3) To overcome the shortcoming of the fixed adjacency matrix, this paper proposes a Learnable Adjacency Matrix (LAM) to learn a reasonable and flexible graph structure. As a result, LAM can automatically model the discriminative relationship between facial landmarks for MER.

4) Based on the strong correlations between AUs and MEs, an Adaptive AU (AAU) loss is proposed to automatically explore a more reasonable and efficient way to introduce AU information. AAU loss can adaptively constrain the geometric features in a multi-scale fashion and emphasize the contributions of AU information at different semantic levels. The experimental results demonstrate that jointing the regular ME loss and AAU loss can build the relationship between facial landmarks, AUs and MEs, improving the performance of MER.

II. RELATED WORK

With the research of MER, increasing researchers have focused on automatic ME analysis, expecting to explore its value and potential in both academic research and commercial applications. Based on traditional machine learning methods, early works [27], [28], [29], [30] classified MEs using hand-crafted features. These hand-crafted features lack adaptability and require effort to design. Later, deep learning methods were well applied in image processing fields, and then some works [31], [32] in MER began to explore deep learning methods and achieved promising performance. In addition to classical deep learning models such as Convolutional Neural Network (CNN), recent works [8], [16], [21], [23], [24], [25] have explored the use of graph-based models for recognizing MEs. The experimental results have proved the effectiveness of the graph-based methods in MER. Next, the related works on input data and graph-based models are presented.

A. Input Data

So far, many works [7], [27], [29], [32] extracted discriminative features from the whole facial RGB images/videos or the feature maps based on these images/videos. However, these extracted features include a lot of redundant information because the discriminative features only exist in a small area of the face. The model consumes excessive energy and computational costs in areas that are not related to ME. Zong et al. [12] proposed the Kernelized Group Sparse Learning (KGSL) model to select effective regions, improving the recognition rate. However, it still requires the extraction of several hierarchical features from the entire face image, thereby increasing computational cost. To overcome the above problem, Polikovsky et al. [30] divided the Regions of Interest (ROIs) from the entire facial region based on facial landmarks. Then the features from these ROIs are input into a classifier to identify ME categories. Some studies [30], [33] showed the effectiveness of using features extracted solely from ROIs, with reduced computational cost. Furthermore, Liu et al. [28] proposed the main direction of optical flow to represent the primary movement direction within ROIs. The more compact features are input to SVM classifier. Utilizing local regions, instead of the global ones, these works largely reduce the costs. However, these local region-based methods still take up computing resources to extract the features in the local area. The computational cost is still huge for some discriminative features, e.g., Optical Flow. In addition, region cropping is also a not trivial issue.

As a comparison, facial landmarks are low-dimension data, and taking it as the input can reduce computational costs. In
MER, facial landmarks need to be extracted for face alignment, interception, and the division of regions, especially for local region-based methods. Thus, extracting facial landmarks is unavoidable for MER. At the same time, when facial muscles move, facial landmarks will move accordingly. Facial muscle movements related to MEs can be captured by the geometric features in facial landmarks. The movement of facial landmarks can represent the movement information of facial muscles in key regions. Therefore, we argue that geometric information is discriminative enough to get comparable even superior performance. Facial landmarks-based methods may not be necessary to perform complicated and costly pre-processing from the raw RGB inputs. Compared with images/videos, the low-dimensional landmarks are much more compact, as well as include the discriminative geometric feature information [34]. Choi et al. [20] employed facial landmarks as the input to recognize MEs instead of original ME images/videos. Specifically, they proposed a 2D landmark feature map by transforming conventional coordinate-based landmark information into 2D image information. Then, CNN and Long Short Term Memory (LSTM) models are employed to process LFM. The work aggregated geometric feature information in facial landmarks and got promising results. However, the model input still is the 2D images. Recently, Kumar et al. [8] proposed a two-stream Graph Convolutional Network (GCN) [35] model to process the optical flow and facial landmark information. Wei et al. [36] constructed Decomposition and Reconstruction-based Graph Representation Learning (DeRe-GRL) model to handle facial landmarks in an interpretable manner. The above works introduce the facial landmark information into the model. However, these works do not focus on facial landmarks or explore their discriminability and efficiency. Thus, this paper focuses on the effectiveness of facial landmarks in terms of both accuracy and efficiency. Furthermore, we explore more effective modules and components to aggregate spatial and temporal information in facial landmarks. Since the landmarks are structure data lying in the non-Euclidean space, a new graph model is designed to learn the geometric feature representations of MEs from the facial landmarks.

B. Graph Models

Recently, graph-based methods have become popular in MER and generally don’t input the whole facial images into models. According to the way to construct the graph, these methods can be divided into two types: landmark-based and AU-based.

Lei et al. [21] first constructed a landmarks-based graph. The magnified shape features around facial landmarks were taken as node features. However, they ignored the special relationships of nodes and only employed Temporal Convolutional Network (TCN) [37] to deal with the node representation. Recently, Kumar et al. [8] took facial landmarks and optical flow features around landmarks as node features, and designed a Graph Attention Convolutional Network (GACN) to process the landmark-based graph. Except for facial landmarks, AU also is suitable for constructing the graph. Lo et al. [25] designed GCN with two layers to deal with an AU-based graph and aggregate AU label information that dots product with spatial-temporal features extracted by 3D convolution. They explored the relationship between AUs and MEs, but the node features are the annotated AU labels, which is not suitable for practical applications. Furthermore, both [16] and [24] took the AU features extracted by an extra model as node features in the AU-based graph. Differently, Zhou et al. [16] connected an inception block and an AU detection module to extract AU features, while Xie et al. [24] employed 3D ConvNet to extract AU features. Lei et al. [23] introduce transformer and AU features to extend their previous work [21] by combining two types of graph. However, on one hand, the existing works with the landmark-based graph aggregated dynamic information, e.g., optical flow and magnified shape feature, which still spends too much computational cost to extract these features. On the other hand, the works with the AU-based graph need the extra learning model to extract AU features as node features, which increases the model complexity and computational cost. Also, AU detection for MER is a very challenging task. Inspired by skeleton-based action recognition works, e.g., [38], instead of providing much complicated and costly appearance features or not easily accessible AU features, we provide a simple and much more efficient way that directly takes landmark coordinate-based geometric features as node features. A new graph network is designed to handle such geometric features from facial landmarks for MER.

Furthermore, the proposed graph-based method involves the relationship between nodes and the AU information. The related works are outlined below:

The Relationship Between Nodes: The relationship between nodes is key to aggregating the feature information in the graph. For a graph, the adjacency matrix can express the node relationship. Although Lei et al. [21] defined a graph, they didn’t give the adjacency matrix in the graph. Namely, the special relationship between nodes is not considered. Most of the above graph-based methods built the relationship between nodes according to some principles. Specifically, based on the human facial structure, Kumar et al. [8] built the relationship between landmark nodes. [24] first connected AU nodes based on the AU correlations of the objective class definition and then adjusted it referring to the training datasets. [16] achieved it by the training set based on common facial expressions and facial anatomy analysis. [23] and [25] defined the adjacency matrix in a data-driven way, and both use the conditional probability of different nodes. Different from these methods, we first pre-defined a fixed adjacency matrix based on facial muscle structure. Then, we define a LAM that is added to the fixed adjacency matrix. LAM can be updated automatically to learn a more reasonable relationship between nodes.

AU Information: As we know, AUs have a close relationship with facial expressions, including MEs. At present, AU recognition received widespread attention, and the researchers in MER have introduced AU information to recognize MEs. In [25] and [23], AUs are fed into the GCN model to learn the AU representations. The learned AU representations are fused with ME representations extracted by other models. However, the two methods need to compute the AU vector in advance. [24] and [16] introduced AU loss to extract AU representations from
original video data or optical flow maps. Then, the learned representations are aggregated by GCN for final classification. These methods need extra models to process AUs or learn AU representations, which increases the model complexity and computational cost. Unlike these works, the proposed AAU loss aggregates AU information without extra models. AAU can constrain multi-scale features in an adaptive way to reasonably construct a strong relationship between facial landmarks, AUs and MEs.

III. THE PROPOSED METHOD

In this section, the proposed facial landmark-based method is introduced in detail. We present the framework of the proposed graph model as shown in Fig. 1. Specifically, this work constructs a geometric movement graph (GM-Graph), designs SS module to deal with GM-Graph, and builds the GTS-GN model. In addition, two key components (LAM and AAU loss) are proposed to improve the module and model. The details are as follows.

A. Geometric Movement Graph

In fact, the extraction of movement features is crucial to MER tasks. For a ME video, the onset, apex and offset frames represent the key process of muscle movement related to MEs. Thus, these three frames contain abundant movement information and remove a lot of redundant information in the entire ME video. Accordingly, the facial landmarks of the onset, apex and offset frames include the discriminative geometric movement information. Thus, in this paper, the landmarks of the onset, apex and offset frames in ME videos are taken as graph nodes to capture ME-related movements as shown in Fig. 2. Furthermore, inspired by the Facial Action Coding System (FACS) [26], we choose the landmarks around the mouth, eyebrows and nose to construct the graph based on their contributions. The landmarks in the eye region are not considered here. The reasons are as follows: 1) some noises in the eye regions exist, e.g., blinking. 2) the changes in the eye regions about MEs can lead to the corresponding changes in eyebrows, e.g., dilation of pupils in a surprise expression. So, the movements in the eye region have a certain amount of noise and redundant information.

As shown in Fig. 2(a), we select 14 key points as the nodes of the spatial graph from facial landmarks. Considering the natural connection of facial regions, we build the spatial-temporal relationship between landmarks for a single frame. Namely, a landmark-based spatial graph $G_o = \{N_o, E_o\}$ is constructed, where $N_o$ and $E_o$ are the node set and edge set in the spatial graph, respectively. Furthermore, movement features need to be extracted from temporal information that is key to recognizing MEs. Thus, based on the $G_o$ of the onset, apex and offset frames, the GM-Graph $G_{GM}$ is constructed to establish the spatial-temporal relationship between facial landmarks. As shown in Fig. 2(b), for modeling temporal information, the corresponding nodes between the three frames also are connected. GM-Graph $G_{GM} = \{N, E\}$ is constructed, where $N$ and $E$ are the node set and edge set in GM-Graph, respectively. Specifically, $N$ is a collection of the $N_o$ for three keyframes, and $E$ denotes the collection of connections between nodes.
The node features are crucial to represent the spatial-temporal information. On one hand, the captured ME videos need high quality due to the subtle facial movement, which ensures high accuracy in detecting facial landmarks; on the other hand, the movements of facial muscles cause the corresponding movement of facial landmarks. Thus, the landmark coordinates include significant movement information in a ME instance. The data dimension of landmark coordinates is much smaller than that of the raw frames. This advantage of landmark coordinates can save computational resources to facilitate practical applications. In this paper, only landmark coordinates $n = (x, y)$ are adopted as node features to study their effectiveness, and higher-order semantic features (distance and angle between landmarks) are added to explore the interaction of low and high-order geometric information.

Fig. 2(c) shows the calculation of distance $D$ and angle $\alpha$ based on landmark coordinates. Suppose that $n_i = (x_i, y_i)$ (the start point of the arrow in Fig. 2(c)) is the current node, and $n_j = (x_j, y_j)$ (the end point of the arrow) is the neighboring node. $D$ and $\alpha$ can be calculated by

$$
\begin{align*}
D &= (x_i - x_j)^2 + (y_i - y_j)^2 \\
\alpha &= \arctan((y_i - y_j)/(x_i - x_j))
\end{align*}
$$

where $\arctan(*)$ is the arctangent function.

According to the description above, we define two types of node features that $(x, y)$ is Type A, and $(x, y, D, \alpha)$ is Type B, where $(x, y)$ is the landmark coordinates.

**B. SS Module**

Inspired by CNN+LSTM [20] that extracts spatial-temporal features separately, we design a SS module to aggregate spatial and temporal information in GM-Graph. The movement of the facial landmarks in a ME instance is small, and the movement of each landmark along the time step is key to recognizing MEs. Thus, the extraction of temporal features is crucial and challenging for MER. GCN can simultaneously extract the spatial and temporal features in the spatial-temporal graph. However, it cannot focus on the extraction of the temporal features, which may neglect some small movement features. Thus, in this work, TCN is adopted to specifically target temporal information in GM-Graph. Furthermore, since the spatial features of each frame also include important information, e.g., geometric and structural information, GCN is employed to aggregate spatial information in GM-Graph. As shown in Fig. 3, SS module adopts GCN to aggregate the spatial information for three frames, respectively. Then, TCN is adopted to aggregate the temporal information between three frames. Like that, every operation focuses on the aggregation of one type of information (spatial or temporal), which facilitates more finely learning geometric movement features.

1) **GCN With Learnable Adjacency Matrix**: SS module first splits GM-Graph into three sub-graphs $\{G_{GM1}, G_{GM2}, G_{GM3}\}$. The three sub-graphs correspond to the spatial graph $G_o$ of three frames. Suppose that $A$ is the pre-defined adjacency matrix that expresses the node connection in $G_o$, and $A \in \mathbb{R}^{14 \times 14}$. By normalizing $A$, we can get $L = I_n - D^{-1/2}AD^{-1/2}$, where $D = \sum_j A_{ij}$ is the degree matrix. Then, $L$ is used for Fourier transform. For $G_{GM1}(f = 1, 2, 3)$ with spatial information, the graph is filtered by $g_\phi$ to get the output node features $Y_f$ as follows:

$$
Y_f = g_\phi(L)X = U g_\phi(\Lambda)U^T X_f,
$$

where $X_f$ is the node features of $G_{GM1}$; $U$ is the Fourier basis and a set of orthonormal eigenvectors for $L$; and $L = UAU^T$ with $\Lambda$ as corresponding eigenvalues. As the calculation of eigenvectors matrix is expensive, a Chebyshev polynomial with $R$th order is employed to well-approximate the filter $g_\phi$ as follows:

$$
Y_f = \sum_{r=1}^{R} \theta_r C_r(\hat{L})X_f,
$$

where $\theta_r$ denotes Chebyshev coefficients, and $C_r(\hat{L})$ denotes Chebyshev polynomial. $\hat{L} = 2L/\lambda_{max} - I_n$ is normalized to $[-1, 1]$, and $C_r(\hat{L}) = 2LC_{r-1}(\hat{L}) - C_{r-2}(\hat{L})$, where $C_0 = 1$ and $C_1 = \hat{L}$. Suggested by [35], $R = 1$, and $\lambda_{max} = 2$. Then,

$$
Y_f = \theta_0 X_f - \theta_1 D^{-1/2}AD^{-1/2}X_f
\begin{align*}
&= \theta(I_n + D^{-1/2}AD^{-1/2})X_f,
\end{align*}
$$

where $\theta_1$ is approximated to a unified $\theta$, and $\theta = \theta_0 = -\theta_1$. Finally, for simplifying expression, we set $L = I_n + D^{-1/2}AD^{-1/2}$ and the final form of GCN is

$$
Y_f = GCN(G_{GM1}) = LX_f \theta,
$$

where $\theta$ is the learnable filter.

**Learnable Adjacency Matrix LAM**: The above pre-defined adjacency matrix $A$ is fixed and expresses a fixed relationship between nodes. The fixed relationship is defined according to some principles (e.g., facial structure [8] and data-driven [23], [25]). These principles are set by the researchers and are sub-optimal. In this paper, we introduce LAM expressed as $A_L$ to learn a more reasonable relationship between nodes. Given that the fixed $A$ has a certain rationality, $A$ is retained. Therefore, the final form of GCN with LAM is as follows:

$$
Y_f = GCN(A_{GM}) = (L + A_L)X_f \theta,
$$

where $A_L$ can be updated and learned automatically in the training stage of model.
2) TCN: After getting the spatial graph representation processed by GCN with LAM, TCN extracts temporal features between \( G_{GM} \). So, the temporal feature \( F^T_n \) of \( n \)th node is
\[
F^T_n = TCN(Y^1_n, Y^2_n, Y^3_n) = Y_nW,
\]
where \( Y_n = [Y^1_n, Y^2_n, Y^3_n] \), and \( W \) is the learnable filter. Thus, SS module can be formed by
\[
TCN'(GGM'(GGM'), GGM'(GGM'), GGM'(GGM')).
\]
As basic blocks, several SS modules are stacked directly to build a SS Graph Network (SS-GN). Also, based on SS modules, GTS-GN is built as follows.

C. Geometric Two-Stream Graph Network

Besides low-order coordinates, the geometric features also include high-order semantic features, e.g., distance \( D \) and angle \( \alpha \). The distance and angle of facial landmarks include semantic information. It can provide more discriminative features to improve performance. However, it is challenging to learn the discriminative features from the information interaction of the low-order and high-order representations due to their differing distributions. Training two models separately to process two types of features cannot take into account their information interaction. Also, two models trained independently are not end-to-end. Thus, an end-to-end model that is able to aggregate low-order and high-order features separately is needed. To meet this, this paper proposes a novel graph model called GTS-GN to process low-order and high-order geometric features in two streams.

Both features processed by GTS-GN belong to geometric features, and there is a certain correlation between the two features. Thus, an earlier fusion of the two features may be more appropriate for information interaction. However, existing two-stream models fuse two features in the last layer, e.g., [8]. Different from these works, GTS-GN tries to fuse the two features at the earlier layer, not limited to the last layer. The whole network is shown in Fig. 1.

First, we normalize the low-order coordinates and high-order semantic features through Batch Normalization (BN) due to their differences in feature distribution. Next, two types of features are inputted into two streams. Two streams adopt the same structure that stacks several SS modules with the same number. After several SS modules, the outputs of two streams are added together. The added features are inputted into several SS modules or a Full Connected (FC) layer to continue aggregate geometric feature information. Finally, the softmax is used to classify features and predict the ME categories.

D. Adaptive AU Loss

Recent studies [16], [24] have verified the strong correlation between AUs and MEs. Using the strong correlation is beneficial to recognize MEs [24]. Furthermore, the geometric movement information in facial landmarks closely relates to the local movement represented by AUs. Thus, our model builds this correlation by learning the geometric movement features related to AUs before the ME classification layer. In this paper, the loss functions are employed to achieve this intention by constraining feature way.

For AU recognition, multi-label AU loss can constrain the learned features in the classification layer so that they are relevant to AUs for classifying AU categories. As opposed to recognizing AUs, our aim is to constrain the learned features before the ME classification layer so that they are relevant to AUs for classifying ME categories. Thus, we introduce multi-label AU loss before the ME classification layer and ME loss in the ME classification layer. On one hand, before the ME classification layer, the multi-label AU loss can map the geometric features of facial landmarks to AU features. On the other hand, in the ME classification layer, the ME loss function can map AU features to high-level ME features.

The multi-scale features in multiple layers are constrained by the proposed AAU loss as shown in Fig. 4. In order to better aggregate low-order and high-order information individually before fusion, AAU loss constrains the geometric features after fusing two streams. Suppose that the features of \( N_L \) layers are constrained, and the corresponding multi-label AU loss is \( L_r (r = 2, \ldots, N_L) \). So, the loss to constrain the multi-scale features is as follows:
\[
L' = \sum_{r=1}^{N_L} L_r,
\]
where \( L_r \) is conducted by MultiLabelSoftMarginLoss package in Pytorch [39]. However, in (9), the constraint strengths are uniform across all layers, which fails to account for differences between the features of different scales. Thus, we introduce the learnable weights to these losses to emphasize the contributions of the features at different semantic levels, as follows:
\[
L'' = \sum_{r=1}^{N_L} W_r L_r,
\]
where \( W_r \) is a learnable weight and can be updated in the training stage. However, (10) makes the model tend to learn the \( W_r \)s with all zero values. Also, positive weights are needed. Thus, we take probability form as the weights to get Adaptive AU (AAU) loss, as follows:
\[
L_{AAU} = \sum_{r=1}^{N_L} \frac{W_r^2 L_r}{\sum_{r=1}^{N_L} W_r^2}.
\]

where \( N_L \geq 2 \) due to AAU loss will degenerate into AU loss when \( N_L = 1 \).

Total Loss: The task for this work is MER. Thus, ME labels are employed to calculate cross-entropy loss as ME loss (\( L_{ME} \)) in the final classification layer. Thus, the total loss (\( L_T \)) is as follows:
\[
L_T = L_{ME} + \beta * L_{AAU},
\]
where \( \beta \) is the trade-off parameter.
IV. EXPERIMENTS

This section reports the experimental results. We study the effectiveness of landmarks and evaluate the performance of the proposed method. First, the effectiveness of facial landmarks is studied for MER. Second, we carry out the ablation analysis with some visualizations to evaluate the proposed model and components, including SS module, GTS-GN, LAM, and AAU loss. Third, the parameter evaluation is carried out; Finally, we compare the proposed method with state-of-the-art (SOTA) methods.

A. Experimental Setting

The experiments are carried out on SAMM [40] and CASME II [41] datasets. In CASME II, all participants are from one ethnicity, and it contains 255 video samples from 26 subjects with seven ME categories based on self-report. In SAMM, participants are from 13 ethnicities, and it includes 159 samples from 32 subjects with eight ME categories based on self-report. The video samples in both datasets are collected by the high-speed cameras at 200 fps. Both datasets can form seven ME categories based on AU annotations;

In this work, ME categories based on AU annotations are adopted to evaluate the proposed method. Like in previous MER tasks [7], [13], due to the long-tail distribution of the samples, we deleted the categories with less than ten samples. As a result, six and four ME categories based on AU annotations (Type 1 in Table I) are selected from CASME II and SAMM, respectively. Furthermore, the existing methods used not only AU annotations but also self-reported annotations. Thus, for a more comprehensive and fair comparison with SOTA methods, we report the results of the proposed method on other two types of ME category settings: under five ME categories based on self-report (Type 3 in Table I) following [8].

As in the existing works, e.g., [21], [43], the onset, apex and offset frames can be found based on database labels, and their detection belongs to another task [13], [44], [45] in ME analysis. The facial muscle movements are magnified by a learning-based motion magnification method [46], such that the movements of landmarks are more obvious. The amplification factor is set to three following the setup in [21]. In addition, the Dlib [47] package is employed to detect facial landmarks. To study the effectiveness of landmarks, the basic graph-based models with landmark are compared to the basic CNN model with image: ResNet18 [48]. All graph-based models used in our experiment have four layers, and the feature dimensions in different layers are shown in Table II.

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
TABLE III
COMPARISON OF RESNET18 (S), RESNET18 (ST), GCN AND SS-GCN

| Methods      | CAMSE II | SAMM  |
|--------------|----------|-------|
|              | ACC      | FI    | ACC   | FI    |
| ResNet18 (S) | 66.13    | 0.588 | 68.57 | 0.649 |
| ResNet18 (ST)| 71.31    | 0.669 | 72.13 | 0.686 |
| GCN [33]     | 70.12    | 0.690 | 70.21 | 0.661 |
| SS-GN        | 72.90    | 0.716 | 75.17 | 0.732 |

(a) The performance comparison.

| Methods      | Input size | Average Time | Parameter |
|--------------|------------|--------------|-----------|
| ResNet18 (ST)| 3×224×224  | 0.0309       | 11,179,590|
| GCN [33]     | 42×128     | 0.0038       | 39,760    |
| SS-GN        | 14×4×2     | 0.0050       | 163,710   |

(b) The costs. Average Time (Unit: sec.)

Data augmentation is important for the training of deep models. In our experiments, ResNet18 takes the magnified apex frame or three keyframes as inputs. For ResNet-based models, data augmentation is performed by rotation, adding noise, cropping randomly, horizontal flip and color jitter. Also, the detected landmarks for the face regions with different sizes have a small difference. Thus, we regard this difference as noise to augment data. First, the aligned frames are cropped to get the frames of different sizes. Then, we detect the facial landmarks of these frames with different sizes to obtain more data. All models are trained on a single GTX 1050 GPU with Pytorch [39] deep learning framework. The learning rate is set as 0.1 and is divided by 10 at the 30th and 50th epoch (70 epochs in total). The network optimizer adopts stochastic gradient descent (SGD), and the weight decay is 0.0006.

B. The Study on the Effectiveness of Landmarks

First of all, by comparing with images, we evaluate the effectiveness of facial landmarks in terms of recognition performance and model efficiency. For a fair comparison, we use vanilla models to process images and landmarks. In terms of image-based methods, basic CNN models are adopted to process video frames that are Euclidean data. Correspondingly, basic graph models are adopted to process facial landmarks that are non-Euclidean data. Image-based ResNet18 is compared with landmark-based GCN and SS-GCN. ResNet18 (S) and ResNet18 (ST) adopt the same model (ResNet18), with different inputs. Specifically, for ResNet18 (S), the apex frame with three channels is the input image. For ResNet18 (ST), the grayscale images of the onset, apex, and offset frames are taken as three channels of the input image, respectively. Thus, ResNet18 (S) aggregates spatial information, while ResNet18 (ST) aggregates spatial and temporal information.

Table III(a) reports the performance comparison between ResNet18 (S), ResNet18 (ST), GCN and SS-GN. Compared with the other three methods, ResNet18 (S) has the poorest performance, which demonstrates that temporal information has an important contribution to improving MER performance. Compared with image-based ResNet18 (S) and ResNet18 (ST), landmark-based SS-GCN gets substantially better performance under both evaluation metrics. For instance, SS-GCN improves ACC by 1.59% and 3.04%, compared with ResNet18 (ST), on CASME II and SAMM, respectively. It demonstrates that the magnified movement of facial landmarks can effectively represent the muscle movements related to MEs. Also, ME images include redundancy information, which may decrease MER performance. Instead, facial landmarks are a more compact modality that can retain discriminative geometric features for MER and achieve promising performance.

For the model efficiency, as shown in Table III(b), SS-GCN is about 6.18 times faster than image-based ResNet18, respectively. In particular, the model size of SS-GCN is only about 1.5% of that of ResNet18 (ST), which is essential to real-time applications. In addition, GCN and SS-GCN have a lower input dimension compared with ResNet18 (ST). These results demonstrate that compared with image-based methods, landmark-based methods have an obvious advantage in computational cost and efficiency.

To summarize, the magnified movements of facial landmarks contain discriminative movement information for MER, and landmark-based methods can effectively recognize MEs. The below experiment (the results comparing with other methods are shown in Tables X and XI) can also demonstrate the discriminability of facial landmarks for MER tasks. Also, geometric features in facial landmarks are more compact representations with a largely reduced computational cost. Overall, compared with the image-based methods, the landmark-based graph methods have much higher computational and parameter efficiency with a competitive recognition rate.

Discussions: Inaccurate facial landmarks significantly affect the methods reliant on facial landmarks. Obviously, our method relies on the accuracy of detecting landmarks. In MER task, existing movement-based methods need to detect landmarks to align faces. The crop of local areas in the local feature-based methods [8], [28], [37] also depends on facial landmarks. Thus, this problem is inevitable in MER task.

For the used datasets, the scenarios are controlled and rather simple, so landmarks can be reliably detected. But when moving to uncontrolled environments with complicated mixed movements and big illumination changes, more powerful landmark detection technology is needed to get accurate landmarks. Fortunately, the detection technologies of landmarks are ongoing research, and some works [49] have promising results, which is technical support for landmark-based methods. Also, image-based methods are greatly affected in complex environments, thus, MER under complex environments is another unsolved and challenging task.

Overall, landmark-based methods face the above-mentioned inevitable problems like other image-based methods. Although image-based methods can aggregate more appearance information, landmark-based methods can avoid the cumulative error caused by cropping ROI areas and extracting features from these areas. In addition, this paper demonstrates the discriminability and efficiency of landmarks. Yet the geometric features from landmarks and appearance features from ME images do not conflict. Therefore, how to better combine appearance features with geometric features is also worthy of further study.
TABLE IV

| Methods            | CAMSE II | SAMM |
|--------------------|----------|------|
|                    | ACC      | F1   | ACC  | F1   |
| GCN+Type A         | 70.12    | 0.699| 70.21| 0.661|
| SS-GN+Type A       | 72.91    | 0.716| 75.17| 0.732|
| SS-GN+Type B       | 72.91    | 0.710| 73.05| 0.693|
| GTS-GN+Type B      | 73.31    | 0.717| 75.89| 0.741|

(a) Graph models with different node features.

| Methods             | CAMSE II | SAMM |
|---------------------|----------|------|
| GTS-GN (layer 1)    | 73.31    | 0.717| 76.60| 0.729|
| GTS-GN (layer 2)    | 72.91    | 0.714| 75.17| 0.723|
| GTS-GN (layer 3)    | 71.71    | 0.690| 73.05| 0.708|
| GTS-GN (layer 4)    | 68.13    | 0.640| 75.89| 0.741|

(b) GTS-GN (n layer), n = 1, 2, 3, 4.

TABLE V

| Components     | CAMSE II | SAMM |
|----------------|----------|------|
| GTS-GN         |          |      |
| LAM            |          |      |
| AAU Loss       |          |      |
| GTS-GN         | 73.31    | 0.717| 75.89| 0.741|
| LAM            | 74.90    | 0.732| 78.01| 0.782|
| AAU Loss       | 76.10    | 0.745| 78.01| 0.756|
| GTS-GN+LAM     | 77.29    | 0.765| 79.43| 0.782|

ACC(%)

C. Ablation Analysis

This section reports and analyzes the results of ablation study for the proposed components. It is divided into three sub-sections: the evaluation on SS-module and GTS-GN, the evaluation on LAM, and the evaluation on AAU loss.

1) The Evaluation on SS-Module and GTS-GN: The Evaluation on SS-module: To evaluate SS module, the vanilla GCN [35] is taken as the baseline and directly processes the whole GM-Graph. As shown in Table IV(a), taking landmark coordinates (Type A) as node features, SS-GN outperforms GCN in terms of both accuracy and F1-score for both datasets. It demonstrates that the features extracted by SS module are more discriminative than those extracted by GCN. Also, extracting spatial and temporal geometric features separately is a better choice for MER.

The Evaluation on GTS-GN: We evaluate the performance of GTS-GN in aggregating low-order and high-order geometric features. Type B (x,y,D,\(\alpha\)) is taken as node features. According to Table IV(a), SS-GN+Type A is superior to SS-GN+Type B, which shows that simply introducing high-order geometric information cannot ensure performance improvement. GTS-GN+Type B outperforms both SS-GN+Type B and SS-GN+Type A. It demonstrates that GTS-GN can aggregate low-order and high-order geometric information more effectively than SS-GN. Also, distance \(D\) and angle \(\alpha\) provide discriminative high-order geometric information.

We explore a better way to fuse two feature flows in GTS-GN. As listed in Table IV(b), interestingly, we find that different datasets prefer different ways to fuse the features. For CASME II, fusing two-stream features at the first layer provides better performance, while for SAMM, fusing at the fourth layer provides a better F1-score and comparable accuracy. This may be caused by the different ME categories in these two datasets. The experiment results demonstrate that fusing two feature flows in the last layer is not optimal, and earlier fusion may result in better performance.

Overall, SS module can effectively aggregate spatial-temporal information in GM-Graph, and GM-Graph with only geometric features as node features can be processed by graph models to get promising results. GTS-GN provides flexibility to fuse and aggregate effectively the low-order and high-order geometric features. Additionally, two stream-based GTS-GN is more effective than one stream-based SS-GN.

2) The Evaluation on Learnable Adjacency Matrix: To evaluate the effectiveness of LAM, GTS-GN with LAM is compared with GTS-GN without LAM. Table V shows the comparison results. It turns out that both GTS-GN without and with AAU loss are improved after using LAM. It demonstrates that LAM is effective for MER to learn a more reasonable adjacency matrix.

Visualization: To illustrate the intuitive advantages of LAM, we visualize the learned LAM on CASME II. Fig. 5 shows the heatmap of the learned LAMs in different layers. From this figure, different layers have different LAMs, and as the layer deepens, the average value of the learned LAM is smaller. It can be interpreted that the node features in earlier layers represent low-level feature information, and have limited information interactions with each other. Thus, there is a high demand for modelling the interaction between nodes. For the deeper layers, every node feature represents higher-level information and has performed certain interactions with each other. Therefore, the demand for modelling the relationships between different nodes have decreased.

We visualize the ten edges with larger values in the LAM of the last layer. As shown in Fig. 6, LAM learns the relationship between different facial muscle regions. These regions do not have connections in the pre-defined adjacency matrix, e.g., the connections between the mouth and eyebrows regions. Thus, introducing LAM allows the model to more effectively capture interaction relationships between node features.

Overall, LAM can consider the difference between different layers to build the relationship between different facial regions in multi-scales. In this way, more reasonable adjacency matrices can be learned to aggregate the node features.

3) The Evaluation on Adaptive AU Loss: This section evaluates GTS-GN, GTS-GN with AAU loss. Table V reports the results that add AU loss to 4 layers, respectively. It turns out that constraining the features in different layers has different performance. Overall, the performance is improved after using AAU loss. In addition, the performance maybe worse after adding AAU loss in some layers, e.g., layer 2 on CASME II.

Next, AAU loss is effective and can improve the performance as shown in Table V. From Table V, both GTS-GN and GTS-GN with LAM achieve a better performance after using AAU.
TABLE VI
THE EVALUATION ON GTS-GN WITH AU LOSS. AU LOSS N DENOTES THAT AU LOSS CONSTRAINS THE FEATURES IN THE NTH LAYER

| Methods          | CAMSII ACC (%) | CAMSII F1 | SAMM ACC (%) | SAMM F1 |
|------------------|----------------|-----------|--------------|---------|
| GTS-GN           | 72.91          | 0.697     | 70.92        | 0.673   |
| GTS-GN with AU loss 1 | 73.31       | 0.712     | 75.18        | 0.734   |
| GTS-GN with AU loss 2 | 71.71       | 0.697     | 74.47        | 0.713   |
| GTS-GN with AU loss 3 | 73.31       | 0.727     | 75.18        | 0.729   |
| GTS-GN with AU loss 4 | 72.91       | 0.708     | 75.18        | 0.724   |

Fig. 5. Visualization results of LAM. (a) to (d) mean the LAMs in layers 1 to 4, respectively.

Fig. 6. Visualization for LAM in layer 1. The line connecting two nodes indicates that the two nodes are strongly correlated, namely, the value in LAM is larger. Red line: eyebrows-nose; Blue line: eyebrows-mouth; Green line: nose-mouth; Black line: eyebrows, where A-B indicates that the connects are between A and B regions.

TABLE VII
THE EVALUATION ON AU LOSS AND AAU LOSS

| Models       | Basic Model ACC (%) | Basic Model F1 | AU Loss ACC (%) | AU Loss F1 | AAU Loss ACC (%) | AAU Loss F1 |
|--------------|---------------------|----------------|-----------------|-----------|------------------|-----------|
| GTS-GN (layer 1) | 72.91               | 0.697          | 73.31           | 0.728     | 75.70            | 0.749     |
| GTS-GN (layer 2) | 74.10               | 0.730          | 74.50           | 0.738     | 77.29            | 0.765     |
| GTS-GN (layer 3) | 74.90               | 0.720          | 75.30           | 0.745     | 75.30            | 0.744     |
| GTS-GN (layer 4) | 70.52               | 0.688          | 71.31           | 0.697     | -                | -         |

(a) On CAMSII

| Models       | Basic Model ACC (%) | Basic Model F1 | AU Loss ACC (%) | AU Loss F1 | AAU Loss ACC (%) | AAU Loss F1 |
|--------------|---------------------|----------------|-----------------|-----------|------------------|-----------|
| GTS-GN (layer 1) | 70.92               | 0.673          | 75.18           | 0.743     | 77.30            | 0.765     |
| GTS-GN (layer 2) | 76.60               | 0.739          | 75.18           | 0.722     | 79.45            | 0.778     |
| GTS-GN (layer 3) | 73.76               | 0.700          | 77.30           | 0.767     | 78.72            | 0.782     |
| GTS-GN (layer 4) | 78.01               | 0.782          | 77.30           | 0.753     | -                | -         |

(b) On SAMM

loss. It demonstrates that AAU loss is helpful in learning more discriminative features for recognizing MEs.

Furthermore, as shown in Table VII, AAU loss is compared with AU loss and basic model (GTS-GN) under four types of GTS-GN. Since GTS-GN (layer 4) only has one constrained layer while $N_L > 1$, GTS-GN (layer 4) with AU loss doesn’t have results. On one hand, AAU loss can greatly improve the performance of GTS-GN in all cases. On the other hand, although AU loss doesn’t promote GTS-GN (layer 2) on SAMM, it still improves the performance of the model in most cases. The above results indicate that before the classification layer, utilizing AU information to constrain features can enhance the performance of MER.

More detailed, although AU loss and AAU loss have similar performance for GTS-GN (layer 3) on CASME II, for all other cases, AAU loss is superior to AU loss. It demonstrates that adaptively constraining multi-scale features across multiple layers is more advantageous than fixed constraints on the features of a single layer.

Visualization: We further analyze AAU loss by showing the learned weight values $\sum_{L} W^2_{L}$ and visualizing the learned features in different layers. GTS-GN (layer 1) is taken as an example, so there are four learnable weights, namely $N_L = 4$. Table VIII shows the learned four weights for 26 subjects on CASME II. It turns out that the weights for the first two layers are much larger than those for the last two layers. AAU loss focuses on limiting the features in the first two layers to aggregate AU information. In fact, the movements of facial landmarks have a direct relationship with AUs. Thus, AAU loss can constrain the features to achieve the aggregation of high-level AU information in more early layers.

To confirm the above points, we visualize the features in different layers. t-SNE [50] is employed to visualize these features to be a scatter plot. The visualization results with and without AAU loss are shown in Fig. 7. Obviously, compared with not using AAU loss, the extracted features have some regularities on the first two layers after using AAU loss. Specifically, for the first layer, the features corresponding to the same category hardly appear cluster when not using AAU loss. While after using AAU loss, those begin to cluster. For the second layer, the clustering effect using AAU loss is better than that not using AAU loss. For instance, for category III, the features not using AAU loss
are distributed on the whole plane, while those using AAU loss are mainly distributed on the left side. Also, Fig. 7(f) shows that the features using AAU loss have some special intersection relationships that correspond to the relationships between AUs and MEs. In fact, a ME is related to single or multiple AUs, and different MEs may share a common AU, e.g., both II and V include AU1, and both III and IV include AU4, AU5, and AU7 (corresponding to Fig. 7(f), the features of II and V or III and IV have some intersections). Above results indicate that AAU loss can constrain the features in the earlier layers to represent high-level AU information, which is consistent with the learned weights.

Based on AU-related features learned in the first two layers, AAU has an obvious advantage for the learned features in the latter two layers. Specifically, for the third layer as shown in Fig. 7(c) and (g), the features using AAU loss obviously are superior to those not using AAU loss whether in terms of intra-class or inter-class. For the last layer, the features using AAU loss maintain the advantage with larger inter-class and smaller intra-class.

Overall, in earlier layers, the AAU loss-constrained model focuses on learning high-level AU features from facial landmarks, while in deeper layers, it focuses on learning high-level ME features from high-level AU features. By introducing AAU loss, the learned features in the earlier layers adhere to some rules derived from AU information, which enhances the discriminative ability of the learned features in the deeper layers for the classification of MEs.

### D. Parameter Evaluation

The affect of the used landmarks and the trade-off parameter $\beta$ are evaluated for the proposed method. We test three types of

---

**TABLE VIII**

| layer | 1    | 2    | 3    | 4    |
|-------|------|------|------|------|
| Subject 1 | 0.5662 | 0.4127 | 0.0010 | 0.0201 |
| Subject 2 | 0.6992 | 0.2246 | 0.0118 | 0.0643 |
| Subject 3 | 0.6732 | 0.2727 | 0.0064 | 0.0477 |
| Subject 4 | 0.5312 | 0.4448 | 0.0055 | 0.0119 |
| Subject 5 | 0.6153 | 0.3296 | 0.0005 | 0.0546 |
| Subject 6 | 0.6404 | 0.3064 | 0.0003 | 0.0529 |
| Subject 7 | 0.6879 | 0.2514 | 0.0111 | 0.0497 |
| Subject 8 | 0.6078 | 0.3593 | 0.0000 | 0.0329 |
| Subject 9 | 0.6376 | 0.2848 | 0.0002 | 0.0774 |
| Subject 10 | 0.6724 | 0.2545 | 0.0025 | 0.0705 |
| Subject 11 | 0.6411 | 0.3087 | 0.0005 | 0.0498 |
| Subject 12 | 0.6625 | 0.2844 | 0.0033 | 0.0499 |
| Subject 13 | 0.6161 | 0.3341 | 0.0002 | 0.0495 |
| Subject 14 | 0.6056 | 0.3657 | 0.0000 | 0.0287 |
| Subject 15 | 0.6035 | 0.3484 | 0.0009 | 0.0417 |
| Subject 16 | 0.6801 | 0.2621 | 0.0080 | 0.0498 |
| Subject 17 | 0.5765 | 0.3853 | 0.0025 | 0.0357 |
| Subject 18 | 0.6387 | 0.3238 | 0.0017 | 0.0358 |
| Subject 19 | 0.6515 | 0.3021 | 0.0022 | 0.0441 |
| Subject 20 | 0.6758 | 0.2470 | 0.0027 | 0.0745 |
| Subject 21 | 0.5488 | 0.4267 | 0.0033 | 0.0211 |
| Subject 22 | 0.6039 | 0.3501 | 0.0007 | 0.0452 |
| Subject 23 | 0.6418 | 0.3126 | 0.0009 | 0.0446 |
| Subject 24 | 0.6341 | 0.3256 | 0.0007 | 0.0396 |
| Subject 25 | 0.6668 | 0.2760 | 0.0055 | 0.0537 |
| Subject 26 | 0.6781 | 0.2538 | 0.0050 | 0.0611 |
| Average | 0.6399 | 0.3172 | 0.0029 | 0.0464 |
landmark point sets as shown in Fig. 8, and $\beta$ are tested to show the balance between AAU loss and ME loss.

The Affect of the Used Landmarks:

Except for the 14 points of the graph defined in the 3.1 section, other two types of landmark point sets are tested, including 1) 31 points: all landmarks of mouth, nose and eyebrows regions; and 2) 68 points: all landmarks of whole face. Table IX reports the comparative results. Obviously, the 14-point set has an obvious advantage and the performance of the 68-point set is the worst. It turns out that under including the key information of eyebrows, noses and mouth, with the point number increases, the performance drops. Compared with the 14-point set, the 31-point set includes redundant information. 68-point set not only contains a lot of redundant information, but also some interference information, e.g., the eyes and facial contours. Also, the 14-point set has fewer points, which reduces the computational cost and improves efficiency. Thus, removing interference and redundant points while retaining key information for eyebrows, nose, and mouth is beneficial to improve performance and efficiency.

The Affect of $\beta$: Fig. 9 shows the evaluation results of $\beta$. The proposed method achieves the highest accuracy under $\beta=0.1$ (78.49%) on CAMSE II and $\beta=0.2$ (80.14%) on SAMM. Furthermore, under $\beta=0.3$ and 0.8, the performances on both datasets reach the valley. In general, the performance under $\beta$ closing to 0.1 or 1 is superior to that under $\beta$ in the middle value from 0.1 to 1. Under $\beta$ closing to 0.1, the highest accuracy can be achieved. It means that it is a better choice to take AU information as auxiliary information and maintain the dominance of ME loss.

E. Comparing With Other Methods

To evaluate the proposed method (GTS-GN with AAU loss and LAM, GTS-GN-AL), we compare it to existing state-of-the-art methods. As shown in Table X, comparing with the existing methods recognizing five ME categories based on AU labels, GTS-GN-AL provides a much better performance on both datasets and achieves a new SOTA performance. Specifically, on CASME II, GTS-GN-AL outperformed TL by 11.53% in ACC and 0.198 in F1-score, and on SAMM, the differences were 1.47% and 0.169, respectively.

For a more comprehensive comparison, we report the comparative results under the mainstream test categories: five ME categories based on self-report labels. According to Table XI, GTS-GN-AL gets a comparable performance as a whole comparing with recent methods. Specifically, compared to the CNN-based methods, GEME and LGCon, GTS-GN-AL provides much better performance on both datasets. Also, DeRe-GRL, GACN and GRAUF are SOTA methods. First, GTS-GN-AL has a slight advantage over DeRe-GRL. Second, compared with GRAUF, GTS-GN-AL get a bit worse result on SAMM and a better result on CASME II, especially in terms of F1-score (0.107 higher than GRAUF on CASME II). Unlike our method...
with a compact landmarks representation, GRAUF aggregated more information, e.g., AUs and the magnified shape. Also, GRAUF employs several models to deal with different features, which increases model complexity and reduces efficiency. Finally, compared with GACN, GTS-GN-AL consistently shows reasonable graph structure that builds the relationship between facial landmarks, AUs and MEs. AAU loss can reasonably learn and build strong correlations between geometric features with graph-based models. This framework is more valuable for practical applications due to its low computational costs and comparable performance.

V. CONCLUSION

This article explored the contribution of facial landmarks and verified the effectiveness of facial landmarks for MER. Notably, only the geometric information of facial landmarks is aggregated by the graph model to achieve MER task. We customized a GM-Graph based on facial landmarks of three keyframes to model geometric and dynamic information in ME videos. Then, SS module was designed to learn deep spatial and temporal features of GM-Graph. The experimental results demonstrate that SS module can aggregate the spatial and temporal information better. Furthermore, geometric information including the low-order coordinates and high-order semantic features are both involved. A new graph model GTS-GN was proposed, which models information interaction and makes better use of complementary information from two types of geometric features. LAM can automatically learn a more reasonable graph structure that builds the relationship between different facial muscle regions and between different nodes. AAU loss can reasonably learn and build strong correlations between facial landmarks, AUs and MEs. AAU loss can adaptively constrain the multi-scale movement features to aggregate AU information in an efficient way, learning more discriminative ME features.

This work encourages further investigation of the framework that takes low-dimensional landmarks as input to extract compact geometric features with graph-based models. This framework is more valuable for practical applications due to its low computational costs and comparable performance.

REFERENCES

[1] S. Porter and L. Brinke, “Reading between the lines: Identifying concealed and falsified emotions in universal facial expressions,” Psychol. Sci., vol. 19, no. 5, pp. 508–514, Jun. 2008.
[2] S. Weinberger, “Airport security: Intent to deceive?,” Nature, vol. 465, pp. 412–415, 2010.
[3] W. Yan, Q. Wu, Y. Chen, J. Liang, and X. Fu, “How fast are the leaked facial expressions: The duration of micro-expressions,” J. Nonverbal Behav., vol. 37, no. 4, pp. 217–230, 2013.
[4] V. Esmaeili, M. M. Feighghi, and S. O. Shahidi, “A comprehensive survey on facial micro-expression: Approaches and databases,” Multimedia Tools Appl., vol. 81, no. 28, pp. 40089–40134, 2022.
[5] H. Chen, H. Shi, X. Liu, X. Li, and G. Zhao, “SMG: A micro-gesture dataset towards spontaneous body gestures for emotional stress state analysis,” Int. J. Comput. Vis., vol. 131, no. 6, pp. 1346–1366, 2023.
[6] Z. Xia, W. Peng, H.-Q. Khor, X. Feng, and G. Zhao, “Revealing the invisible with model and data shrinking for composite-database micro-expression recognition,” IEEE Trans. Image Process., vol. 29, pp. 8590–8605, 2020.
[7] Y.-J. Lu and J. Yan, “A comparative study on movement feature in different directions for micro-expression recognition,” Neurocomputing, vol. 449, pp. 159–171, 2021.
[8] A. J. R. Kumar and B. Bhanu, “Micro-expression classification based on landmark relations with graph attention convolutional network,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops, 2021, pp. 1511–1520.
[9] Y. Li, J. Wei, Y. Liu, J. Kauttonen, and G. Zhao, “Deep learning for micro-expression recognition: A survey,” IEEE Trans. Affective Comput., vol. 13, no. 4, pp. 2028–2046, Fourth Quarter, 2022.
[10] X. Li et al., “4DME: A spontaneous 4D micro-expression dataset with multimodalities,” IEEE Trans. Affective Comput., vol. 14, no. 4, pp. 3031–3047, Fourth Quarter, 2023.
[11] Y.-J. Liu, B.-J. Li, and Y.-K. Lai, “Sparse MDMO: Learning a discriminative feature for spontaneous micro-expression recognition,” IEEE Trans. Affective Comput., vol. 12, no. 1, pp. 254–261, First Quarter, 2021.
[12] Y. Zong, X. Huang, W. Zheng, Z. Cui, and G. Zhao, “Learning from hierarchical spatiotemporal descriptors for micro-expression recognition,” IEEE Trans. Multimedia, vol. 20, no. 11, pp. 3160–3172, Nov. 2018.
[13] Y.-J. Liu, X. Huang, and G. Zhao, “Joint local and global information learning with single apex frame detection for micro-expression recognition,” IEEE Trans. Image Process., vol. 30, pp. 249–263, 2021.
[14] J. Wei, G. Lu, J. Yan, and Y. Zong, “Learning two groups of discriminative features for micro-expression recognition,” Neurocomputing, vol. 479, pp. 22–36, 2022.
[15] W. Peng, X. Hong, Y. Xu, and G. Zhao, “A boost in revealing subtle facial expressions: A consolidated eulerian framework,” in Proc. 14th IEEE Int. Conf. Autom. Face Gesture Recognit., Lille, France, 2019, pp. 296–300.
[16] H.-X. Xie, L. Lo, H.-H. Shuai, and W.-H. Cheng, “Au-assisted graph attention convolutional network for micro-expression recognition,” in Proc. 28th ACM Int. Conf. Multimedia, New York, NY, USA, 2020, pp. 2871–2880.
[17] 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.
[18] J. Wei, G. Lu, J. Yan, and H. Liu, “Micro-expression recognition using local binary pattern from five intersecting planes,” Multimedia Tools Appl., vol. 81, no. 15, pp. 20643–20668, 2022.
[19] Y.-H. Oh, J. See, A. C. Le Ngo, R. C.-W. Phan, and V. M. Baskaran, “A survey of automatic facial micro-expression analysis: Databases, methods, and challenges,” Front. Psychol., vol. 9, p. 1128, 2018.
[20] D. Y. Choi and B. C. Song, “Facial micro-expression recognition using two-dimensional landmark feature maps,” IEEE Access, vol. 8, pp. 121549–121563, 2020.

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
[21] L. Lei, J. Li, T. Chen, and S. Li, “A novel graph-TCN with a graph structured representation for micro-expression recognition,” in Proc. 28th ACM Int. Conf. Multimedia, 2020, pp. 2237–2245.

[22] Y. Liu, X. Zhang, Y. Li, J. Zhou, X. Li, and G. Zhao, “Graph-based facial affect analysis: A review,” IEEE Trans. Affective Comput., vol. 14, no. 4, pp. 2657–2677, Fourth Quarter, 2023.

[23] L. Lei, T. Chen, S. Li, and J. Li, “Micro-expression recognition based on facial graph representation learning and facial action unit fusion,” in Proc. IEEE/CVF Conf. Vis. Pattern Recognit. Workshops, 2021, pp. 1571–1580.

[24] L. Zhong, Q. Mao, and M. Dong, “Objective class-based micro-expression recognition through simultaneous action unit detection and feature aggregation,” 2020, arXiv:2012.13148.

[25] X. Chui, Y.-Q. Chen, and W.-H. Cheng, “MER-GCN: Micro-expression recognition based on relation modeling with graph convolutional networks,” in Proc. IEEE Conf. Multimedia Inf. Process. Retrieval, 2020, pp. 79–84.

[26] P. Ekman and W. Friesen, “Facial action coding system: A technique for the measurement of facial movement,” Consulting Psychologists Press Palo Alto, vol. 3, no. 2, p. 5, Jan. 1978.

[27] X. Li, T. Pfister, X. Huang, G. Zhao, and M. Pietikäinen, “A spontaneous micro-expression database: Induction, collection and baseline,” in Proc. 16th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit., Shanghai, China, 2013, pp. 1–6.

[28] Y.-J. Liu, J.-K. Zhang, W.-J. Yan, S.-J. Wang, G. Zhao, and X. Fu, “A main directional mean optical flow feature for spontaneous micro-expression recognition,” IEEE Trans. Affective Comput., vol. 7, no. 4, pp. 299–310, Fourth Quarter, 2016.

[29] X. Huang, S. Wang, X. Liu, G. Zhao, X. Feng, and M. Pietikäinen, “Discriminative spatiotemporal local binary pattern with revisited integral projection for spontaneous facial micro-expression recognition,” IEEE Trans. Affective Comput., vol. 10, no. 1, pp. 32–47, First Quarter, 2019.

[30] S. Polikovsky, Y. Kameda, and Y. Ohta, “Facial micro-expressions recognition using high speed camera and 3D-gradient descriptor,” in Proc. 3rd Int. Conf. Imag. Crime Detection Prevention, 2010, pp. 1–6.

[31] X. Nie, M. A. Takalkar, M. Duan, H. Zhang, and M. Xu, “GEME: Dual-stream multi-task gender-based micro-expression recognition,” Neurocomputing, vol. 427, pp. 13–28, 2021.

[32] H.-Q. Khor, J. See, S. Liong, R. Phan, and W. Lin, “Dual-stream shallow networks for facial micro-expression recognition,” in Proc. IEEE Int. Conf. Image Process., Taipei, Taiwan, 2019, pp. 36–40.

[33] S. Liong, J. See, R. Phan, K. Wong, and S.-W. Tan, “Hybrid facial regions extraction for micro-expression recognition system,” J. Signal Process. Syst., vol. 90, no. 4, pp. 601–617, 2018.

[34] L. Zhong, C. Bai, J. Li, T. Chen, S. Li, and Y. Liu, “A graph-structured representation with BRNN for static facial-expression recognition,” in Proc. IEEE 14th Int. Conf. Autom. Face Gesture Recognit., Lille, France, 2019, pp. 270–274.

[35] T. Kipf and M. Welling, “Semi-supervised classification with graph convolutional networks,” in Proc. Int. Conf. Learn. Representations, 2017, pp. 1–10.

[36] J. Wei, H. Chen, G. Lu, J. Yan, Y. Xie, and G. Zhao, “Prior information based decomposition and reconstruction learning for micro-expression recognition,” IEEE Trans. Inf. Syst., vol. E106-D, no. 10, pp. 1752–1756, 2023.

[37] S. Bai, J. Z. Kolter, and V. Koltun, “An empirical evaluation of generic convolutional and recurrent networks for sequence modeling,” 2018, arXiv:1803.01271.

[38] S. Yan, Y. Xiong, and D. Lin, “Spatial temporal graph convolutional networks for skeleton-based action recognition,” in Proc. AAAI Conf. Artif. Intell., 2018, Art. no. 912.

[39] A. Paszke et al., “PyTorch: An imperative style, high-performance deep learning library,” in Proc. 33rd Conf. Neural Inf. Process. Syst., Vancouver, Canada, 2019, Art. no. 721.

[40] A. Davison, C. Lansley, N. Costen, K. Tan, and M. H. Yap, “SAMM: A spontaneous micro-facial movement dataset,” IEEE Trans. Affective Comput., vol. 9, no. 1, pp. 116–129, First Quarter, 2018.

[41] W. Yan et al., “CASME II: An improved spontaneous micro-expression database and the baseline evaluation,” PLoS One, vol. 9, no. 1, 2014, Art. no. e86041.

[42] M. Feng, Z. Wu, Z. Zhang, and T. Chen, “From macro to micro expression recognition: Deep learning on small datasets using transfer learning,” in Proc. IEEE 13th Int. Conf. Autom. Face Gesture Recognit., X’ian, China, 2018, pp. 657–661.

[43] Y. Liu, H. Du, L. Zheng, and T. Gedeon, “A neural micro-expression recognizer,” in Proc. IEEE 14th Int. Conf. Autom. Face Gesture Recognit., Lille, France, 2019, pp. 631–634.

[44] T.-K. Tran, Q.-N. Vo, X. Hong, X. Li, and G. Zhao, “Micro-expression spotting: A new benchmark,” Neurocomputing, vol. 443, pp. 356–368, 2021.

[45] S.-J. Wang, Y. He, J. Li, and X. Fu, “MESNet: A convolutional neural network for spotting multi-scale micro-expression intervals in long videos,” IEEE Trans. Image Process., vol. 30, pp. 3956–3969, 2021.

[46] M. Peng, J. Kolter, and V. Koltun, “An empirical evaluation of ensemble of regression trees,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Columbus, OH, 2014, pp. 1867–1874.

[47] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Seattle, WA, 2016, pp. 770–778.

[48] J. Wan, Z. Lai, J. Li, J. Zhou, and C. Gao, “Robust facial landmark detection by multiorder multiconstraint deep networks,” IEEE Trans. Neural Netw. Learn. Syst., vol. 33, no. 5, pp. 2181–2194, May 2022.

[49] W. Yan et al., “CASME II: An improved spontaneous micro-expression database,” in Proc. Proc. IEEE 14th Int. Conf. Autom. Face Gesture Recognit., Xian, China, 2018, pp. 667–674.

Jinsheng Wei received the PhD degree from NUIST, China, in 2022. From 2020 to 2021, he was a visiting student with the Center for Machine Vision and Signal Analysis, University of Oulu, Finland. He is currently a lecturer with the Nanjing University of Posts and Telecommunications (NUIST), Nanjing, China. His current research interests include micro-expression recognition, image processing, pattern recognition, computer vision, and machine learning.

Wei Peng received the MS degree in computer science from Xiamen University, Xiamen, China, in 2016, and the PhD degree from the University of Oulu, Finland, in 2022. He is currently postdoctoral scholar with Stanford University. His articles have published in mainstream conferences and journals, such as CVPR, AAAI, ICCV, ACM Multimedia, IEEE Transactions on Pattern Analysis and Machine Intelligence, and IEEE Transactions on Image Processing. His current research interests include machine learning, medical imaging analysis, and neuroscience.

Guanning Lu received the BE degree in radio engineering and the MS degree in communication and electronic systems from the Nanjing University of Posts and Telecommunications (NUIST), Nanjing, China, in 1985 and 1988, respectively, and the PhD degree in communication and information systems from Shanghai Jiao Tong University, Shanghai, China, in 1999. He is currently a professor with the College of Communication and Information Engineering, NUIST. His current research interests include image processing, affective computing, and machine learning.
Yante Li received the PhD degree in computer science from the University of Oulu, Oulu, Finland, in 2022. She is currently a postdoc with the Center for Machine Vision and Signal Analysis of University of Oulu. Her current research interests include affective computing, micro-expression analysis and facial action unit detection.

Jingjie Yan received the BE degree in electronic science and technology, in 2006 and the MS degree in signal and information processing, in 2009, both from the China University of Mining and Technology, and the PhD degree in single and information processing from Southeast University, Nanjing, China, in 2014. He is currently a associate professor with the College of Communication and Information Engineering, Nanjing University of Posts and Telecommunications (NUPT). His current research interests include pattern recognition, affective computing, computer vision and machine learning.

Guoying Zhao (Fellow, IEEE) is currently an academy professor with the University of Oulu, IAPR fellow, and AAIA fellow. She has authored or co-authored more than 280 papers in journals and conferences with more than 19120 citations in Google Scholar and h-index 65. She is a co-program chair for ICMI 2021, has served as area chair for several conferences, and is an associate editor for Pattern Recognition, IEEE Transactions on Circuits and Systems for Video Technology, and Image and Vision Computing Journals. Her current research interests include image and video descriptors, facial expression and micro-expression recognition, emotional gesture analysis, affective computing, and biometrics.