The Visual Computer
https://doi.org/10.1007/s00371-025-03817-9

RESEARCH

Enhancing facial action unit recognition through topological feature integration and relational learning

Keqi Li[1] · Yaping Wan[1] · Gang Zou[1,2] · Wangxiu Li[1] · Jian Yang[3] · Changyi Xie[4]

Accepted: 17 January 2025

Abstract

Facial action unit (AU) recognition involves predicting the activation states of AUs, which describe facial movements. In complex scenarios, AU relationships and facial features are challenging to capture effectively. This study proposes a novel AU recognition approach that supplements topological features with AU relationship learning. By integrating a channel-topology convolution feature generation structure (CCFG) with a multi-scale attention feature generation structure (MAFG) within a graph neural network, our method models dynamic AU associations and enriches feature representations. Experimental results demonstrate that our approach significantly outperforms state-of-the-art methods on benchmark datasets, achieving average F1 scores of 66.6% and 67.2% on DISFA and EmotioNet, respectively, highlighting its robustness and precision in facial expression recognition tasks. Our code and dataset documentation are available at https://github.com/lkq52110/au-recognition.

Keywords Facial AU recognition · AU relationship learning · Multi-scale learning · Topological feature learning

1 Introduction

Facial action unit (AU) recognition is a domain within facial expression analysis, focusing on predicting the probability of each AU's presence or absence. This field is governed by the facial action coding system (FACS) [1, 2], a highly comprehensive and objective framework for describing subtle facial expressions. Unlike emotion-based categorical models, AUs offer a more comprehensive and objective description of facial expressions [3] and play a significant role in complex applications, such as research into neurodegenerative cognitive disorders and emotion recognition.

AU recognition has made significant progress over time. Early methods laid the foundation for AU recognition by utilizing prior knowledge and deep learning techniques [4, 5]. However, these methods struggled to achieve high-precision recognition in complex applications. To improve accuracy, region-based learning methods emerged, exploiting the spatial relationships between facial keypoints and AUs [6–8]. Despite their progress, these methods were limited by predefined regions and could not handle dynamic AU variations across different samples. As shown in Fig. 1a, traditional methods directly learn features based on predefined binary labels for AU activation or non-activation, without considering the potential correlations between AUs. Recent approaches have turned to graph-based methods [9] to model the interdependencies between AUs, utilizing graph neural networks (GNNs) to capture AU correlations. As described in Fig. 1b, by learning the collaborative mechanisms of AU activation states, graph-based methods can capture dynamic intensity relationships between AUs and apply them to the recognition weights of each AU. Despite these advancements, there is still room for improvement, particularly in detecting AUs in obstructed or complex environments.

In this paper, we summarize the key challenges and issues faced in AU recognition. *Problem 1.* Traditional methods often rely on a single or fixed feature map for AU recognition

✉ Yaping Wan
512828758@qq.com

✉ Gang Zou
37759542@qq.com

[1] School of Computer, University of South China, Hengyang, China

[2] HuNan ZK Help Innovation Intelligent Technology Research Institute, Changsha, China

[3] Hunan Cancer Hospital, Changsha, China

[4] School of Psychological Sciences, The University of Melbourne, Melbourne, Australia

Published online: 31 January 2025

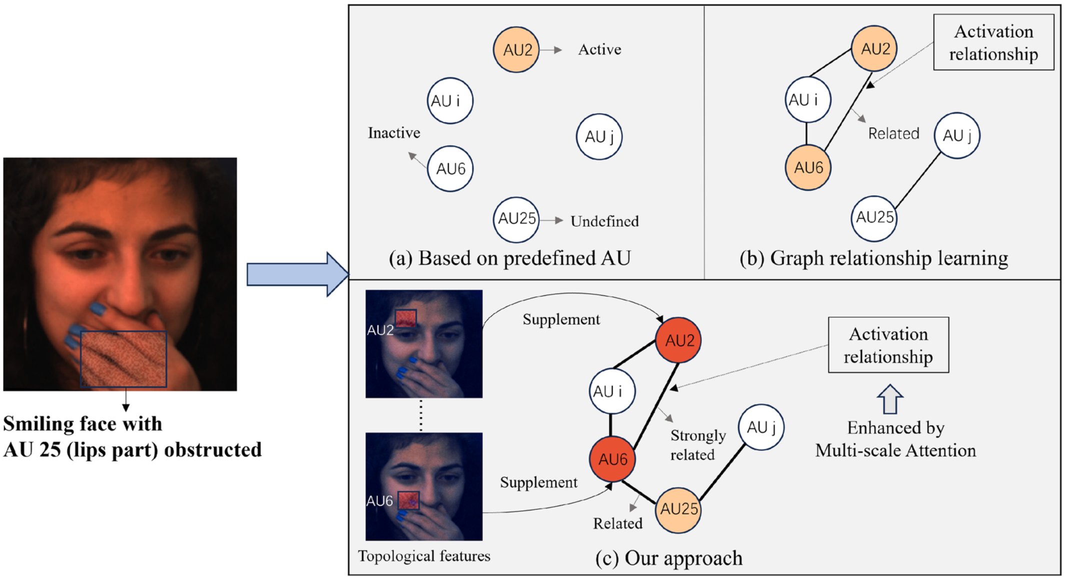


Fig. 1 Comparison of our method with existing methods: deeper colors indicate higher activation intensities. **a** Early deep learning methods directly recognize AUs based on manually defined activation states; in the figure, AU6 (cheek Raiser)and AU25 (lips part)are sometimes not correctly defined as active. **b** Graph relationship learning methods capture the relevance between AU2(outer brow raiser)and AU6 through learned activation relationships, assigning accurate activation weights to AU6. **c** Our method enhances AU activation relationships through multi-scale attention and supplements AU information with topological features, reducing the impact of occluded regions on activation relationships. Based on the spatial and activation relationships learned through our method, a suitable weight is also assigned to the occluded AU25. Just as our daily smiles are usually accompanied by certain lip movements (AU25)

in each sample, which limits their ability to capture the unique characteristics of each AU across different images. These methods do not account for the distinct facial representations of AUs, particularly when they are occluded or undergo dynamic changes. *Problem 2.* Existing techniques in AU relationship learning often overlook the importance of optimizing AU activation states for graph learning. The positive and negative correlations between AUs, influenced by their activation states, play a critical role in their interrelationships. Misidentifying these activation states can severely disrupt relationship learning and negatively impact overall AU recognition performance. *Problem 3.* In scenarios involving occlusions or distorted facial regions, relationship modeling based on incorrect AU data can significantly degrade model performance. This issue arises when AU-related regions are not correctly identified, leading to inevitable information loss unless corrective measures or feature enhancements are applied.

Given the challenges discussed, this paper proposes an AU recognition method that integrates channel-topology convolution with relationship learning, effectively addressing the feature representation of each AU from dual perspectives. This approach encompasses the synergistic movement of each AU and the supplemental information provided by shallow facial features. We propose a channel-topology convolution feature generation structure, named CCFG, which captures the implicit relational features of the facial AU regions spatially and delineates the fine-grained variations caused by facial muscle movements. To address *Problem 1*, this aggregation constructs a topological relationship enabling the integration of local information from various channels into each independently generated AU feature map. The processing of multi-scale features has been proven to effectively increase the receptive field and enhance the initial features [10–13], which potentially benefits the extraction of AU features from facial images. We adopt graph neural networks to learn the relationships among individual AU features. To address *Problem 2*, these AU features are treated as nodes and generated by a multi-scale feature generation module (MAFG) that captures and integrates features across different scale channels. In order to also account for *Problem 1*, this setup focuses on generating each independent AU feature map using multi-scale attention, thereby enabling the graph neural network to more effectively identify the correct AU activation states for relationship learning. Facing potential information loss due to obstructed AU region recognition, as mentioned in *Problem 3*, which impacts AU-related learning, our method suggests concatenating less affected local AU information, which contains spatial relationships, with the corresponding AU representations after relationship learning. This concatenation of CCFG's joint features with AU relationship features achieves feature fusion, balancing the dual-branch information and enabling the model to complement each perspective when handling complex, extreme images. Even if one perspective is limited or assimilates incorrect information, the other can still supplement the necessary information as much as possible, enhancing the overall performance of the model. Figure 1c illustrates how the structure we proposed enhances AU recognition in challenging environments. We employ multi-scale attention to improve the activation states and strengthen the activation relationships in graph learning. Additionally, AU topological features mentioned in Fig. 1c, which captures spatial relationships through topological structures, are adaptively weighted and integrated into the AU representations after relationship learning, thereby achieving the goal of information supplementation.

In summary, the contributions of this paper are as follows:

- We propose a multi-scale attention feature generation structure (MAFG) to enhance AU recognition by generating feature maps at different scales and using graph neural networks for relationship learning.
- We introduce a channel-topology convolution feature generation structure (CCFG) to dynamically aggregate features across different channels, improving AU feature representation.
- We propose a dual-branch feature fusion structure that combines AU relationship learning with topology-based multi-channel aggregation, enhancing performance in complex and occluded scenarios.

Table 1 Summary of representative AU recognition methods from 2022 to 2024

Method	Venue	Category
KDSRL [22]	CVPR 22	Using additional auxiliary information
BG-AU [18]	CVPR 23	
FG-Net [20]	WACV 24	
ME-Graph [16]	IJCAI 22	Relationship learning
FAN-Trans [23]	WACV 23	
SACL [17]	IEEE Trans… 24	
KS [19]	ICCV 23	Knowledge distillation
AUFormer [21]	ECCV 24	Transfer learning

2 Related work

Facial action unit (AU) recognition has seen significant advancements in recent years, with a variety of approaches utilizing deep learning techniques to improve AU detection accuracy. We have listed recent representative AU recognition methods in Table 1. The majority of existing methods treat AU recognition as a multi-label classification problem. Early approaches, such as Ji et al. [4], employed ResNet-34 [14] for initial pre-training on general facial expressions and then fine-tuned the model on AU-specific datasets. Similarly, Shao et al. [5] applied multitask learning strategies, leveraging a convolutional neural network (CNN) [15] to concurrently address facial AU recognition and face alignment. Although these methods have yielded foundational results in facial action unit recognition, they fall short of achieving the high-precision recognition required for more advanced applications.

Region learning approaches in previous research have aimed to enhance the accuracy of AU recognition by exploiting the spatial relationships between facial keypoints and AUs. Jaiswal et al. [6] leveraged facial keypoints to demarcate regions of interest for extracting features specific to each AU. Similarly, Li et al. [7] defined attention-focused areas for each AU to improve feature extraction efficiency, while Ma et al. [8] incorporated object detection tasks by establishing bounding boxes around AUs based on facial keypoints. Despite these advancements, such methods are limited by their dependence on pre-established keypoints, often struggling to accurately identify AUs that deviate from these predefined zones. This limitation becomes particularly evident as AUs can vary dynamically across different images or videos.

Recent advancements in AU recognition have introduced several notable methods. FG-Net [20] enhances AU detection efficiency using StyleGAN2-derived feature maps and a Pyramid CNN Interpreter. KS [19] leverages a lightweight, online semi-supervised framework with Progressive Knowledge Distillation, enabling efficient detection with fewer annotations. Meanwhile, BG-AU [18] employs a biomechanics-guided approach that integrates 3D physics and 2D appearance information to effectively predict AUs with less training data. AUFormer [21] sets a new standard with its parameter-efficient transfer learning and a mixture-of-knowledge expert mechanism, achieving superior AU detection efficiency and minimal data dependency. KDSRL [22] framework utilizes a self-supervised learning approach based on FACS-defined AU rules to efficiently train on unlabeled facial images for AU recognition. FAN-Trans [23] uses a hybrid convolutional and transformer network with online knowledge distillation and an attention mechanism to model AU co-occurrences and dependencies. However, these advanced approaches have not fully utilized the implicit correlations among AUs and have yet to perfectly address the issue of obstructions in AU regions.

Considering the interdependencies among AUs, some methods focus on effectively modeling relational information [16, 17, 24–26]. The interdependencies between AUs significantly influence their recognition. For example, the activation of “Cheek Raiser” (AU6) often impacts the “Lip Corner Puller” (AU12), typically resulting in simultaneous activation (positive correlation), while other AUs might never activate together (negative correlation). Capturing these correlations through relationship learning can enhance AU recognition. Li et al. [9] pioneered the use of graph neural networks for AU relation modeling. Subsequently, Liu et al. [24] utilized Graph Convolutional Networks (GCN) to model AU relationships, improving recognition outcomes. MLCR [27] implemented a multi-label cooperative training strategy via GCN embedding, integrating prior knowledge of AU relationships. MFS [28] introduces a self-supervised facial expression recognition method that leverages a multi-level feature selector and GCNs to learn and refine facial features without labeled data. These methods, however, generally rely on a single feature map for each sample or fixed feature maps embedding prior knowledge, and do not consider the unique facial representations of each AU in images, which can affect their relationships. Luo et al. [16] introduced a method for facial action unit relation graph learning based on multi-dimensional edge features, treating AUs as diverse feature maps for graph learning and using GCN to model these edge features, thereby significantly enhancing AU recognition. Despite achieving promising results in AU recognition, these methods may overlook the importance of ensuring the correct activation states of AUs in relationship learning, which exposes performance issues in recognition when feature extraction is insufficient.

Based on the discussed AU relational graph learning methods, we recognize that each AU involves different facial regions and exhibits unique characteristics and potential muscular variations. Learning pixel-level features for each AU and treating them as independent facial representations may contribute to more refined and differentiated learning of AU activation states. The positive and negative correlations between AUs are influenced by their activation states. Incorrectly defining these activation states can significantly impair the effectiveness of relationship learning. Therefore, ensuring the accuracy of AU activation states is beneficial for correctly modeling the correlations between multiple AUs. However, in specific AU recognition scenarios where activation information is inevitably lost, it is necessary to implement compensatory measures to mitigate the impact of impaired relational learning results.

Our method differs by integrating multi-scale attention and channel-topology convolution, which enables the generation of independent feature maps for each AU that capture multi-scale channel features and topological features involving spatial relationships. This is combined with graph neural networks for relationship learning, which leverages optimized activation states for relational modeling and supplements the results of relationship learning with additional information. Our approach better addresses the limitations of previous methods in relational learning and achieves more robust AU recognition in complex environments.

3 Proposed method

This section describes our AU recognition method that integrates channel-topology convolution with relationship learning, as depicted in Fig. 2. The methodology is structured around four principal components: the backbone network, MAFG, GNN, and CCFG. Initially, a comprehensive facial representation, X, is learned through the backbone network. In the MAFG module, X facilitates the generation of AU feature maps by balancing contextual information via an attention mechanism. These feature maps are then processed by the graph neural network (GNN), which models the relationships between AU activation states end-to-end, resulting in relational AU feature representations v_i^{GNN}. Notably, we propose an innovative approach in the CCFG branch, where X is processed through a channel-topology convolution structure to learn and amalgamate features across channels, producing AU-specific feature representations v_i^{CCFG}. These representations v_i^{CCFG} are combined with the relational features v_i^{GNN} to enhance the final feature representation f, thereby realizing AU recognition.

3.1 Relation learning on AU features

In this section, we detail the MAFG and GNN components of the relational learning structure illustrated in Fig. 2. The approach begins with selective emphasis on feature map generation via a multi-scale attention mechanism. These features are then initialized as graph nodes. Through multiple layers of feature propagation and node state updates in a GNN, the model captures the relational dynamics and dependencies among AUs.

3.1.1 Multi-scale attention for facial generation

In the relationship learning structure, our proposed Multi-scale Attention Feature Generator (MAFG) incorporates multi-scale channel attention and multiple feature extractors to generate independent feature maps for each AU. Multi-scale features are exploitable in AU recognition due to the complex deformations of facial muscles across different facial organs, which vary in size and shape among individuals. The MAFG extracts features at multiple scales, capturing both subtle local variations and broader global information, enabling a more accurate depiction of the AUs’ activation

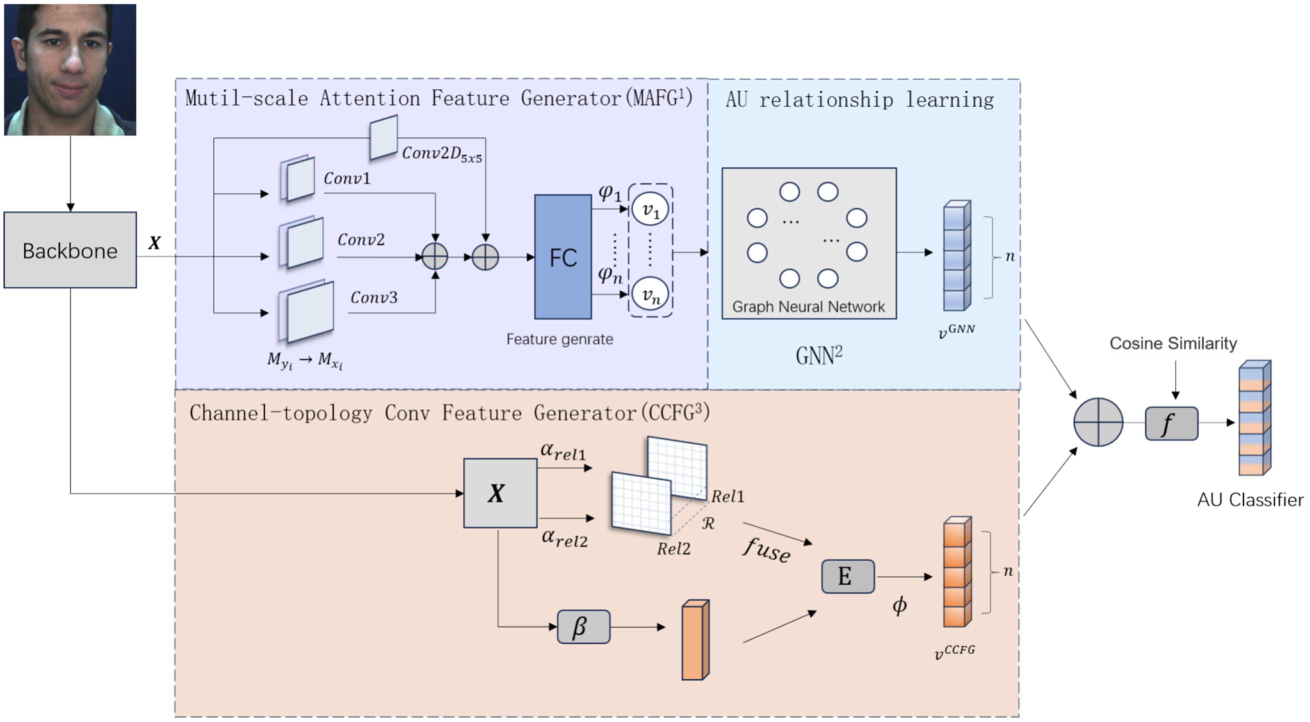


Fig. 2 Overview of our AU detection method, primarily consisting of the backbone network, MAFG[1], GNN[2], and CCFG[3]. Facial images are inputted, and the basic features obtained are enhanced by MAFG to strengthen AU activation representations, with AU activation relationships learned by the GNN. Concurrently, CCFG utilizes a topological convolution structure to aggregate relationship and local features of AUs. Finally, the relational information and channel aggregation information are integrated for AU detection

states. Multi-scale channel features, spanning from detailed to holistic levels, are essential for recognizing spatial structures and shapes. Inspired by SegNext [29] and MCANet [10], our method integrates multi-scale convolutional features from both vertical and horizontal directions using axial attention. This process sequentially aggregates features along these directions, maintaining feature independence, and effectively captures global contextual information. Ultimately, a distinct feature map for each AU is generated from the aggregated multi-scale information. Specifically, the input is a complete facial image, $X \in R^{\mathrm{H}\times\mathrm{W}\times\mathrm{C}}$, with dimensions H × W and C channels, processed through the backbone network ResNet50. X is then encoded in parallel using one-dimensional convolutions at three different kernel sizes and paddings—1×7, 1×11, and 1×21, respectively, known as $\mathrm{Conv1}D_n (n = 0, 1, 2)$, based on configurations from MCANet [10]. The formula can be expressed as follows:

$$M_{y_n} = \mathrm{Conv1}D_n^y(\mathrm{Conv2}D_{5x5}(X)), \tag{1}$$

In this formulation, $\mathrm{Conv1}D_n^y$ denotes the one-dimensional convolution along the y-axis for the given scale. $\mathrm{Conv2}D_{5x5}$ is applied for deep convolution on the input image, serving as the initial step in the multi-scale convolution process. The output from this y-axis convolution at that scale, denoted as M_{y_n}, serves as the input for the subsequent x-axis convolution. The output along the x-axis, denoted as M_{x_n}, is then calculated as follows:

$$M_{x_n} = \mathrm{Conv1}D_n^x\left(M_{y_n}\right), \tag{2}$$

where $\mathrm{Conv1}D_n^x$ represents the one-dimensional convolution along the x-axis for that scale. Given the context information M_{y_n} captured in the vertical direction, the features along the vertical axis are aggregated along the horizontal axis using a convolution kernel of the same size, representing the finalized information for that scale. After processing through deep convolutions of varying sizes, a 1×1 convolution is employed to mix channels and integrate information across different scales. This integration is expressed as follows:

$$\mathrm{Attn} = \mathrm{Conv2}D_{1\times 1}\left(\mathrm{Conv2}D_{5x5}(X) + \sum_{n=0}^{2} M_{x_n}\right), \tag{3}$$

Finally, the original input feature map X is element-wise multiplied with the channel-mixed information, followed by the generation of multiple feature maps. This process implements a multi-scale attention generation mechanism that

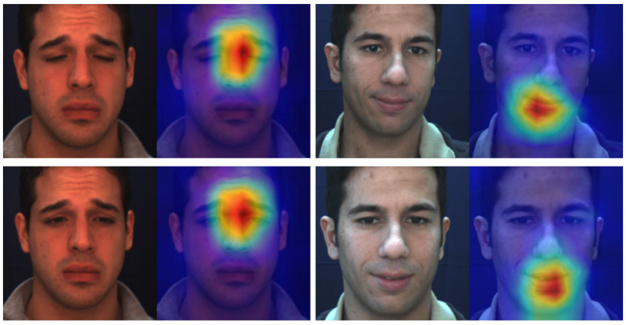

(a) AU 1
(b) AU 12

Fig. 3 Visualization of the attention structures of MAFG using Grad-CAM [30], where **a** corresponds to AU 1, focusing on the feature of raising the inner corners of the eyebrows; **b** corresponds to AU 12, focusing on the feature of pulling the corners of the mouth upward. By focusing on independent feature maps for each type of AU, MAFG is able to provide more reliable AU feature representations for relationship learning

specifically focuses on generating multiple feature maps, as described in the following equation:

$$V = \text{Cat}(\varphi_i(\text{Attn} * X)),\ i = 1, 2, \ldots, \text{Num}, \tag{4}$$

where Num represents the number of target AU categories. In this study, based on the label format of the datasets, Num is set to 8 in the DISFA dataset and 12 in the EmotioNet dataset. The features processed with multi-scale attention are independently learned through Num linear blocks φ, each comprising a fully connected layer (FC) and a normalization layer. This method produces vectors $v_i (i = 1,2, \ldots, \text{Num})$ in V, which include channel feature values C and represent the feature map for the i-th AU. Furthermore, all AU feature maps are aggregated into a unified feature representation V. By extracting multi-scale features from the original feature map and mixing channels, feature maps for each AU category are generated, thereby enhancing the initial representation for the relationship modeling phase. We describe through Fig. 3 the multi-scale attention to the complete feature representation X when the current AU is activated.

3.1.2 Graph neural networks for modeling AU relationships

Graph neural networks (GNNs) exhibit exceptional capabilities in modeling relationships due to their structured network of nodes and edges, providing substantial benefits in managing complex interdependencies. In the context of recognizing AU, GNNs adeptly model the mutual influences and associations among various AUs, including their long-distance latent connections. Drawing from existing GNN frameworks [31], our GNN architecture employs linear transformations between adjacent nodes, where each AU feature map representation is treated as a node. Edges between nodes are dynamically generated by computing correlations among nodes, thus capturing the associations between features. Specifically, we construct a dynamic graph for each AU's feature map using the dot product to measure similarity, resulting in a similarity matrix S, where each element $S_{i,j}$ represents the dot product similarity between the feature vectors of nodes i and j $(i, j \leq \text{Num})$. For each node, we select the top K values from the i-th row of S to identify its K nearest neighbors, where K is set to a maximum of 4 in this study. These selections form an adjacency matrix A, where $A_{i,j}$ is set to 1 if node j is among the top K nearest neighbors of node i, and 0 otherwise. The adjacency matrix A represents the graph structure by indicating the connections between each AU node and its closest counterparts. We then normalize A to obtain the normalized adjacency matrix A', which is used in the subsequent GNN process. The computation formula can be represented as follows:

$$S = \rho(V)^T \rho(V), \tag{5}$$

where $\rho(V)$ is the projected feature matrix of the feature representation V, which has a shape of Num × C in each sample, and maps the generated feature space to a new space through a dot product operation. This matrix S is a square matrix of size Num × Num, where Num is the total number of nodes in the graph, and represents the number of target AU categories. The element $S_{i,j}$ indicates the similarity between node i and node j. In summary, this similarity matrix S encodes the relationships between all pairs of nodes, where larger values in $S_{i,j}$ indicate a stronger similarity between the corresponding nodes. Next, we construct the adjacency matrix A based on the similarity matrix S. For each node i, we identify the top K values in the i-th row of S, denoted as S_i. These top K values correspond to the highest similarity values between node i and other nodes in the graph. We then set the adjacency matrix element $A_{i,j}$ to 1 if node j is one of the top K nodes with the highest similarity values in S_i, and 0 otherwise. This operation can be written as:

$$A_{i,j} = \begin{cases} 1 \text{ if } S_{i,j} \text{ is one of the top } K \text{ highest values in } S_i \\ 0 \text{ otherwise} \end{cases}, \tag{6}$$

This adjacency matrix A represents the graph's structure, where nodes i and j are connected if they are among each other's top K nearest neighbors according to the similarity matrix S. To avoid numerical instability during the training process, the adjacency matrix A is normalized row-wise. The degree matrix D is computed as:

$$D_{ii} = \sum_{j=1}^{\text{Num}} A_{ij},\ i = 1, 2, \ldots, \text{Num}, \tag{7}$$

where D_{ii} represents the degree of node i, which is the number of nodes connected to node i in A. The adjacency matrix A is then normalized as follows:

$$A' = D^{-\frac{1}{2}} A D^{-\frac{1}{2}}, \tag{8}$$

where A' is the normalized adjacency matrix. This normalization helps to stabilize the training process by mitigating the influence of node degrees. Finally, the facial representation for each AU category v_i (i = 1,2, …, Num) and its activation representation updated through the GNN, denoted v_i^{GNN}, are expressed as follows:

$$v_i^{\text{GNN}} = \sigma\left[v_i + g\left(v_i, \sum_{j=1}^{\text{Num}} r\left(v_j, a_{i,j}\right)\right)\right], \tag{9}$$

where σ is a nonlinear activation function, g is a node information update function, Num represents the total number of nodes in the graph, $a_{i,j}$ is an element of the normalized adjacency matrix A', indicating the relevance between nodes i and$j$, and r is a function that reweights the features of v_j based on the adjacency matrix element$a_{i,j}$.

In summary, by representing AU feature maps as nodes, the GNN employs a multi-layer propagation strategy that enables each node to aggregate information from its nearest neighbors. Subsequently, it iteratively learns to integrate potential long-distance neighbor relationships. During each update in the propagation process, the node v_i aggregates features from neighboring nodes v_j, updating its state through the nonlinear activation function and learned weights. This iterative execution ensures that each node's features fully reflect the common characteristics and distinctions pertinent to all associated AUs. Thus, the GNN learns the dynamic and complex relationships between multiple AUs. The updated result v_i^{GNN} for each AU node serves as the feature representation for graph relation learning.

3.2 Multi-channel topological feature fusion

This paper addresses complex facial action unit (AU) recognition scenarios hindered by manual setups such as foreign object obstructions, non-frontal face orientations, and color confusion. Traditional methods reliant on prior knowledge, including manual matrix definitions and keypoint approaches, often fail to adapt comprehensively to these scenarios. In such settings, adaptive learning of relationships may incorrectly propagate information to adjacent nodes. This section proposes that network structures based on relationship learning need to adapt and complement information from another approach. We envision enhancing local features for each AU class through as much relationship modeling information as possible, guiding the correct propagation of AU activation states.

3.2.1 Channel-topology convolution for joint feature map generation

We propose a branched parallel structure, as illustrated in Fig. 2, where AU feature relationship learning serves as the primary perspective, and a secondary perspective utilizes the CCFG to jointly generate feature maps. This enhances the feature representation from the primary perspective through adaptive feature concatenation. Additionally, through Fig. 4, we are able to illustrate this process. Channel-topology structures, proven effective in complex scenario recognitions such as skeletal posture identification [32], can dynamically learn and effectively aggregate features across various channels, encompassing local relational and geometric features. We assume that the regions involved in each action unit (AU) can form a topological structure based on their relative spatial positions, which is largely consistent across different individuals' faces. By leveraging clues from this topological structure, we are able to additionally capture the AU activation information pertaining to corresponding regions. Consequently, we build a non-shared topology structure tailored to the implicit spatial relationships and local information of AUs. This involves multiple defined linear transformation layers processing the aggregated features from the topology convolution, creating independent feature maps for each AU learned from a topological perspective, thus making this topology convolution structure effectively applicable to the multi-label classification task of AU recognition. Unlike traditional graph relationship learning structures that share information, our approach learns relationship information in a non-shared manner, allowing the model to independently obtain AU feature representations containing latent information, addressing environmental challenges where AU feature recognition is impeded. We regard the multi-feature representation output of CCFG as joint feature, adapted to the unique characteristics of AU representations. Specifically, this method first transforms the input feature map X through three one-dimensional convolutional layers, combining relationship and geometric features into a topological relationship to emphasize the channel features of AUs. Structurally, linear transformation functions composed of convolutions, α_{rel1} and α_{rel2}, are defined to extract relationship features, while β_{local} captures local muscle granularity and contour information, performing feature transformation. Features extracted through α_{rel1} and α_{rel2} are subtracted to calculate the relational features between channels, and nonlinear transformations are performed using the tanh activation function to capture the topological structure of features. The formula is as follows:

$$X_{\text{rel}} = \tanh(\alpha_{\text{rel1}}(X) \ominus \alpha_{\text{rel2}}(X)), \tag{10}$$

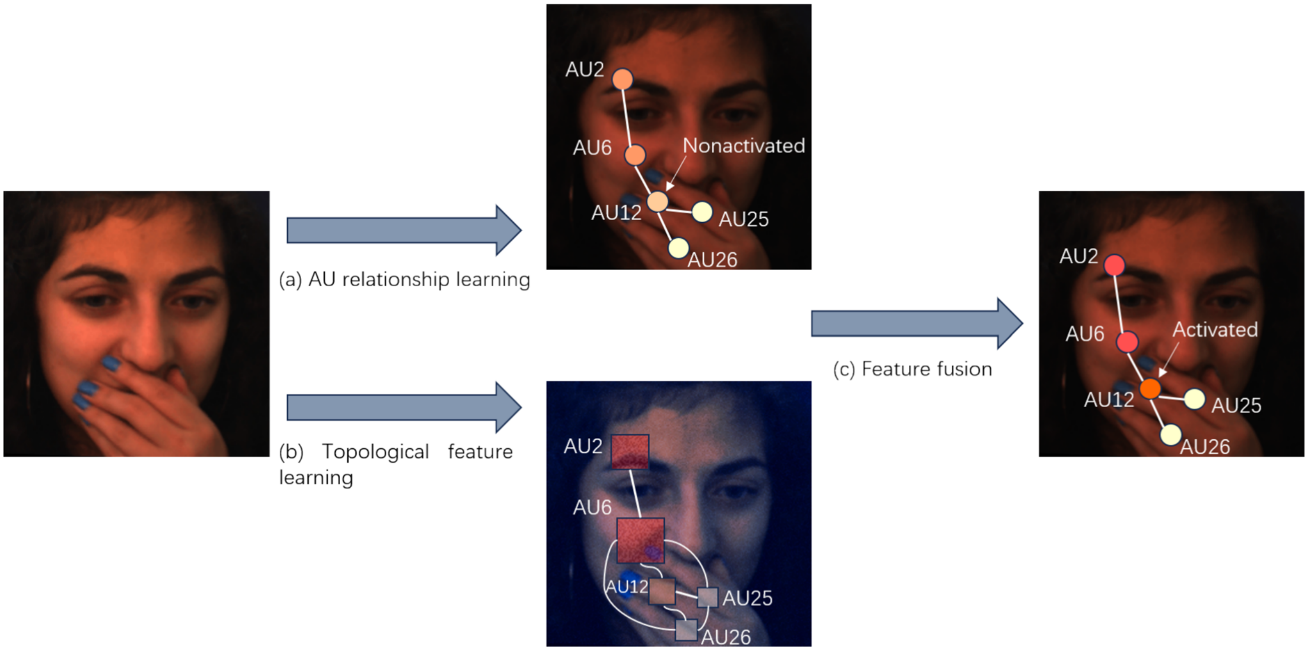


Fig. 4 Visualization of the Structural Impact of CCFG: **a** When some AUs are located in obscured regions, the adjacency relationships established by the current AU graph structure may interfere with the activation representation of unobscured AUs, potentially leading to lower activation scores. **b** The CCFG module improve this by enhancing the activation representation of unobscured AUs and capturing the local muscle information of discernible AUs, based on channel features and relative positional information. **c** This information is then integrated into the graph relation learning stage, which is influenced by the features of recognition-obstructed AUs

Here, $\ominus$ denotes element-wise subtraction between the features. These relational features X_{rel} are then processed through a convolutional layer $\text{Conv}1D_f$, and scaled by an adjustment coefficient λ. They are subsequently integrated with local features β_{local} using the Einstein summation convention, which facilitates the integration of the topological structure. This approach allows each channel's features to complement each other, culminating in a comprehensive topological feature as the output. The formula for this integration is expressed as:

$$E = \sum_{i=1}^{C} \text{Conv}_f(X_{\text{rel}}) \cdot \lambda \ominus \beta_{\text{local}}(X), \tag{11}$$

In this expression, Conv_f represents the convolutional layer used to fuse the relational features, and C denotes the number of channels. Finally, the topological feature E is processed through a linear block ϕ, which consists of an average pooling layer followed by a fully connected layer. This configuration integrates the channel features corresponding to each class of AU feature maps and allows for the learning of additional weights. The feature representation for each AU class is considered independent and sequentially corresponds to our feature representations learned through graph neural networks, v_i^{GNN}. Unlike the adjacency relationships inherent in graph feature representations, we treat the multi-class, completely independent AU features—or joint features—as the final output, v_i^{CCFG}. This output provides the feature representation for the i-th AU, which is expressed as:

$$v_i^{\text{CCFG}} = \phi_i(E),\ i = 1, 2, \ldots, \text{Num}, \tag{12}$$

3.2.2 AU feature fusion and prediction

We have derived AU feature representations through distinct processing methods via GNN and CCFG. For the relationally modeled feature representation, v_i^{GNN}, and the feature representation that integrates local information with a topological structure, v_i^{CCFG}, our objective is to enhance model performance and generalizability. This is achieved through a feature concatenation fusion strategy, which merges feature representations from both branches. A fully connected layer is employed to linearly transform the fused feature tensor, adjusting the channel count to meet the model's output requirements and learning the weights between each output and input channel. Subsequently, the information in each channel undergoes normalization and nonlinear transformations through a Post layer that incorporates batch normalization and ReLU activation functions, resulting in the final AU activation representation f. The formula is as follows:

$$f_i = \mathrm{Post}\big(\mathrm{Concat}\big(v_i^{\mathrm{GNN}}, v_i^{\mathrm{CCFG}}\big)\big), i = 1, 2, \ldots, \mathrm{Num},$$

This paper employs a similarity calculation strategy, as described in reference [16], to predict the i-th AU. This involves a trainable vector s_i that matches the dimension of f_i. The occurrence probability of the i-th AU is obtained by calculating the cosine similarity between f_i and s_i:

$$p_i = \frac{\mathrm{ReLU}(f_i)^T \mathrm{ReLU}(s_i)}{\mathrm{ReLU}(f_i)_2 \mathrm{ReLU}(s_i)_2}, i = 1, 2, \ldots, \mathrm{Num}, \tag{14}$$

Here, ReLU denotes a nonlinear activation function. In summary, our approach effectively balances the strong associative information between AUs and the channel information aggregated from a topological structure for each category of AU output p_i.

The main learning component of the network focuses on the AU activation status and its associated node features for each facial expression. Notably, as a multi-label task, there is a marked imbalance in the labels within existing AU datasets—some AUs occur infrequently, while others are typically inactive in most facial images. To address this imbalance, this paper, drawing on reference [33], introduces a weighted asymmetric loss formula. This formula calculates the loss between the actual values generated by the facial graph generator module and their predictions. It assigns unique weights to the recognition of each AU, thereby reducing the reliance on hyperparameters. Specifically, the weighted asymmetric loss formula is defined as:

$$L_{\mathrm{WA}} = -\frac{1}{\mathrm{Num}} \sum_{i=1}^{\mathrm{Num}} w_i \big[y_i \log(p_i) + (1 - y_i) p_i \log(1 - p_i)\big], \tag{15}$$

In this formula, p_i, y_i, and w_i, respectively, represent the predicted probability, actual value, and weight of the i-th AU. The weight w_i is defined as:

$$w_i = \frac{\mathrm{Num} * \left(\frac{1}{r_i}\right)}{\sum_{j=1}^{N} \left(\frac{1}{r_j}\right)}, i = 1, 2, \ldots, \mathrm{Num}, \tag{16}$$

where r_i is the occurrence rate of the i-th AU calculated from the training set. This weighting ensures that AUs with a higher occurrence rate have a reduced impact on the loss value, prioritizing the loss caused by less frequently occurring AUs. Moreover, for the loss term of inactive AUs, $(1 - y_i) log(1 - p_i)$, when p_i is low, the loss is minimized, which helps in reducing false positives where inactive AUs are incorrectly predicted as active.

The weight w_i, defined in Eq. (16), adjusts based on the occurrence frequency of each AU in the training set, making the model place more emphasis on rare AUs during the learning process. This weighting strategy leads to increased recall for these less common AUs, which in turn boosts the overall F1 score that we set as evaluation metrics. By applying the weighted asymmetric loss defined in Eqs. (15) and (16), we are able to effectively improve the label imbalance issue present in AU recognition tasks. The weighting mechanism ensures that less frequently occurring AUs receive higher importance during training, thereby improving the model's focus on their accurate detection. Consequently, the overall loss of our proposed method can be expressed as:

$$L = L_{\mathrm{WA}} \tag{17}$$

4 Experimental results

4.1 Datasets

This study evaluates the method's performance using widely recognized benchmark datasets. DISFA, as cited in references [34] and [35], is a facial action unit (AU) recognition dataset obtained in a controlled setting. It utilizes dual cameras positioned on the left and right, capturing 4845 frames from 27 participants (12 females and 15 males) who viewed videos designed to elicit seven expressions: anger, disgust, surprise, fear, sadness, happiness, and neutral. Each frame is annotated with the presence of multiple AUs, totaling 130,815 facial images. Additionally, experiments were conducted on EmotioNet, a challenging dataset cited in reference [36]. EmotioNet comprises over 20,000 facial images from uncontrolled natural environments, collected from the internet. These images feature diverse backgrounds, complex angles, and instances of human obstructions. Each image in EmotioNet has been manually annotated by experts for 12 AUs, providing a robust dataset for testing the method under varied conditions.

4.2 Experimental setup and evaluation metrics

The facial action unit recognition method presented in this paper infers the scores of the target AUs from a single image. To minimize the impact of the background environment on recognition accuracy, this study employs MTCNN [37], a widely used tool in facial expression recognition, for face detection and alignment. Subsequently, the aligned facial images are randomly cropped to a size of 224 × 224 pixels to serve as input for the network.

We utilize ResNet50 [14] as the backbone network, which is composed of convolutional layers, pooling layers, and bottleneck convolutional blocks. ResNet50 has been widely employed in previous AU detection methods, as documented in references [38] and [16]. For this research, we deploy a ResNet50 model pre-trained on ImageNet [39], which enhances the model's generalization capability and expedites the training process.

In this study, following the methodologies of previous researchers [7, 16, 40, 41], we employed three-fold cross-validation on each dataset and reported the average results across the three folds. During the training process, the AdamW optimizer was utilized with $\beta 1 = 0.9$, $\beta 2 = 0.999$, and a weight decay of 5e-4. For the dataset DISFA, the number of nearest neighbors, K, for facial graph node learning was set to 3, while for EmotioNet, K was set to 4. The experiments entailed 40 epochs of model training, starting with an initial learning rate of $1e - 4$ and a batch size of 16. We adopted a cosine decay learning rate scheduler to more effectively facilitate convergence to the optimal solution of the model. All experiments were conducted GTX 3080 Ti GPU within the PyTorch deep learning framework.

In terms of evaluation metrics, this study follows the standards of previous AU event recognition research [5, 16, 17, 41], utilizing frame-based F1 scores and average accuracy (Acc) to assess model performance. The F1 score is defined by the formula F1 = 2*P*R/(P + R) where P represents precision and R represents recall. Precision measures the proportion of true positive samples among all samples classified as positive by the model, and is defined as P = TP/(TP + FP), where TP (true positives) represents the number of correctly predicted active AUs, and FP (false positives) represents the number of incorrectly predicted active AUs. Recall measures the proportion of actual positive samples that are correctly identified as positive by the model, and is defined as R = TP/(TP + FN), where FN (false negatives) represents the number of active AUs that the model failed to detect.

In this study, these metrics are calculated for each AU independently based on the model's predictions. Specifically, for each action unit, we determine the number of true positives, false positives, and false negatives using the predicted activation status and the ground truth labels from the dataset. Subsequently, the precision and recall values are computed for each AU, and these are then used to derive the frame-based F1 score. After computing the F1 score for each individual AU, we then calculate the overall mean F1 score by averaging the F1 scores across all AUs to measure model's performance.

4.3 Comparative experiments and analysis

Comparison with State-of-the-Art Methods This section compares the method introduced in this paper with state-of-the-art methods on two datasets. For the DISFA dataset, the state-of-the-art methods evaluated include DRML [40], EAC-Net [7], JAA-Net [42], ARL [43], SEV-Net [25], FAUDT [44], UGN-B [26], HMP-PS [41], SACL [17], ME-Graph [16], BG-AU [18], KS [19], KDSRL [22], FG-Net [20], FAN-Trans [23] and AUFormer [21]. For the EmotioNet dataset, the methods compared are ResNet-34 [14], the backbone network ResNet-50 [14], MLCT [45], Mean-teacher [46], Co-training [47], MLCR [27], and ME-Graph [16]. We have cited partial experimental data from these studies conducted on the same dataset. To ensure a fair comparison, the experiments presented in this paper did not omit any frames from either dataset.

4.3.1 Evaluation on DISFA

Table 2 reports the recognition results for 8 AUs on the DISFA dataset. Our proposed method achieves higher F1 scores compared to all other listed methods, with an average improvement of 0.2% over the previous state-of-the-art method AUFormer. Specifically, our method reaches an average F1 score of 66.6%, demonstrating its effectiveness.

Compared to non-relationship learning methods, our approach shows a significant advantage, improving the F1 score by 7.9% over the most advanced AU region learning method. This substantial improvement is attributed to our MAFG structure, which enhances AU recognition by generating feature maps at different scales and employing graph neural networks for relationship learning. The MAFG allows the model to capture features at multiple resolutions, providing a more comprehensive representation of facial expressions.

When compared to leading graph learning methods such as ME-Graph and SACL, our method exhibits superior performance with an increase in F1 scores of 1.1%. While ME-Graph incorporates the relationship information of independent feature maps for each AU and adds extra relational clues, it does not account for the implicit local information within each AU's feature maps. Our CCFG structure addresses this limitation by dynamically aggregating features across different channels, improving AU feature representation through the exploitation of local topological structures.

KDSRL, FG-Net and AUFormer define additional landmarks and synergy mechanisms to assist in AU detection but require extra auxiliary data. In contrast, our method introduces a dual-branch feature fusion structure that combines AU relationship learning with topology-based multi-channel aggregation without the need for additional data. This fusion enhances performance in complex and occluded scenarios by effectively integrating global relational information with local topological features.

FAN-Trans, which learns AU relationships without using a graph model, does not show better performance. SACL leverages multi-scale features to enrich relational information but fails to provide independent feature representations for each AU class that integrate with the corresponding AU

Table 2 Comparison to state-of-the-art methods on DISFA dataset with 8 AUs (expressed in %)

Method	AU								Avg
	1	2	4	6	9	12	25	26	
Region learning									
DRML [40]	17.3	17.7	37.4	29.0	10.7	37.7	38.5	20.1	26.7
EAC-Net [7]	41.5	26.4	66.4	50.7	**80.5**	**89.3**	88.9	15.6	48.5
JAA-Net [42]	43.7	46.2	56.0	41.4	44.7	69.6	88.3	58.4	56.0
ARL [43]	43.9	42.1	63.6	41.8	40.0	76.2	95.2	66.8	58.7
Other									
BG-AU [18]	41.5	44.9	60.3	51.5	50.3	70.4	91.3	55.3	58.2
KS [19]	53.8	59.9	69.2	54.2	50.8	75.8	92.2	46.8	62.8
KDSRL [22]	60.4	59.2	67.5	52.7	51.5	76.1	91.3	57.7	64.5
FG-Net [20]	–	–	–	–	–	–	–	–	65.4
AUFormer [21]	–	–	–	–	–	–	–	–	66.4
Relationship learning									
SEV-Net [25]	55.3	53.1	61.5	53.6	38.2	71.6	**95.7**	41.5	58.8
UGN-B [26]	43.3	48.1	63.4	49.5	48.2	72.9	90.8	59.0	60.0
HMP-PS [41]	38.0	45.9	65.2	50.9	50.8	76.9	93.3	**67.6**	61.0
FAUDT [44]	46.1	48.6	72.8	**56.7**	50.0	72.1	90.8	55.4	61.5
ME-Graph [16]	54.6	47.1	72.9	54.0	55.7	76.7	91.1	53.0	63.1
FAN-Trans [23]	56.4	50.2	68.6	49.2	57.6	75.6	93.6	58.8	63.8
SACL [17]	62.0	**65.7**	**74.5**	53.2	43.1	76.9	95.6	53.1	65.5
Ours	**63.6**	59.5	72.1	55.2	56.0	76.2	92.32	58.29	**66.6**

The best and second best results of each column are indicated with bold font and underline, respectively. '–' indicates that the method did not report individual F1 scores for AUs. We categorize all methods into region learning, relationship learning, and other

relational information. Our method's combination of MAFG and CCFG modules effectively complements the feature representations obtained through graph relation learning. The MAFG enhances robust activation representations, while the CCFG exploits the local feature information of each AU via its topological structure. This synergy leads to improved recognition performance and demonstrates the effectiveness of our proposed structures.

4.3.2 Evaluation on EmotioNet

According to Table 3, our method also achieves commendable results on the EmotioNet dataset, effectively addressing the challenges of uncontrolled environments. Our method attains an average F1 score of 67.2%, within 0.9% of the best results from the information embedding method MLCR [27], but shows an improvement of 1.1% over the most advanced graph learning methods.

Given the infrequent application of graph relation learning on EmotioNet, we replicated and compared the advanced AU relationship learning method ME-Graph [16] under consistent settings, utilizing the same backbone network, ResNet-50, for feature extraction. This comparison demonstrates the significant enhancement our method brings to relationship learning.

Our method effectively leverages the topological channel features of the branches to capture local angle information through the CCFG. This capability is crucial in addressing extreme cases within complex datasets and integrates seamlessly into the potentially impaired relation learning information for essential feature compensation. The MAFG contributes by generating feature maps at different scales, which is particularly beneficial in handling the variability present in large-scale datasets like EmotioNet.

While the best information embedding methods have demonstrated good adaptability on the EmotioNet dataset, they often rely on extensive auxiliary information. In contrast, our dual-branch feature fusion structure enhances performance without requiring additional data, highlighting the robustness and generalization capability of our approach.

The graph learning method ME-Graph, although robust, is somewhat limited by the dataset's specificity. Its relational clues might rely on AU labels that are ambiguously defined in scenarios with occlusions, leading to a loss of effective relational information. Our method overcomes this limitation

Table 3 Comparison to state-of-the-art methods on EmotioNet Dataset with 12 AUs (expressed in %)

Method	AU												Avg
	1	2	4	5	6	9	12	17	20	25	26	43	
Other													
ResNet-34 [14]	55.6	41.1	70.1	46.6	80.2	59.2	90.7	45.5	45.1	94.2	58.7	60.5	62.3
ResNet-50 [14]	56.6	46.7	66.8	51.7	75.4	61.6	85.1	44.0	52.6	89.6	54.9	64.1	62.5
MLCT[45]	57.8	44.8	73.7	50.1	82.8	58.1	91.8	44.8	37.1	95.1	61.6	63.4	63.4
Mean-teacher [46]	55.5	46.3	71.1	48.6	81.6	61.7	91.0	46.7	43.5	94.7	60.2	63.9	63.7
Co-training [47]	58.3	48.4	70.0	50.4	83.1	64.4	91.7	49.9	47.1	95.0	60.0	66.9	65.5
MLCR [27]	61.4	**49.3**	**75.9**	54.1	**83.5**	**68.3**	**92.0**	50.8	**53.5**	**95.2**	**65.1**	68.1	**68.1**
Relationship learning													
ME-Graph [16]	65.5	47.4	73.4	53.3	79.3	73.0	86.9	47.8	46.0	93.0	61.2	65.8	66.1
Ours	**66.0**	49.1	73.4	**54.9**	80.2	67.5	87.2	**53.7**	52.4	92.6	61.0	**68.9**	67.2

Table 4 Ablation experiments on the DISFA dataset

Methods	F1	Acc
Resnet50	59.1	90.5
Resnet50+Evit [48]	61.7	91.3
Resnet50+MAFG^1	63.0	92.2
Baseline(Resnet50+GNN^2)	62.9	91.8
Baseline+MAFG^1	63.7	92.7
Baseline+CCFG^3	64.9	93.4
Baseline+MAFG^1+CCFG^3	66.6	94.1

F1 and Acc average scores across 8 AUs are used to evaluate the structures MAFG^1, GNN^2, and CCFG^3 as shown in Fig. 2. Additionally, the multi-scale attention structure, Evit [48], was introduced to validate the performance of MAFG

Table 5 Ablation experiments on the EmotioNet dataset

Methods	F1	Acc
Resnet50	62.5	90.5
Baseline(Resnet50+GNN^2)	63.3	91.8
Baseline+MAFG^1	64.8	94.2
Baseline+CCFG^3	65.9	94.8
Baseline+MAFG^1+CCFG^3	67.2	95.6

F1 and Acc average scores across 12 AUs are used to evaluate the structures MAFG^1, GNN^2, and CCFG^3 as shown in Fig. 2

by combining AU relationship learning with topology-based multi-channel aggregation, thereby enhancing performance in complex and occluded scenarios.

These findings underscore the effectiveness of our proposed structures—MAFG, CCFG, and the dual-branch feature fusion—in improving AU recognition performance across various challenging scenarios. By integrating multi-scale attention features with dynamic channel-topology aggregation, our method demonstrates advancements over existing approaches.

4.4 Ablation study

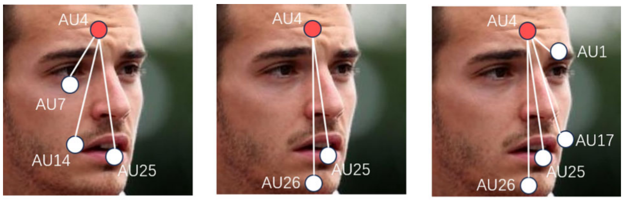
4.4.1 Model structure evaluation

This section evaluates each component of our approach, using a ResNet50 concatenated with GNN as the baseline. Tables 4 and 5 illustrate the impact of each structural component on the average F1 scores and average Acc scores for AU recognition, where 'Acc' refers to recognition accuracy.

Notably, the integration of the MAFG module enhances the performance of the baseline network. This is indicated by the Baseline+MAFG configuration, which captures the detailed layers of facial AUs from a multi-scale perspective, demonstrating its efficacy in recognizing AU activation states as illustrated in Fig. 3. To ensure that the features processed by MAFG are sufficient for classification, we conducted experiments using only the backbone network combined with MAFG (ResNet50+MAFG) for prediction, maintaining the same experimental parameters as in our main experiments. As shown in Table 4, the AU feature classification performance of MAFG with the backbone network is consistent with that of the Baseline, aligning with our assertion that MAFG enhances the recognition of AU activation states. To further demonstrate the advantages of our multi-scale features, we incorporated an efficient multi-scale linear attention structure (EVit) [48] into our experiments for comparison. EVit has demonstrated advantages in global receptive fields and multi-scale learning in the literature. We also set up experiments combining EVit with the backbone network

ResNet50 to leverage its global capabilities, using the same scale parameters as specified in the original paper. As presented in Table 4, our MAFG structure outperforms EVit by 1.3% in F1 score, indicating that for multi-scale attention mechanisms, not only structural differences but also the scale settings vary significantly across different target tasks. The scale settings of EVit (3 × 3 and 5 × 5 with special Conv) may be more suitable for high-resolution target tasks, while the scales of MAFG are more appropriate for AU recognition. This consistency with our methodological details demonstrates that MAFG exhibits superior performance in our AU recognition task. To ensure the stability of the experimental data, these multi-scale feature experiments were conducted only on the DISFA dataset, which was collected under controlled conditions.

Similarly, employing Baseline+CCFG to aggregate AU channel features with a topological structure and integrate this with relational modeling from the graph neural network has proven highly effective. This approach significantly enhances the model's capacity to supplement information about AU local features. In graph relation learning models, the activation of one AU is often linked to the activation of others. However, in complex scenarios involving facial shifts or occlusions, the activation state of a specific AU may not be accurately recognized, which can also misrepresent the activation states of related AUs. This issue may lead to incorrect relation modeling in graph learning networks, affecting the robustness of AU recognition. By focusing on AU region features through CCFG, our method mitigates such informational errors. For instance, as shown in Fig. 4, AU6 (cheek raiser) typically activates simultaneously with AU12 (lip corner puller upward)—a smiling action—and is also slightly more activated in relation to AU2 (outer brow raiser), AU25 (lips part), and AU26 (jaw drop). However, if the regions containing AU12, AU25, and AU26 are occluded, their activation states are nearly ignored. The adjacency relationship weights established by the current graph structure might interfere with the activation representation of AU6, even reducing its activation score. Such misrepresentations of AU activation states may be incorrectly propagated through adjacent relationships. CCFG independently enhances the activation representations of unobscured AUs such as AU6 and captures partial local muscle information of AU12 despite occlusions. Similarly, we also assign a suitable weight to the occluded AU25 based on the spatial and activation relationships. Subsequently, CCFG supplements these features into the affected feature representations from graph relational learning by feature fusion. CCFG's focus on supplementary local features ensures the authentic activation states of related AUs in unobscured regions as much as possible, even when the activation status in occluded areas is compromised. Moreover, by aggregating multiple channel relational features and local features, relation from CCFG places emphasis on the relative positions of key regions between channels than on different graph relations, thus enhancing sensitivity to local edges and contours.



(a) Predefined AU (b) Correlation learning (c) Ours

Fig. 5 Visualization of AU relational information obtained by different methods, showing the relational information for AU4, which has the highest feature representation score in the image: **a** AU correlations obtained based on predefined rules, which typically rely on statistical rules defined by frequently occurring AUs in the data; **b** AU association graph obtained through relationship learning methods; **c** relationship learning method with feature supplementation proposed by us. Compared to existing methods, our approach obtains more AU information and establishes connections even in cases of facial feature asymmetry

The final results demonstrate that the simultaneous application of MAFG for enhanced relationship learning and CCFG for feature supplementation significantly improves AU recognition performance. As depicted in Fig. 5, our method effectively utilizes feature supplementation to extract more AU information in environments where recognition is obstructed, such as unaligned facial images, thereby forming a more comprehensive AU relation graph. These findings underscore two key points: (I) the integration of attention mechanisms through MAFG aids GNNs in learning more accurate AU activation representations, serving as initial inputs for relation modeling; (II) the feature fusion structure of CCFG allows for the supplementation of potentially missing relational information from the main pathway with local feature representations from another perspective.

In summary, compared to the baseline networks, our method achieved an increase of 3.7% and 3.9% in F1 scores on the DISFA and EmotioNet datasets, respectively, and an improvement of 2.3% and 3.8% in accuracy (Acc). Additionally, compared to the backbone network ResNet50, the F1 scores and accuracy were respectively enhanced by 7.5%, 4.7%, and 3.6%, 5.1%. These enhancements clearly demonstrate the effectiveness of our approach in handling complex AU recognition scenarios.

4.4.2 Evaluation of important parameter settings

In this study, the optimal number of nearest neighbors, K, for the facial action unit (AU) relation graph was determined based on its effectiveness in capturing AU dependencies. As indicated in Fig. 6, setting K to 4 yields the highest average F1 score on the EmotioNet validation set; conversely, setting K to 3 results in the best average F1 score on the DISFA validation set. This variation suggests that moderate local

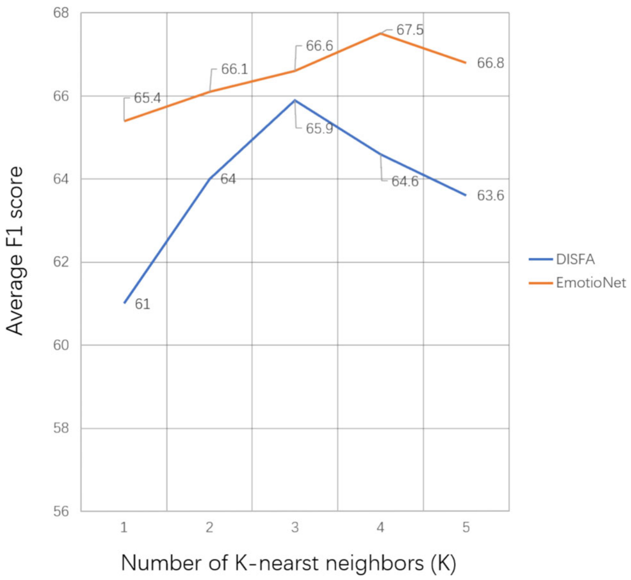


Fig. 6 Average F1 scores corresponding to different numbers of K nearest neighbors on DISFA and EmotioNet

connectivity is crucial in effectively capturing the dependencies among AUs, thereby optimizing the graph learning's expressive capacity for relationship modeling.

For the two datasets, when the value of K is reduced below the optimal setting, a significant decline in F1 scores is observed. A smaller K value restricts the exchange of information between nodes in the graph, resulting in the loss of key implicit relational information and negatively impacting the overall performance of the AU relation graph. Conversely, increasing K beyond the optimal number also leads to a decline in F1 scores. Larger K values may introduce redundant and irrelevant connections, incorporating noise that interferes with the model's ability to learn important AU relationships effectively. The need for a larger K value in EmotioNet compared to DISFA is attributed to the greater complexity of scenarios in EmotioNet, where AU recognition is often obstructed. In such environments, more relational clues are necessary to achieve optimal relation modeling among AUs than in the more controlled settings of DISFA.

Selecting the appropriate K value is crucial for constructing an effective AU relation graph. This study determines that $K = 3$ and $K = 4$ are the optimal nearest neighbor settings for the DISFA and EmotioNet datasets, respectively. These settings balance the integrity of information transmission with the risk of overfitting, ensuring that the model captures essential relational information without incorporating excessive noise that could impair learning.

5 Conclusion

This paper proposes an advanced AU recognition method that integrates channel-topology convolution with relationship learning through an innovative dual-branch feature generation structure. Specifically, we introduced the MAFG module to enhance AU recognition by generating feature maps at different scales and employing graph neural networks for relationship learning. We also developed the CCFG structure to dynamically aggregate features across different channels, improving AU feature representation by capturing local topological structures. The combination of these modules within a dual-branch feature fusion structure allows for effective integration of AU relationship learning with topology-based multi-channel aggregation, enhancing performance in complex and occluded scenarios. Experiments on widely recognized datasets, such as DISFA and EmotioNet, affirm the superior performance of our method. The results demonstrate significant improvements in the robustness of graph relation learning and the precision of facial expression recognition tasks, outperforming most state-of-the-art methods.

In this paper, there are still some limitations to our study. While our method shows strong performance on controlled datasets, its effectiveness in more diverse and unconstrained environments requires further investigation. The computational complexity introduced by the MAFG and CCFG modules may limit real-time applications, especially on devices with limited processing power. Additionally, our approach currently focuses on static images; extending it to video sequences to capture temporal dynamics of facial expressions could enhance its applicability. Looking ahead, we aim to address these limitations by optimizing the computational efficiency of our modules, exploring their applicability in real-time systems, and extending our method to handle dynamic facial expressions in video data. Furthermore, we are interested in applying this multi-branch feature map generation structure to enhance feature representation in other multi-label tasks involving complex relational information among labels.

Acknowledgements This research is partly funded by Natural Science Foundation of Hunan Province of China.

Author contributions The authors confirm their contribution to the paper as follows: K.L and Y.W contributed to study conception and design; K.L and G.Z performed draft manuscript preparation; K.L and J.Y performed analysis and interpretation of results; K.L, C.X, and W.L performed data preparation. All authors reviewed the results and approved the final version of the manuscript.

Funding This project is partly funded by Natural Science Foundation of Hunan Province of China (2024JJ7428).

Data availability The DISFA dataset and EmotioNet dataset used in our study can be requested at http://mohammadmahoor.com/disfa/ and http://cbcsl.ece.ohio-state.edu/EmotionNetChallenge/, respectively.

Declarations

Conflict of interest No conflicts of interest.

References

1. Ekman, P., Friesen, W.V.: Facial action coding system: a technique for the measurement of facial movement. Consulting Psychologists Press, Palo Alto (1978)
2. Ekman, R.: What the face reveals: basic and applied studies of spontaneous expression using the facial action coding system (FACS). Oxford University Press, USA (1997)
3. Martinez, B., Valstar, M.F., Jiang, B., Pantic, M.: Automatic analysis of facial actions: a survey. IEEE Trans. Affect. Comput. **10**(3), 325–347 (2017)
4. Ji, S., Wang, K., Peng, X., et al.: Multiple transfer learning and multi-label balanced training strategies for facial AU detection in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 1657–1661. IEEE, Seattle (2020)
5. Shao, Z., Liu, Z., Cai, J., et al.: Jâa-net: Joint facial action unit detection and face alignment via adaptive attention. Int. J. Comput. Vision **129**(2), 321–340 (2021)
6. Jaiswal, S., Valstar, M.: Deep learning the dynamic appearance and shape of facial action units. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1–8. IEEE (2016)
7. Li, W., Abtahi, F., Zhu, Z., et al.: Eac-net: Deep nets with enhancing and cropping for facial action unit detection. IEEE Trans. Pattern Anal. Mach. Intell. **40**(11), 2583–2596 (2018)
8. Ma, C., Chen, L., Yong, J.: AU R-CNN: Encoding expert prior knowledge into R-CNN for action unit detection. Neurocomputing **355**(AUG.25), 35–47 (2019)
9. Li, G., Zhu, X., Zeng, Y., Wang, Q., Lin, L.: Semantic relationships guided representation learning for facial action unit recognition. Proc. AAAI Conf. Artif. Intell. **33**(01), 8594–8601 (2019)
10. Shao, H., Zeng, Q., Hou, Q., et al.: MCANet: Medical image segmentation with multi-scale cross-axis attention. arXiv preprint arXiv:2312.08866 (2023)
11. Huang, G., Wen, Y., Qian, B., Bi, L., Chen, T., Sheng, B.: Attention-based multi-scale feature fusion network for myopia grading using optical coherence tomography images. Vis. Comput. **40**, 1–12 (2023)
12. Lin, X., Zhou, Y., Li, D., Huang, W., Sheng, B.: Image inpainting using multi-scale feature joint attention model. J. Comput. Aid. Des. Comput. Graph. **34**(8), 1260–1271 (2022)
13. Li, L., Chen, Z., Dai, L., Li, R. and Sheng, B.: MA-MFCNet: mixed attention-based multi-scale feature calibration network for image dehazing. In: IEEE Transactions on Emerging Topics in Computational Intelligence (2024)
14. He, K., Zhang, X., Ren, S., et al.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778. Springer, Las Vegas (2016)
15. Krizhevsky, A., Sutskever, I., Hinton, G.: ImageNet Classification with deep convolutional neural networks. Adv. Neural Inf. Process. Syst. **25**(2) (2012)
16. Luo, C., Song, S., Xie, W., Shen, L., Gunes, H.: Learning multi-dimensional edge feature-based au relation graph for facial action unit recognition. arXiv preprint arXiv:2205.01782 (2022)
17. Liu, Xin, et al.: Multi-scale promoted self-adjusting correlation learning for facial action unit detection. IEEE Trans Affect Comput (2024)
18. Cui, Z., Kuang, C., Gao, T., Talamadupula, K., Ji, Q.: Biomechanics-guided facial action unit detection through force modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8694–8703 (2023)
19. Li, X., Zhang, X., Wang, T., Yin, L.: Knowledge-spreader: Learning semisupervised facial action dynamics by consistifying knowledge granularity. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 20979–20989 (2023)
20. Yin, Y., Chang, D., Song, G., Sang, S., Zhi, T., Liu, J., Luo, L., Soleymani, M.: Fg-net: Facial action unit detection with generalizable pyramidal features. arXiv preprint arXiv:2308.12380 (2023)
21. Yuan, K., Yu, Z., Liu, X., Xie, W., Yue, H., Yang, J.: Auformer: Vision transformers are parameter-efficient facial action unit detectors. In: European Conference on Computer Vision, pp. 427–445. Springer, Cham (2025)
22. Chang, Y., Wang, S.: Knowledge-driven self-supervised representation learning for facial action unit recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022)
23. Jing, Y. et al.: Fan-trans: Online knowledge distillation for facial action unit detection. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision(2023)
24. Liu, Z., Dong, J., Zhang, C., et al.: Relation modeling with graph convolutional networks for facial action unit detection. In: Proceedings of the International Conference on Multimedia Modeling. Daejeon: Springer, pp. 489–501 (2020)
25. Yang, H., Yin, L., Zhou, Y., Gu, J.: Exploiting semantic embedding and visual feature for facial action unit detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10482–10491 (2021)
26. Song, T., Chen, L., Zheng, W., Ji, Q.: Uncertain graph neural networks for facial action unit detection. Proc. AAAI Conf. Artif. Intell. **35**(7), 5993–6001 (2021)
27. Niu, X., Han, H., Yang, S., Huang, Y., Shan, S.: Multi-label co-regularization for semi-supervised facial action unit recognition. Adv. Neural Inf. Process. Syst. **32** (2019)
28. An, H.Y., Jia, R.S.: Self-supervised facial expression recognition with fine-grained feature selection. Vis. Comput. **40**, 1–13 (2024)
29. Guo, M.-H., Lu, C.-Z., Hou, Q., Liu, Z., Cheng, M.-M., Hu, S.-M.: Segnext: Rethinking convolutional attention design for semantic segmentation. Adv. Neural. Inf. Process. Syst. **35**, 1140–1156 (2022)
30. Selvaraju, R. R., et al.: Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision (2017)
31. Kipf, T. N., Welling, M.: Semi-supervised classification with graph convolutional networks. Int. Conf. Learn. Represent. (2016)
32. Chen, Y., et al.: Channel-wise topology refinement graph convolution for skeleton-based action recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2021)
33. Ridnik, T., Ben-Baruch, E., Zamir, N., Noy, A., Friedman, I., Protter, M., Zelnik-Manor, L.: Asymmetric loss for multi-label classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 82–91 (2021)
34. Mavadati, S.M., et al.: Disfa: A spontaneous facial action intensity database. IEEE Trans. Affect. Comput. **4**(2), 151–160 (2013)
35. Mavadati, S. M., et al.: Automatic detection of non-posed facial action units. In: 2012 19th IEEE International Conference on Image Processing. IEEE (2012)
36. Benitez-Quiroz, C. F., Srinivasan, R., Martinez, A. M.: EmotioNet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016)

37. Yin, X., Liu, X.: Multitask convolutional neural network for pose-invariant face recognition. IEEE Trans. Image Process. **27**(2), 964–975 (2017)
38. Chen, Y., Chen, D., Wang, T., Wang, Y., Liang, Y.: Causal intervention for subject-deconfounded facial action unit recognition. Proc. AAAI Conf. Artif. Intell. **36**(1), 374–382 (2022)
39. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. Computer Vision and Pattern Recognition (2009)
40. Zhao, K., Chu, W.-S., Zhang, H.: Deep region and multi-label learning for facial action unit detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3391–3399 (2016)
41. Song, T., Cui, Z., Zheng, W., Ji, Q.: Hybrid message passing with performance-driven structures for facial action unit detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6267–6276 (2021)
42. Shao, Z., Liu, Z., Cai, J., Ma, L.: Deep Adaptive Attention for Joint Facial Action Unit Detection and Face Alignment. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 705–720, (2018)
43. Shao, Z., Liu, Z., Cai, J., Wu, Y., Ma, L.: Facial action unit detection using attention and relation learning. IEEE Trans. Affect. Comput. **13**(3), 1274–1289 (2019)
44. Jacob, G. M., Stenger, B.: Facial action unit detection with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7680–7689 (2021)
45. Xing, Y., Yu, G., Domeniconi, C., Wang, J., Zhang, Z.: Multi-label Co-training. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence, pp. 2882–2888 (2018)
46. Tarvainen, A., Valpola, H.: Mean teachers are better role models: weight-averaged consistency targets improve semi-supervised deep learning results. Adv. Neural Inf. Process. Syst. **30** (2017)
47. Blum, A., Mitchell, T.: Combining labeled and unlabeled data with co-training. In: Proceedings of the Eleventh Annual Conference on Computational Learning Theory, pp. 92–100, (1998)
48. Cai, H., et al.: Efficientvit: Multi-scale linear attention for high-resolution dense prediction. arXiv preprint arXiv:2205.14756 (2022)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Keqi Li is currently pursuing his master's degree at the School of Computer, University of South China. His research interests include machine learning, computer vision, and medical image processing.

Yaping Wan obtained his Ph.D. in System Structure Engineering from the School of Computer, Huazhong University of Science and Technology in 2009. He currently serves as a professor and dean at the School of Computer, University of South China. His research interests include the Internet of Things and intelligent nuclear security, big data causality inference, and machine vision.

Gang Zou earned his Ph.D. from the National University of Defense Technology and is currently a professor at the School of Computer, University of South China. He also conducts research at the HuNan ZK Help Innovation Intelligent Technology Research Institute, focusing on the integration of medical and engineering sciences. His research interests include computer vision, medical image analysis, and neural cognitive analysis.

Wangxiu Li holds a master's degree and is an associate professor at the School of Computer, University of South China. As a full-time teacher, her research interests lie in graphic processing and computer vision.

Jian Yang received his Master of Pharmacy degree from Jinan University and currently works as a chief pharmacist at Hunan Cancer Hospital. His research interests include cognitive assessment, neural cognitive analysis, and medical image analysis.

Changyi Xie obtained his master's degree from Monash University in Australia and is currently pursuing his Ph.D. at The University of Melbourne. His research focuses on big data analysis and cognitive science.