International Journal for Housing Science and Its Applications
Publish August 4, 2025. Volume 46, Issue 3 Pages 2436-2448

https://doi.org/10.70517/ijhsa463203

Research on expression recognition model based on multimodal hierarchical graph comparison learning

Xiaoyao Mo[1], Hairui Wang[1] and Guifu Zhu[2,*]
[1] Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, Yunnan, 650500, China
[2] Information Technology Construction Management Center, Kunming University of Science and Technology, Kunming,Yunnan, 650500, China
Corresponding authors: (e-mail: 17378091293@163.com).

Abstract In response to the limitations of existing methods in dynamic modeling of complex expressions, multimodal data quality optimization, and hierarchical feature fusion, this paper proposes a hierarchical graph comparison learning model based on local and global features. This model integrates graph neural network and contrastive learning techniques. It captures expression details by constructing local graphs, models cross-modal semantic collaboration through global graphs, and introduces an automatic graph enhancement strategy to improve the model's generalization ability. In the multimodal feature extraction stage, key features are accurately obtained from the video, audio, and text modalities respectively, and then the features are integrated through the intra-modal attention and multimodal fusion mechanisms. The experiments use the CMU-MOSI and CMU-MOSEI datasets. The results show that compared with multiple benchmark models, the model proposed in this paper performs better in terms of accuracy, recall rate, F1 score, and other indicators, and its mean square error is at a relatively good level. It can effectively integrate multimodal information, has excellent performance in the expression recognition task, and provides new ideas and methods for the development of this field.

Index Terms Multimodality, Deep Learning, Expression Recognition, Graph Convolutional Network

I. Introduction

With the rapid development of artificial intelligence technology, multimodal deep learning has demonstrated significant potential in the fields of affective computing and expression recognition. Traditional evaluation methods rely on single-modal data or manual observation, making it difficult to comprehensively capture human cognitive states and emotional changes [1]. Therefore, achieving precise and dynamic expression recognition through multimodal data fusion technology has become a critical issue for optimizing evaluations.

In recent years, remarkable progress has been made in multimodal sentiment analysis and expression recognition. Early works, such as the Tensor Fusion Network (TFN) proposed by Zadeh et al. [2], integrated multimodal features via tensor decomposition but failed to fully exploit dynamic inter-modal correlations. Liu et al. [3] introduced Low-rank Multimodal Fusion (LMF), which enhances fusion efficiency using modality-specific factors, yet exhibits limited capability in modeling temporal information. For multi-view sequential learning, Zadeh's team further proposed the Memory Fusion Network (MFN) [4], which models cross-modal interactions through gating mechanisms but still lacks sufficient capture of local features. Wang et al. [5] explored the dynamic adjustment of textual representations using nonverbal behaviors but did not systematically integrate hierarchical relationships in multimodal data. Subsequent studies, such as Tsai et al. [6], proposed Factorized Multimodal Representations (MFM), optimizing semantic representations by combining generative and discriminative objectives. MulT [7] addressed temporal misalignment via cross-modal attention but suffered from high computational complexity. Recent works like MISA [8] improved robustness by separating modality-invariant and modality-specific representations, MAG-BERT [9] integrated nonverbal information into pretrained models, and Self-MM [10] enhanced modality independence through self-supervised tasks. However, existing methods still face limitations in dynamic modeling of complex expressions, multimodal data quality optimization, and hierarchical feature fusion, particularly in educational scenarios characterized by data heterogeneity and insufficient synergy between local and global features.

To address these challenges, this paper proposes a hierarchical graph contrastive learning model based on local and global features, integrating graph neural networks (GNN) and contrastive learning techniques. By constructing local graphs to capture expression details and global graphs to model cross-modal semantic collaboration, the model introduces an automatic graph enhancement strategy to improve generalization. This approach enables efficient recognition of multimodal expressions while providing novel insights for the development of this field.

II. Multimodal Deep Learning and Graph Neural Networks

II. A. Multimodal Deep Learning

Multimodal deep learning is flourishing in facial expression recognition, as it combines multiple data modalities to significantly improve recognition accuracy and robustness. Facial expression recognition is often hindered by factors such as lighting, occlusion, and pose variations. Single-modal data struggles to comprehensively capture expression features, whereas multimodal data effectively compensates for these limitations [11]. Common multimodal fusion approaches include:

(1) Data-level fusion: Direct integration of raw data preserves rich information but imposes heavy computational burdens and is prone to noise.

(2) Feature-level fusion: Features are extracted from each modality separately before fusion, reducing redundancy and improving efficiency. This method is widely adopted.

(3) Decision-level fusion: Combines classification results from individual modalities, offering flexibility but risking information loss.

II. B. Graph Neural Networks

Graph Neural Networks (GNNs) exhibit unique advantages in expression recognition by effectively modeling complex feature relationships. Expressions involve multidimensional information, including facial muscle movements, morphological changes in facial features, and correlations with speech or text. GNNs treat expression-related elements (e.g., facial regions, audio tones, text keywords) as nodes and their relationships as edges to construct graph structures [12], [13]. For example, nodes representing eye closure and mouth opening are connected by edges that reflect their correlations during expression changes.

Through message passing and node representation updates, GNNs capture both global and local features of expressions [14]. For instance, when recognizing surprise, GNNs not only detect local features like widened eyes and open mouths but also analyze coordinated facial changes through node interactions, thereby improving authenticity assessment.

II. C. Multimodal Feature Extraction

Accurate feature extraction from multimodal data is essential for building efficient expression recognition models. Key modalities include video, audio, and text:

(1) Video modality: Extract keyframes at appropriate rates to focus on facial muscle movements and morphological changes. Tools like the Facial Action Coding System (FACS) and FaceNet[15] are used to obtain action units, poses, and gaze features.

(2) Audio modality: Extract pitch, speech rate, volume, pauses, and intonation using Librosa. For example, excitement correlates with faster speech and higher volume, while sadness manifests as lower pitch and slower speech.

(3) Text modality: Process text (e.g., chat logs, comments) using BERT-based embeddings [16] to capture semantic and emotional tendencies (e.g., "happy" or "angry" words indicating positive or negative emotions).

III. Multimodal Fusion Classification Framework

III. A. Intra-Modal Attention

After feature extraction, cross-modal attention mechanisms integrate features from different modalities. For instance, anger may involve frowning (visual), loud speech (audio), and negative keywords (text). Cross-modal attention assigns weights to these correlated features. Figure 1 illustrates the framework.

Let $N^{(j)} \in \square^{l \times d^a}$ denote the input features of the j-th layer in a modality encoder, where d^a is the feature dimension. Queries (Q), keys (K), and values (V) are computed via linear transformations: $Q = N^{(j)} W_Q$, $W_Q \in \square^{d^a \times d^a}$ .The cross-modal attention is formulated as:

$$\text{inter}\left(N^{(j)}, M^t\right) = soft\max\left(\frac{QK^T}{\sqrt{d}}\right)V = soft\max\left(\frac{N^{(j)}W_Q(M^tW_K)^T}{\sqrt{d}}\right)M^tW_V \tag{1}$$

In the above formula, inter is the attention between modes, and softmax is the calculated attention function.

This mechanism enables different modal information to complement and enhance each other, allowing the model to focus on key information, accurately insight into the emotional state behind the expression, take into account the impact of emotional factors on facial expression recognition, achieve deep integration and analysis of multi-modal information, and improve the accuracy of expression recognition.

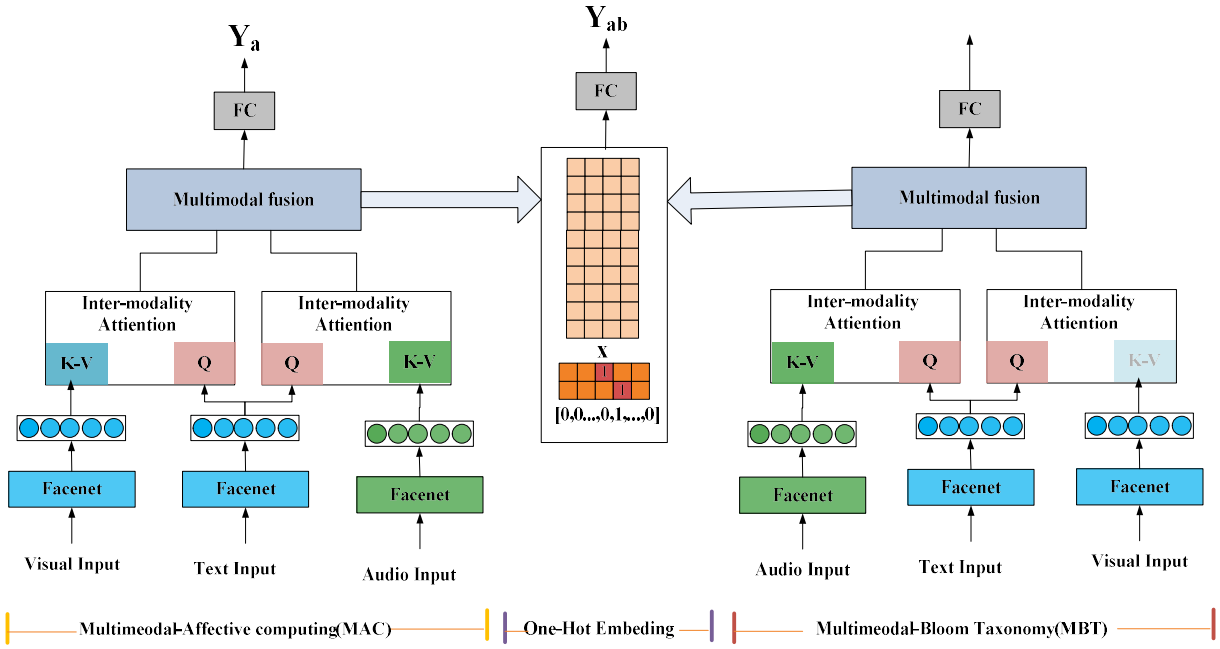

Figure 1: Multimodal Fusion Classification Framework

III. B. Multimodal Fusion

Multimodal fusion in expression recognition integrates the information from the three independent modalities of video, audio, and text into a cross-modal representation R_m rich in semantics. The specific formula is as follows:

$$R_m = F\left(R_t, R_a, R_v\right) \tag{2}$$

where R_a and R_v are the audio and video modal features after incorporating text information, respectively, R_t represents text features, and $F(\cdot)$ is the feature fusion encoding function.

The ot encoding technology is used. Taking the classification of expression categories as an example, in the text modality, if expressions are divided into categories such as happy, sad, and angry, they can be encoded as [1,0,0], [0,1,0], [0,0,1], etc. respectively. In the video modality, encoding is carried out according to the features of facial expressions (such as the degree of mouth corner raising, the size of eye opening, etc.). In the audio modality, encoding is performed according to the emotional features of speech (such as cheerful, low-pitched, passionate, etc.). Through this encoding method, data from different modalities can be converted into numerical forms suitable for calculation, enabling the integration of information from each modality within a unified framework, giving full play to the advantages of different modal data, and improving the model's ability to distinguish complex expression states.

IV. Hierarchical Graph Contrastive Learning Model Based on Local and Global Features

IV. A. Overall Model Framework

In the research of expression recognition, the hierarchical graph contrastive learning model based on local and global features plays a core role. This model consists of five key modules: graph construction, local-level graph contrastive learning, global-level graph contrastive learning, cross-modal graph contrastive learning, and fusion and expression prediction. These modules work together to accurately capture expression features and recognize expression states. Figure 2 shows the hierarchical graph contrastive learning framework based on local and global features.

(1) Graph Construction Module: Elements related to expressions, such as the movements of key facial parts (eyes, mouth, eyebrows, etc.), the associations between expressions and speech, and the connections between expressions and text emotions, are regarded as nodes. The internal logical relationships between them are regarded as edges to construct a graph structure. For example, since the closing action of the eyes is related to the surprised expression, these two elements are taken as nodes, and their association is taken as an edge. The

discrete variational autoencoder (dVAE) is used to mine the global co-occurrence features in the dataset, and the embedding space of each modality is obtained to lay the foundation for subsequent analysis.

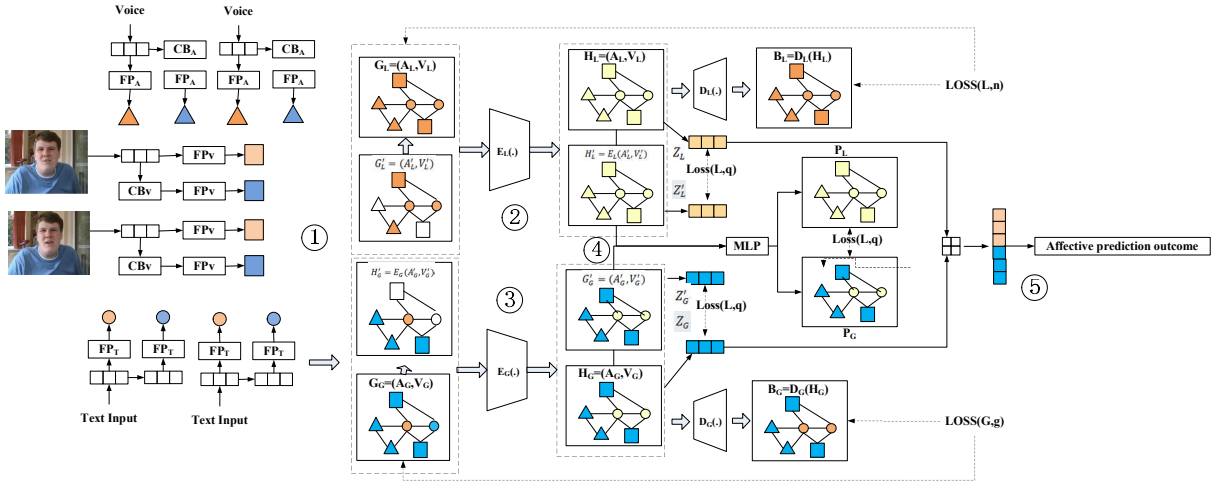

Figure 2: Hierarchical Graph Comparison Learning Framework Based on Local and Global Features

(2) Local-Level Graph Contrastive Learning: For the constructed local graph, an automatic graph enhancement strategy is used to generate an augmented graph. Then, both the original and augmented graphs are input into the graph encoder to obtain potential semantic features. By introducing the read-out function of the global average pooling and the fully-connected layer decoder, the graph-level and node-level invariances are considered respectively, and a special loss function is designed. This module focuses on local expression features, accurately identifying the details of specific expression actions and the instantaneous changes of emotional states, providing a basis for local expression analysis.

(3) Global-Level Graph Contrastive Learning: The sequence features of each modality are mapped and transformed with the help of codebooks to construct a global multimodal graph. It comprehensively explores the information associations at the global level, comprehensively considering the commonalities and differences of expressions in different scenarios and of different individuals, and evaluating expression features from a macroscopic perspective to avoid the one-sidedness of local analysis and improve the accuracy of overall expression recognition.

(4) Cross-Modal Graph Contrastive Learning: The local and global latent graph representations from the same sample are defined as positive-sample pairs. A non-linear projection MLP is used to transform them into the same space for comparative analysis. Through a carefully designed contrastive loss function, the model is prompted to accurately learn the complex associations between different modalities, such as the collaborative relationships between facial expression actions and speech emotions, and text semantics, realizing the in-depth fusion and accurate understanding of cross-modal information and enhancing the ability to distinguish comprehensive expression states.

(5) Fusion and Expression Prediction: The local and global representations are organically connected and input into the classifier as the fusion result to finally predict the expression category and emotional intensity. It integrates multimodal information related to expressions, including facial expression changes, speech intonation, and text emotional tendencies, and comprehensively evaluates the emotional state behind the expression. This can be applied in scenarios like determining customer emotions in intelligent customer service and understanding students' learning emotions in the education field.

IV. B. Graph Construction

In the research of expression recognition, graph construction is a crucial part of the hierarchical graph contrastive learning model based on local and global features. Its purpose is to transform multimodal data related to expressions into a graph structure, explore the information associations therein, and achieve accurate analysis of expression states.

(1) Codebook Construction: The discrete variational autoencoder (dVAE) is used to explore the embedding space from the multimodal dataset related to expressions and obtain the codebook of each modality. Taking the

facial action sequence features in the video modality as an example, given the original action sequence feature x_a, its feature vector set can be expressed as $X_a = \{a_i \mid i = 1, \cdots, T_a\} \in R^{T_a \times d_a}$, where a_i is the i - th vector, T_a is the input sequence length, and d_a is the vector dimension. The dVAE takes the action sequence features of all samples in the training set as input and, through calculation, obtains the action codebook $CB_a = \{cb_a^k \mid k = 1, \cdots, K_a\} \in R^{K_a \times d_a}$, where cb_a^k is the k - th vector and K_a is the size of the discrete space. Similarly, the text codebook CB_t and the audio codebook CB_v can be obtained. These codebooks contain the global co - occurrence features of the dataset and are important bases for subsequent graph structure construction.

(2) Local Graph Construction: The input feature vectors of each modality in the expression video are first processed by a dedicated feed - forward network to transform the feature embeddings of different modalities into the same dimension. Figure 3 shows the edge construction form in the model. When constructing edges, considering the core position of facial expressions in expression recognition, edges are constructed with facial expression features as the core. If the nodes come from the same modality, they are directly connected; if they come from different modalities, such as audio and facial expressions, or text and facial expressions, they are connected according to specific criteria. For example, when positive - emotion words appear in the text, a connection is established with the node of the smiling expression. After the connection operation, the local graph $G_L = (A_L, V_L)$ is obtained, where A_L is the adjacency matrix describing the connection relationship between nodes, and V_L is the node feature carrying specific information.

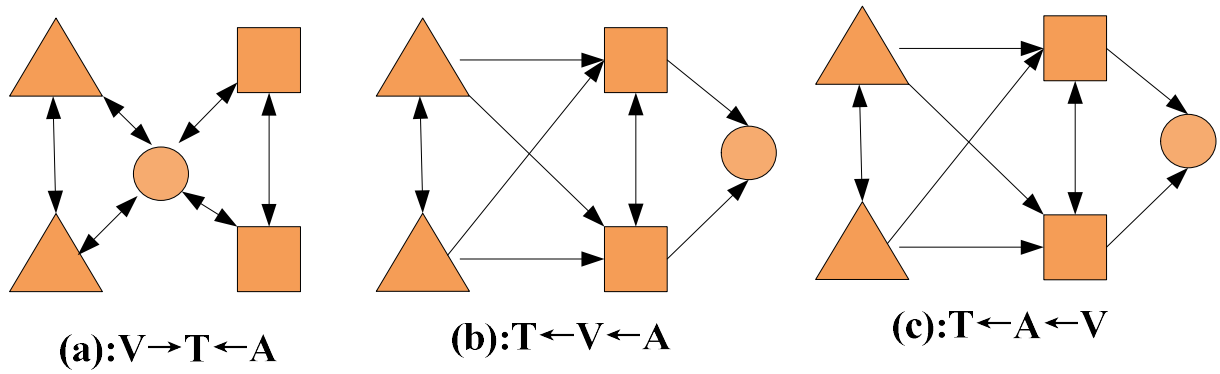

Figure 3: Edge Construction Form in the Model

(3) Global Graph Construction: Using the previously obtained codebooks $CB_m \left(m \in (t, a, v)\right)$, the original sequence features of each modality are mapped and transformed. Taking the audio modality as an example, for the global audio sequence feature x_a^i, through $k = 1, \cdots, K_a$, $id_i = \arg\min \left\| CB_a^k - x_a^i \right\|$, the corresponding CB_a^{id} is found, and $X_a^i = \{CB_a^{id} \mid i = 1, \cdots, T_n\} \in R^{T_a \times d_a}$ is obtained. The original sequence features of text and video are processed in the same way to obtain the global sequence feature X_t' of text and the global sequence feature X_v' of video. Finally, according to the method of constructing the local graph, the global multimodal graph $G_G = A_G$, V_G is constructed, where A_G is the adjacency matrix and V_G is the node feature, to explore the information interaction at the global level and analyze the associations of expression - related information from a macroscopic perspective.

IV. C. Hierarchical Graph Contrastive Learning

In the research of expression recognition, hierarchical graph contrastive learning is a key part of the hierarchical graph contrastive learning model based on local and global features. It improves the model's understanding and recognition ability of expression features through the analysis of local and global graphs.

Given the local graph $G_L = (A_L, V_L)$, to enhance the model's understanding and learning of data, an automatic graph enhancement strategy is used to generate an augmented graph $G_L' = (A_L, V_L)$. Then, G_L and G_L' are input into the graph encoder to obtain their potential semantic features H_L and H_L'.

In this process, to ensure that the model learns invariant feature representations, considerations are made from the graph - level and node - level perspectives. For graph - level invariance, the read - out function of the global average pooling is introduced:

$$z_L = ALLOUTPUT\left(H_t\right) \tag{3}$$

$$z'_L = ALLOUTPUT\left(H'_L\right) \tag{4}$$

The ALLOUTPUT function can aggregate the node features of a graph into a graph - level representation, enabling the model to focus on the overall common features when processing different local graphs. For node - level invariance, a fully - connected layer is used as the decoder, and the expression is:

$$G_{(L,r)} = Decoder\left(H_L\right) = \left(A_L, V_{(L,v)}\right) \tag{5}$$

Based on the above operations, for a mini - batch containing N examples, the local graph contrastive learning loss function is designed as:

$$Loss_{local} = Loss_{(L,n)} + \alpha Loss_{(L,g)} \tag{6}$$

where the calculation methods of each variable are as follows:

$$Loss_{(L,n)} = \frac{1}{N}\sum_{i=1}^{N}\frac{\left\|V_L^i - V_{(L,r)}^i\right\|^2}{\left\|V_L^i\right\|} \tag{7}$$

$$Loss_{(L,g)} = \frac{1}{N}\sum_{i=1}^{N}\left\|z_L^i - z_L^{ri}\right\| \tag{8}$$

In the above formulas, V_L^i and $V_{(L,r)}^i$ represent the original node features and the node features reconstructed by the decoder in the i - th graph, respectively. $Loss_{(L,n)}$ is used to measure the node - level difference, prompting the model to learn stable node feature representations; z_L^i and z_L^{ri} are the graph - level representations of the i - th graph, and $Loss_{(L,g)}$ focuses on the graph - level difference, helping the model capture the overall structure information of the graph. α is a hyper - parameter for adjusting the balance, used to adjust the weights of the two losses in the overall loss. Through this hierarchical graph contrastive learning method, the model can better mine the local features in expression data, improve the recognition ability of expression details and specific expression actions, and lay a foundation for accurate expression recognition.

IV. D. Automatic Graph Data Augmentation Strategy

In the research of expression recognition, the automatic graph data augmentation strategy is crucial for improving the performance of the hierarchical graph contrastive learning model based on local and global features. It expands data diversity by performing specific operations on the graph structure, enabling the model to learn more abundant expression features and enhancing the model's generalization ability.

Figure 4 shows the automatic graph enhancement framework of the model. On the left is the input data (Input), which contains multiple nodes (such as 1, 2, 3, 4, etc.). The input graph is processed and divided into two paths: one is the augmented graph (Augmented), and the other is the original graph (Raw). For the augmented graph, its node and edge structures are different from those of the original graph. For example, some edges may be added or deleted. The augmented graph and the original graph are respectively input into the encoder (Encoder), which consists of multiple stacked modules. After being processed by the encoder, they are then input into the classifier (Classifier). The entire framework shows how to improve the model's performance through automatic graph data augmentation, from the processing of the original graph to the augmented graph, and then to the encoding and classification processes.

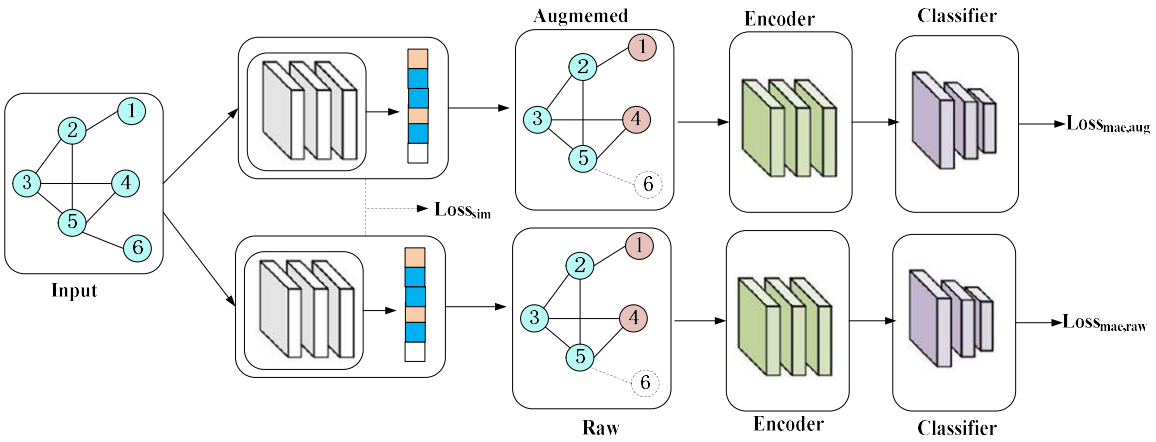

Figure 4: Automatic Graph Enhancement Framework of the Model

In terms of node feature perturbation, for the node features V_L in the local graph $G_L = (A_L, V_L)$, random noise is added for perturbation. Suppose the original node feature vector is as follows:

$$V_L^i = \{v_{i1}, v_{i2}, \ldots, v_{id}\} \tag{9}$$

where d is the feature dimension and i represents the i - th node. The enhanced node feature vector $V_L^{i'}$ can be expressed as:

$$V_L^{i'} = V_L^i + \dot{o} \tag{10}$$

where $\dot{o} = \dot{o}_1, \dot{o}_2, \cdots, \dot{o}_d$ is a random noise vector sampled from a normal distribution with a mean of 0 and a variance of σ^2. This perturbation simulates possible measurement errors or minor changes in actual data, prompting the model not to rely too much on precise node feature values and enhancing its adaptability to feature changes.

For edge operations, they include random addition and deletion of edges. In the process of random edge addition, for the element $a_{i,j}$ (representing the connection relationship between node i and node j) in the adjacency matrix A_L, with a probability of p_{add}, an edge is added between node pairs that were not originally adjacent. That is, if originally $a_{i,j} = 0$ and meets the random condition, for example, the value sampled from a uniform distribution on [0, 1] is less than p_{add}, then $a_{i,j}$ is updated to 1. At the same time, to maintain the symmetry of the graph, $a_{j,i}$ is also updated to 1.

The automatic graph data augmentation strategy significantly improves the performance of the model in the expression recognition task through the above operations on node features and edges, providing more powerful support for accurate expression recognition.

IV. E. Graph Representation Learning

In the hierarchical graph contrastive learning model based on local and global features, graph representation learning is a core step to obtain effective expression feature vectors and is crucial for accurate expression recognition.

For the local graph $G_L = (A_L, V_L)$ and the global graph $G_G = (A_G, V_G)$, a graph neural network (GNN) is employed for feature learning and representation update. During the forward propagation of the GNN, the update of the node feature V_L (or V_G) is based on the aggregation of its own features and those of adjacent nodes. Let the feature of node i at layer k be v_i^k, and its update formula can be expressed as:

$$v_i^k = \sigma\left(\sum_{j \in N(i)} w_k v_j^{k-1} + b_k\right) \tag{11}$$

where N_i represents the set of neighboring nodes of node i, W_k and b_k are the weight matrix and bias vector of layer k respectively, and σ is the activation function. Through the stacking of multiple GNN layers, nodes can continuously fuse multi-hop neighborhood information and gradually form semantic-rich feature representations.

In cross-modal graph contrastive learning, to better understand the relationships between graph representations of different modalities, the local and global latent graph representations H_L and H_G are transformed into the

same space through a non-linear projection MLP for comparison. Let the projected features be Z_L and Z_G, and the projection formulas are as follows:

$$Z_L = MLP(H_L) \tag{12}$$

$$Z_G = MLP(H_G) \tag{13}$$

A contrastive loss function is used to reduce the distance between positive sample pairs (i.e., the local and global graph representations of the same sample) and increase the distance between negative sample pairs. The InfoNCE loss formula is as follows:

$$\zeta_{\inf oNCE} = -\log \frac{\exp\left(sim\left(z_i^L, z_i^G\right)/r\right)}{\sum_{j-1}^{N} \exp\left(sim\left(z_i^L, z_i^G\right)/r\right)} \tag{14}$$

where sim represents the similarity measurement function (such as cosine similarity), r is the temperature parameter, and N is the number of samples. In this way, the model can learn the commonalities and differences of graph representations of different modalities and optimize the graph representation. For example, it can better discover the relationships between facial expression actions and speech emotions, as well as text semantics, enabling the graph representation to more accurately reflect the multimodal information in expressions. This provides high-quality inputs for subsequent fusion and expression prediction modules, thereby enhancing the overall performance of the model in expression recognition.

V. Experiments and Analysis

V. A. Datasets

In this experiment, the CMU-MOSI dataset [17] and CMU-MOSEI dataset [18] released by Zadeh et al. were used. These datasets provide rich and high-quality data resources for multimodal expression recognition research.

(1) CMU-MOSI Dataset: The videos in this dataset are sourced from opinion - sharing video blogs (vlogs) on various topics posted by users on the YouTube platform. The speakers in the videos are usually single individuals, and most of them express themselves facing the camera. The recording environments of the videos vary greatly. Some users use professional high-tech microphones and cameras, while others use more ordinary recording equipment, resulting in differences in video resolution, audio quality, etc. At the same time, the distances between users and the camera are different, and the background and lighting conditions also vary. The video duration ranges from 2 to 5 minutes. This dataset contains 2,199 opinion videos, and each sentence in the videos is assigned an emotion annotation within the range of [-3, 3], covering seven levels. Negative values represent negative emotions, positive values represent positive emotions, and a score of 0 indicates no obvious emotional tendency. These annotations are all manually labeled, ensuring the reliability of the data and providing diverse samples for expression recognition research.

(2) CMU-MOSEI Dataset: As the next - generation of the CMU-MOSI dataset, the CMU-MOSEI dataset is larger in scale and more diverse. It collects 23,453 annotated video clips from 1,000 different speakers, covering 250 topics. All videos are from online video - sharing websites. During the data processing, 5,000 videos were finally selected, and 14 expert reviewers strictly manually checked the quality of the videos, audio, and transcribed content within three months. The annotations of this dataset include not only emotion labels but also mood labels. The emotion labels adopt 2/5/7 classification methods, and the mood labels cover six aspects: happiness, sadness, anger, fear, disgust, and surprise. All annotations are completed by master - level workers with a pass rate of over 98%, ensuring high - quality annotations and providing richer emotional information for model training.

To ensure the scientificity and fairness of the experimental results, the CMU-MOSI dataset and CMU-MOSEI dataset were divided. The CMU-MOSI dataset was divided into a training set, a validation set, and a test set in a ratio of 6:1:3, containing 1,284, 229, and 686 samples respectively. The CMU-MOSEI dataset was divided in a ratio of 7:1:2, with the number of samples in the training set, validation set, and test set being 16,326, 1,871, and 4,659 respectively. Through such division, different subsets are used for parameter learning, performance verification, and effect evaluation of the model, which helps to comprehensively and accurately evaluate the performance of the model in the multimodal expression recognition task.

V. B. Experimental Environment

The hardware used is a high - performance server configured with an Intel Xeon E5 - 2620 v4 processor, 64GB of DDR4 memory, and an NVIDIA Tesla V100 GPU. In terms of software, the operating system is Ubuntu 18.04 LTS,

the deep - learning framework is PyTorch 1.8.0, and relevant scientific computing libraries such as Python 3.7, NumPy, and Pandas are also used.

V. C. Experimental Settings

(1) Experimental Method: The hierarchical graph contrastive learning model based on local and global features proposed in this paper is designed based on existing relevant research. It adopts a multimodal fusion method to integrate and analyze data from video, audio, and text modalities. During the model training process, techniques such as contrastive learning and automatic graph data augmentation are used to optimize the model performance. At the same time, the cross - validation method is used to train and evaluate the model multiple times to ensure the reliability of the experimental results.

(2) Evaluation Metrics: To comprehensively evaluate the performance of the model in the expression recognition task, metrics such as accuracy, recall, F1 - score, and mean square error (MSE) are used. The accuracy is used to measure the proportion of correctly recognized expressions by the model, and the calculation formula is:

$$Accuracy = \frac{Number\ of\ correctly\ predicted\ samples}{Total\ number\ of\ samples} \tag{15}$$

The recall reflects the model's ability to recognize specific expressions, and the calculation formula is:

$$\mathrm{Re}\,call = \frac{Number\ of\ correctly\ recognized\ specific\ \exp ression\ samples}{Actual\ number\ of\ specific\ \exp ression\ samples} \tag{16}$$

The F1 - score comprehensively considers precision and recall and can more comprehensively evaluate the model performance. The calculation formula is:

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{17}$$

where the precision is calculated as follows:

$$\mathrm{Pr}\,ecision = \frac{Number\ of\ correctly\ predicted\ samples}{Number\ of\ predicted\ samples} \tag{18}$$

The mean square error is used to evaluate the deviation between the model's predicted values and the true values, and the calculation formula is:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 \tag{19}$$

where y_i is the true value, $\hat{y}_i$ is the predicted value, and n is the number of samples.

(3) Comparative Experiment Design: To comprehensively evaluate the effectiveness of this model in the expression classification task, multiple representative benchmark models in the field were selected for performance comparison. The selected models cover diverse technical paths, specifically including:

TFN [2]: It integrates deep neural networks and tensor decomposition methods, and constructs multi - dimensional tensors through outer products to capture single - modal, cross - modal, and multimodal correlation features, showing outstanding performance in affective computing and expression classification tasks.

LMF [3]: It explores the correlations and complementarities between modalities through low - rank matrix decomposition, achieving the collaborative optimization and efficient integration of multi - source data, thereby enhancing task performance.

MFN [4]: It continuously models the interactions between specific perspectives and cross - perspectives, and combines the multi - perspective gating memory mechanism to achieve information induction and optimize the model performance.

RAVEN [5]: It simulates the human language cognitive mechanism, analyzes fine - grained audio - visual modal features to construct non - verbal representations, and dynamically adjusts semantic expressions based on non - verbal behaviors.

MFM [6]: It jointly optimizes generative and discriminative objectives, and uses modal discriminative factors and generative factors to extract multimodal semantic representations.

MulT [7]: It expands the multimodal Transformer framework based on the bidirectional cross - attention mechanism to solve the problems of cross - modal temporal misalignment and long - range dependence in an end - to - end manner.

MISA [8]: It learns modality - invariant (reducing modal differences) and modality - specific (retaining unique modal features) representations through dual - projection sub - spaces.

MAG - BERT [9]: It introduces multimodal adapters to transform the BERT architecture and fuses non - verbal data to optimize the fine - tuning effect of the sentiment analysis task.

SELF - MM [10]: It jointly trains multimodal and single - modal tasks based on a self - supervised label generation module to achieve the extraction of independent supervision signals.

V. D. Analysis of Experimental Results

(1) Comparison Results of Different Models in Experiments

Expression recognition experiments were carried out on the CMU-MOSI and CMU-MOSEI datasets for the hierarchical graph contrastive learning model based on local and global features and multiple benchmark models. Table 1 shows the experimental results of the CMU-MOSI dataset. The results show that on the CMU-MOSI dataset, the accuracy of this model is 84.85/86.45, the recall is 84.60/86.47, the F1 - score reaches 86.47, and the mean square error is 0.707. Table 2 shows the experimental results of the CMU-MOSEI dataset. On the CMU-MOSEI dataset, the accuracy is 85.06/85.79, the recall is 84.92/85.68, the F1 - score is 85.68, and the mean square error is 0.540. Compared with benchmark models such as TFN, LMF, MFN, RAVEN, MFM, MulT, MISA, MAG - BERT, and Self - MM, this model performs better in terms of accuracy, recall, and F1 - score, and its mean square error is also at a relatively good level. This indicates that the model has excellent performance in the expression recognition task, can more accurately recognize expressions, effectively integrate multimodal information, and has significant advantages in the field of expression recognition.

Table 1: Experimental Results of the CMU-MOSI Dataset

Model	Accuracy	Recall	F1 - score	Mean Square Error
TFN	-	-	80.7	-
LMF	-	-	82.4	-
MFN	77.4	-	77.3	-
RAVEN	78.0	-	76.6	-
MFM	-	-	81.6	-
MulT	81.5/84.1	80.6/83.9	83.9	0.861
MISA	81.8/83.4	81.7/83.6	83.6	0.783
MAG-BERT	84.2/86.1	84.1/86.0	86.0	0.712
Self-MM	84.00/85.98	84.42/85.95	85.95	0.713
Our Model	84.85/86.45	84.60/86.47	86.47	0.707

Table 2: Experimental Results of the CMU-MOSEI Dataset

Model	Accuracy	Recall	F1 - score	Mean Square Error
TFN	-	-	82.1	-
LMF	-	-	82.1	-
MFN	76.0	-	76.0	-
RAVEN	79.1/-	79.5/-	79.5	0.614
MFM	-	-	84.3	-
MulT	-/82.5	-/82.3	82.3	0.58
MISA	83.6/85.5	83.8/85.3	85.3	0.555
MAG-BERT	84.7/-	84.5/-	-	-
Self-MM	82.81/85.17	82.53/85.30	85.30	0.530
Our Model	85.06/85.79	84.92/85.68	85.68	0.540

(2) Comparison Results of Recognition Accuracy of Different Models for Multiple Typical Expressions

In the research of facial expression recognition, to further evaluate the performance of this model, it was compared with typical models such as MulT, MISA, MAG - BERT, and Self - MM. This experiment used the

CMU-MOSI dataset, and Table 3 shows the comparison results of the recognition accuracy of different models for typical expressions.

Table 3: The accuracy of different models in the recognition of various typical expressions

Expression Category	Our Model	MulT	MISA	MAG-BERT	Self-MM
Surprise	87.24%	75.00%	78.00%	72.00%	76.00%
Fear	86.50%	72.00%	74.00%	70.00%	73.00%
Anger	87.00%	73.00%	76.00%	71.00%	75.00%
Joy	88.00%	78.00%	80.00%	76.00%	79.00%
Neutral	87.50%	76.00%	77.00%	74.00%	77.00%
Disgust	86.80%	74.00%	75.00%	72.00%	75.00%
Sadness	87.30%	73.00%	74.00%	71.00%	74.00%
Contempt	84.00%	68.00%	70.00%	66.00%	69.00%
Confusion	85.00%	70.00%	72.00%	68.00%	71.00%
Fatigue	86.00%	72.00%	73.00%	70.00%	73.00%
Shyness	85.50%	71.00%	72.00%	69.00%	72.00%
Embarrassment	85.80%	71.50%	73.00%	70.00%	72.50%
Pride	86.20%	74.00%	75.00%	72.00%	74.50%
Depression	87.10%	73.50%	74.50%	71.50%	74.00%
Relief	87.60%	76.00%	77.00%	74.00%	76.50%
Curiosity	86.90%	75.50%	76.00%	73.50%	75.00%
Expectation	87.40%	76.50%	77.50%	74.50%	77.00%
Indifference	84.50%	70.00%	71.00%	68.00%	70.50%
Doubt	85.20%	71.50%	72.50%	69.50%	72.00%
Pleasant Surprise	87.80%	77.00%	78.50%	75.50%	78.00%

It can be seen from Table 3 that in the recognition of multiple expressions, this model shows higher recognition accuracy compared with models such as MulT, MISA, MAG - BERT, and Self - MM. This fully verifies that by fusing multimodal features, this model can more accurately capture expression features, improve the accuracy and robustness of expression recognition, and has obvious advantages.

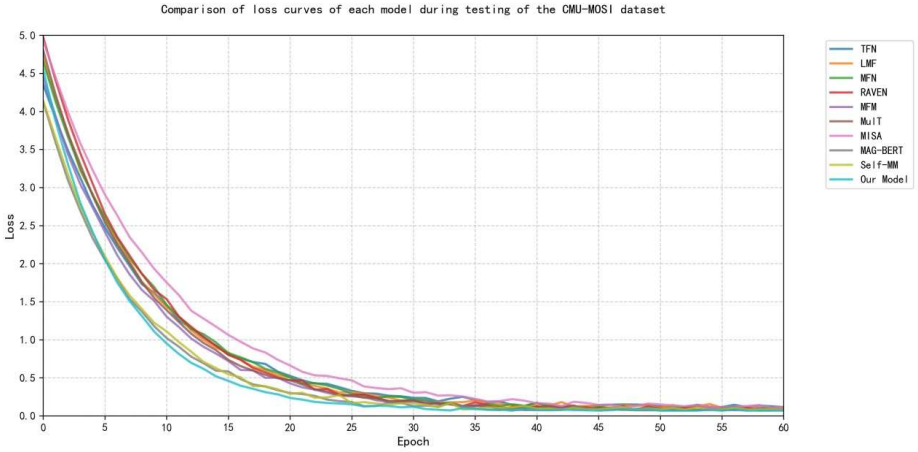

Figure 5: Comparison of loss curves of each model in CMU-MOSI dataset

During the model training process, the loss curve is an important indicator for evaluating the convergence and optimization effect of the algorithm. Figure 5 shows the comparison of the loss curves of this model and nine benchmark models on the CMU-MOSI dataset. The experiment simulated the training dynamics of different models, focusing on examining the loss attenuation trend and the final convergence level of each model within 60 training epochs. This model adopted specific parameter settings (base decay rate 0.85, final loss 0.068), while other models generated curves based on different decay parameters and final loss values.

As can be seen from Figure 5, the loss curve of the model presented in this paper presents the steepest decay trend, rapidly converges to the lowest level of 0.068 after the 35th training cycle, and remains stable in subsequent cycles. Compared with other benchmark models, the loss curve decay rate is generally slower, and the final loss value is concentrated in the range of 0.07-0.12. For example, although the pre-trained models such as MAG-BERT and Self-MM showed a fast downward trend in the early stage, due to the lack of hierarchical feature fusion mechanism for multi-modal data, the loss value tended to be flat after 30 cycles and finally stabilized at about 0.071. MulT and MISA models, due to the complexity of cross-modal attention mechanism, lead to large fluctuations in the training process, and the loss curve presents obvious oscillating characteristics. All the comparison models did not break the threshold of 0.07 in the late training period. However, by introducing the local-global hierarchical graph contrast learning and automatic graph enhancement strategies, the proposed model effectively suppressed the overfitting phenomenon and realized the continuous optimization of the loss value. The results verified the significant advantages of the proposed model in the dynamic modeling and feature fusion of multi-modal data.

VI. Conclusion

The research on expression recognition based on multimodal hierarchical graph contrastive learning proposed in this paper effectively addresses some of the key issues existing in the field of multimodal expression recognition. Through experimental verification, the model demonstrates significant advantages in complex expression recognition, can more accurately integrate multimodal information, and provides new directions for the development of expression recognition technology.

However, there is still room for improvement in this paper. On the one hand, at the data processing level, although multiple data augmentation strategies have been adopted, the adaptability of the model can be further optimized when facing more complex and diverse data in real - world scenarios. On the other hand, the computational complexity of the model is relatively high when processing large - scale data. In the future, it is necessary to explore more efficient algorithms to improve the operating efficiency of the model.

Funding

This work was supported by National Natural Science Foundation of China (62462064), Research and application of lossless compression key technology driven by deep learning on massive observation data of large telescopes; Education Science Planning Project of Yunnan Province in 2024 (BC24019), a Study on College Students' Classroom Behavior Recognition and Learning Outcome Prediction Based on Education Model; Scientific Research Fund Project of Education Department of Yunnan Province (2024J0105), Digital Competence of university teachers based on Deep Knowledge Tracking Evaluation and promotion research.

References

[1] Shao Zhiwen, Zhou Yong, Tan Xin, et al. A review of expression action unit recognition based on Deep learning [J]. Acta Electronica Sinica, 2022, 50(08): 2003 - 2017.

[2] Zadeh, Amir, Minghai Chen, et al. Tensor fusion network for multimodal sentiment analysis[C]// Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017: 1103 - 1114.

[3] Liu Zhun, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, et al. Efficient low - rank multimodal fusion with modality - specific factors[C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018: 2247 - 2256.

[4] Zadeh, Amir, Paul Pu Liang, et al. Memory fusion network for multi - view sequential learning[C]// Proceedings of the Thirty - Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence, 2018: 5634 - 5641.

[5] Wang Y, Shen Y, Liu Z, et al. Words can shift: dynamically adjusting word representations using nonverbal behaviors[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2019, 33(01): 7216 - 7223.

[6] Tsai, Yao - Hung Hubert, et al. Learning factorized multimodal representations[J]. arXiv preprint arXiv: 1806.06176, 2018.

[7] Tsai, Yao - Hung Hubert, Shaojie Bai, et al. Multimodal transformer for unaligned multimodal language sequences[C]// Proceedings of the Conference. Association for Computational Linguistics. Meeting. NIH Public Access, 2019: 2019 - 6558.

[8] D. Hazarika, R. Zimmermann, S. Poria, Misa: modality - invariant and specific representations for multimodal sentiment analysis[C]// Proceedings of the 28th ACM International Conference on Multimedia, 2020: 1122 – 1131.

[9] W. Rahman, K. M. Hasan, S. Lee, et al. Integrating multimodal information in large pretrained transformers[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020: 2359 – 2369.

[10] W. Yu, H. Xu, Z. Yuan, et al. Learning modality - specific representations with self - supervised multi - task learning for multimodal sentiment analysis[C]// Proceedings of the 35th AAAI Conference on Artificial Intelligence, 2021: 10790 – 10797.

[11] Jiao Shuang, Chen Guanghui. Multi - modal Emotion Recognition of Speech Expression based on weighted fusion [J]. Computer Simulation, 2024, 41(07): 417 - 422 + 428.

[12] Peng F, Liao F, Lu X, et al. Revisiting explicit recommendation with DC - GCN: Divide - and - Conquer Graph Convolution Network[J]. Information Systems, 2025, 130: 102513 - 102513.

[13] Lian H, Lu C, Chang H, et al. AMGCN: An adaptive multi - graph convolutional network for speech emotion recognition[J]. Speech Communication, 2025, 168: 103184 - 103184.

[14] Immordino G, Vaiuso A, Ronch D A, et al. Predicting transonic flowfields in non – homogeneous unstructured grids using autoencoder graph convolutional networks[J]. Journal of Computational Physics, 2025, 524: 113708 - 113708.
[15] Pei Yi, Liu Guangyu, Lei Yuanbin, et al. Research on Face recognition system for wearing masks based on RetinaFace and FaceNet algorithm [J]. Journal of Dali University, 2024, 9(12): 51 - 57.
[16] Jiao Yuchao, Yan Gang. Similar case matching based on BERT and factor Extraction [J]. Intelligent Computer and Applications, 2025, 15(01): 130 - 135.
[17] Peng F, Liao F, Lu X, et al. Revisiting explicit recommendation with DC - GCN: Divide - and - Conquer Graph Convolution Network[J]. Information Systems, 2025, 130: 102513 - 102513.
[18] Lian H, Lu C, Chang H, et al. AMGCN: An adaptive multi - graph convolutional network for speech emotion recognition[J]. Speech Communication, 2025, 168: 103184 - 103184.