2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

HyperMoCL: Emotion Recognition via Multimodal Representation Learning and Multi-Level Hypergraph Contrastive Learning

Shuoxin Liu[1], Guocheng An[2] and Xiaolong Wang[2, ✉]

[1] Shanghai Jiao Tong University, Shanghai, China
[2] Artificial Intelligence Research Institute of Shanghai Huaxun Network System Co., LTD., Chengdu, 610074, China
leonlsx1@sjtu.edu.cn

Abstract. Multimodal Emotion Recognition in Conversation (MERC) aims to identify the emotional states of speakers by integrating linguistic, audio, and visual information from dialogues. The core challenge of MERC lies in effectively fusing multimodal information and extracting key features. In recent years, hypergraph-based methods have been explored to construct hypergraphs directly using features output from unimodal encoders. However, due to the heterogeneity across modalities and the propagation of noise and redundant information within the hypergraphs, the modeling of inter-modal relationships often becomes inaccurate. Furthermore, existing approaches that employ node-level hypergraph contrastive learning overlook global structural information, resulting in insufficient modeling of global features. To address these limitations, we propose HyperMoCL, which integrates multimodal representation learning and multi-level hypergraph contrastive learning. First, HyperMoCL obtains higher-quality modal features through multimodal representation learning for hypergraph construction. Subsequently, a multi-level hypergraph contrastive learning framework is employed to comprehensively capture the structural features of the hypergraph, thereby enhancing feature discriminability and model robustness. Experimental results on two widely-used datasets (IEMOCAP, MELD) demonstrate that our method outperforms previous state-of-the-art approaches.

Keywords: multimodal representation learning, multi-level hypergraph contrastive learning, conversation emotion recognition.

1 INTRODUCTION

Emotion represents an innate psychological phenomenon that reflects the mental state of individuals and influences behavioral patterns. With the widespread application of AI and Human-Computer interaction technologies, the recognition of emotional states in conversations has become increasingly important. Multimodal Emotion Recognition in Conversation (MERC) aims to infer the emotional states of speakers by integrating multiple sensory information, such as language, audio and facial expressions. However, the expression and recognition of emotions are influenced by various factors. Firstly, different modalities (e.g., language, audio, and facial expressions) may contain complementary or conflicting information. For instance, a person may exhibit a neutral emotion in speech, while their facial expressions and tone reveal a negative emotion. Emotion recognition also relies on global context, such as the topic or background of the conversation. Additionally, individual characteristics of the speaker can affect the perception of emotions. Therefore, compared to unimodal emotion recognition, multimodal models can integrate multi-source information to comprehensively uncover the underlying emotional states of speakers, thereby improving recognition accuracy and robustness. Nevertheless, the heterogeneity between modalities makes the effective fusion of multimodal information and extraction of key emotional features a core challenge in MERC tasks.

Currently, MERC research primarily employs sequence-based and graph-based methods. Sequence-based methods employ Recurrent Neural Networks (RNNs) [1] and Transformers [2] to capture cross-modal interactions and temporal dynamics in conversational data, but they have limited capabilities in modeling long-range dependencies and struggle to distinguish between intra-utterance and inter-utterance interaction relationships. To address these issues, researchers have introduced Graph Neural Network (GNN) into MERC tasks. Traditional graph-based methods treat the multimodal information of utterances as nodes, connecting nodes of different modalities within the same utterance and nodes of the same modality across different utterances through edges. Graph convolution operations are then used to update the embeddings of nodes and edges, enabling information interaction within and between utterances. However, traditional graphs, limited to binary relationships, struggle to represent complex higher-order relations, potentially leading to insufficient information fusion when dealing with multimodal data. To address this issue, recent studies [5] have introduced Hypergraph Neural Networks (HGNNs) into MERC tasks. By leveraging hyperedges to simultaneously connect multiple nodes, HGNNs enable the modeling of complex high-order relationships across modalities. Through hypergraph convolution, these models effectively aggregate information from both nodes and hyperedges, thereby facilitating a more comprehensive representation of multimodal interactions. HAUCL[6] proposed a variational hypergraph autoencoder to adaptively adjust the hypergraph structure and combined it with hypergraph contrastive learning to enhance model stability. However, existing hypergraph methods still face the following challenges: (1) While these approaches directly construct hypergraphs from unimodal encoder outputs, such a strategy inherently overlooks the modality-specific characteristics and potential cross-modal interactions. The resulting hypergraphs are susceptible to structural degradation caused by modality heterogeneity, as well as the amplification of noise and redundant signals during message propagation, ultimately leading to suboptimal modeling of inter-modal dependencies. (2) Existing hypergraph contrastive learning methods restrict their contrastive objectives solely to node-level comparisons, failing to incorporate the rich structural information inherent in hypergraph topologies.

To address these challenges, we propose a novel model, HyperMoCL, which integrates multimodal representation learning with multi-level hypergraph contrastive learning. Specifically, HyperMoCL first employs unimodal encoders to extract modality-specific features from each input stream. These features are then further decomposed into shared and private components via a multimodal representation learning module. A Transformer encoder is subsequently utilized to capture global contextual information, resulting in high-quality modality representations for hypergraph construction.To optimize the structural quality of the hypergraph, a Variational Hypergraph Autoencoder (VHGAE) is applied, followed by hypergraph convolution to iteratively update both node and hyperedge embeddings. To further enhance the discriminability and robustness of learned features, we introduce a multi-level hypergraph contrastive learning framework, which performs contrastive learning from three complementary perspectives: (1) node-level, (2) hyperedge-level, and (3) association-level, which optimizes the structural dependencies between nodes and hyperedges. Finally, the optimized node features are fed into a classification layer for emotion recognition.

The main contributions of this paper are summarized as follows:

- A novel multimodal representation learning module is proposed, which extracts shared and private features of modalities through a shared-private network architecture and combines a Transformer encoder for global context modeling, providing higher-quality modal representations for hypergraph construction.
- A multi-level hypergraph contrastive learning framework is designed, conducting contrastive learning from three levels: nodes, hyperedges, and their interrelationships, achieving comprehensive modeling of the hypergraph structure and enhancing the robustness of the model.
- Experiments on two mainstream datasets, IEMOCAP and MELD, demonstrate that the proposed method outperforms existing state-of-the-art methods in terms of accuracy and weighted F1 score.

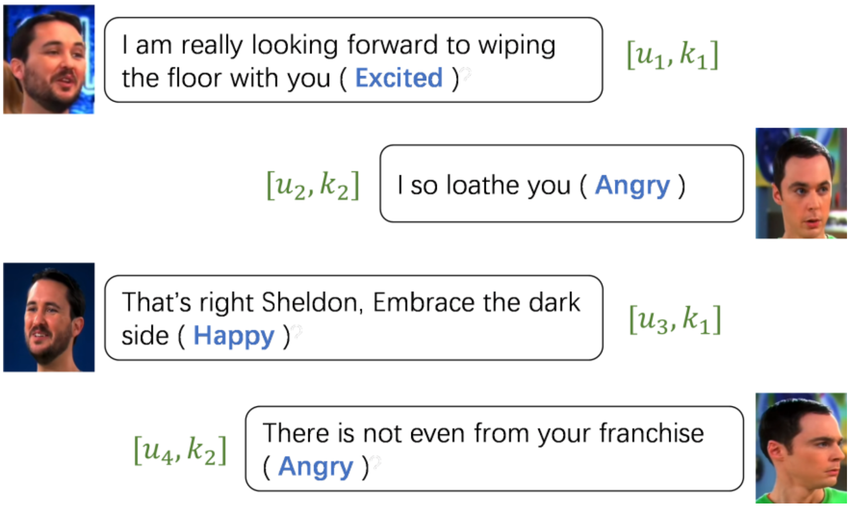
I am really looking forward to wiping the floor with you (Excited)
$[u_1, k_1]$
$[u_2, k_2]$
I so loathe you (Angry)
That's right Sheldon, Embrace the dark side (Happy)
$[u_3, k_1]$
$[u_4, k_2]$
There is not even from your franchise (Angry)

Figure 1. An example of a multimodal emotion dialogue system. Image from The Big Bang Theory

2 RELATED WORK

2.1 Multimodal Emotion Recognition

Multimodal emotion recognition research aims to identify speakers' emotions by integrating information from multiple modalities, such as text, visual, and audio. Early studies were based on Transformer and recurrent neural networks. For example, DialogueRNN [1] employs three GRUs (global GRU, party GRU, and emotion GRU) to model emotions, context, and speaker states in dialogues, enabling efficient information extraction and fusion through recursive connections. MulT [2] introduces a multimodal Transformer model, utilizing directional cross-modal Transformers to achieve pairwise modality feature fusion. Recently, graph neural networks have been introduced into this field. MMGCN [3] constructs fully connected graphs to model multimodal information, but it also introduces data heterogeneity issues. GraphMFT [4] addresses this by creating three bimodal graphs to reduce data heterogeneity and incorporates a graph attention mechanism to optimize information fusion.

However, traditional graphs are limited to binary relationships, making it difficult to represent complex multi-way interactions. To address this issue, M3Net[5] pioneers the use of hypergraph neural networks, leveraging hyperedges to connect multiple nodes and capture high-order relationships, while improving prediction accuracy through frequency filters. HAUCL[6] proposes a variational hypergraph autoencoder to dynamically adjust the hypergraph structure, reduce information redundancy, and enhance model robustness through hypergraph contrastive learning. These approaches provide new perspectives for multimodal emotion recognition.

2.2 Hypergraph Learning

Hypergraph learning extends traditional graph learning methods and uses hypergraph structures to model complex interactions and higher-order relationships between data. Unlike traditional graphs, hypergraphs allow a hyperedge to connect multiple nodes, enabling a more natural representation of multi-order relationships. This capability enhances the integration of multimodal information and the modeling of multi-way relationships. Hypergraph learning has demonstrated significant advantages in various fields, including recommendation systems[7], computer vision[8], and sentiment recognition[9].

With the rapid advancement of AI, research on hypergraph neural networks has deepened. HGNN[10] defines the Laplacian operator on hypergraphs, facilitating information propagation and aggregation between nodes and hyperedges, thereby learning deep feature representations. DHGNN[11] proposes a dynamic hypergraph neural network that dynamically updates the hypergraph structure using k-NN and k-means, adapting to changes in feature embeddings and enhancing the ability to capture complex relationships. To further exploit hypergraph learning, researchers have delved into hypergraph contrastive learning. TriCL[12] introduces a general hypergraph contrastive learning framework, employing three forms of contrast to enhance feature learning.

HyGCL-AdT[13] improves model performance through a dual-level contrastive learning approach, encompassing node-level and community-level contrasts.

3 METHOD

The goal of MERC is to identify the emotional type of each utterance in dialogues. Formally, a dialogue sequence containing n utterances is defined as: $\{[u_1, k_1]$, $[u_2, k_2]$, $[u_3, k_3]$ ......$[u_n, k_n]\}$, where u_i represents the i-th utterance, and k_i denotes the speaker of that utterance. The model takes the utterances as input, with each utterance u_i consisting of three modalities: text, audio and visual, represented as $[u_i^t, u_i^a, u_i^v]$. The objective of the model is to identify the emotional type of the utterance by integrating its multimodal information u_i^t, u_i^a, u_i^v and the speaker information k_i. As shown in Figure 2, the proposed model framework comprises six core modules: unimodal encoding, multimodal representation learning, hypergraph reconstruction, hypergraph convolution, multi-level hypergraph contrastive learning, and emotion classification.

3.1 Unimodal Encoding

In this paper, we first employ unimodal encoders to extract features from the raw input data of each modality. Specifically, following the approach in M3NET[5], we utilize the RoBERTa large model[14] to extract textual features, the OpenSmile toolkit[15] to extract acoustic features, and DenseNet[16] or 3D-CNN[17] to extract visual features. To enhance the fusion of contextual information, we further process the extracted features: two single-layer fully connected networks are used to encode the acoustic and visual features, respectively, while a bidirectional Gated Recurrent Unit (Bi-GRU) is employed to encode the textual features. Additionally, the features of the three modalities are projected into the same dimensional space to facilitate subsequent multimodal information fusion. The mathematical representation is as follows:

$$\begin{aligned} &\mathrm{U_i^t{=}W_t\left(\overleftrightarrow{GRU}(u_i^t,U_{i\text{-}1}^t,U_{i+1}^t)\right)} \\ &\mathrm{U_i^a{=}W_a u_i^a{+}b_i^a} \\ &\mathrm{U_i^v{=}W_v u_i^v{+}b_i^v} \end{aligned} \quad (1)$$

Where $\mathrm{U_i^t}$,$\mathrm{U_i^a}$,$\mathrm{U_i^v}\in\mathrm{R^{d_m}}$, $\mathrm{u_i}$ and $\mathrm{U_i}$ are the input and output of the unimodal encoder respectively, W and b are trainable parameters.

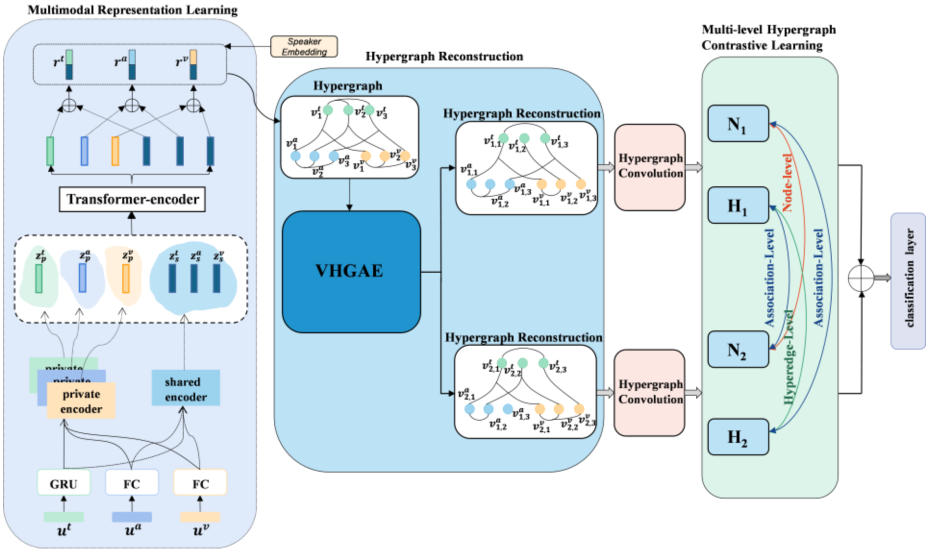
Multimodal Representation Learning
Speaker Embedding
r^t
r^a
r^v
Transformer-encoder
z_p^t
z_p^a
z_p^v
z_s^t
z_s^a
z_s^v
private
private
private encoder
shared encoder
GRU
FC
FC
u^t
u^a
u^v
Hypergraph Reconstruction
Hypergraph
v_1^t v_2^t v_3^t v_1^a v_2^a v_3^a v_1^v v_2^v v_3^v
VHGAE
Hypergraph Reconstruction
$v_{1,1}^t$ $v_{1,2}^t$ $v_{1,3}^t$ $v_{1,1}^a$ $v_{1,2}^a$ $v_{1,3}^a$ $v_{1,1}^v$ $v_{1,2}^v$ $v_{1,3}^v$
Hypergraph Reconstruction
$v_{2,1}^t$ $v_{2,2}^t$ $v_{2,3}^t$ $v_{2,1}^a$ $v_{1,2}^a$ $v_{1,3}^a$ $v_{2,1}^v$ $v_{2,2}^v$ $v_{2,3}^v$
Hypergraph Convolution
Hypergraph Convolution
Multi-level Hypergraph Contrastive Learning
N_1
H_1
N_2
H_2
Node-level
Association-Level
Hyperedge-Level
Association-Level
classification layer

Figure 2. The overall framework of the model we proposed. N_i and H_i represent the node features and hyperedge features of the i-th view of the hypergraph, respectively.

3.2 Multimodal Representation Learning

In the task of MERC, directly constructing hypergraphs based on the modal features output by unimodal encoders poses two main challenges: (1) the heterogeneity between different modal data leads to inaccurate modeling of inter-modal relationships; and (2) noise and redundant information in the data propagate through the hypergraph structure. To address these issues, we design a Multimodal Representation Learning module. This module learns shared information across modalities through a shared network layer, reducing the impact of heterogeneity, while capturing modality-private features through private network layers, ensuring that key information within each modality is preserved. Specifically, the shared network layer E_s^x uses parameters θ_s, while the private network layer E_p^x assigns independent parameters θ_p^x for each modality. Each modal feature U^x is processed by the shared and private network layers to obtain the shared feature z_s^x and the private feature z_p^x, respectively. Both the shared and private network layers are composed of feedforward neural networks. Mathematically:

$$z_s^x = E_s^x[U^x, \theta_s]$$

$$z_p^x = E_p^x[U^x, \theta_p^x] \tag{2}$$

where $x \in \{t, a, v\}$. To further capture inter-modal interactions and perform global context modeling, the six features $[z_s^t, z_s^a, z_s^v, z_p^t, z_p^a, z_p^v]$ are concatenated and fed into a Transformer-Encoder for information fusion. Finally, the fused shared and specific features of each modality are concatenated to form a higher-quality representation r^x of each modality:

$$r^x = z_s^x \oplus z_p^x \tag{3}$$

where $\oplus$ denotes the concatenation operation. This design reduces noise and redundant information, providing a holistic view of the multimodal data. Considering the importance of speaker information, we use one-hot encoding to encode the speaker and add it to the context-aware unimodal encoding to obtain the final output of this module:

$$K_i = \boldsymbol{W}_1 k_i$$
$$v_i^x = r_i^x + K_i \tag{4}$$

where K_i represents the speaker information encoding, v_i^x is the final output, and W_1 is a trainable weight matrix.

Loss: Inspired by [18], this paper introduces three loss functions for this module: similarity loss, diversity loss, and reconstruction loss. The details are as follows:

Similarity Loss: To capture emotion information shared across modalities, we introduce a similarity loss. This loss constrains the shared features of different modalities to be as close as possible in the feature space using the Central Moment Discrepancy (CMD) metric:

$$\mathcal{L}_{sim} = \frac{1}{3}\sum_{(x_1,x_2)} \mathrm{CMD}(z_s^{x_1}, z_s^{x_2}) \tag{5}$$

Diversity Loss: The diversity loss ensures that the shared and private features capture different aspects of the input data by enforcing orthogonality constraints to achieve non-redundancy between these features. The orthogonality constraint is implemented by calculating the Frobenius norm between two features. Additionally, to ensure that the private features of different modalities are as uncorrelated as possible, an orthogonality constraint is added between the private features of different modalities:

$$\mathcal{L}_{diff} = \sum_{x\in\{t,a,v\}} \left\| {z_s^x}^T z_p^x \right\|_F^2 + \sum_{(x_1,x_2)} \left\| {z_p^{x_1}}^T z_p^{x_2} \right\|_F^2 \tag{6}$$

where $\|.\|_\mathrm{F}^2$ denotes the squared Frobenius norm.

Reconstruction Loss: To ensure the effectiveness of the shared and private features and prevent the model from learning invalid information during feature decomposition, we introduce a reconstruction loss that forces the decoder to reconstruct the original modal data. The decoder is defined as:

$$r^x = D(z_s^x \oplus z_p^x) \tag{7}$$

where D consists of linear network layers. The reconstruction loss between u^x and h^x is defined as：

$$\mathcal{L}_{re} = \frac{1}{3}\sum_{x\in\{t,a,v\}} \frac{\|u^x - r^x\|_2^2}{d} \tag{8}$$

where $\|.\|_2^2$ denotes the squared L2 norm. Finally, the three losses are weighted and summed to obtain the final modal representation loss:

$$\mathcal{L}_o = 0.25 * \mathcal{L}_{sim} + 0.25 * \mathcal{L}_{diff} + 0.5 * \mathcal{L}_{re} \tag{9}$$

3.3 Hypergraph Reconstruction

Hypergraph structure initialization. In this paper, we construct a hypergraph Φ= (V, E), where $v \in V$ represents a node and $e \in E$ represents a hyperedge. For a dialogue containing N utterances, the information from each modality of every utterance is treated as a node in the hypergraph, resulting in a total of $3N$ nodes. We define two types of hyperedges: the first type connects all modalities within the same utterance, and the second type connects all utterance nodes of the same modality, resulting in a total of $M = 3 + N$ hyperedges. The incidence matrix H of the hypergraph is defined as:

$$\boldsymbol{H}_{i,j} = \begin{cases} 1, if\ node\ i\ is\ on\ hyperedge\ j \\ 0, if\ node\ i\ is\ not\ on\ hyperedge\ j \end{cases} \tag{10}$$

VHGAE. We adopt the Variational Hypergraph Autoencoder (VHGAE) proposed in [6] to optimize the original hypergraph and generate two enhanced hypergraph structures. VHGAE consists of three steps: encoding, sampling, and decoding.

Encoder: The encoder utilizes Hypergraph Neural Networks (HGNN) to perform hypergraph convolution, extracting feature representations of nodes and hyperedges, denoted as v and e, respectively. These features are then mapped to mean vectors μ and variance vectors σ in the latent space through linear transformations and activation functions:

$$v, e = HGNN(G) \qquad \mu_l = f(l, \theta_u),\ \sigma_l = f(l, \theta_\sigma) \tag{11}$$

Where f denotes linear transformation, θ represents trainable parameters, $l \in \{v, e\}$.

Sampler: The sampler generates new embeddings by sampling representations from the latent space and introduces randomness using the reparameterization trick:

$$m_l = u_l + \sigma_l \odot \delta \tag{12}$$

where $\delta \sim N(0,1)$ is standard normal distribution noise, $\odot$ denotes element-wise multiplication, and m_l represents the new embedding.

Decoder: The decoder computes the initial connection matrix h_i by taking the dot product of node and hyperedge embeddings. The Gumbel-Softmax function is then applied to h_i to introduce randomness, followed by a softmax operation to obtain the final matrix h:

$$h_i = m_v^T m_e$$

$$h = softmax(Gumbel\_Softmax(h_i, \tau) + p) \tag{13}$$

where τ is the temperature coefficient, and p is a constant to prevent numerical overflow. The first column of matrix h corresponds to the incidence matrix of the new hypergraph Φ′= (V, E'), completing the hypergraph structure reconstruction. Through

VHGAE, we effectively reduce redundant information in the hypergraph and generate new hypergraph structures.

The loss function of VHGAE is defined as:

$$\mathcal{L}_g = CE(h_o, h) + KL(m_v, v) + KL(m_e, e) \tag{14}$$

where $CE(h_o, h)$ is the cross-entropy loss, measuring the difference in connection structures between the generated hypergraph and the original hypergraph. $KL()$ denotes the Kullback-Leibler (KL) divergence, which measures the difference between the distribution of latent variables (nodes and hyperedges) and the prior distribution.

3.4 Hypergraph Convolution

After constructing the new hypergraph, we employ hypergraph convolution to propagate information and update the embeddings of nodes and hyperedges. The process consists of two steps: **(1) Node-to-Hyperedge Information Propagation**: The features of nodes connected by hyperedges are used to update the hyperedge embeddings. Mathematically:

$$\boldsymbol{E} = \widehat{\boldsymbol{H}}\boldsymbol{V} \tag{15}$$

$$\widehat{\boldsymbol{H}} = \begin{cases} \gamma_j^i,\ if\ node\ i\ is\ on\ hyperedge\ j \\ 0, if\ node\ i\ is\ not\ on\ hyperedge\ j \end{cases} \tag{16}$$

where $\widehat{\boldsymbol{H}}$ is the weighted adjacency matrix, and γ_j^i represents the contribution weight of node i to hyperedge j. (2) **Hyperedge-to-Node Information Propagation**: The information from hyperedges is propagated back to the nodes connected to them. Mathematically, this is expressed as:

$$\boldsymbol{V} = \boldsymbol{HWE} \tag{17}$$

where **W**=diag$\{\mathrm{w}_1, \mathrm{w}_2, \mathrm{w}_3\ ...\}$ is the hyperedge weight matrix. The overall process can be represented as:

$$\widehat{\boldsymbol{V}} = \boldsymbol{D}^{-1}\boldsymbol{HWB}^{-1}\widehat{\boldsymbol{H}}^T\boldsymbol{V} \tag{18}$$

Here, $\boldsymbol{D}$ and $\boldsymbol{B}$ are the degree matrices of hyperedges and nodes, respectively, and both w and γ_j^i are trainable parameters initialized randomly. Through these two steps of information propagation, hypergraph convolution effectively updates the embeddings of nodes and hyperedges, thereby capturing high-order relationships among the data.

3.5 Multi-level Hypergraph Contrastive Learning

Previous studies [6] have focused solely on node-level hypergraph contrastive learning, overlooking the global structural properties of the hypergraph. To address this limitation, we propose a multi-level hypergraph contrastive learning framework that captures structural information from three complementary perspectives: (a)node-level, (b)hyperedge-level, and (c)association-level. The details of each level are described as follows:

(a) Node-Level: Corresponding node pairs in the two hypergraph views are treated as positive samples, while other node pairs are treated as negative samples. Specifically, for a node $v_{1,i}$ in the first view as the anchor, the corresponding node $v_{2,i}$ in the second view is the positive sample, and the remaining nodes $v_{2,k}$ (where $k \neq i$) in the second view are negative samples. The similarity between samples is measured using cosine similarity $\mathrm{s}(.)$. The node-level contrastive loss is defined as:

$$\mathcal{L}_{\mathrm{n}}\left(v_{1,\mathrm{i}}, v_{2,\mathrm{i}}\right) = -\log\frac{\mathrm{e}^{\mathrm{s}(v_{1,\mathrm{i}},v_{2,\mathrm{i}})/\tau}}{\sum_{\mathrm{k}=1}^{|\mathrm{V}|}\mathrm{e}^{\mathrm{s}(v_{1,\mathrm{i}},v_{2,\mathrm{k}})/\tau}} \tag{19}$$

Considering symmetry, the final node-level loss is obtained by averaging the losses from both views:

$$\widehat{\mathcal{L}_{\mathrm{n}}} = \frac{1}{2|\mathrm{V}|}\sum_{\mathrm{i}=1}^{|\mathrm{V}|}\{\mathcal{L}_{\mathrm{n}}\left(v_{1,\mathrm{i}}, v_{2,\mathrm{i}}\right) + \mathcal{L}_{\mathrm{n}}\left(v_{2,\mathrm{i}}, v_{1,\mathrm{i}}\right)\} \tag{20}$$

(b) Hyperedge-Level: Similarly, for a hyperedge $e_{1,j}$ in the first view as the anchor, the corresponding hyperedge $e_{2,j}$ in the second view is the positive sample, and the remaining hyperedges $e_{2,k}$ (where $k \neq j$) in the second view are negative samples. Using cosine similarity as the scoring function and considering symmetry, the hyperedge-level contrastive loss is defined as:

$$\mathcal{L}_{\mathrm{w}}\left(e_{1,j}, e_{2,j}\right) = -\log\frac{\mathrm{e}^{\mathrm{s}(e_{1,j},e_{2,j})/\tau}}{\sum_{\mathrm{k}=1}^{|\mathrm{E}|}\mathrm{e}^{\mathrm{s}(e_{1,j},e_{2,k})/\tau}}$$

$$\widehat{\mathcal{L}_{w}} = \frac{1}{2|E|}\sum_{j=1}^{|E|}\{\mathcal{L}_{w}(e_{1,j}, e_{2,j}) + \mathcal{L}_{w}(e_{2,j}, e_{1,j})\} \tag{21}$$

(c) Association-Level: To model the structural relationships between nodes and hyperedges, we propose a node-hyperedge association-level contrastive learning method. For a node v_i and a hyperedge e_j connected in the hypergraph (i.e., e_j connects v_i), we take node $v_{1,i}$ in the first view as the anchor and the corresponding hyperedge $e_{2,j}$ in the second view as the positive sample. Negative samples are randomly selected from hyperedges $e_{2,k}$ that are not connected to v_i (where $i \notin k$). Alternatively, hyperedge $e_{2,j}$ can be used as the anchor, with nodes $v_{1,k}$ (where $k \notin j$) as negative samples. A scoring function $R(v, e)$ is designed to evaluate the matching degree of node-hyperedge pairs. Based on this, the association-level contrastive loss is defined as:

$$\mathcal{L}_{\mathrm{m}}\left(v_{\mathrm{i}}, e_{\mathrm{j}}\right) = -\log\frac{e^{R\left(v_i,e_j\right)/\tau}}{e^{R\left(v_i,e_j\right)/\tau}+\sum_{k:i\notin k}e^{R(v_i,e_k)/\tau}} - \log\frac{e^{R\left(v_i,e_j\right)/\tau}}{e^{R\left(v_i,e_j\right)/\tau}+\sum_{k:k\notin j}e^{R\left(v_k,e_j\right)/\tau}} \tag{22}$$

Considering symmetry, the final association-level loss is:

$$\widehat{\mathcal{L}_{m}} = \frac{1}{2|K|}\sum_{i=1}^{|V|}\sum_{j=1}^{|E|}[h_{ij} = 1]\{\mathcal{L}_{m}(v_{1,i}, e_{2,j}) + \mathcal{L}_{m}(v_{2,i}, e_{1,j})\} \tag{23}$$

Finally, the overall loss function for this module is a weighted sum of the three losses:

$$\mathcal{L}_c = \widehat{\mathcal{L}_\mathrm{n}} + 0.5 * \widehat{\mathcal{L}_w} + \widehat{\mathcal{L}_m} \tag{24}$$

3.6 Emotion Classification Layer

After multi-level hypergraph contrastive learning, we aggregate the node features from the two hypergraphs by summing them. Then, the features of the three modalities for the same utterance are concatenated to obtain the final feature representation of the utterance. These fused features are subsequently fed into a classifier composed of fully connected layers to predict the corresponding emotion category. The entire process can be formally expressed as follows:

$$\begin{aligned} p_i &= (v_{1,i}^t + v_{2,i}^t) \oplus (v_{1,i}^a + v_{2,i}^a) \oplus (v_{1,i}^v + v_{2,i}^v) \\ P_i &= softmax(\boldsymbol{W}_2 ReLU(p_i) + b_1) \\ y_i &= argmax(P_i) \end{aligned} \tag{25}$$

where $\boldsymbol{W}_2$ is the weight matrix, b_1 is the bias term, and y_i is the predicted label.
Loss: We employ a cross-entropy loss with L2 regularization for classification, defined as:

$$\mathcal{L}_{erc} = -\frac{1}{\sum_{m=1}^{N} s(m)} \sum_{i=1}^{N} \sum_{j=1}^{s(i)} \log P_{i,j} \left[Y_{i,j}\right] + \lambda \|\theta\|_2 \tag{26}$$

where P and Y represent the predicted and true label probabilities, respectively, N is the number of dialogues, and $s(m)$ is the number of utterances in dialogue m. Combining the multimodal representation loss $\mathcal{L}_o$ from Equation (9), the hypergraph reconstruction loss $\mathcal{L}_g$ from Equation (14), and the multi-level contrastive learning loss $\mathcal{L}_c$ from Equation (24), the final loss function is:

$$\mathcal{L}_{all} = \mathcal{L}_{erc} + \lambda_o \mathcal{L}_o + \lambda_g \mathcal{L}_g + \lambda_c \mathcal{L}_c \tag{27}$$

Where λ_o, λ_g, and λ_c are the weighting parameters for the respective losses.

4 Experiments

4.1 Datasets

This paper evaluates the performance of modal on two widely used datasets for dialogue emotion recognition: IEMOCAP [20] and MELD [21]. The IEMOCAP dataset consists of 151 dyadic dialogues from 10 speakers, comprising a total of 7,433 utterances. Each utterance is annotated with one of six emotion labels: happy, sad, neutral, angry, excited, and frustrated. For experiments, 120 dialogues are used for training and validation, while the remaining dialogues are reserved for testing. MELD is a multi-party dialogue dataset collected from the TV series *Friends*. It contains 1,433 dialogues and 13,708 utterances, with each utterance labeled as one of seven emotion categories: neutral, happy, surprise, sad, anger, disgust, and fear. For experiments, 1,153 dialogues are used for training and validation, and 280 dialogues are used for testing. For both datasets, this paper utilizes information from three modalities: audio, visual, and text, for emotion recognition.

Table 1. Details of hyper-parameters for HyperMoCL

Dataset	Batch size	Dropout	Learning rate	λ_c	λ_o	λ_g	d_m
IEMOCAP	24	0.7	0.0001	0.1	0.1	0.8	512
MELD	12	0.4	0.0001	1	1	0.5	256

4.2 Experimental Details and Baseline Models

Experimental Setup: The model is implemented using the PyTorch framework and trained/tested on an NVIDIA GeForce RTX 4090 machine. The experimental environment is configured with PyTorch 2.4.1 and CUDA 11.8. Adam is used as the optimizer for training, and additional hyperparameter settings are listed in Table 1. Model performance is evaluated using accuracy (ACC) and weighted F1 score (WF1), with the F1 score used as the metric for evaluating each individual emotion category.
Baseline Models: To evaluate our model, we compare it with the following state-of-the-art multimodal emotion recognition models, categorized into three groups:
Non-graph-based deep learning models: bc-LSTM[22], MFN[23], and DialogueRNN [1]. Graph-based deep learning models: DialogueGCN[24], MMGCN[3], GraphMFT [4], COGMEN[25], and GraphCFC[26]. Hypergraph-based deep learning models: HAUCL [6] and M3NET [5].

Table 2. Overall performance of various methods. (*) denotes these baselines are reproduced from References [4, 6]

Method	IEMOCAP		MELD	
	Overall		Overall	
	Acc	WF1	Acc	WF1
BC-LSTM*	59.58	59.10	59.62	56.80
MFN*	60.14	60.32	59.93	57.29
DialogueRNN*	63.40	62.75	60.31	57.66
DialogueGCN*	65.54	65.04	58.62	56.36
MMGCN*	65.56	65.71	59.31	57.82
GraphMFT*	67.90	68.07	61.30	58.37
GraphCFC	68.76	68.31	61.32	58.66
M3NET	69.01	69.09	67.43	65.81
HAUCL	69.44	69.50	67.74	66.30
HyperMoCL	**70.43**	**70.28**	**68.12**	**66.99**

Table 3. F1 scores for six emotion categories of IEMOCAP.

Method	Emotion Categories of IEMOCAP (F1)					
	Happy	Sad	Neutral	Angry	Excited	Frustrated
BC-LSTM*	32.62	70.34	51.14	63.44	67.91	61.06
MFN*	47.19	72.49	55.38	63.04	64.52	61.91
DialogueRNN*	33.18	78.80	59.21	65.28	71.86	58.91
DialogueGCN*	47.10	80.88	58.71	66.08	70.97	61.21
MMGCN*	45.45	77.53	61.99	66.67	72.04	64.12
GraphMFT*	45.99	83.12	63.08	70.30	76.92	63.84
GraphCFC	40.32	84.65	65.26	70.89	78.46	65.33
M3NET	55.37	78.60	68.10	65.48	77.89	63.87
HAUCL	54.24	81.89	69.55	65.69	72.89	66.30
HyperMoCL	51.66	80.72	68.93	69.49	75.99	67.83

4.3 Experimental Results and Analysis

To validate the performance of HyperMoCL, we conduct extensive experiments on two widely used datasets, IEMOCAP and MELD, and compared the results with state-of-the-art models, as shown in Table 2. The experiments demonstrate that HyperMoCL achieves superior performance in terms of accuracy and weighted F1 score. Compared to the state-of-the-art hypergraph model HAUCL, HyperMoCL improves accuracy and weighted F1 score by 0.99% and 0.78%, respectively, on the IEMOCAP dataset, and by 0.38% and 0.69%, respectively, on the MELD dataset. Table 3 provides a detailed breakdown of the F1 scores for each emotion category on the IEMOCAP dataset.

Compared to sequence-based and graph-based methods, HyperMoCL enhances the modeling of long-range dependencies through its hypergraph structure, more effectively capturing high-order relationships among information. In contrast to existing hypergraph methods, the proposed multimodal representation learning module provides higher-quality modal features for hypergraph construction, while the multi-level contrastive learning framework comprehensively models the structural information of the hypergraph. In summary, the experimental results demonstrate that the proposed model, HyperMoCL, outperforms other baseline models on both the IEMOCAP and MELD datasets.

4.4 Parameter Sensitivity Study

To evaluate the impact of model parameters on performance, we conducted experiments on the IEMOCAP and MELD datasets by varying the number of layers K in the Transformer-Encoder and the number of layers L in the hypergraph convolution. The results indicate that setting the number of Transformer-Encoder layers to 1 allows the model to effectively capture contextual information while avoiding overfitting. Similarly, setting the number of hypergraph convolution layers to 1 yield optimal model performance. Increasing the number of layers leads to varying degrees of performance degradation, suggesting that additional layers not only increase computational complexity but also cause performance decline due to over-smoothing. Therefore, selecting an appropriate number of layers is crucial for model optimization.

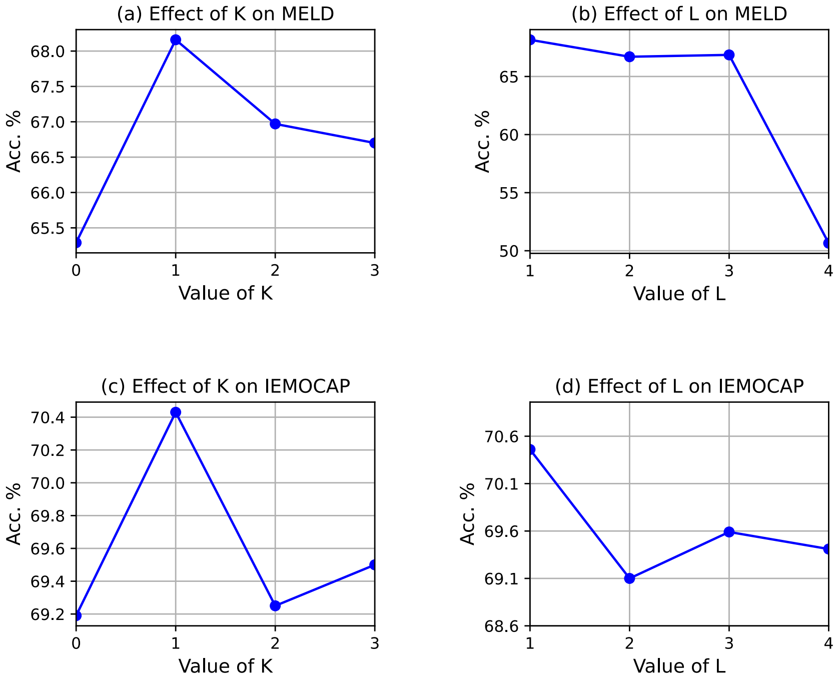

Figure 3. (a) and (b) show the effect of parameters K and L on the performance of MELD dataset, while Figure 3 (c) and (d) show the effect of K and L on the performance of IEMOCAP dataset. During the test, other parameters were fixed to the optimal values.

At the same time, we also conducted sensitivity analysis on the loss weights of contrastive learning at different levels and the loss weights λ_o, λ_g, λ_c of different modules on the MELD, and obtained the most appropriate loss weight parameters. The results are shown in Figures 4 and 5 respectively.

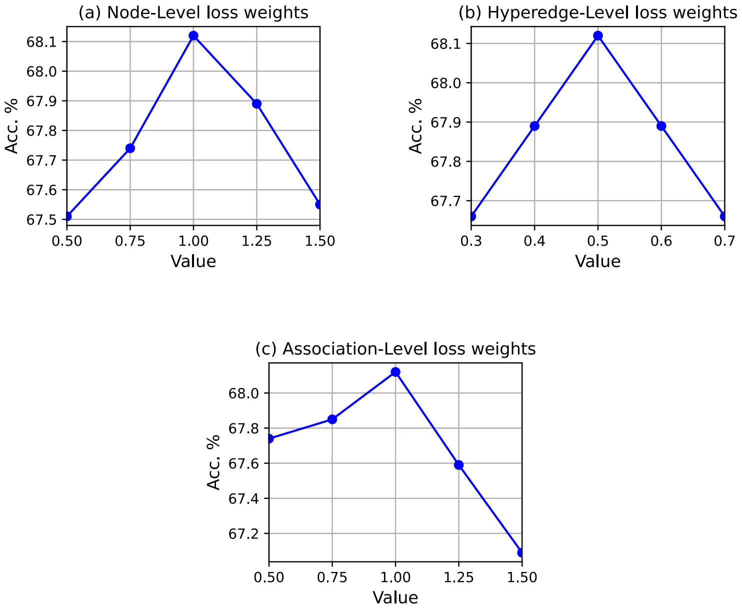

Figure 4. (a), (b), and (c) respectively show the impact of Node-level loss weight, Hyperedge-level loss weight, and Association-level loss weight on the performance of the model on the MELD dataset.

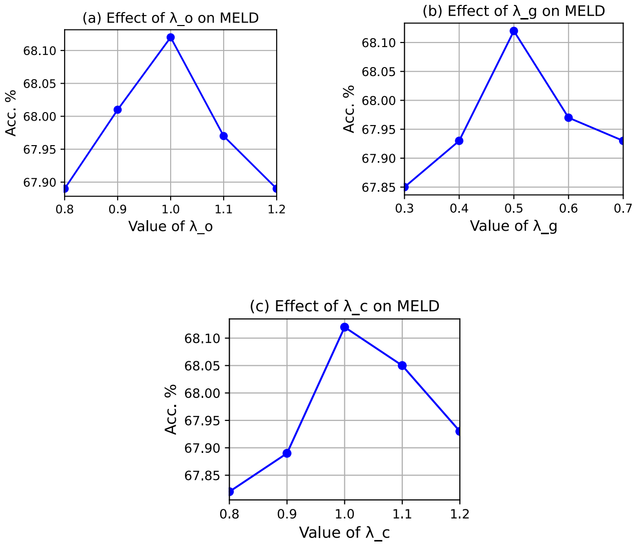

Figure 5. (a), (b), and (c) respectively show the impact of different module loss weights (λ_o, λ_g, λ_c) on the performance of the model on the MELD dataset.

4.5 Ablation Study

To validate the importance of each module in HyperMoCL, we conducted the following ablation experiments, with the results shown in Table 4:

(1) Contrastive Learning Module: The proposed multi-level hypergraph contrastive learning framework performs contrastive learning at three levels: node-level, hyperedge-level, and association-level, to fully exploit high-order dependencies in the data. We validated its effectiveness through the following configurations: 1) removing the contrastive learning module ("w/o CL" in table 4); 2) using only Node-Level contrastive learning ("w/o A_CL&H_CL"); and 3) using Node-Level and Hyperedge-Level contrastive learning, removing the Association-Level contrastive learning module ("w/o A_CL"). The experimental results demonstrate that the multi-level contrastive learning framework significantly enhances model performance.

(2) Multimodal Representation Learning Module: Table 4 compares the performance of our model with and without the multimodal representation learning module ("w/o MRL" in table 4). The results show that removing this module leads to accuracy drops of 1.67% and 0.80% on the IEMOCAP and MELD datasets, respectively. This confirms that the multimodal representation learning module effectively reduces noise and redundant information, improves information fusion, and thereby enhances model performance.

(3) Impact of Speaker Embedding Information: We conducted ablation experiments to evaluate the impact of speaker information by removing it from our model on both datasets (denoted as "w/o SP" in Table 4). The results show a performance drop of 0.74% on the IEMOCAP dataset and 1.87% on the MELD dataset in terms of accuracy. This performance degradation highlights the importance of incorporating speaker features, underscoring their critical role in enhancing the accuracy of emotion recognition models.

Table 4. Ablation experimental results of HyperMoCL.

Method	IEMOCAP		MELD	
	Acc	F1	Acc	F1
w/o CL	68.95	68.94	67.28	65.98
w/o A_CL&H_CL	69.87	69.89	67.36	66.47
w/o A_CL	69.99	69.97	67.82	66.82
w/o MRL	68.76	68.80	67.32	66.13
w/o SP	69.69	69.64	66.25	65.32
HyperMoCL	**70.43**	**70.28**	**68.12**	**66.99**

5 Conclusion

This paper proposes a joint framework model based on multimodal representation learning and multi-level hypergraph contrastive learning to address the task of MERC. Our method introduces a multimodal representation learning module, which reduces noise and redundant information in multimodal data, providing higher-quality modal representations for hypergraph construction. Additionally, we propose a multi-level hypergraph contrastive learning framework to comprehensively model the hypergraph structure, enhancing the model's robustness and feature discriminability. Experimental results on the IEMOCAP and MELD datasets demonstrate the superior performance of HyperMoCL. In future work, we will continue to explore the application of hypergraph models in emotion recognition with missing modalities and investigate integrating our model with recent pre-trained models.

References

1. Majumder, Navonil, et al. "Dialoguernn: An attentive rnn for emotion detection in conversations." *Proceedings of the AAAI conference on artificial intelligence*. Vol. 33. No. 01. 2019.Author, F., Author, S.: Title of a proceedings paper. In: Editor, F., Editor, S. (eds.) CONFERENCE 2016, LNCS, vol. 9999, pp. 1–13. Springer, Heidelberg (2016)
2. Tsai, Yao-Hung Hubert, et al. "Multimodal transformer for unaligned multimodal language sequences." *Proceedings of the conference. Association for computational linguistics. Meeting*. Vol. 2019. 2019.Author, F.: Contribution title. In: 9th International Proceedings on Proceedings, pp. 1–2. Publisher, Location (2010)
3. Hu, Jingwen, et al. "MMGCN: Multimodal fusion via deep graph convolution network for emotion recognition in conversation." *arXiv preprint arXiv:2107.06779* (2021).
4. Li, Jiang, et al. "GraphMFT: A graph network based multimodal fusion technique for emotion recognition in conversation." *Neurocomputing* 550 (2023): 126427.
5. Chen, Feiyu, et al. "Multivariate, multi-frequency and multimodal: Rethinking graph neural networks for emotion recognition in conversation." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.
6. Yi, Zijian, et al. "Multimodal Fusion via Hypergraph Autoencoder and Contrastive Learning for Emotion Recognition in Conversation." *Proceedings of the 32nd ACM International Conference on Multimedia*. 2024.
7. Xia, Lianghao, Chao Huang, and Chuxu Zhang. "Self-supervised hypergraph transformer for recommender systems." *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*. 2022.
8. Shi, Heyuan, et al. "Hypergraph-induced convolutional networks for visual classification." *IEEE transactions on neural networks and learning systems* 30.10 (2018): 2963-2972.
9. Huang, Jian, et al. "Dynamic hypergraph convolutional network for multimodal sentiment analysis." *Neurocomputing* 565 (2024): 126992.
10. Feng, Yifan, et al. "Hypergraph neural networks." *Proceedings of the AAAI conference on artificial intelligence*. Vol. 33. No. 01. 2019.
11. Jiang, Jianwen, et al. "Dynamic hypergraph neural networks." *IJCAI*. 2019.
12. Lee, Dongjin, and Kijung Shin. "I'm me, we're us, and i'm us: Tri-directional contrastive learning on hypergraphs." *Proceedings of the AAAI conference on artificial intelligence*. Vol. 37. No. 7. 2023.

13. Qian, Yiyue, et al. "Dual-level Hypergraph Contrastive Learning with Adaptive Temperature Enhancement." *Companion Proceedings of the ACM Web Conference 2024*. 2024.
14. Liu, Yinhan, et al. "Roberta: A robustly optimized bert pretraining approach." *arXiv preprint arXiv:1907.11692* (2019).
15. Eyben, Florian, Martin Wöllmer, and Björn Schuller. "Opensmile: the munich versatile and fast open-source audio feature extractor." *Proceedings of the 18th ACM international conference on Multimedia*. 2010.
16. Huang, Gao, et al. "Densely connected convolutional networks." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.
17. Yang, Hao, et al. "Asymmetric 3d convolutional neural networks for action recognition." *Pattern recognition* 85 (2019): 1-12.
18. Hazarika, Devamanyu, Roger Zimmermann, and Soujanya Poria. "Misa: Modality-invariant and-specific representations for multimodal sentiment analysis." *Proceedings of the 28th ACM international conference on multimedia*. 2020.
19. Zellinger, Werner, et al. "Central moment discrepancy (CMD) for domain-invariant representation learning." *arXiv preprint arXiv:1702.08811* (2017).
20. Busso, Carlos, et al. "IEMOCAP: Interactive emotional dyadic motion capture database." *Language resources and evaluation* 42 (2008): 335-359.
21. Poria, Soujanya, et al. "Meld: A multimodal multi-party dataset for emotion recognition in conversations." *arXiv preprint arXiv:1810.02508* (2018).
22. Poria, Soujanya, et al. "Context-dependent sentiment analysis in user-generated videos." *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)*. 2017.
23. Zadeh, Amir, et al. "Memory fusion network for multi-view sequential learning." *Proceedings of the AAAI conference on artificial intelligence*. Vol. 32. No. 1. 2018.
24. Hu, Dou, Lingwei Wei, and Xiaoyong Huai. "Dialoguecrn: Contextual reasoning networks for emotion recognition in conversations." *arXiv preprint arXiv:2106.01978* (2021).
25. Joshi, Abhinav, et al. "COGMEN: COntextualized GNN based multimodal emotion recognitioN." *arXiv preprint arXiv:2205.02455* (2022).
26. Li, Jiang, et al. "GraphCFC: A directed graph based cross-modal feature complementation approach for multimodal conversational emotion recognition." *IEEE Transactions on Multimedia* 26 (2023): 77-89.