Revisiting Multimodal Emotion Recognition in Conversation from the Perspective of Graph Spectrum

Wei Ai[1], **Fuchen Zhang**[1], **Yuntao Shou**[1], **Tao Meng**[1*], **Haowen Chen**[2*], **Keqin Li**[3]

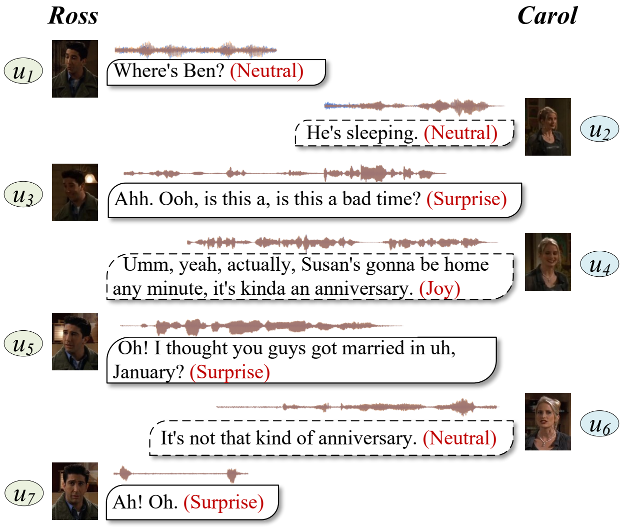
[1] College of Computer and Mathematics, Central South University of Forestry and Technology, 410004, China
[2] College of Computer Science and Electronic Engineering, Hunan University, 410082, China
[3] Department of Computer Science, State University of New York, 12561, USA
{aiwei, mengtao, hwchen}@hnu.edu.cn, fuchen.zhang@csuft.edu.cn, shouyuntao@stu.xjtu.edu.cn, lik@newpaltz.edu

Abstract

Efficiently capturing consistent and complementary semantic features in context is crucial for Multimodal Emotion Recognition in Conversations (MERC). However, limited by the over-smoothing or low-pass filtering characteristics of spatial graph neural networks, are insufficient to accurately capture the long-distance consistency low-frequency information and complementarity high-frequency information of the utterances. To this end, this paper revisits the task of MERC from the perspective of the graph spectrum and proposes a Graph-Spectrum-based Multimodal Consistency and Complementary collaborative learning framework GS-MCC. First, GS-MCC uses a sliding window to construct a multimodal interaction graph to model conversational relationships and designs efficient Fourier graph operators (FGO) to extract long-distance high-frequency and low-frequency information, respectively. FGO can be stacked in multiple layers, which can effectively alleviate the over-smoothing problem. Then, GS-MCC uses contrastive learning to construct self-supervised signals that reflect complementarity and consistent semantic collaboration with high and low-frequency signals, thereby improving the ability of high and low-frequency information to reflect genuine emotions. Finally, GS-MCC inputs the coordinated high and low-frequency information into the MLP network and softmax function for emotion prediction. Extensive experiments have proven the superiority of the GS-MCC architecture proposed in this paper on two benchmark data sets.

Introduction

With the continuous development of Human-Computer Interaction (HCI), the multimodal emotion recognition task in conversation (MERC) has recently received extensive research attention (Majumder et al. 2019; Ghosal et al. 2019; Ai et al. 2024). MERC aims to identify the emotional state of each utterance using textual, acoustic, and visual information in the conversational context (Lian et al. 2023; Yang et al. 2024), which is crucial for multimodal conversational understanding and an essential component for building intelligent HCI systems (Hu et al. 2021; Mai et al. 2022). As shown in Fig. 1, MERC needs to recognize the emotion of each multimodal utterance in the conversation.



Where's Ben? (Neutral)

He's sleeping. (Neutral)

Ahh. Ooh, is this a, is this a bad time? (Surprise)

Umm, yeah, actually, Susan's gonna be home any minute, it's kinda an anniversary. (Joy)

Oh! I thought you guys got married in uh, January? (Surprise)

It's not that kind of anniversary. (Neutral)

Ah! Oh. (Surprise)

Figure 1: An example of a multimodal conversation from the MELD dataset. MERC aims to identify each utterance's emotion label (e.g., *Neutral, Surprise, Joy*).

Unlike traditional unimodal or non-conversational emotion recognition (Gerczuk et al. 2021; Deng and Ren 2021), MERC requires joint conversational context and multimodal information modeling to achieve consistency and complementary semantic capture within and between modalities (Zhang et al. 2024). Fig. 1 gives an example of a multimodal conversation between two people, Ross and Carol, from the MELD dataset. As shown in utterance u_4, Carol has a "joy" emotion, which is vaguely reflected in textual features but more evident in visual or auditory features reflecting the complementary semantics between modalities. In addition, it is difficult to identify the emotion of "Surprise" from the utterance u_7 alone. However, due to the potential consistency of conversational emotions, it can be accurately inferred based on previous utterances. Therefore, the key to multimodal conversational emotion recognition is to capture the consistency and complementary semantics between multimodal information by utilizing the conversational context and emotional dependence between speakers to reveal the speaker's genuine emotion.

[*] Corresponding author

The current mainstream research method uses the Transformer (Lian, Liu, and Tao 2021; Ma et al. 2023; Zou, Huang, and Shen 2023) or GNN (Li et al. 2023c,a; Tu et al. 2024) architecture to model the MERC task. Transformer-based methods mainly learn complex semantic information between multimodal and conversational contexts from global sequence modeling. For example, CTNet (Lian, Liu, and Tao 2021) builds a single Transformer and cross Transformer to capture long-distance context dependencies and realize intra-module and inter-module information interaction to achieve multimodal conversational emotion recognition. Although transformer-based methods have made progress from the perspective of global utterance sequence modeling, this paradigm underestimates the complex emotional interactions between multimodal utterances (Tu et al. 2024) and ignores the multiple relationships between utterances (Chen et al. 2023), which limits the model's emotion recognition performance.

Benefitting from GNN's ability to mine and represent complex relationships (Yin et al. 2023, 2024), recent GNN-based methods (Ai et al. 2024; Hu et al. 2021; Li et al. 2023b) have made significant progress in the MERC task. For instance, MMGCN (Hu et al. 2021) fully connects all utterance nodes of the same modality and connects different modal nodes of the same utterance to build a heterogeneous graph to model the complex semantic relationships between multimodal utterances, then uses a deep spectral domain GNN to capture long-distance contextual information to achieve multimodal conversational emotion recognition. Although these GNN-based methods show promising performance, they still have some common limitations:

(1) Insufficient long-distance dependence perception. Considerable methods (Ghosal et al. 2019; Ai et al. 2024; Li et al. 2023c) using sliding windows to limit the length of fully connected utterances and then using GNN to learn multimodal utterance representations to achieve emotion recognition. However, limited by the over-smoothing characteristics of GNN (Liu et al. 2022; Yi et al. 2024), usually only two layers can be stacked for capturing semantic information, making it difficult for these methods to capture long-distance emotional dependencies. Although the method (Hu et al. 2021; Chen et al. 2023) without a sliding window can enhance the capture of long-distance dependencies, it will cause many nodes with the non-same emotions in the neighborhood, which is not conducive to the representation learning of GNN and puts enormous performance pressure on GNN. Therefore, previous GNN-based methods still have limitations in long-distance dependency capture.

(2) Underutilization of high-frequency features. Many studies have shown that GNN has low-pass filtering characteristics (Nt and Maehara 2019; Chang et al. 2021; Yin et al. 2022), which mainly obtain node representation by aggregating the consistency features of the neighborhood (low-frequency information) and suppressing the dissimilarity features of the neighborhood (high-frequency information). However, consistency and dissimilarity features are equally important in the MERC task. When specific modalities express less obvious emotions, information from other modalities is needed to compensate, thereby revealing the speaker's genuine emotions. Inspired by this, M^{3}Net (Chen et al. 2023) tried to use high-frequency information to improve the MERC task and improved the emotion recognition effect by directly fusing high- and low-frequency features. However, essential differences exist between high and low-frequency features, and direct fusion cannot establish efficient collaboration. Thus, previous GNN-based methods still have limitations in utilizing and collaborating high and low-frequency features.

Inspired by the above analysis, we propose a Graph-Spectrum-based Multimodal Consistency and Complementary feature collaboration framework GS-MCC. The contributions of our work are summarized as follows:

- We propose an efficient long-distance information learning module that designs Fourier graph operators to build a mixed-layer GNN to capture high- and low-frequency information to obtain consistency and complementary semantic dependencies in multimodal conversational contexts.
- We propose an efficient high- and low-frequency information collaboration module that uses contrastive learning to construct self-supervised signals that reflect the collaboration of high- and low-frequency information in terms of complementarity and consistent semantics and improves the ability to distinguish emotions between different frequency information.
- We conducted extensive comparative and ablation experiments on two benchmark data sets, IEMOCAP and MELD. The results show that our proposed method can efficiently capture long-distance context dependencies and improve the performance of MERC.

Related Work

Multimodal conversational context feature capture. In early work, the MERC task mainly adopted GRU (Majumder et al. 2019) or LSTM (Poria et al. 2017) to capture multimodal information in the conversational context. For example, *Poria et al.* (Poria et al. 2017) proposed a multimodal conversation emotion recognition model based on Bidirectional Long Short-Term Memory (Bi-LSTM), which captures multimodal contextual information at each time step to understand conversational context relationships in sequence data better. Although methods based on GRU or LSTM can model multimodal conversation context, they cannot capture long-distance information dependencies due to limited memory capabilities. For instance, *Ma et al.* (Ma et al. 2023) used intra-modal and inter-modal Transformers to capture semantic information in a multimodal conversation context and designed a hierarchical gating mechanism to achieve the fusion of multimodal features. Although Transformer-based methods can capture long-distance semantic information through global sequence modeling, they underestimate the complexity of multimodal dialogue semantics. Due to the superiority of GNN in modeling complex relationships, most existing research chooses to use GNN for global semantic capture and has achieved remarkable results. For example, *Li et al.* (Li et al. 2023c) proposed directed Graph-based Cross-modal Feature Complementation (GraphCFC), which alleviates the heterogeneity gap problem in multimodal fusion by utilizing multiple subspace extractors and pairwise cross-modal complementation strategies. In addition, speaker information is vital in emotion recognition because emotions are usually subjective and individual experiences. Therefore, *Ren et al.* (Ren et al. 2021) built a graph model to incorporate conversational context information and speaker dependencies, and then introduced a multi-head attention mechanism to explore potential connections between speakers.

Multimodal conversational context feature fusion. Choosing an appropriate multimodal feature fusion strategy is another crucial step in multimodal dialogue emotion recognition (Chudasama et al. 2022; Zou, Huang, and Shen 2023). For example, *Zadeh et al.* (Zadeh et al. 2017) proposed Tensor Fusion Network (TFN), has advantages in processing higher-order data structures (such as multidimensional arrays) and is therefore better able to preserve relationships between data when integrating multimodal information. Furthermore, contrastive learning has attracted increasing research attention due to its powerful ability to obtain meaningful representations through alignment fusion. *Wang et al.* (Wang et al. 2022) proposed a multimodal feature fusion framework based on contrastive learning. The framework first improves the ability to capture emotional features through contrastive learning and then uses an attention mechanism to achieve the fusion of multimodal features. Although multimodal conversational emotion recognition has made significant progress by modeling contextual semantic information and feature fusion, the critical role of high-frequency information in MERC has been ignored. To this end, *Hu et al.* (Hu et al. 2021) proposed a Multimodal Fusion Graph Convolution Network (MMGCN). MMGCN can not only capture high and low-frequency information in multimodal conversations, but also utilizes speaker information to model inter-speaker and intra-speaker dependencies. Similarly, *Chen et al.* (Chen et al. 2023) modeled MERC from multivariate information and high- and low-frequency information, further improving the effect of multimodal conversational emotion recognition. Nevertheless, as discussed earlier, these methods do not profoundly explore the uses of high and low-frequency signals, ignoring the consistency and complementary synergy between them.

Preliminary

Multi-modal Feature Extraction

Consistent with previous work (Kim and Vossen 2021; Shen et al. 2021; Chudasama et al. 2022), we employ RoBERTa (Liu et al. 2019), openSMILE (Eyben, Wöllmer, and Schuller 2010) and 3D-CNN (Ji et al. 2012) models for text, audio, and vision feature extraction, yielding respective embeddings φ_t, φ_a, and φ_v.

Speaker Information Embedding

Inspired by previous work (Hu et al. 2021; Chen et al. 2023; Zhang et al. 2024), we incorporate speaker information into each unimodal utterance to obtain an unimodal representation of context and speaker information. The embedding of the i-th speaker is as follows:

$$S_i = W_{speaker} s_i, \tag{1}$$

where $W_{speaker}$ is the trainable weight. In addition, to obtain higher-order feature representation, we utilize bidirectional Gated Recurrent Units (GRU) to encode conversational text features. The specific encoding calculation is as follows:

$$\begin{aligned} u_t &= \overleftrightarrow{GRU}(\varphi_t, u_t^{(+,-)1}), \\ u_a &= W_a \varphi_a + b_a, \\ u_v &= W_v \varphi_v + b_v, \end{aligned} \tag{2}$$

where W_a, b_a, W_v and b_v are the learnable parameters of the auditory and visual encoders, respectively. We then add speaker embeddings to obtain speaker- and context-aware unimodal representations:

$$x_m = u_m + S_i, \quad m \in \{t, a, v\}, \tag{3}$$

where t, a, v represent text, audio, and vision modal, respectively.

Methodology

The proposed Graph-Spectrum-based Multimodal Consistency and Complementary collaborative learning framework GS-MCC contains five modules: feature encoding, multimodal interaction graph construction, Fourier graph neural network, contrastive learning, and emotion classification. The overall process of the GS-MCC is shown in the Fig. 2.

Multimodal Interaction Graph

Given a conversation sequence $U = \{u_1, ..., u_N\}$ with N multimodal utterances, under the restriction of the sliding window k, we can construct a multimodal interaction graph $G^k = \left(V^k, E^k, A^k, X^k\right)$, where the node $v \in V^k$ represents a single-modal utterance and the edge $e \in E^k$ represents two semantic interactive relationships between unimodal utterances, A^k is the adjacency matrix, and X^k is the feature matrix.

Nodes: We treat each modality in each utterance as an independent node, using text modal node x_t^i, auditory modal node x_a^i, and visual modal node x_v^i represents, and uses the corresponding features x_m^i to represent the initial embedding of the node.

Edges: We fully connect the nodes in the same mode within sliding window k. In addition, we connect different modal nodes of the same utterance to construct semantic interactions between modalities.

Fourier Graph Neural Network

Fourier Graph Operator. For a given multimodal interaction graph, $G^k = \left(V^k, E^k, A^k, X^k\right)$, where $A^k \in \mathbb{R}^{3N \times 3N}$ is the adjacency matrix, $X^k \in \mathbb{R}^{3N \times d}$ is the feature matrix, N is the number of multimodal utterances, and d is the dimension of the feature. According to FourierGNN (Yi et al. 2024), we can obtain the Green kernel $\kappa \in \mathbb{R}^{d \times d}$ that meets the conditions based on the adjacency matrix A^k and the weight matrix $W \in \mathbb{R}^{d \times d}$, which needs to satisfy the conditions $\kappa\left[i, j\right] = \kappa\left[i - j\right]$, $\kappa\left[i, j\right] = A_{ij}^k \circ W$, and i and j are

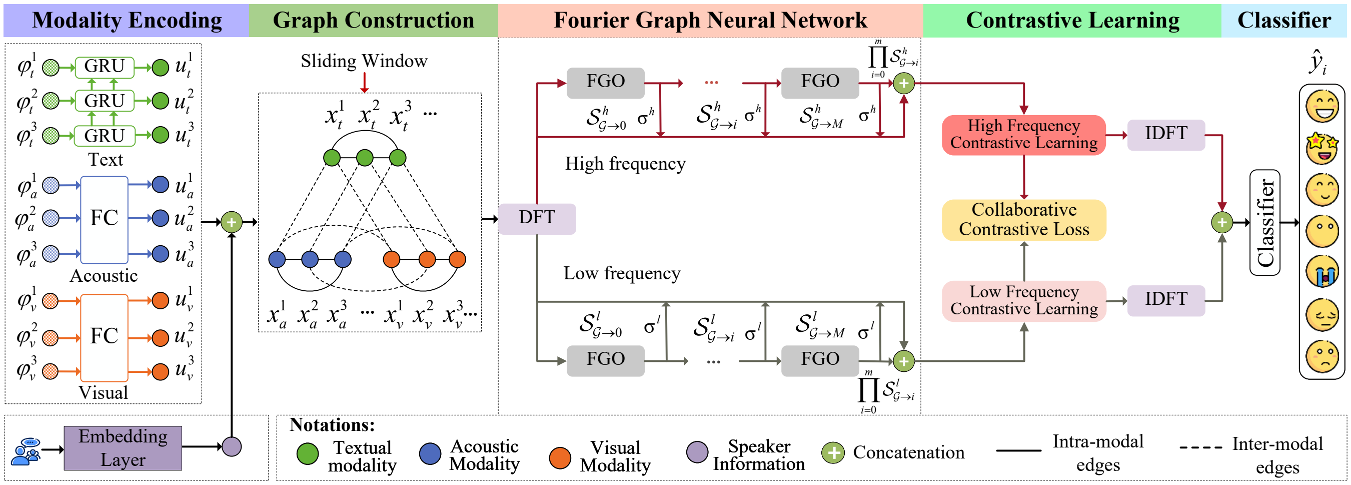


Figure 2: The overall architecture of the proposed model GS-MCC.

fall between 1 and $3N$. Based on the kernel κ, we can obtain the following Fourier graph operator $\mathcal{S}_{\mathcal{G}}$:

$$\mathcal{S}_{\mathcal{G}} = \mathcal{F}(\kappa) \in \mathbb{C}^{3N\times d\times d}, \tag{4}$$

where $\mathcal{F}$ is the Discrete Fourier Transform (DFT). According to the graph convolution theory, we can express the graph convolution operation as follows:

$$F_{\theta_{\mathcal{G}}}\left(X^k, A^k\right) = A^k X^k W = \mathcal{F}^{-1}\left(\mathcal{F}\left(X^k\right)\mathcal{F}(\kappa)\right), \tag{5}$$

where $\theta_{\mathcal{G}}$ is the learnable parameter and $\mathcal{F}^{-1}$ is the Inverse Discrete Fourier Transform (IDFT). According to the convolution theory and the conditions of FGO, we can expand the frequency domain term in Eq. (5) as follows:

$$\begin{aligned}\mathcal{F}\left(X^k\right)\mathcal{F}(\kappa) &= \mathcal{F}\left(\left(X^k * \kappa\right)[i]\right)\\ &= \mathcal{F}\left(X^k[j]\,\kappa[i-j]\right) = \mathcal{F}\left(X^k[j]\,\kappa[i,j]\right)\\ &= \mathcal{F}\left(A^k_{ij}X^k[j]\,W\right) = \mathcal{F}\left(A^k X^k W\right).\end{aligned} \tag{6}$$

As seen from Eq. (6) , the graph convolution operation is implemented through the product of FGO and features in the frequency domain. In addition, according to the convolution theory, the convolution of time-domain signals is equal to the product of frequency-domain signals. The product operation in the frequency domain only requires $O(N)$ time complexity, while the convolution operation in the time domain requires $O\left(N^2\right)$ time complexity. Therefore, an efficient graph neural network can be constructed based on the Fourier graph operator.

To efficiently capture high- and low-frequency information, we perform targeted optimization on FGO and use the high-pass and low-pass filters to extract complementary and consistent semantic information. The specific filter design is as follows:

$$L^l = I + D_{\mathcal{G}}^{-1/2} A^k D_{\mathcal{G}}^{-1/2}, \tag{7}$$

$$L^h = I - D_{\mathcal{G}}^{-1/2} A^k D_{\mathcal{G}}^{-1/2}, \tag{8}$$

where I is the identity matrix, $D_{\mathcal{G}}$ and A^k are the degree matrix and adjacency matrix of the multimodal interaction graph, respectively, and L^l and L^h are the low-pass and high-pass filters, respectively. Based on low-pass and high-pass filters, we can obtain the following low and high-frequency Green kernel and graph Fourier operator:

$$\kappa^{l/h}[i,j] = L^{l/h}_{ij} \circ W, \tag{9}$$

$$\mathcal{S}^{l/h}_{\mathcal{G}} = \mathcal{F}\left(\kappa^{l/h}\right). \tag{10}$$

Finally, we can build an M-layer Fourier graph neural network based on these efficient Fourier graph operators to capture long-distance high and low-frequency dependency information in multimodal interaction graphs:

$$F^{l/h}_{\theta_{\mathcal{G}}}\left(X^k, A^k\right) = \sum_{m=0}^{M} \sigma\left(\mathcal{F}(X^k)\mathcal{S}^{l/h}_{\mathcal{G}\Rightarrow[0:m]} + b_{l/h}\right), \tag{11}$$

$$\mathcal{S}^{l/h}_{\mathcal{G}\Rightarrow[0:m]} = \prod_{i=0}^{m} \mathcal{S}^{l/h}_{\mathcal{G}\to i}, \tag{12}$$

where σ is the activation function, $b_{l/h}$ is the bias parameter, $\mathcal{S}^{l/h}_{\mathcal{G}\to i}$ is the FGO in the i-th layer, l, and h represent low and high frequencies respectively.

By stacking M layers of Fourier graph operators, our model can capture long-distance dependency information and obtain each node's low-frequency feature representation, x^l_m, and high-frequency feature representation, x^h_m, respectively.

Contrastive Learning

Low-frequency features reflect the trend of slow changes in emotion, while high-frequency features reflect the trend of rapid changes in emotion. To synergize these two features, we employ contrastive learning to build self-supervised signals to promote consistent and complementary semantics learning in multimodal utterances.

Inspired by the SpCo (Liu et al. 2022) method, increasing the frequency domain difference between two contrasting views can achieve better contrast learning effects. Unlike SpCo, our contrastive learning is performed directly in the frequency domain and does not rely on data augmentation to generate contrastive views. Specifically, we use a combination of low-frequency contrast learning and high-frequency contrast learning to promote the synergy of the two features. In addition, we only use the strategy of negative sample pairs far away from each other to increase the frequency domain difference between contrasting views and obtain better contrast learning effects.

IFCL: Low Frequency Contrastive Learning. FCL aims to use low-frequency samples as anchor nodes and all high-frequency nodes as negative samples to construct a self-supervised signal to increase the frequency domain difference between contrast views to obtain better contrast learning effects and promote consistent semantics and complementary semantics learning in multimodal conversations. For each low-frequency anchor node, the self-supervised contrast loss can be defined as:

$$\mathcal{L}_{IF} = -\frac{1}{\tau} + \log\left(e^{1/\tau} + \sum_{i=1}^{3N} e^{\left(\left(x_m^l\right)^T x_m^{hi-}\right)/\tau}\right), \quad (13)$$

where τ is the temperature coefficient, x_m^l is the low-frequency anchor node, and x_m^{hi-} is the i-th high-frequency negative sample.

HFCL: High Frequency Contrastive Learning. HFCL is similar to LFCL, except that HFCL uses high-frequency samples as anchor nodes and all low-frequency nodes as negative samples to construct a self-supervised signal to increase the frequency domain difference between contrasting views. The specific contrast loss can be defined as:

$$\mathcal{L}_{HF} = -\frac{1}{\tau} + \log\left(e^{1/\tau} + \sum_{i=1}^{3N} e^{\left(\left(x_m^h\right)^T x_m^{li-}\right)/\tau}\right), \quad (14)$$

where x_m^h is the high-frequency anchor node, and x_m^{li-} is the i-th low-frequency negative sample.

The overall contrastive learning loss is the sum of LFCL and HFCL, which can be expressed as $\mathcal{L}_{CL}$:

$$\mathcal{L}_{CL} = \mathcal{L}_{IF} + \mathcal{L}_{HF}. \quad (15)$$

Finally, we use the inverse discrete Fourier transform to convert the high and low-frequency features into time domain features and concatenation the two parts of features to obtain the final embedding representation of the uni-modal utterance node:

$$v_m = \mathrm{IDFT}\left(x_m^l\right) \oplus \mathrm{IDFT}\left(x_m^h\right), \quad (16)$$

where $m \in \{t, a, v\}$ represents any one of text, auditory and visual modalities.

Emotion Classifier

For modal utterance U_i, we concatenate the features of each modality for emotion classification.

$$U_i = v_t^i \oplus v_a^i \oplus v_v^i, \quad (17)$$

$$\tilde{U}_i = \mathrm{ReLU}(U_i), \quad (18)$$

$$\mathcal{P}_i = \mathrm{softmax}(W_u \tilde{U}_i + b_u), \quad (19)$$

$$\hat{y}_i = \mathrm{argmax}(\mathcal{P}_i[\tau]), \quad (20)$$

where W_u and b_u are learnable parameters, and $\hat{y}_i$ is the predicted emotion label of utterance U_i. Finally, we employ categorical cross-entropy loss and contrastive loss for model training.

Experiments

Experimental Details

Datasets and Evaluation Metrics: In our experiments, we used two benchmark multimodal datasets IEMOCAP (Busso et al. 2008) and MELD (Poria et al. 2019), which are widely used in multimodal emotion recognition. In addition, we record the F1 for each emotion category, as well as the overall weighted weighted average F1 (W-F1).

Baseline Methods: We compare several baselines on the IEMOCAP and MELD datasets, including bc-LSTM (Poria et al. 2017), and A-DMN (Xing, Mai, and Hu 2020) based on RNN architecture, DialogueGCN (Ghosal et al. 2019), LR-GCN (Ren et al. 2021), DER-GCN (Ai et al. 2024), MMGCN (Hu et al. 2021), AdaIGN (Tu et al. 2024), RGAT (Ishiwatari et al. 2020) and CoMPM (Lee and Lee 2022) based on GCN, EmoBERTa (Kim and Vossen 2021) and CT-Net (Lian, Liu, and Tao 2021) based on Transformer architecture.

Experimental Setup: All experiments are conducted using Python 3.8 and PyTorch 1.8 deep learning framework and performed on a single NVIDIA RTX 4090 24G GPU. Our model is trained using AdamW with a learning rate of 1e-5, cross-entropy as the loss function, and a batch size of 32. The optimal parameters of all models were obtained by performing parameter adjustment using the leave-one-out cross-validation method on the validation set. Specifically, in our experiments, the reported results are the averages of 10 runs with different weight initializations. The results are statistically significant (all $p < 0.05$) as determined by paired t-tests. Our code is publicly available at https://github.com/FuchenZhang/GS-MCC.

Comparison with Baselines

Table 1 show the emotion recognition effects of the proposed GS-MCC method and the baseline method on the IEMOCAP and MELD datasets, respectively. Specifically, on the IEMOCAP dataset, GS-MCC has the best emotion recognition effect, outperforming all comparison baselines, and is 3.2% better than AdaIGN on W-F1. In addition, GS-MCC also has significant improvements in F1 values in some emotion categories, such as "happy", "neutral", "excited" and "frustrated". Similarly, on the MELD data set, compared with all comparison baselines, GS-MCC also has the best emotion recognition effect, outperforming AdaIGN by 2.2% on W-F1. Furthermore, GS-MCC is optimal in F1 on "fear", "joy" and "disgust" emotion categories.

Experimental results demonstrate the effectiveness of GS-MCC. The performance improvement may be attributed to the proposed method's ability to fully utilize long-distance contextual semantic information from high- and low-frequency signals while avoiding the over-smoothing phenomenon of GCN. Furthermore, the number of model

Methods	IEMOCAP								MELD							
	Parmas.	Happy	Sad	Neutral	Angry	Excited	Frustrated	W-F1	Neutral	Surprise	Fear	Sadness	Joy	Disgust	Anger	W-F1
bc-LSTM (Poria et al. 2017)	1.28M	34.4	60.8	51.8	56.7	57.9	58.9	54.9	73.8	47.7	5.4	25.1	51.3	5.2	38.4	55.8
DialogueGCN (Ghosal et al. 2019)	12.92M	42.7	**84.5**	63.5	64.1	63.1	66.9	65.6	72.1	41.7	2.8	21.8	44.2	6.7	36.5	52.8
A-DMN (Xing, Mai, and Hu 2020)	-	50.6	76.8	62.9	56.5	77.9	55.7	64.3	78.9	55.3	8.6	24.9	57.4	3.4	40.9	60.4
RGAT (Ishiwatari et al. 2020),	15.28M	51.6	77.3	65.4	63.0	68.0	61.2	65.2	78.1	41.5	2.4	30.7	58.6	2.2	44.6	59.5
EmoBERTa (Kim and Vossen 2021)	499M	56.4	83.0	61.5	69.6	78.0	68.7	69.9	**82.5**	50.2	1.9	31.2	61.7	2.5	46.4	63.3
CTNet (Lian, Liu, and Tao 2021)	8.49M	51.3	79.9	65.8	67.2	78.7	58.8	67.5	77.4	50.3	10.0	32.5	56.0	11.2	44.6	60.2
LR-GCN (Ren et al. 2021)	15.77M	55.5	79.1	63.8	69.0	74.0	68.9	69.0	80.8	57.1	0	36.9	65.8	11.0	54.7	65.6
MMGCN (Hu et al. 2021)	0.46M	42.3	78.7	61.7	69.0	74.3	62.3	66.2	77.1	53.9	0	17.7	56.9	0	42.6	59.4
CoMPM (Lee and Lee 2022)	-	60.7	82.2	63.0	59.9	78.2	59.5	67.3	82.0	49.2	2.9	32.3	61.5	2.8	45.8	63.0
AdaIGN (Tu et al. 2024)	6.3M	53.0	81.5	71.3	65.9	76.3	67.8	70.7	79.8	**60.5**	15.2	**43.7**	64.5	29.3	56.2	66.8
DER-GCN (Ai et al. 2024)	78.59M	58.8	79.8	61.5	**72.1**	73.3	67.8	68.8	80.6	51.0	10.4	41.5	64.3	10.3	**57.4**	65.5
GS-MCC (Our Model)	2.10M	**65.4**	81.2	**70.9**	70.8	**81.4**	**71.0**	**73.9**	81.8	58.3	**23.8**	35.8	**66.4**	**30.7**	54.4	**69.0**

Table 1: Comparison with other baselines on the IEMOCAP and MELD dataset. The best result in each column is in bold.

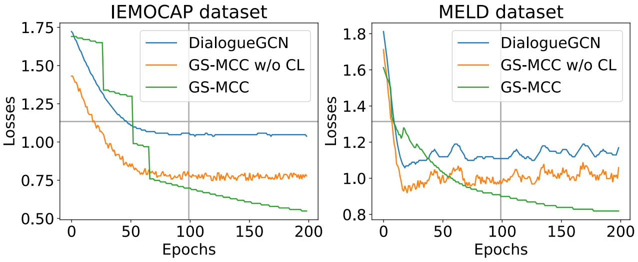


Figure 3: Loss trends on IEMOCAP and MELD datasets.

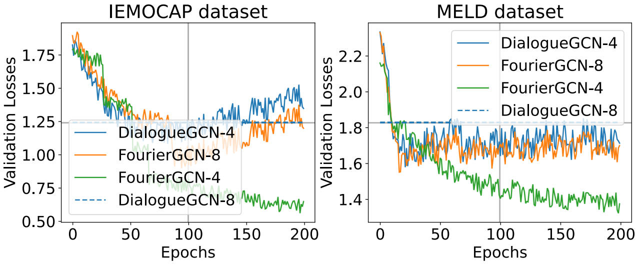


Figure 4: Over-smoothing phenomenon of the model.

parameters of the proposed GS-MCC is only 2.10M, which is far lower than DER-GCN and most other GCN-based emotion recognition methods. Experimental results also demonstrate the potential application of our method in efficient computing.

Trends of Losses

To verify the effectiveness of high- and low-frequency contrastive learning, we observed the loss trends of MMGCN, GS-MCC without contrastive learning, and GS-MCC on the IEMOCAP and MELD datasets. As shown in Fig. 3, GS-MCC has the best convergence, which is significantly better than MMGCN and GS-MCC without contrast learning. Experimental results demonstrate that the contrastive learning mechanism we proposed can coordinate the convergence of high-frequency and low-frequency features.

Over-smoothing Analysis

To verify the inhibitory effect of GS-MCC on over-smoothing, we stacked 4 layers and 8 layers of GCN respectively to explore the comparative effect of the model. The experimental comparison results are shown in Fig. 4. Specifically, when 8 layers are stacked, MMGCN-8 has a serious over-smoothing phenomenal, while GS-MCC-8 is not obvious. When 4 layers are stacked, the convergence of GS-MCC-4 is perfect, which is significantly better than MMGCN-4. The experimental results show that GS-MCC can alleviate the over-smoothing problem of the model to a certain extent.

Ablation Study

Ablation studies for SE, Fourier GNN, CL. Speaker embedding (SE), Fourier graph neural network (Fourier GNN) and contrastive learning (CL) are the three key components of our proposed multimodal emotion recognition model. We only remove one proposed module at a time to verify the effectiveness of the component. It is worth noting that when Fourier GNN is removed, we use DialogueGCN as the backbone of the model. From the emotion recognition results in Table 2 we conclude: (1) All the modules we proposed are useful, because no matter which proposed module is deleted, it will cause the emotion recognition performance of the model to decrease. (2) Speaker embedding has a relatively large impact on the emotion recognition performance of the model, because if the speaker embedding information is removed on the IEMOCAP and MELD data sets, the emotion recognition effect of the model will be greatly reduced. The experimental results show that the embedded information of the speaker is very necessary for the model to understand emotions. (3) On the IEMOCAP and MELD datasets, Fourier GNN is more important than contrastive learning. We speculate that this is because Fourier GNN can capture high and low frequency signals to provide more useful emotional semantic information, and the contrastive learning mechanism mainly assists Fourier GNN to better achieve complementary and consistent semantic information collaboration.

Ablation studies for multimodal features. We conduct ablation experiments on multimodal features to compare the

Methods	IEMOCAP		MELD	
	W-Acc.	W-F1	W-Acc.	W-F1
GS-MCC	**73.1**	**73.3**	**68.1**	**69.0**
w/o SE	70.3$_{(\downarrow 2.8)}$	70.6$_{(\downarrow 2.7)}$	65.4$_{(\downarrow 2.7)}$	64.6$_{(\downarrow 4.4)}$
w/o Fourier GCN	68.7$_{(\downarrow 4.4)}$	67.7$_{(\downarrow 5.6)}$	64.2$_{(\downarrow 3.9)}$	64.1$_{(\downarrow 4.9)}$
w/o CL	70.3$_{(\downarrow 2.8)}$	71.3$_{(\downarrow 2.0)}$	66.1$_{(\downarrow 2.0)}$	65.9$_{(\downarrow 3.1)}$

Table 2: Ablation studies for SE, Fourier GNN, CL on the IEMOCAP and MELD datasets.

performance of single-modal, bi-modal and tri-modal experimental results to explore the importance of each modality. The experimental results are listed in Table 3. We choose W-Acc and W-F1 as evaluation metrics. In single-modal experiments, text modality features achieved the best performance, which shows that text features play a decisive role in MERC. Video features have the worst emotion recognition effect. We speculate that video features have more noise signals, making it difficult for the model to learn effective emotional feature representation. In bi-modal experiments, all bi-modal emotion recognition effects are better than their own single-modal emotion recognition effects. The tri-modal emotion recognition effect is the best among all experiments. The improvement in performance may be attributed to the fact that the effective fusion of multimodal complementary semantic information can improve the feature representation ability of emotions.

Modality	IEMOCAP		MELD	
	W-Acc.	W-F1	W-Acc.	W-F1
T+A+V	**73.8**	**73.9**	**68.1**	**69.0**
T	66.3$_{(\downarrow 7.5)}$	66.0$_{(\downarrow 7.9)}$	63.7$_{(\downarrow 4.4)}$	62.5$_{(\downarrow 6.5)}$
A	57.7$_{(\downarrow 16.1)}$	58.1$_{(\downarrow 15.8)}$	53.8$_{(\downarrow 14.3)}$	53.4$_{(\downarrow 15.6)}$
V	50.4$_{(\downarrow 23.4)}$	50.5$_{(\downarrow 23.4)}$	41.4$_{(\downarrow 26.7)}$	42.3$_{(\downarrow 26.7)}$
T+A	71.6$_{(\downarrow 2.2)}$	71.0$_{(\downarrow 2.9)}$	66.3$_{(\downarrow 1.8)}$	65.9$_{(\downarrow 3.1)}$
T+V	69.5$_{(\downarrow 4.3)}$	68.7$_{(\downarrow 5.2)}$	64.2$_{(\downarrow 3.9)}$	64.1$_{(\downarrow 4.9)}$
V+A	63.7$_{(\downarrow 10.1)}$	63.0$_{(\downarrow 10.9)}$	54.6$_{(\downarrow 13.5)}$	53.4$_{(\downarrow 15.6)}$

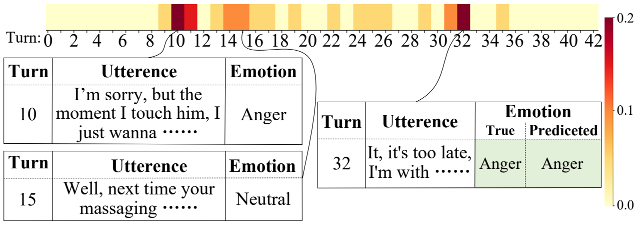
Table 3: The effect of our method using unimodal features and multimodal features, respectively.

Running Time

The traditional GCN method has a quadratic complexity in the time domain, while our method has only a log-linear complexity in the frequency domain. Table 4 shows the inference time of different baseline GCN methods on the IEMOCAP and MELD datasets. The experimental results show that our model GS-MCC has good inference performance and is significantly better than other baseline GCN methods. Specifically, even compared with the frequency domain method MMGCN or the unimodal time domain method DialogueGCN, GS-MCC still has an advantage in inference performance.

Methods	Running time (s)	
	IEMOCAP	MELD
DialogueGCN	58.1s	127.5s
RGAT	68.5s	146.3s
LR-GCN	87.7s	142.3s
MMGCN	93.7s	75.3s
DER-GCN	125.5s	189.7s
GS-MCC (Our Model)	**56.8s**	**71.2s**

Table 4: Inference time on IEMOCAP, and MELD dataset.



Turn	Utterence	Emotion
10	I’m sorry, but the moment I touch him, I just wanna ······	Anger

Turn	Utterence	Emotion
15	Well, next time your massaging ······	Neutral

Turn	Utterence	Emotion	
		True	Prediceted
32	It, it's too late, I'm with ······	Anger	Anger

Figure 5: An illustrative example of long-distance utterance dependency capture in the IEMOCAP dataset.

Long Distance Dependence

To verify the importance of capturing long-distance utterance dependencies, we tested our model GS-MCC on the IEMOCAP dataset. As shown in Fig. 5, the target utterance “32” achieves correct emotion prediction by capturing the semantic dependencies of the long-distance utterance “10” in the context.

Conclusions

In this paper, we rethink the problem of multimodal emotion recognition in conversation from the perspective of the graph spectrum, taking into account some shortcomings of existing work and innovations. Specifically, we propose a Graph-Spectrum-based Multimodal Consistency and Complementary feature collaboration framework GS-MCC. First, we combine sliding windows to build a multimodal interaction graph to model the conversational relationship between utterances and speakers. Secondly, we design efficient Fourier graph operators to capture long-distance utterances’ consistency and complementary semantic dependencies. Finally, we adopt contrastive learning and construct self-supervised signals with all negative samples to promote the collaboration of the two semantic information. Extensive experiments on two widely used benchmark datasets, IEMOCAP and MELD, demonstrate the effectiveness and efficiency of our proposed method.

Acknowledgments

This work is supported by National Natural Science Foundation of China (Grant No. 62472165), Excellent Young Scholars of Hunan Province of China (Grant No. 22B0275), and YueLuShan Center Industrial Innovation (Grant No. 2024YCII0110).

References

Ai, W.; Shou, Y.; Meng, T.; and Li, K. 2024. DER-GCN: Dialog and Event Relation-Aware Graph Convolutional Neural Network for Multimodal Dialog Emotion Recognition. *IEEE Transactions on Neural Networks and Learning Systems*.

Busso, C.; Bulut, M.; Lee, C.-C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J. N.; Lee, S.; and Narayanan, S. S. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42: 335–359.

Chang, H.; Rong, Y.; Xu, T.; Bian, Y.; Zhou, S.; Wang, X.; Huang, J.; and Zhu, W. 2021. Not all low-pass filters are robust in graph convolutional networks. *Advances in Neural Information Processing Systems*, 34: 25058–25071.

Chen, F.; Shao, J.; Zhu, S.; and Shen, H. T. 2023. Multivariate, multi-frequency and multimodal: Rethinking graph neural networks for emotion recognition in conversation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10761–10770.

Chudasama, V.; Kar, P.; Gudmalwar, A.; Shah, N.; Wasnik, P.; and Onoe, N. 2022. M2fnet: Multi-modal fusion network for emotion recognition in conversation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4652–4661.

Deng, J.; and Ren, F. 2021. A survey of textual emotion recognition and its challenges. *IEEE Transactions on Affective Computing*, 14(1): 49–67.

Eyben, F.; Wöllmer, M.; and Schuller, B. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, 1459–1462.

Gerczuk, M.; Amiriparian, S.; Ottl, S.; and Schuller, B. W. 2021. Emonet: A transfer learning framework for multi-corpus speech emotion recognition. *IEEE Transactions on Affective Computing*, 14(2): 1472–1487.

Ghosal, D.; Majumder, N.; Poria, S.; Chhaya, N.; and Gelbukh, A. 2019. DialogueGCN: A Graph Convolutional Neural Network for Emotion Recognition in Conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 154–164.

Hu, J.; Liu, Y.; Zhao, J.; and Jin, Q. 2021. MMGCN: Multimodal Fusion via Deep Graph Convolution Network for Emotion Recognition in Conversation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 5666–5675.

Ishiwatari, T.; Yasuda, Y.; Miyazaki, T.; and Goto, J. 2020. Relation-aware graph attention networks with relational position encodings for emotion recognition in conversations. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, 7360–7370.

Ji, S.; Xu, W.; Yang, M.; and Yu, K. 2012. 3D convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1): 221–231.

Kim, T.; and Vossen, P. 2021. Emoberta: Speaker-aware emotion recognition in conversation with Roberta. *Computing Research Repository-arXiv*, 2021: 1–7.

Lee, J.; and Lee, W. 2022. CoMPM: Context Modeling with Speaker's Pre-trained Memory Tracking for Emotion Recognition in Conversation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 5669–5679.

Li, B.; Fei, H.; Liao, L.; Zhao, Y.; Teng, C.; Chua, T.-S.; Ji, D.; and Li, F. 2023a. Revisiting disentanglement and fusion on modality and context in conversational multimodal emotion recognition. In *Proceedings of the 31st ACM International Conference on Multimedia*, 5923–5934.

Li, J.; Wang, X.; Lv, G.; and Zeng, Z. 2023b. GA2MIF: graph and attention based two-stage multi-source information fusion for conversational emotion detection. *IEEE Transactions on affective computing*.

Li, J.; Wang, X.; Lv, G.; and Zeng, Z. 2023c. Graphcfc: A directed graph based cross-modal feature complementation approach for multimodal conversational emotion recognition. *IEEE Transactions on Multimedia*.

Lian, Z.; Chen, L.; Sun, L.; Liu, B.; and Tao, J. 2023. Gcnet: Graph completion network for incomplete multimodal learning in conversation. *IEEE Transactions on pattern analysis and machine intelligence*.

Lian, Z.; Liu, B.; and Tao, J. 2021. CTNet: Conversational transformer network for emotion recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29: 985–1000.

Liu, N.; Wang, X.; Bo, D.; Shi, C.; and Pei, J. 2022. Revisiting graph contrastive learning from the perspective of graph spectrum. *Advances in Neural Information Processing Systems*, 35: 2972–2983.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ma, H.; Wang, J.; Lin, H.; Zhang, B.; Zhang, Y.; and Xu, B. 2023. A Transformer-Based Model With Self-Distillation for Multimodal Emotion Recognition in Conversations. *IEEE Transactions on Multimedia*.

Mai, S.; Zeng, Y.; Zheng, S.; and Hu, H. 2022. Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis. *IEEE Transactions on Affective Computing*.

Majumder, N.; Poria, S.; Hazarika, D.; Mihalcea, R.; Gelbukh, A.; and Cambria, E. 2019. Dialoguernn: An attentive rnn for emotion detection in conversations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 6818–6825.

Nt, H.; and Maehara, T. 2019. Revisiting graph neural networks: All we have is low-pass filters. *arXiv preprint arXiv:1905.09550*.

Poria, S.; Cambria, E.; Hazarika, D.; Majumder, N.; Zadeh, A.; and Morency, L.-P. 2017. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)*, 873–883.

Poria, S.; Hazarika, D.; Majumder, N.; Naik, G.; Cambria, E.; and Mihalcea, R. 2019. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 527–536.

Ren, M.; Huang, X.; Li, W.; Song, D.; and Nie, W. 2021. Lr-gcn: Latent relation-aware graph convolutional network for conversational emotion recognition. *IEEE Transactions on Multimedia*, 24: 4422–4432.

Shen, W.; Wu, S.; Yang, Y.; and Quan, X. 2021. Directed Acyclic Graph Network for Conversational Emotion Recognition. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1551–1560.

Tu, G.; Xie, T.; Liang, B.; Wang, H.; and Xu, R. 2024. Adaptive Graph Learning for Multimodal Conversational Emotion Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 19089–19097.

Wang, X.; Zhang, D.; Tan, H.-Z.; and Lee, D.-J. 2022. A self-fusion network based on contrastive learning for group emotion recognition. *IEEE Transactions on Computational Social Systems*, 10(2): 458–469.

Xing, S.; Mai, S.; and Hu, H. 2020. Adapted dynamic memory network for emotion recognition in conversation. *IEEE Transactions on Affective Computing*, 13(3): 1426–1439.

Yang, Z.; Li, X.; Cheng, Y.; Zhang, T.; and Wang, X. 2024. Emotion Recognition in Conversation Based on a Dynamic Complementary Graph Convolutional Network. *IEEE Transactions on Affective Computing*.

Yi, K.; Zhang, Q.; Fan, W.; He, H.; Hu, L.; Wang, P.; An, N.; Cao, L.; and Niu, Z. 2024. FourierGNN: Rethinking multivariate time series forecasting from a pure graph perspective. *Advances in Neural Information Processing Systems*, 36.

Yin, N.; Feng, F.; Luo, Z.; Zhang, X.; Wang, W.; Luo, X.; Chen, C.; and Hua, X.-S. 2022. Dynamic hypergraph convolutional network. In *2022 IEEE 38th International Conference on Data Engineering (ICDE)*, 1621–1634. IEEE.

Yin, N.; Shen, L.; Xiong, H.; Gu, B.; Chen, C.; Hua, X.-S.; Liu, S.; and Luo, X. 2023. Messages are never propagated alone: Collaborative hypergraph neural network for time-series forecasting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Yin, N.; Wang, M.; Chen, Z.; De Masi, G.; Xiong, H.; and Gu, B. 2024. Dynamic spiking graph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 16495–16503.

Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; and Morency, L.-P. 2017. Tensor Fusion Network for Multimodal Sentiment Analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 1103–1114.

Zhang, X.; Cui, W.; Hu, B.; and Li, Y. 2024. A Multi-Level Alignment and Cross-Modal Unified Semantic Graph Refinement Network for Conversational Emotion Recognition. *IEEE Transactions on Affective Computing*.

Zou, S.; Huang, X.; and Shen, X. 2023. Multimodal Prompt Transformer with Hybrid Contrastive Learning for Emotion Recognition in Conversation. In *Proceedings of the 31st ACM International Conference on Multimedia*, 5994–6003.