Contextual xLSTM-Based Multimodal Fusion for Conversational Emotion Recognition

Yupeng Qi[1,2], Mayire Ibrayim[1,2*], Turdi Tohti[1,2]

[1*]School of Computer Science and Technolog, Xinjiang University, Urumqi, 830017, Xinjiang, China.

[2] Xinjiang Key Laboratory of Signal Detection and Processing, Urumqi, 830017, Xinjiang, China.

*Corresponding author(s). E-mail(s): mayire401@xju.edu.cn;
Contributing authors: 107552203927@stu.xju.edu.cn; turdy@xju.edu.cn;

Abstract

In real-world dialogue systems, the ability to understand user emotions and engage in human-like interactions is of great significance. Emotion Recognition in Conversation (ERC) is one of the key approaches to achieving this goal and has attracted increasing attention. However, existing methods for ERC often fail to effectively model contextual information and exploit the complementarity of multimodal information. Few approaches are capable of fully capturing the complex correlations and mapping relationships between different modalities. Additionally, classifying minority classes and semantically similar classes remains a significant challenge. To address these issues, this paper proposes an attention-based contextual modeling and multimodal fusion network. The proposed method efficiently combines an extended LSTM network (xLSTM) with an attention mechanism to thoroughly model the contextual information generated during conversations. xLSTM is an enhanced LSTM unit featuring matrix memory and exponential gating mechanisms, which can better capture long-range dependencies and improve recognition performance. Furthermore, a Transformer-based modality encoder is employed to map features from different modalities into a shared feature space, enabling alignment and mutual enhancement among modalities. This facilitates both intra-modal and inter-modal information interaction, thereby maximizing the complementary advantages of multimodal data. A multimodal fusion module based on bidirectional multi-head cross-attention layers is then used to capture cross-modal mapping relationships among text, audio, and visual modalities, effectively integrating multimodal information. Extensive experiments conducted on two benchmark ERC datasets, IEMOCAP and MELD, demonstrate that the proposed method achieves weighted F1 scores of 75.21 and 69.78, respectively, outperforming the current state-of-the-art methods by 2.2 and 3.3 points. It also achieves accuracy rates of 80.59% and 69.16% on the two

datasets, representing improvements of 6.64% and 1.11%, respectively. The codes and models are available at:https://github.com/rhoqwomda/MFCRE

Keywords: Emotion recognition in conversation, Convolutional neural network, Multi-modal interaction, Multi-modal fusion

1 Introduction

Multimodal emotion analysis aims to identify an individual's emotional state by processing different types of data, such as video, audio, and text. Compared to unimodal approaches, multimodal research significantly enhances the understanding of human emotions by exploring the relationships between different modalities. Particularly, multimodal emotion recognition in dialogue, as a complex task within multimodal research, faces even greater challenges. Multilateral dialogues often involve multiple participants simultaneously expressing various emotions, making it crucial to integrate information from multiple modalities such as text, video, and audio to more comprehensively capture emotional expression. Among these modalities, visual information plays a key role, as it often provides direct clues for emotion prediction. Emotion recognition in dialogue has broad applications across various fields, including opinion mining [1], healthcare [2], the development of empathic dialogue systems [3], and behavior recognition in virtual environments [4].In recent years, the rapid development of deep learning technologies has provided strong technical support for multimodal emotion recognition. By leveraging methods such as multimodal Transformers, attention mechanisms, and interactive neural networks, researchers are able to more effectively model the interactions between different modalities. However, challenges remain in this field: the text modality may suffer from contextual ambiguity, sarcasm, and other complexities; the audio modality is limited by background noise, variations in speaking speed, and subtle emotional cues; while the video modality often faces issues such as lighting changes, occlusion, and the difficulty of recognizing subtle facial expressions.

This paper proposes an attention-based framework for extracting contextual information and perceiving the correlations among multimodal data. First, features are extracted from each modality, such as text, vision, and audio. For visual feature extraction, we introduce a feature extractor called VisExtNet, which is based on a multi-task cascaded convolutional network (MTCNN) [5] and the pre-trained ResNet-101 [6] model from VGGFace2 [7]. VisExtNet effectively captures visual cues from the speaker's emotional facial expressions in the video, while avoiding the modeling of irrelevant visual information in the scene. Next, we design a context information modeling module that combines attention mechanisms with xLSTM [8] to extract relevant contextual information from the conversation. Then, a modality encoder based on the Transformer architecture is used to facilitate both intra-modality and inter-modality information exchange, thereby enhancing the sequence representation ability of each modality. Subsequently, a multimodal fusion module based on

a bidirectional multi-head cross-attention layer [9] is employed to effectively integrate multimodal information, capture complex cross-modal correlations, and establish contextual mapping relationships among text, visual, and audio modalities.

The main contributions of our work are set out below:

- A contextual information modeling module was designed using xLSTM and the attention mechanism. The module utilizes the matrix memory mechanism in xLSTM that can capture more complex relationships and dependencies from the input data, and uses the attention mechanism to match relevant contextual information from the global memory, enabling the model to represent contextual information and long-term dependencies more comprehensively.
- A transformer-based modality encoder was designed to learn more efficient modality representations while enhancing both inter-modal and intra-modal information interaction. This enables the model to better leverage the complementarity of multimodal information, thereby enhancing the representation of individual modality sequences.
- Extensive experiments were conducted on the MELD and IEMOCAP benchmark datasets, and the results demonstrate that our proposed model achieved superior performance on both datasets. Moreover, the improvements were particularly notable for a few emotional categories and those with semantically similar emotions.

2 Related work

With the emergence of available dialogue datasets such as IEMOCAP[10], AVEC[11], and MELD[12], as well as Transformer-based architectures like CM-BERT[13] and MAF[14], recent studies have shown promising results in modeling interactions within specific modalities through self-attention and cross-attention mechanisms. Emotion recognition in dialogues has attracted widespread interest from researchers. Recent work typically adopts deep neural network approaches and focuses on modeling context-aware and speaker-sensitive dependencies. Based on whether speaker information is utilized, existing methods can be categorized into speaker-independent and speaker-dependent approaches.

Speaker-independent methods primarily focus on capturing contextual information within dialogues. HiGRU [15] consists of two gated recurrent units (GRUs) that are responsible for modeling the contextual relationships between words and discourse. Li et al.[16] proposed a dual-modality emotion analysis framework based on audio and video, where capsule networks and one-dimensional convolutional layers are introduced on both the video and audio modalities to enhance feature representation capabilities. They also designed a cross-modal attention interaction module to explicitly interact modality information, improving the fusion effect. AGHMN [17] employs a hierarchical memory network to strengthen discourse representation and combines attention-based GRU to model contextual information. Darren et al.[18] proposed a real-time facial feature extraction algorithm for emotion recognition through online games and metaverse avatars. MVN [19] models word-level and discourse-level dependencies in dialogue using a multi-view network. Chen et al.[20] introduced a Contrastive Translation and Hierarchical Fusion Network (CTHFNet) that explores complex relationships both

within and between modalities through cross-layer multimodal fusion networks, and captures modality interactions across different levels and categories using contrastive learning. In contrast, speaker-dependent methods place greater emphasis on modeling dependencies sensitive to both context and speaker information. DialogueRNN[21] employs three different GRUs to update speaker states, conversational context, and emotional states, respectively. However, it is limited by the forgetting mechanism when dealing with long-range dependencies in dialogue. DialogueGCN[22] utilizes a graph convolutional network to model speaker and dialogue sequence information. Although it captures the relationships among participants through a graph structure, it relies heavily on structural priors, making it less adaptable to diverse scenarios. HiTrans[23] consists of two hierarchical transformers designed to capture global contextual information and employs auxiliary tasks to model speaker-sensitive dependencies.

However, most studies primarily focus on the text modality, with less attention given to the potential contributions of other modalities. With the continuous improvement in multimodal recognition performance, some methods have started to concentrate on emotion recognition tasks in multimodal dialogue. For example, DialogueTRM [24] explores emotion features within and between modalities using hierarchical transformers and a multi-granularity interaction fusion module. MMGCN [25] employs fully connected graphs to model multimodal information and long-distance context, and encodes speaker-related information by introducing speaker embeddings. MM-DFN [26] designs a graph-based dynamic fusion module that effectively reduces redundant information and enhances modality complementarity. MMTr [27] preserves the integrity of the primary modality representation and uses multi-head attention mechanisms to strengthen the expression of weak modality features. UniMSE [28] performs modality fusion at the syntactic and semantic levels, incorporating inter-modal contrastive learning techniques to distinguish fused feature representations between samples. MultiEMO [29] captures the cross-modal mapping relationships between text, audio, and visual modalities through a bidirectional multi-head cross-attention layer, thus effectively integrating multimodal cues.

Additionally, ALMT [30] combines an adaptive hyper-modal (AHL) module to learn conflict-suppression representations from visual and audio modalities, guided by different linguistic features. It generates complementary joint representations through multimodal fusion to improve emotion analysis performance. FacialMMT [31] proposes a two-stage multimodal multi-task learning framework, which includes multimodal facial recognition, unsupervised facial clustering, and facial matching. By utilizing extracted facial sequences, this method designs a multimodal facial expression-aware emotion recognition model that enhances discourse-level emotion recognition performance through frame-level facial emotion distributions. GraphMFT [32] uses an improved graph attention network to capture both intra-modal contextual information and complementary inter-modal features. SMFNM [33] leverages additional unlabeled data to extract high-quality intra-modal representations and captures complementary information through cross-modal interactions, thereby enhancing the expression of audio modality features. Moreover, this method employs directed acyclic graphs and gated recurrent units to model dialogue context from both multimodal and primary

modality perspectives, ultimately fusing these two contextual features for emotion recognition. This paper primarily focuses on modeling contextual information in dialogue and facilitating information exchange both within and across modalities to further improve the performance of dialogue emotion recognition.

The application of multimodal dialogue emotion recognition is quite extensive. For example, it can be applied to session-based recommendation systems (SBRS)[34], enhancing user experience and enabling timely understanding of product popularity. It can also be integrated into certain recommendation systems such as DGAT-AEA [35], which adjusts edge features based on user-item-context relationships to improve the capture of complex interactions involving contextual information. Another example is the Deep GraphSAGE-based recommender system [36], which represents real-world scenarios by considering relationships between data and understanding them in a better way. Multimodal dialogue emotion recognition can also be applied to social recommendation systems[37], such as collaborative filtering recommender systems[38], where users can quickly access relevant resources based on their preferences. Moreover, emotion recognition research can be used in online information retrieval[39, 40], automatically generating article recommendations by analyzing users' preferences and historical behaviors.

3 Proposed method

The overall framework of the proposed model is shown in Figure 1, which consists of five core components: unimodal feature extraction, context modeling, modality encoders, multimodal fusion, and emotion classification. In the following sections, we will provide a detailed description of each module.

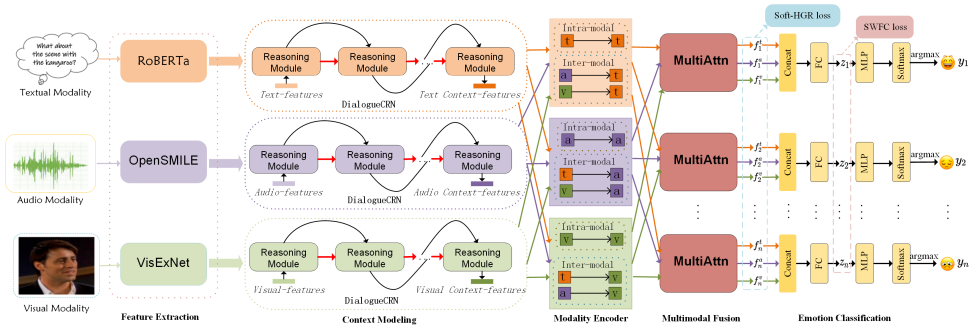

Fig. 1 General framework diagram of the model.

3.1 Feature Extraction

3.1.1 textual modal feature extraction

Current research typically employs two main paradigms to extract contextualized textual features: (1) the two-stage paradigm[41]: first, a pretrained language model is

used to capture local discourse-level textual features by inputting the text sequence. These local features are then passed to another transformer, which integrates contextual information from the dialogue to generate global dialogue-level textual features. (2) The single-stage paradigm[42]: in this approach, a single pretrained language model is fine-tuned to simultaneously capture both local discourse-level features and global dialogue-level textual features. To improve computational efficiency and simplify the process, we adopt the single-stage paradigm for feature extraction in the text modality. Specifically, to effectively encode speaker information, we prepend the speaker's name as a prefix to each textual utterance. This allows the model to identify and distinguish between different speakers. The input sequence for the i-th utterance consists of three main segments: the preceding contextual utterances, the current utterance, and the subsequent contextual utterances. In this way, the model is better able to capture and integrate contextual information between utterances. After being input into the pretrained RoBERTa model, these three segments are transformed into a 256-dimensional textual feature representation that effectively incorporates contextual information, aiding the model in better understanding the emotions and context within the dialogue.

3.1.2 audio modal feature extraction

For audio feature extraction, we use the open-source tool OpenSMILE, which is widely employed in the fields of audio processing and emotion analysis. OpenSMILE is a powerful feature extraction tool that can extract a variety of features from audio signals, including spectral features, energy features, timbre features, and other relevant characteristics of speech signals. One of its key advantages is its high flexibility, allowing users to customize the feature extraction process through configuration files to suit different research or application scenarios. Following the approach proposed by Majumder et al.[21], we used this tool to extract a total of 6373-dimensional audio feature representations for each utterance. These features encompass frequency, energy distribution, and other characteristics that may reflect emotional or tonal variations in the audio signal. To facilitate subsequent processing and analysis, we further reduced the dimensionality of these features using a fully connected neural network, compressing them from 6373 dimensions to 512 dimensions, resulting in a refined audio modality feature representation. This method effectively reduces redundant features while retaining the crucial audio information for emotion recognition, thereby enhancing the performance of the emotion recognition model.

3.1.3 visual modal feature extraction

Currently, many visual feature extraction methods for emotion recognition in dialogue typically encode not only the speaker's facial expressions but also capture scene information associated with each utterance. However, this approach may introduce redundancy in emotion recognition tasks, particularly concerning the visual features of the speaker. The main reason for this is that scene information does not have a strong correlation with the speaker's emotional state. In many cases, even within the same scene, emotional expressions can vary significantly, and the scene itself does not

provide substantial support for the transmission or perception of emotions. Therefore, scene information is often redundant for emotion recognition tasks and may even distract the model from focusing on crucial emotional cues.

To address this issue, this paper introduces a novel visual feature extraction method called VisExtNet [29]. This model effectively extracts the speaker's facial expression features while avoiding the introduction of irrelevant scene information, thereby more accurately capturing visual features relevant to emotion recognition. VisExtNet consists of two key components: first, MTCNN [5] is used to precisely detect the speaker's face in each frame of the image and perform face detection at multiple scales, ensuring accurate recognition of facial features even in complex environments; second, ResNet-101 [6], a deep neural network pretrained on the VGGFace2 [7] dataset, efficiently extracts facial expression features related to emotion, providing strong representational capability. By combining these two components, VisExtNet can effectively capture and process facial expression changes of the speaker while avoiding redundant scene information, thus offering precise visual information support for emotion recognition tasks. The specific architecture of VisExtNet is shown in Figure 2.

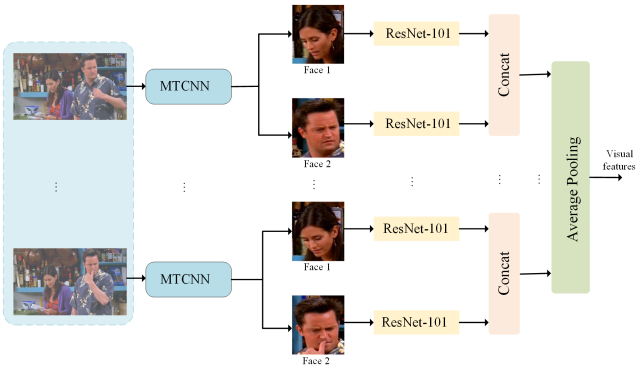

Fig. 2 VisExtNet visual feature extraction network.

The core of VisExtNet lies in effectively integrating facial expressions of speakers across multiple video frames to accurately extract visual information while avoiding the inclusion of irrelevant scene content. For each input dialogue video, VisExtNet performs feature extraction on 20 frames per dialogue segment to capture key visual features. The frame selection interval for each utterance is set to 20 frames to ensure coverage of dynamic changes and emotional fluctuations throughout the dialogue. Specifically, in each frame, MTCNN is first used to detect and locate all speaker facial regions. Each detected face image is then passed through a ResNet-101 network pretrained on the VGGFace2 dataset. This deep convolutional network extracts visual feature vectors rich in emotional information. These vectors capture variations in facial expressions and can effectively correspond to emotional states.

Specifically, in each frame of the image, MTCNN is first used to detect and locate the facial regions of all speakers in the image. MTCNN is capable of accurately identifying the speaker's face, even under complex backgrounds and variations in pose,

effectively locating facial features. Each detected facial image is then passed into a ResNet-101 model pretrained on the VGGFace2 dataset, which extracts visually rich feature vectors containing emotional information through deep convolutional networks. These feature vectors reflect changes in facial expressions and can be effectively aligned with emotional states.

The visual representation of each frame is formed by the collection of facial expression features from all speakers in the image. These feature vectors comprehensively describe the emotional states of the participants in each frame. By repeating the above processing for all 20 frames in the segment, VisExtNet generates a sequence of visual feature vectors based on frame-level facial expression information for each video segment. Finally, these features are aggregated via an averaging operation along the frame axis to generate a 1000-dimensional visual feature vector. This vector consolidates the emotional information from all frames in the segment, ensuring dynamic capture of emotional changes and providing a solid visual feature foundation for subsequent multimodal fusion and emotion recognition.

3.2 Contextual Information Modeling

In multimodal emotion recognition tasks, modeling contextual information is crucial. This is because dialogue is a dynamic interactive process, and emotional states continuously evolve throughout the conversation. Contextual information helps the model capture the trends of these dynamic changes. For example, at the beginning of a conversation, emotions may manifest as joy, but as the conversation progresses, the emotion may gradually shift to anxiety or anger. By modeling contextual information, the model can more accurately capture emotional changes, thus improving recognition accuracy. Furthermore, emotional expressions in dialogue often depend on the contextual environment. Different behavioral cues, such as tone and facial expressions, may convey distinct emotional meanings. The same emotion may appear neutral in some cases, but when combined with specific context, it can take on a clear emotional tendency. By effectively modeling contextual information, the model can gain a deeper understanding of context-dependent relationships, thereby enabling more accurate emotion recognition. To address these limitations, this paper designs a context modeling module named DialogueCRN. Its core component is an xLSTM-based reasoning module, which effectively integrates global and local contextual features through xLSTM and models semantic relationships across modalities using a multi-head attention mechanism. This design compensates for the shortcomings of previous methods in context modeling and multimodal coupling. The purpose of the reasoning module is to iteratively extract and integrate contextual information from text, audio, and visual modalities throughout the dialogue process. The structure of the proposed reasoning module is illustrated in Figure 3(a).

The core of the reasoning module is xLSTM, and its network structure is shown in Figure 3(b). xLSTM enhances the model's ability to make significant changes to the memory cell state by replacing the traditional sigmoid activation function with an exponential activation function. The exponential gating mechanism allows the memory cells to undergo more substantial adjustments, enabling the model to integrate

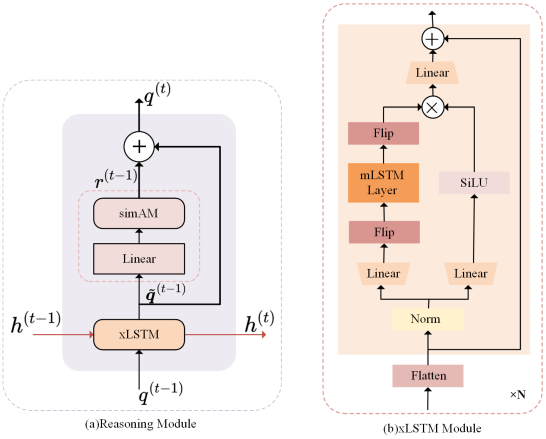

Fig. 3 (a) Inference module structure; (b) xLSTM module structure.

new information more quickly and perform corresponding memory updates. Additionally, xLSTM is equipped with a matrix memory, replacing the traditional scalar storage unit. The matrix memory not only significantly improves the model's ability to store and process high-dimensional information but also better captures complex relationships and dependencies in the input data, allowing the model to more effectively represent context and long-term dependencies. xLSTM also adopts a parallel architecture, which substantially increases training and inference speed. In the first reasoning module, the features of each input modality are first passed through the xLSTM network to learn the contextual information in the dialogue. The output vector of xLSTM is then matched with globally relevant contextual information through a linear layer and the simAM attention mechanism. Finally, the output of xLSTM is concatenated with the results of the attention mechanism to provide input for the next reasoning process.

In the t-th inference module, after passing through the xLSTM network the model can learn the logical order inherent in each modality and integrate the contextual information throughout the conversation. This process can be represented as:

$$\tilde{\mathbf{q}}_i^{(t-1)}, \mathbf{h}_i^{(t)} = \overrightarrow{\mathrm{LSTM}}\left(\mathbf{q}_i^{(t-1)}, \mathbf{h}_i^{(t-1)}\right) \tag{1}$$

where $\tilde{\mathbf{q}}_i^{(t-1)} \in \mathbb{R}^{2d_u}$ is the output vector. $\mathbf{q}^{(t)} \in \mathbb{R}^{4d_u}$ is the initialization of the context-level context $\mathbf{c}_i$ represented by the context of the current discourse, i.e. $\mathbf{q}_i^{(0)} = \mathbf{W}_q\mathbf{c}_i + \mathbf{b}_q$, where $\mathbf{W}_q \in \mathbb{R}^{4d_u \times 2d_u}$ and $\mathbf{b}_q \in \mathbb{R}^{4d_u}$ are learnable parameters. $\mathbf{h}_i^{(t)} \in \mathbb{R}^{2d_u}$ refers to the working memory, which not only stores and updates previous memories $\mathbf{h}_i^{(t-1)}$, but also guides the extraction of the next round of contextual information, and during the sequential flow of the working memory, the implicit logical order between contextual information can be learned. $\mathbf{h}_i^{(t)}$ is initialized to zero, and t is the index indicating how many reasoning modules are to be executed to compute the final state.

In each inference module, we utilize the attention mechanism to match globally relevant contextual information. This process can be represented as:

$$\mathbf{r}_i^{(t-1)} = \sum_{j=1}^{N} \text{sigmoid}\left(1/E\right) \odot \tilde{\mathbf{q}}_i^{(t-1)} \tag{2}$$

The output $\tilde{\mathbf{q}}_i^{(t-1)}$ of the xLSTM is then connected to the output $\mathbf{r}_i^{(t-1)}$ of the attention to form the input $\mathbf{q}_i^{(t)}$ of the next inference module, i.e:

$$\mathbf{q}_i^{(t)} = \left[\tilde{\mathbf{q}}_i^{(t-1)}; \mathbf{r}_i^{(t-1)}\right] \tag{3}$$

The inputs $\mathbf{q}_i^{(t)}$ to the next inference module will be updated under the guidance of working memory $\mathbf{h}_i^{(t)}$ and additional contextual information can be retrieved from the global features.

3.3 Modality Encoder

After modeling contextual information through DialogueCRN, the features from different modalities (such as text, audio, and visual) typically exist in their respective independent feature spaces, which may make direct fusion of modality-specific information challenging. To address this issue, the role of the modality encoder is to map these different modality features into a shared feature space. In this common space, features from different modalities can be effectively aligned and complemented, ensuring that each modality's information is complementary and leverages its unique advantages. Additionally, the modality encoder enables both intra-modal and inter-modal information interactions, enhancing the feature representations of individual modalities. This interaction enriches and adds meaningful context to the features of each modality. The enhanced feature representations thus provide more consistent and complementary information for subsequent multimodal fusion, improving the overall performance of the multimodal emotion recognition model. In this way, the modality encoder not only optimizes the integration of the feature space but also facilitates the synergistic interaction between modalities, further enhancing the model's capability to handle complex emotion recognition tasks.

The structure of the modality encoder is shown in Figure 4. Taking the visual modality encoder as an example, the design of this encoder begins with ensuring that the features from different modalities can be mapped into the same shared space. To achieve this, we input the features from each modality into a 1D convolutional layer, where convolution operations are applied to extract the key information from each modality. Additionally, we introduce position embeddings, which leverage the position and sequential information of the utterances within the dialogue to further enhance the expressive power of the convolutional sequence. This step helps capture the temporal relationships of each utterance within the sequence, thereby improving the model's ability to understand context.

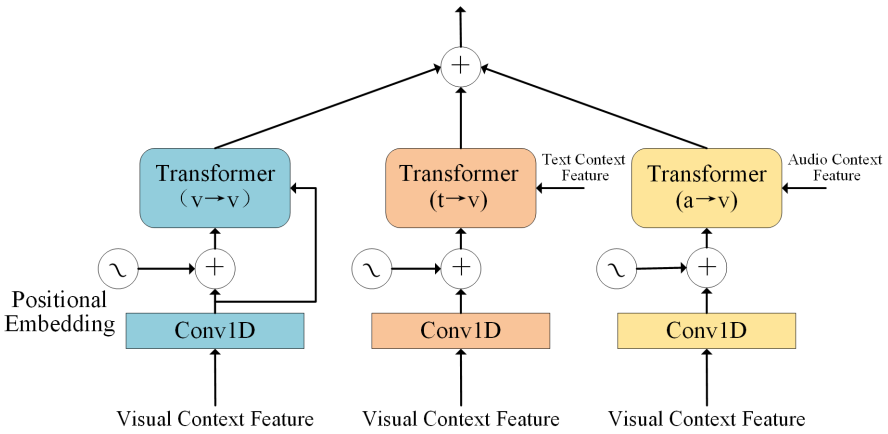

Fig. 4 The structure of the Modality Encoder.

After processing these features, we introduce modality-specific and cross-modality transformers, which are responsible for modeling the intra-modal and inter-modal interactions within the dialogue sequence, respectively. The intra-modal transformer is primarily used to extract the dependencies within each modality, ensuring that the information within each modality is effectively integrated. On the other hand, the inter-modal transformer models the relationships between different modalities, allowing the features from various modalities to be more tightly fused. This structure not only enhances the expressive power of each modality's features but also strengthens their interaction and synergy, providing more robust feature representations for subsequent multimodal fusion and emotion recognition tasks. In this way, the modality encoder is able to better handle complex multimodal information, thus improving the model's accuracy in emotion recognition. The encoder we use [9] consists of three inputs, query $\mathbf{Q} \in \mathbb{R}^{T_q \times d_k}$, key $\mathbf{K} \in \mathbb{R}^{T_k \times d_k}$ and value $\mathbf{V} \in \mathbb{R}^{T_k \times d_v}$. We denote the transformer encoder as Transformer ($\mathbf{Q}$, $\mathbf{K}$, $\mathbf{V}$).

For intra-modal transformer, we will use $\mathbf{H}_m$ as queries, keys and values:

$$\mathbf{H}_{m \to m} = \text{ Transformer } (\mathbf{H}_m, \mathbf{H}_m, \mathbf{H}_m) \in \mathbb{R}^{N \times d} \tag{4}$$

Where $m \in \{t, a, v\}$, the intra-modal converter itself enhances the feature representation of the modality and therefore captures the intra-modal information interactions between discourse sequences.

For the inter-modal transformer, we will use $\mathbf{H}_m$ as a query that will be used $\mathbf{H}_n$ as key and value:

$$\mathbf{H}_{n \to m} = \text{ Transformer } (\mathbf{H}_m, \mathbf{H}_n, \mathbf{H}_n) \in \mathbb{R}^{N \times d} \tag{5}$$

Where $m \in \{t, a, v\}$, and $n \in \{t, a, v\} - \{m\}$. The inter-modal transformer enables the m-modality to obtain information from the n-modality, and therefore captures

inter-modal information interactions between discourses. In summary, n augmented m-modal sequence representations $\mathbf{H}_{n\to m}$ are obtained from the modal encoder module, where $n, m \in \{t, a, v\}$.

3.4 Multi-modal fusion

To more effectively capture the complex relationships and mappings between different modalities, we introduce a novel multimodal fusion network, called MultiAttn [9], which is based on a bidirectional multi-head cross-attention mechanism. Unlike traditional attention mechanisms, in MultiAttn, the query (Query) originates from one modality, while the key (Key) and value (Value) come from other modalities. This allows the query modality to interact with the keys and values from other modalities, uncovering potential relationships and complementary information between the modalities. When calculating the attention distribution, MultiAttn also fully utilizes context information, not only referencing the current modality but also incorporating previous and subsequent contextual information to obtain more comprehensive attention weights. This approach helps the model consider not only the current interaction information but also capture temporal dependencies during modality fusion, thus improving the accuracy and robustness of emotion recognition. The architecture of MultiAttn is shown in Figure 5, illustrating how features from different modalities are fused through the bidirectional multi-head cross-attention mechanism. In this architecture, by integrating the information from different modalities through complex attention calculations, the model can more accurately capture the deep-level relationships between modalities and achieve more efficient multimodal information fusion. This mechanism enables more flexible interaction between modalities and allows for dynamic adjustment of each modality's influence depending on the context, thereby enhancing the overall performance of the emotion recognition task.

The MultiAttn network consists of three main components: MultiAttn-text, MultiAttn-audio, and MultiAttn-visual. The core of each component lies in fusing one modality with complementary information from the other two modalities to enhance the interrelationships between the modalities. Specifically, the architecture design of MultiAttn-text, MultiAttn-audio, and MultiAttn-visual is similar, with the main difference being the sources of the query, key, and value inputs. To better understand the working mechanism of multimodal fusion, we will provide a detailed explanation using MultiAttn-visual as an example.

The MultiAttn-visual employs a three-stage processing method to combine visual modality with audio and text information, thereby capturing the complex relationships across modalities. In the first stage, MultiAttn-visual takes the visual modality as the query input and the text modality as the key and value, performing bidirectional multi-head cross-attention operations to learn the cross-modal correlations and mapping relationships between the visual and text modalities. This process aims to uncover how textual features influence the interpretation of visual information, especially in the interaction related to emotional expression. In the second stage, the output from the first stage is used as a new query. The fused features of the visual and text modalities undergo a cross-attention operation with the audio modality. Specifically, the audio modality is used as the key and value, and the bidirectional multi-head cross-attention

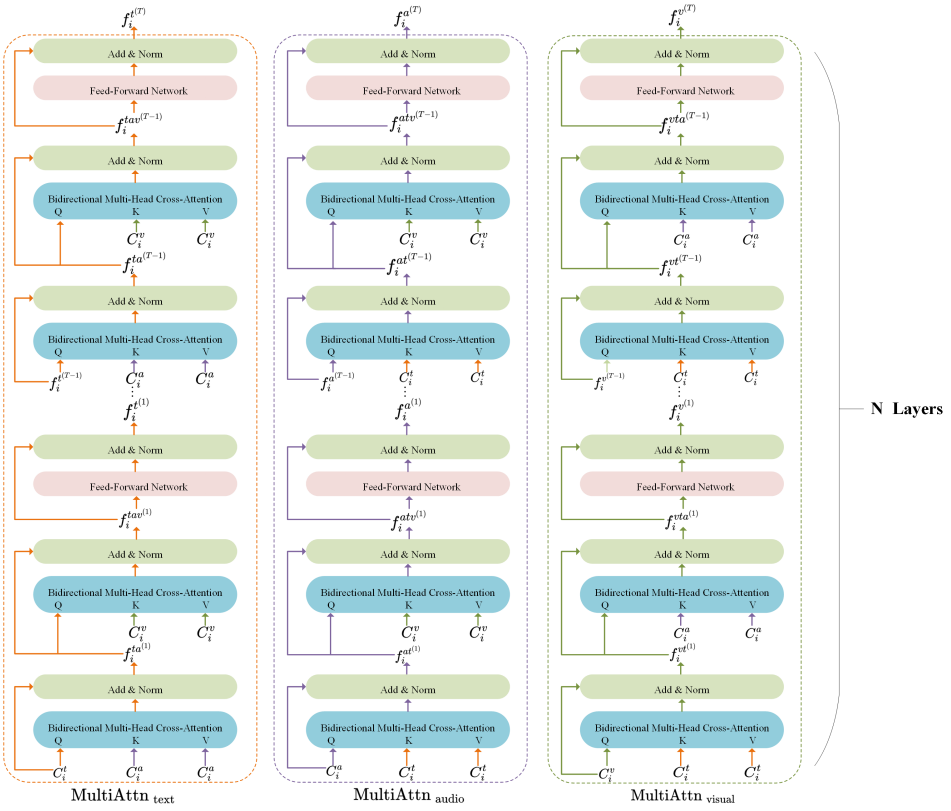

Fig. 5 Multi-modal fusion structure.

mechanism fuses the combined visual-textual information with the audio information, thereby enhancing the complementary interaction among the three modalities in the dialogue. This further improves the emotion recognition performance, especially as the interaction between visual and linguistic features provides more accurate emotional cues. In the third stage, a feed-forward network consisting of two fully connected layers with rectified linear units (ReLU) is used as the key-value memory store. This stage introduces nonlinear transformations to enhance the expressive power of the model, making the fusion of cross-modal information more precise and expressive. To optimize the network's training process, the output of each stage undergoes residual connections and layer normalization, which helps mitigate the vanishing gradient problem, accelerates the model's convergence, and enhances its stability.

To construct a deeper and more powerful multimodal fusion network, MultiAttn stacks multiple layers of MultiAttn-visual, MultiAttn-text, and MultiAttn-audio units. The output of each layer serves as the query input for the subsequent layer, forming a hierarchical processing structure. In this manner, MultiAttn is able to capture cross-modal features at different levels and effectively enhance the interaction between modalities, ultimately improving the accuracy and robustness of multimodal emotion recognition.

3.5 Classification of emotions

After completing multimodal fusion, the resulting multimodal fused features include text features ($\mathbf{f}_i^t$), audio features ($\mathbf{f}_i^a$), and visual features ($\mathbf{f}_i^v$). These features are then concatenated at the feature level to form a richer feature vector. Subsequently, the concatenated features are passed through a fully connected layer for further feature transformation and information fusion. At this stage, all the fused features are fed into a multilayer perceptron (MLP) network with two layers, each equipped with rectified linear unit (ReLU) activation functions, to introduce non-linear transformations and enhance the model's expressive power. After processing by the MLP, the final feature representation is passed through a Softmax layer, which computes the probability distribution for each emotion category. Specifically, the Softmax layer converts the score for each emotion category into a probability value, indicating the likelihood of that category being the correct emotion label. Finally, the system selects the emotion category with the highest probability as the predicted result, thus determining the emotion expressed in the i-th utterance.

The core of this method lies in multimodal fusion, where the complementarity of text, audio, and visual information is leveraged to effectively integrate emotion cues from different modalities, thereby improving the accuracy and robustness of emotion recognition. Through the probability distribution calculation in the Softmax layer, the model can make an optimal selection among multiple emotion categories and predict the most suitable emotion label for each input utterance. The specific calculation method is as follows:

$$\mathbf{f}_i = \mathbf{f}_i^t \oplus \mathbf{f}_i^a \oplus \mathbf{f}_i^v \tag{6}$$

$$\mathbf{z}_i = \mathbf{W}^z \mathbf{f}_i + \mathbf{b}_z \tag{7}$$

$$\mathbf{l}_i = \max\left(0, \mathbf{W}^l \mathbf{z}_i + \mathbf{b}_l\right) \tag{8}$$

$$\mathbf{p}_i = \text{Softmax}\left(\mathbf{W}^{\text{smax}} \mathbf{l}_i + \mathbf{b}_{\text{smax}}\right) \tag{9}$$

$$\hat{y}_i = \underset{t}{\text{argmax}}\left(\mathbf{p}_i[t]\right) \tag{10}$$

Where $\oplus$ denotes concatenation, $\mathbf{W}^z$, $\mathbf{W}^l$ and $\mathbf{W}^{\text{smax}}$ are weight matrices, $\mathbf{b}_z$, $\mathbf{b}_l$ and $\mathbf{b}_{\text{smax}}$ are bias parameters.

To address the challenges of minority class emotions and semantically similar emotion classification in emotion recognition tasks, we use a loss function based on Focal Contrastive Loss (FCL), which we refer to as Sample-Weighted Focal Contrastive (SWFC) loss [43]. The design of this novel loss function aims to solve two key issues: first, the difficulty in recognizing minority class emotion samples, by assigning higher weights to these hard-to-classify minority samples during training, thereby enhancing the model's ability to learn from them; second, to effectively distinguish between semantically similar emotion labels, the SWFC loss maximizes the inter-class distance between different emotion categories, ensuring sufficient separability between samples with different emotion labels.

The core idea of the SWFC loss is to use a focal contrastive mechanism, allowing the model to place more emphasis on those hard-to-classify samples when handling minority class emotion samples. Specifically, the SWFC loss combines the traditional contrastive loss function with a weighting mechanism, assigning higher training weights

to minority class samples and hard-to-classify samples. Additionally, by introducing inter-sample contrast within the loss function, SWFC not only helps improve the model's classification performance but also effectively prevents the model from becoming overly reliant on easily classifiable samples, thereby enhancing its robustness in complex emotion recognition tasks. The mathematical formulation of the SWFC loss is as follows:

$$s_{j,g}^{(i)} = \frac{\exp\left(\mathbf{z}_{i,j}^{\mathrm{T}} \mathbf{z}_{i,g}/\tau\right)}{\sum_{\mathbf{z}_{i,s} \in A_{i,j}} \exp\left(\mathbf{z}_{i,j}^{\mathrm{T}} \mathbf{z}_{i,s}/\tau\right)} \tag{11}$$

$$L_{\mathrm{SWFC}} = -\sum_{i=1}^{M} \sum_{j=1}^{C(i)} \left(\frac{N}{n_{y_{i,j}}}\right)^{\alpha} \frac{1}{|R_{i,j}|} \sum_{\mathbf{z}_{i,g} \in R_{i,j}} \left(1 - s_{j,g}^{(i)}\right)^{\gamma} \log s_{j,g}^{(i)} \tag{12}$$

Where $\mathbf{z}_{i,j}$ is the output of the fully connected layer for utterance j in dialogue i, $A_{i,j}$ is the set of features in the batch other than $\mathbf{z}_{i,j}$, $y_{i,j}$ is the label of utterance j in dialogue i, $R_{i,j} = \{\mathbf{z}_{i,g} \in A_{i,j} \mid y_{i,g} = y_{i,j}\}$ is the set of positive features that share the same label as $\mathbf{z}_{i,j}$, $n_{y_{i,j}}$ is the count of label $y_{i,j}$ in the batch, α is a sample-weight parameter that controls the degree of focus on minority classes, τ is a temperature parameter that controls the strength of penalties on negative samples, γ is a focusing parameter which forces the model to focus on hard-to-classify examples.By maximizing the distance between samples of different emotion categories, the SWFC loss effectively prevents the confusion of semantically similar emotion labels, promoting the model's ability to better distinguish between these emotion categories.

To further enhance the correlation between different modality features in the Multi-Attn model, we introduce a technique called Hirschal-Gebelein-Rényi (Soft-HGR) loss [44]. This loss function is designed to maximize the correlation between text, audio, and visual features, ensuring that the model can effectively capture complementary information between modalities during multimodal fusion.

Specifically, the Soft-HGR loss strengthens the interrelationship between different modalities by measuring and optimizing their similarity. Traditional loss functions often focus on learning individual modalities, while Soft-HGR loss introduces a complex correlation metric to align multimodal features in a shared representation space, thereby enhancing the effectiveness of multimodal fusion. This approach effectively reduces the representation discrepancies between modalities, improves the information interaction between them, and results in more comprehensive fused features, ultimately enhancing the accuracy of emotion recognition. The definition of the Soft-HGR loss is as follows:

$$\begin{aligned} &L_{\text{Soft-HGR}} = -\sum_{\mathbf{Q} \neq \mathbf{V}, \mathbf{Q}, \mathbf{V} \in F} \left(\mathbb{E}\left[\mathbf{Q}^{\mathrm{T}} \mathbf{V}\right] - 1/2 \operatorname{tr}(\operatorname{cov}(\mathbf{Q}) \operatorname{cov}(\mathbf{V}))\right) \\ &\text{s.t. } \mathbb{E}[\mathbf{Q}] = 0, \forall \mathbf{Q} \in F. \end{aligned} \tag{13}$$

Where $F = \{\mathbf{F}^t, \mathbf{F}^a, \mathbf{F}^v\}$, $\mathbf{F}^t = [\mathbf{f}_1^t, \ldots, \mathbf{f}_N^t]^{\mathrm{T}}$, $\mathbf{F}^a = [\mathbf{f}_1^a, \ldots, \mathbf{f}_N^a]^{\mathrm{T}}$, $\mathbf{F}^v = [\mathbf{f}_1^v, \ldots, \mathbf{f}_N^v]^{\mathrm{T}}$. Expectations and covariances are approximated through sample means and sample covariances.

By using this loss function, the model encourages features from different modalities to be closer in semantic space, thereby maximizing their correlation and continuously optimizing the synergy between multimodal features during training. The advantage of adopting the Soft-HGR loss lies in its ability to not only strengthen the intrinsic

relationships between different modality features but also improve the stability of multimodal learning. This, in turn, helps the emotion recognition model better understand and integrate emotional signals from multiple sources of information. This process not only enhances the accuracy of emotion recognition but also improves the model's ability to recognize complex and dynamic emotions.

In addition, we use Cross-entropy loss to measure the difference between the predicted probability and the true label:

$$L_{\mathrm{CE}} = -\sum_{i=1}^{M}\sum_{j=1}^{C(i)} \log \mathbf{p}_{i,j}\left[y_{i,j}\right] \tag{14}$$

Where $\mathbf{p}_{i,j}$ is the probability distribution over the emotion classes for utterance j in dialogue i, $y_{i,j}$ is the ground-truth label of utterance j in dialogue i.

Finally, a linear combination of SWFC loss, Soft-HGR loss and cross-entropy loss is utilized as the loss function of the model:

$$\begin{aligned} L_{\mathrm{Train}} &= 1/N\left(\mu_1 L_{\mathrm{SWFC}} + \mu_2 L_{\mathrm{Soft-HGR}} + \left(1-\mu_1-\mu_2\right) L_{\mathrm{CE}}\right) + \lambda\|\theta\|_2^2, \\ &\mu_1, \mu_2 \in [0,1] \end{aligned} \tag{15}$$

Where μ_1 and μ_2 are tunable hyperparameters, λ is the L_2 regularization weight, θ is the set of all trainable parameters.

4 Experiments

In this section, the dataset, experimental setup and evaluation metrics used in the experiments are first described. Then, our proposed method is compared with the current state-of-the-art multi-modal emotion recognition methods, and detailed ablation experiments are conducted to demonstrate the effectiveness of our proposed method.

4.1 Datasets

IEMOCAP [10]: IEMOCAP contains approximately 12 hours of videos of dyadic conversations, which are segmented into 7433 utterances and 151 dialogues. Each utterance is annotated with one of six emotion labels: happiness, sadness, neutral, anger, excitement and frustration.

MELD [12]: MELD is a multi-party dataset with 13708 utterances and 1433 dialogues from the TV series Friends. Each utterance is annotated with one of seven emotion categories: anger, disgust, fear, joy, neutral, sadness and surprise.

The overall sentiment distribution of the IEMOCAP and MELD datasets is shown in Table 1.

4.2 Experimental setting and evaluation metrics

RoBERTa pretraining settings: We fine-tune the RoBERTa-base model using the AdamW optimizer. The learning rate is set to 2e-5, with a batch size of 16 and a maximum sequence length of 128 tokens. We apply a linear learning rate scheduler with warmup over the first 10% of total training steps. Fine-tuning is conducted for

Table 1 Distribution of Emotions in IEMOCAP and MELD Dataset.

dataset	Emotion Labels									
	Happy	Joy	Sad	Neutral	Anger	Excited	Frustrated	Disgust	Fear	Surprise
IEMOCAP	648	–	1084	1708	1103	1041	1849	–	–	–
MELD	–	2308	1002	6436	1607	–	–	361	358	1636

3 epochs, and early stopping is employed based on the validation loss to prevent overfitting.

The hyperparameter settings are as follows: (1) Dataset-specific settings: Since the MELD dataset is more imbalanced than the IEMOCAP dataset, the batch size for IEMOCAP is set to 64, while the batch size for MELD is set to 100. (2) The total number of training epochs is set to 200, with the Adam optimizer used and the hyperparameters set to $\beta_1 = 0.9$, $\beta_2 = 0.99$, and an initial learning rate of 0.00001. The learning rate decays by a factor of 0.95 every 10 epochs. Additionally, the weight for L_2 regularization, λ, is set to 0.00001. To prevent overfitting, dropout is applied with a rate of 0.1. (3) Hyperparameter settings in MultiEMO: In the MultiAttn module, the number of layers N is set to 6 for the IEMOCAP dataset and 2 for the MELD dataset. For the SWFC loss function, the temperature parameter τ, sample weight parameter α, and focusing parameter γ are set to 0.9, 0.8, and 3, respectively. The merging coefficients μ_1 and μ_2 in the total training loss function L_{Train} are set to 0.4 and 0.3, respectively.

Evaluation metrics: We use accuracy (A_{CC}) and weighted average F1 value (Weighted-F1) for model evaluation.

4.3 Experimental results and analysis

Tables 2 and 3 present the experimental results of the proposed method on the IEMOCAP and MELD datasets, respectively, in comparison with several state-of-the-art approaches. The results show that our method achieves state-of-the-art weighted average F1-scores on both datasets, and demonstrates superior recognition performance across most emotion categories. This is attributed to the model's ability to effectively capture contextual information from dialogues, which enhances its predictive capability. As is widely recognized, contextual information plays a critical role in conversational emotion recognition, especially when emotional states change—context becomes essential for accurately identifying subsequent emotions. On the IEMOCAP dataset, the performance on the Happy category is slightly lower than that of the EmoCaps[45] method, and the performance on the Neutral category is lower than that of the SDT[46] method. However, our method outperforms existing methods across all other categories. This is mainly because the SDT model tends to favor emotion classes with larger sample sizes in the dataset, often overlooking minority classes. In contrast, our approach focuses on improving recognition accuracy across all categories, achieving notable improvements in recognizing both minority emotion classes and semantically similar categories on both datasets. On the MELD dataset, our method only performs slightly worse than the EmoCaps method on the Surprise category,

while surpassing other methods in all remaining emotion categories. These comparisons indicate that the proposed method effectively addresses challenges related to the classification of minority emotion classes and semantically similar emotions, showing significant improvements in these aspects across both datasets.

The proposed method achieves weighted F1-scores that are 1.55 and 5.46 points higher than those of the EmoCaps method on the two similar categories Sad and Frustrated, respectively, and 3.31 and 3.58 points higher than the baseline MultiEMO[29] method, respectively. This indicates that our approach is effective in recognizing semantically similar emotion categories. On the MELD dataset, our method shows noticeable improvements in recognition performance across all emotion categories, with particularly significant enhancements for the minority classes Fear and Disgust. Compared to the baseline MultiEMO method, our approach achieves F1-score improvements of 4.34 and 3.39 points for these two categories, respectively. For the semantically similar emotion categories Anger and Disgust, our method outperforms the SDT method by 2.62 and 9.89 points, and surpasses the baseline MultiEMO method by 3.35 and 3.39 points, respectively. These comparisons further demonstrate that the contextual information extracted from dialogues by our method contributes significantly to the model's performance. In addition, the designed modality encoder facilitates the alignment and complementation of multimodal information and enhances both inter-modal and intra-modal interactions. This further confirms that the proposed method not only aids in distinguishing semantically similar emotion categories, but also substantially improves recognition performance for emotion categories with scarce samples and imbalanced distributions.

Table 2 Experimental results on the IEMOCAP dataset. The best results are highlighted in bold. "-" indicates that the results could not be obtained from the original paper.

Model	IEMOCAP							
	Happy	Sad	Neutral	Anger	Excited	Frustrated		
	F1	F1	F1	F1	F1	F1	Acc	W-F1
SC-LSTM[47]	47.03	79.86	56.41	62.33	71.42	60.68	63.51	63.54
AGHMN[17]	52.16	73.35	58.39	61.94	69.76	62.31	63.52	63.49
DialogueRNN[21]	32.93	78.04	59.18	63.26	73.62	59.46	63.33	62.83
HAUCL[48]	53.57	82.04	68.61	66.44	75.60	68.23	70.30	70.27
AdaIGN[49]	53.04	81.47	71.26	65.87	76.34	67.79	–	70.74
Zhang et al.[50]	57.10	79.90	71.00	71.50	78.40	67.50	72.40	71.60
EmoCaps[45]	**74.31**	85.47	67.03	65.26	80.14	68.38	73.67	73.01
GraphCFC[51]	43.15	85.03	64.72	71.34	78.93	63.74	69.13	68.92
SMFNM[33]	59.52	81.49	70.06	62.91	74.43	70.15	70.82	70.94
MultiEMO[29]	65.24	83.71	67.47	68.47	76.14	70.26	–	71.98
CFN-ESA[52]	53.67	80.60	71.65	70.32	74.82	68.06	70.78	71.04
SDT[46]	66.19	81.84	**74.62**	69.73	80.17	68.68	73.95	74.08
Ours	68.75	**87.02**	71.98	**71.56**	**80.23**	**73.84**	**80.59**	**75.21**

As shown in Tables 2 and 3, most existing methods, including the one proposed in this paper, perform significantly better on the IEMOCAP dataset compared to the MELD dataset. This is because the IEMOCAP dataset has a relatively balanced distribution of samples across different emotion categories and a larger overall data volume, with only the Happiness category having relatively fewer samples. This allows various methods to train more effective models on this dataset. In contrast, the MELD dataset suffers from a highly imbalanced distribution of samples among emotion categories and has a smaller overall data volume. The Neutral category has the highest number of samples, resulting in the best recognition performance for this class. However, Fear and Disgust have the fewest samples, leading to poor recognition performance across all methods for these two categories. Despite these challenges, as shown in Table 3, our method still achieves significantly better recognition performance on the MELD dataset compared to most other approaches. This demonstrates that the proposed method is not only effective in improving recognition of semantically similar emotions, but also capable of addressing the difficulties associated with recognizing minority emotion categories.

Table 3 Experimental results on the MELD dataset. The best results are highlighted in bold. "-" indicates that the results could not be obtained from the original paper.

Model	MELD								
	Neutral	Surprise	Fear	Sad	Joy	Disgust	Angry		
	F1	F1	F1	F1	F1	F1	F1	Acc	W-F1
DialogueRNN[21]	76.56	47.64	0.00	24.65	51.49	0.00	46.01	58.03	56.98
MMGCN[25]	76.96	49.63	3.64	20.39	53.76	2.82	45.23	–	58.41
DialogueTRM[24]	79.41	55.27	17.39	36.48	60.30	20.18	49.79	65.70	63.80
AdaIGN[49]	79.75	60.53	–	43.70	64.54	–	56.15	–	66.79
MVN[19]	76.65	53.18	11.70	21.82	53.62	21.86	42.55	61.30	59.03
HAUCL[48]	–	–	–	–	–	–	–	68.05	66.72
Zhang et al.[50]	78.40	53.50	12.10	35.90	56.90	11.90	44.30	–	62.30
EmoCaps[45]	74.28	**64.74**	2.14	42.52	62.52	7.05	60.26	64.93	63.88
SMFNM[33]	75.06	57.48	16.83	36.79	62.35	25.04	50.33	62.60	62.42
MultiEMO[29]	79.98	60.28	28.24	41.20	62.86	35.28	53.60	–	66.47
CFN-ESA[52]	80.05	58.78	21.62	41.82	66.50	26.92	54.18	67.85	66.70
SDT[46]	80.19	59.07	17.88	43.69	64.29	28.78	54.33	67.55	66.60
Ours	**82.38**	64.29	**32.58**	**43.84**	**66.71**	**38.67**	**56.95**	**69.16**	**69.78**

Although the proposed method achieves excellent recognition performance, there are still some limitations. We analyzed the confusion matrices of the recognition performance on the two datasets. As shown in Figure 6, (1) our model still exhibits a high error rate in classifying similar emotions, such as "Happy" and "Excited", "Anger" and "Frustrated" on the IEMOCAP dataset, and "Surprise" and "Joy" on the MELD dataset. (2) Our model also tends to incorrectly classify other emotion categories in the MELD dataset as "Neutral", due to the dominant proportion of samples in the

"Neutral" category. (3) On the MELD dataset, it is challenging to accurately detect the "Fear" and "Disgust" emotion categories because these categories have very few samples, making it difficult to train effective models. Therefore, recognizing similar emotions and emotions with imbalanced data presents significant challenges.

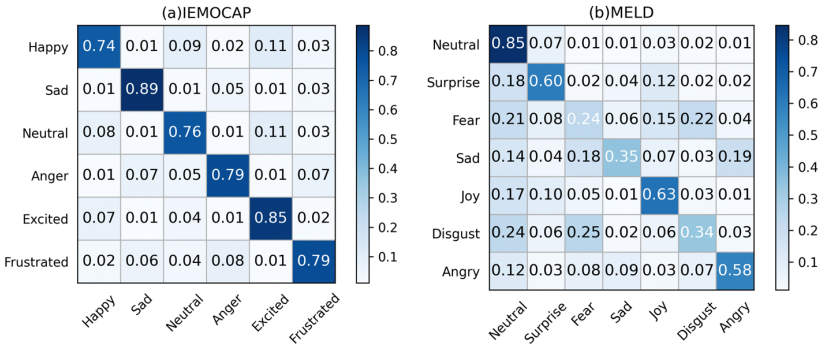

Fig. 6 Confusion matrix on IEMOCAP and MELD datasets.

4.4 Ablation experiments

To investigate the contribution of key components in our proposed model to its overall performance, we conducted ablation studies on both the IEMOCAP and MELD datasets using the same experimental settings. The results are shown in Table 4. As seen from the table, removing the DialogueCRN module—designed for contextual information modeling—has a considerable impact on model performance. On the IEMOCAP dataset, the overall recognition accuracy drops by 4.23%, and the weighted F1-score decreases by 1.48. On the MELD dataset, the accuracy decreases by 4.2%, and the weighted F1-score drops by 1.33. These results indicate that the use of the DialogueCRN module effectively models contextual information, allowing the model to better capture dynamic emotional changes, thereby improving emotion recognition accuracy. Furthermore, excluding the attention mechanism within the DialogueCRN module also negatively affects model performance. This is because, without attention, the model fails to adequately preserve and update the extracted contextual information, leading to insufficient guidance in capturing relevant contextual cues later in the dialogue. In addition, removing the modality encoder also significantly impacts both the accuracy and weighted F1-score on the two datasets. On the IEMOCAP dataset, the overall recognition accuracy drops by 3.01%, and the weighted F1-score decreases by 1.33. On the MELD dataset, accuracy falls by 2.75%, and the weighted F1-score declines by 1.18. This is because the modality encoder maps features from different modalities into a shared feature space, enabling effective alignment and complementation of multimodal features. It also facilitates both intra-modal and inter-modal information interaction, which enhances the representation capability of individual modalities and further improves the overall model performance.

Table 4 Results of ablation experiments on IEMOCAP and MELD datasets.

	IEMOCAP		MELD	
	Acc	W-F1	Acc	W-F1
Ours	**80.59**	**75.21**	**69.16**	**69.78**
w/o DialogueCRN	76.36	73.73	64.96	68.45
w/o simAM attention	78.76	74.49	67.37	68.98
w/o Modality Encoder	77.58	73.88	66.41	68.60
w/o MultiAttn	75.84	70.56	63.82	64.22
w/o SWFC loss	79.25	74.85	66.59	68.15
w/o Soft-HGR loss	79.97	74.92	68.87	69.43

Additionally, to analyze the impact of the multimodal fusion module MultiAttn on model performance, we replaced MultiAttn with simple feature concatenation to fuse the multimodal features. As shown in Table 4, the model's performance drops sharply on both datasets. On the IEMOCAP dataset, overall recognition accuracy decreases by 4.75%, and the weighted F1-score drops by 4.65. On the MELD dataset, the accuracy decreases by 5.34%, and the weighted F1-score declines by 5.56. This demonstrates that using a multimodal fusion network can effectively capture cross-modal correlations and dependencies among textual, audio, and visual modalities, enabling more comprehensive fusion and utilization of multimodal features.We also conducted ablation experiments on the SWFC loss function and the Soft-HGR loss function adopted in this work. The results show that both loss functions have a certain degree of influence on model performance. The SWFC loss function assigns greater weight to hard-to-classify minority emotion classes during training and encourages mutual exclusivity between samples with different emotion labels to maximize inter-class distance. This enhances the recognition of semantically similar emotions, with a more noticeable performance drop observed on the MELD dataset due to its more severe class imbalance compared to IEMOCAP. Furthermore, the Soft-HGR loss function strengthens the correlations among textual, audio, and visual features extracted by MultiAttn, further improving multimodal representation learning.

Table 5 Ablation experiment results for the sample weight parameter α, temperature parameter τ, and focal weight parameter γ in the SWFC loss function.

α	MELD		τ	MELD		γ	MELD	
	Acc	W-F1		Acc	W-F1		Acc	W-F1
0.4	65.19	66.27	0.5	66.54	68.15	1	67.48	69.08
0.5	65.74	67.08	0.6	66.91	68.44	2	67.95	69.12
0.6	66.32	67.63	0.7	67.36	68.67	3	**68.33**	**69.25**
0.7	67.06	68.46	0.8	67.83	68.85	4	68.07	69.16
0.8	**67.53**	**68.84**	0.9	**68.09**	**69.13**	5	67.63	69.05
0.9	66.94	68.69	1.0	67.75	68.97	6	67.29	68.92
0.95	66.18	68.52	1.1	66.98	68.68	7	66.84	68.79

We conducted ablation experiments on the SWFC loss function, focusing on the sample weight parameter α that controls attention to minority classes, the temperature parameter τ that controls the penalty strength for negative samples, and the focal weight parameter γ that enforces model focus on difficult-to-classify samples. The experiments were carried out using the MELD dataset. This is because the MELD dataset not only includes a richer set of emotion categories but also exhibits a more pronounced class imbalance and contains semantically similar emotions. The experimental results are shown in Table 5. As indicated by the data in Table 5, when the sample weight parameter α controlling attention to minority classes is set to 0.8, the model exhibits good recognition performance. Setting the temperature parameter τ controlling the penalty strength for negative samples to 0.9 also results in good performance. When the focal weight parameter γ enforcing the model's focus on difficult-to-classify samples is set to 3, the model achieves optimal performance.

Table 6 Ablation experiment results for linear combination coefficients μ_1.

μ_1	IEMOCAP		MELD	
	Acc	W-F1	Acc	W-F1
0.1	78.38	74.22	68.16	69.18
0.2	78.57	74.39	68.36	69.35
0.3	78.82	74.48	**68.41**	**69.47**
0.4	**79.21**	**74.56**	68.29	69.39
0.5	79.03	74.51	68.05	69.31
0.6	78.75	74.43	67.87	69.23

Table 7 Ablation experiment results for linear combination coefficients μ_2.

μ_2	IEMOCAP		MELD	
	Acc	W-F1	Acc	W-F1
0.1	78.84	74.60	68.22	69.32
0.2	79.08	74.73	68.31	69.39
0.3	**79.26**	**74.82**	68.40	69.47
0.4	79.17	74.81	**68.46**	**69.53**
0.5	79.06	74.75	68.37	69.49
0.6	78.76	74.68	67.28	69.42

We also conducted ablation experiments on the linear combination coefficients μ_1 and μ_2 of SWFC loss, Soft-HGR loss, and cross-entropy loss, as well as on the L_2 regularization weight. This set of experiments was performed on both the MELD and IEMOCAP datasets to validate the effectiveness of the training loss of the model constructed through linear combination. The experimental results are shown in Table 6, Table 7, and Table 8. The experimental data in Table 6 show that setting μ_1 to 0.4 results in the best performance on the IEMOCAP dataset, while setting μ_1 to 0.3

yields the best performance on the MELD dataset. The experimental data in Table 7 show that setting μ_2 to 0.3 yields the best performance on the IEMOCAP dataset, while setting μ_2 to 0.4 results in the best performance on the MELD dataset. The experimental data in Table 8 show that setting the value of λ to 0.00001 results in the best performance of our model on both datasets.

Table 8 Ablation experiment results for the L_2 regularization weight λ.

λ	IEMOCAP		MELD	
	Acc	W-F1	Acc	W-F1
0.1	79.39	74.79	68.23	69.46
0.01	79.48	74.83	68.36	69.52
0.001	79.61	74.88	68.42	69.58
0.0001	79.73	74.94	68.54	69.63
0.00001	**79.85**	**74.98**	**68.62**	**69.67**
0.000001	78.74	74.91	67.57	69.59

Table 9 Ablation experiment results for the number of layers in the multi-modal fusion module MultiAttn.

N	IEMOCAP		MELD	
	Acc	W-F1	Acc	W-F1
1	76.45	74.82	67.65	69.66
2	77.31	74.86	**69.13**	**69.74**
3	78.19	74.93	69.08	69.71
4	78.93	75.01	68.74	69.62
5	79.82	75.10	68.39	69.57
6	**80.47**	**75.18**	67.86	69.49
7	80.16	75.13	67.35	68.41
8	79.77	75.07	66.97	67.33

Finally, we conducted ablation experiments on the number of layers in the multi-modal fusion module MultiAttn. Similarly, the experiments were conducted on the MELD and IEMOCAP datasets to find the optimal number of layers for MultiAttn suitable for both datasets. The results are shown in Table 9. The experimental data in Table 9 show that when the number of layers in MultiAttn is set to 2, our model achieves the best recognition performance on the MELD dataset. Conversely, when the number of layers is set to 6, our model performs best on the IEMOCAP dataset.

5 Conclusion

This paper proposes an attention-based context modeling and multimodal fusion network architecture for multimodal emotion recognition tasks. The core innovation of this architecture lies in the integration of the attention mechanism with the

xLSTM-based context extraction module. This design effectively models the contextual information in the dialogue process, ensuring that the model can capture critical emotional cues within the conversation. Furthermore, we introduce a modality encoder based on the Transformer encoder, which maps features from different modalities (e.g., text, audio, and visual) into a unified feature space. This enables precise alignment and complementarity of features across modalities. This design not only promotes the rich expression of information within each modality but also enhances the information interaction between modalities, further improving the feature representation quality of each modality. In the multimodal fusion phase, we employ a bidirectional multi-head cross-attention layer, which effectively models cross-modal interactions and captures the mapping relationships between different modalities, significantly improving the representation capability of the fused features. Additionally, to address the challenges of recognizing minority class emotional samples and semantically similar emotion categories, we introduce the SWFC loss function. This loss function emphasizes the difficult samples by assigning higher weights to them, enabling the model to focus more on the hard-to-classify emotion categories during training. Especially in imbalanced datasets, this approach significantly enhances the model's ability to recognize minority class emotions. We validate the effectiveness of the proposed method through extensive experiments on two widely used multimodal emotion analysis datasets, IEMOCAP and MELD. The results demonstrate that the proposed method excels in improving both emotion recognition accuracy and model generalization ability.

Future research can further explore the emotional information conveyed by punctuation marks in the text modality and attempt to enhance the use of acoustic features in the audio modality to improve the overall recognition performance of the model. Punctuation marks, as emotional regulators in text, carry rich emotional information, and incorporating these cues can provide a more comprehensive understanding of text emotions. Speech features in the audio modality, such as pitch, speech rate, and emotional tone, often provide crucial clues for emotion analysis, further enhancing the understanding of emotional expression. Additionally, in future work, we plan to extend our model to multilingual conversational datasets and explore emotion transfer across languages. Additionally, we aim to integrate visual emotion cues from static images, such as emojis or comics, to enrich contextual understanding. Another promising direction is applying our model in emotion-aware recommendation systems, especially in social platforms or healthcare applications.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No.62166043, U2003207) and Tianshan Talent Training Project-Xinjiang Science and Technology Innovation Team Program (2023TSYCTD0012).

References

[1] Kumar, A., Dogra, P., Dabas, V.: Emotion analysis of twitter using opinion mining. In: 2015 Eighth International Conference on Contemporary Computing (IC3), pp. 285–290 (2015). IEEE

[2] Pujol, F.A., Mora, H., Martínez, A.: Emotion recognition to improve e-healthcare systems in smart cities. In: Research & Innovation Forum 2019: Technology, Innovation, Education, and Their Social Impact 1, pp. 245–254 (2019). Springer

[3] Zhou, L., Gao, J., Li, D., Shum, H.-Y.: The design and implementation of xiaoice, an empathetic social chatbot. Computational Linguistics **46**(1), 53–93 (2020)

[4] Xiao, Z., Chen, Y., Zhou, X., He, M., Liu, L., Yu, F., Jiang, M.: Human action recognition in immersive virtual reality based on multi-scale spatio-temporal attention network. Computer Animation and Virtual Worlds **35**(5), 2293 (2024)

[5] Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE signal processing letters **23**(10), 1499–1503 (2016)

[6] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)

[7] Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: Vggface2: A dataset for recognising faces across pose and age. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), pp. 67–74 (2018). IEEE

[8] Beck, M., Pöppel, K., Spanring, M., Auer, A., Prudnikova, O., Kopp, M., Klambauer, G., Brandstetter, J., Hochreiter, S.: xlstm: Extended long short-term memory. arXiv preprint arXiv:2405.04517 (2024)

[9] Vaswani, A.: Attention is all you need. arXiv preprint arXiv:1706.03762 (2017)

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

[11] Schuller, B., Valster, M., Eyben, F., Cowie, R., Pantic, M.: Avec 2012: the continuous audio/visual emotion challenge. In: Proceedings of the 14th ACM International Conference on Multimodal Interaction, pp. 449–456 (2012)

[12] Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., Mihalcea, R.: Meld: A multimodal multi-party dataset for emotion recognition in conversations. arXiv preprint arXiv:1810.02508 (2018)

[13] Yang, K., Xu, H., Gao, K.: Cm-bert: Cross-modal bert for text-audio sentiment analysis. In: Proceedings of the 28th ACM International Conference on Multimedia, pp. 521–528 (2020)

[14] Xu, B., Huang, S., Sha, C., Wang, H.: Maf: a general matching and alignment

framework for multimodal named entity recognition. In: Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining, pp. 1215–1223 (2022)

[15] Jiao, W., Yang, H., King, I., Lyu, M.R.: Higru: Hierarchical gated recurrent units for utterance-level emotion recognition. arXiv preprint arXiv:1904.04446 (2019)

[16] Li, H., Guo, A., Li, Y.: Ccma: Capsnet for audio–video sentiment analysis using cross-modal attention. The Visual Computer, 1–12 (2024)

[17] Jiao, W., Lyu, M., King, I.: Real-time emotion recognition via attention gated hierarchical memory network. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 8002–8009 (2020)

[18] Bellenger, D., Chen, M., Xu, Z.: Facial emotion recognition with a reduced feature set for video game and metaverse avatars. Computer Animation and Virtual Worlds **35**(2), 2230 (2024)

[19] Ma, H., Wang, J., Lin, H., Pan, X., Zhang, Y., Yang, Z.: A multi-view network for real-time emotion recognition in conversations. Knowledge-Based Systems **236**, 107751 (2022)

[20] Chen, Q., Xie, S., Fang, X., Sun, Q.: Cthfnet: contrastive translation and hierarchical fusion network for text–video–audio sentiment analysis. The Visual Computer, 1–14 (2024)

[21] Majumder, N., Poria, S., Hazarika, D., Mihalcea, R., Gelbukh, A., Cambria, E.: Dialoguernn: An attentive rnn for emotion detection in conversations. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 6818–6825 (2019)

[22] Ghosal, D., Majumder, N., Poria, S., Chhaya, N., Gelbukh, A.: Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. arXiv preprint arXiv:1908.11540 (2019)

[23] Li, J., Ji, D., Li, F., Zhang, M., Liu, Y.: Hitrans: A transformer-based context-and speaker-sensitive model for emotion detection in conversations. In: Proceedings of the 28th International Conference on Computational Linguistics, pp. 4190–4200 (2020)

[24] Mao, Y., Sun, Q., Liu, G., Wang, X., Gao, W., Li, X., Shen, J.: Dialoguetrm: Exploring the intra-and inter-modal emotional behaviors in the conversation. arXiv preprint arXiv:2010.07637 (2020)

[25] Hu, J., Liu, Y., Zhao, J., Jin, Q.: Mmgcn: Multimodal fusion via deep graph convolution network for emotion recognition in conversation. arXiv preprint arXiv:2107.06779 (2021)

[26] Hu, D., Hou, X., Wei, L., Jiang, L., Mo, Y.: Mm-dfn: Multimodal dynamic fusion network for emotion recognition in conversations. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7037–7041 (2022). IEEE

[27] Zou, S., Huang, X., Shen, X., Liu, H.: Improving multimodal fusion with main modal transformer for emotion recognition in conversation. Knowledge-Based Systems **258**, 109978 (2022)

[28] Hu, G., Lin, T.-E., Zhao, Y., Lu, G., Wu, Y., Li, Y.: Unimse: Towards unified multimodal sentiment analysis and emotion recognition. arXiv preprint arXiv:2211.11256 (2022)

[29] Shi, T., Huang, S.-L.: Multiemo: An attention-based correlation-aware multimodal fusion framework for emotion recognition in conversations. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 14752–14766 (2023)

[30] Zhang, H., Wang, Y., Yin, G., Liu, K., Liu, Y., Yu, T.: Learning language-guided adaptive hyper-modality representation for multimodal sentiment analysis. arXiv preprint arXiv:2310.05804 (2023)

[31] Zheng, W., Yu, J., Xia, R., Wang, S.: A facial expression-aware multimodal multi-task learning framework for emotion recognition in multi-party conversations. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 15445–15459 (2023)

[32] Li, J., Wang, X., Lv, G., Zeng, Z.: Graphmft: A graph network based multimodal fusion technique for emotion recognition in conversation. Neurocomputing **550**, 126427 (2023)

[33] Yang, J., Dong, X., Du, X.: Smfnm: Semi-supervised multimodal fusion network with main-modal for real-time emotion recognition in conversations. Journal of King Saud University-Computer and Information Sciences **35**(9), 101791 (2023)

[34] El Alaoui, D., Riffi, J., Sabri, A., Aghoutane, B., Yahyaouy, A., Tairi, H.: A novel session-based recommendation system using capsule graph neural network. Neural Networks, 107176 (2025)

[35] El Alaoui, D., Riffi, J., Sabri, A., Aghoutane, B., Yahyaouy, A., Tairi, H.: Contextual recommendations: dynamic graph attention networks with edge adaptation. IEEE Access (2024)

[36] El Alaoui, D., Riffi, J., Sabri, A., Aghoutane, B., Yahyaouy, A., Tairi, H.: Deep graphsage-based recommendation system: jumping knowledge connections with ordinal aggregation network. Neural Computing and Applications **34**(14), 11679–11690 (2022)

[37] El Alaoui, D., Riffi, J., Sabri, A., Aghoutane, B., Yahyaouy, A., Tairi, H.: Social recommendation system based on heterogeneous graph attention networks. International Journal of Data Science and Analytics, 1–17 (2024)

[38] El Alaoui, D., Riffi, J., Aghoutane, B., Sabri, A., Yahyaouy, A., Tairi, H.: Collaborative filtering: comparative study between matrix factorization and neural network method. In: Networked Systems: 8th International Conference, NETYS 2020, Marrakech, Morocco, June 3–5, 2020, Proceedings 8, pp. 361–367 (2021). Springer

[39] El Alaoui, D., Riffi, J., Sabri, A., Aghoutane, B., Yahyaouy, A., Tairi, H.: Comparative study of filtering methods for scientific research article recommendations. Big Data and Cognitive Computing **8**(12), 190 (2024)

[40] El Alaoui, D., Riffi, J., Aghoutane, B., Sabri, A., Yahyaouy, A., Tairi, H.: Overview of the main recommendation approaches for the scientific articles. In: International Conference on Business Intelligence, pp. 107–118 (2021). Springer

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

[42] Lee, J., Lee, W.: Compm: Context modeling with speaker's pre-trained memory tracking for emotion recognition in conversation. CoRR **abs/2108.11626** (2021) 2108.11626

[43] Zhang, Y., Hooi, B., Hu, D., Liang, J., Feng, J.: Unleashing the power of contrastive self-supervised visual models via contrast-regularized fine-tuning. Advances in Neural Information Processing Systems **34**, 29848–29860 (2021)

[44] Wang, L., Wu, J., Huang, S.-L., Zheng, L., Xu, X., Zhang, L., Huang, J.: An efficient approach to informative feature extraction from multimodal data. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 5281–5288 (2019)

[45] Shou, Y., Liu, H., Cao, X., Meng, D., Dong, B.: A low-rank matching attention based cross-modal feature fusion method for conversational emotion recognition. IEEE Transactions on Affective Computing (2024)

[46] Ma, H., Wang, J., Lin, H., Zhang, B., Zhang, Y., Xu, B.: A transformer-based model with self-distillation for multimodal emotion recognition in conversations. IEEE Transactions on Multimedia (2023)

[47] Poria, S., Cambria, E., Hazarika, D., Majumder, N., Zadeh, A., Morency, L.-P.: 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), pp. 873–883 (2017)

[48] Yi, Z., Zhao, Z., Shen, Z., Zhang, T.: Multimodal fusion via hypergraph autoencoder and contrastive learning for emotion recognition in conversation. In: Proceedings of the 32nd ACM International Conference on Multimedia, pp. 4341–4348 (2024)

[49] Tu, G., Xie, T., Liang, B., Wang, H., Xu, R.: Adaptive graph learning for multimodal conversational emotion detection. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, pp. 19089–19097 (2024)

[50] Zhang, X., Cui, W., Hu, B., Li, Y.: A multi-level alignment and cross-modal unified semantic graph refinement network for conversational emotion recognition. IEEE Transactions on Affective Computing (2024)

[51] Li, J., Wang, X., Lv, G., Zeng, Z.: Graphcfc: A directed graph based cross-modal feature complementation approach for multimodal conversational emotion recognition. IEEE Transactions on Multimedia **26**, 77–89 (2023)

[52] Li, J., Wang, X., Liu, Y., Zeng, Z.: Cfn-esa: A cross-modal fusion network with emotion-shift awareness for dialogue emotion recognition. IEEE Transactions on Affective Computing (2024)