GraphCFC: A Directed Graph based Cross-modal Feature Complementation Approach for Multimodal Conversational Emotion Recognition

Jiang Li, Xiaoping Wang, *Senior Member, IEEE*, Guoqing Lv, and Zhigang Zeng, *Fellow, IEEE*

Abstract—Emotion Recognition in Conversation (ERC) plays a significant part in Human-Computer Interaction (HCI) systems since it can provide empathetic services. Multimodal ERC can mitigate the drawbacks of uni-modal approaches. Recently, Graph Neural Networks (GNNs) have been widely used in a variety of fields due to their superior performance in relation modeling. In multimodal ERC, GNNs are capable of extracting both long-distance contextual information and inter-modal interactive information. Unfortunately, since existing methods such as MMGCN directly fuse multiple modalities, redundant information may be generated and diverse information may be lost. In this work, we present a directed Graph based Cross-modal Feature Complementation (GraphCFC) module that can efficiently model contextual and interactive information. GraphCFC alleviates the problem of heterogeneity gap in multimodal fusion by utilizing multiple subspace extractors and Pair-wise Cross-modal Complementary (PairCC) strategy. We extract various types of edges from the constructed graph for encoding, thus enabling GNNs to extract crucial contextual and interactive information more accurately when performing message passing. Furthermore, we design a GNN structure called GAT-MLP, which can provide a new unified network framework for multimodal learning. The experimental results on two benchmark datasets show that our GraphCFC outperforms the state-of-the-art (SOTA) approaches.

Index Terms—Emotion Recognition in Conversation, Multimodal Fusion, Graph Neural Networks, Cross-modal Feature Complementation.

I. INTRODUCTION

E MOTIONS pervade our personal and professional lives, shape our relationships and social interactions [1]. Consequently, research on emotion recognition and understanding is crucial. Emotion recognition in conversation (ERC), which aims to automatically determine the emotional state of a speaker during a conversation based on information about human behavior such as text content, facial expressions and audio signals, has received extensive attention and study in recent years [2]–[5]. Emotion recognition can be applied to many practical scenarios such as medical diagnosis [6], opinion mining [7], fake news detection [8] and dialogue generation [9], to provide high-quality and humanized empathetic

services. ERC will play an increasingly vital role as Human-Computer Interaction (HCI) technology advances.

In a multi-person dialogue scenario, each speaker generates a succession of ups and downs in emotional reactions. The majority of prior techniques have been centered on the study of contextual ERC systems. DialogueGCN [2] utilizes a relational Graph ATtention network (GAT) to capture long-distance contextual dependencies in conversations, and leverages self- and inter-dependency of the speakers to improve context understanding for ERC. Shen et al. [10] model the conversation as a directed acyclic graph in an attempt to combine the respective strengths of Graph Neural Networks (GNNs) and Recurrent Neural Networks (RNNs). DialogueCRN [4] designs multi-turn reasoning modules to extract and integrate the emotional clues in order to fully understand the conversational context from a cognitive perspective. Nonetheless, these approaches only take into account information from a single modality. The ERC system will fail if the uni-modality signals do not carry a significant emotional signature. Furthermore, the notorious emotional-shift issue plagues uni-modal emotion recognition systems [3], [10].

In real scenarios, people can instinctively obtain complex emotional cues from multiple information sources including the facial expressions, personality and tone of speaker, as well as the conversation history to infer the real emotions of others. Multimodal ERC follows this idea and attempts to combine simultaneously information from multiple modalities such as textual, acoustic and visual modalities to comprehensively identify emotions in conversations. Fig. 1 shows an instance of a multimodal conversation system. The ERC system takes each modality as input and then performs emotional prediction. bc-LSTM [11] employs textual, visual and acoustic modalities for multimodal emotion recognition, feeding each modality separately into a bidirectional Long Short-Term Memory (LSTM) network to acquire contextual information. CMN [12] uses Gated Recurrent Unit (GRU) and multimodal features for contextual modeling, as well as applies an attention mechanism to pick the most valuable historical utterances. ICON [13] models the contextual knowledge of self- and inter-speaker impacts via a GRU-based multi-hop memory network, while capturing essential emotional cues applying an attention module. DialogueRNN [3] detects current sentiment by tracking the contextual information of the utterance and considering the characteristics of the speaker. These approaches, nevertheless, directly concatenate multimodal information without incorporating the interaction between modalities. In addition,

recurrence-based approaches tend to use recent utterances for modeling, which makes these models difficult to gather long-distant information.

Fig. 1: An instance of a multimodal dialogue system. The utterances contain three modalities, i.e., textual, acoustic and visual modalities.

Recently, GNNs have attracted wide attention in a variety of fields because they can model relationships. MMGCN [5] achieves outstanding performance in multimodal conversational emotion recognition by employing GNNs to capture long-distance contextual information and inter-modal interactive information. However, MMGCN connects the current node directly to all other nodes in the dialogue, perhaps resulting in redundant information. Besides that MMGCN simply divides all edges into two types (i.e., inter-modal edges and intra-modal edges) and leverages the angular similarity to represent edge weight, which can cause the inability of the GNN to accurately select important information when aggregating neighbor information.

There is a heterogeneity gap [14], [15] between distinct modalities in multimodal fusion, which makes it challenging to effectively fuse multimodal features. MMGCN directly puts the utterance of three modalities into the graph as the same type of nodes, and then performs multimodal feature fusion by GNN. This approach not only adds redundant information due to inconsistent data distribution among modalities, but also may risk losing diverse information in the conversational graph. Therefore, we propose a novel graph-based multimodal feature fusion method to alleviate the aforementioned limitations. In the Graph based Cross-modal Feature Complementation (GraphCFC) module, unlike MMGCN treating all utterances as neighbor nodes, we model the conversation as a multimodal directed heterogeneous graph with variable contextual information and extract more than 2 types of edges from the graph based on the perspective of modality type and speaker identity; then, we utilize multiple subspace extrac-

tors to simultaneously preserve the consistency and diversity of multimodal features; finally, we employ the Pair-wise Cross-modal Complementation (PairCC) strategy to gradually achieve feature complementation and fusion. In addition, we propose a novel GNN layer, GAT-MLP, to provide a unified network model for multimodal feature fusion, which can also effectively minimize the over-smoothing problem [16] of GNNs. Our main contributions in this paper are as follows:

① We propose a directed Graph based Cross-modal Feature Complementation (GraphCFC) module. GraphCFC can not only effectively alleviate the heterogeneity gap issue of multimodal fusion, but also sufficiently extract the diverse information from multimodal dialogue graphs.

② A new GNN layer named GAT-MLP is designed, which not only alleviates the over-smoothing problem of GNNs, but also provides a new network framework for multimodal learning.

③ The conversations are represented as a multimodal directed graph with variable contextual utterances and extract distinct types of edges from this graph for encoding, so that GAT-MLP can accurately select the critical contextual and interactive information.

④ Extensive comparative experiments and ablation studies are conducted on two benchmark datasets. The experimental results reveal that our proposed GraphCFC is capable of productive complementation and fusion of multimodal features, attaining optimal performance in comparison to previous SOTA approaches.

The remainder of this paper is organized as follows. The related works of this paper is briefly mentioned in Section II. Section III depicts the proposed graph-based multimodal ERC method. Section IV presents the experimental setup of this work, and the experimental results are analyzed in detail in Section V. Section VI summarizes and prospects to this work.

II. RELATED WORK

A. Emotion Recognition in Conversation

Emotion Recognition in Conversation (ERC), which aims to predict the emotion label of each utterance, plays a crucial role in affective dialogue due to facilitating the understanding of the user's emotions and responding with empathy. This task has been recently attached much importance by numerous NLP researchers for its potential applications in extensive areas such as opinion mining in social media [7], empathy building in dialogue systems [17] and detection of fake news [8]. The emotion of a query utterance is easily influenced by many factors, such as the identity of speaker and the context of conversation. Indeed, the way to model the conversational context is the core of this task [18].

Massive methods have been taken to model the conversation context on the textual modality, which can be divided into two categories: graph-based methods and recurrence-based methods. Besides, models based on multimodal inputs have been proposed, which improve the performance of ERC tasks by leveraging multimodal dependencies and complementarities.

Graph-based Methods. DialogGCN [2] constructs a dialogue graph where each utterance is related with the surrounding utterances. Ishiwatari et al. [19] improves DialogGCN

by taking positional encoding into account. ConGCN [20] constructs a large heterogeneous graph by treating the speakers and utterances as nodes. KET [21] leverages a context-aware affective graph attention mechanism to dynamically capture external commonsense knowledge. DAG-ERC [10] combines the advantages of both graph neural networks and recurrent neural networks, and performs excellently without the aid of external knowledge.

Recurrence-based Methods. ICON [13] and CMN [12] both utilize Gated Recurrent Unit (GRU) and memory networks. HiGRU [22] is made up of two GRUs, one is an utterance encoder and the other is a conversation encoder. DialogRNN [3] is a sequence-based method, where several RNNs model the dialogue dynamically. COSMIC [23] constructs a network that is closely similar to DialogRNN and performs better by adding external commonsense knowledge. DialogueCRN [4] utilizes bidirectional LSTM to build ERC model from a cognitive perspective.

Multimodal-based Methods. CMN [12] leverages multimodal information by concatenating the features from three modalities but fails to consider the interaction between modalities. bc-LSTM [11] adopts an utterance-level LSTM to capture multimodal information. MFN [24] conducts multi-views information fusion and aligns the features of different modalities, but it is unable to model speaker information. MMGCN [5] utilizes an undirected graph to explore a more effective way of multimodal fusion, which outperforms significantly other approaches under the multimodal dialogue setting. There have been a range of works [15], [25]–[27] associated with multimodal learning in sentiment analysis. These efforts, however, do not highlight the social interaction and contextual cues between the speaker and listener in a conversation, thus they do not fall under the purview of ERC. In addition, most sentiment analysis tasks only need to distinguish positive, negative, and neutral opinions. Thus it is difficult to divide emotion into numerous categories like *Happy*, *Excited*, *Sad*, *Angry*, *Frustrated* as in the case of ERC tasks.

B. Multimodal Fusion

Multimodal fusion is one of the most important parts in machine learning, which can integrate information from multiple modalities to predict a result [28]. It can be intuitively assumed that multimodal fusion can provide three benefits: providing more robust prediction results, capturing complementary information, and working without certain modalities [29]. In recent years, multimodal fusion in sentiment analysis [25]–[27] has been researched widely and lots of methods such as multiple kernel learning and various neural networks have been explored to cope with it. However, multimodal sentiment analysis rarely involves multi-person conversational information and focuses mainly on utterance-level prediction. Multimodal ERC is the study of conversations with two or more participants and is a conversation-level emotion prediction. For instances, an individual's emotion is not only derived from self-expression, but is also influenced by the expressions of others. Furthermore, Guo et al. [14] have noted that multimodal fusion faces several challenges, one of which

is the heterogeneity gap [15] between modalities. For this reason, our proposed GraphCFC concentrates on alleviating the heterogeneity gap dilemma of conversational emotion recognition.

C. Graph Neural Networks

In recent years, an increasing number of non-Euclidean data have been represented as graphs. The complexity of graph data has posed significant challenges to existing neural network models. Graph Neural Networks (GNNs) have attracted much attention for their ability to effectively deal with non-Euclidean data. GNNs have been applied in a wide range of applications, including recommendation system, computer vision, natural language processing, biomedicine and traffic forecasting. Graph convolutional networks, a type of GNNs, can be divided into two main streams: the spectral-based and the spatial-based approaches. Spectral-based approaches implement graph convolution by defining filters in a manner similar to graph signal processing. Spatial-based approaches define graph convolution by information propagation, and they have recently gained rapid momentum due to their attractive efficiency, flexibility, and generality. Graph-SAGE [30], GAT [31], and FastGCN [32] are widely-used GNN techniques.

III. PROPOSED METHODS

Given the challenges of multimodal emotion recognition mentioned above, we introduce a novel graph-based multimodal feature fusion approach for ERC in this section. The section consists of four parts, including general overview, uni-modal encoder, Graph based Cross-modal Feature Complementation (GraphCFC) module, and multimodal emotion classifier.

A. General Overview

1) Problem Definition: In an ERC scenario, a dialogue is defined as a sequence of n utterances $[(u_1, s_{u_1}), (u_2, s_{u_2}), ..., (u_n, s_{u_n})]$. Where n denotes the number of utterances, u_i is the i-th utterance in this dialogue sequence, and s_{u_i} indicates the speaker who utters utterance u_i. Each utterance u_i consists of m_i tokens, i.e., u_i can be expressed as $[t_{i1}, t_{i2}, ..., t_{im_i}]$. The number of speaker s in a dialogue system should be greater than or equal to 2; if $s_{u_i} = s_{u_j}(i \neq j)$, then utterance u_i and u_j are uttered by the same participant in the dialogue. Each utterance u involves three modalities, including textual, acoustic and visual modalities, so utterance u_i can also be denoted as $[u_i^t, u_i^a, u_i^v]$, where t, a, v denote textual, acoustic and visual modalities, respectively. Given the defined emotion labels $Y = [y_1, y_2, ..., y_l]$ (y_i is generally represented by one-hot encoding), the objective of the multimodal ERC task is to predict the emotion state label y_i for each utterance u_i based on the available inter-modal interaction and intra-modal context. The quantity of emotion labels in various datasets varies, e.g., 6 for IEMOCAP and 7 for MELD. We also experimented with coarsened emotion labels, which consisted of *Positive*, *Negative* and *Neutral*. For instance, in the IEMOCAP dataset, *Happy* and *Excited* are categorized as *Positive*; *Sad*, *Angry* and *Frustrated* are categorized as *Negative*; and *Neutral* remained unchanged.
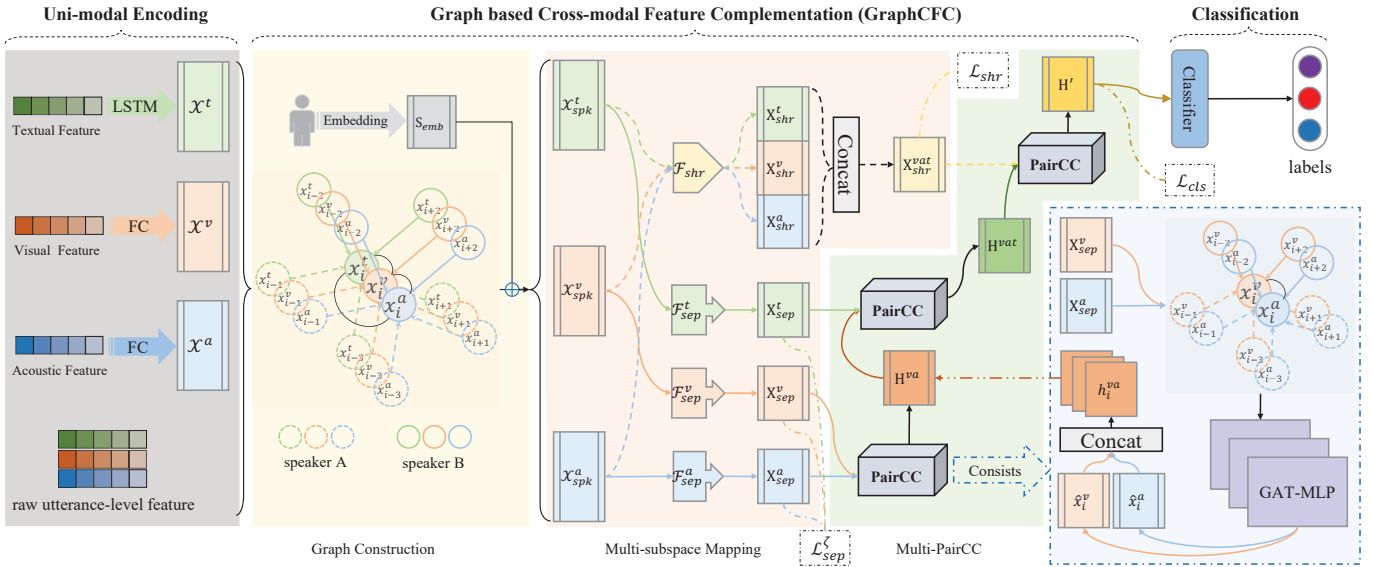
Fig. 2: The illustration of graph-based multimodal ERC, which includes uni-modal encoding, graph based cross-modal feature complementation and multimodal emotion classification.

2) Overall Architecture: Fig. 2 shows the overall architecture of graph-based multimodal ERC in this paper, which mainly consists of uni-modal encoding, Graph based Cross-modal Feature Complementation (GraphCFC) and multimodal emotion classification. Firstly, we encode the uni-modal features by means of three uni-modal encoders. Next, a cross-modal feature complementation module based on GNN is employed for collecting long-distance intra-modal contextual information and inter-modal interactive information. Finally, we utilize multiple loss functions to build multitask learning model for multimodal emotion classification.

B. Uni-modal Encoder

To capture the context-aware feature information of textual modality, referring to MMGCN [5], we leverage a bidirectional Long Short Term Memory (BiLSTM) network. The feature pre-extraction of textual modality can be formulated as:

$$x_i^t, x_{h,i}^t = \overleftrightarrow{\text{LSTM}}(u_i^t; \Theta_{ls}^t), \tag{1}$$

where x_i^t and $x_{h,i}^t$ are the output and hidden vector of pre-extractor, respectively; $\overleftrightarrow{\text{LSTM}}$ and Θ_{ls}^t denote the BiLSTM network and trainable parameter, respectively.

For acoustic and visual modalities, again as with MMGCN, we use a fully connected network for uni-modal feature pre-extraction as follows:

$$x_i^\tau = \text{FC}(u_i^\tau; \Theta_{fc}^\tau), \tau \in \{a, v\}, \tag{2}$$

where x_i^τ is the output vector of pre-extractor; FC and Θ_{fc}^τ are the fully connected network and trainable parameter, respectively; a and v denote acoustic and visual modalities, respectively.

C. Graph based Cross-modal Feature Complementation Module

We propose a Graph based Cross-modal Feature Complementation (GraphCFC) module for efficient multimodal feature fusion. The module includes two types of information complementation, i.e., intra-modal contextual information and inter-modal interactive information. The four primary reasons for which GraphCFC is proposed are as follows. First, to simultaneously preserve the consistency and diversity information of multimodal features. Second, to select crucial intra-modal contextual information and inter-modal interaction information as accurately as possible. Third, to alleviate the heterogeneity gap problem of multimodal ERC. Last, to propose a network model that can be applied to visual, acoustic, and textual modalities simultaneously.

The GraphCFC Module is divided into five main parts. First, we describe how to construct the graph; second, we introduce multiple subspace mappings which are leveraged to simultaneously ensure the consistency and diversity of multimodal features; third, we present a new graph neural network structure named GAT-MLP; fourth, we introduce GAT-MLP based Pair-wise Cross-modal Complementation (PairCC) for alleviating the heterogeneity gap issue of multimodal ERC; finally, we detail the GAT structure of GAT-MLP used in this work.

1) Graph Construction: In uni-modal ERC, a dialogue with n utterances is represented as a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$; where $\mathcal{V}$ is the node set, which denotes the set of utterances, i.e., $\mathcal{V} = \{u_1, u_2, ..., u_n\}$; $\mathcal{E}$ is the set of relational dependencies between nodes; and if an edge exists for two nodes, then $e_{ij} \in \mathcal{E}$ has two key properties: edge weight and edge type. Assuming the existence of two modalities P, Q, we construct the dialogue graph as follows.

Nodes. In a graph, each utterance u_i ($i = 1, 2, ..., n$) is considered as node u_i^P and node u_i^Q, represented as vector x_i^P and vector x_i^Q. If there are n utterances, then $\mathcal{V}$ can be denoted as $\mathcal{V} = \{u_1^P, u_1^Q, u_2^P, u_2^Q, ..., u_n^P, u_n^Q\}$, $|\mathcal{V}| = 2 \times n$. In M modalities, $|\mathcal{V}| = M \times n$, M is the number of modalities and n is the number of utterances.

Edges. In a graph, an edge is defined as a connection be-

tween nodes. In the dialogue graph of multiple modalities, we define edges in two perspectives: the contextual connection of intra-modal utterance, and the interactive connection of inter-modal utterance. Particularly, we term these two types of edges as intra-edge ($\mathcal{E}_{intra}$) and inter-edge ($\mathcal{E}_{inter}$), respectively. The intra-edge is utilized for capturing intra-modal contextual information, whereas the inter-edge is utilized for capturing cross-modal interactive information.

The intra-edge is defined as follows. Assuming the existence of modality P, we connect the current utterance node u_i^P with the previous/past j utterance nodes $u_{i-j}^P, u_{i-j+1}^P, ..., u_{i-1}^P$. Similarly, we connect u_i^P with the next/future k utterance nodes $u_{i+1}^P, u_{i+2}^P, ..., u_{i+k}^P$. Therefore, we can formalize $\mathcal{E}_{intra}$ as follows:

$$\mathcal{E}_{intra} = \begin{cases} \{(u_t^P, u_i^P) | i - j < t < i - 1\} \\ \{(u_i^P, u_t^P) | i + 1 < t < i + k\} \end{cases}, \quad (3)$$

where i, j, k are constants, t is a variable; and i, j, k are less than n; i, j, k, t all belong to $\mathbb{N}_+$.

The inter-edge is defined as follows. In a dialogue, we connect the utterance node u_i^P of modality P to the corresponding utterance node u_i^Q of modality Q. Thus, we can formulate $\mathcal{E}_{inter}$ as follows:

$$\mathcal{E}_{inter} = \{(u_i^P, u_i^Q), (u_i^Q, u_i^P)\}, \quad (4)$$

where $i < n$ and $i \in \mathbb{N}_+$.

Edge Types. Based on the definition of edges above, we may divide all edges into two types: intra-edge type and inter-edge type, labeled as ET_{intra} and ET_{inter}, respectively. It is commonly known that if two utterance nodes in a dialogue has edge, it may or may not be from the same speaker. Therefore, ET_{intra} can be subdivided according to the perspective of speaker. Specifically, suppose that there are 3 speakers (s_1, s_2, s_3) in a dialogue, then the set of existing edge types when only a single modality is considered can be written as:

$$\begin{aligned} \mathrm{ET}_{intra} = \{&\mathrm{et}(s_1, s_1), \mathrm{et}(s_1, s_2), \mathrm{et}(s_1, s_3), \\ &\mathrm{et}(s_2, s_2), \mathrm{et}(s_2, s_3), \mathrm{et}(s_3, s_3)\}. \end{aligned} \quad (5)$$

It can be easily concluded that if there are D speakers in a dialogue, then there are $D \times (D+1)/2$ kinds of edges/relations. If there are M modalities, then there are $M \times (D^2 + D)/2$ elements in ET_{intra}.

ET_{inter} denotes the set of inter-modal edge types. Suppose that there are 3 modalities (mod_1, mod_2, mod_3) of the same utterance, then ET_{inter} can be formalized as:

$$\begin{aligned} \mathrm{ET}_{inter} = \{&\mathrm{et}(mod_1, mod_2), \mathrm{et}(mod_1, mod_3), \\ &\mathrm{et}(mod_2, mod_3)\}. \end{aligned} \quad (6)$$

Therefore, if there are M modalities in the same utterance, then existing $M \times (M-1)/2$ kinds of edges/relations. In this work, we consider three modalities of an utterance, so that there are a total of three elements in ET_{inter}.

Edge Weights. Edge weights are utilized to identify the relevance of distinct neighboring nodes when GNNs aggregates information. We employ a learnable attention module with edge features, which is detailed in Section III-C5.

2) Multi-subspace Extractor: Inspired by MMGCN [5], we consider the speaker information is of importance. The embedding of Multi-speaker S_{emb} can be formalized as:

$$S_{emb} = \mathrm{Embedding}(\mathrm{S}, D), \quad (7)$$

where S denotes the set of speakers, D is the number of speakers. To encode the speaker identity information, we add the speaker embedding to the features of utterance nodes:

$$\mathcal{X}_{spk}^\zeta = \mu S_{emb} + \mathcal{X}^\zeta, \quad (8)$$

where $\mathcal{X}^\zeta$ ($\zeta \in \{t, a, v\}$) is the feature matrix from uni-modal encoder, and $x_i^\zeta \in \mathcal{X}^\zeta$; $\mathcal{X}_{spk}^\zeta$ denotes the feature matrix adding the speaker embedding; $\mu \in [0, 1]$ is the ratio of the speaker embedding.

Currently, one of the quandaries we confront in multimodal fusion is the existence of heterogeneity gap [14]. In other words, the distribution of data is inconsistent across modalities. Therefore, before performing cross-modal feature complementation, we map the features of each modality into a shared subspace to maintain the consistency of feature representation across modalities. But yet the more similar the feature representations of multiple modalities are, the less complementary the feature between modalities are. To put it another way, we want to preserve the diversity of feature representations of multiple modalities so that the features of one modality can complement those of others. In view of this, we map the features of each modality into separate subspaces for capturing the diversity of feature representations across modalities. We argue that capturing the diversity and consistency information of multiple modalities simultaneously facilitates the complementation and fusion between modalities.

In order to capture the consistency of multimodal information in the shared subspace, we use three mapping functions $\mathcal{F}_{shr}$ with the same trainable parameter Θ_{shr}. For the separate subspace, we use three mapping functions $\mathcal{F}_{sep}$ with different trainable parameters Θ_{sep}^ζ to capture the variety of multimodal information. The two kinds of mapping methods are shown in Fig. 2, and are formulated as follows:

$$\begin{aligned} \mathrm{X}_{shr}^\zeta &= \mathcal{F}_{shr}(\mathcal{X}_{spk}^\zeta; \Theta_{shr}), \\ \mathrm{X}_{shr}^{vat} &= \mathrm{Lin}([\mathrm{X}_{shr}^v \| \mathrm{X}_{shr}^a \| \mathrm{X}_{shr}^t]; \Theta_{shr}'), \\ \mathrm{X}_{sep}^\zeta &= \mathcal{F}_{sep}(\mathcal{X}_{spk}^\zeta; \Theta_{sep}^\zeta), \end{aligned} \quad (9)$$

where $\|$ denotes concatenation operation; X_{shr}^ζ (X_{sep}^ζ) denotes the consistency (diversity) feature matrices, and $\zeta \in \{t, a, v\}$; Θ_{shr}, Θ_{shr}' and Θ_{sep}^ζ are the trainable parameters. In this paper, we define the mapping function $\mathcal{F}$ as follows:

$$\begin{aligned} \mathcal{F}(\mathcal{X}; \Theta) = \\ \mathrm{Norm}(\mathrm{Drop}(\mathrm{Lin}(\mathrm{Drop}(\sigma(\mathrm{Lin}(\mathcal{X}; \Theta_0)));\Theta_1))), \end{aligned} \quad (10)$$

the mapping function $\mathcal{F}$ is actually two fully connected layers; where $\mathcal{X}$ is the input of $\mathcal{F}$; Lin, σ, Drop and Norm denote the linear, non-linear activation, dropout and normalization functions, respectively; Θ denotes the learnable parameter.

Despite the fact that the shared mapping function $\mathcal{F}_{shr}$ and the separate mapping function $\mathcal{F}_{sep}$ are utilized to extract consistency and diversity features, they should have the equivalent
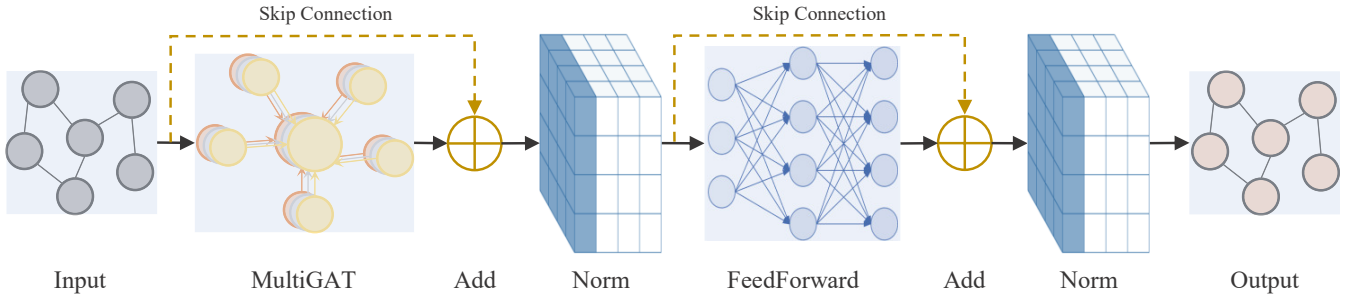
Fig. 3: The structure of the designed GAT-MLP, where the Norm operation can be placed before MultiGAT and FeedForward. MultiGAT denotes the multi-head graph attention network; the layer normalization function is used for the Norm in our work.

learning goal, i.e., the features of the same utterance mapped by different functions should correspond to the same emotion label. Therefore, we utilize four subspace loss functions to limit the features extracted by muti-subspace extractor such that they do not deviate from the ultimate goal task. The shared subspace loss function is computed as:

$$
\begin{aligned}
v_i' &= \mathrm{ReLU}(W_0' x_{shr,i}^{vat} + b_0'), \\
p_i' &= \mathrm{Softmax}(W_1' v_i^s + b_1'),
\end{aligned}
\tag{11}
$$

$$
\mathcal{L}_{shr} = -\frac{1}{\sum_{k=0}^{N-1} n(k)} \sum_{i=0}^{N-1} \sum_{j=0}^{n(i)-1} y_{ij} \log p_{ij}' + \lambda|\Theta_{re}'|, \tag{12}
$$

where $x_{shr,i}^{vat} \in X_{shr}^{vat}$; N is the number of dialogues, $n(i)$ is the number of utterances in dialogue i; y_{ij} denotes the ground truth label of the j-th utterance in the i-th dialogue, p_{ij}' denotes the probability distribution of predicted emotion label of the j-th utterance in the i-th dialogue; λ is the L2-regularizer weight, and W_0', W_1', b_0', b_1', Θ_{re}' are the trainable parameters. Similarly, the separate subspace loss function is computed as:

$$
\begin{aligned}
v_i^\zeta &= \mathrm{ReLU}(W_0^\zeta x_{sep,i}^\zeta + b_0^\zeta), \\
p_i^\zeta &= \mathrm{Softmax}(W_1^\zeta v_i^\zeta + b_1^\zeta),
\end{aligned}
\tag{13}
$$

$$
\mathcal{L}_{sep}^\zeta = -\frac{1}{\sum_{k=0}^{N-1} n(k)} \sum_{i=0}^{N-1} \sum_{j=0}^{n(i)-1} y_{ij} \log p_{ij}^\zeta + \lambda|\Theta_{re}^\zeta|, \tag{14}
$$

where $x_{sep,i}^\zeta \in X_{sep}^\zeta$, $\zeta \in \{t, a, v\}$ is the type of modalities, i.e., textual, acoustic and visual modalities; W_0^ζ, W_1^ζ, b_0^ζ, b_0^ζ, Θ_{re}^ζ are the learnable parameters.

3) GAT-MLP Layer: It is well known that the dilemma of over-smoothing exists in GNNs. Experimental studies have shown that the performance of the model begins to deteriorate dramatically as the number of layers in the GNN reaches a specific threshold. This is due to the impact of graph convolutions in that it inherently makes representations of adjacent nodes closer to each other [16]. Thus, after multiple graph convolutions, the node features within the same connected component of the network tend to be similar and the model degenerates.

Inspired by the ResNet [33] model, ResGCN [34] was proposed to address the over-smoothing of GNNs. ResGCN and most of the extant others, however, concatenate the output of each layer, which potentially limit the expressiveness of the

model. Transformer has seen success in the fields of computer vision [35], [36], natural language processing [37], and speech recognition [38] in recent years, and its network structure is regarded as excellent. As depicted in Fig. 3, we design a new GNN layer called GAT-MLP based on the ideas of ResNet and Transformer. The GAT-MLP layer can be formulated as:

$$
\begin{aligned}
X_{gat} &= \mathrm{Norm}(\mathrm{MultiGAT}(\mathcal{E}, X_{in}; \Theta_{gat}) + X_{in}), \\
X_{out} &= \mathrm{Norm}(\mathrm{FeedForward}(X_{gat}; \Theta_{fed}) + X_{gat}),
\end{aligned}
\tag{15}
$$

where X_{in} (X_{out}) denotes the input (output) matrix of node features; $\mathcal{E}$ denotes the set of edge; MultiGAT is the multi-head graph attention network; Θ_{gat} and Θ_{fed} are the trainable parameters; FeedForward and Norm are the feedforward and normalization functions, respectively. The layer normalization function is used as Norm in this work. The feedforward function is computed as follows:

$$
\begin{aligned}
\mathrm{FeedForward}(X_{gat}, \Theta_{fed}) = \\
\mathrm{Drop}(\mathrm{Lin}(\mathrm{Drop}(\sigma(\mathrm{Lin}(X_{gat}; \Theta_0)))); \Theta_1)),
\end{aligned}
\tag{16}
$$

where Drop and Lin are the dropout and linear functions, respectively; σ is the non-linear activation function (e.g., Relu); Θ_0 and Θ_1 are the trainable parameters. The MultiGAT is designed as follows:

$$
\begin{aligned}
\mathrm{MultiGAT}(\mathcal{E}, X_{in}; \Theta_h) &= \Theta_h[\mathrm{head}_1 \,\|\, ... \,\|\, \mathrm{head}_h], \\
\textit{where } \mathrm{head}_i &= \mathrm{SingleGAT}(\mathcal{E}, X_{in}; \Theta_i),
\end{aligned}
\tag{17}
$$

where SingleGAT in this paper will be described in detail in Section III-C5.

If the Norm operation is placed before MultiGAT and FeedForward, then it can be modified as follows:

$$
\begin{aligned}
X_{gat} &= \mathrm{MultiGAT}(\mathcal{E}, \mathrm{Norm}(X_{in}); \Theta_{gat}) + X_{in}, \\
X_{out} &= \mathrm{FeedForward}(\mathrm{Norm}(X_{gat}); \Theta_{fed}) + X_{gat}.
\end{aligned}
\tag{18}
$$

We argue intuitively that unlike the textual and acoustic tasks which rely on sequence-level context modeling, the visual task relies more on the features directly expressed in the current image. A model that is capable of both sequence-level context modeling and feature-level modeling is desired for the multimodal feature fusion. The MultiGAT sublayer in GAT-MLP can capture sequence-level contextual information, while the FeedForward sublayer compensates for the failure to efficiently capture feature-level information. Therefore, the combination of the MultiGAT and FeedForward in the GAT-MLP layer can mutually compensate for encoding disparities of various modalities in the multimodal task.

4) GAT-MLP based PairCC: If the features of multiple modalities are concatenated together directly, then it will not only be challenging to fuse due to heterogeneity gap, but it will also neglect cross-modal interactive information. What's worse, the larger the number of modalities is, the more serious the problem of heterogeneity gap between modalities is. Therefore, we propose the strategy of GAT-MLP based Pair-wise Cross-modal Complementation (PairCC) for cross-modal feature interaction and minimizing the heterogeneity gap. The process of GAT-MLP based PairCC is shown in Fig. 2, which mainly consists of GAT-MLP and concatenation layer. Specifically, we first feed the feature matrices of visual and acoustic modality into GAT-MLP layer for intra-modal context and inter-modal interaction encoding, and concatenate the visual and acoustic encoding results to obtain the v-a (visual-acoustic) feature matrix H^{va}; then we perform the same encoding operation between H^{va} and the textual feature matrix to obtain the v-a-t (visual-acoustic-textual) feature matrix H^{vat}; finally, H^{vat} and the feature matrix of shared subspace are encoded similarly to obtain the final feature matrix. The above steps can be formulated simply as follows:

$$\begin{aligned} \mathrm{H}^{va} &= \mathrm{PairCC}(\mathrm{X}^{v}_{sep}, \mathrm{X}^{a}_{sep}; \Theta^{va}_{sep}), \\ \mathrm{H}^{vat} &= \mathrm{PairCC}(\mathrm{X}^{t}_{sep}, \mathrm{H}^{va}; \Theta^{vat}_{sep}), \\ \mathrm{H}' &= \mathrm{PairCC}(\mathrm{X}^{vat}_{shr}, \mathrm{H}^{vat}; \Theta'), \end{aligned} \quad (19)$$

where H' is the final output of feature matrix; Θ^{va}_{sep}, Θ^{vat}_{sep} and Θ' are the trainable parameters, respectively. PairCC indicates GAT-MLP based PairCC function.

5) SingleGAT: We will present the graph attention network of this paper in this part. Graph Neural Networks (GNNs) typically involve two processes: aggregating information with the aggregation function and updating state with the combination function. Following that, we'll describe our SingleGAT (single-head graph attention) sublayer in terms of the aggregation function AGG and combination function COM. The two processes can be formalized as follows:

$$\begin{aligned} \mathrm{x}_{agg,i} &= \mathrm{AGG}(\{\mathrm{x}_j | w_j \in \mathcal{N}(w_i)\}; \Theta_{agg}), \\ \mathrm{x}_{com,i} &= \mathrm{COM}(\mathrm{x}_i, \mathrm{x}_{agg,i}; \Theta_{com}), \end{aligned} \quad (20)$$

where AGG and COM are the aggregation and combination functions, respectively; $\mathrm{x}_i \in \mathrm{X}$ denotes the feature vector of node w_i, and $w_i \in \mathcal{V}$; x_j is the feature vector of w_i's neighbor w_j; Θ_{agg} and Θ_{com} denote the learnable parameters.

Aggregation. It is well known that computing the importance of neighbor information is crucial when GNNs aggregate information. So we utilize the attention mechanism to implement the aggregation function AGG. The output of aggregation is expressed as follows:

$$\mathrm{x}_{agg,i} = \sum_{w_j \in \mathcal{N}(w_i)} \alpha_{ij} \mathrm{W}_{agg} \mathrm{x}_j, \quad (21)$$

where α_{ij} is attention coefficient, as well as the edge weight between node w_i and w_j; w_j is the neighboring node of w_i; x_j denotes the feature vector of w_j, and $\mathrm{x}_j \in \mathrm{X}$; W_{agg} denotes the learnable parameter.

GATv2 based Edge Weights. We use the attention module of GATv2 [39] to learn edge weights for characterizing the

relevance of diverse neighbor information. We define the attention coefficient α_{ij} as follows:

$$\alpha_{ij} = \frac{\exp\left(\mathrm{a}^\top \sigma\left(\Theta_{\mathrm{att}}[\mathrm{x}_i \,\|\, \mathrm{x}_j]\right)\right)}{\sum_{w_k \in \mathcal{N}(w_i)} \exp\left(\mathrm{a}^\top \sigma\left(\Theta_{\mathrm{att}}[\mathrm{x}_i \,\|\, \mathrm{x}_k]\right)\right)}, \quad (22)$$

where σ denotes the non-linear activation function, such as LeakyReLU; $\|$ denotes the concatenation operation; Θ_{att} is the learnable parameter; x_i is the feature representation of the current node w_i; both x_j and x_k are the representations of neighboring node of w_i. In this work, the neighboring node is either an intra-modal contextual node or an inter-modal interactive node of w_i.

Embedding of Edge Types. We assume that different types of edge/relation involve different implicit dependency information in the dialogue. Here are two conjectures:

① Suppose w_j is an intra-modal contextual neighbor node of w_i (w_j is a long-distance contextual node). w_j and w_i may have similar semantics when they are uttered by the same speaker. At this moment, w_j is more critical relative to others that have different speakers from w_i.

② Suppose w_j is the inter-modal interactive neighbor node of w_i. When the semantics of w_i does not match the ground-truth emotion label, w_j can be semantically complementary to w_i. w_j is more important at this time relative to other neighbor nodes.

Therefore, we encode the edge types as vector representations, and put them into the attention module to aid in the computation of the attention coefficient. We consider that the edge weight is affected not only by the nodes, but also by the edge types. The embedding of edge types, i.e., feature of edge type, can be formalized as follows:

$$\begin{aligned} ET_{emb} &= \mathrm{Embedding}(\mathrm{ET}, DM), \\ where\ DM &= M \times (D^2 + D + M - 1)/2, \end{aligned} \quad (23)$$

where $\mathrm{ET} = \mathrm{ET}_{intra} \cup \mathrm{ET}_{inter}$ denotes the set of edge types, and DM is the number of edge types in a dialogue with D speakers and M modalities. The attention coefficient with the addition of edge feature is computed as follows:

$$\alpha'_{ij} = \frac{\exp\left(\mathrm{a}^\top \sigma\left(\Theta_{\mathrm{att}}[\mathrm{x}_i \,\|\, \mathrm{x}_j \,\|\, \mathrm{et}_{ij}]\right)\right)}{\sum_{w_k \in \mathcal{N}(w_i)} \exp\left(\mathrm{a}^\top \sigma\left(\Theta_{\mathrm{att}}[\mathrm{x}_i \,\|\, \mathrm{x}_k \,\|\, \mathrm{et}_{ik}]\right)\right)}, \quad (24)$$

where $\mathrm{et}_{ij} \in ET_{emb}$ denotes the edge feature of between utterance node w_i and w_j.

Combination. The combination function COM combines $\mathrm{x}_{agg,i}$ with x_i. We employ GRU as the combination function, which is inspired by GraphSage [30] but different from it. The output of the graph attention is expressed as follows:

$$\mathrm{x}^1_{com,i} = \mathrm{GRU}(\mathrm{x}_i, \mathrm{x}_{agg,i}; \Theta^1_{com}), \quad (25)$$

where $\mathrm{x}_{com,i}$, x_i and $\mathrm{x}_{agg,i}$ are the output, input and hidden state of GRU, respectively; Θ_{com} is the trainable parameter. The neighbor information $\mathrm{x}_{agg,i}$ (including intra-modal contextual information and inter-modal interactive information) is employed as the hidden state of GRU, and it may not be completely exploited. Therefore, we reverse the order of x_i

and $\mathrm{x}_{agg,i}$, i.e., $\mathrm{x}_{agg,i}$ and x_i are respectively utilized as the input and hidden state of GRU:

$$\mathrm{x}^2_{com,i} = \mathrm{GRU}(\mathrm{x}_{agg,i}, \mathrm{x}_i; \Theta^2_{com}), \qquad (26)$$

where $\mathrm{x}'_{com,i}$ is the output, input and hidden state of GRU, and Θ'_{com} is the trainable parameter. The final output of the single-head graph attention SingleGAT as follows:

$$\mathrm{x}_{com,i} = \mathrm{x}^1_{com,i} + \mathrm{x}^2_{com,i}, \qquad (27)$$

By calculating the average of multiple single-head graph attentions, we can obtain the following result:

$$\mathrm{x}_{\mathrm{gat},i} = \frac{1}{K}\sum_{k=1}^{K}\mathrm{x}^k_{com,i}, \qquad (28)$$

where $\mathrm{x}_{\mathrm{gat},i} \in \mathrm{X}_{\mathrm{gat}}$ is the output of the multi-head graph attention network, and K denotes the number of heads.

D. Multimodal Emotion Classifier

After encoding with the GAT-MLP based PairCC, the feature vector $h_i \in H'$ of utterance u_i can be obtained. It is then fed to the fully connected layer to predict the emotion label $\hat{y}_i$ for the utterance u_i:

$$
\begin{aligned}
v_i &= \mathrm{ReLU}(\mathrm{W}_0 h_i + \mathrm{b}_0), \\
p_i &= \mathrm{Softmax}(\mathrm{W}_1 v_i + \mathrm{b}_1), \\
\hat{y}_i &= \underset{\mathrm{k}}{\mathrm{argmax}}(p_i[\mathrm{k}]),
\end{aligned}
\qquad (29)
$$

We employ cross-entropy loss along with L2-regularization as classification loss function to train the model:

$$\mathcal{L}_{cls} = -\frac{1}{\sum_{k=0}^{N-1}n(k)}\sum_{i=0}^{N-1}\sum_{j=0}^{n(i)-1} y_{ij}\log p_{ij} + \lambda|\Theta_{re}|, \quad (30)$$

where N is the number of dialogues, $n(i)$ is the number of utterances in dialogue i; y_{ij} denotes the ground truth label of the j-th utterance in the i-th dialogue, p_{ij} denotes the probability distribution of predicted emotion label of the j-th utterance in the i-th dialogue; λ is the L2-regularizer weight, and Θ_{re} is the trainable parameter.

Finally, combining the shared subspace loss $\mathcal{L}_{shr}$, separate subspace losses $\mathcal{L}^{\zeta}_{sep}$ ($\zeta \in \{a, v, t\}$) and classification loss $\mathcal{L}_{cls}$ together, the final objective function is computed as:

$$\mathcal{L} = \mathcal{L}_{cls} + \beta\mathcal{L}_{shr} + \gamma^a\mathcal{L}^a_{sep} + \gamma^v\mathcal{L}^v_{sep} + \gamma^t\mathcal{L}^t_{sep}, \qquad (31)$$

where β, γ^a, γ^v, γ^t are the trade-off parameters.

IV. EXPERIMENT

A. Datasets and Evaluation Metrics

1) Datasets: We evaluate our GraphCFC model on two multimodal benchmark datasets: IEMOCAP [40] and MELD [18], which are subjected to raw utterance-level feature extraction according to MMGCN [5]. The statistics of them are shown in Table I.

IEMOCAP is a multimodal dataset of two-way conversations from ten professional actors. It contains 151 conversations, a total of 7433 dyadic utterances. Emotion labels of IEMOCAP include *Neutral*, *Happy*, *Sad*, *Angry*, *Frustrated*

TABLE I: The statistics of IEMOCAP and MELD

Dataset	Dialogues			Uterances			Classes	Speakers in a Dialogue
	train	valid	test	train	valid	test		
IEMOCAP	120		31	5810		1623	6	2
MELD	1039	114	280	9989	1109	2610	7	3 or more

and *Excited*. As in previous works [2], [5], we utilize the first 80% of the data as the training set and the remaining data as the test set, with the 10% of the training set used as the validation set. IEMOCAP is one of the most popular datasets in ERC task, with high quality and multimodal information.

MELD is a multimodal dataset, containing videos of multi-party conversations from Friends TV series. It involvs over 1433 conversations, a total of 13708 utterances by 304 speakers. Distinct from IEMOCAP, each conversation in MELD includes three or more speakers. Emotion labels include *Anger*, *Disgust*, *Sadness*, *Joy*, *Neutral*, *Surprise* and *Fear*. The conversations in this dataset involve many backgrounds knowledge, which makes it challenging to recognize the right emotion.

2) Metrics: Following the previous methods [2], [5], we chose weighted-average F1 score as the evaluation metric due to the class imbalanced problem. F1 score is reported for each class to allow for a more comprehensive comparison with the baselines. We also record the average accuracy score in addition to the weighted-average F1 score.

B. Baselines

To verify the effectiveness of our proposed GraphCFC model, we compare it with several previous baselines. The baselines include bc-LSTM [11], CMN [12], ICON [13], DialogueRNN [3], DialogueGCN [2], DialogueCRN [4] and MMGCN [5]. The details of these models are listed as follows.

bc-LSTM encodes context-aware information through a bidirectional LSTM network, but without taking speaker-related information into account. **CMN** models utterance context through speaker-dependency GRUs, but it can only work when the conversation includes two speakers. **ICON** has improved CMN by modeling distinct speakers. A global GRU is utilized to model the variance of emotion status in a conversation. Nevertheless, ICON still can't be applied in scenario with more than two speakers. **DialogueRNN** leverages three GRUs to model information of speakers and sequence in conversations, which contain Global GRU, Speaker GRU and Emotion GRU. The goals of three GRUs are to extract context information, model identity information of the speakers and detect emotion of utterances, respectively. **DialogueGCN** focuses on the function of GCN, i.e., aggregating neighbor information, to improve the performance of ERC tasks. We extend DialogueGCN by directly concatenating features of each modality to implement multimodal setting. **DialogueCRN** extracts and integrates emotional clues by devising multi-turn reasoning modules to sufficiently model the situation-level and speaker-level context in a conversation. In order to achieve multimodal setting, we concatenate features of three modalities simply. **MMGCN** adopts a graph-based approach for multimodal feature fusion. MMGCN is currently

TABLE II: The overall performance of all models on both IEMOCAP and MELD datasets under the multimodal setting. IEMOCAP includes 6 labels, and MELD includes 7 labels (2 of them, i.e., *Fear* and *Disgust*, are not reported due to their statistically insignificant results). Evaluation metrics contain accuracy score (Accuracy (%)) and weighted-average F1 score (wa-F1 (%)). Best performances are highlighted in bold.

Model	IEMOCAP								MELD						
	Happy	*Sad*	*Neutral*	*Angry*	*Excited*	*Frustrated*	Accuracy	wa-F1	*Neutral*	*Surprise*	*Sadness*	*Joy*	*Anger*	Accuracy	wa-F1
bc-LSTM	32.63	70.34	51.14	63.44	67.91	61.06	59.58	59.10	75.66	48.47	22.06	52.10	44.39	59.62	56.80
CMN	30.38	62.41	52.39	59.83	60.25	60.69	56.56	56.13	-	-	-	-	-	-	-
ICON	29.91	64.57	57.38	63.04	63.42	60.81	59.09	58.54	-	-	-	-	-	-	-
DialogueRNN	33.18	78.80	59.21	65.28	71.86	58.91	63.40	62.75	76.79	47.69	20.41	50.92	45.52	60.31	57.66
DialogueCRN	**51.59**	74.54	62.38	67.25	73.96	59.97	65.31	65.34	76.13	46.55	11.43	49.47	44.92	59.66	56.76
DialogueGCN	47.10	80.88	58.71	66.08	70.97	61.21	65.54	65.04	75.97	46.05	19.60	51.20	40.83	58.62	56.36
MMGCN	45.45	77.53	61.99	66.67	72.04	**64.12**	65.56	65.71	75.16	48.45	25.71	**54.41**	45.45	59.31	57.82
GraphCFC	43.08	**84.99**	**64.70**	**71.35**	**78.86**	63.70	**69.13**	**68.91**	**76.98**	**49.36**	**26.89**	51.88	**47.59**	**61.42**	**58.86**

significantly superior to most baselines for multimodal ERC, which provides a new idea for multimodal fusion.

C. Implementation Details

We implement the GraphCFC model through the PyTorch framework, and all experiments are executed on NVIDIA Tesla A100. The optimizer is AdamW, the L2 regularization parameter is 0.00001, and the Dropout rate is 0.1. For IEMOCAP dataset, the number of GAT-MLP layers is 5, the learning rate is 0.00001, the ratio of the speaker embedding μ is 1.0, and the batch size is 8. For MELD dataset, the number of GAT-MLP layers is 3, the learning rate is 0.00001, the ratio of the speaker embedding μ is 0.7, and the batch size is 32. We utilize the method proposed by Kendall et al. [41] to set the trade-off parameters (β, γ^a, γ^v and γ^t) of multiple loss functions as learnable parameters instead of setting them manually.

V. RESULTS AND ANALYSIS

In this section, we report and discuss the results of all comparative experiments and ablation studies. In addition, we provide three case studies on the IEMOCAP dataset at the end of this section.

A. Overall Performance

We compare our proposed GraphCFC with the baseline models on the IEMOCAP and MELD datasets. The overall performance of all models is shown in Table II. Based on the experimental findings, we can see that the accuracy and weighted-average F1 score of our proposed model is 3.57% and 3.20% higher than those of the best baseline model (i.e., MMGCN) on the IEMOCAP dataset. The GraphCFC model achieves higher F1 scores than MMGCN in the most emotions when each emotion is observed separately. The F1 scores of *Sad* and *Excited* are respectively 84.99% and 78.86% on the IEMOCAP dataset, which are higher than those of other emotions. For *Sad* and *Excited* emotions, the F1 scores of GraphCFC are far higher than those of MMGCN, which are 7.46% and 6.82% higher than those of MMGCN, respectively. Overall, the GraphCFC model outperforms the others in terms of accuracy and weighted-average F1 scores. Therefore, we can conclude that our method can more adequately extract

long-distance intra-modal contextual information and inter-modal interactive information in comparison to the baselines such as MMGCN.

We note that while DialogueCRN can achieve excellent performance in uni-modal setting [4], direct concatenation of the results from multiple modalities is not as effective. One probable reason is that direct concatenation generates redundant information and fails to capture the inter-modal interactive information. GraphCFC, in contrast, extracts the interactive information through a GNN-based approach while also reducing redundant information, resulting in superior performances.

Although GraphCFC model outperforms other approaches on the MELD dataset, its improvement was not very significant. It can be observed that the improvements in accuracy and weighted-average F1 scores of GraphCFC are 2.11% and 1.04%, respectively, relative to those of MMGCN. The reason for these results may be that the utterance sequences of a dialogue on the MELD dataset aren't from a continuous conversation in real scenes. Therefore, the graph-based models do not take advantage of their ability to capture contextual information. Another potential reason is that the MELD dataset contains a lot of background noise that is unrelated to emotion due to the camera setup and recording conditions. In addition, we observe that the F1 score of the *Sadness* class is low in the results of the MELD dataset. By looking at the distribution of classes in the MELD dataset, we find that the dataset suffers from the class-imbalanced problem. And the *Sadness* class belongs to the minority class, thus resulting in its low F1 score.

B. Comparison under Various Modality Settings

Table III shows the performance of GraphCFC model in different modality settings. Overall, the performance of multimodal settings outperforms that of uni-modal settings. The textual modality has the best performance in the uni-modal settings, whereas the visual modality has the lowest result. One probable explanation for the poor result is that the visual modality contains a lot of noise due to the effects of camera position, background, lighting, etc. The combination of textual and acoustic modalities produces the best performance in the two-modal settings, whereas the combination of visual and acoustic modalities produces the worst result. As expected, the

TABLE III: The performance of GraphCFC under various modality setting. A, V, and T indicate acoustic, visual, and textual modalities, respectively. Note that shared space mapping and $\mathcal{E}_{inter}$ is not performed in uni-modal settings.

Modality Setting	IEMOCAP		MELD	
	Accuracy	wa-F1	Accuracy	wa-F1
A	54.16	53.85	47.55	41.62
V	31.61	27.67	47.59	33.26
T	59.95	60.09	60.77	56.81
A + V	54.17	53.89	47.61	41.67
A + T	64.20	64.74	59.96	57.46
V + T	63.15	62.96	59.46	57.29
A + V + T	**69.13**	**68.91**	**61.42**	**58.86**

fusion of acoustic, visual and textual modalities can improve the performance of GraphCFC.

C. Effect of Various Components in GAT-MLP Layer

TABLE IV: The results of GraphCFC when $\mathrm{MultiGAT}$ or $\mathrm{FeedForward}$ is not used in the GAT-MLP Layer. w/o and w are respectively $\mathrm{without}$ and with, where w/o (w) indicates the non-use (use) of $\mathrm{MultiGAT}$ or $\mathrm{FeedForward}$ sublayer.

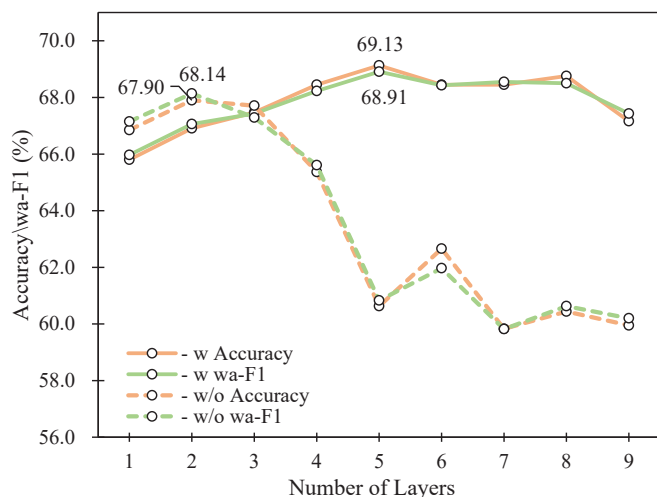| MultiGAT	FeedForward	IEMOCAP		MELD	
		Accuracy	wa-F1	Accuracy	wa-F1
- w/o	- w	64.20	64.55	60.50	58.03
- w	- w/o	67.04	67.24	60.84	58.43
- w	- w	**69.13**	**68.91**	**61.42**	**58.86**

Fig. 4: The effects of the number of GAT-MLP layer and skip connection on our model. The figure shows the results on the IEMOCAP dataset. w (w/o) indicates the use (non-use) of skip connection.

We report the effects of $\mathrm{MultiGAT}$ and $\mathrm{FeedForward}$ in the GAT-MLP layer in Table IV. The performance of our model is noticeably degraded when the $\mathrm{MultiGAT}$ or $\mathrm{FeedForward}$ sublayer is not adopted. When the $\mathrm{MultiGAT}$ sublayer is not utilized on the IEMOCAP dataset, the accuracy and weighted-average F1 scores of GraphCFC decrease by 4.93% and

4.36%, respectively. The accuracy and F1 scores respectively decrease by 2.09% and 1.67% when the $\mathrm{FeedForward}$ sublayer is not applied. As a result, we can deduce that the effect of $\mathrm{MultiGAT}$ in the GAT-MLP layer is more significant than that of $\mathrm{FeedForward}$.

The effects of different numbers of GAT-MLP layer and skip connection on the GraphCFC model are shown in Fig. 4. We can see that if we remove the skip connection, the performance of the model will drop sharply on the IEMOCAP dataset as the number of GAT-MLP layer increases when a certain threshold is exceeded. On the contrary, if we keep the skip connection, the performance of the proposed GraphCFC decreases slowly. Therefore, skip connection can help to mitigate the problem of over-smoothing to a certain extent.

D. Effect of Multi-subspace Loss

The impacts of the multi-subspace loss functions are seen in Table V. After eliminating the shared subspace loss or separate subspace loss, both accuracy and weighted-average F1 scores decline, as seen in the Table V. The experimental results suggest that setting the loss function in the multi-subspace extractor can effectively improve the performance of our GraphCFC.

TABLE V: The effect of the multi-subspace loss functions. $\mathcal{L}_{shr}$ and $\mathcal{L}_{sep}^{\zeta}$ ($\zeta \in \{a, v, t\}$) denote shared and separate subspace losses, respectively.

$\mathcal{L}_{shr}$	$\mathcal{L}_{sep}^{\zeta}$	IEMOCAP		MELD	
		Accuracy	wa-F1	Accuracy	wa-F1
- w/o	- w	68.70	68.35	61.00	58.39
- w	- w/o	67.53	67.56	60.27	57.99
- w/o	- w/o	68.70	68.36	60.38	58.09
- w	- w	**69.13**	**68.91**	**61.42**	**58.86**

E. Effect of Multi-speaker and Edge Types

TABLE VI: The influence of speakers and edge types on our GraphCFC model. S_{emb} and E_{emb} indicate the embeddings of multi-speaker and edge types, respectively.

S_{emb}	E_{emb}	IEMOCAP		MELD	
		Accuracy	wa-F1	Accuracy	wa-F1
- w/o	- w	68.02	68.04	60.69	58.35
- w	- w/o	65.26	65.91	60.46	57.91
- w	- w	**69.13**	**68.91**	**61.42**	**58.86**

The influence of speakers and edge types on our GraphCFC model is seen in Table VI. The performance of GraphCFC will be compromised if the embedding of multi-speaker or edge types is not employed. The weighted-average F1 score drops to 65.91% on the IEMOCAP dataset when the embedding of edge types is not utilized, which amply proves our hypothesis that edge types affects the relevance of neighbor information. We note that without adding speaker information, the results of GraphCFC show only a slight degradation, which is still higher than the results of baseline models. The phenomenon demonstrates that GrpahCFC is not heavily dependent on

speaker and has a certain degree of generalization capability. Generally speaking, the performance of our proposed method can be improved by adding the embeddings of multi-speaker and edge types.

F. Effect of the Past j and Future k Utterance Nodes

As shown in Fig. 5, we discuss the effect of past j nodes and future k nodes on our proposed GraphCFC model. We set j and k to multiple combinations (the combination can be denoted as (j, k)), such as $(0, 0)$, $(2, 2)$, $(4, 4)$, $(6, 6)$, ..., $(40, 40)$. From Fig. 5, it can be concluded that the accuracy and weighted-average F1 scores increase on the IEMOCAP dataset with increasing values of j and k. When a certain threshold combination (i.e., $(j, k) = (18, 18)$) are reached, however, the accuracy and F1 scores gradually decrease. In particular, GraphCFC performs worst when the conversational context is not available (i.e., setting both j and k set to 0). Therefore, we can draw the conclusion that the conversational context is a crucial parameter for the proposed method.
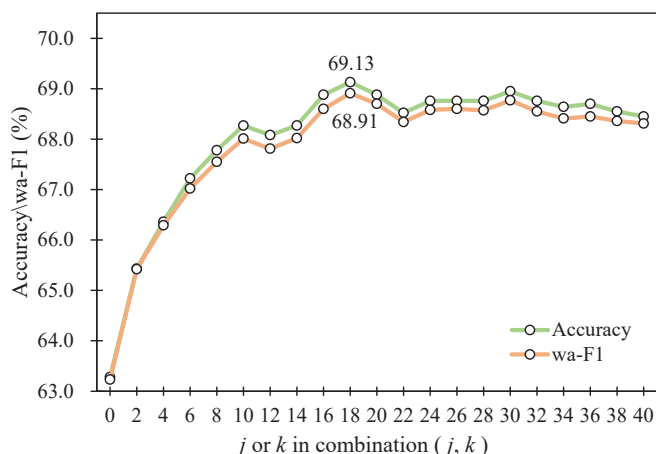

Fig. 5: The effects of j nodes in the past and k nodes in the future on the proposed GraphCFC model. The figure shows the results on the IEMOCAP dataset.

G. Overall Performance of Three-emotion

In this part, we conduct comparative experiments of the three-emotion. Prior to model training, we merge the original emotion labels into three categories (i.e., *Positive*, *Neutral*, and *Negative*), while the proposed GraphCFC is transformed into a three-classification model. Specifically, Table VII shows the statistics of the merged emotion labels.

TABLE VII: The statistics of the merged emotion labels.

New Label	IEMOCAP	MELD
Positive	*Happy, Excited*	*Joy*
Negative	*Sad, Angry, Frustrated*	*Surprise, Fear, Sadness, Disgust, Anger*
Neutral	*Neutral*	*Neutral*

The experimental results of our three-emotion are recorded in Table VIII. We can find that the experimental results are similar to those of the previous experiments with six or seven-emotion, with improved accuracy and weighted-average F1 scores for all models. It can be seen that the accuracy and F1 scores of GraphCFC improve 0.61% and 1.20% relative to those of MMGCN on the MELD dataset, respectively. Similarly, there are slight improvements of accuracy and F1 scores on the IEMOCAP dataset. It may be attributable to the fact that the emotion labels are coarsened (similar emotions like *Excited* and *Happy* are merged) after converting the dataset into three-emotion labels, so most of the models are capable of performing the task of emotion classification easily.

H. Case Studies

As shown in Fig. 6, we conduct several case studies in this part. In text-modal ERC models such as DialogueGCN and DialogueCRN, several utterances with non-*Neutral* emotion labels, such as "*okay.*", "*yeah.*" or "*What's the meaning of that?*", are usually recognized as *Neutral* directly. In contrast, multimodal ERC models such as GraphCFC make integrated judgments based on multiple modalities, which, for example, may eventually be recognized as *Sad*. Therefore, visual and acoustic modalities can compensate for such lapses. Fig. 6a depicts the above-mentioned case on the IEMOCAP dataset.

Fig. 6b and Fig. 6c show that the cases of emotional-shift on the IEMOCAP dataset. In Fig. 6b, when a speaker's emotion is *Neutral* for several preceding consecutive utterances, most of the models (e.g., MMGCN) tend to identify the speaker's next utterance as *Neutral*. In Fig. 6c, when a speaker's emotion was *Neutral* for several consecutive utterances, the majority of models trend towards recognizing the next utterance spoken by another speaker as *Neutral*. Unlike approaches such as MMGCN, our proposed GraphCFC can accurately identify the emotion of utterance as *Excited* in the above two cases.

VI. CONCLUSION

In this paper, we propose a directed Graph based Cross-modal Feature Complementation (GraphCFC) method for reducing the multimodal heterogeneity gap and compensating the inadequacies of earlier SOTA methods such as MMGCN. Concretely, we model the multimodal dialogue as a directed graph with variable context and extract distinct types of edges from the graph for graph attention learning, thus ensuring that GNNs can select accurately critical intra-modal contextual and inter-modal interactive information; meanwhile, we also address the heterogeneity gap using multiple subspace mapping functions and PairCC strategy. In addition, we design a graph-based network structure named GAT-MLP, which provides a feasible solution for multimodal interaction. Experimental results on the IEMOCAP and MELD datasets show that our proposed GraphCFC outperforms other SOTA methods and is capable of effectively modeling long-distance intra-modal contextual information and inter-modal interactive information.

As we can see, some challenges of multimodal machine learning remain. In future work, we hope to further explore the methodologies of multimodal fusion and evaluate the validity of the GAT-MLP layer on other multimodal tasks. Another work in the future is how to alleviate the problem of class

TABLE VIII: The overall performance after converting the dataset into three-emotion labels under the multimodal setting.

| Model	IEMOCAP					MELD				
	Positive	*Neutral*	*Negative*	Accuracy	wa-F1	*Positive*	*Neutral*	*Negative*	Accuracy	wa-F1
bc-LSTM	**90.58**	55.63	84.04	79.54	79.10	36.97	75.12	61.46	65.13	64.26
DialogueRNN	88.36	57.99	83.81	78.87	78.94	40.29	74.95	62.10	65.52	64.93
DialogueCRN	79.39	61.51	83.09	75.66	76.97	40.80	74.40	62.87	65.98	65.32
DialogueGCN	84.22	56.88	83.66	77.57	77.48	32.92	**75.64**	63.96	66.67	64.80
MMGCN	85.20	**64.21**	83.73	79.36	79.95	43.32	75.5	65.57	67.93	66.92
GraphCFC	88.48	62.03	**84.35**	**79.91**	**80.20**	**50.66**	75.12	**66.26**	**68.54**	**68.12**
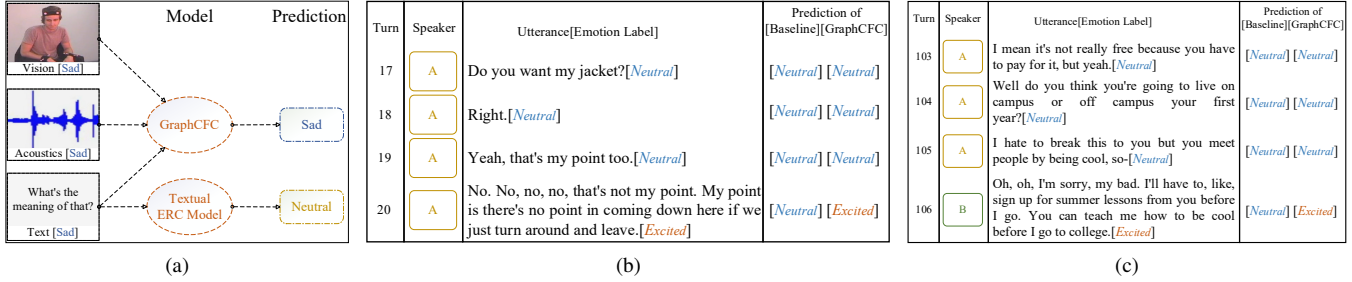
Fig. 6: The cases of ERC on the IEMOCAP. (a) An example shows that multi-modality can be used to compensate for the shortcoming of single-textual modality. (b) Emotional-shift in one-speaker scenario. (c) Emotional-shift in two-speaker scenario.

imbalance and background noise on the MELD dataset in the conversational emotion recognition task.

REFERENCES

[1] G. A. van Kleef and S. Côté, "The social effects of emotions," *Annual review of psychology*, vol. 73, pp. 629–658, 2022.

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

[3] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, and E. Cambria, "Dialoguernn: An attentive rnn for emotion detection in conversations," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 6818–6825.

[4] D. Hu, L. Wei, and X. Huai, "Dialoguecrn: Contextual reasoning networks for emotion recognition in conversations," *arXiv preprint arXiv:2106.01978*, 2021.

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

[6] Z. Huang, J. Epps, and D. Joachim, "Speech landmark bigrams for depression detection from naturalistic smartphone speech," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5856–5860.

[7] A. Chatterjee, K. N. Narahari, M. Joshi, and P. Agrawal, "Semeval-2019 task 3: Emocontext contextual emotion detection in text," in *Proceedings of the 13th international workshop on semantic evaluation*, 2019, pp. 39–48.

[8] C. Guo, J. Cao, X. Zhang, K. Shu, and H. Liu, "Dean: Learning dual emotion for fake news detection on social media," *arXiv e-prints*, pp. arXiv–1903, 2019.

[9] C. Huang, O. R. Zaiane, A. Trabelsi, and N. Dziri, "Automatic dialogue generation with expressed emotions," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 2018, pp. 49–54.

[10] W. Shen, S. Wu, Y. Yang, and X. Quan, "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*. Association for Computational Linguistics, 2021, pp. 1551–1560.

[11] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.-P. Morency, "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)*, 2017, pp. 873–883.

[12] D. Hazarika, S. Poria, A. Zadeh, E. Cambria, L.-P. Morency, and R. Zimmermann, "Conversational memory network for emotion recognition in dyadic dialogue videos," in *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, vol. 2018. NIH Public Access, 2018, p. 2122.

[13] D. Hazarika, S. Poria, R. Mihalcea, E. Cambria, and R. Zimmermann, "Icon: Interactive conversational memory network for multimodal emotion detection," in *Proceedings of the 2018 conference on empirical methods in natural language processing*, 2018, pp. 2594–2604.

[14] W. Guo, J. Wang, and S. Wang, "Deep multimodal representation learning: A survey," *IEEE Access*, vol. 7, pp. 63 373–63 394, 2019.

[15] D. Hazarika, R. Zimmermann, and S. Poria, "Misa: Modality-invariant and-specific representations for multimodal sentiment analysis," in *Proceedings of the 28th ACM international conference on multimedia*, 2020, pp. 1122–1131.

[16] Q. Li, Z. Han, and X.-M. Wu, "Deeper insights into graph convolutional networks for semi-supervised learning," in *Thirty-Second AAAI conference on artificial intelligence*, 2018.

[17] N. Majumder, P. Hong, S. Peng, J. Lu, D. Ghosal, A. Gelbukh, R. Mihalcea, and S. Poria, "Mime: Mimicking emotions for empathetic response generation," *arXiv preprint arXiv:2010.01454*, 2020.

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

[19] T. Ishiwatari, Y. Yasuda, T. Miyazaki, and J. Goto, "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)*, 2020, pp. 7360–7370.

[20] D. Zhang, L. Wu, C. Sun, S. Li, Q. Zhu, and G. Zhou, "Modeling both context-and speaker-sensitive dependence for emotion detection in multi-speaker conversations," in *IJCAI*, 2019, pp. 5415–5421.

[21] P. Zhong, D. Wang, and C. Miao, "Knowledge-enriched transformer for emotion detection in textual conversations," *arXiv preprint arXiv:1909.10681*, 2019.

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

[23] D. Ghosal, N. Majumder, A. Gelbukh, R. Mihalcea, and S. Poria, "Cosmic: Commonsense knowledge for emotion identification in conversations," *arXiv preprint arXiv:2010.02795*, 2020.

[24] S. Sahay, S. H. Kumar, R. Xia, J. Huang, and L. Nachman, "Multimodal

relational tensor network for sentiment and emotion classification," *arXiv preprint arXiv:1806.02923*, 2018.

[25] Y.-H. H. Tsai, P. P. Liang, A. Zadeh, L.-P. Morency, and R. Salakhutdinov, "Learning factorized multimodal representations," *arXiv preprint arXiv:1806.06176*, 2018.

[26] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proceedings of the conference. Association for Computational Linguistics. Meeting*, vol. 2019, 2019, p. 6558.

[27] Z. Sun, P. Sarma, W. Sethares, and Y. Liang, "Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, 2020, pp. 8992–8999.

[28] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 2, pp. 423–443, 2018.

[29] S. K. D'mello and J. Kory, "A review and meta-analysis of multimodal affect detection systems," *ACM computing surveys (CSUR)*, vol. 47, no. 3, pp. 1–36, 2015.

[30] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," *Advances in neural information processing systems*, vol. 30, 2017.

[31] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," *stat*, vol. 1050, p. 20, 2017.

[32] J. Chen, T. Ma, and C. Xiao, "Fastgcn: Fast learning with graph convolutional networks via importance sampling," in *6th International Conference on Learning Representations*. OpenReview.net, 2018.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[34] G. Li, M. Müller, G. Qian, I. C. D. Perez, A. Abualshour, A. K. Thabet, and B. Ghanem, "Deepgcns: Making gcns go as deep as cnns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[35] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[36] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 012–10 022.

[37] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[38] Q. Zhang, H. Lu, H. Sak, A. Tripathi, E. McDermott, S. Koo, and S. Kumar, "Transformer transducer: A streamable speech recognition model with transformer encoders and rnn-t loss," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7829–7833.

[39] S. Brody, U. Alon, and E. Yahav, "How attentive are graph attention networks?" in *International Conference on Learning Representations*, 2022.

[40] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[41] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7482–7491.
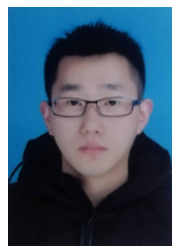
Jiang Li received the bachelor's degree in software engineering in 2018, and master's degree in computer science in 2021. He is a doctoral candidate in artificial intelligence, currently studying at the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, China, and also at the Key Laboratory of Image Processing and Intelligent Control of the Education Ministry of China. His current research interests include the theory and application of graph neural networks, multimodal machine learning and affective computing.

Xiaoping Wang received the B.S. and M.S. degrees in automation from Chongqing University, Chongqing, China, in 1997 and 2000, respectively, and the Ph.D. degree in systems engineering from Huazhong University of Science and Technology, Wuhan, China, in 2003. Since 2011, she has been a Professor at the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, and also at the Key Laboratory of Image Processing and Intelligent Control of the Education Ministry of China. Her current research interests include affective computing, memristors and its applications to memory storage, modeling, and simulation.
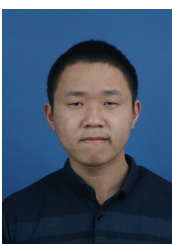
Guoqing Lv received the B.E. degree in automation from Shandong University, Jinan, China, in 2021. He is currently pursuing the M.S. degree in control science and engineering at the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, China, and also at the Key Laboratory of Image Processing and Intelligent Control of the Education Ministry of China. His current research interests include graph neural networks, multimodal emotion recognition and empathic response generation.

Zhigang Zeng received the Ph.D. degree in systems analysis and integration from Huazhong University of Science and Technology, Wuhan, China, in 2003. He is currently a Professor at the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, China, and also at the Key Laboratory of Image Processing and Intelligent Control of the Education Ministry of China, Huazhong University of Science and Technology. He has published over 200 international journal articles. His current research interests include the theory of functional differential equations and differential equations with discontinuous right-hand sides, and their applications to dynamics of neural networks, memristive systems, and control systems.

Dr. Zeng has been a member of the Editorial Boards of Neural Networks since 2012, Cognitive Computation since 2010, and Applied Soft Computing since 2013. He was an Associate Editor of the IEEE Transactions on Neural Networks from 2010 to 2011. He has been an Associate Editor for the IEEE Transactions on Cybernetics since 2014 and the IEEE Transactions on Fuzzy Systems since 2016.