



PDF Download
3786343.pdf
02 February 2026
Total Citations: 0
Total Downloads: 91

Latest updates: <https://dl.acm.org/doi/10.1145/3786343>

SURVEY

A Comprehensive Survey on Multi-modal Conversational Emotion Recognition with Deep Learning

Published: 28 January 2026
Online AM: 23 December 2025
Accepted: 09 September 2025
Revised: 25 June 2025
Received: 16 December 2024

[Citation in BibTeX format](#)

YUNTAO SHOU, Central South University of Forestry and Technology, Changsha, Hunan, China

TAO MENG, Central South University of Forestry and Technology, Changsha, Hunan, China

WEI AI, Central South University of Forestry and Technology, Changsha, Hunan, China

FANGZE FU, Central South University of Forestry and Technology, Changsha, Hunan, China

NAN YIN, Mohamed Bin Zayed University of Artificial Intelligence, Abu Dhabi, Abu Dhabi, United Arab Emirates

KEQIN LI, SUNY New Paltz, New Paltz, NY, United States

Open Access Support provided by:

Central South University of Forestry and Technology

Mohamed Bin Zayed University of Artificial Intelligence

SUNY New Paltz

A Comprehensive Survey on Multi-modal Conversational Emotion Recognition with Deep Learning

YUNTAO SHOU, TAO MENG, WEI AI, and FANGZE FU, College of Computer and Mathematics, Central South University of Forestry and Technology, Changsha, China

NAN YIN, Department of Computer Science, Mohamed bin Zayed University of Artificial Intelligence, Masdar City, United Arab Emirates

KEQIN LI, Department of Computer Science, State University of New York at New Paltz, New Paltz, New York, USA

Multi-modal Conversation Emotion Recognition (MCER) aims to recognize and track the speaker's emotional state using text, speech, and visual information. Compared with traditional single-utterance multi-modal emotion recognition or single-modal conversation emotion recognition, MCER is more challenging. It requires modeling complex emotional interactions and learning consistent and complementary semantics across multiple modalities. Although many deep learning-based approaches have been proposed for MCER, there is still a lack of systematic reviews summarizing existing modeling methods. Therefore, a timely and comprehensive overview of MCER's recent advances in deep learning is of great significance. In this survey, we provide a comprehensive overview of MCER modeling methods and roughly divide MCER methods into four categories, i.e., context-free modeling, sequential context modeling, speaker-differentiated modeling, and speaker-relationship modeling. Unlike conventional taxonomies based on modality combinations or task-stage decomposition, our framework focuses on how models structurally capture conversational dynamics, speaker roles, and emotional dependencies. In addition, we further discuss MCER's publicly available popular datasets, multi-modal feature extraction methods, application areas, existing challenges, and future development directions. We hope this review provides valuable insights into the current state of MCER research and inspires the development of more effective models.

CCS Concepts: • **General and reference** → **Surveys and overviews**; • **Human-centered computing** → *Natural language interfaces*;

This work is supported by the National Natural Science Foundation of China (Grant Nos. 62372478, 62472165), the YueLuShan Center Industrial Innovation (Grant No. 2026YCH0126), the Research Foundation of Education Bureau of Hunan Province of China (Grant No. 22B0275), the Hunan Provincial Natural Science Foundation General Project (Grant No. 2025JJ50380), and the Hunan Provincial Natural Science Foundation Youth Project (Grant No. 2025JJ60420).

Authors' Contact Information: Yuntao Shou, College of Computer and Mathematics, Central South University of Forestry and Technology, Changsha, China; e-mail: shouyuntao@stu.xjtu.edu.cn; Tao Meng (corresponding author), College of Computer and Mathematics, Central South University of Forestry and Technology, Changsha, China; e-mail: mengtao@hnu.edu.cn; Wei Ai, College of Computer and Mathematics, Central South University of Forestry and Technology, Changsha, China; e-mail: aiwei@hnu.edu.cn; Fangze Fu, College of Computer and Mathematics, Central South University of Forestry and Technology, Changsha, China; e-mail: fufangze@csuft.edu.cn; Nan Yin, Department of Computer Science, Mohamed bin Zayed University of Artificial Intelligence, Masdar City, United Arab Emirates; e-mail: nan.yin@mbzuai.ac.ae; Keqin Li, Department of Computer Science, State University of New York at New Paltz, New Paltz, New York, USA; e-mail: lik@newpaltz.edu.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](https://www.acm.org/permissions).

© 2026 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM 1558-2868/2026/1-ART47

<https://doi.org/10.1145/3786343>

Additional Key Words and Phrases: Multi-modal conversational emotion recognition, Deep Learning, Multi-modal datasets, Multi-modal feature fusion, Multi-modal feature extraction

ACM Reference format:

Yuntao Shou, Tao Meng, Wei Ai, Fangze Fu, Nan Yin, and Keqin Li. 2026. A Comprehensive Survey on Multi-modal Conversational Emotion Recognition with Deep Learning. *ACM Trans. Inf. Syst.* 44, 2, Article 47 (January 2026), 48 pages.

<https://doi.org/10.1145/3786343>

1 Introduction

With the development of the mobile Internet, social media has become the main platform for people to communicate with each other [80]. Users can fully express their emotions through multi-modal data such as text, voice, image, and video. Building a **Multi-modal Conversational Emotion Recognition (MCER)** model using multi-modal data is of vital practical significance for understanding users' true emotional intentions [22]. Therefore, researchers have been trying to give machines the ability to understand emotions in recent years [56, 157, 163].

Before the emergence of MCER, early methods [45, 68, 116, 125, 161] primarily relied on single-modal data, such as text or speech. These approaches mainly focused on modeling contextual dependencies within the same modality and leveraging the semantic content of words or audio signals to recognize emotions [77, 104, 135, 141]. However, relying solely on textual information may be insufficient for accurately interpreting a speaker's emotional state, as speakers often express their opinions in a reserved or implicit manner [21, 87, 164]. For example, a speaker may be veiled in expressing his anger, which may result in a more neutral utterance. In response to the above problems, MCER technology was proposed to solve the problem of insufficient expression of text semantic information [87, 102, 103, 141]. As shown in Figure 1, MCER aims to extract semantic information complementary within and between modalities and identify the emotions expressed by speakers in text, audio, and video. One major advantage of MCER is its ability to enhance emotion understanding when the emotional polarity conveyed by text alone is insufficient [60]. In such cases, the model can leverage visual cues (e.g., facial expressions) and acoustic signals (e.g., tone of voice) to supplement and enrich the emotional representation [121, 137–140]. As illustrated in Figure 2(a), the text modality alone predicts the speaker's emotion to be "Neutral," whereas the audio and visual modalities correctly predict the speaker's emotion to be "Sad," highlighting the limitations of relying solely on text in emotional context inference. Furthermore, to validate the effectiveness of multi-modal fusion, we provide LR-GCN latent space visualizations for both unimodal and multi-modal settings in Figure 2(b) and (c). It is evident that the multi-modal feature space yields better inter-class separation,

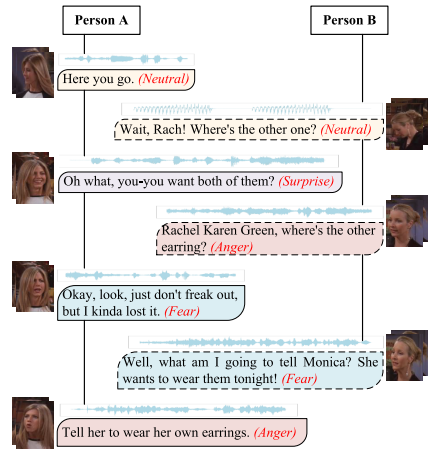


Fig. 1. An example of an MCER dataset which contains three modal features: video, audio, and text. The task of MCER is to identify the emotion label of each speaker at the current moment based on the utterance content (e.g., neutral, angry, and surprised).

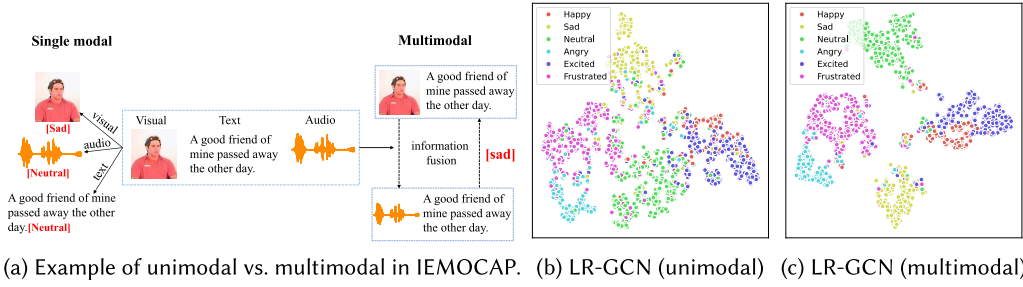


Fig. 2. Illustration of the advantage of multi-modal fusion in emotion recognition. (a) Example from the IEMOCAP dataset showing that the textual modality alone may fail to capture emotional intent (“Neutral”), while audio and visual modalities correctly identify the emotion as “Sad.” (b) Latent space visualization of GS-MCC with unimodal (text-only) input shows overlapping clusters and poor separation between emotion classes. (c) The same visualization under multi-modal fusion shows significantly improved class separability, demonstrating the effectiveness of incorporating audio-visual information.

especially among subtle emotions such as “Neutral,” “Frustrated,” and “Sad,” demonstrating superior discriminative capability.

However, unlike traditional single-utterance multi-modal emotion recognition or single-modal conversation emotion recognition, MCER is a more challenging issue. It requires consideration of factors such as multi-modal context, dialogue scenarios, the speaker’s emotional inertia, and the interlocutor’s stimulation [9, 152]. Powerful deep learning technology [158] enables MCER to recognize emotion by fusing semantic features with complex emotional interactions. Feature fusion in MCER mainly considers intra-modal contextual semantic information fusion and inter-modal complementary semantic information fusion [38]. On the one hand, intra-modal contextual semantic information fusion refers to extracting the temporal and spatial dependencies of speaker feature representations in each modality. On the other hand, complementary semantic information fusion between modalities refers to using the interactive information between different modalities to enhance the emotional understanding ability of the model. MCER synergistically improves the effect of emotion recognition by fusing the characteristics of various modal data, which has important theoretical significance for processing and understanding multi-modal data [31, 131, 157].

Despite the growing number of researchers focusing on new models and methods for MCER [9, 19, 65], there is still a lack of understanding regarding the theoretical and methodological classification of MCER, particularly those based on deep learning. To the best of our knowledge, this survey is the first comprehensive survey focusing on deep learning in MCER. Existing reviews mainly focus on multi-modal emotion recognition with modal combination [155] or multi-modal emotion recognition with task stage decomposition (i.e., feature extraction, feature fusion, and classification) [164], without fully considering conversational dynamics, speaker roles, and emotional dependencies.

Different from previous taxonomies, we propose a novel classification framework that emphasizes how methods characterize and model conversational dynamics. Specifically, we categorize MCER methods into four distinct types: context-free modeling, sequential context modeling, speaker-differentiated modeling, and speaker-relationship modeling, as illustrated in Figure 3.

Based on the above framework, this survey systematically reviews the key research efforts in MCER. First, we introduce several widely used and publicly available datasets, along with commonly adopted feature extraction methods across modalities. Next, we detail the proposed

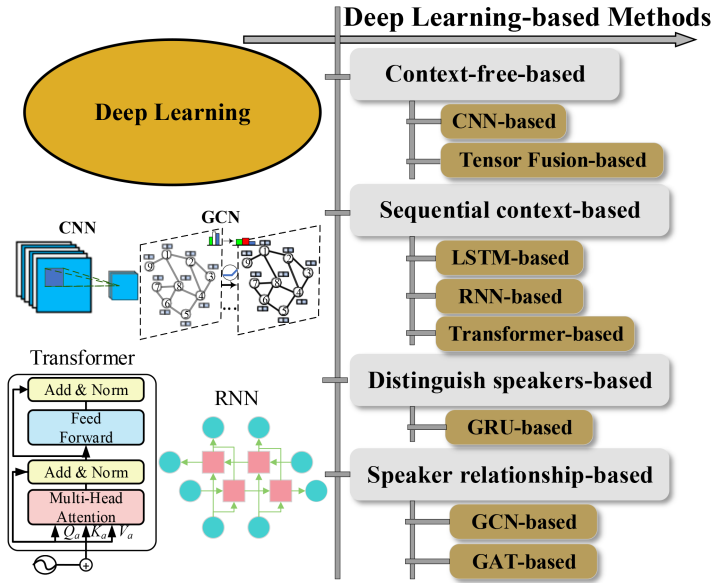


Fig. 3. A taxonomy of modeling approaches for MCER in conversation. We categorize existing MCER methods into four categories, i.e., context-free modeling, sequential context modeling, distinguishing speaker modeling, and speaker-relationship modeling.

modeling taxonomy and comprehensively analyze representative methods within each category. We then discuss evaluation metrics frequently used in MCER experiments. Following that, we examine real-world applications and key challenges faced in this field. Finally, we outline promising directions for future research.

The contributions made in this article are summarized as follows:

- *New Taxonomy*: We provide a new taxonomy for MCER. Specifically, we classify existing MCER methods into four groups: context-free modeling, sequential context modeling, distinguishing speaker modeling, and speaker-relationship modeling.
- *Comprehensive Review*: This article provides the most comprehensive review of deep learning and machine learning algorithms for MCER. For each modeling approach, we provide representative models and make corresponding comparisons.
- *Abundant Resources*: We collect relevant resources about MCER, including state-of-the-art models and publicly available datasets. This article can serve as a practical guide for learning and developing different emotion recognition algorithms.
- *Future Directions*: We analyzed the limitations of existing MCER methods and proposed possible future research directions in many aspects, such as the collaborative generation of multi-modal data, the deep fusion of multi-modal features, and the unbiased learning of multi-modal emotions.

The article is organized as follows: Section 2 summarizes the publicly available and popular datasets in the field of MCER. Section 3 illustrates the background, definitions, and commonly used feature extraction techniques for MCER. Section 4 broadly divides MCER methods into four categories and analyzes their advantages and disadvantages. Section 5 summarizes some commonly used evaluation metrics for MCER tasks. Section 6 gives the performance of different algorithms on the IEMOCAP and MELD datasets. Section 7 discusses the real-life applications of MCER. Section 9

Table 1. Publicly Available Benchmark Datasets in MCER

Datasets	Year	Modality	Available at
IEMOCAP [4]	2008	Text, Video, Audio	https://sail.usc.edu/iemocap/
MELD [86]	2019	Text, Video, Audio	https://web.eecs.umich.edu/~mihalcea/downloads/MELD.Raw.tar.gz
DailyDialog [57]	2017	Text	https://huggingface.co/datasets/daily_dialog
EmoryNLP [149]	2017	Text	https://github.com/emorynlp/character-mining
SEMAINE [76]	2012	Text, Video, Audio	https://semaine-db.eu/
EmotionLines [32]	2018	Text	https://doraemon.iis.sinica.edu.tw/emotionlines/index.html
EmoContext [7]	2019	Text	https://www.humanizing-ai.com/emocontext.html

Table 2. Distribution of Seven Conversational Emotion Recognition Datasets on Different Emotion Labels

Labels	IEMOCAP	MELD	EmoContext	EmotionLines	EmoryNLP	DailyDialog	SEMAINE
Neutral	1,708	6,436	-	6,530	15,104	855,72	-
Happiness/Joy	648	2,308	4,669	1,710	11,020	12,885	93
Surprise/Powerful	-	1,636	-	1,658	4,252	1,823	-
Sadness	1,084	1,002	5,838	498	3,376	1,150	58
Anger/Mad	1,103	1,607	5,954	772	5,328	1,022	41
Disgust	-	361	-	338	-	354	7
Fear/Scared	-	358	-	255	6,584	74	3
Frustrated	1,849	-	-	-	-	-	-
Excited	1,041	-	-	-	-	-	-
Other	-	-	21,960	-	4,760	-	197

illustrates the problems of existing research, and Section 10 gives directions for future research. Finally, we conclude the work of this article.

2 Popular Benchmark Datasets

Table 1 presents seven publicly available emotion recognition benchmark datasets. We counted the release time, modality, and open source URL for each dataset. As shown in Table 2, we also counted the distribution of the dataset on different emotional labels, and the data showed a long-tail distribution.

2.1 IEMOCAP

The interactive emotional dyadic motion capture database (IEMOCAP) dataset [4] was released in 2008 and contains 12.46 hours of conversations. The IEMOCAP dataset contains three modal features, i.e., video, audio, and text, and it is the first multi-modal dataset for MCER. In the IEMOCAP dataset, 10 theater actors express specific emotion categories (i.e., sad, neutral, frustrated, anger, happy, and excited) through binary dialogue. To ensure the consistency and accuracy of annotation, each sentence is annotated by multiple experts.

2.2 MELD

The multi-modal emotionLines (MELD) dataset [86] is from the classic TV series Friends, which contains text, video, and audio data. The MELD dataset contains a total of 13,709 video clips, and each sentence is labeled as a specific emotion (i.e., anger, neutral, fear, disgust, surprise, joy, and disgust). In addition, the MELD dataset is also annotated by neutral, negative, and positive three-category emotion. To ensure the consistency and accuracy of annotation, each sentence is annotated by multiple experts.

2.3 DailyDialog

The DailyDialog dataset [57] is a multi-turn dialogue dataset about daily chat scenarios, which only contains text modalities. The DailyDialog dataset contains 13,000 dialogues and labels each sentence with intention (i.e., inform, commissive, directives, and questions) and emotion (surprise, sadness, fear, happiness, disgust, and anger). Each sentence is annotated jointly by three experts.

2.4 EmoryNLP

EmoryNLP [149] is a unimodal dataset, containing only text modalities. The EmoryNLP dataset contains 12,606 utterances, and each utterance is annotated with seven emotions: peaceful, scared, crazy, powerful, sad, happy, and neutral. EmoryNLP dataset is divided into training set, testing set, and validation set.

2.5 SEMAINE

The sustained emotionally colored machine–human interaction (SEMAINE) [76] is a multi-modal conversation dataset, which contains four binary conversations between robots and humans. The SEMAINE dataset has 95 dialogues with a total of 5,798 sentences. Four emotional dimensions are marked: Valence, Arousal, Expectancy, and Power. Valence, Arousal, and Expectancy are continuous values in the range $[-1, 1]$, and the size of the SEMAINE dataset is small.

2.6 EmotionLines

The EmotionLines dataset [32] comes from binary conversations between Friends and Facebook, and it only contains text data. The EmotionLines dataset contains 1,000 dialogues with a total of 29,245 sentences. Seven categories of emotions are marked: neutral, fear, surprise, sadness, anger, happiness, and disgust. The EmotionLines dataset is rarely used in conversational emotion recognition.

2.7 EmoContext

The emotion contextual detection (EmoContext) dataset [7] only contains text data. It has a total of 38,421 dialogues and a total of 115,263 sentences. Three types of emotions are marked: happiness, sadness, and anger. Although the EmoContext data is large, it is rarely used in conversational emotion recognition because it only contains text data.

2.8 CH-SIMS

The Chinese single-label multi-modal sentiment analysis (CH-SIMS) [143] is a benchmark dataset designed for Chinese multi-modal sentiment analysis tasks. This dataset is constructed from real Chinese video conversations, covering three modalities: text, audio, and vision, and has a good foundation for multi-modal fusion and alignment. CH-SIMS contains a total of about 10,000 Chinese sentence-level samples, all of which are accompanied by manually annotated continuous sentiment intensity labels (ranging from -1 to $+1$) and single-label classifications (positive, neutral, and negative). Compared with mainstream English multi-modal sentiment analysis datasets such as IEMOCAP and MELD, CH-SIMS is more in line with the characteristics of Chinese language expression, especially when faced with semantic ambiguity (e.g., irony and sarcasm), there may be significant inconsistencies between modalities.

2.9 MuSE

Multi-modal Sentiment dataset (MuSE) [105] is a multi-lingual and multi-modal emotion recognition dataset, which aims to study the performance of multi-modal emotion modeling in different

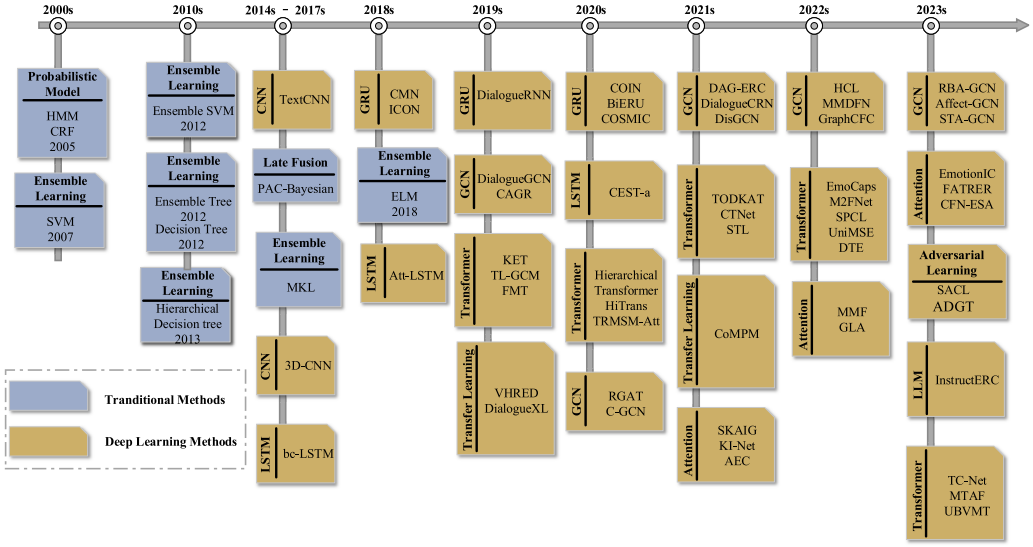


Fig. 4. Timeline of MCER algorithms.

languages and cross-cultural contexts. The dataset contains natural speech videos from English and Spanish, with about 2,800 samples, covering three modalities: text, audio, and vision. Each video is recorded by real participants expressing freely around a specific topic, and provides continuous emotion labels, including valence, arousal, and dominance. The labels are derived from the fusion of self-reports and third-party manual evaluations to enhance the objectivity and consistency of annotations. At the modality level, MuSE provides high-quality speech features, expressions, and body posture information, and is equipped with precisely aligned text transcription data, making it suitable for research tasks such as multi-modal fusion, modal alignment analysis, and modal inconsistency modeling. In addition, since the dataset has both multi-lingual and multi-modal characteristics, it provides an important experimental platform for cross-lingual emotion transfer learning, multi-modal collaborative modeling, and emotion recognition in low-resource languages.

3 Background, Definition, and Feature Extraction

3.1 Background

As shown in Figure 4, we counted MCER algorithms from 2000 to 2023. As can be seen from the figure, before 2018, traditional machine learning algorithms were mainly used, and then deep learning algorithms gradually became the main ones. Next, we briefly return to the main development history of the MCER algorithm.

3.1.1 Brief History of Conversational Emotion Recognition. The emotion recognition method based on the dictionary is the earliest used for emotion recognition [25], which motivated early work on Naive Bayes method [16]. With the widespread application of machine learning algorithms in classification tasks, representative algorithms such as the **Support Vector Machine (SVM)** [37, 94] and binary decision tree [10, 51, 67] have also gained prominence in emotion recognition. The above-mentioned method determines the category of emotion by learning the polarity and occurrence frequency of emotional words in the text, which is difficult to extract the semantic information and context information.

Encouraged by the success of **Convolutional Neural Networks (CNNs)** in computer vision, CNNs began to be migrated to text classification tasks and received extensive research attention [43, 46, 48]. In 2017, Poria et al. [84] used **Long Short-term Memory (LSTM)** for the first time to resolve dependencies between contexts. Since then, improvements, extensions, and applications of LSTMs and **Gated Recurrent Units (GRUs)** have increased [26, 27, 75, 90]. Until recently, many **Graph Neural Networks (GNNs)**-based methods (e.g., [21, 40, 55, 99, 148]) emerged. Apart from CNNs, **Recurrent Neural Networks (RNNs)**, and GNNs, many alternative Transformer-based methods (e.g., [54, 161, 163]) have been developed in the past decades. We detail the categories to which these algorithms belong in Section 4.

3.1.2 MCER versus Traditional Machine Learning. MCER methods based on traditional machine learning [10, 37, 51, 62, 67, 94] are closely related to hand-extracted features, which have attracted increasing attention from the data mining and emotion recognition communities. These methods aim to learn the feature embeddings of raw data for subsequent downstream tasks such as classification and clustering. The classic conversational emotion recognition method based on machine learning is to use SVM to map emotional features to a hyperplane and classify them [3, 94]. However, these methods require a large amount of high-quality labeled data.

3.1.3 Conversational Emotion Recognition versus CNN. The CNN-based emotion recognition methods [46, 48, 69] are the first deep learning method to solve the emotion classification problem historically [45]. These CNN-based methods employ convolutional filters to extract semantic features of text so that the model can use supervised learning to understand the meaning of text. Similar to machine learning algorithms, CNN can also map emotional features into vector space through mapping functions. The difference is that this mapping function is learned in an end-to-end manner. Since the convolution kernel extracts local receptive field information, CNN cannot contain contextual semantic information.

3.1.4 MCER versus RNN. The RNN-based emotion recognition methods [46, 70, 75, 112] are developed on the basis of CNN, but they believe that contexts should be mutually influential and interdependent [70, 112]. These RNN-based methods usually use LSTM or GRU (to avoid gradient disappearance or gradient explosion) to extract semantic features including context. Similar to CNN, RNN can also map emotional features into vector space through mapping functions in an end-to-end manner.

3.1.5 MCER versus Transformer. Similar to the RNN-based emotion recognition method, the Transformer-based emotion recognition methods [38, 61, 89, 114] also extract semantic information including context, and completes subsequent emotion classification based on this [61]. However, unlike RNN, Transformer's sequential context modeling ability is better than RNN. Therefore, the accuracy of the Transformer-based emotion recognition methods are significantly better than RNNs.

3.1.6 MCER versus GNN. The GNN-based emotion recognition methods [40, 55, 63, 99, 157] inherit the idea of the RNN method, i.e., the contexts should interact and depend on each other [21]. On the basis of RNN, GNNs believe that there is also a relationship of mutual influence between speakers. Therefore, GNNs model the dialogue relationship between speakers through the inherent properties of the graph structure.

3.2 Definitions and Preliminaries

The symbols used in this article are listed in Table 3. Now, we define the sets needed to understand this article. In particular, we use uppercase letters for matrices and lowercase letters for vectors.

Table 3. Some Symbols Commonly Used in the Article

Notations	Descriptions
$ \cdot $	The length of the set.
\odot	Element-wise product.
\mathcal{G}	A graph.
\mathcal{V}	A set of nodes in a graph.
v	A node $v \in V$.
\mathcal{E}	A set of edges in a graph.
e_{ij}	An edge $e_{ij} \in E$.
$N(v)$	The neighbors of a node v .
S	A speaker.
u	An utterance.
K	The context window size.
M	The number of the speakers.
L	The number of utterances in a dialogue.
U	The set of contextual utterance.
\mathcal{R}	The type of edge.
\mathcal{W}	Learnable parameters.
\mathbf{A}	The adjacency matrix of a graph.
m	The node properties of the graph.
$x^t \in \mathbb{R}^d$	d -dimensional text feature vectors.
$x^a \in \mathbb{R}^k$	k -dimensional audio feature vectors.
$x^v \in \mathbb{R}^h$	h -dimensional video feature vectors.
x	Concatenated video, audio, and text feature vectors.

Table 4. We Assume That There Are Three Speakers in a Dialogue, and the Window Size K of the Dialogue is Set to 6

Speaker	Utterances	Description
C_a, C_b, C_c	$u_1^a, u_3^a, u_2^b, u_5^b, u_4^c, u_6^c$	Contextual utterances
S_b	u_7^b	Predicted utterance

The dialogue process is as shown above.

Definition 1 (Utterances Context). The MCER task aims to recognize the emotional changes (e.g., happiness and sadness) of speakers $\{S_1, S_2, \dots, S_M\}$ at the current moment t in a dialogue. L represents the number of utterances in a dialogue, U represents a set of contextual utterances, and $U = \{u_1, u_2, \dots, u_L\}$.

The MCER task aims to correctly classify each utterance by incorporating contextual information. At the current moment t , the model needs to infer the speaker's emotion based on the context information $\{u_1, u_2, \dots, u_{t-1}\}$. We assume that the context window size is set to K . The set of contextual utterances is defined as follows:

$$C_\lambda = \{u_i \mid i \in [t - K, t - 1], u_i \in U_\lambda, |C_\lambda| \leq K\}. \quad (1)$$

When the context window size is 6, the speaker's contextual utterances and predicted utterances are shown in Table 4.

Definition 2 (Dialogue Graph). A dialogue graph is represented as $\mathcal{G} = \{\mathcal{V}, \mathcal{E}, \mathcal{R}, \mathcal{W}\}$, where \mathcal{V} is a set of nodes in the graph, \mathcal{E} is a set of edges, $v_i \in \mathcal{V}$ represents the i th node, $e_{ij} = (v_i, v_j) \in \mathcal{E}$ represents a directed edge from v_i to v_j , and the relationship $r_{ij} \in \mathcal{E}$ represents that there is a

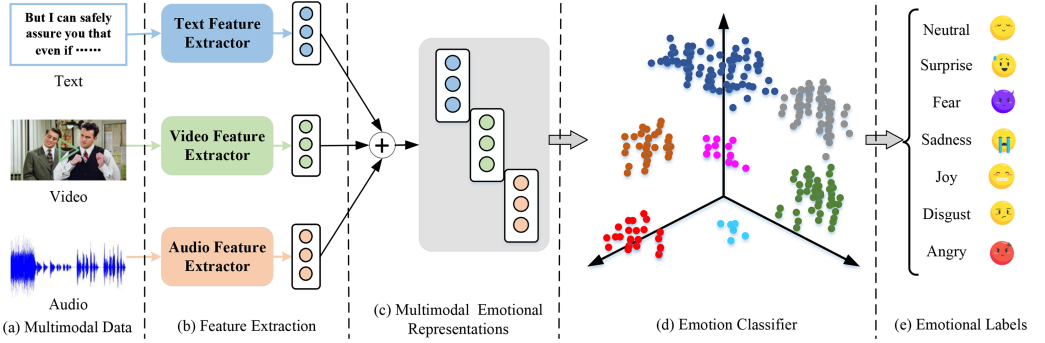


Fig. 5. The proposed MCER methods mainly include multi-modal feature extraction, multi-modal emotion representation, and emotion classifier.

dialog relationship between nodes v_i and v_j . The neighbor nodes of node v are represented as $N(v) = \{u \in \mathcal{V} | (v, u) \in \mathcal{E}\}$. $\mathbf{A} \in \mathbb{R}^{n \times n}$ means the adjacency matrix with $\mathbf{A}_{ij} = 1$ if $e_{ij} \in \mathcal{E}$, otherwise $\mathbf{A}_{ij} = 0$. $\mathbf{X} \in \mathbb{R}^{m \times m}$ represents the node properties of the graph. For the MCER task based on GCN, the speaker's utterance information is regarded as the node of the graph, and the dialogue relationship information between speakers is regarded as the edge of the graph.

Definition 3 (Problem Definition). For a given multi-modal utterance sequence U , the MCER task requires using the utterance context information to determine a deep neural network $F(u_i)$ so that the output emotion label \hat{y}_i is as close as possible to the real emotion label y_i , $i \in \{1, \dots, L\}$. Deep neural networks can solve the optimal parameters by minimizing loss, and its loss is defined as:

$$\min_F \frac{1}{L} \sum_{i=1}^L \mathcal{L}(\hat{y}_i = F(u_i), y_i), \quad (2)$$

where L represents the number of utterances in the dialogue, \mathcal{L} is an indicator function.

From the development history and related preliminary definitions of MCER, it can be seen that the process of MCER mainly includes three aspects: multi-modal feature extraction, multi-modal feature fusion representation, and emotion classification. The overall process is shown in Figure 5, and we will provide a comprehensive overview of these three aspects below.

3.3 Multi-modal Feature Extraction

Multi-modal feature extraction (e.g., text, video, and audio) is one of the important techniques for emotion analysis. In this section, we introduce the process of using feature extraction methods to perform data preprocessing on text, video, and audio, and list some commonly used feature extraction methods. As shown in Table 5, we count the multi-modal feature extraction techniques used by many deep learning methods.

3.3.1 Text Feature Extraction. With the rapid development of deep learning, word embedding has also been widely used to extract text features. Before embedding, raw text data typically undergo several preprocessing steps to improve the quality and consistency of the input. These steps often include tokenization, lowercasing, punctuation, and stop-word removal, and in some cases, lemmatization or stemming. Some studies also perform syntactic or dependency parsing to capture the structural relationships between words, which can further enhance the semantic representation in downstream tasks. Word embedding technology then uses a shallow neural network to learn the semantic information of words and uses Euclidean distance to measure

Table 5. Feature Extraction Methods for Text, Video, and Audio Features Used by Different Emotion Recognition Techniques

Methods	Text	Video	Audio	Methods	Text	Video	Audio
THMM [78]	Polarized words	OKAO Vision	OpenEAR	CMN [27]	TEXT-CNN	3D-CNN	openSMILE
SVM [81]	Bag-of-words	CERT	OpenEAR	Att-BiLSTM [84]	TEXT-CNN	3D-CNN	openSMILE
MKL [83]	Word2vec	CLM-Z	openSMILE	ICON [26]	TEXT-CNN	3D-CNN	openSMILE
SAL-CNN [118]	Word2vec	CLM-Z	COVAREP	DialogueRNN [75]	TEXT-CNN	3D-CNN	openSMILE
TFN [145]	GLOVE	Facet	COVAREP	DialogueGCN [21]	TEXT-CNN	3D-CNN	openSMILE
LMF [66]	GLOVE	Facet	COVAREP	COIN [151]	TEXT-CNN	3D-CNN	openSMILE
HEFN [72]	GLOVE	Facet	COVAREP	CESTa [123]	TEXT-CNN	3D-CNN	openSMILE
LMFN [73]	GLOVE	Facet	COVAREP	EmoCaps [58]	BERT	3D-CNN	openSMILE
GME-LSTM [73]	GLOVE	Facet	COVAREP	MM-DFN [34]	TEXT-CNN	3D-CNN	openSMILE
MARN [147]	GLOVE	Facet	COVAREP	M2FNet [9]	RoBERTa	Mel Spectrograms	MTCNN
MFN [146]	GLOVE	Facet	COVAREP	GraphCFC [55]	TEXT-CNN	openSMILE	3D-CNN
RAVEN [122]	GLOVE	Facet	COVAREP	UniMSE [36]	T5	openSMILE	3D-CNN
SWRM [126]	BERT	Facet	COVAREP	EmotionC [142]	TEXT-CNN	openSMILE	3D-CNN
MCTN [82]	GLOVE	Facet	COVAREP	SACL-LSTM [33]	RoBERTa	openSMILE	3D-CNN
MuT [114]	GLOVE	Facet	COVAREP	HyCon [74]	BERT	Facet	COVAREP
MAG [89]	BERT	Facet	COVAREP	HGraph-CL [63]	BERT	Facet	COVAREP
ICDN [154]	GLOVE	Facet	COVAREP	bc-LSTM [84]	TEXT-CNN	3D-CNN	openSMILE
AMOA [59]	BERT	OpenFace 2.0	openSMILE	MMMU-BA [19]	GLOVE	Facet	COVAREP
ICCN [109]	BERT	Facet	COVAREP	MISA [29]	BERT	Facet	COVAREP

the similarity between words. Unlike traditional one-hot encoding methods, word embedding technology maps high-dimensional sparse feature vectors to low-dimensional dense vectors. This reduces computational resource usage and addresses the issue of one-hot encoding's inability to capture the semantic gap between words. A commonly used word embedding method is **Word to Vector (Word2vec)** [8], which contains two different forms: **Continuous Bag of Words (CBOW)** [22] and Skip-gram [13]. CBOW predicts the central word based on the surrounding words, and Skip-gram predicts the surrounding words based on the central word. Although the above methods can capture the semantic similarity between words, they require large datasets for training.

Some recent studies use TextCNN [45] and global vectors for word representation (GLOVE) [17] to extract text features. In addition, large-scale predictive pre-training models such as **Bidirectional Encoder Representations from Transformers (BERT)** [71] and **Robustly Optimized BERT Approach (RoBERTa)** [44] are often used to capture contextual information through attention mechanisms.

3.3.2 Video Feature Extraction. Visual feature extraction is mainly used to extract information such as facial expressions and gestures that contain the speaker's emotions from the video. In recent years, deep neural networks have been able to extract deep features from images in an end-to-end learning manner, avoiding the tedious manual feature extraction. For example, Tran et al. [113] proposed an effective and efficient 3D-CNN to process video frames containing spatio-temporal features.

In most modern multi-modal emotion recognition systems, visual preprocessing begins with frame sampling, where video is typically downsampled to a fixed frame rate (e.g., 25 or 30 frames per second) to reduce redundancy and maintain temporal resolution [42, 120]. Each frame is then resized (commonly to 224×224 pixels) and normalized using per-channel mean subtraction and standard deviation scaling (e.g., ImageNet normalization settings) to ensure consistency across inputs. For facial region detection and alignment, face detectors (e.g., dlib or MTCNN) are employed to locate the face in each frame, and affine transformations are applied to align facial landmarks to a canonical pose, improving robustness to head movements and variations in scale or rotation.

Various open source toolkits are used to extract deep visual features from the aligned facial regions. For instance, OpenFace 2.0 [2] detects 68 facial landmarks, estimates head pose, facial action units, gaze direction, and eye-blink frequency, which are strongly correlated with affective states. Facet further extracts features such as facial muscle activation, histograms of oriented gradients, emotion intensity scores, and micro-expressions. These features are typically calculated on a frame-by-frame basis and then aggregated over time using statistical functions (e.g., mean, standard deviation, and max) or temporal models such as LSTMs. OKAO Vision is a commercial toolkit capable of estimating smile intensity (ranging from 0 to 100) and eye gaze orientation, while CERT adaptively captures head pose dynamics and subtle expressions over short temporal segments. These visual feature extractors are configured with default or task-optimized parameters depending on the specific dataset conditions (e.g., lighting, camera angle, and facial occlusion).

3.3.3 Audio Feature Extraction. Deep learning has increasingly attracted attention in the field of audio feature extraction, enabling automatic modeling of acoustic patterns associated with human emotion. For instance, LSTM networks [127] have been widely applied to model temporal dependencies in speech, while Poria et al. [84] used CNNs to extract local patterns from audio signals, followed by feeding the extracted features into emotion classification models.

In recent years, an increasing number of emotion recognition models [75, 116, 161] have adopted open source toolkits for systematic and standardized audio feature extraction. Commonly used tools include COVAREP [14], openSMILE [47], LibROSA [106], and OpenEAR [97]. These toolkits provide frame-level acoustic descriptors based on well-established speech analysis techniques. Specifically, OpenEAR is capable of extracting a comprehensive set of low-level descriptors, such as prosodic (e.g., pitch and energy), spectral, and cepstral features, and applies Z-score normalization to ensure feature comparability across samples. The openSMILE toolkit is often configured with the INTERSPEECH 2010 or eGeMAPS feature set, extracting **Mel-frequency Cepstral Coefficients (MFCC)**, pitch, zero-crossing rate, voice intensity, and other prosody-based features. Audio signals are typically resampled to 16 kHz mono-channel, then segmented using sliding windows (e.g., 25 ms with 10 ms stride) to generate frame-wise features. LibROSA, a widely used Python-based audio analysis library, is used to extract 33 frame-level acoustic features, including 20-dimensional MFCCs, chroma features, and Constant-Q Transform, which captures tonal energy variations. Similarly, COVAREP provides features such as 12-dimensional MFCC, Maxima Dispersion Quotient, Normalized Amplitude Quotient, and Liljencrants–Fant glottal model parameters, which are valuable for capturing subtle vocal tract changes related to emotional state.

4 Taxonomy of MCER Algorithms

In this section, we present a taxonomy of MCER modeling approaches. We categorize existing work into context-free modeling, sequential context modeling, distinguishing speaker modeling, and speaker relation modeling. We briefly introduce each method in the following.

4.1 Context-free Modeling

These are mostly pioneering works on conversational emotion recognition. Context-free modeling methods aim to learn a feature representation for each sentence, which does not exploit the contextual information of the sentence [68, 98, 152]. For example, some traditional machine learning methods (e.g., SVMs [62, 94] and decision trees [10, 51]) is used to extracts the feature representation of each sentence, and utilize the extracted sentence features to complete emotion classification. The above process assumes that each sentence is independent and does not influence each other. We introduce several common context-free modeling methods based on feature fusion below.

4.1.1 Add. The early fusion method based on addition operation obtains the final emotional feature representation by weighted summation of different modality features [11]. This fusion method is simple to operate and requires only a small amount of calculation. However, its shortcomings are also obvious. It cannot model the context information in a fine-grained manner, and the information that can be utilized is limited. The formula for implementing the context-free modeling method using the additive approach is defined as follows:

$$h_e = x^t + x^a + a^v, \quad (3)$$

where h_e represents the fused emotional vectors, x^t, x^a, x^v represent the text, audio, and video vectors, respectively. The Add method is essentially the direct accumulation of different modal information in the same semantic space, which retains the weighted contribution of each modality in the corresponding feature dimension. Through summation, the model can automatically adjust the numerical expression of each modal feature during the learning process, making the information complementary and improving the overall representation ability.

The Add method is relatively simple and easy to understand and implement. By directly merging the features of different modalities, the feature information of different modalities can be fully utilized. For modalities with strong complementarity, the Add method can effectively capture the correlation between them. However, the features of different modalities may have different scales or importances. The Add method does not consider the difference in importance between the features of different modalities, which may lead to information loss or imbalance. Therefore, the Add method can be considered in scenarios where there is strong complementarity between modalities, limited computing resources, or requirements for model complexity.

4.1.2 Concatenation. The early fusion method based on the concatenation operation obtains the final emotion feature representation by concatenating and merging different modal features [5]. Although this fusion method does not introduce additional calculations, it leads to very high dimensionality of the data, which makes calculations difficult. Furthermore, it also fails to capture intra-modal and inter-modal semantic information that is complementary:

$$h_e = \text{Concat}([x^t, x^a, a^v]), \quad (4)$$

where $\text{Concat}(\cdot)$ represents the concatenation operation. The concatenation method does not perform any information compression or mapping on the features of each modality, but directly concatenate the original information by dimension, theoretically retaining the complete expression of each modality.

The concatenate method only requires simple vector concatenation operations, without complex parameter learning or model design. The concatenate method avoids possible information loss in early fusion. However, the feature dimension after concatenation is the sum of each modality, which may lead to sparsity and overfitting, especially on small-scale datasets. In addition, high-dimensional features may increase the computational complexity of subsequent models and simple concatenation cannot explicitly model the nonlinear relationship between modalities. Therefore, when the information quality of each modality is high and complementary to each other, concatenation can effectively retain information.

4.1.3 SVM. SVM is a machine learning algorithm for classification and regression whose optimization goal is to find a hyperplane (a straight line in two-dimensional space, and a hyperplane in high-dimensional space) that separates samples of different classes. Based on the above research, Perez-Rosas et al. [81] concatenate multi-modal features as input vectors and use SVM to classify utterances for emotion. SVM works better for binary classification problems, but is less effective in multi-classification problems, and is only suitable for training small-scale datasets. The formula of

SVM is defined as follows:

$$f(x) = \text{sign} \left(\sum_{i=1}^N \alpha_i^* y_i \exp \left(-\frac{\|x - z\|^2}{2\sigma^2} \right) + b^* \right), \quad (5)$$

where $\text{sign}(x > 0) = 1$, $\text{sign}(x = 0) = 0$, $\text{sign}(x < 0) = -1$, α_i^* , b^* represent the learnable parameters, $\exp \left(-\frac{\|x - z\|^2}{2\sigma^2} \right)$ represents the kernel function, N is the number of samples. SVM uses a nonlinear kernel function to implicitly model the nonlinear interaction information between modalities.

Through the kernel function, SVM can find the best classification hyperplane in the high-dimensional feature space, effectively perform nonlinear classification, and is suitable for high-dimensional data. When there are fewer training samples, SVM has good generalization ability, but it is time-consuming for large-scale datasets.

4.1.4 Multiple Kernel Learning. After preprocessing the features of three different modalities, Poria et al. [83] constructed two different feature selectors to achieve feature dimensionality reduction. One of the feature selectors is based on circular correlated feature subset selection, and the other is based on principal component analysis. The above two feature selectors cannot only eliminate redundant information and noise information but also improve the running speed of the model. After feature selection and dimensionality reduction, the researchers spliced and merged the processed feature vectors and trained a classifier using a **Multi-kernel Learning (MKL)** algorithm [83]. Based on the previous research work, the authors further propose the convolutional recurrent MKL [85] model. Convolutional recurrent multi-kernel learning uses a convolutional RNN for emotion detection, which can extract contextual information. The formula of MKL is defined as:

$$\begin{aligned} \max_{\alpha, \beta} & \left[\sum_{i=1}^N \alpha_i - \sum_{i,j=1}^N \alpha_i \alpha_j y_i y_j K_{mkl}(x_i, x_j) \right] \\ & \sum_{i=1}^N \alpha_i y_i = 0 \\ & 0 \leq \alpha_i \leq C \\ & K_{mkl} = \sum_k^M \beta_k K_k > 0, \end{aligned} \quad (6)$$

where y_i is the true label, α, β are the learnable parameters, M is the feature dimension. MKL achieves flexible fusion of multi-modal information at the kernel space level through multi-core combination and weight optimization.

The multiple kernel learning method can combine multiple different kernel functions according to different data characteristics, which helps to process complex data structures. It is suitable for processing a variety of heterogeneous data or multi-modal data and can combine information from different modalities. However, calculating the combination of multiple kernel functions may result in high computational costs.

4.1.5 Select-additive Learning (SAL) CNN. CNN is a classic neural network in visual tasks and cannot be directly used for emotion recognition. To solve this problem, Kim et al. [45] proposed the TextCNN model, and its overall process is shown in Figure 6. To perform multi-modal emotion recognition, Wang et al. [118] proposed the SAL-CNN model, which first uses multi-modal data to fully train the CNN, and then uses SAL to improve its versatility and prevent the model from overfitting during training. The SAL method consists of two phases (i.e., selection and addition). In the selection phase, SAL preserves important features and removes noisy information from

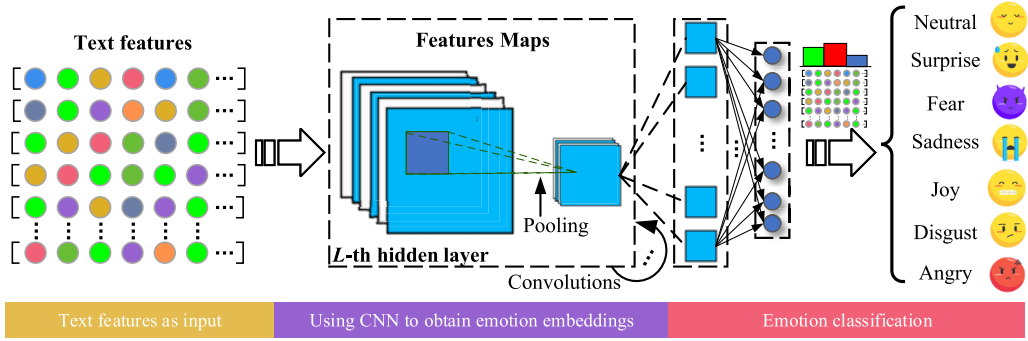


Fig. 6. The flowchart of the proposed TextCNN method. Specifically, given text features, the TextCNN employs convolution filters of different sizes to generate feature maps, and uses 1D-max pooling to expand the receptive field of feature maps, and further utilizes a MLP to complete emotional prediction. MLP, Multi-layer Perceptron.

the latent feature representations learned from neurons. In the addition phase, SAL improves the model's noise immunity by adding Gaussian noise to the feature representation. The SAL method improves the generalization performance of deep fusion models.

The formula for extracting text features by CNN is defined as follows:

$$\begin{aligned} x_{1:n}^t &= x_1 \oplus x_2 \oplus \dots \oplus x_n \\ c_i &= f(\omega \cdot x_{p:p+q-1} + b), \end{aligned} \quad (7)$$

where \oplus represents concatenation operator, ω represents convolution filter, c_i represents the feature representation within a window, $f(\cdot)$ represents activation function. Convolutional filters are used to extract features from all sentences to generate feature maps:

$$c = \text{maxpooling}[c_1, c_2, \dots, c_{n-h+1}]. \quad (8)$$

The max pooling operation is used to capture the most critical semantic information in the sentence.

It can be seen from the processing flow of the convolutional neural network that using CNN to extract text features does not contain contextual information, i.e., it is assumed that each sentence is independent of each other.

The CNN model is relatively simple and has a fast training speed. It can effectively extract local features in text, especially when processing long texts. Although CNN handles long texts well, it has poor adaptability to short texts and structured data and may lose some word order information when processing texts with sequential order.

4.1.6 Tensor Fusion Network (TFN). As shown in Figure 7, the tensor-based feature fusion method mainly calculates the tensor product of different modal feature representations through the Cartesian product to obtain the fused tensor representation [79]. Therefore, the above methods need to first map the input multi-modal feature representation into a

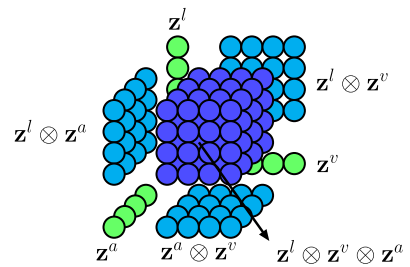


Fig. 7. Illustration of the TFN for tri-modal fusion. The feature vectors from the language (z^l), acoustic (z^a), and visual (z^v) modalities are first augmented with a constant term and then combined via tensor outer product. This operation explicitly captures unimodal, bimodal (e.g., $z^l \otimes z^a$, and $z^a \otimes z^v$), and trimodal ($z^l \otimes z^v \otimes z^a$) interactions in a structured tensor space.

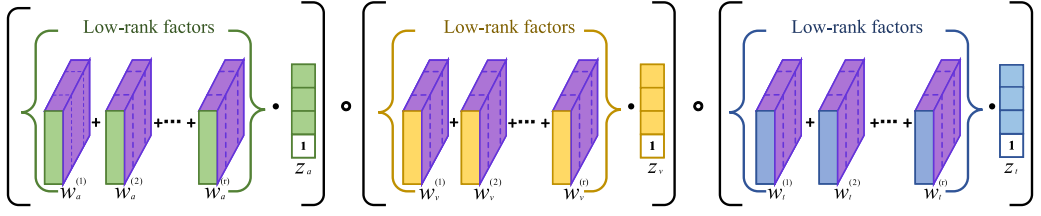


Fig. 8. The overall flow chart of LFM. LFM mainly performs low-rank decomposition of the learnable parameters of specific factors of the mode.

high-dimensional space, and then map it back to a low-dimensional tensor space for emotion representation. Tensor-based methods are able to capture important high-order interaction information across time, space, and modality. However, the computational complexity of tensor methods is very high and grows exponentially, and there is no fine-grained semantic information interaction between modalities. Zadeh et al. [145] proposed the multi-modal TFN. TFN adopts the method of tensor fusion, which can simulate the interaction process between the three modalities of text, audio and video, and effectively fuse multi-modal features. Although TFN can effectively model information interaction within and between modalities, the model complexity of the TFN method is related to the dimensionality of multi-modal features and grows exponentially. The formula of TFN is defined as follows:

$$\left\{ (x^t, x^v, x^a) \mid x^t \in \begin{bmatrix} \mathbf{x}^t \\ 1 \end{bmatrix}, x^v \in \begin{bmatrix} \mathbf{x}^v \\ 1 \end{bmatrix}, x^a \in \begin{bmatrix} \mathbf{x}^a \\ 1 \end{bmatrix} \right\}, \quad (9)$$

where the extra dimension with 1 is used to perform modal interaction. The Cartesian product is then used to fuse the three modal features as follows:

$$\begin{aligned} \mathbf{x}^m &= \mathbf{x}^l \otimes \mathbf{x}^v \otimes \mathbf{x}^a \\ &= \begin{bmatrix} 1 & \mathbf{z}_a^\top \\ \mathbf{z}_v & \mathbf{z}_v \mathbf{z}_a^\top \end{bmatrix} \begin{bmatrix} \mathbf{z}_l & \mathbf{z}_l \mathbf{z}_a^\top \\ \mathbf{z}_l \mathbf{z}_v^\top & \mathbf{z}_l \mathbf{z}_v^\top \mathbf{z}_a^\top \end{bmatrix}, \end{aligned} \quad (10)$$

where \otimes represents the outer product, \mathbf{x}^m represents fused vectors. With the help of tensor outer products, the interaction information of all levels can be systematically preserved.

Tensor fusion methods use tensor decomposition and high-dimensional fusion techniques to map multi-modal information into a unified space and capture high-order relationships between modalities through tensor operations. However, the computational overhead of processing tensor operations is high, which may lead to computational bottlenecks during training.

4.1.7 Low-rank TFN. On the basis of TFN, in order to more efficiently fuse multi-modal data, Liu et al. [66] proposed a **Low-rank Multimodal Fusion (LMF)** method to achieve dimensionality reduction of multi-modal features, so as to improve the fusion efficiency of multi-modal features as shown in Figure 8. **Low-rank Fusion Multimodal (LFM)** has achieved high performance on many different tasks:

$$\begin{aligned} \mathbf{x}^m &= \left(\sum_{i=1}^r \mathbf{w}_a^{(i)} \otimes \mathbf{w}_v^{(i)} \otimes \mathbf{w}_t^{(i)} \right) \cdot \mathbf{x} \\ &= \left(\sum_{i=1}^r \mathbf{w}_a^{(i)} \cdot x_a \right) \circ \left(\sum_{i=1}^r \mathbf{w}_v^{(i)} \cdot x_v \right) \circ \left(\sum_{i=1}^r \mathbf{w}_t^{(i)} \cdot x_t \right), \end{aligned} \quad (11)$$

where \mathbf{w}_a , \mathbf{w}_v , and \mathbf{w}_t represent the decomposed low-rank learnable tensor. LFM essentially retains the multi-modal high-order information expression capability of tensor outer product and explicitly

models the joint distribution and deep semantic dependencies of different modalities through low-rank approximation.

Low-rank TFN reduces the complexity of the model by low-rank decomposition of tensors, which can significantly reduce the computational overhead of tensors and reduce memory requirements. However, although low-rank decomposition can reduce model complexity, it may lose some information.

4.1.8 Data Augmentation with Generative Adversarial Networks (GANs). Multi-modal emotion recognition based on adversarial learning is an advanced direction in this field, which combines the principles of adversarial learning to improve the accuracy and robustness of emotion recognition [92, 144]. Next, we introduce the existing overall process of data augmentation based on adversarial generative networks.

(1) **Conditional Generative Adversarial Network (cGAN).** cGAN [110] is a variant of GAN that introduces conditional information to more precisely control the output of the generator. The core idea of cGAN is to pass additional condition information to the generator and discriminator during the generation process, thereby generating specific types of data based on given conditions. The main advantage of cGAN is its ability to precisely control the generation process in order to generate data that meets the conditional information. The optimization goal of cGAN is defined as:

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{\text{data}}(x)} \{\log D([x, y])\} + \mathbb{E}_{z \sim p_z(z)} \{\log(1 - D([G([z, y]), y]))\}, \quad (12)$$

where x represents real data, and y represents extra information:

$$\begin{aligned} \mathcal{L}_D^{(cGAN)} &= -\mathbb{E}_{x \sim p_{\text{data}}(x)} \{\log D([x, y])\} - \mathbb{E}_{z \sim p_z(z)} \{\log(1 - D([G([z, y]), y]))\} \\ \mathcal{L}_G^{(cGAN)} &= -\mathbb{E}_{z \sim p_z(z)} \{\log(D([G([z, y]), y]))\}. \end{aligned} \quad (13)$$

(2) **Adversarial Autoencoders (AAEs).** AAE [50] combines the ideas of autoencoder and GAN. The main goal of AAE is to make this encoding space more continuous and have better data generation capabilities while learning a compressed representation of data. The training objective function of AAE usually includes two parts: one is the reconstruction error of the autoencoder, which ensures the quality of the encoding, and the other is the GAN loss, which makes the encoding distribution more continuous and closer to the real distribution. The formula is defined as follows:

$$\begin{aligned} \mathcal{L}_D^{(AAE)} &= -\mathbb{E}_{z \sim p_z(z)} \{\log D(z)\} - \mathbb{E}_{x \sim p_{\text{data}}(x)} \{\log(1 - D(E(x)))\} \\ \mathcal{L}_E^{(AAE)} &= -\mathbb{E}_{x \sim p_{\text{data}}(x)} \{\log(D(E(x)))\} \\ \mathcal{L}_R^{(AAE)} &= \mathbb{E}_{x \sim p_{\text{data}}(x)} \{\|x - R(E(x))\|^2\}, \end{aligned} \quad (14)$$

where $p_z(z)$ represents the prior distribution.

(3) **Adversarial Data Augmentation Network (ADAN).** ADAN [119] includes the following components: autoencoder $R(E(x))$, auxiliary classifier $C(E(x))$, generator $G(z, y)$ and discriminator $D(h)$. First, ADAN aims to learn a latent representation of the input data x to preserve the emotional information in it. Second, it attempts to ensure that the generated latent representation is consistent with the emotional information of the input data by matching the posterior distribution $p(h|z, y)$ with $p(h|x)$. Third, ADAN simultaneously strives to minimize the reconstruction error between the input data x and its reconstructed version \hat{x} to ensure high-quality data reconstruction. The generator $G(z, y)$ accepts a sample z drawn from an M -dimensional Gaussian distribution and a one-hot encoding of the emotion label y as input, and the goal is to generate samples in the latent space such that they are indistinguishable from real samples. The discriminator $D(h)$ is optimized

to distinguish whether the latent vector h comes from real data or from the generator:

$$\begin{aligned}
 \mathcal{L}_D^{(\text{ADAN})} &= -\mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} \{\log D(E(\mathbf{x}))\} - \mathbb{E}_{\mathbf{z} \sim p_z(\mathbf{z})} \{\log(1 - D(G(\mathbf{z}, \mathbf{y})))\} \\
 \mathcal{L}_C^{(\text{ADAN})} &= -\mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} \left\{ \sum_{k=1}^K y_{\text{emo}}^{(k)} \log C(E(\mathbf{x}))_k \right\} \\
 \mathcal{L}_R^{(\text{ADAN})} &= \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} \{ \|\mathbf{x} - R(E(\mathbf{x}))\|^2 \} \\
 \mathcal{L}_E^{(\text{ADAN})} &= \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} \left\{ \|\mathbf{x} - R(E(\mathbf{x}))\|^2 - \sum_{k=1}^K y_{\text{emo}}^{(k)} \log C(E(\mathbf{x}))_k \right\} \\
 \mathcal{L}_G^{(\text{ADAN})} &= \mathbb{E}_{\mathbf{z} \sim p_z(\mathbf{z})} \left\{ \log(1 - D(G(\mathbf{z}, \mathbf{y}))) - \alpha \sum_{k=1}^K y_{\text{emo}}^{(k)} \log C(G(\mathbf{z}, \mathbf{y}))_k \right\},
 \end{aligned} \tag{15}$$

where α determines the contribution of classification error to model optimization. ADAN improves the authenticity of fused features and the ability to align cross-modal distributions through the confrontation between the discriminator and the generator.

GANs are able to generate high-quality new samples that are very close to real data through an adversarial training process. This enables the model to generate new samples with greater diversity and authenticity, thereby improving the model's generalization ability. For tasks with scarce data, especially when there are fewer samples of a specific category, GANs can be used to enhance the dataset by generating new samples, avoiding the limitations of traditional data augmentation methods on data diversity and complexity. However, the training process of GANs is often unstable, and the adversarial process between the generator and the discriminator may cause the gradient to vanish or explode, resulting in unstable quality of the generated samples.

4.2 Sequential Context Modeling

Context-free modeling is conceptually important and has inspired later research on sequential context modeling [115]. In particular, sequential context modeling methods consider that contextual sentences are mutually influential. Sequential context modeling approaches [70, 112, 127] consider each sentence influenced by its surrounding utterances. The main idea is to generate a feature representation with rich contextual semantic information by combining its own utterance representation x_i with the surrounding contextual sentence representation $\{x_{i-k}, \dots, x_{i-1}, x_{i+1}, \dots, x_{i+k}\}$, where k represents the context window size. Different from the context-free modeling method, the sequential context modeling method obtains a better feature representation by setting a memory network to preserve the context information of the sentence. Taking Figure 9 as an example, a LSTM or Transformer is used to extract contextual information for three modalities of video, audio, and text. The sequential context modeling approach plays an important role for many other MCER modeling approaches.

Tri-modal Hidden Markov Model is a sequence context modeling method, which relies on the previous state and can effectively capture local dependencies. For example, Morency et al. used text, video, and audio features for the task of tri-modal emotion analysis, and designed a model to extract useful information in different modal features [78]. After extracting multi-modal features, the three modal features are connected and input into a **Hidden Markov Chain (HMM)** classifier [78] to learn the emotional state of the input signal. HMM believes that the state of the current moment is only related to the information of the previous moment, which enables the model to use

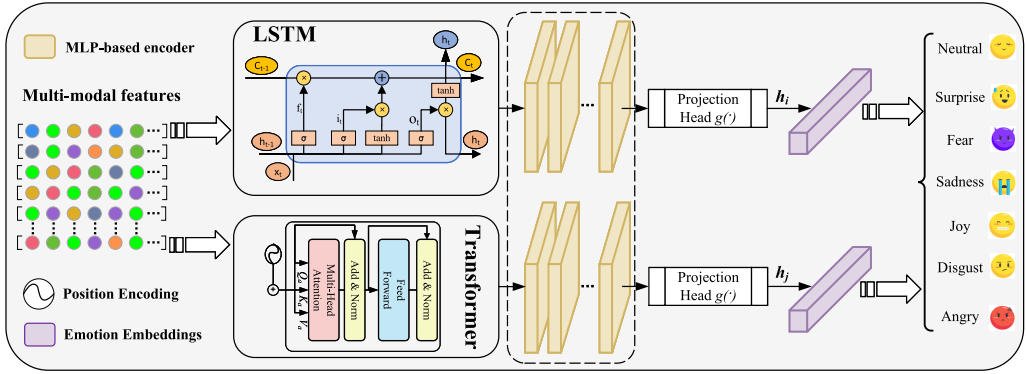


Fig. 9. The flowchart of the proposed contextual modeling approach. Sequential context modeling methods use LSTMs or Transformer to capture high-level features with rich contextual semantic information among different modal features.

the context information of the utterance. The formula of HMM is defined as follows:

$$\begin{aligned}
 P(w|x^a, x^v, x^t) &= \sum_{i=1}^C \sum_{j=1}^D \sum_{k=1}^M P(w, \lambda_i^a, \lambda_j^v, \lambda_k^t | x^a, x^v, x^t) \\
 &= \sum_{i=1}^C \sum_{j=1}^D \sum_{k=1}^M P(w | \lambda_i^a, \lambda_j^v, \lambda_k^t, x^a, x^v, x^t) \\
 &\quad \times P(\lambda_i^a, \lambda_j^v, \lambda_k^t | x^a, x^v, x^t),
 \end{aligned} \tag{16}$$

where C, D, M represent feature vector dimensions for audio, video, and text, w represents the emotional class, $P(\lambda_i^a, \lambda_j^v, \lambda_k^t | x^a, x^v, x^t)$ represents the confidence of the emotion classification.

Since the true class label is based on the output of the predicted class label \hat{w}_b , the formula of HMM can be expanded as follows:

$$\begin{aligned}
 &P(w | \lambda_i^a, \lambda_j^v, \lambda_k^t, x^a, x^v, x^t) \\
 &= \sum_{b=1}^B P(w, \hat{w}_b | \lambda_i^a, \lambda_j^v, \lambda_k^t, x^a, x^v, x^t) \\
 &= \sum_{i=1}^B P(w | \hat{w}_b, \lambda_i^a, \lambda_j^v, \lambda_k^t, x^a, x^v, x^t) \\
 &\quad \times P(\hat{w}_b | \lambda_i^a, \lambda_j^v, \lambda_k^t, x^a, x^v, x^t),
 \end{aligned} \tag{17}$$

where $P(\hat{w}_b | \lambda_i^a, \lambda_j^v, \lambda_k^t, x^a, x^v, x^t)$ represents the probability of predicted label. HMM captures cross-modal temporal dependencies and synchronization relationships through temporal dependency structures.

HMM can effectively process sequence data, especially when there are unobservable hidden states at each timestep. By modeling the hidden state, HMM can capture the time dependency in the data. Therefore, HMM has strong capabilities in sequence modeling. However, HMM is based on linear models and cannot effectively handle complex nonlinear relationships in sequence data. For nonlinear sequence data, the performance of HMM may be limited. In addition, when processing large-scale data, the computational complexity of HMM training is high, especially when the parameter space is large, computational bottlenecks may be encountered.

LSTM is a variant of RNN that can remember contextual information. Specifically, LSTM models long-distance dependent context through cellular units and can solve the vanishing gradient problem. Each LSTM consists of input gate j_t , output gate O_t , cell state C_t , and forget gate f_t :

$$\begin{bmatrix} \tilde{C}_t \\ O_t \\ j_t \\ f_t \end{bmatrix} = \begin{bmatrix} \tanh \\ \sigma \\ \sigma \\ \sigma \end{bmatrix} W_T \begin{bmatrix} x_t \\ h_{i-1}^t \end{bmatrix} \quad (18)$$

$$C_t = C_t \odot j_t + C_{t-1} \odot f_t$$

$$h_i^t = O_t \odot \tanh(C_t),$$

where σ represents activation function. LSTM extracts dynamic information within the modality by capturing time dependencies.

LSTM can remember information over long time spans and is particularly suitable for modeling long-term dependencies. However, LSTM still performs worse than other more advanced models (such as Transformer) in some extremely long sequence tasks. LSTM's ability to retain memory over long time spans is also limited. Therefore, LSTM is suitable for most tasks that need to capture time dependencies, but for very long dependencies or large-scale data, modern models such as Transformer may provide better performance.

After LSTM was used in MCER, many other works were proposed to extract contextual emotional information. Lu et al. [69] proposed a multi-scale LSTM multi-modal emotion recognition model, which uses LSTM to extract low-level and high-level local emotional features in multi-modal features. This method can capture subtle changes in complex expressions in a more fine-grained manner and implement an information feedback mechanism. However, it cannot capture the status information of the utterance and the status information of the speaker.

Existing models ignore modal alignment and directly fuse information on different modal features. Modal alignment can eliminate the heterogeneity of single-modal features and obtain accurate emotional representations of different modal features. Based on this current situation, Hou et al. [31] proposed a semantic alignment network based on multi-space learning, which uses LSTM to extract emotional features of different modalities and obtains high-level emotional representations as supervisory signals for modal alignment. This method can capture the global correlation between different modalities and achieve feature fusion between modalities.

Transformers are another way of modeling sequential context [38, 54, 163]. Transformer's long-distance modeling capabilities are far superior to RNNs, and Transformer can achieve parallel computing. Therefore, existing research on multi-modal emotion recognition based on sequential context modeling often regards Transformer as an important technology. The implementation details of Transformer are as follows.

Firstly, video, audio, and text features (i.e., x^t, x^a, x^v) are concatenated into a fusion vector. The formula is defined as follows:

$$Q, K, V = \text{Concat}(x^t, x^a, x^v), \quad (19)$$

where Q, K, V represent the query vector, key vector, and value vector of multi-modal features, respectively. Transformer automatically captures key modal and key area information by dynamically assigning weights.

Secondly, we use a feedforward neural network to perform multiple linear transformations on Q , K , and V . The formula is defined as follows:

$$\begin{aligned}\tilde{Q} &= \text{Concat}(QW_1^Q, \dots, QW_i^Q, \dots, QW_m^Q) \\ \tilde{K} &= \text{Concat}(KW_1^K, \dots, KW_i^K, \dots, KW_m^K) \\ \tilde{V} &= \text{Concat}(VW_1^V, \dots, VW_i^V, \dots, VW_m^V),\end{aligned}\tag{20}$$

where m represents the number of linear transformations.

We then perform multi-head attention in parallel to obtain emotion feature representation:

$$\begin{aligned}\text{head}_i &= \frac{\text{softmax}\left((QW_i^Q)(KW_i^K)^T\right)}{VW_i^V} \\ H_{\text{head}} &= \text{Concat}(\text{head}_1, \dots, \text{head}_m),\end{aligned}\tag{21}$$

where H_{head} represents the emotion feature vectors.

Finally, position encoding is used to obtain the position information of the emotion sequence:

$$\begin{aligned}PE_{(pos, 2i)} &= \sin\left(\frac{pos}{10000^{2i/d}}\right) \\ PE_{(pos, 2i+1)} &= \cos\left(\frac{pos}{10,000^{2i/d}}\right).\end{aligned}\tag{22}$$

where pos is the index of the i th sentence, position encoding information is fused into Q , K , and V .

Transformer uses a self-attention mechanism to process input data, which is independent of the order of the sequence. Compared with traditional RNNs and LSTMs, Transformer can calculate the relationship between each input position in parallel, which significantly speeds up the training process. Due to its parallelism, Transformer is more suitable for training large-scale datasets and can greatly reduce training time. In addition, Transformer's Multi-head Self-attention allows the model to focus on different parts of the input from multiple perspectives, which helps capture multi-level information. However, since Transformer involves interactions with all positions when calculating self-attention, its time complexity and space complexity are both high, especially when processing long sequences, the amount of calculation increases quadratically. For long sequences, Transformer's computational and memory requirements are very large. In addition, when the amount of data is limited, Transformer may not be able to fully exert its advantages, requiring a large amount of pre-training data and careful tuning.

After Transformer was proposed, many Transformer-based MCER methods were proposed to model long-distance context dependencies [18, 28]. Previous works failed to model long-distance dependencies between different modal features. To address this, Yang et al. [131] proposed a multi-modal speech emotion recognition method using a context Transformer, which enhances the emotional representation of the current utterance by embedding contextual information. This method can adaptively learn feature fusion between modalities.

Existing methods struggle to dynamically identify subtle emotional changes in multi-modal and multi-scale features. To address this, Liu et al. [65] proposed a multi-scale self-attention fusion emotion recognition method, which uses the self-attention mechanism to extract context-related dependencies in multi-modal features. Therefore, there is potential to use Transformers to model long-distance context dependencies. This method combines bc-LSTM and a multi-head attention mechanism to achieve fine-grained emotional information mining, and uses feature-level fusion and decision-level fusion methods to experiment with cross-modal feature fusion.

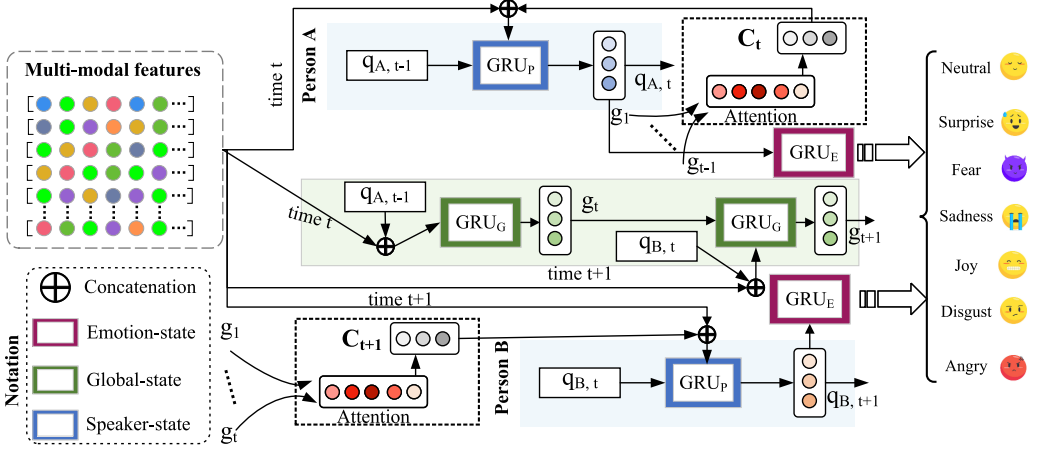


Fig. 10. The flowchart of the proposed distinguishing speakers modeling approach. The distinguishing speakers modeling approach designs three GRU states, i.e., a global GRU, an emotional GRU, and a speaker GRU, which are used to update global context information, emotion category information, and speaker information, respectively.

4.3 Distinguishing Speaker Modeling

The distinguishing speakers modeling method considers that the speaker's emotion is not only related to the global context, but also related to the speaker's own emotional state. Take Figure 10 as an example, there are three GRU states (i.e., a global GRU, an emotional GRU, and a speaker GRU). The global GRU is utilized to extract global multi-modal information and speaker's emotional state information. The speaker GRU is used to fuse the semantic information with context captured by the attention mechanism and the speaker's emotional state information. The emotion GRU combines the speaker's emotional state information and global context information to complete the final emotion classification.

Global GRU captures the contextual semantic information of an utterance by modeling the utterance and speaker states. Each speaker state is used to memorize a speaker-specific representation of an utterance. By distinguishing the subordination relationship between speakers and utterances, it is beneficial to model the dependency relationship between speakers and utterances, thereby enhancing the semantic representation ability of context. The formula of Global GRU is defined as:

$$g_t = GRU_G(x_{t-1}, (x_t \oplus q_{s(x_t), t-1})), \quad (23)$$

where g_t represents the latent feature representation of the global state, and $q_{s(x_t)}$ represents the speaker state of the current utterance x_t .

Speakers typically reply to conversations based on contextual information from other. Therefore, speaker GRU extracts the context c_t related to the utterance x_t . The formula is defined as follows:

$$\begin{aligned} \beta &= \text{softmax}(x_t^T W_\beta [g_1, g_2, \dots, g_{t-1}]) \\ c_t &= \beta [g_1, g_2, \dots, g_{t-1}]^T, \end{aligned} \quad (24)$$

where W_β is the learnable parameters. First, calculate the attention score of the global state in the previous $t-1$ time. The attention score assigns higher weight to utterances related to utterance x_t . The final context vector c_t is obtained by the dot product of the attention score β and the global state g_t :

$$q_{s(u_t), t} = GRU_P(q_{s(u_t), t-1}, (u_t \oplus c_t)). \quad (25)$$

The emotional representation e_t of the utterance u_t is obtained by combining the speaker's state $q_{s(ut),t}$ and the utterance e_{t-1} at time $t - 1$. One underlying intuition is that context has a greater impact on utterance u_t , and e_{t-1} integrates emotional contextual information from other parties' states into the emotional representation e_t . Therefore, we use the Emotion GRU unit to model e_{t_1} , and the formula is defined as follows:

$$e_t = GRU_E(e_{t-1}, q_{s(u_t),t}). \quad (26)$$

The emotion representation e_t that combines context information and speaker status information is used for the final emotion classification. Distinguishing speaker modeling method combines global context with individual information to improve understanding of complex information such as semantics and emotions.

Similar to traditional RNN and LSTM models, distinguishing speakers modeling methods have the ability to capture long-range dependencies. In multi-turn conversations, the previous conversation content often affects subsequent understanding and generation. Distinguishing speakers modeling methods can effectively learn these long-term dependencies and help understand contextual information. However, in scenarios where the amount of data is small or the conversation data is simple, distinguishing speakers modeling methods may suffer from overfitting. In particular, when there are not many conversation rounds in the training dataset, the model tends to remember specific patterns in the training set and cannot effectively generalize to new data.

In modeling methods based on distinguishing between speakers, Ghosal et al. [20] proposed **Commonsense Knowledge for Emotion Identification in Conversations (COSMIC)**, which clarifies the relationship between the speaker and the utterance. It also introduces common sense knowledge to enhance the emotional understanding of the model. COSMIC can learn a variety of different prior knowledge (e.g., event relationships and causal relationships), and can distinguish speaker information and dynamically detect the speaker's emotional changes.

In view of the fact that existing methods cannot pay attention to the correlation between utterances and speakers and the lack of interaction between speakers, Zhang et al. [151] proposed a conversational interaction model, which extracts contextual semantic information and state interaction information of utterances through stacked global interaction modules. In addition, this method also implements adversarial feature representation of the model by introducing noise information. Experimental results prove that adversarial learning can improve the performance of emotion recognition.

4.4 Speaker Relationship Modeling

4.4.1 GNN for Speaker Relationship Modeling. The speaker relationship modeling method innovatively introduces GNN to capture the speaker's dialogue relationship information while extracting sequential context information. Taking Figure 11 as an example, it extracts dialogue relationships between speakers and inter-speaker dependencies by constructing a speaker relationship graph.

GCN extends convolution operations into graph-structured data to extract structural information. GCN performs first-order neighbor information aggregation and spectral domain estimation. The formula of GCN is defined as follows:

$$H_i^{(l+1)} = ReLU \left(\tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}} H^{(l)} W^{(l)} \right), \quad (27)$$

where $W^{(l)}$ is the learnable parameters, $\tilde{A} = A + I_n$, I_n is the identity matrix, $\tilde{D}_{ii} = \sum_j \tilde{a}_{ij}$. $H^{(l+1)}$ represents the latent feature representations of layer $l + 1$.

The steps to apply GCN to the field of multi-modal emotion recognition are as follows. First, each utterance is represented as a node in the graph, and edge relationships are constructed based on the context between utterances. We then apply GCN to the constructed dialogue graph for

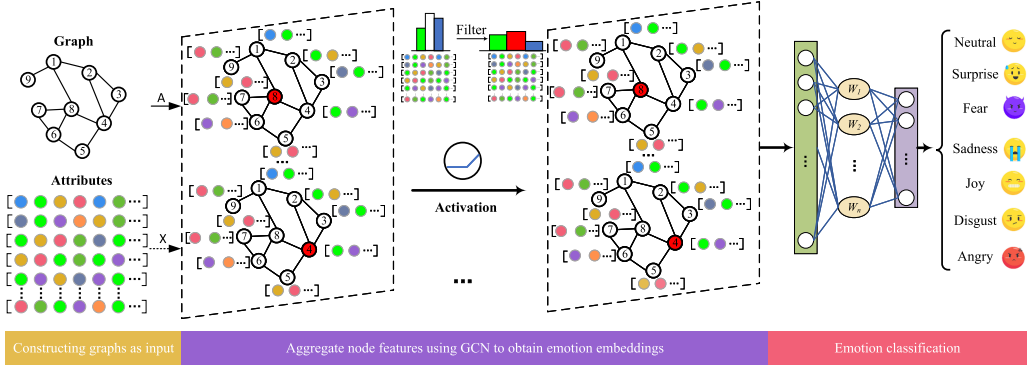


Fig. 11. The flowchart of the proposed speaker relationship modeling approach. The speaker modeling method aggregates dialogue relationship information between speakers by using a graph CNN.

speaker-level information extraction. Through the above process, the model can dynamically learn the correlation between sentences. According to the definition of Equation (27), our formula for aggregating surrounding contextual utterance information is deformed as follows:

$$H_i^{(l+1)} = ReLU \left(\sum_{r \in \mathcal{R}} \sum_{j \in \mathcal{N}_i^r} \frac{1}{|\mathcal{N}_i^r|} \left(W_{\theta_1}^{(l)} H_j^{(l)} + W_{\theta_2}^{(l)} H_i^{(l)} \right) \right), \quad (28)$$

where W_{θ_1} and W_{θ_2} are the learnable parameters, \mathcal{N}_i^r represents the neighbor node under the relationship $r \in \mathcal{R}$.

The multi-layer convolutional structure of GCN can effectively integrate feature information from different modalities and enhance the expressiveness of emotional information. Especially when the amount of information is large, GCN can effectively aggregate the features of each modality, which helps the accuracy of emotional classification or regression tasks. However, GCN has difficulties in modeling long-distance dependencies. As the depth of the graph increases, the problem of over-smoothing will occur during the information propagation process, causing the representation of nodes to become similar and lose their original distinguishability.

Graph Attention Network (GAT) is a variant of GCN that aggregates surrounding neighbor node features through learnable weights with an attention mechanism. GAT captures the more important node features in the graph by calculating the degree of similarity between nodes. The formula for GAT is defined as follows:

$$H_i^{(l+1)} = ReLU \left(\sum_{j \in N(w_i)} \alpha_{ij}^{(l+1)} W^{(l+1)} h_j^{(l)} \right), \quad (29)$$

where α_{ij} is the edge weight between node i and node j .

Similarly, the formula for using GAT to extract conversational relationships between speakers is defined as follows:

$$H_i^{+(l+1)} = ReLU \left(\sum_{r \in \mathcal{R}} \sum_{j \in \mathcal{N}_i^r} \frac{1}{|\mathcal{N}_i^r|} \left(\alpha_{ij}^{(l)} W_{\theta_1}^{(l)} H_j^{+(l)} + \alpha_{ii}^{(l)} W_{\theta_2}^{(l)} H_i^{+(l)} \right) \right). \quad (30)$$

GAT dynamically distinguishes the importance of neighboring nodes, giving the model stronger nonlinear expression and local pattern capture capabilities.

In multi-modal emotion recognition, the relationship between modalities may be heterogeneous, that is, the data of different modalities may be different in nature. GAT can effectively handle

such heterogeneous graphs because it gives different weights to each type of node through the attention mechanism, thereby better representing the heterogeneous relationship between different modalities. However, although the attention mechanism can improve the flexibility of information aggregation, it may also lead to excessive focus on local information, especially in some tasks that need to consider the global context. The local weighting of GAT may limit the global learning ability of the model.

The multi-modal method based on GNN is the current mainstream research, which can consider context information and speaker relationship information simultaneously [21]. To jointly learn sequential context, multi-modal interaction, and multi-task representation, Zhang et al. [157] designed the **Multi-modal, Multi-task Interactive Graph Attention Network (M3GAT)**. M3GAT simultaneously models context dependencies, multi-modal emotional interactions, and speaker dependencies. It enables cross-modal feature interaction, captures sequential contextual semantic information, and establishes task correlations.

Existing graph fusion methods often cause the model to lose important semantic information and fail to eliminate redundancy. To address this, Li et al. [55] proposed a graph network based on cross-modal feature complementarity. This method effectively extracts the speaker's context and interaction information using multiple hypothesis spaces in the graph. This method eliminates the heterogeneity between modalities and fuses modal information by performing different message aggregation on different nodes and edge relationships in the graph, thereby extracting contextual information and speaker relationship information.

Although existing MCER methods use GCN to model conversational relationships between speakers. In particular, the most competitive methods model the dependence of conversational relations between speakers and the importance between conversational relations by using relational GATs. However, existing GCN-based MCER methods do not consider conversational relationships and sequential information in contextual relationships. Based on the above problems, Ishiwatari et al. [40] introduced relational position coding in **Relationship Graph Attention Network (RGAT)** to provide sequence information. The specific flow chart of RGAT is shown in Figure 12.

The position encoding formula used by RGAT is defined as follows:

$$PE_{ijr} = \begin{cases} \max(-p, \min(p, j - i)) & r = 1, \text{ where } j \in \mathcal{N}^1(i) \\ \max(-p, \min(p, j - i)) & r = 2, \text{ where } j \in \mathcal{N}^2(i) \\ \max(-f, \min(f, j - i)) & r = 3, \text{ where } j \in \mathcal{N}^3(i) \\ \max(-f, \min(f, j - i)) & r = 4, \text{ where } j \in \mathcal{N}^4(i) \end{cases}, \quad (31)$$

where PE_{ijr} represents the relative position distance between node i under relationship type r and its surrounding neighbor nodes j . The maximum relative position distance between nodes is clipped to p or 4, which represents the context window size. $\mathcal{N}^r(i)$ represents the neighborhood of node i under relationship type r . To make the position encoding information learnable, **Feedforward Network (FFN)** is used to obtain position embeddings.

RGAT improves emotion recognition performance by explicitly modeling complex relationships within and between modalities. However, if the prior relationship definition is unreasonable (e.g., incorrectly connecting irrelevant modalities), the performance may deteriorate.

4.5 Emotion Classification

After obtaining the multi-modal emotion feature representation, the MCER task uses a multi-layer perceptron and a softmax layer to achieve the final emotion classification. The probability

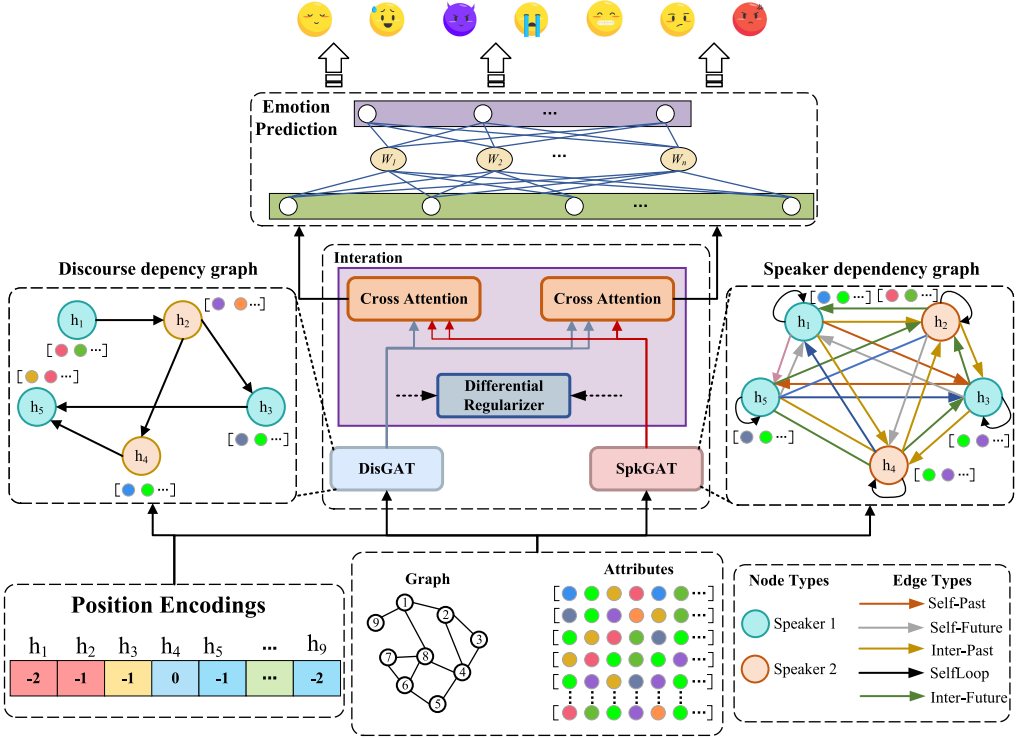


Fig. 12. The flow chart of RGAT. RGAT mainly includes dialogue relationship dependency graph, speaker dependence graph and position coding information. RGAT, Relationship Graph Attention Network.

distribution of emotion categories is as follows:

$$\begin{aligned}
 l_t &= \text{ReLU}(W_l e_t + b_l) \\
 \mathcal{P}_t &= \text{softmax}(W l_t + b) \\
 \hat{y}_t &= \arg\max_i (\mathcal{P}_t[i]),
 \end{aligned} \tag{32}$$

where W_l, W, b_l, b are the learnable parameters, \mathcal{P}_t is the probability distribution of emotion categories, \hat{y}_t is the predicted labels.

5 Evaluation Metrics

For MCER tasks, there are four commonly used evaluation indicators, i.e., accuracy rate, weighted average accuracy rate (WA), F1 value, and weighted average F1 value (WF1). These four indicators are defined as follows.

We assume that N is the number of emotion labels in the dialogue emotion dataset, E_j represents the total number of samples of emotion labels in the j th, $j \in [1, N]$.

(1) Accuracy represents the emotion recognition accuracy of the model, and the formula is defined as follows:

$$\text{Accuracy}_j = \frac{\sum_{n=1}^{\partial_2} E_j^i}{\sum_{m=1}^{\partial_1} S_j^m}, \tag{33}$$

where ϑ_1 is the number of labels on a certain category of emotion. ϑ_2 is the number that the model predicts on a certain category of emotion. E_j^i means that the i th sample in the j th emotionally predicted correctly. $E_j^i \in [0, 1]$. S_j^m represents the m th sample of the j th emotion. The larger the value of $Accuracy_j$, the better the recognition effect of the model on the j th type of emotion.

(2) The F1 value is the F1-score of each emotion, and the formula is defined as follows:

$$F1_j = \frac{2 \times \text{Recall}(E_{TP}^j, E_{FP}^j) \times \text{Precision}(E_{TP}^j, E_{FN}^j)}{\text{Recall}(E_{TP}^j, E_{FP}^j) + \text{Precision}(E_{TP}^j, E_{FN}^j)}, \quad (34)$$

and

$$\begin{aligned} \text{Precision}(E_{TP}^j, E_{FN}^j) &= \frac{|E_{TP}^j|}{|E_{TP}^j \cup E_{FN}^j|} \\ \text{Recall}(E_{TP}^j, E_{FP}^j) &= \frac{|E_{TP}^j|}{|E_{TP}^j \cup E_{FP}^j|}, \end{aligned} \quad (35)$$

where E_{TP}^j is the number of samples that the model predicts correctly on the j th category of emotion, E_{FP}^j is the number of samples that the model predicts incorrectly on the j th category of emotion, and E_{FN}^j is the number of emotions from other categories that the model predicts as the j th category of emotion. $\text{Precision}(E_{TP}^j, E_{FN}^j)$ is the model's precision on the j th category of emotion, and $\text{Recall}(E_{TP}^j, E_{FP}^j)$ is the recall of the model on the j th emotion. f1 value combines the effects of both precision and recall metrics. Usually, the larger the value of f1, the better the prediction of the model.

(3) Weight accuracy (WA) is the weighted average of the classification accuracy of all emotion categories. The more samples of the j th emotion, the smaller the weight of the sample. The formula is defined as follows:

$$WA = \frac{\sum_{m=1}^{\vartheta_1} S_j * \text{Accuracy}_j}{\sum_{j=1}^N \sum_{m=1}^{\vartheta_1} S_j^m}. \quad (36)$$

WA is the classification accuracy of the model combining all emotions. The larger the WA, the better the model performs on average across all classes.

(4) Weighted F1 (WF1) is the weighted F1 value of all emotion categories. The more samples of the j th emotion, the smaller the weight of the sample. The formula is defined as follows:

$$WF1 = \frac{\sum_{m=1}^{\vartheta_1} S_j * F1_j}{\sum_{j=1}^N \sum_{m=1}^{\vartheta_1} S_j^m}. \quad (37)$$

WF1 is the F1 value where the model integrates all emotions. WF1 is another effective index to evaluate the model effect. In general, the larger the WF1, the better the average performance of the model across all classes.

6 Experimental Results

To comprehensively evaluate the performance of different MCER methods, this experiment systematically analyzes the existing representative methods from multiple dimensions. First, from the perspective of overall performance, we statistically analyze the comprehensive performance of each method under the weighted F1. This section focuses on comparing the differences in the

effects of different categories of methods (e.g., context-free, sequential context, distinguishing speakers, and speaker relationship modeling), reflecting the advantages of introducing context information and speaker dependence mechanisms in improving recognition performance. Secondly, from the perspective of fine-grained performance, the comparison results of each method in terms of precision, recall, and AUC are further listed, and the stability and generalization ability of the model under different evaluation indicators are comprehensively measured. Meanwhile, to deeply analyze the actual application efficiency of the model, the parameter scale, inference time, and classification accuracy and F1 score of each method for different refined emotion categories (e.g., happy, sad, angry, excited, and fear) are statistically analyzed. This section reveals the tradeoff between parameter complexity and performance of different methods, and highlights the differences in the ability to maintain high-precision classification under small parameters and low latency conditions.

As shown in Table 6, we present the emotion recognition effects of different algorithms on multiple datasets. In particular, each algorithm uses multi-modal data, and we distinguish different MCER algorithms according to our classification method. Experimental results show that context-free-based algorithms have the worst performance because they contain the least semantic information and cannot obtain good emotional feature representation. The MCER algorithm based on sequential context has significant performance improvement compared to the context-free algorithm. The performance improvement may be attributed to the sequential context algorithm's ability to model the dependencies between contexts and its ability to utilize context information to improve the feature representation of emotions. The emotion recognition effects of modeling methods based on distinguishing speaker relationships and sequential context modeling methods are similar, and both are better than context-independent modeling methods. The performance improvement may be attributed to the ability of the distinguishing speakers modeling method to dynamically capture the speaker status information of the utterance and integrate it into the emotion representation information. The modeling method based on speaker relationship has the best performance and is currently the most popular modeling method. The modeling method based on speaker relationship mainly constructs the dialogue relationship between speakers through the inherent properties of the graph structure, and extracts the dialogue relationship representation between speakers through GCN. In addition, the speaker relationship modeling method can also consider the dependency information of the sequential context simultaneously.

To better understand the performance differences among various MCER algorithms, we provide a detailed analysis of each model's architecture, focusing on how structural design, parameter complexity, and feature fusion strategy influence emotion recognition effectiveness. TextCNN and LFM are classic context-free baselines. TextCNN uses convolutional layers on word embeddings without any sequential modeling or multi-modal interaction mechanisms. LFM extends this by incorporating limited modality fusion but still lacks temporal modeling. Their poor performance (e.g., <50% F1 on MELD) confirms that models ignoring context and interaction structures struggle to capture emotional semantics. bc-LSTM and bc-LSTM+Att represent sequential context models based on Bi-LSTM architectures. They process utterances in temporal order, enabling the capture of inter-utterance dependencies. The addition of attention mechanisms improves performance by focusing on emotionally salient parts. This explains the consistent performance improvement over context-free models, especially on IEMOCAP. A-DMN leverages a dynamic memory network, which not only captures sequential information but also performs iterative attention-based reasoning. Its moderate parameter size (7.39M) and good F1 performance indicate its strength in temporal reasoning while maintaining efficiency. DialogueRNN, a foundational distinguishing speakers model, models each speaker's emotional state over time using GRUs and a global attention mechanism. It explicitly distinguishes speaker roles, which improves emotion tracking in multi-speaker conversations.

Table 6. We Count the Performance of Different Types of Emotion Recognition Algorithms on Publicly Available Datasets

Approaches	Category	Inputs	Database	Performance (%)
SAL [118]	Context free	T+A+V	IEMOCAP/MELD	49.2/58.8
SVM [94]	Context free	T+A+V	IEMOCAP/MELD	48.7/56.4
TFN [145]	Context free	T+A+V	IEMOCAP/MELD	54.2/56.7
LFM [66]	Context free	T+A+V	IEMOCAP/MELD	55.3/56.7
UniMSE [36]	Sequential context	T+V+A	IEMOCAP/MELD	70.7/65.5
bc-LSTM+Att [84]	Sequential context	T+V+A	IEMOCAP/MELD	55.0/56.4
M2FNet [9]	Sequential context	T+V+A	IEMOCAP/MELD	69.9/66.7
CESTa [123]	Sequential context	T+V+A	IEMOCAP/DailyDialog/MELD	67.1/63.1/58.4
CMN [27]	Sequential context	T+V+A	IEMOCAP	56.2
SACL-LSTM [33]	Sequential context	T+A+V	IEMOCAP/MELD/EmoryNLP	69.2/66.5/39.7
Att-BiLSTM [116]	Sequential context	T+V+A	IEMOCAP	62.9
DialogueCRN [35]	Sequential context	T+A+V	IEMOCAP/MELD	66.2/58.39
EmoCaps [58]	Sequential context	T+V+A	IEMOCAP/MELD	71.8/64.0
ICON [26]	Sequential context	T+V+A	IEMOCAP	63.5
DialogueRNN [75]	Distinguishing speakers	T+V+A	IEMOCAP/MELD	62.8/56.8
EmotionIC [142]	Distinguishing speakers	T+V+A	IEMOCAP/DailyDialog/MELD/EmoryNLP	69.5/59.8/66.4/40.0
COIN [151]	Distinguishing speakers	T+V+A	IEMOCAP	65.4
COSMIC [20]	Distinguishing speakers	T+A+V	IEMOCAP/DailyDialog/MELD/EmoryNLP	65.3/58.5/65.2/38.1
RGAT [40]	Speaker relationship	T+A+V	IEMOCAP/DailyDialog/MELD/EmoryNLP	65.2/54.3/60.9/34.4
DialogueGCN [21]	Speaker relationship	T+V+A	IEMOCAP/MELD	64.2/58.1
DAG-ERC [99]	Speaker relationship	T+A+V	IEMOCAP/DailyDialog/MELD/EmoryNLP	68.0/59.3/63.7/39.0
MM-DFN [34]	Speaker relationship	T+V+A	IEMOCAP/MELD	68.2/59.5
GraphCFC [55]	Speaker relationship	T+V+A	IEMOCAP/MELD	68.9/58.9

The weighted F1 score is chosen as evaluation metric.

Its solid performance on both datasets confirms the benefit of dynamic speaker state modeling. DialogueGCN further advances this by applying graph convolution on utterance nodes, capturing both temporal and speaker-specific dependencies. It constructs directed conversation graphs and updates node representations via GCN layers. This architecture achieves strong overall F1, especially in emotions requiring long-term relational modeling (e.g., frustration and sadness). MM-DFN and M2ETNet are multi-modal transformer-based models that focus on advanced modality fusion strategies. MM-DFN integrates modality-specific feature streams using deep fusion networks, while M2ETNet adds temporal and modality-level attention. These models excel at learning fine-grained multi-modal interactions, resulting in top-tier performance across most emotion categories. Their performance, however, comes at the cost of larger parameter sizes and inference time. EmoCaps introduces a capsule network-based structure to model intra-modal hierarchies and inter-modal routing, capturing subtle semantic features. It shows strong performance in excited and angry categories, which benefit from complex vocal and textual cues. CT-Net integrates context and speaker-specific cues using temporal attention and cross-modal interactions. It performs stably across emotions with moderate complexity (8.49M). **Latent Relation-Graph Convolutional Network (LR-GCN)**, the most recent and complex model, constructs a speaker relationship graph with relational GCN layers. It jointly models sequential context, speaker identity, and emotion transition. Despite its large parameter size (15.77M) and long runtime (up to 147s on MELD), it consistently achieves the best F1-scores across datasets and emotion categories. Its strength lies in combining graph reasoning, speaker modeling, and contextual information in a unified framework. DisGCN focuses on distinguishing speakers via graph structure, while SumAggGIN applies hierarchical aggregation over utterances. These methods show reasonable performance, but are often outperformed by speaker relationship models, suggesting that modeling explicit speaker interaction via GCN yields more robust features. Models such as ICON and CNN focus only on sequential context or modality fusion without graph reasoning. Their lower performance suggests that lack of explicit speaker modeling limits their ability to capture dialogue dynamic.

Table 7. We Count the Performance of Different Types of Emotion Recognition Algorithms on Publicly Available Datasets

Approaches	Database	Precision (%)	Recall (%)	AUC (%)
SAL [118]	IEMOCAP/MELD	51.3/60.2	50.3/57.4	70.9/76.6
SVM [94]	IEMOCAP/MELD	49.1/59.4	47.6/57.9	65.3/73.2
TFN [145]	IEMOCAP/MELD	57.3/57.0	55.4/56.6	76.1/72.9
LFM [66]	IEMOCAP/MELD	56.2/58.4	57.3/56.1	78.4/71.7
UniMSE [36]	IEMOCAP/MELD	68.8/65.2	65.4/64.7	83.6/79.0
bc-LSTM+Att [84]	IEMOCAP/MELD	56.7/58.9	57.4/56.3	76.4/71.7
M2FNet [9]	IEMOCAP/MELD	67.6/67.4	65.3/66.1	85.1/79.3
CESTa [123]	IEMOCAP/DailyDialog/MELD	68.5/64.3/59.2	67.2/65.7/62.3	87.9/76.7/79.2
CMN [27]	IEMOCAP	58.4	57.3	75.3
SACL-LSTM [33]	IEMOCAP/MELD/EmoryNLP	69.1/65.1/43.8	67.3/64.7/45.6	88.4/79.0/60.1
Att-BiLSTM [116]	IEMOCAP	64.5	62.7	80.3
DialogueCRN [35]	IEMOCAP/MELD	67.4/61.4	65.3/62.3	84.0/79.3
EmoCaps [58]	IEMOCAP/MELD	70.1/64.5	71.2/62.7	86.8/78.5
ICON [26]	IEMOCAP	64.3	62.0	83.3
DialogueRNN [75]	IEMOCAP/MELD	65.5/58.3	63.7/59.6	80.9/73.6
EmotionIC [142]	IEMOCAP/DailyDialog/MELD/EmoryNLP	70.0/62.5/65.6/43.1	68.6/63.3/64.3/46.8	87.2/72.1/75.2/55.3
COIN [151]	IEMOCAP	67.8	66.5	85.0
COSMIC [20]	IEMOCAP/DailyDialog/MELD/EmoryNLP	66.3/60.2/63.1/43.3	64.2/63.1/60.9/44.5	85.2/74.8/82.2/60.9
RGAT [40]	IEMOCAP/DailyDialog/MELD/EmoryNLP	67.8/55.2/61.3/36.3	65.7/56.3/62.4/39.9	84.2/66.9/69.7/52.3
DialogueGCN [21]	IEMOCAP/MELD	65.3/57.1	54.2/58.5	83.9/69.2
DAG-ERC [99]	IEMOCAP/DailyDialog/MELD/EmoryNLP	69.5/60.7/64.2/41.3	68.4/61.3/62.2/43.8	87.9/72.3/66.1/62.0
MM-DFN [34]	IEMOCAP/MELD	67.2/62.5	66.4/63.1	87.2/77.6
GraphCFC [55]	IEMOCAP/MELD	69.3/60.4	68.6/58.0	89.1/81.5

The precision, recall, and AUC are chosen as evaluation metric.

To provide a more comprehensive assessment of model performance, we incorporate multiple evaluation metrics including Precision, Recall, and AUC, in addition to the Weighted F1-score. As shown in Table 7, we observe that Graph-based models generally outperform traditional context-free and sequential models across all three metrics. For instance, GraphCFC achieves the highest AUC (89.18%) and also maintains strong Precision (69.96%) and Recall (60.48%), indicating both robust classification ability and balanced detection across emotion categories. Among the sequential context models, architectures like SACL-LSTM and DialogueGCN show competitive performance, with SACL-LSTM achieving a Recall of 76.47% and AUC of 88.47%. This suggests that incorporating temporal dependencies and attention mechanisms enhances the model's sensitivity to subtle emotional cues. Speaker modeling methods (e.g., EmotiCon, RGAT, DialogueRNN) also perform well. Notably, EmotiCon achieves a high Recall of 81.63% but slightly lower Precision, suggesting that the model favors recall-oriented decisions, potentially useful in applications where missing emotional signals is more critical than occasional false alarms. Interestingly, simpler models such as CNN or UniMSE display relatively low AUC and Precision, indicating that they struggle to make accurate decisions across varied threshold settings, reinforcing the importance of context and multi-modal integration in MCER tasks.

In addition, we also counted the emotion recognition effects of different MCER algorithms on different emotion categories. As shown in Table 8, on the IEMOCAP dataset, the performance effects of each algorithm on various emotions are consistent with the overall results introduced previously. The method based on context-free modeling has the worst effect, with the recognition effect on the “happy” emotion being less than 50%. In comparison, most of the other three types of algorithms have exceeded 60%, and some categories of emotions have exceeded 80%. The performance of each algorithm on the MELD dataset is shown in Table 9. The recognition effects of each algorithm on most categories of emotions are similar to those on the IEMOCAP dataset. It is important to note that we found that all emotion recognition methods have poor performance in identifying “fear” and “disgust” emotions, and the accuracy of some algorithms is even 0%. When we observe the

Table 8. On the IEMOCAP Dataset, We Counted the Parameters, Running Time, and Emotion Recognition Effects of Different MCER Algorithms on Different Emotion Categories

Methods	IEMOCAP							
	Parmas.	Running time	Happy		Sad		Neutral	
			Acc. F1	Acc. F1	Acc. F1	Acc. F1	Acc. F1	Acc. F1
TextCNN [45]	0.47M	0.96 s	27.73	29.81	57.14	53.83	34.36	40.13
bc-LSTM [84]	1.28M	2.16 s	29.16	34.49	57.14	60.81	54.19	51.80
bc-LSTM+Att [84]	2.17M	2.59 s	30.56	35.63	56.73	62.09	57.55	53.00
CMN [27]	3.85M	4.14 s	25.01	30.34	55.96	62.45	52.81	52.36
LFM [66]	6.24M	6.23 s	25.63	33.14	75.71	78.83	58.52	59.21
A-DMN [129]	7.39M	6.69 s	43.15	50.64	69.47	76.88	63.05	62.92
DialogueRNN [75]	15.17M	20.05 s	25.63	33.11	75.14	78.85	58.56	59.24
DialogueGCN [21]	12.92M	14.18 s	40.63	42.71	89.14	84.45	61.97	63.54
DialogueCRN [35]	6.57M	6.44 s	71.47	51.93	66.17	59.86	78.53	64.16
SumAggGIN [100]	14.28M	18.79 s	56.74	54.22	86.85	79.17	62.95	65.32
DisGCN [108]	14.16M	17.85 s	71.17	56.92	68.65	76.47	66.63	57.41
MM-DFN [34]	6.29M	6.27 s	40.17	42.22	74.27	78.98	69.13	66.42
M2FNet [9]	9.31M	10.37 s	65.92	60.00	79.18	82.11	65.80	65.88
EmoCaps [58]	13.41M	16.28 s	70.34	72.86	77.39	82.45	64.27	65.10
CT-Net [61]	8.49M	9.06 s	47.97	51.36	78.01	79.94	69.08	65.82
LR-GCN [91]	15.77M	21.06 s	54.24	55.51	81.67	79.14	59.13	63.84

The best result in each column is in bold.

Table 9. On the MELD Dataset, We Counted the Parameters, Running Time, and Emotion Recognition Effects of Different MCER Algorithms on Different Emotion Categories

Methods	MELD							
	Parmas.	Running time	Neutral		Surprise		Fear	
			Acc. F1	Acc. F1	Acc. F1	Acc. F1	Acc. F1	Acc. F1
TextCNN [45]	0.34M	15.25 s	76.23	74.91	43.35	45.51	4.63	3.71
bc-LSTM [84]	1.07M	20.29 s	78.45	73.84	46.82	47.71	3.84	5.46
bc-LSTM+Att [84]	1.66M	24.28 s	70.45	75.55	46.43	46.35	0.00	0.00
A-DMN [129]	6.25M	51.13 s	76.54	78.92	56.24	55.35	8.22	8.61
DialogueRNN [75]	14.28M	141.31 s	72.12	73.54	54.42	49.47	1.61	1.23
CT-Net [61]	7.69M	61.46 s	75.61	77.45	51.32	52.76	5.14	10.09
DialogueCRN [35]	4.78M	40.35 s	70.91	75.73	47.32	47.18	0.00	0.00
SumAggGIN [100]	13.26M	122.69 s	78.19	77.82	52.27	54.11	2.17	2.31
DisGCN [108]	13.17M	108.71 s	70.84	76.67	42.71	46.13	1.17	1.55
MM-DFN [34]	5.33M	45.63 s	78.17	77.76	52.15	50.69	0.00	0.00
M2FNet [9]	8.15M	66.54 s	72.88	67.98	72.76	58.66	5.57	3.45
EmoCaps [58]	12.31M	101.77 s	75.24	77.12	63.57	63.19	3.45	3.03
LR-GCN [91]	14.97M	147.38 s	81.51	80.83	55.42	57.11	0.00	0.00

The best result in each column is in bold.

distribution of the dataset, we can find that the MELD dataset has a serious data imbalance problem. This results in the model's very poor emotion recognition performance on minority classes.

To comprehensively evaluate the practicality of different MCER models, we compare their parameter sizes and inference time in addition to their recognition performance. Tables 8 and 9 report the number of parameters and average running time per inference on the IEMOCAP and MELD datasets, respectively. The results show that lighter models such as TextCNN and bc-LSTM possess significantly fewer parameters (0.47M and 1.28M, respectively) and exhibit faster inference times (0.96 s and 2.16 s). However, their performance across most emotion categories tends to be lower compared to more complex architectures. In contrast, recent transformer-based or graph-enhanced models, such as DialogueGCN, MM-DFN, M2ETNet, and especially LR-GCN, require substantially more parameters (up to 15.77M) and longer inference times (up to 21.06 s for IEMOCAP and 147.38 s for MELD), but consistently achieve better recognition performance. This

observation highlights a common tradeoff between model complexity and computational efficiency. While heavier models are better suited for applications where accuracy is the priority, lightweight models may be more favorable for real-time or resource-constrained deployment scenarios. In particular, LR-GCN achieves the best overall performance on both datasets, ranking highest in multiple emotion categories, albeit with the largest parameter count and slowest inference time. This suggests that while model complexity improves expressive power, there is a pressing need to explore model compression, pruning, and quantization to improve the feasibility of deploying such models in real-world applications.

7 Applications of Multi-modal Conversational Emotion Analysis

Emotion recognition is a method of applying natural language processing, machine learning, and deep learning techniques to multi-modal data such as text, video, and audio to identify and analyze the emotional state expressed in multi-modal data [11]. Therefore, analyzing and studying the problem of emotion recognition has broad application value in many practical application scenarios.

7.1 Social Media Analysis

MCER has many broad applications in social media analysis [153]. The most typical application is product improvement and innovation, that is, by analyzing user comments and feedback on social media, companies can understand users' preferences and dissatisfaction with products. This helps companies tweak product designs, improve functionality, and develop products that better meet user needs. Therefore, businesses can employ emotion analysis techniques to improve their products. In addition, emotion analysis can also help advertisers understand users' emotional attitudes towards advertisements, thereby optimizing advertisement content and strategies, and improving advertisement effectiveness. For example, FaceReader [53] can measure people's emotional responses to different advertising creatives, providing valuable insights into the effectiveness of emotional appeals, humor, or shock value. With an 89% recognition rate for static images and an 80% recognition rate for animated expressions, FaceReader provides a reliable method for assessing the emotional impact of advertising imagery.

7.2 Public Opinion Analysis

Multi-modal conversational emotion analysis also has a wide application value in opinion mining, which can help mine and analyze people's opinions and emotions expressed in text, video, and audio [111]. For example, emotion analysis of online public opinion on emergencies can better understand public emotion and conduct effective crisis management. ECR-BERT [117] proposed a BERT-based model that integrates emotion-cognitive reasoning mechanisms, enabling more accurate emotional understanding in complex scenarios such as public opinion analysis during sudden events. Compared with the standard BERT model, ECR-BERT achieved absolute average accuracy improvements of 0.82%, 1.74%, 0.98%, and 1.37% across different datasets. This enhanced emotional recognition capability helps to more precisely capture public sentiment dynamics, providing valuable support for timely public opinion monitoring and effective crisis management.

7.3 Recommendations Systems

Multi-modal conversational emotion analysis in recommender systems can help personalize recommendations more in line with users' emotions and preferences [12]. For example, the recommendation system can recommend products that users are more interested in according to the emotional changes of consumers, and can perform emotion analysis on multi-modal data of user evaluations to realize real-time early warning and disposal of negative product evaluations. For

example, Agent4Rec [150] displays four movies on each recommendation page, and the agent will decide whether to continue to the next recommendation page or exit the recommendation system based on their satisfaction. After the agent exits, the system will ask him to give a satisfaction score of the recommendation system, ranging from 1 to 10. Ratings above 3 are regarded as signals of liking. After the entire simulation is completed, the following multi-faceted indicators are collected: average viewing rate, average number of likes, average like ratio, average number of exit pages, and average user satisfaction score. The satisfaction score of random recommendations is 2.93, while the satisfaction score of the recommendation algorithm based on emotional preferences is 3.85. The experimental results prove the effectiveness of emotion recognition technology in the field of social media.

7.4 Medical Care

Multi-modal conversational emotion analysis plays an important role in many aspects in the field of health care [95]. It can help medical institutions and doctors better understand the current emotional state of patients, so as to give better treatment plans. For example, by performing emotion analysis on unstructured data such as patients' medical records, consultation conversations, online consultation texts, or social media posts, the medical system can identify patients' potential emotional states, thereby providing supplementary psychological information for clinical diagnosis [107]. In addition, in mental health assessments, emotion analysis can be used to detect early signs of depression tendencies or anxiety symptoms, which helps to achieve early screening and early intervention. In chronic disease management, the system can continuously monitor patients' emotional responses to treatment plans and help doctors dynamically adjust treatment strategies.

7.5 Financial Field Analysis

In the field of financial analysis, emotion analysis can help financial practitioners and investors better understand the emotional state of the market and predict market trends, thereby helping investors make correct investment decisions [18]. For example, the improved GPT model [128] combined with the emotion analysis module achieved an accuracy of 88.34% in the financial emotion classification task, which is significantly better than the original GPT model (74.57% accuracy) that did not use emotion information. The results show that emotion factors can provide the model with richer semantics and market tendency judgment basis, further verifying the actual effectiveness and research value of emotion analysis in financial text understanding and trend prediction.

7.6 Social Robot

MCER has many potential applications on social robots, which can enhance the capabilities of social robots and make them more intelligent and humane [52]. Social robots can use multi-modal emotion recognition to sense the emotional state of the users they interact with. This includes identifying users' facial expressions, voice emotions, text emotions, and other modal emotional signals [49]. The robot can then adjust its interaction to better meet the user's emotional needs, providing support, comfort or entertainment. In addition, social robots can use MCER to better understand users' needs and emotional states to provide personalized suggestions and assistance. For example, emotion analysis can detect in real time the user's anxiety, frustration, or fatigue during the interaction process. The chatbot can then adjust its language response strategy and provide more soothing tone and suggestions, thereby significantly improving the user's trust and satisfaction. After the introduction of the emotion perception mechanism, the user's positive feedback rate on the chatbot has been significantly improved [41].

8 Privacy and Security of Multi-modal Data

With the widespread application of MCER, privacy concerns have become increasingly critical [136]. Unlike unimodal data, multi-modal data often involves sensitive personal information spanning facial expressions, voice patterns, textual content, and physiological signals. These data types may reveal not only the user's identity but also intimate emotional states, behavioral tendencies, and even mental health conditions. Consequently, the risk of personal emotional information leakage poses a significant challenge to the secure deployment of these technologies. To address these issues, privacy protection must be considered at every stage of system design and data processing. First, data anonymization techniques should be applied to remove or obfuscate identifying information, such as name, face, or unique voice features [24, 134]. Second, data encryption techniques [93, 159] should be used to ensure protection against unauthorized access to sensitive data, whether in storage, transmission, or computation. In a distributed environment, federated learning [160] provides an effective framework in which model training is performed locally on the user's device and only encrypted model updates are shared with a central server, thereby protecting the privacy of the original data. Moreover, differential privacy [130] can be introduced to inject calibrated noise into feature representations or model outputs, reducing the possibility of individual re-identification. In scenarios where multiple institutions or agents collaborate, secure multi-party computation and homomorphic encryption offer mechanisms for privacy-preserving joint model training or inference, albeit at the cost of computational efficiency [130]. Finally, privacy-preserving representation learning is gaining traction, where adversarial training or disentangled learning techniques are used to suppress sensitive attributes (e.g., user identity) while preserving task-relevant emotional information [15].

9 Research Challenges

Although deep learning technology has promoted the prosperity of MCER tasks, many scholars have proposed many state-of-the-art algorithms. However, building an accurate MCER model still faces challenges.

9.1 Scarcity of Training Data

MCER models require sufficient and comprehensive emotional samples as a basis to achieve accurate prediction or classification of emotions. The existing multi-modal benchmark datasets IEMOCAP, MELD, and SEMAINE have only 11,098, 5,810, and 394 utterances, respectively. Unfortunately, although we can easily collect large amounts of multi-modal conversation data from channels such as social media, the emotion labeling process is often expensive and time-consuming. In addition, the collected multi-modal data inevitably has problems such as ambiguous labels or multiple labels, which makes it a great challenge to obtain sufficient multi-modal labeled data, which in turn leads to the scarcity of multi-modal training data. Therefore, the scarcity of training data limits the effectiveness of current MCER models.

9.2 Data Is Heterogeneous and Noisy

MCER models need to fully eliminate heterogeneity and noise information between modalities to achieve accurate prediction or classification of emotions. Multi-modal data is naturally heterogeneous, and features of different modalities have huge differences in processing methods and representation forms. Additionally, multi-modal conversation data often contains a large amount of redundant or noisy information. The emotion is typically determined by a small amount of consistent key information, such as specific words in a sentence, a particular frequency band in speech, or a distinct expression in a video. Even in some extreme cases, part of the modal information is

basically unavailable under noise interference, such as ambiguous sentence expressions, noise in the speech, blocked expressions, and so on. Therefore, the heterogeneity and noise of data limit the effectiveness of current MCER models.

9.3 Unbalanced Data Distribution

Multi-modal dialogue data samples have serious imbalance problems, and the unbiased learning of the model is seriously interfered with. The MCER model is based on cross-modal feature fusion, driven by emotion category sample data, and is easily affected by the number of emotion category samples. However, multi-modal conversation emotion data naturally suffers from the problem of category sample imbalance. A few emotion category samples account for a larger proportion, while most emotion category samples account for a small proportion. For example, in the MELD dataset, the “fear” emotion only accounts for 1.91% of the total samples, and the “disgust” emotion only accounts for 2.61% of the total samples. A similar sample distribution also exists on the benchmark dataset SEMAINE. Small samples are difficult to drive unbiased learning of the model, which seriously affects the model’s prediction accuracy for small sample emotional categories. Therefore, the unbalanced sample distribution limits the effectiveness of current MCER models.

9.4 Consistent Semantic Association

MCER requires the model to learn the consistent semantics across modalities to filter out noise and eliminate heterogeneity. This is essential for building an accurate multi-modal emotion recognition model. However, the consistent semantic association in multi-modal conversation is more complex. It is not only related to the multi-modal context but also influenced by factors such as the conversation scene, the speaker’s emotional inertia, and their responses. In addition, multi-modal data are heterogeneous, each modality has differentiated representation and distribution characteristics in space, and some consistent semantic associations are hidden in the feature distribution space between modalities. Therefore, efficiently performing consistent semantic association is the primary issue that needs to be considered at the model level.

9.5 Complementary Semantic Capture

MCER models need to establish accurate and consistent semantic associations and capture complementary semantic features between modalities, which can expand the emotional representation capabilities of a single modality. However, unlike consistency semantics, complementary semantics represent differences between modalities, and this difference may contain noise components. Therefore, consistency semantics and complementarity semantics are a pair of game entities, and how to balance the relationship between them is another issue that needs to be considered at the model level.

9.6 Multi-model Collaboration

Multi-model collaboration is the third challenge faced at the model level in building accurate MCER models. MCER often requires the collaboration of multiple models to complete tasks, such as feature extraction models and feature fusion models. However, existing methods often perform task collaboration from the data level and ignore the collaborative relationship between models. Therefore, in order to achieve ideal synergistic results, not only the respective characteristics of the modes and their interrelationships need to be considered, but also the synergistic relationships between models need to be considered.

10 Future Work

10.1 Multi-modal Conversation Data Generation

MCER models require sufficient and comprehensive emotional samples as a basis. When sample data is scarce, training MCER models without causing overfitting or underfitting problems is extremely challenging. However, the sample size of existing benchmark datasets is relatively small, and there is a common problem of data scarcity. Multi-modal dialogue data generation can effectively alleviate this problem. However, the distribution of multi-modal conversation data is more complex, and traditional single-modal data generation or cross-modal data generation models cannot meet the requirements. Therefore, there is an urgent need to solve the problem of collaborative generation of multi-modal conversation data.

To solve the problem of ensuring strong correlation and synergy among modalities in Multi-modal Conversation Data Generation, researchers have designed advanced generative frameworks that explicitly model the cross-modal dependencies during the generation process. One representative approach leverages **Variational Autoencoders (VAEs)**, which map different modalities into a shared latent semantic space, allowing the model to capture deep inter-modal relationships and generate coherent multi-modal conversational data through joint reconstruction. This shared representation ensures that generated text, audio, and visual signals are semantically aligned and contextually consistent. In addition, GAN-based frameworks introduce modality-specific discriminators alongside a joint discriminator to constrain both the individual quality and overall coherence of generated modalities. By adversarially optimizing the generator to produce modality-consistent outputs, these methods effectively enhance cross-modal correlation in generated conversations. Recently, diffusion models have shown strong potential for multi-modal data generation by modeling the complex joint distribution of multiple modalities through iterative denoising steps. Diffusion-based methods [30] can incorporate cross-modal conditional signals at each generation step, ensuring that the evolving text, audio, and visual outputs remain temporally synchronized and semantically coupled.

10.2 Multi-modal Feature Deep Fusion

Multi-modal feature fusion is crucial to the MCER task. The fused feature vector can represent the consistent semantics and complementary information between modalities. However, many different information interactions exist between multi-modalities, and many consistent or complementary features are hidden in multiple time series or local spatial correlations. Since multi-modal conversation data is heterogeneous and contains noise, there are significant differences in the temporal period and spatial distribution of different modal features, and the spatiotemporal importance between modalities is dynamic. Currently, few works consider this difference, and more efforts are still needed for deep fusion of multi-modal features.

To solve the above problems, on the one hand, the deformable temporal convolution can be used to allow each modality to dynamically sample the most relevant timestep based on its own features. Then, the locality-aware attention is utilized to focus on the strong correlation of local areas in space. The time period misalignment and local information loss caused by heterogeneity are solved before fusion. On the other hand, to achieve cross-modal alignment, a dynamic weighted alignment mechanism can be introduced to calculate the dynamic weights of consistency and complementarity between modalities for each moment or local spatial area. Through the cross-modal attention module, the consistency score and complementarity score of each modality at the current spatiotemporal point are calculated. Then, the dynamic gating mechanism is used to adaptively adjust the contribution weight of each modal feature during fusion according to the score to avoid information redundancy or key feature loss caused by static weighting.

10.3 Unbiased Emotional Learning

Many benchmark datasets in the field of MCER suffer from serious sample category imbalance, that is, the minority emotion category contains a large amount of data, while the majority category emotion only contains a small amount of data. In the case of unbalanced data, the existing models tend to be biased towards fitting the minority emotion with a large amount of data, and the learning is insufficient on the majority emotion with a small amount of data, which leads to the model being in a small sample emotion category, resulting in the recognition accuracy is poor. Thus, the small-sample problem in multi-modal dialogue emotion recognition urgently requires further research.

To effectively solve the problem of small samples in MCER, we can start from three dimensions: enhancing samples, optimizing model structure, and adjusting training strategies, and build an integrated framework of data augmentation, prototype modeling, and category balance training. Specifically, at the data augmentation level, the emotional sentences of small sample categories are combined with context to reconstruct new samples, and the diversity of small sample categories is improved while ensuring the rationality of the context. At the prototype modeling level, a multi-modal prototype center is constructed for each emotion category, and the distance between samples and prototypes of the same type is shortened by contrastive loss, and the distance from prototypes of other types is pushed away to alleviate the problem of inter-class imbalance caused by differences in sample size. At the category balance training level, the category balance focal loss or label distribution-aware margin loss is used to dynamically adjust the loss penalty items of each category. Without changing the overall training process, the model's attention to small sample categories is improved.

10.4 Incomplete MCER

Each modality is not always available in real-world scenarios, which can lead to modal incompleteness problems. For example, the voice contains much noise, the expression is blocked, the light is dim, and so on. At this moment, some modal information becomes unavailable due to noise interference. Modal integrity requirements reduce the applicability of MCER methods. Therefore, cross-modal content recovery methods based on deep learning should continue to be developed to achieve MCER in missing modalities.

To solve the common problem of missing modalities in MCER, researchers have proposed a variety of modal restoration methods in recent years, aiming to maintain the discriminative ability of the model when some modalities are unavailable. A common type of method is based on autoencoders or VAEs, which reconstruct the representation of the missing modality using the available modalities by learning the mapping relationship between different modalities. Another type of method uses GANs to constrain the generated features by introducing discriminators to improve the authenticity and diversity of modality completion. There are also studies that use knowledge distillation strategies to guide student models to maintain performance in modality-incomplete scenarios with the help of teacher models trained under complete modalities. Recently, the development of generative models has also provided stronger modeling capabilities for modality restoration. Diffusion models have strong distribution modeling and step-by-step optimization capabilities by modeling the reverse generation process from noise to data. They perform well in conditional generation tasks and are suitable for high-quality restoration of missing modalities. Meanwhile, flow-based models accurately model the joint distribution of modalities through reversible transformations and support conditional sampling to achieve missing modality completion.

10.5 Zero-shot MCER

Affected by factors such as the complexity of emotions and the high cost of labeling, it is difficult to fully label some emotional samples. Furthermore, with the rapidly growing personal emotion annotation space, real-world emotion recognition systems may frequently encounter unseen emotion labels. Therefore, improving the generalization performance of emotion recognition models is an issue that needs to be considered. Deep methods utilizing zero-shot learning are expected to achieve better multi-modal dialogue emotion recognition.

In recent years, large-scale pretrained models, especially cross-modal models such as CLIP [88], Flamingo [1], GPT-4V [132], and so on, have brought a new paradigm for zero-shot multi-modal emotion recognition. These models usually have strong multi-modal alignment capabilities and natural language understanding capabilities, and can achieve category expansion through contrastive learning or generative modeling of text-image or text-audio. For example, CLIP builds a shared embedding space through image-text contrastive learning, which can convert emotion labels into natural language descriptions to achieve open-class emotion recognition. Audio-text models such as Whisper [6] and AudioCLIP [23] are also used to map speech emotion embeddings to language space to achieve zero-shot emotion transfer. In addition, the generative capabilities of multi-modal large models [64] can transform emotion recognition tasks into natural language generation or question-answering problems, and complete the understanding and prediction of new emotion categories under unsupervised or weak supervision. Combined with strategies such as prompt engineering, instruction tuning, or emotionally informed prompts, the model can generalize to new emotion categories with only label descriptions.

10.6 Multi-modal Conversation Multi-label Emotion Recognition

In multi-modal conversation scenarios, existing emotion recognition models usually use a single-label supervised learning. Due to the ambiguity of emotions, emotion recognition in real life is often a multi-label task. The single-label requirement greatly limits the application scenarios of MCER. Therefore, the multi-label emotion recognition problem in multi-modal conversation scenarios should be considered in future work.

To address the multi-label expression problem in MCER, researchers have proposed a range of targeted solutions to capture the complexity and coexistence of emotions within conversations. A common approach is to replace traditional Softmax classification with Sigmoid activation, enabling the model to independently predict the probability of each emotional label and naturally support the coexistence of multiple emotions such as sadness and anger. In addition, label dependency modeling is widely adopted, where the statistical co-occurrence patterns of emotions are explicitly captured using techniques such as label graphs or GNNs, ensuring semantic consistency and reducing contradictory label outputs. To further improve robustness in real-world noisy environments, uncertainty-aware mechanisms have been integrated into multi-label frameworks, allowing the model to dynamically adjust label confidence based on the reliability of different modalities. Moreover, some studies introduce auxiliary tasks, such as emotion intensity regression or label quantity estimation, to provide richer supervision signals and enhance the model's ability to represent complex emotional states. Recently, researchers have also explored label-specific attention mechanisms, dynamically modulating the contribution of each modality for different emotional labels, which effectively improves the multi-label recognition accuracy under modality imbalance or incomplete scenarios.

10.7 Multi-modal Emotion Recognition in Dynamic Dialogue Scenarios

Real-world conversations are inherently dynamic, with speaker roles changing depending on the context and interpersonal relationships evolving over time. However, most current MCER methods

rely on static modeling frameworks that fail to capture these real-time dynamics. To address this gap, future research should focus on developing models that can adapt to such temporal and structural variability in conversation.

One promising direction is the use of time-series modeling techniques. Methods such as Temporal Convolutional Networks [39, 101] and Hierarchical RNNs [96, 162] can be employed to model frame-level and utterance-level emotional fluctuations across a dialogue sequence. These models can maintain temporal dependencies and detect sudden or gradual emotional transitions by capturing long-range contextual signals. For example, by feeding sequential utterance embeddings into a TCN layer with causal convolutions, the model can learn how earlier statements influence emotional progression. Another critical aspect is the dynamic modeling of speaker interactions. Dynamic GNNs [124, 133, 156] provide a powerful framework to represent evolving dialogue structures. Here, each node in the graph represents a speaker utterance or a speaker entity, and edges represent context-aware relationships (e.g., speaker interactions and turn-taking). Unlike static GCNs, dynamic GNNs update node and edge embeddings over time based on incoming utterances and relational changes. For instance, Temporal Graph Networks and EvolveGCN can learn temporal node representations by integrating recurrent modules that evolve the graph state as the dialogue progresses. These approaches allow the system to adjust for speaker role shifts and emotional influence propagation between participants. Additionally, speaker-role modeling mechanisms can be introduced by using speaker-aware encoders that condition emotion predictions on dynamic role embeddings. For example, using attention modules that incorporate speaker identity, conversational history, and position in the dialogue tree can help model role-specific emotional behaviors. Coupled with adaptive memory modules (e.g., Transformer-based memory networks), the system can retain evolving emotional cues and speaker traits across multiple turns. Finally, hybrid systems that combine reinforcement learning for emotion trend tracking and graph-based relational reasoning can adaptively adjust prediction strategies based on dialogue context evolution. This provides a more robust mechanism for handling emotional ambiguity and inter-speaker dynamics in realistic, multi-turn conversations.

10.8 Lightweight and Efficient MCER

Although most existing MCER models have demonstrated strong emotion recognition performance, their computational complexity and resource requirements hinder deployment in practical applications, especially on resource-constrained platforms such as mobile devices, wearable devices, and embedded systems. In real-world scenarios, real-time emotion recognition is essential for applications such as mobile assistants, edge-based healthcare monitoring, and socially interactive robots. Therefore, designing lightweight and efficient MCER models is critical for practical adoption.

To achieve real-time emotion recognition, the deployment efficiency of the model on resource-constrained devices (such as mobile terminals and embedded systems) must be solved. One of the key paths is to compress and optimize the emotion recognition model so that it can significantly reduce the computational overhead while maintaining high recognition accuracy. First, model compression can transfer the predictive ability of a large model (i.e., teacher model) trained on a high-performance platform to a smaller model (student model) with a more compact structure through methods such as knowledge distillation. During the distillation process, the student model not only learns the original labels of the training data, but also learns the “soft labels” output by the teacher model, that is, the probability distribution of each emotion category. This method allows the small model to learn richer feature representations. The compressed model greatly reduces the number of parameters and inference delay without sacrificing too much accuracy. Secondly, pruning technology mainly evaluates the importance of certain neurons or channels in the network to the final output and prunes the less influential parts. Pruning is usually divided into two

categories: structured pruning and unstructured pruning. Structured pruning can remove the entire convolution kernel, channel or layer. This method is more hardware-friendly and easy to accelerate and parallelize. Unstructured pruning performs sparse processing at the weight level. Although it has more detailed control over precision loss, it requires special sparse matrix acceleration support in actual deployment. Through multiple rounds of pruning and fine-tuning, the model complexity can be further reduced and the inference speed can be improved. Finally, the quantization method converts the original floating-point model parameters and intermediate activation values into a lower bit-width representation, such as compressing from 32-bit floating points to 8-bit integers or even lower bit widths. Quantization can significantly reduce the storage requirements and memory bandwidth consumption of the model, while fully utilizing the integer operation acceleration capabilities of the hardware on many mobile and edge computing devices. In order to avoid the decline in recognition effect due to reduced accuracy, strategies such as quantization-aware training or post-training quantization are usually used to fine-tune the weights and activation distributions to maintain the accuracy of emotion recognition.

11 Conclusion

This article reviews the latest research results in the field of MCER. To allow readers to implement emotion recognition tasks better, we have collected popular datasets in this field and given relevant download links. Since text, video, and audio are unstructured data that cannot be directly input into a computer for computation, we summarize some publicly available feature extraction methods. We divide emotion recognition methods into four categories, i.e., context-free modeling, sequential context modeling, distinguishing speaker modeling, and speaker relationship modeling. This article further discusses the challenges faced by existing methods and future research directions. According to the review of existing work, it is found that multi-modal emotion recognition mainly improves the effect of emotion recognition by modeling intra-modal and inter-modal complementary semantic information. We hope this review can shed some light on developments in this field.

References

- [1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: A visual language model for few-shot learning. In *Proceedings of the Advances in Neural Information Processing Systems*, Vol. 35, 23716–23736.
- [2] Tadas Baltrušaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. In *Proceedings of the 2018 13th IEEE International Conference on Automatic Face and Gesture Recognition (FG '18)*. IEEE, 59–66.
- [3] Anjali Bhavan, Pankaj Chauhan, and Rajiv Ratn Shah. 2019. Bagged support vector machines for emotion recognition from speech. *Knowledge-Based Systems* 184 (2019), 104886.
- [4] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation* 42 (2008), 335–359.
- [5] Erik Cambria, Devamanyu Hazarika, Soujanya Poria, Amir Hussain, and R. B. V. Subramanyam. 2018. Benchmarking multimodal sentiment analysis. In *Proceedings of the 18th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing '17), Revised Selected Papers, Part II*. Springer, 166–179.
- [6] Nan Cao, Yu-Ru Lin, Xiaohua Sun, David Lazer, Shixia Liu, and Huamin Qu. 2012. Whisper: Tracing the spatiotemporal process of information diffusion in real time. *IEEE Transactions on Visualization and Computer Graphics* 18, 12 (2012), 2649–2658.
- [7] Ankush Chatterjee, Umang Gupta, Manoj Kumar Chinnakotla, Radhakrishnan Srikanth, Michel Galley, and Puneet Agrawal. 2019. Understanding emotions in text using deep learning and big data. *Computers in Human Behavior* 93 (2019), 309–317.
- [8] Yi-Chen Chen, Sung-Feng Huang, Hung-Yi Lee, Yu-Hsuan Wang, and Chia-Hao Shen. 2019. Audio Word2vec: Sequence-to-sequence autoencoding for unsupervised learning of audio segmentation and representation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27, 9 (2019), 1481–1493.

- [9] Vishal Chudasama, Purbayan Kar, Ashish Gudmalwar, Nirmesh Shah, Pankaj Wasnik, and Naoyuki Onoe. 2022. M2FNet: Multi-modal fusion network for emotion recognition in conversation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4652–4661.
- [10] Jarosław Cichosz and Krzysztof Slot. 2007. Emotion recognition in speech signal using emotion-extracting binary decision trees. In *Proceedings of Affective Computing and Intelligent Interaction*, 1–10.
- [11] Jiawen Deng and Fuji Ren. 2021. A survey of textual emotion recognition and its challenges. *IEEE Transactions on Affective Computing* 14, 1 (2021), 49–67.
- [12] Yi Ding, Neethu Robinson, Su Zhang, Qiuhaio Zeng, and Cuntai Guan. 2022. TSception: Capturing temporal dynamics and spatial asymmetry from EEG for emotion recognition. *IEEE Transactions on Affective Computing* 14, 3 (2022), 2238–2250.
- [13] Xingbo Du, Junchi Yan, Rui Zhang, and Hongyuan Zha. 2020. Cross-network skip-gram embedding for joint network alignment and link prediction. *IEEE Transactions on Knowledge and Data Engineering* 34, 3 (2020), 1080–1095.
- [14] Sri Harsha Dumpala, Katerina Dikaos, Sebastian Rodriguez, Ross Langley, Sheri Rempel, Rudolf Uher, and Sageev Oore. 2023. Manifestation of depression in speech overlaps with characteristics used to represent and recognize speaker identity. *Scientific Reports* 13, 1 (2023), 11155.
- [15] Chenyuan Feng, Daquan Feng, Guanxin Huang, Zuozhu Liu, Zhenzhong Wang, and Xiang-Gen Xia. 2024. Robust privacy-preserving recommendation systems driven by multimodal federated learning. *IEEE Transactions on Neural Networks and Learning Systems* 36, 5 (2024), 8896–8910.
- [16] Eibe Frank, Leonard Trigg, Geoffrey Holmes, and Ian H. Witten. 2000. Naive Bayes for regression. *Machine Learning* 41 (2000), 5–25.
- [17] Leilei Gan, Zhiyang Teng, Yue Zhang, Linchao Zhu, Fei Wu, and Yi Yang. 2022. SemGloVE: Semantic co-occurrences for glove from BERT. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 30 (2022), 2696–2704.
- [18] Maurice Gerczuk, Shahin Amiriparian, Sandra Ottl, and Bjorn W. Schuller. 2021. EmoNet: A transfer learning framework for multi-corpus speech emotion recognition. *IEEE Transactions on Affective Computing* 14, 2 (2021), 1472–1487.
- [19] Deepanway Ghosal, Md Shad Akhtar, Dushyant Chauhan, Soujanya Poria, Asif Ekbal, and Pushpak Bhattacharyya. 2018. Contextual inter-modal attention for multi-modal sentiment analysis. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 3454–3466.
- [20] Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. COSMIC: COMmonSense knowledge for eMotion Identification in Conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2470–2481.
- [21] Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019. DialogueGCN: A graph convolutional neural network for emotion recognition in conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
- [22] Shreya Ghosh and Tarique Anwar. 2021. Depression intensity estimation via social media: A deep learning approach. *IEEE Transactions on Computational Social Systems* 8, 6 (2021), 1465–1474.
- [23] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. 2022. AudioCLIP: Extending CLIP to image, text and audio. In *Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '22)*. IEEE, 976–980.
- [24] Simon Hanisch, Patricia Arias-Cabarcos, Javier Parra-Arnau, and Thorsten Strufe. 2025. Anonymization techniques for behavioral biometric data: A survey. *ACM Computing Surveys* 57, 11 (2025), 1–54.
- [25] Tanvi Hardeniya and Dilipkumar A. Borikar. 2016. Dictionary based approach to sentiment analysis-a review. *International Journal of Advanced Engineering, Management and Science* 2, 5 (2016), 239438.
- [26] Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, Erik Cambria, and Roger Zimmermann. 2018. ICON: Interactive conversational memory network for multimodal emotion detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2594–2604.
- [27] Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann. 2018. Conversational memory network for emotion recognition in dyadic dialogue videos. In *Proceedings of the Annual Meeting of the North American Chapter of the Association for Computational Linguistics*, Vol. 2018. NIH Public Access, 2122.
- [28] Devamanyu Hazarika, Soujanya Poria, Roger Zimmermann, and Rada Mihalcea. 2021. Conversational transfer learning for emotion recognition. *Information Fusion* 65 (2021), 1–12.
- [29] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. MISA: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*, 1122–1131.
- [30] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In *Proceedings of the Advances in Neural Information Processing Systems*, Vol. 33, 6840–6851.

- [31] Mixiao Hou, Zheng Zhang, Chang Liu, and Guangming Lu. 2023. Semantic alignment network for multi-modal emotion recognition. *IEEE Transactions on Circuits and Systems for Video Technology* 33, 9 (2023), 5318–5329.
- [32] Chao-Chun Hsu, Sheng-Yeh Chen, Chuan-Chun Kuo, Ting-Hao Huang, and Lun-Wei Ku. 2018. EmotionLines: An emotion corpus of multi-party conversations. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC '18)*.
- [33] Dou Hu, Yinan Bao, Lingwei Wei, Wei Zhou, and Songlin Hu. 2023. Supervised adversarial contrastive learning for emotion recognition in conversations. arXiv:2306.01505. Retrieved from <https://arxiv.org/abs/2306.01505>
- [34] Dou Hu, Xiaolong Hou, Lingwei Wei, Lianxin Jiang, and Yang Mo. 2022. MM-DFN: Multimodal dynamic fusion network for emotion recognition in conversations. In *Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '22)*. IEEE, 7037–7041.
- [35] Dou Hu, Lingwei Wei, and Xiaoyong Huai. 2021. DialogueCRN: Contextual reasoning networks for emotion recognition in conversations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, 7042–7052.
- [36] Guimin Hu, Ting-En Lin, Yi Zhao, Guangming Lu, Yuchuan Wu, and Yongbin Li. 2022. UniMSE: Towards unified multimodal sentiment analysis and emotion recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 7837–7851.
- [37] Hao Hu, Ming-Xing Xu, and Wei Wu. 2007. GMM supervector based SVM with spectral features for speech emotion recognition. In *Proceedings of the 2007 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '07)*, Vol. 4. IEEE, IV–413.
- [38] Jian Huang, Jianhua Tao, Bin Liu, Zheng Lian, and Mingyue Niu. 2020. Multimodal transformer fusion for continuous emotion recognition. In *Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '20)*. IEEE, 3507–3511.
- [39] Altaf Hussain, Tanveer Hussain, Waseem Ullah, Samee Ullah Khan, Min Je Kim, Khan Muhammad, Javier Del Ser, and Sung Wook Baik. 2025. Big data analysis for industrial activity recognition using attention-inspired sequential temporal convolution network. *IEEE Transactions on Big Data* 11, 4 (2025), 1840–1851.
- [40] Taichi Ishiwatari, Yuki Yasuda, Taro Miyazaki, and Jun Goto. 2020. Relation-aware graph attention networks with relational position encodings for emotion recognition in conversations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 7360–7370.
- [41] Rahul Islam and Sang Won Bae. 2024. Revolutionizing mental health support: An innovative affective mobile framework for dynamic, proactive, and context-adaptive conversational agents. arXiv:2406.15942. Retrieved from <https://arxiv.org/abs/2406.15942>
- [42] Teja Kattenborn, Jens Leitloff, Felix Schiefer, and Stefan Hinz. 2021. Review on convolutional neural networks (CNN) in vegetation remote sensing. *ISPRS Journal of Photogrammetry and Remote Sensing* 173 (2021), 24–49.
- [43] Smith K. Khare and Varun Bajaj. 2020. Time–frequency representation and convolutional neural network-based emotion recognition. *IEEE Transactions on Neural Networks and Learning Systems* 32, 7 (2020), 2901–2909.
- [44] Byoungjae Kim, Jungyun Seo, and Myoung-Wan Koo. 2021. Randomly wired network based on RoBERTa and dialog history attention for response selection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), 2437–2442.
- [45] Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.
- [46] Dimitrios Kollias and Stefanos Zafeiriou. 2020. Exploiting multi-CNN features in CNN-RNN based dimensional emotion recognition on the omg in-the-wild dataset. *IEEE Transactions on Affective Computing* 12, 3 (2020), 595–606.
- [47] Devesh Kumar, Pavan Kumar V. Patil, Ayush Agarwal, and S. R. Mahadeva Prasanna. 2022. and Fake speech detection using OpenSMILE features. In *Proceedings of the International Conference on Speech and Computer*. Springer, 404–415.
- [48] Mustaqeem and Soonil Kwon. 2021. MLT-DNet: Speech emotion recognition using 1D dilated CNN based on multi-learning trick approach. *Expert Systems with Applications* 167 (2021), 114177.
- [49] Guy Laban, Val Morrison, Arvid Kappas, and Emily S. Cross. 2022. Informal caregivers disclose increasingly more to a social robot over time. In *Proceedings of the CHI Conference on Human Factors in Computing Systems Extended Abstracts*, 1–7.
- [50] Siddique Latif, Rajib Rana, Sara Khalifa, Raja Jurdak, Julien Epps, and Björn W. Schuller. 2020. Multi-task semi-supervised adversarial autoencoding for speech emotion recognition. *IEEE Transactions on Affective Computing* 13, 2 (2020), 992–1004.
- [51] Chi-Chun Lee, Emily Mower, Carlos Busso, Sungbok Lee, and Shrikanth Narayanan. 2011. Emotion recognition using a hierarchical binary decision tree approach. *Speech Communication* 53, 9–10 (2011), 1162–1171.
- [52] Christine P. Lee, Bengisu Cagiltay, and Bilge Mutlu. 2022. The unboxing experience: Exploration and design of initial interactions between children and social robots. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 1–14.

- [53] Hairong Li and Nan Zhang. 2024. Computer vision models for image analysis in advertising research. *Journal of Advertising* 53, 5 (2024), 771–790.
- [54] Jingye Li, Donghong Ji, Fei Li, Meishan Zhang, and Yijiang Liu. 2020. HiTrans: A transformer-based context-and speaker-sensitive model for emotion detection in conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, 4190–4200.
- [55] Jiang Li, Xiaoping Wang, Guo QingLv, and Zhigang Zeng. 2023. GraphCFC: A directed graph based cross-modal feature complementation approach for multimodal conversational emotion recognition. *IEEE Transactions on Multimedia* 26 (2023), 77–89.
- [56] Wei Li, Wei Shao, Shaoxiong Ji, and Erik Cambria. 2022. BiERU: Bidirectional emotional recurrent unit for conversational sentiment analysis. *Neurocomputing* 467 (2022), 73–82.
- [57] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the 8th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 986–995.
- [58] Zaijing Li, Fengxiao Tang, Ming Zhao, and Yusen Zhu. 2022. EmoCaps: Emotion capsule based model for conversational emotion recognition. In *Findings of the Association for Computational Linguistics: ACL 2022*, 1610–1618.
- [59] Ziming Li, Yan Zhou, Weibo Zhang, Yaxin Liu, Chuanpeng Yang, Zheng Lian, and Songlin Hu. 2022. AMOA: Global acoustic feature enhanced modal-order-aware network for multimodal sentiment analysis. In *Proceedings of the 29th International Conference on Computational Linguistics*, 7136–7146.
- [60] Zheng Lian, Lan Chen, Licai Sun, Bin Liu, and Jianhua Tao. 2023. GCNet: Graph completion network for incomplete multimodal learning in conversation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 7 (2023), 8419–8432.
- [61] Zheng Lian, Bin Liu, and Jianhua Tao. 2021. CTNet: Conversational transformer network for emotion recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), 985–1000.
- [62] Yi-Lin Lin and Gang Wei. 2005. Speech emotion recognition based on HMM and SVM. In *Proceedings of the 2005 International Conference on Machine Learning and Cybernetics*, Vol. 8. IEEE, 4898–4901.
- [63] Zijie Lin, Bin Liang, Yunfei Long, Yixue Dang, Min Yang, Min Zhang, and Ruifeng Xu. 2022. Modeling intra-and inter-modal relations: Hierarchical graph contrastive learning for multimodal sentiment analysis. In *Proceedings of the 29th International Conference on Computational Linguistics*, 7124–7135.
- [64] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in Neural Information Processing Systems* 36 (2023), 34892–34916.
- [65] Yang Liu, Haoqin Sun, Wenbo Guan, Yuqi Xia, and Zhen Zhao. 2022. Multi-modal speech emotion recognition using self-attention mechanism and multi-scale fusion framework. *Speech Communication* 139 (2022), 1–9.
- [66] Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh, and Louis-Philippe Morency. 2018. Efficient low-rank multimodal fusion with modality-specific factors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2247–2256.
- [67] Zhen-Tao Liu, Min Wu, Wei-Hua Cao, Jun-Wei Mao, Jian-Ping Xu, and Guan-Zheng Tan. 2018. Speech emotion recognition based on feature selection and extreme learning machine decision tree. *Neurocomputing* 273 (2018), 271–280.
- [68] Reza Lotfian and Carlos Busso. 2019. Curriculum learning for speech emotion recognition from crowdsourced labels. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27, 4 (2019), 815–826.
- [69] Yuanyuan Lu, Zengzhao Chen, Qiuyu Zheng, Yanhui Zhu, and Mengke Wang. 2024. Exploring multimodal data analysis for emotion recognition in teachers’ teaching behavior based on LSTM and MSCNN. *Soft Computing* 28 (2024), 699.
- [70] Jiaxin Ma, Hao Tang, Wei-Long Zheng, and Bao-Liang Lu. 2019. Emotion recognition using multimodal residual LSTM network. In *Proceedings of the 27th ACM International Conference on Multimedia*, 176–183.
- [71] Tinghuai Ma, Qian Pan, Huan Rong, Yurong Qian, Yuan Tian, and Najla Al-Nabhan. 2021. T-BERTSum: Topic-aware text summarization based on BERT. *IEEE Transactions on Computational Social Systems* 9, 3 (2021), 879–890.
- [72] Sijie Mai, Haifeng Hu, and Songlong Xing. 2019. Divide, conquer and combine: Hierarchical feature fusion network with local and global perspectives for multimodal affective computing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 481–492.
- [73] Sijie Mai, Songlong Xing, and Haifeng Hu. 2019. Locally confined modality fusion network with a global perspective for multimodal human affective computing. *IEEE Transactions on Multimedia* 22, 1 (2019), 122–137.
- [74] Sijie Mai, Ying Zeng, Shuangjia Zheng, and Haifeng Hu. 2022. Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis. *IEEE Transactions on Affective Computing* 14, 3 (2022), 2276–2289.
- [75] Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. DialogueRNN: An attentive RNN for emotion detection in conversations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33, 6818–6825.

- [76] Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schroder. 2011. The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing* 3, 1 (2011), 5–17.
- [77] Tao Meng, Yuntao Shou, Wei Ai, Jiayi Du, Haiyan Liu, and Keqin Li. 2024. A multi-message passing framework based on heterogeneous graphs in conversational emotion recognition. *Neurocomputing* 569 (2024), 127109.
- [78] Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi. 2011. Towards multimodal sentiment analysis: Harvesting opinions from the web. In *Proceedings of the 13th International Conference on Multimodal Interfaces*, 169–176.
- [79] Sandeep Kumar Pandey, Hanumant Singh Shekhawat, and S. R. M. Prasanna. 2022. Attention gated tensor neural network architectures for speech emotion recognition. *Biomedical Signal Processing and Control* 71 (2022), 103173.
- [80] Sunghyun Park, Han Suk Shim, Moitreyee Chatterjee, Kenji Sagae, and Louis-Philippe Morency. 2016. Multimodal analysis and prediction of persuasiveness in online social multimedia. *ACM Transactions on Interactive Intelligent Systems* 6, 3 (2016), 1–25.
- [81] Verónica Pérez-Rosas, Rada Mihalcea, and Louis-Philippe Morency. 2013. Utterance-level multimodal sentiment analysis. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 973–982.
- [82] Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabás Póczos. 2019. Found in translation: Learning robust joint representations by cyclic translations between modalities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33, 6892–6899.
- [83] Soujanya Poria, Erik Cambria, and Alexander Gelbukh. 2015. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2539–2544.
- [84] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 873–883.
- [85] Soujanya Poria, Iti Chaturvedi, Erik Cambria, and Amir Hussain. 2016. Convolutional MKL based multimodal emotion recognition and sentiment analysis. In *Proceedings of the 2016 IEEE 16th International Conference on Data Mining (ICDM)*. IEEE, 439–448.
- [86] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.
- [87] Ruiyang Qin, Ryan Cook, Kai Yang, Ahmed Abbasi, David Dobolyi, Salman Seyedi, Emily Griner, Hyeokhyen Kwon, Robert Cotes, Zifan Jiang, et al. 2024. Language models for online depression detection: A review and benchmark analysis on remote interviews. *ACM Transactions on Management Information Systems* 16, 2 (2024), 1–35.
- [88] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*. PMLR, 8748–8763.
- [89] Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, Amir Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. 2020. Integrating multimodal information in large pretrained transformers. In *Proceedings of the Annual Meeting of the North American Chapter of the Association for Computational Linguistics*, Vol. 2020. NIH Public Access, 2359.
- [90] Srividya Tirunellai Rajamani, Kumar T. Rajamani, Adria Mallol-Ragolta, Shuo Liu, and Björn Schuller. 2021. A novel attention-based gated recurrent unit and its efficacy in speech emotion recognition. In *Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '21)*. IEEE, 6294–6298.
- [91] Minjie Ren, Xiangdong Huang, Wenhui Li, Dan Song, and Weizhi Nie. 2021. LR-GCN: Latent relation-aware graph convolutional network for conversational emotion recognition. *IEEE Transactions on Multimedia* 24 (2021), 4422–4432.
- [92] Minjie Ren, Xiangdong Huang, Jing Liu, Ming Liu, Xuanya Li, and An-An Liu. 2023. MALN: Multimodal adversarial learning network for conversational emotion recognition. *IEEE Transactions on Circuits and Systems for Video Technology* 33, 11 (2023), 6965–6980.
- [93] Sakib Anwar Rieyan, Md Raisul Kabir News, A. B. M. Muntasir Rahman, Sadia Afrin Khan, Sultan Tasneem Jawad Zaarif, Md Golam Rabiul Alam, Mohammad Mehedi Hassan, Michele Ianni, and Giancarlo Fortino. 2024. An advanced data fabric architecture leveraging homomorphic encryption and federated learning. *Information Fusion* 102 (2024), 102004.
- [94] Viktor Rozgić, Sankaranarayanan Ananthkrishnan, Shirin Saleem, Rohit Kumar, and Rohit Prasad. 2012. Ensemble of SVM trees for multimodal emotion recognition. In *Proceedings of the 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference*. IEEE, 1–4.

- [95] Stanislaw Saganowski, Bartosz Perz, Adam Polak, and Przemyslaw Kazienko. 2022. Emotion recognition for everyday life using physiological signals from wearables: A systematic literature review. *IEEE Transactions on Affective Computing* 14, 3 (2022), 1876–1897.
- [96] S. Onur Sahin and Suleyman S. Kozat. 2024. Nonlinear regression with hierarchical recurrent neural networks under missing data. *IEEE Transactions on Artificial Intelligence* 5, 10 (2024), 5012–5025.
- [97] Henning Schepker, Florian Denk, Birger Kollmeier, and Simon Doclo. 2020. Acoustic transparency in hearables—perceptual sound quality evaluations. *Journal of the Audio Engineering Society* 68, 7/8 (2020), 495–507.
- [98] Kah Phooi Seng, Li-Minn Ang, and Chien Shing Ooi. 2016. A combined rule-based & machine learning audio-visual emotion recognition approach. *IEEE Transactions on Affective Computing* 9, 1 (2016), 3–13.
- [99] Weizhou Shen, Siyue Wu, Yunyi Yang, and Xiaojun Quan. 2021. Directed acyclic graph network for conversational emotion recognition. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, 1551–1560.
- [100] Dongming Sheng, Dong Wang, Ying Shen, Haitao Zheng, and Haozhuang Liu. 2020. Summarize before aggregate: A global-to-local heterogeneous graph inference network for conversational emotion recognition. In *Proceedings of the 28th International Conference on Computational Linguistics*, 4153–4163.
- [101] Ziyu Sheng, Yuting Cao, Yin Yang, Zhong-Kai Feng, Kaibo Shi, Tingwen Huang, and Shiping Wen. 2024. Residual temporal convolutional network with dual attention mechanism for multilead-time interpretable runoff forecasting. *IEEE Transactions on Neural Networks and Learning Systems* 36, 5 (2024), 8757–8771.
- [102] Yuntao Shou, Wei Ai, and Tao Meng. 2023. Graph information bottleneck for remote sensing segmentation. arXiv:2312.02545. Retrieved from <https://arxiv.org/abs/2312.02545>
- [103] Yuntao Shou, Wei Ai, Tao Meng, and Keqin Li. 2023. CZL-CIAE: CLIP-driven zero-shot learning for correcting inverse age estimation. arXiv:2312.01758. Retrieved from <https://arxiv.org/abs/2312.01758>
- [104] Yuntao Shou, Tao Meng, Wei Ai, Sihan Yang, and Keqin Li. 2022. Conversational emotion recognition studies based on graph convolutional neural networks and a dependent syntactic analysis. *Neurocomputing* 501 (2022), 629–639.
- [105] Lukas Stappen, Alice Baird, Lukas Christ, Lea Schumann, Benjamin Sertolli, Eva-Maria Messner, Erik Cambria, Guoying Zhao, and Björn W. Schuller. 2021. The MuSe 2021 multimodal sentiment analysis challenge: Sentiment, emotion, physiological-emotion, and stress. In *Proceedings of the 2nd on Multimodal Sentiment Analysis Challenge*, 5–14.
- [106] Shubham Suman, Kshira Sagar Sahoo, Chandramouli Das, N. Z. Jhanjhi, and Ambik Mitra. 2022. Visualization of audio files using Librosa. In *Proceedings of 2nd International Conference on Mathematical Modeling and Computational Science (ICMMCS '21)*. Springer, 409–418.
- [107] Guangzhi Sun, Xiao Zhan, and Jose Such. 2024. Building better AI agents: A provocation on the utilisation of persona in LLM-based conversational agents. In *Proceedings of the 6th ACM Conference on Conversational User Interfaces*, 1–6.
- [108] Yang Sun, Nan Yu, and Guohong Fu. 2021. A discourse-aware graph neural network for emotion recognition in multi-party conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2949–2958.
- [109] Zhongkai Sun, Prathusha Sarma, William Sethares, and Yingyu Liang. 2020. 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, 8992–8999.
- [110] Zhe Sun, Hehao Zhang, Jiatong Bai, Mingyang Liu, and Zhengping Hu. 2023. A discriminatively deep fusion approach with improved conditional GAN (im-cGAN) for facial expression recognition. *Pattern Recognition* 135 (2023), 109157.
- [111] Liang Tan, Keping Yu, Long Lin, Xiaofan Cheng, Gautam Srivastava, Jerry Chun-Wei Lin, and Wei Wei. 2021. Speech emotion recognition enhanced traffic efficiency solution for autonomous vehicles in a 5G-enabled space-air-ground integrated intelligent transportation system. *IEEE Transactions on Intelligent Transportation Systems* 23, 3 (2021), 2830–2842.
- [112] Fei Tao and Gang Liu. 2018. Advanced LSTM: A study about better time dependency modeling in emotion recognition. In *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2906–2910.
- [113] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3D convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, 4489–4497.
- [114] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the Annual Meeting of the North American Chapter of the Association for Computational Linguistics*, Vol. 2019. NIH Public Access, 6558.
- [115] Geng Tu, Jintao Wen, Cheng Liu, Dazhi Jiang, and Erik Cambria. 2022. Context-and sentiment-aware networks for emotion recognition in conversation. *IEEE Transactions on Artificial Intelligence* 3, 5 (2022), 699–708.

- [116] Efthymios Tzinis, Georgios Paraskevopoulos, Christos Baziotis, and Alexandros Potamianos. 2018. Integrating recurrence dynamics for speech emotion recognition. In *Proceedings of the Interspeech 2018*, 927–931.
- [117] Bingtao Wan, Peng Wu, Chai Kiat Yeo, and Gang Li. 2024. Emotion-cognitive reasoning integrated BERT for sentiment analysis of online public opinions on emergencies. *Information Processing & Management* 61, 2 (2024), 103609.
- [118] Haohan Wang, Aaksha Meghawat, Louis-Philippe Morency, and Eric P. Xing. 2017. Select-additive learning: Improving generalization in multimodal sentiment analysis. In *Proceedings of the 2017 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 949–954.
- [119] Ning Wang, Hui Cao, Jun Zhao, Ruilin Chen, Dapeng Yan, and Jie Zhang. 2022. M2R2: Missing-modality robust emotion recognition framework with iterative data augmentation. *IEEE Transactions on Artificial Intelligence* 4, 5 (2022), 1305–1316.
- [120] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros. 2020. CNN-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8695–8704.
- [121] Xin Wang, Hong Chen, Zirui Pan, Yuwei Zhou, Chaoyu Guan, Lifeng Sun, and Wenwu Zhu. 2025. Automated disentangled sequential recommendation with large language models. *ACM Transactions on Information Systems* 43, 2 (2025), 1–29.
- [122] Yansen Wang, Ying Shen, Zhun Liu, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2019. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33, 7216–7223.
- [123] Yan Wang, Jiayu Zhang, Jun Ma, Shaojun Wang, and Jing Xiao. 2020. Contextualized emotion recognition in conversation as sequence tagging. In *Proceedings of the 21st Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 186–195.
- [124] Ye Wang, Wei Zhang, Ke Liu, Wei Wu, Feng Hu, Hong Yu, and Guoyin Wang. 2025. Dynamic emotion-dependent network with relational subgraph interaction for multimodal emotion recognition. *IEEE Transactions on Affective Computing* 16, 2 (2025), 712–725.
- [125] Zhiyuan Wen, Jiannong Cao, Jiaying Shen, Ruosong Yang, Shuaiqi Liu, and Maosong Sun. 2024. Personality-affected emotion generation in dialog systems. *ACM Transactions on Information Systems* 42, 5 (2024), 1–27.
- [126] Yang Wu, Yanyan Zhao, Hao Yang, Song Chen, Bing Qin, Xiaohuan Cao, and Wenting Zhao. 2022. Sentiment word aware multimodal refinement for multimodal sentiment analysis with ASR errors. In *Findings of the Association for Computational Linguistics: ACL 2022*, 1397–1406.
- [127] Yue Xie, Ruiyu Liang, Zhenlin Liang, Chengwei Huang, Cairong Zou, and Björn Schuller. 2019. Speech emotion classification using attention-based LSTM. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27, 11 (2019), 1675–1685.
- [128] Frank Xing. 2025. Designing heterogeneous LLM agents for financial sentiment analysis. *ACM Transactions on Management Information Systems* 16, 1 (2025), 1–24.
- [129] Songlong Xing, Sijie Mai, and Haifeng Hu. 2020. Adapted dynamic memory network for emotion recognition in conversation. *IEEE Transactions on Affective Computing* 13, 3 (2020), 1426–1439.
- [130] Honghui Xu, Wei Li, Daniel Takabi, Daehee Seo, and Zhipeng Cai. 2025. Privacy-preserving multimodal sentiment analysis. *IEEE Internet of Things Journal* 12, 11 (2025), 15467–15478.
- [131] Dingkang Yang, Shuai Huang, Yang Liu, and Lihua Zhang. 2022. Contextual and cross-modal interaction for multimodal speech emotion recognition. *IEEE Signal Processing Letters* 29 (2022), 2093–2097.
- [132] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023. 2023. The dawn of LLMs: Preliminary explorations with GPT-4V(ision). arXiv:2309.17421. Retrieved from <https://arxiv.org/abs/2309.17421>
- [133] Zhenyu Yang, Xiaoyang Li, Yuhu Cheng, Tong Zhang, and Xuesong Wang. 2024. Emotion recognition in conversation based on a dynamic complementary graph convolutional network. *IEEE Transactions on Affective Computing* 15, 3 (2024), 1567–1579.
- [134] Mang Ye, Wei Shen, Junwu Zhang, Yao Yang, and Bo Du. 2024. SecureReID: Privacy-preserving anonymization for person re-identification. *IEEE Transactions on Information Forensics and Security* 19 (2024), 2840–2853.
- [135] Ziyi Ye, Xiaohui Xie, Qingyao Ai, Yiqun Liu, Zhihong Wang, Weihang Su, and Min Zhang. 2024. Relevance feedback with brain signals. *ACM Transactions on Information Systems* 42, 4 (2024), 1–37.
- [136] Lihua Yin, Sixin Lin, Zhe Sun, Simin Wang, Ran Li, and Yuanyuan He. 2024. PriMonitor: An adaptive tuning privacy-preserving approach for multimodal emotion detection. *World Wide Web* 27, 2 (2024), 9.
- [137] Nan Yin, Li Shen, Baopu Li, Mengzhu Wang, Xiao Luo, Chong Chen, Zhigang Luo, and Xian-Sheng Hua. 2022. DEAL: An unsupervised domain adaptive framework for graph-level classification. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22)*. ACM, New York, NY, 3470–3479. DOI: <https://doi.org/10.1145/3503161.3548012>

- [138] Nan Yin, Li Shen, Mengzhu Wang, Long Lan, Zeyu Ma, Chong Chen, Xian-Sheng Hua, and Xiao Luo. 2023. CoCo: A coupled contrastive framework for unsupervised domain adaptive graph classification. arXiv:2306.04979. Retrieved from <https://arxiv.org/abs/2306.04979>
- [139] Nan Yin, Li Shen, Mengzhu Wang, Xiao Luo, Zhigang Luo, and Dacheng Tao. 2023. OMG: Towards effective graph classification against label noise. *IEEE Transactions on Knowledge and Data Engineering* 35, 12 (2023), 12873–12886. DOI: <https://doi.org/10.1109/TKDE.2023.3271677>
- [140] N. Yin, L. Shen, H. Xiong, B. Gu, C. Chen, X. Hua, S. Liu, and X. Luo. 2024. Messages are never propagated alone: Collaborative hypergraph neural network for time-series forecasting. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46, 4 (Nov. 2024), 2333–2347. DOI: <https://doi.org/10.1109/TPAMI.2023.3331389>
- [141] RunKai Ying, Yuntao Shou, and Chang Liu. 2021. Prediction model of Dow Jones index based on LSTM-Adaboost. In *Proceedings of the 2021 International Conference on Communications, Information System and Computer Engineering (CISCE)*. IEEE, 808–812.
- [142] Liu Yingjian, Li Jiang, Wang Xiaoping, and Zeng Zhigang. 2023. EmotionIC: Emotional inertia and contagion-driven dependency modelling for emotion recognition in conversation. arXiv:2303.11117. Retrieved from <https://arxiv.org/abs/2303.11117>
- [143] Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. 2020. CH-SIMS: A Chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. ACL, 3718–3727. DOI: <https://doi.org/10.18653/v1/2020.acl-main.343>
- [144] Ziqi Yuan, Yihe Liu, Hua Xu, and Kai Gao. 2023. Noise imitation based adversarial training for robust multimodal sentiment analysis. *IEEE Transactions on Multimedia* 26 (2023), 529–539.
- [145] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 1103–1114.
- [146] Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Memory fusion network for multi-view sequential learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.
- [147] Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. 2018. Multi-attention recurrent network for human communication comprehension. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.
- [148] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2236–2246.
- [149] Sayyed M. Zahrir and Jinho D. Choi. 2017. Emotion detection on TV show transcripts with sequence-based convolutional neural networks. arXiv:1708.04299. Retrieved from <https://arxiv.org/abs/1708.04299>
- [150] An Zhang, Yuxin Chen, Leheng Sheng, Xiang Wang, and Tat-Seng Chua. 2024. On generative agents in recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1807–1817.
- [151] Haidong Zhang and Yekun Chai. 2021. COIN: Conversational interactive networks for emotion recognition in conversation. In *Proceedings of the 3rd Workshop on Multimodal Artificial Intelligence*, 12–18.
- [152] Jianhua Zhang, Zhong Yin, Peng Chen, and Stefano Nichele. 2020. Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review. *Information Fusion* 59 (2020), 103–126.
- [153] Ke Zhang, Yuanqing Li, Jingyu Wang, Erik Cambria, and Xuelong Li. 2021. Real-time video emotion recognition based on reinforcement learning and domain knowledge. *IEEE Transactions on Circuits and Systems for Video Technology* 32, 3 (2021), 1034–1047.
- [154] Qiongan Zhang, Lei Shi, Peiyu Liu, Zhenfang Zhu, and Liancheng Xu. 2023. ICDN: Integrating consistency and difference networks by transformer for multimodal sentiment analysis. *Applied Intelligence* 53, 12 (2023), 16332–16345.
- [155] Shiqing Zhang, Yijiao Yang, Chen Chen, Xingnan Zhang, Qingming Leng, and Xiaoming Zhao. 2024. Deep learning-based multimodal emotion recognition from audio, visual, and text modalities: A systematic review of recent advancements and future prospects. *Expert Systems with Applications* 237 (2024), 121692.
- [156] Tong Zhang, Yelin Chen, Shuzhen Li, Xiping Hu, and C. L. Philip Chen. 2024. TT-GCN: Temporal-tightly graph convolutional network for emotion recognition from gaits. *IEEE Transactions on Computational Social Systems* 11, 3 (2024), 4300–4314.
- [157] Yazhou Zhang, Ao Jia, Bo Wang, Peng Zhang, Dongming Zhao, Pu Li, Yuexian Hou, Xiaojia Jin, Dawei Song, and Jing Qin. 2023. M3GAT: A multi-modal multi-task interactive graph attention network for conversational sentiment analysis and emotion recognition. *ACM Transactions on Information Systems* 42, 1 (2023), 1–32.

- [158] Yabin Zhang, Zhenlei Wang, Wenhui Yu, Lantao Hu, Peng Jiang, Kun Gai, and Xu Chen. 2024. Soft contrastive sequential recommendation. *ACM Transactions on Information Systems* 42, 6 (2024), 1–28.
- [159] Yancheng Zhang, Mengxin Zheng, Yuzhang Shang, Xun Chen, and Qian Lou. 2024. HEPPrune: Fast private training of deep neural networks with encrypted data pruning. In *Proceedings of the Advances in Neural Information Processing Systems*, Vol. 37, 51063–51084.
- [160] Zixin Zhang, Fan Qi, and Changsheng Xu. 2024. Enhancing storage and computational efficiency in federated multimodal learning for large-scale models. In *Proceedings of the 41st International Conference on Machine Learning*.
- [161] Peixiang Zhong, Di Wang, and Chunyan Miao. 2019. Knowledge-enriched transformer for emotion detection in textual conversations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 165–176.
- [162] Kun Zhou, Sung-Kwon Oh, Jianlong Qiu, Witold Pedrycz, Kisung Seo, and Jin Hee Yoon. 2024. Design of hierarchical neural networks using deep LSTM and self-organizing dynamical fuzzy-neural network architecture. *IEEE Transactions on Fuzzy Systems* 32, 5 (2024), 2915–2929.
- [163] Lixing Zhu, Gabriele Pergola, Lin Gui, Deyu Zhou, and Yulan He. 2021. Topic-driven and knowledge-aware transformer for dialogue emotion detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 1571–1582.
- [164] Linan Zhu, Zhechao Zhu, Chenwei Zhang, Yifei Xu, and Xiangjie Kong. 2023. Multimodal sentiment analysis based on fusion methods: A survey. *Information Fusion* 95 (2023), 306–325.

Received 16 December 2024; revised 25 June 2026; accepted 9 September 2025