融合大语言模型与跨域图结构的多模态对话情感识别方法

黄 辰[1,2,3,4],马浩博[1,2,3,4],张 龑[1,2,3,4*],杨 超[1,2,3,4],宋建华[2,3,4,5]

(1. 湖北大学计算机学院,湖北武汉 430062;2. 智能感知系统与安全教育部重点实验室,湖北武汉 430062;
3. 大数据智能分析与行业应用湖北省重点实验室,湖北武汉 430062;4. 湖北省高校人文社科重点研究基地-绩效评价信息管理研究中心,湖北武汉 430062;5. 湖北大学网络空间安全学院,湖北武汉 430062)

摘 要: 多模态对话情感识别(Multimodal Emotion Recognition in Conversation, MERC)通过融合文本、语音、视觉等多模态信息来识别对话中的情感状态。随着对话式人工智能和情感计算的快速发展,MERC 成为情感计算和人机交互领域的研究热点。相比传统单一模态情感识别,多模态方法能够更全面、精确地捕捉情感的多维特征,如文本传递显性情感内容,语音提供音调、语速等隐性情感线索,视觉信息(如面部表情)则反映情感的非语言表现。这些模态信息相互补充,有助于提高情感识别的准确性和鲁棒性。然而,多模态情感识别面临诸多挑战:首先,不同模态的数据在信息表示上存在显著差异,传统的特征拼接或加权平均方法无法充分捕捉模态间复杂的交互关系,容易导致信息丢失;其次,情感识别任务常常遭遇局部噪声和离群样本干扰,影响模型稳定性;最后,情感识别的准确性与对话上下文的综合利用密切相关,情感往往受到前后文的影响,因此,如何有效提取和利用上下文信息是提高准确性的一大挑战。为应对这些问题,本文结合大语言模型(Large Language Model,LLM)与全局-局部跨域图结构,提出了 LLM-EmoGraph 方法,旨在实现多模态数据的精确融合与高效建模。该方法引入多模态掩码机制来处理不同模态之间的缺失和不一致信息,确保模型在信息不完整时依然保持较好性能。通过大规模跨域多图预训练,LLM-EmoGraph 提升了多模态间及图结构间的迁移能力,增强了模型的鲁棒性。其创新的自适应双尺度特征融合策略实现了文本、语音和视觉信息的高效对齐,提升了情感识别精度,尤其在多模态高度交互的情境下表现优异。此外,结合大语言模型的弱监督层次化情感分类方案,通过逐层引导情感信息提取,有效避免了全局情感模式的干扰,使得即使在有限标注数据下,模型也能准确学习情感特征。实验结果表明,LLM-EmoGraph 在多个基准数据集上显著超越现有主流方法,验证了其在多模态情感识别中的有效性和先进性。总体而言,LLM-EmoGraph 通过创新的多模态融合策略、大规模预训练和弱监督学习方法,解决了多模态情感识别中的一系列问题,为提升情感识别系统的准确性和稳定性提供了有力支持。

关键词: 多模态情感识别;对话系统;大语言模型;图神经网络;特征融合

基金项目: 湖北省重大攻关项目(No.2023BAA018);湖北省科技计划重大科技专项(No.2024BAA008)

中图分类号: TP391;TP399 **文献标识码:** A **文章编号:** 0372-2112(2026)01-0340-12

电子学报 URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20250772

Fusing Large Language Models with Cross-Domain Graphs for Multimodal Emotion Recognition in Dialogue

HUANG Chen[1,2,3,4], MA Haobo[1,2,3,4], ZHANG Yan[1,2,3,4*], YANG Chao[1,2,3,4], SONG Jianhua[2,3,4,5]

(1. *School of Computer Science*, *Hubei University*, *Wuhan*, *Hubei* 430062, *China*;
2. *Key Laboratory of Intelligent Sensing System and Security*, *Ministry of Education*, *Wuhan*, *Hubei* 430062, *China*;
3. *Hubei Key Laboratory of Big Data Intelligent Analysis and Application*, *Wuhan*, *Hubei* 430062, *China*;
4. *Hubei Province Project of Key Research Institute of Humanities and Social Sciences at Universities*, *Wuhan*, *Hubei* 430062, *China*;
5. *School of Cybersecurity*, *Hubei University*, *Wuhan*, *Hubei* 430062, *China*)

Abstract: Multimodal emotion recognition in conversation (MERC) refers to the identification of emotional states in conversations by integrating various modalities such as text, speech, and visual information. With the rapid development of conversational AI and affective computing, MERC has become a research hotspot in the fields of affective computing and human-computer interaction. Compared to traditional unimodal emotion recognition, multimodal approaches can capture the multifaceted characteristics of emotions more comprehensively and accurately. For instance, text conveys explicit emotional content, speech provides subtle emotional cues like tone, speed, and intonation, while visual information (such as fa-

收稿日期:2025-09-05;录用日期:2026-01-08;责任编辑:覃怀银

*通信作者:张龑

cial expressions) reflects non-verbal emotional expressions. These multimodal signals complement each other, enhancing the accuracy and robustness of emotion recognition. However, multimodal emotion recognition faces several challenges. First, there are significant differences in the representation of information across different modalities, and traditional methods like feature concatenation or weighted averaging fail to fully capture the complex interactions between modalities, which can lead to information loss. Second, emotion recognition tasks often suffer from local noise and outlier samples, which can degrade model stability. Lastly, the accuracy of emotion recognition is closely tied to the effective use of contextual information in a conversation, as emotions are often influenced by preceding and succeeding dialogue. Thus, how to effectively extract and utilize contextual information becomes a major challenge in improving accuracy. To address these issues, this paper proposes a novel emotion recognition method, LLM-EmoGraph, which combines large language model (LLM) with global-local cross-domain graph structures to achieve precise fusion and efficient modeling of multimodal data. This method introduces a multimodal masking mechanism to handle missing and inconsistent information across modalities, ensuring that the model maintains good performance even with incomplete or low-quality information. Through large-scale cross-domain multi-graph pretraining, LLM-EmoGraph enhances the model's transferability between modalities and graph structures, further improving its robustness. The innovative adaptive dual-scale feature fusion strategy aligns textual, speech, and visual semantic features efficiently, improving emotion recognition accuracy, particularly in scenarios involving high interaction among modalities. Additionally, the paper designs a weakly supervised hierarchical emotion classification scheme based on LLM. This approach guides the extraction of emotional information layer by layer, effectively preventing interference from global emotional patterns, and allows the model to learn emotional features accurately, even with limited annotated data. Experimental results show that LLM-EmoGraph significantly outperforms existing mainstream methods on multiple benchmark datasets, demonstrating its effectiveness and advancement in multimodal emotion recognition tasks. In summary, LLM-EmoGraph, through its innovative multimodal fusion strategies, large-scale pretraining, and weakly supervised learning methods, provides effective solutions to a series of challenges in multimodal emotion recognition, offering strong support for improving the accuracy and stability of emotion recognition systems.

Keywords: multimodal emotion recognition; dialogue system; large language model; graph neural network; feature fusion

Foundation Item(s): Major Project of Hubei Province (JD) (No.2023BAA018); Major Science and Technology Special Project of Hubei Science and Technology Plan (No.2024BAA008)

0　引言

情感在社会交往中发挥关键作用,本质上具有动态性和交互性,能够深刻地反映人类的内在心理活动,是理解决策行为的重要参考[1]。随着人工智能技术的发展,情感识别技术[2-4]在多个领域展现了广泛的应用前景,特别是在现实人机交互场景中具有重要意义。为了实现更具反馈性、自然个性化且富有共情能力的交互体验,智能系统需要具备洞察和识别用户情绪状态的能力,从而获得更佳效果。

对话情感识别(Emotion Recognition in Conversation,ERC)旨在自动检测对话中的情绪,随着深度学习技术的发展,已广泛应用于情感分析、推荐系统、医疗诊断等多个场景[5-7]。由于对话数据通常具有时间序列特性,近年来,许多ERC方法采用了序列建模机制[8]。然而,单一模态(如文本)的识别方法逐渐暴露出局限性,因此,融合语音、文本和面部表情等多模态数据已成为提升情感识别准确率的关键手段[9]。多模态情感识别利用不同模态间的互补信息,显著提升了识别性能。

然而,复杂的情绪依赖关系仍然是挑战,尤其是在多方对话中。情感状态不仅受到说话人情绪的驱动,还受参与者互动的影响,增加了识别难度[10]。图神经网络(Graph Neural Network,GNN)在建模复杂关系和上下文依赖方面表现出优势,因此被广泛应用于情感识别任务。相比传统的递归神经网络,GNN擅长捕捉长距离上下文关系,适用于多模态数据与复杂情绪依赖共存的场景。例如,ConGCN[11]采用异构图建模语境与说话人依赖,I-GCN[12]融合话语级与说话人级图卷积网络,COGMEN[13]结合R-GCN与图Transformer建模自我与人际依赖,MMGCN[14]构建大型异构图以整合多模态特征。近期的优化研究,如GraphMFT[15]通过增强GAT结构提高依赖建模能力,GraphCFC[16]优化特征一致性与多样性,GA2MIF[17]采用多源信息融合策略,更有效地挖掘模态间的互补信息。尽管图结构方法在上下文传播方面取得了显著进展,当前仍存在以下主要问题:其一,现有方法通常仅适用于特定模态组合,缺乏跨模态迁移能力,难以在无需重新训练或微调的条件下实现模型泛化;其二,由于多模态情感识别模型的训练样本不足,伪标签的选择不够准确,导致不同模态中的情感特征与真

实标签之间相关性较低，从而影响整体识别性能。

为此，本文提出了一种新型多模态情感识别框架LLM-EmoGraph，该方法基于全局-局部跨域图结构，并引入大语言模型（Large Language Model，LLM）以增强建模能力。具体而言，LLM-EmoGraph设计了多模态掩码策略，并采用大规模跨域多图预训练算法，以实现异构图域和多模态间的有效迁移学习。同时，框架还引入了自适应双尺度特征融合策略，以确保模态间语义一致性，从而更好地整合来自文本、语音与视觉的情绪信息。最后，LLM-EmoGraph构建了一套由大语言模型增强的弱监督层次化情感识别机制，旨在提升伪标签的可靠性，进一步增强识别系统的鲁棒性与准确性。

1 相关工作

在本节中，主要回顾了一些关于多模态表示与图学习、多模态融合的相关研究工作。

1.1 多模态表示与图学习

近年来，面向多模态情感识别任务，构建通用的多模态表示学习模型逐渐成为研究热点。研究者致力于在视觉、语言和音频等异构模态之间实现统一的特征对齐与情感语义建模。早期的代表性进展主要体现在图文预训练（Vision-Language Pre-training，VLP）模型的发展，该类方法主要关注图像与文本对的建模，并通过对比学习与掩码语言建模等任务实现跨模态语义对齐。典型代表包括CLIP[18]与ALIGN[19]，它们充分展示了大规模预训练在跨模态理解中的潜力。随着统一架构（ViT[20]与Perceiver IO[21]）的提出以及多样化预训练策略（如BEiT[22]和MAE[23]）的应用，近期研究进一步拓展了跨模态语义对齐的边界，从传统的图文配对延伸至音频与视频模态[24]，为更加全面的多模态情感表示学习奠定了基础。

在上述研究趋势的推动下，多模态图学习作为一种建模多模态间复杂关系与上下文交互的重要范式，逐渐受到关注。然而，当前大多数多模态图学习方法仍集中于特定领域，如知识图谱[25-26]、自然科学中的分子[27]和脑连接图[28]等。这类方法通常依赖领域先验知识，并服务于特定的任务目标，因此在模态泛化性和应用场景的可扩展性方面存在一定局限。此外，这些方法多侧重于图中节点的结构连接关系，而较少关注如何在模态间学习统一的情感语义表示，导致其难以适应来自多模态场景的异构图数据结构。

相较而言，MMGL[29]的提出为多模态图结构中引入不同模态的大型基础模型提供了新的思路，支持面向情感的生成类任务。然而，MMGL方法主要聚焦于生成应用，并未明显解决跨模态语义融合与对齐等在情感识别中亟待突破的核心问题。

1.2 多模态融合

多模态融合是多模态情感识别的核心，关键在于高效整合来自文本、语音和视觉等模态的信息。由于模态间在时间分辨率、特征分布和语义表示上的差异[30]，如何克服这些差异以构建统一且具判别性的情感表示，是提升情感预测系统准确性与鲁棒性的关键。为应对这些挑战，研究者提出了多种融合策略。例如，MIA-Net引入多模态交互注意力机制，选择性增强关键信息以提升语义表示能力[31]；MCWSA-CMHA结合多头注意力和深度神经网络，精细建模模态间依赖关系[32]；GraphCFC通过子空间特征提取器和跨模态互补对齐策略，缩小模态间的特征差异[33]；MM-DFN利用门控图卷积机制聚合上下文信号，减少冗余信息[34]；MultiEMO引入双向多头注意力机制，促进语义交互与深层次建模[35]。

尽管现有研究在特征融合方面已取得显著进展，但多数方法仍未充分考虑情绪依赖关系以及模态间动态说话人交互等因素，限制了对细微情绪表达的感知能力，从而影响整体识别性能。

2 模型构建

本文提出的多模态情感识别框架LLM-EmoGraph主要由三个关键模块构成：多模态掩码策略、自适应双尺度特征融合模块以及LLM增强的层次化学习模块。图1展示了LLM-EmoGraph的整体流程与结构示意图。

2.1 问题定义

为有效建模多模态特征间的复杂交互关系，以及话语与说话人之间的上下文依赖关系，本文将情感识别中的多模态信息形式化为多模态图结构（MultiModal Graph，MMG）。该图结构定义为 $G=(V,E,M,W)$，其中，V表示图中的节点集合，E表示边的集合，$M\to 2^{W}$是一个映射函数，用于将每个节点$v\in V$映射到一个模态子集$W_v\in W$。W表示所有可能模态的集合，如文本、音频、图像等。每个节点v可包含多个来自不同模态的特征，但并不要求所有节点都具备所有模态的数据。

在特定的多模态属性图（Multimodal Attributed Graph，MAG）$G_{\mathrm{TAG}}=(V,E,M,\mathrm{text},\mathrm{audio},\mathrm{image})$中，每个节点均包含文本$t_v\in T_v$、音频$a_v\in A_v$及图像特征$i_v\in I_v$。我们定义MMG的模态映射函数如式（1）所示：

$$M(v)=\{\mathrm{text},\mathrm{audio},\mathrm{image}\},\quad \text{for all } v\in V \tag{1}$$

如图2所示，来自不同来源的多模态特征首先被投射到一个共享的嵌入空间中，随后进行结构化建图

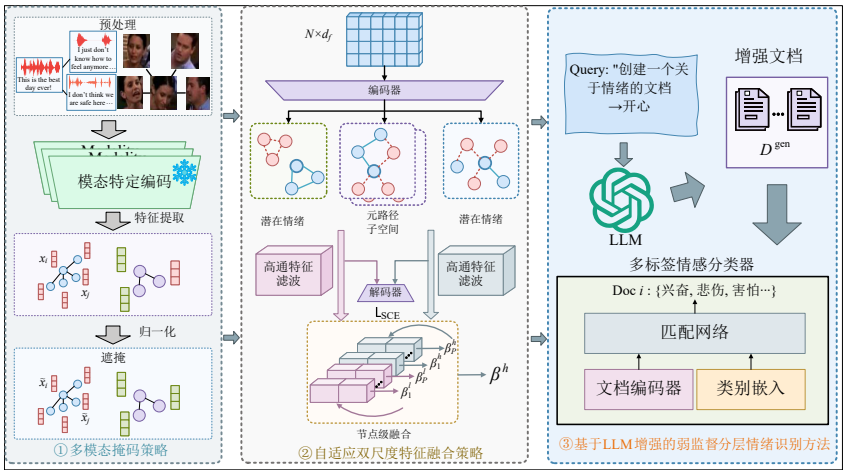


图 1　所提出的 LLM-EmoGraph 框架的整体架构

Figure 1　The overall architecture of the proposed LLM-EmoGraph framework

并用于后续情感识别等下游任务。在该建模框架下，$W=\{\text{text, audio, image}\}$ 表示系统所覆盖的模态集合。

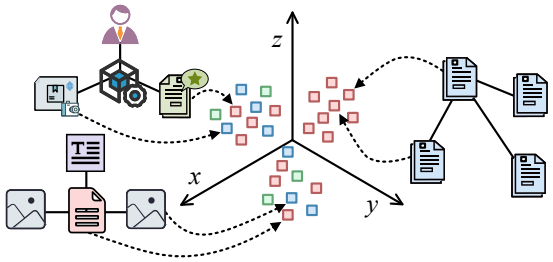


图 2　统一多模态图以支持多样化下游任务

Figure 2　Unifying multimodal graphs for diverse downstream tasks

2.2　多模态掩码策略

在多模态情绪识别的背景下，掩蔽策略对于多模态图起着至关重要的作用。目标是掩蔽部分多模态特征，并要求模型重建这些特征，从而促使模型有效地捕捉情绪线索和多模态信息，如音频、视觉和文本数据，这对于准确的情绪识别至关重要。

模态特定编码。在应用掩蔽过程之前，使用模态特定编码器将来自不同模态的原始数据映射为特征向量。在多模态图 $G=(V,E,M)$ 的背景下，其中每个节点 $v\in V$ 可以具有来自模态子集 $W_v\subseteq W$ 的特征，原始特征通过针对每种模态的编码器进行转换(如文本使用语言模型，图像使用视觉变换器)。

设 E_ω 表示模态 $\omega\in W$ 的编码器，且 $\boldsymbol{x}_i^{(\omega)}\in\mathbb{R}^{d_m}$ 表示通过模态 ω 获得的节点 v_i 的特征向量。模态特定编码可以表示为

$$\boldsymbol{x}_i^{(\omega)}=E_\omega\left(v_i^{(\omega)}\right) \tag{2}$$

在多模态情绪识别的背景下，节点 v_i 的特征 $\boldsymbol{x}_i\in\mathbb{R}^{d_m}$ 是通过对与该节点相关的所有模态 W_v 的特征进行平均得到的，即

$$\boldsymbol{x}_i=\frac{1}{|W_v|}\sum_{\omega\in W_v}\boldsymbol{x}_i^{(\omega)} \tag{3}$$

其中，$\boldsymbol{x}_i^{(\omega)}$ 表示从模态 ω 获得的节点 v_i 的特征向量；W_v 是与节点 v_i 相关的模态子集。通过这种方式，模型可以有效地结合来自不同模态(如音频、视觉和文本)的信息，从而增强情绪识别的准确性。

节点特征掩码处理。在模态特定编码器将原始数据转换为统一的特征向量后，LLM-EmoGraph 进一步引入了特征掩码策略，以提升模型的泛化能力和鲁棒性。具体而言，从图中的节点集合 V 中随机抽取一个子集 S，采用不放回均匀采样的方式选取约 75% 的节点，并将这些节点的原始特征替换为一个可学习的掩码向量。该比例的选择是通过实验验证得到的。在实验中，我们尝试了不同的节点采样比例(如 50%、75%、90% 等)，并通过交叉验证评估了模型在情绪识别任务中的表现。结果表明，当约 75% 的节点被掩码时，模型的性能达到最佳，表现出更好的鲁棒性和泛化能力。具体实验内容如表 1 所示。

表 1　节点比例选取

Table 1　Node sampling ratio

采样比例/%	加权准确率/%	加权F1分数	模型鲁棒性	泛化能力
50	82.3	0.79	较差	较差
75	86.7	0.83	较好	较好
90	84.5	0.81	一般	一般

这一掩码机制模拟了现实中模态信息不完整或噪声干扰的情况，促使模型依赖图结构和上下文信息进行情感推理，从而增强其对未知或缺失模态的适应能力。

对于任意节点 $v_i\in V_i$，其被掩码后的特征表示为

$$\tilde{\boldsymbol{x}}_i=\begin{cases}\boldsymbol{x}^{[M]}, \text{if } v_i\in\tilde{V}\\ \boldsymbol{x}_i, \quad \text{otherwise}\end{cases} \tag{4}$$

所有经过掩码处理的特征向量$\tilde{\boldsymbol{x}}_i$随后被输入至自适应双尺度特征融合模块(Adaptive Dual-Scale Feature Fusion,ADSFF)。该模块设计旨在全局与局部两个尺度上,自适应地整合来自不同模态与图域的信息。在局部尺度上,ADSFF关注节点邻域内的短程依赖与模态间的细粒度交互;而在全局尺度上,模块则捕捉跨模态、跨说话人之间的长程依赖与情感传播关系,从而实现更丰富和结构化的语义融合。这一融合机制显著增强了情绪表示的语义完整性和判别性,有效提升了模型在模态不完整、特征缺失或噪声干扰等复杂情况下的推理能力与情感识别鲁棒性。

2.3 自适应双尺度特征融合策略

在多模态情绪识别的背景下,由于不同模态之间存在大量异质结构和复杂的关系,我们引入高通特征滤波,以增强相邻模态之间特征的区分能力,从而避免不同模态区域的特征混淆。相比于传统依赖监督信号进行复杂滤波器学习的方法,这一设计更适用于多模态情绪识别中的无监督场景。为此,我们提出了一种简单且具有可解释性的机制:采用$\tilde{A}^{\phi_p}_{\text{sym}}$作为低通滤波器,用于提取低频信息并保留模态间的共性;同时,采用$\tilde{L}^{\phi_p}_{\text{sym}}$作为高通滤波器,用于提取高频信息并增强情绪特征的表达。上述图滤波操作被应用于情绪特征嵌入上。在多模态情绪识别的背景下,由于不同模态之间存在大量异质结构和复杂的关系,我们引入了高通滤波器和低通滤波器,用于提取低频与高频特征。式(5)中,低频特征和高频特征的提取涉及对每个模态的邻域结构的处理,具体来说,邻域结构的确定是通过图卷积或K近邻算法(k-NN)来实现的。在本研究中,邻域结构是通过K近邻算法来获取的,具体通过计算每个样本的欧几里得距离来选择其最近的k个邻域点。低频特征通过平滑操作从邻域结构中提取,而高频特征则通过强化邻域内的局部变化来获取。该邻域结构的计算和使用确保了模态间的共享信息得到保留,同时增强了情绪特征的表达能力。

$$H^{\phi_p,l}=\left(\tilde{A}^{\phi_p}_{\text{sym}}\right)^r f(X) \tag{5}$$

$$H^{\phi_p,\boldsymbol{h}}=\left(\tilde{L}^{\phi_p}_{\text{sym}}\right)^r f(X) \tag{6}$$

$$\tilde{H}_i^{\phi_p,l}=\text{norm}\left(\tilde{H}_i^{\phi_p,l}\right),\tilde{H}_i^{\phi_p,l}=\text{norm}\left(\tilde{H}_i^{\phi_p,\boldsymbol{h}}\right) \tag{7}$$

其中,$f(\cdot):\mathbb{R}^{d_f}\to\mathbb{R}^d$表示多模态特征编码器,在本研究中通过多层感知机实现;$H^{\phi_p,l}$和$H^{\phi_p,\boldsymbol{h}}$分别表示在不同模态邻域结构下提取的低频平滑特征和高频增强特征;r为滤波器的阶数。为消除特征幅值对情绪识别效果的影响,我们对每个模态的特征向量进行L2归一化,即$\text{norm}(\boldsymbol{h})=(\boldsymbol{h}/\|\boldsymbol{h}\|)$。

随后,将每个模态邻域视图下的低频平滑特征和高频增强特征共同输入到共享解码器$g(\cdot):\mathbb{R}^{d_f}\to\mathbb{R}^d$中。为了保证各邻域表示能够保留充足的信息,我们采用缩放余弦误差(Scaled Cosine Error,SCE)作为重构损失函数。通过调整锐化参数γ,可以控制简单样本在损失函数中的贡献。具体来说,锐化参数γ的取值范围为[0,1],并通过交叉验证进行调优。实验表明,锐化参数的调整对情绪识别任务的性能具有显著影响,特别是当γ较大时,高频特征的增强效果更为显著,有助于提升情绪特征的表达,而当γ较小时,低频特征的贡献增强,更强调模态间的共性信息。

$$\mathcal{L}_{\text{SCE}}=\frac{1}{N\cdot P}\sum_{p=1}^{P}\sum_{p=1}^{N}\left(1-\frac{X_i^{\mathrm{T}}\hat{X}_i^P}{\|X_i\|\|\hat{X}_i^P\|}\right)^{\gamma} \tag{8}$$

其中,$[\cdot\|\cdot]$表示按行拼接操作。本文中的解码器同样采用多层感知机实现。式(9)中表示的损失函数中,锐化参数γ的设置依据实验数据,经过交叉验证选择最优值,以实现低频与高频特征之间的平衡,从而优化情绪识别的准确性。

实验发现,在多模态情绪识别中,来自不同模态的特征存在显著的结构和语义差异。因此,有必要采用自适应的编码机制来应对多样化的模态特征。针对这一需求,本文引入了一种自适应双频特征融合方法,以实现多模态的多尺度建模和情绪特征的补全。

$$\omega_{i,p}^{l}=\sigma\left(q_l^{\mathrm{T}}\tilde{H}_i^{\phi_p,l}\right),\omega_{i,p}^{\boldsymbol{h}}=\sigma\left(q_l^{\mathrm{T}}\tilde{H}_i^{\phi_p,\boldsymbol{h}}\right) \tag{9}$$

$$\beta_{i,p}^{l}=\frac{\exp\left(\omega_{i,p}^{l}\right)}{\sum_{j=1}^{p}\exp\left(\omega_{i,j}^{l}\right)+\sum_{j=1}^{p}\exp\left(\omega_{i,j}^{\boldsymbol{h}}\right)} \tag{10}$$

$$\beta_{i,p}^{\boldsymbol{h}}=\frac{\exp\left(\omega_{i,p}^{\boldsymbol{h}}\right)}{\sum_{j=1}^{p}\exp\left(\omega_{i,j}^{l}\right)+\sum_{j=1}^{p}\exp\left(\omega_{i,j}^{\boldsymbol{h}}\right)} \tag{11}$$

$$Z_i=\sum_{p=1}^{P}\beta_{i,p}^{l}\tilde{H}_i^{\phi_p,l}+\sum_{p=1}^{P}\beta_{i,p}^{l}\tilde{H}_i^{\phi_p,\boldsymbol{h}} \tag{12}$$

其中,$\boldsymbol{q}_1$和$\boldsymbol{q}_{\boldsymbol{h}}$分别表示与低频和高频信息相关的可学习注意力向量;$\beta_{i,p}^{l}$和$\beta_{i,p}^{h}$分别为节点$v_i$在$\phi_p\sigma(\cdot)$下低通和高通特征的自适应融合权重;$v_i$是非线性激活函数;$Z_i$表示节点$v_i$的最终特征表示,可用于后续下游任务。与许多现有的无监督异构图表示学习方法为所有节点采用共享的语义融合权重不同[36-37],我们的方法针对不同邻域模式的节点,实现了自适应的双频融合。

2.4　基于LLM增强的弱监督分层情绪识别方法

在多模态情绪识别的最后识别模块中,我们训练一个层次化的情绪分类器,使用自信优化后的核心类特征。直接将核心类作为伪标签进行训练的方法效率低,因为核心类不能覆盖所有情绪类别,尤其是细粒度和长尾类,它们因低频率而未被选为核心类。为了弥补这一不足,我们提出了基于路径的文档生成策略,通过LLM生成少量增强文档(每个路径生成5个文档)。这些文档被添加到伪标签数据中,确保每个情绪类别至少作为正类出现在5个文档中。通过使用路径而非单一类别来引导生成,确保低级别类别的意义依赖于父类别,如"情感关怀"→"快乐情绪"路径生成关于快乐情绪的文本。为了增加数据多样性,每条路径生成多个不同文档。我们将这些生成的文档表示为D_{gen},并在接下来的情绪识别任务中使用这些增强数据。

现在,拥有两组数据,伪标签文档D_{core}和LLM生成的文档D_{gen},我们可以介绍分类器架构和训练过程。将融合后的多模态特征输入到一个分类器模型中。在此基础上,结合低通和高通滤波器的输出,引入了一个文本匹配网络,以进一步优化情绪识别过程。我们采用类似于文献[38]中提出的简单文本匹配网络架构,其中包含一个通过预训练BERT-base模型初始化的文档编码器[39],并使用对数双线性匹配网络进行分类。

该架构的设计原理与之前的多模态特征提取过程相辅相成,尽管文本匹配网络的核心是基于文本信息,但它同样能够与其他模态融合,通过特征编码器从每个模态中提取有价值的情绪信息,从而实现更准确的情绪分类。

具体来说,分类器预测文档d_i属于特定类别c_j的概率,通过式(13)表示为

$$p\left(c_j|d_i\right)=P\left(y_i=1|d_i\right)=\sigma\left(\exp\left(c_j^{\mathrm{T}}\boldsymbol{W}d\right)\right) \tag{13}$$

其中,σ是sigmoid激活函数;$\boldsymbol{W}$是一个可学习的交互矩阵;c_j和d_i分别是类别和文档的编码表示。

对于每个标记为核心类c_j的文档d_i,我们根据其与核心类的关系以及在标签层级中的位置,定义其正类和负类,来为模型训练提供标签。具体而言,文档的正类$C_{\text{core},i,+}$被定义为其核心类及其在标签层级中的所有祖先类的并集。假设核心类的祖先类更有可能是真正的标签,而后代类可能并非完全是负类,尤其是在自动生成的核心类可能不是最优的情况下。因此,正类可以表示为

$$C_{\text{core},i,+}=C_i\cup\left(\bigcup\nolimits_{c\in\mathbb{C}_i}\text{Anc}(c)\right) \tag{14}$$

其中,$\text{Anc}(c)$表示类c的所有祖先类集合。

负类$C_{\text{core},i,-}$则定义为不属于正类且不属于任何核心类后代的类。我们认为,核心类的后代类并不一定是负类,因此,负类可以表示为

$$C_{\text{core},i,-}=C-C_{\text{core},i,+}-U_{c\in\mathbb{C}_i}\text{Dec}(c) \tag{15}$$

其中,$\text{Dec}(c)$表示类c的所有后代类集合。

对于由LLM生成的文档,由于我们对其伪标签有更高的信心,直接将对应路径上的所有类视为正类,将其他所有类视为负类。

$$\mathbb{C}_{p,+}^{\text{gen}}=\mathbb{C}_p,\ \mathbb{C}_{p,+}^{\text{gen}}=C-\mathbb{C}_p \tag{16}$$

然后,我们使用二元交叉熵损失训练一个多标签分类器:

$$\mathcal{L}_{\text{core}}=-\sum_{d_i\in D^{\text{core}}}\left(\sum_{c_j\in\mathbb{C}_{i,+}^{\text{core}}}\log p\left(c_j|d_i\right)+\sum_{c_j\in\mathbb{C}_{i,-}^{\text{core}}}\log\left(1-p\left(c_j|d_i\right)\right)\right) \tag{17}$$

$$\mathcal{L}_{\text{gen}}=-\sum_{d_i\in D^{\text{gen}}}\left(\sum_{c_j\in\mathbb{C}_{i,+}^{\text{gen}}}\log p\left(c_j|d_i^p\right)+\sum_{c_j\in\mathbb{C}_{i,-}^{\text{gen}}}\log\left(1-p\left(c_j|d_i^p\right)\right)\right) \tag{18}$$

$$\mathcal{L}=\mathcal{L}_{\text{core}}+\frac{\left|D^{\text{core}}\right|}{\left|D^{\text{gen}}\right|}\cdot\mathcal{L}_{\text{gen}} \tag{19}$$

两个数据集的损失项按它们的相对大小$\frac{\left|D^{\text{core}}\right|}{\left|D^{\text{gen}}\right|}$进行加权。

3　仿真实验

3.1　数据集与评估标准

本文在3个具有代表性的多模态情感识别基准数据集上对所提出的LLM-EmoGraph模型进行了系统评估,分别为IEMOCAP[40]、MELD[41]和MOSEI[42]。其中,IEMOCAP包含大量双人角色扮演对话,涵盖音频、文本和视频模态,适合评估模型对语音情绪线索的建模能力;MELD则来源于美剧《老友记》的多说话人对话片段,具有强上下文依赖、模态切换频繁及情绪动态明显等特点,能够有效检验模型在真实场景中的泛化能力;MOSEI是一个大规模的多模态数据集,包含超过23 000个视频片段,涵盖文本、音频和视频模态,适用于评估模型在更大规模数据集上的表现及情绪识别能力。

在特征提取阶段,音频模态使用OpenSmile工具提取prosodic与spectral特征,文本模态采用TextCNN网络获取句子级语义表示,视觉模态则通过DenseNet网络提取面部表情和姿态变化等高层特征表示。评估指标方面,本文分别报告每类情绪的F1分数,用于衡量模型在各类别上的识别效果,并采用加权平均F1值(W-F1)作为总体性能指标,以充分体现模型在

类间不平衡条件下的综合识别能力。所有实验均在统一的训练策略与参数配置下进行，确保与主流基线方法的公平对比。

3.2 实验设置

所有实验均在 Python 3.8 环境下，基于 PyTorch 深度学习框架完成，并在单块 NVIDIA RTX 4090 24 GB GPU 上进行训练与测试。所提出的 LLM-EmoGraph 模型采用 Adam 优化器进行参数更新，其中 IEMOCAP 数据集的初始学习率设为 10^{-4}，MELD 数据集则设为 5×10^{-5}，以适应不同数据集的训练规模与收敛速度。对应的批次大小分别为 16（IEMOCAP）和 8（MELD），训练损失函数采用标准交叉熵，用于多类别情绪分类任务。为充分捕捉对话上下文信息，IEMOCAP 的上下文窗口大小设为 20，MELD 为 5，分别优化长对话与短对话场景。为防止过拟合并增强模型的泛化能力，实验过程中统一采用 0.5 的 dropout 概率，并设置 L2 权重衰减系数为 1.0×10^{-5}。所有实验在不同权重初始化条件下独立运行 10 次，并取其平均性能作为最终结果，以缓解单次训练中随机性的影响并提高结果的可靠性。

为了全面评估 LLM-EmoGraph 的性能，我们选取了多种具有代表性的多模态情感识别方法作为对比基线模型，包括传统时序建模方法、图神经网络方法以及最新的多模态融合策略，如 DialogueRNN[43]、DialogueGCN[44]、MMGCN[14]、MM-DFN[34]、COGMEN[13]、MultiEMO[35]、SDT[45]、GraphCFC[16]和 RL-EMO[46]，以评估 LLM-EmoGraph 在多模态情感识别任务中的优势。

3.3 与基线比较结果

表 2、表 3、表 4 分别展示了 LLM-EmoGraph 方法在 IEMOCAP、MELD 和 MOSEI 3 个多模态情感识别数据集上的优异性能，并与多种主流方法进行了比较。在 IEMOCAP 数据集上，LLM-EmoGraph 在加权准确率上提升了 6.2 个百分点，并在多个情绪类别（如高兴、愤怒等）上表现出色。在 MELD 数据集上，模型在加权准确率上提升了 0.26 个百分点，特别是在中性、愤怒和厌恶情绪的识别上最优，展现了其在多说话人交互中的稳定性。在 MOSEI 数据集上，LLM-EmoGraph 在加权准确率上提升了 0.38 个百分点，且在中性、愤怒、喜悦等情绪上的表现最优，验证了其在多模态情绪建模中的优势。实验结果证明，LLM-EmoGraph 的优异性能源于大规模语言模型的语义建模能力、跨模态图结构的融合以及图神经网络中的结构性改进，有效提升了情感识别的准确性与稳定性。

3.4 情绪转变分析

为进一步评估模型在处理复杂情绪分类任务中的能力，我们对 IEMOCAP 和 MELD 数据集的混淆矩阵进行了可视化，如图 3 所示。图中左侧子图对应 IEMOCAP 数据集，右侧子图展示的是 MELD 数据集的结果。

表 2 IEMOCAP 数据集性能比较与结果标注

Table 2 Performance comparison and result annotation on the IEMOCAP dataset

模型	高兴	悲伤	中性	愤怒	兴奋	沮丧	加权准确率/%	加权 F1 分数
DialogueRNN	52.58	81.13	69.14	59.20	64.12	59.33	66.68	65.50
DialogueGCN	46.12	77.90	61.92	54.31	68.22	59.43	64.48	62.75
MMGCN	41.60	81.15	70.51	54.36	73.85	59.41	66.54	66.18
MM-DFN	45.29	<u>81.83</u>	70.45	64.07	73.65	59.42	66.52	66.22
COGMEN	47.12	77.60	68.72	61.88	69.82	59.86	66.56	67.08
MultiEMO	53.67	**83.10**	70.54	65.23	75.72	<u>66.13</u>	67.61	66.51
SDT	<u>63.95</u>	81.20	**74.21**	<u>69.41</u>	<u>78.98</u>	61.02	67.70	66.80
GraphCFC	52.02	78.42	70.91	64.38	75.25	61.03	<u>69.30</u>	<u>68.27</u>
RL-EMO	47.40	79.42	67.92	61.72	75.46	63.15	67.39	66.57
LLM-EmoGraph	**67.10**	81.40	<u>74.18</u>	**71.05**	**79.61**	**70.88**	**75.50**	**75.02**

注：加粗表示最优结果，下划线表示次优结果。

在 IEMOCAP 数据集的混淆矩阵中，模型在识别“中性”和“高兴”等情绪上表现良好，但仍存在如“高兴”被误判为“中性”“愤怒”与“惊讶”或“厌恶”混淆的现象，反映出情绪表达的模糊性和情绪间感知线索的重叠。在 MELD 数据集中，模型在“中性”和“沮丧”类别上的识别准确率较高，但也存在如“兴奋”与“高兴”之间的误判，表明情绪在语言或声学特征上的相似性加大了区分难度。总体而言，混淆矩阵揭示了模型在主导情绪类别上的优良表现，但在区分相似情绪时仍面临挑战，尤其是在情绪微小转变或多情绪依赖的情境中，这与已有研究一致，进一步凸显了情绪识别任务的复杂性。

3.5 消融实验

本研究提出的多模态情感识别框架由 3 大核心组件构成：多模态遮蔽策略（Multimodal Masking Strategies，MMS）、ADSFF 以及基于大语言模型的增强型层次学习模块（LLM-Enhanced Hierarchical Learning，LLM-EHL）。为验证各模块的有效性，我们分别移除其中一个模块进行消融实验，并对模型性能进行评估。实验结果如表 5 所示，3 项模块均对最终性能有正向贡献，移除任意一个组件都会造成准确率和 F1 分数的下降，说明各模块在整体架构中具备互补性。

其中，MMS 模块对性能的影响最为显著。在 IEMOCAP 与 MELD 数据集上去除该模块后，模型性能均有明显下降，表明其在捕捉多模态情感线索中的关键作用。相比之下，ADSFF 的作用强于 LLM-EHL，可能是由于其具备融合多尺度特征的能力，能够同时

表3　MELD数据集性能比较与结果标注

Table 3　Performance comparison and result annotation on the MELD dataset

模型	中性	悲伤	喜悦	厌恶	愤怒	加权准确率/%	加权F1分数
DialogueRNN	79.12	40.38	60.78	26.04	53.12	64.80	65.02
DialogueGCN	78.67	41.61	61.57	14.02	53.44	64.47	64.44
MMGCN	78.75	**43.75**	61.25	19.05	55.62	65.20	65.75
MM-DFN	78.97	42.32	62.42	25.91	54.88	65.31	65.56
MultiEMO	78.81	40.30	62.53	24.88	53.45	64.89	65.12
SDT	79.26	40.52	61.58	25.94	54.86	64.95	65.34
GraphCFC	78.32	37.88	60.87	24.84	53.98	64.45	64.79
RL-EMO	78.96	40.97	**62.56**	23.72	54.62	64.85	65.25
LLM-EmoGraph	**79.31**	43.12	62.55	**26.47**	**55.69**	**65.57**	**65.94**

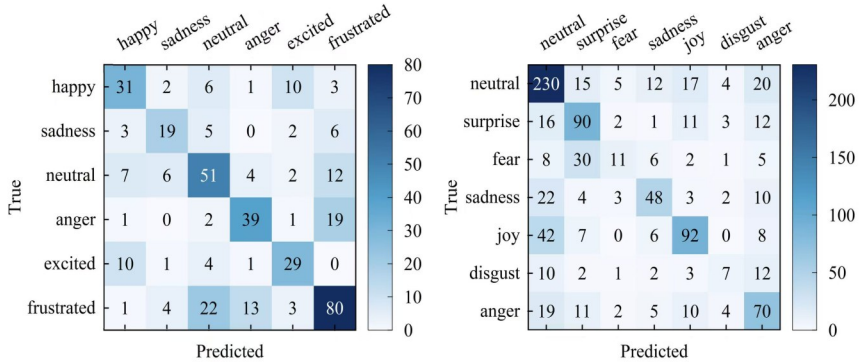
注:加粗表示最优结果,下划线表示次优结果。

表4　MOSEI数据集性能比较与结果标注

Table 4　Performance comparison and result annotation on the MOSEI dataset

模型	中性	悲伤	喜悦	厌恶	愤怒	加权准确率/%	加权F1分数
DialogueRNN	79.80	45.20	62.50	30.50	55.60	70.40	69.80
DialogueGCN	78.50	47.10	60.90	28.00	54.50	71.20	70.10
MMGCN	79.00	48.30	63.20	32.40	57.20	72.00	71.70
MM-DFN	80.10	46.90	64.50	34.20	58.10	73.50	72.90
MultiEMO	79.50	47.80	63.90	33.10	56.80	73.10	72.50
SDT	80.20	48.50	65.00	35.00	59.00	74.10	73.30
GraphCFC	79.30	46.00	62.80	31.50	55.80	72.70	71.90
RL-EMO	80.50	49.10	**65.60**	36.00	**60.20**	75.00	74.20
LLM-EmoGraph	**81.10**	**50.00**	65.50	**37.50**	59.30	**76.80**	**75.50**

注:加粗表示最优结果,下划线表示次优结果。



注:矩阵的行和列分别表示真实标签和预测标签

图3　LLM-EmoGraph在IEMOCAP和MELD数据集中的情绪转变样本上的混淆矩阵

Figure 3　Confusion matrix of emotion transition samples in LLM-EmoGraph on the IEMOCAP and MELD datasets

捕捉高频与低频的情绪信号,从而丰富情感表示。而LLM-EHL则进一步提升语义一致性与上下文完整性,同时承担最终的情绪分类任务。

3.6　不同窗口大小设置的影响分析

如图4所示,我们通过改变每个子图中所包含的上下文语句数量,分析不同对话窗口大小对模型性能的影响。窗口大小为k表示在目标语句的基础上,分别引入其前后各k条语句作为上下文信息。当窗口大小为0,即完全不使用上下文信息时,模型性能最差。

随着窗口大小的增加,模型性能逐渐提升,并在某一特定窗口大小下达到最优。以IEMOCAP数据集为例,当窗口大小为20时性能最佳;而在MELD数据集中,最佳窗口大小为5。这一差异主要来源于两个

表5 对LLM-EmoGraph关键组件的消融实验研究

Table 5 Ablation study on key components of LLM-EmoGraph

方法	IEMOCAP	IEMOCAP	MELD	MELD
	加权准确率/%	加权F1分数	加权准确率/%	加权F1分数
LLM-EmoGraph	75.50	75.02	65.57	65.94
去除 MMS	71.18	71.74	64.39	64.72
去除 ADSFF	69.63	68.88	63.27	63.06
去除LLM-EHL	71.34	72.18	64.05	64.01

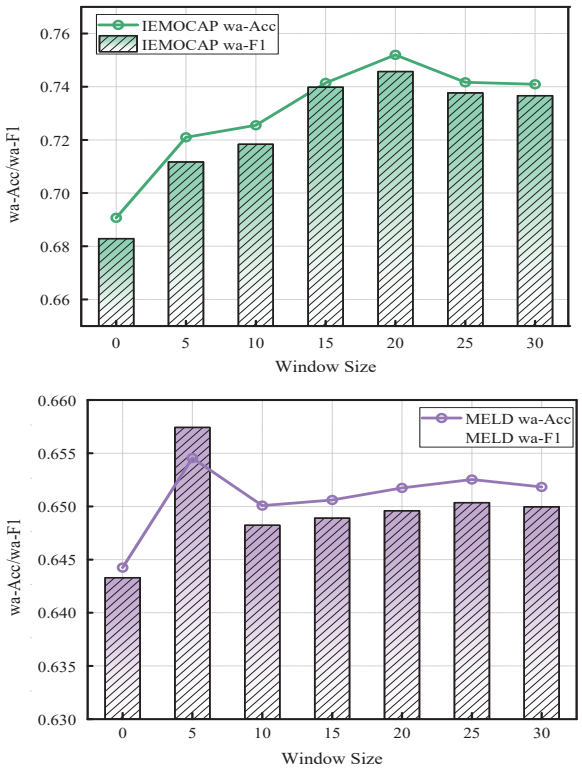


图4 不同窗口大小对LLM-EmoGraph模型性能的影响

Figure 4 The impact of different window sizes on the performance of the LLM-EmoGraph model

数据集中对话长度的不同:IEMOCAP包含较长的多轮对话,更广泛的上下文有助于理解情感表达;而MELD中的对话相对较短,较小的窗口便足以提供必要的语境支持。

需要注意的是,虽然更大的窗口有助于提升情感识别的准确性,但也会显著增加计算开销,从而影响模型的整体运行效率。因此,在实际应用中需根据数据特性与资源限制,在性能与效率之间进行权衡。

3.7 对比的多模态对话情感识别方法

如表6所示,LLM-EmoGraph在IEMOCAP和MELD数据集上展示了优越的性能,尤其在加权准确率和加权F1分数上显著高于基于GPT-4o和GPT-4V的模型。与多模态大模型相比,我们的框架在情感识别任务中的优势主要体现在精确的情感依赖建模和较强的鲁棒性。LLM-EmoGraph通过跨域多图结构和自适应双尺度特征融合机制,有效捕捉情感的动态转变和复杂的模态间依赖关系。同时,LLM-EmoGraph能够在模态缺失、噪声干扰等复杂情境下保持较高的识别精度,展现出较强的鲁棒性。

多模态大模型(GPT-4o/ GPT-4V)尽管在跨模态理解上有较好的表现,但它们在细粒度情感分类和处理复杂情感依赖方面相对逊色。特别是在面对多说话人对话或复杂情感变化时,LLM-EmoGraph展现出了更高的准确性和更强的适应能力。

表6 与多模态大模型(GPT-4o / GPT-4V)对比的多模态对话情感识别方法

Table 6 Multimodal emotion recognition methods compared with multimodal large models (GPT-4o / GPT-4V)

模型	IEMOCA加权准确率/%	IEMOCAP加权F1分数	MELD加权准确率/%	MELD加权F1分数
LLM-EmoGraph	75.50	75.02	65.57	65.94
GPT-4o	70.12	69.35	63.47	63.85
GPT-4V	71.45	70.25	64.52	64.80
DialogueRNN	66.68	65.50	64.80	65.02
DialogueGCN	62.75	62.05	64.44	64.10
MMGCN	66.18	65.18	65.75	65.40

4 结论

本文提出了一种新型多模态情感识别框架——LLM-EmoGraph,将大语言模型与全局-局部跨域图结构深度融合,以应对模态异构性、上下文歧义性及信号噪声等关键挑战。该框架包括3大核心模块:跨域多图预训练实现多模态语义对齐,双尺度特征融合捕捉全局与局部情绪线索,弱监督层次化识别增强语义一致性与标签利用效率。在IEMOCAP、MELD等主流

数据集上,LLM-EmoGraph在加权准确率与加权F1分数等多个指标上均优于现有方法,且在模态缺失与类别不平衡等实际干扰条件下展现出强鲁棒性,为构建具备情绪感知能力的对话系统提供了有力支持。

参考文献

[1] Van Kleef G A, Côté S. The social effects of emotions[J]. Annual Review of Psychology, 2022, 73: 629-658.

[2] Zhu T, Li L D, Yang J F, et al. Multimodal sentiment analysis with image-text interaction network[J]. IEEE Transactions on Multimedia, 2023, 25: 3375-3385.

[3] Zhu T, Li L D, Yang J F, et al. Multimodal emotion classification with multi-level semantic reasoning network[J]. IEEE Transactions on Multimedia, 2023, 25: 6868-6880.

[4] Nie W Z, Bao Y R, Zhao Y, et al. Long dialogue emotion detection based on commonsense knowledge graph guidance[J]. IEEE Transactions on Multimedia, 2024, 26: 514-528.

[5] Wei L W, Hu D, Zhou W, et al. Modeling both intra- and inter-modality uncertainty for multimodal fake news detection[J]. IEEE Transactions on Multimedia, 2023, 25: 7906-7916.

[6] Liu K, Xue F, Guo D, et al. Multimodal graph contrastive learning for multimedia-based recommendation[J]. IEEE Transactions on Multimedia, 2023, 25: 9343-9355.

[7] Wen J Y, Qin F W, Du J, et al. MsgFusion: Medical semantic guided two-branch network for multimodal brain image fusion[J]. IEEE Transactions on Multimedia, 2024, 26: 944-957.

[8] Poria S, Majumder N, Mihalcea R, et al. Emotion recognition in conversation: Research challenges, datasets, and recent advances[J]. IEEE Access, 2019, 7: 100943-100953.

[9] Wang Y X, Liu M, Li Z, et al. Unlocking the power of multimodal learning for emotion recognition in conversation[C]// Proceedings of the 31st ACM International Conference on Multimedia. New York: ACM, 2023: 5947-5955.

[10] Wang P, Ganushchak L, Welie C, et al. The dynamic nature of emotions in language learning context: Theory, method, and analysis[J]. Educational Psychology Review, 2024, 36(4): 105.

[11] Zhang D, Wu L Q, Sun C L, et al. Modeling both context- and speaker-sensitive dependence for emotion detection in multi-speaker conversations[C]//Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence. International Joint Conferences on Artificial Intelligence Organization, 2019: 5415-5421.

[12] Nie W Z, Chang R H, Ren M J, et al. I-GCN: Incremental graph convolution network for conversation emotion detection[J]. IEEE Transactions on Multimedia, 2022, 24: 4471-4481.

[13] Joshi A, Bhat A, Jain A, et al. COGMEN: Contextualized GNN based multimodal emotion recognition[C]//Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: ACL, 2022: 4148-4164.

[14] Hu J W, Liu Y C, Zhao J M, et al. MMGCN: Multimodal fusion via deep graph convolution network for emotion recognition in conversation[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Stroudsburg: ACL, 2021: 5666-5675.

[15] Li J, Wang X P, Lv G Q, et al. GraphMFT: A graph network based multimodal fusion technique for emotion recognition in conversation[J]. Neurocomputing, 2023, 550: 126427.

[16] Li J, Wang X P, Lv G Q, et al. GraphCFC: A directed graph based cross-modal feature complementation approach for multimodal conversational emotion recognition[J]. IEEE Transactions on Multimedia, 2024, 26: 77-89.

[17] Li J, Wang X P, Lv G Q, et al. GA2MIF: Graph and attention based two-stage multi-source information fusion for conversational emotion detection[J]. IEEE Transactions on Affective Computing, 2024, 15(1): 130-143.

[18] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[C/OL]//Proceedings of the 38th International Conference on Machine Learning, PMLR 139, 2021: 8748-8763. https://proceedings.mlr.press/v139/radford21a.

[19] Jia Chao, Yang Yinfei, Xia Ye, et al. Scaling up visual and vision-language representation learning with noisy text supervision[C/OL]//Proceedings of the 38th International Conference on Machine Learning, PMLR 139, 2021: 4904-4916. https://proceedings.mlr.press/v139/jia21b.html.

[20] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16 × 16 words: Transformers for image recognition at scale[PP/OL]. V2.arXiv (2021-06-03)[2025-09-05]. https://doi.org/10.48550/arXiv.2010.11929.

[21] Jaegle A, Borgeaud S, Alayrac J B, et al. Perceiver IO: A general architecture for structured inputs & outputs[PP/OL]. V3.arXiv (2022-03-15)[2025-09-05]. https://doi.org/

10.48550/arXiv.2107.14795.

[22] Bao H B, Dong L, Piao S H, et al. BEiT: BERT pre-training of image transformers[PP/OL]. V2. arXiv (2022-09-03)[2025-09-05]. https://doi.org/10.48550/arXiv.2106.08254.

[23] He K M, Chen X L, Xie S N, et al. Masked autoencoders are scalable vision learners[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 15979-15988.

[24] Girdhar R, El-Nouby A, Liu Z, et al. ImageBind one embedding space to bind them all[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2023: 15180-15190.

[25] Chen X, Zhang N Y, Li L, et al. Hybrid transformer with multi-level fusion for multimodal knowledge graph completion[C]//Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2022: 904-915.

[26] Zeng Y W, Jin Q, Bao T F, et al. Multi-modal knowledge hypergraph for diverse image retrieval[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2023, 37(3): 3376-3383.

[27] Jin W G, Yang K, Barzilay R, et al. Learning multimodal graph-to-graph translation for molecular optimization[PP/OL]. V3.arXiv (2019-01-28)[2025-09-05]. https://doi.org/10.48550/arXiv.1812.01070.

[28] Wang M L, Shao W, Huang S, et al. Hypergraph-regularized multimodal learning by graph diffusion for imaging genetics based Alzheimer's Disease diagnosis[J]. Medical Image Analysis, 2023, 89: 102883.

[29] Yoon M, Koh J Y, Hooi B, et al. Multimodal graph learning for generative tasks[PP/OL]. V2. arXiv (2023-10-12) [2025-09-05]. https://doi.org/10.48550/arXiv.2310.07478.

[30] Sahu G, Vechtomova O. Adaptive fusion techniques for multimodal data[PP/OL]. V2. arXiv (2021-01-26) [2025-09-05]. https://doi.org/10.48550/arXiv.1911.03821.

[31] Li S Z, Zhang T, Chen B N, et al. MIA-net: Multi-modal interactive attention network for multi-modal affective analysis[J]. IEEE Transactions on Affective Computing, 2023, 14(4): 2796-2809.

[32] Zheng J H, Zhang S, Wang Z L, et al. Multi-channel weight-sharing autoencoder based on cascade multi-head attention for multimodal emotion recognition[J]. IEEE Transactions on Multimedia, 2023, 25: 2213-2225.

[33] Zhao S L, Liu Y C, Jiao Q, et al. Mitigating modality discrepancies for RGB-T semantic segmentation[J]. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35(7): 9380-9394.

[34] Hu D, Hou X L, Wei L W, et al. MM-DFN: Multimodal dynamic fusion network for emotion recognition in conversations[C]//ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE, 2022: 7037-7041.

[35] Shi T, Huang S L. MultiEMO: An attention-based correlation-aware multimodal fusion framework for emotion recognition in conversations[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2023: 14752-14766.

[36] Wang X, Liu N, Han H, et al. Self-supervised heterogeneous graph neural network with co-contrastive learning[C]//Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. New York: ACM, 2021: 1726-1736.

[37] Yu J X, Li X. Heterogeneous graph contrastive learning with meta-path contexts and weighted negative samples[C]//Proceedings of the 2023 SIAM International Conference on Data Mining (SDM). Philadelphia: PASociety for Industrial and Applied Mathematics, 2023: 37-45.

[38] Shen J M, Qiu W D, Meng Y, et al. TaxoClass: Hierarchical multi-label text classification using only class names[C]//Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: ACL, 2021: 4239-4249.

[39] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Kerrville: Association for Computational Linguistics 2019: 4171-4186.

[40] Busso C, Bulut M, Lee C C, et al. IEMOCAP: Interactive emotional dyadic motion capture database[J]. Language Resources and Evaluation, 2008, 42(4): 335-359.

[41] Poria S, Hazarika D, Majumder N, et al. MELD: A multimodal multi-party dataset for emotion recognition in conversations[PP/OL]. V6. arXiv (2019-06-04) [2025-09-05]. https://doi.org/10.48550/arXiv.1810.02508.

[42] Zadeh A, Zellers R, Pincus E, et al. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos[PP/OL]. V2.arXiv (2016-08-12)[2025-09-05]. https://arxiv.org/abs/1606.06259.

[43] Majumder N, Poria S, Hazarika D, et al. DialogueRNN:

An attentive RNN for emotion detection in conversations[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(1): 6818-6825.

[44] Ghosal D, Majumder N, Poria S, et al. DialogueGCN: A graph convolutional neural network for emotion recognition in conversation[PP/OL]. V1.arXiv (2019-08-30)[2025-09-05]. https://doi.org/10.48550/arXiv.1908.11540.

[45] Ma H, Wang J, Lin H F, et al. A transformer-based model with self-distillation for multimodal emotion recognition in conversations[J]. IEEE Transactions on Multimedia, 2024, 26: 776-788.

[46] Zhang C W, Zhang Y H, Cheng B. RL-EMO: A reinforcement learning framework for multimodal emotion recognition[C]//ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE, 2024: 10246-10250.

作者简介

黄　辰　男,1983年8月生,福建龙岩人。现为湖北大学计算机学院教授。主要研究方向为人工智能、脑机接口。

E-mail: huang@hubu.edu.cn

杨　超　男,1982年9月生,湖北武汉人。现为湖北大学计算机学院教授。主要研究方向为智能计算、信息安全等。

E-mail: stevenyc@hubu.edu.cn

马浩博　男,2000年5月生,河北秦皇岛人。现为湖北大学计算机学院硕士研究生。主要研究方向为人工智能、脑科学、情感分析。

E-mail: 202321116012629@stu.hubu.edu.cn

宋建华　女,1973年3月生,湖北襄阳人。现为湖北大学网络空间安全学院教授。主要研究方向为网络与信息安全。

E-mail: sjhhubu@126.com

张　龑　男,1974年6月生,湖北宜昌人。现为湖北大学计算机学院教授。主要研究方向为信息安全、大数据分析。中国电子学会会员编号:E190197582M。

E-mail: zhangyan@hubu.edu.cn