计 算 机 研 究 与 发 展
Journal of Computer Research and Development
https://crad.ict.ac.cn
63(2): 305−323, 2026

对抗软对比调制动态图学习的多模态会话情绪识别方法

王顺杰　蔡国永　吕光瑞
（桂林电子科技大学计算机与信息安全学院　广西桂林　541004）
（广西可信软件重点实验室（桂林电子科技大学）　广西桂林　541004）
（wsj12130@foxmail.com）

Adversarial Soft-Contrast Modulated Dynamic Graph Learning for Multimodal Emotion Recognition in Conversations

Wang Shunjie, Cai Guoyong, and Lü Guangrui
（*School of Computer and Information Security*, *Guilin University of Electronic Technology*, *Guilin*, *Guangxi* 541004）
（*Guangxi Key Laboratory of Trusted Software* (*Guilin University of Electronic Technology*), *Guilin*, *Guangxi* 541004）

Abstract　Multimodal emotion recognition in conversations (MERC) has emerged as a prominent research focus in the domain of human-computer intelligent interaction. It demonstrates extensive applicability across diverse scenarios, including affective dialogue systems and conversational recommendation systems, thereby substantially improving user's experience through enhanced emotional engagement. Identifying abstract emotional semantics in multimodal conversation scenarios poses a significant challenge. Current research predominantly relies on contrastive learning frameworks to extract discriminative feature. Although these approaches effectively enforce intra-class feature consistency, they inherently constrain fine-grained feature diversity. This limitation consequently impairs the model's generalization capacity, particularly when processing samples from minority classes. In addition, current dialogue context graph learning only captures the dependence with a fixed window, neglecting the dynamic cross-correlation of the conversation flow, which can lead to either redundancy or insufficient of extracted context information. To address the aforementioned limitations, the adversarial soft-contrast modulated dynamic graph learning is proposed for MERC. Specifically, a dynamic discourse graph is firstly constructed according to the number of utterances emitted by the speaker in the conversation, which allows for dynamically modeling the conversational dependency range of the utterance in each modality to extract rich contextual information more precisely. Secondly, the adversarial soft-contrast training mechanism is designed to enhance the discrimination and robustness of the network learning. Among them, the class sample space is expanded by adding perturbations to the hidden layers of different modal feature extractors to generate adversarial samples, and the soft-contrast learning is exploited to maximize label semantic consistency of the original samples and the adversarial samples. Finally, a bi-stream graph learning strategy for different modalities is constructed, which not only explores the cross-modal consistent feature interaction guided by text semantics, but also learns the context dynamic dependencies within multimodal dialogue flow to collaboratively

收稿日期：2024-11-18；**修回日期**：2025-06-30
基金项目：国家自然科学基金项目（62366010）；广西自然科学基金项目（2024GXNSFAA010374）；广西可信软件重点实验室项目（kx202207）；广西研究生教育创新计划项目（YCBZ2025158）；广西高校中青年教师科研基础能力提升项目（2024KY0218）
This work was supported by the National Natural Science Foundation of China (62366010), the Guangxi Natural Science Foundation (2024GXNSFAA010374), the Project of Guangxi Key Laboratory of Trusted Software (kx202207), the Innovation Project of Guangxi Graduate Education (YCBZ2025158), and the Research Foundation Ability Enhancement Project for Young and Middle Aged Techers in Guangxi Universities (2024KY0218).
通信作者：蔡国永（ccgycai@guet.edu.cn）

facilitates the complementary fusion of multimodal conversations data. Extensive experiments conducted on the IEMOCAP and MELD benchmark datasets have demonstrated that our proposed model achieves competitive results in terms of effectiveness for MERC, compared with current state-of-the-art methods.

Key words multimodal emotion recognition in conversations (MERC); dynamic discourse graph; adversarial soft-contrast; bi-stream graph learning; semantic consistency

摘 要 多模态会话情绪识别（multimodal emotion recognition in conversations, MERC）已成为人机智能交互的研究热点，目前被广泛应用于情感对话机器人及对话推荐等多个场景。在多模态会话场景中识别抽象的情感语义是困难的，大多现有研究利用对比学习来提取判别特征，其尽管约束了类内特征的一致性，但损害了精细的多样性表征，导致模型泛化性降低，尤其不利于少类样本的学习。此外，目前的会话上下文学习多建模固定窗口的语境依赖，忽略了会话信息流的动态互相关。为了解决上述限制，提出对抗软对比调制动态图学习（adversarial soft-contrast modulated dynamic graph learning, ASDG）方法。具体地：首先根据会话场景中说话人的话语数量构建话语图实现在各自模态内动态地建模话语的会话依赖范围，以精准地提取丰富的语境信息；其次，设计对抗性软对比训练机制，通过在不同模态特征提取器的隐藏层中添加扰动生成对抗样本来扩展类样本空间，并使用软对比学习在原始样本和对抗样本上最大化类扩展样本的标签语义一致性以增强模型学习的判别性与鲁棒性；最后，构建针对不同模态的双流图学习策略，以协同地促进各模态会话数据的互补融合。在 IEMOCAP 和 MELD 的多模态会话基准数据集上进行广泛的实验，结果表明提出的方法与目前先进的方法相比，在 MERC 任务上取得了具有竞争力的效果。

关键词 多模态会话情绪识别；动态话语图；对抗性软对比；双流图学习；语义一致性

中图法分类号 TP391

DOI: 10.7544/issn1000-1239.202440896 **CSTR:** 32373.14.issn1000-1239.202440896

随着智能设备的发展以及会话应用场景的增加，增强机器对人类情感的感知和识别能力使得对话系统能够产生连贯且共情响应，进而提高人机交互的舒适感。近年来，多模态会话情绪识别（multimodal emotion recognition in conversations, MERC）在人工智能领域引起了越来越多研究者的兴趣[1]，旨在识别说话人在文本、视频和音频等多模态信息下的情绪及衍变。目前，MERC 的应用正扩展到对话生成[2]、推荐系统[3]、意见挖掘[4]、医疗诊断[5]等领域。尽管 MERC 取得了一定的进展，但会话情景中的多模态数据相互关联，情感表达的隐晦性更强，使得 MERC 具有极大的挑战性。首先，由于情感是一种语用级的抽象表达，从数据中提取与情感标签相关的泛化表示是困难的[6]，且会话情绪识别涉及不同会话者，使得在会话情绪识别任务中提取判别且泛化的特征更加严峻；其次，由于不同的话语所依赖的会话语境范围是有差异的，然而较多研究设计固定会话窗口的规则来学习上下文依赖关系，其不能随着话语的改变而动态调整，导致获取的上下文信息缺失或者冗余[7-9]；此外，会话情境中不同的模态通常表达相同的情感，但有时由于情感的刻意隐藏而使会话中异构模态的表达是矛盾的，因而如何有效地建模会话场景中异构模态的语义关联是必要且挑战的[10]。

泛化的模态表示学习对于提取稳健的判别性特征是至关重要的。Ma 等人[11]使用迁移学习来提高模型的泛化性，通过在相同的数据集中训练会话生成模型，并将生成模型的中间层参数迁移到会话情感判断任务，但该策略可能会过度适应特定的任务，从而影响模型的稳定性。Ren 等人[12]利用生成对抗网络来学习数据的泛化表示，该方法可以缓解模态间的异构问题，但会使得模态间的判别性降低，不利于挖掘模态间的互补性。Abonizio 等人[13]通过对原始数据添加噪声的方式生成对抗样本以学习更加鲁棒的特征表示，但是在会话数据中添加噪声会破坏会话语义的连贯性。此外，结合交叉熵损失的软对比学习能够捕获类内样例的相似性，并将其与不同类的样例进行对比以学习特征的泛化表示[14]，但是由于相同的情感具有相似的上下文和重叠的特征空间，直接压缩每个类的特征空间可能会损坏类内的细粒度特征，并且大大降低对少样本的分类能力，从而限制了模型的泛化性。因此，为了既保护会话上下文的语义连贯性又提高模态特征的泛化性，本文

提出对抗性软对比训练机制。具体地，通过将噪声施加在特征提取的隐藏层以生成对抗样例，不仅能避免破坏会话上下文的语义关联，而且可扩展每个模态的类表示空间，并在迭代学习中联合原始样本和对抗样本进行软对比优化以最大化模态内相同标签样本的一致性，从而尽可能地保留细粒度的类内特征，以达到更好的泛化效果。

在人际交谈中，情绪随语境流动，并通过语言表述、音频特征（语调、语速等）及面部表情等多模态方式而体现[15-17]，因此针对会话中多模态数据的上下文建模也是一个关键挑战[18]。起初，针对多模态会话上下文建模的研究多通过构建循环神经网络来实现[19-22]，尽管这些工作取得了令人鼓舞的成果，但这些方法不利于远距离的会话上下文依赖关系的学习。目前，研究者更多关注利用图神经网络（graph neural network, GNN）来建模多模态会话的远程上下文依赖性[6-9,23-25]，这是由于图结构能够依据边关系从局部到全局逐步地传播图中任意关联节点间的信息。例如 Shou 等人[6] 和 Hu 等人[25] 通过构建密集连接的图结构来自适应学习上下文关联，尽管构建了会话中任一节点与其他所有节点间的连接，但是容易挖掘到冗余的特征。此外，一些方法[8-9] 通过设置超参数来构建固定的窗口以建模会话的语境信息，然而这也不太适合会话的动态特性，同样限制了从会话场景中自适应地获取精确上下文信息的能力。为此，本文提出能够精确建模上下文的动态话语图构建机制，如图 1 所示，使得当前话语能够根据会话中的讲话者密度来动态调整窗口，不仅能够关注话语较少的说话者所表达的内容，且不会对话语较多的说话者产生过多的关注，以此获得充足的语境信息，且能有效避免上下文信息的冗余。

除了动态建模会话场景中任一话语的各自模态上下文涵盖范围，不同模态的数据交互融合对 MERC 而言也是至关重要的。一些研究基于 GNN 能够捕捉节点间复杂关系结构的优势，实现 MERC 中的多模态交互融合。如 Hu 等人[25] 构建了一个大范围多模态密集图来学习多模态交互和模态内的上下文信息，节点表示不同的模态特征，则每个节点的邻居节点同质或异质，尽管实现了不同模态数据的信息融合，但忽略了不同模态数据中存在的异质性。为了缓解不同模态的异构性，Li 等人[8-9] 构造了 3 个异构图，每个异构图仅融合 2 种模态，但当中心节点与邻居节点是异质模态时，在信息传播过程中异构信息仍然影响不同模态信息的有效交互。此外，Li 等人[26] 使用跨模态注意机制实现多模态密集交互，但其不能精确界定多模态话语级的上下文依赖范围，这会导致用于情绪判断的多模态特征的冗余。为了缓解以上问题，本文将首先构建动态话语图来准确建模各自模态内的语境依赖范围，然后设计文本语义引导的一致性跨模态图学习以降低不同模态信息的异质性并实现模态间信息的高效传递与融合，同时，构建会话流动态图学习来更精确地挖掘多模态话语级上下文依赖范围以进一步减轻多模态特征的冗余。

综上，本文提出对抗软对比调制动态图学习（adversarial soft-contrast modulated dynamic graph lear-

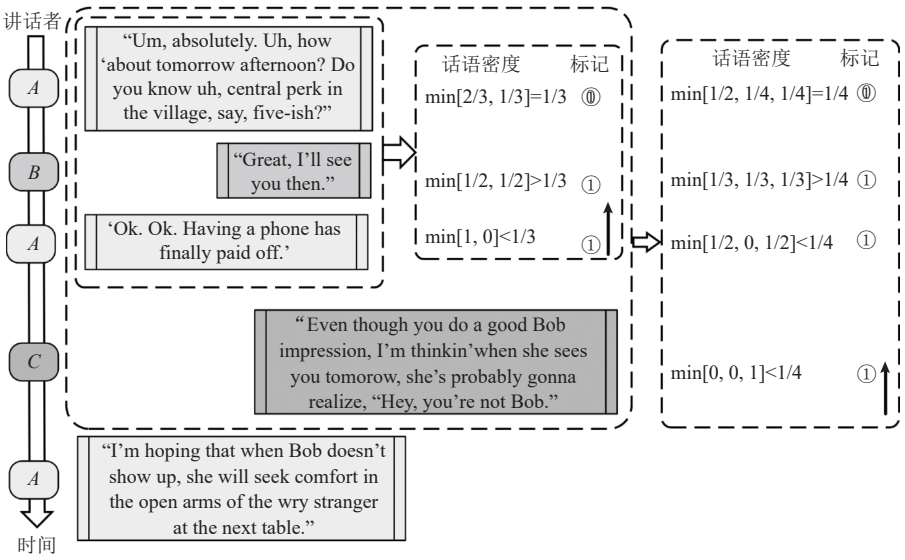


Fig. 1 The process of constructing discourse graph

图 1 构建话语图的过程

ning, ASDG)方法用于多模态会话情绪识别, 所设计的总体结构如图 2 所示。为了提高会话语境依赖的上下文学习的泛化性, 设计对抗训练策略自适应地生成对抗样本, 通过将扰动$\boldsymbol{r}$作用在模态特征提取器的隐藏层, 不仅能避免直接生成对抗样例干扰话语之间的相关性, 还能丰富有限的会话数据样本, 使得模型能够提取更多样化的特征。尤其, 在原始样本和对抗样本上分别进行软对比学习, 使得模型能够学习到更鲁棒的情感特征。同时, 为了动态建模各自模态的上下文依赖信息和模态间的交互信息, 依次设计单模态的会话上下文动态图学习和互补双流图学习。单模态的会话上下文动态图学习根据会话中的讲话者密度分别构建 3 个模态的话语图, 精确建模会话中每个话语的语境依赖范围, 实现模态内上下文语义信息的有效传递。而在互补双流图学习中, 文本语义引导的一致性跨模态图学习用于实现文本语义为中心约束的跨模态特征交互融合, 以缓解异质模态特征的鸿沟; 而会话流动态图学习仍基于动态话语图的策略建模多模态话语级的动态上下文依赖, 以减轻多模态特征的冗余, 然后聚合双流图学习分支用于 MERC 任务。在 2 个公共基准数据集上进行实验, 实验结果一致表明本文所设计的方法优于其他基线方法。

本文的主要贡献概括如下:

1)提出对抗性软对比训练机制。在模型迭代优化时从上下文依赖的会话中生成每个模态特征的对抗样例, 并在原始样例和对抗样例上利用软对比学习最大化标签一致性以丰富多样化的类特征, 从而增强模型的鲁棒性。

2)提出单模态的会话上下文动态图学习策略。会话场景中每个话语根据其所在的位置以及讲话者数量计算其话语密度, 动态地调整话语所关注的前向与后向的窗口大小, 从而构建不同模态的动态话语图, 以准确地提取会话上下文的语义依赖。

3)提出互补双流图学习机制。文本语义引导的一致性跨模态图学习通过在文本语义主导的约束下挖掘文本-音频跨模态交互和文本-视频跨模态交互, 而会话流动态图学习自适应地建模多模态话语级

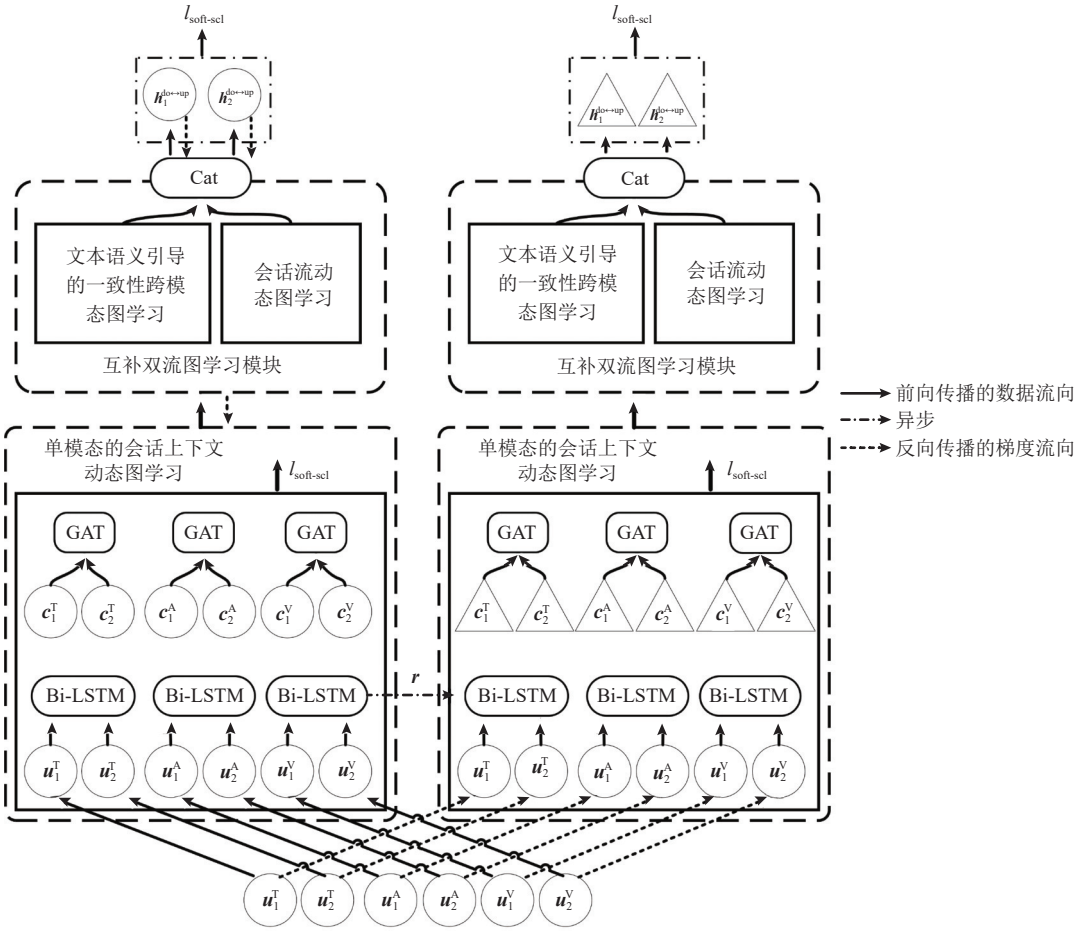


Fig. 2 Overall architecture of adversarial soft-contrast modulated bi-stream dynamic graph learning

图 2 对抗软对比调制双流动态图学习的总体架构

的上下文依赖，两者协同地促进多模态特征的有机融合。

1 相关工作

1.1 多模态会话上下文建模

多模态数据的上下文建模是 MERC 任务的一个关键挑战，通常当前话语的语境对情绪感知至关重要。最初的方法多使用基于循环神经网络（recurrent neural network, RNN）的模型来建模远程上下文依赖用于提取整体语义信息。随着 Transformer 结构的出现，一些研究者开始利用其强大的全局上下文信息建模能力以实现模态内或模态间的特征融合。此外，由于图结构具有复杂的关系建模能力，使得研究者也多关注通过 GNN 来构建多模态的相互依赖关系。

在基于 RNN 的多模态会话情绪识别研究中，Huan 等人[19]提出了一种基于双向门控循环单元和注意力融合的视频多模态情绪识别方法。Wang 等人[20]设计了一个基于长短期记忆网络（long short-term memory, LSTM）的双层模型来处理从音频中提取的每个话语的梅尔频率倒谱数据以建模时序上的语境依赖关系。Li 等人[21]提出了基于注意力的多模态双向 LSTM 网络，该方法将自注意力机制与 LSTM 相结合自适应地学习上下文依赖的多模态特征。Ho 等人[22]提出了一种结合多头注意力机制的 RNN 多层结构来分别建模语音和文本的上下文依赖用于多模态语音情绪识别。尽管基于 RNN 的相关方法在上下文语义建模的情绪识别任务中取得了较好的效果，但是其对远程上下文信息建模能力不足，且当会话的轮次较多时，此类型的网络结构难以满足会话情绪分析的需求，因此，最近的一些研究工作通常将循环神经网络（如 LSTM）作为初步的上下文依赖特征提取器。

多模态的表示和融合是多模态机器学习的关键任务[27]。最近，许多研究利用 Transformer 结构来建模远程上下文交互以及用于多模态会话情绪识别任务，如 Xie 等人[28]利用 Transformer 结构实现了以各自模态为中心的跨模态交互。Wen 等人[29]也利用 Transformer 结构实现了模态间的两两双向交互，并进一步使用 Transformer 来两两聚合跨模态交互的特征用于情绪识别。Huang 等人[30]首先利用 Transformer 结构分别获得各自模态的长期依赖关系，然后构建跨模态 Transformer 结构来实现模态间的融合。Ren 等人[12]通过使用传统的 Transformer 和跨模态注意层来交换不同模态之间的信息以捕获模态间的互补信息。综上，虽然 Transformer 结构已被广泛用于模态内的上下文依赖学习和模态间的跨模态特征交互以提取丰富的语义信息，但是仅仅使用这些密集的 Transformer 结构，或不可避免地导致模态特征的冗余。

目前，研究者多关注基于 GNN 的多模态会话情绪识别研究，如 Hu 等人[25]基于相同模态内节点两两互连、同一话语的不同模态节点也相连的原则，构建了一个音频、视频、文本 3 模态的密集连接图，尽管能充分挖掘模态内与模态间的信息，但这会导致模型学习到冗余的语境信息，且忽略了多模态融合中的异构问题。随后，Li 等人[9]基于上述构图原则构建了 3 个异质图，每个图只考虑 2 种模态的信息，通过减少图中的异构节点，一定程度上缓解了异构模态间的差异。与此同时，Li 等人[8]也使用相同构图方式构建了一个异质图和一个同质图，但却采用固定的超参数来控制上下文窗口的大小，且在构图过程中定义了讲话者内和讲话者间以及模态内和模态间的边的概念，一定程度上减少了冗余特征的学习。此外，Li 等人[26]提出了一种 2 阶段的多模态融合方法，首先为每个模态构建密集连接的图结构来提取模态内的上下文信息，并使用固定的超参数设定上下文窗口的依赖范围，然后利用多头跨模态注意网络实现不同模态的交互和融合。Fu 等人[24]提出了一个基于层次化融合的全局-局部图融合网络，以讲话者关系（讲话者内和讲话者间）构建了 3 个双向的单模态图，在每个单模态图中，采用固定窗口的图网络来渐进地学习模态内的上下文依赖。而 Wei 等人[31]基于事件关系构建了图结构以建模在不同的事件关系下多模态话语级特征之间的关联，并引入了自监督掩码的自编码器来提高图节点特征和结构的鲁棒性。2024 年，Shen 等人[23]提出了一个知识增强的多模态网络，基于讲话者的互动关系构建图，并挖掘话语间的常识知识来丰富文本模态的上下文表示，并使用 Transformer 结构来实现跨模态交互，但其仅建立了一个针对文本的知识聚合图来捕获话语之间的常识知识依赖，却忽略了对视频和音频的上下文依赖建模。显然，以上研究均通过构建不同的图结构来实现模态内的上下文依赖或模态间的特征交互，但这些方法在构建图结构时往往忽略了会话的动态性，多通过设计基于任意节点或固定前后向窗口的密集连接来实现话语级的上下文关联聚合，导致当前话语关联的上下文信息出现缺失或冗余。此外，一些方法还忽略了语音、视频模态内的上下文依赖挖掘，显然不利于实现多模态信息的有效交互与融合。然而会话

是动态的且每个话语所依赖的上下文范围是不同的，则如何自适应地学习图关系以及有效挖掘模态内与模态间的关系是必要的。本文将构建同质的动态图来自适应学习模态内以及话语级多模态特征的上下文依赖，且构建双流图模块来缓解模态间的异构性。

1.2 对比学习

在会话情绪识别任务中，会话所表达出的情绪语义较为相似[32]，而对比学习能够通过不断缩小正样本之间以及扩大负样本之间的差异来学习不同类别之间可区分的类别边界信息，是一种探索判别嵌入的表示学习技术[33]。为了缓解数据分布不平衡的问题，Wei 等人[31]采用欧氏距离将话语的正类样本和负类样本进行对比学习，缩小正样例与该话语的差距，并扩大负样例与该话语的差距，以学习更具判别性的类边界信息。为了考虑会话者自身的历史影响，Xu 等人[34]利用余弦相似度将与该话语是同一讲话者的会话内容与其他话语进行对比以区分当前说话人与之前说话人是否为同一人。Tu 等人[35]利用余弦相似度将该话语的伪标签正负样本进行对比学习以捕获话语间的细微相关性和判别性。Shou 等人[6]利用互信息将该样例的正负类样例进行对比学习以获得更具判别性的嵌入表示。Hu 等人[36]通过乘积操作将多模态融合特征与各个模态进行对比，以增强模态间的相互影响，并放大样本与融合表征的差异。这些对比学习方法通过利用标签信息将属于同一类样例的特征聚集，而不属于同一类的远离。但是这类方法往往通过直接压缩类内特征空间，不可避免地损坏细粒度的类内特征，从而降低模型的泛化能力。另外，通常仅使用分类交叉熵损失函数用于 MERC[8]来学习模态的判别性，难以有效应对数据中的类别不平衡的问题。为了缓解上述问题，本文将利用由对比损失和交叉熵损失构成的软对比损失来学习更具判别性的模态表示，并结合对抗学习策略来提高模型的泛化能力。

1.3 对抗训练

泛化的多模态表示一直是 MERC 任务中的一个严峻挑战。对抗训练是一种广泛使用的正则化方法，往往通过使用最小-最大的博弈训练来提高模型的鲁棒性[37]。例如，Szegedy 等人[38]通过在原始样例上添加扰动生成对抗样例，并在对抗样例和原始样例的混合样例上以对抗的方式训练神经网络。为了缓解一些神经网络通常错误地分类扰动样例，Goodfellow 等人[39]进一步提出了一种快速梯度符号方法，该方法基于模型的梯度信息生成对抗样本。而 Miyato 等人[40]将对抗性和虚拟对抗性训练扩展到文本域，主要在词嵌入层上应用扰动。生成对抗网络也被引入多模态会话情绪识别任务中，如 Ren 等人[12]为每个模态设置一个编码器和判别器，所有模态共享 1 个生成器，在每个模态上进行对抗训练来挖掘多模态信息的共性和多样性。而 Meng 等人[41]提出了一个类边界增强表示学习方法，首先在各个模态上加噪生成对抗样例，然后利用生成器和反向生成器间的一致性损失来确保每个模态的对抗样例和原始样例的一致性，但对模态的异构性考虑不足。为了缓解模态的异构问题，Shou 等人[6]为 3 个模态分别构建了生成器和判别器，其中每个模态的生成器输入是其他 2 种模态，而对应模态的判别器输入是这 3 种模态的特征，以此通过对抗训练来缓解不同模态特征之间的数据分布差异，进而增强模型的鲁棒性。显然，扰动对抗和生成对抗是从不同的角度来进行对抗训练以提高模型的鲁棒性，但考虑到会话上下文的连贯属性，虽然生成对抗网络缓解了不同模态的异构性，但是会损害会话的连贯性。因此，本文将通过采用扰动对抗来提高模型的泛化，以在扰动对抗学习过程中确保会话的连贯性，具体地，通过将扰动应用于单模态特征提取器的隐层中以生成对抗样本，并在原始样本和对抗样本上联合使用软对比损失来有效地学习类扩展表征和标签一致性特征，以提高模型的泛化能力。

2 方法描述

形式上，一个会话 U 是由 N 个连续的语句 $\{u_1, u_2, \cdots, u_N\}$ 和 S 个说话人 $\{p_1, p_2, \cdots, p_S\}$ 组成。每个话语 u_i 由参与方 $p_{\phi(u_i)} \in \{p_1, p_2, \cdots, p_S\}$ 说出，其中 ϕ 是将语句映射到对应的说话人。此外，u_i 涉及文本(T)、音频(A)和视频(V)模态，其特征表示分别为 $\boldsymbol{u}_i^{\mathrm{T}} \in \mathbb{R}^{d_{\mathrm{T}}}$，$\boldsymbol{u}_i^{\mathrm{A}} \in \mathbb{R}^{d_{\mathrm{A}}}$ 和 $\boldsymbol{u}_i^{\mathrm{V}} \in \mathbb{R}^{d_{\mathrm{V}}}$。会话中所有话语的文本模态、音频模态和视频模态的序列表示为 $\boldsymbol{U}_{\mathrm{T}} = (\boldsymbol{u}_1^{\mathrm{T}}, \boldsymbol{u}_2^{\mathrm{T}}, \cdots, \boldsymbol{u}_N^{\mathrm{T}}) \in \mathbb{R}^{N \times d_{\mathrm{T}}}$，$\boldsymbol{U}_{\mathrm{A}} = (\boldsymbol{u}_1^{\mathrm{A}}, \boldsymbol{u}_2^{\mathrm{A}}, \cdots, \boldsymbol{u}_N^{\mathrm{A}}) \in \mathbb{R}^{N \times d_{\mathrm{A}}}$ 且 $\boldsymbol{U}_{\mathrm{V}} = (\boldsymbol{u}_1^{\mathrm{V}}, \boldsymbol{u}_2^{\mathrm{V}}, \cdots, \boldsymbol{u}_N^{\mathrm{V}}) \in \mathbb{R}^{N \times d_{\mathrm{V}}}$。给定一个话语，MERC 旨在基于该话语的多模态表示 $\{\boldsymbol{u}_i^{\mathrm{T}}, \boldsymbol{u}_i^{\mathrm{A}}, \boldsymbol{u}_i^{\mathrm{V}}\}$ 以及会话上下文预测每个话语的情感标签 y_i。通常，每个话语代表一种情绪，由于不同数据集的会话主题不一致，其包含的情感丰富度也不同，因此，不同的数据集具有不同的情绪类别，MELD 数据集[42]包含 7 种情绪标签，IEMOCAP 数据集[43]包含 6 种情绪标签。

图 3 给出了本文提出的 ASDG 方法的模型细节

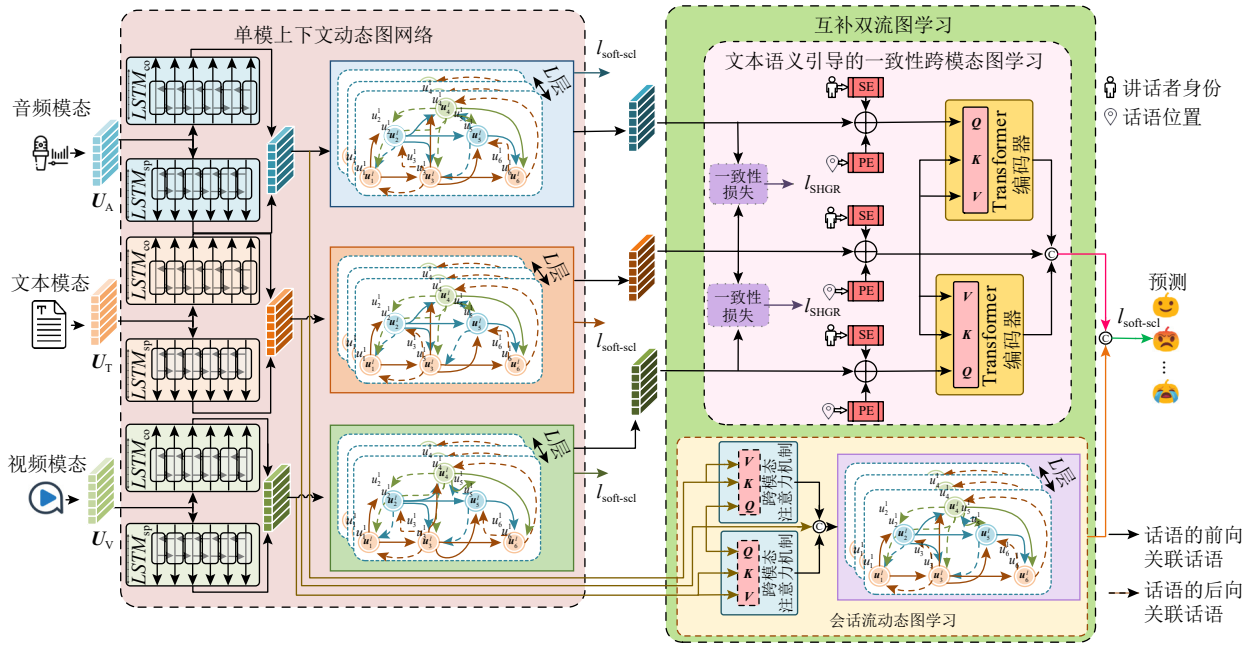


Fig. 3 Model details of adversarial soft-contrast modulated dynamic graph learning method

图 3 对抗软对比调制动态图学习方法的模型细节

展示。具体地，基于预训练模型提取话语级的单模态特征表示后，利用双向 LSTM 来初步建模有关文本、音频和视频的会话上下文依赖关系以及讲话者依赖关系，然后同时构建文本话语图、语音话语图以及视频话语图来自适应地建模文本、音频、视频的模态内动态依赖，且设计文本语义引导的一致性约束的跨模态图学习来探索不同模态间特征的有机融合，主要利用 Transformer 结构实现文本-音频、文本-视频的跨模态交互；同时设计会话流动态图学习来发掘多模态话语级语境的动态依赖，具体使用跨模态注意力机制来分别实现文本对语音的引导以及文本对视频的引导，并进一步构建话语图来实现信息的传播，最后集成双流图的输出用于会话情绪判断。

此外，针对模型的优化，提出使用软对比训练策略来学习相同类别的一致性以及不同类别的判别性，并将其作用在 3 个模态的话语图输出上以及互补双流图的输出上。另外，利用对抗训练策略生成每个话语的对抗样本以扩大类空间表示，在对抗样本上也使用软对比学习来进行训练，并使对抗样本的模型训练和原始样本保持一致，这种联合原始样本和对抗样本进行软对比学习使得类内特征更具判别性，从而提高模型的泛化性和鲁棒性。接下来，将详细阐述 ASDG 模型的细节。

2.1 单模态的会话上下文动态图学习

为了捕获会话的语境感知依赖信息和说话者感知依赖信息，以及处理维度不一致的模态数据，使用双重双向 LSTM 对文本、音频和视频数据分别进行会话上下文特征编码和讲话者依赖特征编码。以文本为例来阐述特征编码的过程。

在会话中讲话者的情感状态不仅受到其他讲话者的影响，还受自身影响。为了建模说话人自身的上下文依赖，首先分别抽取各个讲话者在会话中说的话语，然后使用双向 LSTM 来学习说话人感知特征。但是如果说话人的话语数小于预定义的整数阈值ξ，这些说话人的特征直接由共享的通用说话人向量 $\boldsymbol{o}$ 表示。因此，说话人感知特征$\boldsymbol{c}_i^{\mathrm{sp}}$的计算过程如下：

$$\boldsymbol{c}_i^{\mathrm{sp}}, \boldsymbol{h}_{s,j}^{\mathrm{sp}} = \begin{cases} \boldsymbol{o}, \text{ 无}, & \text{if } |\boldsymbol{U}_s| < \xi, \\ \overrightarrow{LSTM}_{\mathrm{sp}}(\boldsymbol{u}_i, \boldsymbol{h}_{s,j-1}^{\mathrm{sp}}), & j \in [1, |\boldsymbol{U}_s|], \text{ 其他}, \end{cases} \tag{1}$$

其中$\overrightarrow{LSTM}_{\mathrm{sp}}$表示通过双向 LSTM 获取说话人感知依赖特征，$\boldsymbol{h}_{s,j}^{\mathrm{sp}}$是说话人$p_s$发出的第$j$个话语的隐状态。$s=\phi(\boldsymbol{u}_i)$是讲话者身份。$\boldsymbol{U}_s$指一个会话中$p_s$说的所有话语。语境感知特征$\boldsymbol{c}_i^{\mathrm{co}}$的计算过程为

$$\boldsymbol{c}_i^{\mathrm{co}}, \boldsymbol{h}_i^{\mathrm{co}} = \overrightarrow{LSTM}_{\mathrm{co}}(\boldsymbol{u}_i, \boldsymbol{h}_{i-1}^{\mathrm{co}}), \tag{2}$$

其中$\overrightarrow{LSTM}_{\mathrm{co}}$表示通过双向 LSTM 获取语境感知特征，$\boldsymbol{h}_i^{\mathrm{co}}$是话语$\boldsymbol{u}_i$的隐状态。

然后将说话人感知特征和语境感知特征进行拼接以形成话语的文本上下文表示$\boldsymbol{c}_i^{\mathrm{T}} = [\boldsymbol{c}_i^{\mathrm{sp}}; \boldsymbol{c}_i^{\mathrm{co}}]$。如此，针对音频和视频模态经过同样操作，以生成音频的会话上下文表示$\boldsymbol{c}_i^{\mathrm{A}}$以及视频的会话上下文表示$\boldsymbol{c}_i^{\mathrm{V}}$。

为了提高 ASDG 的泛化能力，通过加噪的方式生成对抗样例来扩展每个类的表示空间以缓解类分

布不平衡的问题,这尤其对少类样例的训练是很有必要的。但是对于会话上下文依赖的场景,直接在样本上添加扰动噪声会很大程度上摧毁不同话语之间的连贯性,因此将扰动噪声作用于经过原始样本训练后的双向 LSTM 的隐藏层,从而借助上下文依赖的隐层输出以得到平滑的表示空间。特别强调的是,原始样本和对抗样本都需经过 ASDG 来学习,以联合的形式有效地学习类内的一致性特征,并通过软对比学习保留细粒度的精细类内特征,以达到更有效的泛化效果,具体见 2.4 节。

当前基于 GNN 的会话情绪识别研究往往以任一节点为中心构建密集连接的图[6-25]或设置固定超参数窗口大小的图[7-9],然而这种固定形式的构图方式会导致获取的上下文语境信息冗余或者不足,且难以适应动态的会话场景。为了精准地获取任一话语关联的上下文信息并模拟实际会话中上下文的相互影响,设计一个动态话语图来实现各自模态话语间的情感信息扩散,使得每个话语可以动态地学习上下文依赖范围。单模态的会话上下文动态图学习包含文本话语图、音频话语图、视频话语图。由于每个话语图都有一一对应的 3 个模态信息,因此该 3 个话语图的图结构是一致的。下面以文本话语图的构建过程来展开阐述。

在话语图的构建过程中,话语之间的边是根据话语人的话语密度来动态建立的,其定义是说话人所说的话语数占整个话语总数的比例。根据该规则,对任意一个会话计算每个说话人的话语密度,并记为 $\boldsymbol{D}=(d_1,d_2,\cdots,d_p)$,其中 p 是该会话的参与者数量,d_i 是说话人 i 的话语密度。在图 1 所示的对话中,第 5 个话语(讲话者 A 所说)的前向语境是其前面 3 个说话人形成的 4 个话语,讲话者 A, B, C 的话语数量为 2, 1, 1,此时他们的话语密度为 1/2, 1/4, 1/4。同理,第 4 个话语的前向语境是其前面 2 个说话人形成的 3 个话语,讲话者 A, B 的话语数量为 2, 1,此时他们的话语密度为 2/3, 1/3。

动态话语图表示为 $G=(\boldsymbol{C},\boldsymbol{E},\boldsymbol{W})$,节点是会话中的语句,即 $\boldsymbol{C}=(\boldsymbol{c}_1^{\mathrm{T}},\boldsymbol{c}_2^{\mathrm{T}},\cdots,\boldsymbol{c}_N^{\mathrm{T}})$,且边 $(i,j,e_{i,j})\in\boldsymbol{E}$ 表示信息从节点 $\boldsymbol{c}_i^{\mathrm{T}}$ 传播到节点 $\boldsymbol{c}_j^{\mathrm{T}}$,其中 $e_{i,j}\in\{0,1\}$ 表示边是否存在,其中 $\boldsymbol{W}$ 是边权重,构建话语图的过程如下:

1)节点。话语 $\boldsymbol{c}_i^{\mathrm{T}}$ 代表话语图中的初始节点。

2)边。以下规则确定动态话语图中边的连接。

①基于目标话语在会话序列中的相对位置,可将其上下文划分为前向语境和后向语境。

②针对这 2 个语境片段,需要分别统计各说话人在每个片段中的话语分布密度,$d_{\min}^{\mathrm{f}}=\min(d_1^{\mathrm{f}},d_2^{\mathrm{f}},\cdots,d_{p1}^{\mathrm{f}})$,$d_{\min}^{\mathrm{b}}=\min(d_1^{\mathrm{b}},d_2^{\mathrm{b}},\cdots,d_{p2}^{\mathrm{b}})$,其中 $d_{\min}^{\mathrm{f}}$ 和 $d_{\min}^{\mathrm{b}}$ 分别是前向和后向会话中的最小话语密度值,其中 $p1$ 表示前向会话涉及的讲话者数量,$p2$ 代表后向会话包含的讲话者数量。如图 1 所示,在前向对话分析中,第 5 句话的最小话语密度为 1/4,而第 4 句话的对应数值为 1/3。

③以前向注意过程为例来阐述目标话语需要关注的前向话语。首先,初始化一个空列表 h,并依次将目标话语之前的且按时间顺序最邻近的语句加入该列表,然后计算 h 中出现的讲话人的话语密度,列表 h 中的最小话语密度值用 $d_{t,\min}^{\mathrm{f}}$ 表示,如果某个说话人的话语未出现在 h 中,则该说话人的话语密度记为 0。随着话语连续添加到列表 h, $d_{t,\min}^{\mathrm{f}}$ 的值也随之而改变。如果 h 中的最小话语密度值满足 $d_{t,\min}^{\mathrm{f}}<d_{\min}^{\mathrm{f}}$,则继续搜索。而当一个话语添加在 h 中后并且 $d_{t,\min}^{\mathrm{f}}\geqslant d_{\min}^{\mathrm{f}}$ 时,则停止搜索,此时列表 h 中的话语是目标话语需要关注的上下文。如图 1 左侧所示,在计算第 5 个话语的前向会话语境时,首先将第 4 个话语添加到 h,此时 3 个说话人的最小话语密度是 0,低于阈值 1/4,因此第 4 个话语被标记为需要关注的内容,并继续执行搜索过程,直到将讲话者 B 所说的话语添加到列表 h 时,3 个说话人的最小话语密度为 1/3,超过了阈值 1/4,此时搜索过程终止。那么,第 5 句话要关注的前向语境是第 4 句话、第 3 句话以及第 2 句话。类似地,如图 1 右侧所示,第 4 个话语要关注的话语是第 3 和第 2 个话语。动态话语图的构建过程如算法 1 所示。

算法 1. 动态话语图构建算法。

输入:完整会话列表 $\{c_1,c_2,\cdots,c_n\}$,会话中每个会话的说话人身份;

输出:所要关注的前向上下文列表 H。

① 初始化空列表 H;

② for $i\leftarrow 1$ to $n-1$ do /*话语 0 没有前向语境,从话语 1 开始遍历*/

③ 空列表 h;

④ $P_i=p_{0\sim i-1}$; /*在轮次 0~i−1 出现的讲话者*/

⑤ $NP_i=np_{P_i}^{0\sim i-1}$; /*在轮次 0~$i$−1 会话中每个讲话者的话语数量*/

⑥ $d_{\min}^{\mathrm{f}}\leftarrow\min(NP_i/i)$; /*话语 i 的前向会话列表中的最小讲话者密度*/

⑦ for $j\leftarrow i-1$ to 0 do

⑧ $h.append(c_j)$; /*话语 c_j 添加到列表 h*/

⑨ $NP_i^t=np_{P_i}^{j\sim i-1}$; /*时刻 t 时,h 中每个讲话者的话语数量*/

⑩ $d^{f}_{t,\min} \leftarrow \min\ (NP^{t}_{i}/i-j)$; /*时刻 t 时，h 列表中的最小讲话者密度*/

⑪ if $d^{f}_{t,\min} < d^{f}_{\min}$ then /*时刻 t 时如果满足条件则继续搜索*/

⑫ 继续;

⑬ else

⑭ 中止;

⑮ end if

⑯ end for

⑰ $H.\ append(h)$;

⑱ end for

⑲ return H。

通过这种方法来构建动态话语图，可以使得每个话语感知到前向会话中所有出现的说话人表达的内容，每个话语能够充分获取上下文信息且能避免上下文信息的冗余。

3)边的权重。由于图注意力网络(graph attention network, GAT)可以自适应地学习节点之间的强弱关联性。因此，本文利用该网络来更新节点表示，其中边权重的计算过程为：

$$\alpha_{i,j}=\frac{\exp(LeakyReLU(\boldsymbol{a}^{\mathrm{T}}[\boldsymbol{W}^{\mathrm{T}}\boldsymbol{c}_{i}^{\mathrm{T}}\|\boldsymbol{W}^{\mathrm{T}}\boldsymbol{c}_{j}^{\mathrm{T}}]))}{\sum_{g\in H_i}\exp(LeakyReLU(\boldsymbol{a}^{\mathrm{T}}[\boldsymbol{W}^{\mathrm{T}}\boldsymbol{c}_{i}^{\mathrm{T}}\|\boldsymbol{W}^{\mathrm{T}}\boldsymbol{c}_{g}^{\mathrm{T}}]))}, \tag{3}$$

其中$\boldsymbol{W}^{\mathrm{T}}$是线性变换的权值。$H_i$表示图中第 i 个话语所要关注的节点集合。$LeakyReLU$ 是激活函数，$\boldsymbol{a}$是一个可训练的注意力权重。

4)图学习。将图节点和边权重以及边关系作为 GAT 的输入，并执行可学习的线性变换和多头注意力机制逐层更新节点。其中第 i 个节点经过 l 层 GAT 的聚合更新过程表示如下：

$$\boldsymbol{x}_{i}^{\mathrm{T}(l)}=\mathop{\Big\|}_{head=1}^{Head}\sigma\left(\sum_{g\in H_i}\alpha_{i,g}^{head}\boldsymbol{W}\boldsymbol{c}_{g}^{\mathrm{T}(l-1)}\right), \tag{4}$$

其中l代表第l层 GAT。$Head$ 是注意力机制的头的数量。$\|$为级联运算，σ为非线性激活函数。通过以上操作，每个话语能够动态地感知语境信息，从而获得精准且充分的文本模态上下文特征$\boldsymbol{X}^{\mathrm{T}}$。

与此类似，针对音频模态和视频模态，利用与文本模态相同的话语图构建方式以及图注意力学习方式，来生成音频模态的上下文特征$\boldsymbol{X}^{\mathrm{A}}$以及视频模态的上下文特征$\boldsymbol{X}^{\mathrm{V}}$。

2.2 互补双流图学习

在 2.1 节中，我们仅充分挖掘了各自不同模态的上下文依赖特征，但没有关注会话上下文中不同模态的有效融合。为此，进一步设计互补双流图学习架构，该架构主要由 2 个分支流构成，如图 3 所示，即文本语义引导的一致性跨模态图学习分支和会话流动态图学习分支。一方面从模态的一致性角度来挖掘以文本语义为中心的跨模态交互，缓解不同模态特征融合中的异质鸿沟，另一方面从会话的动态性角度建模会话中多模态话语级信息流的动态关联，随后集成双流分支的输出进一步促进模态间的交互。

2.2.1 文本语义引导的一致性跨模态图学习

由于语、视、音模态信息存在本质上的异构特性，使得各个模态的数据分布不一致，显然如何挖掘 3 者之间的一致性语义是相对困难的。目前研究者倾向于使用 Transformer 结构实现跨模态交互[44]，但其往往忽略了模态间相关性挖掘的主导性探索。另外，一些相关研究已经证明文本模态的语义信息比音频模态或视频模态更丰富[8,26]，因此，以文本为锚点来分别探索文本与音频的相关性以及文本与视频的相关性，具体地，使用改进的 SHGR(soft hirschfeld Gebelein-Rényi)一致性损失[45]来探索模态间的相关性：

$$l_{\mathrm{SHGR}}=-\sum_{M\in\{\mathrm{A,V}\}}E[\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}^{M}]-\frac{1}{2}\mathrm{tr}(cov(\boldsymbol{X}^{\mathrm{T}})cov(\boldsymbol{X}^{M})), \tag{5}$$

其中$\boldsymbol{X}^{\mathrm{T}}$, $\boldsymbol{X}^{\mathrm{A}}$, $\boldsymbol{X}^{\mathrm{V}}$分别表示从动态图中获取的文本特征、音频特征以及视频特征。E 代表样本均值，cov 代表样本方差。

在一致性损失式(5)的约束下，实现文本与音频的跨模态交互以及文本与视频的跨模态交互。由于话语序列是有时间信息存在的且不同的话语是由不同的说话人发出，则明确话语的位置信息以及赋予话语的说话者属性是非常有必要的。因此，利用话语的位置信息和讲话者嵌入信息来增强模态的上下文特征，具体地，位置嵌入的计算为：

$$\begin{aligned} pe_{(pos,2i)}&=\sin\left(\frac{pos}{10\,000^{2i/d}}\right),\\ pe_{(pos,2i+1)}&=\cos\left(\frac{pos}{10\,000^{2i/d}}\right), \end{aligned} \tag{6}$$

其中 pos 是话语索引，i 是维度索引。同时，将说话人 p_j映射为一个向量，如式(7)所示：

$$\boldsymbol{se}_{j}=\boldsymbol{V}_{\boldsymbol{se}}one(p_j),\ j=1,2,\cdots,S, \tag{7}$$

其中$\boldsymbol{V}_{\boldsymbol{se}}$是一个可训练的讲话者嵌入矩阵，$one(p_j)$是说话人$p_j$的独热编码。因此，一个会话的讲话者嵌入可以表示为$\boldsymbol{SE}=(\boldsymbol{se}_{\phi(u_1)},\boldsymbol{se}_{\phi(u_2)},\cdots,\boldsymbol{se}_{\phi(u_N)})$。

由此可得到模态 M 的话语位置嵌入和说话人嵌入的上下文表征序列$\boldsymbol{Q}^{M}$，如式(8)所示。

$$\boldsymbol{Q}^{M}=\boldsymbol{X}^{M}+\boldsymbol{PE}+\boldsymbol{SE}。 \tag{8}$$

然后基于$\boldsymbol{Q}^{M}$使用 Transformer 编码器来实现跨模态交互,具体将$\boldsymbol{Q}^{\mathrm{T}}$作为查询,$\boldsymbol{Q}^{M}$作为键和值,计算过程为

$$\boldsymbol{Q}^{\mathrm{T}\to M}=Transformer(\boldsymbol{Q}^{\mathrm{T}},\boldsymbol{Q}^{M},\boldsymbol{Q}^{M})\in\mathbb{R}^{N\times d}, \tag{9}$$

其中$\boldsymbol{Q}^{\mathrm{T}}$代表文本模态,$M\in\{\mathrm{A},\mathrm{V}\}$。以此实现文本-音频以及文本-视频的跨模态交互。最后将$\boldsymbol{Q}^{\mathrm{T}}$,$\boldsymbol{Q}^{\mathrm{T}\to\mathrm{A}}$,$\boldsymbol{Q}^{\mathrm{T}\to\mathrm{V}}$进行拼接得到最后的输出表征$\boldsymbol{Q}^{\mathrm{up}}$。

2.2.2 会话流动态图学习

文本语义引导的一致性跨模态图学习更多关注跨模态交互融合,却忽略了会话中话语级多模态信息的流动性,因此,进一步设计会话流动态图学习来建模话语级多模态特征间的动态语境关联。

针对会话流动态图的构建,多模态会话流动态图$G=(\boldsymbol{Z},\boldsymbol{E},\boldsymbol{W})$,节点$\boldsymbol{Z}$是每个话语的文本、音频、视频 3 模态注意力拼接后的特征,且边$(i,j,e_{i,j})\in\boldsymbol{E}$表示 2 节点之间的信息传播,其中$e_{i,j}\in\{0,1\}$表示边是否存在,$\boldsymbol{W}$是边的权重,构建会话流动态图的规则如下:

1)节点。多模态会话流动态图中的每个节点是话语级多模态表示,若直接将 3 模态表示进行拼接,既不能实现话语中细粒度的模态间对齐与融合,也难以有效应对融合时出现的冲突问题,显然不利于后续关于会话流的上下文建模。此外,由于文本比音频和视频的语义更丰富,因此,将构建以文本为中心的跨模态注意力机制来分别实现文本-音频的语义对齐融合以及文本-视频的语义对齐融合。

具体而言,利用跨模态注意力机制来自适应地学习话语级文本模态引导音频的关联对齐以及话语级文本模态引导视频的关联对齐。将$\boldsymbol{C}^{\mathrm{T}}$作为查询,$\boldsymbol{C}^{M}$作为键和值,计算如下所示:

$$\boldsymbol{Z}^{M}=\mathop{\Big\|}_{head=1}^{Head} softmax\left(\frac{(\boldsymbol{W}_1^{head}\boldsymbol{C}^{\mathrm{T}})^{\mathrm{T}}(\boldsymbol{W}_2^{head}\boldsymbol{C}^{M})}{\sqrt{d/Head}}\right)(\boldsymbol{W}_3^{head}\boldsymbol{C}^{M})^{\mathrm{T}}, \tag{10}$$

其中$M\in\{\mathrm{A},\mathrm{V}\}$,$Head$是跨模态注意头的数量,$\boldsymbol{W}_1^{head}$,$\boldsymbol{W}_2^{head}$,$\boldsymbol{W}_3^{head}\in\mathbb{R}^{d/Head\times N}$是查询、键和值的权重矩阵。$\boldsymbol{C}^{\mathrm{T}}$,$\boldsymbol{C}^{M}$是双向 LSTM 的输出。然后再聚合$\boldsymbol{C}^{\mathrm{T}}$,$\boldsymbol{Z}^{\mathrm{A}}$,$\boldsymbol{Z}^{\mathrm{V}}$作为会话流动态图的节点表示。

2)边。根据单模态动态图学习模块中的算法 1 创建节点之间的连接,以动态地建模每个话语级多模态特征的依赖范围。

3)边的权重。根据单模态动态图学习模块中的方法计算 2 节点之间的权重,如式(3)所示。

4)图学习。根据单模态动态图学习模块中的图聚合方式来逐层更新节点表示,如式(4)所示。

通过会话流动态图的学习,得到了文本-音频-视频的上下文动态关联特征$\boldsymbol{P}^{\mathrm{do}}$。

文本语义引导的一致性跨模态图学习和会话流动态图学习从不同的角度实现会话中的多模态交互。为了更有效地发挥双分支流图的功能,进一步将其整合以挖掘它们之间的协同效应与互补优势,如下:

$$\boldsymbol{H}^{\mathrm{do}\leftrightarrow\mathrm{up}}=[\boldsymbol{P}^{\mathrm{do}}\|\boldsymbol{Q}^{\mathrm{up}}]\in\mathbb{R}^{N\times 6\times d}。 \tag{11}$$

至此,获得了针对会话场景的多模态序列特征表示$\boldsymbol{H}^{\mathrm{do}\leftrightarrow\mathrm{up}}=(\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_N)$。

2.3 情绪分类

为了计算会话中每个多模态话语的情绪类别,将$\boldsymbol{H}^{\mathrm{do}\leftrightarrow\mathrm{up}}$输入到带有全连接层和 *softmax* 层的分类器:

$$\hat{\boldsymbol{Y}}=softmax(\boldsymbol{W}_{\mathrm{e}}\boldsymbol{H}^{\mathrm{do}\leftrightarrow\mathrm{up}}+\boldsymbol{b}_{\mathrm{e}}), \tag{12}$$

其中$\boldsymbol{W}_{\mathrm{e}}$和$\boldsymbol{b}_{\mathrm{e}}$是可训练参数。$\hat{\boldsymbol{Y}}=(\hat{\boldsymbol{y}}_1,\hat{\boldsymbol{y}}_2,\cdots,\hat{\boldsymbol{y}}_N)$,$\hat{\boldsymbol{y}}_i$是话语$\boldsymbol{u}_i$的情感概率。最后,选择$\arg\max(\hat{\boldsymbol{y}}_i)$作为$\boldsymbol{u}_i$的预测的情绪标签。

2.4 优化目标

学习泛化且鲁棒的判别性表征一直是情感分析相关任务的难点。由于 MERC 任务的情绪类别较多,且情感语义较为相似,因此,需要设计更为合适的优化策略来提取更具判别性的多模态融合表征,以提高模型的泛化性。软对比学习通过捕捉同类样本间的相似性,将其与其他类的样本进行对比来提高模态特征的判别性。然而,多模态会话数据集存在严重的类别分布不平衡问题,仅用软对比学习会强制压缩每个类的特征空间,进而损害细粒度的多样性特征,尤其给少类样本的学习增加了困难。因此,考虑构建对抗训练的方式来扩展类空间,以增强模型的鲁棒性。综上,提出对抗性软对比训练机制,即首先在原始样本上利用软对比学习训练模型,再基于单模态特征提取器的隐层梯度来获取扰动信息,然后通过扰动对抗训练生成对抗样本以丰富样本特征,并在对抗样本上利用软对比学习进一步优化模型以提高泛化能力。

软对比学习的损失是交叉熵损失和带有权衡参数λ的对比学习损失之和。给定一个小批量样本集Ω,定义$\vartheta(i)=\{e\in\Omega\backslash\{i\}:y_e=y_i\}$是$\Omega$中与第$i$个样例标签相同的样本集合,但不包含$i$样本,软对比学习的过程如下所示:

$$l_{\text{soft-scl}}=l_{\mathrm{ce}}+\lambda l_{\mathrm{scl}}, \tag{13}$$

$$l_{\mathrm{ce}}=-\frac{1}{|\Omega|}\sum_{i=1}^{\Omega}\sum_{j=1}^{K}\boldsymbol{y}_{i,j}\lg\left(\hat{\boldsymbol{y}}_{i,j}^{\mathrm{soft}}\right), \tag{14}$$

$$\widehat{\boldsymbol{Y}}^{\mathrm{soft}}=softmax(\boldsymbol{W}^{\mathrm{soft}}\boldsymbol{\Psi}+\boldsymbol{b}^{\mathrm{soft}}), \tag{15}$$

$$l_{\mathrm{scl}}=-\sum_{i\in\Omega}\frac{-1}{|\vartheta(i)|}\sum_{e\in\vartheta(i)}\lg\frac{\exp(sim(\boldsymbol{\varphi}_i,\boldsymbol{\varphi}_e)/\tau)}{\sum_{j\in\Delta(i)}\exp(sim(\boldsymbol{\varphi}_i,\boldsymbol{\varphi}_j)/\tau)},\tag{16}$$

其中$\boldsymbol{W}^{\mathrm{soft}}\in\mathbb{R}^{d\times K}$和$\boldsymbol{b}^{\mathrm{soft}}\in\mathbb{R}^{K}$是可训练的参数，$K$是情绪标签的类别数。$\boldsymbol{y}_{i,j}$和$\hat{\boldsymbol{y}}_{i,j}^{\mathrm{soft}}$分别表示类别索引$j$下的独热向量和概率向量。$\Delta(i)=\Omega\backslash\{i\}$，$\boldsymbol{\varphi}_i$是网络模块输出的第$i$个样本的特征表示。$sim(\cdot,\cdot)$是相似性函数，$\tau$是控制类分散的一个标量。

然后，将软对比损失分别作用于单模态会话上下文动态图网络的3个输出上（如图3所示）来充分挖掘各自模态的判别性语义，便于后续在文本语义主导的一致性损失下，跨模态图学习能够充分且有效地挖掘不同模态的交互融合特征。此外，为了进一步使得话语级的多模态特征更具判别性，将软对比损失作用于互补双流图学习的融合输出上，进一步挖掘更为判别的话语级多模态融合特征。总体而言，本文提出的模型在原始样本上训练的总损失为

$$l_{\mathrm{or}}=l_{\mathrm{SHGR}}+l_{\text{soft-scl}}^{\mathrm{T}}+l_{\text{soft-scl}}^{\mathrm{A}}+l_{\text{soft-scl}}^{\mathrm{V}}+l_{\text{soft-scl}}^{\mathrm{TAV}},\tag{17}$$

其中$l_{\text{soft-scl}}^{\mathrm{T}}$，$l_{\text{soft-scl}}^{\mathrm{A}}$，$l_{\text{soft-scl}}^{\mathrm{V}}$，$l_{\text{soft-scl}}^{\mathrm{TAV}}$分别代表文本话语图输出、音频话语图输出、视频话语图输出以及分类器上的软对比损失。

在每个批训练步过程，首先在原始样本上利用l_{o}损失来训练网络，然后采用扰动对抗训练策略生成关于原始样本的对抗样本特征以扩展类样本空间。现有研究将噪声扰动信息放在原始样本上[41]，然而这会摧毁上下文的语义连贯性，使得语义难以理解。为了在扰动对抗过程中保护会话的语义连贯性，受到PGD对抗训练策略[46]的启发，本文将扰动$\boldsymbol{r}$放在双向LSTM的上下文感知的隐藏层，其包括3个门控层和1个记忆细胞层。给定L_q范式球和$\boldsymbol{r}$的半径ε，通过线性近似，并在软对比损失目标下，对抗样例的训练过程如下所示：

$$\begin{aligned}&\min_{\theta}E_{(u,y)\sim\Omega}\max_{\|\boldsymbol{r}\|_q\leqslant\varepsilon}l_a^{\mathrm{adv}}(u+\boldsymbol{r},y;\theta),\\&\boldsymbol{g}=\nabla_u\lg p(y|u;\hat{\theta}),\\&\boldsymbol{r}=-\varepsilon\boldsymbol{g}/\|\boldsymbol{g}\|_q。\end{aligned}\tag{18}$$

式(18)中的内部最大化是生成上下文级别的最坏情况对抗样本，外部最小化是针对对抗样本训练一个鲁棒的网络。在引入PGD对抗之后，软对比损失可以从上下文依赖的输入中进一步学习更多样化的特征和平滑的表示空间，并增强模型对上下文扰动的鲁棒性。对抗样例上的优化损失$l_{\mathrm{ad}}^{\mathrm{adv}}$与原始样例上的优化损失$l_{\mathrm{or}}$计算一致，在对抗样本上的总损失$l_{\mathrm{ad}}^{\mathrm{adv}}$如下所示：

$$l_{\mathrm{ad}}^{\mathrm{adv}}=l_{\mathrm{SHGR}}^{\mathrm{adv}}+l_{\text{soft-scl}}^{\mathrm{T,adv}}+l_{\text{soft-scl}}^{\mathrm{A,adv}}+l_{\text{soft-scl}}^{\mathrm{V,adv}}+l_{\text{soft-scl}}^{\mathrm{TAV,adv}},\tag{19}$$

其中$l_{\mathrm{SHGR}}^{\mathrm{adv}}$，$l_{\text{soft-scl}}^{\mathrm{T,adv}}$，$l_{\text{soft-scl}}^{\mathrm{A,adv}}$，$l_{\text{soft-scl}}^{\mathrm{V,adv}}$，$l_{\text{soft-scl}}^{\mathrm{TAV,adv}}$分别是对抗样例上一致性损失、文本话语图输出、音频话语图输出、视频话语图输出以及互补双流图输出上的软对比损失。

综上，本文模型ASDG的优化总目标函数为

$$l=l_{\mathrm{or}}+l_{\mathrm{ad}}^{\mathrm{adv}}。\tag{20}$$

3 实　　验

3.1 数据集和评估指标

使用MELD[42]和IEMOCAP[43]这2个数据集来验证所提模型ASDG的有效性，这些数据集目前在MERC的研究中得到了广泛的使用[8,25]，数据集的详细信息如表1所示，为了更公平地比较ASDG的性能，本文采用现有MERC研究使用的数据集划分设置。

Table 1　Statistics of Datasets

表1　数据集的统计

数据集	会话数量		话语数量		类别	会话者人数
	训练+验证	测试	训练+验证	测试		
IEMOCAP	120	31	5 810	1 623	6	10
MELD	1 153	280	11 098	2 610	7	9

1）IEMOCAP。该数据集由151个对话和7 433个话语组成，每句话都被标记为6种情绪快乐、悲伤、中性、愤怒、兴奋和沮丧中的1种，每个会话涉及2个讲话者。

2）MELD。该数据集是从电视剧《老友记》中搜集的9人对话数据集，包含1 433个对话和13 708个句子，包含中性、惊讶、恐惧、悲伤、喜悦、厌恶和愤怒7种情绪，每句话被标注为其中的一种情绪。

为了评估ASDG在多模态会话情绪识别任务上的性能，参照现有MERC研究中[8,25]广泛使用的评估标准，具体使用以下评估指标：1）准确率（accuracy, ACC）；2）加权平均准确率（weighted average accuracy, WAA）；3）$F1$-score；4）加权平均$F1$-score（weighted average $F1$-score, WAF1）。

3.2 特征抽取

根据文献[47]，采用RoBERTa大语言模型提取文本特征。RoBERTa是一个使用多层Transformer编码器架构的预训练模型，将最后一层[CLS]嵌入来获得文本特征。此外，根据文献[25]，使用特征提取工具包openSMILE进行音频特征提取，然后通过全连

接层来降低音频特征维度,以及使用在面部表情识别数据集上预训练的稠密卷积网络 DenseNet 进行视觉特征提取,DenseNet 由多个密集块组成,每个块包含多层 CNN。

3.3 基线方法

为了评估 ASDG 的性能,本文选择一些当前较有竞争力的 MERC 模型作为基线用于对比分析,大致将模型分为 3 类,分别是基于循环神经网络的模型、基于图神经网络的模型以及基于对比学习的模型,具体描述如下:

1)EmoCaps(emotion capsule)[48]。EmoCaps 通过 Emoformer 提取不同模态的特征,并将不同模态进行维度拼接形成话语级多模态表示,然后输入 LSTM 网络学习上下文依赖。

2)GraphCFC(graph based cross-modal feature complementation)[8]。GraphCFC 是一个基于有向图的跨模态特征互补模型,将每个话语的语、视、音特征分别作为一个节点,构建了一个包含讲话者信息以及多种边关系的异构图,使得 GNN 能够同时获取模态内上下文信息和模态间的交互信息。

3)GA2MIF(graph and attention based two-stage multi-source information fusion)[26]。GA2MIF 是基于图注意力的 2 阶段多模态信息融合方法,将会话中的每个模态的所有上下文分别构建一个图,避免了将异构图作为模型输入的问题。

4)HiMul-LGG(hierarchical decision fusion-ba- sed local-global graph neural network)[24]。HiMul-LGG 采用层次化的决策融合策略,构建局部-全局图神经网络架构来增强说话人在模态间和模态内的依赖。

5)AdaIGN(adaptive interactive graph neural network)[49]。AdaIGN 采用 Gumbel softmax 技术在所构建的异构图中自适应选择节点和边,增强模态内和跨模态交互,并设计特定任务的损失函数以优先考虑文本模态和说话人自身依赖的上下文选择。

6)UniMSE(unified multimodal sentiment analysis and emotion recognition)[36]。UniMSE 将情感分类和情绪识别任务相结合,相互提供多模态情感知识,并引入对比学习来放大样例间的差异。

7)FacialMMT(facial expression aware multimo-dal multi-task learning)[44]。FacialMMT 是 2 阶段学习框架。第 1 阶段利用无监督人脸聚类和人脸匹配算法提取每个会话的人脸序列;第 2 阶段利用面部表情识别来获取情绪分布,从而辅助情绪识别任务。

8)CBERL(class boundary enhanced represent-ation learning)[41]。CBERL 利用生成对抗网络的数据增强方式生成新样本,并将原始样本和新样本输入变分自编码器进行跨模态融合,然后将多模态特征输入到多任务图神经网络中优化类边界的分布表示,以缓解类分布不平衡的问题。

9)MALN(multimodal adversarial learning net-work)[12]。MALN 是一个多模态对抗学习网络,以对抗的方式提取多模态的共性信息,同时学习多样性信息,从而对多模态数据进行综合挖掘。

10)AGF-IB(adversarial alignment and graph fusion via information bottleneck)[6]。AGF-IB 是一种基于对抗对齐和图融合的会话情感识别方法,通过对抗学习消除模态间的异构性,并引入图对比学习来捕获模态内和模态间的互补语义信息。

11)DER-GCN(dialog and event relation-aware graph convolutional neural network)[31]。DER-GCN 构建了一个带权的多关系图来同时捕捉对话中说话人与事件关系之间的依赖关系,引入了自监督掩码图自编码器来提高特征和结构的融合表示能力,提出一种基于对比学习的损失优化策略,以增强少数类特征的表示能力。

3.4 实现细节

使用 Pytorch 实现所提出的 ASDG 模型,并且使用 AdamW 作为优化器在 IEMOCAP 数据集上的学习率为 6E−5,批大小为 1,对应的文本、音频和视觉模态的维度分别设置为 1 024, 1 582, 342。在 MELD 数据集上的学习率为 8E−5,批大小为 12,对应的文本、音频和视觉模态的维度分别设置为 1 024, 300, 342。对于 Transformer 编码器,隐藏层的维度大小、注意力头数量和层数分别设置为 1 024, 1, 1。对于 GAT 网络,注意力头和层数分别设置为 4, 2。为了防止过拟合,将 L_2 权重衰减设置为 1E−5,并将神经元丢失率 dropout 设置为 0.5。

3.5 实验分析与比较

表 2 和表 3 分别展示了在 IEMOCAP 和 MELD 数据集上 ASDG 方法与 3.3 节中采用的一系列基线方法在各类情绪上的 ACC 和 *F*1 值,以及总体分类的 WAA 和 WAF1。

如表 2 所示,在 IEMOCAP 数据集上,ASDG 方法相较于基线方法取得了更具优势的识别效果,其 WAA 和 WAF1 值分别达到了 72.09% 和 72.25%,且在“兴奋”和“中性”2 个类别下的 *F*1 值也是最高的,这表明本文构建的对抗软对比调制动态图学习的 MERC 方法是有效的。其中基于循环神经网络的 EmoCaps

Table 2 Performance of ASDG and Baseline Models on IEMOCAP Dataset

表 2 在 IEMOCAP 数据集上 ASDG 和基线模型的性能 %

方法	情感类别												WAA	WAF1
	喜悦		悲伤		中性		愤怒		兴奋		沮丧			
	ACC	*F*1	ACC	*F*1	ACC	*F*1	ACC	*F*1	ACC	*F*1	ACC	*F*1		
EmoCaps		**71.91**		85.06		64.48		68.99		78.41		66.76		71.77
UniMSE													70.56	70.66
GraphCFC		43.08		**84.99**		64.70		71.35		78.86		63.70	69.13	68.91
GA2MIF		46.15		84.50		68.38		70.29		75.99		66.49	69.75	70.00
HiMul-LGG		53.95		79.92		71.66		67.56		72.00		68.46	70.12	70.22
MALN		55.4		81.8		64.1		69.1		78.0		71.4		70.8
CBERL	58.84	67.34	63.31	72.84	56.42	60.75	**75.32**	**73.51**	70.32	70.77	78.21	**71.19**	69.36	69.27
AdaIGN		53.04		81.47		71.26		65.87		76.34		67.79		70.74
AGF-IB	**71.88**	69.96	74.64	81.58	67.25	63.80	73.79	68.37	**82.66**	79.15	60.45	63.95	70.46	70.36
DER-GCN	60.7	58.8	75.9	79.8	66.5	61.5	71.3	72.1	71.1	73.3	**66.1**	67.8	69.7	69.4
ASDG（本文）	64.58	59.62	**76.73**	81.74	**77.08**	**75.51**	65.88	64.74	78.93	**79.33**	64.30	65.42	**72.09**	**72.25**

注：黑体数值表示最优结果。

Table 3 Performance of ASDG and Baseline Models on MELD Dataset

表 3 在 MELD 数据集上 ASDG 和基线模型的性能 %

方法	情绪类别														WAA	WAF1
	中性		惊讶		恐惧		悲伤		喜悦		厌恶		愤怒			
	ACC	*F*1	ACC	*F*1	ACC	*F*1	ACC	*F*1	ACC	*F*1	ACC	*F*1	ACC	*F*1		
EmoCaps		77.12		**63.19**		3.03		42.52		57.50		7.69		57.54		64.00
UniMSE															65.09	65.51
FacialMMT		80.13		59.63		19.18		41.99		64.88		18.18		**56.00**		66.58
GraphCFC		76.98		49.36				26.89		51.88				47.59	61.42	58.86
GA2MIF		76.92		49.08				27.18		51.87				48.52	61.65	58.94
HiMul-LGG															66.21	65.18
MALN		**82.0**		58.6		21.2		43.0		64.3		17.6				**66.9**
AdaIGN																66.79
AGF-IB	81.10	81.19	56.16	57.24	6.90	5.06	47.41	37.32	**65.92**	**65.92**	2.94	2.94	42.28	45.22	64.14	64.01
DER-GCN	76.8	80.6	50.5	51.0	14.8	10.4	56.7	41.5	69.3	64.3	17.2	10.3	52.5	57.4	66.8	66.1
ASDG（本文）	**82.88**	80.11	55.87	59.58	22.00	**23.16**	39.42	**43.97**	64.43	65.90	**32.35**	**29.53**	**53.04**	52.97	**67.24**	66.83

注：黑体数值表示最优结果。

方法取得的效果仅次于 ASDG 方法，其 WAF1 值为 71.77%，但是 EmoCaps 在各个情绪类别的 ACC 相对较差，这是由于 EmoCaps 方法尽管能够学习话语级的上下文关联，但是其难以自适应挖掘当前话语的上下文信息，且不能关注判别性特征的学习，而 ASDG 方法不仅构建动态话语图来自适应地建模上下文依赖，还设计对抗性软对比训练机制来提高模型的泛化性以及判别性。针对基于图结构的模型，如 GraphCFC 方法取得的 WAA 和 WAF1 值分别为 69.13% 和 68.91%，其虽然利用独立和共享编码器考虑了模态的异构问题，但是其构建图过程中的窗口大小是固定的，不能自适应地确定上下文的依赖范围。虽然 GA2MIF 和 HiMul-LGG 通过分阶段融合来缓解模态间的异构性，但是同样面临固定窗口的限制，导致性能提升有限。DER-GCN 利用图重构的思想来自适应地丢弃或增加节点与边，而 AdaIGN 利用 Gumbel softmax 自适应地选择节点和边，尽管 DER-GCN 和 AdaIGN 都采用一定程度的自适应构图的思路，但是

它们均缺乏考虑模态表征的泛化设计，使得效果明显劣于 ASDG 方法。此外，ASDG 也明显优于最近提出的基于对比思想的方法，例如，UniMSE 仅实施不同模态特征间的对比来缓解异构问题，但是并未考虑情感语义上的标签级的对比。AGF-IB 不仅考虑了标签级的对比，并对文本、音频、视频 3 模态分别构建生成器和判别器，而 MALN 仅对 3 模态构建一个共同的生成器，虽然 AGF-IB 方法和 MALN 方法能不同程度地缓解模态间的异构问题，但是过于强调模态间的相似性，不利于挖掘模态间的互补性。CBERL 方法通过在 3 模态的特征级上加噪来提高模态的泛化性，但是这会摧毁会话的上下文语义连贯性，而 ASDG 方法通过在特征抽取的 LSTM 隐藏层加噪来生成对抗样本，并在原始样本和对抗样本上进行联合软对比学习，来扩展每个模态的类表示空间和最大化模态内具有相同标签样本的一致性，从而达到更可靠的泛化效果。

不同于 IEMOCAP 数据集，MELD 数据集的每个会话的轮次更少，但每个会话中涉及的说话者更多，且 MELD 数据集呈现出极其不平衡的类分布特点，这使得在 MELD 数据集上开展 MERC 任务更加困难。如表 3 所示，ASDG 方法在 MELD 数据集上的 WAA 和 WAF1 值分别为 67.24%，66.83%，而对比的 EmoCaps 方法的 WAF1 值仅达到了 64.00%，这是由于 MELD 数据集牵扯 9 个讲话者，而 EmoCaps 方法没有考虑讲话者特征，但 ASDG 方法既涵盖讲话者依赖特征又嵌入讲话者身份的独热编码，能有效地帮助识别讲话者的话语语义关联。此外，ASDG 方法要优于所列举的不同形式的基于 GNN 的方法，如 GraphCFC，GA2MIF，DER-GCN，AdaIGN，这是因为 MELD 数据集涉及的讲话者多但会话轮次少，导致能够利用的上下文信息较少，因此，精准地学习上下文依赖范围显得尤为重要，这更表明 ASDG 方法中设计的动态会话图学习策略更适合于精准界定上下文关联范围。尤其，ASDG 在少类样本上取得了最好的效果，如在“恐惧”“悲伤”“厌恶”上 $F1$ 值分别达到了 23.16%，43.97%，29.53%，相较于基于对比思想的 MALN 方法，分别提高了 1.96 个百分点，0.97 个百分点，11.93 个百分点。这说明 ASDG 方法中的对抗性软对比训练机制对少类样本的处理较有优势，能够获得更具判别的特征。

综上实验可知，ASDG 方法在样本量、会话长度、讲话者数量等分布差异较大的 IEMOCAP 数据集和 MELD 数据集上均能有效识别多模态会话中的情绪类别，这表明本文所设计的基于密度的自适应图学习策略能有效地捕获模态内以及跨模态的动态语境依赖，且在原始样例和设计的基于加噪策略生成的对抗样例上联合进行对抗性软对比训练以调制构建的双流图结构，最终使得构建的模型能够提取更泛化和鲁棒的模态表征。

3.6 消融实验

为了评估 ASDG 中各模块对情绪识别效果的影响，具体在 IEMACAP 和 MELD 这 2 个数据集上从模型组件、训练方式以及不同模态贡献度这 3 个角度展开模型的消融实验，表 4 给出了不同消融设置下的实验结果，并选择 WAA 和 WAF1 作为评价指标。

Table 4 Results of Ablation Studies on IEMOCAP and MELD

表 4 在 IEMOCAP 和 MELD 上的消融实验结果 %

方法		IEMOCAP		MELD	
		WAA	WAF1	WAA	WAF1
ASDG（本文）		72.09	72.25	67.24	66.83
ASDG-based model	-w/o CCGL	68.70	69.13	65.94	65.69
	-w/o CFDG	72.02	72.03	66.55	66.23
	-w/o DG	70.67	70.55	65.29	65.56
	-w/o PE	70.36	70.35	66.74	66.54
	-w/o SE	71.1	71.34	66.32	65.69
训练机制	-w/o ASCL	69.99	70.08	66.21	65.58
	-w/o AT	69.75	69.77	66.55	65.45
	-w/o SHGR	71.90	71.96	66.39	65.94
Modality	文本	68.44	68.47	65.25	64.96
	音频	59.09	57.84	46.02	34.95
	视频	38.20	30.81	44.33	32.31
	文本 + 音频	70.43	70.57	65.82	65.45
	文本 + 视频	68.52	68.83	66.74	65.89
	音频 + 视频	62.85	62.23	47.39	32.46

3.6.1 关于 ASDG 方法中各个组件的消融分析

1）-w/o CCGL（consistent cross-modal graph learning guided by textual semantics）。移除文本语义引导的一致性跨模态图学习，但保留一致性损失。

2）-w/o CFDG（conversation flow dynamic graph learning）。移除会话流动态图学习分支。

3）-w/o DG（density graph）。移除单模上下文动态图网络和会话流动态图学习中的图学习，直接进行特征维度的拼接。

4）-w/o PE（positional embedding）。在 CCGL 模块中移除位置嵌入。

5)-w/o SE（speaker embedding）。在 CCGL 模块中移除讲话者嵌入。

以上组件是 ASDG 的 5 个关键组成部分，通过每次只移除 1 个组件来评估该组件的有效性。从表 4 中可以看出：1)本文所提出的这些组件都是有效的，因为任意移除其中 1 个组件都会导致性能的下降。2)文本语义引导的一致性跨模态图学习对模型性能的影响最大，当移除该模块时，模型性能在 2 个数据集上 WAA 分别下降了 3.39 个百分点，1.3 个百分点，这说明在语义一致性损失的约束下实现跨模态交互可以充分挖掘不同模态的语义信息。3)会话流动态图学习在 IEMOCAP 数据集上的影响不是很大，但是在 MELD 数据集上影响较大，这是由于 MELD 数据集中的会话轮次少且信息量少，通过充分地挖掘多模态话语级的会话动态上下文依赖，是能显著提高模型对多模态数据语义的理解能力。综合 2)和 3)可以看出以双流结构的学习方式进行跨模态交互可以增强对会话中抽象情感语义的学习能力，并有效地缓解模态异构问题。4)动态话语图对 2 个数据集的情绪识别都有相当大的影响，这说明自适应地建模上下文范围可以精准地界定会话流的依赖范围，从而减轻特征冗余或者信息不足的情况。5)位置嵌入和说话者嵌入对 2 个数据集的情绪识别也有一定的影响，这意味着捕获会话场景中的话语顺序信息和讲话者信息是有价值的。

3.6.2　关于训练机制的消融分析

1)-w/o ASCL（adversarial soft-contrast learning）。为了验证对抗性软对比的有效性，移除对抗性软对比训练机制，但将软对比损失替换为交叉熵损失。

2)-w/o AT（adversarial training）。为了验证对抗性训练的有效性，在对抗性软对比训练机制中，移除对抗性训练，仅保留软对比学习。

3)-w/o SHGR。移除 SHGR 一致性损失。

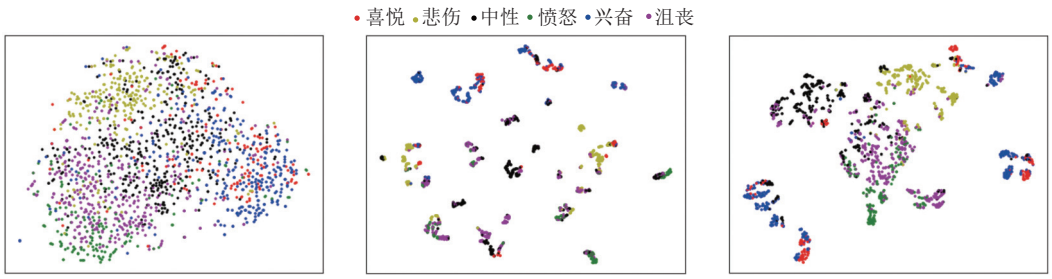
从表 4 中可以看出：1)本文所提出的对抗性软对比训练机制是有效的，因为当移除对抗性软对比训练机制时，评估指标 WAF1 在 IEMOCAP 数据集上下降了 2.17 个百分点，且在 MELD 数据集上下降了 1.25 个百分点，这说明了本文设计的对抗性软对比训练机制比仅用交叉熵损失来协同优化，更能使得学习到的多模态特征具有判别性。2)本文设计对抗训练机制也有较大的效用，因为当仅移除对抗训练时，评估指标 WAF1 在 IEMOCAP 数据集和 MELD 数据集上分别下降了 2.48 个百分点和 1.38 个百分点，样本和对抗样本进行联合优化，使得模型能学习到更泛化的模态表示，且能有效地避免生成的对抗样本面临语义不连贯的问题。3)当不使用 SHGR 损失时，模型的 WAF1 指标在 2 个数据集上分别下降了 0.29 个百分点和 0.89 个百分点，这说明了使用本文设计的文本语义引导的跨模态一致性学习来迫使音频模态和视频模态分别与文本模态的语义保持一致性是合理的，能够促进跨模态特征的有效交互。

3.6.3　关于不同模态贡献度的消融分析

为了体现多模态会话情绪识别中不同模态的贡献度，每次移除 1 个或者 2 个模态来进行实验。从表 4 中可以得出 3 个结论：1)对于单模态会话情绪识别的实验结果，文本模态的效果要远远优于其他 2 种模态，表明文本模态比音频模态和视频模态的语义要丰富，这更说明本文所设计的文本语义引导的一致性跨模态图学习中以文本为中心来引导和集成音频和视频的设计是合理的。2)在双模态的会话情绪识别中，文本+音频的效果要好于文本+视频的效果，这说明在情绪识别中音频信息要比视频信息更重要，语音中含有的语调等信息可能比图像更能反映人的情感，而图像中的背景噪声往往较大。3)同时使用文本、音频和视频这 3 种模态可以获得最佳的会话情绪识别性能，这说明结合多种模态的信息更能反映人的情绪及其变化，但是合理地融合多模态信息进行情绪识别是至关重要的，因为本文提出的多模态会话信息融合模型相比其他的融合模型的情绪识别效果更好，如表 2 和表 3 所示。

3.7　多模态表示可视化

为了更直观地验证本文所提出的对抗性软对比训练机制的有效性，分别在去除和保留对抗性软对比损失的情况下训练模型，并在 IEMOCAP 的测试集中可视化模态特征的分布情况，如图 4(b)(c)所示。首先图 4(a)直观展示了原始数据中不同情绪类别的分布状况。具体地，从 IEMOCAP 的测试集中提取每个话语的不同模态表征，并整合这些不同模态表征以形成初始的话语级多模态表征，然后利用 t-SNE 算法[50]将这些多模态表征降维并投影至 2 维空间进行可视化。通过比较图 4(a)(c)，可以看出相较于原始的多模态表征，本文提出的基于对抗性软对比调制双流图学习模型所习得的表征在聚类效果上有了显著提升。通过分析图 4(b)，可以看出在没有应用对抗性软对比策略学习的情况下，多模态特征显得过于离散，即便是同一情绪类别的多模态特征也难以有效聚合。通过对比分析图 4(b)(c)，可以清晰地观察到，在引入对抗性软对比学习策略后，在图 4(c)中相同



(a) 原始特征表示 (b) 无ASCL (c) 有ASCL

Fig. 4 t-SNE visualization of multimodal representations of different emotion classes on IEMOCAP dataset

图 4 在 IEMOCAP 数据集上不同情绪类别的多模态表示的 t-SNE 可视化

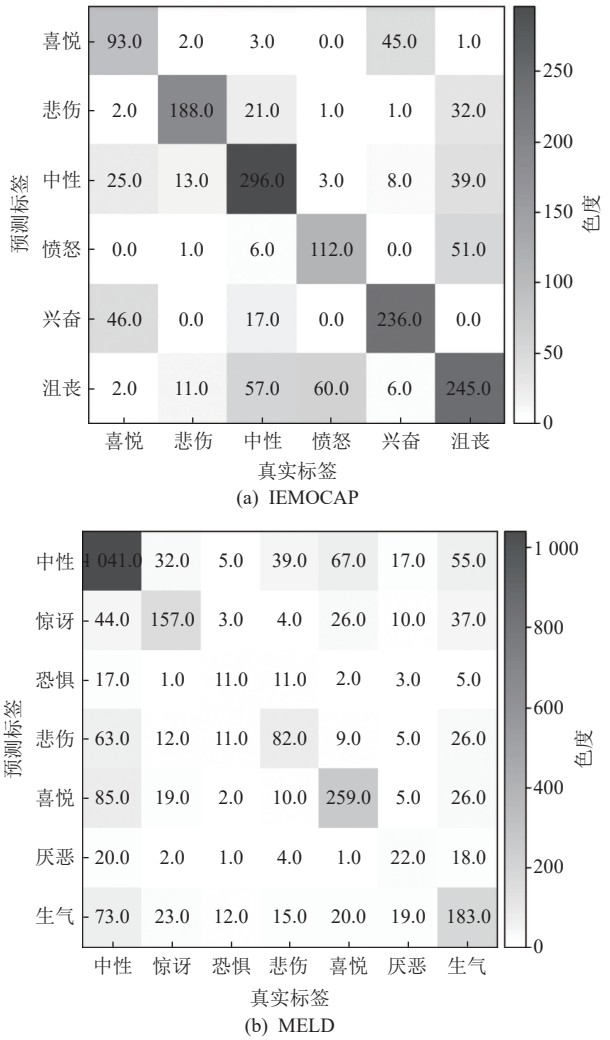
情绪类别的多模态表征能够更加紧密地聚集在一起，而不同情绪类别的表征则较少地混杂在一起。通过可视化结果表明，本文设计的对抗性软对比的调制策略能够帮助构建的模型学习到更鲁棒且泛化的模态表征。

3.8 混淆矩阵

表 2 和表 3 可知，ASDG 方法相较对比方法在 IEMOCAP 数据集(如“悲伤”“兴奋”)和 MELD 数据集(如“恐惧”“悲伤”“厌恶”“愤怒”)的少样例类别上取得了较好的结果，这表明 ASDG 方法能够提取更具泛化且鲁棒的特征表示。尽管 ASDG 在 MERC 任务上展现出了相对优越的性能，但它仍面临一些局限，尤其在处理情绪类别语义相近的样例上识别效果仍存在一定的不足。为此，在 IEMOCAP 和 MELD 的测试集上通过混淆矩阵展开具体情绪类别上的性能分析。具体地，如图 5 所示，ASDG 方法在处理语义边界模糊的情绪时仍存在样例的混淆，例如，在 IEMOCAP 数据集中，模型倾向于将“喜悦”类的样本误判为“兴奋”，以及将“愤怒”类的样本误识别为“沮丧”。类似地，在 MELD 数据集中，模型对“恐惧”情绪的识别也常常偏离，错误地将其归类为“悲伤”。诸如这种高级抽象语义的混淆表明 ASDG 方法在区分极其细微情绪差异时存在挑战。此外，ASDG 方法还展现出对“中性”情绪类别的偏好，这与数据集中存在的情绪类别不平衡问题有关，尤其“中性”情绪的样本在数据集中往往占据较大的比例，使得模型在学习过程中可能过度拟合这类情绪的特征，进而影响对少类样例的情绪识别。

4 结 论

本文针对当前多模态会话研究难以提取泛化且鲁棒的特征表示以及不能动态调整会话窗口导致提取的语境信息冗余或不足的问题，提出对抗软对比调制动态图学习的方法，不仅能实现会话语境的动态建模，还能有效融合多模态异构特征，并显著提升模型的鲁棒性与泛化性。ASDG 首先通过 LSTM 初步提取各自模态的上下文特征，并构建基于动态话语图的 GAT 来动态地建模各自模态内的精准上下文依赖，然后，设计互补双流图学习机制来分别建模跨模态交互和话语级多模态上下文的交互关联，其中，



(a) IEMOCAP

(b) MELD

Fig. 5 Confusion matrix on IEMOCAP and MELD test sets

图 5 IEMOCAP 和 MELD 测试集上的混淆矩阵

文本语义引导的跨模态流图学习通过 Transformer 架构来集成 3 个模态的图特征表示以探索跨模态的异构融合，而会话流动态图学习利用跨模态注意力来学习文本引导的音频和视频表示，并将其融合作为话语图 GAT 的输入以自适应地学习多模态话语级的上下文依赖。尤其，在模型优化上设计了对抗性软对比训练机制，通过在 LSTM 隐藏层加入扰动来生成对抗样本，并在原始样本和对抗样本上利用软对比学习来扩展类样本空间，以学习到更泛化的类特征表示。在 IEMOCAP 和 MELD 这 2 个不同代表性的数据集上进行了实验，实验结果表明 ASDG 方法的性能优于相关的对比方法，并通过消融实验证明了 ASDG 中的不同模块以及优化设置的有效性。在未来的研究中，将在会话场景上下文学习中探索因果关联式的自适应上下文的学习，从而进一步探索多模态会话情绪识别的可解释性。

作者贡献声明：王顺杰负责研究方案的构建，并完成实验和撰写论文；蔡国永提出了指导意见并修改论文；吕光瑞提出了算法思路和实验方案。

参 考 文 献

[1] Singh G V, Firdaus M, Ekbal A, et al. EmoInt-Trans: A multimodal transformer for identifying emotions and intents in social conversations[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2022, 31: 290−300

[2] Zhou Hao, Huang Minlie, Zhang Tianyang, et al. Emotional chatting machine: Emotional conversation generation with internal and external memory[C]//Proc of the 32nd AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2018: 730−738

[3] Rosa R L, Schwartz G M, Ruggiero W V, et al. A knowledge-based recommendation system that includes sentiment analysis and deep learning[J]. IEEE Transactions on Industrial Informatics, 2018, 15(4): 2124−2135

[4] Chatterjee A, Narahari K N, Joshi M, et al. SemEval−2019 task 3: EmoContext contextual emotion detection in text[C]//Proc of the 13th Int Workshop on Semantic Evaluation. Stroudsburg, PA: ACL, 2019: 39−48

[5] Bhavan A, Chauhan P, Shah R R. Bagged support vector machines for emotion recognition from speech[J]. Knowledge-Based Systems, 2019, 184(C): 104886

[6] Shou Yuntao, Tao Meng, Wei Ai, et al. Adversarial alignment and graph fusion via information bottleneck for multimodal emotion recognition in conversations[J]. Information Fusion, 2024, 112: 102590

[7] Wang Ruobing, Zha Daren, Zhao Qingfei, et al. RW-MMDCG: Muti-modal via rolling-Window directed graph network for conversational emotion recognition[C]//Proc of the 26th Int Conf on Computer Supported Cooperative Work in Design. Piscataway, NJ: IEEE, 2023: 606−611

[8] Li Jiang, Wang Xiaoping, Lv Guoqing, et al. GraphCFC: A directed graph based cross-modal feature complementation approach for multimodal conversational emotion recognition[J]. IEEE Transactions on Multimedia, 2023, 26: 77−89

[9] Li Jiang, Wang Xiaoping, Lv Guoqing, et al. GraphMFT: A graph network based multimodal fusion technique for emotion recognition in conversation[J]. Neurocomputing, 2023, 550(suppl-C): 126427

[10] Lu Nannan, Han Zhiyuan, Han Min, et al. Bi-stream graph learning based multimodal fusion for emotion recognition in conversation[J]. Information Fusion, 2024, 106: 102272

[11] Ma Hongchao, Wang Zhongqing, Zhou Xiabing, et al. Emotion recognition with conversational generation transfer[J]. Transactions on Asian and Low-Resource Language Information Processing, 2022, 21(4): 1−17

[12] Ren Minjie, Huang Xiangdong, Liu Jing, et al. MALN: Multimodal adversarial learning network for conversational emotion recognition[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(11): 6965−6980

[13] Abonizio H Q, Paraiso E C, Barbon S. Toward text data augmentation for sentiment analysis[J]. IEEE Transactions on Artificial Intelligence, 2021, 3(5): 657−668

[14] Gunel B, Du Jingfei, Conneau A, et al. Supervised contrastive learning for pre-trained language model fine-tuning[C]//Proc of the 9th Int Conf on Learning Representations. Washington: ICLR, 2021: 1−15

[15] Huang Faliang, Li Xuelong, Yuan Changan, et al. Attention-emotion-enhanced convolutional LSTM for sentiment analysis[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 33(9): 4332−4345

[16] Lu Yi, Mak M W. Improving speech emotion recognition with adversarial data augmentation network[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 33(1): 172−184

[17] Essa I A. Analysis, interpretation and synthesis of facial expressions[D]. Cambridge, MA: Massachusetts Institute of Technology, 1995

[18] Zhang Hanqing, Song Dawei. Towards contrastive context-aware conversational emotion recognition[J]. IEEE Transactions on Affective Computing, 2022, 13(4): 1879−1891

[19] Huan Ruohong, Shu Jia, Bao Shenglin, et al. Video multimodal emotion recognition based on Bi-GRU and attention fusion[J]. Multimedia Tools and Applications, 2021, 80(6): 8213−8240

[20] Wang Jianyou, Xue M, Culhane R, et al. Speech emotion recognition with dual-sequence LSTM architecture[C]//Proc of the 2020 IEEE Int

Conf on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2020: 6474−6478

[21] Li Chao, Bao Zhongtian, Li Linhao, et al. Exploring temporal representations by leveraging attention-based bidirectional LSTM-RNNs for multi-modal emotion recognition[J]. Information Processing & Management, 2020, 57(3): 102185

[22] Ho N H, Yang H J, Kim S H, et al. Multimodal approach of speech emotion recognition using multi-level multi-head fusion attention-based recurrent neural network[J]. IEEE Access, 2020, 8: 61672−61686

[23] Shen Xudong, Huang Xianying, Zou Shihao, et al. Multimodal knowledge-enhanced interactive network with mixed contrastive learning for emotion recognition in conversation[J]. Neurocomputing, 2024, 582(C): 127550

[24] Fu Changzeng, Qian Fengkui, Su Kaifeng, et al. HiMul-LGG: A hierarchical decision fusion-based local–global graph neural network for multimodal emotion recognition in conversation[J]. Neural Networks, 2025, 181: 106764

[25] Hu Jingwen, Liu Yuchen, Zhao Jinming, et al. MMGCN: Multimodal fusion via deep graph convolution network for emotion recognition in conversation[C]//Proc of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th Int Joint Conference on Natural Language Processing. Stroudsburg, PA: ACL, 2021: 5666−5675

[26] Li Jiang, Wang Xiaoping, Lv Guoqing, et al. GA2MIF: Graph and attention based two-stage multi-source information fusion for conversational emotion detection[J]. IEEE Transactions on Affective Computing, 2023, 15(1): 130−143

[27] Huang Xuejian, Ma Tinghuai, Wang Gensheng. Multimodal learning method based on intra- and inter-sample cooperative representation and adaptive fusion[J]. Journal of Computer Research and Development, 2024, 61(5): 1310−1324 (in Chinese)
(黄学坚, 马廷淮, 王根生. 基于样本内外协同表示和自适应融合的多模态学习方法[J]. 计算机研究与发展, 2024, 61(5): 1310−1324)

[28] Xie B, Sidulova M, Park C H. Robust multimodal emotion recognition from conversation with transformer-based crossmodality fusion[J]. Sensors, 2021, 21(14): 4913

[29] Wen Huanglu, You Shaodi, Fu Ying. Cross-modal dynamic convolution for multi-modal emotion recognition[J]. Journal of Visual Communication and Image Representation, 2021, 78(Suppl C): 103178

[30] Huang Jian, Tao Jianhua, Liu Bin, et al. Multimodal transformer fusion for continuous emotion recognition[C]//Proc of the 2020 IEEE Int Conf on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2020: 3507−3511

[31] Wei Ai, Shou Yuntao, Meng Tao, et al. DER-GCN: Dialog and event relation-aware graph convolutional neural network for multimodal dialog emotion recognition[J]. IEEE Transactions on Neural Networks and Learning Systems, 2024, 36(3): 4908−4921

[32] Du Jinming, Sun Yuanyuan, Lin Hongfei, et al. Conversational emotion recognition incorporating knowledge graph and curriculum learning[J]. Journal of Computer Research and Development, 2024, 61(5): 1299−1309 (in Chinese)
(杜金明, 孙媛媛, 林鸿飞, 等. 融入知识图谱和课程学习的对话情绪识别[J]. 计算机研究与发展, 2024, 61(5): 1299−1309)

[33] Kalantidis Y, Sariyildiz M B, Pion N, et al. Hard negative mixing for contrastive learning[C]//Proc of the 34th Int Conf on Neural Information Processing Systems. Cambridge, MA: MIT, 2020: 21798−21809

[34] Xu Yuan, Yang Meng. MCM-CSD: Multi-Granularity context modeling with contrastive speaker detection for emotion recognition in real-time conversation[C]//Proc of the 2024 IEEE Int Conf on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2024: 11956−11960

[35] Tu Geng, Xiong Feng, Ling Bin, et al. Multimodal eemotion recognition calibration in conversations[[C]//Proc of the 32nd ACM Int Conf on Multimedia. New York: ACM, 2024: 9621−9630

[36] Hu Guimin, Lin Tingen, Zhao Yi, et al. UniMSE: Towards unified multimodal sentiment analysis and emotion recognition[C]//Proc of the 2022 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2022: 7837−7851

[37] Lu Jiajun, Sibai H, Fabry E. Adversarial examples that fool detectors[J]. arXiv preprint, arXiv: 1712.02494, 2017

[38] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks[J]. arXiv preprint, arXiv: 1312.6199, 2013

[39] Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples[J]. arXiv preprint, arXiv: 1412.6572, 2014

[40] Miyato T, Dai A M, Goodfellow I. Adversarial training methods for semisupervised text classification[C]//Proc of the 5th Int Conf on Learning Representations. Washington: ICLR, 2017: 24−26

[41] Meng Tao, Shou Yuntao, Ai Wei, et al. Deep imbalanced learning for multimodal emotion recognition in conversations[J]. IEEE Transactions on Artificial Intelligence, 2024, 5(12): 2691−4581

[42] Poria S, Hazarika D, Majumder N, et al. MELD: A multimodal multi-party dataset for emotion recognition in conversations[C]//Proc of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2019: 527−536

[43] 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

[44] Zheng Wenjie, Yu Jianfei, Xia Rui, et al. A facial expression-aware multimodal multi-task learning framework for emotion recognition in multi-party conversations[C]//Proc of the 61st Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL,

2023: 15445−15459

[45] Wang Lichen, Wu Jiaxiang, Huang Shaolun, et al. An efficient approach to informative feature extraction from multimodal data[C]//Proc of the 2019 AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2019: 5281−5288

[46] Mądry A, Makelov A, Schmidt L, et al. Towards deep learning models resistant to adversarial attacks[J]. arXiv preprint, arXiv: 1706.06083, 2017

[47] Ghosal D, Majumder N, Gelbukh A, et al. COSMIC: Commonsense knowledge for emotion identification in conversations[C]//Proc of the 2020 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2020: 2470–2481

[48] Li Zajing, Tang Fengxiao, Zhao Ming, et al. EmoCaps: Emotion capsule based model for conversational emotion recognition[C]//Proc of the 60th Association for Computational Linguistics. Stroudsburg, PA: ACL, 2022: 1610−1618

[49] Tu Geng, Xie Tian, Liang Bin, et al. Adaptive graph learning for multimodal conversational emotion detection[C]//Proc of the 38th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2024: 19089−19097

[50] Van der Maaten L, Hinton G. Visualizing data using t-SNE[J]. Journal of Machine Learning Research, 2008, 9(11): 2579−2605



Wang Shunjie, born in 1996. PhD candidate. Her main research interests include deep learning and multi-modal affective computing.

王顺杰, 1996 年生。博士研究生。主要研究方向为深度学习、多模态情感计算。



Cai Guoyong, born in 1971, PhD, professor, PhD supervisor. Distinguished member of CCF. His main research interests include multi-modal affective computing, trustable AI theory and techniques.

蔡国永, 1971 年生。博士，教授，博士生导师。CCF 杰出会员。主要研究方向为多模态情感计算、可信 AI 理论技术。



Lü Guangrui, born in 1989. PhD, master supervisor. Member of CCF. His main research interests include computer vision and multi-modal information processing.

吕光瑞, 1989 年生。博士，硕士生导师。CCF 会员。主要研究方向为计算机视觉、多模态信息处理。