arXiv:2302.02447v1 [eess.AS] 5 Feb 2023

# 文本与语音话语级情绪识别的跨模态融合技术

*Jiachen Luo*[1]*,Huy Phan*[2*]*, Joshua Reiss*[1]

伦敦玛丽女王大学数字音乐中心，英国[2]亚马逊 Alexa，美国马萨诸塞州剑桥

**摘要**

多模态情感识别（MER）是一项基础且复杂的研究问题，其原因在于人类情绪表达具有不确定性，且不同模态之间存在异质性差距。对于人类参与者理解情绪而言，音频和文本模态尤为重要。尽管已有许多成功的尝试为 MER 设计了多模态表征，但仍有多个挑战亟待解决：1）弥合多模态特征之间的异质性差距，并对多种模态的模态间和模态内交互进行建模；2）有效且高效地对对话序列中的上下文动态进行建模。本文提出跨模态 RoBERTa（CM-RoBERTa）模型，用于从口语音频及其对应的转录文本中检测情绪。作为 CM-RoBERTa 的核心单元，该模型采用并行的自注意力和交叉注意力，动态捕捉音频和文本的模态间及模态内交互。具体而言，模型采用中层融合和残差模块，以建模长期上下文依赖关系并学习模态特定模式。我们在 MELD数据集上评估了该方法，实验结果表明，所提出的方法在该数据集上达到了最先进的性能。

索引词— 语音情感识别、多模态融合、深度学习

## 1. 引言

对话情绪识别（ERC）至关重要，但也极具挑战性，应用于自然的人机交互 [1], 智能教育辅导 [2], 和心理健康分析等领域 [3]。在日常生活中，人们以自然的方式进行多轮对话，通过语言和非语言内容（例如面部表情和肢体语言）传达情绪状态 [4]。不同模态都包含与情绪相关的信息，如何高效、有效地融合多个模态的异构输入，一直是活跃的研究课题。

在日常生活中，文本模态通常伴随着音频模态。文本特征中包含的情感内容，可由基频、强度、停顿、响度以及其他频率相关测量指标等多种信息加以补充 [5,6]。文本与音频数据之间的有效交互能够呈现更全面的内容，并提供更多情绪信息 [7∼9]。如图1所示，单词“Okay”含义模糊，可以表达不同的情绪。仅根据一段文本话语来判断其对应的情绪颇具挑战。加入相应的低沉声音和啜泣声后，便不难辨别出这句话表达的是负面情感。然而，如何有效融合不同模态，是研究人员经常面临的一项并非易事的任务。

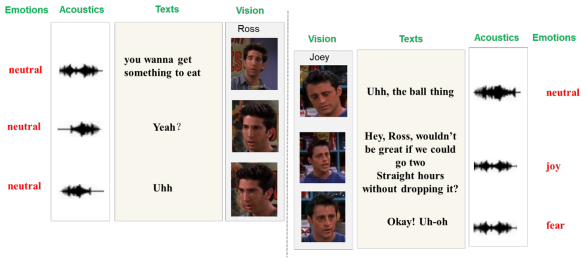


**图 1**。对话中说话者情绪动态的比较。

以往方法已取得良好表现 [8,9], , 但多模态情感识别仍面临一些关键挑战：由于处于异构空间，不同模态依赖各自独立的预处理和特征挖掘设计；为了构建适用于单个模态及融合模型、具有泛化能力的模型，有必要学习模态内和跨模态交互，以揭示判别性情绪内容；此外，情绪是一种主观概念 [10∼12]。此外，文本信息及其相关上下文在推断说话者的情绪时提供了更具影响力的信息 [13], , 并在推断对话中的情绪方面发挥着关键作用。遗憾的是，现有方法无法在融合过程中高效捕捉文本情绪相关内容，而且模态内和跨模态信息的学习往往会损失部分语义内容 [12,14]。

为解决上述问题，我们提出了跨模态 RoBERTa，在架构中通过微调预训练的 RoBERTa 提取器来整合音频模态的信息。作为 CM-RoBERTa 的基本单元，交叉注意力层和自注意力层旨在学习模态间和模态内的信息，并能够自适应地扩展来自不同来源的信息之间的交互与相关性。此外，我们设计了一个时间校准模块，以有效保留单一模态（文本和音频）的局部信息，以及在上下文层面生成的整体信息。我们在公开的多模态情感数据集 MELD [15] 上对其进行了评估。实验结果表明，模型性能和效率均有显著提升。本文的主要贡献可总结如下：

•我们提出了一种 CM-RoBERTa 模型，引入音频模态信息，以辅助预训练 RoBERTa 模型的文本模态微调。

*H. Phan 在伦敦玛丽女王大学数字音乐中心工作期间完成了这项工作，之后才加入亚马逊。感谢中国国家留学基金管理委员会和伦敦玛丽女王大学提供资助。

• 我们采用并行的自注意力和交叉注意力，通过音频与文本之间的交互动态调整注意力权重，并捕捉模态间和模态内的信息。

• 我们利用中层融合和残差连接来捕捉长期上下文依赖，并学习特定于模态的模式。

• 我们证明，所提出的方法在对话情感识别方面优于当前最先进的方法。

## 2. 相关工作

对话情绪识别因其在各个领域的应用而成为热门研究方向。由于人类情绪有多种表达方式，且单一模态的情绪识别存在一些局限，多模态情感识别逐渐成为一种新趋势。多模态情感识别面临的一项基本挑战是如何有效整合多模态信息。

迄今为止，多模态情感识别大致分为四类 [14~21]：早期融合、后期融合、模型级融合和混合融合。早期融合是通过拼接等方式融合来自不同模态的输入特征。Poria 等人利用上下文感知话语表示进行情绪识别 [15]。由于输入层面的模态间动态关系复杂，这种特征融合方法无法捕捉模态内信息。相比之下，后期融合为每种模态分别训练独立的分类模型，并采用决策投票或其他分类器组合策略。Xie 等人采用跨模态 Transformer 融合，对音频、文本和视频进行后期融合，从而避免了融合异构信息的困难[22]。然而，此类方法忽视了多个模态之间的关联，未能对彼此间的相关性进行建模。

另一方面，模型级融合是对前两种方法的折中，其融合发生在多模态特征的中间表示之间。Akhtar 等人设计了一种上下文级跨模态注意力模块，用于学习话语的文本、音频和视觉信息之间的联合关联[20]。混合融合结合了早期融合与后期融合。Wollmer 等人将音频和视频信息与双向长短期记忆网络相融合，用于视听情感识别[14]。在多模态情感识别中，如何设计有效且高效的结构来捕捉模态间和模态内特征至关重要。

随着注意力机制的普及，它凭借融合多模态信息的能力，近来受到该领域的关注。Devamanyu 等人采用多注意力循环网络框架来学习多模态交互 [23]。此外，近年来，利用预训练网络提取特征的迁移学习技术也逐渐流行 [24]。基于 BERT 的模型 RoBERTa 通过利用预训练权重进行微调，在情绪识别任务中取得了更好的性能 [25]。在我们的工作中，我们提出了一种有效且高效的微调融合方法，将异质非语言信息（音频）与 RoBERTa 的语言表达相融合，以对其进行补充。

## 3. 方法论

我们利用多模态信息和上下文信息来辨别话语中的判别性情绪。模型架构如图 2 所示。

### 3.1. 音频和文本特征提取

CM-RoBERTa 模型的输入由从数据中提取的每段音频和文本的有效异构表示（手工特征和瓶颈特征）组成（见图 2）。对于每段音频话语，我们使用 openSMILE 工具包 [26] 提取 6552 个低级手工特征（$a^{LLF}$），这些特征由多个描述符以及针对各种声音和韵律特征计算的统计函数构成。此外，我们采用预训练OpenL3模型 [27] 提取瓶颈嵌入（$a^{OpenL3}$）。随后将 $a^{LLF}$ 和 $a^{OpenL3}$ 拼接，得到音频表示 $x^a = (a^{LLF} \oplus a^{OpenL3})$。其中，$\oplus$ 表示拼接。对于文本模态，RoBERTa 编码器（$t^{RoBERTa}$）接收话语的转录文本，并生成最后四层的激活值。随后将这四层的激活值拼接，得到维度为 4096 [25] 的上下文无关话语特征向量 $x^t$。

### 3.2. 跨模态 RoBERTa

如图 2 所示，CM-RoBERTa 首先使用双向 LSTM 对声学和文本特征进行编码，分别记为 $H_a$ 和 $H_t$。我们将它们（$H_a$ 和 $H_t$）输入自注意力与交叉注意力单元（SCA），以捕捉每对声学帧和文本词语之间的模态内与模态间交互。SCA 模块由两个堆叠的并行交叉注意力层、自注意力层和前馈层组成。该模块的核心思想是学习模态关联和模态特定模式，然后根据注意力权重从模态内和模态间两个角度传播信息。

接下来， [28], 我们以交叉方式估计音频和文本模态之间的关联，其查询（$Q_m$）、键（$K_m$）和值（$V_m$）分别是模态 $m$、$H_m$在不同投影空间下的表示，其中 $m$ 可以是音频（$a$）或文本（$t$）。具体公式如下：

$$\nabla H_{a-t} = softmax(Q_a K_t^{\mathsf{T}}/\sqrt{d})V_t, \quad (1)$$

$$\nabla H_{t-a} = softmax(Q_t K_a^{\mathsf{T}}/\sqrt{d})V_a, \quad (2)$$

其中， $\nabla H_{a-t}$ 和 $\nabla H_{t-a}$ 分别表示从音频传播至文本、从文本传播至音频的信息。 $d$ 表示键维度。随后，我们利用从另一模态传播而来的信息更新其中一个模态的特征：

$$H^{L_a^N} = LayerNorm(H_a \oplus \nabla H_{t-a}), \quad (3)$$

$$H^{L_t^N} = LayerNorm(H_t \oplus \nabla H_{a-t}). \quad (4)$$

为进一步增强表示能力，我们将学习到的加权信息输入全连接层，以获得音频和文本模态之间最终的交叉注意力信息：

$$H^{c_a} = LayerNorm(H^{L_a^N} \oplus \nabla FeedForward(H^{L_a^N})), \quad (5)$$

$$H^{c_t} = LayerNorm(H^{L_t^N} \oplus \nabla FeedForward(H^{L_t^N})). \quad (6)$$

我们分别将音频和文本模态之间最终的交叉注意力信息表示为 $H_a^c$ 和 $H_t^c$。其中， $\oplus$ 是拼接操作。 ”

自注意力机制与交叉注意力机制的原理相似。唯一的区别是，查询、键和值来自同一模态。单个

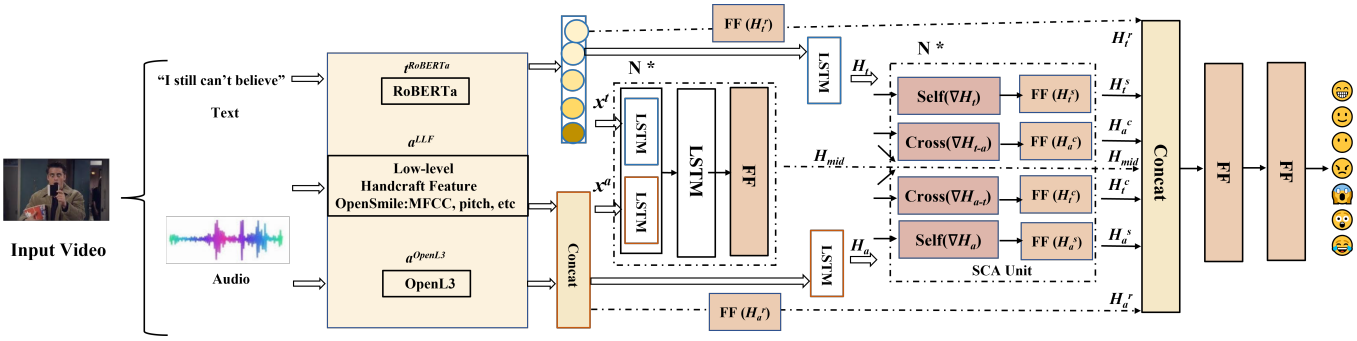


**图 2**。跨模态 RoBERTa 融合网络的整体架构。N 表示两层，前两个并行 LSTM 与后两个并行 LSTM 完全相同。

堆叠层中的自注意力模块可概括如下：

$$\nabla H_m = softmax(Q_m K_m^{\mathsf{T}}/\sqrt{d})V_m, \quad (7)$$

$$H_m^{LN_s} = LayerNorm(H_m \oplus \nabla H_m), \quad (8)$$

$$H_m^s = LayerNorm(H_m^{LN_s} \oplus \nabla FeedForward(H_m^{LN_s})), \quad (9)$$

其中，$\nabla H_m$ $m$ $m \in \{a, t\}$ 是模态，内传播的信息。我们分别将两个自注意力模块中最后一个堆叠层的输出记为 $H_a^s$ 和 $H_t^s$。与此同时，我们对具有模态特定模式的联合特征进行中层融合，以帮助捕捉模态间和模态内的动态变化。各个模态（$x_a, x_t$）分别并行通过两个堆叠的 LSTM 层，以高效存储单模态特征（$H_a^m, H_t^m$），随后通过双向 LSTM 进行整合，以探索模态关联：

$$H_{mid} = FeedForward(LSTM([H_a, H_t])). \quad (10)$$

遵循先前的研究 [13], ，我们在音频和文本表示上采用残差连接，以保留数据的原始结构（$H_a^r, H_t^r$）。随后将其输入线性层和归一化层。最后，我们将它们拼接起来（$H_a^c, H_t^c, H_a^s, H_t^s, H_{mid}, H_a^r, H_t^r$），作为聚合表示，并将其输入两个全连接层进行分类。

## 4. 实验

### 4.1. 数据集与特征集

#### *4.1.1.* 数据集

多模态情感数据集（MELD）是一个多模态（视觉、音频和文本）、多说话者情绪识别数据集，包含电视剧《老友记》中的对话 [15]。MELD 包含超过 13000 条话语和 1400 段对话。每条话语均采用以下离散类别进行标注：愤怒、喜悦、悲伤、中性、厌恶、恐惧和惊讶。该数据集分为训练集、验证集和测试集，分别包含 9989、1109 和 2610 条话语 [15]。由于类别分布高度不平衡，因此使用加权 F1 分数评估性能。

#### *4.1.2.* 音频特征

每段话语的音频特征由两类特征组成：手工提取特征和瓶颈特征。作为手工提取特征，我们使用 openSMILE 工具包 [26]提取了 6552 个特征。这些特征涵盖了梅尔频率倒谱系数、基频等。此外，我们使用 OpenL3 模型 [27]提取了基于深度学习的瓶颈特征。OpenL3 是从 AudioSet [27]中学习得到的自监督嵌入。OpenL3 网络包含四个堆叠在一起的模块，每个模块由三个卷积层和一个池化层组成。我们对音频数据进行处理，使其符合网络输入要求，并提取了 6144 个高层瓶颈特征。

#### *4.1.3.* 文本特征

我们将预训练的 RoBERTa Large 模型（未区分大小写：隐藏层维度为1024、注意力头数为24、层数为24）微调为语言模型，用于从文本模态中提取特征 [25]。RoBERTa 以话语的转录文本作为输入，并从最后四层生成丰富的上下文文本表示。这样，对于输入中的每个词元，都会生成四个1024维向量。随后，我们将这些向量连接起来，得到维度为4096的与上下文无关的话语特征向量。

### 4.2. 基线方法

为进行全面评估，我们将模型与多模态情绪识别领域中的以下基线方法和最先进模型进行了比较（表1）。

• SMIN 将半监督学习与多模态交互相结合，用于对话情感识别 [29]。 • EmoCaps 利用 Transformer结构进行上下文情绪分类 [17]。 • EmbraceNet 采用基于Transformer的跨模态融合来估计情绪 [22]。 • CMN 使用基于注意力的跳跃连接来捕捉说话者间依赖关系 [23]。 • ICON 采用记忆网络存储上下文摘要，以进行分类 [30]。 • DialogueRNN 使用三个门控循环单元对说话者、上下文以及先前话语的情绪进行建模 [16]。 • bc-LSTM 执行上下文相关情感分析和情绪识别 [15]。

### 4.3. 实验设置

模型采用批量大小为 32 进行训练，共训练 200 轮；若验证损失连续 15 轮未下降，则停止训练。模型的隐藏神经元数量为 128。我们采用 Adam优化器，学习率设为 1e-4。$L_2$ 权重衰减设为 3e-4。网络训练使用交叉熵损失，并在最后一层采用 softmax激活函数。

表 1：所提出的方法、最先进水平的方法与基线方法之间的总体性能比较。评估指标采用加权 F1 分数。请注意，我们的单模态音频模型和文本模型均由 LSTM 层和残差块组成。

| 模型 | 音频 | text | 双模态 | 双模态 - 文本 | 参考文献 |
|---|---|---|---|---|---|
| SMIN | 45.2 | 63.65 | 64.5 | 0.85 | [29] |
| EmoCaps | 31.26 | 63.51 | 63.73 | 0.22 | [17] |
| EmbraceNet | 40 | 61 | 63.1 | 2.1 | [22] |
| CMN | 38.3 | 54.5 | 55.5 | 1 | [23] |
| ICON | 37.7 | 54.6 | 56.3 | 1.7 | [30] |
| DialogueRNN | 34 | 55.1 | 57 | 1.9 | [16] |
| bc-LSTM | 36.4 | 54.3 | 56.8 | 2.5 | [15] |
| **不含SCA单元的所提方法** | **43.10** | **63.83** | **65.66** | **1.83** | |
| **所提方法** | **43.10** | **63.83** | **66.28** | **2.45** | |

表2：MELD数据集上各类别的加权F1分数

| 所提方法 | 中性 | 惊讶 | fear | 悲伤 | joy | 厌恶 | 愤怒 | 加权平均F1 |
|---|---|---|---|---|---|---|---|---|
| 音频模态 | 62.9 | 20.12 | 0.0 | 6.7 | 22.65 | 0.0 | 53.21 | 43.10 |
| 文本模态 | 76.39 | 60.47 | 30.5 | 42.21 | 63.21 | 31.74 | 45.82 | 63.83 |
| 双模态 | 78.67 | 63.42 | 28.7 | 42.19 | 63.65 | 30.91 | 53.56 | 66.28 |

## 5. 结果与讨论

表 1 总结了最先进水平的基于文本的情绪识别对话系统的单模态和多模态变体，以及所提出的 CM-RoBERTa 模型的性能。值得注意的是，文本模态的表现明显优于音频模态（见表 1 和表 2）。然而，对于 MELD数据集而言，有几个因素使得利用音频进行对话中的情绪检测相当困难。首先，背景噪声很多。其次，话语长度比其他基准数据集（例如 IEMOCAP [31]）更短。更糟的是，大多数对话中有超过 5 位说话人。

文本通过在联合表示中提供丰富的情感特征，往往能有效缓解上述问题。词级、随时间变化的文本内容蕴含着大量信息，例如主题、观点、意图等。这些知识对于捕捉说话人情绪动态的本质及其变化过程至关重要。在简短交流中，某个特定词语（例如“amazing”）可能直接体现情绪状态，并主导最终决策。在我们的工作中，我们采用经过微调的 RoBERTa 模型提取话语级文本特征。与其他文本嵌入方法相比（见表 2），RoBERTa 的结果略好一些[25], ，这可能归功于动态变化的掩码模式，以及使用更多数据对模型进行更长时间的训练，从而使其与大规模语言模型更好地适配。

然而，对话情绪识别有时依赖音频而非文本，来推断对话参与者的情绪。我们发现，基于深度学习的高层声学特征——OpenL3嵌入，在区分对话情绪方面明显优于基频、梅尔频率倒谱系数和其他音频特征等传统低层特征（表3）。在七分类任务中，使用预训练OpenL3音频嵌入的模型性能比基线bc-LSTM模型高4.9%。预训练OpenL3模型适用于规模有限且类别不平衡的MELD数据集，因为知识迁移应能带来更好的高层情绪音频特征。在存在噪声和多个说话人的真实场景中，基频和梅尔频率倒谱系数可能并不适用于对话情绪识别。此外，在我们的研究中，通过结合低层描述符与高层OpenL3声学表示，相较于基于文本的单模态模型，惊讶和愤怒的F1分数分别提升了2.95%和7.74%（表2）。研究表明，愤怒最容易通过语音识别，其特点是基频和能量较高。

将音频模态的信息引入文本模态可带来显著优势，因为它能够融合额外的音频特征、消除语言含义的歧义，并弥合与真实环境之间的差距，从而提升情绪检测效果。多模态方法从各自的模态来源中提取并融合关键信息，捕捉到比单一模态更丰富的表示，并取得更好的性能。多种模态之间的模态内、模态间及跨模态交互、相关性和关系能够释放更多情绪相关内容，从而提升性能和预测效果。

表3：MELD数据集上的音频特征消融研究

| 方法 | 加权平均F1 |
|---|---|
| 不含低层描述符 | 37.3 |
| 不含openL3嵌入 | 41.3 |
| 我们的音频特征 | 43.1 |

我们提出的多模态融合框架在F1分数上比基于文本的单模态模型提升了2.45%，而SCA单元的改进使F1分数提升了约1.83%（见表1）。CM-RoBERTa的优异性能可能得益于以下原因。首先，早期的多模态交叉注意力融合了文本和音频特征之间的底层时间交互。我们通过时间感知引入交叉注意力校准，以推断异构知识之间相互排斥且相互依赖的关系。通过赋予相关音频模态更高的权重，我们能够引入音频模态的信息，帮助文本模态有效调整词语的权重。我们推测，交叉注意力融合单元能够有效地将不同知识映射到新的空间中，从而强化单一模态中嵌入的细节，并自适应地融合隐含的互补内容，以放大交互和相关性。其次，我们还采用了自注意力机制，在每个模态内部传播信息。通过对模态内交互进行建模，我们能够捕获并学习模态特定模式，并在每个模态的整个对话序列上分配自注意力权重。第三，中层融合和残差块旨在显式建模音频与文本之间的模态间和模态内交互。这一方面能够融合异构空间，保留上下文模态内特定表示，并促进同步。

## 6. 结论

本文提出了CM-RoBERTa模型，通过对用于语言处理的RoBERTa模型进行微调，并统一手工提取的音频特征与瓶颈音频特征，有效实现了多模态情感识别。得益于并行的自注意力和交叉注意力模块、中层融合及残差块，该模型能够显式建模音频与文本内部及两者之间的模态间和模态内交互，捕获长期上下文依赖关系，并学习模态特定模式。在MELD数据集上的评估结果表明，该模型实现了最先进水平的话语级识别性能。

我们的方法存在一些局限性。本研究依赖于受控数据库，其录音条件并不真实，且文本转录没有错误。因此，在实践中，需要针对对话中的多模态情绪识别问题设计更稳健的模型，以应对以下情况：(i) 自动语音识别出现意料之外的错误；(ii) 在模态含噪或缺失的情况下进行推理。

## 7. 参考文献

[1] Farhoudi Z、Setayeshi S，《利用深度学习特征与脑情感学习混合模型进行视听情感识别》，*Speech Communication*，2021年，127:92-103。[2] Yang K、Xu H、Gao K，《用于文本-音频情感分析的跨模态BERT》，发表于第*28*届*ACM*国际多媒体会议论文集，2020年。[3] Lee S、Han D、Ko H，《采用异构特征统一化的BERT多模态情感识别融合分析》，*IEEE Access*，2021年，9:94557-94572。[4] Sun L 等，《用于语音情感识别的多模态交叉与自注意力网络》，*IEEE*国际声学、语音与信号处理会议（*ICASSP*），2022年，4275-4279。[5] Chi P-H 等，《Audio Albert：用于音频表示自监督学习的轻量级BERT》，发表于*2021*年*IEEE*口语技术研讨会（*SLT*），2021年。[6] Ancilin J、Milton A，《利用梅尔频率幅度系数改进语音情感识别》，*Applied Acoustics*，2021年，179。[7] Aneesh M、Martin R，《使用四元数卷积神经网络进行语音情感识别》，*IEEE*国际声学、语音与信号处理会议（*ICASSP*），2021年，6309-6313。[8] Kumar P 等，《基于端到端三元组损失的语音情感识别情感嵌入系统》，发表于第*25*届国际模式识别会议（*ICPR*），2021年。[9] Neil S 等，《SERAB：用于语音情感识别的多语言基准》，*IEEE*国际声学、语音与信号处理会议（*ICASSP*），2022年，7697 7701-。[10] Sailunaz K 等，《文本和语音中的情绪检测：综述》，*Social Network Analysis and Mining*，2018年，8。[11] Khalil R 等，《使用深度学习技术进行语音情感识别：综述》，*IEEE Access*，2019年，7 117327 117345：-。[12] K 话语情绪识别中的 RoBERTa”，*arXiv*，*https://arxiv.org/abs/2108.12009*，2021。[13] KrishnaD N，Ankita P，“使用跨模态注意力和一维卷积神经网络进行多模态情绪识别”，*Interspeech*，2020，4243-4247。[14] Wang Y 等，“情感计算综述：情绪模型、数据库与最新进展”，信息融合，2022，19-52。[15] Soujanya P 等，“MELD：用于对话情绪识别的多模态多方数据集”，计算语言学协会第*57*届年会论文集，2019，527-536。[16] Navonil M 等，“DialogueRNN：用于对话情绪检测的注意力循环神经网络”，*AAAI*人工智能会议论文集，2019，6818-6825。[17] Li Z 等，“EmoCaps：基于情绪胶囊的对话情感识别模型”，计算语言学协会：*ACL*论文集，2022，1610–1618。[18] Abbaschian BJ，Sierra-Sosa D，Elmaghraby A，“语音情感识别的深度学习技术：从数据库到模型”，*Sensors*（巴塞尔），2021，21(4)。[19] Zhou J 等，“从对话语音数据推断情绪：一种半监督多路径生成神经网络方法”，第三十二届*AAAI*人工智能会议，2018，32:579-586。[20] Akhtar MS 等，“用于多模态情绪识别和情感分析的多任务学习”，*arXiv*，*https://arxiv.org/abs/1905.05812*，2019。[21] Zadeh A 等，“用于理解人类交流的多注意力循环网络”，*AAAI*人工智能会议论文集，2018。[22] Xie B，Sidulova M，Park CH，“基于Transformer的跨模态融合，实现对话中的稳健多模态情绪识别”，*Sensors*，2021，21(14):4913-4932。[23] Devamanyu H 等，“用于双人对话视频情绪识别的对话记忆网络”，*2019*年计算语言学协会北美分会会议：人类语言技术论文集，2019，1:2122-2132。[24] Feng K，Chaspari T，“自动情绪识别中可泛化迁移学习综述”，*Frontiers in ComputerScience*，2020，2。[25] Ghosal D 等，“COSMIC：用于对话情绪识别的常识知识”，计算语言学协会论文集（*EMNLP*），2020。[26] Eyben F，Wllmer M，Schuller B，“openSMILE：慕尼黑推出的灵活快速开源音频特征提取器”，*ACM*多媒体国际会议，2010。[27] Cramer J 等，“看、听，了解更多：深度音频嵌入的设计选择”，*IEEE*声学、语音与信号处理国际会议（*ICASSP*），2019。[28] Vaswani A 等，“注意力就是你所需要的一切”，神经信息处理系统进展第*30*卷，2017。

[29] Zheng Lian BL 和 Jianhua Tao，“SMIN：用于对话情感识别的半监督多模态交互网络”，*IEEE*情感计算汇刊，2022，1-15。[30] Hazarika D 等，“ICON：用于多模态情绪检测的交互式对话记忆网络”，*2018*年自然语言处理实证方法会议论文集，2018。[31] Busso C 等，“IEMOCAP：交互式情绪双人动作捕捉数据库”，《语言资源与评估》期刊，2008，42(4)，335-359。