arXiv:2303.01694v1 [cs.SD] 3 Mar 2023

DWFormer：用于语音情感识别的动态窗口 Transformer

Shuaiqi Chen[1], *Xiaofen Xing*[1*], *Weibin Zhang*[2],*Weidong Chen*[1],*Xiangmin Xu*[3 4]

[1]华南理工大学电子与信息工程学院，中国[2] VoiceAI Technologies，中国深圳 [3] 琶洲实验室，中国[4] 华南理工大学未来技术学院，中国

摘要

语音情感识别对于人机交互至关重要。代表不同情感的时间区域在语音的不同局部部分中分散出现。此外，重要信息的时间尺度在语音片段内部及不同语音片段之间可能存在很大差异。尽管基于 Transformer 的模型在该领域取得了进展，但现有模型无法在不同时间尺度上精准定位重要区域。为了解决这一问题，我们提出了动态窗口 Transformer（DWFormer），这是一种通过依据时间重要性动态地将样本拆分为窗口的新型架构。在窗口内部应用自注意力机制，以细粒度方式捕获局部的时间重要信息。同时，模型还考虑了跨窗口信息交互，以实现全局通信。我们在 IEMOCAP 和 MELD 数据集上对 DWFormer 进行了评估。实验结果表明，与此前最先进方法相比，所提出的模型取得了更优的性能。

索引词— 语音情感识别、Transformer、语音信号处理

1. 引言

语音情感识别（SER）是人机交互的关键。为了使人机交互更加自然，机器必须准确捕捉情绪并做出恰当的回应。

几十年来，人们一直在研究语音情感识别。近年来，基于 Transformer 的模型推动语音情感识别领域取得了巨大进步 [1, 2, 3, 4]。标准 Transformer [5] 在建模语音序列中的长程依赖方面表现出色。然而，其核心机制——全局自注意力机制——容易受到噪声干扰，也可能无法聚焦于与情绪所在位置相同的区域 [1]。这限制了 Transformer 模型的有效性。能够显著体现情绪的声音事件，例如语调和语速的变化、笑声和叹息声，都位于局部区域。此外，重要信息的时间尺度在不同语音片段内部及片段之间变化范围很大（见图 1）。 [2] 采用局部窗口注意力机制

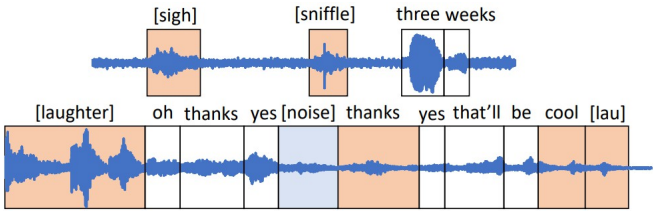

图 1。示例选自 IEMOCAP[6]。 [lau]表示笑声。语音的局部区域中存在着指示不同情绪的重要声音事件，例如笑声、叹息、抽鼻声以及积极语义等，其持续时间各不相同。

使模型能够更加关注局部变化。然而，固定不变的窗口长度限制了这些模型捕捉随不同时间尺度变化的情感信息的能力。

在计算机视觉领域，动态设计使模型能够拥有灵活的感知范围，从而捕捉不同形状的目标。在语音情感识别领域， [10,11, 12] 提出了局部-全局架构，以捕捉重要的时间信息。受此启发，本文提出了一种名为动态窗口 Transformer（DWFormer）的新架构，以解决上述问题。该架构的核心 DWFormer 块由动态局部窗口变换器（DLWT）模块和动态全局窗口 Transformer（DGWT）模块组成。DLWT 将输入特征动态划分为多个尺度的窗口，并在每个窗口中捕捉局部重要信息。DGWT 在 DLWT 处理后重新衡量窗口之间的重要性。DLWT 与 DGWT 的结合有助于模型发现与任务相关的区域。本文的主要贡献如下：

(i) 提出了一种名为动态窗口变换器（DWFormer）的新架构，以深入探究语音情感识别中如何捕获不同长度的重要时间信息这一问题。

(ii) 我们在 IEMOCAP 和 MELD [13]数据集上评估了 DWFormer，并证明其性能显著优于标准 Transformer 和固定窗口 Transformer。此外，DWFormer 的表现与传统研究及最先进方法相当

*通讯作者。xfxing@scut.edu.cn

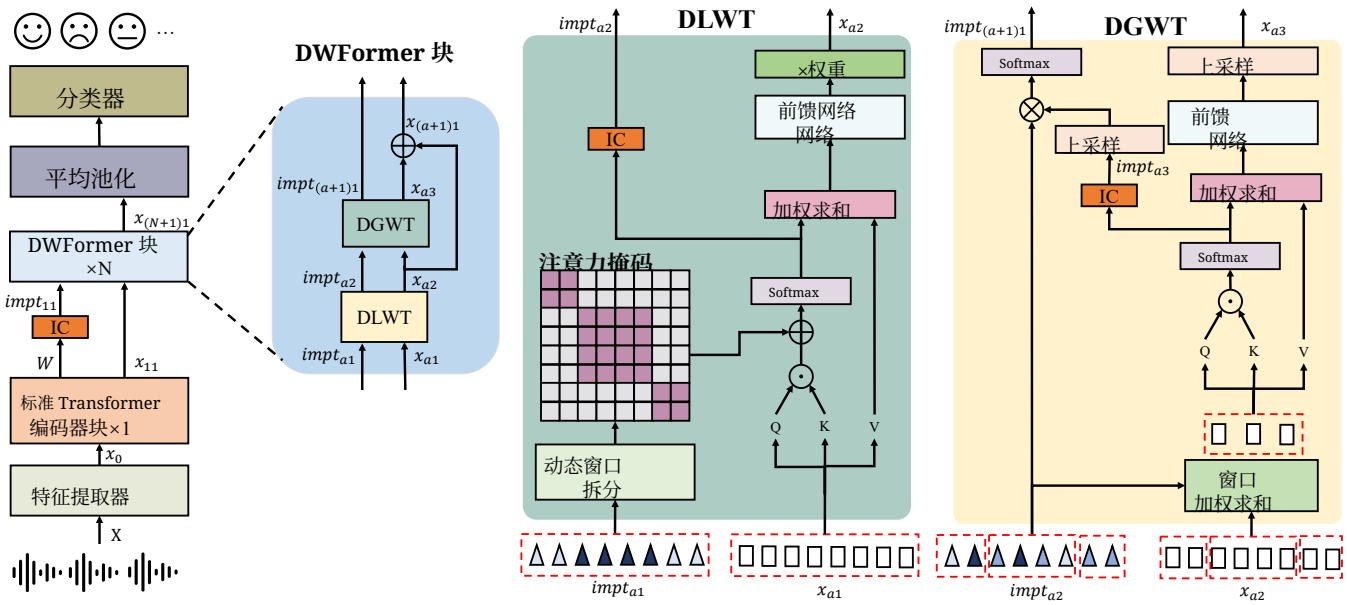

图 2。DWFormer 的模型架构。为简洁起见，图中未绘制残差连接和层归一化。IC 表示重要性计算模块（详见图 3）。三角形序列 $impt$ 表示重要性权重。颜色越深，词元越重要。矩形序列 x 表示特征图。

。代码将公开发布。[1]

2. 方法

DWFormer 的架构如图 2 所示。模型架构的核心是 DWFormer 块，它由动态局部窗口 Transformer 模块和动态全局窗口 Transformer 模块组成。下面将介绍模型的各个组成部分。

输入音频信号首先送入特征提取器以提取特征 $x_0 \in \mathbb{R}^{T\times D}$，其中 T 表示特征词元的数量，D 表示特征维度。随后，x_0 经过一个标准 Transformer 编码器层。编码器层的输出包括隐藏特征 x_{11} 和注意力权重 $W \in \mathbb{R}^{H\times T\times T}$，其中 H 表示注意力头的数量。W 被送入重要性计算模块，以获得时间重要性估计，这是第一个 DWFormer 块所必需的。

2.1. 重要性计算模块

提出了重要性计算（IC）模块，用于衡量 token 的重要性。受 [1]，启发，IC 模块利用从 Transformer 中获取的注意力权重进行计算。该过程如图 3 所示，具体描述如下：

$$impt = Softmax[\sum_{1}^{T_1}(\frac{1}{H}\sum_{s=1}^{H} aw_s)] \qquad (1)$$

其中，aw_s 表示来自 s头的注意力权重，T_1 表示平均矩阵 aw_{avg}的行长度。Softmax 函数

[1]https://github.com/scutcsq/DWFormer

用于归一化。

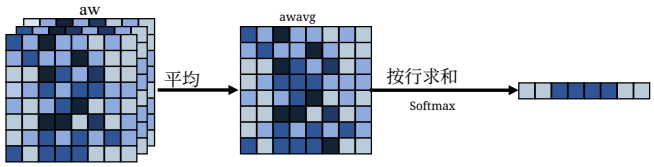

图 3。IC 模块根据注意力权重计算重要性。

IC 模块获得的每个令牌的重要性分数 $impt_{11}$ 与隐藏特征 x_{11} 随后被传递至 N 个堆叠的 DWFormer 块中，以进行进一步评估。

2.2. DWFormer 块

2.2.1. 动态局部窗口变换器模块

动态局部窗口变换器（DLWT）模块会为输入特征动态划分区域，并通过局部关系建模捕获重要信息。具体流程如下所述。

首先，利用**动态窗口划分**（DWS）操作，根据从 IC 组件获得的重要性值，将特征词元动态拆分为长度不等的窗口。如图4所示，根据前一模块计算的重要性分数，按时间顺序将重要性分数高于/低于阈值的词元分别归入若干强/弱情感关联窗口。阈值设为所有重要性值的中位数。 A 强情感关联窗口

并且， B 强情感关联窗口由 x_{a1}获得。

为了按批次处理数据，窗口划分结果通过注意力掩码机制实现：

$$M_{ij} = \begin{cases} 0, & (b_{w_k} \le i \le e_{w_k}, b_{w_k} \le j \le e_{w_k} \\ & k = 1, ..., A+B) \\ -\infty, & otherwise \end{cases} \quad (2)$$

其中， M_{ij} 是注意力掩码 $M \in \mathbb{R}^{T\times T}$中第 i行第 j列的值。 b_{w_k}和 e_{w_k} 分别是第 k个窗口的行列起始索引和结束索引。

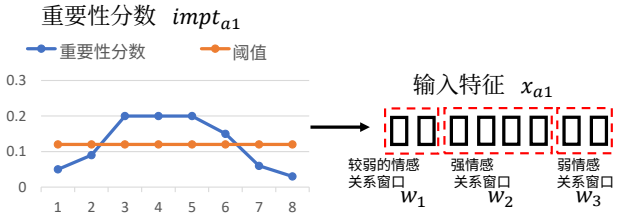

图4。动态窗口划分的操作。

随后，每个窗口都会经过一个 Transformer 编码器，以进行窗口内信息交互，其定义如下：

$$DLWT(x_{a1}) = FFN(Softmax(\frac{Q_{a1}K_{a1}^{\mathrm{T}}}{\sqrt{d_h}} + M)V_{a1}) \quad (3)$$

其中，FFN 表示前馈网络， Q_{a1}、 K_{a1}、V_{a1} 是特征 x_{a1} 的投影映射，T 表示转置操作， d_h 是缩放因子。

对于弱情感相关窗口，前一模块学习到的先验知识表明，这些窗口很可能对情感识别而言是冗余的，因此其中词元的特征会乘以权重 λ（≤ 1）；而强情感关联窗口中的特征则乘以 1。输出定义为 $x_{a2} \in \mathbb{R}^{T\times D}$。

每个窗口内各词元的时间重要性由 IC 模块计算。随后，将所有窗口的计算结果按照时间顺序拼接成一个序列，记为 $impt_{a2} \in \mathbb{R}^{T}$。 x_{a2} 和$impt_{a2}$ 被传入动态全局窗口 Transformer 模块进行进一步处理。

2.2.2. 动态全局窗口 Transformer

动态全局窗口 Transformer（DGWT）模块采用整体方法，在 DLWT 之后重新衡量窗口之间的重要性关系。具体而言，每个窗口首先通过**窗口加权求和**操作生成一个词元，其定义为：

$$wt_k = \sum_{p=b_{w_k}}^{e_{w_k}} impt_{a2_p} \times x_{a2_p} \quad (4)$$

其中， p 是词元的索引。 $wt \in \mathbb{R}^{(A+B)\times D}$ 是窗口词元序列。

随后，将序列 wt 传入 Transformer 编码器以进行全局交互，其定义如下：

$$DGWT(wt) = FFN(Softmax(\frac{Q_{wt}K_{wt}^{\mathrm{T}}}{\sqrt{d_h}})V_{wt}) \quad (5)$$

其中， Q_{wt}、 K_{wt}、 V_{wt} 是 wt的投影映射。

接下来，通过复制窗口的向量，将每个窗口词元上采样至对应窗口的相同长度。然后，将这些词元连接成一个序列，记为 $x_{a3} \in \mathbb{R}^{T\times D}$。DWFormer 块 $x_{(a+1)1}$ 的输出是 x_{a2} 和 x_{a3} 的求和，从而使每个词元同时获得局部和全局信息。

窗口 $impt_{a3} \in \mathbb{R}^{A+B}$ 之间的重要性分数通过 IC 计算。经过一个 DWFormer 块后，每个词元 $impt_{(a+1)1}$ 的重要性通过以下方式重新测量：

$$impt_{(a+1)1} = Softmax(impt_{a2} \times Upsampling(impt_{a3})) \quad (6)$$

上采样操作与上述所述相同。

在下一个 DWFormer 块中， $x_{(a+1)1}$ 根据 $impt_{(a+1)1}$被划分为窗口。最后，对 N个 DWFormer 块的输出特征 $x_{(N+1)1}$ 应用时间平均池化层，再通过多层感知机分类器进行情感分类。

3. 实验

3.1. 实验设置

我们在 IEMOCAP 和 MELD 数据集上评估 DWFormer。在 IEMOCAP 数据集上，使用 5 折留一部分交叉验证评估 DWFormer。选取 4 种情绪（快乐与兴奋、愤怒、悲伤和中性）进行分类。评估指标为加权准确率（WA）和非加权准确率（UA）。对于包含 7 种情绪（愤怒、厌恶、恐惧、喜悦、中性、悲伤、惊讶）的 MELD 语料库，报告测试集上的加权 F1（WF1）分数。

将预训练WavLM-Large[14] 模型的第 12 个 Transformer 编码器层的输出特征用作音频特征。IEMOCAP 使用 3 个 DWFormer 块，MELD 使用 2 个。注意力头数为 8。激活函数为 ReLU。批量大小为 32。IEMOCAP 的学习率初始化为 3e-4，MELD 的学习率则初始化为 5e-4。 λ 的值为 0.85。我们使用 SGD优化器训练 120 个训练轮次，并采用带余弦预热调度器的余弦衰减学习率调度器。优化函数为交叉熵损失。

3.2. 与基线网络的比较

标准 Transformer，以及将输入特征划分为等长窗口

并在每个窗口内应用自注意力的固定窗口 Transformer，被选作基线网络。基线网络的参数与 DWFormer 相同。固定窗口 Transformer 的窗口长度与 DWFormer 中窗口的平均长度相同。为验证 DWFormer 各模块的有效性，还进行了消融实验。

表 1 中的结果表明，DWFormer 在 IEMOCAP 和 MELD 数据集上的表现均优于标准 Transformer 和固定窗口 Transformer。同时，从 DWFormer 中移除 DLWT 或 DGWT 模块都会导致性能显著下降。

模型	IEMOCAP		MELD
	加权准确率（%）	非加权准确率（%）	加权 F1 分数（%）
标准 Transformer	70.7	71.9	47.1
固定窗口 Transformer	71.2	72.3	47.6
DWFormer（我们的方法），不含 DLWT	71.5	72.4	47.8
DWFormer（我们的方法），不含 DGWT	71.5	72.7	47.7
DWFormer（我们的方法）	**72.3**	**73.9**	**48.5**

表 1。与基线网络的比较结果。

3.3. 与传统研究的比较与可视化

由于我们的模型采用局部-全局架构，因此我们开展了与传统研究的对比实验。 [10] 是开源的，因此我们首先复现 [10] 的结果，以确保代码的正确性，然后在我们的实验设置下测试该模型（**实验1**）。由于 [11, 12] 的代码未公开，我们按照其论文中描述的实验设置测试模型（**实验2**：随机划分为 80% 的训练集和 20% 的测试集）。实验结果如表2所示，表明与其他传统研究相比，我们的模型表现更优。

实验设置	模型	IEMOCAP		MELD
		加权准确率（%）	非加权准确率（%）	加权 F1 分数（%）
实验1	[10]	59.6	60.5	39.8
	DWFormer（我们的方法）	**72.3**	**73.9**	**48.5**
实验2	[11]	69.4	70.2	-
	[12]	70.3	70.8	-
	DWFormer（我们的方法）	**76.3**	**77.2**	-

表2 。与其他传统研究的对比结果 。

此外，可视化结果如图5所示。如图5所示，标准 Transformer、固定窗口 Transformer 和 ATDA 在定位重要的时间信息方面均不如我们的方法。

3.4. 与先前最先进方法的比较

表 3 展示了以往最先进方法与 DWFormer 的对比结果。实验结果证明，我们的方法优于以往最先进方法。

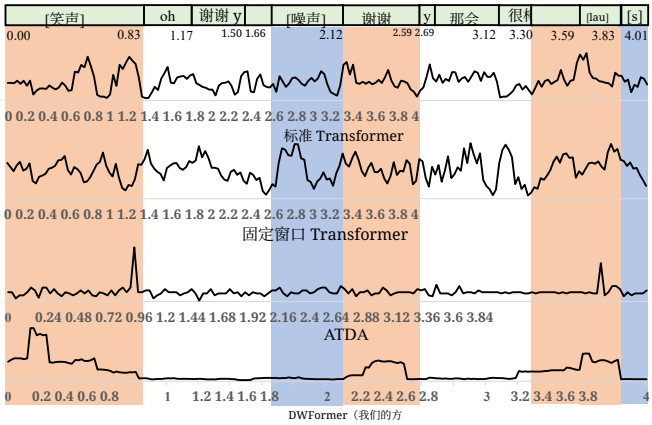

图5。图中展示了标准 Transformer、固定窗口 Transformer、ATDA 和 DWFormer 的可视化结果。横轴表示时间顺序，纵轴表示重要性分数。 [s] 表示静音， [lau] 表示笑声，y 表示“yes”。黄色边框标出了需要重点关注的区域，例如笑声、口音（“thanks”）和积极语义（“cool”）。蓝色边框表示不应关注的区域，例如噪声和静音。

数据集	模型	加权准确率（%）	非加权准确率（%）	加权 F1 分数（%）
IEMOCAP	[Chen 等人， 2022][2]	62.9	64.5	-
	[Li 等人 2022][15]	68.0	68.2	-
	[Zou 等人， 2022][16]	69.8	71.1	-
	DWFormer（我们的方法）	**72.3**	**73.9**	-
MELD	[Chudasama 等人， 2022][17]	-	-	35.8
	[Chen 等人， 2022][2]	-	-	41.9
	[Lian 等人， 2022][18]	-	-	45.2
	DWFormer（本文方法）	-	-	**48.5**

表3. 与先前最先进方法的比较结果。

4. 结论

我们提出了一种新的基于 Transformer 的框架 DWFormer，旨在语音情感识别领域内及不同样本之间，以不同尺度捕捉重要的时间区域。我们通过实验证明，DWFormer 优于以往最先进方法。消融研究证明了 DLWT 和 DGWT 模块的有效性。凭借定位重要信息的能力，我们计划将 DWFormer 应用于病理语音识别领域，帮助研究人员了解疾病对发音的影响。

5. 致谢

本研究部分获以下项目资助：广东省自然科学基金（2022A1515011588）；国家重点研发计划（2022YFB4500600）；广州科技计划项目（202103010002）；广东省科技计划项目（2022B0101010003）；国家自然科学基金（U1801262）；广东省人体数字孪生重点实验室（2022B1212010004）

6. 参考文献

[1] Weidong Chen、Xiaofeng Xing、Xiangmin Xu、Jichen Yang 和 Jianxin Pang，“用于多模态语音情感识别的键稀疏 Transformer”，载于*ICASSP 2022-2022 IEEE* 国际声学、语音与信号处理会议（*ICASSP*）。IEEE，2022 年，第 6897–6901 页。

[2] Weidong Chen、Xiaofen Xing、Xiangmin Xu、Jianxin Pang 和 Lan Du，《SpeechFormer：一种融合语音特征的高效分层框架》，载于*Proc. Interspeech 2022*，2022 年，第 346–350 页。

[3] Xianfeng Wang、Min Wang、Wenbo Qi、Wanqi Su、Xi- angqian Wang 和 Huan Zhou，“一种具有堆叠 Transformer 层的新型端到端语音情感识别网络”，载于*ICASSP 2021-2021 IEEE* 声学、语音与信号处理国际会议 *(ICASSP)*。IEEE，2021，第 6289–6293 页。[4] Yuhua Wang、Guang Shen、Yuezhu Xu、Jiahang Li 和 Zhengdao Zhao，“在用于语音情感识别的多模态 Transformer 中学习互相关”，载于*Interspeech*，2021，第 4518–4522 页。[5] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、 Ł ukasz Kaiser 和 Illia Polosukhin，“Attention is all you need”，神经信息处理系统进展，第 30 卷，2017。[6]Carlos Busso、Murtaza Bulut、Chi-Chun Lee、Abe Kazemzadeh、Emily Mower、Samuel Kim、Jeannette N Chang、Sungbok Lee 和 Shrikanth S Narayanan，“Iemocap：交互式情绪双人动作捕捉数据库”，语言资源与评估，第 42 卷，第 4 期，第 335–359 页，2008。[7] Pengzhen Ren、Changlin Li、Guangrun Wang、Yun Xiao、Qing Du、Xiaodan Liang 和 Xiaojun Chang，“超越固定注视：动态窗口视觉 Transformer”，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2022，第 11987–11997 页。[8] Yulin Wang、Kangchen Lv、Rui Huang、Shiji Song、Le Yang 和 Gao Huang，“一瞥与聚焦：一种减少图像分类中空间冗余的动态方法”，载于神经信息处理系统进展，H. Larochelle、M. Ranzato、R. Hadsell、M.F. Balcan 和 H. Lin 编，2020，第 33 卷，第 2432–2444 页，Curran Associates, Inc.。[9] Sheng Wan、Chen Gong、Ping Zhong、Bo Du、Lefei Zhang 和 Jian Yang，“用于高光谱图像分类的多尺度动态图卷积网络”，*IEEE* 地球科学与遥感汇刊，第 58 卷，第 5 期，第 3162–3177 页，2020。

[10] Lu-Yao Liu、Wen-Zhe Liu、Jian Zhou、Hui-Yuan Deng 和 Lin Feng，“ATDA：用于语音情感识别的注意力时序动态激活”，知识库系统，第 243 卷，第 108472 页，2022。[11] Jiaxing Liu、Zhilei Liu、Longbiao Wang、Lili Guo 和 Jianwu Dang，“通过局部-全局感知深度表征学习进行语音情感识别”，载于*ICASSP 2020-2020 IEEE* 声学、语音与信号处理国际会议 *(ICASSP)*。IEEE，2020，第 7174–7178 页。[12] Jiaxing Liu、Zhilei Liu、Longbiao Wang、Yuan Gao、Lili Guo 和 Jianwu Dang，“基于潜在表征的语音情感识别时序注意力卷积网络”，载于*INTERSPEECH*，2020，第 2337–2341 页。[13] Soujanya Poria、Devamanyu Hazarika、Navonil Majumder、Gautam Naik、Erik Cambria 和 Rada Mihalcea，“MELD：用于对话情感识别的多模态多方数据集”，*arXiv* 预印本 *arXiv:1810.02508*，2018。[14] Sanyuan Chen、Chengyi Wang、Zhengyang Chen、Yu Wu、Shujie Liu、Zhuo Chen、Jinyu Li、Naoyuki Kanda、Takuya Yoshioka、Xiong Xiao 等，“WavLM：面向全栈语音处理的大规模自监督预训练”，*IEEE*信号处理专题期刊，2022。[15] Jinchao Li、Shuai Wang、Yang Chao、Xunying Liu 和 Helen Meng，“面向情感识别的上下文感知多模态融合”，载于*Proc. Interspeech 2022*，2022，第 2013–2017 页。[16] Heqing Zou、Yuke Si、Chen Chen、Deepu Rajan 和 Eng Siong Chng，“基于协同注意力的多层级声学信息语音情感识别”，载于*ICASSP 2022 - 2022 IEEE* 声学、语音与信号处理国际会议 *(ICASSP)*，2022，第 7367–7371 页。[17] Mixiao Hou、Zheng Zhang 和 Guangming Lu，“采用自引导模态校准的多模态情感识别”，载于*ICASSP 2022 - 2022 IEEE* 声学、语音与信号处理国际会议 *(ICASSP)*，2022，第 4688–4692 页。[18] Zheng Lian、Bin Liu 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，第 1–1 页，2022。