带时间对齐的多模态自注意力音视频情感识别网络

Inyong Koo[1] *Yeeun Seong*[2] *Minseok Son*[1] *Jaehyuk Jang*[1] *Changick Kim* [1]

韩国科学技术院（KAIST）[1]电气工程学院 [2]绿色增长与可持续发展研究生院 大田，韩国

arXiv:2603.11095v1 [cs.MM] 11 Mar 2026

摘要

视听情感识别（AVER）方法通常会融合话语级特征，即使是帧级注意力模型也很少处理不同模态之间的帧率不匹配问题。本文提出了一种基于Transformer的框架，重点关注多模态特征的时间对齐。我们设计了一种多模态自注意力编码器，在共享特征空间中同时捕获模态内和模态间的依赖关系。为解决异构采样率问题，我们引入时间对齐旋转位置嵌入（TaRoPE），以隐式同步音频和视频令牌。此外，我们提出跨时间匹配（CTM）损失，强制时间邻近对之间保持一致性，引导编码器实现更好的对齐。在CREMA-D和RAVDESS数据集上的实验表明，与近期基线方法相比，该方法取得了稳定的性能提升，说明显式处理帧率不匹配有助于保留时间线索并增强跨模态融合。

索引词— 多模态情感识别、Transformer 编码器、时间对齐、旋转位置嵌入（RoPE）、跨时间匹配损失

1. 引言

从语音和面部表情等多模态信号中理解人类情感，是情感计算和人机交互领域的一项核心挑战 [1]。视听情感识别（AVER）受益于不同模态间的互补线索：语音中的韵律和语调常与视频中的面部动作单元及表情同时出现。然而，有效建模并同步这些异构信号仍是一个长期存在的问题。

早期的 AVER 系统通常在话语级别提取模态特定表示，然后通过简单拼接进行融合以完成分类 [2, 3, 4, 5, 6]。这种设计简单且稳健，但会将模态特定的时间动态压缩为全局嵌入，从而限制细粒度的跨模态交互。较新的研究尝试利用各种交叉注意力架构来挖掘音频和视频序列之间的跨模态依赖关系 [7, 8, 9, 10],，但通常仅依赖特征级相似度，而未纳入时间线索。具体而言，基于 Transformer 的架构 [11, 12] 通过位置嵌入捕捉各模态内部的顺序，但对模态间的相对时间结构缺乏考量，因此未能充分建模跨模态同步。

当我们考察音频和视频表征不同的时间粒度时，这一挑战便显而易见。音频特征以更高的时间分辨率提取（例如 50 FPS），而视频特征通常较为粗略（例如 30 FPS），从而形成不同步的令牌序列。若缺乏同步，跨模态注意力可能分散到无关位置，削弱细粒度的多模态关联。

本文从时间对齐的角度研究多模态融合。我们提出了一种基于 Transformer 的框架，将音频和视频特征投影到共享嵌入空间中，并通过统一的多模态自注意力编码器同时捕捉模态内和模态间的依赖关系。为解决不同流之间的帧率差异，我们引入时间对齐旋转位置嵌入（TaRoPE），这是 RoPE [13] 的一种变体，旨在沿一致的时间轴隐式同步词元。此外，我们提出跨时间匹配（CTM）损失，利用时间戳感知的高斯相似度，鼓励时间上接近的视听配对具有更高相似度。

我们在两个基准数据集 CREMA-D [14] 和 RAVDESS [15]上验证了所提方法。实验结果表明，我们的方法优于表现强劲的基于注意力的基线方法，证明显式建模帧率不匹配能够实现更有效的多模态融合。

2. 方法论

本节介绍用于提取和融合视听特征的架构、用于隐式同步的时间对齐 RoPE（TaRoPE）

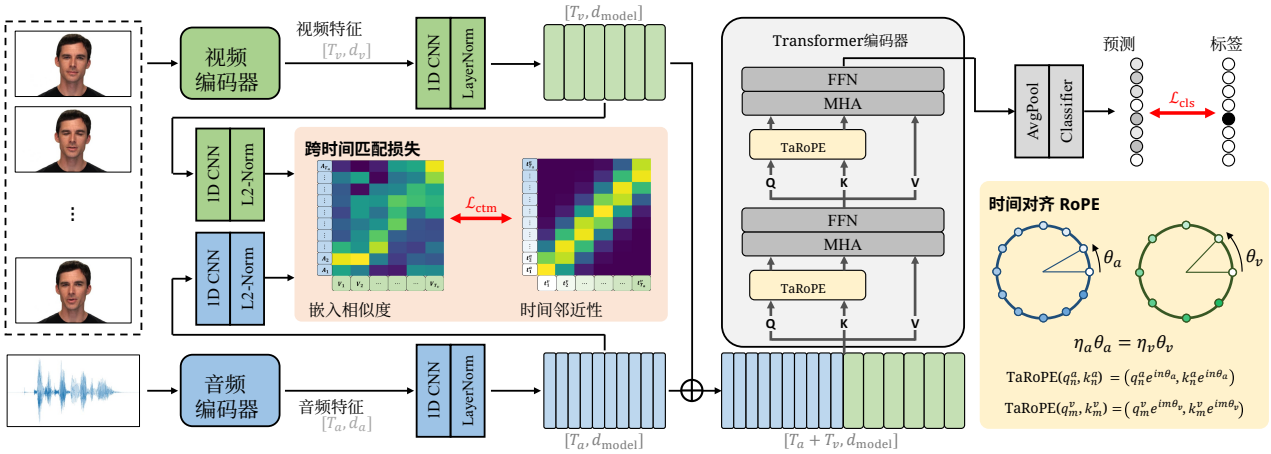


图1。音频和视频特征首先由特定模态编码器提取，并投影到共享嵌入空间中。时间对齐 RoPE（TaRoPE）通过使旋转位置嵌入适应不同的时间分辨率，在统一的Transformer编码器内隐式同步异构的音视频序列。跨时间匹配（CTM）损失通过沿共享时间轴对齐时间上邻近的音视频特征对，显式地强制时间一致性。最终表示被池化用于情绪分类。

以及用于显式时间一致性的跨时间匹配（CTM）损失。图 1 展示了所提出框架的概览。

2.1. 架构

该框架首先使用专用编码器处理原始音频和视频输入，以提取特定模态的特征。对于音频模态，我们采用预训练的 xlsr-Wav2Vec 2.0 编码器 [16], ，其生成 1024 维的帧级嵌入。对于以 16kHz 采样的输入，编码器内部 320 倍的下采样率使有效帧率达到 $\eta_a = 50$ FPS。对于视频模态，我们使用 OpenFace 库 [18]，从每张以 $\eta_v = 30$ FPS 采样的人脸帧中提取 35 个动作单元特征描述符。这些基于肌肉的特征由面部动作编码系统（FACS） [17], 定义。

为便于多模态集成，首先将音频和视频特征线性投影到共享的 d_{model} 维嵌入空间：

$$\mathbf{F}^a = [f_1^a, f_2^a, \ldots, f_{T_a}^a] \in \mathbb{R}^{T_a \times d_{\text{model}}}, \tag{1}$$

$$\mathbf{F}^v = [f_1^v, f_2^v, \ldots, f_{T_v}^v] \in \mathbb{R}^{T_v \times d_{\text{model}}}, \tag{2}$$

其中$\mathbf{F}^a$ 和$\mathbf{F}^v$ 分别是长度为 T_a 和 T_v的帧级音频和视频特征序列。这些序列作为词元嵌入输入统一的Transformer编码器，该编码器包含两个结合TaRoPE的多模态自注意力模块，以确保注意力过程中的时间对齐。随后应用时间平均池化，以获得话语级表示。最后，将该表示传入Softmax分类器，以预测目标情感。

2.2. 时间对齐 RoPE

多模态融合中一个基础但尚未得到充分研究的问题，是音频和视频序列之间时间分辨率的不匹配。为解决这一问题，我们提出了时间对齐旋转位置嵌入（TaRoPE），通过 [13] 调整RoPE，以同步异构帧率。

在标准 RoPE 公式中，位置信息通过根据索引旋转查询向量和键向量 q_n， $k_n \in \mathbb{R}^{d_{\text{model}}}$（由词元嵌入 x_n 投影得到）注入其中：

$$(q_n, k_n) \mapsto (q_n e^{in\theta_{\text{base}}},\ k_n e^{in\theta_{\text{base}}}), \tag{3}$$

其中 n 表示词元位置， θ_{base} 为基频。该构造确保位置 n 与 m之间的注意力分数 ⟨RoPE(q_n),RoPE(k_m)⟩仅取决于它们的索引距离（$n - m$），从而在自注意力机制中直接编码相对位置信息。

设 f_n^a和 f_m^v 分别表示第 n个音频帧与第 m个视频帧的特征，查询向量和键向量由此导出。TaRoPE施加模态特定旋转，即

$$(q_n^a, k_n^a) \mapsto (q_n^a e^{in\theta_a},\ k_n^a e^{in\theta_a}), \tag{4}$$

$$(q_m^v, k_m^v) \mapsto (q_m^v e^{im\theta_v},\ k_m^v e^{im\theta_v}), \tag{5}$$

与 $\theta_v = \frac{\eta_a}{\eta_v}\theta_a$。这会将视频位置重新映射到音频时间轴，确保跨模态注意力取决于一致的时间距离。

2.3. 跨时间匹配损失

为显式地强化时间一致性，我们引入跨时间匹配（CTM）损失，鼓励沿共享物理时间轴在时间上相近的音频和视频帧具有相似的表征。

在 Transformer编码器之前，两个模态都被投影到 L2归一化嵌入空间 $\mathbb{R}^{d_{\text{emb}}}$。给定嵌入 $\mathbf{E}^a = [A_1, A_2, \ldots, A_{T_a}]$ 和 $\mathbf{E}^v = [V_1, V_2, \ldots, V_{T_v}]$，以及定义在共同时间线上的时间戳 t_i^a 和 t_j^v，我们定义时间高斯亲和度如下：

$$g_{ij} = \exp\Big(-\tfrac{(t_i^a - t_j^v)^2}{2\sigma^2}\Big), \tag{6}$$

其中，σ 控制对错位的容忍度。特征相似度表示为 $s_{ij} = A_i^\top V_j$，并通过温度缩放的softmax（$\tau > 0$）进行归一化。因此，预测分布 p 由 $\{s_{ij}\}$得到，而目标分布 q 则根据 $\{g_{ij}\}$导出：

$$p_{ij}^{a\to v} = \frac{\exp(s_{ij}/\tau)}{\sum_{j'} \exp(s_{ij'}/\tau)}, \qquad q_{ij}^{a\to v} = \frac{g_{ij}}{\sum_{j'} g_{ij'}}, \tag{7}$$

$$p_{ij}^{v\to a} = \frac{\exp(s_{ij}/\tau)}{\sum_{i'} \exp(s_{i'j}/\tau)}, \qquad q_{ij}^{v\to a} = \frac{g_{ij}}{\sum_{i'} g_{i'j}}. \tag{8}$$

CTM损失通过双向交叉熵对齐这些分布：

$$\mathcal{L}_{a\to v} = \tfrac{1}{T_a} \sum_{i=1}^{T_a} \mathrm{CrossEntropy}\big(q_{i\cdot}^{a\to v},\, p_{i\cdot}^{a\to v}\big), \tag{9}$$

$$\mathcal{L}_{v\to a} = \tfrac{1}{T_v} \sum_{j=1}^{T_v} \mathrm{CrossEntropy}\big(q_{\cdot j}^{v\to a},\, p_{\cdot j}^{v\to a}\big), \tag{10}$$

$$\mathcal{L}_{\text{ctm}} = \tfrac{1}{2}\big(\mathcal{L}_{a\to v} + \mathcal{L}_{v\to a}\big). \tag{11}$$

此处，符号 $q_{i\cdot}^{a\to v}$和 $p_{i\cdot}^{a\to v}$表示在所有视频帧上的第 i 行分布，而 $q_{\cdot j}^{v\to a}$和 $p_{\cdot j}^{v\to a}$表示在所有音频帧上的第 j 列分布。

最终的训练目标将此项与分类损失结合起来：

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + \lambda_{\text{ctm}}\, \mathcal{L}_{\text{ctm}}, \tag{12}$$

其中，λ_{ctm} 表示CTM损失的权重。

3. 实验

3.1. 数据集与评估指标

CREMA-D数据集 [14] 包含7,442个由91名演员（48名女性、43名男性）说出的短句片段，这些演员具有多样的年龄和族裔背景。每条话语被标注为六种基本情绪之一：愤怒、厌恶、恐惧、快乐、中性和悲伤。在评估中，我们采用 [19]提供的说话人独立的训练/验证划分。

表 1。CREMA-D和RAVDESS数据集上的视听情感识别性能（%）对比。

方法	Year	CREMA-D	RAVDESS
Ghaleb等人 [20]	2019	74.00	67.70
MATER [21]	2020	67.20	76.30
TA-AVN [4]	2021	84.00	78.70
Luna-Jim´enez等人 [2]	2021	-	80.08
Luna-Jim´enez等人 [5]	2021	-	86.70
V8 +A4 [6]	2022	-	86.00
Mocanu等人 [8]	2023	84.57	87.85
Lei等人 [3]	2023	85.06	-
AttA-Net [7]	2024	-	88.00
HiCMAE [9]	2024	84.91	87.96
VQ-MAE-AV [10]	2025	80.40	84.80
ATTSF-Net [12]	2025	-	88.67
Ours	2025	**89.49**	**89.25**

RAVDESS数据集 [15] 由24名专业演员（12名女性、12名男性）产生的1,440条话语组成，涵盖八种类别情感：中性、平静、快乐、悲伤、愤怒、恐惧、惊讶和厌恶。在评估中，我们遵循 [2]采用说话人无关的5折交叉验证协议。所报告的性能为五折上的平均准确率。

3.2. 实现细节

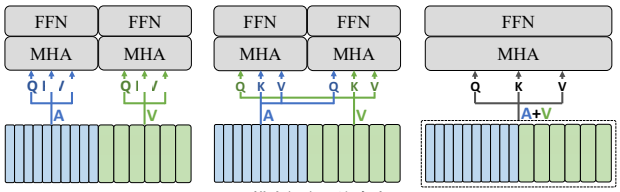
我们使用 AdamW优化器，以 4 的批大小训练所有模型 50 个训练轮数。学习率初始设为 5×10^{-5}，并在训练过程中线性衰减至零。我们的 Transformer编码器配置为$d_{\text{model}} = 512$，投影层则将特征映射到 $d_{\text{emb}} = 128$ 维嵌入空间。对于跨时间匹配损失，我们将高斯带宽设为 $\sigma = 0.5$，温度设为 $\tau = 0.07$，加权系数设为 $\lambda_{\text{ctm}} = 0.5$。

3.3. 与最先进方法的比较

表 1 比较了 CREMA-D 和 RAVDESS 数据集上的视听情感识别表现。我们的方法在 CREMA-D 和 RAVDESS 上分别达到 89.49% 和 89.25%，在两个数据集上均取得新的最先进水平。在 CREMA-D 上，我们比此前表现最佳的方法 [3]（85.06%）高出 4.43 个百分点；在 RAVDESS 上，我们比 ATTSF-Net [12]（88.67%）提升了 0.58 个百分点。

3.4. 消融研究

为了验证统一的Transformer编码器能够作为有效的融合模块，我们实验了图2所示的若干设计变体。先前方法要么依赖模态内自注意力（ISA），要么依赖模态间交叉注意力（ICA），而我们的设计引入了多模态



(a) 模态内自注意力 (b) 模态间交叉注意力 (c) 多模态自注意力

图 2。使用注意力机制的不同融合策略示意图。

表2。在 CREMA-D 上对不同融合策略进行的消融研究。

融合模块	参数量	Acc.
拼接	-	85.71
ISA +ISA	12.61M	87.98
ICA + ICA		87.49
ISA + ICA		87.71
ICA + ISA		88.31
MSA + MSA（我们的）	6.83M	**88.95**

自注意力（MSA），它在单个Transformer 块内将二者统一起来。为了与基于堆叠的模型（如ISA +ICA或 ICA + ISA）进行公平比较，每个变体均配置两个块，以确保模型容量相当。

表2在基于 TaRoPE 的位置编码这一通用设置下比较了不同的融合策略。虽然拼接提供了一种轻量级基线，但堆叠的 ISA 和 ICA 变体通过建模更丰富的依赖关系提升了性能，同时也带来了更高的复杂度。相比之下，我们的模型采用两个 MSA 层，以更少的参数取得了最佳准确率，展现了在统一模块中联合建模模态内和模态间依赖关系的效率。此外，ISA、ICA 和 MSA 层的计算开销相近，对于 2 秒输入，每种结构均额外引入约 0.5 十亿次浮点运算。

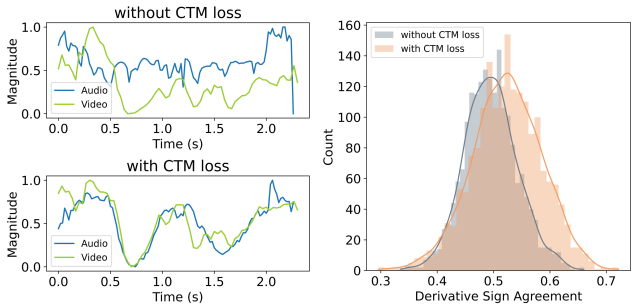
表3全面比较了位置编码策略以及CTM损失的贡献。所提出的TaRoPE优于正弦、可学习和原始 RoPE 嵌入，凸显了显式时间对齐在多模态融合中的有效性。此外，引入CTM损失在所有变体上均带来一致提升，证实其作为互补对齐目标的作用。总体而言，结果凸显了时间对齐作为实现有效多模态整合的关键设计原则的重要性。

3.5. 时间对齐分析

我们进一步分析所提出的 CTM损失如何影响模态之间的时间一致性。图3展示了

表3。在CREMA-D上进行的位置编码策略与CTM损失消融研究。

位置编码	无 $\mathcal{L}_{\text{ctm}}$	有 $\mathcal{L}_{\text{ctm}}$
正弦	88.09	88.79
可学习	87.44	88.79
RoPE	87.76	89.00
TaRoPE	88.95	**89.49**



(a) 逐帧特征幅值 (b) 导数符号一致性分布

图 3。CTM损失对特征动态的影响。

采用和未采用 CTM损失时的特征动态与一致性统计数据。

图 3（a）展示了 CREMA-D 测试集中的一个示例（语句 1087 MTIANG XX），其中音频和视频特征经过最小-最大归一化，并在同一时间轴上进行可视化。应用 CTM损失后，音频和视频特征幅值随帧变化呈现出相似的轨迹，这表明辅助损失促使不同模态间的时间动态保持一致。

为确保这一观察结果并非仅限于单个案例，图 3（b）报告了整个测试集上的导数符号一致性分布。我们发现，使用 CTM损失时，分布更集中于较高的一致性水平，这表明在大规模数据上，音频和视频流之间的时间趋势能够更加一致地同步。这些结果证实，CTM损失不仅提高了识别准确率，还通过强化跨模态动态的一致性，改善了时间对齐。

4. 结论

本文提出了一种用于视听情感识别的 Transformer 框架，该框架采用时间对齐 RoPE 和跨时间匹配损失。我们的框架在 CREMA-D 和 RAVDESS 上取得了最先进水平的表现，而消融研究则凸显了统一架构的效率及时间对齐的优势。尽管本研究聚焦于受控基准测试，但将所提出的框架扩展到大规模自然场景数据集仍是未来工作的重要方向。

5. 参考文献

[1] Hailun Lian、Cheng Lu、Sunan Li、Yan Zhao、Chuan- gao Tang 和 Yuan Zong，《深度学习多模态情感识别综述：语音、文本与人脸》，*Entropy*，第 25 卷，第 10 期，第 1440 页，2023 年。

[2]Cristina Luna-Jim´enez、David Griol、Zoraida Calle- jas、Ricardo Kleinlein、Juan M Montero和Fernando Fern´andez-Mart´ınez，“基于迁移学习的RAVDESS数据集多模态情感识别”，*Sensors*，第21卷，第22期，第7665页，2021年。

[3] Yuanyuan Lei和Houwei Cao，“基于意图标签与多模态感知标签的偏好学习的视听情感识别”，*IEEETransactionsonAffectiveComputing*，第14卷，第4期，第2954–2969页，2023年。

[4] AnamariaRadoi、AndreeaBirhala、Nicolae-Catalin Ris- tea和Liviu-Cristian Dutu，“一种基于多模态信息时序聚合的端到端情感识别框架”，*IEEE Access*，第9卷，第135559–135570页，2021年。

[5] Cristina Luna-Jim´enez、Ricardo Kleinlein、David Griol、Zoraida Callejas、Juan M Montero 和 Fernando Fern´andez-Mart´ınez，“基于听觉Transformer和动作单元的RAVDESS数据集多模态情感识别方案”，*Applied Sciences*，第 12 卷，第 1 期，第 327 页，2021 年。

[6]Asif Iqbal Middya、Baibhav Nag 和 Sarbani Roy，“基于深度学习的多模态情感识别：音频–视觉模态的模型级融合”，*Knowledge-based Systems*，第 244 卷，第 108580 页，2022 年。

[7] Ruijia Fan、Hong Liu、Yidi Li、Peini Guo、Guoquan Wang 和 Ti Wang，“AttA-Net：用于视听情感识别的注意力聚合网络”，载于 *IEEEInternational Conferenceon Acoustics,Speech and Sig-nal Processing(ICASSP)*。IEEE，2024 年，第 8030–8034 页。

[8] Bogdan Mocanu、Ruxandra Tapu 和 Titus Zaharia，“利用注意力和深度度量学习进行跨模态音视频融合的多模态情感识别”，*Image and Vision Computing*，第 133 卷，页码 104676，2023 年。

[9] Licai Sun、Zheng Lian、Bin Liu 和 Jianhua Tao，“Hic-mae：用于自监督视听情感识别的层次对比掩码自编码器”，信息融合，第 108 卷，页码 102382，2024 年。

[10] Samir Sadok、Simon Leglaive 和 Renaud S´eguier，“用于视听语音情感识别的矢量量化掩码自编码器”，*Computer Vision and Im-ageUnderstanding*，第 257 卷，第 104362 页，2025 年。

[11] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、LukaszKaiser 和 Illia Polosukhin，“Attention is all you need”，神经信息处理系统进展（*NeurIPS*），第 30 卷，2017 年。[12] Jiaming Zhang、Zhijia Zhang 和 Zhaojie Ju，“Attsf-net：用于视听情感识别的基于注意力的相似性融合网络”，*IEEETransactionson AffectiveComputing*，2025 年。

[13] Jianlin Su、Murtadha Ahmed、Yu Lu、Shengfeng Pan、Wen Bo 和 Yunfeng Liu，“Roformer：采用旋转位置嵌入的增强型 Transformer”，神经计算，第 568 卷，页码 127063，2024 年。

[14]Houwei Cao、David G Cooper、Michael K Keutmann、Ruben C Gur、Ani Nenkova 和 Ragini Verma，“Crema-d：众包情感多模态演员数据集”，*IEEETransactionson AffectiveComputing*，第 5 卷，第 4 期，页码 377–390，2014 年。

[15] Steven R Livingstone和Frank A Russo，“瑞尔森情感语音与歌曲视听数据库（RAVDESS）：一套动态的、多模态的北美英语人脸与声音表达数据集，”*PloS one*，第13卷，第5期，页码e0196391，2018。

[16] Alexis Conneau、Alexei Baevski、Ronan Collobert、Abdelrahman Mohamed和Michael Auli，“面向语音识别的无监督跨语言表示学习，”*arXiv*预印本 *arXiv:2006.13979*，2020。

[17] E Friesen和Paul Ekman，“面部动作编码系统：一种测量面部运动的技术，”帕洛阿尔托，第3卷，第2期，第5页，1978。

[18] Tadas Baltrusaitis、Amir Zadeh、Yao Chong Lim和Louis-Philippe Morency，“OpenFace 2.0：面部行为分析工具包”，载于*2018*年第*13*届*IEEE*自动人脸与手势识别国际会议（*FG 2018*），2018，第59–66页。

[19] Ziyang Ma、Mingjie Chen、Hezhao Zhang、Zhisheng Zheng、Wenxi Chen、Xiquan Li、Jiaxin Ye、Xie Chen 和 Thomas Hain，“Emobox：多语言多语料库语音情感识别工具包与基准测试”，载于*INTERSPEECH* 会议论文集，2024。

[20] Esam Ghaleb、Mirela Popa 和 Stylianos Asteriadis，“用于情感识别的视听线索的多模态与时间感知”，载于*2019*年第*8*届国际情感计算与智能交互会议（*ACII*）。IEEE，2019，第552–558页。

[21] Esam Ghaleb、Jan Niehues 和 Stylianos Asteriadis，“用于时间情感识别的多模态注意力机制”，载于*2020*年 *IEEE* 国际图像处理会议（*ICIP*）。IEEE，2020，第251–255页。