arXiv:2603.11468v1 [cs.MM] 12 Mar 2026

面向连续效价-唤醒度估计的阶段自适应可靠性建模

YubeenLee* Sangeun Lee*成均馆大学 电子与电信研究院
yubeenlee@skku.edu sange1104@etri.re.kr

Junyeop Cha 成均馆大学 jycha95@g.skku.edu

Eunil Park†成均馆大学 eunilpark@skku.edu

摘要

真实环境中的连续效价-唤醒度估计具有挑战性，原因在于模态可靠性不一致，以及视听信号存在依赖于交互的变异性。现有方法主要侧重于对时间动态进行建模，往往忽略了模态可靠性会在不同交互阶段发生显著变化这一事实。为解决这一问题，我们提出*SAGE*，一种阶段自适应可靠性建模框架，在多模态融合过程中显式估计并校准模态级置信度。*SAGE*引入一种可靠性感知融合机制，根据音频与视觉表征在不同阶段的信息量动态重新平衡二者，防止不可靠信号主导预测过程。通过将可靠性估计与特征表征分离，所提出的框架能够在跨模态噪声、遮挡以及变化的交互条件下实现更稳定的情绪估计。在*Aff-Wild2*基准上的大量实验表明，与现有多模态融合方法相比，*SAGE*持续提升一致性相关系数得分，凸显了可靠性驱动建模在连续情感预测中的有效性。

1. 引言

情绪识别已成为机器智能领域日益重要的研究课题，因为它能够推动开发出可理解人类情感状态的智能系统 [13, 15, 18, 24, 31]。尽管许多传统方法将情绪识别表述为

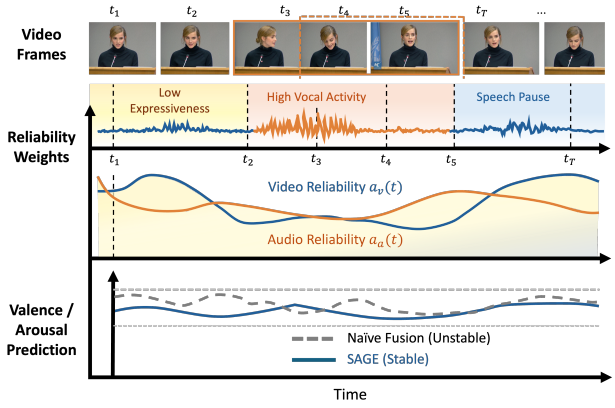


图 1. 由于富有表现力的面部线索和语音活动的变化，各模态的时间可靠性会有所不同。SAGE 会随时间自适应地重新加权模态贡献，从而实现稳定且准确的效价与唤醒度预测。

分类式分类问题 [41, 42], , 另一种范式则使用效价和唤醒度（VA）在连续空间中表示情绪 [3]。在这种表述中，效价描述愉悦程度，而唤醒度反映激活水平。在连续 VA 空间中对情绪建模，使系统能够捕捉细微且模糊的情感变化，而这些变化难以用离散类别表示。

野外情感行为分析（ABAW）挑战赛已成为推动自然环境中情感识别发展的重要基准。本研究参与第十届ABAW竞赛，该竞赛建立在此前一系列挑战赛的基础上 [4, 5, 8, 10–12, 14, 16, 19, 20, 22, 23, 44]。具体而言，效价和唤醒度估计任务使用了扩展版 Aff-Wild2 数据集 [6, 8, 10, 14, 17, 21], , 目标是预测

*贡献相同。
†通讯作者。

基于多模态音频–视觉线索的逐帧效价和唤醒度分数。

为应对多模态情绪识别的复杂性，近期研究[36, 39, 40]探索了各种融合策略，以结合多个模态的信息。基于交叉注意力和门控的融合模型通过实现模态间的细粒度交互和自适应特征选择，展现出良好的性能。然而，这些方法通常侧重于建模特征交互，而非显式估计各模态的可靠性。在真实环境中，音频和视觉信号的质量可能会随时间显著变化，原因包括噪声、遮挡或模态失衡。如图1所示，不同模态的可靠性会随时间波动，具体取决于富有表现力的面部线索或间歇性语音活动等因素。如果不考虑这些可靠性变化，多模态融合可能会纳入不可靠的信号，导致预测值不稳定。

为应对这一挑战，我们提出了 SAGE，这是一种用于连续 VA 估计的阶段自适应多模态框架。该方法引入了一种可靠性引导的加权策略，用于估计跨模态置信度，并在融合过程中动态调整模态贡献。通过显式建模模态可靠性，SAGE 能够在噪声或失衡条件下稳健地整合多模态信号。

在 ABAW 基准上的大量实验表明，与强基线方法相比，SAGE 始终能够提升一致性相关系数（CCC）性能。我们的工作主要贡献总结如下：

- 我们提出了 SAGE，一种用于连续效价-唤醒度估计的阶段自适应可靠性建模框架。
- 我们设计了一种可靠性引导的加权策略，用于量化跨模态置信度，从而在噪声和模态不平衡的情况下实现稳健融合。
- 在 ABAW 基准上的大量实验表明，与强基线方法相比，SAGE 始终能够提升一致性相关系数（CCC）性能。

2. 相关工作

2.1. 效价-唤醒度估计

视听估计以 Russell [37], 提出的情感环形模型为基础，该模型在由效价（即愉悦—不愉悦）和唤醒度（即激活程度）定义的二维空间中表示情绪。这一表述为将情感建模为连续坐标而非离散类别奠定了理论基础。

早期关于视听估计的研究依赖于从多模态信号中提取的手工设计特征。Nicolaou 等人[29]利用面部表情、肩部手势和音频线索，并使用双向长短期记忆网络对时间依赖关系进行建模，以实现连续预测。Savran 等人 [38]将视觉纹理描述符与韵律特征、音频频谱特征相结合，通过粒子滤波整合多模态情感指标。

随着深度学习的兴起，特征提取从手工设计转向深度表征学习。Kollias 和 Zafeiriou [9] 采用多CNN架构提取低层、中层和高层视觉特征，随后使用基于RNN的时序模块对这些特征进行建模，以实现视听估计。Liu 等人 [26] 提出了一种集成框架，将六个预训练视觉编码器与多个低层和深层音频提取器相结合，证明了聚合异构深度表征对于实现稳健视听估计的有效性。

基于注意力的策略越来越多地用于增强连续视听估计中的多模态整合和序列建模。MAVEN [1]通过双向跨模态注意力与模态专属编码器整合视觉、音频和文本模态。Park 等人 [30]和 Meng 等人 [27] 将基于长短期记忆网络的时序建模与Transformer 编码器相结合，利用姿态特征或视听表征捕捉序列情感动态。

2.2. 多模态融合

现有多模态情绪识别研究大致可分为模态重加权方法和基于交叉注意力的交互方法。

早期研究主要侧重于加权整合或后期融合策略。Toisoul 等人 [40]提出了一种三级融合框架，分别对音频和视觉特征进行回归，并迭代地组合其输出。Sharafi 等人 [39] 采用直接的特征级融合策略，将视觉空间特征、时间表征以及音频梅尔频率倒谱系数和能量特征拼接成统一表征，用于分类。

近期研究更加注重通过交叉注意力机制实现细粒度的模态间交互。FedSER-XAI [2] 利用视觉Transformer提取上下文表征，并基于图卷积网络提取结构表征，再通过交叉注意力进行融合，以捕捉相互依赖关系。Mocanu 等人 [28]首先利用注意力模块优化模态内表征，随后应用视听交叉注意力来建模模态间关系。Praveen 和 Alam [33]提出了递归联合跨模态注意力，该方法根据联合模态表征与单模态表征之间的相关性计算注意力权重，并通过递归逐步优化多模态特征。

在递归交叉注意力的基础上，后续研究引入门控机制，进一步调节

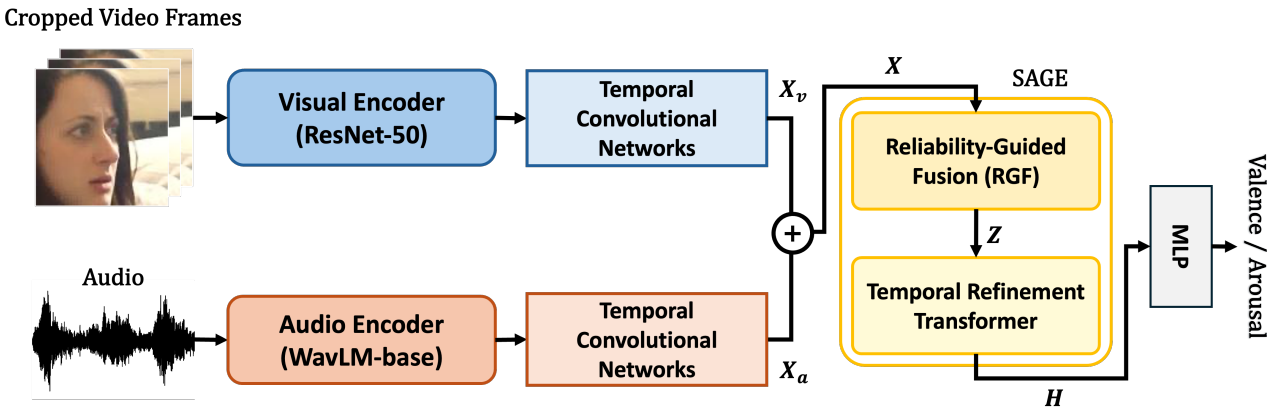


图2. 所提出的用于连续效价-唤醒度估计的 SAGE框架整体架构。利用预训练编码器提取视觉和音频特征，并通过时间卷积网络进行时序编码。融合后的表示由 SAGE模块处理，该模块执行可靠性引导融合和时间细化，随后通过回归头进行帧级效价-唤醒度预测。

信息流。门控递归联合交叉注意力 [36]通过引入自适应门控来扩展联合注意力框架，根据模态互补性的强弱调节交叉注意力特征的贡献。与此相应，时间感知门控融合 [25] 整合了时间门控模块，以突出情绪显著片段，同时抑制瞬时噪声。然而，对时变模态可靠性的显式建模仍然有限。SAGE 通过将可靠性感知自适应重加权引入多模态融合，旨在弥补这一不足。

3. 方法

3.1. 概述

图 2 展示了所提出的用于连续效价-唤醒度估计的 SAGE 模型整体架构。我们设计了一个由四个阶段组成的可靠性感知自适应融合框架。

(i) 多模态特征提取。 给定一个短视频片段，将预处理后的裁剪视频帧及其对应的原始音频波形作为输入。视觉流采用在 ImageNet 上预训练的 ResNet-50 提取帧级视觉表示。对于音频流，我们采用预训练的 WavLM-base 模型，直接从原始波形中获取自监督声学嵌入。因此，每种模态的时序表征可表示为：

$$X_v = f_v(V) \in \mathbb{R}^{T\times D_v}, \quad X_a = f_a(A) \in \mathbb{R}^{T\times D_a}, \quad (1)$$

其中，T 表示时间步的数量，D_v，D_a分别表示视觉和音频嵌入的特征维度。

（ii）时间编码。为捕捉短期时间依赖关系，我们采用时间卷积网络（TCN）：

$$\tilde{X}_v = \mathrm{TCN}_v(X_v), \quad \tilde{X}_a = \mathrm{TCN}_a(X_a). \quad (2)$$

将时间编码特征拼接起来：

$$X = [\tilde{X}_v; \tilde{X}_a] \in \mathbb{R}^{T\times D}, \quad (3)$$

其中 $D = D_v + D_a$。

(iii) 阶段自适应可靠性建模。 随后将融合表示 X 传入SAGE 模块：

$$Z = \mathrm{RGF}(X), \quad H = \mathrm{Transformer}(Z). \quad (4)$$

SAGE 首先执行可靠性引导融合（RGF），在每个时间步动态调整各模态的贡献。随后，通过时间细化 Transformer 对可靠性调整后的表示进行细化，以增强模态不平衡和噪声条件下的跨模态交互。

(iv) 回归头。 最后，将细化后的表示映射为连续的 VA 预测值：

$$\hat{y}_t = \Phi(H_t), \quad \hat{y}_t \in \mathbb{R}^2, \quad (5)$$

其中，$\Phi(\cdot)$ 表示逐帧多层感知机（MLP）。回归头针对每个时间步独立生成预测值。

3.2. 阶段自适应可靠性建模

SAGE 会在模态不平衡的情况下动态调整时序特征贡献。它由两个部分组成：RGF 和时间细化变换器，如图3所示。

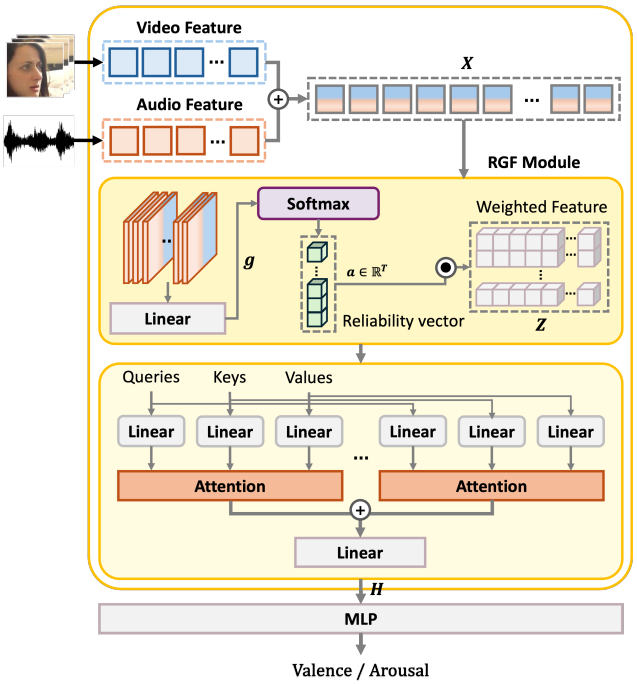


图3. 所提出的 SAGE 模块的详细架构。RGF 计算随时间变化的可靠性分数，以自适应地重新加权时间特征。随后，基于自注意力的时间变换器对可靠性调整后的表示进行细化，以捕获长程依赖关系，然后进行最终回归。

3.2.1. 可靠性引导融合

给定时间编码的多模态表示$X \in \mathbb{R}^{T\times D}$，我们估计随时间变化的可靠性分数，以动态调整模态贡献。对于每个时间步 t，计算一个标量可靠性逻辑值：

$$g_t = W_r X_t + b_r, \tag{6}$$

其中，$X_t \in \mathbb{R}^D$ 表示时间步 t处的多模态特征，$W_r \in \mathbb{R}^{1\times D}$ 和 $b_r \in \mathbb{R}$为可学习参数。将所有 logits 堆叠后得到：

$$\mathbf{g} = XW_r^\top + b_r, \tag{7}$$

其中 $\mathbf{g} \in \mathbb{R}^T$。为获得沿时间维度归一化的可靠性分布：

$$\alpha_t = \frac{\exp(g_t)}{\sum_{k=1}^{T}\exp(g_k)}, \tag{8}$$

从而得到满足以下条件的可靠性向量 $\boldsymbol{\alpha} \in \mathbb{R}^T$：

$$\sum_{t=1}^{T}\alpha_t = 1. \tag{9}$$

可靠性调整后的表示计算如下：

$$Z_t = \alpha_t X_t, \tag{10}$$

或等价地用矩阵形式表示：

$$Z = \mathrm{diag}(\boldsymbol{\alpha})X, \tag{11}$$

其中 $Z \in \mathbb{R}^{T\times D}$。

3.2.2. 时间细化变换器

可靠性加权表示 Z 通过基于自注意力的变换器进一步细化，以捕捉长程时间依赖关系。多头自注意力定义如下：

$$\mathrm{Attention}(Q,K,V) = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V, \tag{12}$$

其中

$$Q = ZW_Q, \quad K = ZW_K, \quad V = ZW_V. \tag{13}$$

细化后的表示计算如下：

$$H = \mathrm{Transformer}(Z). \tag{14}$$

通过处理经过可靠性调整的特征，该变换器增强了模态不平衡情况下的跨模态交互。

4. 实验

4.1. 数据集描述

4.1.1. Aff-Wild2 数据集

我们在第十届ABAW竞赛提供的 Aff-Wild2 [7] 数据集上进行了实验。Aff-Wild2 是一个用于多模态情感行为分析的大规模真实场景数据集。该数据集包含 594 个从 YouTube 收集的视频，共计约 200 万帧标注数据，采集条件多样，涵盖姿势、光照、年龄和种族等方面的变化。

针对 VA 估计挑战赛，训练集和验证集提供了帧级标注。每个标注文件对应一段视频，包含每帧在 $[-1,1]$ 范围内的连续效价和唤醒度值。标注值为 -5 的帧被视为无效，并在训练和评估期间被排除。所有裁剪对齐的人脸图像的分辨率为 $112\times 112\times 3$。我们遵循比赛提供的官方主体独立划分。此外，为评估鲁棒性和泛化性能，我们在官方训练集内进行 K 折交叉验证。最终结果按照挑战赛协议在验证集和测试集上报告。

4.1.2. 数据预处理

在本研究中，我们基于 Aff-Wild2 数据集构建具有时间结构的多模态序列。每段视频被分割为由 300 个连续帧组成的固定长度片段，步长为 200，以生成部分重叠的片段。标注无效或人脸检测失败的帧在训练期间被排除，以确保标签可靠性和一致性。

对于视觉模态，我们使用挑战赛组织方提供的裁剪并对齐的人脸图像，输入分辨率为 48×48。训练期间，采用随机水平翻转和随机裁剪，以提升泛化性能。相比之下，验证期间仅采用中心裁剪，以保持实验一致性。对于音频模态，我们从原始视频中提取原始波形信号，并将其重采样至 16 kHz。为与视觉流保持时间对齐，我们根据视频帧率对音频序列进行时间同步。

4.2. 评估指标

为定量评估连续情绪估计的性能，我们采用一致性相关系数（CCC）作为主要评估指标。CCC 通过综合考量相关性和分布相似性，衡量预测值与真实标注之间的一致性。其定义如下：

$$\mathrm{CCC} = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2}, \tag{15}$$

其中，ρ 表示预测值 x 与真实值 y之间的皮尔逊相关系数，μ_x 和 μ_y分别表示它们的均值，σ_x^2 和 σ_y^2 表示它们的方差。

不同于均方误差（MSE）在每个时间步独立评估逐点差异，一致性相关系数（CCC）会显式考虑整个时间序列中的相关性一致性以及均值-方差对齐。这一特性使CCC特别适用于连续情感估计任务，因为在此类任务中，保持全局时间动态和结构一致性与最小化瞬时预测误差同等重要。

为保持优化目标与评估指标之间的一致性，在训练过程中采用基于一致性相关系数的损失函数。具体而言，训练损失定义为：

$$\mathcal{L}_{CCC} = 1 - \mathrm{CCC}. \tag{16}$$

通过直接最大化一致性，模型不仅能够提升预测结果的数值准确性，还能促使预测情绪轨迹与真实情绪轨迹在结构动态上保持一致。

表1。按照官方挑战赛划分，在Aff-Wild2验证集（fold-0）上对CCC性能进行比较。

模型	验证集		
	效价一致性相关系数	唤醒度 CCC	CCC 平均值
MM-CV-LC [45]	0.469	0.649	0.559
网易伏羲虚拟人 [46]	0.464	0.640	0.552
CtyunAI [48]	0.550	**0.681**	0.616
HFUT-MAC [47]	0.554	0.659	0.607
Situ-RUCAIM3 [27]	0.588	0.668	**0.628**
JCA [34]	**0.663**	0.584	0.623
RJCA [35]	0.443	0.639	0.541
DCA [32]	0.451	0.647	0.549
GRJCA [36]	0.459	0.652	0.556
HGRJCA [36]	0.464	0.660	0.562
TAGF [25]	0.427	0.676	0.552
Ours	0.509	0.674	0.591

4.3. 实验结果

4.3.1. 验证性能

首先，使用官方 Aff-Wild2 验证集划分评估所提出模型的验证性能。遵循标准的 ABAW 评估协议，采用一致性相关系数作为 VA 预测的主要评估指标。为评估所提出方法的有效性，我们将其验证性能与 ABAW 挑战赛相关文献中报告的若干近期最先进方法进行比较。如表 1所总结，所提出模型在效价和唤醒度上的一致性相关系数分别达到 0.509 和 0.674，在验证集上取得了 0.591 的平均一致性相关系数。尽管一些以往方法借助复杂架构或额外资源获得了略高的分数，但所提出模型凭借相对精简的框架展现出可靠的性能。

4.3.2. 测试性能

为评估模型在 Aff-Wild2 测试集上的性能，将验证性能最佳的模型提交至官方 ABAW 挑战赛评估服务器。提交过程中采用了按目标的折选择策略，即为每个目标选择验证性能最佳的折。

表 2 将所提方法与 ABAW 挑战赛文献中报告的若干近期多模态方法进行比较。根据挑战赛评估协议，测试集上效价和唤醒度的单项一致性相关系数得分并未公开披露；因此仅报告平均一致性相关系数得分。所提方法在测试集上取得了 0.58 的平均一致性相关系数。

与 JCA [34], RJCA [35], 和 DCA [32], 等早期多模态方法相比，所提出的模型在平均一致性相关系数得分方面展现出具有竞争力的性能。具体而言，我们的方法优于 MM-CV-LC [45]和 HFUT-MAC [47],等方法，同时取得了与 GRJCA [36] 和 HGRJCA [36]相当的结果。尽管性能最佳的方法

表 2. 通过官方挑战赛服务器评估的 Aff-Wild2 测试集 CCC 性能比较。我们的提交基于验证性能，采用按目标选择折的策略。

模型	测试集		
	效价一致性相关系数	唤醒度 CCC	CCC 平均值
MM-CV-LC [45]	0.46	0.49	0.47
网易伏羲虚拟人 [46]	**0.64**	0.62	**0.63**
CtyunAI [48]	0.50	**0.63**	0.56
HFUT-MAC [47]	0.52	0.54	0.53
Situ-RUCAIM3 [27]	0.61	0.59	0.60
JCA [34]	0.37	0.36	0.36
RJCA [35]	0.53	0.57	0.55
DCA [32]	0.54	0.58	0.56
GRJCA [36]	0.56	0.60	0.58
HGRJCA [36]	0.56	0.62	0.59
USTC-IAT-United [43]	0.62	0.57	0.60
TAGF [25]	0.51	0.57	0.54
Ours	-	-	0.58

取得了略高的得分，但所提出的模型无需依赖额外的外部数据集或集成策略，依然展现出具有竞争力的性能。

5. 结论

本文从模态可靠性的角度研究了多模态效价-唤醒度估计，而非仅关注架构复杂度。我们提出了 SAGE，即一种阶段自适应可靠性建模框架，可在多模态融合过程中显式估计并调节模态级置信度。在第十届 ABAW竞赛的设置下开展的大量实验表明，真实世界情绪识别的性能局限往往源于模态贡献不稳定，而非时序建模能力不足。SAGE 通过在不同交互阶段动态校准跨模态影响，在噪声大、不均衡且不受约束的条件下生成更加稳定的情感轨迹。这些结果表明，可靠性感知建模是实现稳健多模态情绪估计的一项基本设计原则。该框架在第十届ABAW竞赛的官方评测中取得了有竞争力的表现，进一步验证了其在大规模真实场景基准测试中的实际效果。

参考文献

[1]Vrushank Ahire、Kunal Shah、Mudasir Khan、Nikhil Pakhale、Lownish Sookha、Mudasir Ganaie 和 Abhinav Dhall。MAVEN：面向效价-唤醒度情绪网络的多模态注意力。载于计算机视觉与模式识别会议论文集，第 5789–5799 页，2025 年。2[2] Eman Abdulrahman Alkhamali、Arwa Abdulaziz Allinjawi、Rehab Bahaaddin Ashari 和 Mohammed Tawfik。Fedser-xai：用于可解释联邦语音情绪识别的 PSO优化的多流交叉注意力变换器与图特征。《科学报告》，2025 年。2[3] Lisa A Feldman。效价关注与唤醒度关注：个体情感体验结构的差异。Journal of Personality and Social Psychology, 69(1):153, 1995. 1[4] Dimitrios Kollias. ABAW：效价-唤醒度估计、表情识别、动作单元检测与多任务学习挑战。见IEEE/CVF计算机视觉与模式识别会议论文集，第2328–2336页，2022年。1[5] Dimitrios Kollias. ABAW：从合成数据学习与多任务学习挑战。见欧洲计算机视觉会议，第157–172页。Springer，2023年。1[6] Dimitrios Kollias. 多标签复合表情识别：C-expr数据库与网络。见IEEE/CVF计算机视觉与模式识别会议论文集，第5589–5598页，2023年。1[7] Dimitrios Kollias与Stefanos Zafeiriou. Aff-Wild2：扩展Aff-Wild数据库以用于情感识别。arXiv预印本 arXiv:1811.07770，2018. 4[8] Dimitrios Kollias与Stefanos Zafeiriou. 表情、情感与动作单元识别：Aff-Wild2、多任务学习与ArcFace。arXiv预印本 arXiv:1910.04855，2019. 1[9] Dimitrios Kollias与Stefanos Zafeiriou. 在OMG真实场景数据集上利用多CNN特征进行基于CNN-RNN的维度情绪识别。IEEE Transactions on Affective Computing, 12(3):595–606, 2020. 2[10] Dimitrios Kollias与Stefanos Zafeiriou. 真实场景情感分析：效价-唤醒度、表情、动作单元与统一框架。arXiv预印本 arXiv:2103.15792，2021. 1[11] Dimitrios Kollias与Stefanos Zafeiriou. 第二届ABAW2竞赛中的情感行为分析。见IEEE/CVF国际计算机视觉会议论文集，第3652–3660页，2021.[12] D Kollias、A Schulc、E Hajiyev与S Zafeiriou. 第一届ABAW 2020竞赛中的情感行为分析。见2020年第15届IEEE自动人脸与手势识别国际会议（FG 2020）（FG），第794–800页。1[13]Dimitrios Kollias、Viktoriia Sharmanska与Stefanos Zafeiriou. 按需面部行为：单一网络中的表情、情感与动作单元。arXiv预印本 arXiv:1910.11111，2019. 1[14] Dimitrios Kollias、Panagiotis Tzirakis、Mihalis A Nicolaou、Athanasios Papaioannou、Guoying Zhao、Bj¨orn Schuller、Irene Kotsia与Stefanos Zafeiriou. 真实场景深度情感预测：Aff-Wild数据库与挑战、深度架构及更多。《国际计算机视觉杂志》，第1–23页，2019. 1[15]Dimitrios Kollias、Viktoriia Sharmanska与Stefanos Zafeiriou. 面向异构多任务学习的分布匹配：一项大规模人脸研究。arXiv预印本 arXiv:2105.03790，2021. 1[16] Dimitrios Kollias、Panagiotis Tzirakis、Alice Baird、Alan Cowen与Stefanos Zafeiriou. ABAW：效价-唤醒度估计、表情识别、动作单元检测与情绪反应强度估计挑战。见IEEE/CVF计算机视觉与模式识别会议论文集，第5888–5897页，2023年。1

[17] Dimitrios Kollias、Chunchang Shao、Odysseus Kaloidas与Ioannis Patras. Behaviour4all：真实场景面部行为分析工具包。arXiv预印本 arXiv:2409.17717，2024. 1[18]Dimitrios Kollias、Viktoriia Sharmanska与Stefanos Zafeiriou. 面向分类任务多任务学习的分布匹配：关于人脸及其他的大规模研究。见AAAI人工智能会议论文集，第2813–2821页，2024. 1[19] Dimitrios Kollias、Panagiotis Tzirakis、Alan Cowen、Stefanos Zafeiriou、Irene Kotsia、Alice Baird、Chris Gagne、Chunchang Shao与Guanyu Hu. 第六届真实场景情感行为分析（ABAW）竞赛。见IEEE/CVF计算机视觉与模式识别会议论文集，第4587–4598页，2024. 1[20] Dimitrios Kollias、Stefanos Zafeiriou、Irene Kotsia、Abhinav Dhall、Shreya Ghosh、Chunchang Shao与Guanyu Hu. 第七届ABAW竞赛：多任务学习与复合表情识别。见欧洲计算机视觉会议，第31–45页。Springer，2024. 1[21]Dimitrios Kollias、Damith C Senadeera、Jianian Zheng、Kaushal KK Yadav、Greg Slabaugh、Muhammad Awais与Xiaoyun Yang. DVD：用于推进真实世界场景暴力检测的综合数据集。arXiv预印本 arXiv:2506.05372，2025. 1[22] Dimitrios Kollias、Panagiotis Tzirakis、Alan Cowen、Stefanos Zafeiriou、Irene Kotsia、Eric Granger、Marco Pedersoli、Simon Bacon、Alice Baird、Chris Gagne等. 情感与行为分析的进展：第八届ABAW研讨会与竞赛。见计算机视觉与模式识别会议论文集，第5572–5583页，2025. 1[23] Dimitrios Kollias、Stefanos Zafeiriou、Irene Kotsia、Greg Slabaugh、Damith Chamalke Senadeera、Jianian Zheng、Kaushal Kumar Keshlal Yadav、Chunchang Shao与Guanyu Hu. 从情绪到暴力：第九届ABAW的多模态细粒度行为分析。见IEEE/CVF国际计算机视觉会议论文集，第1–12页，2025. 1[24] Kevin LaGrandeur. 情绪、人工智能与伦理。载于《超越人工智能：正在消失的人机分界》，第97–109页。Springer，2015. 1[25] Yubeen Lee、Sangeun Lee、Chaewon Park、Junyeop Cha与Eunil Park. 用于跨模态效价-唤醒度估计的动态时间门控网络。见IEEE/CVF国际计算机视觉会议论文集，第61–70页，2025. 3, 5, 6[26] Xiaolong Liu、Lei Sun、Wenqiang Jiang、Fengyuan Zhang、Yuanyuan Deng、Zhaopei Huang、Liyu Meng、Yuchen Liu与Chuanhe Liu. Evaef：真实场景中的集成效价-唤醒度估计框架。见IEEE/CVF计算机视觉与模式识别会议论文集，第5863–5871页，2023年。2[27] Liyu Meng、Yuchen Liu、Xiaolong Liu、Zhaopei Huang、Wenqiang Jiang、Tenggan Zhang、Chuanhe Liu与Qin Jin. 基于多模态时序感知特征的真实场景视频效价与唤醒度估计。见IEEE/CVF计算机视觉与模式识别会议论文集，第2345–2352页，2022年。2, 5, 6

[28] Bogdan Mocanu、Ruxandra Tapu 和 Titus Zaharia。采用注意力机制和深度度量学习进行跨模态音视频融合的多模态情绪识别。Image and Vision Computing，133:104676，2023。2[29] Mihalis A Nicolaou、Hatice Gunes 和 Maja Pantic。在效价-唤醒空间中，基于多种线索和模态对自发情感进行连续预测。IEEE Transactions on Affective Computing，2(2):92–105，2011。2[30] Ho-min Park、Ilho Yun、Ajit Kumar、Ankit Kumar Singh、Bong Jun Choi、Dhananjay Singh 和 Wesley De Neve。利用姿态特征工程和 Transformer 编码器实现时间连续情绪的多模态预测。在第3届国际多模态情感分析研讨会与挑战赛论文集，第47–54页，2022。2[31] Rosalind W Picard。情感计算。MIT出版社，2000。1[32] R Gnana Praveen 和 Jahangir Alam。交叉注意力并非总是必要：用于音视频维度情绪识别的动态交叉注意力。载于 2024 IEEE International Conference on Multimedia and Expo (ICME)，第1–6页。IEEE，2024。5，6[33] R Gnana Praveen 和 Jahangir Alam。用于维度情绪识别多模态融合的递归联合跨模态注意力。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第4803–4813页，2024。2[34] R Gnana Praveen、Wheidima Carneiro de Melo、Nasib Ullah、Haseeb Aslam、Osama Zeeshan、Th´eo Denorme、Marco Pedersoli、Alessandro L Koerich、Simon Bacon、Patrick Cardinal 等。用于维度情绪识别音视频融合的联合交叉注意力模型。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第2486–2495页，2022。5，6[35] R Gnana Praveen、Eric Granger 和 Patrick Cardinal。用于基于回归的情绪识别音视频融合的递归联合注意力。载于 ICASSP 2023-2023 IEEE 声学、语音与信号处理国际会议（ICASSP），第1–5页。IEEE，2023。5，6[36] Gnana Praveen Rajasekhar、Jahangir Alam 和 Eric Charton。团结则存，分裂则亡：处理效价-唤醒空间中音视频情绪识别的弱互补性。载于计算机视觉与模式识别会议论文集，第5741 5751页，2025 2，3，5，6 – 。[37] James A Russell。情感环形模型。Journal of Personality and Social Psychology，39(6):1161，1980。2[38] Arman Savran、Houwei Cao、Miraj Shah、Ani Nenkova 和 Ragini Verma。通过粒子滤波，在自发对话中结合视频、音频和词汇情感指标。载于第14届 ACM 多模态交互国际会议论文集，第485–492页，2012。2[39] Masoumeh Sharafi、Mohammadreza Yazdchi、Reza Rasti 和 Fahimeh Nasimi。一种新颖的时空卷积神经网络框架，用于多模态情绪识别。Biomedical Signal Processing and Control，78 103970，2022 2： 。

[40] Antoine Toisoul、Jean Kossaifi、Adrian Bulat、Georgios Tz-imiropoulos 和 Maja Pantic。在自然情境下从人脸估计连续效价和唤醒度。Nature Machine Intelligence，3(1):42–50，2021。2[41] Samarth Tripathi、Shrinivas Acharya、Ranti Sharma、Sudhanshi Mittal 和 Samit Bhattacharya。利用深度卷积神经网络准确分类 DEAP 数据集中的情绪。载于 AAAI 人工智能会议论文集，第4746–4752页，2017。1[42] Yue Xie、Ruiyu Liang、Zhenlin Liang、Chengwei Huang、Cairong Zou 和 Bj¨orn Schuller。使用基于注意力的长短期记忆网络进行语音情绪分类。IEEE/ACM 音频、语音与语言处理汇刊，27(11):1675–1685，2019。1[43] Jun Yu、Yongqi Wang、Lei Wang、Yang Zheng 和 Shengfan Xu。用于情绪识别的交互式多模态时序建模框架。载于计算机视觉与模式识别会议论文集，第5699–5706页，2025。6[44] Stefanos Zafeiriou、Dimitrios Kollias、Mihalis A Nicolaou、Athanasios Papaioannou、Guoying Zhao 和 Irene Kotsia。Aff-Wild：真实场景中的效价和唤醒度挑战赛。载于计算机视觉与模式识别研讨会（CVPRW），2017 IEEE 会议，第1980–1987页。IEEE，2017。1[45] Su Zhang、Yi Ding、Ziquan Wei 和 Cuntai Guan。采用音视频领随式注意力融合进行连续情绪识别。载于 IEEE/CVF国际计算机视觉会议论文集，第3567 3574页，2021 5，6 – 。[46] Wei Zhang、Bowen Ma、Feng Qiu 和 Yu Ding。基于掩码自编码器的多模态面部情感分析。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第5793–5802页，2023。5，6[47] Ziyang Zhang、Liuwei An、Zishun Cui、Ao Xu、Tengteng Dong、Yueqi Jiang、Jingyi Shi、Xin Liu、Xiao Sun 和 Meng Wang。ABAW5 挑战赛：采用 Transformer 编码器和视听融合的面部情感识别方法。载于 IEEE/CVF 计算机视觉与模式识别会议论文集，第 5725–5734页，2023。5，6[48] Weiwei Zhou、Jiada Lu、Zhaolong Xiong 和 Weifeng Wang。结合时序卷积网络和 Transformer，实现连续情绪识别中的高效视听融合。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第5756–5763页，2023。5，6