arXiv:2304.07958v1 [cs.CV] 17 Apr 2023

回归式情绪识别中的递归联合注意力音视频融合

R Gnana Praveen、*Eric Granger* 和 *Patrick Cardinal*

图像、视觉与人工智能实验室（LIVIA）´高等技术学院，蒙特利尔，加拿大

摘要

在基于视频的情绪识别（ER）中，在保留各个模态的模态内特征的同时，有效利用音频（A）和视觉（V）模态之间的互补关系至关重要。本文提出了一种递归联合注意力模型，并结合长短期记忆网络（LSTM）模块，用于基于回归的情绪识别中语音和面部表情的融合。具体而言，我们研究了利用联合交叉注意力模型以递归方式结合长短期记忆网络，挖掘音频和视觉模态的互补特性，从而捕获相同模态内的时间依赖关系以及A-V特征表征之间的时间依赖关系的可行性。通过将长短期记忆网络与递归联合交叉注意力相结合，我们的模型能够有效利用模态内和模态间关系，实现音频和视觉模态的融合。在具有挑战性的Affwild2和疲劳（私有）数据集上[1] 开展的大量实验结果表明，所提出的音频-视频融合模型能够显著优于当前最先进的方法。

索引词— 情绪识别、音频-视频融合、注意力机制、长短期记忆网络。

1. 引言

自动情绪识别（情绪识别）是一项颇具挑战的任务，因为不同个体和文化中的表情复杂且极其多样。在大多数现实应用中，除了六种基本类别的表情——愤怒、厌恶、恐惧、快乐、悲伤和惊讶之外，情绪还会呈现出广泛的情绪状态。例如，情绪状态可以表示为疲劳、压力和疼痛在离散等级上的强度。同样，范围广泛的连续情绪状态通常以维度情感识别的形式进行描述，其中多样而复杂的人类情绪沿着效价和唤醒度两个维度表示。效价表示与愉悦程度相关的连续情绪状态范围，从情绪状态从极度悲伤（负向）延伸至极度快乐（正向）。同样，唤醒度涵盖与强度相关的情绪状态范围，从非常被动（困倦）到极其活跃（高度兴奋）。本文重点在于开发一种稳健的模型，用于效价-唤醒度空间中的基于回归的情感识别，以及疲劳识别。

音频和视觉模态之间通常存在互补关系，充分利用这些关系对于构建高效的基于回归的情感识别音频-视频融合系统至关重要。除了音频和视觉模态之间的模态间关系外，视频中的时间动态还包含与面部表情和声音表达随时间演变相关的重要信息。因此，有效利用音频和视觉模态之间的模态间关联，以及音频和视觉模态内部的时间动态（模态内），对于构建稳健的音频-视频识别系统发挥着重要作用。本文探讨了在统一框架中利用音频和视觉模态的这些模态间和模态内特征的可能性。在现有的大多数基于回归的情感识别方法中，使用长短期记忆网络对视频中的模态内时间动态进行建模[1, 2]，这是因为它们能够高效捕捉长期时间动态 [3]。另一方面，交叉注意力模型[4] 已被用于对音频和视觉模态的模态间特征进行建模，以实现维度情感识别。

在本研究中，我们提出了一个用于音频-视频融合的统一框架，分别利用长短期记忆网络和联合交叉注意力，有效利用视频中的模态内信息和模态间信息。为了进一步提升联合交叉注意力模型的A-V特征表征，我们还探索了递归注意力机制。对联合交叉注意力模型进行递归训练，可以细化A和V的特征表示，从而提升系统性能。本文的主要贡献如下。(1) 提出了一种用于音频-视频融合的递归联合交叉注意力模型，在有效利用跨模态互补关系的同时，采用递归机制进一步细化A-V特征表征。(2) 进一步集成长短期记忆网络，以有效捕捉各个独立模态内部以及A-V特征表征中的时间动态。(3) An

本研究得到了加拿大自然科学与工程研究委员会以及魁北克自然与技术研究基金会的资助。[1]代码可在 GitHub 上获取：https://github.com/praveena2j/RecurrentJointAttentionwithLSTMs。

在具有挑战性的Affwild2和Fatigue（私有）数据集上开展了大量实验，结果表明，所提出的视听融合模型在基于回归的情绪识别方面优于相关的先进模型。

2. 相关工作

Tzirakis 等人提出了一种早期的基于深度学习、用于音频-视频融合的维度情感识别方法[5], ：将深度特征（分别使用 Resnet-50 提取视频特征、使用一维卷积神经网络提取音频特征）拼接后输入长短期记忆网络。最近，Vincent 等人 [3] 研究了注意力模型的有效性，并将其与循环网络进行了比较。他们表明，在维度情感识别中，与注意力模型相比，长短期记忆网络能够相当有效地捕捉时间依赖关系。Kuhnke 等人 [2]提出了一种双流视听网络，使用深度模型提取音频和视频特征，再将其拼接以进行维度情感识别。这些方法大多无法有效捕捉音频和视频模态之间的跨模态语义。在 [6] 和 [7], 中，作者聚焦于使用 Transformer 进行跨模态注意力，以利用音频和视频模态在维度情感识别中的模态间关系。Rajasekhar 等人 [4] 探索了交叉注意力模型，利用基于音频和视频特征互相关的模态间特征。他们通过将联合特征表示引入交叉注意力模型来改进方法，从而保留模态内特征 [8, 9]。这些方法大多无法有效利用模态内关系。Chen 等人[10] 使用长短期记忆网络对音频和视频特征建模，并根据基于长短期记忆网络的条件注意力，利用注意力权重融合单模态预测。Priyasad 等人 [11] 也探索了使用长短期记忆网络处理视频特征，并在拼接后的音频和视频模态特征上使用基于深度神经网络的注意力，以生成最终预测。Beard 等人 [12] 提出了一种递归循环注意力模型，通过增加共享记忆状态来扩展长短期记忆网络，从而以递归方式捕捉多模态关系。与这些方法不同，我们着重对音频-视频关系进行建模：在特征拼接之前，使音频和视频特征相互作用，并以递归方式衡量模态间及模态内的语义相关性。我们采用长短期记忆网络对单模态和多模态特征进行时间建模，进一步增强所提出的框架。因此，我们提出的模型能够有效利用模态内关系以及互补的模态间关系，从而取得更高的性能。

3. 所提出的方法

A）问题形式化：给定一个输入视频子序列 S，均匀采样出 L 个互不重叠的视频片段，并分别通过预训练网络从音频和视觉模态中提取深度特征向量 $\boldsymbol{X}_{\mathbf{a}}$ 和 $\boldsymbol{X}_{\mathbf{v}}$ 。设 $\boldsymbol{X}_{\mathbf{a}} = \{\boldsymbol{x}_{\mathbf{a}}^1, \boldsymbol{x}_{\mathbf{a}}^2, ..., \boldsymbol{x}_{\mathbf{a}}^L\} \in \mathbb{R}^{d_a \times L}$ 和 $\boldsymbol{X}_{\mathbf{v}} = \{\boldsymbol{x}_{\mathbf{v}}^1, \boldsymbol{x}_{\mathbf{v}}^2, ..., \boldsymbol{x}_{\mathbf{v}}^L\} \in \mathbb{R}^{d_v \times L}$，其中 d_a 和 d_v 分别表示音频和视觉特征表示的维度，$\boldsymbol{x}_{\mathbf{a}}^l$ 和 $\boldsymbol{x}_{\mathbf{v}}^l$ 分别表示视频片段的音频和视觉特征向量，对应第 $l = 1, 2, ...$，L 个片段。该问题的目标是利用训练数据 $\boldsymbol{X}$ 估计回归模型 F： $\boldsymbol{X} \rightarrow \boldsymbol{Y}$，其中 $\boldsymbol{X}$ 表示输入视频片段的音频和视觉特征向量集合，$\boldsymbol{Y}$ 表示相应视频片段的回归标签。

B）音频和视觉网络：研究发现，频谱图与多种二维卷积神经网络（ResNet-18[13]）结合用于情绪识别 [14, 15]时效果良好。因此，我们在所提出的框架中探索了频谱图。为了有效利用 A 模态中的时间动态，我们还探索了在 A 序列的各个时间片段上使用长短期记忆网络。最后，L 视频片段的 A 特征向量表示为 $\boldsymbol{X_a} = (\boldsymbol{x}_{\mathbf{a}}^1, \boldsymbol{x}_{\mathbf{a}}^2, ..., \boldsymbol{x}_{\mathbf{a}}^L) \in \mathbb{R}^{d_a \times L}$。

面部表情包含大量与视频中的视觉外观和时间动态相关的信息。研究发现，长短期记忆网络能够有效捕获长期时间动态，而三维卷积神经网络则擅长捕获短期时间动态 [16]。因此，我们将长短期记忆网络与 3D CNN（R3D [17]）结合，以提取用于融合模型的 V 特征。在大多数现有方法中，最后一个卷积层的输出为 512 x 7 x 7，随后经过池化操作，将空间维度缩减至 1（7→ 1）。由于步幅较大，这种空间维度缩减会丢失有用信息。因此，受 [18], 的启发，我们使用 A 特征表示，平滑地缩减每个视频片段原始 V 特征的空间维度，这与 [18]类似。最后，我们将视频片段的 V 特征向量组成一个矩阵，表示为$\boldsymbol{X_v} = (\boldsymbol{x}_{\mathbf{v}}^1, \boldsymbol{x}_{\mathbf{v}}^2, ..., \boldsymbol{x}_{\mathbf{v}}^L) \in \mathbb{R}^{d_v \times L}$。

C) 递归联合注意力模型：给定 A 和 V 特征 $\boldsymbol{X_a}$ 和 $\boldsymbol{X_v}$，通过拼接 A 和 V 特征向量 $\boldsymbol{J} = [\boldsymbol{X}_{\mathbf{a}}; \boldsymbol{X}_{\mathbf{v}}] \in \mathbb{R}^{d \times L}$，其中 $d = d_a + d_v$ 表示拼接后特征的维度。现在，使用视频子序列（$\boldsymbol{S}$）的拼接音视频特征表示（$\boldsymbol{J}$）来关注单模态特征表示 $\boldsymbol{X}_{\mathbf{a}}$ 和$\boldsymbol{X}_{\mathbf{v}}$。A 特征$\boldsymbol{X}_{\mathbf{a}}$ 与组合音视频特征 $\boldsymbol{J}$ 之间的联合相关矩阵 $\boldsymbol{C}_{\mathbf{a}}$ 定义如下：

$$\boldsymbol{C}_{\mathbf{a}} = \tanh\left(\frac{\boldsymbol{X}_{\mathbf{a}}^T \boldsymbol{W}_{\mathbf{ja}} \boldsymbol{J}}{\sqrt{d}}\right) \tag{1}$$

其中，$\boldsymbol{W}_{\mathbf{ja}} \in \mathbb{R}^{L \times L}$ 表示 A 模态与组合后的音视频特征之间的可学习权重矩阵，T 表示转置操作。类似地，V 特征的联合相关矩阵表示为：

$$\boldsymbol{C}_{\mathbf{v}} = \tanh\left(\frac{\boldsymbol{X}_{\mathbf{v}}^T \boldsymbol{W}_{\mathbf{jv}} \boldsymbol{J}}{\sqrt{d}}\right) \tag{2}$$

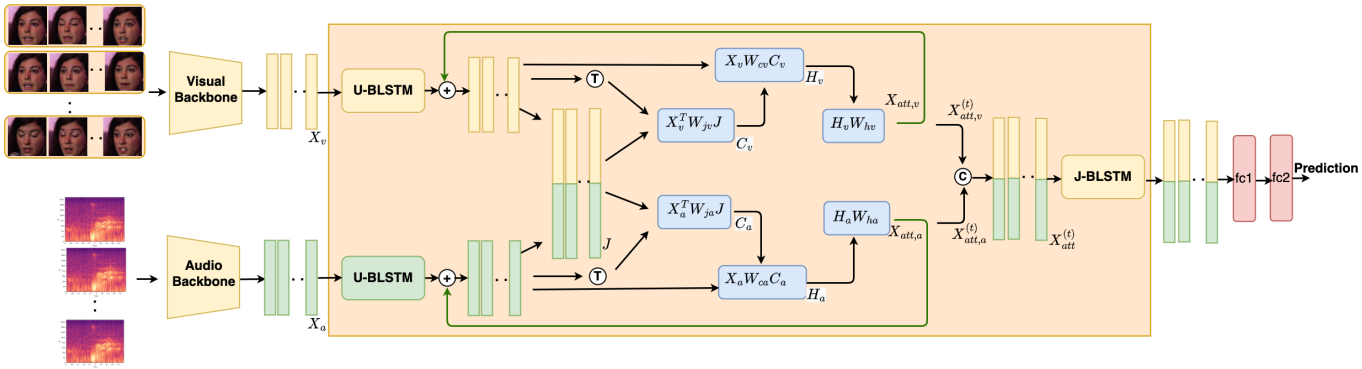

图 1. 所提出的基于双向长短期记忆网络的递归联合注意力模型框图。

联合相关矩阵捕获了连续视频片段之间音频和视觉模态的语义相关性，以及同一模态内部的语义相关性，有助于有效利用模态内和模态间关系。计算联合相关矩阵后，估计音频和视觉模态的注意力权重。对于 A 模态，将联合相关矩阵 $\boldsymbol{C_a}$ 与相应的 A 特征 $\boldsymbol{X_a}$ 结合，并使用可学习权重矩阵 $\boldsymbol{W_{ca}}$ 计算 A 模态的注意力权重，表示为 $\boldsymbol{H_a} = ReLU(\boldsymbol{X_a W_{ca} C_a})$，其中 $\boldsymbol{W_{ca}} \in \mathbb{R}^{d_a \times d_a}$ 和 $\boldsymbol{H_a}$ 表示 A 模态的注意力图。类似地，V 模态的注意力图（$\boldsymbol{H_v}$）表示为 $\boldsymbol{H_v} = ReLU(\boldsymbol{X_v W_{cv} C_v})$，其中 $\boldsymbol{W_{cv}} \in \mathbb{R}^{d_v \times d_v}$。随后，使用注意力图计算音频和视觉模态的注意力加权特征：

$$\boldsymbol{X_{att,a}} = \boldsymbol{H_a W_{ha}} + \boldsymbol{X_a} \tag{3}$$

$$\boldsymbol{X_{att,v}} = \boldsymbol{H_v W_{hv}} + \boldsymbol{X_v} \tag{4}$$

其中，$\boldsymbol{W_{ha}} \in \mathbb{R}^{d \times d_a}$ 和 $\boldsymbol{W_{hv}} \in \mathbb{R}^{d \times d_v}$ 分别表示 A 和 V 的可学习权重矩阵。获得注意力加权特征后，再将其输入联合交叉注意力模型，以计算新的 A 和 V 特征表示：

$$\boldsymbol{X^{(t)}_{att,a}} = \boldsymbol{H^{(t)}_a W^{(t)}_{ha}} + \boldsymbol{X^{(t-1)}_a} \tag{5}$$

$$\boldsymbol{X^{(t)}_{att,v}} = \boldsymbol{H^{(t)}_v W^{(t)}_{hv}} + \boldsymbol{X^{(t-1)}_v} \tag{6}$$

其中，$\boldsymbol{W^{(t)}}_{\mathbf{ha}} \in \mathbb{R}^{d \times d_a}$ 和 $\boldsymbol{W^{(t)}}_{\mathbf{hv}} \in \mathbb{R}^{d \times d_v}$ 分别表示第 t^{th} 次迭代中 A 和 V 的可学习权重矩阵。最后，将经过 t 次迭代后的注意力加权 A 和 V 特征进一步拼接，并输入双向长短期记忆网络，以提取精炼后的 A-V 特征表示中的时间依赖关系；随后将这些特征输入全连接层，得到最终预测。

4. 结果与讨论

A）数据集： Affwild2 是情感计算领域最大的数据集之一，包含从 YouTube 收集的 564 个真实场景视频 [19]。标注由四位专家使用操纵杆完成，最终标注取四位评分者的平均值。数据集总计包含 2,816,832 帧，涉及 455 名受试者，其中男性 277 名，女性 178 名。效价和唤醒度的标注以连续值形式给出，范围为 $[-1, 1]$。数据集以受试者独立方式划分，确保每位受试者的数据仅出现在一个子集中。划分后，训练集、验证集和测试集分别包含 341、71 和 152 个视频。

Fatigue数据集（私有）采集自康复中心的 18 名参与者，他们患有会引发 Fatigue 的退行性疾病。共录制了 27 场视频会话，每场持续 40–45 分钟，并以 0 到 10 的量表按序列级进行标注，每 10 到 15 分钟标注一次。我们将 80% 的数据用作训练数据（50,845 个样本），20% 用作验证数据（21,792 个样本）。

B) 消融研究： 表1展示了消融研究在验证集上进行的实验结果。我们使用一致性相关系数（CCC）评估该方法的性能。本节分析了双向长短期记忆网络的作用，并分别进行了使用和不使用双向长短期记忆网络的实验。首先，我们在单独的音频和视频表征以及A-V特征表征上进行了不使用双向长短期记忆网络的实验。随后，在将音频和视频模态输入联合注意力融合模型之前，仅对各自的模态使用双向长短期记忆网络，即仅使用单模态双向长短期记忆网络（U-BLSTMs）。通过使用U-BLSTMs捕捉各个模态内部的时间依赖关系，我们观察到性能有所提升。因此，与联合注意力模型中基于相关性的模态内建模相比，双向长短期记忆网络更有望捕捉模态内的时间动态。之后，我们还加入了联合双向长短期记忆网络（J-BLSTM），以捕捉联合A-V特征表征中的时间动态，从而进一步提升了系统性能。请注意，在上述实验中，我们没有进行递归注意力。除了U-BLSTM和J-BLSTM的影响之外，

我们还进行了更多实验，以探究联合注意力模型递归行为的影响。首先，我们在不使用长短期记忆网络的情况下进行递归，发现递归带来了一定的性能提升。随后，我们加入长短期记忆网络，并通过改变融合模型中的递归（迭代）次数进行了多项实验。随着递归次数增加，模型性能逐渐提升，并在达到一定次数后开始下降。在测试集上也观察到了类似的性能变化趋势。因此，这可能是因为递归也起到了正则化器的作用，从而提升了模型的泛化能力。在实验中，我们发现 $t = 2$的性能最佳，即模型在进行两次递归迭代时取得了最佳性能。

表1. 我们的方法结合双向长短期记忆网络和递归注意力，在 Affwild2 数据集上的性能。M

ethod	Valence	A_{rousa} l
不含递归的 JA 融合		
不含 U-BLSTM 的融合	0.670	0.590
不含 J-BLSTM 的融合	0.691	0.646
结合 U-BLSTM 和 J-BLSTM 的融合	0.715	0.688
带递归的 JA 融合		
不含双向长短期记忆网络的 JA 融合 , $t = 2$	0.703	0.623
含双向长短期记忆网络的 JA 融合 , $t = 2$	**0.721**	**0.694**
含双向长短期记忆网络的 JA 融合, $t = 3$	0.706	0.652
含双向长短期记忆网络的 JA 融合 , $t = 4$	0.685	0.601

C）与当前最佳方法的比较：表 2 展示了我们的方法与 Affwild2 数据集上相关先进音频-视频融合模型的结果对比。Affwild2 数据集近期已广泛用于野外情感行为分析（ABAW）挑战赛 [20, 21]。因此，我们将所提方法与 ABAW 挑战赛中的相关先进方法进行比较。Kuhnke 等人 [2]使用简单的特征拼接方法，并分别采用 ResNet-18 处理音频、R3D 处理视频模态，其唤醒度表现优于效价。Zhang 等人[22] 提出了一种领导者-跟随者注意力模型用于融合，并提升了 Kuhnke 等人提出的模型 [2]在唤醒度方面的表现。Rajasekhar 等人 [4] 探索了交叉注意力模型，仅利用音频和视频之间的模态间关系，并在效价方面取得了提升，但唤醒度方面没有提升。Rajasekhar 等人 [8]进一步通过在交叉注意力模型中引入联合特征表示，提升了模型性能。所提模型通过引入长短期记忆网络以及递归注意力机制，表现甚至优于原始 JCA [8]。在测试集结果方面，最新一届 ABAW 挑战赛 [23]的获胜者通过使用音频-视频融合取得了性能提升，但同时使用了三个外部数据集和多个骨干网络。我们还将音频和视频骨干网络与长短期记忆网络及 Transformer 集成模型 [23] 在验证集上的表现进行了比较。Vincent [3] 使用长短期记忆网络来

表 2. 所提出方法与最先进方法在 Affwild2 数据集上进行音频-视频融合的 CCC 性能。

方法	融合类型	效价	唤醒度
验证集			
Kuhnke 等人 [2]	特征拼接	0.493	0.613
Zhang 等人 [22]	领导者-跟随者注意力	0.469	0.649
Rajasekhar 等人 [4]	交叉注意力	0.541	0.517
Rajasekhar 等人 [8]	联合交叉注意力	0.657	0.580
Ours	长短期记忆网络 + Transformer	0.628	0.654
Ours	递归联合注意力 + 双向长短期记忆网络	**0.721**	**0.694**
测试集			
Meng 等人 [23]	长短期记忆网络 + Transformer	0.606	0.596
Vincent 等人 [3]	长短期记忆网络 + Transformer	0.418	0.407
Rajasekhar 等人 [8]	联合交叉注意力	0.451	0.389
Ours	递归联合注意力 + 双向长短期记忆网络	**0.467**	**0.405**

表 3. 疲劳数据集上的 CCC 性能。

方法	疲劳等级
仅音频（二维卷积神经网络：ResNet-18）	0.312
仅视觉（三维卷积神经网络：R3D）	0.415
特征拼接	0.378
交叉注意力 [4]	0.421
递归联合注意力 +双向长短期记忆网络（我们的方法）	**0.447**

捕捉模态内依赖关系，并探索了用于跨模态注意力的 Transformer；然而，它们无法有效捕捉连续视频片段之间的模态间关系。Rajasekhar 等人 [8] 进一步通过联合交叉注意力提升了效价方面的性能。所提出的模型优于 [3] 和 [8]。

表 3 展示了所提出方法在疲劳数据集上的表现。我们展示了各单一模态以及特征拼接和交叉注意力 [4]的表现。所提出的方法优于交叉注意力 [4] 和基线特征拼接。

5. 结论

本文提出了一种递归联合注意力模型，并结合双向长短期记忆网络，以实现基于回归的情感识别中的有效时空音视频融合。具体而言，该联合注意力模型以递归方式训练，从而能够不断优化音视频特征。我们还进一步研究了双向长短期记忆网络在捕捉基于回归的情感识别中各个音频和视觉模态的模态内时间动态，以及音视频特征的时间动态方面的作用。通过利用双向长短期记忆网络有效捕捉模态内关系，并利用递归联合注意力捕捉模态间关系，所提出的方法能够优于相关的最先进方法。

6. 参考文献

[1]Liam Schoneveld、Alice Othmani 和 Hazem Abdelkawy，《利用深度学习的最新进展进行视听情感识别》，《模式识别快报》，第 146 卷，2021年。

[2]Felix Kuhnke、Lars Rumberg 和 J¨orn Ostermann，“自然环境中的双流听觉-视觉情感分析”，载于*FG*研讨会，2020年。

[3] Vincent Karas、Mani Kumar Tellamekala、Adria Mallol- Ragolta、Michel Valstar 和 Bj¨orn W. Schuller，“采用循环机制与注意力机制进行自然环境中的情感识别的时间连续音视频融合”，载于*CVPRW*，2022。

[4] Gnana Praveen Rajasekhar、Eric Granger 和 Patrick Cardinal，“用于维度情绪识别的交叉注意力音视频融合”，载于*FG*，2021年。

[5]Panagiotis Tzirakis、George Trigeorgis、Mihalis A. Nicolaou、Bj¨orn W. Schuller 和 Stefanos Zafeiriou，“使用深度神经网络的端到端多模态情感识别”，*IEEE* 信号处理精选主题期刊，第 11 卷，第 8 期，第 1301–1309 页，2017 年。

[6] Panagiotis Tzirakis、Jiaxin Chen、Stefanos Zafeiriou 和 Bj¨orn Schuller，“真实环境中的端到端多模态情感识别”，《信息融合》，第68卷，2021年。

[7] Srinivas Parthasarathy 和 Shiva Sundaram，“使用多模态Transformer检测表情”，发表于*2021 IEEE*口语技术研讨会（*SLT*），2021年，第636–643页。

[8] R Gnana Praveen、Patrick Cardinal 和 Eric Granger，“利用联合交叉注意力在效价-唤醒度空间中进行情绪识别的音视频融合”，*IEEE*生物识别、行为与身份科学汇刊，2023年。

[9] 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 和 Eric Granger，“用于维度情感识别的音视频融合联合交叉注意力模型”，发表于*CVPRW*，2022。

[10] S Chen 和 Q Jin，《用于维度情感预测的多模态条件注意力融合》，载于*ACM* 会议论文集，2016年。

[11] Priyasad Darshana、F Tharindu、S Simon 和 F Clin- ton，《利用循环神经网络和基于注意力的融合学习多模态情感识别的显著特征》，载于*AVSP* 论文集，2019年。

[12] R Beard、R Das、R. W. M. Ng、P. G. K Gopalakrish- nan、L Eerens、P Swietojanski 和 O Miksik，《通过递归注意力进行多模态序列融合以实现情绪识别》，载于*CoNLL*，2018年。[13] 何恺明、张祥雨、任少卿和孙剑，《用于图像识别的深度残差学习》，载于*CVPR*，2016年，第770–778页。[14] Anwer Slimi、Mohamed Hamroun、Mounir Zrigui 和 Henri Nicolas，《利用频谱图和浅层神经网络从语音中识别情绪》，载于*ICAMCM*，2020年，第35–39页。

[15] Samuel Albanie、Arsha Nagrani、Andrea Vedaldi 和 Andrew Zisserman，《通过跨模态迁移在真实场景中进行语音情绪识别》，载于第*26*届*ACMM*，2018年，第292–301页。

[16] 樊寅、卢祥菊、李典和刘元柳，《使用卷积神经网络-循环神经网络与C3D混合网络进行基于视频的情绪识别》，载于*ACM ICMI*，2016年，第445–450页。

[17] Du Tran、Heng Wang、Lorenzo Torresani、Jamie Ray、Yann LeCun 和 Manohar Paluri，《深入探究用于动作识别的时空卷积》，载于*CVPR*，2018年，第6450–6459页。

[18] 段斌、唐浩、王伟、宗子亮、杨国伟和闫岩，“通过联合协同注意力递归融合进行视听事件定位”，载于*WACV*，2021年。

[19] Dimitrios Kollias、Panagiotis Tzirakis、Mihalis A.Nico- laou、Athanasios Papaioannou、赵国英、Bjorn Schuller、Irene Kotsia 和 Stefanos Zafeiriou，“自然场景中的深度情感预测：Aff-Wild数据库与挑战赛、深度架构及其他”，*IJCV*，第127卷，第907–929页，2019年。

[20]D Kollias、A Schulc、E Hajiyev 和 S Zafeiriou，“首届ABAW挑战赛中的情感行为分析”，载于*FG 2020*，2020年。

[21] Dimitrios Kollias 和 Stefanos Zafeiriou，“第二届ABAW2竞赛中的情感行为分析”，载于*ICCVW*，2021年，第3652–3660页。

[22] 张苏、丁毅、魏子全和关存泰，“基于视听引导者-跟随者注意力融合的连续情绪识别”，发表于*ICCVW*，2021年。

[23] Liyu Meng、Yuchen Liu、Xiaolong Liu、Zhaopei Huang、Wenqiang Jiang、Tenggan Zhang、ChuanheLiu 和 Qin Jin，“基于多模态时序感知特征的真实场景视频效价和唤醒度估计”，发表于*CVPRW*，2022，页码 2344–2351。