绪论

根据世界卫生组织（世界卫生组织）ss59]的数据，抑郁症是一种常见而严重的疾病，对超过2.8亿人的日常生活产生负面影响。抑郁症的严重表现被称为重度抑郁症（MDD）或重度抑郁症（MD），其定义为一种普遍和持续的情绪低落的精神状态，伴有对活动的厌恶ss5，40]。

MD很难诊断：与大多数身体疾病相比，常见的症状，如悲观、自卑和愤世嫉俗的行为，更主观，因此更难发现。因此，大约33%的抑郁症患者在临床诊断程序中没有被识别出来，只有不到40%的抑郁症患者得到了适当的治疗ss29]。幸运的是，抑郁症在某些情况下是可以治疗的，例如早期诊断ss24，34]，但让大多数人都能获得如此大规模的诊断将大大增加心理学家的工作压力，这表明了辅助工具对MD端到端自动估计的重要性。痛苦分析访谈语料库—奥兹巫师（DAIC—WOZ）数据集的发布ss12]使基于学习的抑郁症评估方法得以系统开发和评估。在这个数据集中，抑郁线索是从音频、视觉和文本数据中学习的，显示出对MD估计的良好性能，而音频是过去工作中最常见的模态（例如，DepAudioNet s39]）。MD的不同线索可以从不同类型的数据中获得：例如，抑郁症患者倾向于单调、虚弱或焦虑地说话，头部运动、眼神交流或微笑较少ss19，58]。鉴于不同数据源的互补性，多模态对于改进自动抑郁评估模型至关重要，但如何融合信息成为一个重要的研究方向，也是我们工作的主要动力。

文献综述

受益于数据驱动方法的快速发展，通过基于学习的技术进行抑郁评估引起了人们的极大关注。最近提出的针对抑郁症估计的方法可以分为分为四组：基于文本、基于音频、基于视觉和多模态的方法，通过发布多模态抑郁症估计数据集DAIC—WOZs12]来实现。根据s49]，即使心理状态的微小差异也有可能导致声学领域的显著变化，这使得音频数据成为一种竞争性的模态，可以用作抑郁症估计框架的输入s9，10，17，31，39，49，61]。Mel倒谱特征和共振峰频率轨迹是Williamson等人s61]提出的两种主要的觉醒表示，他们从声道共振频率和频谱动力学中获取有区别的觉醒线索。Scherer等人s49]介绍了四种基于语音的心理困扰特征，随后由支持向量机（SVM）进行分类由于近年来深度学习逐渐接管了模式识别领域，在深度神经网络而不是手工制作的基于特征的方法的帮助下取得了显著进展，He等人提供了基于音频数据的此类方法的概述s28]。Ma等人s39]报告了出色的性能，他们提出了DepAudioNet，这是一种使用卷积神经网络（CNN）和长短期记忆网络（LSTM）的端到端抑郁估计模型。Saidi等人s46]旨在通过多元回归进行抑郁程度估计融合。Sardari等人s48]提出了一种使用卷积自动编码器的基于音频的抑郁检测架构。除了音频，文本数据是另一个流行的抑郁线索来源。Dinkel等人s15]提出了一种用于凹陷估计的基于文本的多任务双向门循环单元（BGRU）网络。Salimath等人s47]提出了一种通过使用否定句来量化抑郁症严重程度的指标。视觉线索主要从面部关键点s7，23，25，44，64]或原始视频数据s1，3，41，42]中提取，也通过捕捉轻微的面部表情变化来用于抑郁估计。Facial Action Units（FAU）、面部标志、头部姿势和凝视方向被用作基于视觉数据的方法的CNN输入s16]。Xie等人s62]利用视频数据，使用3D CNN从有问题 的长期视频记录中解释抑郁症。为了结合不同模态类型的线索，近年来研究 人员还提出了几种融合架构s2，20，30，45，51，56]。s54]介绍了一种用于MD评 估的深度多模态网络。他等人s27]通过结合视觉和音频线索实现了多模态抑 郁估计。Gong等人s20]提出了一种更深层次的因果神经网络，用于融合不同 的模态。从传统图像分类的最新进展中获得灵感s11]，我们通过基于跨模态 注意力的融合来解决抑郁估计任务。此外，现有的大多数单模态和多模态方 法都依赖于大量数据预处理技术，这违反了端到端原则。与这些方法相比， 我们通过直接使用几乎原始的数据集和较少的数据清理技术来训练和测试所 研究的基线和我们提出的模型的性能，从而解决了MD评估问题。

绪论

抑郁症是最常见的精神障碍之一，表现为一系列生理症状，如体重减轻和失眠，严重病例可能导致自杀或药物滥用d1]。检测抑郁症面临两大挑战： （1）患者人数稳步增长， （2）人工诊断成本高昂。因此，迫切需要开发一种高效的抑郁症检测系统。

近年来，基于多模态的方法整合了音频、视频和文本模态的信息，在抑郁症检测方面显示出了有前景的结果。这些方法主要关注多模态融合，可分为三种类型：特征级、决策级和模型级融合