arXiv:2601.14651v1 [cs.CV] 21 Jan 2026

READ-Net：通过自适应特征重校准消除视听抑郁症检测中的情绪歧义

Chenglizhao Chen1，Boze Li1，Mengke Song[1]†，Dehao Feng1，Xinyu Liu1，Shanchen Pang1，Jufeng Yang2，Hui Yu[3]1中国石油大学（华东）计算机科学与技术学院，中国[2]南开大学计算机学院，中国[3]格拉斯哥大学计算科学学院，苏格兰，英国

摘要—**抑郁症是一种严重的全球性心理健康问题，会损害日常功能并降低整体生活质量。尽管近年来音视频方法提升了自动抑郁检测的效果，但忽略情绪线索的方法往往无法捕捉隐藏在情绪表达中的细微抑郁信号。相反，纳入情绪信息的方法常常在特征表示中将短暂情绪表达与稳定的抑郁症状混淆，这种现象称为情绪歧义，进而导致检测错误。为解决这一关键问题，我们提出了 READ-Net，这是首个专门通过自适应特征重校准（AFR）来消除情绪歧义的视听抑郁症检测框架。AFR 的核心思想是动态调整情绪特征的权重，以增强抑郁相关信号。READ-Net 并非简单地忽略或盲目结合情绪信息，而是创新性地识别并保留情绪特征中与抑郁相关的线索，同时自适应地滤除无关的情绪噪声。这种重校准策略显著澄清了特征表示，并有效缓解了长期存在的情绪干扰难题。此外，READ-Net 可轻松集成到现有框架中，以提升性能。在三个公开数据集上的大量评估表明，READ-Net 的表现优于最先进方法，准确率平均提高 4.55%，F1分数平均提高 1.26%，证明了其对情绪干扰的鲁棒性，并提升了视听抑郁症检测的效果。**

索引词—**抑郁症检测、情绪歧义、自适应特征重校准、多模态学习。**

一、引言

抑郁症EPRESSION是21世纪最严峻的公共卫生挑战之一，影响超过2.8亿——全球人口。世界卫生组织（WHO）预测，到2030年，抑郁症将成为全球疾病负担的首要因素。传统临床诊断主要依赖自评量表和结构化访谈等主观工具，这些工具固有地存在时效性有限、主观性强，以及容易受到患者隐瞒症状影响等问题 [1]。

为解决这些不足，近期研究日益重视基于客观抑郁症自动检测

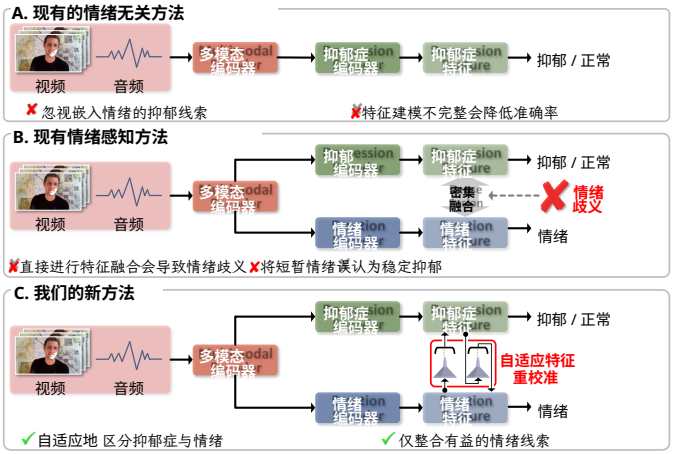

图1. 现有方法与所提出的音视频抑郁检测方法对比。(A) 情绪无关方法忽视了嵌入情绪中的抑郁线索，导致建模不完整。(B) 直接融合情绪特征与抑郁特征可能会混淆短暂情绪与稳定的抑郁信号。(C) 我们提出的方法采用自适应特征重校准，有选择地整合有益的情绪线索，从而提升准确率和鲁棒性。

生物标志物的抑郁症检测。音频-视觉模态，包括面部表情、声调 [2],和身体动作，因能够捕捉与抑郁状态相关的行为和生理线索而受到广泛关注 [3]。深度学习的进步推动研究从单模态分析转向多模态融合框架，通过整合互补信息源来建模复杂的情感模式。然而，许多多模态方法仍受制于简单的融合策略，也难以区分短暂的情绪波动与持续存在的抑郁指标，从而限制了现有方法的诊断效度和泛化能力。现有研究还表明，抑郁个体在面部行为、声音特征和身体动作方面呈现出特征性模式。

当前的视听抑郁症检测框架通常分为情绪无关方法和情绪感知方法。情绪无关方法 [4], [5], [6], [7]（图1-**A**）排除情绪线索，专注于抑郁症特有的生物标志物，例如语音停顿、回避注视和情感平淡。虽然避免了情绪噪声，但这类方法也舍弃了重要信息，因为抑郁症往往会

前两位作者对本研究的贡献相同。

†通讯作者：Mengke Song（songsook@163.com）。

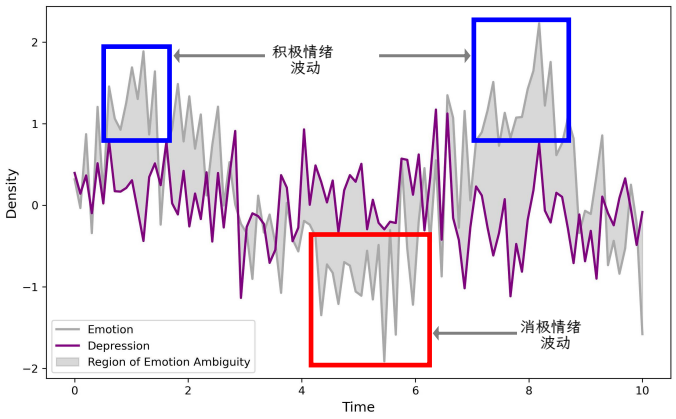

图2. 情绪歧义的可视化。该图展示了随时间变化的短暂情绪波动（积极和消极）与稳定抑郁信号之间的重叠。阴影区域表示情绪歧义的范围，在此范围内，情绪波动会使抑郁症状难以准确识别。

改变情绪反应性，包括对积极刺激反应迟钝，或出现与情境不符的表情，例如在痛苦时短暂微笑。忽略这些受情绪调节的行为会导致特征表示不完整。

情绪感知方法 [8], [9], [10] （图1-**B**）明确融入情绪特征，其前提是抑郁症会改变典型的情感模式。尽管此类方法能够捕捉与抑郁症相关的情绪波动，却也会引入情绪歧义（图2），即将短期情绪（如访谈中的瞬时悲伤或快乐）与持续性抑郁症状（如慢性情绪麻木）混淆。简单融合会让短暂情感污染稳定的抑郁线索，从而导致错误，例如将健康个体短暂的悲伤误判为抑郁症，或因偶尔出现的正性情感而掩盖抑郁个体的状态，最终损害表征质量和泛化能力。

一个尚未解决的关键挑战，是区分反映抑郁病理的情绪表达与受情境因素驱动的情绪表达。现实世界中的情绪反应受到访谈内容、社会互动和环境刺激的影响，常常产生与临床抑郁症无关的短暂情感状态。例如，谈论一段愉快的回忆可能会引发快乐，而回忆一件困难的事情则可能引发与病理无关的悲伤。这种情境依赖性引出了一个问题：模型检测到的是抑郁特质，还是瞬时反应？要解决这一问题，就必须认识到情绪与抑郁是动态交织的，而非简单叠加的信号。否则，受情境驱动的情绪可能会被误解为抑郁症，从而降低可解释性和诊断效度。

为缓解情绪歧义，我们提出了 READ-Net，这是一种新颖的音视频框架，可自适应地重校准情绪与抑郁的相互作用（图 1-**C**）。READ-Net 既不会去除情绪信息，也不会不加区分地将其整合进来。相反，它通过自适应特征重校准，动态地区分与抑郁相关的情绪线索和短暂情绪噪声。这种重校准弥补了情绪无关设计的局限，并避免将转瞬即逝的情绪与稳定的抑郁特质混为一谈，从而提高准确率和泛化能力。

在三个基准数据集 LMVD [11], D-vlog [12], 和 DAIC-WOZ [13], 上的实验表明，READ-Net 超越了最先进方法，同时表现出对情绪歧义的强大鲁棒性。其模块化架构还可无缝集成到现有框架中，为抑郁检测系统提供了广泛适用的增强方案。

总而言之，本文的主要贡献包括：1）作为该方向的首项研究，我们将抑郁症检测中的情绪歧义 概念形式化，并通过实验证明其对现有音视频方法的负面影响。2）我们提出了 READ-Net，这是一种统一的音视频抑郁症检测框架，通过自适应特征重校准机制应对情绪歧义。该框架可作为即插即用模块集成到现有系统中。3）我们进一步提出分层特征分离和双重一致性正则化，以将抑郁相关特征与情绪及噪声表示分离。4）在公开数据集上的实验表明，READ-Net 的表现优于最先进方法，准确率和 F1 分数平均分别提升 4.55% 和 1.26%，体现出对情绪歧义的强大抵御能力。代码、数据集和结果可在此获取[1]。

II. 相关工作

自动抑郁检测（ADD）借助音视频信号以及日益增多的语言信号取得了快速进展。AVEC'13/14 和 AVEC'17 等基准确立了情感与抑郁估计的研究规范，并推动了多模态处理流程的发展 [14], [15], [13]。除受控访谈外，D-vlog [12] 等真实场景资源以及 LMVD [11] 等大规模数据集提升了模型在现实条件下的鲁棒性，而半结构化临床语料库则支持筛查导向研究 [16]。标准声学工具包（openSMILE [17]）和早期深度音频模型（DepAudioNet [18]）提供了强有力的基线；鉴于临床量表的普遍使用，BDI 等标签可靠性问题也促使研究者谨慎设计实验方案 [1]。

A. 与情绪无关的视听方法

许多研究在没有明确情绪监督的情况下，对抑郁症的副语言和非语言标记进行建模。在声学方面，基于频谱图或原始波形的卷积神经网络和长短期记忆网络编码器，以及时空堆叠，能够捕捉抑郁语音中的韵律和发音线索 [2], [18]。在标签有限的情况下，自监督语音编码器（如 wav2vec 2.0 和 HuBERT ）进一步提升了鲁棒性 [19]。在视觉方面，深度网络对面部外观、光流、三维动态和多尺度时间差异进行建模，以检测表达能力减弱和微观动态的变化 [20], [6], [21]。身体层面的信号，包括自我适应动作和坐立不安，也提供了互补线索 [22]。

[1]https://github.com/READNet-DemoCode/READ-Net

多模态变体采用层级式或桥接式策略融合视听流，同时不依赖情绪信息 [23], [24]。这些流程虽有效，但可能会忽略更高层次的情感情境和细微的抑郁模式。

B. 情绪感知视听方法

为捕捉抑郁与情感的相互作用，情绪感知方法通过辅助标签、中间情感表征或关注具有情绪显著性的片段来引入情绪监督。时序Transformer联合学习抑郁症与情绪 [8]，，多任务框架则在多模态融合中整合情感倾向或情感线索 [9]。一些方法显式跟踪跨模态的情绪变化，以稳定预测结果 [25]。然而，对情绪线索进行统一利用，可能会将短暂情感（如短暂的喜悦或悲伤）与稳定的抑郁特质（如快感缺失）混为一谈，从而产生情绪歧义。相关证据来自情感标注的不确定性建模 [26] 以及对抑郁者面部行为的模态特定观察 [27]。在语音文本场景中，如果情境信息或风格信息未被解耦，跨模态注意力可能会使抑郁特征产生偏差 [28]。因此，朴素的情绪整合可能引入噪声并削弱鲁棒性，尤其是在访谈和视频博客中的长弱标注序列上。

C. 特征分离与歧义缓解

近期研究转向采用选择性 机制，突出抑郁相关证据，同时抑制干扰因素。在视觉建模中，双重注意力和逐元素重校准能够突出信息通道和时间区域。在语音和跨模态领域，分层注意力、单调注意力以及加性跨模态对齐能够定位显著的声学片段，并对齐音频文本 [29], [28]。对于长篇视频博客，时间感知融合和联合注意力融合可在弱监督条件下改进信息聚合 [30], [31], [4]。为扩展至长情境，Mamba变体等状态空间模型和长程时序模型支持高效的跨模态融合 [32], [5]。因果融合和公平融合方法还可应对混杂和人口统计偏差 [33]。

基于这些方向，我们明确地通过歧义感知门控、自适应特征重校准和跨模态对齐来处理情绪歧义 [34]。该设计保留抑郁相关情绪证据，同时抑制语境驱动的瞬时情感，并能够扩展到AVEC式数据和视频博客数据中常见的长序列弱标注序列 [14], [13], [12], [11]。它融合了情绪无关线索的鲁棒性 [2], [24], [23], 、注意力和重校准机制的选择性 [35], [29], [28], [36],以及时间感知融合的时间敏感性 [30], [4],，从而实现更细粒度且更解耦的多模态抑郁检测。

III. 所提方法

A. 深入探究情绪歧义

情绪歧义是指在视听抑郁分析中，难以区分瞬时情绪状态与稳定的抑郁症状。这是因为面部表情、声调和身体动作等行为线索，往往会在瞬时情绪与临床抑郁症之间重叠。情绪具有动态性且依赖情境，而抑郁症则体现为持久、稳定的模式。将情绪特征直接整合到抑郁模型中，可能会把短期波动与长期抑郁信号混杂在一起，从而引入噪声并降低准确率和泛化能力。为阐明这一问题，我们重点介绍两个典型场景：

非抑郁个体的短暂悲伤：如图所示，2 （红框），非抑郁个体可能因外部因素而短暂表现出悲伤，表现为嘴角下垂、语速减慢或精力下降。这些线索与抑郁指标相似，若解读有误，可能导致假阳性。

抑郁个体的瞬时积极情绪：相反，如图所示，2 （蓝框），抑郁个体可能会短暂表现出快乐或投入，例如在互动时微笑。这类转瞬即逝的积极表达可能掩盖其潜在的抑郁状态，若过分强调，可能导致假阴性。

在这两种情况下，瞬时情绪表达都可能与受试者真实的临床状况相冲突。这些不匹配正是抑郁检测模型中情绪歧义产生的原因。

情绪歧义的形式化。令 $Y \in \{0, 1\}$ 表示序列级抑郁标签，其中 $Y = 1$表示抑郁受试者，$Y = 0$ 表示非抑郁受试者。令 $\mathbf{E}t \in \mathbb{R}^{d_e}$表示时间步 t 的情绪特征，$\mathbf{E}1{:}T = \{\mathbf{E}1, \ldots, \mathbf{E}T\}$ 表示情绪轨迹。我们考虑一个仅使用情绪信息、输出后验概率 $p_\theta(Y \mid \mathbf{E}t)$ 的情绪分类器。

当瞬时情绪表达相较于真实状态，强烈倾向于相反的抑郁症标签时，就会出现情绪歧义。给定间隔 $\tau \in (0,0.5)$，情绪歧义时间步的集合定义为

$$\mathcal{A}_\tau = \left\{ t \in \{1, \ldots, T\} : p_\theta(Y = 1 - Y \mid \mathbf{E}_t) \geq 0.5 + \tau \right\}, \quad (1)$$

其中，$1 - Y$ 表示相反类别。因此，$\mathcal{A}_\tau$ 包含仅情绪分类器强烈倾向于错误标签的时刻。根据 $\mathcal{A}_\tau$，序列级情绪歧义率为

$$\mathrm{EA}_{\mathrm{err}}(\mathbf{E}_{1:T}, Y) = \frac{1}{T} \sum_{t=1}^{T} \mathbb{I}[t \in \mathcal{A}_\tau], \quad (2)$$

其中 $\mathbb{I}[\cdot]$ 是指示函数。较高的情绪歧义率$_{\text{误差}}$ 表明短暂情绪表达经常误导分类器。如图 2所示，这包括：(i) 非抑郁受试者表现出短暂的负面情绪或情感迟钝，呈现出类似抑郁行为的特征；(ii) 抑郁受试者暂时表现出积极或中性情感，从而掩盖其状况。这些片段会破坏特征表示，并导致不可靠预测。

挑战在于，情绪表达会随着情境迅速变化，而抑郁症则持续存在且普遍存在。当不加区分地使用情绪特征时，

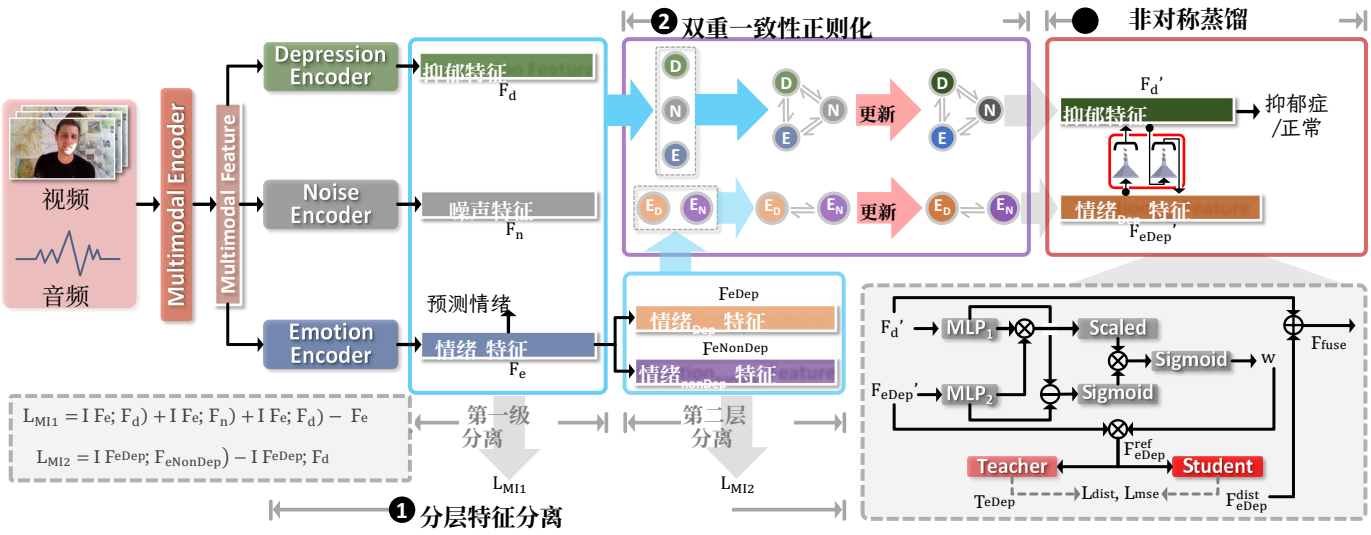

图3。READ-Net 概览。该网络融合了分层特征分离（1）、双重一致性正则化（2）和非对称蒸馏（3），以有效处理情绪歧义。通过动态重新校准与情绪相关的特征，并利用多模态数据，READ-Net 显著提升了抑郁症检测的准确率和可靠性，优于传统方法。

算法1READ-Net 的训练与推理流程

要求： 数据集 $\mathcal{D} = \{(x_{ai}, x_{vi}, y_i)\}$ **确保：** 预测 $\hat{y}$

阶段0：预训练与提取

训练情绪识别分支以获得提取器$\mathcal{E}_{emo}$。为每个样本提取原始音频–视觉特征F_{raw}。

阶段1：分层特征分离（HFS）

通过三条流对 Fraw 进行编码，得到 Fd、Fn 和 Fe。计算第一级互信息损失 LMI1。将 Fe 分解为 →{FDepe , FNonDepe }，并计算 LMI2。

阶段2：双重一致性正则化（DCR）

在 {Fd, Fn, Fe} 上构建时间图，并更新特征
通过图卷积网络。

计算图一致性损失 Lgraph。

在 {FDepe ,FNonDepe } 上构建子图，并计算信息流损失 Lflow。

阶段3：非对称蒸馏（AD）

应用注意力，将 F_{e}^{Dep} 优化为 F_{e}^{f}。通过教师–学生门控循环单元，将 Frefe 蒸馏为 Fdiste。融合 Fd 和 Fdiste，得到融合特征 Ffuse。

阶段4：优化

根据融合特征 Ffuse 预测抑郁评分 $\hat{y}$。计算 Lcls 和总损失 Ltotal。通过反向传播更新所有参数。

推理

使用预训练的 $\mathcal{E}_{emo}$ 和训练完成的 READ-Net，以前向模式运行阶段1–3，得到$\hat{y}$。

短暂情感波动会污染稳定抑郁相关的生物标志物（例如长期情感迟钝或回避注视），使表征质量下降，并因掩盖持久抑郁信号与暂时性情绪噪声之间的区别而阻碍可靠的视听抑郁症检测。

B. 我们消除情绪歧义的方法

为解决情绪歧义，READ-Net采用自适应特征重校准机制 [35], [20]，评估情绪特征与抑郁症的相关性，在保留情感迟钝等抑郁症相关线索的同时抑制情境性情绪噪声。

自适应特征重校准（第 III-C节）包含三个组成部分。分层特征分离通过保留抑郁相关线索并避免直接融合短暂的情绪波动，将情绪特征与抑郁特征相结合。双重一致性正则化（第III-D节）优化情绪特征与抑郁特征之间的关系，确保仅保留与抑郁特质一致的持续性情绪模式。非对称蒸馏（第 III-E节）进一步增强与抑郁相关的情绪线索，通过从混合情绪信号中选择性蒸馏有用信息，减少噪声影响并提高预测精度。

这些组成部分共同使 READ-Net 能够区分与抑郁相关的情绪线索和短暂情绪，并提高鲁棒性。算法1 概述了整个流程：阶段0对情绪分支进行预训练并提取视听特征；阶段1–3应用 HFS、DCR 和非对称蒸馏，以解耦并融合抑郁相关表征，生成融合特征 Ffuse 和预测$\hat{y}$；阶段4使用总损失 Ltotal 对所有模块进行联合优化，推理时则使用训练好的模块，不进行更新。

如算法 1 所示，各阶段分别对应第 III-C–III-E 节详述的模块。阶段1实现分层特征分离模块，其目标基于互信息损失 LMI1 和 LMI2。阶段2执行双重一致性正则化，引入基于图的损失 Lgraph 和 Lflow，以增强类型内一致性并抑制跨类型干扰。阶段3实现非对称蒸馏，对与抑郁症相关的情绪特征进行细化和蒸馏，并将其注入抑郁分支。阶段4随后将分类损失 Lcls 与所有正则化项结合，构成

后续实验中使用的总训练损失 Ltotal。

C. 分层特征分离

如图 3 (1)所示，音频和视频等多模态抑郁检测数据通常同时包含与抑郁相关的情绪线索和无关的情绪波动。关键挑战在于将抑郁特异性信号与噪声或非抑郁情绪区分开来。若缺乏有效的解耦，模型可能会误解短暂情绪，或无法捕捉真实的抑郁状态。为此，我们提出分层特征分离，以从情绪噪声中分离出抑郁相关特征，并提高预测精度。

其核心思路是先从多模态数据中提取整体情绪信号，再进一步区分与抑郁相关的情绪线索和一般情绪变化。这种两级分离至关重要。第一级从一般情绪和噪声成分中分离出抑郁特异性模式，使模型能够专注于相关信号。第二级则进一步将与抑郁相关的情绪与无关情绪区分开来，从而避免细微的非抑郁情绪干扰检测。这种分层策略能够增强模型对抑郁核心证据的关注，同时滤除短暂信号。

1）特征提取：给定多模态输入 Fraw ∈ℝT×D，我们构建三个并行流，作为独立的特征提取器，在相同证据上运行但承担不同的作用。**抑郁流**通过多模态架构生成 Fd，该架构结合了多尺度时间卷积网络与双向跨模态注意力，随后进行分层时间聚合，以获得丰富的抑郁相关时频表征。

噪声流使用相同的骨干网络估计 Fn，但采用不同的训练目标。梯度反转层和正交约束促使表征偏离抑郁流形。具体而言，我们施加 Lorth = |⟨Fd,Fn⟩|2，以最小化 Fd 与 Fn 的内积，使 Fn 张成一个正交特征空间。

情绪流采用与抑郁分支相同的网络结构生成 Fe，但使用六分类情绪头（快乐、悲伤、愤怒、恐惧、惊讶、厌恶），以捕捉

为后续评估提取通用的情绪表达。

2) 预训练与知识迁移： 为了将情绪特征有效地整合到抑郁症检测模型中，我们首先使用经过仔细标注的数据集对情绪识别分支进行预训练。在 DAIC-WOZ 数据集上，我们利用面部动作单元（面部动作单元） [37]为每一帧标注六种基本情绪。经过特征提取、筛选和标准化后，我们构建了一个包含 1,200 个样本的均衡情绪训练集。为解决不同数据集之间的差异（D-vlog 缺少面部动作单元，而 LMVD 缺少原始音频数据），我们从 DAIC-WOZ 中提取了兼容的情绪特征，确保不同数据集之间的一致性。

我们采用了两阶段训练策略：首先，独立预训练情绪识别分支，使验证准确率达到 70% 以上。接着，我们冻结其参数，并将其用作固定的特征提取器，以完成

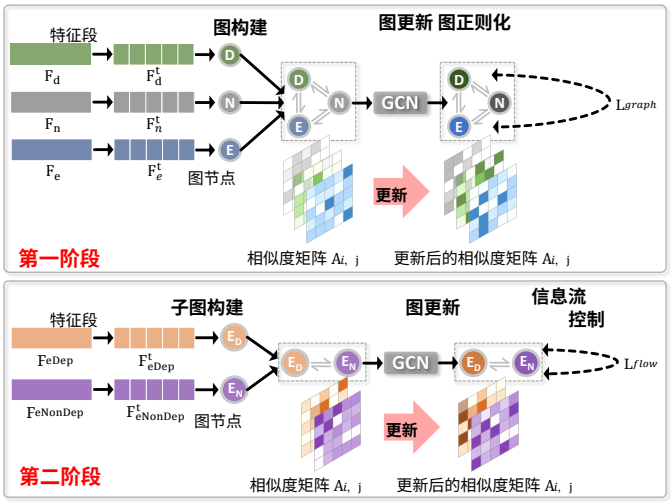

图 4. 双重一致性正则化（DCR）工作流程。展示了 DCR 的两个阶段：第一阶段通过图构建、更新和正则化（Lgraph）分离 F_d、F_n 和 F_e；第二阶段使用子图构建、更新和信息流控制（L_{flow}）细化 F^{eDep} 和 $F^{eNonDep}$，从而增强可区分性和鲁棒性。

抑郁症检测任务。这种情绪知识迁移机制通过捕捉抑郁症特有的情绪表达模式，提升了模型在结构各异的数据集上的抑郁症检测性能。

3）第一层分离：第一层分离旨在通过最大化这些特征集合之间的互信息，将抑郁特征、噪声特征和情绪特征解耦。此步骤可确保模型隔离相关的抑郁相关特征，并去除噪声或无关的情绪信号。损失函数定义如下：

$$L_{MI1} = \lambda_0 \cdot I(F_d; F_n) + \lambda_1 \cdot I(F_e; F_d) + \lambda_2 \cdot I(F_e; F_n) - \lambda_3 \cdot H(F_e) \quad (3)$$

其中，I(·;·) 表示神经互信息估计器，H(·) 表示熵正则项，而 λ_0、λ_1、λ_2、λ_3 是加权超参数。

4）第二层分离：在第一层分离出情绪特征 Fe 后，第二层分离着重于进一步将抑郁相关情绪特征 FeDep 与非抑郁相关情绪特征 FeNonDep 分离。此步骤可提高模型的特异性，确保短暂情绪不会干扰对稳定抑郁特征的识别。

计算抑郁相关情绪特征 FeDep 与抑郁特征 Fd 之间的互信息，同时最小化抑郁相关情绪特征与非抑郁相关情绪特征 FeNonDep 之间的互信息。损失函数定义如下：

$$L_{MI2} = I(F_{eDep}; F_{eNonDep}) - I(F_{eDep}; F_d) \quad (4)$$

通过结合分层特征分离、情绪预训练和知识迁移，我们确保抑郁症检测模型能够有效分离抑郁相关特征，同时处理情绪歧义，从而在各种数据集上实现更准确、更稳健的性能。

D. 双重一致性正则化

如图 3 （2）所示，双重一致性正则化（DCR）作用于分层

特征分离（HFS）模块的输出。HFS 首先将原始多模态特征解耦为抑郁特征 Fd、噪声特征 Fn 和情绪特征 Fe，随后进一步将 Fe 细化为与抑郁相关的 FeDep 和与非抑郁相关的 FeNonDep。然而，由于时间动态的重叠和复杂的多模态交互，这些特征通常仍保留细微的相互依赖关系，从而削弱分离的清晰度和模型的鲁棒性。

为此，我们提出 DCR，这是一种针对批次优化的方法，通过增强特征类型内部的一致性并尽量减少不同类型间的相关性，来提升特征判别性。如图4所示，DCR 分两个阶段运行：动态图构建与正则化（第III-D1节），以及子特征动态正则化（第 III-D2 节），从而在保持计算效率的同时实现精确分离。

1）动态图构建与正则化：如图4-第一阶段所示，该阶段利用特征 Fd、Fn 和 Fe 之间的时间关系，每个特征均表示为批张量 $\mathbb{R}T\times D$，其中 T 为序列长度，D 为特征维度。

图构建。为捕获时间动态，这些特征被视为动态图中的节点，并将每个序列划分为 K 个片段（例如，K = 5）。每个片段对应一个图节点 Vt ={Ftd, Ftn, Fte}，t = 1. . . , K，从而能够高效捕获批次在不同时间步的动态。

计算相似度矩阵 $A \in \mathbb{R}K\times K$，以建模批次内片段之间的关系：

$$A_{ij} = \text{softmax}\left(\frac{\text{MLP}(F_i)^T\text{MLP}(F_j)}{\sqrt{D}} \cdot \exp(-\gamma \cdot |t_i - t_j|)\right). \tag{5}$$

此处，多层感知机将特征投影到较低维空间（例如，从 256 维降至 128 维），指数衰减项 exp(−γ · |ti −tj|)（γ经验上设为 1.0，以平衡局部时间相关性与全局信息。）通过降低对远距离片段的注意力，优先保证局部时间一致性。这确保图能够反映情境相关的特征交互。

图更新。为高效地在图中传播信息，我们采用批并行图卷积：

$$H^{(l+1)} = \text{ReLU}(\hat{A} \cdot H^{(l)}W^{(l)}), \tag{6}$$

其中，ˆA = $D^{-1}/2$(A +I)$D^{-1}/2$ 是带自环的归一化邻接矩阵，W(l) 是特定于层的权重矩阵。此更新通过聚合各片段间的时间依赖关系来优化特征。ReLU 是用于在特征更新中引入非线性的激活函数。

图正则化。为促进特征的一致性与独立性，我们应用正则化损失：

$$L_{\text{graph}} = \frac{1}{B}\sum_{b=1}^{B}(S_b - \lambda \cdot C_b), \tag{7}$$

其中，Sb = ∑i diag(Fbi (Fbi)T) 用于衡量每种特征类型 Fd、Fn、Fe 内部的自相关，而 Cb = ∑i≠j diag(Fbi (Fbj)T) 则用于惩罚跨类型相关性。二者的平衡由λ控制。图更新后，特征通过一个共享层和残差连接投影回原空间，并通过 α （默认值为 0.7）加权，以整合图增强信息与原始信息。

2）子特征动态正则化：如图4-第二阶段所示，正则化的第二阶段利用批处理，高效地优化抑郁相关情绪特征 FeDep 与非抑郁相关情绪特征 FeNonDep 之间的分离。二者均表示为维度为 $\mathbb{R}T\times D$ 的张量。

子图构建。我们将 FeDep 和 FeNonDep 都划分为 K 个片段。每个片段的特征都通过共享的双层多层感知机进行编码，以获得低维表示：

$$F^{\text{enc}}_{\text{eDep}}, F^{\text{enc}}_{\text{eNonDep}} = \text{MLP}(F^{\text{seg}}_{\text{eDep}}), \text{MLP}(F^{\text{seg}}_{\text{eNonDep}}). \tag{8}$$

计算跨模态交互矩阵：

$$A_{\text{cross}} = \text{softmax}\left(\frac{F^{\text{enc}}_{\text{eDep}} \cdot (F^{\text{enc}}_{\text{eNonDep}})^T}{\sqrt{D_{\text{enc}}}}\right). \tag{9}$$

该矩阵通过对批次内的成对交互进行建模，捕捉 FeDep 与 FeNonDep 之间的关系，并按特征维度进行缩放以确保稳定性，从而突出二者的差异性。

图更新。与采用独立图卷积网络（GCN）的传统方法不同，我们采用共享权重图卷积网络同时更新这两种特征类型，确保特征转换一致：

$$F^{\text{upd}}_{\text{eDep}} = \text{ReLU}(\hat{A}_{\text{norm}} \cdot F^{\text{enc}}_{\text{eDep}}W_e) \tag{10}$$

$$F^{\text{upd}}_{\text{eNonDep}} = \text{ReLU}(\hat{A}_{\text{norm}} \cdot F^{\text{enc}}_{\text{eNonDep}}W_e) \tag{11}$$

其中，ˆAnorm = $D^{-1}/2$(Across +I)$D^{-1}/2$ 是带自环的归一化邻接矩阵，We 是共享可学习权重矩阵，采用 Xavier 均匀初始化，以确保梯度流稳定。

信息流控制。我们设计了一种针对批次优化的信息流控制损失：

$$L_{\text{flow}} = -\sqrt{\text{Var}(F_{\text{eDep}})} - \sqrt{\text{Var}(F_{\text{eNonDep}})} + \mu \cdot R_{\text{dn}} \tag{12}$$

其中，Rdn 表示展平后的 FeDep 与 FeNonDep 之间的标准化协方差，计算方法为协方差除以两者标准差的乘积。这有助于提升类型内多样性，同时抑制跨类型相似性，其中μ = 0.1 用于控制权衡。

残差集成。经过图处理后，更新后的最终特征 FfinaleDep 通过线性层投影回原始维度，并通过残差连接与原始特征集成：

$$F^{\text{final}}_{\text{eDep}} = \alpha \cdot F_{\text{eDep}} + (1 - \alpha) \cdot \text{Proj}(F^{\text{upd}}_{\text{eDep}}). \tag{13}$$

这种残差结构结合 α = 0.7，在利用基于图的特征精炼优势的同时保留了重要的原始特征信息，从而促进特征在训练迭代中的稳定演变。

总体而言，这种批次优化方法增强了特征判别性，减少了冗余，并提升了 READ-Net 在抑郁症检测中的鲁棒性和可解释性。

E. 非对称蒸馏

如图 3 (3)所示，在 READ-Net 框架中，前序模块（例如分层特征分离和双重一致性正则化）能够有效精炼特征表示，生成抑郁特征 F′d 和抑郁相关情绪特征 F′eDep。然而，F′eDep 中可能仍存在细微的情绪噪声，而情绪信号中的抑郁线索可能未能充分融入 F′d，从而可能削弱鲁棒性和准确率。对称融合方法存在双向干扰的风险，可能使噪声污染抑郁特征。为此，我们提出了 **非对称蒸馏（AD）**模块，该模块选择性地将与抑郁相关的情绪线索融入 F′d，同时抑制无关噪声，从而形成鲁棒的融合表示 Ffuse。通过确保信息从情绪特征单向流向抑郁特征，AD 提升了抑郁症检测的准确率和鲁棒性。AD 模块包括三个步骤：非对称特征精炼、特征蒸馏，以及融合与分类。

1) 非对称特征精炼： 非对称性是指情绪特征单向增强抑郁症特征，从而避免反向噪声干扰。此步骤确保 FrefeDep 通过注意力选择性地放大与抑郁症相关的信息。该步骤从 F′eDep $\in \mathbb{R}^{T\times D_{eDep}}$ 中识别并增强与抑郁症相关的子特征，同时滤除无关噪声。自适应注意力权重 wt 的计算方式如下：

$$w_t = \mathrm{softmax}\Big(\frac{\mathrm{MLP}_1(F'_{eDep}[t])^{T}\cdot \mathrm{MLP}_2(F'_{d}[t])}{\sqrt{D_{eDep}}} \cdot \mathrm{sigmoid}(\mathrm{Linear}(F'_{eDep}[t]-F'_{d}[t]))\Big), \quad (14)$$

其中，MLP1（128 个单元）和 MLP2（64 个单元）对特征进行编码，Sigmoid门则选取相关成分。

精炼后的特征为：

$$F^{ref}_{eDep} = W \odot F'_{eDep}, \quad W = [w_1, w_2, \ldots, w_T]. \quad (15)$$

2) 特征蒸馏： 此步骤利用知识蒸馏将精炼后的特征压缩为低维且信息丰富的表示，从而提高效率并减少冗余。教师网络（三层门控循环单元，每层 256 个单元）生成：

$$T_{eDep} = \mathrm{GRU}_3(\mathrm{GRU}_2(\mathrm{GRU}_1(F^{ref}_{eDep}))) \in \mathbb{R}^{T\times 128}. \quad (16)$$

轻量级学生网络（单层门控循环单元，64 个单元）生成：

$$F^{dist}_{eDep} = \mathrm{GRU}_{light}(F^{ref}_{eDep}) \in \mathbb{R}^{T\times 64}. \quad (17)$$

学生网络通过 KL 散度和均方误差损失模仿教师网络，温度参数设为 $\tau = 2.0$，如下所示：

$$L_{dist} = \mathrm{KL}\Big(\mathrm{softmax}(T_{eDep}/\tau),\ \mathrm{softmax}(F^{dist}_{eDep}/\tau)\Big). \quad (18)$$

$$L_{mse} = \Big\|\mathrm{Linear}(F^{dist}_{eDep}) - T_{eDep}\Big\|_2^2. \quad (19)$$

$$L_{total} = L_{dist} + 0.5\,L_{mse}. \quad (20)$$

这能高效地强化单向增强。

融合与分类。 这一步确保蒸馏后的情绪特征与抑郁症特定特征无缝整合，从而生成全面且稳健的特征表示用于分类。蒸馏后的特征通过与 F′d 融合，与抑郁特征空间对齐：

$$F_{fuse} = F'_d + \beta \cdot F^{align}_{eDep}, \quad (21)$$

其中 β 是用于平衡情绪特征与抑郁特征贡献的权重。消融实验（表 IV-C）确定 0.5 为最优权重。最后，通过全连接层根据融合特征 Ffuse 对抑郁症进行分类。

IV. 实验

A. 实验设置、数据集与指标

实验设置。 READ-Net 使用 PyTorch 实现，并在 NVIDIA RTX 4090 GPU 和一颗 16 核 Intel(R) Xeon(R) Platinum 8352V CPU 上训练。我们使用 Adam优化器，学习率为 1×10^{-5}，训练 300 轮，批次大小为 16，并采用早停以防止过拟合。为提升泛化能力，我们采用了包括随机裁剪在内的标准数据增强方法。推理时，该模型处理每个样本约需 30 毫秒，计算量为 2 十亿次浮点运算。

数据集。 我们的实验在三个抑郁症检测数据集上进行：LMVD [11],D-vlog [12],和 DAIC-WOZ [13]。LMVD 数据集包含来自 1,475 名参与者的 1,823 个视频（214 小时），其中有 908 个抑郁样本和 915 个非抑郁样本，并提供了预先提取的特征，如面部动作单元、面部关键点和眼动。D-vlog 数据集包含来自 816 名说话者的 961 个 YouTube 视频博客（160 小时），其中有 555 个抑郁样本和 406 个非抑郁样本。DAIC-WOZ 包含通过虚拟访谈系统“Ellie”采集的 189 场访谈（733 分钟，平均 16 分钟），并提供 PHQ-8 评分及 COVAREP音频、面部动作单元等多模态特征。

指标。 遵循以往研究 [32], ，我们采用四项经典指标全面评估模型性能：准确率、精确率、召回率和 F1分数。

B. 定量评估

在 LMVD 数据集上的评估将 READ-Net 与最先进方法进行了比较，包括 Xception [38], SERes-Net [39], 视觉变换器 [40],MDDformer [11], 和 DepMamba [32]。如表 I所示，READ-Net 的表现优于所有竞争方法，取得了最高的准确率（77.59%）、精确率（78.10%）、召回率（77.59%）和 F1分数（77.47%）。表现次优的方法 LMVD 的准确率为 73.79%。在 LMVD 上，报告结果为 10 折交叉验证的均值。此外，READ-Net 与 MDDformer 之间的配对t检验得到 p = 0。023 < 0.05，表明这些改进具有统计显著性。READ-Net 的性能提升源于其自适应特征重校准机制，该机制能够有效处理情绪歧义并保留抑郁相关线索。

在 D-vlog 数据集上，READ-Net 与 STFT [41],TAMFN [30], JAMFN [31],DepMSTAT [4],进行了比较

表 I LMVD、D-vlog 和 DAIC-WOZ 集合上与最先进水平代表性方法的定量比较。每列中的最佳结果以粗体标出。

		LMVD 数据集						D-vlog 数据集							DAIC-WOZ 数据集				
Set	度量	Xcept [2017]	SERes [2017]	ViT [2020]	MMD [2024]	DepM [2025]	OURS	TAMFN [2022]	JAMFN [2023]	DepMS [2024]	STFT [2024]	ABDM [2024]	DepM [2025]	OURS	Perc [2021]	FUSE [2024]	RBF [2024]	HiQuE [2024]	OURS
LMVD	准确率	71.38	72.54	73.03	73.79	72.13	**77.59**	-	-	-	-	-	-	-	-	-	-	-	-
	精确率	71.94	73.13	73.52	74.20	70.18	**78.10**	-	-	-	-	-	-	-	-	-	-	-	-
	召回率	71.38	72.54	73.03	73.79	76.56	**77.59**	-	-	-	-	-	-	-	-	-	-	-	-
	F1	71.19	72.36	72.90	73.68	73.20	**77.47**	-	-	-	-	-	-	-	-	-	-	-	-
D-vlog	精确率	-	-	-	-	-	-	68.08	68.18	71.53	64.67	67.27	69.64	**73.94**	-	-	-	-	-
	召回率	-	-	-	-	-	-	**82.93**	68.39	75.60	77.51	67.77	69.79	79.27	-	-	-	-	-
	F1	-	-	-	-	-	-	74.75	68.25	73.51	69.77	66.64	68.91	**76.52**	-	-	-	-	-
DAIC-WOZ	精确率	-	-	-	-	-	-	-	-	-	-	-	-	-	75.00	52.00	68.00	78.00	**82.88**
	召回率	-	-	-	-	-	-	-	-	-	-	-	-	-	59.00	60.00	66.00	**80.00**	78.72
	F1	-	-	-	-	-	-	-	-	-	-	-	-	-	58.00	57.00	67.00	79.00	**79.53**

表 II 现有抑郁症检测模型将 READ-Net 用作即插即用模块后的性能提升。

数据集指标	ViT	ViT+	Xcep	Xcep+	MMD	MMD+	DepM	DepM+
准确率	74.17	**76.37**	73.07	**74.17**	73.79	**75.17**	68.40	**70.98**
F1	75.64	**75.97**	75.37	**76.38**	73.68	**74.46**	68.36	**70.10**
准确率	63.20	**64.62**	66.50	**68.39**	68.86	**70.31**	69.79	**71.35**
F1	65.48	**74.57**	**75.43**	71.72	**75.91**	69.83	68.91	**70.69**
准确率	**70.21**	**70.21**	74.46	**76.59**	72.34	**74.46**	**74.46**	**74.46**
F1	68.58	**69.50**	69.05	**70.69**	67.44	**70.74**	72.05	**73.86**

ABDM [42], 和 DepMamba [32]。READ-Net 的精确率达到 73.94%，比排名第二的方法 DepMSTAT（71.53%）高出 2.41%。尽管 TAMFN 的召回率（82.93%）高于 READ-Net（79.27%），但 READ-Net 的 F1分数高于 TAMFN（76.52% 对比 74.75%）。

在 DAIC-WOZ 数据集上，READ-Net 与 FAIRREFUSE [33], Perceiver [43], RBF [44],和 HiQuE [45] 等方法进行了比较。READ-Net 的表现优于所有方法，准确率达到 82.88%，比次优方法 HiQuE（78%）高出 4.88%。虽然 HiQuE 的精确率略高（80%），但 READ-Net 的 F1分数更优（79.53% 对比 79%）。这些结果证实，READ-Net 提升了抑郁症检测性能，相较于现有方法提供了更稳健、准确的解决方案。

C. READ-Net 作为现有抑郁症检测模型的性能增强器

如表II所示，READ-Net 不仅作为独立模型表现出色，还可作为即插即用模块，增强现有的抑郁症检测框架。 为便于说明，表*I* 列出了原始论文报告的数据，而表*II* 则展示了我们的复现结果，因此略有偏差在所难免。集成 READ-Net 后（记为“+”），其在所有数据集上的性能均显著提升。平均而言，准确率提高了 1.48% （从 70.77% 提升至 72.25% ），F1 分数提高了 1.05%（从 71.32% 提升至 72.37%）。具体来说，在 LMVD 数据集上，将 READ-Net 与视觉变换器集成后，准确率从 74.17% 提升至 76.37%，F1分数从 75.64% 提升至 75.97%。在 D-vlog 数据集上，DepMamba 与 READ-Net 结合后，准确率从 69.79% 提升至 71.35%，F1分数从 68.91% 提升至 70.69%。在 DAIC-WOZ 数据集上，FAIRREFUSE 与 READ-Net 结合后，准确率从 76.50% 提升至 78.23%，F1分数从 77.12% 提升至 78.45%。这些结果证实了 READ-Net 的

表 III

LMVD 数据集上各组件的消融研究。“NOISE”移除噪声特征提取分支；“EMO.”移除情绪特征提取路径；“1ST”和“2SEC”分别移除第一级和第二级分离过程；“UP3”和“UP2”移除

动态图构建和正则化中的图更新过程，以及子特征动态正则化；“REG3”和“REG2”分别消除了图正则化和信息流控制机制；“DISTILL”和“FUSE”展示了从非对称蒸馏模块中移除特征蒸馏和融合后的影响。

非对称蒸馏模块。

No.	HFS（第 3.3 节）				DCR（第 3.4 节）				非对称蒸馏（第 3.4 节）		指标			
	噪声Emo.	1st	2sec	up3	up2	REG3	REG2	DISTILL	FUSE	准确率	Pre.	Rec.	F1	
0											71.42	74.07	65.93	69.76
1		✓	✓	✓	✓	✓	✓	✓	✓	✓	76.92	77.52	75.82	76.66
2	✓										72.52	73.56	70.32	71.91
3	✓	✓	✓		✓	✓	✓	✓	✓	✓	75.27	73.95	78.02	75.93
4	✓	✓		✓	✓	✓	✓	✓	✓	✓	73.07	72.34	74.72	73.51
5	✓	✓	✓	✓		✓	✓	✓	✓	✓	75.82	76.40	74.72	75.55
6	✓	✓	✓	✓	✓		✓	✓	✓	✓	76.92	77.52	75.82	76.66
7	✓	✓	✓	✓	✓	✓		✓	✓	✓	74.17	73.91	74.72	74.31
8	✓	✓	✓	✓	✓	✓	✓		✓	✓	75.27	73.46	79.12	76.19
9	✓	✓	✓	✓	✓	✓	✓	✓		✓	76.37	77.90	73.62	75.70
10	✓	✓	✓	✓	✓	✓	✓	✓	✓		75.82	74.73	76.02	76.34
11	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	**77.59**	**78.10**	**77.59**	**77.47**

模块化设计有效增强了现有模型的特征提取和情绪歧义处理能力，使其成为抑郁症检测中提升性能的通用模块。此外，不同骨干模型和数据集上持续取得的改进，凸显了 READ-Net 的泛化能力和可迁移性。其模块化架构能够无缝集成，同时不会引入显著的计算开销，因此可作为现实世界抑郁症检测系统的实用增强方案。

D. 组件评估

对 READ-Net 核心阶段开展的消融研究如表III所示。

在分层特征分离（第 III-C节）中，移除噪声分支的影响很小（F1 = 76.66%）。取消情绪预训练后，F1值降至 71.91%，召回率降至 70.32%，表明情绪预训练至关重要。移除第一阶段分离后，F1值降至 73.51%；移除第二阶段分离后，F1值降至 75.93%，精确率降至 73.95%，凸显了细粒度情绪特征分离的必要性。

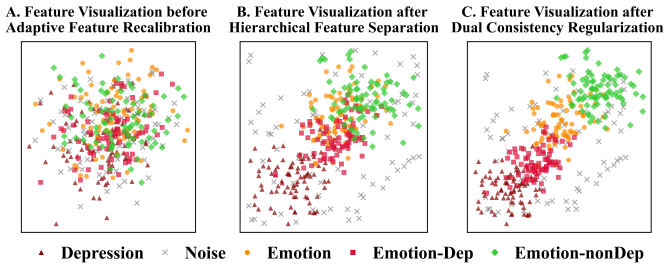

图5. HFS 和 DCR 后的渐进式特征解耦。该可视化展示了 AFR 如何在潜在空间中细化与抑郁症相关的情绪特征和噪声特征。

表 IV LMVD 数据集上针对超参数 λ、μ 和 β 的消融研究。

数值	Acc.	F1
0.1	75.12	74.87
0.5	76.29	76.18
1.0	**77.59**	**77.47**
1.5	76.02	75.89

数值	Acc.	F1
0.05	75.88	75.57
0.1	**77.59**	**77.47**
0.3	76.22	76.03
0.5	75.37	75.14

数值	Acc.	F1
0.3	75.87	75.61
0.5	**77.59**	**77.47**
0.7	76.84	76.72
0.9	75.15	74.93

C. 方程 18 中β的消融研究 A. 方程 5 中λ的消融研究 B. 方程 10 中μ的消融研究

在双重一致性正则化（第 III-D节）中，移除第一阶段图更新后，F1值降至 75.55%，而第二阶段更新的影响较小（76.66%）。移除第二阶段图正则化后，F1值降至 76.19%；移除第一阶段正则化后，F1值进一步降至 74.31%，凸显了其对稳定性的重要作用。

在非对称蒸馏（第 III-E节）中，移除蒸馏学习使 F1 值降至 75.70 %，而移除特征融合则使精确率降至 74.73%，F1 值降至 76.34%。

特征可视化。为提供定性视角，图 5展示了分层特征分离（HFS）和双重一致性正则化（DCR）阶段后的特征分布。如图所示，HFS 能有效分离与抑郁症相关的成分、情绪成分和噪声成分，而 DCR 则进一步细化这些子空间之间的边界。这些可视化结果展现了自适应特征重校准（AFR）解耦并净化潜在表示的能力，为定量消融结果提供了补充。

总体而言，特征解耦和情绪预训练的影响最大，其次是双重一致性正则化，最后是蒸馏与融合模块的组合。这些结果进一步验证了 READ-Net 的架构，尤其凸显了两阶段特征解耦和第二阶段图正则化的重要性。

E. 消融实验

我们分析了超参数 λ、 μ和 β对模型准确率和F1分数的影响。首先， λ （式：7、表IV-A中的特征正则化参数）控制第一层分离中抑郁特征与噪声特征之间的正交性，其中 λ = 1.0时性能最佳，准确率（77.59%）和F1分数（77.47%）均达到最高。其次， μ （式：12、表IV-B中的信息流控制参数）用于平衡不同特征类型之间的相关性，其中 μ = 0.1时效果最佳，并能防止过拟合。最后， β （式：21、表IV-C中的融合权重）控制情绪的贡献

表 V 情感

在我们的集合上比较 READ-Net 与基线最先进模型的情绪歧义测试准确率。

强度	READ-Net	DepMamba	ConvBiLSTM
None	**86.5**	81.2	79.80
Low	**84.3**	78.4	76.5
中等	**82.1**	74.6	72.7
High	**79.2**	69.8	67.3

以及抑郁症特征，而 β = 0.5 提供了最佳平衡，进一步提升了模型性能。总体而言，最优配置为 λ = 1.0、 μ = 0.1 和 β = 0.5。

F. 情绪歧义测试

为评估READ-Net应对情绪模糊性的效果，我们设计了一项实验，通过模拟不同强度的短暂情绪波动，比较其与基线模型DepMamba和ConvBiLSTM的性能，并评估其在复杂情绪场景中的鲁棒性。参照D-vlog数据集的采集方法，我们从YouTube收集了100个视频博客（约17小时），其中包含50段标注为抑郁和50段标注为非抑郁的视频，因为现有抑郁识别数据集缺少可公开获取的原始视频数据。我们使用基于Aff-Wild2的情绪识别模型 [46]从同一视频中提取悲伤片段。每段视频插入三个10秒片段，分别呈现低、中、高强度的悲伤，并通过调整语音和面部表情来控制强度。我们重新提取了多模态特征，以兼容READ-Net。数据按80:20的比例划分为训练集和测试集，实验结果取五次运行的平均值。我们使用准确率进行评估。结果（表V）表明，READ-Net能够减轻悲伤的干扰，在不同强度下保持稳定性能，并优于基线模型。由于缺乏处理模糊性的能力，DepMamba和ConvBiLSTM会将高强度悲伤下的非抑郁样本误分类，导致准确率下降。READ-Net在高强度下的误分类率为20.8%，低于DepMamba（30.2%）和ConvBiLSTM（32.7%），证实了其鲁棒性。

G. 用于验证功能解耦的反事实特征重组

为了验证 READ-Net 是否实现了抑郁表征与功能 情绪表征之间真正的解耦，我们设计了一项基于特征重组的反事实实验。其核心思想是：如果我们在潜在空间中，将抑郁受试者 A 的抑郁核心 与健康受试者 B 的情绪外壳 重新组合，那么一个解耦模型的预测应该仅依据抑郁核心。

我们构建了 100 对样本，每对均包含一名抑郁受试者 A 和一名健康对照 B。使用经过充分训练的 READ-Net 处理两人的视频后，我们从 HFS 模块中提取各自的抑郁和情绪特征。随后形成四种潜在空间组合：原始抑郁（A：核心 = A，外壳 = A）、原始健康（B：核心 = B，外壳 =B）、嵌合体 A（核心 = A，外壳 = B）和嵌合体 B（核心 = B，外壳 = A）。这些组合经过 AFR 和分类器处理，并记录所得的抑郁症概率。

表 VI 特征重组与反事实预测实验结果。该表报告了 READ-Net 接收不同抑郁症核心特征与情绪外壳特征组合时的预测指标。准确率随核心特征来源而变化，这证实了抑郁表征与情绪表征之间的功能解耦。

输入组合	抑郁症核心特征来源	情绪外壳特征来源	准确率	精确率	召回率	F1
原始抑郁 (A)	患者 A	患者 A	86.82	86.48	87.13	86.80
原始健康 (B)	健康者 B	健康者 B	86.28	85.66	86.52	86.09
嵌合体 A	患者 A	健康者 B	79.27	78.84	79.23	79.03
嵌合体 B	健康者 B	患者 A	78.71	78.10	78.65	78.37

如表VI所示，嵌合体的预测结果与核心来源一致：嵌合体 A 的准确率达到 79.27%，嵌合体 B 达到 78.71%。置信度与原始样本略有不同，但最终决策主要由抑郁核心表征 Fd 决定，而重新组合的情绪外壳不会改变病理识别。

这些结果清楚地表明，情绪通道的变化不会干扰抑郁检测通路。因此，READ-Net 同时实现了数学层面和功能层面的解耦，在提供情绪语境的同时不会污染抑郁表征，并有效减少情绪歧义。

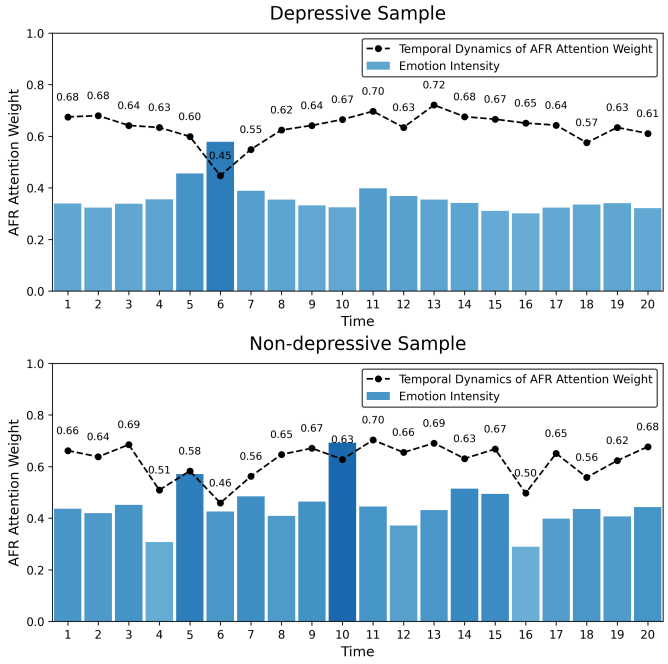

图6. AFR机制的可视化。该图展示了AFR注意力权重（黑色虚线）与检测到的情感强度（蓝色柱形）随时间的变化。上方面板对应抑郁样本，下方面板对应非抑郁样本。AFR会自适应地提高稳定抑郁线索的权重，同时抑制短暂、依赖语境的情绪。

H. 自适应特征重校准（*AFR*）的可视化与定量分析

为考察 AFR 的内部行为，我们可视化其在不同情绪条件下如何调整特征权重。我们分析了两个具有代表性的视频片段：一段来自临床确诊患者的抑郁案例，该患者持续表现出情绪平淡（快感缺失），偶尔会露出社交性微笑；另一段来自健康参与者的非抑郁 案例，表现为回忆一件令人痛苦的往事时，出现了由情境诱发的强烈悲伤。我们追踪非对称蒸馏模块生成的注意力权重 wt，这些权重引导情绪特征与抑郁特征的融合，并将其与帧级情感强度对齐。

如图6所示，AFR展现出明显的自适应性和判别能力。在抑郁样本（上方）中，稳定的情感迟钝使权重始终较高，而短暂的社交性微笑会导致权重骤降，从而抑制非病理性波动。在非抑郁样本（下方）中，尽管出现明显但由情境引发的悲伤，AFR仍维持较低权重。这些模式证实，READ-Net能够区分情境驱动的短暂悲伤与病理驱动的持续低落情感。总体而言，AFR充当情境感知滤波器，动态调节特征显著性以减少情绪歧义，确保抑郁相关线索在最终预测中占主导地位。

I. 通过反向验证保留的身份信息分析噪声通道

READ-Net 中的分层特征分离（HFS）模块引入了一个专门的噪声流，并在正交约束下学习独立于抑郁特征空间 Fd 的表征。我们预期这一被舍弃的通道能够捕获与抑郁病理无关的冗余信息。为验证这一点并更好地理解 HFS，我们设计了一项反向验证实验，考察噪声表征 Fn 中编码了哪些信息。我们的假设是，Fn 主要包含稳定的身份相关线索，例如音色、面部结构和基线说话风格；这些线索在抑郁症检测中起到混杂因素的作用。

为评估这一点，我们开展了一项辅助说话人识别任务。在使用抑郁症检测任务训练 READ-Net 后，我们冻结所有参数，提取每个测试样本的 Fd 和 Fn，并训练两个轻量级多层感知机分类器来预测说话人身份。如图7所示，结果有力地支持了我们的假设。基于 Fn 的分类器在我们的数据集上取得了最高准确率（73.19%），表明 Fn 保留了丰富且稳定、与抑郁症状无关的个性化信息。基于 Fd 的分类器准确率较低（69.44%），说明身份线索已从 Fd 中有效去除。作为对比，我们还测试了 MMD、Xception、视觉变换器和 DepMamba 等代表性骨干网络提取的特征，但它们在说话人识别任务上的表现均不及 Fn。这些发现证实，HFS 模块在 READ-Net 中实现了有效解耦。

表 VII 不同特征/模型的性能指标

特征	准确率	精确率	召回率	F1
F_n	73.19	74.53	73.14	73.82
F_d	69.44	67.70	68.25	67.97
MMD	70.27	71.23	71.18	71.20
Xception	65.31	66.89	64.36	65.60
ViT	66.94	67.46	66.83	67.14
DepMamba	69.31	70.03	70.12	70.07

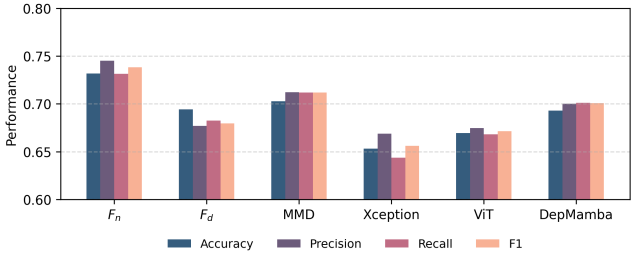

图7. 使用从不同通道和模型中提取的特征进行说话人识别性能对比。准确率越高，表明身份信息保留得越充分。噪声通道 F_n 的准确率最高，证实其能够有效捕获与身份相关但与抑郁症无关的信息。

J. 结论

本文提出了 READ-Net，这是一种用于视听抑郁症检测的新型框架，旨在解决情绪模糊性问题。其核心创新是“自适应特征重校准”，能够保留与抑郁症相关的情绪线索，同时滤除短暂情绪噪声。在三个公开数据集上开展的大量实验表明，该方法显著优于最先进方法。READ-Net 弥补了现有方法在建模短暂情绪与稳定抑郁特质之间差异方面的不足，从而提升了抑郁检测系统的可靠性和泛化能力。其即插即用的设计便于与现有框架集成，可拓展其在心理健康监测和临床支持中的应用。

参考文献

[1] P. C. Kendall、S. D. Hollon、A. T. Beck、C. L. Hammen 和 R. E. Ingram，《关于使用贝克抑郁量表的问题与建议》，《认知治疗与研究》，卷 11，299，1987年 289页。–。[2] L. He 和 C. Cao，《利用卷积神经网络从语音中自动分析抑郁症》，《生物医学信息学杂志》，卷 83，第 103–111 页，2018年。[3] L. Yang、D. Jiang 和 H. Sahli，《融合深层与浅层模型进行多模态抑郁症分析——混合架构》，*IEEE*情感计算汇刊，卷 12，期 1，第 239–253 页，2021年。[4] Y. Tao、M. Yang、H. Li、Y. Wu 和 B. Hu，《Depmstat：用于抑郁症检测的多模态时空注意力变换器》，《*IEEE*知识与数据工程汇刊》，卷 36，期 7，第 2956–2966 页，2024年。[5] M. Liu、Z. Wang 和 F. Liu，《Mddmamba：一种基于 Mamba、采用节省内存的跨模态注意力机制的多模态抑郁症检测模型》，载于*2024*年*IEEE*生物信息学与生物医学国际会议（*BIBM*），2024年，第 1600–1605 页。[6] W. C. de Melo、E. Granger 和 A. Hadid，《用于通过面部动态评估抑郁症的深度多尺度时空网络》，*IEEE*情感计算汇刊，卷 13，1592，2022年 3，1581 页。–。[7] Z. Huang、J. Epps 和 D. Joachim，《抑郁症检测中的语音地标模式研究》，*IEEE*情感计算汇刊，卷 13，期 2，第 666–679 页，2022年。[8] W. Zheng、L. Yan 和 F.-Y. Wang，《一石二鸟：用于抑郁症检测与情绪识别的知识嵌入式时序卷积变换器》，*IEEE*情感计算汇刊，卷 14，期 4，第 2595–2613 页，2023年。

[9] S. Teng、S. Chai、J. Liu、T. Tateyama、L. Lin 和 Y.-W. Chen，《借助情感辅助的多模态多任务抑郁症检测》，载于*2024*年*IEEE*国际消费电子会议（*ICCE*），2024年，第 1–5 页。[10] S. A. Khowaja、L. Nkenyereye、P. Khuwaja、H. Al Hamadi 和 K. Dev，《利用情绪感知编码器和基于模糊逻辑的对比网络检测社交媒体帖子中的抑郁症》，*IEEE*模糊系统汇刊，卷 33，期 1，第 43–53 页，2025年。[11] L. He、K. Chen、J. Zhao、Y. Wang、E. Pei、H. Chen、J. Jiang、S. Zhang、J. Zhang、Z. Wang 等，《Lmvd：用于自然场景下抑郁症检测的大规模多模态视频博客数据集》，*arXiv* 预印本 *arXiv:2407 00024*，2024年。。[12] J. Yoon、C. Kang、S. Kim 和 J. Han，《D-vlog：用于抑郁症检测的多模态视频博客数据集》，*AAAI*人工智能会议论文集，卷 36，期 11，第 12 226–12 234 页，2022年。[13] F. Ringeval、B. Schuller、M. Valstar、J. Gratch、R. Cowie、S. Scherer、S. Mozgai、N. Cummins、M. Schmitt 和 M. Pantic，《Avec 2017：现实生活中的抑郁症与情感识别研讨会及挑战赛》，载于第七届音频/视觉情绪挑战年度研讨会论文集，2017年，第 3–9 页。[14] M. Valstar、B. Schuller、K. Smith、F. Eyben、B. Jiang、S. Bilakhia、S. Schnieder、R. Cowie 和 M. Pantic，《Avec 2013：连续音频/视觉情绪与抑郁症识别挑战赛》，载于第三届*ACM*音频/视觉情绪挑战国际研讨会论文集，2013年，第 3–10 页。[15] M. Valstar、B. Schuller、K. Smith、T. Almaev、F. Eyben、J. Krajewski、R. Cowie 和 M. Pantic，《Avec 2014：三维情感与抑郁症识别挑战赛》，载于第四届音频/视觉情绪挑战国际研讨会论文集，2014年，第 3–10 页。[16] B. Zou、J. Han、Y. Wang、R. Liu、S. Zhao、L. Feng、X. Lyu 和 H. Ma，《面向抑郁障碍自动初筛的半结构化访谈式中文多模态抑郁症语料库》，*IEEE*情感计算汇刊，卷 14，2838，2023年 4，2823 页。–。[17] F. Eyben 和 B. Schuller，《openSMILE：慕尼黑开源大型多媒体特征提取器》，*SIGMultimedia* 召回率，卷 6，期 4，第 4–13 页，2015年。[18] X. Ma、H. Yang、Q. Chen、D. Huang 和 Y. Wang，《Depaudionet：一种用于基于音频的抑郁症分类的高效深度模型》，载于第六届音频/视觉情绪挑战国际研讨会论文集，2016年，第 35–42 页。[19] W.-N. Hsu、B. Bolte、Y.-H. H. Tsai、K. Lakhotia、R. Salakhutdinov 和 A. Mohamed，《HuBERT：通过隐藏单元掩码预测进行自监督语音表征学习》，*IEEE/ACM*音频、语音与语言处理汇刊，卷 29，第 3451–3460 页，2021 年。[20] Y. Zhu 、 Y. Shang 、 Z. Shao 和 G. Guo，《基于深度网络编码面部外观与动态的自动抑郁症诊断》，*IEEE*情感计算汇刊，卷 9，584，2018年 4，578页。[21] S. Zhang、X. Zhang、X. Zhao、J. Fang、M. Niu、Z. Zhao、J. Yu 和 Q. Tian，《Mtdan：用于自动视频抑郁症检测的轻量级多尺度时序差分注意力网络》，*IEEE*情感计算汇刊，卷 15，期 3，第 1078–1089 页，2024年。[22] W. Lin、I. Orton、Q. Li、G. Pavarini 和 M. Mahmoud，《观察身体：心理困扰中身体动作与自适应动作的自动分析》，*IEEE*情感计算汇刊，卷 14，期 2，第 1175–1187 页，2023年。[23] Z. Zhou、Y. Guo、S. Hao 和 R. Hong，《通过音频响应级建模进行分层多特征融合以检测抑郁症》，*IEEE*计算社会系统汇刊，卷 10，期 5，第 2797–2805 页，2023年。[24] M. Hu、L. Liu、X. Wang、Y. Tang、J. Yang 和 N. An，《用于音视频自动抑郁症评估的并行多尺度桥接融合网络》，*IEEE*计算社会系统汇刊，卷 11，期 5，第 6830–6842 页，2024年。[25] J. Ye、Y. Yu、L. Lu、H. Wang、Y. Zheng、Y. Liu 和 Q. Wang，《Dep-fer：基于情绪变化、利用面部表情和音频特征进行多模态抑郁症识别》，*IEEE*视频技术电路与系统汇刊，卷 35，期 3，第 2087–2100 页，2025年。[26] D. Bose、V. Sethu 和 E. Ambikairajah，《连续情绪模糊性预测：基于贝塔分布的建模》，*IEEE*情感计算汇刊，卷 15，期 3，第 1684–1695 页，2024年。[27] J. Ye、Y. Yu、Y. Zheng、Y. Liu 和 Q. Wang，《Dep-fer：基于自主面部表情模仿的抑郁症患者面部表情识别》，*IEEE*情感计算汇刊，卷 15，期 3，第 1725–1738 页

，2024年。

[28] N. K. Iyortsuun、S.-H. Kim、H.-J. Yang、S.-W. Kim 和 M. Jhon，“基于音频和文本特征的抑郁症检测加性跨模态注意力网络（acma）”，*IEEE Access*，卷 12，页 20 479– 20 489，2024年。[29] Z. Zhao、Z. Bao、Z. Zhang、J. Deng、N. Cummins、H. Wang、J. Tao 和 B. Schuller，“通过层次注意力迁移网络和注意力自编码器自动评估语音中的抑郁症”，*IEEE*信号处理精选主题期刊，卷 14，期 2，页 423–434，2020年。[30] L. Zhou、Z. Liu、Z. Shangguan、X. Yuan、Y. Li 和 B. Hu，“TAMFN：用于抑郁症检测的时间感知注意力多模态融合网络”，*IEEE*神经系统与康复工程汇刊，卷 31，页 669–679，2023年。[31] L. Zhou、Z. Liu、Z. Shangguan、X. Yuan、Y. Li、B. Hu，“JAMFN：用于抑郁症检测的联合注意力多尺度融合网络”，载于 *Interspeech 2023* 会议论文集，2023年，页 3417–3421。[32] J. Ye、J. Zhang 和 H. Shan，“DepMamba：用于多模态抑郁检测的渐进式融合 Mamba”，载于 *IEEE*国际声学、语音与信号处理会议，2025年，页 1–5。[33] J. Cheong、S. Kalkan 和 H. Gunes，“FAIRREFUSE：裁判引导融合在抑郁症检测中的多模态因果公平性”，载于 第三十三届国际人工智能联合会议论文集，2024年。[34] Y. Liu、X. Zhou、H. Kou、Y. Zhao、X. Xu、X. Zhang 和 L. Qi，“面向地质旅行的基于简化图卷积网络的隐私保护兴趣点推荐”，*ACM*智能系统与技术汇刊，卷 15，期 4，页 1–17，2024年。[35] M. Niu、Z. Zhao、J. Tao、Y. Li 和 B. W. Schuller，“用于自动抑郁程度预测的双重注意力和元素重校准网络”，*IEEE*情感计算汇刊，卷 14，期 3，页 1954–1965，2023年。[36] Y. Liu、H. Wu、K. Rezaee、M. R. Khosravi、O. I. Khalaf、A. A. Khan、D. Ramesh 和 L. Qi，“面向旅行企业连续兴趣点推荐的交互增强型时间感知图卷积网络”，*IEEE*工业信息学汇刊，卷 19，期 1，页 635–643，2022年。[37] Y.-I. Tian、T. Kanade 和 J. Cohn，“面部表情分析中的动作单元识别”，*IEEE*模式分析与机器智能汇刊，卷 23，期 2，页 97–115，2001年。[38] F. Chollet，“Xception：采用深度可分离卷积的深度学习”，载于*2017*年 *IEEE*计算机视觉与模式识别会议（*CVPR*），2017年，页 1800–1807。[39] J. Hu、L. Shen 和 G. Sun，“压缩激励网络”，载于*IEEE*计算机视觉与模式识别会议论文集（*CVPR*），2018年6月。[40] A. Dosovitskiy，“一张图像胜过 16×16 个词：面向大规模图像识别的 Transformer”，*arXiv*预印本 *arXiv:2010.11929*，2020年。[41] Y. Tao、M. Yang、Y. Wu、K. Lee、A. Kline 和 B. Hu，“通过视频博客面部与语音流的时空 Transformer 感知抑郁语义”，《数字通信与网络》，卷 10，期 3，页 577–585，2024年。[42] Z. Shangguan、X. Li、Y. Dong 和 X. Yuan，“使用基于注意力的深度多实例学习进行自动抑郁症检测”，载于异构网络质量、可靠性、安全性与鲁棒性国际会议，2023年，页 40–51。[43] A. Jaegle、F. Gimeno、A. Brock、O. Vinyals、A. Zisserman 和 J. Carreira，“Perceiver：通过迭代注意力实现通用感知”，载于第*38*届国际机器学习会议论文集，卷 139，2021年，页 4651–4664。[44] D. Gimeno-Gómez、A.-M. Bucur、A. Cosma、C.-D. Martínez-Hinarejos 和 P. Rosso，“读懂帧间信息：从视频中的非语言线索进行多模态抑郁检测”，载于欧洲信息检索会议，2024年，页 191–209。[45] J. Jung、C. Kang、J. Yoon、S. Kim 和 J. Han，“HiQuE：用于多模态抑郁检测的层次问题嵌入网络”，载于第*33*届 *ACM*国际信息与知识管理会议论文集，2024年，页 1049–1059。[46] D. Kollias 和 S. Zafeiriou，“Aff-Wild2：扩展用于情感识别的 Aff-Wild数据库。arXiv 2018”，*arXiv*预印本 *arXiv:1811.07770*，卷 2，2018年。