DOI: 10.13382/j.jemi.B2609228

基于多尺度时序表征与掩码自适应融合的个体心理异常状态检测*

田中雨 姜 囡

(中国刑事警察学院公安信息技术与情报学院 沈阳 110854)

摘 要:在公共安全场景中,个体极端行为往往具有突发性、危害性和较强的不确定性,在非侵入条件下对相关异常状态进行及时识别,对风险预警和前置干预具有重要意义。针对现有研究中专用数据不足、单一模态表征有限、长短时动态难以统一建模以及模态缺失和质量波动易导致融合失稳等问题,提出一种基于多尺度时序表征与掩码自适应融合的个体心理异常状态检测方法。首先设计了基于公开心理健康数据预训练到自建数据微调的迁移学习方案,利用可迁移的心理状态表征知识改善模型初始化质量,缓解小样本条件下训练不足与过拟合问题,增强模型对目标任务的适应能力。在模型设计上,结合视频、音频与远程光电容积脉搏波描记法(remote photoplethysmography,rPPG)多源信息,针对心理异常状态短时波动与长程演化并存的特征,提出多尺度时序表征方法,以提升对跨模态动态线索的协同建模能力;进一步针对真实场景下模态缺失、信号质量波动及模态贡献不均衡等问题,提出掩码自适应融合机制,以增强模型在复杂条件下的稳定性与鲁棒性。实验结果表明,所提方法在 AVEC2014 数据集上平均绝对误差(MAE)为 5.14、均方根误差(RMSE)为 6.37,在自建数据集上准确率(Acc)为 0.76、F1 为 0.82,验证了所提方法的有效性。

关键词:心理异常状态检测;掩码自适应融合;多尺度时序表征;迁移学习

中图分类号:TN911.7 **文献标识码**:A **国家标准学科分类代码**:510.4010

Individual psychological abnormal state detection via multi-scale temporal representation and mask-adaptive fusion

Tian Zhongyu Jiang Nan

(College of Public Security Information Technology and Intelligence, Criminal Investigation Police University of China, Shenyang 110854, China)

Abstract: In public safety scenarios, extreme individual behaviors are often sudden, harmful, and highly uncertain. Timely identification of relevant abnormal states in a non-intrusive manner is therefore of great significance for risk warning and early intervention. To address the limitations of existing studies, including the lack of dedicated datasets, insufficient unimodal representations, difficulty in jointly modeling short- and long-term dynamics, and unstable fusion caused by missing modalities and signal quality fluctuations, this paper proposes an individual psychological abnormal state detection method based on multi-scale temporal representation and mask-adaptive fusion. First, a transfer learning scheme is designed by pretraining on a public mental health dataset and fine-tuning on a self-built dataset, so as to leverage transferable psychological state knowledge, improve model initialization, alleviate undertraining and overfitting under small-sample conditions, and enhance adaptation to the target task. In model design, multi-source information from video, audio, and rPPG is integrated, and a multi-scale temporal representation method is introduced to better capture the coexistence of short-term fluctuations and long-term evolution in psychological abnormal states, thereby improving collaborative modeling of cross-modal dynamic cues. Furthermore, a mask-adaptive fusion mechanism is proposed to address modality absence, signal quality variation, and unequal modality contributions in real-world scenarios, thus improving model stability and robustness under complex conditions. Experimental results show that the proposed method achieves an MAE of 5.14 and an RMSE of 6.37 on the AVEC2014 dataset, and an accuracy of 0.76 and an F1-score of 0.82 on the self-built dataset, demonstrating its effectiveness.

Keywords: psychological abnormal state detection; mask-adaptive fusion; multi-scale temporal representation; transfer learning

收稿日期:2026-03-28 Received Date:2026-03-28

*基金项目:中央引导地方科技资金计划项目(2025090005JH61004)、国家重点研发计划项目(2017YFC0821005)、辽宁省研究生教育教学改革研究项目(LNYJG2024310)、中国刑事警察学院研究生创新能力提升项目(2025YCZD04)资助

0 引 言

在公共安全治理不断强调风险前移和源头防控的背景下,个体极端行为预警已成为公共安全研究中的重要议题。此类事件具有突发性强、危害后果重和社会影响广等特征;但从既有威胁评估与针对性暴力研究来看,相关行为并非完全无征兆,而是常伴随一系列先兆性异常表现,如情绪失衡、认知偏差、社会联系弱化、言语异常以及行为控制减弱等[1-2]。这类异常信号一旦能够在早期阶段被及时捕捉,便有可能为风险识别、分级研判和干预处置争取必要时间窗口。因此,如何面向真实场景,在低干扰、非侵入条件下对个体心理异常状态开展连续、低干扰的自动化识别,不仅关系到极端事件早期发现能力的提升,也关系到公共安全治理模式由事后处置向事前预警转变的技术支撑水平[1-2]。

面向公共安全场景的个体心理异常状态检测面临较大挑战。心理异常状态本身具有较强的隐蔽性、阶段性和个体差异性,其外在表达通常不是由某一单一情绪、单次行为或孤立线索所决定,而更表现为多类细微信号在具体情境中的综合呈现。传统依赖人工观察、经验判断或量表筛查的方式,往往存在主观性较强、连续性不足和标准一致性有限等问题;而接触式生理测量虽然能够提供更直接的内部状态信息,却又在开放环境、自然交互和持续感知等应用场景中面临部署成本高、干扰性强和适用性受限等不足。Quiram 等[3]在 2025 年的综述中指出,远程心理健康筛查虽然显示出可行性,但在技术限制、用户因素、伦理法律约束以及实际场景整合等方面仍存在明显障碍;Garzón-Partida 等[4]综述也表明,数字生物标志物在情绪障碍识别与监测中具有潜力,但其临床转化和真实环境应用仍需进一步完善。因此仅凭单一外显行为难以作出稳定判断,更需要构建面向自然场景、能够持续感知并综合多源线索的低干扰检测方法。

从可观测信息来源看,面部表情、头部姿态与局部动作变化能够反映个体的外显行为特征,语音韵律、停顿模式与发声状态能够体现情绪表达及交流负荷,而生理节律变化则可在一定程度上补充紧张、激惹和唤醒水平等内在状态信息。Jiang 等[5]基于远程访谈开展的研究表明,联合面部、语音、语言和心血管模式能够更有效揭示心理健康状态相关信息;Wang 等[6]在 2025 年的系统综述与 Meta 分析中也指出,AI 辅助的多模态生理与行为信息在抑郁筛查中的总体表现优于单模态方法。与此同时,多模态情感识别与心理状态分析领域的最新研究认为,融合视觉、语音、文本和生理信号,较单一模态更有助于形成完整、稳健的状态刻画[7-8]。心理异常状态并非静态表征,而是伴随短时波动、局部突变和阶段性演化的动态过程。Tao 等[9]在面向异常情绪状态识别的研究中指出,时空信息在多模态心理状态分析中具有重要作用,若忽略时序动态,仅依赖静态拼接或单尺度建模,往往难以充分捕捉异常状态的复杂演变规律。因此,针对该任务还需要在时序表征层面进一步强化对短时与长程动态的联合建模。

从近年的相关研究看,现有多模态心理状态分析方法已在抑郁评估、情绪识别和远程心理健康监测等方向取得积极进展,但若直接服务于公共安全语境下的个体心理异常状态检测,仍存在进一步完善的空间。首先,现有公开研究的任务设置与数据资源更多围绕临床辅助分析、通用情感计算或线上心理状态识别展开,而与公共安全预警场景高度贴合、兼顾自然互动情境与异常状态表达的数据基础仍相对薄弱,这在一定程度上限制了模型的针对性建模与实际迁移能力[1,5,9]。其次,已有方法虽然普遍重视多模态信息融合,但对于心理异常状态短时波动与长程演化并存的特点,仍缺乏更加细致的多尺度时序刻画;相关系统综述也指出,当前多模态情感识别仍面临跨场景泛化不足、时序动态建模不充分和复杂场景适配性有限等问题[7-8]。此外,在真实应用环境中,遮挡、噪声、失帧、采集失败以及隐私约束等因素广泛存在,模态缺失、信号质量波动和模态贡献不均衡往往会直接影响融合稳定性。Zhan 等[10]在关于不完整多模态学习的系统综述中指出,传感器限制、成本约束、隐私因素和数据丢失,是多模态模型在训练和测试阶段出现缺失模态问题的主要原因,而如何在不完整输入条件下保持模型鲁棒性,已成为多模态学习的重要研究方向。基于此,本文面向公共安全风险预警任务,构建兼顾多尺度时序建模与缺失模态鲁棒融合能力的个体心理异常状态检测方法。本文创新点如下:

1)面向公共安全风险预警中的个体心理异常状态检测需求,构建多模态短视频数据集并明确任务定义;

2)针对心理异常状态短时波动与长程演化并存的特点,设计多尺度时序表征模型,实现视频、音频与 rPPG 线索的协同建模;

3)针对真实场景下模态缺失、信号质量波动及模态贡献不均衡等问题,提出了多动态多区域(region of interest,ROI)双通道质量感知融合方法进行心率估计;在多模态融合方面,提出掩码自适应融合方法,提高多模态判别的鲁棒性,增强模型对复杂场景的适应能力;

4)针对目标域样本规模有限问题,构建公开数据预训练—自建数据微调的迁移学习方案,提升模型在目标任务上的适应能力。

1 数据集构建与预处理

为兼顾模型通用表征学习能力与目标任务适配能

力,采用“公开数据集预训练—自建数据集微调”的两阶段数据使用方案。首先利用 AVEC2014 数据集对视频、音频与远程光电容积脉搏波描记法(remote photoplethysmography,rPPG)三模态分支进行预训练,使模型能够在较大规模、标注相对规范的公开数据上学习稳定的多模态时序表征;随后在自建心理异常状态检测数据集上进行微调,使模型由通用心理状态表征进一步迁移到公共安全语境下的心理异常状态检测任务。

1.1　心理异常状态检测数据集构建

为支撑心理异常状态检测研究,构建了一个面向自然交互场景的多模态短视频数据集。

1)心理异常定义

在标注规则上,本研究对心理异常状态的定义,参考了 Nevid 等[11]提出的异常行为评估框架。当个体的行为或状态符合以下多项标准的组合时,即可被界定为异常:(1)统计学上的不寻常性或罕见性;(2)违反社会规范或为社会所不容;(3)对现实存在错误的感知或认知解释偏差;(4)伴随显著的个人负性痛苦体验;(5)表现出适应不良性或自我挫败性;(6)具有潜在的危险性。在实际二分类标注过程中,两名标注者以此定义为依据进行标注。重点考察片段中的目标人物是否表现出上述 6 项指标的综合线索交织(例如在影视高压冲突中表现出极度的现实认知偏差、严重的负性情绪痛苦以及潜在的危险行为倾向),而非依据单一的瞬间情绪或孤立动作作出片面判断。标注时“异常”并非医学诊断意义上的精神障碍类别,而是指心理异常状态线索组合;“正常”则主要指不满足上述组合判据的日常或相对稳定互动状态。此外,同时记录片段对应的细粒度语义标签(如冲动、受害经历等),作为该异常状态组合的具体表型信息。

2)数据集构建

数据来源上,源于中英文影视素材,片源覆盖电视剧、剧集与电影等多种类型,选取能够体现自然对话、冲突升级与高压互动的片段,保证样本既包含较丰富的行为与情绪线索,又能够满足视频、音频和 rPPG 的同步处理需求。需要说明的是,在真实的公共安全突发场景中,个体极端行为往往伴随着强烈的情绪爆发、冲突对抗和高压互动(属于高唤醒度、强张力状态)。影视剧素材因其丰富的情境冲突,在情绪唤醒度和行为对抗性上,与公共安全防范中需要拦截的突发行为先兆具有高度的情境相似性。因此,本研究将该自建数据集定位为公共安全预警任务下的高仿真替代域,用于验证模型在复杂对抗情境下的异构线索整合能力。

样本组织上,数据集共包含 126 段视频片段,每段时长约为 5~20 s,其中正常样本 50 段、异常样本 76 段。为减小身份特征对分类结果的干扰,围绕 16 名能够明确对应到固定人物或角色的个体构建样本,每名人物均尽量同时包含典型片段与异常片段,能够形成基本的人物内对照结构。

1.2　数据预处理

为使公开数据预训练与自建数据微调采用一致的输入形式,对视频、音频与 rPPG 三模态信号进行统一预处理,并以时间锚点为基础建立跨模态对齐关系。为解决 AVEC2014 数据集样本数量有限且视频时长差异显著的问题,采用固定时间窗口(20 s)与滑动步长(15 s)进行分段处理,以扩充训练样本并保证片段级输入在时间维度上的可比性;在自建心理异常状态数据集中,沿用同一处理思想,将原始片段映射为固定长度或统一组织形式的时序样本,以便后续 3 个模态分支在同一时间尺度上完成编码与融合。

1)视频预处理

在视频模态预处理中,首先对原始视频进行均匀采样,获得时间分布相对稳定的帧序列;随后利用人脸检测与关键点定位方法提取人脸区域,并完成裁剪与对齐,以减弱头部姿态变化、拍摄角度差异及背景噪声对表征学习的干扰。经处理后的视频序列能够更加突出面部运动、局部行为变化与细微表情线索,为后续短 clip 时空编码和多尺度时序建模提供规范化输入[12]。

2)音频预处理

在音频模态预处理中,首先将原始音频统一转换为单通道信号并重采样至 16 kHz;其后并行构建两类输入:一方面提取 Log-Mel 语谱图以捕捉语音的低层时频纹理特征;另一方面,利用预训练的 Wav2Vec 2.0 模型提取高层自监督学习(semi-supervised learning,SSL)特征,以捕捉更加鲁棒的副语言学信息[13]。同时保留音频时间序列索引,与视频片段及 rPPG 序列在同一时间锚点上完成对齐,从而保证多模态输入在片段级具有一致的时间对应关系。

3)rPPG 模态预处理

针对 POS(plane-orthogonal-to-skin)算法在开放交互环境下易受面部表情等非刚性运动形变干扰,以及 OMIT 算法在低噪平稳情境下显色特征增益受限、易导致信号质量剧烈波动的问题,本文提出了多动态多 ROI 双通道质量感知融合方法进行心率估计,解决了单通道提取时由局部肌肉形变、运动伪影重叠以及视频高压缩率块效应引发的生理信噪比下滑问题[14-16]。

对人脸检测裁剪后的连续帧序列进行基于增强相关系数(enhanced correlation coefficient,ECC)的图像帧间对齐以平滑动作位移,有效抑制由于呼吸或语言沟通带来的轻微位移仿射失真。随后,引入肤色掩膜排除非皮肤背景像素,并对人脸网格进行密集空间解耦划分。系统计算每个子区域绿色通道时间序列的频带能量、主峰锐度与局部稳定性指标并进行综合打分,筛选出评分最高

的 N 个最佳动态 ROI 集合(如额头),经形态学平滑与合并后构建为最终的空间动态 ROI 集合 $\{R_k\}$。对于集合内的每个动态 ROI,计算其时间序列上的空间均值 RGB 信号。应用两类无监督提取算法,在每个区域上并行重构出双通道风格互补的时序视图集合 $\{s_{posk}, s_{omit_k}\}$。最后,对信号统一进行一维去趋势核心心率频带带通滤波与标准正态化处理,消除生理基线漂移。

2 多模态特征提取与心理异常检测模型构建

2.1 模型总体框架

提出了一种多模态心理异常状态识别模型(multi-scale dilated encoding and cross- attentive fusion network, MDCA-Net),其框架如图 1 所示。首先将视频、音频与 rPPG 等多源数据按时间顺序构建为多个视图样本,并通过严格对齐与掩码机制处理模态缺失,保证不同模态在片段级别具有一致的组织形式与可比性。在特征学习阶段,模型的核心在于引入多尺度膨胀时序编码器:视频分支以短 clip 为基本单元,利用 R(2+1)D 编码器提取局部动态与微表情变化,再以多尺度膨胀 TCN 在片段序列上建模从短时波动到长程行为模式的动态演化,并采用 PMA(pooling by multihead attention)时间池化获得样本级表示;音频分支对数梅尔频谱与自监督表征融合,在频谱流后端引入 MS-TCN 与 PMA,并结合自注意力和门控融合形成稳健的样本级音频表示;rPPG 分支则采用多尺度膨胀时序卷积编码脉搏节律变化,并结合时域 PMA 与频域 PSD 统计构建样本级生理表示。

在融合与预测阶段,MDCA-Net 将各模态样本级表示投影到统一维度空间,采用带掩码的自适应加权机制完成视图内模态级融合;随后,对同一样本的多个视图融合表示通过 PMA 进行样本内聚合,得到融合表征并输出样本级心理异常状态判别结果。对于同一被试的多个样本预测,进一步采用中位数聚合得到最终人级 心理异常状态判别结果。为提升小样本场景下的优化稳定性,训练时采用两阶段策略:先冻结各模态分支,仅训练融合头与目标域判别头;再逐步解冻分支并进行端到端微调,以获得更一致的跨模态表征空间与更强的泛化能力。

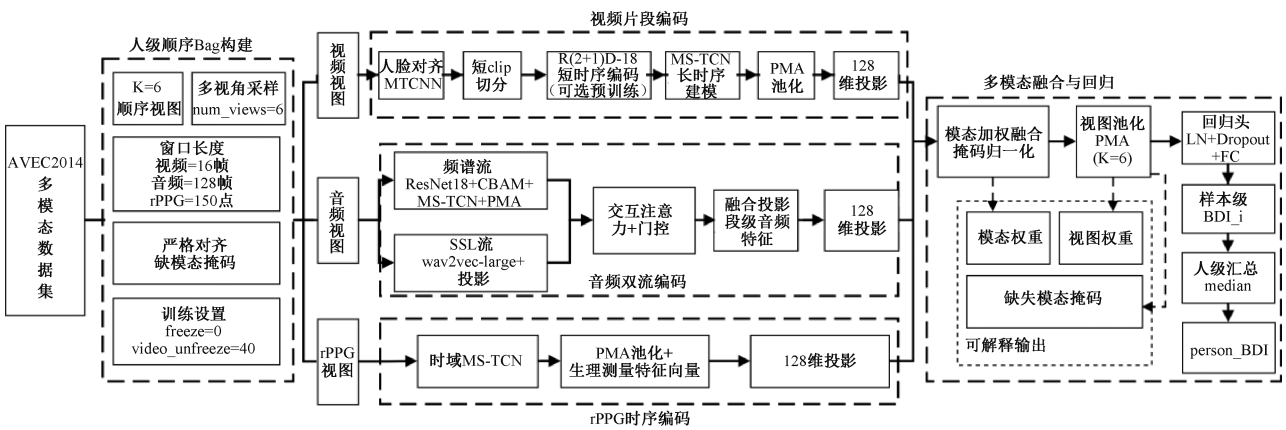


图 1 MDCA-Net 模型框架

Fig. 1 Framework of the MDCA-Net model

2.2 视频模态特征提取

心理异常状态在公共安全语境下往往伴随表情失衡、目光回避、面部紧张和局部动作异常等动态线索,且具有短时波动与阶段性演化并存的特点。传统基于视频模态的心理状态建模有以下问题:1)依赖光流或局部纹理等手工描述时,易受噪声、姿态变化与光照扰动影响,难以稳定刻画细微面部动态;2)直接采用重型三维卷积或大规模时空网络时参数量与训练难度较高,在 AVEC2014 这类样本规模有限的数据上容易出现过拟合[17]。针对上述不足,在视频分支构建了短 clip 时空表征多尺度膨胀时序编码的视频特征提取方法如图 2 所示。将视图序列按滑窗切分为多个短 clip,以 R(2+1)D-18 提取局部时空动态特征并形成 clip 级时序序列;随后通过轻量的多尺度膨胀 TCN 增强长程依赖建模能力;通过多头注意力机制对时序特征进行注意力聚合,形成视频模态样本级表示。该方法能够同时建模短时间尺度下的面部微动态与较长时间尺度下的行为变化,从而提升视频模态对异常状态的表征能力,又能通过分解卷积与预训练权重提升小样本场景下的优化稳定性。

在时空特征提取阶段,选用 R(2+1)D-18 视频编码器。R(2+1)D 通过将三维卷积因子化为“空间卷积+时间卷积”,在控制复杂度的同时强化对局部动态线索的刻画能力,并可引入 Kinetics 预训练权重以提升泛化性能。设一段视频视图包含 T 帧,批量大小为 B,单帧大小为

$H \times W$。将视频帧序列按长度为 L、步长为 S 的滑窗切分为 N 个短 clip，则第 n 短 clip 记为 $c_n \in \mathbf{R}^{L\times3\times H\times W}$，其中：

$$N = \lfloor \frac{T-L}{S} \rfloor + 1 \tag{1}$$

对每个短 clip 进行时空编码，并对其时空特征图进行全局池化得到 clip 级语义向量：

$$e_n = \mathrm{GAP}(\Phi_{R(2+1)D}(c_n)), e_n \in \mathbf{R}^{512}, n = 1,\cdots,V \tag{2}$$

将所有 clip 级特征按时间顺序堆叠得到 $\boldsymbol{E} = [\boldsymbol{e}_1, \cdots, \boldsymbol{e}_N] \in \boldsymbol{R}^{B\times N\times 512}$。为降低后续时序编码的计算开销并统一通道维度，使用线性映射将其投影到 256 维时序空间：

$$\boldsymbol{X} = \boldsymbol{E}\boldsymbol{W}_p + \boldsymbol{b}_p, \quad \boldsymbol{X} \in \mathbf{R}^{B\times N\times 256} \tag{3}$$

其中，$\boldsymbol{X}$ 作为 MS-TCN 的输入序列。

在时序编码阶段，采用多尺度膨胀 TCN 编码器以捕捉不同时间跨度的动态变化。编码器堆叠 3 个 Block，膨胀率集合为 $\{1,2,4\}$，在不显著增加参数的前提下扩大感受野并建模长程依赖。每个 Block 内部由“多分支卷积—拼接融合—门控抑制冗余—残差稳定训练”构成：设输入为 $\boldsymbol{X} \in \mathbf{R}^{B\times C\times N}$，多分支卷积得到 $\{U_i\}$，通道拼接为 $\mathrm{Concat}(\boldsymbol{U}_1,\cdots,\boldsymbol{U}_m)$，再经 1×1 卷积投影到 2C 通道并分解为 $\boldsymbol{A}$、$\boldsymbol{B}$，用 GLU 完成门控：

$$\boldsymbol{Y} = \boldsymbol{A} \odot \sigma(\boldsymbol{B}) \tag{4}$$

$$\boldsymbol{H} = \mathrm{LN}(\boldsymbol{Y} + \boldsymbol{R}(\boldsymbol{X})) \tag{5}$$

其中，$\boldsymbol{R}(*)$ 为残差分支的通道映射；LN 表示层归一化，该结构在保留关键信息的同时抑制冗余响应，并通过残差与归一化提升训练稳定性。经过 3 个 Block 后输出序列特征 $\boldsymbol{h} \in \mathbf{R}^{B\times N\times 256}$。

由于异常线索往往只在少量关键时刻显著（如短暂的表情迟滞或眼神回避），简单的时间均值会稀释有效信息。为此，采用 PMA 多头注意力池化对 $\boldsymbol{H}$ 进行自适应聚合。设可学习种子向量为 $\boldsymbol{s}$，则样本级表示可表示为：

$$\boldsymbol{f} = \mathrm{PMA}(\boldsymbol{H}) = \mathrm{MHA}(\boldsymbol{s},\boldsymbol{H},\boldsymbol{H}), \quad \boldsymbol{f} \in \mathbf{R}^{256} \tag{6}$$

其中，MHA 表示多头注意力操作，对样本级表示 $\boldsymbol{f}$ 采用 Dropout 进行正则化，在源域预训练阶段，将其输入目标域判别头以学习通用心理状态表征；在目标域检测阶段，则作为视频模态特征参与后续心理异常状态判别。

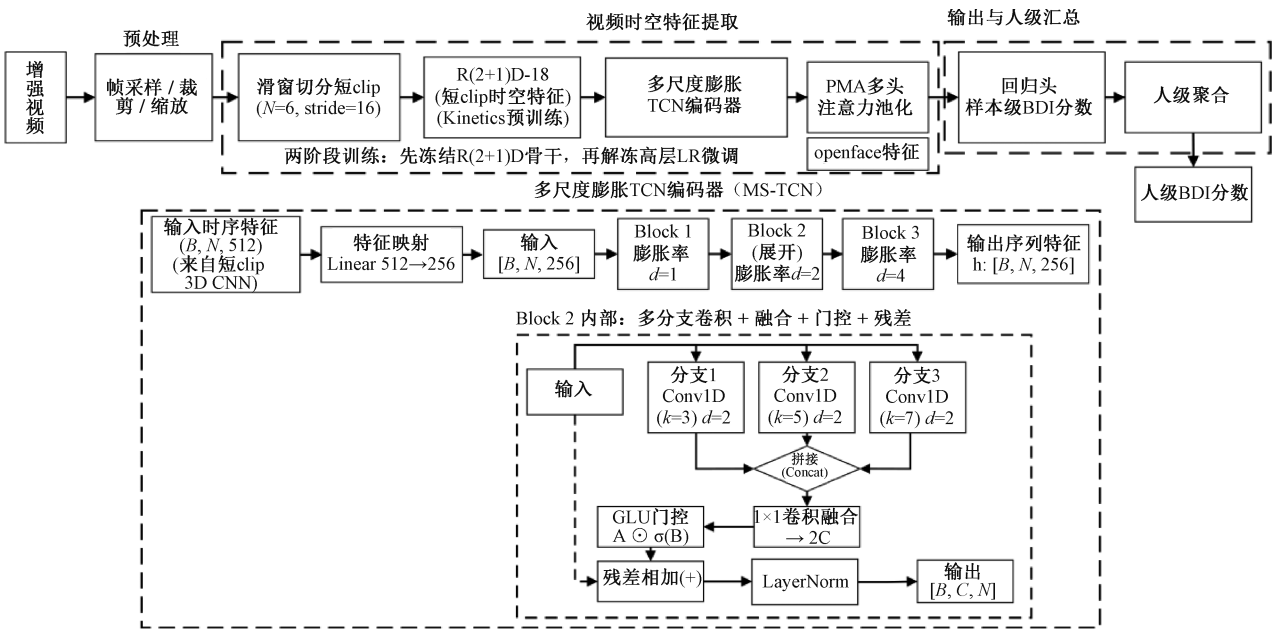


图 2　视频特征提取流程

Fig. 2　Video feature extraction process

2.3　音频模态特征提取

心理异常状态不仅体现在面部行为层面，也可能通过语速变化、能量波动和韵律失衡等语言学特征体现出来。传统基于音频的心理状态建模有以下问题：一方面，依赖能量、共振峰等手工声学描述时，特征设计受经验限制且跨说话人泛化较弱；另一方面，仅使用单一路径的深度特征时，容易偏向局部频谱纹理或忽略更高层语义线索，导致对心理异常相关韵律与语义异常的刻画不充分[18]。因此，本文提出双流声学表征学习框架如图 3 所示，在不使用手工特征的前提下，联合建模对数梅尔频谱与自监督语音表征，并通过自注意力机制和门控融合实现互补信息聚合，得到稳健的段级音频表示用于后续多模态融合与状态预测。

在频谱流中，首先对音频片段进行时间切窗，并将其转换为对数梅尔频谱及其一阶差分、二阶差分构成的三通道张量；引入带注意力增强的 ResNet18+CBAM 网络提

升对关键频带与时频区域的聚焦能力。与原有单纯频谱主干不同,进一步在频谱流后端引入多尺度膨胀时序卷积编码器(MS-TCN)进行时序建模,并采用 PMA 多头注意力池化对时序特征进行自适应聚合,最后再使用线性映射将其投影到共享嵌入维度,形成频谱表征。该过程可写为:

$$\boldsymbol{h}_{spec} = \mathrm{PMA}(\mathrm{MS-TCN}(\mathrm{Backbone}_{spec}(\boldsymbol{S}))) \tag{7}$$

$$\boldsymbol{z}_{spec} = \boldsymbol{W}_s \boldsymbol{h}_{spec} + \boldsymbol{b}_s \tag{8}$$

式中:$\boldsymbol{S}$ 为对数梅尔频谱及其差分特征构成的三通道输入;$\mathrm{Backbone}_{spec}(\cdot)$ 表示频谱主干网络;MS-TCN(·)表示频谱流的多尺度时序编码器;PMA(·)表示多头注意力池化 $\boldsymbol{W}_s$ 和 $\boldsymbol{b}_s$ 分别表示线性投影参数。

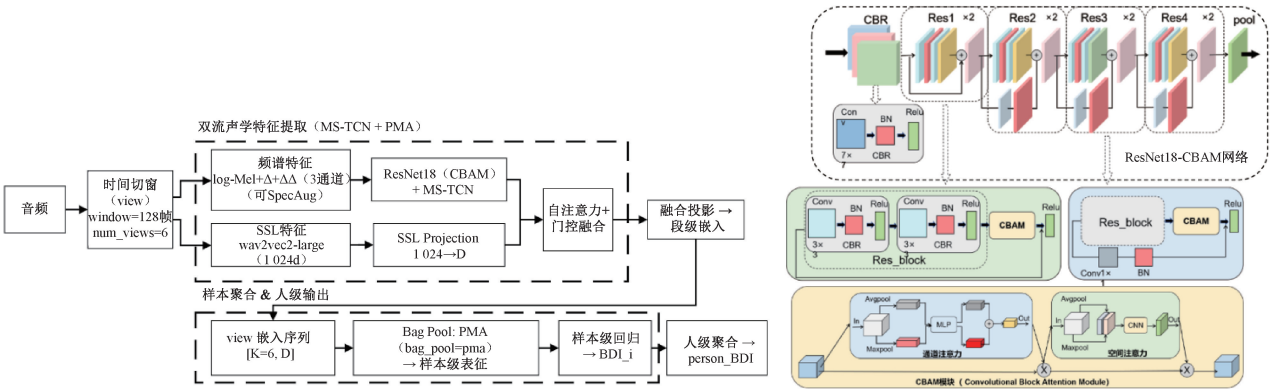


图 3 音频特征提取流程

Fig. 3 Audio feature extraction process

在自监督语音流中,以离线提取的自监督语音模型(wav2vec2-large)特征向量作为输入,并通过含批归一化与非线性的感知机映射到共享维度,得到 $\boldsymbol{z}_{ssl}$。为充分利用两路信息的互补性,构建自注意力机制和门控融合模块:经多头自注意力与前馈网络,并配合残差结构与层归一化得到更新后的特征表示;随后由门控子网络为每个特征生成标量重要性并进行归一化加权求和,得到声学综合表征:

$$[\tilde{\boldsymbol{z}}_{spec}, \tilde{\boldsymbol{z}}_{ssl}] = \mathrm{LN}(\mathrm{FFN}(\mathrm{MHA}([\boldsymbol{z}_{spec}, \boldsymbol{z}_{ssl}])) + [\boldsymbol{z}_{spec}, \boldsymbol{z}_{ssl}])$$

$$\alpha_m = \frac{\exp(g_m)}{\sum_{j \in \{spec, ssl\}} exp(g_j)}, g_m = \mathrm{MLP}_g(\tilde{\boldsymbol{z}}_m) \tag{9}$$

$$\boldsymbol{z}_a = \alpha_{spec} \tilde{\boldsymbol{z}}_{spec} + \alpha_{ssl} \tilde{\boldsymbol{z}}_{ssl} \tag{10}$$

式中:MHA(·)表示多头自注意力;FFN(·)表示前馈网络;LN(·)表示层归一化;$\mathrm{MLP}_g(\cdot)$ 为小型门控感知机;α_m 为归一化后的权重系数。目的是使模型能够在频谱细粒度线索与自监督语义线索之间自适应分配注意力,抑制冗余并突出关键声学线索。

在段级融合与输出阶段,将 $\boldsymbol{z}_{spec}$、$\boldsymbol{z}_{ssl}$ 与自注意力增强后的融合表示在特征维进行组合,经融合投影网络映射回共享维度,得到最终段级音频表示 $\boldsymbol{z}_a$。最后,将 $\boldsymbol{z}_a$ 输入目标域判别头,输出预测;同时,也可将 $\boldsymbol{z}_a$ 作为音频模态的段级特征提供给后续多模态融合模块使用。

2.4 rPPG 特征提取

在高压、激惹或强烈负性情绪状态下,个体生理唤醒水平可能发生波动,进而反映在远程脉搏节律变化中。传统基于 rPPG 的心理状态建模通常存在不足:首先依赖峰谷统计、频域能量等手工描述时,特征设计受经验限制且对噪声、光照与个体差异敏感;其次,直接采用较重的序列模型时,容易在小样本场景出现过拟合与训练不稳定[19]。为此提出多动态多 ROI 双通道质量感知融合的生理信号测量方法和基于多尺度卷积时序编码的 rPPG 特征提取方法如图 4 所示。

1)多动态 ROI 双通道质量感知融合方法

为彻底消除外显行为伪影及环境光噪引发的偏差干扰,本文设计了一种轻量化的基于自适应质量门控的双通道 rPPG 协同调制融合方法。在空间皮肤区域 $\{R_k\}$ 上获取生理视图 $\{s_{pos_k}, s_{omit_k}\}$,系统实时提取信号的时频域可解释特征向量,显式构建生理信号质量描述符向量 $\boldsymbol{X}(t) \in \mathbf{R}^8$。该向量的表达式为:

$$\boldsymbol{X}(t) = [1, SNR_{pos}, SNR_{omit}, P_{pos}, P_{omit}, r_{abs}, -\Delta f_{peak}, 1 - E_{penalty}]^{\mathrm{T}} \tag{11}$$

式中:SNR_{pos} 和 SNR_{omit} 分别表征两通道信号在心率核心频带内的谱信噪比;P_{pos} 和 P_{omit} 为自相关函数的周期性显著性评分;r_{abs} 表示双通道时域信号的绝对皮尔逊相关系数;Δf_{peak} 表示双通道信号功率谱密度主峰频率的绝对偏差值;$E_{penalty}$ 为局部环境运动与亮度剧烈度惩罚项。

门控融合层利用决策参数向量 $\boldsymbol{\theta} \in \mathbf{R}^8$ 与全局增益系数 g 对特征描述符进行投影组合,并通过 *Sigmoid* 函数动态为双通道分配互补的自适应调制权重如式(12)~(13)所示,最终的去噪重构生理信号如式(14)所示。

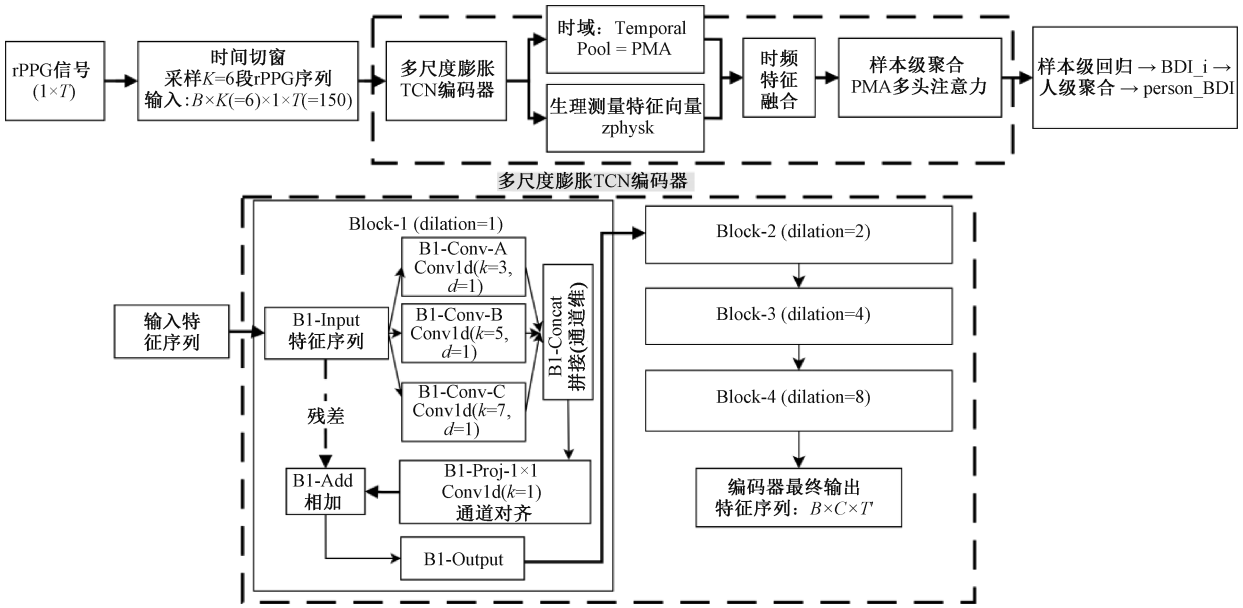


图 4　rPPG 特征提取流程

Fig. 4　rPPG feature extraction process

$$w_{pos} = \frac{1}{1 + e^{-(\sum_{m=1}^{8} \theta_m \times X_m(t)) \times g}} \tag{12}$$

$$w_{omit} = 1 - w_{pos} \tag{13}$$

$$\hat{s} = w_{pos} S_{pos} + w_{omit} S_{omit} \tag{14}$$

2) rPPG 生理测量特征

利用自监督策略输出的生理信号,构建多维生理测量学特征描述向量。从心率变异性(heart rate variability, HRV)时域、HRV 频域以及非线性几何、功率谱分布统计和脉搏波形形态 5 个特征维度构建生理测量特征向量如表 1 所示。

3) rPPG 深度网络分支时序编码

在时序编码器的前端,对每个样本构造 K 段长度为 150 的 rPPG 子序列,输入为 $\boldsymbol{X} \in \mathbf{R}^{B\times K\times 1\times T}$。将一维 rPPG 序列送入多尺度膨胀 TCN 编码器,在保持时序分辨率的同时增强对脉搏波细粒度形态和长程节律变化的建模能力。

在编码器主干层面,构建多尺度膨胀时序卷积块以提升对不同时间跨度模式的表达能力。每个卷积块包含若干并行分支,分支卷积核取{3,5,7},并在同一膨胀率下并行提取多尺度响应;各分支输出在通道维进行拼接,再通过 1×1 卷积投影回目标通道数,并引入残差连接以缓解深层训练退化问题。设输入为 $\boldsymbol{X}$,第 i 个分支输出为 $\boldsymbol{Y}_i$,拼接并融合后有:

$$\boldsymbol{H} = \phi(\mathrm{BN}(\mathrm{Conv}_{1\times 1}(\mathrm{Concat}(\boldsymbol{Y}_1, \boldsymbol{Y}_2, \boldsymbol{Y}_3))) + \boldsymbol{R}(\boldsymbol{X})) \tag{15}$$

式中:$\boldsymbol{R}(\boldsymbol{X})$ 为残差支路;$BN(\cdot)$ 表示批量归一化;$\phi(\cdot)$ 表示非线性激活函数。

为进一步扩大感受野并建模长程节律变化,将上述卷积块堆叠为 4 个阶段,并采用指数型膨胀率逐层扩展时间覆盖范围,膨胀率依次设置为 d = 1,2,4,8。编码器

表 1　rPPG 生理测量特征

Table 1　rPPG physiological measurement features

特征种类	rPPG 特征	特征解释
HRV 时域	PNN50(%);RR 间期范围(ms);SDNN	1. PNN50(%):相邻 RR 间期差值大于 50 ms 的个数占总间期个数的百分比 2. RR 间期范围(ms):通过计算时间窗内相邻搏动间期最大值与最小值的差值 3. SDNN:RR 间期标准差(ms) 其中 RR 序列为连续相邻的搏动间期序列
HRV 频域	LF;HF;总功率;LF/HF 比值	1. LF 功率:在低频分带(0.04~0.15 Hz)进行能量积分 2. HF 功率:在高频分带(0.15~0.40 Hz)进行能量积分 3. 总功率:对全生理频段(0.04~0.40 Hz)进行能量积分 4. LF/HF 比:计算低频功率与高频功率的比值
功率谱分布统计	谱质心;谱熵;软峰值比	1. 谱质心:用于表征功率谱分布的质心位置 2. 谱熵:用以定量度量瞬时频谱的能量分布均匀度与非平稳随机化程度 3. 软峰值比:定义为心率主峰频带能量相对于背景基线噪声能量的集中度比值
非线性几何	庞加莱图 SD2/SD1 比	SD2/SD1 比:以(rr_i,rr_{i+1})空间坐标轴构建二维空间拓扑散点图。沿恒等线的长程变异指标 SD2 与垂直于恒等线的短程变异指标 SD1 之比
脉搏波形形态	脉搏振幅均值	脉搏振幅均值

输出时序特征后,采用 PMA 多头注意力池化,以突出关键时间位置的脉搏动态;

设第 k 段 rPPG 序列经编码器得到的时序特征为 $\mathrm{PMA}(\boldsymbol{H}_k)$,为了建立深层时序拓扑与显式多物理维度指标间的互补关联,将该高阶时序特征与 2.4 中第 2)小节中提取的生理测量特征向量 $z_{phy}s_k$ 进行融合,重构为段级综合时频表征:

$$\boldsymbol{u}_k = \mathrm{Fuse}(\mathrm{PMA}(\boldsymbol{H}_k), z_{phy}s_k) \tag{16}$$

其中,Fuse(·)表示时域与频域特征的融合操作,由此可得到每段子序列对应的表示 $\boldsymbol{u}_k$。

在样本级聚合阶段,将同一样本内 K 段 rPPG 子序列的表示 $\{\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots, \boldsymbol{u}_K\}$ 进一步通过 PMA 多头注意力进行样本级聚合,得到样本级 rPPG 嵌入向量:

$$\boldsymbol{z}_r = \mathrm{PMA}([\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots, \boldsymbol{u}_K]) \tag{17}$$

最后,将 $\boldsymbol{z}_r$ 输入目标域判别头以输出预测;同时,也可将 $\boldsymbol{z}_r$ 作为 rPPG 模态的样本级特征输入后续多模态融合模块完成融合与最终预测。

2.5 掩码自适应融合与状态判别

为提高多模态心理异常状态识别在模态质量不均衡、局部缺失及视图数量不一致条件下的鲁棒性,针对上述问题设计了带掩码的两级自适应融合框架如图 5 所示。首先,在视图内对视频、音频和 rPPG 三模态特征进行动态加权,得到视图级融合表示;其次,在样本内对多个视图的融合表示进一步汇聚,形成样本级融合特征;最后,通过目标域判别头输出样本级 BDI 预测,并对同一被试的多个样本预测采用中位数聚合,得到最终人级 BDI 结果。

设第 k 个片段的三模态段级特征为 $\boldsymbol{z}_k^{(v)}$、$\boldsymbol{z}_k^{(a)}$ 和 $\boldsymbol{z}_k^{(r)}$。首先将三模态投影到共享空间:

$$\tilde{\boldsymbol{z}}_k^{(m)} = \boldsymbol{W}_m \boldsymbol{z}_k^{(m)} + \boldsymbol{b}_m, \quad m \in \{v, a, r\} \tag{18}$$

然后通过轻量评分网络结合模态掩码计算权重,并完成视图内加权融合:

$$\alpha_k^{(m)} = \frac{\exp(s_k^{(m)}) M_k^{(m)}}{\sum_{j \in \{v,a,r\}} \exp(s_k^{(j)}) M_k^{(j)}}$$

$$f_k = \sum_{m \in \{v,a,r\}} \alpha_k^{(m)} \tilde{\boldsymbol{z}}_k^{(m)} \tag{19}$$

将同一样本的多个视图融合表示 $\{\boldsymbol{f}_1, \boldsymbol{f}_2, \cdots, \boldsymbol{f}_K\}$ 通过 PMA 进行样本内聚合,得到样本级表示:

$$\boldsymbol{h} = \mathrm{PMA}([\boldsymbol{f}_1, \boldsymbol{f}_2, \cdots, \boldsymbol{f}_K], \boldsymbol{m}_{view}) \tag{20}$$

其中,$\boldsymbol{m}_{view}$ 为视图掩码。最后,将 $\boldsymbol{h}$ 输入目标域判别头得到样本级预测 $\hat{y}_{p,M_p}$,并对同一被试的多个样本预测采用中位数聚合:

$$\hat{y}_p = \mathrm{median}(\hat{y}_{p,1}, \hat{y}_{p,2}, \cdots, \hat{y}_{p,M_p}) \tag{21}$$

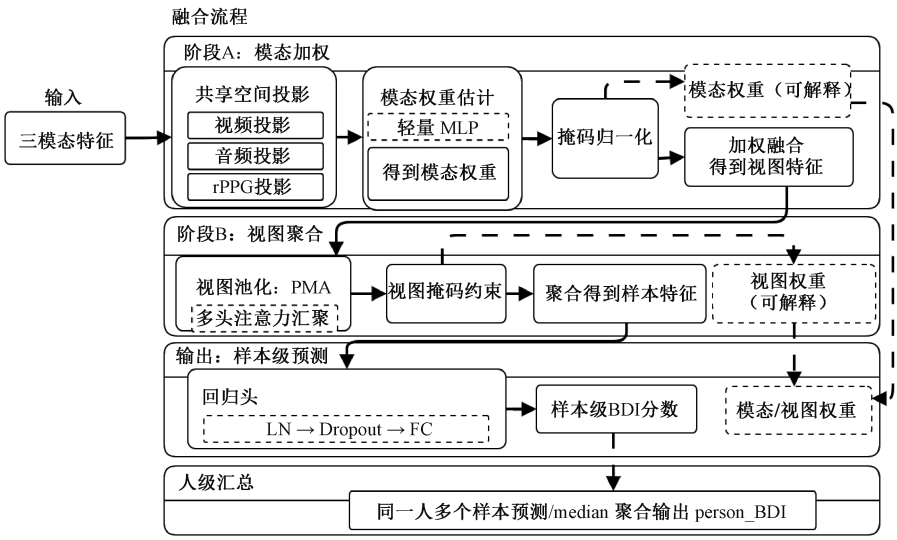


图 5 特征融合流程

Fig. 5 Feature fusion process

3 实验分析

3.1 实验设置

为验证所提方法的有效性,在 AVEC2014 及自建心理异常状态数据集上开展实验。其中,AVEC2014 采用贝克抑郁量表 II(beck depression inventory-II)评分作为标注,自建数据集用于最终的心理异常状态识别。自建数据集实验则重点考察模型迁移后的分类性能。

实验平台为戴尔 DESKTOP-P0GUV,硬件配置为 NVIDIA GeForce RTX 4090 D 显卡(24 GB 显存),驱动版本为 580.105.08,操作系统为 Windows 11,软件环境为

Python 3.8.10、PyTorch 2.1.2 和 CUDA 12.1。除特别说明外，训练阶段统一采用 batch_size 大小为 3、训练轮数为 30 的设置，学习率设为 2×10^{-4}；源域回归任务采用均方误差和平均绝对误差的等权组合损失，以同时兼顾整体拟合精度与对离群误差的鲁棒性。计算公式如式(22)~(24)所示。

$$L_{\mathrm{MSE}}=\frac{1}{N}\sum_{i=1}^{N}(\tilde{y}_i-\hat{y}_i)^2 \tag{22}$$

$$L_{L1}=\frac{1}{N}\sum_{i=1}^{N}|\tilde{y}_i-\hat{y}_i| \tag{23}$$

$$L=\frac{1}{2}(L_{\mathrm{MSE}}+L_{L1}) \tag{24}$$

其中，$\tilde{y}_i$ 表示真实值；$\hat{y}_i$ 表示模型预测值。目标域自建心理异常状态数据集上，将异常状态标签编码为{0，1}输出单一连续分数。训练阶段采用均方误差与平均绝对误差的加权组合作为监督；测试阶段将连续输出映射为类别标签。

评价指标公开数据集实验采用平均绝对误差(MAE)和均方根误差(RMSE)评价回归性能；自建心理异常状态数据集实验采用准确率(Acc)、精确率(precision)、召回率(recall)和 F1 值评价分类性能。

3.2　实验结果与分析

首先，在自建心理异常状态数据集上比较不同训练策略的性能，以验证迁移学习的有效性。结果如表 2 所示，最佳分类混淆矩阵如图 6 所示。

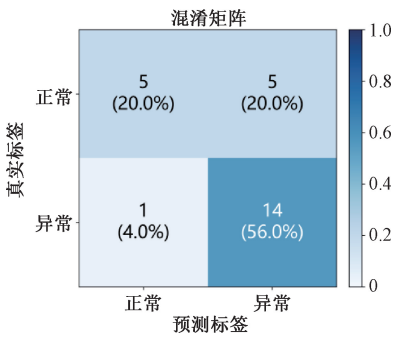


图 6　自建数据集混淆矩阵

Fig. 6　Confusion matrix of self-built dataset

表 2　自建心理异常状态数据集上不同训练策略的结果

Table 2　Results of different training strategies on the self-built risk dataset

训练策略	Acc	Precision	Recall	F1
仅自建数据上训练	0.64	0.63	1.00	0.77
AVEC2014 预训练后微调	0.68	0.67	1.00	0.80
AVEC2014 预训练后部分冻结微调	0.76	0.74	0.93	0.82

从表 2 可以看出，迁移学习对于本研究的心理异常状态识别任务具有明显的有效性。虽 AVEC2014 抑郁数据集与自建数据并非完全同一任务，但二者都属于面向情绪与心理状态的视听时序建模问题，在表情变化、语音特征以及生理信号波动等底层表征上具有一定共性。因此，先在公开数据集上学习通用的多模态时序特征，再迁移到自建数据上进行适配，是较为合理的技术路径。实验表明，AVEC2014 预训练后微调的性能整体优于仅在自建数据上随机初始化训练，说明预训练模型能够提供更稳定、更有效的参数初始化，缓解小样本条件下模型训练不充分的问题。在此基础上，采用部分冻结微调策略进一步取得了最佳结果，Acc 和 F1 分别达到 0.76 和 0.82。该结果说明，部分冻结不仅能够保留预训练阶段学到的通用知识，还能避免在小规模目标数据上过度更新参数，从而实现泛化能力与任务适应性的更好平衡。

为验证提出的多视图多 ROI 双通道质量感知融合算法(图 7)的性能，在数据集 UBFC-RPPG(包含静态少动的 UBFC1 与带有心理应激运动的 UBFC2)进行比较，结果如表 3 所示。

表 3　UBFC 数据集上不同 rPPG 提取方法对比

Table 3　Comparison of different rPPG extraction methods on the UBFC dataset

信号提取方法	UBFC1		UBFC2	
	MAE(min^{-1})	PCC	MAE(min^{-1})	PCC
CHROM	2.26±0.97	0.76	4.06±1.21	0.89
POS	1.82±0.43	0.92	4.18±7.87	0.61
OMIT	2.27±4.04	0.91	3.66±2.36	0.87
本文	1.38±0.51	0.99	2.05±0.92	0.94

由表 3 可知，本文提出的双通道质量感知融合框架方法超越了单独的 POS 或 OMIT 算法，并在存在心理应激与表情肌肉形变的复杂场景(UBFC2)下能够动态评估质量并进行通道优选，取得了较低的误差。

为验证模型在源域任务上的表征学习能力及跨数据集泛化性能，将其与公开数据集上的已有方法进行比较，结果如表 4 所示。

表 4　公开数据集上的模型结果与现有方法对比

Table 4　Comparison of model results on public datasets with existing methods

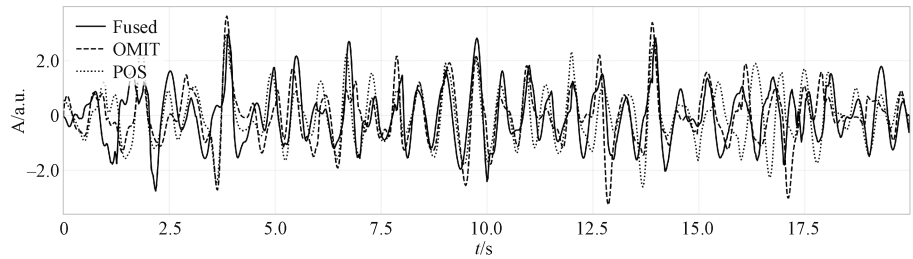
方法	MAE	RMSE
文献[20]	8.19	10.00
文献[21]	6.01	7.60
文献[22]	5.21	7.03
文献[23]	5.50	7.29
本文	5.14	6.37

由表 4 可知，本文方法在 AVEC2014 上取得了 5.14 的 MAE 和 6.37 的 RMSE，整体性能处于现有方法较优水平，尤其 RMSE 具有一定优势，表明模型在误差控制和预

测稳定性方面表现较好。该结果说明所提模型能够较充分地学习与心理状态相关的多模态动态特征,为后续迁移到目标任务奠定表征基础。

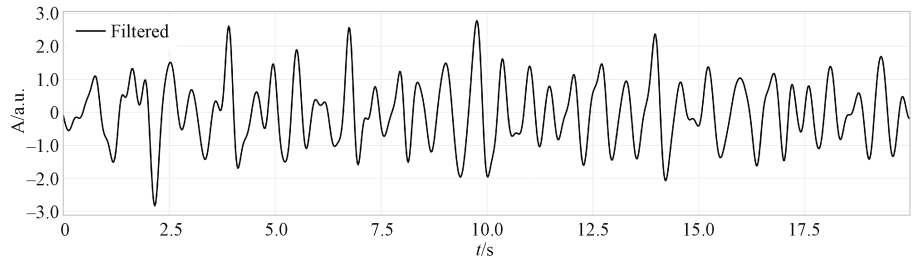
3.3 消融实验与对比分析

为进一步分析模型中关键模块的作用,从单模态特征提取分支和多模态融合方式两个方面开展对比实验。前者通过替换不同模态的特征提取网络考察时序编码方式对性能的影响;后者则在统一单模态特征输出的基础上比较不同融合策略,以验证掩码约束与自适应加权机制的有效性。



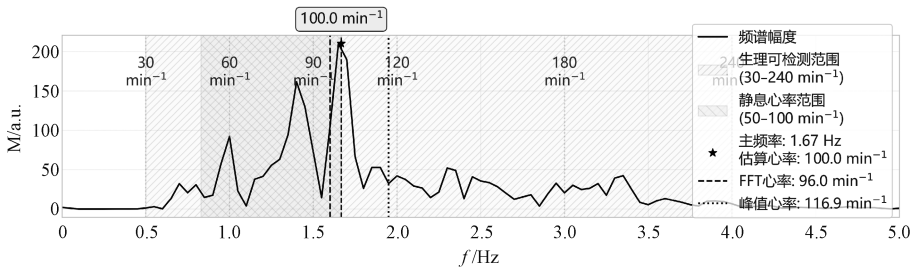
(a) POS、OMIT 与融合算法时序波形

(a) Time series waveforms of POS, OMIT and fusion algorithms



(b) 滤波后心率脉搏时域波形

(b) Time-domain waveform of filtered heart rate and pulse



(c) 频域分析图

(c) Frequency domain analysis diagram

图 7 rPPG 融合算法示例图

Fig. 7 Example diagram of rPPG fusion algorithm

表 5 不同特征提取网络

Table 5 Different feature extraction networks

模态	网络结构	AVEC2014		DAIC-WOZ	
		MAE	RMSE	MAE	RMSE
视频	ResNet18+BiLSTM	6.88	8.64	/	/
	ResNet18+Transformer	6.34	7.39	/	/
	R(2+1)D-18	6.73	8.69	/	/
	R(2+1)D-18+多尺度空洞 TCN(GLU)	5.85	7.49	/	/
	R(2+1)D-18+多尺度空洞 TCN(GLU)+统计测量特征	6.28	8.34	/	/
音频	ResNet18_CBAM	8.50	10.41	5.25	6.59
	ResNet18_CBAM+SSL+MLP	7.91	10.63	4.73	5.86
	ResNet18_CBAM+SSL+双流自注意力融合	7.39	9.19	3.94	5.28
rPPG	Transformer	8.17	10.47	/	/
	多尺度 TCN+生理测量特征	7.54	8.96	/	/

由表 5 可见,引入更有效的时序建模后,各模态性能均有不同程度提升。对于视频分支,R(2+1)D-18 结合多尺度空洞 TCN 后取得最优结果,说明多尺度时序编码能够更有效地刻画面部动态中的局部变化与中程依赖。在视频模态中加入 OpenFace 统计测量特征后,MAE 和 RMSE 未出现明显降低。本文认为自然场景下的关键点抖动会放大高阶导数的噪声,且片段级的统计池化抹杀了微观的连续时序动态,从而干扰了深度时空网络的特征表达。对于音频分支,在引入自监督语音表征后,模型性能较仅使用谱图特征明显提升;采用双流自注意力融合后,在两个数据集上均取得最佳结果,表明低层时频特征与高层语音语义信息具有较强互补性。对于 rPPG 分支,多尺度 TCN 与生理测量特征的组合优于直接使用 Transformer,说明在噪声较大、局部波动明显的远程生理信号上,基于局部节律建模的方式更具稳健性。

表 6　不同特征融合方式性能对比

Table 6　Performance comparison of different feature fusion methods

融合方法	MAE	RMSE
Cat(特征级拼接+MLP)	6. 75	8. 29
Weighted Borda(决策级加权 Borda 投票融合)	6. 21	7. 45
本文(掩码模态加权融合)	5. 14	6. 37

在多模态融合方面,比较了特征拼接、决策级加权投票和本文方法在 AVEC2014 上的表现,结果如表 6 所示。特征拼接难以有效处理模态之间的信息冗余与质量差异,因而误差较大;决策级加权投票虽能在一定程度上利用模态互补信息,但缺乏对模态可用性和动态相关性的显式约束。相比之下,在融合阶段引入模态可用性掩码,并在权重归一化时对缺失模态进行屏蔽,使融合权重仅在有效模态间分配;同时结合视角注意力池化完成片段级聚合,降低低质量模态对预测结果的干扰。结果表明,该方法在 AVEC2014 上取得了最优的 MAE 和 RMSE,说明本文方法能够有效提升多模态融合性能。

综合上述结果可以看出,性能提升主要来源于两方面:1)多尺度时序编码增强了各模态对状态动态演化过程的表征能力;2)掩码自适应融合提高了跨模态信息整合的稳健性。二者共同作用,使模型不仅能够在公开抑郁数据上获得较好的回归性能,而且能够在迁移到自建心理异常状态数据集后保持较强的分类能力。

3.4　模型可解释性分析

为了定量评估不同模态对模型决策的贡献程度,引入了基于 SHAP(SHapley additive exPlanations)的可解释性分析。将每种模态的特征向量视为一个模态子空间,计算其对最终预测值的平均 SHAP 值作为该模态的重要性指标。

对全量 AVEC2014 测试集进行聚合后结果如图 8 所示,其中视频权重最高达到了 0. 513,音频模态平均权重为 0. 310,rPPG 模态的平均权重最低仅为 0. 177。分析各模态权重的标准差。视频模态与音频模态表现出了极高的离散度。这种强烈的波动性源于外显行为特征的个体异质性与情境敏感性:在实际应用中,不同受试者的面部表情丰富度、言语表达习惯、实验环境中的光照变动与背景噪声,都会导致视频和音频特征信息发生剧烈波动。而 rPPG 生理模态展现出了极低的统计标准差,揭示了 rPPG 特征作为一种由自主神经系统调控的非自愿性内部生理信号,无论外显的行为模态因个体差异如何剧烈波动,模型都能从 rPPG 子空间中提取出高鲁棒性、高泛化性的基线生理表征,确保整体多模态决策系统在面对异质性样本时不会发生灾难性的预测失效。

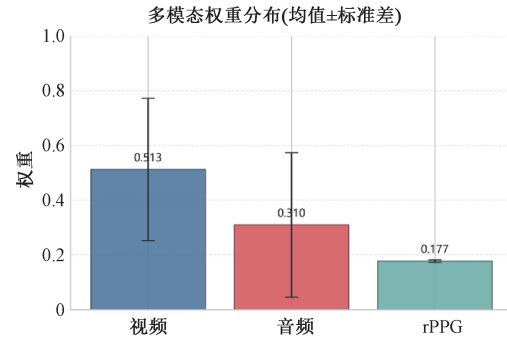


图 8　模态重要性分析

Fig. 8　Analysis of modal importance

4　结　论

本文为公共安全场景下个体心理异常状态的非接触式识别提供了一种可行的技术路径。其面向复杂开放环境中的风险预警需求,通过融合视频、音频与 rPPG 等多源异构信息,实现对个体心理异常状态的早期感知与辅助识别,从而为公共安全领域中潜在极端行为相关线索的前置发现提供客观、连续的技术支撑。需要指出的是,本文所提出的方法旨在服务于风险预警中的辅助分析与状态研判,并非对个体进行医学诊断或风险定性判断。相关研究对于推动心理异常状态识别技术在公共安全预警场景中的应用具有一定参考意义,也可为后续的人机协同研判与早期干预研究提供方法基础。

参考文献

[1] Gu A Q. An exploratory assessment model for preventing individual extreme violent crimes from a social control perspective—a qualitative study of four Chinese cases[J]. Frontiers in Psychology, 2025:1593232.

[2] Adiarte S, Carbon C C, Orr R, et al. The critical role of threat detection and movement behavior assessment: Identifying key concepts, research scope, and gaps-a scoping review [J]. Frontiers in Psychology, 2025, 16: 1627066.

[3] Quiram N, Salam T, Sadjadpour F, et al. A literature review of remote mental health screening: Barriers, potential solutions, and tools [J]. Frontiers in Digital Health, 2025, 7: 1670691.

[4] Garzón-Partida A P, Padilla-Gómez C B, Martínez-Fernández D E, et al. The implementation of digital biomarkers in the diagnosis, treatment and monitoring of mood disorders: A narrative review [J]. Frontiers in Digital Health, 2025, 7: 1595243.

[5] Jiang Z F, Seyedi S, Griner E, et al. Multimodal mental health digital biomarker analysis from remote interviews using facial, vocal, linguistic, and cardiovascular patterns [J]. IEEE Journal of Biomedical and Health Informatics, 2024, 28(3): 1680-1691.

[6] Wang L Y, Wang C H, Li C Y, et al. AI-assisted multi-modal information for the screening of depression: A systematic review and meta-analysis [J]. npj Digital Medicine, 2025, 8(1): 523.

[7] Wu Y, Mi Q W, Gao T H. A comprehensive review of multimodal emotion recognition: Techniques, challenges, and future directions [J]. Biomimetics, 2025, 10(7): 418.

[8] Zhang S Q, Yang Y J, Chen C, et al. Deep learning-based multimodal emotion recognition from audio, visual, and text modalities: A systematic review of recent advancements and future prospects [J]. Expert Systems with Applications, 2024, 237: 121692.

[9] Tao Y F, Yang M Q, Wu Y S, et al. Depressive semantic awareness from vlog facial and vocal streams via spatio-temporal transformer [J]. Digital Communications and Networks, 2024, 10(3): 577-585.

[10] Zhan Y F, Yang R, You J X, et al. A systematic literature review on incomplete multimodal learning: Techniques and challenges [J]. Systems Science & Control Engineering, 2025, 13(1): 2467083.

[11] Nevid J S, Rathus S A, Greene B S. Abnormal psychology in a changing world [M]. 11th ed. New York: Pearson, 2022.

[12] 李泽慧, 张琳, 山显英, 等. 基于多尺度时序建模与动态空间特征融合的视频摘要模型 [J]. 液晶与显示, 2025, 40(11): 1729-1743.
Li Zehui, Zhang Lin, Shan Xianying, et al. Video summarization model based on multi-scale temporal modeling and dynamic spatial feature fusion [J]. Chinese Journal of Liquid Crystals and Displays, 2025, 40(11): 1729-1743.

[13] 李云峰, 闫祖龙, 高天, 等. 基于多任务学习的语音情感识别 [J]. 数据采集与处理, 2024, 39(2): 424-432.
Li Yunfeng, Yan Zulong, Gao Tian, et al. Speech emotion recognition based on multi-task learning [J]. Journal of Data Acquisition and Processing, 2024, 39(2): 424-432.

[14] Álvarez Casado C, Bordallo López M. Face2PPG: An Unsupervised pipeline for blood volume pulse extraction from faces [J]. IEEE Journal of Biomedical and Health Informatics, 2023, 27(11): 5530-5541.

[15] Zheng X J, Zhang C, Chen H, et al. Remote measurement of heart rate from facial video in different scenarios [J]. Measurement, 2022, 188: 110243.

[16] Huang B, Hu S, Liu Z M, et al. Challenges and prospects of visual contactless physiological monitoring in clinical study [J]. Npj Digital Medicine, 2023, 6: 231.

[17] 周先春, 吕梦楠, 芮旸, 等. 基于注意力机制的双卷积图像去噪网络 [J]. 电子测量与仪器学报, 2025, 39(2): 60-71.
Zhou Xianchun, Lyu Mengnan, Rui Yang, et al. Dual convolution image denoising network based on attention mechanism [J]. Journal of Electronic Measurement and Instrumentation, 2025, 39(2): 60-71.

[18] 沈心旸, 王善敏, 孙玉宝. 基于语音语料对齐与自适应融合的抑郁症识别 [J]. 计算机科学, 2025, 52(6): 219-227.
Shen Xinyang, Wang Shanmin, Sun Yubao. Depression recognition based on speech corpus alignment and adaptive fusion [J]. Computer Science, 2025, 52(6): 219-227.

[19] 徐展宇, 陈兆学. 基于改进三维卷积网络的非接触式生理参数检测方法 [J]. 中国医学物理学杂志, 2025, 42(4): 479-488.
Xu Zhanyu, Chen Zhaoxue. Non-contact physiological parameter detection method based on improved 3D convolutional network [J]. Chinese Journal of Medical Physics, 2025, 42(4): 479-488.

[20] 何浪. 基于 3D-CNN 和时空注意力-卷积 LSTM 的抑郁症识别研究 [J]. 首都师范大学学报(自然科学版), 2021, 42(2): 17-25.
He Lang. Research on depression recognition based on 3D-CNN and spatio-temporal attention- convolutional LSTM [J]. Journal of Capital Normal University (Natural Science Edition), 2021, 42(2): 17-25.

[21] Li Y T, Liu Z Y, Zhou L, et al. A facial depression recognition method based on hybrid multi-head cross attention network[J]. Frontiers in Neuroscience, 2023, 17: 1188434.

[22] Uddin M A, Joolee J B, Sohn K A. Deep multi-modal network based automated depression severity estimation[J]. IEEE Transactions on Affective Computing, 2023, 14(3): 2153-2167.

[23] Niu M Y, Tao J H, Liu B, et al. Multimodal spatiotemporal representation for automatic depression level detection[J]. IEEE Transactions on Affective Computing, 2023, 14(1): 294-307.

作者简介

田中雨，2023 年于中国刑事警察学院获得学士学位，现为中国刑事警察学院硕士研究生，主要研究方向为智能警务识别。
E-mail: 205214745@qq.com

Tian Zhongyu received his B. Sc. degree from the Criminal Investigation Police University of China in 2023. He is currently a M. Sc. candidate at the same university. His main research interest includes intelligent policing recognition.

姜囡（通信作者），2001 年于东北大学获得学士学位，2004 年于东北大学获得硕士学位，2007 年于东北大学获得博士学位，2012 年于东北大学完成博士后研究工作，现为中国刑事警察学院教授，主要研究方向为智慧警务、安全防控技术、智能识别与信息处理。
E-mail: zgxj_jiangnan@126.com

Jiang Nan (Corresponding Author) received her B. Sc. degree, M. Sc. degree and Ph. D. degree from Northeastern University in 2001, 2004 and 2007. She completed her postdoctoral research at Northeastern University in 2012. She is currently a professor at the Criminal Investigation Police University of China. Her main research interests include smart policing, security prevention and control technology, intelligent recognition and information processing.