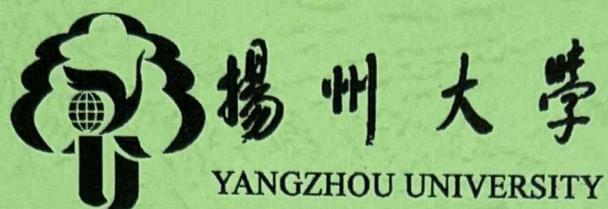


分类号 \_\_\_\_\_  
UDC \_\_\_\_\_

学 号 MZ120220933  
密 级 \_\_\_\_\_



## 硕士学位论文

(全日制专业学位)

### 基于多模态数据的心理健康测评模型研究与系统实现

学 院 : 信息工程学院

专 业 学 位 类 别 : 电子信息

专 业 学 位 领 域 : 计算机技术

研 究 生 : 潘 登

校 内 指 导 教 师 : 朱俊武 教授

校 外 指 导 教 师 : 朱 涛 高级工程师

扬州森科科技有限公司

答 辩 委 员 会 主 席 : 李 斌 教 授

答 辩 日 期 : 2025 年 5 月 23 日

# **Research and System Implementation of a Multimodal Data-Based Mental Health Assessment Model**

A thesis submitted to  
Yangzhou University  
in partial fulfillment of the requirements  
for the degree of  
Electronic Information

By  
Deng Pan  
Supervisor: Prof. Junwu Zhu  
Co-supervisor: Senior Engineer Tao Zhu  
Computer Technology  
2025

# 目 录

第1章 绪论.....	1
1.1 研究背景与意义.....	1
1.2 国内外研究现状.....	2
1.2.1 基于提示学习的图文分类.....	2
1.2.2 基于多模态访谈记录的心理病症等级评估 .....	3
1.3 研究内容.....	5
1.3.1 基于多模态提示学习的心理健康问题分类预测 .....	5
1.3.2 基于多模态访谈记录的心理病症等级评估 .....	5
1.3.3 多模态心理健康检测评估系统设计与实现 .....	6
1.4 章节设计 .....	6
第2章 基础理论与知识 .....	9
2.1 基于多模态提示学习的模型与方法 .....	9
2.1.1 提示学习 .....	9
2.1.2 Transformer .....	10
2.1.3 CLIP .....	13
2.2 基于音视频的模型与方法 .....	15
2.2.1 CLNF.....	15
2.3 本章小结.....	16
第3章 基于多模态提示学习的心理健康问题分类预测 .....	19
3.1 引言 .....	19
3.2 数据集和数据预处理.....	20
3.2.1 数据集 .....	20
3.2.2 MENTAL5 数据集预处理 .....	21
3.2.3 心理健康量表预处理.....	22
3.3 模型与算法.....	23
3.3.1 多尺度浅层视觉增强模块.....	24
3.3.2 跨模态协同提示生成器.....	28
3.3.3 心理健康联合预测模块.....	31
3.4 实验与分析.....	31
3.4.1 实验环境.....	32
3.4.2 实验配置 .....	32
3.4.3 评价指标.....	32
3.4.4 基线模型.....	34
3.4.5 对比实验.....	36

3.4.6 消融实验.....	38
3.4.7 参数实验.....	41
3.5 本章小结.....	42
第4章 基于多模态访谈记录的心理病症等级评估 .....	45
4.1 引言.....	45
4.2 数据集与数据预处理.....	45
4.2.1 数据集.....	45
4.3 模型与方法.....	46
4.3.1 访谈记录音频特征提取.....	47
4.3.2 访谈记录视觉特征提取.....	48
4.3.3 访谈记录文本特征提取.....	49
4.3.4 跨尺度信息构建.....	49
4.3.5 复合式金字塔协同网络.....	51
4.3.6 心理健康等级评估.....	56
4.4 实验与分析.....	57
4.4.1 实验环境.....	57
4.4.2 实验配置.....	57
4.4.3 评价指标.....	58
4.4.4 基线模型.....	59
4.4.5 对比实验.....	60
4.4.6 消融实验.....	64
4.4.7 参数实验.....	65
4.5 本章小结.....	66
第5章 多模态心理健康检测评估系统设计与实现 .....	67
5.1 引言.....	67
5.2 系统开发环境与技术.....	67
5.3 系统需求分析.....	67
5.3.1 功能性需求分析.....	68
5.3.2 非功能性需求分析.....	69
5.4 系统设计.....	69
5.4.1 系统总体设计.....	69
5.4.2 系统功能模块设计.....	70
5.5 系统功能实现.....	71
5.5.1 用户登录.....	71
5.5.2 系统首页.....	72
5.5.3 心理状况检测.....	72
5.5.4 病症等级评估.....	73
5.5.5 量表测试.....	74

5.5.6 知识科普 .....	75
5.5.7 视频浏览 .....	75
5.5.8 健康咨询 .....	76
5.5.9 个人中心 .....	77
5.6 系统测试 .....	77
5.6.1 功能测试 .....	78
5.6.2 界面测试 .....	79
5.6.3 性能测试 .....	79
5.7 本章小结 .....	80
第 6 章 总结与展望 .....	81
参考文献 .....	83
附录 .....	89

## 摘要

心理健康问题日益成为全球范围内的重要公共卫生挑战，如何利用先进技术实现高效、精准的心理健康检测与评估已成为当前研究的热点和难点。传统的测评手段大体分类两种：第一种采用心理健康量表的测评方式对待测人群进行标准化和系统化心理健康状态的评估。第二种采用专家访谈的问答方式对被访谈者进行深入、专业的心灵测评。然而，上述传统的测评方式存在模态单一性、低效性、信息利用不充分性等缺点。因此，基于计算机技术的心理健康测评成为当前前沿的研究课题。对于上述两种传统方式，目前已经提出了众多基于深度学习的模型与算法。然而，兼顾心理检测任务中准确性、高效性、信息利用充分性依旧是一个巨大的挑战。针对这些问题，本文对国内外关于多模态心理健康检测评估方法进行了详细的梳理和总结，并进行了以下内容的研究：

(1) 提出了一种名为 VSCP-Net 的多模态提示学习模型。该模型通过引入基于提示学习技术的多尺度浅层视觉增强模块和跨模态协同提示生成器，提高了对心理健康问题分类的精度和解释性。此外，本文采用 MENTAL5 数据集验证了模型的性能，实验结果表明，VSCP-Net 在多项评估指标上显著优于现有方法，为心理问题的早期分类预测提供了技术支持。

(2) 提出了基于多尺度语义构建和金字塔注意力机制的 MPCN 模型。该模型在音视频数据中捕捉跨尺度信息，并通过语义分层建模和复合式金字塔协同网络有效评估心理疾病等级。实验证明了模型在预测精度和效率上的优越性，为常见心理健康疾病的自动化等级评估提供了一种高效方法。

(3) 构建了一套集成心理健康问题分类、病症等级评估、心理量表分析与个性化健康建议的多模态心理健康检测评估系统。该系统采用模块化设计，功能覆盖心理健康检测、健康咨询、量表测试、知识科普等功能，测试结果显示其具有良好的用户体验与实用性。

**关键词：**心理健康；多模态分析；多尺度融合；提示学习

## Abstract

Mental health issues have increasingly become a critical global public health challenge. How to leverage advanced technologies for efficient and accurate mental health detection and assessment has emerged as both a research hotspot and challenge. Traditional assessment methods primarily fall into two categories: the first employs standardized psychological scales to systematically evaluate the mental health status of subjects, while the second relies on expert interviews for in-depth professional psychological evaluations. However, these conventional approaches suffer from limitations such as unimodal nature, inefficiency, and inadequate information utilization. Consequently, computer-based mental health assessment has evolved into a cutting-edge research field. Although numerous deep learning-based models and algorithms have been proposed to address these traditional methods, achieving a balance between accuracy, efficiency, and comprehensive information utilization in mental health detection remains a significant challenge. To address these issues, this study conducts a comprehensive review of multimodal mental health assessment methodologies worldwide and presents the following research contributions:

(1) VSCP-Net: A multimodal prompt learning model is proposed, which enhances classification accuracy and interpretability for mental health issues through a multi-scale shallow visual enhancement module and a cross-modal collaborative prompt generator based on prompt learning techniques. Validated on the MENTAL5 dataset, experimental results demonstrate that VSCP-Net significantly outperforms existing methods across multiple evaluation metrics, offering technical support for early-stage classification of psychological disorders.

(2) MPCN: A framework integrating multi-scale semantic construction and pyramid attention mechanisms is developed. This model captures cross-scale information in audiovisual data, enabling effective assessment of mental disorder severity through hierarchical semantic modeling and a composite pyramid collaborative network. Experiments confirm its superiority in prediction accuracy and efficiency, providing an automated approach for grading common mental health conditions.

(3) Integrated System: A multimodal mental health assessment system is constructed,

incorporating psychological disorder classification, severity evaluation, scale analysis, and personalized health recommendations. Designed with modular architecture, the system covers functions such as mental health screening, counseling, scale testing, and educational resources. Testing results indicate favorable user experience and practical utility.

**Keywords:** mental health; multimodal analysis; multi-scale fusion; prompt learning

## 第1章 绪论

### 1.1 研究背景与意义

心理健康问题已成为 21 世纪全球公共卫生领域的重大挑战。根据世界卫生组织最新流行病学调查，全球范围内约 13.0% 的疾病负担可归因于精神障碍，其中抑郁症位列致残原因首位<sup>[1]</sup>。2019 年新型冠状病毒疾病（Corona Virus Disease 2019，COVID-19）大流行更引发心理健康危机加剧。有研究证实，2020-2022 年全球重度抑郁症和焦虑症患病率分别激增 27.6% 和 25.6%，其中青年群体（18-29 岁）发病率到达基线水平的 2.3 倍<sup>[2]</sup>。大规模、长时间的居家隔离是阻断疫情传播的有效手段，但同时这也引起了人们对于疾病本身以外的担忧。研究表明，人们在隔离解除之后会产生许多心理后遗症，包括压力、抑郁、烦躁、失眠、恐惧、困惑、沮丧等<sup>[3]</sup>。由新冠疫情这一全球事件而引发的相关心理健康事件可能会演变成长期的心理健康问题<sup>[4]</sup>。除了采取必要的卫生措施，将现代化的计算机技术与传统的心理评估方式相结合来解决人们潜在的心理健康问题，是一个值得研究的课题。

心理健康问题的评估一直以来都是一项难度较大的任务。传统的测评手段大致分为两类：其一是采用心理健康量表的测评方式，针对待测人群展开标准化且系统化的心理健康状态评估。该方法一般包含一系列精心设计的问题或陈述，要求被测者依据自身实际情况或感受给予回答。心理健康量表可以是自评式的，也可由专业人员进行他评。常用的量表包含广泛用于评估个体心理症状和心理健康状态的症状自评量表-90 项<sup>[5]</sup>（Symptom Check List-90，SCL-90），以及广泛用于筛查和评估抑郁症严重程度的患者健康问卷-9 项<sup>[6]</sup>（Patient Health Questionnaire-9，PHQ-9）等。其二是运用专家访谈的问答方式，对被访谈者实施深入且专业的心理评估。这种方法一般涉及与经验丰富的心理健康专家进行面对面交流，借助问答形式探讨和评估个体心理状况。但是上述传统测评方式存在主观性偏差以及时间滞后性等缺点，一方面，不管是心理量表还是专家访谈，每一步评估都需要患者主动进行自我审查，若期间患者拒绝诊断甚至中途放弃，那么整体测评效果会大幅降低。另一方面，这两种方式本质上是对患者过往一段时间心理状态的评估，无法反映患者的即时心理状态，这可能致使患者错过

最佳治疗时机甚至加重病情。故而，采用新兴计算机技术与传统方式相结合进行心理问题评估与诊断，有极大的研究价值。

现代化计算机技术取得了巨大的进步，这为心理健康诊断给予了全新的可能性。借助机器学习、人工智能以及大数据分析等前沿技术，系统可处理数量庞大的多维数据，而且还可以借助复杂的算法识别出传统评估方法很难发现的模式。比如运用自然语言处理技术来分析患者的语言模式，可有效地识别出抑郁、焦虑等情绪问题<sup>[7]</sup>；采用计算机视觉，可以捕捉面部表情<sup>[8]</sup>、眼动<sup>[9]</sup>等非语言信号，以非侵入性的方式迅速采集数据并加以分析，极大地提升了心理健康评测的客观性和实时性。另外数字化心理健康应用程序以及远程诊断平台越来越多地被用于远程监控和评估心理健康状态<sup>[10]</sup>。这些技术的应用让心理健康评估变得更为精准和高效，促进了基于计算机技术的心理健康评估与诊断的发展。然而，由于对心理健康状态的检测是一项复杂且艰巨的任务，兼顾心理评估的多模态性、准确性、高效性、信息利用充分性依旧是一个巨大的挑战。

本课题将基于多模态数据的心理健康问题分类预测以及等级评估当作研究目标，探索基于多模态提示学习的心理健康问题分类预测方法以及基于多模态访谈记录的心理病症等级评估方法，同时结合真实的应用场景，最终设计并实现一套多模态心理健康检测评估系统。

## 1.2 国内外研究现状

### 1.2.1 基于提示学习的图文分类

近年来，预训练的视觉语言模型在图文分类任务里呈现出强大的通用性与灵活性。该类模型借助训练图像与文本的嵌入空间，让二者可在同一语义空间中达成对齐。这类模型可直接把自然语言当作分类标签，极大减少了对标注数据的依赖，并且有跨领域迁移的能力。当这类模型与提示学习<sup>[11]</sup>技术相结合时，能提升分类性能以及泛化能力。提示学习的核心观念是依靠设计恰当的文本提示，将分类任务转变为模型更易于理解的形式。在心理健康相关的图文检测分类任务中，预训练的视觉语言模型以及提示学习技术有着关键应用价值，此类任务一般涉及对图像与文本内容的情感、心理状态或者潜在风险信号的联合分析。近几年，基于提示学习技术的分类识别模型取得了较大发展。

Zhou 等人<sup>[12]</sup>提出了上下文优化方法（Context Optimization, CoOp），此方法把

上下文词建模为可学习的向量，自动化视觉-语言模型的提示生成过程，以解决传统提示工程在图像分类任务中面临的挑战。CoOp 方法借助冻结的预训练模型参数，仅优化上下文词的可学习向量，大幅提升了在数据稀缺情况下的性能，特别在多个数据集上证实了其有效性，显示出超过 15% 的性能增益。

之后，Zhou 等人<sup>[13]</sup>的团队提出条件上下文优化方法（Conditional Context Optimization, CoCoOp），依靠动态生成与每个输入图像对应的提示，提高了模型在未见类别上的识别准确性。这一方法在多项实验中呈现出更强的迁移能力，且在多个数据集上获得了十分突出的准确率提升。

Khattak 等人<sup>[14]</sup>提出多模态提示学习方法（Multi-modal Prompt Learning, MaPLe），凭借强化视觉与语言提示之间的协同作用，提升预训练视觉-语言模型在多个下游任务中的适应性。MaPLe 依靠联合提示学习策略，在多个编码层中实现视觉与语言的有效对齐。实验结果说明 MaPLe 在新类别的泛化和跨数据集评估中均呈现出优越的性能，超越了现有的一些基准方法。

Guo 等人<sup>[15]</sup>提出一种新颖的提示调优方法，将文本视为图像，有效利用文本描述的丰富性来提升模型性能，同时降低对标注数据的依赖。这一方法依靠双粒度提示调优策略，在多个主流数据集上超越了基线模型，为多标签图像识别问题提供了新的解决方案。

Jia 等人<sup>[16]</sup>在研究中提出了一种视觉提示微调的高效模型（Visual Prompt Tuning, VPT），借助引入任务特定的可学习提示参数，达成对大型预训练视觉模型的有效迁移。在多个下游识别任务里，VPT 减少了存储需求以及计算开销，在数据稀缺情形下的表现超过了传统的全量微调方法，较大程度提升了模型的适应能力与使用效率。

Xing 等人<sup>[17]</sup>引入了双模态提示调优的方法（Dual-modality Prompt Tuning, DPT），依靠同时学习视觉和文本提示来提高视觉-语言模型在下游任务中的表现。DPT 利用类感知视觉提示，借助交叉注意力机制把视觉和文本信息结合，提升模型对目标类的关注能力。实验结果显示，DPT 在多个数据集上的性能表现超过了依赖单一文本提示的调优方法，体现了其在多模态学习中的关键性与优越性。

## 1.2.2 基于多模态访谈记录的心理病症等级评估

近年来，深度学习在医疗领域的应用取得了显著的发展<sup>[18-25]</sup>。在心理健康问题检测方面，基于计算机技术的智慧医疗方案有传统评估方法所不具备的优势。在现实世

界当中存在着形式不一样、结构也各有差异的数据，研究者们运用深度学习技术去探索了各式各样的预测模型以及学习方法。这些模型和方法所运用到的数据囊括了面部图像、微表情、肢体动作、社交媒体数据、音视频、脑电图、心理量表测试记录、访谈记录等单一模态或者是多个模态的组合，其中包含多种模态数据的患者访谈记录可有效地对潜在患者的细致心理状态进行预测以及综合分析。

Dinkel 等人<sup>[26]</sup>提出一种用于心理问题检测的自监督音频嵌入方法。他们采用类似于 Word2Vec 的预训练框架，借助编码器-解码器网络来提取音频嵌入。在实验里，模型在数据稀疏场景下比传统音频特征表现得更好。在分类和回归性能方面，验证了跨领域预训练是有效的。

Sun 等人<sup>[27]</sup>设计了一个基于多层感知机（Multilayer Perceptron, MLP）的多模态特征处理框架，用于情感分析和心理病症检测。该网络依靠在序列、模态和通道三个维度上对多模态特征进行混合，达成了高效的特征融合。模型在多个基准数据集上取得了领先的性能，同时也大幅降低了计算成本。

Ray 等人<sup>[28]</sup>提出了一种基于多层次注意力机制的多模态心理问题预测网络，该方法借助融合文本、音频和视频特征，提高了病症的预测准确性。研究显示，模型在不同模态中动态选择和加权最具影响力的特征，且优化了决策过程，在均方根误差上超越了现有基线 17.52%。该研究为自动化心理症诊断提供了新的思路与方法，推动了情感计算领域的发展。

Chen 等人<sup>[29]</sup>提出一种基于深度学习的方法，该方法凭借显式建模访谈中的层次化问题结构，提高了问题语义与多模态特征的交互表示。在公开心理健康数据集上的实验说明，模型在抑郁等心理问题检测任务中实现了先进性能，呈现了对未见问题的强泛化能力，为临床访谈中的情绪分析提供了新的视角。

Wang 等人<sup>[30]</sup>提出了一种名为非均匀扬声器解缠法的方法，来降低对扬声器身份的依赖，有效减少隐私泄露风险。借助对抗性损失的非均匀加权，该方法提升了心理疾病检测的准确性，也强调了隐私保护的关键性。在公开心理数据集上的实验结果显示其 F1 得分达到了 0.735。

Qureshi 等人<sup>[31]</sup>针对基于注意力机制的深度神经网络模型展开研究，将声学、文本以及视觉信息加以融合，以此对心理疾病评估能力给予优化。实验得出的结果说明，该模型于均方根误差以及平均绝对误差这两方面，相较于现有技术而言，分别提升了 7.17% 与 8.08%，充分呈现出多模态特征融合在心理健康检测环节的有效性。

Ghadiri 等人<sup>[32]</sup>推出一种创新的多模态框架，借助把音频信号和文本信息进行结合的方式，运用深度学习技术提升常见心理问题的预测精准度。凭借提取低级音频特征以及图特征，该模型于心理健康数据集上实现了最高 86.6% 的准确率，有效弥补了传统单模态心理检测方法存在的不足。

Zhao 等人<sup>[33]</sup>搭建了一个情感音频文本抑郁语料库，为抑郁检测方法的发展奠定关键基础。他们对 162 名志愿者的音频和文本数据展开收集与处理工作，填补了当下中文抑郁症数据集研究方面的空白，并且利用此数据集开展了一系列深度学习应用研究。

Liu 等人<sup>[34]</sup>提出一个结合双向门控递归单元和多任务学习的模型，着重关注文本心理检测里的稀疏数据问题。该研究对大量文本输入进行优化，成功提高了疾病检出率。实验结果说明该模型在公开心理数据集的开发集上取得了平均宏 F1 分数为 0.84 的出色表现。

Gimeno-Gómez 等人<sup>[35]</sup>探寻了一种基于视频中多模态特征的心理问题识别新办法，重点关注面部表情、声音以及空间信息等非语言线索。在多个数据集上，该模型在音频与视频特征联合运用方面达到了较好的检测效率，F1 分数最高可达到 0.78，为心理健康领域的早期识别提供了有效的工具。

## 1.3 研究内容

### 1.3.1 基于多模态提示学习的心理健康问题分类预测

心理健康问题筛查的传统办法是运用标准化、结构化的心理量表，近些年发展出一些基于深度学习的多模态检测模型对心理问题进行评估筛查，然而二者并不能很好地平衡效率和精度两大核心指标。针对这一问题，本文提出一种基于多模态提示学习的心理健康问题分类预测模型 VSCP-Net（Visual-Scale-Caption Prompt Network）。该模型借助多模态数据进行融合分析，依靠多尺度浅层视觉增强模块提高视觉模态的特征提取能力，并且采用跨模态协同提示生成器达成模态间特征的深度协同融合。模型还设计了心理健康联合预测模块，提升分类结果的准确性与可解释性。VSCP-Net 把常见心理疾病当作研究对象，结合多模态图文数据与结构化量表信息，提高了心理健康问题的预测性能。

### 1.3.2 基于多模态访谈记录的心理病症等级评估

患者访谈记录含有丰富的心理信息，传统做法是凭借医患面对面交流的形式来评估心理状况。但是由于不同医生的专业能力有差异，传统评测方式存在主观性、信息未充分利用等问题。故本文提出基于多模态访谈记录的心理病症等级评估模型 MPCN (Multimodal Pyramid Cooperation Network)，充分利用音频、视频等多种数据模态进行综合心理评估。该模型凭借融合不同尺度的信息，可有效捕捉多模态数据中深层次的时间动态特征和模态间交互信息。同时模型以多种心理量表分数作为参考标准，设计复合评价模块对心理等级进行定性定量评估。实验显示，该方法在预测准确性和模型解释性方面都比传统方法好，为心理健康评估提供了一种精确且高效的技术支持。

### 1.3.3 多模态心理健康检测评估系统设计与实现

本文将基于多模态提示学习的心理健康问题分类预测模型和基于多模态访谈记录的心理病症等级评估进行融合，设计并实现多模态心理健康检测评估系统。该系统可对用户心理状态作实时评估与动态反馈。系统的功能点覆盖用户登录模块、健康检查模块、心理疏导模块和个人中心模块。借助用户上传的图像、音视频、文本数据，系统能初步筛查心理健康问题、定量评估心理等级，还可以给出个性化健康管理建议。另外系统整合了交互性佳的用户界面与智能算法，在功能性和易用性上表现良好，为心理健康监测及干预提供了高效便捷方案。

## 1.4 章节设计

论文整体章节设计如图 1-1 所示。

第 1 章：绪论。该部分主要介绍本论文的基本信息情况。首先阐述论文研究背景及其意义，其次介绍的国内外现状，接着阐述本文的主要研究内容，最后概括整体章节安排。

第 2 章：基础理论与知识。本章系统梳理心理健康检测与评估领域的相关理论与技术基础。

第 3 章：基于多模态提示学习的心理健康问题分类预测。本章提出一种心理健康问题分类预测模型 VSCP-Net，针对心理健康问题的多模态特征设计了基于提示学习技术的多尺度浅层视觉增强模块与跨模态协同提示生成器，并在 MENTAL5 数据集上进行实验验证，为心理健康的早期检测提供技术方案。

第4章：基于多模态访谈记录的心理病症等级评估。本章设计出一种心理病症等级评估模型 MPCN，它将音视频、文本等多模态数据与金字塔协同模块相结合，对音视频、文本等访谈信息展开特征提取以及跨尺度建模，再运用基于多种心理量表评价指标的复合评价模块对心理问题给予细致评测，构建出更为精准的心理健康等级评估办法。

第5章：多模态心理健康检测评估系统设计与实现。该章依据前面的研究成果，设计并实现了一套多模态心理健康检测评估系统。系统采用模块化架构，把心理健康问题分类预测、心理等级评估、心理量表分析以及个性化建议等功能集成在一起，还详细阐述了系统需求分析、功能设计、技术实现以及测试结果。

第6章：总结和展望。首先对本文所提的心理健康检测评估工作进行了总结，然后对心理健康检测评估以及拓展领域的目标识别的未来工作进行展望。

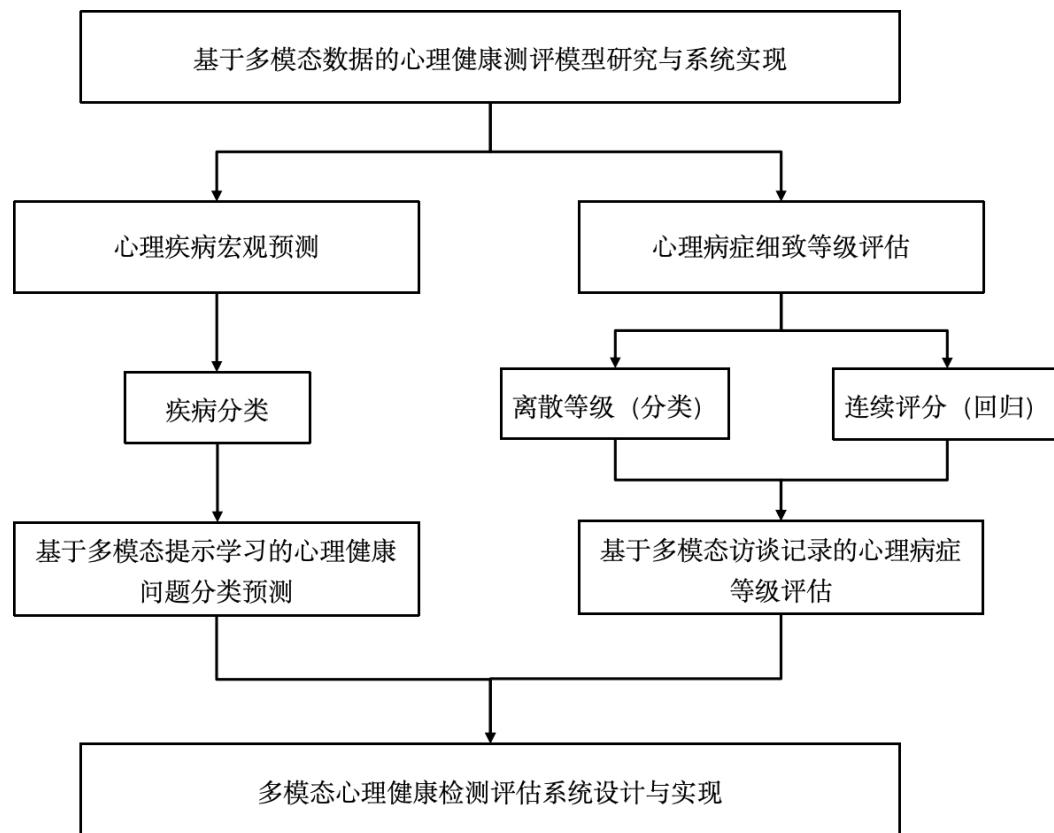


图 1-1 章节设计结构图



## 第 2 章 基础理论与知识

本文旨在探索基于深度学习的心理健康疾病的分类预测和严重等级评估，研究内容包括基于多模态提示学习的心理健康问题分类预测和基于多模态访谈记录的心理病症等级评估两个方面。本文在第 1 章中各自总结了两方面的国内外研究现状与进展，本章将对 1.2 节中所提及的相关方法与基准模型进行介绍。

### 2.1 基于多模态提示学习的模型与方法

本节是所介绍的基础模型与方法是第 3 章研究工作的预备知识。第 3 章的研究内容是基于多模态提示学习的心理健康问题分类预测研究，实质是采用提示学习技术的视觉语言模型的分类问题研究。本节将介绍提示学习相关技术与主流的视觉语言模型，为第 3 章的工作提供借鉴与对比的模型方法。

#### 2.1.1 提示学习

提示学习<sup>[11]</sup>（Prompt-based Learning）作为自然语言处理研究工作里的一种新型范式，基于提示学习的框架可让语言模型在大量原始文本上开展预训练。借助定义新的提示函数，模型可迅速实现各种自然语言处理任务之间的迁移与扩展，如情感分析、问答系统、文本生成等下游子任务。提示学习的核心优势是不用大量额外数据以及标注就能让模型掌握新技能，它突破了传统深度学习模型依赖大规模标记数据的限制，为快速适应新任务、提高泛化能力提供了全新途径。凭借这种办法，模型可调用其内部的知识库，达成对新情境的理解与应对。

传统的监督学习模型输入  $x$  并将输出  $y$  预测为  $P(y|x; \theta)$ ，其中  $\theta$  为可学习参数，而基于提示的语言模型直接对文本概率进行建模。此类模型执行预测任务使用模板将原始输入  $x$  修改为一段含有未填充槽的文本字符串  $x'$ ，然后使用语言模型概率地将未填充槽进行填充以生成答案集  $z$ ，之后映射为最终输出  $y$ 。这一机制使得模型能够更加灵活、通用，提升了其在面对多样性和不确定性时的表现。在不断演进的过程中，提示学习正逐渐成为连接语言模型与现实世界应用的重要桥梁。具体而言，提示学习

分为三步得到最终输出  $y$  中的最高得分  $\hat{y}$ ，分别为提示附加、答案搜索和答案映射。

第一步操作为提示附加（Prompt Addition）。在此步骤中，采用提示函数  $f_{prompt}(\cdot)$  将原始输入文本  $x$  修改为提示  $x' = f_{prompt}(x)$ 。在之前的大部分工作中<sup>[31-34]</sup>，该函数由两步过程组成：

- (1) 应用一个模板。该模板是具有两个槽的文本字符串：用于输入  $x$  的输入槽  $[X]$  和用于中间生成的答案文本  $z$  的答案槽  $[Z]$ ，答案文本  $z$  之后映射到  $y$ ；
- (2) 使用输入文本  $x$  填充输入槽  $[X]$ 。

第二步为答案搜索（Answer Search），即从答案集  $z$  中搜索得分最高的文本  $\hat{z}$ 。首先将  $Z$  定义为  $z$  的一组允许值。在生成性任务的情况下， $Z$  的范围可以是整个语言，或者在分类的情况下， $Z$  可以是语言中单词的一个小子集。然后，定义一个函数  $f_{fill}(x', z)$ ，该函数用潜在答案  $z$  填充提示  $x'$  中的位置  $[Z]$ 。所有经过此过程的提示称为已填充提示。具体地说，如果提示填充了真实答案，则将该提示称为已回答提示（Tab2 显示了一个示例）。最后，使用预先训练的语言模型的预测函数  $P(\cdot; \theta)$  来计算相应填充提示的概率来搜索潜在答案集  $z$ ，如式 2-1 所示。

$$\hat{z} = \{\text{search } P(f_{fill}(x', z); \theta) | z \in Z\} \quad (2-1)$$

该搜索函数可以是搜索得分最高输出的 *argmax* 函数，或者可以是按照语言模型的概率分布随机生成输出的 *sampling* 函数。

第三步是答案映射（Answer Projection）。构建从得分最高答案到得分最高输出的映射。某些情况下，答案本身就是输出，但也存在多个答案产生相同输出的情况。因此，需要在搜索的答案和输出值之间构建合适的映射函数，如式 2-2 所示。

$$\hat{y} = f_{projection}(\hat{z}) \quad (2-2)$$

以往研究<sup>[12-17]</sup>显示，对大型视觉语言模型开展提示微调（Prompt Tuning），在维持预训练参数权重的情形下，可有效提升模型性能。于心理健康检测分析范畴，针对给定的心理健康患者数据，巧妙地设计心理线索提示，可促使模型更全面地融合心理特征，提高心理疾病预测的准确程度。

## 2.1.2 Transformer

本文第三章所提出的模型基于大型预训练视觉语言模型，而 CLIP<sup>[36]</sup>（Contrastive Language-Image Pre-training）模型是将视觉和语言两种模态进行表征对齐的优秀之作，其后众多的模型均以 CLIP 作为基准。CLIP 模型分别有文本和视觉两个编码器，其中

文本编码器的骨干网络和视觉编码器的部分结构都是基于 Transformer<sup>[37]</sup>而组成的预训练模型，故首先对 Transformer 模型的核心架构进行介绍。

Transformer 是深度学习领域中一种极具创新性的架构，由 Google 公司于 2017 年提出，最早在自然语言处理领域表现出卓越的性能。近些年，基于 Transformer 的一些模型在图像分类、目标检测等多个领域得到广泛的应用。相比于以往处理序列数据时所经常使用的循环神经网络<sup>[38]</sup>（Recurrent Neural Network, RNN）模型，Transformer 模型采用多头自注意力机制（Multi-Head Self Attention, MHSA）和前馈神经网络（Feed-Forward Neutral Network, FFN）直接捕捉输入序列中任意两个位置之间的关系，有效克服了传统模型处理序列数据时存在的难以并行化和长期依赖问题。

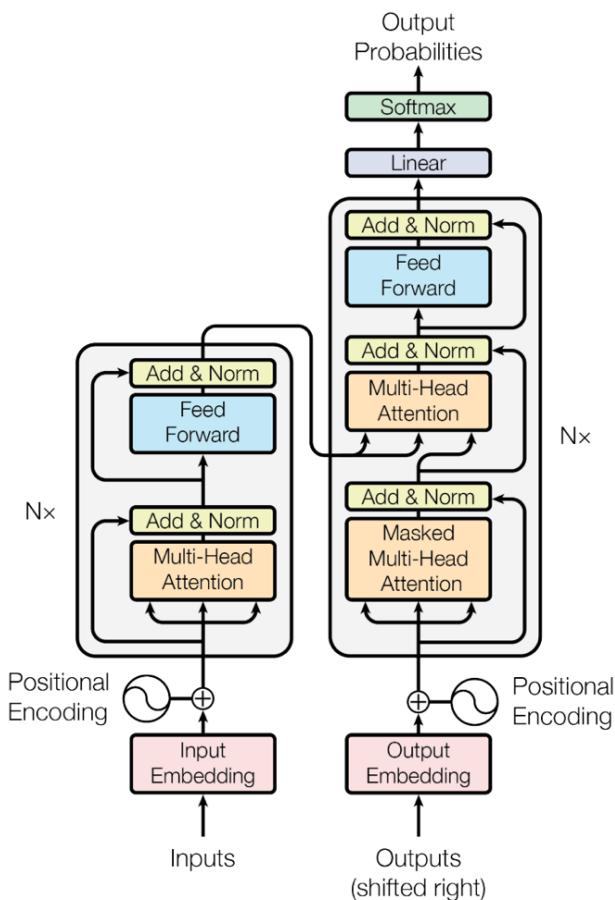


图 2-1 Transformer 整体架构<sup>[37]</sup>

Transformer 模型主要由输入层、编码层、解码层以及输出层这四个部分构成，其完整的架构呈现于图 2-1 之中。输入层是由词嵌入层以及位置编码层共同组建而成的，这样模型便可更有效地处理原始数据。编码层（Encoder Layer）以及解码层（Decoder Layer）是 Transformer 模型的关键部分，这两者都囊括了多头自注意力层、残差连接

层以及前馈神经网络层。多头自注意力层可让模型同时去关注输入序列里的不同位置，捕捉长距离依赖关系，关于这一点会在后续内容里展开详细说明。残差连接层使得模型可学习到输入和输出之间的增量变化，并非直接去学习一个复杂的非线性映射，如此便有效地缓解了神经网络在反向传播过程中所引发的梯度消失问题，加快了网络收敛速度并且提升了模型整体性能。前馈神经网络层会针对每个位置的向量开展非线性变换，以此来提取更为高级别的特征。解码层相较于编码层还额外增加了掩码注意力层，以此保证模型在生成输出序列时可遵循正确的因果关系，避免信息泄露，同时还允许并行计算，这样就提高了模型的训练效率以及生成质量。模型最后的输出层会把上一层的输出转变为概率分布预测出最终结果。

Transformer 模型的核心部分是编码层和解码层，而这两者的最重要的部分是多头自注意力层。多头自注意力是经典自注意力机制的延伸与发展，它通过多个“头”（并行的注意力）计算来捕捉输入序列不同表示子空间中的信息，从而更全面地理解输入序列的语义关系。

具体而言，设输入序列为  $X \in \mathbb{R}^{n \times d_{model}}$ ，其中  $n$  是序列长度， $d_{model}$  是输入向量的维度。定义  $h$  为“头”的数量，对于第  $i \in 1, 2, \dots, h$  个头，首先对输入序列进行线性变换得到查询 (Query) 向量  $Q^i$ 、键 (Key) 向量  $K^i$  和值 (Value) 向量  $V^i$ ，如 2-3 至式 2-5 所示。 $Q^i$ 、 $K^i$ 、 $V^i$  分别由各自可学习的权重矩阵  $W_Q^i \in \mathbb{R}^{d_{model} \times d_k}$ 、 $W_K^i \in \mathbb{R}^{d_{model} \times d_k}$ 、 $W_V^i \in \mathbb{R}^{d_{model} \times d_v}$  与  $X$  相乘得到， $d_k$  和  $d_v$  是  $d_{model}$  经过线性变换之后得到的维度，满足  $d_k = d_v = \frac{d_{model}}{h}$ 。

$$Q^i = XW_Q^i \quad (2-3)$$

$$K^i = XW_K^i \quad (2-4)$$

$$V^i = XW_V^i \quad (2-5)$$

其次，计算第  $i$  个头的注意力分数  $A^i$ ，如式 2-6 所示。式中所用的  $softmax$  函数是对最后一个维度进行操作，将分数归一化到  $[0, 1]$  区间且每行的和为 1。 $\sqrt{d_k}$  的作用是为了防止内积过大导致  $softmax$  函数的梯度消失或梯度爆炸，提升模型在反向传播时的稳定性。

$$A^i = softmax\left(\frac{Q^i(K^i)^T}{\sqrt{d_k}}\right) \quad (2-6)$$

接着，将上述注意力分数乘以相应的值向量，得到对应  $Q^i$ 、 $K^i$ 、 $V^i$  的 *Attention* 结果，如式 2-7 所示。

$$\text{Attention}(Q^i, K^i, V^i) = \text{softmax}\left(\frac{Q^i(K^i)^T}{\sqrt{d_k}}\right)V^i \quad (2-7)$$

最后，将所有头的结果进行拼接，并与可学习的参数矩阵  $W^o \in \mathbb{R}^{h \times d_v \times d_{model}}$  相乘得到最终的输出，如式 2-8 所示，其中  $head_i$  由式 2-9 得到。

$$\text{MultiHead}(X) = \text{Concat}(head_1, head_2, \dots, head_h)W^o, i \in h \quad (2-8)$$

$$head_i = \text{Attention}(Q^i, K^i, V^i), i \in h \quad (2-9)$$

### 2.1.3 CLIP

CLIP<sup>[36]</sup>是 2021 年由 OpenAI 推出的一种具有开创性的深度学习模型，它能够构建图像与文本之间的强大联系。该模型借助对比学习的方法，在大规模未标注数据上进行预训练，达成了图像与文本的统一表征空间，使模型可理解并生成与人类感知相符的视觉概念。近年来，一些与视觉语言模型相关的工作，如 CoOp<sup>[12]</sup>、MaPLe<sup>[14]</sup>、DPT<sup>[17]</sup>等模型，都将 CLIP 模型当作基准。

CLIP 模型旨在对齐图像特征空间和文本特征空间，这使得模型具有零样本迁移到下游任务的能力。其核心在于它的双重编码器架构：一个文本编码器和一个图像编码器。文本编码器基于 Transformer<sup>[37]</sup>架构，能够处理各种长度的文本输入；图像编码器则利用 ResNet<sup>[39]</sup>等卷积神经网络或者是 ViT<sup>[40]</sup>（Vision Transformer）结构，来捕获图像的细节和语义。两个编码器在共享的向量空间中协同优化，通过最小化图像和对应文本描述之间的距离，最大化非匹配对之间的余弦距离，实现了跨模态的对齐与统一。

CLIP 拥有 4 亿对海量图文样本，在对比学习框架下进行训练，其中关联的图像和文本被视为正样本，而非关联的样本被视为负样本。预训练的 CLIP 模型的所有参数都将被冻结用于下游任务，无需任何微调。在下游任务中，将手工制作的提示（prompt）输入到文本末尾，通过嵌入目标数据集的类名称来合成零样本线性分类器。以分类任务为例，“[CLASS]”标记可以首先通过模板进行扩展，例如“a photo of a [CLASS]”。然后，将该句子视为提示，并由文本编码器进行编码以导出文本表示与图像编码器得到图像表示进行余弦相似度（cosine similarity）计算，得到特定类别分类的预测概率。

CLIP 模型的架构如图 2-2 所示。第一步，CLIP 对图像  $I \in \mathbb{R}^{H \times W \times 3}$  和相应的文本描述进行编码。对于图像编码，以基于 ViT 架构的骨干网络为例，具有  $K$  个 transformer 层的图像编码器  $V = \{V_i\}_{i=1}^K$ ，将图像  $I$  分割成  $M$  个固定大小的块，这些块

被投影到块嵌入  $E_0 \in \mathbb{R}^{M \times d_v}$  中。随后块嵌入  $E_i$  与可学习的类 (CLS) 标记  $c_i$  一起输入到第  $(i + 1)$  个 transformer 块 ( $V_i + 1$ )，并通过  $K$  个 transformer 块进行顺序处理，如式 2-10 所示。

$$[c_i, E_i] = V_i([c_{i-1}, E_{i-1}]), i = 1, 2, \dots, K \quad (2-10)$$

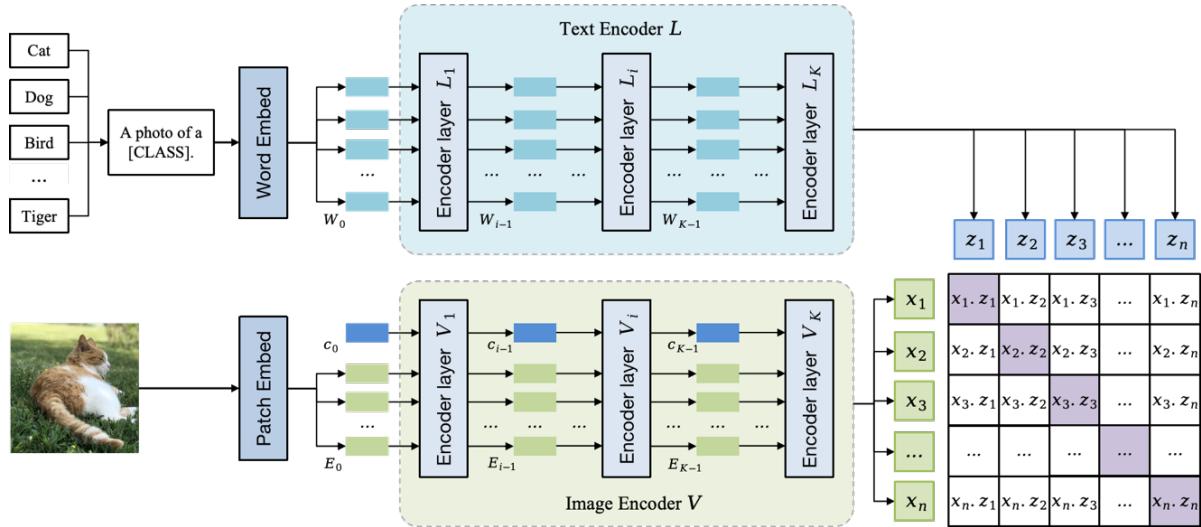


图 2-2 CLIP 整体架构

为了获得最终的图像表示  $x$ ，最后一个变换层 ( $V_K$ ) 的类标记  $c_K$  通过 *ImageProj* 方法投影到公共 V-L 潜在嵌入空间，如式 2-11 所示。

$$x = \text{ImageProj}(c_K)x \in \mathbb{R}^{d_{vl}} \quad (2-11)$$

对于文本编码，将手工制作的提示与目标数据集的类名称组合而成的文本描述（例如“a photo of a [CLASS]”）输入到 CLIP 模型的文本编码器进行标记并将其投影到文本嵌入  $W_0 = [w_0^1, w_0^2, \dots, w_0^N] \in \mathbb{R}^{N \times d_l}$  来生成文本描述的特征表示。在每个阶段， $W_i$  被输入到文本编码分支 ( $L_i + 1$ ) 的第  $(i + 1)$  个变换层，如式 2-12 所示。

$$[W_i] = L_i([W_{i-1}]), i = 1, 2, \dots, K \quad (2-12)$$

文本表示  $z$  是通过将最后一个转换器块  $L_K$  的最后一个 token 对应的文本嵌入通过 *TextProj* 投影到公共 V-L 潜在嵌入空间来获得的，如式 2-13 所示

$$z = \text{TextProj}(w_K^N)x \in \mathbb{R}^{d_{vl}} \quad (2-13)$$

第二步，将第一步中得到的图像特征  $x$  与文本表示  $z$  进行对齐，预测概率  $y$  的计算如式 2-14 所示。

$$p(y|x) = \frac{\exp(\text{sim}(x, z_y)/\tau)}{\sum_{i=1}^C \exp(\text{sim}(x, z_i)/\tau)} \quad (2-14)$$

其中  $C$  是目标数据集中的类别总数， $sim(\cdot, \cdot)$  表示余弦相似度计算， $\tau$  是 CLIP 的学习参数。

## 2.2 基于音视频的模型与方法

### 2.2.1 CLNF

CLNF<sup>[41]</sup>（Constrained Local Neural Fields）是一种运用于面部特征点检测（Facial Landmark Detection, FLD）的算法，它主要是把局部特征学习以及全局形状约束结合起来，以此提升检测的准确性与鲁棒性。CLNF 可精确追踪面部特征点，像眉眼、嘴角以及肌肉的细微运动等，还可以有效地捕捉和焦虑、抑郁等心理状态相关的微表情变化，比如短暂皱眉、嘴角下垂等。它在亚像素级定位方面的能力比传统方法要好，可量化情绪表达的细微差别，为心理健康评估提供可靠的数据支撑。该算法的全局形状模型可以根据头部姿态的变化进行自适应调整，比如偏转、倾斜等情况，而局部神经场对于光照不均以及部分遮挡，像眼镜、手势等问题，有着较强的容错能力，能保证在真实场景下有稳定的表现。另外 CLNF 的形态学参数（例如嘴角开合度、眼周肌肉活动等）和心理学量表（如 PHQ-9<sup>[6]</sup>）存在显式关联，其基于形状模型的特征提取更容易和临床指标相契合，契合医疗领域对算法透明性的要求。

CLNF 算法的核心原理覆盖局部特征学习（Local Neural Fields）、全局形状约束（Shape Constraints）以及迭代优化这三个部分。CLNF 先是针对每个特征点周围的局部区域运用神经网络来学习像素级的特征表示，网络输出每个位置的响应值（Response Map），以此表示该位置属于目标特征点的概率。接着凭借点分布模型（Point Distribution Model, PDM）来描述特征点的全局几何关系，保证检测结果符合人脸解剖结构，算法运用非参数级联回归（比如高斯过程回归），逐步优化特征点位置，把局部响应和全局形状对齐。最后借助坐标下降法或者牛顿法迭代调整特征点位置。

对于人脸输入图像  $I \in \mathbb{R}^{H \times W \times 3}$ ，采用 CLNF 算法输出形状向量  $s = [x_1, y_1, \dots, x_n, y_n]^T \in \mathbb{R}^{2n}$ ，其中  $n$  为特征点坐标的数量。

具体而言，给定参数化的人脸形状点分布模型表示和目标能量函数，如式 2-15 和 2-16 所示。

$$s = s_0 + Pp + Qq \quad (2-15)$$

式 2-15 中的  $s \in \mathbb{R}^{2n}$  表示形状向量（由  $n$  个特征点的  $(x, y)$  坐标组成）， $s_0 \in \mathbb{R}^{2n}$  为平均形状（训练集对齐后的均值）， $P \in \mathbb{R}^{2n \times k}$  和  $Q \in \mathbb{R}^{2n \times j}$  分别代表形状主成分矩阵（PCA 分解的基向量）和姿态变换矩阵（仿射变换参数）。 $p \in \mathbb{R}^k$  和  $q \in \mathbb{R}^j$  分别代表形状参数（控制全局变形）和姿态参数（如旋转、平移、缩放）。

$$E(p, q) = \sum_{i=1}^n [1 - f_i(s_i(p, q))] + \lambda p^T \Sigma^{-1} p \quad (2-16)$$

式 2-16 中的  $f_i(\cdot)$  表示第  $i$  个特征点的局部神经网络响应函数， $\lambda$  为正则化权重， $\Sigma$  是形状参数的协方差矩阵。

CLNF 算法的步骤流程如下：

第一步通过人脸检测器获取初始人脸区域并设置初始参数。

第二步进行局部响应计算，对每个特征点使用指定神经网络模型在领域  $\mathcal{N}(s_i^{(t)})$  内计算响应图，如式 2-17 所示。

$$f_i(x, y) = NeuralNet(I(x, y)), \forall (x, y) \in \mathcal{N}(s_i^{(t)}) \quad (2-17)$$

第三步为局部位移搜索，对每个特征点  $i$ ，求解最优位移，如式 2-18 所示。

$$\Delta s_i^{(t)} = argmin_{\Delta s} [1 - f_i(s_i^{(t)} + \Delta s)] \quad (2-18)$$

第四步对全局参数进行更新，通过加权最小二乘更新形状和姿态参数，如式 2-19 和 2-20 所示，其中  $W$  为对角权重矩阵，矩阵内元素为各个特征点的响应置信度。

$$p^{(t+1)} = p^{(t)} + (P^T W P + \lambda \Sigma^{-1})^{-1} P^T W (\Delta s - s_0) \quad (2-19)$$

$$q^{(t+1)} = q^{(t)} + (Q^T W Q)^{-1} Q^T W (\Delta s - s_0) \quad (2-20)$$

最后进行结果计算，能量函数的结果表示如式 2-21 所示。若  $|E^{(t+1)} - E^{(t)}| \geq \epsilon$ ，则需要重复步骤二至步骤四直至结果收敛。

$$E^{(t+1)} = \sum_{i=1}^n [1 - f_i(s_i^{(t+1)})] + \lambda (p^{(t+1)})^T \Sigma^{-1} p^{(t+1)} \quad (2-21)$$

## 2.3 本章小结

这一章节对心理健康检测与评估领域所涉及的各类基础知识展开了系统梳理，以此为后续研究构建起理论与技术方面的基础支撑。此章节对基于图片以及文本数据的深度学习技术于心理健康评估里的应用进行了详细说明，其中覆盖提示学习、多模态建模以及注意力机制等关键知识内容。针对多模态访谈记录的特点，还介绍了与之相关的模态提取算法，经由对这些基础知识的归纳总结，可为后续章节所提出的模型设

计以及系统实现给予必要的背景方面的支持以及技术上的指导。



## 第3章 基于多模态提示学习的心理健康问题分类预测

### 3.1 引言

心理健康数据涉及多种形式的信息，如图像、文本、访谈记录以及测评记录等。一方面，传统心理健康量表进行可从不同心理维度对个体状态给予精准评估，但由于量表题目数量普遍较多，这类测评手段存在时效性低、诊断周期长等问题。另一方面，患者自身的图像和文本数据，如面部表情、生活场景照片、心理自述等，暗藏着潜在的心理问题线索。但是这类数据在面对心理检测等复杂任务时，由于缺少合适引导普遍存在可解释性差、主观性偏见等状况，使得预测结果与真实情形出现误差甚至并不相符。因此，采用提示学习技术将量表信息与多模态图文数据加以引导融合，可以有效兼顾二者的诊断效率和精度，切实提高心理健康问题的预测效果。

本章设计基于提示学习的多模态心理健康问题分类预测模型，运用图像、文本、量表等多模态数据来为心理健康诊断给予辅助评估。此章选取常见的五种心理健康疾病当作分类对象，分别是抑郁症（Depression）、广泛性焦虑障碍（Generalized Anxiety Disorder，GAD）、睡眠障碍（Sleep Disorder）、双相情感障碍（Bipolar Disorder）和精神分裂症（Schizophrenia），且在此基础之上构建数据集。在模型算法方面，本章提出 VSCP-Net，这个模型涉及如下三个模块：

(1) 多尺度浅层视觉增强模块。由于预训练模型的权重参数被冻结，其内部图像编码器对多尺度特征的提取能力受到限制。针对这一问题，本章提出多尺度浅层视觉增强模块。该方法对于原始心理图片分别使用不同的尺寸大小进行分割，以多尺度维度进行特征提取，之后子分支的特征经过设计好的网络转化为多尺度提示，并与主分支的特征在浅层进行非侵入性融合，增强了视觉侧数据模态内的特征表达，弥补了单一尺度下模型捕捉心理线索能力的欠缺。

(2) 跨模态协同提示生成器。由于图文等非结构化数据容易受到主观偏差的影响，对于心理信息的理解性不足。为解决这个问题，本文提出跨模态协同提示生成器（Scale-Text Prompt Generator，STPG）。该模块将心理健康量表的结构化指标与非结

构化文本类别信息进行融合生成协同提示，增强视觉侧对于多模态数据的特征表达，提高模型对心理问题模态间的理解能力。

(3) 心理健康联合预测模块。由前两大模块得到的多尺度、跨模态的复合视觉信息与文本信息相结合，对指定心理疾病进行联合预测。

此外，本章将提出的 VSCP-Net 模型进行多项对比试验、消融实验和参数实验，验证了该模型对于心理健康问题预测分类的有效性。同时，VSCP-Net 模型将集成于本文第 5 章的多模态心理健康检测评估系统中，为心理问题的早期预测提供辅助判断依据。

## 3.2 数据集和数据预处理

### 3.2.1 数据集

在先前研究使用的心理健康数据集中，无效样本即阴性样本所占比例，要远远高于有效样本即阳性样本，使得数据集呈现出高度不平衡的特性。比如 NHANES<sup>[42]</sup> (National Health and Nutrition Examination Survey)，它是一项用于评估美国国民健康状况的全国性调查。在 2005 年到 2016 年期间，该数据集中阳性患者的占比不超过 10%。KNHANES<sup>[43]</sup> (Korea National Health and Nutrition Examination Survey) 是一项源自韩国普通民众的健康调查。在 2014 年、2016 年以及 2018 年，其阴性样本接近 95%。因此，本小节提出 MENTAL5 数据集，以此缓解样本分布极度不均衡的状况。

MENTAL5 数据集里为常见心理疾病的图片文本对数据。数据采集是由十位计算机科学与技术专业的本科生来完成，时间范围从 2023 年 9 月一直到 2024 年 2 月。其中图像数据规定为用户当天带有本人面部或者身体形态的生活图片，大小不超过 1M；文本内容是用户的内心自述或者心情日记，长度少于 50 字。MENTAL5 数据集覆盖五种心理疾病，分别是抑郁症、广泛性焦虑障碍、睡眠障碍、双向情感障碍以及精神分裂症，数据标注由三位计算机科学与技术专业的本科生完成，标注的依据是每种疾病的代表性表征。在先前的研究中，Menon 等人<sup>[44]</sup>利用大模型把目标类别分解成多种代表性特征，提升了模型的图像分类能力。本研究结合 ChatGPT<sup>[45-46]</sup>、天工 AI 以及多项心理健康量表对五种心理健康疾病进行拆解，挑选前三项强阳性特征作为标注依据，如表 3-1 所示。

表3-1 五种心理疾病的强阳性指标选择

疾病名称	量表选择	强阳性指标
抑郁症	PHQ-9 <sup>[6]</sup>	感到情绪低落、沮丧或绝望
		对做事失去兴趣或乐趣
		感觉自己毫无价值或有过度内疚感
广泛性焦虑障碍	GAD-7 <sup>[47]</sup>	感到紧张、焦虑或烦躁
		担忧过多且难以控制
		感到坐立不安或难以放松
睡眠障碍	ISI <sup>[48]</sup>	入睡困难（超过30分钟）
		夜间易醒或睡眠片段化
		早醒且无法再次入睡
双相情感障碍	YMRS <sup>[49]</sup>	情绪高涨或易激惹
		活动和精力增加
		言语急促和思维奔逸
精神分裂症	BPRS <sup>[50]</sup>	存在幻觉（如听到不存在的声音）
		存在妄想（如被害妄想）
		思维形式紊乱（逻辑混乱）

### 3.2.2 MENTAL5 数据集预处理

在深度学习所涉及的各类任务当中，数据预处理属于保证模型可呈现出良好性能的一项关键步骤。借助合理的方式来开展数据预处理工作，可切实有效地提升模型对于数据的理解水平，为训练流程提供一个有更强鲁棒性以及泛化能力的输入数据集。接下来将针对 MENTAL5 数据集的预处理展开介绍。

首先要对 MENTAL5 展开数据清洗工作。平台用户上传的心理数据大多时候会有缺失、重复以及异常值等状况，需要开展数据清洗来保证数据有可靠性与一致性。第一步是排除数据里的异常值，对不符合本章 3.2.1 节上传要求的图文以及文本描述当中的异常字符进行排除。第二步是检查冗余数据，把重复图片以及文本描述里的重复句段删除。第三步是补全部分文本数据里缺失的字段。最后一步是替换文本数据中直接表述病症名称的语句，以此保证模型可有效运行。

完成数据清洗工作以后，着手对 MENTAL5 开展数据标准化操作，来使其符合模型的输入格式要求。标准化步骤包括对输入图片的标准化和对文本语句的标准化。图片标准化包含尺寸变换、数据增强和归一化处理，具体方法如表 3-2 所示。文本标准化的流程为语言转换、小写化、分词和词形还原。

表 3-2 输入图片标准化处理

尺寸变换	数据增强	像素值归一化
使用双三次插值法（bicubic）将输入图片调整为 $224 \times 224$ 的分辨率	使用随机裁剪（random resized crop）和随机水平翻转（random flip）算法	使用的均值为 [0.48145466, 0.4578275, 0.40821073] 和标准差为 [0.26862954, 0.26130258, 0.27577711] 以匹配模型输入标准

### 3.2.3 心理健康量表预处理

在本研究中，引入心理健康量表有助于增强非结构化数据的可解释性。由于每个量表具有宏观结构相似，微观标准不同的特点，故需要对心理量表进行预处理以匹配模型输入需求。本章 3.2.1 节中提到，MENTAL5 是包含五种心理疾病的数据集，所以对于每种心理疾病选择相应的心理量表，如表 3-3 所示。量表的选择满足“一般人群适用”、“问题项（条目）数量不超过 15”、“无阴性条目”三个条件。

表 3-3 五种心理疾病对应量表选择

疾病名称	量表选择	问题项（条目）数量
抑郁症	PHQ-9	9
广泛性焦虑障碍	GAD-7	7
睡眠障碍	ISI	7
双相情感障碍	YMRS	10
精神分裂症	BPRS	8

本章的研究目标是对心理疾病进行宏观分类预测，因此仅对量表内容中的问题项部分进行处理。首先将原始量表中的问题项数量调整为 10 条，将具有近似语义的词汇进行合并，不同含义的行为进行拆分。例如，将 BPRS 量表中的第 5 项“激动或攻击

行为”拆分为“感到激动”和“具有攻击行为”。之后对所有文本进行数据标准化操作，具体操作为语言转换、小写化和分词以匹配模型输入。

### 3.3 模型与算法

本章的研究内容是基于多模态提示学习的心理健康问题分类预测。多模态数据含有潜在且丰富的心理信息，为这些数据构造合适的提示能够有效提高心理问题的识别。因此，本章提出基于多模态提示学习的心理问题分类预测模型 VSCP-Net 模型。模型的整体架构如图 3-1 所示，该模型由三个部分组成：

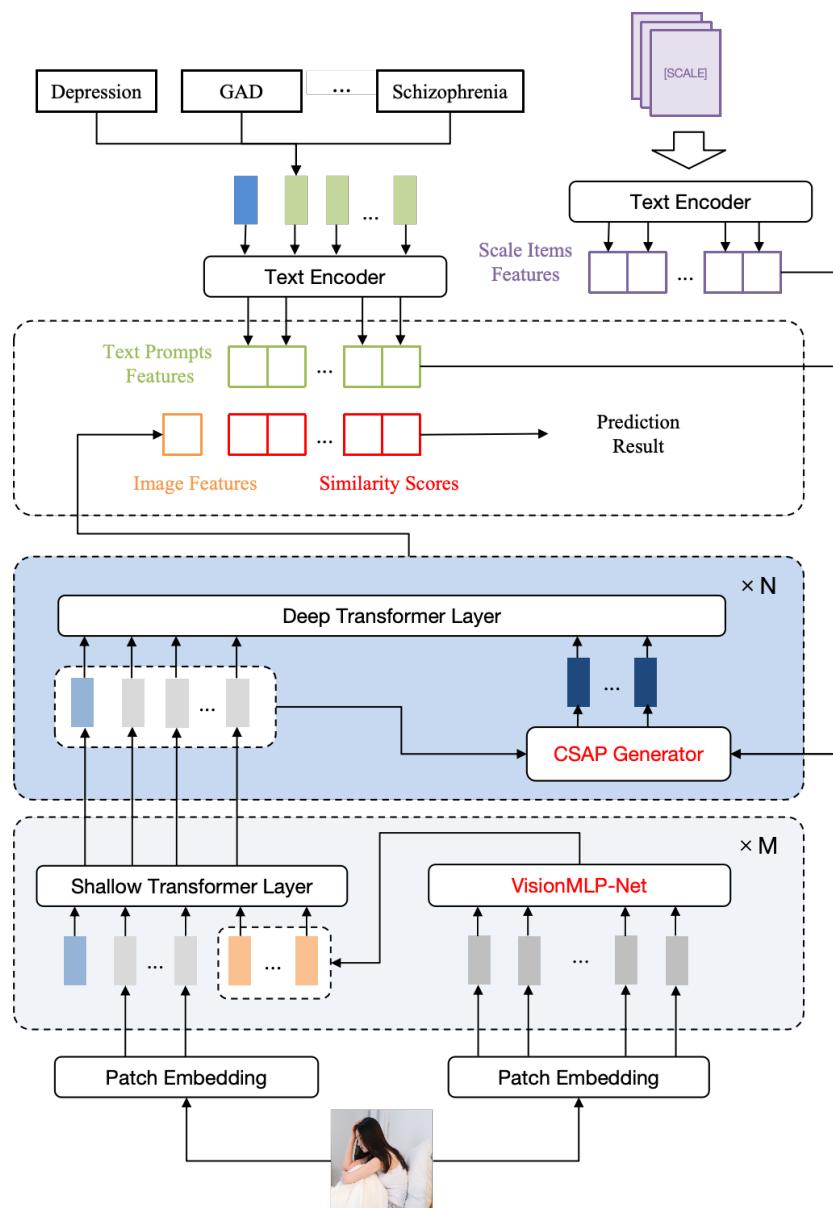


图 3-1 VSCP-Net 整体结构

(1) 多尺度浅层视觉增强模块。针对图像编码器在单一尺度下特征捕捉能力欠缺的问题，本章提出多尺度浅层视觉增强模块。该模块以多尺度形式对视觉浅层表示进行特征增强。在原有视觉分支基础上，额外引入基于多层次感知机网络的视觉分支，每个分支以不同的补丁块（patch）规模运行，最后基于提示学习的思想将额外视觉信息对原本分支的浅层视觉输入进行增强。该模块弥补单一视觉分支所不具备的多尺度视觉信息，扩大视觉模态内的感受野，提升视觉编码器对于视觉侧心理问题描述的模态内表达能力。

(2) 跨模态协同提示生成器 STPG。为了提高模型在识别心理图像时对于其他模态的理解性，本文提出跨模态协同提示生成器 STPG。该模块通过引入经典心理健康量表中具有代表性的心理疾病局部表征和基于类别信息的全局特征的动态文本提示，将经过多尺度浅层视觉增强模块得到的视觉增强表达与其他多模态数据生成的深层协同提示进行中期融合，扩展视觉信息与不同模态间的协同能力，提高模型对于心理问题模态间特征的识别效果。

(3) 心理健康联合预测模块。由多尺度浅层视觉增强模块和跨模态协同提示生成器 STPG 两大模块得到的复合视觉信息与文本表示相融合，对心理疾病进行联合分析预测。

### 3.3.1 多尺度浅层视觉增强模块

多尺度浅层视觉增强模块旨在增强视觉输入初期的特征表达。该模块涉及到的图像编码器，其骨干网络是 Transformer<sup>[37]</sup>的变体，即 ViT<sup>[40]</sup>模型。在先前的工作中，研究人员证明 patch 大小的粒度影响 ViT 的准确性和复杂性。回顾本文 2.2 节中对 Transformer 模型的介绍，ViT 同样是基于多头注意力机制。使用细粒度的 patch 会使 ViT 能够捕捉更加细微的特征，这对于心理健康问题的评估有更大的帮助。因此，本小节提出多尺度浅层视觉增强模块，在利用更细粒度补丁大小优势的同时，增强视觉模态内的特征表达，提高图像编码器对于心理问题特征的捕获和解释能力。

多尺度浅层视觉增强模块的整体结构如图 3-2 所示。该模块采用 ViT 模型作为骨干网络的图像编码器由  $K$  层编码层堆叠而成，首先基于 MLP 网络以不同尺度对原始视觉输入进行额外的特征提取，随后与原有的  $L_\alpha$  ( $1 \leq L_\alpha < K$ ) 层编码层之间的视觉嵌入进行浅层提示拼接，从而弥补视觉数据在单一尺度输入时因视角受限而损失的复合尺度信息。模块由两个视觉分支组成：基于 ViT 的主分支 Br-M 和基于 MLP 的互补分

支 Br-S。Br-M 采用粗粒度 patch 大小  $P_l$  和更大的嵌入维度进行更加复杂地编码；Br-S 以细粒度 patch 大小  $P_s$  运行，具有较少的编码器和更小的嵌入维度，最后两个分支进行浅层融合进行增强。

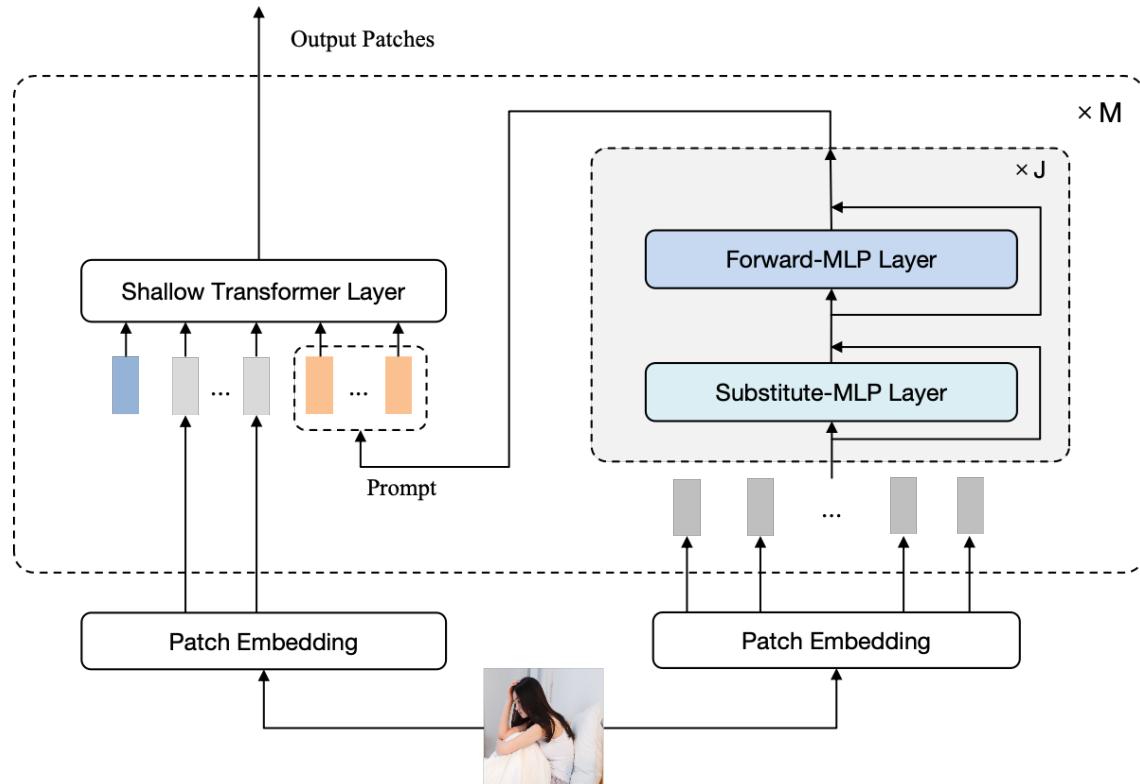


图 3-2 多尺度浅层视觉增强模块整体结构

图 3-2 所示的模块整体架构由两部分组成：基于 MLP 网络的细粒度视觉提示生成和多尺度视觉输入增强。接下来对各个部分作详细说明：

(1) 基于 MLP 的细粒度视觉提示生成。由于多尺度复合视觉信息在 Br-M 进行融合，因此优先介绍多尺度信息的提取过程。MLP 网络相比于之前的深度学习网络，具有轻量便捷的特点且能够提高模型的特征识别能力<sup>[47]</sup>，因此本小节在原有主干分支 Br-M (Main Branch) 的基础上进行改进，提出互补分支 Br-S (Substitute Branch)，称为 VP (VisionMLP) 网络。

VP 网络  $VP = \{VP_i\}_{i=1}^J$  由  $J$  个 Substitutor 层组成。Substitutor 层包含两个核心组件：SP (Substitute-MLP) 层和 FP (Forward-MLP) 层。他们分别负责空间信息编码和通道信息编码。FP 层与 Transformer 中的前馈神经网络层相类似，由两个全连接层和一个激活层组成，如图 3-3 所示。

为了充分利用图像的空间信息，SP 层分别沿高度、宽度和通道对输入数据进行处理。给定一个维度为  $C$  的输入  $X \in \mathbb{R}^{H \times W \times C}$ ， $O_{SP}$  是经过 SP 层后的输出， $O_{FP}$  是整个编码层的最后输出。Substitutor 层数据处理流程如式 3-1 和式 3-2 所示：

$$O_{SP} = SP(LayerNorm(X)) + X \quad (3-1)$$

$$O_{FP} = FP(LayerNorm(O_{SP})) + O_{SP} \quad (3-2)$$

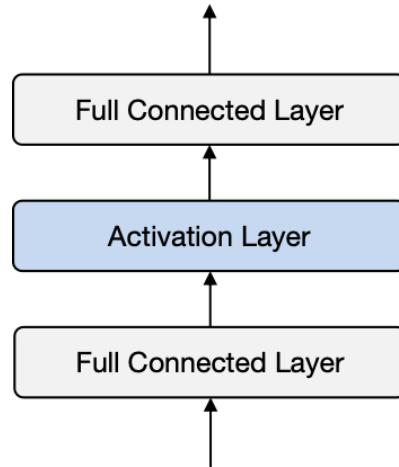


图 3-3 Forward-MLP 结构

与 Transformer 类似，第  $i$  个 Substitutor 层的输出作为  $X_{i+1}$  输入到第  $(i+1)$  个 Substitutor 层  $VP_{i+1}$ ，如式 3-3 所示：

$$X_{i+1} = VP_i(X_i) \quad (3-3)$$

在经过  $J$  个 Substitutor 层之后，VP 网络最终输出为  $O_{VP}$ ，并基于提示的思想将  $O_{VP}$  输入到前  $M$  层主分支中，作为额外提示增强原有视觉输入。

下面对 Substitutor 层中的 SP 模块进行详细的介绍。先前的研究<sup>[48]</sup> 将高度  $H$  和宽度  $W$  合并为统一的令牌信息，接收  $HW \times C$  的二维输入，而 SP 模块将空间信息拆分，接收  $H \times W \times C$  的三维表示。如图 3-4 所示，SP 模块由三个分支组成，每个通道沿高度  $H$ 、宽度  $W$  和通道  $C$  对信息进行编码，最后通过一个权重为  $\mathcal{W}_C \in \mathbb{R}^{C \times C}$  的全连接层生成相对于输入  $X$  的投影  $X_C$ ，并与经过高度分支和宽度分支处理得到  $X_H$  和  $X_W$  相加后投影得到最终输出。具体来说，给定一个通道维度为  $C$  的输入  $X \in \mathbb{R}^{H \times W \times C}$ ，以宽度分支为例，首先沿通道维度将  $X$  分割为  $S$  段， $X$  被切割成  $[X_{W_1}, X_{W_1}, \dots, X_{W_S}]$ ，其中  $X_{W_i} \in \mathbb{R}^{H \times W \times U}$  并满足  $C = U * S$ ， $U$  在宽度分支中表示  $W$ ，在高度分支中表示  $H$ 。接着对于每一个  $X_{W_i}$  将其宽度与通道维度进行置换得到  $[X'_{W_1}, X'_{W_2}, \dots, X'_{W_S}]$  并沿着通道维度进行串联作为置换操作的输出。然后使用一个权重为  $\mathcal{W}_W \in \mathbb{R}^{C \times C}$  的全连接层来混和宽度信息，同时将宽度与通道维度进行重新置换以恢复原始维度信息并输出  $X_W$ 。对于高度分支，采取与

宽度分支类似的置换操作并输出 $X_H$ 。最后，三个分支的输出进行相加并通过一个权重为 $\mathcal{W}_{SP} \in \mathbb{R}^{C \times C}$ 的全连接层得到 SP 模块的最终输出 $Z$ ，如式 3-4 所示：

$$Z = (X_H + X_W + X_C)\mathcal{W}_{SP} \quad (3-4)$$

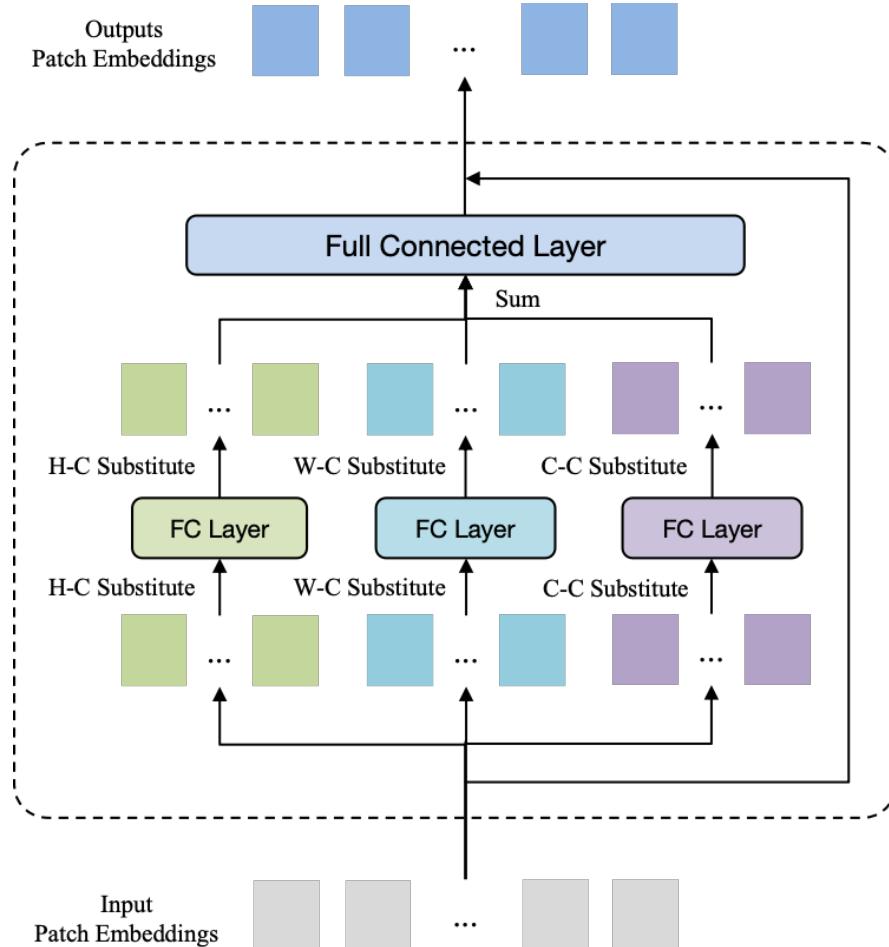


图 3-4 Substitute-MLP 结构

(2) 多尺度视觉输入增强。对于主分支 Br-M，具有 $K$ 层 transformer 层的视觉编码器 $V = \{V_i\}_{i=1}^K$ ，首先使用粗粒度 patch 大小 $P_l$ 将图像 $I \in \mathbb{R}^{H \times W \times 3}$ 切割成 $L \times L$ 个 patch 块，并将这 $L^2$ 个块以指定维度 $d_{v_l}$ 投影到块嵌入 $E_0 \in \mathbb{R}^{L^2 \times d_{v_l}}$ ，接着块嵌入 $E_i$ 与可学习的类（CLS）令牌标记 $c_i$ 进行拼接操作并为每个嵌入附加位置编码，最后将得到的复合嵌入输入到第 $i$ 个 transformer 层 $V_i$ ，对前 $L_\alpha$  ( $1 \leq L_\alpha < K$ ) 层 transformer 块进行顺序处理，如式 3-5 所示：

$$[c_i, E_i] = V_i([c_{i-1}, E_{i-1}]), i = 1, 2, \dots, K \quad (3-5)$$

在本小节所提出的模块中，经过互补分支 Br-S 处理得到输出 $output_{VP}$ 经过投影得到提示嵌入 $p_{VP}$ ，视觉嵌入 $E_i$ 和类令牌标记 $c_i$ 一并输入到第 $i$ 个 transformer 层 $V_i$ 。由于本小节提出的模块仅对浅层，即前 $L_\alpha$  ( $1 \leq L_\alpha < K$ ) 层进行处理，因此式 3-5 所表示的

流程需要在前  $L_\alpha$  层进行调整，如式 3-6 所示：

$$[c_i, E_i, \_] = V_i([c_{i-1}, E_{i-1}, p_{VP}]), i = 1, 2, \dots, L_\alpha \quad (3-6)$$

### 3.3.2 跨模态协同提示生成器

跨模态协同提示生成器的目标在于提高视觉信息对于其他模态数据的理解能力。之前的研究采用通用动态提示的方式加强语言信息对视觉信息的识别与解释<sup>[13]</sup>，但却未对具体的任务制作更加强关联性的提示<sup>[14]</sup>。因此，本小节提出 STPG 模块，通过使用不同模态之间的特征信息来制作任务强相关提示以达到不同数据源协同合作的目的。该生成器首先筛选与视觉信息具有正相关的前  $K$  个全局类别表征，接着提取相关心理量表中前  $J$  个强阳性局部特性，然后将二者进行中期融合生成模态间协同提示，最后与上一小节模块所生成的视觉增强特征进行深层融合，加强不同模态之间的协同能力。

跨模态协同提示生成器由两个部分构成：基于异构数据的双向提示组设计和跨模态感知提示融合生成。下面对每个部分作详细的说明：

(1) 基于异构数据的双向提示组设计。回顾本文 2.2.3 节的介绍，预训练的 CLIP 将所有模型参数进行冻结，仅通过微调手工制作的提示应用于各种下游任务中。近几年的很多工作都是基于 CLIP 模型的提示微调策略进行展开研究<sup>[12-17]</sup>。受到这些工作的启发，本小节提出基于异构数据构建粗细粒度的双向提示组。由于原始心理量表以不同的结构呈现，因此首先对相关量表进行统一格式预处理，这在本章的 3.2 节已进行相关介绍。之后给定含有  $H$  个心理健康病症的分类集  $C = \{c_i\}_{i=1}^H$ ，对于每个  $c_i$ ，经过预处理得到对应的格式化心理量表测评项集  $R_i = \{r_i^s\}_{s=1}^G$ ，其中  $G$  表示每个量表预处理得到的特征数量。最后，为心理病症分类集  $C$  和量表测评项集  $R$  构造具有全局性和局部性的双向提示组。提示组的设计采用两种方案： $t_{unified}$  和  $t_{specific}$ 。下面分别对两种方案进行介绍。

$t_{unified}$  是为同一集合中所有类别构造相同提示的方案。 $t_{unified}$  一共有两种模式，默认模式 EM (End Mode) 将类别置于最末端，其设计采用“双向动态提示+分类”组合方式，这种方式是 CLIP 所提出“手工提示+分类”方法的二次改进，在保留原本大分类的全局信息的情况下，扩展了基于量表测评项的细粒度表达，提高局部模块对于心理问题预测分类的准确性与可解释性。 $t_{unified}$  的 EM 模式设计如式 3-7 和式 3-8 所示：

$$t_{unified}^C = [V]_1 [V]_2 \cdots [V]_{M_C} [GLB_C LS] \quad (3-7)$$

$$\boldsymbol{t}_{unified}^R = [\nu]_1[\nu]_2 \cdots [\nu]_{M_R} [SUB_R EP] \quad (3-8)$$

式 3-7 的  $[GLB_C LS]$  是心理病症分类集  $C$  所表示的向量,  $[V] = [V]_{i=1}^{M_C}$  是一组可学习的与  $[GLB_C LS]$  具有相同维度的向量集合,  $M_C$  是一个超参数, 用于指定  $[V]$  中的向量数量。式 3-8 的  $[SUB_R EP]$  是心理量表测评项集  $R$  代表的向量,  $[\nu] = [\nu]_{i=1}^{M_R}$  与  $[V]$  类似, 具有与  $[SUB_R EP]$  相同维度的可学习向量集,  $M_R$  是规定  $[\nu]$  中的向量数量的超参数。

$\boldsymbol{t}_{unified}$  的另一种模式 MM (Middle Mode) 将类别放在文本中间, 如式 3-9 和式 3-10 所示。CLIP 所构造的“提示+分类”模板限定了提示与类别的相对位置。然而正如人们日常说话, 其关键词可能出现在文本中的任意位置。因此构造如式 3-9 和式 3-10 的设计, 增加动态提示的灵活性。

$$\boldsymbol{t}_{specific}^C = [V]_1 \cdots [V]_{\frac{M_C}{2}} [GLB_C LS] [V]_{\frac{M_C}{2}+1} \cdots [V]_{M_C} \quad (3-9)$$

$$\boldsymbol{t}_{specific}^R = [\nu]_1 \cdots [\nu]_{\frac{M_R}{2}} [SUB_R EP] [\nu]_{\frac{M_R}{2}+1} \cdots [\nu]_{M_R} \quad (3-10)$$

本小节提出的另一种提示构造设计  $\boldsymbol{t}_{specific}$  有助于弥补  $\boldsymbol{t}_{unified}$  的缺陷。 $\boldsymbol{t}_{unified}$  的提示设计通用于所有类别, 但可能会导致特定类别的分类准确度下降, 因此本小节构建类特定提示设计  $\boldsymbol{t}_{specific}$ 。 $\boldsymbol{t}_{specific}$  的设计灵感源自于 CoOp 模型<sup>[12]</sup>, 它针对不同的类别应用不同的可学习提示。具体来说, 对于第  $i$  和  $k$  ( $i \neq k$ ) 个大分类,  $\boldsymbol{t}_{specific}$  相对应的全局提示  $[U] = [U]_{n=1}^{N_C}$  和局部提示  $[u] = [u]_{n=1}^{N_R}$  存在如下关系:  $[U]_1^i [U]_2^i \cdots [U]_{N_C}^i \neq [U]_1^k [U]_2^k \cdots [U]_{N_C}^k$  以及  $[u]_1^i [u]_2^i \cdots [u]_{N_R}^i \neq [u]_1^k [u]_2^k \cdots [u]_{N_R}^k$ 。 $\boldsymbol{t}_{specific}$  与  $\boldsymbol{t}_{unified}$  同样具有 EM 和 MM 两种模式。

(2) 跨模态感知提示融合生成。本小节提出基于类别特征与量表表征的双向感知提示生成器 (Class-Scale Aware Prompt Generator, CSAP)。该生成器将由  $\boldsymbol{t}_{unified}$  或  $\boldsymbol{t}_{specific}$  策略生成的基于类别的全局提示与基于量表的局部提示进行融合并生成基于上述异构数据的跨模态感知提示, 进而对视觉侧深层信息进行协同增强。该模块的设计结构如图 3-5 所示。

CSAP 接受两个输入源: 基于类别特征的粗粒度提示  $\boldsymbol{t}_C$  和基于量表表征的细粒度提示  $\boldsymbol{t}_R$ 。由于每个病症分类  $c_i$  与对应的量表测评项集  $R_i$  有关, 与其他测评项集  $R_j$  ( $i \neq j$ ) 无关, 因此首先对  $\boldsymbol{t}_R$  进行人工分组, 提取与每个  $c_i$  有关的  $h_i$  个测评项  $r_i$  进行聚合生成  $\boldsymbol{t}'_R$ , 如式 3-11 所示:

$$\boldsymbol{t}'_R = GROUP(\boldsymbol{t}_R) \quad (3-11)$$

接着, 由于不同量表测评项集  $R_i$  含有测评项的数量  $h_i$  存在分布不均的情况 (如 SCL-90 量表含有 90 道测评项, 而 PHQ-9 量表仅有 9 道), 并且 CSAP 的计算复杂度

随着 $h_i$ 数量的提高而线性增加，故对分组后的 $t'_R$ 筛选出前 $K_N$ 个正相关的提示 $\hat{t}_j$ （ $1 \leq j \leq K_N$ ），如式 3-12 所示。式中的 $SELECT$ 采用人工筛查的方式进行判别。

$$\hat{t}_j = SELECT(t'_R), 1 \leq j \leq K_N \quad (3-12)$$

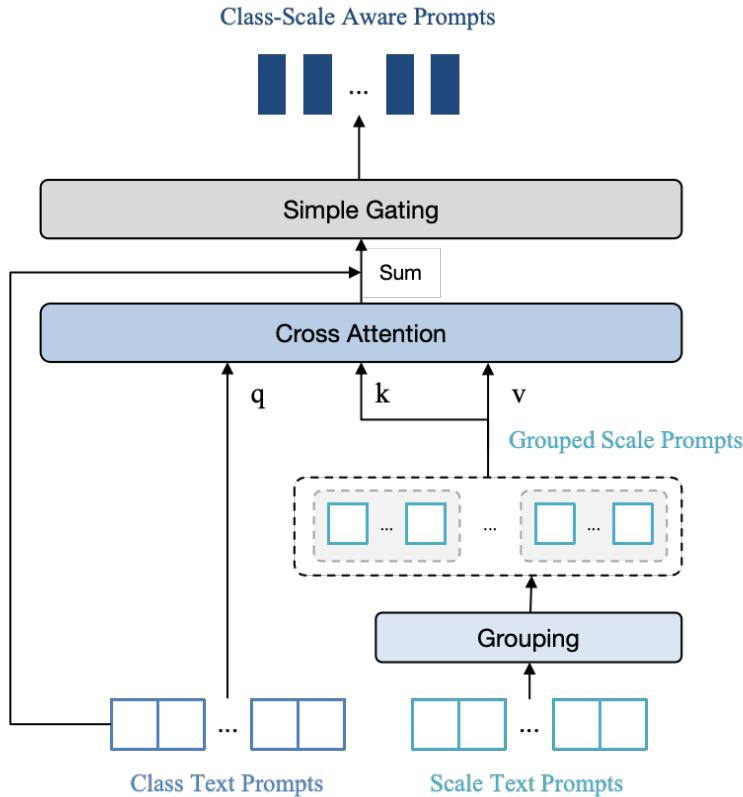


图 3-5 CSAP 生成器设计结构

然后，将分组筛选后得到的局部提示 $\hat{t}_j$ 经过一个全连接层得到交叉注意力机制的查询输入向量 $q_j \in \mathbb{R}^d$ （ $1 \leq j \leq K_N$ ）。键向量和值向量 $k_C \in \mathbb{R}^{d_k}$ 均由类别全局提示经过全连接层变换得到。对于每个 $q_j$ ，输出提示由式 3-13 计算得到：

$$t_l^j = softmax\left(\frac{q_j W_q (k_C W_k)^T}{\sqrt{d_k}}\right) k_C W_v, 1 \leq j \leq K_N \quad (3-13)$$

$$\tilde{p}_l^j = FC(t_l^j + q_j), 1 \leq j \leq K_N \quad (3-14)$$

式 3-13 中的 $W_q \in \mathbb{R}^{d \times d_k}$ 、 $W_k \in \mathbb{R}^{d \times d_k}$ 和 $W_v \in \mathbb{R}^{d \times d}$ 为本文 2.2.2 小节中介绍的注意力机制中的参数。式 3-14 中 $FC(\cdot)$ 的代表全连接层。

最后，将 CSAP 生成的最终跨模态感知提示 $\tilde{p}_{csap}$ 与视觉嵌入 $E_i$ 和类令牌标记 $c_i$ 输入到视觉主干 $L - Branch$ 的深层，如式 3-15 所示：

$$[c_i, E_i, \_] = V_i([c_{i-1}, E_{i-1}, \tilde{p}_{csap}]), i = L_\alpha, \dots, K \quad (3-15)$$

### 3.3.3 心理健康联合预测模块

以 CLIP<sup>[36]</sup>为基础的众多图文多模态模型<sup>[12-17]</sup>表现出在分类任务中的性能优势。本章所提出的 VSCP-Net 模型在此基础上，通过将视觉信息、文本信息、量表信息进行有机结合，提升模型对于多模态数据的特征表达。本模块将 3.3.2 小节得到的复合视觉信息与文本信息进行融合，得到心理健康疾病的预测分类。

形式上，由 3.3.2 小节得到的视觉特征  $x$  与文本类别表示  $z$  进行相似度计算，分类预测概率  $y$  的计算结果如式 3-16 所示。

$$p(y|x) = \frac{\exp(\cosine(x, z_y)/\tau)}{\sum_{i=1}^C \exp(\cosine(x, z_i)/\tau)} \quad (3-16)$$

其中  $C$  代表 MENTAL5 数据集中的疾病类别总数， $\cosine(\cdot, \cdot)$  为余弦相似度函数， $\tau$  是模型可学习参数。

## 3.4 实验与分析

为了对本章提出的 VSCP-Net 模型进行有效性评估，采用自建 MENTAL5 数据集并与多个基线模型进行实验。下面从实验环境、实验配置、评价指标、基线模型选择、多项实验结果与分析几个方面进行详细介绍。

表 3-4 实验环境配置

设备名称	参数配置
操作系统	Ubuntu 22.04.3
CPU	12th Gen Intel(R) Core(TM) i9-12900KF
主频	5.2GHz
内存	64G
显卡	NVIDIA GeForce RTX 3090
硬盘	500G
编程语言	Python 3.8
编程框架	Pytorch
编程环境	Anaconda 24.1.2

### 3.4.1 实验环境

本章实验在 CPU 为 Intel(R) Core(TM) i9 处理器、内存为 64G 的服务器上进行，实验所配置的操作系统为 Ubuntu 22.04.3，相关参数如表 3-4 所示。

### 3.4.2 实验配置

本小节对模型的基本参数配置和数据集配置进行介绍。

#### (1) 模型的基本参数配置

视觉编码器的骨干网络使用 ViT-B/16 架构，并选取  $16 \times 16$  的块大小（patch size）。训练集和测试集分别采用不同的批次大小（batch size），训练集初始设为 32，测试集设为 500，以平衡计算效率与内存占用。数据加载线程数设置为 8，以充分利用多核 CPU 资源，提升数据加载速度。优化器选用随机梯度下降法（SGD），初始学习率（learning rate）设置为 0.01，初始 epoch 配置为 50 次。学习率调度采用余弦退火策略（cosine），并在训练初始阶段设置了 1 个预热周期，预热策略为恒定学习率（constant），其值为  $1e-5$ 。

#### (2) 数据集配置

实验在 MENTAL5 自建心理数据集上完成。回顾 3.2.1 的介绍，传统心理数据集呈现正负样本高度不平衡的问题。因此本章提出 MENTAL5 数据集，采用人工采集的方式，保持常见心理疾病的样本分布情况相对均衡，同时增加罕见病症的样本数量，以提升模型训练的鲁棒性与准确性。训练集、验证集和测试集按照 7:1:2 的比例进行划分。数据集配置如表 3-5 所示。

表 3-5 数据集配置

参数	抑郁症	广泛性焦虑障碍	睡眠障碍	双相情感障碍	精神分裂症
样本数	1905	1575	723	504	423
样本占比	37%	31%	14%	10%	8%

### 3.4.3 评价指标

基于多模态提示学习的心理健康问题分类预测是一个多分类任务，并且数据集总

存在数据分布不平衡问题（常见疾病样本多，罕见疾病样本少）。因此本研究采用多维指标来评估模型的性能。定义疾病类别集合为 $C = \{c_1, c_2, \dots, c_h\}$ ,  $h$ 表示类别数量，测试集包含 $N$ 个样本，类别 $c_k$ 的样本数量为 $n_k$ 。评价指标定义如下：

(1) 准确率 (Accuracy) :

准确率是深度学习中最常使用的衡量指标之一，尤其适用于分类任务，它的计算公式如如 3-17:

$$ACC = \frac{TP}{TP + FP + FN + TN} \quad (3-17)$$

(2) 加权 F1 分数 (weighted-F1 Score)

定义各类别 F1 值的样本量加权平均，如式 3-18:

$$F1_{weighted} = \sum_{k=1}^h \left( \frac{n_k}{N} \cdot \frac{2 \cdot TP_k}{2 \cdot TP_k + FP_k + FN_k} \right) \quad (3-18)$$

其中 $TP_k$ ,  $FP_k$ ,  $FN_k$ 分别表示类别 $c_k$ 的真阳性、假阳性和假阴性样本数。该指标通过加权策略平衡类别分布不均的影响，更关注样本量较大类别的性能表现。

(3) 宏平均召回率 (macro-Recall)

计算各类别召回率的算术平均值，如式 3-19:

$$Recall_{macro} = \frac{1}{h} \sum_{k=1}^h \left( \frac{n_k}{N} \cdot \frac{TP_k}{TP_k + FN_k} \right) \quad (3-19)$$

该指标平等对待所有类别的漏检风险，在心理健康预测任务中，确保不同类别患者的检出率均被充分评估。

(5) 马修斯相关系数<sup>[53]</sup> (Matthews Correlation Coefficient, MCC)

马修斯相关系数为全局性评估指标，定义为式 3-20:

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (3-20)$$

该系数的值域为[-1,1]，其中 1 表示完美预测，0 等价于随机猜测。MCC 综合考量所有类别混淆矩阵元素，在类别分布高度不平衡时比准确率 (Accuracy) 更具鲁棒性。

(6) 归一化混淆矩阵 (Normalized Confusion Matrix)

构建行归一化矩阵 $M$ ，元素 $m_{ij}$ 表示真实类别为 $c_i$ 的样本被预测为 $c_j$ 的比例，如式 3-21:

$$m_{ij} = \frac{\sum_{s=1}^N \Pi(y_s^{true} = c_i \wedge y_s^{pred} = c_j)}{\sum_{s=1}^N \Pi(y_s^{true} = c_i)} \quad (3-21)$$

该矩阵直观揭示模型对症状相似疾病（如抑郁症与双相情感障碍）的误判模式，

为误差分析提供可视化依据，其中 $\Pi(\cdot)$ 表示指示函数。

### 3.4.4 基线模型

为验证模型有效性，本章提出的 VSCP-Net 模型与六个基线模型展开对比实验。依据采用提示学习技术形式的不同，所有基线模型被分为基于静态提示的方法以及基于动态提示的方法这两类。

#### (1) 基于静态提示的方法

**CoOp<sup>[12]</sup>**: CoOp 模型是一种用于预训练视觉语言模型的上下文优化办法，凭借自动化提示工程来提高下游任务的适应能力。它的核心想法是把提示模板里的上下文词换成可学习的连续向量，以此避免手动设计提示模板时出现的低效情况。CoOp 采用可学习提示的设计，能以提示微调的方式从一般物体识别转移到心理健康症状识别。

#### (2) 基于动态提示的方法

**CoCoOp<sup>[13]</sup>**: CoCoOp 属于对 CoOp 的改进举措，主要针对静态提示于新类别方面泛化能力欠缺的状况加以处理。它的核心创新之处在于引入条件提示学习，借助轻量级神经网络为每一幅输入图像生成动态的、与实例相关的提示向量，以此提高模型对于类别变化的适应能力。CoCoOp 是在 CoOp 的基础之上，把上下文向量和图像特征相互结合，生成输入条件化的提示，让提示可依据不同实例进行动态调整。CoCoOp 模型的整体设计情况如图 3-6 所示。

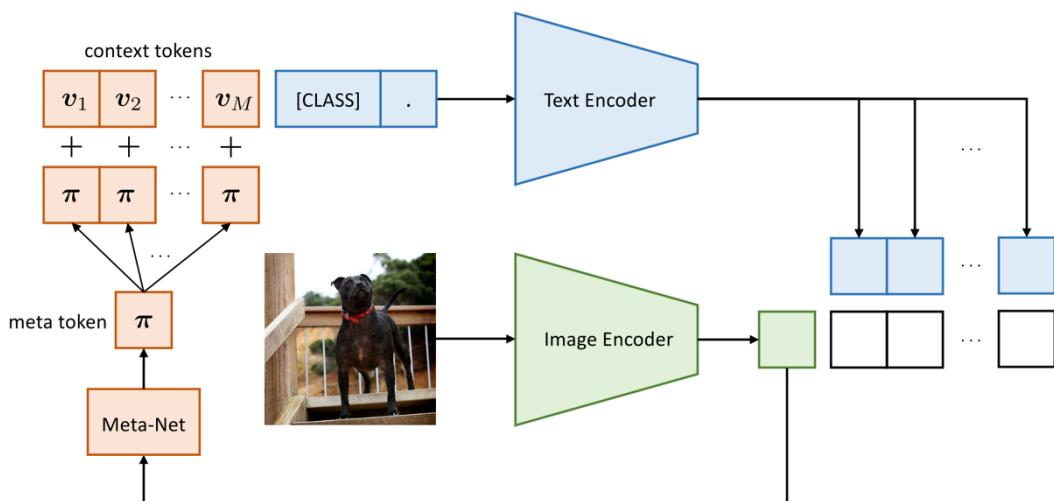
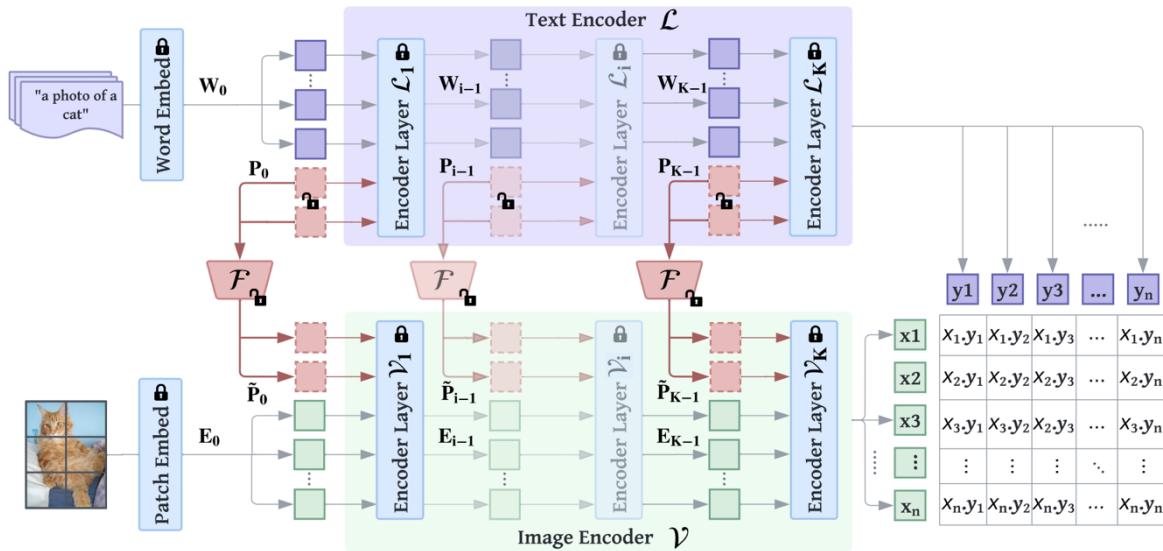


图 3-6 CoCoOp 模型<sup>[13]</sup>

**MaPLe<sup>[14]</sup>**: MaPLe 借助在 CLIP 的视觉以及语言分支里同步开展协同提示学习，

以此提高跨模态对齐能力。其核心设计囊括以下两方面：其一为深度分层提示，也就是在视觉和语言编码器的多个层级中各自引入可学习的提示。其二是视觉-语言提示耦合函数，该函数借助线性投影把语言提示转化为视觉提示，推动模态间的梯度传播以及协同优化。此方法的创新之处在于运用多模态联合提示，可提升模型对于像心理健康状况这类复杂视觉概念的适应能力。MaPLe 模型的整体设计如图 3-7 所示。

图 3-7 MaPLe 模型<sup>[14]</sup>

IVLP<sup>[14]</sup> (Independent Vision-Language Prompting)：IVLP 属于 MaPLe 框架里的一种多模态提示学习变体。此模型会在视觉编码器的输入处或者中间层插入视觉提示向量，同时在语言编码器中给文本上下文引入独立的语言提示向量。IVLP 为后续模型所提出的视觉-语言提示耦合函数提供了对比方面的依据，突出了多模态联合提示在提高模型泛化能力当中的关键作用。

**VPT<sup>[16]</sup>：** VPT 模型是一种针对 ViT<sup>[40]</sup>的轻量级适配方式，该模型的核心思路是在预训练模型的输入端引入可进行训练的提示向量，而不必对整个模型权重作出调整。VPT 借助在 ViT 输入序列里插入一组额外的可学习参数，让原有的 Transformer 结构可以适应新的任务，并且维持原始权重不变。这种方法有效降低了微调成本，在多个视觉任务中呈现出较好的适应性。

**DPT<sup>[17]</sup>：** DPT 模型可提升 Transformer 在密集预测任务方面的能力。DPT 采用经过预训练的 ViT 当作主干网络，以此来提取全局特征，同时对不同模态的数据使用交叉注意力机制，提高视觉语言的特征融合能力。

### 3.4.5 对比实验

本章提出的 VSCP-Net 在 MENTAL5 自建数据集上比基于浅层提示的方法（CoOp、CoCoOp）和基于深层提示的方法（VPT、MaPLe、IVLP、DPT）表现出更好的性能。

表 3-6 提示参数配置

参数模式	提示长度	类别标记位置	是否采用定制提示
800	8	end	False
801	8	end	True
810	8	middle	False
811	8	middle	True
1600	16	end	False
1601	16	end	True
1610	16	middle	False
1611	16	middle	True

#### (1) 基于静态提示方法的对比实验

静态提示指的是模型有关提示部分的设计由提示（prompt）相关的超参数（hyper parameter）配置。回顾本章 3.3.2 小节中基于异构数据的双向提示组设计，VSCP-Net 模型在这一部分设计与基于静态提示的相关模型类似。因此本节首先与基于静态提示的 CoOp 模型进行实验对比，验证 VSCP-Net 提示设计的有效性。提示有关参数配置如表 3-6 所示。各组实验结果如表 3-7 所示。

本实验对 CoOp 与 VSCP-Net 模型在多种参数模式下的性能进行了系统性对比，涉及 Accuracy、weighted-F1、macro-Recall 及 MCC 四项核心评价指标。实验结果表明，VSCP-Net 在多数参数配置下展现出更优的综合性能。

VSCP-Net 在八组参数模式里的六组获得了最高的 Accuracy，其中最优 Accuracy 达到了 83.5%，相比 CoOp 的最高值 80.9% 高出不少。在模型稳定性这个方面，VSCP-Net 的 Accuracy 标准差是 3.0%，比 CoOp 的 4.7% 更优，它对于参数调整有着更强的鲁棒性。

在 weighted-F1 这个指标方面，VSCP-Net 在七组实验里比 CoOp 表现更出色。VSCP-Net 在参数模式 811 中达到了 82.5% 的峰值，相比 CoOp 的 73.4% 提高了 9.1 个百分点。

分点。针对类别均衡敏感的 macro-Recall 指标，CoOp 在参数模式 800 以及 1611 中的表现较为突出，不过 VSCP-Net 借助参数优化可缩小这一差距。

在 MCC 这一指标上，VSCP-Net 在五组参数模式里处于领先地位，其最高值是 74.7%，此为模式 811 的结果，相较于 CoOp 最优值 73.1%，即模式 1611 的情况，提升了 1.6%。在参数模式 1611 中，CoOp 的 MCC 为 73.1%，VSCP-Net 的 MCC 为 73.8%，二者相近，在该模式下两类模型的判别能力趋于相同。

表 3-7 各组参数模式的实验结果

参数模式	模型	Accuracy	weighted-F1	macro-Recall	MCC
800	CoOp	0.732	0.734	0.789	0.647
	VSCP-Net	<b>0.768</b>	<b>0.757</b>	0.735	<b>0.659</b>
801	CoOp	0.727	0.731	0.744	0.631
	VSCP-Net	<b>0.771</b>	<b>0.747</b>	0.645	<b>0.640</b>
810	CoOp	0.727	0.731	0.744	0.631
	VSCP-Net	<b>0.808</b>	<b>0.795</b>	0.712	<b>0.705</b>
811	CoOp	0.729	0.734	0.725	0.632
	VSCP-Net	<b>0.835</b>	<b>0.825</b>	<b>0.736</b>	<b>0.747</b>
1600	CoOp	0.668	0.669	0.727	0.577
	VSCP-Net	<b>0.735</b>	<b>0.726</b>	0.703	<b>0.664</b>
1601	CoOp	0.718	0.723	0.715	0.606
	VSCP-Net	<b>0.779</b>	<b>0.773</b>	0.724	<b>0.664</b>
1610	CoOp	0.726	0.728	0.804	0.652
	VSCP-Net	<b>0.782</b>	<b>0.761</b>	0.667	<b>0.659</b>
1611	CoOp	0.809	0.811	0.811	0.731
	VSCP-Net	<b>0.821</b>	<b>0.812</b>	0.696	<b>0.738</b>

## (2) 基于动态提示方法的对比实验

动态提示是指提示参与模型的训练过程，并与不同的输入特征进行融合。在本小节里，挑选了 CoCoOp、MaPLe、IVLP、VPT、DPT 这几个模型来开展对比实验，验证 VSCP-Net 模型视觉动态提示部分有效性，实验结果呈现于表 3-8。

本研究于相同实验环境之中，针对 VSCP-Net 以及其他五种模型的分类性能展开

了对比，所涉及的关键指标囊括准确率、加权 F1 值、宏观召回率以及马修斯相关系数。从表 3-8 可看出，VSCP-Net 跟 DPT 模型在多数指标方面呈现出一定优势，然而传统方法 VPT 却存在着较为十分突出的性能局限。

从综合判别能力方面进行分析，VSCP-Net 在准确率为 83.5% 以及 MCC 值为 0.747 的情况下处于首位。其 MCC 值相较于第二名 DPT 的 0.729 提升了 1.8 个百分点，这显示出在类别不平衡场景中，VSCP-Net 有更出色的预测一致性。DPT 的准确率为 81.5%，虽与 VSCP-Net 相近，但其 macro-Recall 为 79.3%，相比 VSCP-Net 的 73.6% 低 5.7 个百分点，这体现出 VSCP-Net 在少数类识别上存在相对优势。

表 3-8 动态提示对比实验结果

模型	Accuracy	weighted-F1	macro-Recall	MCC
CoCoOp	0.688	0.686	0.776	0.607
MaPLe	0.809	0.813	0.828	0.717
IVLP	0.701	0.706	0.742	0.601
VPT	0.625	0.688	0.525	0.451
DPT	0.815	0.816	0.793	0.729
VSCP-Net	<b>0.835</b>	<b>0.825</b>	0.736	<b>0.747</b>

在分类均衡性这个方面，MaPLe 凭借 82.8% 的 macro-Recall 值处于领先位置，不过它的 Accuracy 为 80.9%，MCC 是 0.717，这两个指标都比 VSCP-Net 低，它较高的召回率或许是以牺牲一部分分类精度作为代价的。与之相对比，VSCP-Net 在维持较高的 macro-Recall 的情况下，达成了 Accuracy 与 MCC 指标的同步优化，呈现出更为均衡的类别判别能力。

实验最终得出的结果清晰显示，VSCP-Net 借助融合视觉与语义协同提示机制这一方式，于 Accuracy 以及 MCC 这两项核心指标方面，均达成了当前阶段的最优水准，该设计切实有效地平衡了分类精度以及类别均衡性的需求。

### 3.4.6 消融实验

为了验证 VSCP-Net 模型各子网络的性能，本小节对 VSCP-Net 进行消融实验。实验分为对提示深度的消融验证和各模块消融验证。

### (1) 提示深度 (prompt depth) 消融验证

根据先前的研究<sup>[14-17]</sup>, 编码器提示深度  $L_\alpha$  的选择对模型的性能具有显著的影响。VSCP-Net 模型在  $l_{shallow}$  ( $1 \leq l_{shallow} < L_\alpha$ ) 层中对视觉侧数据进行模态内增强, 在  $l_{deep}$  ( $L_\alpha \leq l_{deep} < K$ ) 层中对跨模态信息采用模态间融合, 因此本小节选取不同的提示深度  $L_\alpha$  进行实验, 实验选择性能指标的最好的 811 参数模式。从图 3-8 中可以看出, 当  $7 \leq L_\alpha \leq 9$  时各项性能指标得到显著的提升, 尤其是 Accuracy、weighted-F1 和 macro-Recall 三个指标。然而, 随着  $L_\alpha$  的深度加大, 模型的性能敏感性提高, 这说明提示深度仅在合理区间才能有效提高模型性能。

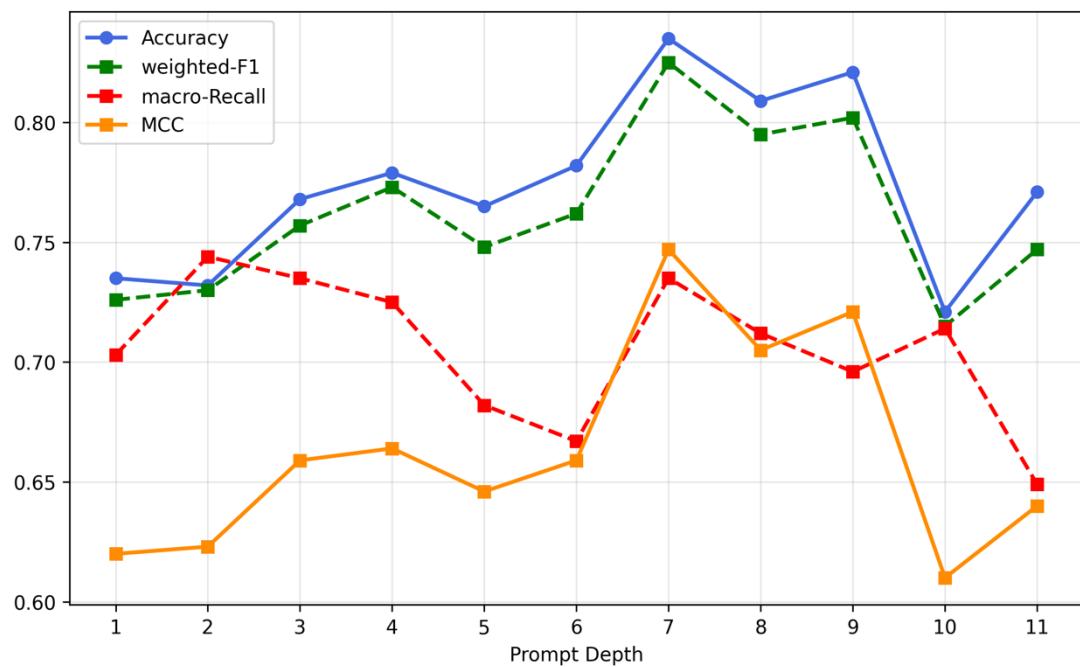


图 3-8 提示深度对于评价指标的影响

### (1) 提示深度 (prompt depth) 消融验证

根据先前的研究<sup>[14-17]</sup>, 编码器提示深度  $L_\alpha$  的选择对模型的性能具有显著的影响。VSCP-Net 模型在  $l_{shallow}$  ( $1 \leq l_{shallow} < L_\alpha$ ) 层中对视觉侧数据进行模态内增强, 在  $l_{deep}$  ( $L_\alpha \leq l_{deep} < K$ ) 层中对跨模态信息采用模态间融合, 因此本小节选取不同的提示深度  $L_\alpha$  进行实验, 实验选择性能指标的最好的 811 参数模式。从图 3-8 中可以看出, 当  $7 \leq L_\alpha \leq 9$  时各项性能指标得到显著的提升, 尤其是 Accuracy、weighted-F1 和 macro-Recall。然而, 随着  $L_\alpha$  的深度加大, 模型的性能敏感性提高, 这说明提示深度仅在合理区间才能有效提高模型性能。

### (2) 各模块消融验证

为了进一步验证本文提出的 VSCP-Net 模型中各模块的效用，下面对 VSCP-Net 模型进行消融实验。实验选择具有最高性能指标的 811 模式，共分为五组，介绍如下：

- ALL：包含所有子网络的完整 VSCP-Net 模型。
- V：去掉多尺度浅层视觉增强模块中的 VisionMLP 网络，视觉特征仅由骨干 ViT 提取，保留跨模态协同提示生成器 STPG。
- S：去掉 STPG 中的量表特征融合，仅使用文本类别信息参与协同提示生成，保留多尺度浅层视觉增强模块。
- SC：去掉协同提示生成器 STPG，仅保留多尺度浅层视觉增强模块生成浅层提示。
- VSC：去掉 VisionMLP 网络，去掉协同提示生成器 STPG，采用原生视觉特征和文本特征进行融合预测。

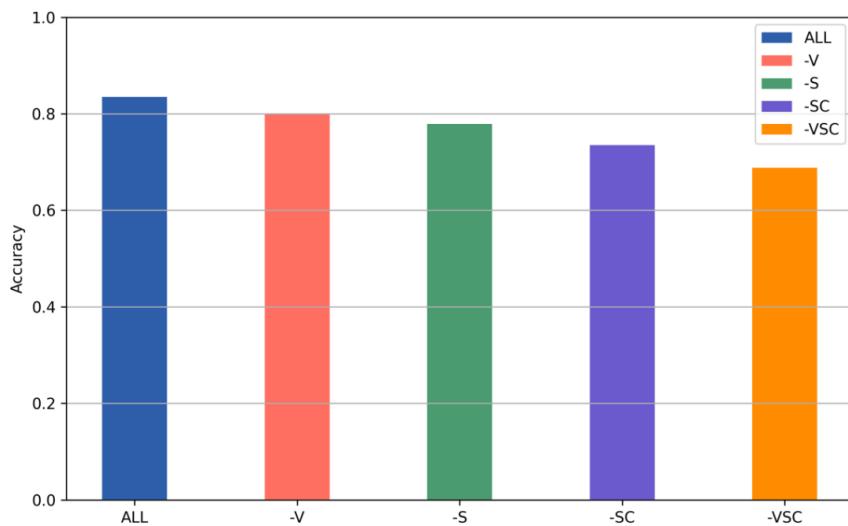


图 3-9 模块消融对于 Accuracy 的影响

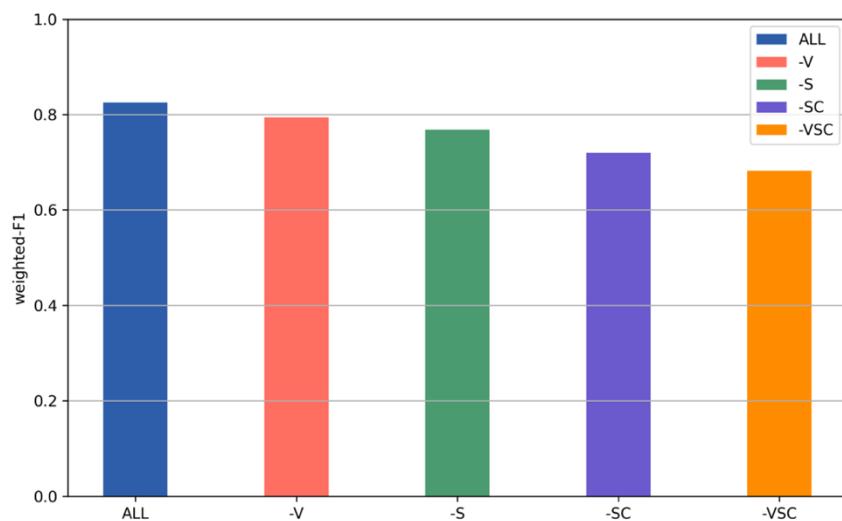


图 3-10 模块消融对于 weighted-F1 的影响

从图3-9至3-11中可以看出，VSCP-Net模型加入包含VisionMLP网络的多尺度浅层视觉增强模块和跨模态协同提示生成器STPG后，模型的Accuracy、weighted-F1和MCC三个指标得到有效的性能提升。

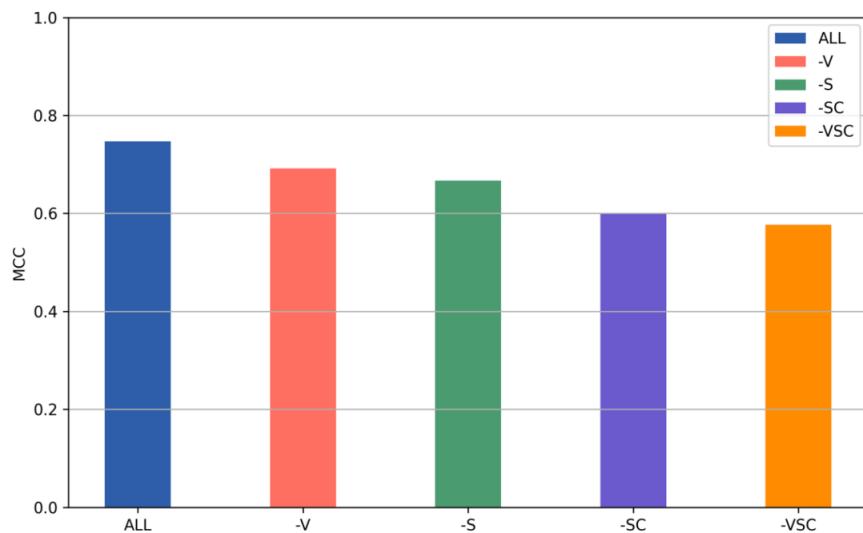


图3-11 模块消融对于MCC的影响

### 3.4.7 参数实验

在参数实验环节当中，本小节研究借助对学习率以及批次大小这两个关键超参数进行系统性对比，来剖析它们对模型性能所产生的影响，以此验证参数设置是否合理且有鲁棒性。各项参数实验所呈现的结果具体如下：

(1) 学习率 (learning rate)：学习率是梯度下降算法里的关键超参数，直接对模型参数更新步长起到控制作用。如果学习率设置得比较大，就会对模型学习稳定性造成影响，使得模型在不同结果间出现震荡情况；而要是学习率设置得比较小，那么模型的训练时间就会增加，致使收敛速度变得过慢。

(2) 批次大小 (batch size)：批次大小可决定单次参数更新时所使用的样本数量，其对梯度估计的准确性以及计算资源的利用率均会产生直接影响。如果将批次大小设置得过大，那么模型训练所需的计算成本便会明显增多；若批次大小设置得过小，模型则有可能陷入局部最优解或者无法实现收敛。

图3-12展示了不同学习率对于VSCP-Net性能的影响。可以看出，当学习率设置为0.01时，VSCP-Net各项指标最优。

图 3-13 是批次大小对 VSCP-Net 的影响。当批次大小设置为 32 时，各项性能指标达到峰值。由于模型接受图片和文本数据，因此批次大小过小（设置为 8, 16）会导致模型学习不充分；而设置过大则会降低服务器的计算效率，这对准确率等指标也是一种影响因素。

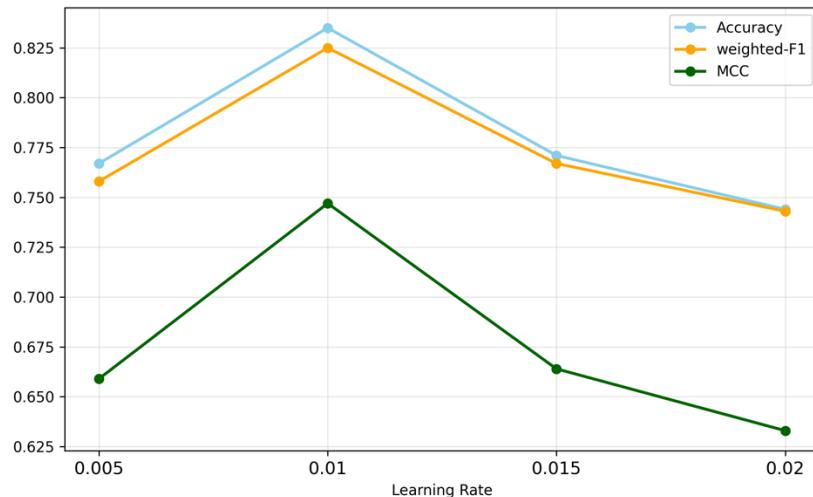


图 3-12 学习率对于模型性能的影响

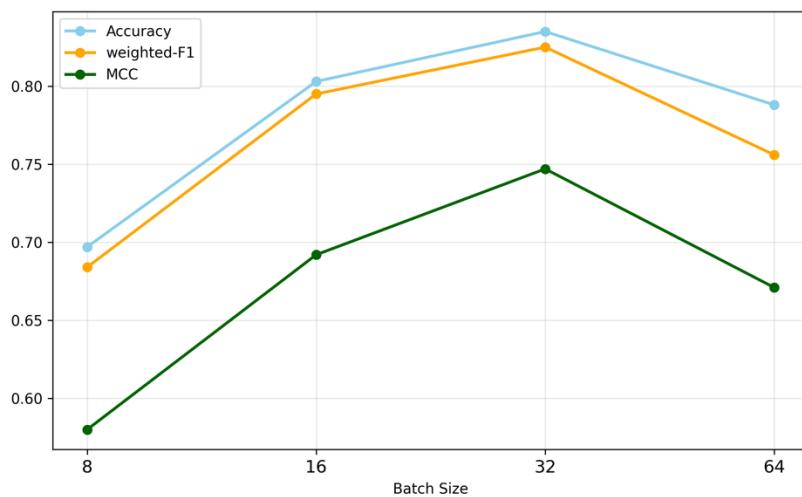


图 3-13 批次大小对于模型性能的影响

### 3.5 本章小结

本章提出了一种基于提示学习的多模态心理健康问题分类预测模型 VSCP-Net。通过引入多尺度浅层视觉增强模块和跨模态协同提示生成器，本模型有效提升了多模态数据特征提取和融合的能力。同时，设计的心理健康联合预测模块进一步提高了分类的精度和模型的解释性。本章详细阐述了模型的设计思想、技术实现以及实验结果，

实验表明，VSCP-Net 在心理健康问题的分类预测中具有较好的性能优势，为多模态心理健康研究提供了一种创新的解决方案。



## 第 4 章 基于多模态访谈记录的心理病症等级评估

### 4.1 引言

心理病症的等级评估对于识别个体心理健康风险以及制定科学的干预策略而言是一项十分关键的工作，尽早发现并准确诊断是改善患者生活质量的关键所在。然而传统的评估方法主要依靠临床面谈以及问卷调查，如症状自评量表 SCL-90<sup>[5]</sup>。这些方法虽说有一定效果，可能存在时间成本较高以及主观性较强等限制。

近些年来，多模态数据分析以及深度学习技术开始兴起，这为心理病症评估给予了全新的解决办法。音频信号里的语音特征像语调、语速、音量等，以及视频信号中的面部表情、姿态变化，都被视作与心理状况紧密相关<sup>[54]</sup>，多模态方法凭借融合多种维度的特征，对提高心理病症等级评估的鲁棒性和准确性有帮助<sup>[55]</sup>。

在本章节的研究中，以多个公开数据集作为基础，以最具代表的抑郁症和焦虑症<sup>[1-2]</sup>作为检测目标，通过分析多模态患者访谈记录，提出了一种心理病症等级评估模型 MPCN。此模型集成了深度学习以及多模态特征融合等相关技术，可为心理诊断与干预给予相应的技术支撑，对心理健康领域的研究以及应用起到推动作用。

在本章之中，于 4.4 小节针对所提出的 MPCN 开展了一系列对比实验、消融实验以及参数实验，以此来验证该模型针对抑郁症、广泛性焦虑障碍等心理病症严重程度评估的有效性。MPCN 模型也被应用于第 5 章所提出的多模态心理健康检测评估系统里，可为潜在心理患者给予诊断建议以及个性化心理疏导方案。

### 4.2 数据集与数据预处理

#### 4.2.1 数据集

许多研究人员和医院致力于研究并发布数量更大和质量更高的心理数据集<sup>[56]</sup>。然而，适合使用基于深度学习方法的公开数据集却十分有限，过往研究中最广泛使用的心理数据集是 DAIC-WOZ<sup>[57,58]</sup> 和 E-DAIC<sup>[59]</sup>。

DAIC-WOZ 数据集包括从 142 名患者的 189 次临床访谈中收集的 50 小时数据，涵

盖音频、视频和文本等多种模态。受试者在模拟的人机交流场景中与虚拟助手进行对话，虚拟助手对患者进行采访并识别精神疾病的语言性和非语言性指标。该助手是具有动画效果的虚拟机器人而不是临床医生，并且该机器人由另一个房间的真人控制。该数据集此前曾用于 AVEC 2017<sup>[60]</sup>（Audio/Visual Emotion Challenge 2017），该比赛采用 PHQ-8<sup>[61]</sup>（Patient Health Questionnaire-8，PHQ-8）评分对数据集进行抑郁分类。

另一个广泛使用的公开数据集是 E-DAIC，它是 DAIC-WOZ 的扩展版本。该数据集于 AVEC 2019<sup>[62]</sup>（Audio/Visual Emotion Challenge 2019）中提出，包含焦虑、抑郁和创伤后应激障碍等心理困扰状况的临床访谈的音视频记录，并且所有的访谈数据都是由基于 AI 的机器人采集而不是人类进行。与 DAIC-WOZ 相比，E-DAIC 的数据规模有所增大，受试者数量增多，包含了不同性别、年龄以及文化背景的多样化人群。数据集中一共有 275 名受试者，其中 163 名受试者被用于训练集，56 名受试者被用于开发和测试集。

近几年的研究显示，DAIC-WOZ 可用于 GAD 等其他心理病症的评估<sup>[63]</sup>。GAD 患者的语言模式大多时候表现为高频使用像“可能”或者“也许”这样的不确定性词汇，以及灾难化表述，并且这些患者大多时候随着着频繁的小动作，比如摸脸、抖腿，以及紧张性微笑。这样的特征可依靠 DAIC-WOZ 的音频、视频数据来进行分析捕捉。对于抑郁症，是用本文数据集标记的 PHQ-8 分数来进行评估，对于 GAD，使用 GAD-7 量表对访谈记录进行评分，数据标注是由五位计算机专业的本科生完成的。

### 4.3 模型与方法

本章的研究工作是基于多模态访谈记录的心理病症等级评估。回顾本文第三章的介绍，采用图文、心理量表等多模态数据可以对常见心理疾病进行宏观分类，但却不足以对疾病作更细致的测评。因此本章提出基于多模态访谈记录的心理病症等级评估模型 MPCN，旨在对常见心理疾病的严重程度作诊断，并基于多项经典量表分数作为结果预测。如图 4-1 所示，整体模型包含如下六个模块：

（1）音频特征提取模块：针对访谈记录的音频信息，本模型采用 openSMILE<sup>[64]</sup> 提取音频特征。

（2）视觉特征提取模块：针对访谈记录的视觉信息，本模型使用 CLNF<sup>[41]</sup> 算法提取视觉特征。

（3）文本特征提取模块：针对访谈记录的文本信息，本模型运用 RoBERTa<sup>[65]</sup>

(Robustly Optimized BERT Pretraining Approach) 模型提取文本特征。

(4) 跨尺度信息构建：针对多模态数据的序列特点，本方法提出 MSCM 模块，用于生成不同模态的多语义表示，构建后续特征融合的多层次输入结构。

(5) 复合式金字塔协同网络：针对心理访谈数据集的结构特点，本模型提出复合式金字塔协同网络，对多模态信息特征进行跨尺度、多层次的深度融合。

(6) 心理病症等级评估：该模块结合多种分类与回归指标，并基于经典量表分数给出心理疾病等级的预测结果。

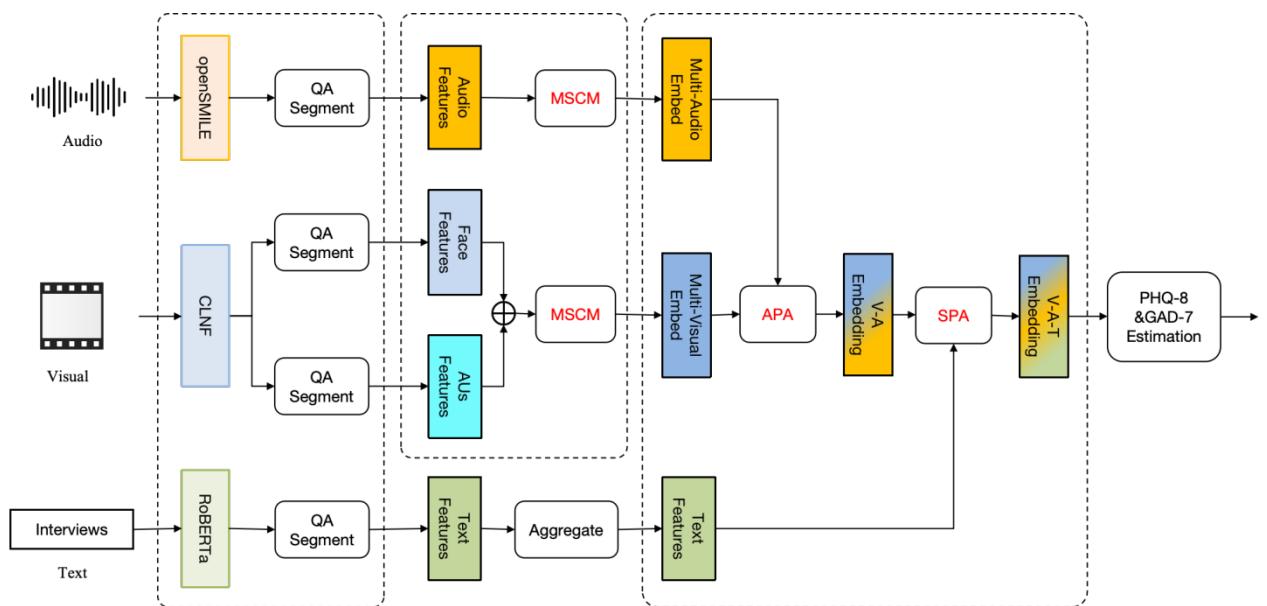


图 4-1 MPCN 整体结构

### 4.3.1 访谈录音音频特征提取

在这一小节当中，采用 openSMILE<sup>[64]</sup>工具来对访谈记录的音频数据展开特征提取工作。openSMILE 是由德国奥尔登堡大学以及慕尼黑工业大学共同研发的音频信号提取工具，其被设计用来处理语音、进行情感分析、开展音乐信息检索等诸多与音频相关的任务。这一小节先是运用 openSMILE 工具输出访谈录音音频模态的低级特征，随后依据数据集转录文本文件把低级特征划分成多个答案片段。

具体来说，给定音频信号序列  $S = [s_1, s_2, \dots, s_N]$ ， $N$  表示音频  $S$  的总采样点数。第一步使用预定义特征集  $\mathcal{F}$  输出帧级特征矩阵  $F \in \mathbb{R}^{T_a \times D_a}$ ，其中  $T_a$  代表访谈记录的总时间步， $D_a$  表示帧级维度，特征集  $\mathcal{F}$  用于指定提取的低层描述符。

第二步，按照 DAIC-WOZ 数据集中转录文本文件问答起止时间，将  $F_a$  分割成  $M$  组问答对，并选择其中的回答片段得到答案特征  $R_a \in \mathbb{R}^{M \times L_a \times D_a}$ ，如式 4-1 和式 4-2 所示， $f_i$  为第  $i$  个问答对特征， $r_i$  表示第  $i$  个答案特征，且满足  $r_i \subset f_i$ 。

$$F = [f_1, f_2, \dots, f_M] \quad (4-1)$$

$$R_a = [r_1, r_2, \dots, r_M], r_i \subset f_i \quad (4-2)$$

openSMILE 有多种预定义的特征集，在本小节当中使用的是 eGeMAPS<sup>[66]</sup>，即扩展的日内瓦简约声学参数集。eGeMAPS 是专门针对语音生理学以及情感分析所设计的一种音频特征集，在情感计算、心理健康评估以及人机交互等诸多领域都有着广泛的应用。

### 4.3.2 访谈记录视觉特征提取

考虑到受访者隐私安全方面的因素，本小节运用 CLNF<sup>[41]</sup> 算法所提取的视觉特征。CLNF 是一种将局部神经场与全局形状模型相结合的面部关键点检测算法，其借助对局部纹理以及面部形状信息进行建模，可在图像、视频等各类视觉数据当中精准地检测和跟踪面部关键点。在心理健康检测任务里，CLNF 是一种有高效性与可靠性的特征获取途径。本模型借助 CLNF 算法分别提取视频帧序列的多维人脸关键点信息以及面部动作信息。

人脸关键点指的是分布于人脸区域的特定的点，其作用在于描述面部的几何结构以及形状方面的信息，这些关键点一般囊括面部特征像眼睛、眉毛、鼻子、嘴巴等部位的边缘点以及中心点。动作单元（Action Units, AUs）是面部表情编码系统（Facial Action Coding System, FACS）里的基本单元，动作单元用来表示特定面部肌肉所呈现出的激活状态。

具体而言，给定视频访谈记录单帧图像  $I \in \mathbb{R}^{H \times W \times C}$  和面部初始检测框  $B = (x, y, w, h)$ ，其中  $H$ 、 $W$  和  $C$  分别表示图像的高度、宽度和通道数， $(x, y)$  表示检测框  $B$  的左上角坐标， $w$  和  $h$  代表检测框  $B$  的宽度和高度。

针对人脸关键点信息，本文使用 CLNF 提取 3D 关键点坐标，如式 4-3 所示。

$$X_{3D} = \{(x_k, y_k, z_k)\}_{k=1}^K, X_{3D} \in \mathbb{R}^{K \times 3} \quad (4-3)$$

式 4-3 中的  $K$  表示面部关键点的数量， $(x_k, y_k, z_k)$  代表第  $k$  个关键点在 3D 空间中的位置信息。

对于面部表情动作单元，采用 CLNF 算法输出每帧图像的动作单元强度向量，如

式 4-4 所示。

$$\mathbf{AU} = [AU_1, AU_2, \dots, AU_n], \mathbf{AU} \in \mathbb{R}^n \quad (4-5)$$

式 4-4 中表示提取的动作单元的数量,  $AU_i$  表示第  $i$  个动作单元的强度, 即面部肌肉的激活程度。本模型筛选关键动作单元作为视觉特征输入, 选取列表见附录。同时对面部关键点特征和面部动作特征进行初步融合得到第  $j$  个时间步的视觉特征, 如式 4-5 所示。

$$R_v^{(j)} = \text{Concat}(X_{3D}, \mathbf{AU}) \quad (4-5)$$

视觉模态采样频率与音频模态保持一致, 二者在时间步长度这一维度上保持对齐与同步。因此本小节采用与 4.3.1 小节类似的处理方式, 将帧级视觉表示变换为片段级特征。

### 4.3.3 访谈记录文本特征提取

本小节使用 RoBERTa<sup>[65]</sup> 模型对访谈记录的文本信息进行特征获取。RoBERTa 是由 Facebook AI 提出的自然语言处理 (Natural Language Processing, NLP) 模型, 它是基于 BERT<sup>[67]</sup> (Bidirectional Encoder Representations from Transformers) 的改进版本。RoBERTa 在各种 NLP 任务中强大地捕获上下文信息和语义细微差别的优势。本小节首先将访谈记录文本分割成对应于每个问题的答案集合, 随后将分段后的访谈文本输入到 RoBERTa 并提取模型最后一层的表示, 为每个访谈记录生成片段级文本特征。

形式上, 第一步将访谈记录的转录文本按照文件中的起止时间分割成  $M$  个答案文本。如式 4-6 所示,  $\mathcal{V}$  为指定语言词汇表集,  $L_t$  表示文本  $T_i$  的长度。

$$\mathbf{T} = [T_1, T_2, \dots, T_M], \mathbf{T} \in \mathcal{V}^{M \times L_t} \quad (4-6)$$

第二步, 对于每一个答案文本输入序列  $T = [t_1, t_2, \dots, t_{L_t}]$ , 将其输入到预训练的 RoBERTa 生成文本特征矩阵  $H = [h_1, h_2, \dots, h_{L_t}]$ , 如式 4-7 所示。其中  $T \in \mathcal{V}^{L_t}$  且  $H \in \mathbb{R}^{L_t \times d_t}$ ,  $h_j$  是第  $j$  个位置的上下文表示向量,  $d_t$  为 RoBERTa 的隐藏层维度。

$$[h_1, h_2, \dots, h_{L_t}] = \text{RoBERTa}([t_1, t_2, \dots, t_{L_t}]) \quad (4-7)$$

最后进行特征聚合, 取 [CLS] 标记的输出  $h_{CLS}$  作为对应于文本序列  $T$  的全局特征。

### 4.3.4 跨尺度信息构建

心理访谈记录数据是由众多“问答对”构成的, 受试人员依据采访者所提出的问

题，给出符合自身情况的各异回答。其中每段回答的帧级特征可提供细粒度的动态信息，可捕捉受访者心理状况的瞬间变化，回答片段的高维特征可提供语义以及模式总结，可用于判断该回答期间患者的情绪状态。而提取多个片段特征所组成的针对访谈记录的全局特征，可从不同模态角度反映患者整体的心理状况，基于这些特点，模型可从不同的时间尺度以及模态间的信息交互里提取综合特征，提升心理健康任务的检测精度。本小节提出跨尺度信息构建模块 MSCM，此模块将短期性的情绪变化与长期性的行为趋势相结合，对跨尺度、多模态的时序特征进行建模。MSCM 对回答片段内的帧级特征进行逐层降采样与层次融合，可将长时间依赖建模任务转变为低分辨率上的短距离依赖，在有效降低计算复杂度的同时提高模型对心理健康状态特征的捕捉能力，还为后续模块提供高效便捷的输入，MSCM 的设计结构如图 4-2 所示。

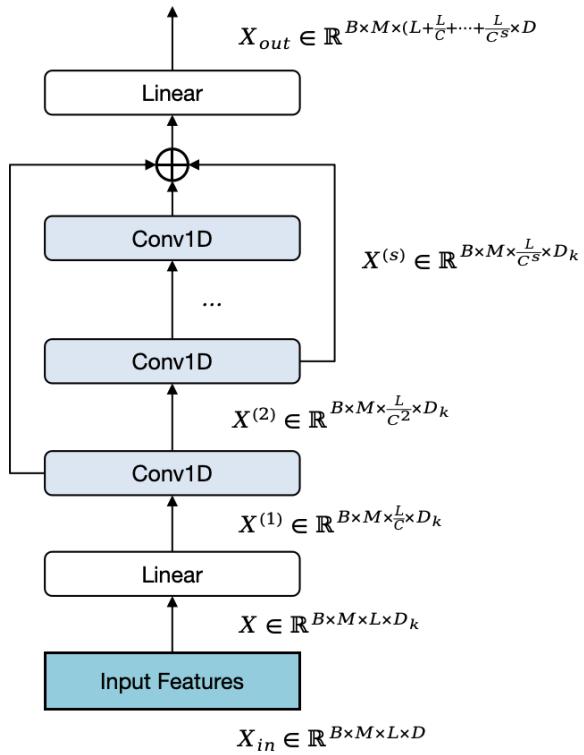


图 4-2 MSCM 设计结构

由于音频与视频信息的采样率相同，二者在时间步维度上具有对齐的特点。下面以音频信息为例进行介绍，视频数据仅第四维信息有差别，其余维度与音频保持相同。

形式上，给定帧级音频特征序列  $X \in \mathbb{R}^{B \times M \times L \times D_a}$ ，其中  $B$  为批次大小， $M$  是访谈记录的回答片段数量， $L$  为每一片段的帧数， $D_a$  代表每一个时间步的维度。第一步，通过多个降采样层生成不同尺度的节点。MSCM 对每层的输入序列应用核大小和步长均

为  $C$  的一维卷积操作，每次卷积操作将序列长度从  $L^{(s-1)}$  降低到  $L^{(s)} = \frac{L}{C^s}$ 。如式 4-8 所示，其中  $s$  表示当前尺度， $X^{(s)} \in \mathbb{R}^{B \times M \times \frac{L}{C^s} \times D_a}$  为第  $s$  层生成的节点。

$$X^{(s)} = Conv1D(X^{(s-1)}) \quad (4-8)$$

第二步，构建瓶颈结构。为了降低计算开销以及防止过拟合，MSCM 在卷积前降低了输入特征的维度，并在输出卷积之后，将特征通过线性变换恢复维度，如式 4-9 和式 4-10 所示。

$$X_{reduced} = XW_{in} \quad (4-9)$$

$$X_{expanded} = X_{reduced}W_{out} \quad (4-10)$$

最后，生成多尺度表示。如式 4-11 所示，所有节点按照细到粗的顺序进行拼接，形成最终的多分辨率特征表示。

$$X_f = [X^{(1)}, X^{(2)}, \dots, X^{(S)}], X_f \in \mathbb{R}^{B \times M \times (L + \frac{L}{C} + \frac{L}{C^2} + \dots + \frac{L}{C^S}) \times D_a} \quad (4-11)$$

相异片段的跨尺度序列形成一个 C 叉树，便于后续的模块网络进行信息交换。

### 4.3.5 复合式金字塔协同网络

本小节提出复合式金字塔协同网络 HPCN，它可达成多层次以及跨尺度的信息表示，给多模态数据的联合建模给予强有力的支撑。卷积神经网络<sup>[68]</sup>（Convolutional Neural Network, CNN）和循环神经网络<sup>[38]</sup>等传统方法，虽说有线性计算的复杂度，然而因为信号传递的路径过长，很难捕捉到长时间的依赖关系。Transformer<sup>[37]</sup>尽管借助多头注意力机制缩短了信号路径的长度，可是因其计算复杂度过高，故而难以扩展到长时间心理访谈记录的处理上。在本小节里，采用了基于协同金字塔注意力机制（Co-Pyramid Attention, Co-PA）的复合式金字塔协同网络 HPCN，来对不同模态的多粒度特征构建高效的信息交互，以此提升模型对于多模态心理访谈记录的特征融合能力。

HPCN 整体设计如图 4-3 所示。直观上，对于一段含有丰富心理线索的健康访谈记录，心理医生往往会寻找多种模态间共通的心理倾向，同时也会顾及不同问题回答所呈现的相异心理特性，来综合考量患者当前的心理状况。因此，本章提出一种创新融合结构，称为 HPCN。该网络的核心部分包含一个聚合金字塔注意力（Aggregating Pyramid Attention, APA）模块和一个稳定金字塔注意力（Stable Pyramid Attention, SPA）模块。

在过往的研究中，Gimeno-Gómez 等人<sup>[35]</sup>直接将各种模态的帧级特征直接输入到

Transformer 模型进行心理预测；Chen 等人<sup>[29]</sup>预先将低级特征融合成高维特征并经过多种注意力机制来评估患者的心理状态。这些方式所提取的低级帧级特征能够捕获瞬时面部表情、语气语调、心理状态等，但对长距离时序数据所表现出心理趋势线索的分析能力不足。因此，本小节提出 APA 和 SPA 模块，旨在对 4.3.4 小节中跨尺度、多模态的长短粒度数据进行联合分析，以此提高模型对于访谈记录的特征表达与预测评估能力。

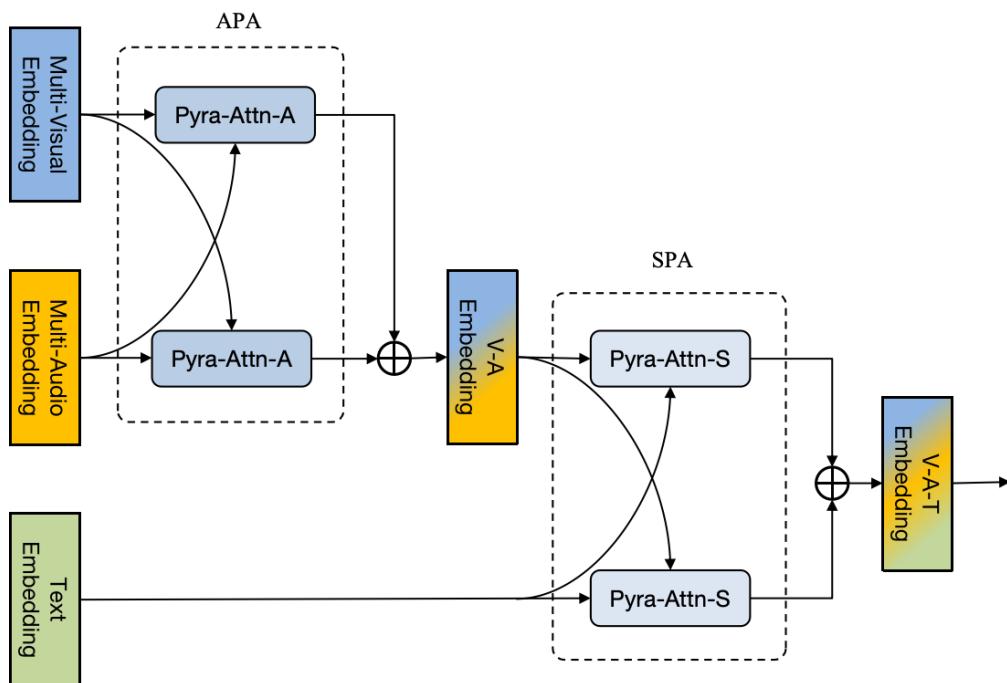


图 4-3 HPCN 整体结构

由于 APA 和 SPA 均是基于 Co-PA 的模块，因此首先对 Co-PA 层进行介绍。Co-PA 是基于金字塔注意力机制 PyraFormer<sup>[69]</sup>中 PA 层的变体。Co-PA 层的内部由金字塔注意力算法（Pyramid Attention Method, PAM）和 FFN 组成，如图 4-4 所示。下面简要介绍一下 PAM 的算法流程：

形式上，给定序列特征  $R \in \mathbb{R}^{L_{total} \times D}$ ，其中  $D$  为每个节点的维度， $L_{total}$  是所有尺度节点的总长度，如式 4-12 所示。

$$L_{total} = L + \frac{L}{C} + \frac{L}{C^2} + \dots + \frac{L}{C^s} \quad (4-12)$$

首先构建金字塔图结构，金字塔图结构分为跨层连接（Inter-scale Connections）和层内连接（Intra-scale Connections）两类连接。对于跨层连接，PAM 构建一个 C 叉树，父节点聚合子节点的信息，如式 4-13 所示，其中表示第  $s$  层中的第  $l$  个节点的父节点。

$$P_l^{(s)} = \{n_j^{(s+1)} : j = [l/C]\} \quad (4-13)$$

对于层内连接, PAM 连接每个节点及其在同一层的邻居节点, 如式 4-14 所示, 其中表示邻近  $A$  个节点的集合。

$$A_l^{(s)} = \{n_j^{(s)} : |j - l| \leq \frac{A - 1}{2}\} \quad (4-14)$$

跨层和层内连接的邻居集合为如式 4-15 所示, 其中为第  $s$  层中的第  $l$  个节点的子节点集合。

$$N_l^{(s)} = A_l^{(s)} \cup P_l^{(s)} \cup C_l^{(s)} \quad (4-15)$$

然后对金字塔图结构进行注意力机制计算。对于金字塔图中的每个节点  $N_l^{(s)}$ , 注意力机制仅计算其邻居节点的权重, 如式 4-16 所示。

$$y_i^{(s)} = \sum_{j \in N_l^{(s)}} \frac{\exp(q_i^{(s)} k_j^{(s)\top} / \sqrt{d_k})}{\sum_{j' \in N_l^{(s)}} \exp(q_i^{(s)} k_{j'}^{(s)\top} / \sqrt{d_k})} v_j^{(s)} \quad (4-16)$$

注意力机制中的查询、键和值分别如式 4-17 所示,  $W_Q, W_K, W_V$  分别为查询、键和值的线性变换矩阵,  $d_k$  为键的特征维度。

$$q_i^{(s)} = X_i^{(s)} W_Q, \quad k_i^{(s)} = X_i^{(s)} W_K, \quad v_i^{(s)} = X_i^{(s)} W_V \quad (4-17)$$

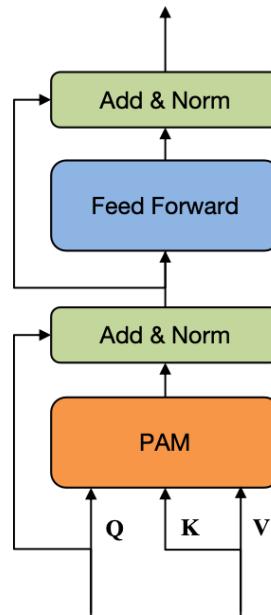


图 4-4 Co-PA 层整体结构

APA 作为基于 Co-PA 的模块, 是由两个并行连接的 Pyra-Attn-A 共同构成 APA 层的, 其中每一个 Pyra-Attn-A 内部是由多个 Co-PA 层以及其变体堆叠形成的。APA 块的功能是生成视频和音频模态相互融合的表示, 在这个过程中, 借助赋予两个 Pyra-Attn-A 不一样的输入特征, APA 层会计算每个 Pyra-Attn-A 的查询、键以及值, 这一

点和 Transformer 中的 MSA 有相似之处。随后两个 Pyra-Attn-A 的输出会被连接在一起，用作 SPA 的输入，APA 层借助交换输入模态之间的信息，对输入模态之间的密集交互展开建模。

与采用金字塔注意力机制的 PyraFormer 有所不同，APA 增设了高维聚合层，其作用在于获取全局特征信息。对于一段长达几十分钟的访谈记录数据，以往模型<sup>[35]</sup>处理帧级规模的特征信息具有高损耗、低效率的特点。因此，APA 在内部添加聚合层，将原本冗长的低级特征逐层提取成具有心理倾向的高级特征，同时保持原有金字塔图结构不变。

APA 块的设计结构如图 4-5 所示，内部由多个启动层（Starting Layer, SL）、高维聚合层（High Aggregation Layer, HAL）和增强层（Enhancement Layer, EL）按顺序堆叠组成。SL 对输入的多尺度特征数据进行初步权重分配；HAL 的设计目标是对经过 SL 后的优化特征进行层次化聚合；而 EL 将融合的高维全局信息进行平滑处理并输出。

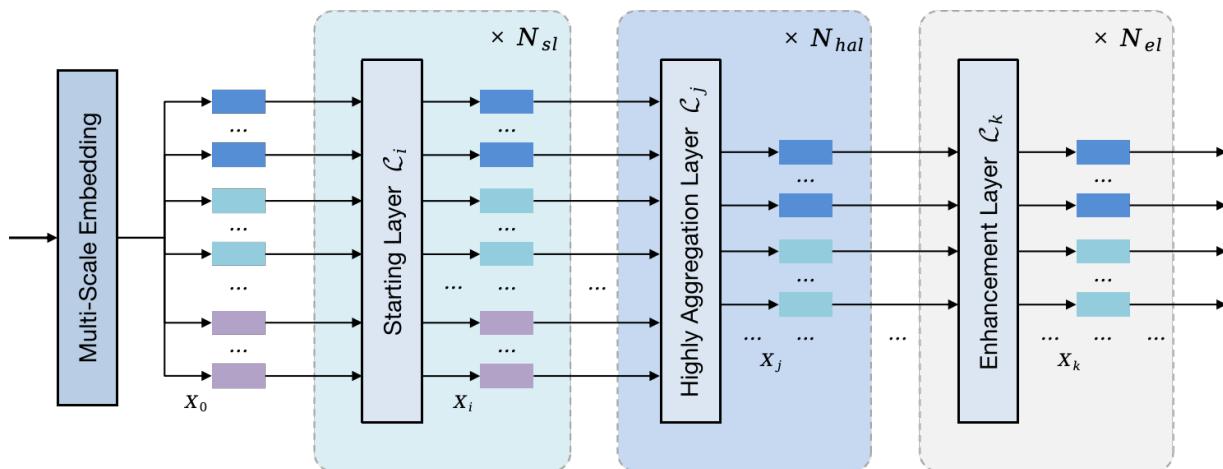


图 4-5 APA 整体结构

跨尺度特征首先输入到 SL 并生成初步融合的多模态表示传递到 HAL，HAL 将最低一层的低级特征进行均值计算，并与上一层的特征进行融合以生成基于低级特征的高维表示，最后的 EL 将得到的具有细粒度特性的全局信息进行进一步学习优化并输出。

形式上，给定跨尺度视觉特征序列  $R_v \in \mathbb{R}^{B \times M \times L_{total} \times D_v}$  和跨尺度频域特征  $R_a \in \mathbb{R}^{B \times M \times L_{total} \times D_a}$ ， $L$  为原始数据序列长度， $C$  代表卷积步长与卷积核大小， $S$  是尺度层数， $L_{total}$  是跨尺度构建后的序列总长度，如式 4-18 所示。

$$L_{total} = L + \frac{L}{C} + \frac{L}{C^2} + \cdots + \frac{L}{C^S} \quad (4-18)$$

回顾本小节对于 PAM 的介绍，序列长度具有  $L_{total}$  结构的特征可以表示为一个金字塔图。首先将  $R_v$  与  $R_a$  输入到两个并联的 APA 层的 SL 中得到频域感知的层次视觉特征  $R_{v \leftarrow a} \in \mathbb{R}^{B \times M \times L_{total} \times D_v}$  和视觉感知的层次频域特征  $R_{a \leftarrow v} \in \mathbb{R}^{B \times M \times L_{total} \times D_a}$ ，如式 4-19 和 4-20 所示。

$$R_{v \leftarrow a} = SL(R_v, R_a, R_a) \quad (4-19)$$

$$R_{a \leftarrow v} = SL(R_a, R_v, R_v) \quad (4-20)$$

随后将  $R_{v \leftarrow a}$  和  $R_{a \leftarrow v}$  馈送到 HAL 并将第  $s$  层的低级特征  $R_{v \leftarrow a}^{(s)}$  和  $R_{a \leftarrow v}^{(s)}$  按照序列长度这一维度分割成  $H$  段，每一小段进行均值计算，同时满足条件  $L = H \times C$ ，如式 4-21 所示。

$$R_{v \leftarrow a}^{(s)'} = MEAN(R_{v \leftarrow a}^{(s)}) \quad (4-21)$$

之后将得到的第  $s$  层视觉特征  $R_{v \leftarrow a}^{(s)'}$   $\in \mathbb{R}^{B \times M \times \frac{L}{C} \times D_v}$  和频域特征  $R_{a \leftarrow v}^{(s)'}$   $\in \mathbb{R}^{B \times M \times \frac{L}{C} \times D_a}$  与第  $s+1$  层进行相加，得到加权特征  $R_{v \leftarrow a}^{(s+1)'}$  和  $R_{a \leftarrow v}^{(s+1)'}$ ，如式 4-22 和 4-23 所示，其中  $W_{v \leftarrow a}$  和  $W_{a \leftarrow v}$  代表可学习权重。

$$R_{v \leftarrow a}^{(s+1)'} = W_{v \leftarrow a} R_{v \leftarrow a}^{(s)'} + (1 - W_{v \leftarrow a}) R_{v \leftarrow a}^{(s+1)} \quad (4-22)$$

$$R_{a \leftarrow v}^{(s+1)'} = W_{a \leftarrow v} R_{a \leftarrow v}^{(s)'} + (1 - W_{a \leftarrow v}) R_{a \leftarrow v}^{(s+1)} \quad (4-23)$$

最终， $R_{v \leftarrow a}$  和  $R_{a \leftarrow v}$  经过堆叠 HAL 的输出特征可以表示为  $R'_{v \leftarrow a} \in \mathbb{R}^{B \times M \times L_{hal} \times D_v}$  和  $R'_{a \leftarrow v} \in \mathbb{R}^{B \times M \times L_{hal} \times D_a}$ ，如式 4-24 和 4-25 所示。其中  $L_{hal}$  是经过 HAL 变换后的序列长度， $N_{hal}$  代表堆叠 HAL 的层数，如式 4-26 所示。

$$R'_{v \leftarrow a} = SL(R_{v \leftarrow a}, R_{a \leftarrow v}, R_{a \leftarrow v}) \quad (4-24)$$

$$R'_{a \leftarrow v} = SL(R_{a \leftarrow v}, R_{v \leftarrow a}, R_{v \leftarrow a}) \quad (4-25)$$

$$L_{hal} = \frac{L}{C^{N_{hal}}} + \frac{L}{C^{N_{hal}+1}} + \cdots + \frac{L}{C^S} \quad (4-26)$$

最后，聚合的高维特征输入 EL 并将结果进行拼接，经过一个线性变换生成与文本向量对齐的特征，如式 4-27 至 4-29 所示，其中  $W_{APA}$  表示全连接层的学习权重。

$$R''_{v \leftarrow a} = SL(R'_{v \leftarrow a}, R'_{a \leftarrow v}, R'_{a \leftarrow v}) \quad (4-27)$$

$$R''_{a \leftarrow v} = SL(R'_{a \leftarrow v}, R'_{v \leftarrow a}, R'_{v \leftarrow a}) \quad (4-28)$$

$$R_{APA} = (R''_{v \leftarrow a} \oplus R''_{a \leftarrow v}) W_{APA} \quad (4-29)$$

在 APA 块中，查询向量  $Q$  来自一种模态，而键向量  $K$  与值向量  $V$  来自另一种模态，并且  $Q$  被用作注意力层之后的剩余项。APA 块为一种模态产生以另一种模态为条件的注意力特征。如果  $Q$  来自视觉信息， $K$  和  $V$  来自音频，则使用  $Q$  和  $K$  计算的关注值可以用作视频和音频之间的相似性度量，最后对音频进行加权。对于包含复杂信息的健康

访谈记录，医生会更多地关注患者面部神情与回答自述有较高关联的部分，以便准确把握受试者的心理状态。

SPA 同样是基于 Co-PA 的模块，两个并联的 Pyra-Attn-S 来组成 SPA 层，每一个 Pyra-Attn-S 由多个 Co-PA 层组成。SPA 块的设计目的是将基于音视频模态的复合感知特征与访谈记录的文本信息进行双向交互。由于心理访谈记录采用“一问一答”的形式进行记录，文本模态数据呈现分段式特点。帧级规模的音视频信息输入到 APA 块进行高维聚合，并经过一个全连接层与文本信息对齐，之后一并馈送到 SPA 模块获得基于三种模态的心理健康全局表示。

形式上，给定来自 APA 的音视频融合特征  $R_{APA} \in \mathbb{R}^{B \times M \times D_{a-v}}$  和文本序列  $R_t \in \mathbb{R}^{B \times M \times D_t}$ ，经过 SPA 的最终输出特征  $R_{final} \in \mathbb{R}^{B \times M \times D_{a-v-t}}$ ，如式 4-30 至 4-32 所示， $W_{SPA}$  为可学习向量。

$$R_{APA \leftarrow t} = SPA(R_{APA}, R_t, R_t) \quad (4-30)$$

$$R_{t \leftarrow APA} = SPA(R_t, R_{APA}, R_{APA}) \quad (4-31)$$

$$R_{final} = (R_{APA \leftarrow t} \oplus R_{t \leftarrow APA}) W_{SPA} \quad (4-32)$$

### 4.3.6 心理健康等级评估

在多模态心理健康检测任务中，为了更精细地评估个体的心理健康状态，可以通过结合多元分类和回归任务的多任务学习框架实现双重目标的建模。具体而言，分类任务的目标是对个体心理健康状态的宏观等级进行粗粒度的分类判断；而回归任务的目标是进一步预测个体的具体量表分数，为心理健康疾病的细粒度分级提供支持。本章所提出 MPCN 模型支持的心理疾病和评估分级指标如表 4-1 所示。

表 4-1 支持的心理疾病和评估分级指标

疾病名称	评测量表	分级指标
抑郁症	PHQ-8	3
广泛性焦虑障碍	GAD-7	3

在该框架中，分类任务通过交叉熵（CE）损失函数对模型的分类能力进行优化，其损失函数定义如式 4-33 所示。

$$\mathcal{L}_c = -\frac{1}{N} \sum_{i=1}^N \left( y_c^{(i)} \log \hat{y}_c^{(i)} + (1 - y_c^{(i)}) \log (1 - \hat{y}_c^{(i)}) \right) \quad (4-33)$$

与此同时，回归任务通过均方误差（MSE）损失函数对模型的回归性能进行优化，其损失函数定义如式 4-34 所示。

$$\mathcal{L}_r = \frac{1}{N} \sum_{i=1}^N \left( y_r^{(i)} - \hat{y}_r^{(i)} \right)^2 \quad (4-34)$$

模型的整体优化目标是通过加权组合的多任务损失函数，将分类损失和回归损失统一优化，从而实现两个任务的协同学习。最终的损失函数定义如式 4-35 所示。

$$\mathcal{L} = \alpha \mathcal{L}_c + \beta \mathcal{L}_r \quad (4-35)$$

其中， $\alpha$  和  $\beta$  是可调的权重系数。

这种设计契合对于心理状态的离散等级进行宏观判断，同时还可对疾病的预测分数给予精细化量化，两类任务共享主干网络的多模态特征表示。此框架可充分运用多模态信息，提升特征学习的效率以及泛化能力，为临床干预与个性化治疗提供更具针对性的参考。

## 4.4 实验与分析

为了评估 MPCN 模型的性能，本小节在 DAIC-WOZ 和 E-DAIC 数据集上与多个基线模型展开实验。下面从实验环境、实验配置、评价指标、基线模型选择、实验结果与分析几个方面进行详细介绍。

### 4.4.1 实验环境

本章实验在 CPU 为 Intel(R) Core(TM) i9 处理器、内存为 128G 的主机上进行，服务器的操作系统为 Ubuntu 22.04.3，相关参数如表 4-2 所示。

### 4.4.2 实验配置

训练集和测试集分别采用不同的批次（batch）大小，训练集设为 2，测试集设为 4，以平衡计算效率与内存占用。数据加载线程数设置为 8，以充分利用多核 CPU 资源，提升数据加载速度。优化器选用随机梯度下降法（SGD），初始学习率（lr）设

置为 0.005，最大 epoch 配置为 10 次。学习率调度采用余弦退火策略（cosine），并在训练初始阶段设置了 1 个预热周期，预热策略为恒定学习率（constant），其值为 1e-5。数据集中的训练集、验证集和测试集按照 7:1:2 的比例进行划分。

表 4-2 实验环境配置

设备名称	参数配置
操作系统	Ubuntu 22.04.3
CPU	12th Gen Intel(R) Core(TM) i9-12900KF
主频	5.2GHz
内存	128G
显卡	NVIDIA GeForce RTX 3090
硬盘	1T
编程语言	Python 3.8
编程框架	Pytorch
编程环境	Anaconda 24.1.2

#### 4.4.3 评价指标

基于多模态访谈记录的病症等级评估是一个融合任务，它包含对于病症离散等级的分类和对疾病严重程度评分的预测。评价指标定义如下：

##### (1) 准确率 (Accuracy)

准确率用于评估模型对疾病严重程度等级分类的整体性能，定义为正确预测的样本数与总样本数的比例。设有  $N$  个样本，真实标签为  $y_i = \{L_1, L_2, \dots, L_K\}$ ，其中  $K$  表示严重程度等级数。则模型预测标签的计算公式为：

$$ACC = \frac{1}{N} \sum_{i=1}^N \Pi(y_i = \hat{y}_i) \quad (4-36)$$

其中  $\Pi(\cdot)$  为指示函数，当预测与真实标签一致时取值为 1，否则为 0。该指标适用于多分类任务（如抑郁症的轻度、中度、重度等级划分），反映模型对整体类别的判别能力。

##### (2) F1 分数 (F1-Score)

F1 分数综合了精确率 (Precision) 和召回率 (Recall)，适用于类别不平衡的心

理数据集。其计算方式为：

$$Recall = \frac{TP}{TP + FN} \quad (4-37)$$

$$Precision = \frac{TP}{TP + FP} \quad (4-38)$$

$$F1\ Score = \frac{2 \times Recall \times Precision}{Recall + Precision} \quad (4-39)$$

其中 TP、FP、FN 代表真正例、假正例和假反例。

### (3) 平均绝对误差 (MAE)

MAE 衡量模型对疾病严重程度连续评分的预测偏差，定义为预测值与真实值的绝对差异均值。MAE 定义如下：

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i| \quad (4-40)$$

MAE 对异常值具有鲁棒性，适用于评估心理健康评分的整体预测稳定性，如抑郁症严重程度的总分估计。

### (4) 均方根误差 (RMSE)

RMSE 通过平方误差放大较大预测偏差的影响，定义为：

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2} \quad (4-41)$$

该指标更强调严重预测错误的惩罚，适用于需严格控制高分误差的场景（如重度抑郁症患者的误判）。

## 4.4.4 基线模型

为验证模型有效性，本章提出的 MPCN 模型与多个基线模型开展对比实验，基于多模态访谈记录的病症等级评估属复合任务，包括对病症离散等级的评估以及对病症严重程度评分的预测，下面介绍基线模型：

DEPA<sup>[26]</sup>是一种自监督的音频嵌入方式，借助对音频片段中心频谱图预测的任务来实施预训练，以此学习高层次的序列级语音表征。此方法将领域内数据以及领域外数据相结合开展预训练，可有效地捕捉和心理问题有关的长期语音特征。

CubeMLP<sup>[27]</sup>属于一种多模态融合框架，它是基于多层感知机构建而成的。该框架借助序列、模态以及通道这三个维度的特征混合方式，把文本、音频以及视觉信息给

予整合。CubeMLP 舍弃了传统 Transformer 里的自注意力机制，运用三个独立的 MLP 单元分别针对多模态张量开展轴对齐的仿射变换，并且借助残差连接来强化特征交互。

MAN<sup>[28]</sup>这项研究设计出一种多级注意力网络，借助跨模态特征选择以及融合的方式来提高心理评估的预测性能。该网络在模态内部（如音频的 MFCC 或者 eGeMAPS、视频的 FAU 或者 BoVW 等方面）引入了注意力机制，以此动态地对关键特征进行加权，同时抑制噪声。此工作运用多级注意力机制达成了可解释的特征融合，为临床访谈里的多模态心理评估给予了高效的解决办法。

HiQuE<sup>[29]</sup>，即层次化问题嵌入网络，它与以往研究有所不同，以往研究把访谈数据当作单一序列看待，而 HiQuE 则将访谈中的主问题与跟进问题的层次关系进行了结合，还借助跨模态注意力机制来提取关键信息。

CombAtt<sup>[31]</sup>，它属于基于注意力机制的深度神经网络这一类型，其运行方式是将音频、文本以及视觉信息加以融合，以此来对心理疾病的严重程度进行回归分析。研究得出的结果证实了多模态融合在心理检测里的关键性，语言信息所起到的作用，同时还显示出把多种模态信号进行融合可提升对疾病严重等级预测的准确程度。

ITGF<sup>[32]</sup>作为一种多模态心理检测方法有创新性，它会把语音信号转变为图结构，同时结合基于预训练模型的文本分析，其具体是借助低级音频特征、基于图的语音信号特征提取以及文本分类来实现。与单纯基于语音或者文本的方法相比较，结合了文本与图结构的多模态方法在检测心理症状方面可更为精准。

MMDN<sup>[35]</sup>，此模型属于基于多模态 Transformer 的心理健康检测方法，它的核心贡献在于提出了一种灵活的多模态时序架构，这种架构可从视频的非语言线索里提取高层次的语义特征，还可以借助跨模态注意力机制融合多源信息。该模型借助引入音频语音嵌入、情感感知的面部嵌入、面部/身体/手部关键点以及注视与眨眼信息，全面捕捉心理运动特征。在模型设计方面，作者采用分窗口采样策略来处理不同时长和帧率的视频，并且借助位置嵌入与模态条件向量对齐多模态时序信息，运用 Transformer 编码器达成局部帧级与全局时序特征的联合建模。

#### 4.4.5 对比实验

本章提出的 MPCN 在 DAIC-WOZ 和 E-DAIC 数据集上进行实验。相比本章 4.4.4 小节介绍的基线模型，MPCN 在多个性能指标上表现出更好的性能。实验根据任务类别（分类和回归）和数据集（DAIC-WOZ 和 E-DAIC）的不同共分为四组。

第一组针对抑郁症以及广泛性焦虑症离散等级检测任务，运用 MPCN 和几种主流基线模型在 DAIC-WOZ 数据集上开展了系统性对比实验。如表 4-3 所示，MPCN 模型在这两类任务上的综合性能比其他对比模型都要好。针对抑郁症分类任务，MPCN 的准确率、精确率、召回率以及 F1 分数都是第一名，其 ACC 和 F1 分别比次优模型 HiQuE 提高了 2.2% 和 2.3%，并且精确率与召回率的差值只有 2.6%，模型在降低误诊和漏检之间达到了良好的平衡。相比之下，DEPA 模型因为召回率偏低存在漏检风险，而 ITGF 模型因为特征冗余使得精确率最低，突出了分类边界模糊的问题。在 GAD 任务中，虽然分类难度更高，MPCN 还是以 ACC = 0.820 和 F1 = 0.803 保持领先，其精确率与召回率只相差 2.5%，明显优于 CubeMLP 和 MMDN 等模型，验证了它对症状重叠场景有很强的适应性。分析发现，MPCN 的优异性能得益于多模态动态交互机制对语音、文本和视频的自适应融合，以及层级特征对齐策略对模态冲突的缓解，同时结合正则化技术有效抑制了小规模数据集上的过拟合现象。MPCN 凭借创新的多模态融合与优化策略，在复杂心理障碍分类任务中实现了鲁棒性、判别性与泛化性的协同提升。

表 4-3 DAIC-WOZ 数据集对比实验结果（离散等级）

模型	Depression				GAD			
	ACC	P	R	F1	ACC	P	R	F1
DEPA	0.792	0.776	0.691	0.731	0.739	0.721	0.724	0.723
CubeMLP	0.784	0.755	0.798	0.776	0.748	0.692	0.704	0.698
HiQuE	0.810	0.801	0.786	0.793	0.798	0.786	0.801	0.793
ITGF	0.767	0.688	0.789	0.735	0.694	0.663	0.702	0.681
MMDN	0.719	0.684	0.666	0.675	0.701	0.679	0.676	0.677
MPCN	<b>0.832</b>	<b>0.829</b>	<b>0.803</b>	<b>0.816</b>	<b>0.820</b>	<b>0.816</b>	0.791	<b>0.803</b>

第二组针对抑郁症和广泛性焦虑症离散等级检测任务，于 E-DAIC 数据集上开展了系统性对比实验。由表 4-4 可见，本章所提 MPCN 模型在这两类任务里都保持着最优表现。对于抑郁症分类任务，MPCN 凭借 ACC = 0.797、P = 0.771、R = 0.762 以及 F1 = 0.766 的综合性能领先所有对比模型，其 ACC 相较于次优模型 HiQuE 提升了 1.5%，同时精确率与召回率的差值缩小到 0.9%，明显优于 CubeMLP 的指标失衡状况。ITGF

与 MMDN 由于特征冗余与模态冲突表现不佳处于垫底位置，证实了多模态动态交互机制的必要性。在 GAD 任务中，MPCN 以  $ACC = 0.784$  和  $F1 = 0.755$  位居首位，相比 HiQuE 提升了 1.3% 和 0.7%，并且精确率与召回率较为均衡，而 MMDN 因模态对齐失效致使召回率大幅下降，暴露出其对症状异质性数据的适应性欠佳。MPCN 在两类任务中  $ACC$  与  $F1$  的差距均低于 1.5%，其对不同数据分布有较强的鲁棒性，而 CubeMLP 在抑郁症任务中虽召回率较高，但因精确率偏低使得误诊风险增加。MPCN 的跨数据集优势可归结于其层级特征对齐策略缓解了模态语义冲突，在复杂、异构的心理健康评估场景中实现了性能与泛化性的双重突破。

表 4-4 E-DAIC 数据集对比实验结果（离散等级）

模型	Depression				GAD			
	ACC	P	R	F1	ACC	P	R	F1
DEPA	0.734	0.727	0.688	0.707	0.701	0.675	0.677	0.676
CubeMLP	0.769	0.712	0.763	0.737	0.720	0.656	0.691	0.673
HiQuE	0.782	0.764	0.744	0.754	0.771	0.746	0.750	0.748
ITGF	0.678	0.634	0.641	0.637	0.655	0.612	0.638	0.624
MMDN	0.621	0.593	0.583	0.587	0.606	0.572	0.509	0.538
MPCN	<b>0.797</b>	<b>0.771</b>	<b>0.762</b>	<b>0.766</b>	<b>0.784</b>	<b>0.763</b>	0.749	<b>0.755</b>

第三组是关于抑郁症以及广泛性焦虑症的连续评分评估任务，于 DAIC-WOZ 数据集之上五种模型的性能展开了对比，具体情况如表 4-5 所示。从实验结果可看出，本文所提出的 MPCN 模型在这两项任务里都有出色表现：在抑郁症预测方面，MAE 为 2.88，RMSE 为 3.65，相较于最优基线模型 CombAtt，MAE 降低了 5.9%，RMSE 降低了 9.7%，在焦虑症预测方面，MAE 是 3.29，RMSE 是 3.98，相较于次优模型 CubeMLP，MAE 降低了 9.8%，RMSE 降低了 14.2%，这证实了该模型多视角特征融合机制有鲁棒性。

在抑郁症评估任务里，MPCN 的 MAE 为 2.88，比 CombAtt 的 3.06 以及 CubeMLP 的 3.29 都要低，并且它的 RMSE 是 3.65，相比 CombAtt 的 4.04 降低了 0.39，这显示出该模型可减少平均误差，还可以有效抑制异常大误差的出现。在焦虑症评估方面，MPCN 的 RMSE 是 3.98，相较于 CubeMLP 的 4.64 降低了 14.2%，同时 MAE 为 3.29，

比 CombAtt 的 3.69 降低了 10.8%，这证明了它有跨任务泛化能力。CombAtt 在抑郁症 MAE 上接近 MPCN，分别是 3.06 和 2.88，但它的 RMSE 为 4.04，与 MPCN 的 3.65 有较大差距，注意力机制可能过度拟合部分样本，而没有解决误差分布的长尾问题。实验结果证实，MPCN 在提升预测稳定性的同时降低了临床误判风险，为基于量表的心理健康评估提供了更可靠的自动化解决办法。

表 4-5 DAIC-WOZ 数据集对比实验结果（连续评分）

模型	Depression		GAD	
	MAE	RMSE	MAE	RMSE
DEPA	5.06	6.53	5.30	6.24
CubeMLP	3.29	3.88	3.65	4.64
MAN	4.37	4.83	5.06	6.53
CombAtt	3.06	4.04	3.69	4.79
MPCN	<b>2.88</b>	<b>3.65</b>	<b>3.29</b>	<b>3.98</b>

第四组为针对两种疾病展开的连续评分评估任务，于 E-DAIC 数据集里对五种模型的性能做了对比，具体情况如表 4-6 所示。从实验结果可看出，本文所提出的 MPCN 模型在这两项任务里都取得了最优性能：在抑郁症预测方面，其 MAE 为 3.06，RMSE 为 3.79，相较于次优模型 CombAtt，MAE 降低了 3.8%，RMSE 降低了 8.2%。在 GAD 预测方面，MAE 是 3.49，RMSE 是 4.31，相较于 CubeMLP，MAE 降低了 4.4%，RMSE 降低了 17.4%，这验证了该模型有跨数据集的泛化能力。

表 4-6 E-DAIC 数据集对比实验结果（连续评分）

模型	Depression		GAD	
	MAE	RMSE	MAE	RMSE
DEPA	5.46	6.58	5.72	6.65
CubeMLP	3.33	4.50	3.65	5.22
MAN	4.64	5.54	5.26	6.30
CombAtt	3.18	4.13	3.89	4.95
MPCN	<b>3.06</b>	<b>3.79</b>	<b>3.49</b>	<b>4.31</b>

#### 4.4.6 消融实验

为了进一步验证本文提出的 MPCN 模型中各模块的效用，下面对 MPCN 模型进行消融实验。实验共分为三组，第一组为保留全部模块的 MPCN 网络，其他为去掉部分模块的子网络。每一个消融子网络介绍如下：

- ALL：包含所有子网络的完整 MPCN 模型。
- MSCM：去掉跨尺度信息构建模块 MSCM，所有输入特征以原始方式输入，HPCN 模块仅进行常规融合。
- HPCN：去掉复合式金字塔协同网络 HPCN，多尺度信息每一层进行均值相加计算后融合输出。

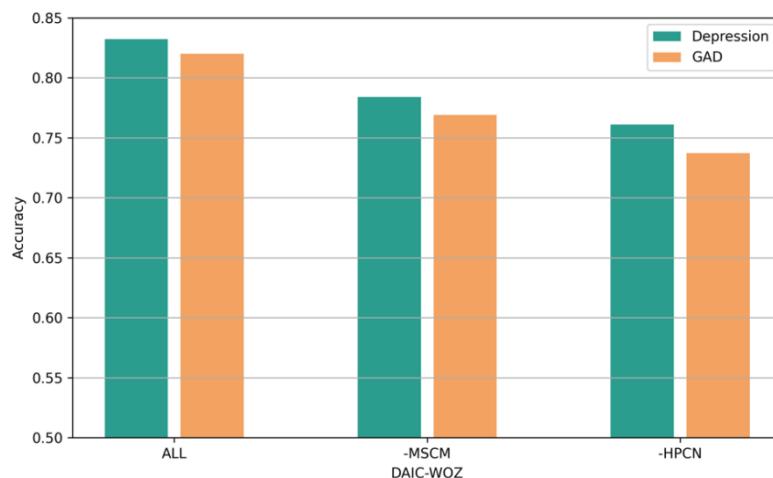


图 4-6 DAIC-WOZ 数据集上消融结果

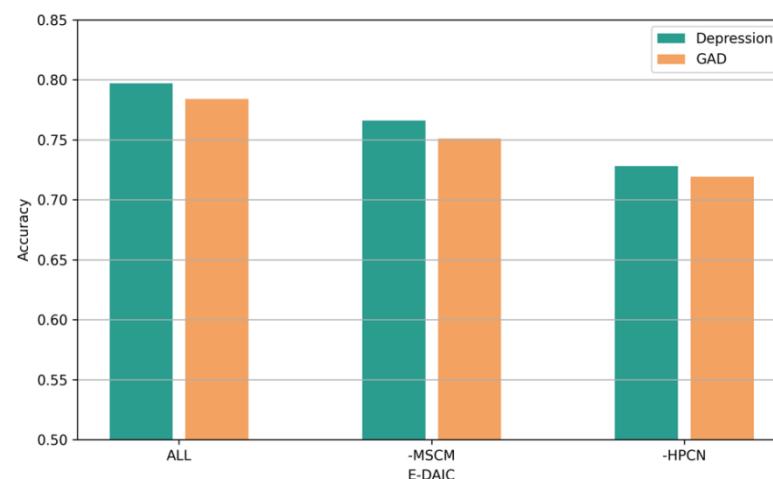


图 4-7 E-DAIC 数据集上消融结果

从图 4-6 至 4-7 中可以看出，VSCP-Net 模型加入跨尺度信息构建模块 MSCM 和复合式金字塔协同网络 HPCN 后，模型的 Accuracy 指标得到有效的性能提升。

#### 4.4.7 参数实验

在参数实验环节当中，本小节研究借助对学习率以及批次大小这两个关键超参数进行系统性对比，来剖析它们对模型性能所产生的影响，以此验证参数设置是否合理且有鲁棒性。各项参数实验所呈现的结果具体如下：

(1) 学习率 (learning rate)：学习率是梯度下降算法里的关键超参数，直接对模型参数更新步长起到控制作用。如果学习率设置得比较大，就会对模型学习稳定性造成影响，使得模型在不同结果间出现震荡情况；而要是学习率设置得比较小，那么模型的训练时间就会增加，致使收敛速度变得过慢。

(2) 批次大小 (batch size)：批次大小可决定单次参数更新时所使用的样本数量，其对梯度估计的准确性以及计算资源的利用率均会产生直接影响。如果将批次大小设置得过大，那么模型训练所需的计算成本便会明显增多；若批次大小设置得过小，模型则有可能陷入局部最优解或者无法实现收敛。

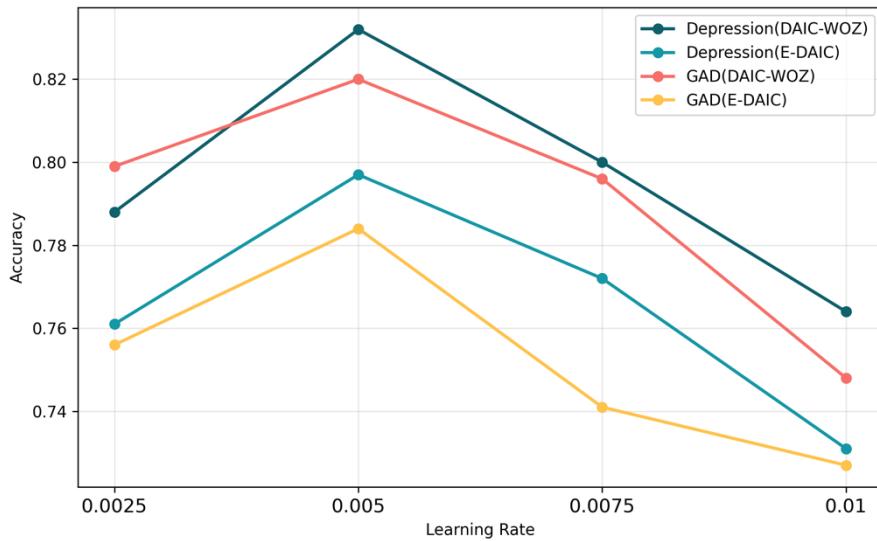


图 4-8 学习率对于模型性能的影响

图 4-8 展示了不同学习率对于 MPCN 模型性能的影响。可以看出，当学习率设置为 0.005 时，MPCN 在不同数据集和不同任务上的 Accuracy 指标为最优。

图 4-9 是批次大小对 MPCN 的影响。从图 4-9 中可以看出，当批次大小设置为 2 时，Accuracy 指标达到峰值。由于心理访谈记录是一段长达几分钟至几十分钟的数据，

因此训练集的批次大小过大会极大降低服务器的训练效率，而过小可能会存在训练数据学习不充分的问题，这些因素都会显著影响准确率等性能分数。

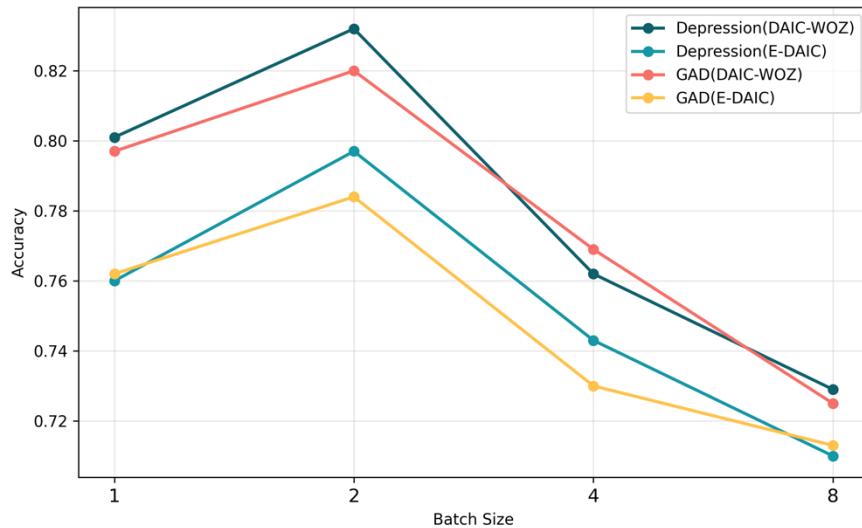


图 4-9 批次大小对于模型性能的影响

## 4.5 本章小结

本章聚焦于使用多模态访谈记录数据的病症等级评估，提出一种结合金字塔注意力机制的多尺度特征建模方法 MPCN。借助对音频、视频以及文本数据展开跨尺度建模，该模型可有效地捕捉短期动态以及长期语义依赖。复合式金字塔网络凭借其稀疏注意力机制使得计算复杂度得以降低，并且借助多层次的特征融合提高了评估精度。实验结果证实，所提方法对于心理问题中最具代表性的抑郁症（Depression）和广泛性焦虑障碍（GAD）两种疾病的评估任务，具备有效性与鲁棒性，为心理健康领域给予了更为精准且高效的技术支持。

## 第 5 章 多模态心理健康检测评估系统设计与实现

### 5.1 引言

心理健康属于个体整体健康里相当关键的一部分。传统的心理健康评估手段以及近几年提出的基于深度学习的检测方式，尽管在诊断过程里发挥了关键作用，但在评估的精度、效率、实时性等方面存在一定的限制。针对这些情况，本文第 3 章与第 4 章分别给出了基于多模态提示学习的心理健康问题分类以及基于多模态访谈记录的心理病症等级评估。在本章节之中，依据上述两种模型设计并研发了基于多模态数据的心理健康检测评估系统。该系统依照第 3 和第 4 章的方法，开发出了心理状况检测以及疾病等级评估功能，同时也拥有量表测试、知识科普、健康咨询、个人中心等功能模块。本系统可对心理问题展开早期发现并进行干预，以契合用户心理评测的需求。

### 5.2 系统开发环境与技术

为了构建一个有高效性以及鲁棒性的心理健康检测评估系统，在本章中运用了现代化的开发工具链，同时采用了先进的技术架构，以此来保障系统在复杂场景里可稳定运行，实现精准评估。

操作系统采用的是 Ubuntu 22.04 LTS 版本，它可提供稳定的服务器环境。开发工具选用 PyCharm 作为主要的开发集成环境（IDE）。编程语言挑选的是 Python 3.9，运用它来实现模型训练、数据处理以及系统集成等工作。服务器的配置是 NVIDIA A100 GPU、Intel Xeon 处理器以及 128GB 内存，以此保证系统在大规模数据处理以及实时分析过程中可以高效运行。深度学习框架是基于 PyTorch 来进行模型训练与部署的，借助其高效的 GPU 加速能力以及丰富的生态工具完成多模态数据建模与评估。数据库管理方面采用 MySQL 来存储用户数据与模型评估结果。前端开发是基于 iOS 系统来开发移动端 App。依靠上述对开发环境与技术的选择，系统在性能、易用性与可靠性之间达成了良好的平衡，为心理健康检测与评估提供了高效且智能的技术支持。

### 5.3 系统需求分析

### 5.3.1 功能性需求分析

多模态心理健康检测评估系统具有多个功能。其中包括针对潜在患者的生活照片以及内心自述开展分类工作，以此判断出其可能存在的常见心理疾病类型。另外，对于那些有可能患有心理疾病的用户的音视频访谈记录进行深入分析，评估其症状等级，并提供相应的心理干预以及疏导方案。本系统的功能模块需求具体如下：

**系统首页：** 首页呈现出本心理健康检测评估系统的常用功能，用户可点击自身感兴趣的模块去展开体验。

**心理状况检测：** 心理状况检测模块属于本系统的关键功能范畴，此模块乃是针对第三章所提方法的实际运用情形，当用户上传近期患者的健康信息之后，系统可针对用户上传的数据展开检测以及分析工作，依据预测结果给出有针对性的建议，为后续的心理干预提供相应的数据支撑。

**病症等级评估：** 病症等级评估模块属于本系统核心功能范畴，此部分将常用心理疾病评估方法以及第四章所提及的方法给予可视化呈现，借助选择算法与相关参数，用户可上传近期访谈记录用以评估病症等级，而系统会依据测试结果给出相应的疏导方案以及治疗计划。

**量表测试方面：** 量表测试功能在疏导以及治疗方案里占据着关键位置，该功能可提供经典心理量表的测试途径，其中囊括了多个经过验证的心理测量量表，当用户完成答题之后，系统可对心理状况进行诊断，给出合理化的建议。

**知识科普：** 知识科普功能属于疏导和治疗方案里的关键内容，其作用是运用简洁明了的方式来普及心理健康方面的知识，帮助用户更有效地理解心理健康的基礎概念、常见心理问题以及针对这些问题的预防和应对办法。

**视频浏览：** 视频浏览功能在疏导以及治疗方案里占据着相当关键的部分，借助经过精心挑选的心理健康教育视频，系统可帮助用户更加深入学习情绪调节的相关技巧。

**健康咨询：** 健康咨询功能属于疏导以及治疗方案里较为关键的部分，该功能可给用户给予专业的心理咨询方面的服务。

**个人中心：** 于这一模块当中，用户可对个人的基本信息加以查看，也可对个人头像进行修改，还可执行退出登录的操作。

### 5.3.2 非功能性需求分析

在需求分析环节中，非功能性需求是系统可于实际应用里实现稳定运行、保障安全以及达成高效运作的基础所在，非功能性需求主要囊括以下几个方面的内容：

性能需求方面，当用户提交数据之后，系统要可迅速做出响应并给予反馈，同时还需支持多个用户同步开展心理评估工作，以此保证在高并发的场景状况下，系统依然可以稳定地运行。

可用性方面，系统需要拥有简洁的用户界面，并且有明确的操作步骤，以此来保证用户可便捷地使用相关功能。

准确性方面，系统需要保证在进行多模态数据处理以及分析工作时，可有较高的准确程度，以此来降低出现误诊以及漏诊情况的发生频率，保证心理诊断结果拥有较高的可信度。

可靠性方面，系统需要始终维持长期运行时的稳定状态，尽可能降低出现崩溃以及失效的次数，并且系统应当有备用恢复机制，当出现如硬件故障、数据错误这类异常情况的时候，要保障用户体验。

可维护性方面，系统需要做好相关工作，也就是要记录下错误日志以及用户活动，这样做的目的在于方便后续阶段进行排查以及维护工作。

可拓展性方面，系统的整体架构需要利于功能扩展，要让新增的模块、算法以及数据输入等可较为便捷地快速集成到现有的系统当中。

## 5.4 系统设计

### 5.4.1 系统总体设计

多模态心理健康检测评估系统为心理健康领域的研究人员和潜在心理患者用户提供了一个便捷、友好的心理健康检测与评估平台，旨在帮助用户了解心理状态，识别潜在问题，并提供个性化的改善建议。系统设计分为四大模块：用户登录模块、健康检查模块、心理疏导模块和个人中心模块。系统的总体设计如图 5-1 所示，其中，健康检查模块包括心理状况检测与疾病等级评估；心理疏导模块包含量表测试、健康咨询、知识科普和视频浏览。

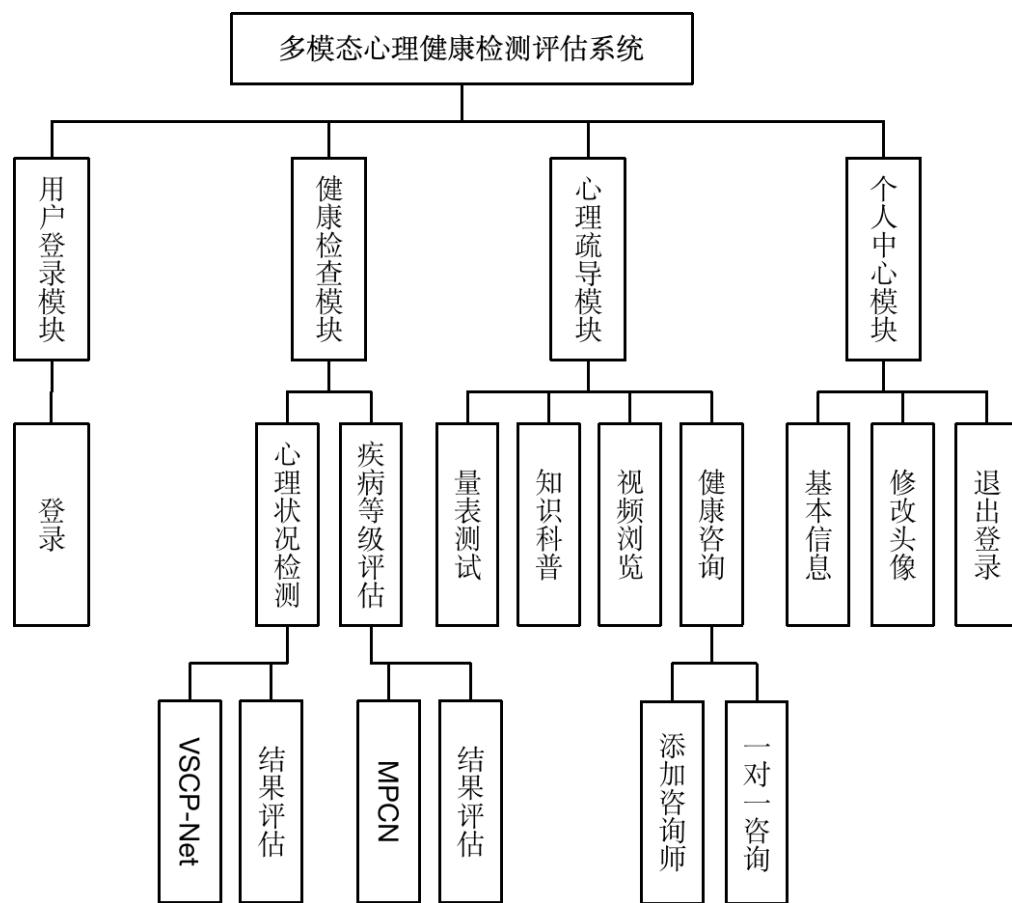


图 5-1 系统总体功能设计图

#### 5.4.2 系统功能模块设计

1. 用户登录模块:

(1) 用户登录

2. 健康检查模块:

(1) 心理状况检测

(2) 痘症等级评估

3. 心理疏导模块:

(1) 量表测试

(2) 知识科普

(3) 视频浏览

(4) 健康咨询

4. 个人中心模块:

- (1) 基本信息
- (3) 修改头像
- (4) 退出登录

## 5.5 系统功能实现

### 5.5.1 用户登录

系统的登录功能旨在为用户提供便捷、安全的访问入口。若用户处于离线状态，可以在文本框输入用户名和对应的登录密码，并点击“登录”按钮进行验证。如若信息正确，系统将自动跳转至系统首页面；当校验失败时，系统会提示错误原因并允许重新输入。登录页面如图 5-2（a）所示。



图 5-2 用户登录和系统首页功能

### 5.5.2 系统首页

系统首页是系统的门户页面，包含系统常用模块并为用户提供快捷的访问入口。用户登录成功之后会自动跳转至系统首页。系统首页提供心理状况检测、疾病等级评估、量表测试、知识科普、视频浏览、健康咨询共 6 个功能的按钮，通过点击不同的图标，用户可以访问使用不同的功能模块。系统首页如图 5-2（b）所示。

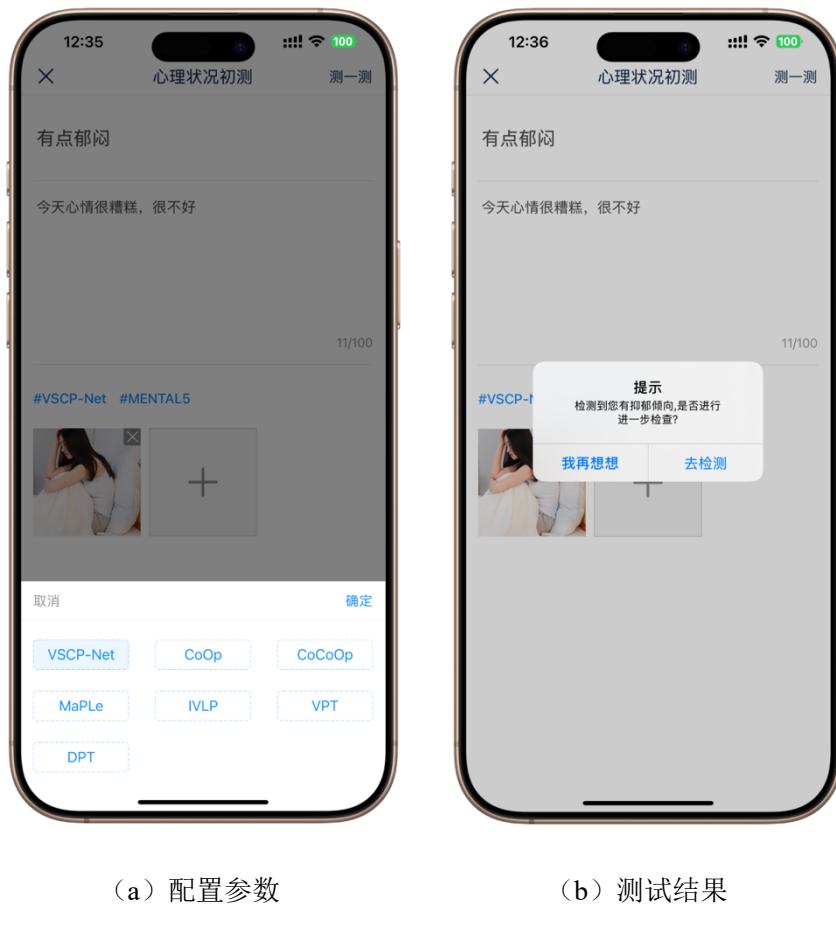


图 5-3 心理状况检测功能

### 5.5.3 心理状况检测

心理状况检测是本系统的核心模块，该功能采用经典方法和本文第三章提出 VSCP-Net 方法对用户的心理状态进行初步筛查，识别出潜在疾病问题，并给出相关下一步的治疗建议。经典方法提供 CoOp、CoCoOp、MaPLe、IVLP、VPT 和 DPT 共 6 种方法。用户在上传近期情绪不佳时的生活照片以及内心自述后，点击“测一测”按

钮进行诊断，如果存在潜在的心理问题，系统会推荐用户进行深入的心理评估。心理状况检测功能如图 5-3 所示。

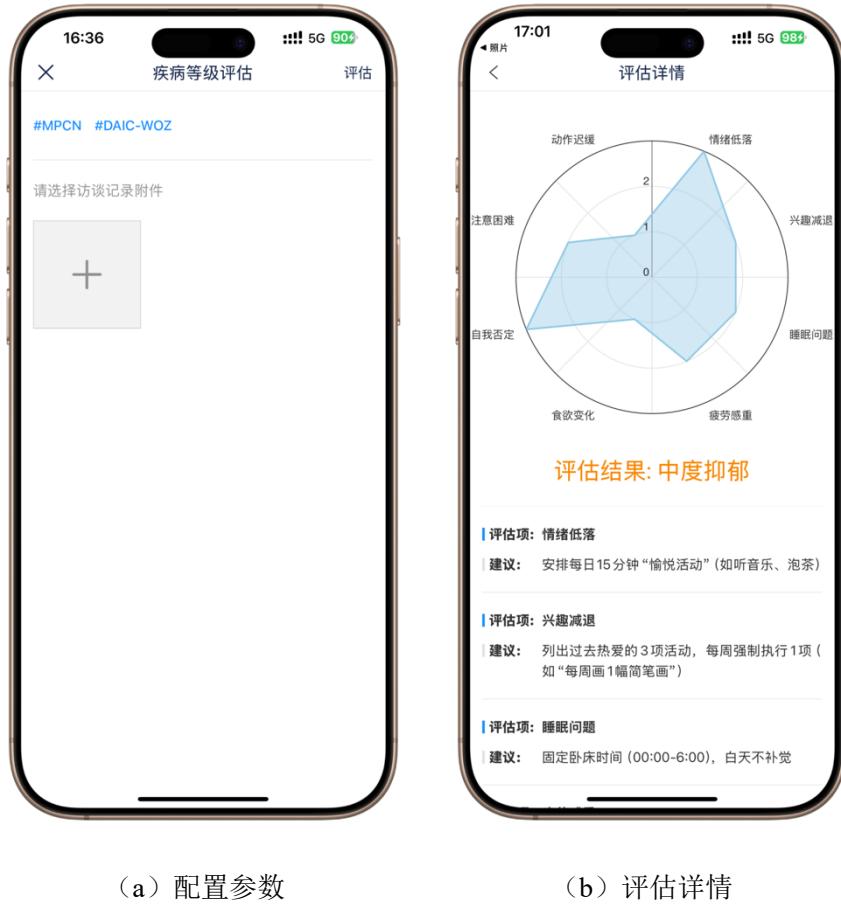


图 5-4 疾病等级评估功能

#### 5.5.4 痘症等级评估

病症等级评估是心理状况检测的进阶模块，该部分属于本系统的核心功能。用户可以根据心理状况检测模块的测评结果或者直接进行常见病症等级评估。除了本文第四章提出的 MPCN 方法，该模块还提供 DEPA、CubeMLP、HiQuE 等共 6 种方法以及 DAIZ-WOZ 和 E-DAIC 共 2 个心理健康公共数据集。用户首先上传近期的心理访谈记录并选择不同的方法和数据集，系统会根据用户的输入和配置参数进行评测，给出评估详情以及个性化心理疏导方案。评估详情界面如图 5-4 (b) 所示，它包含整体评估结果、不同评估指标的多维极地图和根据不同评估的针对性治疗建议。该模块整体结构如图 5-4 所示。

### 5.5.5 量表测试

量表测试功能为用户提供科学、便捷的心理评估工具，该功能是心理疏导的重要步骤。用户通过专业的心理量表完成自评问卷，系统会根据答题结果生成分析报告，帮助用户了解心理状况，识别潜在问题，并提出合理化建议。量表测试整体测试流程如图 5-5 所示。具体来说，用户首先通过系统首页点击想要测试的心理量表进入测试，如图 5-5（a）所示。然后量表测评介绍会在第一页给出，方便用户对量表的具有初步的了解，如图 5-5（b）所示。接着，点击“测一测”按钮进行量表测评，测评页面上方显示的是已完成题目数量和量表测试进度（百分制）；页面中间部分是测试题目和可选答案；界面最下方用户可以现在进行回答下一题或者修改之前的题目。最后，系统会根据用户的回答情况结合量表的评分标准进行测评，测评界面的整体样式与图 5-4（b）一致。



图 5-5 量表测试功能

### 5.5.6 知识科普

知识科普功能为用户提供了易于理解、内容丰富的心理健康知识，该功能是心理疏导的重要组成部分。在系统首页找到“知识”模块，用户可以点击进入该功能列表界面，之后选择任一感兴趣的知识进行详细了解，知识科普功能如图 5-6 所示。知识详情页由包含图片文字的心理知识组成，同时用户还可以查看该篇知识的已浏览人数，如图 5-6（b）所示。



图 5-6 知识科普功能

### 5.5.7 视频浏览

视频浏览功能是心理疏导的重要部分。通过该功能，用户可以观看经过专业心理学家和健康教育专家筛选的视频内容。视频浏览功能如图 5-7 所示，用户首先在系统首页下拉找到“推荐视频”模块，点击任意感兴趣的视频进行详细浏览，如图 5-7

(a)。在视频详情页，用户可以点击屏幕进行暂停、快进、快退、全屏播放等多种操作，如图 5-7 (b) 所示。



(a) 视频浏览入口

(b) 视频详情

图 5-7 视频浏览功能

### 5.5.8 健康咨询

健康咨询功能于心理疏导而言有着相当关键的地位，此功能借助一对一聊天的形式，给用户打造出一个有安全、私密以及便捷特性的心理支持平台。健康咨询功能如图 5-9 所示，用户可在系统首页寻找到“问答”模块进入健康咨询功能。健康咨询首页呈现的是用户跟所有心理咨询师聊天的咨询列表，列表中的每一项囊括了咨询师姓名、咨询师头像、最后咨询时间以及聊天记录，用户可点击列表项跟咨询师展开交流，如图 5-9 (a) 所示。用户可点击咨询首页右上角的“添加”按钮，进入心理咨询师列表去挑选其他的咨询师，咨询师列表里的每一项包含咨询师姓名、咨询师头像以及咨询师个人简介，用户可点击“咨询”按钮跟感兴趣的咨询师进行沟通，获取个性化指导与帮助，如图 5-9 (b) 所示。



图 5-8 健康咨询功能

### 5.5.9 个人中心

个人中心功能可给予用户基础信息呈现以及常用系统功能，在“基本信息”这个栏目当中，用户可查看与之相关的个人基础信息内容。

该系统为用户供给了修改头像的功能，当用户点击头像按钮之后，用户可借助拍摄或者上传自身喜爱的图片来当作自己的展示头像，并且系统也拥有退出登录功能，此功能可让用户安全地终结当前会话，以此保障个人信息以及隐私的安全。

## 5.6 系统测试

系统测试的目的是要保证整个系统在功能、性能、安全性以及可用性等方面可符合设计要求，并且可在实际环境里稳定地运行，它是心理健康系统开发过程中的一个

关键部分。测试涉及了功能测试、界面测试以及性能测试等方面。本小节借助模拟用户实际使用时的场景，对系统的核心功能给予验证，以此保证系统可精确且可靠地处理用户输入，提供科学的心理健康建议，同时保护用户的隐私以及数据安全。

### 5.6.1 功能测试

功能测试在系统测试里占据着核心位置，主要是用来查验系统的各项功能是否依照设计要求准确达成。其测试内容涉及多模态心理健康检测评估系统的关键功能模块，如心理状况检测模块以及病症等级评估模块。测试用例的设计情况在表 5-1 和表 5-2 当中有所呈现。

表 5-1 使用本文算法进行心理状况检测测试

用例编号	001	功能名称	本文算法心理状况检测
功能描述	使用本文提出的算法进行心理状况检测		
测试步骤	输入	输入模型和数据集，输入模型参数，上传近期生活照片，上传心理自述	
	输出	心理状况检测结果（抑郁症倾向、焦虑症倾向等）	
测试结果	通过	发现问题	无

表 5-2 使用本文算法进行病症等级评估测试

用例编号	002	功能名称	本文算法病症等级评估
功能描述	使用本文提出的算法进行病症等级评估		
测试步骤	输入	输入模型和数据集，输入模型参数，上传与心理咨询师的健康访谈记录	
	输出	病症等级评估结果（轻度抑郁症、重度焦虑症等）	
测试结果	通过	发现问题	无

在本小节之中，借助设计各种各样不同的测试用例，以此来模拟用户实际操作的具体情况。经过这样的操作之后，系统在各种不同的使用场景之下均可实现稳定运行，并且输出符合预期的结果，完全契合功能测试的相关要求。

### 5.6.2 界面测试

界面测试的主要目的在于验证系统的用户界面设计是否契合需求以及设计规范，以此来保证系统界面有美观性与易用性。其测试内容包括用户体验、视觉与功能、兼容性和本地化四个方面。借助模拟用户的实际操作路径，对界面的布局、交互、功能、适配等进行测试，保证其契合设计文档的要求，同时维持导航的流畅性以及逻辑的一致性，符合界面设计规范。界面测试的用例和结果如表 5-3 所示。

表 5-3 界面测试用例和测试结果

测试类别	测试指标	描述	测试结果
用户体验	布局合理性	界面元素（按钮、文本、图表）位置是否符合用户直觉	通过（关键按钮位置符合 90% 用户预期）
	交互响应一致性	点击、滑动等操作的响应是否一致且无延迟	通过（所有操作响应时间 $\leq 0.3$ 秒）
视觉与功能	视觉一致性	字体、颜色、图标在不同页面中是否统一	通过
	功能完整性	测评流程、数据提交、结果展示等核心功能是否正常	通过
兼容性	多设备适配	主流设备（手机/平板）和分辨率（1080P/2K）下的显示效果	通过
	系统版本兼容性	在 iOS 14.3-18.2 系统下的运行情况	通过（无崩溃或闪退）
本地化	多语言支持	中/英文切换后界面文本是否完整显示	通过

### 5.6.3 性能测试

性能测试属于系统测试里的关键部分，主要针对系统在不同负载状况下的运行效率以及稳定性展开。其测试涵盖响应速度、资源占用、流量消耗、流畅度和稳定性五个方面。在本小节当中，针对多模态心理健康检测评估系统开展了性能测试工作，该系统在整个测试过程里呈现出较好的反馈表现，可契合相应的性能指标要求。性能测试的用例和测试结果如表 5-4 所示。

表 5-4 性能测试用例和测试结果

测试类别	测试指标	描述	测试结果
响应速度	冷启动时间	从点击图标到主页加载完成的时间（目标≤2秒）	通过（平均 1.8 秒）
	操作响应延迟	提交测评、切换页面等操作的响应时间（目标≤1秒）	通过（平均 0.5 秒）
资源占用	内存占用率	运行期间内存峰值（目标≤200MB）	通过（峰值 185MB）
	CPU 占用率	运行期间 CPU 峰值（目标≤30%）	通过（测评结果渲染时峰值为 25%）
流量消耗	单次测评数据上传流量	完成一次测评的流量消耗（目标≤500KB）	通过（平均 480KB）
流畅度	帧率 (FPS)	页面滑动/动画的帧率（目标≥50 FPS）	通过（所有页面帧率≥50FPS）
稳定性	高负载压力测试	同时 50 用户并发操作时的崩溃率	通过（未发生崩溃或闪退）

## 5.7 本章小结

在这一章节之中详细地阐述了多模态心理健康检测评估系统的设计以及实现过程。本章一开始对开发环境以及相关技术给予了描述，紧接着依据实际应用场景展开了系统需求分析，并且对系统的总体框架进行了概括说明，随后着手开展系统主要功能的设计与开发工作，实现了登录、心理状况检测、病症等级评估、量表测试、知识科普、视频浏览、健康咨询以及个人中心等主要功能模块。此章节还开展了相关的系统测试，涉及了功能测试、界面测试以及性能测试等方面。

## 第 6 章 总结与展望

心理健康问题乃是对全球人口有着关键影响的公共健康问题。传统的心理健康评估方法存在主观性较强、对资源要求较高、难以开展长期监测等问题。近些年来，随着人工智能、深度学习等技术不断发展，目前已经研究出诸多基于深度学习的心理健康检测方法和模型，但依旧存在精度不够高、未能充分运用多模态信息等问题，鉴于现有方法存在的这些不足，本文展开了如下主要工作：

(1) 提出一种运用提示学习的多模态心理健康问题分类预测模型 VSCP-Net。该模型凭借多尺度浅层视觉提高模块以及跨模态协同提示生成器的创新设计，使得其分类性能和对心理健康问题的理解能力得到提升。在实验环节构建 MENTAL5 数据集，借助一系列对比实验与消融实验来验证模型的优越性，提高心理健康问题的预测性能。

(2) 提出了一种疾病等级评估模型 MPCN。该模型是基于多模态访谈记录构建的，它借助音视频访谈记录里的多模态特征，来构建跨尺度信息，还设计了金字塔协同模块用以进行特征深度交互。经过在 DAIC-WOZ 等心理数据集上开展实验验证，发现 MPCN 在疾病等级评估任务里的精度以及计算效率比现有方法更加优秀，能为心理健康检测领域的自动化等级评估提供新的检测方式。

(3) 设计并完成一个多模态心理健康检测评估系统。该系统运用模块化架构，结合本文第三章与第四章的研究内容，整合了心理健康问题分类预测、疾病等级评估、心理量表分析以及健康咨询等功能，同时详细阐述了系统需求分析、功能设计、技术实现以及测试结果。

在今后的工作中，我们将聚焦如下几个方向展开研究：

(1) 基于半监督学习的心理健康预测分类。本文第三章聚焦常见心理健康疾病的预测分类，但现实生活里存在部分潜在人群患有罕见疾病的情况。未来研究方向是借助半监督学习对心理健康数据集给予扩展，以此提升模型对未见病症的识别以及泛化能力。

(2) 基于提示学习的疾病等级评估。本文第四章聚焦基于音视频的疾病等级评估，借助多尺度分析以及注意力机制来剖析音视频等模态数据，但目前能够检测的疾病仅限于抑郁症和焦虑症。未来的研究方向为运用提示学习，以类似于图文数据的“提示微调”方法，把提示融入到音视频信息内，以此提升模型对于除了抑郁症、焦

虑症以外，其他多种疾病检测评估性能和精度。

## 参考文献

- [1] World Health Organization. World mental health report: Transforming mental health for all[R]. Geneva: WHO Press, 2022.
- [2] Santomauro D F, Mantilla Herrera A M, Shrestha S, et al. Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic[J]. The Lancet, 2021, 398(10312): 1700-1712.
- [3] Pfefferbaum B, North C S. Mental health and the Covid-19 pandemic[J]. New England Journal of Medicine, 2020, 383(6): 510-512.
- [4] Torales J, O' Higgins M, Castaldelli-Maia J M, et al. The outbreak of COVID-19 coronavirus and its impact on global mental health[J]. International Journal of Social Psychiatry, 2020, 66(4): 317-320.
- [5] Derogatis L , Derogatis L .SCL-90-R: Administration, scoring, and procedures manual I for the R(evised) version[J].Clinical Psychometric Research, 1977.
- [6] Kroenke K, Spitzer R L, Williams J B W. The phq - 9: Validity of a brief depression severity measure [J]. Journal of General Internal Medicine, 2001, 16(9): 606-613.
- [7] Shatte A B R, Hutchinson D M, Teague S J. Machine learning in mental health: A scoping review of methods and applications[J]. Psychological Medicine, 2019, 49(9): 1426-1448.
- [8] Soleymani M, Pantic M, Pun T. Multimodal emotion recognition in response to videos[J]. IEEE Transactions on Affective Computing, 2011, 3(2): 211-223.
- [9] Liu Y, Sourina O, Nguyen M K. Real-time EEG-based human emotion recognition and visualization [C]//2010 International Conference on Cyberworlds. IEEE, 2010: 262-269.
- [10] Torous J, Kiang M V, Lorme J, et al. New tools for new research in psychiatry: A scalable and customizable platform to empower data driven smartphone research[J]. JMIR Mental Health, 2016, 3(2): e5165.
- [11] Liu P, Yuan W, Fu J, et al. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing[J]. ACM Computing Surveys, 2023, 55(9): 1-35.
- [12] Zhou K, Yang J, Loy C C, et al. Learning to prompt for vision-language models[J]. International Journal of Computer Vision, 2022, 130(9): 2337-2348.

- [13] Zhou K, Yang J, Loy C C, et al. Conditional prompt learning for vision-language models[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 16816-16825.
- [14] Khattak M U, Rasheed H, Maaz M, et al. Maple: Multi-modal prompt learning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 19113-19122.
- [15] Guo Z, Dong B, Ji Z, et al. Texts as images in prompt tuning for multi-label image recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 2808-2817.
- [16] Jia M, Tang L, Chen B C, et al. Visual prompt tuning[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 709-727.
- [17] Xing Y, Wu Q, Cheng D, et al. Dual modality prompt tuning for vision-language pre-trained model[J]. IEEE Transactions on Multimedia, 2023.
- [18] Litjens G, Kooi T, Bejnordi B E, et al. A survey on deep learning in medical image analysis[J]. Medical Image Analysis, 2017, 42: 60-88.
- [19] Kooi T, Litjens G, Van Ginneken B, et al. Large scale deep learning for computer aided detection of mammographic lesions[J]. Medical Image Analysis, 2017, 35: 303-312.
- [20] Spampinato C, Palazzo S, Giordano D, et al. Deep learning for automated skeletal bone age assessment in X-ray images[J]. Medical Image Analysis, 2017, 36: 41-51.
- [21] Jiao Z, Gao X, Wang Y, et al. A deep feature based framework for breast masses classification[J]. Neurocomputing, 2016, 197: 221-231.
- [22] Kumar N, Verma R, Sharma S, et al. A dataset and a technique for generalized nuclear segmentation for computational pathology[J]. IEEE Transactions on Medical Imaging, 2017, 36(7): 1550-1560.
- [23] Nie D , Wang L , Gao Y ,et al. Deep convolutional neural networks for multi-modality isointense infant brain image segmentation.[J]. Proc IEEE Int Symp Biomed Imaging, 2015.DOI:10.1109/j.neuroimage.2014.12.061.
- [24] Carneiro G, Nascimento J C, Freitas A. The segmentation of the left ventricle of the heart from ultrasound data using deep learning architectures and derivative-based search methods[J]. IEEE Transactions on Image Processing, 2011, 21(3): 968-982.
- [25] Avendi M R, Kheradvar A, Jafarkhani H. A combined deep-learning and deformable-model approach to fully automatic segmentation of the left ventricle in cardiac MRI[J]. Medical Image Analysis, 2016, 30: 108-119.

- [26] Zhang P, Wu M, Dinkel H, et al. Depa: Self-supervised audio embedding for depression detection[C]//Proceedings of the 29th ACM International Conference on Multimedia. 2021: 135-143.
- [27] Sun H, Wang H, Liu J, et al. CubeMLP: An mlp-based model for multimodal sentiment analysis and depression estimation[C]//Proceedings of the 30th ACM International Conference on Multimedia. 2020: 3722-3729.
- [28] Ray A, Kumar S, Reddy R, et al. Multi-level attention network using text, audio and video for depression prediction[C]//Proceedings of the 9th International on Audio/visual Emotion Challenge and Workshop. 2019: 81-88.
- [29] Jung J, Kang C, Yoon J, et al. HiQuE: Hierarchical Question Embedding Network for Multimodal Depression Detection[C]//Proceedings of the 33rd ACM International Conference on Information and Knowledge Management. 2024: 1049-1059.
- [30] Wang J, Ravi V, Alwan A. Non-uniform speaker disentanglement for depression detection from raw speech signals[C]//Interspeech. NIH Public Access, 2023, 2023: 2343.
- [31] Qureshi S A, Hasanuzzaman M, Saha S, et al. The verbal and non-verbal signals of depression--Combining acoustics, text and visuals for estimating depression level[J]. arXiv preprint arXiv:1904.07656, 2019.
- [32] Ghadiri N, Samani R, Shahrokh F. Integration of text and graph-based features for detecting mental health disorders from voice[J]. arXiv preprint arXiv:2205.07006, 2022.
- [33] Shen Y, Yang H, Lin L. Automatic depression detection: An emotional audio-textual corpus and a gru /bilstm-based model[C]//ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022: 6247-6251.
- [34] Dinkel H, Wu M, Yu K. Text-based depression detection on sparse data[J]. arXiv preprint arXiv:1904.05154, 2019.
- [35] Gimeno-Gómez D, Bucur A M, Cosma A, et al. Reading between the frames: Multi-modal depression detection in videos from non-verbal cues[C]//European Conference on Information Retrieval. Cham: Springer Nature Switzerland, 2024: 191-209.
- [36] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]//International Conference on Machine Learning. PMLR, 2021: 8748-8763.
- [37] Vaswani A , Shazeer N , Parmar N ,et al. Attention is all you need[J].arXiv, 2017.DOI:10.48550/arXiv.1706.03762.
- [38] Jordan M I. Serial order: A parallel distributed processing approach[J].ICS-Report 8604 Institute for

- Cognitive Science University of California, 1986, 121:64.DOI:doi:<http://dx.doi.org/>.
- [39] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 770-778.
- [40] Dosovitskiy A. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929, 2020.
- [41] Baltrusaitis T, Robinson P, Morency L P. Constrained local neural fields for robust facial landmark detection in the wild [C]//Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW). IEEE, 2013: 354-361. DOI:10.1109/ICCVW.2013.54.
- [42] Centers for Disease Control and Prevention (CDC). National health and nutrition examination survey data[J]. Hyattsville, MD: US Department of Health and Human Services, Centers for Disease Control and Prevention, 2013, 2014.
- [43] Kweon S, Kim Y, Jang M, et al. Data resource profile: The Korea national health and nutrition examination survey (KNHANES)[J]. International Journal of Epidemiology, 2014, 43(1): 69-77.
- [44] Menon S, Vondrick C. Visual classification via description from large language models[J]. arxiv preprint arxiv:2210.07183, 2022.
- [45] Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training[J]. OpenAI Technical Report, 2018.
- [46] Brown T, Mann B, Ryder N, et al. Language models are few-shot learners[J]. Advances in Neural Information Processing Systems, 2020, 33: 1877-1901.
- [47] Spitzer R L, Kroenke K, Williams J B W, et al. A brief measure for assessing generalized anxiety disorder: the GAD-7.[J]. Archives of Internal Medicine, 2006, 166(10):1092-1097.DOI:10.1124/jpet.108.149989.
- [48] Charles M, Morin, Geneviève, et al. The insomnia severity index: Psychometric indicators to detect insomnia cases and evaluate treatment response.[J]. Sleep, 2011.
- [49] Young R C, Biggs J T, Ziegler V E, et al. A rating scale for mania: Reliability, validity and sensitivity.[J]. Br J Psychiatry, 1978, 133(5):429-435.DOI:10.1192/bjp.133.5.429.
- [50] Faustman W O, Overall J E. The brief psychiatric rating scale.[J]. Psychological Reports, 1962, 10(3):799-.DOI:10.2466/PR0.10.3.799-812.
- [51] Hou Q, Jiang Z, Yuan L, et al. Vision permutator: A permutable mlp-like architecture for visual recognition[J]. 2021.DOI:10.48550/arXiv.2106.12368.
- [52] Tolstikhin I O, Houlsby N, Kolesnikov A, et al. Mlp-mixer: An all-mlp architecture for vision[J].

- Advances in Neural Information Processing Systems, 2021, 34: 24261-24272.
- [53] Matthews B W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme.[J].Biochim Biophys Acta, 1975, 405(2):442-451.DOI:10.1016/0005-2795(75)90109-9.
- [54] Scherer S, Stratou G, Gratch J, et al. Investigating voice quality as a speaker-independent indicator of depression and PTSD[C]//Interspeech. 2013: 847-851.
- [55] Tsai Y H H , Bai S , Liang P P ,et al. Multimodal transformer for unaligned multimodal language sequences[J]. 2019.DOI:10.18653/v1/P19-1656.
- [56] Cummins N, Scherer S, Krajewski J, et al. A review of depression and suicide risk assessment using speech analysis[J]. Speech Communication, 2015, 71: 10-49.
- [57] Gratch J, Artstein R, Lucas G M, et al. The distress analysis interview corpus of human and computer interviews[C]//LREC. 2014: 3123-3128.
- [58] DeVault D, Artstein R, Benn G, et al. Sim sensei kiosk: A virtual human interviewer for healthcare decision support[C]//Proceedings of the 2014 International Conference on Autonomous Agents and Multi-agent Systems. 2014: 1061-1068.
- [59] Gratch J , Arstein R , Lucas G ,et al. The distress analysis interview corpus of human and computer interviews[J]. 2014.
- [60] Ringeval F, Schuller B, Valstar M, et al. Avec 2017: Real-life depression, and affect recognition workshop and challenge[C]//Proceedings of the 7th Annual Workshop on Audio/visual Emotion Challenge. 2017: 3-9.
- [61] Kroenke K, Strine T W, Spitzer R L, et al. The phq-8 as a measure of current depression in the general population[J]. Journal of Affective Disorders, 2009, 114(1-3): 163-173.
- [62] Ringeval F, Schuller B, Valstar M, et al. AVEC 2019 workshop and challenge: state-of-mind, detecting depression with AI, and cross-cultural affect recognition[C]//Proceedings of the 9th International on Audio/visual Emotion Challenge and Workshop. 2019: 3-12.
- [63] Malgaroli M, Hull T D, Calderon A, et al. Linguistic markers of anxiety and depression in Somatic Symptom and Related Disorders: Observational study of a digital intervention[J]. Journal of Affective Disorders, 2024, 352: 133-137.
- [64] Eyben F, Wöllmer M, Schuller B. Opensmile: The munich versatile and fast open-source audio feature extractor[C]//Proceedings of the 18th ACM International Conference on Multimedia. 2010: 1459-1462.
- [65] Liu Y ,Ott M ,Goyal N , et al. Roberta: A robustly optimized bert pretraining Approach.[J].CoRR,2019,abs/1907.11692

- [66] Eyben F, Scherer K R, Schuller B W, et al. The geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing[J]. IEEE Transactions on Affective Computing, 2015, 7(2): 190-202.
- [67] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, volume 1 (long and short papers). 2019: 4171-4186.
- [68] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.
- [69] Liu S, Yu H, Liao C, et al. Pyraformer: Low-complexity pyramidal attention for long-range time series modeling and forecasting[C]//# PLACEHOLDER\_PARENT\_METADATA\_VALUE#. 2022.

## 附录

本文 4.3.2 小节采用 CLNF 算法对心理健康访谈记录的视频帧序列进行处理，所提取的动作单元（Action Units, AUs）在面部表情编码系统（FACS）中分别对应一种面部肌肉运动或表情变化，如表 A-1 所示。

表 A-1 FACS 动作单元（AU）代码、含义及描述

序号	AU 代码	含义	描述
1	AU01	内眉提升	表现为眉毛内侧向上抬起
2	AU02	外眉提升	表现为眉毛外侧向上抬起
3	AU04	眉毛下拉	表现为眉毛向下和内侧拉动，通常与愤怒或困惑相关
4	AU05	上眼睑提升	表现为上眼睑的打开，通常表示惊讶或兴奋
5	AU06	面颊提升	表现为眼角的轻微挤压，常见于微笑或眯眼
6	AU09	鼻翼上拉	表现为鼻子附近的皱褶，通常与厌恶相关
7	AU10	上唇提升	表现为上唇向上拉动
8	AU12	唇角拉伸	表现为嘴角向上拉动，常见于微笑表情
9	AU14	唇角压缩	表现为嘴角内收或压缩
10	AU15	唇角下拉	表现为嘴角向下拉动，常见于悲伤或不满
11	AU17	下唇拉伸	表现为下巴向上收紧，下唇抬起
12	AU20	嘴唇拉紧	表现为嘴唇水平拉伸，常与恐惧相关
13	AU23	嘴唇紧闭	表现为嘴唇紧闭
14	AU25	嘴唇分开	表现为嘴唇自然打开，无需特定肌肉用力
15	AU26	下颌下拉	表现为下颌放松或自然下垂
16	AU28	唇咬合	表现为嘴唇卷入嘴内，通常与紧张犹豫有关
17	AU45	眨眼	表现为快速闭合和打开眼睑