

# 融合三维卷积与时序网络的情感唇读系统

周新民<sup>1, 2)\*</sup>, 余焕杰<sup>3)</sup>, 徐云凯<sup>1)</sup>, 江华晋<sup>4)</sup>

<sup>1)</sup>(湖南工商大学人工智能与先进计算学院 长沙 410205)

<sup>2)</sup>(湘江实验室 长沙 410205)

<sup>3)</sup>(湖南工商大学计算机学院 长沙 410205)

<sup>4)</sup>(韶关学院信息工程学院 韶关 512005)

(zhouxinmin2699@163.com)

**摘要:**本文针对情感语音导致唇读性能下降的问题,提出一种融合三维卷积与时序网络的两级情感唇读系统。系统采用“先情感,后内容”的级联策略:首先通过3DCNN-TCN网络从视觉序列中识别说话者的情感状态;随后,依据判别出的情感类别,调用为该状态专门优化的3DCNN-BiLSTM模型进行唇部运动到文本的识别。本文创新性地提出了唤醒度分级、发音动作干扰模式及情感动机维度三种情感建模策略,以指导专用模型的构建。在IEMOCAP和MELD两个公开情感语料库上的实验表明,所提方法能有效建模情感引起的发音变异,显著提升了情感语音的识别准确率,全面优于使用中性或混合情感数据训练的基线模型,其中基于唤醒度分级的策略取得了最优性能。

**关键词:**情感唇读; 视觉语音识别; 三维卷积神经网络; 时序卷积网络; 双向长短时记忆网络

中图法分类号: TP391 DOI: 10.3724/SP.J.1089.2026. 论文编号

## Integrating 3DCNN and Temporal Networks for Emotion Lip-Reading

Zhou Xinmin<sup>1, 2)\*</sup>, Yu Huanjie<sup>3)\*</sup>, Xu Yunkai<sup>1)</sup>, and Jiang Huajin<sup>4)</sup>

<sup>1)</sup>(School of Artificial Intelligence and Advanced Computing, Hunan University of Technology and Business, Changsha 410205)

<sup>2)</sup>(Xiangjiang Laboratory, Changsha 410205)

<sup>3)</sup>(School of Computer Science, Hunan University of Technology and Business, Changsha 410205)

<sup>4)</sup>(School of Information Engineering, Shaoguan University, Shaoguan 512005)

**Abstract:** The paper introduces a two-stage emotional lip-reading system to address performance degradation caused by emotional speech. The system follows an "emotion-first, content-second" cascade: a 3DCNN-TCN network first identifies the speaker's emotional state from visual input, then a dedicated 3DCNN-BiLSTM model—optimized for that emotion—performs lip-reading. Three novel emotion modeling strategies are proposed: arousal grading, articulatory interference patterns, and motivational dimensions. Experiments on IEMOCAP and MELD corpora show the method effectively models emotional variations, outperforming baselines trained on neutral or mixed data, with the arousal-graded strategy performing best.

**Key words:** Emotion Lip Reading; Visual Speech Recognition; 3DCNN; BiLSTM

收稿日期: 20\*\*-\*\*-\*\*; 修回日期: 20\*\*-\*\*-\*\*. 基金项目: 国家社会科学基金资助(21BGL231); 湘江实验室重大项目(24XJ01001). 周新民(1977—), 男, 博士, 教授, 博士生导师, CCF 杰出会员, 论文通信作者, 主要研究方向为智慧交通、商务智能与大数据、AI 大模型; 余焕杰(2003—), 男, 硕士研究生, CCF 学生会员, 主要研究方向为智慧农业、图像识别、目标检测; 徐云凯(2002—), 男, 硕士研究生, CCF 学生会员, 主要研究方向为唇语识别; 江华晋(2003—), 男, 本科, 主要研究方向为智慧农业。

当人类进行言语交流时,由发音器官同步产生的听觉与视觉信号之间存在着强相关性<sup>[1]</sup>。这一现象在听觉受干扰(如环境嘈杂)时尤为明显,人们会自然地注视说话者的唇部运动以辅助理解,即所谓的“唇读”<sup>[2]</sup>。受此启发,视觉语音识别技术应运而生,其目标是从纯粹的唇部运动序列中识别出对应的语音内容。近年来,随着深度学习的突破,尤其是在有限词汇任务上,自动唇读系统的性能已超越人类<sup>[3]</sup>,并广泛应用于人机交互<sup>[4]</sup>与辅助通信等领域。

然而,当前领先的唇读方法在情感中性的语音识别任务中表现出色,却在真实场景中面临严峻挑战:当语音承载强烈情感时,其识别性能会显著下降<sup>[5]</sup>。情感会系统性改变语音的声学特性(如音色、音高)和视觉特性(如口型、发音时长与力度),这使得情感鲁棒的视觉语音识别成为一个极具挑战性的课题。尽管语音情感识别技术已取得长足进步,但如何构建能够理解情感状态下语音内容的视觉识别系统,仍缺乏有效的方法论和充足的标注数据。在视觉语音识别领域,其技术演进主要遵循从手工特征到数据驱动深度学习模型的路径。早期方法关注于从嘴唇区域提取鲁棒的特征,并探索有效的融合策略。例如,主动表观模型(AAM)<sup>[6]</sup>通过联合建模唇部形状与纹理信息,显著提升了传统方法的识别精度。随着深度学习兴起,研究范式转向端到端的特征学习与识别。Noda 等<sup>[7]</sup>率先将卷积神经网络(CNN)应用于唇读,验证了其有效性。为捕捉发音的动态时序信息,3DCNN<sup>[8]</sup>和卷积神经网络-循环神经网络(CNN-RNN)<sup>[9]</sup>混合架构成为主流,其中 CNN 负责提取空间特征,而 RNN 或其变体(如 LSTM、GRU)则用于序列建模。Assael 等<sup>[10]</sup>提出的 LipNet 框架实现了端到端的语句级别识别,是该领域的一个重要里程碑。后续研究从不同粒度进行优化:在音素层面,Helen 等<sup>[11]</sup>探索了音素-视素映射与隐马尔可夫模型(HMM)的结合;在模型结构层面,注意力机制<sup>[12-15]</sup>的引入增强了对长序列关键帧的聚焦能力,Chung 等<sup>[12, 15]</sup>提出的“Watch,

Attend and Spell”(WAS) 框架便是一个典型代表。然而,上述所有进展均在输入语音为情感中性的预设下取得。当语音携带高兴、愤怒、悲伤等情感时,其对应的唇部运动在时空模式上会产生系统性扰动(如幅度、速度、肌肉紧张度的变化),这导致了以中性数据训练的最优模型性能显著衰退。因此,如何使视觉语音识别系统具备情感鲁棒性,即如何建模并适应情感所带来的发音变异仍是一个开放性问题。现有尝试多停留于小规模实验,缺乏能够指导模型设计的有效方法论。其根本瓶颈之一在于数据:目前广泛使用的大规模唇读数据集(如 LRW、LRS、GRID)几乎不包含情感标注,而具有情感标注的多模态语料库(如 IEMOCAP、MELD)其视觉数据规模与多样性通常不足以支撑深度模型的充分训练与评估。

然而,当前领先的唇读方法在情感中性的语音识别任务中表现出色,却在真实场景中面临严峻挑战:当语音承载强烈情感时,其识别性能会显著下降。情感会系统性改变语音的声学特性(如音色、音高)和视觉特性(如口型、发音时长与力度),这使得情感鲁棒的视觉语音识别成为一个极具挑战性的课题。尽管语音情感识别技术已取得长足进步,但如何构建能够理解情感状态下语音内容的视觉识别系统,仍缺乏有效的方法论和充足的标注数据。

针对上述问题,受人类多模态感知与深度学习进展的启发,本文提出一种融合三维卷积与时序网络的情感唇读系统。其核心创新在于采用“先情感,后内容”的级联策略:系统首先从视觉信号中识别说话者的情感状态,随后根据判别出的情感类别,调用一个为该情感专门训练的唇读模型进行短语识别。这一策略本质上是为不同的情感状态建立专用的视觉发音模型,从而解决多情感场景下的唇读难题。本文的主要贡献如下:

1. 提出了一个 3DCNN-TCN 情感识别网络与 3DCNN-BiLSTM 唇读网络的两级系统框架。
2. 引入了三种基于情感内在影响机制(唤醒度、发音动作干扰、行为动机)的新型情感建模

策略, 为情感唇读提供了新的理论视角.

3. 在公开情感语料库上的实验表明, 所提方法显著提升了情感语音的唇读准确率.

4. 提供了全面的实验分析, 详细探讨了不同情感、效价及强度对唇读性能的具体影响.

## 1 相关工作

### 1.1 视觉情感识别的先进方法

过去五年, 随着大规模数据集的发布和机器学习技术的飞速发展, 视觉情感识别 (VER) 领域取得了显著进步. 情感信息可从面部表情<sup>[16]</sup>、身体姿态<sup>[17]</sup>、手势<sup>[18-20]</sup>等多种视觉线索中提取. 当前, 基于深度学习的方法已全面超越传统方法, 其中卷积神经网络与循环神经网络及其变体 (如 LSTM、GRU) 构成了主流架构的核心.

卷积神经网络 (CNN)<sup>[21]</sup>擅长从静态图像或视频帧中提取层次化空间特征. 研究者常利用在大规模数据集 (如 ImageNet) 上预训练的 2D CNN 模型 (如 VGG<sup>[22]</sup>、ResNet<sup>[23]</sup>) 作为骨干网络, 并通过微调以适应情感识别任务. 为建模情感的动态演变, RNN 及其变体被用于处理由 CNN 提取的帧级特征序列, 以捕捉时间维度上的依赖关系. 近年来, 注意力机制与 Transformer 架构被广泛引入, 以增强模型对关键表情帧或面部区域的聚焦能力. 例如, 王金伟等人<sup>[24]</sup>提出的时间分段网络结合了注意力机制来加权不同视频片段的情感贡献. 此外, 三维 CNN 被直接用于从视频片段中同步提取时空特征, 在视频情感识别中展现出优势, 如 Carreira 等人<sup>[25]</sup>提出的 Inflated 3D ConvNet (I3D) 模型. 对于多模态情感识别 (如融合音频与视觉), 研究者探索了在特征层、决策层或通过注意力机制进行信息融合的不同策略. 在特征层, 早期工作如 Noroozi 等人<sup>[26]</sup>通过拼接或基于核的方法融合音视觉特征; 在决策层, 则通常对单模态分类器的输出进行加权或投票<sup>[27]</sup>. 近期研究更多采用基于注意力机制的融合, 以动态捕捉模态间的互补信息, 例如, Tsai 等人<sup>[28]</sup>提出的跨模态

Transformer 和 Hazarika 等人<sup>[29]</sup>开发的记忆融合网络, 均能有效学习模态间的交互并提升情感识别性能.

### 1.2 唇读技术的先进方法

视觉语音识别 (VSR) 或唇读技术的发展同样深受深度学习驱动. 早期系统依赖于手工设计的视觉特征 (如唇部几何特征), 随后处理流程逐渐被端到端的深度网络所取代.

现代唇读系统通常遵循“视觉编码-时序建模-分类/解码”的范式. 视觉编码器多采用 2D 或 3D CNN, 例如使用在大型图像数据集上预训练的 ResNet 或 VGG 作为骨干网络进行微调<sup>[30]</sup>, 负责从每帧或视频片段中提取鲁棒的视觉特征. 时序建模层则普遍采用 RNN (如 LSTM、GRU) 以捕捉帧间依赖<sup>[12]</sup>或更先进的 Transformer 利用自注意力机制建模全局上下文<sup>[31]</sup>, 以捕获发音过程中的长距离协同发音效应. 训练策略上, 除了标准的交叉熵损失, 知识蒸馏、自蒸馏等技术被用于提升模型性能. 此外, 大规模在野唇读数据集 (如 LRW、LRS) 的出现<sup>[32, 33]</sup>, 为训练更强大的模型提供了数据基础. 最新的研究趋势表明, 纯视觉的 Transformer 架构 (如 ViT 的变体应用于视频)<sup>[34]</sup>以及结合卷积与注意力的混合模型 (如 Conformer) 开始在唇读任务中展现卓越性能<sup>[35]</sup>. 尽管进展迅速, 但现有研究大多假设语音为中性, 专门针对情感语音的鲁棒唇读研究尚处于起步阶段, 仅有少数工作开始探索如何通过数据增强、对抗训练或多任务学习来缓解情感变异带来的影响<sup>[36]</sup>.

## 2 语料库

构建一个能够从视觉层面识别情感语音的系统, 面临的主要困难是缺少相关数据集. 目前, 公开的唇读或视听语音识别资源大多只收录中性情感语音. 这些数据难以用于探究情感如何影响发音时的视觉特征. 为了全面评估研究方法, 一个合适的语料库需要满足几项基本要求. 第一, 它

应包含多种清晰标注的情感类型。第二，它需提供高质量、正对说话人脸部的视频。第三，它必须具备足够多的说话人和语句，以支持数据驱动型深度模型的训练。为此，我们对近二十个公开的情感多模态语料库进行了逐一审查。这些语料库的核心信息汇总在表 1 中。

表 1 现有情感视听语料库概览

语料库	语言	讲话人	话语数	情绪
FABO <sup>[37]</sup>	—	23	246	9
eINTERFACE <sup>[38]</sup>	英语	42	1166	6
IEMOCAP <sup>[39]</sup>	英语	10	3060	9
SAVEE <sup>[40]</sup>	英语	4	480	7
RML	多语种	—	720	6
SEMAINE <sup>[41]</sup>	英语	150	959	27
AEFEW <sup>[42]</sup>	英语	330	1809	7
RECOLA <sup>[43]</sup>	法语	46	—	—
CREMA-D <sup>[44]</sup>	英语	91	7442	6
NNIME <sup>[45]</sup>	中文	44	—	6
RAVDESS <sup>[46]</sup>	英语	24	4904	8
RAMAS <sup>[47]</sup>	俄语	10	581	9
PolishDB <sup>[48]</sup>	波兰语	16	560	7
MELD <sup>[49]</sup>	英语	—	1433	7
CHEAVD <sup>[50]</sup>	中文	527	7030	6–8
HEU Emotion <sup>[51]</sup>	多语种	8984	16569	10
AVSP-ESD <sup>[52]</sup>	多语种	—	13285	12
CelebV-HQ <sup>[53]</sup>	英语	15653	35666	8
MimicME <sup>[54]</sup>	多语种	4700	280000	20
EmoSet <sup>[55]</sup>	多语种	—	118102	6

通过对现有资源进行系统评估和细致筛选，本研究最终选定 IEMOCAP 与 MELD 语料库作为实验数据基础。这一选择主要考虑了四个方面的因素。选择这两个语料库的目的，是为了建立一个全面的评估框架。该框架将覆盖从受控的实验室环境到自然的真实场景。

第一，考虑到数据质量与标注信息。IEMOCAP 是情感计算领域的一个标准数据集。它提供了高质量且同步的多模态数据，其中包括高清视频、音频、面部动作捕捉数据和文本转写。它还提供了精细的情感标注。这些标注既包含离散的情感类别，也包含连续的维度评分，例如效价、唤醒度和支配度。这种多层次的标注体系十分重要。它使得深入研究不同情感状态如何影响发

音的时空特征成为可能。

第二，考虑任务的相关性与挑战性。我们选择的两个语料库都以对话为单位组织，而非孤立的短语。IEMOCAP 包含即兴对话和剧本朗读。MELD 则来源于真实的多轮电视剧对白。因此，两者都产生了具有完整语言学上下文的长语句。这对唇读系统提出了更高要求。系统必须能够建模更长时段的时间依赖关系，并理解对话的语境。这增加了任务的现实意义与研究价值。

第三，两个语料库在场景上具有很强的互补性。这有助于全面检验模型的泛化能力。IEMOCAP 是在受控实验室内录制的。其数据纯净，标注可靠。这适合用于准确评估算法处理情感因素的能力。相比之下，MELD 代表了真实世界的自然场景。它包含了多变的自然光线、多样的说话人姿态、自发的情感表达以及复杂的背景。这为检验模型在实际应用中的鲁棒性提供了严格的测试环境。

最后，考虑社区的认可度与可比性。IEMOCAP 和 MELD 都是情感识别与多模态学习领域广泛使用的基准数据集。它们享有较高的学术社区认可度。使用这些数据集便于我们将研究成果与已有工作进行比较。这有助于保证我们实验结果的可靠性，以及最终结论的可比性。

## 2.1 IEMOCAP 语料库

IEMOCAP 语料库由 10 位专业演员参与构建。其中包括 5 位男性和 5 位女性。他们以两两配对的形式进行对话。对话共计 5 幕，内容包含即兴发挥或基于预设脚本的会话。整个数据库总计产生了约 12 小时的多模态数据。在每幕对话中，每位演员扮演特定的角色。这种设计旨在引发丰富且真实的情感互动。数据库对每一个话语片段都进行了精细的标注。标注内容涵盖了 9 种情感类别，例如高兴、悲伤、愤怒、中性、兴奋、沮丧、恐惧、厌恶和惊讶。同时，也标注了情感维度上的连续数值。

针对本研究的视觉唇读任务，我们主要使用该库的高清面部视频流。为了确保评估过程严谨，

我们采用了说话人无关的五折交叉验证方案。具体来说, 在每一折划分中, 我们随机选取 7 位说话人的全部对话数据用于训练。同时, 选取 1 位说话人的数据用于验证。剩下的 2 位说话人的数据则用于测试。这一方案严格遵守了一个原则: 模型在训练阶段从未接触过测试说话人的任何信息, 包括其视觉或语言特征。因此, 该方案能够客观地评估模型对于全新说话人的泛化能力。在预处理阶段, 所有视频被切割为包含完整语句的独立片段。样本的选取则严格依据其高质量的情感标注来进行。

对于情绪状态, 本文将对其进行标签化, 每种情绪对应的标签如表 2 所示。

表 2 本文情绪标签及其释义

标签	情绪	备注
HA	Happy	开心
FE	Fearful	恐惧
AN	Angry	生气
SA	Sad	难过
SU	Surprised	惊讶
NE	Neutral	中性

根据本系统的情绪标签, IEMOCAP 语料库各情绪类别的样本数量如图 1 所示。

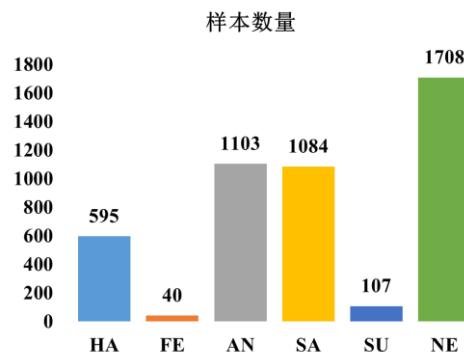


图 1 IEMOCAP 语料库各情绪类别的样本数量

## 2.2 MELD 语料库

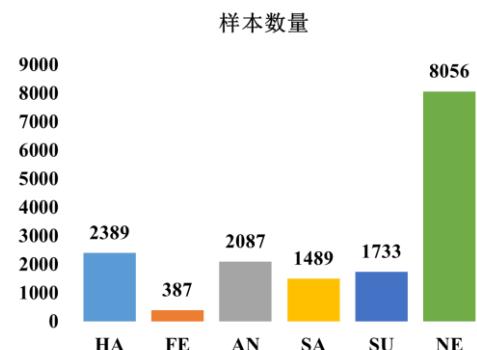
MELD 语料库的数据来源于经典电视剧《老友记》。它是一个大规模、多模态的多人对话数据集。该数据集包含了大约 13000 个独立的话语。这些话语来自于 1433 个不同的对话场景。每个话语都经过了多位众包人员的详细标注。标注内容包括 7 种离散的情感类别, 具体是高兴、悲伤、愤怒、中性、惊讶、恐惧和厌恶。同时, 也标注了 3 种情

感极性, 即积极、消极和中性。

MELD 的核心价值在于其高度的真实性与自然度。其中所有情感表达均源自演员在剧情中的自发表演。这些表达涵盖了从微妙到强烈的各种情绪状态。并且, 情感都紧密地嵌入在复杂的多轮对话上下文之中。视频画面完全保留了原始的电视拍摄效果。例如, 其中包含镜头切换、不同的人物景别以及日常化的环境背景。

在本研究中, 我们严格遵循数据集的官方划分方案进行实验。具体而言, 我们使用其训练集进行模型训练, 该集合包含 9989 个话语。同时, 使用开发集进行超参数调优, 该集合包含 1109 个话语。模型的最终性能在测试集上报告, 该集合包含 2610 个话语。这一划分方式确保了训练集、验证集和测试集来源于不同的剧集片段, 从而有效避免了内容上的重叠, 保证了评估的公正性。

根据本系统的情绪标签, MELD 语料库各情绪类别的样本数量如图 2 所示。



综合来看, IEMOCAP 与 MELD 语料库从两个关键维度为研究提供了数据支持。一个维度是高精度且受控的实验室环境, 另一个维度是复杂多变的真实世界场景。它们共同为本研究验证情感鲁棒唇读方法的有效性奠定了基础。同时, 它们也为评估该方法的泛化能力以及未来应用潜力提供了依据。整体上, 这两个语料库构建了一个全面、坚实且富有挑战性的数据基础。

### 3 情感唇读系统

本文提出了一种新型的两级情感唇读系统。该系统的设计灵感来源于人类感知机制。它采用了一种级联策略来处理情绪的影响。该策略首先识别说话者的情绪状态。然后，它依据识别出的情绪来分析言语内容。这一策略的目的是应对情绪变化对唇读模型性能产生的干扰。

如图 3 所示, 该方法首先对视觉语音数据进行预处理。随后, 处理过程分为两个核心层级。第一级负责完成情感识别任务。第二级则根据识别出的情感类别, 调用专用的视觉语音识别模型进行唇读。不同于现有研究, 本文基于情感影响语音产生的内在机制, 提出三种新的情感建模策略。这三种策略分别为唤醒度分级策略(AW)、发音动作干扰模式策略(PN)以及情感动机维度策略(AC)。图 3 展示了这三种策略对应的分类模型。

图 3 三种策略的分类模型

### 3.1 数据预处理

原始视频数据需进行规范化处理。该处理的核心目标在于使数据能够适配后续深度神经网络的输入标准。考虑到情感识别与唇读两个任务的差异，这两个任务对时空信息的需求并不相同。基于这一差异，本系统设计了并行的双分支预处理流程。

当输入一个原始视频序列时，首先要执行的步骤是人脸与面部特征点检测。该检测过程依托 MediaPipe<sup>[56]</sup>框架完成，检测对象覆盖视频中的每一帧图像。这一检测步骤会同步生成两个图像序列。这两个图像序列均经过对齐处理。第一个序列是包含完整面部区域的人脸图像序列，该序列用于情感分类任务；第二个序列是经精确裁剪得

到的嘴部区域图像序列，该序列服务于后续的唇读分析工作。

在情感识别分支中，存在双重核心需求。第一需求是保留表情动态信息，第二需求是控制模型复杂程度。为同时满足这两个需求，需对输入视频实施时间维度的降采样处理。具体的操作方式为，将视频的帧率统一调整至每秒 5 帧。以一个持续时长为 2 秒的典型语音片段作为实例，经过上述降采样操作后，可得到一个子序列。该子序列由 10 帧关键图像构成。完成时间维度的降采样处理后，需对人脸图像进行空间维度的处理。处理方式为将每张人脸图像进行空间缩放，缩放后的尺寸统一为  $224 \times 224$  像素<sup>[57]</sup>。缩放操作完成后，还需对图像的像素值执行归一化处理。

在唇读识别分支中，其核心需求与情感识别分支存在区别。该分支需要捕捉唇部与下颌的运动信息。这些运动具有细微且快速的特点，且发生于发音过程中。为充分捕捉这些运动信息，系统选择保留视频的原始时间分辨率，不进行时间维度的降采样。同样以 2 秒的语音片段作为实例，在原始时间分辨率下，该片段通常包含的图像帧数约为 60 帧。对于这 60 帧图像中的每一帧，都需要单独对其中的嘴部区域进行裁剪。裁剪得到的嘴部区域图像，需经历缩放与填充操作。在缩放过程中，需保持图像原有的比例关系。经过缩放与填充操作后，嘴部区域图像的尺寸被统一调整为  $88 \times 88$  像素<sup>[58]</sup>。尺寸调整完成后，对嘴部区域图像执行与人脸图像相同的归一化处理。上述差异化的预处理策略，可保障每个子任务均能从视觉输入数据中，精准提取与其任务目标最相关的特征信息。这一设计为后续子任务的高效执行提供了可靠的视觉数据基础。

### 3.2 基于时序卷积网络的情感识别

情感识别模块旨在从预处理后的人脸图像序列中，分类出说话者的离散情绪状态或情感效价。本文构建了一个 3DCNN-TCN 的深度网络架构，其核心是一个三维卷积神经网络与一个时序卷积网络的级联结构，如图 4 所示。

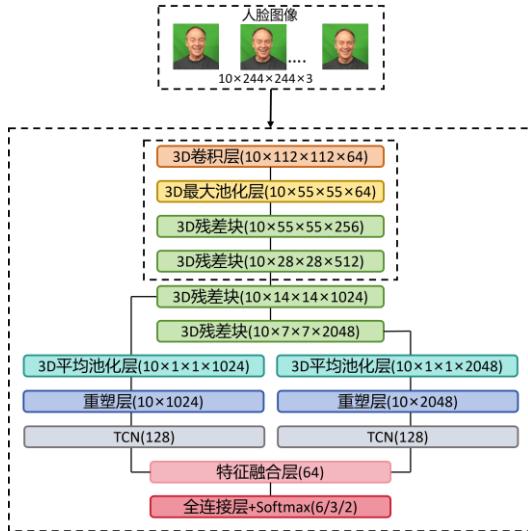


图 4 情感识别模型网络架构

该模型首先利用三维卷积神经网络(3D-CNN)提取包含空间与时间信息的视觉特征。本系统所使用的3D-CNN以ResNet-50<sup>[23]</sup>为基础结构,其权重预先在规模较大的人脸数据集上进行训练。为了全面捕捉面部表情的多层次特性,网络分别从第三和第四个残差模块中提取中等层次与高层次的语义特征。在面向特定情感数据集进行优化时,采用分层训练策略:仅对网络深层参数进行更新,同时固定浅层参数不变<sup>[59]</sup>。这一做法既能够借助预训练模型学到的通用视觉表示,又可以避免在标注数据有限的情感任务上出现模型过拟合。

提取的双流特征随后被馈入时序卷积网络进行时序建模。TCN<sup>[60]</sup>利用堆叠的空洞因果卷积层处理序列数据,其感受野随层数加深呈指数扩大,从而能够有效捕捉长距离的时序依赖。给定一个长度为L、维度为D的输入序列  $X \in \mathbb{R}^{(L \times D)}$ , TCN中单个卷积层的运算可描述为:

$$TCN(X)_{(t,c)} = \sum_{k=1}^K W_{c,k} \cdot X_{t+d \cdot (k-1)}$$

其中, K 表示卷积核尺寸, d 为空洞系数, W 为可训练权重。通过多层 TCN 的级联, 模型能够聚合整个视频片段的表情信息, 输出一个固定维度的情感表征向量。相比于循环神经网络(RNN), TCN 在训练中具有稳定的梯度流和更强的并行计算能力, 因而训练效率更高。

经由 TCN 处理后的两路特征将通过一个可学习的门控融合机制进行整合。该机制动态生成对应每路特征的权重, 并通过加权求和方式将其合并为综合的情感判别特征。该特征最终被馈送至全连接分类器, 以生成各目标情绪类别的概率分布。

### 3.3 基于双向长短时记忆网络的唇读识别

唇读识别模块依据情感识别部分输出的情绪类型, 自适应选取相应的专用模型, 将嘴部区域视频片段转化为对应的文本短语。该组件基于3DCNN与双向长短时记忆网络(BiLSTM)的编码器-解码器框架构建, 如图 5 所示, 主要包含视觉特征编码与序列解码两个部分。

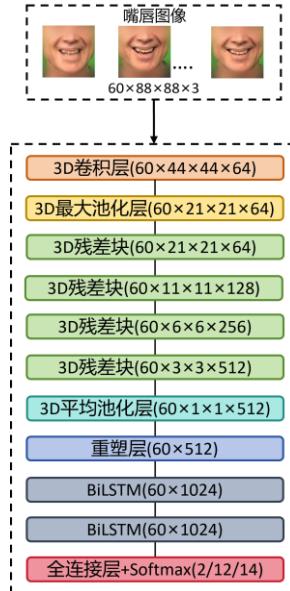


图 5 唇读识别模型网络架构

视觉编码部分采用一个轻量化的三维卷积神经网络实现, 其主干结构为基于 ResNet-18<sup>[23]</sup>设计的3D卷积版本。该编码器接收预处理后的嘴部视频片段, 该网络接收经预处理的嘴部区域视频, 通过多层三维卷积与池化操作, 逐帧提取视觉特征。对于一个包含 K 帧的输入视频, 编码器输出一个视觉特征序列  $R = \{r_1, r_2, \dots, r_K\}$ , 其中每个  $r_i \in \mathbb{R}^N$  表征了第 i 帧的视觉内容。

时序解码器由一个双向长短时记忆网络实现, 负责对视觉特征序列进行建模并输出分类结果。BiLSTM 通过同时运行前向与后向两个 LSTM 层, 使每一时刻的隐藏状态都能聚合整个序列的上下

文信息,这对于解码具有强时序依赖性的语音内容至关重要<sup>[61]</sup>. 前向 LSTM 按时间顺序处理序列,而后向 LSTM 按逆序处理序列. 在每一时间步 t, BiLSTM 的输出是前向隐藏状态与后向隐藏状态的拼接:

$$h_t = \left[ \overrightarrow{LSTM}(r_t, \overleftarrow{h}_{t-1}); \overleftarrow{LSTM}(r_t, \overleftarrow{h}_{t-1}) \right]$$

通过上述两级级联架构,情感唇读系统实现了对带有情感的语音视觉识别. 系统首先通过 3DCNN-TCN 模块辨识情绪上下文,随后利用为该情绪环境专门优化的 3DCNN-BiLSTM 模型进行精准的唇读,有效缓解了情绪因素引起的发音变异对识别性能的负面影响.

## 4 实验结果与分析

本章节对提出的情感唇读系统进行系统性评估. 实验基于 IEMOCAP 与 MELD 两个公开情感视听语料库展开,围绕视觉情感识别、不同情感建模策略下的唇读性能以及系统整体效能三个维度进行. 除使用准确率 (Acc)、未加权平均召回率 (UAR) 和 F1 分数 (F1) 等通用指标外,为全面评估唇读性能,引入了情感类别条件下的平均准确率 (mAcc),即在特定情感策略划分下,系统对短语识别的整体准确率.

### 4.1 视觉情感识别

情感识别模块的可靠性是整个两级系统的基础. 我们基于 3DCNN-TCN 模型,评估了其在三种新型情感划分策略上的识别性能. 这三种策略旨在从不同理论视角解构情感对语音产生的影响,分别为: 唤醒度分级策略(AW)、发音动作干扰模式策略(PN)与情感动机维度策略(AC).

表 3 展示了该模块在三种策略下的分类性能. 三种策略中,模型在唤醒度分级策略上取得了最优的综合性能,这表明以生理激活水平划分的情感类别在视觉上具有更高的区分度. 发音动作干扰模式策略的识别性能居中,这验证了从发音器官运动角度进行划分的有效性. 情感动机维度策略的识别挑战最大,其“趋近”与“回避”动机类

别间存在一定混淆,这或与不同动机可能对应相似面部表情有关. 所有实验均表明,采用 TCN 作为时序聚合器,模型训练过程稳定,收敛速度快,未出现明显的梯度不稳定问题.

表 3 三种策略下的分类性能

策略	IEMOCAP			MELD		
	UAR	F1	ACC	UAR	F1	ACC
AW	78.5	79.2	79.0	76.8	77.5	77.2
PN	76.2	77.0	76.5	74.5	75.3	74.8
AC	74.8	75.5	75.0	73.2	74.0	73.5

### 4.2 情感策略下的唇读识别

为探究不同情感建模策略对唇读性能的实际影响,我们针对每种策略下的各个情感类别,训练了专用的唇读模型,并在相应测试集上评估其短语识别准确率.

#### 4.2.1 唤醒度分级策略

如表 4 所示,基于唤醒度分级策略训练的专用唇读模型在 IEMOCAP 语料库上,高唤醒状态下的短语识别准确率提升最为显著,其次是中唤醒状态. 这一结果清晰地表明,情绪唤醒度是影响发音清晰度与规律性的关键因素. 高唤醒情绪引发的强烈、夸张的发音动作,虽然偏离了中性模式,但其本身具有较高的类内一致性,使得专用模型能够有效学习并实现高精度识别. 相比之下,低唤醒状态下的发音变化较为细微,模型提升幅度相对较小.

表 4 唤醒度分级策略的模型性能

模型	IEMOCAP			MELD		
	UAR	F1	ACC	UAR	F1	ACC
高唤醒	88.5	89.1	88.8	86.2	87.0	86.5
中唤醒	86.0	86.7	86.3	84.0	84.8	84.3
低唤醒	84.5	85.2	84.8	82.5	83.3	82.8

#### 4.2.2 发音动作干扰模式策略

表 5 呈现了基于发音动作干扰模式策略的实验结果. 该策略下的专用模型在不同干扰类型上均表现出性能增益. 其中,针对 B 型-唇齿主导干扰的专用模型取得了最高准确率,这表明愤怒、厌恶等情绪引起的唇部紧绷、龇牙等特征对唇读模型构成了明确且可被建模的干扰模式. A 型-嘴型主导干扰的识别率也有大幅提升,证明大尺度嘴

型变化能被有效建模。C 型-韵律节奏干扰的建模最具挑战, 其准确率提升有限, 暗示纯粹的时序节奏变化可能更需要音频信息的辅助判断。此策略的价值在于, 它为理解情绪如何从生理层面具体地改变发音提供了可计算的视角。

表 5 发音动作干扰模式策略的模型性能

模型	IEMOCAP			MELD		
	UAR	F1	ACC	UAR	F1	ACC
A	87.0	87.6	87.3	85.0	85.7	85.3
B	87.8	88.4	88.1	85.8	86.5	86.1
C	83.0	83.7	83.3	81.0	81.8	81.3
N	86.5	87.1	86.8	84.5	85.2	84.8

#### 4.2.3 情感动机维度策略

表 6 展示了基于情感动机维度策略的唇读识别结果。从数据来看, 趋近动机与回避动机类别下的专用模型性能提升明显, 而攻击动机与抑制动机的提升相对温和。一个有趣的发现是, 属于不同离散情绪但共享相同动机类别(例如, “开心”与“惊讶”同属趋近动机)的语音数据, 在训练同一个专用模型后, 均能获得较好的识别效果。这支持了我们的假设: 驱动行为的动机可能比情绪标签本身更直接地塑造了说话者的表达方式, 包括其发音特征。这为构建更精简、更泛化的情感鲁棒唇读模型提供了新思路。

表 6 情感动机维度策略的模型性能

模型	IEMOCAP			MELD		
	UAR	F1	ACC	UAR	F1	ACC
趋近	84.5	85.2	84.8	82.5	83.3	82.8
回避	83.0	83.7	83.3	81.0	81.8	81.3
攻击	82.0	82.7	82.3	80.0	80.8	80.3
抑制	80.5	81.3	80.8	78.5	79.3	78.8
基准	75.2	76.0	75.5	72.8	73.6	73.0

#### 4.3 系统整体性能

前述分析基于情感标签已知的理想条件。本小节评估完整的、端到端的两级系统性能。系统首先通过情感识别模块预测输入视频的情感策略类别, 然后自动路由至对应的专用唇读模型进行短语识别。

表 7 对比了采用三种不同情感策略的完整系统与两个基线模型的性能, 其中, Baseline(NE)是仅使用中性数据训练的唇读模型; Baseline(All)是

使用所有情感数据混合训练的统一唇读模型。

实验结果表明, 采用唤醒度分级策略的系统在 IEMOCAP 与 MELD 语料库上均取得了最优的整体性能, 显著优于两个基线模型。这证明, 以唤醒度为依据的情感划分策略, 不仅在识别阶段更可靠, 而且其划分出的类别对唇读模型的专用化训练也最为有效。发音动作干扰模式策略与情感动机维度策略的系统同样展现了稳定的性能提升, 但幅度略低于唤醒度策略。这可能是由于后两种策略的类别定义更为复杂, 对第一级情感识别模块的准确性提出了更高要求。

综合来看, 所有三种新型情感策略下的系统均一致地、显著地超越了基线模型。这强有力地证实了本文的核心思想: 通过对情感进行合理建模并据此构建专用模型, 可以系统性地提升视觉语音识别在情感语音上的鲁棒性。其中, 唤醒度分级策略凭借其概念简洁性和在两级任务上的卓越表现, 被视为实现情感鲁棒唇读的一种高效且实用的解决方案。

表 7 不同策略下的模型 mAcc 表现

模型	IEMOCAP	MELD
Baseline(NE)	82.3	80.5
Baseline(All)	81.5	79.6
AW	85.2	82.8
PN	83.5	81.0
AC	82.5	80.6

## 5 结语

本文针对情感语音对自动唇读系统造成性能干扰, 提出了一个新颖的两级情感唇读系统。该系统模拟人类“先察情, 后辨音”的感知逻辑, 首先通过一个基于 3DCNN-TCN 架构的视觉情感识别模块判定输入语音的情感状态, 随后依据该判断, 调用一个为该情感类别专门训练的、基于 3DCNN-BiLSTM 架构的唇读模型进行短语识别。与以往工作不同, 本文创新性地从唤醒度、发音动作干扰模式和情感动机三个理论维度, 提出了全新的情感建模策略, 为理解并建模情感对视觉语

音的影响提供了更本质的视角。

在 IEMOCAP 与 MELD 两个标准情感语料库上的综合实验验证了所提方法的有效性。实验结果表明，系统能够显著提升情感状态下的唇读准确率，其性能优于仅使用中性数据或混合情感数据训练的基线模型。其中，基于唤醒度分级策略的系统展现了最佳的综合性能。此外，对情感识别模块的详细评估也确认了其相对于现有方法的竞争优势。

## 参考文献(References):

- [1] 王阳, 吴岩. 听障者书面语阅读困难的根源: 字词解码缺陷还是语言知识缺陷? [J]. 中国特殊教育, 2024, (02): 30–9.
- [2] 刘丽, 隋金坪, 丁丁, et al. 深度视觉语音生成研究进展与展望 [J]. 国防科技大学学报, 2024, 46(02): 123–38.
- [3] 陈小鼎, 盛常冲, 匡纲要, et al. 唇读研究进展与展望 [J]. 自动化学报, 2020, 46(11): 2275–301.
- [4] 宁培阳, 史景伦, 张荣锋, et al. 基于深度双向模型和特征融合的视频转文字研究 [J]. 计算机应用研究, 2020, 37(01): 317–20.
- [5] 姚鸿勋, 高, 王瑞, 郎咸波. 视觉语言——唇读综述 [J]. 电子学报, 2001, (02): 239–46.
- [6] COOTES T F, EDWARDS G J, TAYLOR C J. Active appearance models [J]. IEEE Transactions on pattern analysis and machine intelligence, 2002, 23(6): 681–5.
- [7] NODA K, YAMAGUCHI Y, NAKADAI K, et al. Audio-visual speech recognition using deep learning [J]. Applied intelligence, 2015, 42(4): 722–37.
- [8] JI S, XU W, YANG M, et al. 3D convolutional neural networks for human action recognition [J]. IEEE transactions on pattern analysis and machine intelligence, 2012, 35(1): 221–31.
- [9] KIM B, LEE J. A deep-learning based model for emotional evaluation of video clips [J]. International Journal of Fuzzy Logic and Intelligent Systems, 2018, 18(4): 245–53.
- [10] ASSAEL Y M, SHILLINGFORD B, WHITESON S, et al. Lipnet: End-to-end sentence-level lipreading [J]. arXiv preprint arXiv:161101599, 2016.
- [11] BEAR H L, HARVEY R W, THEOBALD B-J, et al. Which phoneme-to-viseme maps best improve visual-only computer lip-reading?; proceedings of the International symposium on visual computing, F, 2014 [C]. Springer.
- [12] SON CHUNG J, SENIOR A, VINYALS O, et al. Lip reading sentences in the wild; proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, F, 2017 [C].
- [13] 王豪森. 基于轻量卷积和注意力机制唇读模型研究 [D]; 北京邮电大学, 2020.
- [14] 石浩泽. 基于深度学习的唇读系统设计与实现 [D]; 北京邮电大学, 2021.
- [15] CHUNG J, ZISSERMAN A. Lip reading in profile; proceedings of the British Machine Vision Conference, 2017, F, 2017 [C]. British Machine Vision Association and Society for Pattern Recognition.
- [16] TARNOWSKI P, KOŁODZIEJ M, MAJKOWSKI A, et al. Emotion recognition using facial expressions [J]. Procedia Computer Science, 2017, 108: 1175–84.
- [17] SHI H, PENG W, CHEN H, et al. Multiscale 3D-shift graph convolution network for emotion recognition from human actions [J]. IEEE Intelligent Systems, 2022, 37(4): 103–10.
- [18] 付倩, 沈俊辰, 张茜颖, et al. 面向手语自动翻译的基于 Kinect 的手势识别 [J]. 北京师范大学学报(自然科学版), 2013, 49(06): 586–92.
- [19] WU J, ZHANG Y, SUN S, et al. Generalized zero-shot emotion recognition from body gestures [J]. Applied Intelligence, 2022, 52(8): 8616–34.
- [20] RYUMIN D, IVANKO D, AXYONOV A. Cross-language transfer learning using visual information for automatic sign gesture recognition [J]. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2023, 48: 209–16.
- [21] YENIGALLA P, KUMAR A, TRIPATHI S, et al. Speech emotion recognition using spectrogram & phoneme embedding; proceedings of the Interspeech, F, 2018 [C].
- [22] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition [J]. arXiv preprint arXiv:14091556, 2014.
- [23] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition; proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, F, 2016 [C].
- [24] 王金伟, 孙华志. 基于时间分段网络并融合上下文信息的视频情感识别 [J]. 天津师范大学学报(自然科学版), 2021, 41(02): 74–80.
- [25] CARREIRA J, ZISSERMAN A. Quo vadis, action recognition? a new model and the kinetics dataset; proceedings of the proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, F, 2017 [C].
- [26] NOROOZI F, MARJANOVIC M, NJEGUS A, et al. Audio-visual emotion recognition in video clips [J]. IEEE Transactions on Affective Computing, 2017, 10(1): 60–75.
- [27] PORIA S, CAMBRIA E, BAJPAI R, et al. A review of affective computing: From unimodal analysis to multimodal fusion [J]. Information fusion, 2017, 37: 98–125.
- [28] TSAI Y-H H, BAI S, LIANG P P, et al. Multimodal transformer for unaligned multimodal language sequences; proceedings of the Proceedings of the conference Association for computational linguistics Meeting, F, 2019 [C].
- [29] HAZARIKA D, ZIMMERMANN R, PORIA S. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis; proceedings of the Proceedings of the 28th ACM international conference on multimedia, F, 2020 [C].
- [30] KAREN S. Very deep convolutional networks for large-scale image recognition [J]. arxiv preprint arxiv: 14091556, 2014.
- [31] VASWANI A, SHAZER N, PARMAR N, et al. Attention is all you need [J]. Advances in neural information processing systems, 2017, 30.
- [32] CHUNG J S, ZISSERMAN A. Out of time: automated lip sync in the wild; proceedings of the Asian conference on computer vision, F, 2016 [C]. Springer.
- [33] AFOURAS T, CHUNG J S, SENIOR A, et al. Deep audio-visual speech recognition [J]. IEEE transactions on pattern analysis and machine intelligence, 2018, 44(12): 8717–27.
- [34] ARNAB A, DEHGHANI M, HEIGOLD G, et al. Vivit: A video vision transformer; proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, F, 2021 [C].
- [35] GULATI A, QIN J, CHIU C-C, et al. Conformer: Convolution-augmented transformer for speech recognition [J]. arXiv preprint arXiv:200508100, 2020.
- [36] MA P, MARTINEZ B, PETRIDIS S, et al. Towards practical

- lipreading with distilled and efficient models; proceedings of the ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), F, 2021 [C]. IEEE.
- [37] GUNES H, PICCARDI M. A bimodal face and body gesture database for automatic analysis of human nonverbal affective behavior; proceedings of the 18th International conference on pattern recognition (ICPR'06), F, 2006 [C]. IEEE.
- [38] MARTIN O, KOTSIA I, MACQ B, et al. The eINTERFACE'05 audio-visual emotion database; proceedings of the 22nd international conference on data engineering workshops (ICDEW'06), F, 2006 [C]. IEEE.
- [39] BUSSO C, BULUT M, LEE C-C, et al. IEMOCAP: Interactive emotional dyadic motion capture database [J]. Language resources and evaluation, 2008, 42(4): 335–59.
- [40] HAQ S. Speaker-dependent audio-visual emotion recognition [J]. personal ee surrey ac uk, 2009.
- [41] MCKEOWN G, VALSTAR M, COWIE R, et al. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent [J]. IEEE transactions on affective computing, 2011, 3(1): 5–17.
- [42] DHALL A, GOECKE R, LUCEY S, et al. Collecting large, richly annotated facial-expression databases from movies [J]. IEEE multimedia, 2012, 19(3): 34–41.
- [43] RINGEVAL F, SONDEREGGER A, SAUER J, et al. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions; proceedings of the 2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG), F, 2013 [C]. IEEE.
- [44] CAO H, COOPER D G, KEUTMANN M K, et al. Crema-d: Crowd-sourced emotional multimodal actors dataset [J]. IEEE transactions on affective computing, 2014, 5(4): 377–90.
- [45] CHOU H-C, LIN W-C, CHANG L-C, et al. NNIME: The NTHU-NTUA Chinese interactive multimodal emotion corpus; proceedings of the 2017 Seventh international conference on affective computing and intelligent interaction (ACII), F, 2017 [C]. IEEE.
- [46] LIVINGSTONE S R, RUSSO F A. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English [J]. PloS one, 2018, 13(5): e0196391.
- [47] PEREPELKINA O, KAZIMIROVA E, KONSTANTINOVA M. RAMAS: Russian Multimodal Corpus of Dyadic Interaction for studying emotion recognition [R]: PeerJ Preprints, 2018.
- [48] SAPIŃSKI T, KAMIŃSKA D, PELIKANT A, et al. Multi-modal database of emotional speech, video and gestures; proceedings of the International Conference on Pattern Recognition, F, 2018 [C]. Springer.
- [49] PORIA S, HAZARIKA D, MAJUMDER N, et al. Meld: A multimodal multi-party dataset for emotion recognition in conversations; proceedings of the Proceedings of the 57th annual meeting of the association for computational linguistics, F, 2019 [C].
- [50] LI Y, TAO J, SCHULLER B, et al. Mec 2017: Multimodal emotion recognition challenge; proceedings of the 2018 first Asian conference on affective computing and intelligent interaction (ACII Asia), F, 2018 [C]. IEEE.
- [51] CHEN J, WANG C, WANG K, et al. HEU Emotion: a large-scale database for multimodal emotion recognition in the wild [J]. Neural Computing and Applications, 2021, 33(14): 8669–85.
- [52] LANDRY D, HE Q, YAN H, et al. ASVP-ESD: A dataset and its benchmark for emotion recognition using both speech and non-speech utterances [J]. Global Scientific Journals, 2020, 8: 1793–8.
- [53] ZHU H, WU W, ZHU W, et al. CelebV-HQ: A large-scale video facial attributes dataset; proceedings of the European conference on computer vision, F, 2022 [C]. Springer.
- [54] PAPAOANNOU A, GECER B, CHENG S, et al. Mimicme: A large scale diverse 4d database for facial expression analysis; proceedings of the European Conference on Computer Vision, F, 2022 [C]. Springer.
- [55] YANG J, HUANG Q, DING T, et al. Emoset: A large-scale visual emotion dataset with rich attributes; proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, F, 2023 [C].
- [56] 陈雅茜, 吴非, 赵丁皓. 基于身体姿态关键点检测及算法融合的连续手语识别 [J]. 西南民族大学学报(自然科学版), 2023, 49(02): 165–72.
- [57] 陈零壹, 胡奕然, 徐宁, et al. 融合动态分组注意力与可变形对比的文物图像修复 [J]. 重庆邮电大学学报(自然科学版): 1–13.
- [58] 马金林, 郭兆伟, 马自萍, et al. 多尺度门控时空增强的唇语识别方法 [J]. 计算机辅助设计与图形学学报, 2025, 37(07): 1228–38.
- [59] 马梓博, 米悦, 张波, et al. 面向异质性医学图像处理的深度学习算法综述 [J]. 软件学报, 2023, 34(10): 4870–915.
- [60] BAI S. An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling [J]. arXiv preprint arXiv:180301271, 2018.
- [61] ZHANG Y-M, WANG H. Multi-head attention-based probabilistic CNN-BiLSTM for day-ahead wind speed forecasting [J]. Energy, 2023, 278: 127865.