



李征, 博士, 四川大学计算机学院教授, 博士研究生导师, 计算机学院党委副书记。1997 年 6 月毕业于四川大学计算机科学系, 获得工学学士学位; 2000 年 6 月毕业于四川大学计算机学院, 获得工学硕士学位; 2009 年 6 月毕业于四川大学数学学院, 获得理学博士学位。2000 年 7 月至今, 在四川大学计算机学院从事教学科研工作, 主要研究领域包括数字图像处理、计算机视觉、计算机图形学, 擅长的子领域包括信号重采样理论、视觉特征提取与分析、图像配准、三维表面重建, 作为第一作者或通信作者发表学术论文 20 余篇。

AFAT-MER: 基于解剖学引导的自适应 Transformer 微表情识别

圣高杰¹, 李 征¹, 任印波², 刘 侠², 李 婧³,
熊 婧³, 王 黎³, 廖 艳⁴

(1. 四川大学计算机学院, 成都 610065; 2. 贵州省戒毒管理局, 贵阳 550003;
3. 贵州省女子强制隔离戒毒所, 贵阳 550003; 4. 贵州合谷信息科技有限公司, 贵阳 550081)

摘 要: 现有基于深度学习的微表情识别方法通常引入面部解剖结构先验以辅助特征建模, 但仍存在先验信息参与特征学习不充分、利用方式较为固定, 以及缺乏样本自适应调节能力等局限。针对上述问题, 本文提出一种基于面部动作编码系统(Facial Action Coding System, FACS)的自适应解剖引导 Transformer 微表情识别方法(Adaptive FACS-guided Anatomical Transformer for Micro-expression Recognition, AFAT-MER), 旨在依据表情强度将 FACS 先验自适应地融入特征学习与自注意力建模, 从而增强模型对微表情关键区域及细微变化的表征能力。首先, 利用 FACS 相关动作单元构建粗粒度面部解剖区域先验, 并将其嵌入特征学习过程; 其次, 在自注意力建模中引入软先验引导, 并结合样本表情强度自适应调节先验权重, 以提高先验利用的灵活性和样本适应性。实验结果表明, AFAT-MER 在多个主流微表情数据集上取得了较好的识别性能, 相较基线方法 HTNet 具有稳定改进, 并在与多种代表性方法的比较中表现出一定优势。

关键词: 微表情识别; FACS 面部动作编码系统; 先验引导; 深度学习

中图分类号: TP391.4 **文献标志码:** A **DOI:** 10.19907/j.0490-6756.260029

AFAT-MER: Anatomy-guided adaptive transformer for micro-expression recognition

SHENG Gaojie¹, LI Zheng¹, REN Yinbo², LIU Xia², LI Jing³,
XIONG Jing³, WANG Li³, LIAO Yan⁴

(1. College of Computer Science, Sichuan University, Chengdu 610065, China;
2. Guizhou Province Drug Rehabilitation Administration, Guiyang 550003, China;

收稿日期: 2026-02-04 修回日期: 2026-03-28 接受日期: 2026-03-31

基金项目: 国家重点研发计划项目(2020YFA0714003); 国家自然科学基金(61471250)

作者简介: 圣高杰(2003—), 男, 江苏高邮人, 博士研究生, 研究方向为数字图像处理. E-mail: shenggaojie@stu.scu.edu.cn

通信作者: 李征. E-mail: lizheng@scu.edu.cn

3. Guizhou Women's Compulsory Isolation Drug Rehabilitation Center, Guiyang 550003, China;

4. Guizhou Hegu Information Technology Co., Ltd., Guiyang 550081, China)

Abstract: Although existing deep learning-based micro-expression recognition (MER) methods often incorporate facial anatomical structure priors to facilitate feature modeling, such prior knowledge is typically underutilized, applied in a rigid manner, and lacks sample-level adaptability. To address these limitations, this paper proposes AFAT-MER (Adaptive FACS-guided Anatomical Transformer for Micro-expression Recognition), a novel framework that adaptively integrates Facial Action Coding System (FACS) priors into feature learning and self-attention mechanisms guided by expression intensity. Specifically, the proposed method first constructs coarse facial anatomical region priors inspired by FACS-related Action Units (AUs) and embeds them into the feature learning pipeline. Subsequently, a soft prior guidance mechanism is introduced into the self-attention modeling, where prior weights are dynamically adapted based on expression intensity, thereby enhancing the flexibility and sample-specific suitability of prior utilization. Extensive experiments conducted on several mainstream MER datasets demonstrate that the proposed method achieves competitive recognition performance, yields stable improvements over the HTNet baseline, and exhibits distinct advantages compared with multiple state-of-the-art representative methods.

Keywords: micro-expression recognition; FACS facial action coding system; prior guidance; deep learning

微表情是一种持续时间极短(通常为 0.04~0.2 s)^[1]的自发面部表情,被认为是人类真实情绪的重要表征^[2]。与宏观表情相比,微表情具有持续时间长、强度低、区域性强等特点,往往难以被肉眼捕捉。然而,正由于其具有较强的情绪真实性,微表情识别在情绪分析、心理健康评估、安全检测和人机交互等情感计算领域^[3]具有重要研究价值和应用前景。

微表情识别方法的发展经历了从传统手工特征到深度学习的演进过程。传统方法依赖手工设计的特征,如局部二值模式(Local Binary Pattern, LBP)及其时空变体 LBP-TOP^[4]和光流特征^[5]等,对复杂外观变化与跨个体泛化能力有限。随着深度学习发展,基于卷积神经网络(Convolutional Neural Network, CNN)的方法通过端到端学习提升了微表情特征表示能力,ResNet^[6]、DenseNet^[7]等结构被广泛应用于该任务,但卷积操作以局部感受野为主,对远距离依赖关系的建模能力仍然有限。近年来, Vision Transformer(ViT)^[8]、HTNet^[9]、Micron-BERT^[10]等方法将全局自注意力引入微表情识别,在建模长程依赖与区域交互方面展现出较大潜力。但由于微表情数据通常样本量小、类别分布不均匀且运动信号微弱,Transformer 在该场景下面临归纳偏置不足和过拟合风险。因此需引入与面部解剖一致的结构先验以约束解空间,以增强模型对关键局部区域的有效建模能力,

而非简单堆叠层数或参数量。

面部动作编码系统(Facial Action Coding System, FACS)^[10]将面部运动表述为若干面部动作单元(Action Unit, AU)及其组合,提供了可检验的解剖与表情的对应关系,是融入先验的常用理论依据。已有工作表明,利用 FACS 相关先验有助于增强模型对关键区域的关注,但现有方法在先验利用方式上仍存在若干局限。一方面,部分方法依赖额外的 AU 标注数据或借助跨模态对齐、知识蒸馏、后期特征融合等方式引入先验,使先验主要作用于中高层语义表示阶段,难以充分参与浅层纹理与细微运动特征的形成过程。另一方面,一些方法通过在注意力或特征选择过程中施加固定区域约束来突出解剖相关区域,虽然能够强化局部关注,但也可能限制跨区域的信息交互,难以兼顾局部解剖一致性与全局依赖建模。此外,不同样本在表情强度、动作幅度和显著区域分布上存在明显差异,而现有方法通常采用统一的先验作用方式,缺乏随样本特征动态调节的能力,从而限制了模型在复杂场景下的适应性。由此可见,如何将 FACS 先验以更早期、更柔性且更具样本自适应性的方式融入 Transformer 特征学习过程,仍是值得进一步研究的问题。

针对上述问题,本文提出了一种基于 FACS 理论的自适应解剖引导 Transformer 微表情识别方法(Adaptive FACS-guided Anatomical Trans-

former for Micro-expression Recognition, AFAT-MER),旨在依据表情强度将FACS先验以更早期且更具样本自适应性的方式融入特征学习过程,以克服先验参与不足、跨区域信息交互受限以及样本适应性不强等问题。本文以分层Transformer作为骨干,以光流表征为输入,并结合人脸关键点构建粗粒度解剖区域先验。首先,利用FACS相关区域协同信息构建解剖引导表示,并在特征提取早期与主干网络进行融合,以增强模型对微弱局部运动区域的初始感知能力。其次,在自注意力建模过程中引入软先验引导机制,并结合样本表情强度自适应调节先验权重,使模型在保留全局依赖建模能力的同时,更有效地突出与微表情相关的解剖区域,从而提升对细微区域协同运动的识别能力。

本文的主要贡献如下:1)提出一种面向微表情识别的FACS先验引导早期融合框架,在分层Transformer骨干中将粗粒度区域协同先验引入浅层表示学习。2)设计软先验引导的自注意力机制,在保留全局依赖建模能力的同时融入解剖约束,缓解固定区域约束对跨区域信息交互的限制。3)提出样本级自适应先验强度调节策略,根据表情强度动态控制先验引导程度,提高模型对不同微表情样本的适应能力。

1 相关工作

1.1 微表情识别

早期的微表情识别方法主要依赖手工设计的特征和传统机器学习技术。局部二值模式(Local Binary Pattern, LBP)及其变体LBP-TOP^[4]被广泛用于提取时空纹理特征。光流法用于捕获面部运动信息,TV-L1光流^[5]因其鲁棒性被广泛采用。这些方法虽然能够提取一定的特征,但在复杂场景下的识别性能有限。

随着深度学习技术的发展,基于卷积神经网络(Convolutional Neural Network, CNN)的方法通过端到端学习自动提取特征,取得了显著进展。三维卷积神经网络(3D-CNN)^[11]能够同时捕获空间和时间信息,适合处理视频序列。ResNet^[6]通过残差连接解决了深度网络的退化问题,DenseNet^[7]通过密集连接实现了特征的重复利用。这些深度网络结构被应用到微表情识别中,显著提升了识别性能。一些方法尝试结合多尺度特征或注意力

机制来增强模型的识别能力。

近年来,Vision Transformer(ViT)^[8]在图像识别任务中表现出色,也逐渐被引入微表情识别领域。Transformer的核心优势在于弱归纳偏置,能够学习任意距离的依赖关系,突破了CNN局部感受野的限制。ViT将图像分割成固定大小的patches,并将每个patch视为一个token,通过自注意力机制学习图像表示,实现了全局建模能力。HTNet^[9]采用分层Transformer架构,通过多尺度特征提取提升了识别性能,在不同层级捕获不同粒度的特征。Micron-BERT^[9]将BERT架构应用于微表情识别,取得了较好的性能。MOL^[12]、CSAFEM^[13]等基于光流-地标联合估计或通道自注意力的方法也取得较好效果。然而,在数据稀缺的微表情识别场景中,Transformer的弱归纳偏置可能导致过拟合,需要领域先验知识的补充。微表情数据集通常较小(如SMIC^[14]仅包含164个样本),这使得Transformer模型难以充分学习有效的特征表示。此外,微表情具有持续时间短、强度低、区域性强等特点,现有的通用深度学习方法难以充分利用微表情的解剖结构先验知识,限制了识别性能的进一步提升。

1.2 FACS面部动作编码系统

面部动作编码系统(Facial Action Coding System, FACS)^[10]定义了44个面部动作单元(Action Units, AUs),描述了面部肌肉激活模式与表情的关系。FACS系统地提供了经过验证的解剖-表情映射关系,为表情识别提供了重要的先验知识。例如,AU1(Inner Brow Raiser)和AU2(Outer Brow Raiser)与surprise表情相关,AU12(Lip Corner Puller)与happiness表情相关。不同表情通常涉及多个AUs的组合,如happiness需要AU12(唇部)和AU6(眼部周围)的组合。

基于AU的方法试图利用这些先验知识提升识别性能。Zhao等^[15]提出的深度区域和多标签学习(Deep Region and Multi-label Learning, DRML)方法通过学习区域特征和AU标签之间的关系来提升检测性能。Li等^[14]结合区域适应、多标签学习和最优时序融合进行AU检测,通过多任务学习同时处理多个AUs的检测。近年来,AC2D^[16]框架使用自适应约束的自注意力机制结合因果干预来应对AU的细微性和多样性。AUFormer^[17]利用Vision Transformer,引入知识专家混合

(Mixture-of-Knowledge Expert) 机制进行个性化 AU 学习。这些方法在 AU 检测任务上取得了不错的效果,其优点在于能够显式利用 FACS 的解剖-表情映射关系。然而,这些基于 AU 的方法往往需要额外的 AU 标注数据,标注成本高昂且耗时,在实际应用中受到限制。此外,这些方法通常采用两阶段设计(先检测 AU 再分类),可能导致误差传播,且难以实现端到端优化。

为了在无需 AU 标注的情况下利用 FACS 先验,一些方法尝试通过外部知识库或后处理方式引入 FACS 先验。MER-CLIP^[18] 使用视觉-语言对齐方法,将 AU 标签转换为详细的面部肌肉运动文本描述,通过对齐视觉动态与基于 AU 的文本表示来指导细粒度时空微表情学习。该方法还引入了情感推理模块来捕获微表情模式与情感之间的细微关系。这些方法的优点在于避免了昂贵的 AU 标注,能够利用大规模预训练的视觉-语言模型。但缺点在于先验知识的引入往往采用后处理方式或通过跨模态对齐实现,难以与模型的特征学习过程深度融合,可能造成信息瓶颈。例如, MER-CLIP 通过跨模态对齐引入 FACS 先验,但先验强度是固定的,无法根据样本的表情强度动态调整。此外,这些方法通常使用固定的先验强度,无法适应不同样本的表情强度差异。

1.3 先验引导

在 Transformer 架构中集成先验知识是一个重要研究方向。Vaswani 等^[19] 提出的自注意力机制奠定了 Transformer 的基础。位置编码是最常见的先验知识集成方式,其提供了空间位置信息。ViT^[8] 使用可学习的 1D 位置编码,将 2D 图像位置映射到 1D 序列位置。DETR^[20] 使用 2D 位置编码,更好地保留了空间结构信息。这些方法的优点在于实现简单、计算开销小,但缺点在于只能提供位置信息,难以引入更复杂的领域先验知识。

在微表情识别领域,如何将 FACS 先验知识深度集成到 Transformer 架构中是一个重要研究方向。现有方法往往使用固定的先验强度,通过固定的权重或掩码来应用先验知识。这些方法包括硬掩码、软掩码以及基于知识注入的方法。硬掩码方法通过将非关联区域的 Attention score 设为 $-\infty$ 来约束注意力范围,其优点在于能够强制模型关注特定区域,但缺点在于完全禁止跨区域交互,可能丢失重要的全局信息。软掩码方法通过

固定的权重矩阵来引导注意力,允许跨区域交互,但使用固定的先验强度,假设所有样本对先验引导的需求是相同的。

然而,在实际应用中,不同样本的表情强度、区域激活程度是动态变化的。固定先验强度的方法可能无法适应这种动态变化。例如,强表情样本(如明显的 happiness)的表情强度较高,面部肌肉运动明显,数据驱动能力较强,可能需要较少的先验引导,让模型更多地依赖从数据中学习到的特征;弱表情样本(如 subtle 的 surprise)的表情强度较低,面部肌肉运动细微,数据驱动能力不足,可能需要更多的先验引导,让模型更多地依赖 FACS 先验知识来辅助识别。如何在保持 Transformer 全局性的同时,根据样本特征自适应调整先验引导强度,实现先验引导和数据驱动的动态平衡,仍是一个未解决的问题。

2 模型

本文提出的 AFAT-MER 的整体架构如图 1 所示。给定输入视频序列,模型首先通过 TV-L1 算法^[5] 提取起始帧和峰值帧之间的光流图。然后通过解剖引导的 Patch Embedding 将光流图分割成 $G \times G$ 个 patches(本文中 $G=4$),并根据每个 patch 所属的解剖区域注入 FACS 先验。接着经过 3 层自适应解剖引导的分层 Transformer 处理,在每层的自注意力机制中通过基于表情强度的自适应软先验引导增强同解剖区域 patches 的交互。最后通过解剖感知的特征融合模块整合分层特征,并输出分类结果。

2.1 光流预处理

给定输入视频序列,首先采用 D&C-ROIs(Divide and Conquer-Region of Interest)方法选择峰值帧。建立起始帧与后续帧之间的关联,通过计算起始帧和后续帧的直方图余弦相似度来选择峰值帧:

$$d = \frac{\sum_{i=1}^{N_{bin}} h_1(i) \times h_2(i)}{\sqrt{\sum_{i=1}^{N_{bin}} h_1(i)^2} \times \sqrt{\sum_{i=1}^{N_{bin}} h_2(i)^2}} \quad (1)$$

其中, N_{bin} 为直方图的分箱数, h_1 和 h_2 分别为起始帧和后续帧在面部感兴趣区域(ROI)内的灰度直方图。余弦相似度 $d \in [0, 1]$, d 越小表示两帧的外观差异越大。在微表情识别中,峰值帧对应表情强度达到顶点的时刻,此时面部相对起始帧的变化最大,故选择 d 最小的帧作为峰值帧,以捕获微

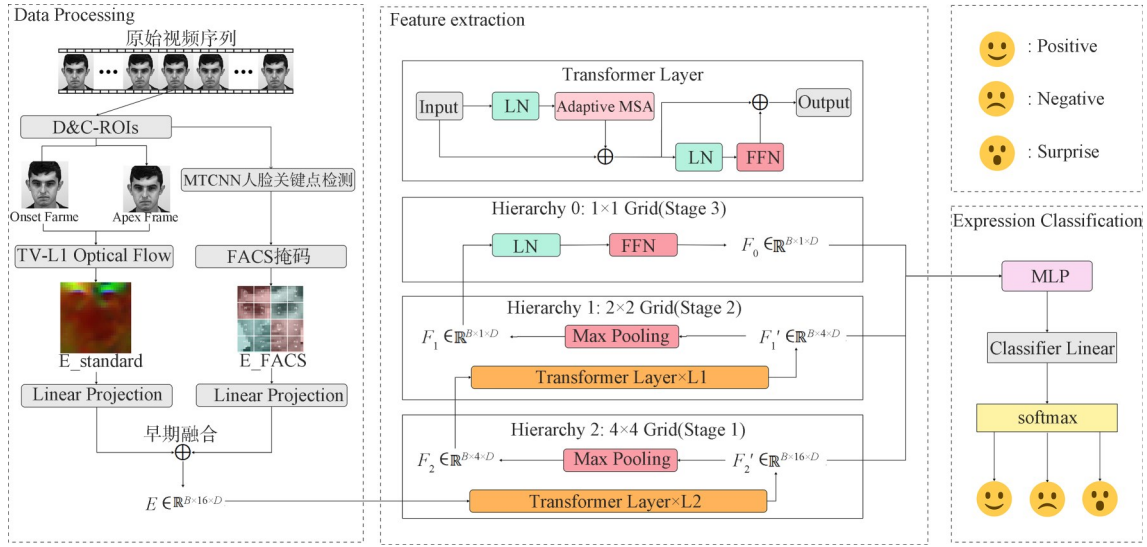


图1 AFAT-MER模型框架图

Fig. 1 Framework diagram of the AFAT-MER model

表情最显著的运动信息。提取起始帧(Onset Frame) $I_{\text{onset}} \in \mathbb{R}^{H \times W \times 3}$ 和峰值帧(Apex Frame) $I_{\text{apex}} \in \mathbb{R}^{H \times W \times 3}$, 其中 H 和 W 分别是图像的高度和宽度。使用 TV-L1 算法^[5] 计算帧间光流, 得到光流特征图 $V \in \mathbb{R}^{H \times W \times 2}$, 其表述如下:

$$V = (u(x, y), v(x, y)) \mid x \in [1, W]; y \in [1, H] \quad (2)$$

其中, $u(x, y)$ 和 $v(x, y)$ 是光流特征图 V 在位置 (x, y) 处的水平和垂直分量, x 和 y 分别是像素的列索引和行索引。

2.2 解剖引导的 Patch Embedding

将第 2.1 节得到的光流特征图 V 统一缩放到固定空间分辨率 $P \times P$, 然后分割成 $G \times G$ 个 patches。同时使用 MTCNN^[21] 在峰值帧上检测面部关键点, 获取左眼(LE)、右眼(RE)、左唇(LM)和右唇(RM)共 4 个区域的关键点坐标, 生成对应的解剖学掩码 $M \in \mathbb{R}^{4 \times P \times P}$ 。通过计算 patch 与解剖区域的加权重叠度, 将每个 patch 映射到对应的解剖区域, 得到区域标签 $A_{\text{FACS}}(x) \in \{1, 2, 3, 4\}^{G^2}$, 其中 x 表示当前输入样本。标准 Patch Embedding 通过线性投影得到 $E_{\text{standard}} \in \mathbb{R}^{B \times G^2 \times D}$, 基于区域标签, 通过可学习嵌入层得到 FACS 先验嵌入 $E_{\text{FACS}}(x)$, 每个区域的嵌入编码了该解剖部位在 FACS 中的典型运动模式。早期融合通过将标准嵌入和 FACS 先验嵌入相加实现:

$$E = E_{\text{standard}} + \text{Project}(E_{\text{FACS}}(x)) \in \mathbb{R}^{B \times G^2 \times D} \quad (3)$$

其中, Project 是线性投影层, 将 FACS 先验嵌入维

度映射到标准嵌入维度 D 。采用相加而非拼接的融合方式, 使解剖先验在特征空间内与运动特征直接叠加, 从 Patch Embedding 阶段即参与表示学习, 避免后融合可能造成的信息瓶颈。

2.3 自适应解剖引导的分层 Transformer

2.3.1 基于 FACS 的解剖先验矩阵

基于 MTCNN^[21] 检测的关键点和掩码 M_k ($k \in \text{LE, RE, LM, RM}$), 计算每个解剖区域的加权质心 c_k , 进而得到区域间的欧氏距离 $d(k, j)$ 。基于 FACS 中 AU 的协同激活规律, 定义功能相似度 $S_{\text{func}}(k, j)$, 其具体数值通过超参数获得: 同一区域为 1.0; 同侧眼-唇(如 LE 与 LM)因常协同表达(如微笑时 AU12 与 AU6)设为 0.8; 对侧眼-唇(如 LE 与 RE)因跨半脸协调较弱设为 0.6; 其余跨区域组合设为 0.3。由此得到:

$$S_{\text{func}}(k, j) = \begin{cases} 1.0, & \text{if } k = j \\ 0.8, & \text{if } \{k, j\} \in \{\{\text{LE, LM}\}, \{\text{RE, RM}\}\} \\ 0.6, & \text{if } \{k, j\} \in \{\{\text{LE, RE}\}, \{\text{LM, RM}\}\} \\ 0.3, & \text{otherwise} \end{cases} \quad (4)$$

将空间邻近性与功能协同性结合, 生成区域级解剖学偏置矩阵:

$$B_{\text{anat}}^{k, j} = \exp\left(-\frac{d(k, j)^2}{2\sigma^2}\right) \cdot S_{\text{func}}(k, j) \quad (5)$$

其中, $d(k, j)$ 为区域 k 与 j 质心间的欧氏距离, σ 控制空间衰减速率(通常取 patch 网格对角线长度的 0.3~0.5 倍)。高斯项使空间邻近区域获得更高

先验, S_{func} 项使 FACS 上功能相关的区域(即便空间较远)也获得适度增强,符合微表情中多区域协同激活的解剖学特性。将 B_{anat} 按 patch 所属区域扩展到 patch 级,得到 Hierarchy 2 的样本自适应先验 $P_{\text{FACS}}^{(2)}(x) \in \mathbb{R}^{G^2 \times G^2}$, 其中 $P_{\text{FACS}}^{(2)}(x)[i, j] = B_{\text{anat}}^{r_i, r_j}$ 为 patch i 所属区域。为 Hierarchy 1 构建中粒度先验 $P_{\text{FACS}}^{(1)}(x) \in \mathbb{R}^{(G/2)^2 \times (G/2)^2}$, 通过块池化 $P_{\text{FACS}}^{(2)}(x)$ 实现:

$$P_{\text{FACS}}^{(1)}(x)[a, b] = \frac{1}{|\mathcal{G}_a| \cdot |\mathcal{G}_b|} \sum_{i \in \mathcal{G}_a} \sum_{j \in \mathcal{G}_b} P_{\text{FACS}}^{(2)}(x)[i, j] \quad (6)$$

其中, $\mathcal{G}_a$ 为 Hierarchy 1 中 token a 对应的 $(G/2) \times (G/2)$ 块在 Hierarchy 2 中的 patch 索引集合($G=4$ 时为 $2 \times 2=4$ 个 patch)。该池化将细粒度 patch 级先验聚合为区域级先验,与 Hierarchy 1 的粗粒度 token 对应,实现多尺度先验引导。

2.3.2 基于表情强度的自适应先验强度

对于先验引导强度,本文通过每个样本的表情强度动态调整。对于每个解剖区域 r ,对属于该区域的 tokens 进行平均池化得到区域特征 $\bar{f}_r(x) \in \mathbb{R}^{B \times D}$ 。在微表情情境下, $\bar{f}_r(x)$ 的 L2 范数 $\|\bar{f}_r(x)\|_2$ 可视作该区域运动/激活强度:范数越大,表示该区域在光流特征上的响应越强,对应更明显、更易从数据中学习的表情变化。定义表情强度 $I_{\text{expression}}(x) = \frac{1}{4} \sum_r \|\bar{f}_r(x)\|_2$ 为四区域特征范数的平均,用于刻画样本整体的可辨识程度。为实现弱表情样本获得更多先验引导、强表情样本允许更多数据驱动的设计目标,采用反比映射:

$$\lambda(x) = \lambda_{\max} - (\lambda_{\max} - \lambda_{\min}) \cdot \text{clip}\left(\frac{I_{\text{expression}}(x) - I_{\min}}{I_{\max} - I_{\min} + \epsilon}, 0, 1\right) \quad (7)$$

其中, $\lambda_{\min} = 0.1$ 和 $\lambda_{\max} = 1.0$ 为先验强度的下限与上限, $I_{\min}$ 和 $I_{\max}$ 由训练集统计得到, ϵ 为防除零小常数。clip 将标准化后的强度限制在 $[0, 1]$ 内。该映射使: $I_{\text{expression}}$ 较大(强表情)时 $\lambda(x)$ 较小,先验引导减弱,模型更依赖数据; $I_{\text{expression}}$ 较小(弱表情)时 $\lambda(x)$ 较大,先验引导增强,弥补数据信号的不足。 $\lambda(x)$ 在每样本上计算一次,并在 Hierarchy 2 与 Hierarchy 1 的所有 Transformer Blocks 中共享,保持同样样本内先验强度的一致性。

2.3.3 软先验引导的自注意力

本文在 Hierarchy 2 和 Hierarchy 1 中采用软先验引导的自注意力机制,在保持 Transformer 全局

建模能力的同时利用 FACS 先验知识。对于输入特征 $x \in \mathbb{R}^{B \times N \times D}$, 通过线性投影得到查询、键和值矩阵 $Q = xW_Q$, $K = xW_K$, $V = xW_V$ 。自适应软先验引导自注意力为:

$$\text{Attention}_{\text{adaptive}}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}} + \lambda(x) \cdot \tilde{P}_{\text{FACS}}(x)\right)V \quad (8)$$

其中, $d_k = D/h$ 为每个注意力头的维度, h 为注意力头数, $\lambda(x)$ 为式(7)的自适应先验强度, $\tilde{P}_{\text{FACS}}(x)$ 是对 $P_{\text{FACS}}(x)$ 按样本线性归一化到 $[0, 1]$ 后的先验矩阵(Hierarchy 2 用 $P(2)\text{FACS}(x)$, Hierarchy 1 用 $P(1)\text{FACS}(x)$)。归一化后的先验矩阵 $\tilde{P}_{\text{FACS}}$ 作为加性 bias 添加到注意力 logits 中, $\lambda(x)$ 控制其相对影响强度。相比硬掩码约束(将非关联区域置为 $-\infty$),软先验引导允许跨区域交互,通过先验矩阵增强同解剖区域 patches 的交互强度。多头自注意力通过将多个注意力头的结果拼接得到最终输出:

$$\text{MHSA}(x) = \text{Concat}(\text{Attention}_1, \dots, \text{Attention}_h)W_O \quad (9)$$

2.3.4 Transformer Block

每个 Transformer Block 包含多头自注意力(MHSA)和前馈网络(FFN),采用残差连接和层归一化。本文采用 Post-LN(Post Layer Normalization)结构。对于输入特征 $x \in \mathbb{R}^{B \times N \times D}$, Transformer Block 的计算过程为:

$$\begin{cases} x' = \text{LayerNorm}(x + \text{MHSA}(x)), \\ x'' = \text{LayerNorm}(x' + \text{FFN}(x')) \end{cases} \quad (10)$$

其中, LayerNorm 是层归一化, $\text{FFN}(x) = \text{ReLU}(xW_1 + b_1)W_2 + b_2$ 是前馈网络。残差连接确保梯度能够有效传播,层归一化稳定训练过程。

2.3.5 分层特征提取

AFAT-MER 采用 3 层分层 Transformer 架构,每层通过 Transformer Block 提取特征,并通过聚合层进行空间下采样。

1) 底部 Hierarchy 2 层输入特征 $E \in \mathbb{R}^{B \times G^2 \times D}$ 经过 L_2 个自适应软先验引导的 Transformer Block 处理,得到 $F_2' \in \mathbb{R}^{B \times G^2 \times D}$, 然后通过聚合层下采样到 $F_2 \in \mathbb{R}^{B \times (G/2)^2 \times D}$ 。该层级使用细粒度 FACS 先验 $P_{\text{FACS}}^{(2)}(x)$ 。2) 中部 Hierarchy 1 层:输入特征 F_2 经过 L_1 个自适应软先验引导的 Transformer Block 处理,得到 $F_1' \in \mathbb{R}^{B \times (G/2)^2 \times D}$, 然后通过聚合层下采样

到 $F_1 \in \mathbb{R}^{B \times 1 \times D}$ 。该层级使用中粒度 FACS 先验 $P_{\text{FACS}}^{(1)}(x)$ 。3) 顶部 Hierarchy $0(1 \times 1)$ 层输入特征 F_1 经过 L_0 个前馈网络层处理, 得到 $F_0 \in \mathbb{R}^{B \times 1 \times D}$ 。由于该层级仅有一个 token, Self-Attention 操作退化为恒等映射, 因此仅使用前馈网络进行特征变换。聚合层通过平均池化实现空间下采样, 将 $H_{\text{in}} \times W_{\text{in}}$ 的特征图下采样到 $H_{\text{out}} \times W_{\text{out}}$, 其中 $H_{\text{out}} = H_{\text{in}}/2$, $W_{\text{out}} = W_{\text{in}}/2$ 。

2.4 特征融合与分类

从 3 个层级提取的特征分别为 $F_2 \in \mathbb{R}^{B \times (G/2)^2 \times D}$ 、 $F_1 \in \mathbb{R}^{B \times 1 \times D}$ 和 $F_0 \in \mathbb{R}^{B \times 1 \times D}$ 。将 F_2 通过全局平均池化得到 $F_2' \in \mathbb{R}^{B \times 1 \times D}$, 然后将 F_2' 、 F_1 和 F_0 在特征维度上拼接, 通过线性变换融合得到 $F_{\text{fused}} \in \mathbb{R}^{B \times 1 \times D}$ 。最后, 融合特征经过线性分类器得到情绪预测结果如式(11)所示。

$$y = \text{Softmax}(F_{\text{fused}} W_{\text{cls}} + b_{\text{cls}}) \in \mathbb{R}^{B \times 3} \quad (11)$$

其中, $W_{\text{cls}} \in \mathbb{R}^{D \times 3}$ 和 $b_{\text{cls}} \in \mathbb{R}^3$ 分别是分类器的权重矩阵和偏置向量。

2.5 损失函数

由于微表情数据集中情绪类别分布不均衡(如消极类样本通常多于积极类、惊讶类), 本文采用加权交叉熵损失进行训练如式(12)所示。

$$L = -\frac{1}{B} \sum_{i=0}^{B-1} w_{y_i} \log P(y_i | x_i) \quad (12)$$

其中, B 是批次大小, w_{y_i} 是类别 y_i 的权重, 按训练

集类别频率的倒数设置, 以缓解类别不平衡对模型的影响。

3 实验结果及分析

3.1 数据集

实验在 3 个公开的微表情数据集上评估本文方法: CASME2^[22]、SAMM^[23] 和 SMIC^[14]。如表 1~3 所示, 这些数据集是微表情识别领域广泛使用的基准数据集, 具有不同的采集条件和样本规模: 1) CASME2^[22]: 包含 26 个受试者的 255 个微表情视频样本, 视频帧率为 200 fps, 分辨率为 640×480 像素。2) SAMM^[23]: 包含 32 个受试者的 159 个微表情视频样本, 视频帧率为 200 fps, 分辨率为 2040×1088 像素。3) SMIC^[14]: 包含 16 个受试者的 164 个微表情视频样本, 分为 3 类情绪(消极、积极、惊讶)。视频帧率为 100 fps, 分辨率为 640×480 像素。

为与主流跨库对比工作一致, 本文将数据集 CASME2、SAMM 的类别标签映射到与数据集 SMIC 对齐的 3 类: 消极、积极和惊讶。其中, 将愤怒、蔑视、厌恶、恐惧、压抑、悲伤等归入消极类; 将开心归入积极类; 将惊喜归入惊讶类。此外, 设置融合数据集 3DB, 由 SMIC、CASME2 和 SAMM 等 3 个数据集在统一三分类标签下合并为同一样本池构成, 以验证本文方法的鲁棒性与识别性能。

表 1 数据集 CASME2 与 SAMM 详细信息

Tab. 1 Detailed information on the CASME2 and SAMM datasets

数据集	受试者/ 人	样本量/ 段	采样帧 率/fps	开心/段	愤怒/段	蔑视/段	厌恶/段	恐惧/段	压抑/段	惊喜/段	悲伤/段	其他/段
CASME2	26	255	200	32	—	—	63	2	27	28	4	99
SAMM	32	159	200	26	57	12	9	8	—	15	6	26

表 2 数据集 SMIC 详细信息

Tab. 2 Detailed information on the SMIC datasets

数据集	受试者/人	样本量/段	采样帧率/fps	消极/段	积极/段	惊讶/段
SMIC	16	164	100	70	51	43

3.2 实施细节

实验在 Ubuntu22.04 操作系统上使用 PyTorch 和 Python3.9 进行, 优化器使用 Adam^[24], 初始学习率为 5×10^{-5} , 批次大小(batch size)为 256。采用余弦退火学习率调度策略, 训练轮数(epochs)为 800。损失函数采用加权交叉熵损失, 类别权重根据训练集的类别分布计算。

3.2.1 实验设置

采用留一受试者交叉验证(LOSOVC)策略进行评估, 即每次选择一个受试者的数据作为测试集, 其余受试者的数据作为训练集。该策略能够更好地评估模型的泛化能力, 避免受试者间的数据泄露。

3.2.2 评估指标

如表 3 所示,统一后的各个数据集用于实验的有效样本合计为 453 个,其中消极类占比 56.95%,存在类别分布不均衡问题,故本文以非加权 $F1$ 分数(UF1)和非加权平均召回率(UAR)作为主要评价指标,正确率(ACC)作为补充参考指标,各指标均在 LOSOCV 全部折上汇总预测后计算。

表 3 数据集类别映射后类别数据(单位:段)

Tab.3 Category-wise data after dataset label mapping (Unit:sample)

标签类别	CASME2	SAMM	SMIC	3DB
积极	32	26	51	109
消极	96	92	70	258
惊讶	28	15	43	86
总计	156	133	164	453

1) 非加权 $F1$ 分数(UF1):也称为宏平均 $F1$ 分数,是评估类别不平衡多分类任务性能的常用指标。UF1 的计算需要统计留一受试者交叉验证(LOSO)所有折(fold)中每个类别(c)的假阳性数(FP_c)、真阳性数(TP_c)和假阴性数(FN_c)。每个类别的 $F1$ 分数计算为:

$$F1_c = \frac{2 \times TP_c}{2 \times TP_c + FP_c + FN_c} \quad (13)$$

未加权 $F1$ 分数为所有类别 $F1$ 分数的平均值:

$$UF1 = \frac{1}{C} \sum_{c=1}^C F1_c \quad (14)$$

2) 非加权平均召回率(UAR):评估类别不平衡场景下模型有效性。 n_c 表示每个类别(c)的总样本数。UAR 定义为:

$$UAR = \frac{1}{C} \sum_{c=1}^C \frac{TP_c}{n_c} \quad (15)$$

3) 正确率(ACC):表示所有测试样本中预测正确的比例。 N 表示对应评估设定下的测试样本总数, N_{correct} 表示预测标签与真实标签一致的样本数,ACC 定义如下:

$$ACC = \frac{N_{\text{correct}}}{N} \quad (16)$$

4) 3 数据集算术平均(Mean):表示在单一标量上汇总 SMIC、CASME2 和 SAMM 等 3 个单数据集通过留一受试者交叉验证(LOSOCV)后的结果,本文对 3 个数据集分别得到的指标取算术平均,

$$\text{例如 } UF1_{\text{mean}} = \frac{UF1_{\text{SAMM}} + UF1_{\text{CASME2}} + UF1_{\text{SMIC}}}{3},$$

UAR_{mean} 、 ACC_{mean} 同理。

3.3 实验结果与分析

由表 4 可知,在本文采用的 LOSOCV 协议与统一三分类映射设定下,AFAT-MER 在融合数据集 3DB、三数据集算术平均 Mean 以及 SMIC 数据集上取得了较为突出的结果,相关 ACC、UF1 和 UAR 指标整体处于表中领先水平;相较于基线方法 HTNet,上述设置下各项指标均有不同程度提升。说明在统一评价协议下,本文方法能够在跨库合并场景以及部分单库场景中获得稳定的性能改进。在 CASME2 数据集上,AFAT-MER 的 3 项指标均为次优,结果略低于 CausalNet,差距约为 0.06~0.57 个百分点。在 SAMM 数据集上,本文方法的 UF1 和 UAR 略低于 HFA^[25],而 ACC 高于 HFA,表明不同方法在该数据集上的各项评价指标存在一定交错,且总体差距较小。

总体而言,本文方法 AFAT-MER 在数据集 SMIC、融合数据集 3DB 及三数据集算术平均 Mean 上取得了最优结果,并在数据集 CASME2 和 SAMM 上达到最优或次优水平,说明所引入的 FACS 先验与自适应机制在不同数据分布和采集条件下具有较好的稳定性,其整体性能处于当前主流深度学习微表情识别方法的较优水平。结合第 3.4 节基于 HTNet 骨干开展的消融实验,可以进一步说明:在固定骨干与评价协议下,本文所引入的解剖先验融合与自适应机制能够带来一致可观察的性能增益。

3.4 消融实验

本节消融实验的目的在于在统一骨干和训练设定下,对所提模块的独立贡献进行归因分析,而非重新比较不同完整模型的整体优劣。因此,所有消融均固定为与表 4 中基线方法 HTNet 一致的骨干结构、输入形式、数据划分、LOSOCV 协议及训练超参数,仅逐步引入本文提出的早期融合、软先验引导与自适应先验强度机制,以分析各组成部分在固定骨干下的作用。

由于本文方法 AFAT-MER 是以 HTNet 为骨干构建的,因此本文选择 HTNet 作为渐进式消融底座。此外,方法 CausalNet 虽以 HTNet 为编码器,但已额外叠加因果建模等机制,不再属于仅包含共享骨干的统一对照平台;若在其完整模型上再叠加本文模块,性能将同时受两类增强及其交互影响,难以单独反映本文所提出的 FACS 与 $\lambda(x)$ 的独立贡献。同时,HFA、Micron-BERT 等

表 4 对比实验表(单位:%)
Tab. 4 Comparison experiment(Unit:%)

方法	SMIC			CASME2			SAMM			Mean			3DB		
	ACC	UF1	UAR	ACC	UF1	UAR	ACC	UF1	UAR	ACC	UF1	UAR	ACC	UF1	UAR
CSAFEM	78.20	76.05	76.39	94.00	92.54	92.98	79.10	77.32	78.51	83.77	81.97	82.63	83.10	82.39	83.00
HTNet	82.80	80.49	79.05	96.40	95.32	95.12	82.50	81.31	81.24	87.23	85.71	85.14	86.20	86.03	84.75
Micron-Bert	85.50	84.15	83.84	91.20	90.34	89.14	85.20	83.86	84.75	87.63	86.57	85.91	86.80	86.21	85.97
MOL	76.50	81.00	72.34	91.80	90.08	89.92	89.60	87.72	87.00	85.97	86.93	83.09	87.00	87.79	85.42
HFA	79.50	77.86	77.86	98.10	97.38	97.54	91.00	90.02	89.38	89.53	87.92	87.78	88.40	88.00	86.60
CausalNet	84.27	84.05	84.33	98.21	97.48	97.82	90.83	86.62	83.04	91.10	89.38	88.40	90.20	89.30	89.41
AFAT-MER(Ours)	86.26	84.29	86.26	98.15	97.40	97.25	91.20	89.08	87.31	91.87	90.26	90.27	91.60	89.73	91.47

注: 、 、 分别表示最优、次优、第三优。
Note: 、 、 respectively represent the best, second-best, and third-best results.

方法与本文在整体架构、输入形式或训练策略上存在差异,缺乏与本文模块一一对应的自然插入位置,不利于严格归因。

上述消融均在单数据集 SMIC、CASME2、SAMM 及融合数据集 3DB 上进行,鉴于 UF1 和 UAR 更适合类别不平衡场景,本节消融主要报告 UF1 和 UAR,ACC 变化趋势与其基本一致,故不

再重复列出。

3.4.1 组件影响分析

表 5 展示了从 HTNet 基线逐步添加组件的渐进式验证结果。基线仅使用标准分层 Transformer,不包含 FACS 先验。依次添加早期融合、软先验引导和自适应先验强度机制。

表 5 核心组件逐步有效性分析(单位:%)
Tab. 5 Ablation study on core components(Unit:%)

配置	3DB	3DB	SMIC	SMIC	CASME2	CASME2	SAMM	SAMM
	UF1	UAR	UF1	UAR	UF1	UAR	UF1	UAR
Baseline (HTNet)	86.03	84.75	80.49	79.05	95.32	95.12	81.31	81.24
+ Early Fusion	87.10	86.50	81.50	81.80	95.85	95.70	83.50	82.80
+ Soft Prior	88.20	88.30	82.80	84.20	96.50	96.35	86.20	85.50
+ Adaptive Lambda (Full Model)	89.73	91.47	84.29	86.26	97.40	97.25	89.08	87.31

注:其中粗体表示该列最优结果。
Note: boldface denotes the best result in each column.

由表 5 可知,相比基线,早期融合在 SAMM 数据集上 UF1 和 UAR 分别提升约 2.19 和 1.56 个百分点,在 SMIC 数据集上分别提升约 1.01 和 2.75 个百分点,在 CASME2 数据集上分别提升约 0.53 和 0.58 个百分点,在 3DB 数据集上分别提升约 1.07 和 1.75 个百分点。

在早期融合基础上,软先验引导进一步提升了性能,在 SAMM 数据集上 UF1 和 UAR 分别提升约 2.70 和 2.70 个百分点,在 SMIC 数据集上分别提升约 1.30 和 2.40 个百分点,在 CASME2 数据集上分别提升约 0.65 和 0.65 个百分点,在 3DB 数据集上分别提升约 1.10 和 1.80 个百分点。

相比固定先验强度,自适应机制在 SAMM 数据集上 UF1 和 UAR 分别提升约 1.60 和 1.30 个百分点,在 SMIC 数据集上分别提升约 0.70 和 1.00

个百分点,在 CASME2 数据集上分别提升约 0.45 和 0.45 个百分点,在 3DB 数据集上分别提升约 0.60 和 1.20 个百分点。早期融合带来的提升相对更明显,后续模块的增量可能因先验已部分注入而减弱;在当前实验设定下,各组件均带来一致的正向增益。

3.4.2 先验矩阵构建策略分析

表 6 对比了 3 种先验注入策略:硬掩码禁止跨区域交互,固定几何先验采用平均脸型关键点而忽略个体差异,自适应几何先验(本文)结合 FACS 功能协同与样本自适应几何距离。结果表明自适应几何先验最优,支持引入解剖学先验的必要性。

3.4.3 自适应先验强度机制分析

表 7 对比了 $\lambda(x)$ 的不同设定,包括固定 $\lambda(\lambda=0.3, \lambda=0.5$ 或 $\lambda=0.8)$ 、可学习标量、正比映射与

反比映射(本文)。固定 λ 难以适应样本间表情强度差异, $\lambda=0.5$ 时相对较优,但在各列上仍低于自适应反比映射约0.5~1.5个百分点;可学习标量虽能拟合全局强度,但无法按样本动态调整。正比映射与本文假设相反,性能低于反比映射约

0.6~1.6个百分点,与“弱表情更依赖先验、强表情更依赖数据”的设计一致。在本消融设定及表中所列各数据集列上,反比映射均为最优,表明自适应先验强度机制有效。

表 6 先验矩阵构建策略分析(单位:%)

Tab. 6 Analysis of prior matrix construction strategies(Unit:%)

策略	3DB	3DB	SMIC	SMIC	CASME2	CASME2	SAMM	SAMM
	UF1	UAR	UF1	UAR	UF1	UAR	UF1	UAR
Hard Mask	87.80	87.90	82.50	83.20	96.40	96.25	86.50	85.30
Fixed Geometry	88.60	89.20	83.50	84.80	96.80	96.65	87.90	86.70
Adaptive Anatomical Prior (Ours)	89.73	91.47	84.29	86.26	97.40	97.25	89.08	87.31

注:其中粗体表示该列最优结果。

Note: boldface denotes the best result in each column.

表 7 自适应先验强度机制分析(单位:%)

Tab. 7 Analysis of the adaptive prior strength mechanism(Unit:%)

方法	3DB	3DB	SMIC	SMIC	CASME2	CASME2	SAMM	SAMM
	UF1	UAR	UF1	UAR	UF1	UAR	UF1	UAR
$\lambda=0.3$	88.40	89.00	83.20	84.80	96.70	96.55	87.20	86.20
$\lambda=0.5$	88.70	89.40	83.50	85.20	96.85	96.70	87.60	86.60
$\lambda=0.8$	88.50	89.20	83.30	85.00	96.75	96.60	87.40	86.40
Learnable Scalar	88.80	89.50	83.60	85.30	96.90	96.75	87.80	86.80
Proportional Mapping	88.60	89.30	83.40	85.10	96.80	96.65	87.50	86.50
Inverse Mapping (Ours)	89.73	91.47	84.29	86.26	97.40	97.25	89.08	87.31

注:其中粗体表示该列最优结果。

Note: boldface denotes the best result in each column.

4 结 论

本文提出了一种基于FACS理论的自适应解剖引导Transformer微表情识别方法AFAT-MER。该方法在HTNet分层Transformer架构的基础上,将FACS系统定义的解剖-表情映射关系融入模型中,有效提升了微表情识别性能。为充分利用FACS先验知识,本文在Patch Embedding阶段通过早期融合注入解剖先验,实现了先验信息与特征表示的有效融合。在自注意力机制中,本文采用软先验引导增强同解剖区域patches的交互,在保持Transformer全局建模能力的同时充分利用FACS先验知识。此外,本文提出了基于表情强度的自适应先验强度机制,根据样本的表情强度动态调整先验引导强度,解决了固定先验无法适应动态数据的问题。在SMIC、CASME2、SAMM及融合数据集3DB上的实验结果表明,AFAT-MER 相较 HTNet 基线在 ACC、UF1 和 UAR 等指标上取得了稳定提升,并在多种代表性

方法对比中表现出较强竞争力。需要指出的是,本文方法在CASME2和SAMM的部分指标上仍略低于个别最新方法,说明不同先验建模策略在特定数据分布下各有优势,后续仍需通过更广泛的跨骨干与跨设置验证进一步提升泛化性与稳健性。

参考文献:

- [1] Yan W J, Wu Q, Liang J, *et al.* How fast are the leaked facial expressions: The duration of micro-expressions[J]. J Nonverbal Behav, 2013, 37(4): 217-230.
- [2] Ekman P, Friesen W V. Nonverbal leakage and clues to deception[J]. Psychiatry, 1969, 32(1): 88-106.
- [3] Picard R W. Affective computing[M]. Cambridge, Mass: MIT Press, 2000.
- [4] Zhao G, Pietikainen M. Dynamic texture recognition using local binary patterns with an application to facial expressions[J]. IEEE Trans Pattern Anal Mach

- Intell, 2007, 29(6): 915-928.
- [5] Zach C, Pock T, Bischof H. A duality based approach for realtime TV-L 1 optical flow [M]//Pattern recognition. Berlin-Heidelberg: Springer, 2007: 214-223.
- [6] He K, Zhang X, Ren S, *et al.* Deep residual learning for image recognition [C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE, 2016: 770-778.
- [7] Huang G, Liu Z, Van Der Maaten L, *et al.* Densely connected convolutional networks [C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI: IEEE, 2017: 2261-2269.
- [8] Dosovitskiy A, Beyer L, Kolesnikov A, *et al.* An image is worth 16 $\times$ 16 words: transformers for image recognition at scale [PP/OL]. V1. [2025-12-20]. <https://arxiv.org/abs/2010.11929>
- [9] Wang Z, Zhang K, Luo W, *et al.* HTNet for micro-expression recognition [J]. Neurocomputing, 2024, 602: 128196.
- [10] Ekman P, Friesen W V. Facial action coding system: A technique for the measurement of facial movement [M]. Palo Alto: Consulting Psychologists Press, 1978.
- [11] Tran D, Bourdev L, Fergus R, *et al.* Learning spatiotemporal features with 3D convolutional networks [C]//2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE, 2015: 4489-4497.
- [12] Shao Z, Cheng Y, Li F, *et al.* MOL: Joint estimation of micro-expression, optical flow, and landmark via transformer-graph-style convolution [J]. IEEE Trans Pattern Anal Mach Intell, 2025, 47(10): 8756-8768.
- [13] Zhao S, Li S, Zhang Y, *et al.* Channel self-attention residual network: Learning micro-expression recognition features from augmented motion flow images [J]. IEEE Trans Affect Comput, 2025, 16(4): 3215-3230.
- [14] Li X, Pfister T, Huang X, *et al.* A spontaneous micro-expression database: Inducement, collection and baseline [C]//2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), 2013: 1-6.
- [15] Zhao K L, Chu W S, Zhang H G. Deep region and multi-label learning for facial action unit detection [C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016: 3391-3399.
- [16] Shao Z, Zhu H, Zhou Y, *et al.* Facial action unit detection by adaptively constraining self-attention and causally deconfounding sample [J]. Int J Comput Vis, 2025, 133(4): 1711-1726.
- [17] Yuan K, Yu Z, Liu X, *et al.* AUFormer: Vision transformers are parameter-efficient facial action unit detectors [M]//Computer vision-ECCV 2024. Gewerbestrasse, Cham, Switzerland: Springer Nature Switzerland, 2024: 427-445.
- [18] Liu S, Mao X, Zhao S, *et al.* MER-CLIP: AU-guided vision-language alignment for micro-expression recognition [J]. IEEE Trans Affective Comput, 2025, 16(4): 3028-3042.
- [19] Vaswani A, Shazeer N, Parmar N, *et al.* Attention is all you need [C]//Advances in Neural Information Processing Systems 30, 2017: 5998-6008.
- [20] Carion N, Massa F, Synnaeve G, *et al.* End-to-end object detection with transformers [C]//European Conference on Computer Vision. Cham: Springer, 2020: 213-229.
- [21] Zhang K, Zhang Z, Li Z, *et al.* Joint face detection and alignment using multitask cascaded convolutional networks [J]. IEEE Signal Process Lett, 2016, 23(10): 1499-1503.
- [22] Yan W J, Li X, Wang S J, *et al.* CASME II: An improved spontaneous micro-expression database and the baseline evaluation [J]. PLoS One, 2014, 9(1): e86041.
- [23] Davison A K, Lansley C, Costen N, *et al.* SAMM: A spontaneous micro-facial movement dataset [J]. IEEE Trans Affect Comput, 2018, 9(1): 116-129.
- [24] Kingma D P, Ba J. Adam: A method for stochastic optimization [PP/OL]. V9. (2017-01-30) [2025-12-20]. <https://arxiv.org/abs/1412.6980>.
- [25] Zhang M, Yang W, Wang L, *et al.* HFA-Net: Hierarchical feature aggregation network for micro-expression recognition [J]. Complex & Intelligent Systems, 2025, 11(3): 169.

(责任编辑: 伍少梅)