捕捉人脸宏-微观变化的自监督人脸动作单元检测

刘柯君[1] 刘袁缘[1] 李康林[1] 唐 厂[2] 覃 杰[3] 罗 威[4]

[1](中国地质大学(武汉)计算机学院 武汉 430074)

[2](华中科技大学软件学院 武汉 430074)

[3](安徽光智科技有限公司 安徽 滁州 239000)

[4](中国舰船研究设计中心 武汉 430064)

摘 要 人脸动作单元(Facial Action Unit, FAU)检测对情感识别和人脸生成等任务至关重要。现有的全监督 FAU 方法依赖高质量标注数据,费时费力且泛化性差。因此,如何利用大量无标签人脸数据的自监督 FAU 检测成为研究热点。现有的基于人脸对比学习的自监督 FAU 方法过于关注人脸全局信息,难以精准获取人脸局部单元的运动信息。因此,本文提出了一种捕捉人脸宏-微观变化的自监督人脸动作单元检测方法(Macro-Micro Changes based Self-supervised Facial Action Unit Detection, MC-SFAU)。该方法主要包括三个模块:宏观人脸双流对比、微观人脸区域重建和宏-微观变化交互,从而实现鲁棒的自监督 FAU 检测。首先,宏观人脸双流对比模块引入宏观人脸流和引导人脸流,学习宏观人脸变化;然后,微观人脸区域重建模块引入区域重建损失,精准捕捉细微人脸运动局部变化;最终,引入宏-微观变化交互模块,强化宏观和微观人脸运动知识交互,从而获得精准 FAU 表征。该方法在无需 FAU 标注的人脸数据集(VoxCeleb)上进行预训练,在两个 FAU 数据集上(BP4D 和 DISFA)评估,结果证明了 MC-SFAU 方法有效地提升了现有自监督 FAU 检测的精度,分别获得了 1.4%和 5.2%的相对提升。

关键词 动作单元检测;自监督学习;对比学习;Landmark 检测;重建学习

中图法分类号 TP18 **DOI 号** 10.11897/SP.J.1016.2026.01061

Macro-Micro Changes based Self-Supervised Facial Action Unit Detection

LIU Ke-Jun[1] LIU Yuan-Yuan[1] LI Kang-Lin[1] TANG Chang[2] QIN Jie[3] LUO Wei[4]

[1](School of Computer Science, China University of Geosciences (Wuhan), Wuhan 430074)

[2](School of Software Engineering, Huazhong University of Science and Technology, Wuhan 430074)

[3](Anhui Guangzhi Technology, Chuzhou, Anhui 239000)

[4](China Ship Development and Design Center, Wuhan 430064)

Abstract Facial Action Unit (FAU) detection is essential for emotion recognition and face generation. Existing fully supervised FAU methods rely on costly annotated data and lack generalization. Self-supervised FAU detection using unlabeled facial data has become a key research area. Current methods based on contrastive learning focus too much on global facial information, missing precise local motion details. This paper proposes Macro-Micro Changes based Self-supervised FAU Detection (MC-SFAU), which includes three modules: macro facial dual-stream contrast, micro facial region reconstruction, and macro-micro change interaction. First, the macro facial

收稿日期:2025-07-15;在线发布日期:2026-02-24。本课题得到国家自然科学基金(No. 62076227)、湖北省自然科学基金(No. 2023AFB572)、湖北省自然科学基金项目和智能地球信息处理湖北省重点实验室(No. KLIGIP-2022-B10)资助。**刘柯君**,博士研究生,主要研究领域为情感计算、机器学习、计算机视觉。E-mail:liukejun@cug.edu.cn。**刘袁缘(通信作者)**,博士,副教授,中国计算机学会(CCF)会员,主要研究领域为情感计算、机器学习与模式识别、计算机视觉。E-mail:liuyy@cug.edu.cn。**李康林**,硕士,主要研究领域为情感计算、机器学习、计算机视觉。**唐 厂**,博士,副教授,中国计算机学会(CCF)高级会员,主要研究领域为情感计算、机器学习与模式识别、计算机视觉。**覃 杰**,硕士,主要研究领域为情感计算、机器学习与模式识别、计算机视觉。**罗 威**,博士,高级研究员,主要研究领域为机器学习与模式识别、计算机视觉。

dual-stream contrast module uses global and guide facial streams to learn macro facial changes. Next, the micro facial region reconstruction module introduces a region reconstruction loss to precisely capture subtle local facial movements. Finally, the macro-micro change interaction module strengthens the interaction between macro and micro facial motion knowledge, leading to accurate FAU representations. The method is pre-trained on the unlabeled VoxCeleb dataset and evaluated on BP4D and DISFA, MC-SFAU shows relative improvements of 1.4% and 5.2%, respectively, demonstrating its effectiveness in enhancing self-supervised FAU detection accuracy.

Keywords action unit detection; self-supervised training; contrast learning; Landmark detection; reconstruction

1 引 言

人脸动作单元(Facial Action Unit,FAU)是描述人脸肌肉激活的行为单元,是分析人脸表情的基础。人脸动作编码系统对常见 FAU 进行了注释[1],如 AU12 表示拉动嘴角的动作。FAU 检测能够揭示人脸不同区域的运动激活,进而分析情感状态。FAU 检测在情绪识别、人机交互、智能机器人、健康医疗和刑侦检测等领域[2-3]具有重要应用价值。

早期的 FAU 检测研究主要集中于全监督学习,这些方法依赖大量标注数据来训练模型。例如,EAC-Net[4]和 ROI[5]通过提取人脸 Landmark 周围的特征,捕捉了具有一定通用性的 FAU 特征。Wang 等人[6]提出了一种新型多尺度时间差分网络,通过自适应加权块建模不同空间尺度的跨帧面部动态特征,并采用两阶段策略对 AU 之间的关系进行分层建模。然而,这些方法仍然依赖高质量标注的特定场景图像,标注过程费时费力,且限制了其在无标注数据的实际场景中的应用。近年来,自监督学习方法逐渐受到关注。与传统的全监督学习不同,自监督学习能够从无标注数据中有效提取特征,避免了对人工标签的依赖,并在计算机视觉和自然语言处理等领域取得了显著进展[7-8]。现有的自监督方法主要包括基于对比和基于重建的方法。基于对比的方法通过对比样本之间的相似性和差异性来增强模型的表征能力,例如,Li 等人[7]提出了跨身份重建和帧时间对比的 FAU 表示方法,捕捉了面部视频中的时间一致性与演化特性。基于重建的方法则通过分解特征并重建或融合它们,利用特征一致性进行训练,例如,TAE[8]所提出的通过解耦姿态和非姿态特征来学习姿态不变的 FAU 表示。尽管取得了一定的进展,然而,现有的自监督学习方法,仍面临了以下两个挑战:一是人脸对比学习方法分离出的 FAU 表征过于关注人脸整体的宏观变化,难以精准提取人脸局部关键区域的细节信息,而 FAU 本质上是微观变化的过程;二是人脸重建方法虽然能够聚焦于细粒度的微观人脸变化,却忽略了对于宏观人脸变化的建模。事实上,FAU 的宏观与微观变化信息具有互补性,二者的结合对于提升 FAU 任务的性能至关重要。因此,为了协调这两个挑战,本文提出了捕捉人脸宏-微观变化的自监督人脸动作单元检测方法(Macro-Micro Changes based Self-supervised Facial Action Unit Detection, MC-SFAU)。该方法在大规模无标签人脸数据上预训练,能够有效捕捉宏观和微观人脸变化,提供鲁棒的 FAU 表征,进而提升 FAU 检测的泛化能力。首先,MC-SFAU 方法引入宏观人脸双流对比模块通过宏观人脸流和引导人脸流从人脸图像中提取宏观 FAU 表征,以捕捉宏观人脸变化。然后,MC-SFAU 进一步引入微观人脸区域重建模块,聚焦于人脸微观运动区域,利用区域重建损失精准捕捉微观的面部细节变化,从而获得微观 FAU 特征。此外,为了弥补宏观和微观 FAU 表征的不足,我们引入宏-微观变化交互模块,利用 Kullback-Leibler 损失函数增强宏观和微观人脸变化,强化两者的知识交互,提升 FAU 检测精度。最终 MC-SFAU 方法在无标注数据上预训练,在不同场景下的 FAU 检测任务中展现了显著性能提升。本文的主要贡献在于:(1)提出了一种捕捉人脸宏-微观变化的自监督人脸动作单元检测方法,该方法能够在无需 FAU 标注的情况下,获得鲁棒的 FAU 表征,并有效泛化到不同场景的 FAU 检测任务。(2)引入宏观人脸双流对比模块,通过对比学习从人脸图像中学习宏观人脸变化,得到宏观 FAU 表征;引入微观人脸区域重建模块,通过学习人脸 Landmark 重建微观人

脸变化,得到细致的微观 FAU 表征;引入宏-微观变化交互模块,强化宏观和微观人脸运动知识交互,提升 FAU 表征的鲁棒性。(3)我们在 VoxCeleb 数据集上进行预训练,并在 BP4D 和 DISFA 两个广泛使用的 FAU 数据集上评估。与现有方法相比,MC-SFAU 显著提高了 FAU 检测精度,在某些情况下达到全监督水平,验证了其在 FAU 检测任务中的优势。

2　相关工作

2.1　人脸动作单元检测

最近随着深度学习和卷积神经网络(CNN)的不断发展,FAU 检测取得了显著的进展。由于 FAU 与人脸肌肉的运动相对应,许多方法致力于基于位置的 FAU 检测。SEV-Net[9] 利用 AU 的语义描述作为辅助信息来提升 AU 检测性能。AU-RCNN[10] 则将先验专家知识嵌入到关键人脸区域中,以实现更精准的 AU 检测。为了避免手动预定义固定的注意力,ARL[11] 提出了一种端到端的学习方法,用于学习 AU 检测的信道和空间注意力。最近,Jacob 和 Stenger[12] 使用基于 Transformer 的编码器捕捉 FAU 之间的关系,然后进行 FAU 检测。尽管这些方法取得了一定的成功,但未能解决身份特异性对 FAU 检测模型的影响问题,因为人脸动作单元的表达在不同个体之间可能存在较大差异,甚至在同一人之间也可能变化。有一些文献提出了学习主体独立的 FAU 特征方法。其中,Zhang 等人[13] 提出了对抗性训练框架 ATF,通过最小化 FAU 损失和最大化人脸识别损失,使得学到的特征对 FAU 检测具有有效性,并且对受试者的变化保持不变。Almaev 等人[14] 通过将容易引发的 FAU 知识转移到难以引发的 FAU 上,构建了一种身份特异性的 FAU 检测模型。Tu 等人[15] 使用包含大量受试者的 FAU 和身份注释数据集,提取身份相关的图像特征,并在测试阶段进行身份感知的 FAU 检测。Yin 等人[16] 提出了 FG-Net,通过从预训练的 StyleGAN2 模型中提取注意力图,并使用 Pyramid CNN Interpreter 检测 AU,实现了强大的可泛化能力和基于注意力图的 AU 检测。Cakir 等人[17] 采用生成对抗网络作为数据增强方法,来进一步提高 AU 检测。SONet[18] 通过向特征空间和标签空间添加集合运算,增强了训练数据,使得这些特征能够推广到看不见的主题。与这些监督方法旨在消除特定于个体的影响不同,我们提出的 MC-SFAU 方法旨在通过跨身份重建,从大量未标记的人脸视频中学习独立于个体的 AU 表示。

2.2　半监督学习

近年来,越来越多的研究者尝试在标签有限的情况下进行 FAU 检测,并提出了一系列半监督检测方法。例如,Shao 等人[19] 通过使用注释充足的数据集进行训练,通过将人脸特征解耦为人脸特征和动作特征进行 FAU 检测。而 Lee 等人[20] 则通过将图像嵌入域特定属性空间和域不变内容空间,捕获不同域之间的共享信息,从而解开图像到图像转换的复杂表示。这里特征解耦的目的是将特征分解为不同的子空间[21],用于构建具有鲁棒性的 FAU 特定表示。最终,这些方法在其他 FAU 数据集上进行人脸特征提取,以进行 AU 检测。然而,这种方法容易受到不同数据集之间的域偏移影响,从而影响模型性能。为了缓解域偏移问题,Csurka 等人[22] 提出了域自适应的方法,将注释充足的数据集作为源域,将没有注释的数据作为目标域,并通过拉近源域和目标域的特征,减小域偏移的影响。Li 等人[23] 提出了一个轻量级的在线半监督框架,通过渐进式知识蒸馏学习 AU 动态稀疏注释,并将跨域信息传播到未标记数据,完成时空知识学习。然而,这种方法仍然会受到一小部分域偏移的影响,从而影响模型性能。同时,这种方法不能完全消除对标签数据的依赖。为了进一步提高模型性能并解决域偏移问题,本文提出了一种新颖的方法 MC-SFAU,旨在在标签有限情况下实现更鲁棒和高效的 FAU 检测。

2.3　自监督学习

最近,自监督学习在人脸情感领域取得了显著的进展,它利用数据本身的监督信号从大量未标记的数据中学习表示。目前,自监督人脸情感检测主要分为基于对比的方法和基于重建的方法。基于对比的方法通常采用对比学习范式进行学习,例如 Chen 等人[24] 引入了一个简单的框架(SimCLR),其中基本编码器网络和投影头被训练,以使用对比损失最大化一致性。Shang 等人[25] 提出了一个对比学习框架,通过负样本重新加权和采样技术解决类别不平衡和噪声标签问题,从而提升 AU 检测的准确性和判别特征的学习。He 等人[26] 提出了 MoCo,它包含一个动量网络来存储大量负样本的队列,以进行有效的对比学习。Chen[24] 和 He[26] 在没有负样本对的情况下学习了一般表示,并通过停止

梯度操作避免了不希望的坍塌解。这些对比学习方法学习的一般表示可能不是 AU 判别式的，因此在 FAU 任务上效果不是很理想。基于重建的方法则着重于特征恢复或者图像像素级别恢复，例如 Zhang 等人[27]提出了一种多模态混合网络，通过早期融合和掩码自动编码器重建丢弃通道，减少冗余并促进鲁棒的 AU 特征学习。PCFRL[28]提出通过解耦姿态和非姿态然后利用重建来学习姿态不变的 FAU 表示。然而这种方法强调重建过程中的微观人脸表示，弱化了宏观人脸表示在动作单元检测中的作用。与这些方法相比，我们提出的 MC-SFAU 方法融合了宏观对比学习和微观重建的细微信息，使网络能够更加准确地关注 FAU 相关的重要区域，最终获得了精确的适用于 FAU 检测的人脸表示。

3 方法原理

3.1 总体框架概述

本文提出一种新的捕捉人脸宏-微观变化的自监督人脸动作单元检测方法，称为 MC-SFAU。MC-SFAU 的总体架构如图 1 所示。MC-SFAU 包含三个模块：宏观人脸双流对比模块（Macro Facial DualStream Contrastive Module，MFDC）、微观人脸区域重建模块（Micro Facial Region Reconstruction Module，MFRR）和宏-微观变化交互模块（MacroMicro Change Interaction Module，MCI）。首先，MC-SFAU 以人脸图像 P 为输入，在 MFDC 中通过宏观人脸流和引导人脸流进行对比学习以捕

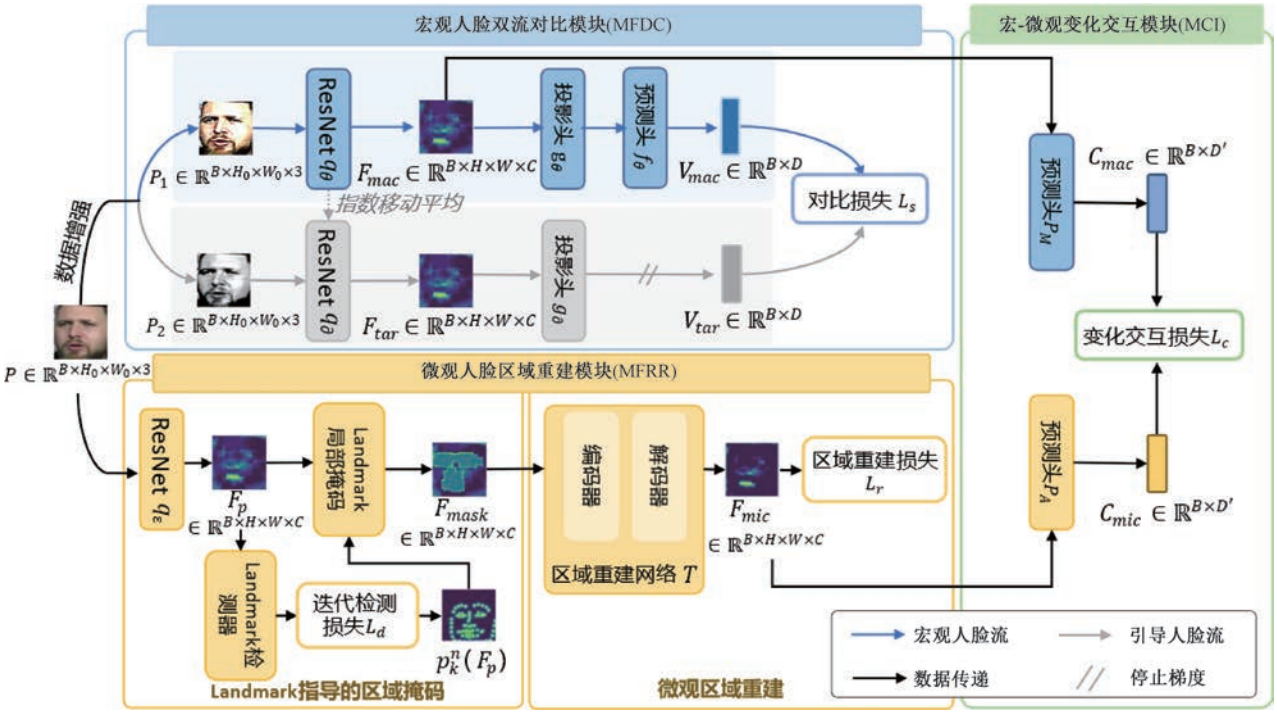

图 1 MC-SFAU 方法框架图（在宏观人脸双流对比模块（MFDC）中，输入人脸图像 $P\in\mathfrak{R}^{B\times H_0\times W_0\times 3}$ 经不同增强后分别输入宏观人脸流与引导人脸流，得到 F_{mac}，$F_{tar}\in\mathfrak{R}^{B\times H\times W\times C}$ 并分别通过映射头得到 V_{mac}，$V_{tar}\in\mathfrak{R}^{B\times D}$，二者在共享空间中进行相似度优化。在微观人脸区域重建模块（MFRR）中，输入 $F_{mask}\in\mathfrak{R}^{B\times H\times W\times C}$ 经过 Transformer 重建生成 $F_{mic}\in\mathfrak{R}^{B\times H\times W\times C}$。在宏-微观变化交互模块（MCI）中，$C_{mac}$，$C_{mic}\in\mathfrak{R}^{B\times D'}$ 被映射到统一空间，并通过 KL 散度约束实现协同优化。这里，B 表示 batch size，H_0，W_0 表示输入图像的空间分辨率，H，W 表示卷积特征图的空间分辨率，C 表示通道数，D，D'表示不同阶段的嵌入维度。）

捉宏观人脸变化，得到宏观 FAU 表征 Fmac。接着，MFRR 通过掩码重建网络聚焦于微观人脸运动的关键区域，捕捉微观 FAU 表征 Fmic。为了进一步弥补宏观和微观 FAU 表征的不足，MCI 通过 Kullback-Leibler 损失函数增强宏观和微观层次的人脸变化，强化两者的知识交互，最终获得鲁棒的 FAU 表征，提升检测精度。接下来，将详细介绍这三个模块。

3.2 宏观人脸双流对比模块（MFDC）

传统的对比学习方法依赖正负样本对进行训

练,但在 FAU 检测任务中,容易引入假负样本,导致学习偏差,从而妨碍模型捕捉准确的宏观人脸运动变化。受到 BYOL 等工作[29]的启发,我们采用双流结构并引入 EMA 机制,仅依赖正样本即可获得稳定目标,从而避免表示塌陷并提升宏观人脸运动的建模能力。宏观人脸流提取图像特征并捕捉宏观 FAU 变化,引导人脸流则通过指数移动平均(EMA)[30]机制平滑更新宏观人脸流的参数,提供稳定的学习目标。两者协同作用,确保模型高效学习宏观 FAU 表征,避免假负样本带来的偏差,提升对宏观人脸运动的建模能力。

详细地,如图 1 所示,我们将宏观人脸流的主干 ResNet 网络记作 q_θ,投影头记作 g_θ,预测网络头记作 f_θ,引导人脸流主干 ResNet 网络记作 q_∂,投影头记作 g_∂。假定输入的人脸图像为 P,为了学习到不受变换影响的稳定 FAU 特征,我们首先应用随机裁剪、随机旋转、颜色抖动等数据增强方法将图像 P 增强为一对不同视图,记作 P_1 和 P_2。于是我们记宏观人脸流和引导人脸流主干 ResNet 网络提取的人脸表征分别为宏观 FAU 表征 $F_{mac}=q_\theta(P_1)$和目标 FAU 表征 $F_{tar}=q_\partial(P_2)$。然后,宏观 FAU 表征 F_{mac} 通过投影头 g_θ 进行映射,并利用预测头 f_θ 生成宏观人脸流表征 $V_{mac}=f_\theta(g_\theta(F_{mac}))$,以预测引导人脸流表征实现自监督学习,从不同视角学习宏观人脸变化。目标 FAU 表征 F_{tar} 则通过投影头 g_∂ 得到引导人脸流表征 $V_{tar}=g_\partial(F_{tar})$,为宏观人脸流的训练提供稳定的目标。

在训练过程中,宏观人脸流的参数 θ 通过梯度反向传播不断更新,而引导人脸流的参数∂则通过对宏观人脸流参数 θ 进行指数移动平均来更新:

$$\partial \leftarrow \tau\partial+(1-\tau)\theta \tag{1}$$

其中,$\tau=0.99$,用于控制目标分支的更新速率。该双流设计为模型提供稳定的训练目标,避免受当前梯度波动干扰。在无负样本监督条件下,该设计能够有效缓解表示塌陷问题,保证学习到判别性强且稳定的人脸表示。我们将引导人脸流表征与宏观人脸流表征作为正样本进行训练,仅依赖正样本间的相似性,消除了负样本生成的复杂性。通过最大化两者之间的相似度,促使模型学习到有意义的宏观 FAU 表征,有效捕捉宏观人脸变化。最终,我们设计的对比损失 L_s 如下:

$$L_s=2-2\left(\frac{V_{mac}\cdot V_{tar}}{||V_{mac}||_2\cdot||V_{tar}||_2}\right) \tag{2}$$

该损失项是宏观人脸流表征和引导人脸流表征的 L_2 归一化均方误差。

3.3　微观人脸区域重建模块(MFRR)

通过 MFDC 获得的初步宏观 FAU 表征有助于捕捉宏观人脸变化,但难以精准获取人脸局部单元的微观运动信息。已有工作(如 BG[31]、PCFRL[28])证明了区域重建在建模局部动态中的有效性。基于此,我们引入了微观人脸区域重建模块(MFRR),专注于细微的人脸关键区域,捕捉微观人脸变化,从而提升 FAU 检测性能。MFRR 包括两个子模块:Landmark 指导的 FAU 掩码和细微的区域重建。首先,Landmark 指导的 FAU 掩码定位于 FAU 变化相关的关键区域,并进行掩码操作。接着,细微的区域重建对这些掩码区域进行重建,从而获得专注于细微区域的微观 FAU 表征,有效捕捉细微的人脸变化。

(1)Landmark 指导的区域掩码

正如前文所述,挖掘精细的微观人脸变化,对于提高 FAU 检测任务的性能至关重要。因此,我们在 MFRR 模块中引入了预训练的人脸 Landmark 检测器 Face Alignment[32],专注于微观人脸变化,如眼睛和嘴巴区域的运动。基于宏观 FAU 表征,检测器定位与 FAU 变化相关的关键区域,并对这些区域进行掩码操作,从而捕捉更多细节信息。具体而言,我们将输入图像 P 传入主干 ResNet 网络 q_ε,进而得到人脸图像表征 $F_p=q_\varepsilon(P)$,以捕获人脸的宏观信息。q_ε 与 MFDC 中的宏观人脸流主干 ResNet 网络共享参数。随后,将 F_p 输入检测器进行 k 次检测迭代。对于第一次迭代,我们直接通过 Face Alignment 检测器,获得初始的 49 个 Landmark 位置坐标 $p_1^n(F_p)(n=1,2,\cdots,49)$,并作为下一次迭代的学习目标,更新位置坐标;重复第 k 次迭代后,我们更新的 49 个 Landmark 点的坐标记为 $p_k^n(F_p)(n=1,2,\cdots,49)$。在迭代更新学习中,我们引入 Landmark 迭代检测损失 L_d,并通过最小化该损失来逐步优化 Landmark 位置更新。通过这种方式,模型能够更好地捕捉人脸关键区域的微小运动,并挖掘其结构和形状的微观变化。在数学上,我们采用均方误差(MSE)损失作为 L_d:

$$L_d=\sum_{m=2}^{k}\sum_{n=1}^{49}MSE(p_m^n(F_p),p_{m-1}^n(F_p)) \tag{3}$$

其中,$p_m^n(F_p)$是 F_p 在第 m 次迭代中对第 n 个 Landmark 点的预测,$p_{m-1}^n(F_p)$则是 F_p 在第 $m-1$ 次迭代中对第 n 个 Landmark 点的预测。每次迭代都使用前次迭代的预测结果作为当前目标,传递模型

知识和预测，逐步优化，帮助捕捉人脸微小运动变化。

基于经过 k 次迭代后检测到的 49 个人脸 Landmark 坐标点 $p_k^n(F_p)(n=1,2,\cdots,49)$，我们分析并建模每个 Landmark 点周围的关键区域，包括眉毛、眼睛、鼻子和嘴巴，这些与 FAU 变化高度相关的敏感区域。对于每个检测到的 Landmark 关键点 p_k^n，我们以其为中心生成边长为 r 像素的矩形掩码 $B(p_k^n,\gamma,r)$，其中概率参数 γ 决定该掩码是否被应用。最终，通过掩码操作 $M(\cdot)$ 将这些矩形掩码施加到人脸图像表征 F_p 上，得到掩码后的人脸表征：

$$F_{mask}=M(F_p,\{B(p_k^n,\gamma,r)\}_{n=1}^{49}) \tag{4}$$

其中，$M(\cdot)$ 的作用是将掩码覆盖区域的特征置零，仅保留未被掩码的部分，从而迫使模型依赖上下文信息进行区域重建。

(2)微观区域重建

在得到人脸图像掩码表征 F_{mask} 后，我们进一步提出基于 Transformer 结构的区域重建网络，用于重建被掩码的人脸图像特征，从而使得模型聚焦于细微的人脸关键区域，提取更加细粒度、细微的微观 FAU 表征。具体来说，对于 F_{mask}，我们采用一个由编码器和解码器组成的经典 Transformer 模型[33]作为区域重建网络 $T(\ )$，以生成能够聚焦于人脸细节信息的微观 FAU 表征 $F_{mic}=T(F_{mask})$。最后，为了实现图像重建的目标，我们利用人脸图像表征 F_p 和生成的微观 FAU 表征 F_{mic} 之间的差异作为损失来指导模型的训练，使得模型能够生成与原始图像尽可能相似的重建图像。区域重建损失 L_r 的计算方式如下：

$$L_r=\|F_p-F_{mic}\|_2 \tag{5}$$

其中，$\|\ \|_2$ 表示 L2 损失函数[34]。

3.4 宏-微观变化交互模块(MCI)

尽管 MFDC 和 MFRR 分别从宏观和微观层面提取了不同的 FAU 特征，但这两种特征存在互补性，单独的宏观或微观特征难以全面捕捉复杂的 FAU 信息。为此，我们提出宏-微观变化交互模块(MCI)，通过强化宏观和微观人脸运动知识的交互，从而获得精准的 FAU 表征，提升 FAU 检测精度。具体来说，为了将宏观 FAU 表征和微观 FAU 表征转化到相同的空间，我们在宏观 FAU 表征 F_{mac} 之后加上一个映射头 P_M，将宏观 FAU 表征映射到共享的特征空间，从而得到宏观变化表征为 $C_{mac}=P_M(F_{mac})$。经过微观人脸区域重建模块(MFRR)提取的微观 FAU 表征 F_{mac}，同样通过一个映射头 P_A 映射到相同的空间，产生的微观变化表征为 $C_{mic}=P_R(F_{mic})$。每个映射头均由两层全连接层构成，并在中间层引入非线性激活函数，最后输出固定维度的嵌入向量。该设计能够在保持宏观与微观表征差异性的同时，确保它们在共享空间中的维度一致性，为后续的分布约束和交互优化提供基础。

为了优化宏观和微观变化表征之间的协同关系，我们引入 Kullback-Leibler(KL)散度[35]损失函数，用于衡量二者在语义分布上的一致性。KL 散度并非强制两种表征相同，而是在保持层次差异的前提下，引导其在语义层面建立一致的情感表达，从而提升 FAU 表征的判别性与鲁棒性。变化交互损失表示为

$$L_c=KL(C_{mic},C_{mac}) \tag{6}$$

3.5 整体优化目标

总体而言，我们总体学习目标 L_{total} 包含了 4 个目标函数，即，对比损失 L_s、人脸 Landmark 迭代检测损失 L_d、区域重建损失 L_r，以及变化交互损失 L_c。形式上，我们描述总的学习目标如下：

$$L_{total}=L_s+L_d+L_r+\alpha\times L_c \tag{7}$$

其中，$\alpha=0.05$ 为变化交互损失权重。

4 实验设置

4.1 实验数据集设置

为确保实验结果的公平性与可比性，我们遵循 CLP[7]的设置，采用 VoxCeleb[36]数据集进行自监督预训练，并在两个主流的 FAU 检测数据集 BP4D[37]和 DISFA[38]上进行评估。数据集详细信息如下。

VoxCeleb[36]拥有 5994 位名人的视频，共 145569 段长视频，并截成 1092009 段视频片段，数据量极大。该数据集源自 YouTube，涵盖复杂多变的现实环境因素，如光照变化、姿态变化和背景干扰，具备高度的视觉多样性。我们利用该数据集进行大规模自监督预训练，以提升模型在非受控场景下的泛化能力。

BP4D[37]在受控实验环境下采集，包含 41 名受试者的 328 段视频，配有丰富的 AU 和 Landmark 注释，图像质量较高，面部姿态稳定。该数据集适合评估模型对典型 FAU 模式(如皱眉、张嘴等宏观面部动作)的建模能力。我们将数据划分为一个包含 28 名受试者 100767 张图像的训练集、一个包含 7 名受试者 24869 张图像的验证集和一个包含 6 名受试者 20940 张图像的测试集。

DISFA[38]由 27 名受试者组成,包含自然状态下的面部微表情变化。每帧图像标注了 66 个 Landmark 和 AU 强度(0-5),表现出明显的低强度 AU、个体差异大等挑战性特征。根据 CLP[7]中的设置,我们将等于或大于 2 的 AU 强度视为发生,而将其他 AU 强度视为未发生。这些帧被划分为一个训练集,包含 18 名受试者的 82971 张图像,一个验证集,包含 4 名受试者的 19275 张图像,一个测试集,包含 5 名受试者的 23898 张图像。

4.2　实验训练细节设置

我们的方法基于 PyTorch 框架,在 NVIDIA RTX4090 GPU 上实现。实验中使用 ResNet50 作为主干网络,将 Landmark 掩码概率 γ 设置为 0.6,即每次训练迭代中有 60%的概率进行掩码操作以增强特征鲁棒性。Landmark 人脸检测器的迭代次数设为 $k=50$,掩码矩形长度 r 为 40,确保覆盖足够图像区域,提升泛化能力。自监督训练中,图像大小统一为 256×256,预训练在 VoxCeleb 训练集上进行 400 个 epoch,使用动量 0.9、权重衰减 0.0005 的 SGD 优化器,批量大小 256,初始学习率 0.001,采用余弦衰减策略。

在下游 FAU 检测任务中,我们冻结 MFDC 中的宏观人脸流 ResNet 主干网络进行特征提取,并微调线性分类器进行分类,得到最终的自监督 FAU 检测结果。

4.3　对比方法设置

为全面验证所提出 MC-SFAU 方法的有效性与鲁棒性,本文选取了三类具有代表性的对比方法进行实验评估,涵盖全监督方法与自监督 AU 检测方法,具体说明如下:

全监督方法包括 SEV-Net[9]、HMP-PS[39]、ANFL[40]和 MDHR[6]等。这些方法均依赖真实 AU 标签进行训练,通常在模型架构中引入注意力机制、层次结构建模或 AU 依赖关系建模等策略,以提升面部动作单元的检测性能,代表了当前监督范式下的主流方法设计。

自监督方法包括 SimCLR[24]、MoCo[26]、CVC[41]、TAE[8]、CLP[7]、KSRL[42]、MCM[27]与 PCFRL[28]等。这些方法无需使用 AU 标签,通过引入图像级增强构建对比样本对(SimCLR、MoCo、CVC 等),或结合时序建模、Landmark 结构约束、图结构推理与区域重建机制(TAE、KSRL、MCM、PCFRL 等),增强模型对面部宏观与微观动作的表征能力,代表了当前自监督 AU 检测任务的主要研究方向。

为确保实验设置的公平性,本文统一采用 ResNet50 作为主干网络,并严格遵循 CLP 的设置,将 BP4D 和 DISFA 数据集划分为训练集、验证集和测试集。在所有预训练阶段,统一使用 VoxCeleb 作为无标签训练数据,采用标准的 InfoNCE 损失函数,设置对比温度参数 τ 为 0.07,batch size 为 256,预训练学习率设为 0.001。其中 SimCLR 和 MoCo 由我们基于公开实现复现,其余方法使用作者公布的代码或直接引用原文结果。

4.4　实验结果分析

4.4.1　MC-SFAU 在 BP4D 数据集上的结果分析

BP4D 拍摄环境稳定、图像质量高,且包含丰富的 AU 注释,适合评估模型对典型面部表情模式的建模能力。在该数据集上,我们将 MC-SFAU 与先进的全监督和自监督方法进行了比较,如表 1 所示。实验结果表明,MC-SFAU 在平均 F1 分数上达到 66.9%,相较于最优全监督方法 MDHR[6]相对提升 0.5%;同时,相较于最优自监督方法 MCM 相对提升 2.1%。这表明 MC-SFAU 在无标签学习范式下依然具备接近甚至超越全监督方法的判别能力。从具体 AU 的表现来看,MC-SFAU 在 AU6、AU12、AU14 和 AU23 等动作单元上表现突出,尤其在 AU12 和 AU14 这种与口部运动和情感表达紧密相关的单元上,相比自监督方法均实现了显著提升,说明模型在捕捉典型微表情变化方面具有较强优势。这得益于 MC-SFAU 的宏-微观特征交互机制,使其能够同时关注整体表情运动和关键区域的细粒度动态。尽管如此,MC-SFAU 在部分局部幅度较小或容易受干扰的 AU(如 AU1 和 AU7)上未取得最优结果。这类 AU 的表情变化幅度有限,极易受到光照、姿态和细微运动噪声的干扰,从而限制了模型的检测能力。换言之,宏观与微观建模在这些低幅度 AU 上难以充分发挥优势。但整体而言,MC-SFAU 在大多 AU 上均优于典型的自监督方法,并在平均性能上取得稳定领先。这些结果反映了 MC-SFAU 在无标签学习范式下所具备的强判别性特征提取能力与鲁棒泛化能力,特别是在面向复杂表情结构的任务中,有效缓解了对大量人工标注的依赖。

4.4.2　MC-SFAU 在 DISFA 数据集上的结果分析

在 DISFA 数据集上,我们的 MC-SFAU 方法与全监督 FAU 检测方法进行了比较,如表 2 所示。MC-SFAU 超越了大部分全监督方法,但在与 ANFL[40]、KS[23]、MDHR[6]的比较中稍显落后,这一

表 1 VoxCeleb→BP4D 上，本方法与全监督、自监督方法的对比

	方法/AU	1	2	4	6	7	10	12	14	15	17	23	24	平均
全监督	SEV-Net[9]	58.2	50.4	58.3	**81.9**	73.9	**87.8**	87.5	61.6	52.6	62.2	44.6	47.6	63.9
	HMP-PS[39]	53.1	46.1	56.0	76.5	76.9	82.1	86.4	64.8	51.5	63.0	**49.9**	54.5	63.4
	ANFL[40]	52.7	44.3	60.9	79.9	80.1	85.3	**89.2**	69.4	55.4	64.4	49.8	55.1	65.5
	MDHR[6]	**58.3**	**50.9**	**58.9**	78.4	**80.3**	84.9	88.2	**69.5**	**56.0**	**65.5**	49.5	**59.3**	**66.6**
自监督	SimCLR[24]	38.0	36.4	37.2	66.6	64.7	76.2	76.2	51.1	29.8	56.1	27.5	37.7	49.8
	MoCo[26]	30.8	41.3	42.1	70.2	70.4	78.7	82.5	53.3	25.2	59.1	31.5	34.3	51.6
	CVC[41]	43.9	47.8	38.7	67.0	70.4	81.8	84.4	57.5	39.5	49.3	27.1	43.6	54.2
	CLP[7]	47.4	50.9	49.5	75.8	78.7	80.2	84.1	67.1	52.0	62.7	45.7	54.8	62.4
	PCL[43]	52.3	**63.5**	51.3	**82.5**	**80.8**	**87.1**	88.7	65.6	47.1	57.6	43.8	51.0	63.6
	Lu et al.[44]	42.3	24.3	44.1	71.8	67.8	77.6	83.3	61.2	31.6	51.6	29.8	38.6	52.0
	TAE[8]	47.0	45.9	50.9	74.7	72.0	82.4	85.6	62.3	48.1	62.3	45.9	46.3	60.3
	KSRL[42]	53.3	47.4	56.2	79.4	80.7	85.1	89.0	67.4	55.9	61.9	48.5	49.0	64.5
	RRL[45]	42.0	35.7	34.0	67.4	67.8	79.1	80.6	63.9	28.6	48.6	26.5	32.4	50.2
	MCM[27]	**54.4**	48.5	60.6	79.1	77.0	84.0	89.1	61.7	59.3	**64.7**	53.0	**60.5**	66.0
	本文方法	52.2	54.6	**56.5**	80.7	73.0	86.2	**90.9**	**72.0**	**61.3**	59.9	**56.1**	59.2	**66.9***

注：表中 * 表示结果相较于最佳基线方法在配对 t 检验（以所有 AU 的结果为样本）中具有统计显著性差异（$p<0.05$）。

表 2 VoxCeleb→DISFA 上，本方法与全监督自监督方法的对比

	方法/AU	1	2	4	6	9	12	25	26	平均
全监督	SRERL[46]	45.7	47.8	59.6	47.1	45.6	73.5	84.3	43.6	55.9
	SO-Net[18]	33.8	44.5	70.3	**57.6**	39.7	**78.2**	86.7	57.3	58.5
	UGN[47]	43.3	48.1	63.4	49.5	48.2	72.9	90.8	**59.0**	60.0
	ANFL[40]	54.6	47.1	72.9	54.0	**55.7**	76.7	91.1	53.0	63.1
	KS[23]	53.8	59.9	69.2	54.2	50.8	75.8	92.2	46.8	62.8
	MDHR[6]	**65.4**	**60.2**	**75.2**	50.2	52.4	74.3	**93.7**	58.2	**66.2**
自监督	MoCo[26]	22.7	18.2	45.9	45.4	34.1	72.9	83.4	54.5	47.1
	SimSiam[48]	35.5	25.5	58.1	53.8	32.4	74.4	79.0	55.7	51.8
	CVC[41]	30.3	20.9	56.4	49.5	26.3	**75.5**	79.1	51.8	48.6
	CLP[7]	42.4	38.7	63.5	**59.7**	38.9	73.0	85.0	**58.1**	57.4
	PCL[43]	53.8	44.9	58.1	37.2	**53.2**	73.1	86.5	31.3	54.8
	Lu et al.[44]	18.7	27.4	35.1	33.6	20.7	67.5	68.0	43.8	39.4
	TAE[8]	21.4	19.6	**64.5**	46.8	44.0	73.2	85.1	55.3	51.5
	RRL[45]	15.4	15.9	49.5	48.8	22.1	70.3	81.4	46.8	43.8
	PCFRL[28]	54.5	62.1	60.3	36.6	47.4	73.6	86.0	32.6	57.8
	本文方法	**59.5**	**75.7**	51.5	49.9	43.4	67.3	**90.1**	49.2	**60.8***

注：表中 * 表示结果相较于最佳基线方法在配对 t 检验（以所有 AU 的结果为样本）中具有统计显著性差异（$p<0.05$）。

现象主要归因于 MC-SFAU 当前未显式建模不同人脸动作单元之间的语义依赖关系，而这些依赖对面部微动作区分较为细微的 AU 识别尤为关键。值得注意的是，DISFA 数据集中面部表情更加自然，AU 激活频率不均衡，且多数 AU 以低强度形式出现，个体间差异显著，识别难度显著提升。在缺乏标签监督的条件下，MC-SFAU 仍取得了 60.8%的平均 F1 分数，接近全监督方法水平，展现出良好的判别能力与泛化性能，表明其具备处理复杂真实场景的能力。从具体 AU 的表现来看，MC-SFAU 在 AU1、AU2、AU25 上取得了最优结果，尤其在 AU25（嘴唇分离）这类与口部动态相关的动作单元上，相比自监督基线表现出了显著优势，说明方法在建模情感表达区域方面能力突出。此外，在 AU12 也取得接近最优的结果，进一步证明其对口部相关动作的建模能力。需要指出的是，MC-SFAU 在 AU4、AU6、AU26 等 AU 上效果相对不佳。这主要源于这些 AU 样本数量明显不足，或动作幅度较小，导致模型难以学习鲁棒表征。但总体而言，MC-SFAU 在多数 AU 上实现了稳定提升，且整体性能优于所有自监督方法。与典型自监督 FAU 检测方法的比较结果也表明，MC-SFAU 的 FAU 表示在 DISFA 数据集上相较最优自监督方法相对提高了 5.2%，充分验证了 MC-SFAU 架构的可靠性。

4.5 消融性实验与分析

4.5.1 MC-SFAU 的不同模块对 FAU 检测的影响

为系统评估 MC-SFAU 各模块的有效性，我们在 BP4D 数据集上进行了消融实验，结果如表 3 所

示。以仅包含 MFDC 模块的模型为基线,初始 $F1$ 分数为 48.7%。引入 MFRR 模块后,性能显著提至 61.2%,相较基线提升 12.5%,表明该模块在增强微观区域建模方面发挥了关键作用。进一步加入 MCI 模块后,模型性能提升至 66.9%,验证了宏-微观信息协同的有效性。值得注意的是,MFRR 模块单独带来的性能提升远高于 MCI,反映出微观区域建模在 FAU 检测中更为关键。这主要归因于 BP4D 数据集中多数 FAU 标签对应精细的人脸肌肉动作,而 MFRR 通过 Landmark 引导与区域重建,显式对这些区域建模,有效补足了宏观特征的局部表达能力。

表 3　BP4D 数据集上,不同模块对 MC-SFAU 检测结果的影响

MFDC	MFRR	MCI	$F1$(↑)
√			48.7
√	√		61.2
√	√	√	**66.9**

4.5.2　微观人脸区域重建模块(MFRR)中关键参数对 FAU 检测的影响

为了系统评估 MC-SFAU 方法中关键超参数的设置对模型性能的影响,我们对掩码矩形边长 r 和掩码概率 γ 进行了广泛的消融实验。具体而言,为了确定最优的 r 取值,我们在 BP4D 数据集上进行了测试,如表 4 所示,我们分别测试了 $\gamma=0.6$ 的实验条件下 r 在 20、30、40、50 的设置下的实验结果。实验结果表明,当 r 的取值为 40 时候,MC-SFAU 的实验结果取得了 66.9%的最佳的水平。我们猜测,当 r 过小时候,人脸关键特征掩码得不够全面,而当 r 过大时,人脸过于缺失,可能导致模型学习不到有用的人脸信息,导致模型效果较差。通过对 r 的消融研究,我们能够确定最佳的参数设置,从而提高了 MC-SFAU 方法在 AU 检测任务上的性能表现。这对于进一步优化和改进 AU 检测算法具有重要的指导意义,同时也为相关领域的研究者提供了有价值的经验。

表 4　BP4D 数据集上,当 $\gamma=0.6$ 时参数 r 对 MC-SFAU 检测结果的影响

参数 r/AU	1	2	4	6	7	10	12	14	15	17	23	24	平均
$r=20$	48.8	54.9	54.5	82.5	75.3	85.7	89.5	69.6	58.3	62.1	51.1	58.3	65.9
$r=30$	51.6	55.1	48.0	82.3	80.0	87.4	89.8	65.0	60.9	61.1	59.1	56.4	66.4
$r=40$	52.2	54.6	56.5	80.7	73.0	86.2	90.9	72.0	61.3	59.9	56.1	59.2	66.9
$r=50$	50.7	53.2	53.9	80.8	75.0	83.2	88.9	73.2	56.3	57.8	55.2	43.1	64.3

为了找到最佳的 γ 取值,我们在 BP4D 数据集上进行了一系列实验。如表 5 所示,我们在 $r=40$ 的实验条件下测试了不同 γ 取值(1、0.8、0.6、0.4)的实验结果。实验结果表明,在 γ 的取值为 0.6 时,MCSFAU 方法取得了最佳水平,达到了惊人的 66.9%的性能水平,而当 γ 取其他值的时候,其实验结果均呈现下降趋势,这进一步证明了将 γ 设置为 0.6 时是最合理的选择。我们推测当 γ 过小时,例如 γ 取 0.4 时,人脸关键特征掩码得不够全面;而当 γ 过大时,例如 γ 取值为 1 时,掩码部位不够随机,导致模型泛化效果较差,容易出现过拟合现象。因此,在选择 γ 的取值时,需要在人脸特征掩码的全面性和随机性之间取得平衡,以确保模型在不同情况下都能取得良好的泛化性能,这一发现为调整 MC-SFAU 方法的关键参数提供了实质性的指导。最后,我们进一步进行了交叉组合实验(如[$\gamma=0.4$, $r=30$],[$\gamma=0.6$, $r=40$],[$\gamma=0.8$, $r=50$]),如表 6 结果与单参数实验一致,即[$r=40$, $\gamma=0.6$]的组合最优。这表明我们的参数选择具有鲁棒性,并为相关研究提供了有价值的参考。

表 5　BP4D 数据集上,当 $r=40$ 时参数 γ 对 MC-SFAU 检测结果的影响

参数 γ/AU	1	2	4	6	7	10	12	14	15	17	23	24	平均
$\gamma=1$	50.3	51.4	52.5	73.3	71.1	80.6	83.3	66.8	62.4	57.8	57.1	56.6	63.3
$\gamma=0.8$	51.1	52.3	54.4	59.9	70.9	84.6	85.2	71.2	61.8	60.2	58.2	59.4	64.1
$\gamma=0.6$	52.2	54.6	56.5	80.7	73.0	86.2	90.9	72.0	61.3	59.9	56.1	59.2	66.9
$\gamma=0.4$	52.0	53.1	57.2	70.0	73.2	84.2	86.4	73.1	60.7	58.8	55.4	58.3	65.2

表 6　BP4D 数据集上,不同参数组合对 MC-SFAU 检测结果的影响

参数 γ,r/AU	1	2	4	6	7	10	12	14	15	17	23	24	平均
$\gamma=0.4$, $r=30$	52.1	**63.7**	51.3	**82.5**	**80.8**	87.1	88.7	65.6	47.1	57.6	43.8	51.0	64.3
$\gamma=0.6$, $r=40$	**52.2**	54.6	**56.5**	80.7	73.0	86.2	**90.9**	**72.0**	**61.3**	59.9	**56.1**	59.2	**66.9**
$\gamma=0.8$, $r=50$	41.1	57.5	44.3	75.4	73.2	**89.1**	89.5	69.3	58.1	**61.5**	44.5	**61.7**	63.8

此外，为了确定最佳的迭代次数 k 的取值，我们在 BP4D 数据集上对不同迭代次数 k 对 MC-SFAU 检测结果的影响进行了对比。实验结果如图 2 所示，当 $k=50$ 时，MC-SFAU 达到了最佳性能。当迭代次数 k 过小时，模型的训练过程尚未充分进行，导致其无法有效地适应不同场景和面部变化，从而影响检测效果。而当 k 值过大时，准确率轻微下降，未能带来显著的性能提升。这表明，适当的迭代次数对于模型性能的优化至关重要，过少或过多的迭代都可能导致性能瓶颈。

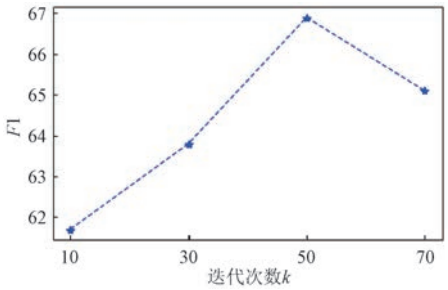

图 2 BP4D 数据集上，不同迭代次数 k 对 MCSFAU 检测结果的影响

4.5.3 MC-SFAU 与不同方法检测效率的对比

为全面评估所提 MC-SFAU 方法在计算成本与检测性能上的综合优势，本文在 BP4D 数据集上选取 SimCLR[24]、MoCo[26]、CVC[41] 和 CLEF[7] 四种代表性方法进行对比，比较指标包括训练参数量、训练速度、推理速度与 $F1$ 分数，结果如表 7 所示。从训练参数量来看，MC-SFAU 为 144.37M，虽然略高于 SimCLR 和 MoCo 以及 CVC，但低于 CLEF 的 156.10M，整体处于适中水平，能够在保证模型容量的同时控制冗余开销。在训练速度方面，受益于模型结构中宏-微观模块的协同学习机制，MCS-FAU 在训练过程中虽具有一定计算开销，但信息建模能力更强，训练时间为 5.1456s，相较 CLEF 的 5.3216s 仍具一定优势。在推理阶段，MC-SFAU 展现出良好的执行效率，单张图像的平均推理时间为 0.0061s，与 MoCo 相当，优于 CVC 和 CLEF。这得益于模型在推理过程中仅保留了预训练阶段获得的 ResNet 主干网络，未引入额外复杂计算，有助于实际部署。在检测性能方面，MC-SFAU 的 $F1$ 分数达到 66.9%，明显优于 SimCLR、MoCo 与 CVC，并略高于 CLEF，展现出优越的表征能力与任务适应性。

表 7 BP4D 数据集上，MC-SFAU 与不同方法计算效率的对比

方法	训练参数量（↓）	训练速度（↓）	推理速度（↓）	$F1$（↑）
SimCLR[24]	**24.62M**	**1.9712s**	0.0064s	49.8
MoCo[26]	65.56M	2.5600s	0.0061s	51.5
CVC[41]	83.96M	3.2768s	0.0073s	54.2
CLEF[7]	156.10M	4.1216s	0.0077s	65.9
本文方法	144.37M	5.1456s	**0.0061s**	**66.9**

4.5.4 MC-SFAU 的不同数据增强方式对 FAU 检测的影响

为进一步分析数据增强策略对模型性能的影响，我们在 BP4D 数据集上对多种增强方式进行了系统对比实验，结果如表 8 所示。具体而言，我们分别评估了单一增强（随机裁剪、随机旋转、颜色抖动）与本文方法的混合增强（将上述增强方式组合，用于生成两幅独立视图）的效果。实验结果表明，单一增强方式能够在一定程度上提升模型性能，其中随机裁剪与随机旋转在捕捉宏观人脸变化时效果较为稳定；然而，单一颜色扰动对模型的贡献有限，过强的扰动甚至可能引入噪声，导致检测性能下降。相比之下，我们的混合增强在 BP4D 数据集的平均 $F1$ 分数以及多数 AU 上均取得了最优结果（66.9%），显著优于任何单一增强方式。这说明混合增强能够在提升模型泛化能力的同时，增强其对宏微观信息的综合建模能力。基于上述分析，本文在所有主实验中均采用混合增强作为默认设置，以兼顾模型的稳定性与鲁棒性。

表 8 BP4D 数据集上，不同数据增强方法对 MC-SFAU 检测结果的影响

方法/AU	1	2	4	6	7	10	12	14	15	17	23	24	平均
随机剪裁	48.5	64.2	50.4	81.8	76.4	88.2	89.9	71.2	32.2	63.5	45.4	58.2	64.1
随机旋转	48.3	58.6	48.6	74.5	76.4	88.1	90.4	70.4	56.9	56.6	55.9	56.1	65.1
颜色抖动	47.8	63.9	33.7	79.8	65.5	87.6	89.5	66.9	53.6	55.1	33.1	49.2	60.5
本文方法	52.2	54.6	56.5	80.7	73.0	86.2	90.9	72.0	61.3	59.9	56.1	59.2	66.9

4.6 可视化实验结果分析

4.6.1 宏观和微观 FAU 表征的注意力可视化

为了验证宏观和微观 FAU 表征在 AU 建模中的作用，我们在 BP4D 数据集测试集上进行了注意力可视化实验，比较了缺少宏-微观变化交互模块（MCI）的模型与完整模型的表现，如图 3 所示。结果显示，缺少 MCI 的模型中，宏观 FAU 表征的注意力集中在人脸整体区域，能够建模整体变化，但未

能充分捕捉细节特征,导致细粒度信息缺失。而微观 FAU 表征的注意力则集中于面部局部区域,如眼睛和嘴唇边缘,表现出对细微变化的高度敏感性。在引入 MCI 模块的完整模型中,宏观 FAU 表征不仅保持对整体变化的关注,还融合了微观表征提供的细节信息,进一步增强了对情感相关关键局部区域的感知能力。同时,借助 MCI 模块中的情感语义对齐约束,微观 FAU 表征相比于缺少 MCI 时更广泛关注与情感表达密切相关的区域,表现出更加合理且丰富的注意力分布。该结果表明,MCI 模块有效促进了宏-微观特征的互补协同,使二者在多层次下共同提升了对 AU 相关情感信息的建模能力与判别能力。

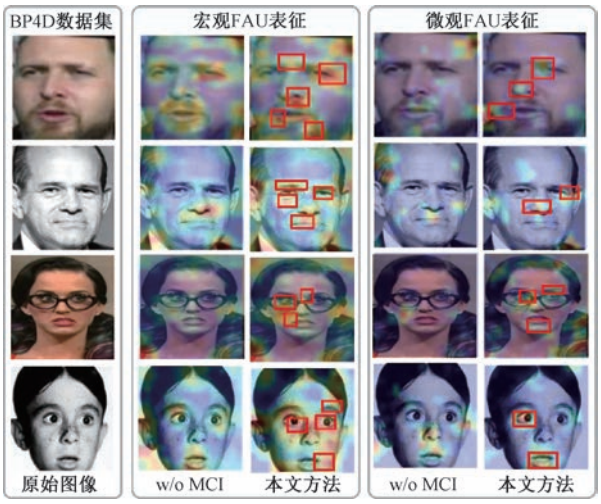

图 3　宏观和微观 FAU 表征注意力可视化结果(颜色映射表示注意力强度:暖色(红/黄)为高注意力,冷色(蓝)为低注意力(数值经归一化至[0,1])。引入 MCI 后,宏观表征在保持全局关注的同时,增强了对情感相关局部区域的感知,微观表征则更全面覆盖情感表达区域。红色框为新增或增强的关注区域,突出 MCI 对宏观-微观人脸变化语义关联建模的帮助。)

4.6.2　不同模块组合的特征分布可视化

为了直观展示 MC-SFAU 的效果,我们将 MFDC 作为对比基线模型,并可视化了基线模型、基线模型加细微微观重建模块(MFRR)的模型(MFDC＋MFRR)以及最终 MC-SFAU 方法的 FAU 特征。我们在 BP4D 数据集的测试集上使用 t-SNE 在 2D 特征空间中可视化自监督提取的特征。由于正负样本不平衡,我们仅选择与正样本数量相同的负样本进行可视化,结果如图 4 所示(注:不同颜色代表不同类别的特征分布,红色点表示当前类别存在,绿色点表示不存在)。相比基线模型,MC-SFAU 方法获得了更具鉴别力的 FAU 表示,特别是在难以区分的类别情况下,我们的方法产生的特征具有更紧密的类内距离。

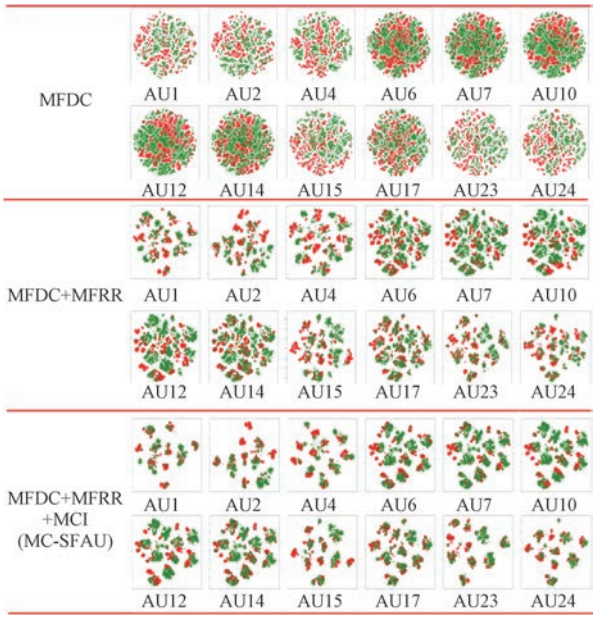

图 4　不同模块组合的特征分布可视化结果

4.6.3　微观人脸区域重建模块(MFRR)中区域掩码和重建的特征可视化

为了深入分析 MC-SFAU 模型中微观人脸区域重建模块(MFRR)在区域掩码和重建任务中的效果,我们通过可视化,在 BP4D 数据集的测试集上分别展示了人脸 Landmark 检测器的检测结果、掩码前提取的完整图像表征、区域掩码后的掩码图像表征以及区域重建后的重建图像表征,如图 5 所示。从图中可以明显看出,我们的模型能够依据人脸 Landmark 检测结果,准确定位并掩码 AU 敏感的关键面部区域得到掩码图像表征。这些区域主要集中在与表情变化密切相关的部位(如眉毛、眼睛、鼻子和嘴部)。

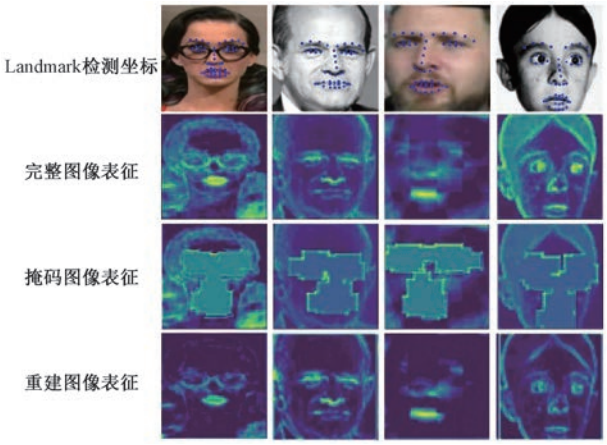

图 5　微观人脸区域重建模块(MFRR)中进行区域掩码和重建的可视化结果

此外，区域重建后的重建图像表征展示了模型对掩码区域的强大重建能力。尽管与掩码前的完整图像表征相比，重建后的特征在掩码区域内略显模糊，但可以看出，模型能够成功重建出被掩码操作遮盖的关键区域，同时保留了与 AU 相关的细粒度信息。这一结果表明，MC-SFAU 模型能够充分利用未掩码区域的上下文信息进行特征推理，从而实现对关键区域的精准重建。整体而言，这些实验结果清晰地验证了 MC-SFAU 模型在区域掩码和重建任务中的有效性。通过区域掩码和重建，我们的模型能够更加聚焦于与 AU 密切相关的关键区域，从而进一步提升了人脸表征的质量和 FAU 检测的性能。

5 总 结

本文提出了一种自监督 FAU 检测方法(MCS-FAU)，成功应对 FAU 检测中标注不足的挑战。MC-SFAU 通过宏观人脸双流对比模块对无标签人脸图像进行对比学习，获得宏观 FAU 特征，实现初步检测。接着，引入微观人脸区域重建模块，聚焦于关键区域，捕捉细致的微观 FAU 表征。结合这两个模块，MC-SFAU 同时获得宏观和微观 FAU 表征。最后，通过宏-微观变化交互模块增强两者，强化宏观和微观人脸运动知识，获得鲁棒的 FAU 表征。未来，MC-SFAU 有望在更广泛的人脸分析任务中展现强大泛化能力，推动领域发展。

参 考 文 献

[1] Ekman P, Friesen W V. Facial action coding system. Environmental Psychology & Nonverbal Behavior, 1978

[2] MacKenzie I S. Human-computer interaction: An empirical research perspective. Amsterdam, The Netherlands: Elsevier, 2024

[3] Reddy S, Allan S, Coghlan S, et al. A governance model for the application of ai in health care. Journal of the American Medical Informatics Association, 2020, 27(3): 491-497

[4] Li W, Abtahi F, Zhu Z, et al. Eac-net: A region-based deep enhancing and cropping approach for facial action unit detection//Proceedings of the 12th IEEE International Conference on Automatic Face & Gesture Recognition. Washington, USA, 2017: 103-110

[5] Li W, Abtahi F, Zhu Z. Action unit detection with region adaptation, multi-labeling learning and optimal temporal fusing//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 1841-1850

[6] Wang Z, Song S, Luo C, et al. Multi-scale dynamic and hierarchical relationship modeling for facial action units recognition//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2024: 1270-1280

[7] Li Y, Shan S. Contrastive learning of person-independent representations for facial action unit detection. IEEE Transactions on Image Processing, 2023, 32: 3212-3225

[8] Li Y, Zeng J, Shan S. Learning representations for facial actions from unlabeled videos. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 44(1): 302-317

[9] Yang H, Yin L, Zhou Y, et al. Exploiting semantic embedding and visual feature for facial action unit detection//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual, 2021: 10482-10491

[10] Ma C, Chen L, Yong J. Au r-cnn: Encoding expert prior knowledge into r-cnn for action unit detection. Neurocomputing, 2019, 355: 35-47

[11] Shao Z, Liu Z, Cai J, et al. Facial action unit detection using attention and relation learning. IEEE Transactions on Affective Computing, 2019, 13 (3): 1274-1289

[12] Jacob G M, Stenger B. Facial action unit detection with transformers//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual, 2021: 7680-7689

[13] Zhang Z, Zhai S, Yin L, et al. Identity-based adversarial training of deep cnns for facial action unit recognition// Proceedings of the 29th British Machine Vision Conference. Newcastle, Australia, 2018: 226

[14] Almaev T, Martinez B, Valstar M. Learning to transfer: transferring latent task structures and its application to person-specific facial action unit detection//Proceedings of the IEEE International Conference on Computer Vision. Santiago, Chile, 2015: 3774-3782

[15] Tu C H, Yang C Y, Hsu J Y. Idennet: Identity-aware facial action unit detection//2019 14th IEEE International Conference on Automatic Face & Gesture Recognition. Lille, France, 2019: 1-8

[16] Yin Y, Chang D, Song G, et al. Fg-net: Facial action unit detection with generalizable pyramidal features//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA, 2024: 6087-6096

[17] Cakir D, Yilmaz G, Arica N. Boosting facial action unit detection with cgan-based data augmentation. Decision Making in Healthcare Systems, 2024, 513: 323-335

[18] Yang H, Wang T, Yin L. Set operation aided network for action units detection//Proceedings of the 15th IEEE International Conference on Automatic Face and Gesture Recognition. Virtual, 2020: 229-235

[19] Shao Z, Cai J, Cham T J, et al. Unconstrained facial action

unit detection via latent feature domain. IEEE Transactions on Affective Computing, 2021, 13(2): 1111-1126

[20] Lee H Y, Tseng H Y, Huang J B, et al. Diverse image-to-image translation via disentangled representations//Proceedings of the European Conference on Computer Vision. Munich, Germany, 2018: 35-51

[21] Lv T, Wen Y, Sun Z, et al. A continuous emotional editing model for talking head videos based on decoupling texture and geometry. Sci Sin Inform, 2023, 53(12): 2423-2439

[22] Csurka G. Domain adaptation for visual applications: A comprehensive survey. arXiv preprint arXiv:1702.05374, 2017

[23] Li X, Zhang X, Wang T, et al. Knowledge-spreader: Learning semi-supervised facial action dynamics by consistifying knowledge granularity//Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France, 2023: 20922-20932

[24] Chen T, Kornblith S, Norouzi M, et al. A simple framework for contrastive learning of visual representations//Proceedings of the 37th International Conference on Machine Learning. Virtual, 2020: 1597-1607

[25] Shang Z, Liu B, Lv F, et al. Learning contrastive feature representations for facial action unit detection. Pattern Recognition, 2026, 173: 112746

[26] He K, Fan H, Wu Y, et al. Momentum contrast for unsupervised visual representation learning//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual, 2020: 9729-9738

[27] Zhang X, Yang H, Wang T, et al. Multimodal channel-mixing: Channel and spatial masked autoencoder on facial action unit detection//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA, 2024: 6065-6074

[28] Liu Y, Feng S, Liu S, et al. Sample-cohesive pose-aware contrastive facial representation learning. International Journal of Computer Vision, 2025, 133(6): 3727-3745

[29] Grill J B, Strub F, Altché F, et al. Bootstrap your own latent-a new approach to self-supervised learning. Advances in Neural Information Processing Systems, 2020, 33: 21271-21284

[30] Tarvainen A, Valpola H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results//Proceedings of the 31st Conference on Neural Information Processing Systems. Long Beach, USA, 2017: 1195-1204

[31] Cui Z, Kuang C, Gao T, et al. Biomechanics-guided facial action unit detection through force modeling//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023: 8694-8703

[32] Bulat A, Tzimiropoulos G. How far are we from solving the 2d & 3d face alignment problem? //Proceedings of the IEEE/CVF International Conference on Computer Vision. Venice, Italy, 2017: 1021-1030

[33] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA, 2017: 6000-6010

[34] Barron J T. A general and adaptive robust loss function//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 4326-4334

[35] Van Erven T, Harremos P. Rényi divergence and kullback-leibler divergence. IEEE Transactions on Information Theory, 2014, 60(7): 3797-3820

[36] Nagrani A, Chung J S, Zisserman A. Voxceleb: A large-scale speaker identification dataset. arXiv preprint arXiv: 1706.08612, 2017

[37] Zhang X, Yin L, Cohn J F, et al. Bp4dspontaneous: A high-resolution spontaneous 3d dynamic facial expression database. Image and Vision Computing, 2014, 32(10): 692-706

[38] Mavadati S M, Mahoor M H, Bartlett K, et al. Disfa: A spontaneous facial action intensity database. IEEE Transactions on Affective Computing, 2013, 4(2): 151-160

[39] Song T, Cui Z, Zheng W, et al. Hybrid message passing with performance-driven structures for facial action unit detection//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual, 2021: 6267-6276

[40] Luo C, Song S, Xie W, et al. Learning multidimensional edge feature-based au relation graph for facial action unit recognition//Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence. Vienna, Austria, 2022: 1239-1246

[41] Wu H, Wang X. Contrastive learning of image representations with cross-video cycle-consistency//Proceedings of the IEEE/CVF International Conference on Computer Vision. Virtual, 2021: 10149-10159

[42] Chang Y, Wang S. Knowledge-driven selfsupervised representation learning for facial action unit recognition//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 20385-20394

[43] Liu Y, Wang W, Zhan Y, et al. Pose-disentangled contrastive learning for self-supervised facial representation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023: 9717-9728

[44] Lu L, Tavabi L, Soleymani M. Self-supervised learning for facial action unit recognition through temporal consistency//Proceedings of the 31th British Machine Vision Conference. Virtual, 2020

[45] Song J, Liu Z. Self-supervised facial action unit detection with region and relation learning//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Rhodes Island, Greece, 2023: 1-5

[46] Li G, Zhu X, Zeng Y, et al. Semantic relationships guided representation learning for facial action unit recognition// Proceedings of the AAAI Conference on Artificial Intelligence. Honolulu, USA, 2019: 8594-8601

[47] Song T, Chen L, Zheng W, et al. Uncertain graph neural networks for facial action unit detection//Proceedings of the AAAI Conference on Artificial Intelligence. Virtual, 2021: 5993-6001

[48] Chen X, He K. Exploring simple siamese representation learning//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual, 2021: 15750-15758

LIU Ke-Jun, Ph. D. candidate. Her main research interests include affective computing, machine learning, and computer vision.

LIU Yuan-Yuan, Ph. D., associate professor. Her main research interests include affective computing, machine learning and pattern recognition, and computer vision.

LI Kang-Lin, M. S. His main research interests include affective computing, machine learning, and computer vision.

TANG Chang, Ph. D., associate professor. His main research interests include affective computing, machine learning and pattern recognition, and computer vision.

QIN Jie, M. S. His main research interests include affective computing, machine learning and pattern recognition, and computer vision.

LUO Wei, Ph. D., senior researcher. His main research interests include machine learning and pattern recognition, and computer vision.

Background

Facial Action Unit (FAU) detection aims to recognize fine-grained facial muscle activations that constitute facial expressions, and it serves as a fundamental component for downstream affective computing applications such as emotion recognition and human-computer interaction. Although modern FAU detectors have achieved strong performance under fully supervised learning, their effectiveness typically hinges on large-scale, high-quality AU annotations. Such annotations are expensive and time-consuming to acquire, and models trained on limited labeled data often suffer from degraded robustness when deployed in unconstrained real-world conditions (e.g., large variations in identity, pose, illumination, and background).

To alleviate the reliance on manual labels, self-supervised FAU representation learning has become an active research direction, leveraging abundant unlabeled facial images and videos for pre-training. Existing self-supervised approaches largely fall into two paradigms: contrastive learning and reconstruction learning. Contrastive methods typically improve representation discriminability by enforcing invariance across augmented views, yet they may over-emphasize global facial appearance consistency and miss the subtle, localized motions that are essential for AU understanding. Conversely, reconstruction-based methods can encourage attention to fine-grained regions, but they often under-model the global (macro) facial dynamics that provide stable contextual cues for AU-related changes. As a result, the imbalance between macro dynamics and micro muscle-related motions becomes a key bottleneck for learning AU-sensitive representations in a self-supervised manner.

This paper addresses the above gap by proposing Macro-Micro Changes based Self-supervised Facial Action Unit Detection (MC-SFAU), a unified framework that jointly captures complementary macro-level and micro-level facial motion cues. Specifically, MC-SFAU integrates (i) a macro facial dual-stream contrast module that learns global facial change representations via a global stream and a guide stream, (ii) a micro facial region reconstruction module that introduces region-wise reconstruction supervision to better preserve subtle local movements, and (iii) a macro - micro change interaction module that strengthens knowledge exchange between the two levels to obtain more accurate and robust FAU representations. Empirically, the model is pre-trained on large-scale unlabeled facial data (VoxCeleb) and then evaluated on standard FAU benchmarks (BP4D and DISFA), where it yields consistent improvements over prior self-supervised baselines.

The significance of this study lies in providing theoretical and technical support for scalable FAU modeling in real-world, label-scarce scenarios. By enabling self-supervised learning of robust AU representations from large-scale unlabeled data while explicitly coordinating macro and micro facial changes, the proposed design reduces heavy reliance on expensive AU annotations and improves robustness for practical deployment in affective computing systems.

The research group has previously made sustained progress in affective computing and facial behavior analysis, including robust facial representation learning, weakly/self-supervised learning for limited-label settings, and fine-grained facial motion modeling, which provides a solid foundation for this study.