中图法分类号:TB18　　**文献标识码**: A　　**文章编号**: 1006-8961(XXXX)XX-0001-15

论文引用格式:Huang Hailin, Wang Jiajun. Spatiotemporal fusion network for facial action unit detection[J/OL]. Journal of Image and Graphics, XXXX:1-15. DOI: 10.11834/jig.250503. (黄海琳, 王加俊. 面部动作单元检测的时空融合网络[J/OL]. 中国图象图形学报, XXXX:1-15. DOI: 10.11834/jig.250503.)[DOI:10.11834/jig.250503]

面部动作单元检测的时空融合网络

黄海琳,王加俊

苏州大学电子信息学院,江苏省苏州市　215001

摘　要:**目的**　面部动作单元(AU)检测是情感计算与计算机视觉中的重要研究问题,现有方法往往在空间关系建模和时间动态建模方面存在不足,导致检测的准确率与鲁棒性受限。为此,提出一种时空融合的统一AU检测框架,以同时捕捉AU间空间依赖与跨帧时间演化。**方法**　本文在利用ResNet-18进行特征提取的基础上,设计了空间关系建模(SRM)模块与时间关系建模(TRM)模块。SRM通过图神经网络显式建模帧内AU的协同激活与对抗模式,TRM结合帧间差分与图建模以捕捉动态变化。进一步提出时空特征融合(SFF)策略,自适应平衡空间与时间特征的重要性。最终利用基于余弦相似度的分类模块完成AU识别,并采用加权交叉熵损失解决类别不平衡问题。**结果**　在BP4D和DISFA两个公开数据集上的实验表明,所提方法在保持低计算与存储开销的同时,取得了先进水平的性能。在BP4D数据集上,平均F1-score达到66.00%,优于最新方法;在DISFA数据集上,平均F1-score为65.34%,接近最优结果。消融实验验证了SRM、TRM和SFF三者协同作用的重要性,不同AU的检测结果也表明该方法在动态AU识别方面优势显著。**结论**　本文提出的时空融合AU检测框架,能够有效整合空间与时间两方面信息,提升检测的准确率和鲁棒性,同时具备轻量化的计算优势。该研究为复杂动态场景下的面部动作单元检测提供了一种高效的解决方案,对情绪识别及人机交互等应用具有积极意义。

关键词:面部动作单元;AU检测;时空建模;图神经网络;特征融合

Spatiotemporal fusion network for facial action unit detection

Huang Hailin, Wang Jiajun

School of Electronic and Information Engineering, Soochow University, Suzhou, Jiangsu 215001, China

Abstract: **Objective** Facial Action Unit (AU) detection is a fundamental yet challenging task in affective computing and computer vision. AU, defined in the Facial Action Coding System (FACS), corresponds to the subtle activation of specific facial muscles, providing objective measurements that reveal human emotions, cognitive states, and mental health conditions. Automatic AU detection plays a critical role in a wide array of applications, including advanced emotion recognition systems, non-intrusive mental health monitoring, and natural human-computer interaction. However, despite significant progress, existing methods face two primary limitations. First, spatial relationship modeling often relies on fixed, predefined priors or learns only shallow feature correlations. These approaches are insufficient to capture the complex, nonlinear co-activation and antagonistic patterns inherent in facial expressions, such as the synergistic relationship between AUs during a smile or the mutually exclusive nature of certain brow movements. Second, temporal dynamic modeling typically depends on computationally intensive techniques like optical flow or manually crafted constraints. These methods are not only prone to noise and tracking errors but are also difficult to scale for real-time applications due to their high computa-

收稿日期:2025-10-15;**修回日期**:2026-03-19

* **通信作者**:王加俊,通信作者,男,教授,主要研究方向为医学图像处理。E-mail:jjwang@suda.edu.cn

tional cost. To overcome these issues, we propose a unified and lightweight spatiotemporal fusion framework for AU detection that jointly captures intra-frame spatial dependencies and inter-frame temporal dynamics in an efficient and scalable manner. **Method** Our framework employs a ResNet-18 architecture as the backbone to extract robust per-frame facial feature representations. Building upon these features, the model integrates two core, lightweight modules: Spatial Relationship Modeling (SRM) and Temporal Relationship Modeling (TRM). The SRM module leverages Graph Neural Networks (GNN) to explicitly encode intra-frame AU structures. Instead of relying on static priors, it dynamically computes pairwise AU feature similarities via cosine similarity. A Top-K pruning mechanism is then applied to the resulting graph, retaining only the strongest connections to suppress spurious correlations and ensure a compact, interpretable, and efficient AU graph structure. This dynamic graph construction requires no additional anatomical labels. The TRM module focuses on temporal dynamics by first generating motion-sensitive features through simple frame-difference operations. This step effectively reduces redundant texture information, highlighting regions of change. A temporal GNN is then applied to these motion features to propagate information across a sequence of frames. This allows the model to distinguish transient noise from genuine AU activations that evolve smoothly over time. To integrate the complementary strengths of spatial context and temporal motion, we propose a Spatiotemporal Feature Fusion (SFF) strategy. This module concatenates the features from the SRM and TRM branches and employs an adaptive attention mechanism to dynamically balance their contributions on a per-instance basis. Finally, AU recognition is performed using a cosine-similarity-based classification module, which is robust to feature scale. A weighted cross-entropy loss function is employed during training to effectively alleviate the pervasive class imbalance problem by emphasizing the learning of infrequent AUs. **Result** Extensive experiments were conducted to validate the proposed framework on two benchmark datasets, BP4D and DISFA. The results demonstrate that our approach achieves state-of-the-art or highly competitive performance while maintaining low computational and memory costs. On the BP4D dataset, the model reached an average F1-score of 66.00%, outperforming recent methods such as RE-Net and AC2D. On the DISFA dataset, it achieved an average F1-score of 65.34%, closely approaching the best-reported result. Comprehensive ablation studies confirmed the necessity and effectiveness of each module: removing the SRM module significantly impaired the detection of weak or infrequent AU, removing the TRM module degraded performance on dynamic AU recognition, and removing the SFF fusion strategy reduced the model's adaptability and overall feature integration efficiency. The synergy of SRM, TRM, and SFF was essential for achieving superior performance. Notably, the framework excelled in detecting dynamic AU such as AU6 (cheek raiser) and AU7 (lid tightener), demonstrating its particular strength in modeling temporal consistency and motion patterns. Statistical tests further verify the significance of our improvements over baselines. These findings highlight that the model is particularly robust in handling varying illumination conditions and subtle expression changes, showcasing its strong generalization capability for unconstrained environments. **Conclusion** In conclusion, the proposed spatiotemporal fusion framework effectively integrates spatial and temporal modeling to enhance AU detection accuracy and robustness, while maintaining a lightweight and efficient design suitable for real-world applications. It provides an effective solution for AU detection in complex, dynamic environments and shows strong potential for deployment in affective computing and human-computer interaction. The model's ability to dynamically balance spatial and temporal features based on the input context is a key contribution. Future work will explore fine-grained local feature modeling around specific AU, multi-scale temporal fusion strategies, and the integration of multimodal data (e.g., audio and text) to further improve generalization and performance in unconstrained real-world scenarios. We also aim to validate the framework across diverse demographic groups and clinical populations.

Key words: facial action units; AU detection; spatiotemporal modeling; graph neural networks; feature fusion

0 引 言

面部动作单元(Facial Action Units, AU)是指面部肌肉在情绪表达过程中所表现出的不同活动模式(Liu等,2024)。在20世纪70年代,为了更客观、更准确地描述人类面部表情,Ekman等人提出了面部动作编码系统(Facial Action Coding System, FACS)(Ekman等,1978),该系统基于面部肌肉解剖结构定义了44种面部动作单元,并为每种动作单元设置了

五个强度等级。然而，手工标注AU是一项繁琐且耗时的任务。成为一名专业的AU标注员需要至少100小时的培训，而仅仅标注一分钟的视频就至少需要耗时2小时(Yacoob等，2002)。因此，通过机器学习实现AU的自动检测逐渐成为研究热点。尤其是基于深度学习技术的AU检测方法得到了广泛发展，例如BLSTM(Jaiswal等，2016)、EAC-Net(Li等，2017)以及JAA-Net(Shao等，2018)。

面部AU可以通过静态图像或视频进行检测。在从静态面部图像中检测AU时，通常会采用跨AU关系建模方法。跨AU关系建模，是通过挖掘AU之间的相关性，可以提取潜在信息，以捕捉微小且难以检测的AU运动。例如，Zhao等人(2016)提出的DRML方法通过建立AU标签与面部区域之间的联系来捕获AU之间的关系；Yang等人(2019)提出的RE-Net通过域自适应的方式在参数空间中建模AU相关性，以提升AU检测效果。除此之外，Li等人(2020)使用了结构化关系图与门控图序列神经网络相结合的方式来生成增强的AU特征。同样的，Tian等人(2022)利用面部表情的弱标签来建立AU之间的相关性，同时采用主网络和辅助网络相结合的方式来融合深层与浅层特征，以提升AU检测性能。除了通过AU相关性，还有研究人员提出了自学习模块，用于在无需辅助信息的情况下精确定位与肌肉运动相关的面部区域。例如AUGAIN(Li等，2025)通过ROI分割模块和注意力挖掘模块生成目标AU的注意力图，准确定位AU的空间位置。然而，该方法可能会引入冗余的特征表示。为了减少冗余，AC2D(Shao等，2025)使用因果干预模块来消除训练样本中的偏差和无关的AU干扰。

由于静态AU检测方法无法充分捕捉面部肌肉运动的动态信息，许多研究团队致力于开发基于动态面部图像的AU检测方法。这些方法(Zhou等，2024)依赖于AU运动的时间连续性，其中被广泛使用的是光流法。例如，JAO(Shao等，2025)利用光流来捕捉与AU相关的运动动态，LTI(Yang等，2019)通过在单张图像中引入隐式光流层，由光流网络估计面部外观变化并与AU网络联合训练。尽管光流能够有效表示运动变化，但其计算开销极大。在实时应用中处理高分辨率视频序列时，这种计算负担尤为突出，通常需要专用硬件加速以维持实际可行的处理速度。为了解决这一局限性，研究人员提出了一些时间建模技术，利用神经网络自动学习时空特征表示。例如Yang等人(2023)提出的AUNet则利用输入视频中相邻帧的变化所产生的时间约束来优化模型。通过同时融合空间与时间信息，这些模型能够更全面地捕捉动态面部动作单元之间的复杂相互关系。这些方法不仅提高了AU识别的准确率，还增强了模型区分相似但时间上存在差异的面部表情的能力，例如自发笑与刻意笑的区别。

尽管面部单元的检测技术已经取得了长足的进展，但现有方法仍存在不足：(1) AU关系依赖固定先验知识，泛化性和稳定性不足；(2) 时间建模容易受到冗余图像信息干扰。针对这些问题，本文提出了一种时空图神经网络，通过空间图神经网络(graph neural networks，GNN)建模AU之间的动态相关性，结合差分机制的时间GNN去除冗余信息，从而实现端到端的时空特征学习。本文的贡献点如下：

1) 提出一种时间建模方法，利用帧间差分操作分解AU运动，捕捉连续帧间的动态变化，强调运动敏感特征。

2) 提出一种空间建模方法，利用AU间差分操作分解AU关系，显式建模同一帧内不同AU的相关性提升对局部肌肉交互的理解。

3) 提出一种时空自适应融合策略，能够根据任务需求自适应平衡空间与时间特征的贡献，形成统一的表示。

1 方 法

1.1 预处理

动态的视频序列天然包含了更丰富的上下文信息，但同样的也会参杂更多的冗余信息(李等，2020)。人脸图像的预处理可以有效减少无关信息

对人脸AU检测的干扰。如图1所示，预处理流程主要包括三个关键步骤：人脸裁剪、关键点检测和图像对齐。

1) 人脸裁剪：首先使用dlib工具包(Lugaresi等，2019)在视频帧中检测人脸，获取一个包围人脸的正方形框。然后将人脸裁剪成256×256像素的图像。

2) 关键点检测：在训练阶段，使用Mediapipe工具包(Lugaresi等，2019)检测人脸关键点，用于图像

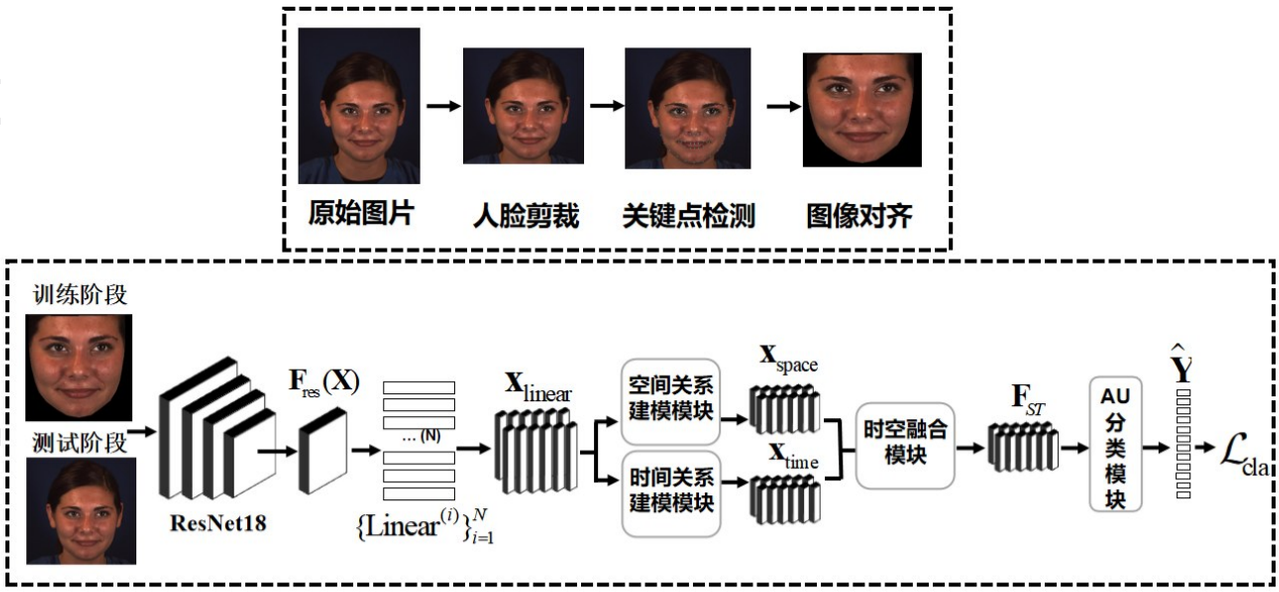

图1　整体网络结构

Fig. 1　The overall network architecture

对齐(而不是感兴趣区域的提取)。

3) 图像对齐:通过仿射变换调整视频片段中的人脸,使每一帧中双眼之间的连线处于水平方向。

除了上述三个预处理步骤外,还对输入的人脸图像进行了数据归一化处理,针对不同的颜色通道,消除了不同样本间像素值分布的差异。需要指出的是,关键点检测和图像对齐步骤仅在训练阶段需要,而在测试阶段仅保留必要的人脸裁剪操作。

1.2　模型概述

面部表情识别是计算机视觉领域中的重要任务之一(崔鑫宇 等,2024),本文提出了一个统一的面部动作单元检测架构,能够协同集成空间和时间关系学习。如图1所示,该框架由三个部分组成:时间关系建模(temporal relation modeling,TRM)模块、空间关系建模(spatial relation modeling,SRM)模块、时空特征融合(spatio-temporal feature fusion,SFF)模块以及AU分类模块(AU classification module, ACM)。

输入图像首先通过ResNet-18主干网络进行特征提取。为了保证对AU的感知能力,ResNet-18提取的特征通过专门的线性变换投影到相应于每个AU的特定特征空间,从而将共享的视觉表示显式解耦成对AU信息敏感的特征。这些相应于特定AU的特征随后依次输入TRM和SRM模块,分别进行时间和空间推理。一方面,TRM模块利用时间GNN组构建基于图神经网络的语义传播网络,通过帧间差分操作显式建模时间依赖关系。特别地,时间建模突出AU在连续帧中的动态激活,使网络能够捕捉细微的运动变化和长期的时间一致性,这对区分细粒度的面部表情至关重要。与此同时,SRM模块利用空间GNN组构建另一套基于图神经网络的语义传播网络,通过帧内AU之间的差异操作分解AU的相关性。该空间建模框架鼓励网络捕捉AU之间的共现和对抗关系,从而丰富语义表示并反映底层的面部肌肉交互。随后,SRM模块通过时空自适应融合策略对得到的空间和时间特征进行融合。该策略根据AU识别的具体需求,动态平衡空间和时间模态的重要性,从而生成一个统一的表示,该表示同时保留了局部的AU结构依赖和AU激活的时间演化模式。最后,来自SRM模块的时空特征用于基于余弦相似度的AU分类,这种方式不仅提供了具有高区分度的判别空间,还增强了类间可分性和类内紧凑性

1.3　空间关系建模模块

面部AU很少单独出现。相反,它们由共享的肌肉驱动,因此表现出明显的、帧特定的协同激活和对抗模式。仅依靠时间推理不足以消除视觉上相似的AU之间的歧义,恢复在遮挡或姿态变化下的弱信号,或抑制虚假检测。因此,本文引入了一个空间关系建模模块,该模块通过数据驱动的图显式编码帧内AU结构。具体而言,给定来自ResNet-18的特

征图 $\mathbf{F}_{\text{res}}(\mathbf{X})$，该模型将 $\mathbf{F}_{\text{res}}(\mathbf{X})$ 投影到相应于每个AU的特定特征空间，以生成每个AU的局部语义表示。通过一组独立的全连接层 $\{\text{Linear}^{(i)}\}_{i=1}^{N}$，将特征映射到特定的子空间：$\mathbf{F}_{\text{AU}}^{(i)}(\mathbf{X}) = \text{Linear}^{(i)}(\mathbf{F}_{\text{res}}(\mathbf{X})) \in \mathbb{R}^{T\times d}, \quad i = 1, 2, \cdots, N\mathbf{F}_{\text{AU}}(\mathbf{X})$，其中 N 为AU类别数。每个线性层执行特征学习以捕捉相应AU的判别模式，所有学习到的AU特征存储在结构化特征集 $\mathbf{x}_{\text{linear}} \in \mathbb{R}^{T\times N\times d}$ 中。这里，T 表示帧的数量。

如图2所示，输入 $\mathbf{x}_{\text{linear}}$ 首先沿时间轴展开为相应于 T 帧的 T 个AU特征集 $\{\mathbf{x}_{\text{frame}}^{t}\}_{t=1}^{T}$，其中每个 $\mathbf{x}_{\text{frame}}^{t} \in \mathbb{R}^{N\times d}$ 包含第 t 帧的 N 个AU嵌入。在具体操作中，一个帧张量作为 $x \in \mathbb{R}^{N\times d}$ 输入到GNN中。对于每一帧，本文使用余弦相似度度量动态构建AU之间的图

$$\mathbf{A}_t = \left(\text{Norm}(\mathbf{x}_{\text{frame}}^{i})\right)^{\top}\text{Norm}(\mathbf{x}_{\text{frame}}^{j}) \tag{1}$$

其中，$\mathbf{x}_{\text{frame}}^{t}$ 是节点特征矩阵，$\text{Norm}(\cdot)$ 是 L_2 归一化函数。在图构建之前分离特征，可以通过阻止梯度流向动态邻域选择操作，避免训练过程中过度干扰邻域结构，从而稳定训练。因此本文对于每一帧 t，通过阈值化 $\mathbf{A}_t$ 的对应行将其修剪到Top-3邻域，从

而得到一个稀疏的邻接矩阵 $\mathbf{A}_t' \in \{0, 1\}^{N\times N}$，其中仅保留三个最相似的邻居。这种Top-K（Fagin等，2003）修剪将图从 $O(N^2)$ 条边减少到 $O(NK)$，抑制了虚假相关性，并生成紧凑、可解释的邻域。在得到的二进制 $\mathbf{A}_t'$ 后，通过归一化图处理获得用于消息传递的归一化邻接矩阵 $\mathbf{A}_t''$。给定归一化图 $\mathbf{A}_t''$ 和当前特征 $\mathbf{x}_{\text{frame}}^{t}$，GNN的输出 $\mathbf{x}_{gnn}^{i}$ 可按如下公式计算：

$$\mathbf{x}_{\text{gnn}}^{t} = \text{ReLU}\left(\text{BN}\left(V(\mathbf{x}_{\text{frame}}^{t}) + \mathbf{A}_t'' \cdot U(\mathbf{x}_{\text{frame}}^{t})\right)\right) \tag{2}$$

其中，BN表示批量归一化（batch normalization）操作，ReLU是线性修正单元（rectified linear unit）激活函数。U 和 V 是可学习参数。V 表示对每个节点邻接特征的变换，提供了它们的变换表示。U 表示对每个节点自身特征的变换，提供一个基础表示，与聚合的邻居特征结合以更新节点特征。。这种帧内图消息传递机制鼓励语义和解剖学相关的AU交换信息，从而通过其强邻居对弱但一致的线索进行放大，对帧特定的协同激活模式进行建模。经过 T 个GNN后，T 个特征 $\mathbf{x}_{\text{gnn}}^{t}$ 直接拼接为 $\mathbf{x}_{\text{space}} \in \mathbb{R}^{T\times N\times d}$。图神经网络模拟AU之间的协同激活模式，最终的 $\mathbf{x}_{\text{space}}$ 通过基于图的消息传递增强了解剖学相关AU的特征表示。

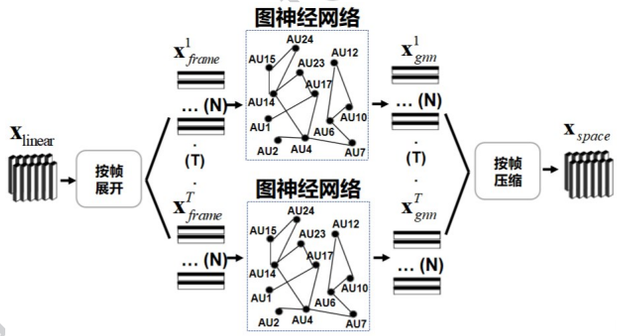

图2　空间关系建模模块的网络结构

Fig. 2　The framework of the Spatial Relationship Modeling module

1.4　时间关系建模模块

虽然SRM有效地捕捉了AU之间的帧内空间依赖关系，但它忽略了跨帧交互，因此缺乏时间感知。为了补充SRM，而不是冗余地重新学习AU特定的空间语义，本文引入了一个时间关系建模模块。为了提高视频中动作识别的准确度，提出基于动作切分（罗会兰 等，2017）的思想，从而专注于AU无关的帧间关系以提供缺失的时间上下文。如图3所示，采用解耦的双路径架构分别对AU的时间动态关系进行建模。

在TRM模块中，输入 $\mathbf{x}_{\text{linear}}$ 被展开为 N 个AU的特征 $\mathbf{x}_{\text{AU}}^{i} \in \mathbb{R}^{T\times d}, (i = 1, 2, \cdots, N)$。每个 $\mathbf{x}_{\text{AU}}^{i}$ 随后进行差分操作以获得 $\mathbf{x}_{\text{diff}}^{i} \in \mathbb{R}^{(T-1)\times d}$。差分操作突出瞬时运动模式，并减少与面部纹理特征相关的冗余信息。在获得 $\mathbf{x}_{\text{diff}}^{i}, (i = 1, 2, \cdots, N)$ 后，这些特征被输入到 N 个独立图神经网络中。

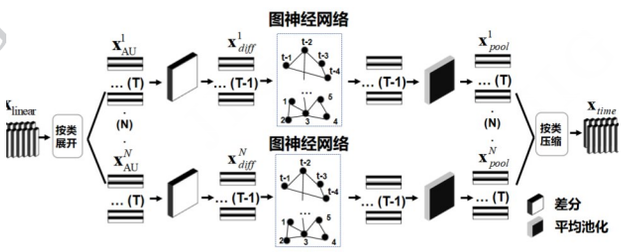

图3　时间关系建模模块的网络结构

Fig. 3　The framework of the Temporal Relationship Modeling module

本文基于帧间余弦相似度构建每个GNN $_i$ 的邻接矩阵 $\mathbf{A}_i \in \mathbb{R}^{(T-1)\times(T-1)}$，余弦相似度通过测量两个个体向量之间的内积空间夹角余弦值来度量它们之

间的相似性(陈大力 等,2014):

$$\mathrm{sim}(\mathbf{x}^p,\mathbf{x}^q)=\frac{\mathbf{x}^p\cdot\mathbf{x}^q}{\|\mathbf{x}^p\|_2\cdot\|\mathbf{x}^q\|_2},\quad \forall p,q\in[1,T-1] \tag{3}$$

其中,$\mathbf{x}^p$和$\mathbf{x}^q$是从$\mathbf{x}^i_{\mathrm{diff}}$中提取的两个帧向量,余弦相似度用于度量它们的邻接关系。通过计算$\mathbf{x}^i_{\mathrm{diff}}$中所有帧向量之间的余弦相似度,可以到邻接矩阵$\mathbf{A}_i$。一旦确定了帧间邻接矩阵$\mathbf{A}_i$,便选择最相似的Top-3邻接帧进行修剪,得到稀疏邻接矩阵$\mathbf{A}'_i$。利用稀疏邻接矩阵$\mathbf{A}'_i$,随后根据公式(2)的传播规则,经GNN传播获得更新特征后将其输入到一个平均池化层,以恢复其维度为$\mathbf{x}^i_{\mathrm{pool}}\in\mathbb{R}^{T\times d}$。平均池化来聚合局部特征(董宏翔 等,2024)的信息,从而提供更丰富的上下文信息,帮助恢复在差分操作中丢失的一些绝对值信息。

将所有N个AU的$\mathbf{x}^i_{\mathrm{AU}}$,$(i=1,2,\cdots,N)$进行上述转换得到$\mathbf{x}^i_{\mathrm{pool}}$。然后,将这些特征沿通道维度拼接为$\mathbf{x}_{\mathrm{time}}\in\mathbb{R}^{T\times N\times d}$。最终得到的动态时间特征$\mathbf{x}_{\mathrm{time}}$捕捉了AU的长期运动趋势。

1.5 时空融合模块

为了以一种简单而高效的方式融合空间和时间信息,输入特征$\mathbf{x}_{\mathrm{space}}$和$\mathbf{x}_{\mathrm{time}}$首先在空间维度上进行拼接,然后输入到一个共享的全连接层和激活函数中,如公式(4)所示:

$$\mathbf{x}_{\mathrm{ST}}=\mathrm{ReLU}\left(\mathrm{Linear}\left([\mathbf{x}_{\mathrm{space}},\mathbf{x}_{\mathrm{time}}]\right)\right) \tag{4}$$

其中,$[\cdot,\cdot]$表示在数据维度上的拼接操作,Linear表示全连接层操作。$\mathbf{x}_{\mathrm{ST}}\in\mathbb{R}^{T\times N\times(2d/r)}$是编码融合信息的特征图,其中$r$为缩减比率,用于降低模型复杂度。随后,$\mathbf{x}_{\mathrm{ST}}$被分割为两个独立张量$\mathbf{x}_{\mathrm{S}}\in\mathbb{R}^{T\times N\times(d/r)}$和$\mathbf{x}_{\mathrm{T}}\in\mathbb{R}^{T\times N\times(d/r)}$。这两个张量分别对应空间特征和时间特征。通过另一个全连接层,这两个张量被转换为与输入张量在数据维度上相同长度的张量。然后,这两个结果张量分别经过批归一化和sigmoid激活,得到动态融合权重$\mathbf{W}_{\mathrm{space}}\in\mathbb{R}^{T\times N\times d}$和$\mathbf{W}_{\mathrm{time}}\in\mathbb{R}^{T\times N\times d}$。最终,时空特征$\mathbf{F}_{ST}\in\mathbb{R}^{T\times N\times d}$可以表达为:$\mathbf{F}_{ST}=\mathbf{W}_{\mathrm{space}}\odot\mathbf{x}_{\mathrm{S}}+\mathbf{W}_{\mathrm{time}}\odot\mathbf{x}_{\mathrm{T}}$。其中,$\odot$表示逐元素相乘。注意力机制能够帮助推荐模型为输入的各个部分分配不同的权重(高广尚,2022),而本文的自学习注意力机制能够自动捕捉时空特征的相对重要性:对于快速变化的AU如AU4眉毛下压,时间路径的权重可能占主导。而对于静态或共激活的AU如与微笑相关的AU12+AU6,空间路径的权重往往更强。

1.6 AU分类模块

在本文的工作中,AU的分类由AU分类模块完成,该模块基于从时空特征融合模块输出时空特征$\mathbf{F}_{ST}\in\mathbb{R}^{T\times N\times d}$进行分类。对于第$t$帧的第$i$个AU(其中$t=1,2,\cdots,T$以及$i=1,2,\cdots,N$),其发生概率$\hat{y}_{i,t}$通过计算时空特征表示$\mathbf{F}^{i,t}_{ST}\in\mathbb{R}^d$与一个可学习向量$\mathbf{s}_i\in\mathbb{R}^d$的余弦相似度来确定,如公式(5)所示:

$$\hat{y}_{i,t}=\frac{\left(\mathrm{ReLU}(\mathbf{F}^{i,t}_{ST})\right)^{\top}\mathrm{ReLU}(\mathbf{s}_i)}{\|\mathrm{ReLU}(\mathbf{F}^{i,t}_{ST})\|_2\cdot\|\mathrm{ReLU}(\mathbf{s}_i)\|_2} \tag{5}$$

其中$\|\cdot\|_2$表示向量的L_2范数。最终,得到分类输出$\hat{\mathbf{Y}}=\{\hat{y}_{i,t}\}\in\mathbb{R}^{T\times N}$。为了应对AU检测任务中的数据不平衡问题,本文引入了加权交叉熵损失(Rezaei-Dastjerdehei等,2020),如公式(6)所示:

$$\mathcal{L}_{\mathrm{cla}}=-\frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{N}w_i\left[y_{i,t}\log(\hat{y}_{i,t})+(1-y_{i,t})\log(1-\hat{y}_{i,t})\right] \tag{6}$$

其中,$\mathcal{L}_{\mathrm{cla}}$表示分类损失,$y_{i,t}\in\{0,1\}$是AU的真实标签,指示第$t$帧中第$i$个AU的存在$(y_{i,t}=1)$或不存在$(y_{i,t}=0)$。$w_i=N(1-r_i)$是第$i$个AU的损失权重,$r_i$为从训练集中计算的第$i$个AU的出现率。通过这种加权方式,网络被迫更加关注出现率较低的AU。

2 结 果

2.1 数据集

为了评估所提出模型的性能,在BP4D和DISFA两个公开的人脸AU检测数据集上进行了训练和测试。这两个数据集的详细信息如下:

BP4D数据集(Zhang等,2013)包含41名18至29岁的年轻个体在与实验员进行八种不同交互会话过程中表现出的各种情绪反应的2D和3D视频。参与者包括来自不同种族的23名女性和18名男性。其中,包括11名亚洲人、6名非裔美国人、4名西班牙裔和20名高加索人。数据采集过程由专业演员和表演艺术指导监督,使用了一系列有效任务来激发被试的真实情绪反应。这些任务包括陌生人之间的社交访谈、预设活动、电影片段观看、导致疼痛的冷压测试、激怒行为、引发疼痛的刺激以及引发厌恶的嗅觉刺激。最终从328段视频中获得了大约140000张有效人脸图像。这些图像或视频由有经验并获得

认证的FACS标注专家为视频标注了27个AU的存在情况。

DISFA数据集(Mavadati等,2013)包含27名成年人在观看能激发自发表情的电影片段时,通过立体相机采集的视频记录。参与者包括12名女性和15名男性,年龄范围为18至50岁。人口统计学分布包括3名亚洲人、21名高加索人、2名西班牙裔和1名非裔美国人。大多数供参与者观看的视频片段来自YouTube。每位参与者的人脸视频时长为4分钟,包含4844帧,总共生成130,788张图像。这些图像由两位认证的FACS专家标注了12个AU。

BP4D和DISFA数据集均存在严重的AU不平衡问题(Liang等,2024)。例如BP4D中的AU13、AU18、AU19仅出现在几百帧中。为了解决这一问题,遵循数据集作者的方法,从BP4D中选择12个AU,从DISFA中选择8个AU用于训练和评估。本文采用三折交叉验证策略,确保训练集和测试集中的被试互不重叠。在每一轮中,两折用于训练,一折用于测试。最终性能取三折的平均值。

2.2 实现细节

所有实验都在相同配置的软件/硬件平台上进行,以确保可比性。硬件平台包括NVIDIA GeForce RTX 2080Ti GPU(11GB显存)和Intel Core I7-9700 CPU,软件环境基于Python 3.8和PyTorch深度学习框架。

在训练过程中,使用动量系数为0.9的随机梯度下降优化器,采样帧为16,批大小为2。整个框架训练了240个轮次。为防止初始学习率过大而导致无法收敛到全局最优,采用余弦衰减策略动态调整学习率,公式如下:

$$\mathrm{lr}_\theta = \mathrm{lr}_{\mathrm{init}} \times \left[\frac{1 + \cos\left(\frac{\theta\pi}{20}\right)}{2} \times (1-\gamma) + \gamma \right] \tag{7}$$

其中,θ是迭代轮次的索引,γ是衰减因子。经过大量实验,初始学习率$\mathrm{lr}_{\mathrm{init}}$和$\gamma$均设置为0.01。公式(4)中的缩减比率r取值为1。

2.3 消融实验

为了分析所提出模型中不同模块对AU检测性能的影响,首先在BP4D数据集上进行了消融实验,并在不同实验场景下进行了测试。为了评估时空建模模块对AU检测模型的影响,在BP4D测试集上评估了包含/不包含空间建模模块、时间建模模块以及时空融合模块的模型。对于缺少空间建模模块或时间建模模块的模型,通过直接去除GNN和池化层且只保留简单的折叠和映射,并将结果直接送入SFF模块。对于缺少时空融合模块的模型,采用最简单的策略,即将空间特征与时间特征相加后直接输入ACM模块进行分类。评估结果如表1所示,其中"√"和"×"分别表示模型中包含与不包含对应的模块或策略。

表1 不同模块对模型性能的影响(平均值±标准差(%))

Table 1 Impact of different blocks on model performance (mean ± std(%))

模型	TRM	SRM	SFF	F1-分数	准确率	参数量	FLOPs
1	×	×	×	63.56 ± 0.55	75.89 ± 0.33	11.69M	2.382G
2	√	×	×	64.20 ± 0.95	75.78 ± 1.95	13.02M	2.489G
3	×	√	×	64.34 ± 1.18	76.67 ± 0.35	13.02M	2.489G
4	×	×	√	63.67 ± 0.75	75.97 ± 0.65	12.75M	2.483G
5	×	√	√	65.04 ± 0.88	**76.77 0.52**	13.05M	2.489G
6	√	×	√	65.11 ± 0.72	76.69 ± 0.67	13.05M	2.489G
7	√	√	×	<u>65.36 ± 0.34</u>	76.37 ± 0.41	13.28M	2.495G
8	√	√	√	**66.00 0.64**	<u>76.74 ± 0.70</u>	13.31M	2.495G

注:加粗数据表示最优结果,下划线数字表示次优结果,数据格式为"均值±标准差"是基于3折交叉验证的结果

如表1所示,首先从基线模型来看,当空间建模、时间建模以及特征融合模块均未引入时,模型仅依赖简单的特征折叠与映射,其性能较低,F1分数仅为63.56%,准确率只到达75.89%,为后续分析提

供了性能下限。其次，分别考察单一模块的作用可以发现：引入时间建模后，F1分数提升至64.20%，但准确率基本不变，说明时间建模对AU检测有一定帮助，但提升有限。相比之下，加入空间建模后带来了更明显的增益，F1分数提升至64.34%，准确率相比基线提高了0.78%，表明AU检测更依赖于面部区域间的空间依赖关系。而仅引入特征融合模块并未带来显著改善，其性能与基线基本持平，说明特征融合模块必须依托于时空特征建模才能发挥作用。尽管特征融合模块单独带来的性能提升并不显著，但其所需的参数量和FLOPs都非常接近于基线模型。这说明特征融合模块的计算代价极低，在与空间建模或时间建模结合时能以较小的额外开销换取稳定的性能提升，因此在整体框架中依然是不可或缺的组成部分。

进一步地，结合两个模块时性能均有显著提升。模型5（SRM+SFF）在F1-score上表现最差，说明仅仅依靠空间建模与融合策略对AU的召回效果不

佳。模型7（TRM+SRM）的表现最佳，这表明在AU检测中，同时对空间信息与时间信息进行联合建模与增强，能够有效提升模型的判别能力与整体性能。值得注意的是，SFF与TRM/SRM的配合均能带来性能提升，验证了特征融合对时/空信息的补充的有效性。当三个模块同时引入时，模型在F1分数上达到最高值，比基线提升了2.44个百分点，准确率也保持在较高水平。同时，该配置下参数量与计算量的增加十分有限，仅比基线增加约1.6M参数和0.11G FLOPs，证明所提出框架在性能与效率之间取得了良好的平衡。

如表2所示，完整模型在F1分数和准确率上的性能分别达到了66.00%以及76.74%，均取得了最佳整体性能，明显优于分别去除各个模块后的消融版本。这充分说明了空间关系建模、时间关系建模以及时空特征融合三者的协同作用对于提升AU检测准确率的必要性。以下从整体性能与单个AU表现两个层面进行分析。

从整体性能来看，完整模型相较于w/o SRM、w/o TRM和w/o SFF三个版本均有明显提升。移除SRM时，F1分数下降至65.11%，准确率下降至76.69%，说明帧内AU间的语义交互对于消除歧义、增强鲁棒性至关重要，例如AU6与AU12之间的共激活关系。移除TRM时，F1分数下降至65.05%，准确率降至76.77%，这表明跨帧的动态一致性在时序信息捕捉方面不可或缺，特别是涉及随时间逐渐增强的AU，如AU1和AU7。而移除SFF模块时，F1分数下降到65.36%，准确率则降为76.37%，说明如果没有自适应融合策略，空间与时间特征的权重难以动态调整，从而在某些情况下丢失关键信息。整体而言，完整模型的结果优于三者，证明了模块间的互补性。

从具体AU类别来看，动态性较强的AU，如AU6（脸颊抬起）和AU17（下巴抬起），完整模型的F1分数分别达到82.27%和68.54%，明显优于w/o TRM与w/o SRM的版本。这说明TRM能够有效捕捉动态AU在时间维度上的变化规律，而SRM进一步通过帧内相关性增强弱信号，二者结合使得模型在连续帧的变化检测中表现最优。相较于动态AU，静态或幅度较小的AU在本文的模型中表现相对有限，例如AU1眉毛内侧抬起、AU2眉毛外侧抬起。F1分数分别为52.65%和48.18%。这类AU变化幅度小、信号弱，很容易受到光照、姿态等干扰，仅依赖时空同时建模难以完全解决。相比之下，保留了SRM的模型在这类AU上的表现相对更优，其优势来源于对AU间相关性的显式建模，即通过强相关AU来增强弱AU的检测。

接下来，我们从不同角度对TRM与SRM模块分别进行了进一步的补充实验分析。针对TRM模块，我们将原有的TRM结构替换为标准的Transformer编码器，以验证基于差分的时序建模策略是否会在特征压缩过程中引入额外噪声，或导致关键语义信息的丢失。通过与Transformer时序建模方式的直接对比，可以更清晰地评估TRM在动态信息建模上的有效性与合理性。针对SRM模块，我们进一步对其中GNN的Top-K邻域参数进行了系统性的消融实验，分析不同K值对模型性能的影响。该实验旨在验证模型对邻域规模的敏感性，并说明在AU相关性建模过程中，合理的K取值如何在捕捉有效关联与抑制冗余噪声之间取得平衡。

结合表3的定量结果可以发现，不同AU对时序建模与空间关系建模的依赖程度存在显著差异。对于以明显肌肉运动为主导的动态AU6、AU7、AU10和AU12，GNN模型在F1分数上整体优于或至少不劣于Transformer方案。这表明，TRM通过差分建模显式刻画相邻帧之间的运动变化，在突出时

表2　单一模块对模型性能的影响(平均值±标准差(%))

Table 2　Impact of single blocks on model performance (mean ± std(%))

AU (F1-分数)	Model w/oTRM	Model w/oSRM	Model w/oSFF	Model
AU1	**56.04　0.92**	51.74±0.24	55.93±0.43	52.65 ± 0.80
AU2	46.22±0.38	45.93±0.65	**52.82　0.35**	48.18 ± 0.25
AU4	**64.25　0.23**	63.79±0.88	60.94±0.86	61.15 ± 1.24
AU6	75.88±0.59	76.11±0.29	81.67±0.39	**82.27　0.05**
AU7	82.02±1.40	**82.37　0.61**	78.55±0.46	80.97 ± 0.34
AU10	85.10±1.93	84.94±1.00	84.45±0.97	**89.55　0.81**
AU12	87.43±0.66	87.25±0.93	87.52±1.02	**88.02　0.78**
AU14	66.55±0.95	64.18±0.52	65.09±0.36	**68.79　0.09**
AU15	43.92±0.29	**49.40　0.17**	44.37±0.71	44.44 ± 0.68
AU17	67.72±0.14	65.94±0.29	66.07±0.98	**68.54　0.94**
AU23	48.41±0.71	**54.33　0.33**	49.41±1.03	50.00 ± 1.02
AU24	57.01±0.93	55.36±0.47	**58.50　0.22**	57.39 ± 0.74
Avg	65.05±0.26	65.11±0.17	65.36±0.21	**66.00　0.64**
AU (准确率)	**Model w/oTRM**	**Model w/oSRM**	**Model w/oSFF**	**Model**
AU1	**82.49　0.85**	80.74±0.41	81.28±0.51	74.28 ± 0.48
AU2	78.33±0.49	72.28±0.83	80.85±0.49	**81.19　0.29**
AU4	80.89±0.37	**83.99　0.45**	81.78±0.97	80.53 ± 0.60
AU6	74.25±0.70	75.85±0.58	71.54±0.53	**77.10　1.10**
AU7	**77.71　1.03**	74.89±0.84	68.95±0.59	74.64 ± 0.91
AU10	79.55±1.72	82.18±1.11	79.83±1.20	**84.80　0.84**
AU12	82.39±0.82	82.32±0.78	82.09±0.77	**84.50　1.49**
AU14	57.19±1.18	55.07±0.78	57.91±0.46	**59.31　1.11**
AU15	73.27±0.44	**73.71　0.48**	72.69±0.62	72.96 ± 0.27
AU17	70.96±0.25	69.21±0.39	71.61±1.02	**73.14　0.02**
AU23	81.41±0.81	**85.11　0.61**	82.06±1.22	76.32 ± 0.65
AU24	82.83±1.20	84.91±0.55	**85.82　0.48**	82.15 ± 0.70
Avg	**76.77　0.26**	76.69±0.20	76.37±0.23	76.74 ± 0.70

注:加粗数据表示最优结果,下划线数字表示次优结果,,数据格式为“均值±标准差”是基于3折交叉验证的结果以及w/o表示移除相应模块

间维度上的局部动态信息方面具有明显优势。从整体均值来看,Transformer与GNN在F1分数上表现接近,但GNN模型在多次实验中的标准差略低,表明其在AU级识别任务中具有更好的稳定性。此外,在保持主干、输入帧数及损失函数一致的前提下,GNN方案在参数量与计算复杂度方面显著低于基于Transformer的时序建模方式。

结合图5的可视化结果,可以更加直观地理解

表3　不同实验设置下模型性能对比(平均值±标准差(%))

Table 3　The performance of different settings on model performance (mean ± std(%))

AU F1-分数	Transformer	GNN (Base)	AU 准确率	Transformer	GNN (Base)
AU1	**60.40 0.61**	52.65 ± 0.80	AU1	**81.11 0.77**	74.28 ± 0.48
AU2	45.75 ± 0.35	**48.18 0.25**	AU2	**85.39 0.39**	81.19 ± 0.29
AU4	59.44 ± 0.85	**61.15 1.24**	AU4	75.36 ± 0.48	**80.53 0.60**
AU6	79.42 ± 0.37	**82.27 0.05**	AU6	**80.91 1.32**	77.10 ± 1.10
AU7	79.47 ± 0.77	**80.97 0.34**	AU7	**77.43 0.25**	74.64 ± 0.91
AU10	82.97 ± 0.24	**89.55 0.81**	AU10	79.13 ± 0.74	**84.80 0.84**
AU12	86.42 ± 0.66	**88.02 0.78**	AU12	84.32 ± 1.25	**84.50 1.49**
AU14	**68.95 0.39**	68.79 ± 0.09	AU14	**62.61 1.01**	59.31 ± 1.11
AU15	**46.79 0.62**	44.44 ± 0.68	AU15	**81.93 0.38**	72.96 ± 0.27
AU17	68.45 ± 0.89	**68.54 0.94**	AU17	**76.02 0.40**	73.14 ± 0.02
AU23	**57.83 1.33**	50.00 ± 1.02	AU23	**82.00 0.22**	76.32 ± 0.65
AU24	**60.88 1.25**	57.39 ± 0.74	AU24	**86.30 0.97**	82.15 ± 0.70
Mean	66.40 ± 0.67	66.00 ± 0.64	Mean	79.38 ± 0.64	76.74 ± 0.70

注:加粗数据表示最优结果,数据格式为"均值±标准差"是基于3折交叉验证的结果

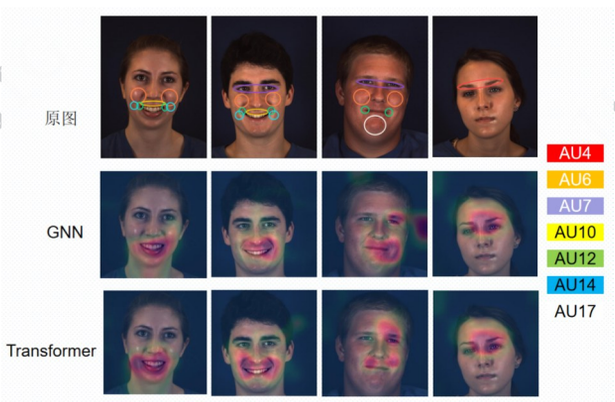

图5　TRM 模块实验结果可视化

Fig. 5　Visualization of Experimental Results for the TRM Module

TRM 模块在时序建模过程中的作用机制。第一行展示了原始输入图像及对应的 AU 关键区域示意,可以看到不同样本在面部肌肉激活位置与强度上存在明显差异。第二行与第三行分别给出了基于 GNN 和 Transformer 编码器的特征响应热力图。

从图中可以观察到,TRM 模型的响应区域更加集中于与 AU 激活直接相关的局部区域,例如 AU6 和 AU12 对应的脸颊抬起区域、AU10 对应的上唇抬起区域,以及 AU7、AU17 所涉及的眼周和下巴区域。其激活分布呈现出明显的局部性和一致性,且在不同个体之间具有较好的稳定性。这表明,TRM 通过差分操作有效抑制了静态背景与个体差异的干扰,使模型更加专注于跨帧变化所带来的关键动态

信息。相比之下，Transformer的激活区域虽然覆盖范围更广，但在部分样本中呈现出较为分散的响应模式，部分与目标AU无直接关联的区域也被赋予了较高的响应权重。这一现象说明，全局自注意力机制在捕捉整体时序一致性的同时，也可能引入冗余的时序依赖，从而削弱对局部动态变化的聚焦能力。

值得注意的是，对于动态性较强的AU，GNN模型在时间维度上的响应更加连贯，体现出对连续帧中微小形变的敏感性；而Transformer的响应则更倾向于对整体外观变化进行建模。这一差异与表3中GNN在动态AU上取得更优或相当F1分数的定

表4　不同实验设置下模型性能的F1分数对比（平均值±标准差（%））

Table 4　The F1-Scoreof different settings on model performance (mean ± std(%))

AU / K	K=1	K=2	Base	K=4	K=5
AU1	53.17 ± 0.74	50.31 ± 0.68	52.65 ± 0.80	**53.42　0.92**	50.01 ± 0.67
AU2	48.35 ± 0.22	45.11 ± 0.27	48.18 ± 0.25	**51.75　0.03**	46.25 ± 0.14
AU4	58.97 ± 1.45	59.45 ± 1.65	**61.15　1.24**	60.38 ± 1.12	59.20 ± 1.57
AU6	79.19 ± 0.30	79.52 ± 0.27	**82.27　0.05**	79.97 ± 1.22	81.77 ± 0.65
AU7	79.91 ± 0.93	80.44 ± 0.79	80.97 ± 0.34	**81.78　0.50**	77.30 ± 0.22
AU10	81.08 ± 0.70	85.18 ± 0.41	**89.55　0.81**	88.56 ± 1.24	89.04 ± 1.25
AU12	85.18 ± 0.66	88.58 ± 0.52	88.02 ± 0.78	85.42 ± 0.73	**90.80　0.48**
AU14	**71.43　0.57**	69.71 ± 0.11	68.79 ± 0.09	65.84 ± 0.65	69.71 ± 0.75
AU15	**56.46　0.15**	50.68 ± 1.03	44.44 ± 0.68	46.58 ± 0.74	48.65 ± 0.32
AU17	66.27 ± 1.02	68.20 ± 0.82	68.54 ± 0.94	65.64 ± 0.82	**68.96　0.37**
AU23	47.48 ± 0.74	50.73 ± 0.81	50.00 ± 1.02	**52.89　1.19**	52.80 ± 0.84
AU24	52.25 ± 0.26	60.72 ± 0.35	57.39 ± 0.74	**61.29　0.26**	57.95 ± 0.96
Mean	64.98 ± 0.71	65.72 ± 0.62	66.00 ± 0.64	66.04 ± 0.82	66.03 ± 0.76

结果高度一致。综上所述，图5的可视化结果从直观层面验证了TRM模块设计的合理性：基于差分的时序建模并未引入明显噪声，反而通过聚焦跨帧变化增强了对动态AU的判别能力。同时，与Transformer相比，TRM在保持识别性能的前提下，实现了更具针对性的动态区域建模和更高的计算效率。

如表4与表5所示，当K从1增加到3时，模型的识别性能呈现单调上升趋势，并在K=3时接近峰值。这验证了我们的假设：聚合2-3个最显著的动态区域对于全面捕捉微表情至关重要。当K > 3时，性能出现稳定，，但并未继续显著提升，且在部分AU上出现轻微波动甚至下降。这一现象表明，过大的K值会引入更多与目标情绪表达相关性较弱的区域特征，从而稀释关键动态区域的贡献，并可能带来额外噪声干扰。这与微表情通常由有限数量的核心肌肉群协同驱动这一生理事实相一致。

基于上述系统的参数敏感性实验，我们最终选择K=3作为模型的默认参数。

2.4　对比实验与结论

为了验证本文所提模型的优势，将其与一些最新的基于区域的人脸动作单元检测方法进行了比较。这些对比方法包括DRML（Zhao等，2016）、RE-ne（Yang等，2020）、SRERL（Li等，2019）、EWRBAL（Tian等，2022）、AUGAIN（Li等，2025）、AC2D（Shao等，2025）、JAO（Zhou等，2024）、LTI（Yang等，2019）、AUNET（Yang等，2023）。在BP4D和DISFA数据集上，不同算法在F1分数上的性能别列于表3和表4中。由于采用了三折交叉验证来检验模型的泛化能力，因此在这些表格中也给出了不同折之间的标准差。为了提高可信度，其它方法的评估结果直接引用自其原始论文。由于表3和表4中的大多数引用论文未提供准确率指标下的性能，因此本文也没有对该指标下的性能进行对比分析。

表5　不同实验设置下模型性能的准确率对比(平均值±标准差(%))

Table 5　The Accuracy of different settings on model performance (mean ± std(%))

AU / K	K=1	K=2	Base	K=4	K=5
AU1	78.18 ± 0.63	71.33 ± 0.98	74.28 ± 0.48	77.93 ± 0.74	**78.90　1.36**
AU2	73.04 ± 0.33	81.19 ± 0.47	81.19 ± 0.29	**81.97　0.22**	81.52 ± 0.65
AU4	**84.89　0.87**	78.28 ± 0.97	80.53 ± 0.60	76.85 ± 0.71	80.10 ± 0.48
AU6	79.28 ± 0.40	78.23 ± 0.54	77.10 ± 1.10	77.57 ± 1.27	**80.53　1.32**
AU7	75.32 ± 1.25	73.03 ± 1.42	74.64 ± 0.91	73.55 ± 1.12	**78.48　1.47**
AU10	74.86 ± 0.53	79.54 ± 0.66	**84.80　0.84**	80.52 ± 0.63	80.10 ± 1.41
AU12	82.28 ± 1.24	85.77 ± 1.15	84.50 ± 1.49	83.28 ± 1.32	**86.64　0.69**
AU14	64.30 ± 1.12	63.68 ± 1.10	59.31 ± 1.11	60.50 ± 0.94	**67.83　1.73**
AU15	78.52 ± 0.75	80.90 ± 0.79	72.96 ± 0.27	74.41 ± 1.33	**81.84　1.18**
AU17	70.53 ± 0.68	**74.51　0.32**	73.14 ± 0.02	68.63 ± 0.47	72.50 ± 0.51
AU23	74.43 ± 0.09	76.99 ± 0.48	76.32 ± 0.65	78.03 ± 0.15	**80.18　1.00**
AU24	77.67 ± 1.31	**83.64　0.85**	82.15 ± 0.70	83.30 ± 1.02	82.48 ± 0.66
Mean	76.11 ± 0.88	77.26 ± 0.65	76.74 ± 0.70	76.38 ± 0.73	79.26 ± 0.62

注:加粗数据表示最优结果,下划线数字表示次优结果。数据格式为"均值±标准差"是基于3折交叉验证的结果

如表3所示,本文的模型在BP4D数据集上的F1分数达到了66.00%,为当前最优水平。其中,在12个AU中有5个取得了最高的F1分数,另有2个达到次优水平。这表明本模型在检测涉及大幅度肌肉运动的AU时具有显著优势。例如,AU6与AU7在表情变化过程中伴随着明显的肌肉收缩,而收缩过程往往随时间动态演变,这对于本文所采用的时空特征建模尤为有利。相比之下,诸如AU1/2/4这类集中于眉部的细微运动,其图像变化极为微弱,因此难以被准确感知,这也从侧面反映出本文模型在局部细节检测上的不足。与之对比,RE-Net强调AU之间的相关性并引入了邻域自适应机制,尽管缺乏时间特征建模,但其在AU关联性建模上表现突出,从而提高了对弱AU或低频AU的检测能力。此外,AC2D通过因果解耦增强了每个AU的独立性,在个别AU的检测上取得了明显提升,但在整体面部AU的建模中仍然缺乏全局特征的把握。而AUGAIN则依赖分割网络对单个AU的感兴趣区域进行学习,并通过landmark定位来生成软掩模标签,从而达到较好的检测效果。然而,这种方法训练开销大、模型规模庞大,在效率和资源利用上不及本文的模型。

为了体现出模型的轻量化优势,除了在F1分数上的比较之外,该模型还与一些公开源代码的方法在计算效率和模型规模方面进行了对比。与现有的AC2D和AUGAIN方法相比,该模型计算开销显著更小,但仍然实现了最先进的检测性能。具体来说,AC2D的FLOPs数为15.10G,模型参数量为13.12 M。而AUGAIN的FLOPs数为62.93G,模型参数量为89.10M。相比之下,该模型在仅需2.50GFLOPs和13.31M参数的情况下,就能达到具有竞争力的精度,这表明该架构在计算和存储需求方面都更加高效。综上所述,本文提出的模型在轻量化设计的前提下,仍然优于上述三类方法,并在整体AU检测上表现出色。尽管在部分细微AU的检测上仍有提升空间,但整体性能已展现出较强的实用性与优势。

如表4所示,本文的模型在DISFA数据集上的F1分数达到了65.34%,为当前的次优水平。在8个AU中,本模型在4个AU上取得了最佳结果,另有2个AU达到次优表现。尽管整体上未能超越表现最佳的RE-Net,但与其差距不足0.1个百分点,且在单独AU检测上,本模型仍然占据了近一半的最优结果。这一现象与数据集特性密切相关。相较于BP4D,DISFA数据集中的人脸图像质量更低,

存在一定的模糊性,这在一定程度上削弱了本文模型对于运动轨迹捕捉的优势。而RE-Net更加

表3　不同模型在BP4D数据集上的表现(平均值±标准差(%))

Table 3　Performance of different models on BP4D dataset (mean ± std(%))

AU	DRML	LTI	EWRBAL	SRERL	AUGAIN	JAO	AUNET	AC2D	RE-Net	Our
AU1	36.4	50.8	45.6	47.7	44.2	54.4	54.2	54.2	**57.7**	52.65 ± 0.80
AU2	41.8	45.3	41.8	50.9	42.0	50.1	44.9	54.7	**59.0**	48.18 ± 0.25
AU4	43.0	56.6	54.6	49.5	52.6	57.7	61.6	56.5	**66.9**	61.15 ± 1.24
AU6	55.0	75.9	78.5	75.8	80.9	79	76.8	77.0	76.3	**82.27　0.05**
AU7	67.0	75.9	73.4	78.7	71.1	76	76.6	76.2	77.0	**80.97　0.34**
AU10	66.3	80.9	82.0	80.2	85.4	83.7	83.6	84.0	88.9	**89.55　0.81**
AU12	65.8	88.4	87.7	84.1	88.0	87.7	88.8	89.0	**89.8**	88.02 ± 0.78
AU14	54.1	63.4	62.2	67.1	74.2	64.8	63.9	63.6	**70.9**	68.79 ± 0.09
AU15	33.2	41.6	38.9	52.0	44.6	47.9	52.3	**54.8**	42.0	44.44 ± 0.68
AU17	48.0	60.6	61.7	62.7	65.4	62.3	65.7	63.6	62.8	**68.54　0.94**
AU23	31.7	39.1	43.6	45.7	**52.1**	44.1	48.5	46.5	44.8	50.00 ± 1.02
AU24	30.0	37.8	47.3	54.8	54.0	48.4	48	54.8	49.3	**57.39　0.74**
Avg	48.3	59.7	59.8	62.4	62.9	63.0	63.8	64.6	65.5	**66.00　0.64**

表4　不同模型在DISFA数据集上的表现(平均值±标准差(%))

Table 4　Performance of different models on DISFA dataset (mean ± std(%))

AU	DRML	LTI	RE-Net	SRERL	AUGAIN	AUNET	AC2D	Our
AU1	17.3	30.9	38.8	45.7	49.2	**59.3**	57.8	53.60 ± 0.34
AU2	17.7	34.7	31.1	47.8	30.9	55.3	**59.2**	**59.21　1.38**
AU4	37.4	63.9	57.2	59.6	70.4	69.4	70.1	**74.63　0.20**
AU6	29.0	44.5	50.1	47.1	46.5	49.0	50.1	**54.50　0.58**
AU9	10.7	31.9	50.2	45.6	54.0	45.9	**54.4**	40.41 ± 1.04
AU12	37.0	78.3	75.5	73.5	73.8	77.0	75.1	**83.84　0.93**
AU25	38.5	84.7	86.6	84.3	80.5	**91.8**	90.3	90.57 ± 0.27
AU26	20.1	60.5	50.6	43.6	56.5	60.0	**66.2**	65.98 ± 0.45
Avg	26.7	30.9	55.0	55.9	57.7	63.5	**65.4**	65.34 ± 0.27

注：加粗数据表示最优结果，下划线数字表示次优结果，，数据格式为“均值±标准差”是基于3折交叉验证的结果

注重AU之间的相关性建模，这使其在弱AU或低频AU的检测上具备更强的表现，从而在DISFA这种低清晰度场景下取得了更优的整体结果。同样地，AUNET也强调AU相关性，也取得了较好的性能。因此，在DISFA数据集上相关性建模能力较强的模型往往能获得更佳表现。然而，需要指出的是，本模型在设计上注重轻量化与时空建模，其在BP4D等高质量数据集上表现尤为突出。因此，未来的工作应当进一步探索如何在低分辨率、非实验室环境下增强对运动轨迹的鲁棒性建模，使模型在更具挑战性的真实场景中依旧能够实现与BP4D类似的优异性能。

3　结　论

本文提出了一种统一的AU检测框架，能够在时空两个维度上协同建模，有效提升AU检测的准确性和鲁棒性。通过时间关系建模(TRM)、空间关

系建模(SRM)以及时空特征融合(SFF)模块的协同作用,模型能够捕捉AU的动态变化及其全局依赖关系。实验结果表明,所提出的方法在保持较低计算和内存开销的同时,实现了对细粒度表情动态的准确识别。然而,现有框架在局部区域细粒度特征提取方面仍存在不足。尽管模型通过ResNet-18主干与图神经网络捕捉了全局的空间-时间依赖关系,但对特定面部肌肉群的微小运动模式利用不足,可能导致对轻微AU激活或受遮挡区域的识别性能下降。

未来的研究将从以下几个方向展开:其一、对于局部区域特征建模。在整体框架中引入局部区域注意力机制或区域原型学习方法,更精细地刻画局部肌肉群的动态变化。其二、多尺度融合。结合局部区域和全局表征,探索多尺度的特征融合策略,以增强模型对细微AU的敏感性。其三、跨模态信息融合。尝试引入语音、心率等多模态信号,与视觉AU特征进行联合建模,从而进一步提升检测的稳定性和泛化性。

综上所述,本研究为人脸AU检测提供了一种高效的时空建模方案,但局部区域特征的深度挖掘仍是未来亟需突破的方向。

参考文献(References)

Chen D L, Shen Y T, Xie B Z and Wang L. 2014. Optimal coach selection algorithm based on cosine similarity model. Journal of Northeastern University (Natural Science), 35(12): 1697 (陈大力, 沈岩涛, 谢槟竹, 王磊. 2014. 基于余弦相似度模型的最佳教练遴选算法. 东北大学学报(自然科学版), 35(12): 1697. [DOI: 10.3969/j.issn.1005-3026.2014.12.006]

Cui X Y, He C, Zhao H K and Wang M L. 2024. Facial expression recognition based on ViT and contrastive learning. Journal of Image and Graphics, 29(01): 123-133 (崔鑫宇, 何翀, 赵宏珂, 王美丽. 2024. 融合ViT与对比学习的面部表情识别. 中国图象图形学报, 29(01): 123-133) [DOI: 10.11834/jig.230043]

Dong H X, An Y, Xie L R and Liu Z. 2024. Large-scale semantic segmentation of mine laser point clouds based on local feature enhancement. Chinese Journal of Lasers, 51(17): 1710002-12 (董宏翔, 安毅, 谢丽蓉, 刘泽. 2024. 基于局部特征增强的矿山大规模激光点云语义分割. 中国激光, 51(17): 1710002-12). [DOI: 10.3788/cjl231425]

Ekman P, Friesen W V and Hager J C. 1978. Facial action coding system (FACS). San Francisco: Consulting Psychologists Press: 22

Fagin R, Kumar R and Sivakumar D. 2003. Comparing top k lists. SIAM Journal on Discrete Mathematics, 17(1): 134-160 [DOI: 10.1137/s0895480102412856]

Gao G S. 2022. Attention mechanism in deep learning recommendation models: a survey. Journal of Computer Engineering & Applications, 58(9): 9-18 (高广尚. 2022. 深度学习推荐模型中的注意力机制研究综述. 计算机工程与应用, 58(9): 9-18) [DOI: 10.3778/j.issn.1002-8331.2112-0382]

Jaiswal S and Valstar M. 2016. Deep learning the dynamic appearance and shape of facial action units//Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV). Lake Placid: IEEE: 1-8 [DOI: 10.1109/wacv.2016.7477625]

Li G, Zhu X, Zeng Y and Wang Y. 2019. Semantic relationships guided representation learning for facial action unit recognition//Proceedings of the AAAI Conference on Artificial Intelligence. 33(1): 8594-8601 [DOI: 10.1609/aaai.v33i01.33018594]

Li K, Liang C, Zou W and Liu M. 2025. Facial AU detection based on a guided attention inference network with embedded regional segmentation branch. Neural Computing and Applications, 1-25 [DOI: 10.1007/s00521-025-11254-x]

Li W, Abtahi F, Zhu Z and Yin L. 2017. 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 (FG). Washington: IEEE: 103-110 [DOI: 10.1109/fg.2017.136]

Liang C, Zou W, Hu D and Wang F. 2024. Facial action unit recognition based on self-attention spatiotemporal fusion//Proceedings of the 5th International Conference on Computing, Networks and Internet of Things. 2024: 600-605 [DOI: 10.1145/3670105.3670210]

Lin ZH, Mo XG, Li HX and Li HB. 2002. Comparison of three spatial interpolation methods for climate variables in China. Acta Geographica Sinica, 57(1): 47-56 (林忠辉, 莫兴国, 李宏轩, 李海滨. 2002. 中国陆地区域气象要素的空间插值. 地理学报, 57(1): 47-56) [DOI: 10.3321/j.issn:0375-5444.2002.01.006]

Liu T, Li J, Wu J and Zhang Y. 2024. Confusable facial expression recognition with geometry-aware conditional network. Pattern Recognition, 148: 110174 [DOI: 10.1016/j.patcog.2023.110174]

Lugaresi C, Tang J, Nash H and Fan L. 2019. Mediapipe: A framework for building perception pipelines. arXiv preprint arXiv: 1906.08172

Mavadati S M, Mahoor M H, Bartlett K and Cohn J. 2013. DISFA: A spontaneous facial action intensity database. IEEE Transactions on Affective Computing, 4(2): 151-160 [DOI: 10.1109/t-affc.2013.4]

Melgani F and Bruzzone L. 2004. Classification of hyperspectral remote sensing images with support vector machines. IEEE Transactions on Geoscience and Remote Sensing, 42(8): 1778-1790 [DOI: 10.1109/TGRS.2004.831865]

Rezaei-Dastjerdehei M R, Mijani A and Fatemizadeh E. 2020. Addressing imbalance in multi-label classification using weighted cross entropy loss function//Proceedings of the 27th National and 5th

International Iranian Conference on Biomedical Engineering (ICBME). Tehran: IEEE: 333-338 [DOI: 10.1109/icbme51989.2020.9319440]

Shao Z, Liu Z, Cai J and Zhang M. 2018. Deep adaptive attention for joint facial action unit detection and face alignment//Proceedings of the European Conference on Computer Vision (ECCV). Munich: Springer: 705-720 [DOI: 10.1007/978-3-030-01261-8_43]

Shao Z, Zhou Y, Li F and Wang H. 2024. Joint facial action unit recognition and self-supervised optical flow estimation. Pattern Recognition Letters, 181: 70-76 [DOI: 10.1016/j.patrec.2024.03.022]

Shao Z, Zhu H, Zhou Y and Chen J. 2025. Facial action unit detection by adaptively constraining self-attention and causally deconfounding sample. International Journal of Computer Vision, 133(4): 1711-1726 [DOI: 10.1007/s11263-024-02258-6]

Tian M, Zhu H, Wang Y and Chen R. 2022. Facial action unit detection by exploring the weak relationships between AU labels//Proceedings of the International Conference on Collaborative Computing: Networking, Applications and Worksharing. Cham: Springer: 478-495 [DOI: 10.1007/978-3-031-24386-8_26]

Yang H and Yin L. 2019. Learning temporal information from a single image for AU detection//Proceedings of the 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG). Lille: IEEE: 1-8 [DOI: 10.1109/fg.2019.8756556]

Yang H and Yin L. 2020. Re-net: A relation embedded deep model for AU occurrence and intensity estimation//Proceedings of the Asian Conference on Computer Vision (ACCV). 2020: 137-153 [DOI: 10.1007/978-3-030-69541-5_9]

Yang J, Hristov Y, Shen J and Li S. 2023. Toward robust facial action units' detection. Proceedings of the IEEE, 111(10): 1198-1214 [DOI: 10.1109/jproc.2023.3257542]

Yacoob Y and Davis L S. 2002. Recognizing human facial expressions from long image sequences using optical flow. IEEE Transactions on Pattern Analysis and Machine Intelligence, 18(6): 636-642 [DOI: 10.1109/34.506414]

Zhao K, Chu W S, Zhang H. Deep region and multi-label learning for facial action unit detection [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 3391-3399. [DOI: 10.1109/cvpr.2016.369]

Zhang X, Yin L, Cohn J F and Canavan S. 2013. A high-resolution spontaneous 3D dynamic facial expression database//Proceedings of the 10th IEEE International Conference on Automatic Face & Gesture Recognition (FG). Shanghai: IEEE: 1-6 [DOI: 10.1016/j.imavis.2014.06.002]

Zhou J, Liu X, Wang H and Guo Y. 2024. Seeing through the mask: Recognition of genuine emotion through masked facial expression. IEEE Transactions on Computational Social Systems, 11(6): 7159-7172 [DOI: 10.1109/tcss.2024.3404611]

Luo LH, Lai ZY and Kong FS. 2017. Video action recognition based on action segmentation and manifold metric learning. Journal of Image and Graphics, 22(8): 1106-1119 (罗会兰, 赖泽云, 孔繁胜. 2017. 动作切分和流形度量学习的视频动作识别. 中国图象图形学报, 22(8): 1106-1119) [DOI: 10.11834/jig.170032]

作者简介

黄海琳，女，硕士研究生，主要研究方向为人脸动作单元识别。E-mail: 20234228016@stu.suda.edu.cn