基于双向光流和小波注意力机制的微表情识别

解为成[1),2),4)] 肖 航[1)] 范玮嘉[1)] 汪子晗[1)] 余梓彤[5)] 沈琳琳[3),4)]

[1)](深圳大学计算机与软件学院 广东 深圳 518060)

[2)](人工智能与数字经济广东省实验室(深圳) 广东 深圳 518083)

[3)](深圳大学人工智能学院 广东 深圳 518060)

[4)](深圳大学广东省智能信息处理重点实验室 广东 深圳 518060)

[5)](大湾区大学信息科学技术学院 广东 东莞 523000)

摘 要 微表情是人在极短时间内不自觉产生的面部表情,能够揭示个体的真实情感状态。现有的微表情识别方法在准确性和鲁棒性方面仍存在一定的局限性,尤其是在噪声干扰和图像序列信息丢失的情况下,传统的光流方法难以有效捕捉微表情的细微变化。此外,深度学习模型在下采样过程中可能会丢失重要的高频信息,进而影响微表情特征的识别能力。针对这些问题,本文提出了一种基于双向光流和小波注意力机制的微表情识别方法,以提高识别准确性和鲁棒性。具体来说,首先分析了传统光流方法在微表情识别中的局限性,针对光流法对噪声的敏感性和连续帧间信息忽视的问题,提出了双向光流模块(Bidirectional Optical Flow Module, BOFM)。利用峰值到起始帧和结束帧的信息,以减少噪声干扰,并强化与微表情相关的面部肌肉运动特征。其次,为了缓解深度学习模型在下采样过程中可能丢失细节信息的问题,并提高模型对高频和低频信息的有效利用,本文引入了小波注意力下采样模块(Wavelet-based Attention Downsampling Module, WADM),以提升模型对微表情特征的表征能力。在MEGC2019数据集上的实验结果表明,本文提出的方法在未加权F1分数(UF1)和未加权平均召回率(UAR)两个指标上均优于现有方法,分别高出最好的方法1.87%和2.61%。同时,在CAS(ME)^{3}和DFME数据集上的实验也验证了该方法在不同场景与人群下的广泛适用性与优越性能。此外,通过Grad-CAM可视化分析,验证了双向光流和小波注意力下采样模块在特征表征过程中的有效性。该算法的前期版本在第四届中国情感计算大会(CCAC2024)微表情自动识别任务上取得了第二名的成绩。代码已公开在https://github.com/LeoDerekh/BOWAM。

关键词 微表情识别;双向光流;小波变换;注意力机制;频域特征

中图法分类号 TP18 **DOI号** 10.11897/SP.J.1016.2026.00328

Micro-Expression Recognition Based on Bidirectional Optical Flow and Wavelet-Based Attention Mechanism

XIE Wei-Cheng[1),2),4)] XIAO Hang[1)] FAN Wei-Jia[1)] WANG Zi-Han[1)]

YU Zi-Tong[5)] SHEN Lin-Lin[3),4)]

[1)](College of Computer Science & Software Engineering, Shenzhen University, Shenzhen, Guangdong 518060)

[2)](Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ), Shenzhen, Guangdong 518083)

[3)](School of Artificial Intelligence, Shenzhen University, Shenzhen, Guangdong 518060)

[4)](Guangdong Provincial Key Laboratory of Intelligent Information Processing, Shenzhen University, Shenzhen,

收稿日期:2025-03-11;在线发布日期:2025-10-31。本课题得到国家自然科学基金面上项目(Nos. 62276170, 62576216, 62576076),国家自然科学基金青年科学基金项目(No. 62306061)、广东省科技计划项目(Nos. 2023A1515011549, 2023A1515010688)、人工智能与数字经济广东省实验室(深圳)开放课题(No. GML-KF-24-11)以及广东省重点实验室(No. 2023B1212060076)资助。**解为成**,博士,副教授,中国计算机学会(CCF)高级会员,主要研究领域为面部表情分析和鲁棒网络设计。E-mail: wcxie@szu.edu.cn。**肖 航**,硕士研究生,主要研究领域为深度学习和微表情识别。**范玮嘉**,硕士研究生,主要研究领域为视觉识别任务和多模态学习。**汪子晗**,硕士研究生,主要研究领域为面部动作单元识别和面部表情编辑。**余梓彤**,博士,副教授,中国计算机学会(CCF)高级会员,主要研究领域为以人为中心的视觉计算、多媒体安全、多模态基础模型。**沈琳琳**(通信作者),博士,教授,中国计算机学会(CCF)会员,主要研究领域为深度学习、面部识别、分析/合成和医学图像处理。E-mail: llshen@szu.edu.cn。

Guangdong 518060）
[5)]（School of Computing and Information Technology, Great Bay University, Dongguan, Guangdong 523000）

Abstract Micro-expression is a facial expression produced unconsciously in a very short period of time, which can reveal the real emotional state of an individual. Existing micro-expression recognition methods still have some limitations in terms of accuracy and robustness, especially in the case of noise interference and loss of image sequence information, where the traditional optical flow method is difficult to effectively capture the subtle changes of micro-expressions. In addition, deep learning models may lose important high-frequency information during the downsampling process, which affects the recognition ability of micro-expression features. To address these problems, this paper proposes a novel micro-expression recognition method based on bidirectional optical flow and a wavelet-based attention mechanism to effectively improve recognition accuracy and robustness. Specifically, the limitations of the traditional optical flow method in micro-expression recognition are first analyzed, and a Bidirectional Optical Flow Module (BOFM) is then proposed to address the sensitivity of the optical flow method to noise as well as the problem of neglecting the information between consecutive frames. Unlike conventional single-directional optical flow, the proposed strategy captures richer temporal dynamics and mitigates motion inconsistency across frames. The information from the apex frame to the onset frame and offset frame is used to reduce noise interference and enhance the facial muscle movement features associated with micro-expressions. By combining forward and backward optical flows, the proposed BOFM captures the entire temporal evolution of facial motions, ensuring that both the generation and attenuation stages of expressions are effectively modeled while suppressing global noise caused by head movement. Secondly, in order to alleviate the problem that deep learning models may lose detailed information during downsampling and to improve the effective utilization of both high-frequency and low-frequency information, this paper introduces a robust and innovative Wavelet-based Attention Downsampling Module (WADM) to enhance the model′s representation capacity of micro-expression features. The WADM integrates wavelet transform and attention mechanisms to decompose image features into multi-scale frequency sub-bands. Low-frequency components preserve overall facial structures, while high-frequency components retain fine-grained details of micro-movements. By introducing spatial and channel attention mechanisms into different frequency branches, the model adaptively focuses on discriminative areas and assigns higher weights to the most informative sub-bands, effectively reducing the loss of subtle but crucial details and improving the robustness of feature learning under complex conditions. Experimental results on the MEGC2019 dataset show that the proposed method outperforms existing methods in terms of both unweighted F1 score (UF1) and unweighted average recall (UAR), which are 1.87% and 2.61% higher than the current best methods, respectively. Meanwhile, experiments conducted on CAS(ME)[3] and DFME datasets also demonstrate the wide applicability and superior performance of the proposed method across different scenes and subjects. In addition, the effectiveness of the bidirectional optical flow and wavelet-based attention downsampling modules in the feature representation process is verified by Grad-CAM visualization analysis, which shows that the model pays more attention to facial regions related to muscle activity and suppresses irrelevant background noise. The previous version of this algorithm has achieved the second place in the automatic micro-expression recognition task at the 4th Chinese Conference on Affective Computing (CCAC2024). The code has been made publicly available at https://github.com/LeoDerekh/BOWAM.

Keywords micro-expression recognition; bidirectional optical flow; wavelet transform; attention mechanism; frequency domain feature

1 引 言

面部表情在人类沟通中起着至关重要的作用，通过体验、行为和生理因素的复杂互动揭示隐藏的情感[1]。通常，这些表情可以分为宏表情和微表情。

其中微表情是细微且不自觉的快速动作，仅持续5毫秒到0.5秒[2]。虽然持续时间短且强度低，但它们提供了对真实情感的重要洞察，常常隐藏在有意识表情背后。识别这些表情具有挑战性，通常需要专家分析。

由于这些特性，微表情在心理学、人机交互和安全领域有重要应用[3]。例如，它们可以帮助心理学家研究情感、提升人与计算机的交互体验，以及在法律审讯中识别谎言。

在微表情识别任务中，传统方法[4-6]通常利用起始帧和峰值帧之间的单向光流来捕捉面部的细微变化。这些方法依赖于光流的局部特征（如局部运动模式和纹理变化）来识别表情的微小变化。传统光流方法在微表情识别中易受噪声干扰，尤其在头部姿态变化下难以区分面部肌肉运动与头部位移引起的像素变化，这可能会干扰模型对鲁棒性特征的捕获。现有光流增强算法已尝试从多角度改善这一问题，如Liong等人[4]提出双加权定向光流（Bi-WOOF）特征提取器，通过编码光流方向、大小和光流应变（Optical Strain）来强调表情变化；Gan等人[7]结合光流特征与卷积神经网络，通过两阶段特征融合增强表情细节；Liu等人[8]提出一种稀疏主方向平均光流（MDMO）方法，利用流形结构提升特征辨别能力。然而，这些方法仍存在明显局限，即仅计算起始到峰值帧的单向运动，忽略了微表情从峰值到结束帧的动态恢复过程。这种单向处理方式导致时序信息利用不充分，且难以消除头部姿态变化引入的全局噪声。因此，本文提出双向光流方法，通过时序对称性建模，以提升光流特征的鲁棒性和判别力。

对于增强微表情识别鲁棒性的特征提取方法，频域特征的引入为捕捉面部动态变化提供了新的视角。Tao等人[9]提出了一种用于自然场景动态面部表情识别任务的Freq-HD方法。它包含两个核心模块：空间-时间频率分析（STFA）模块和多频段互补选择（MBC）模块。STFA模块负责从视频中提取片段的空间-时间频率特征，并计算出每个片段的动态值。而MBC模块对这些动态值进行分析，以修正由于非表情噪声引起的不适当反应，从而识别出与表情变化相关的片段。通过这种方法，Freq-HD能够有效地从复杂的自然场景视频中筛选出富含表情变化的关键片段。

尽管频域特征在捕捉微表情的细微高频变化方面具有独特的优势，相对于传统的频域特征提取方法，此类方法在挖掘信号的时空特征方面存在局限。具体表现在其无法充分捕获信号在空间维度上的位置依赖关系和结构特征。近年来，深度学习在视觉任务中的突破性进展[10-14]为微表情识别提供了新的思路。注意力机制作为一种重要的深度学习技术，通过聚焦于图像中的关键区域，能提高特征提取的精度和效率[15-19]。例如，Li等人[15]提出了一种基于双分支网络的框架，主分支通过连续注意力（CA）模块提取运动模式特征，子分支通过基于Vision Transformer（ViT）的位置校准（PC）模块生成面部位置嵌入，最终融合两种特征用于微表情分类。然而，这些基于深度学习的工作大多集中于空间域或时间域的特征提取，忽略了频域信息在微表情识别中的潜在作用。而频域信息能够捕捉微表情的细微高频变化，如眼角的轻微抽动和嘴角的微小上扬等，这些高频细节揭示了真实情感的自然流露。同时，低频信息提供了整体面部结构和表情轮廓的基础信息，为高频信息的分析提供了驱动和参考。因此，将注意力机制与频域特征相结合，可以更全面地提取微表情特征。

然而，频域信息的有效利用还面临着传统下采样方法导致的信息损失问题。频域信息的重要性在于微表情的细微差异往往体现在高频细节中，而传统卷积神经网络（CNN）中广泛使用的最大池化（Max-Pooling）方法尽管在降低计算复杂度和增强特征不变性方面表现出色，却可能导致这些关键高频细节的丢失[20]。特别是在复杂的表情分析任务中，这种信息丢失可能导致模型对微表情特征的语义理解能力产生显著下降。因此，为进一步提升模型对高频细节的捕获能力，更有效地利用高频信息和低频信息，同时有效缓解传统下采样方法导致的信息损失问题，我们提出了小波注意力下采样模块。该模块不仅能够有效保留细节信息，还能够通

过频域分析增强模型对微表情的辨别能力。

总结来说，本文的贡献如下。

（1）提出了一种新的双向光流方法，旨在缓解传统光流方法存在的噪声大和微表情视频时序信息利用不充分等问题。

（2）提出了小波注意力下采样模块，其结合了小波变换和注意力机制的优势，能够在频域中更精细地处理高频和低频信息。通过小波变换分解图像特征，该模块能够有效捕捉和保留微表情中的细微高频细节，同时利用注意力机制增强关键区域的特征表示，减少传统下采样方法带来的信息丢失。

（3）本文方法在MEGC 2019数据集上与当前最好的方法（CVPR 2023）相比，UF1和UAR分别提高了1.87%和2.61%。

2　相关工作

2.1　视觉情感计算领域的研究进展

视觉情感计算旨在理解和建模人类在视觉信息中的情绪感知与表达机制。近年来，研究在大规模数据集构建、情绪分布建模、多模态融合与情感可解释性方面均取得显著进展。

Zhao等人[21]回顾了视觉情感图像分析的研究，指出其已由手工特征方法发展为融合深度语义的情感特征学习，未来应强化上下文理解与观者－图像交互。Pei等人[22]系统分析了1997-2023年全球情感计算研究，指出该领域正由“情绪识别”向“情绪理解与生成”转变，重点方向包括多模态融合与人机共情交互等。

在数据与解释性建模方面，Achlioptas等人[23]提出Affection框架，使模型能够从视觉数据中学习可解释的情感推理过程，并生成自然语言层面的情绪解释，从“感知”走向“理解”。Yang等人[24]构建了EmoSet数据集，提供了百万级样本与多层次语义标签，在多文化、多场景条件下显著提升了模型的泛化性能。

在多模态情绪推理方面，Haydarov等人[25]提出Affective Visual Dialog（AVD），首次将视觉情感理解与多轮对话情绪推理结合，使模型在图像语境中实现共情式情感响应。Wu等人[26]通过跨模态知识迁移，利用大规模图文对实现语义对齐，有效缓解了“情感缺口（Affective Gap）”问题。

在情绪分布建模方面，Xu等人[27]提出MFR-Net，基于情感分布学习（Emotion Distribution Learning，EDL）引入特征精炼机制，实现情绪标签从离散到连续的过渡。而在情感生成与表达控制方向，Wang等人[28]的InstructAvatar框架则将视觉情感识别与生成式建模结合，通过文本指令驱动虚拟角色的表情与动作，实现从语义到视觉的情绪生成控制。

总体来看，视觉情感计算已由单模态静态识别发展为多模态语义理解与可解释生成，但情绪建模仍偏宏观，难以刻画不同时间尺度的细微差异。研究重心因此转向宏表情与微表情的分层建模与动态分析，为多层次表情识别奠定基础。

2.2　表情识别领域的研究进展

表情识别是视觉情感计算的重要分支，旨在通过分析面部纹理变化与动态运动模式来识别个体的情绪状态。根据表情的持续时间与强度，可分为宏表情识别与微表情识别两类。前者面向明显且持续时间较长的表情变化，后者则聚焦于短暂、细微的情绪泄露。但两者的性能都随着深度学习的发展得到了显著的提升。

宏表情识别　在结构设计方面，Zheng等人[29]提出基于金字塔交叉融合结构的Transformer网络（POSTER），在多层特征间实现语义一致性与全局依赖建模。Wu等人[30]提出Patch-Aware表征学习，通过局部补丁注意力机制增强显著区域（眼、眉、嘴角等）的特征表达，提升了遮挡与光照扰动下的鲁棒性。

在表征学习方面，Xie等人[31]提出跨层对比学习框架，通过层间正负样本约束增强语义一致性；Zhang等人[32]提出通用特征提取框架，引入域不变编码器与样本均衡策略以提升跨人群与跨场景的稳定性。Wang等人[33]提出四元组交叉相似性（Quadruplet Cross Similarity）机制，联合建模类内紧致性与类间分离性，实现更精确的特征区分。

在时序特征建模方面，Li等人[34]提出PTH-Net，突破了传统表情识别依赖人脸对齐的限制，通过时空堆叠与注意力机制自适应捕捉面部局部区域的细微动态变化；Zhang等人[35]首次将社会心理学中的动态刻板印象理论（Dynamic Stereotype Theory）引入微表情识别，构建了基于定向形变的模型，用以建模不同人群的面部运动规律。

微表情识别　与宏表情识别相比，微表情识别需要在极短时间内捕捉幅度微小的面部运动特征，因此对模型的时空敏感性和特征判别性提出了更高要求。

Chen等人[36]设计块划分卷积网络，通过面部区域划分与隐式特征增强机制强化了局部肌肉运动与纹理变化的动态表示。Zhou等人[37]提出Inceptr网络，将Inception多尺度结构与CBAM注意力及Vision Transformer结合，以同时捕获面部区域的细粒度纹理变化与跨区域的全局语义依赖关系。Nguyen等人[38]提出Micron-BERT模型，引入基于Transformer的帧级上下文建模，显著提升跨帧依赖的表达能力。Wang等人[39]的HTNet通过层次化时序特征提取与注意力门控结构，有效建模表情演化过程。Ma等人[40]的双分支运动网络（EDMDBN）融合全局运动趋势与局部细节变化，实现表情运动特征的互补建模。

现有研究在模型算法设计与泛化能力提升方面取得了显著进展，且多数方法能对空间域的静态特征进行有效建模，但这些方法在时间维度上对细微动态与跨帧一致性的捕捉仍存在不足，其限制了毫秒级情绪变化的识别精度。

因此，当前研究开始引入能表征像素级运动变化的特征表示，例如光流特征，以便更精准地捕捉微小肌肉位移与动态模式。

2.3 表情识别中的光流

光流[41]（Optical Flow）是指在连续图像序列中，描述物体表面像素点运动的向量场，它捕捉了像素点随时间的运动信息。其经常用于检测和描述图像序列中物体表面像素运动的模式，已广泛应用于运动检测、对象跟踪等任务。

传统的光流估计方法可分为稠密光流和稀疏光流。稠密光流，为图像中每个像素点计算光流向量，提供详尽的运动信息；而稀疏光流仅在关键点或特征点上计算光流，以减少计算量。其中，Lucas-Kanade算法[42]，是一种基于特征点的稀疏光流估计方法，特别适用于跟踪图像序列中的特征点，因其计算效率高而在实时应用中广受欢迎。而在稠密光流算法，Farneback算法[43]通过多项式展开对图像区域进行近似建模，并利用相邻帧间的全局运动信息计算每个像素的光流向量，能够生成连续且完整的光流场，但相对需要更高的计算资源。

随着深度学习技术的发展，一些基于此技术的光流估计算法[44-46]因其强大的特征提取能力和对复杂场景的适应性而成为研究热点。光流法能够捕获像素级的偏移，因此在微表情识别领域得到了广泛应用。例如，Liong等人[4]提出了一种仅利用视频中的峰值帧和起始帧进行微表情识别的方法，并引入了双加权定向光流（Bi-WOOF）特征提取器来突出表情变化，通过编码光流的方向、大小和光流应变来实现。Liong等人[5]提出了一种浅层三流三维卷积神经网络，它在保持高计算效率的同时，能够从光流应变、水平和垂直光流信息中提取微表情关键特征。Gan等人[7]提出了一种基于两个关键帧（起始帧和峰值帧）的微表情识别方法，通过结合光流特征和卷积神经网络，有效地提取和增强表情细节以改进情绪状态的预测。Liu等人[8]提出了一种稀疏主方向平均光流（MDMO）特征提取方法，该方法基于数据点的稀疏性来定义新的距离度量以揭示流形结构，利用无监督学习的稀疏编码技术，只需少量训练数据，并在保持特征紧凑性的同时提高辨别能力。值得注意的是，当前研究正致力于通过时空特征融合和注意力机制进一步提升光流特征的表征能力.

上述方法突出了准确捕捉面部细微变化对微表情识别的重要性。然而，传统光流特征提取方法仅关注起始到峰值帧的单向光流信息，忽视了微表情动态过程的对称性以及起始帧和结束帧之间的信息互补性。这使得传统光流方法在处理细微表情时易受噪声干扰，而这些噪声可能源于环境变化、面部特征微小移动或算法对细微变化的敏感度不足。为克服这些挑战，我们提出了一种双向光流方法，即通过改进特征表征以减少噪声影响，更精确捕捉面部微小变化。

2.4 频域特征与注意力机制的融合

频域特征。低频信息在图像处理中指的是图像中大尺度且变化缓慢的部分，通常涵盖了图像的主要结构和整体形状。在微表情识别领域，这类信息极为重要，主要包括整体面部结构、表情的基本形态，以及光线和阴影的变化。

如图1所示，整体面部结构主要是指面部的基础形状和重要特征如眼睛、鼻子、嘴巴的位置。表情的基本形态则涉及面部表情所展示的基本情绪，比如快乐、悲伤或愤怒。此外，光线和阴影的分布有助于揭示面部的三维结构，对于准确解读和理解微表情具有重要影响。高频信息在微表情识别中起着至关重要的作用，因为它关联到图像中迅速变化且细节丰富的部分，如纹理、边缘及细小线条等，这些往往局限在特定的面部肌肉区域，有助于区分那些宏观上看起来非常相似的微表情。

微表情序列的分析可以通过傅里叶变换或小波变换将其转换到频域，从而提取与频率相关的特征（例如振幅和相位信息）。这种转换能够识别微表情

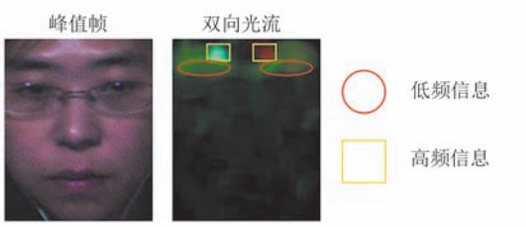


图1　微表情样本的高低频信息

中特定的局部几何特征，如面部轮廓的角点和线条，而这些特征通常容易被传统特征描述方法所忽略。Oh等人[47]采用Riesz小波变换提取了图像的幅度、相位和方向信息，并通过傅里叶变换进一步揭示微表情的二维局部结构(i2D)。Oh等人[48]利用泊松拉普拉斯(LOP)带通滤波器结合高阶Riesz变换，以恢复i2D的相位和方向特征，并从中提取LBP-TOP特征及其相应的特征直方图。这种方法的优势在于能够捕捉传统方法容易遗漏的局部复杂结构，例如面部的细节轮廓。与i2D类似，i1D[49]特征是通过一阶Riesz变换提取的，用于进一步分析微表情中的方向信息。此外，Zhang等人[50]采用Gabor滤波器提取微表情的关键频率纹理特征，并有效地抑制其它不相关的纹理信息，从而实现对方向统计结构的深度挖掘。这些方法表明，频域分析与特征提取技术的结合能够有效挖掘微表情的复杂局部结构信息。

基于注意力机制的微表情识别方法。一些工作将注意力机制[51]拓展到视觉识别任务中[52-53]。对于微表情识别，Chen等人[36]提出块划分卷积网络(BDCNN)，基于起始帧与峰值帧计算的四种光流特征进行学习，将图像划分为小块并依次执行卷积与池化，同时在深特征空间采用改进的隐式语义数据增强以缓解小样本问题。Wang等人[17]提出了新型注意力机制，其与残差网络结合，无需显著增加参数即可提高微表情识别的准确率，并关注面部特定区域的动作单元。Li等人[18]提出了一种结合空间和通道注意力的微表情动作单元检测方法，通过高阶统计量捕捉面部区域间的空间关系和通道间特征变化，提高在有限样本上的检测鲁棒性，并采用多任务学习框架进行AU检测。Zhou等人[19]提出了Dual-ATME框架，其结合基于先验知识的手工特征区域选择(HARS)和基于注意力机制的自动特征区域选择(AARS)，从而提取微表情的多尺度特征。

然而，现有基于注意力机制的方法仍存在两个主要局限：首先，这些方法在处理动作程度较弱的微表情时，面临关键特征定位不准确的问题；其次，它们未深入探讨频域中高频和低频信息在微表情识别中的功能区别。特别是对于下采样过程，高频细节信息的丢失会导致模型对微表情细微变化的捕获能力下降[20]，从而影响模型的整体泛化能力。

为此，我们提出了小波注意力下采样模块，以缓解上述问题。

3　本文方法

本节详细介绍基于双向光流和小波注意力下采样的微表情识别模型，该模型是一种“捕获-增强”的解决方案。双向光流模块基于微表情的时序演变(起始帧→峰值帧→结束帧)，从正反两个时间维度提取完整运动信息，实现有效捕获。频域变换模块则通过小波变换分解光流特征，并结合注意力机制，针对性增强微表情特征，同时抑制噪声和无关信息，完成特征增强。二者协同作用，形成“捕获→分解→增强→整合”的处理流程。整体框架如图2所示。

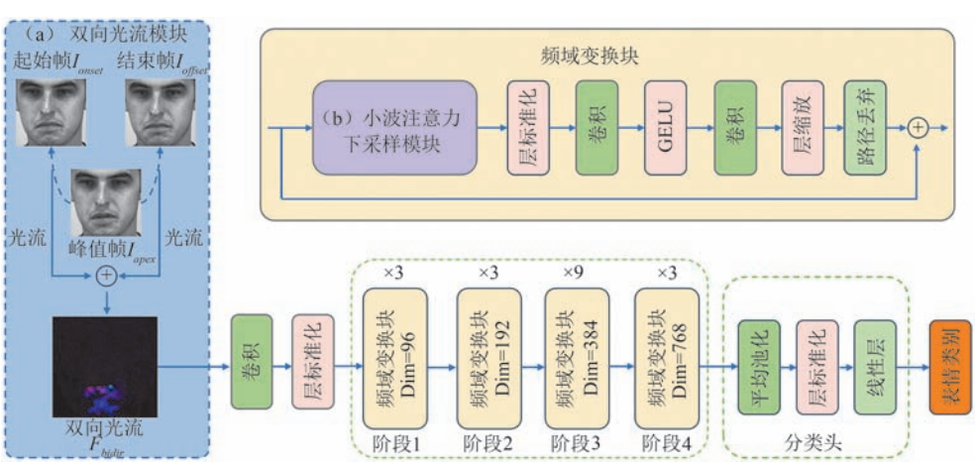


图2　本文模型整体框架图

具体来说，对于微表情识别，传统方法仅计算起始到峰值帧的单向光流，忽略了微表情动态过程的对称性以及起始和结束帧之间的信息互补性。这种方法难以区分面部肌肉运动与头部位移引起的噪声，导致特征提取的不完整。为了克服这些局限性，我们采用双向光流模块，利用起始帧、峰值帧和结束帧构建完整的双向光流特征。这一方法较完整地捕捉了微表情的动态过程，增强与微表情相关的面部肌肉运动特征，并通过方向相反的光流相互抵消头部位移引入的噪声；进而通过四个阶段，分别对应四个频域变换块（Frequency Domain Transform Block，FDTB），以增强高低频信息的特征表示，其中每个块主要包含本文构建的小波注意力下采样模块（在3.2节详述）等；最后经过分类头得到微表情类别。

3.1 双向光流

3.1.1 动 机

对于微表情识别，研究人员希望通过光流获取视频序列帧与帧之间的细微像素级变化。然而，通过从起始到峰值帧提取的光流数据往往会受到噪声的影响，特别是在存在头部姿态变化的情况下。这些噪声可能源于光照变化、背景运动或摄像头抖动等因素。

头部姿态变化对光流的影响尤为显著，因为它会在图像中引入大范围的像素移动，使得算法难以分辨这些变化是由面部肌肉运动还是头部位置变化引起的。

为了缓解这些问题，研究人员结合面部关键点检测技术，通过跟踪和对齐面部关键点来校正头部姿态的影响。此外，他们也使用更加鲁棒的光流算法，如稀疏光流或基于深度学习的光流估计方法，从而帮助减少噪声干扰。

本文研究特别关注面部肌肉运动的对称性。考虑到微表情通常表现为快速的面部肌肉运动，这些运动会在达到峰值后迅速恢复到初始状态。因此，与微表情分析相关的面部肌肉运动的检测与识别，成为了研究的重点。

然而，与这些快速面部运动不同，头部姿态的变化通常具有单向性和较长的持续时间，这种姿态变化可能会混淆微表情的检测和分析。因此，为了减少头部姿态变化对微表情检测的干扰，本文提出了一种双向光流方法，利用峰值帧分别到起始帧和结束帧的光流信息，增强与微表情相关的局部肌肉运动特征，并通过方向相反的光流相互抵消头部位移引入的噪声，从而更准确地捕捉微表情特征。

如图3所示，每一行代表一个微表情片段的样本，第1至第5列分别展示了起始帧、峰值帧、结束帧、传统光流和双向光流的情况。与传统光流相比，双向光流的噪声有明显减少，光流信息更加集中于眉毛、眼睛和嘴角等关键局部区域。而传统光流的信息则在整个面部均有分布，可能会包含更多与微表情无关的信息。

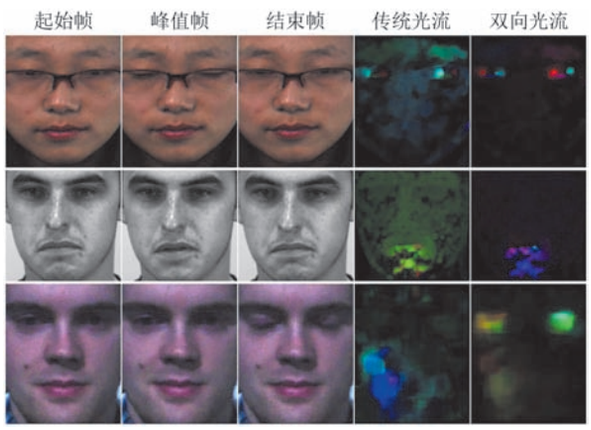


图3 传统光流和双向光流的对比。传统光流的噪声较大，光流在整个面部均有分布；双向光流的噪声显著减少，光流更加集中于眉毛、眼睛和嘴角等关键局部区域。

3.1.2 方 法

如图4所示，传统方法提取起始帧到峰值帧的光流，而本文方法提取峰值帧到起始帧以及结束帧的光流，并将这两组光流相加作为模型的输入。具体来说，设I_{onset}、I_{apex}和I_{offset}分别表示起始帧、峰值帧和结束帧，传统光流和双向光流的计算方法如下所示：

$$\begin{cases} F_{trad}=Farneback(I_{onset},I_{apex}) \\ F_{bidir}=Farneback(I_{apex},I_{onset})+ \\ \qquad Farneback(I_{apex},I_{offset}) \end{cases} \tag{1}$$

其中，F_{trad}、F_{bidir}分别表示传统光流和双向光流，$Farneback(prev,next)$表示用Farneback算法[43]计算前一帧（prev）和后一帧（next）之间的光流。

这种处理方式能够强化与微表情相关的面部肌肉运动，同时有效地抵消头部姿态变化带来的影响，从而更准确地捕捉到微表情特征。这主要归因于其独特的信号处理特性，根据图3所示的实验结果，传统光流在整个面部区域上有较大的干扰信号，而双向光流展现出明显的噪声抑制效果；同时，双向光流信息更加集中于眉毛、眼睛和嘴角等微表情最常出现的关键局部区域，突出了真正的微表情信号。此

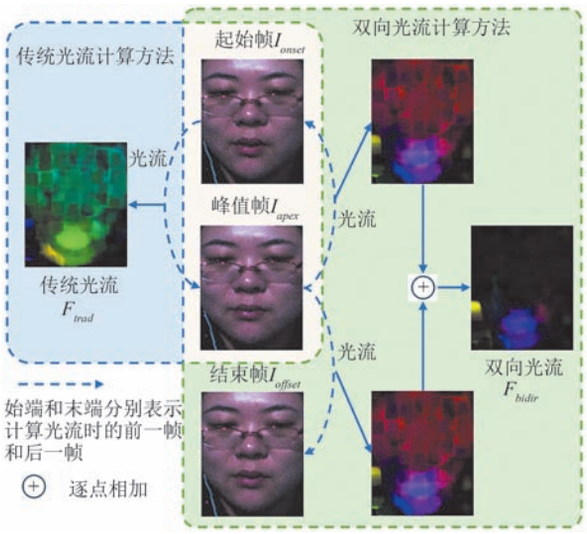


图4　双向光流与传统光流的计算方法。与传统光流计算方法相比，双向光流计算方法引入了结束帧，并且利用了峰值帧分别到起始帧和结束帧之间光流的和。

外，双向光流同时考虑微表情从峰值帧到起始帧的反向运动和从峰值帧到结束帧的正向运动，双向光流创建了一个更完整的运动表示，能够捕捉微表情整个生命周期的运动一致性，从而区分出真正的微表情运动和非微表情的头部姿态变化。

为确保微表情动态的完整性，我们选取微表情标准序列中已标注的终止帧作为结束帧，该帧对应表情完全消退、面部肌肉恢复至中性状态的时刻。由于其代表着从峰值状态回落的过程，因此在光流轨迹上呈现与起始帧至峰值帧相反的运动趋势。例如，若表情表现为眉毛上扬，则起始帧至峰值帧为向上运动，峰值帧至结束帧则为向下回落。

尽管如图4所示，双向光流（即峰值帧与起始帧、峰值帧与结束帧之间）均存在一定噪声，但这并不削弱其结构有效性。微表情本身具有幅度小、变化局部、持续时间短的特点，其运动信号常被噪声掩盖，但正因如此，其前后运动轨迹在空间上通常表现出区域一致、方向相反的对称性。相对而言，头动、光照变化等非表情噪声缺乏这种对称性，往往在双向光流中呈现为随机或不一致模式。

基于此特性，我们提出双向光流融合策略：通过增强区域一致、方向相反的真实表情运动，抑制缺乏一致性的背景噪声，从而提高整体信噪比，突出有效表情区域（如眉毛、眼角、嘴角等）的动态特征。

3.2　小波注意力下采样模块

3.2.1　动　机

在微表情识别中，高频信息对捕捉细微差异至关重要。传统最大池化方法虽降低了计算复杂度并增强了特征不变性，但可能导致高频细节丢失。为此，我们引入了小波注意力下采样模块，结合小波变换的频域分析能力和注意力机制的特征选择优势，在降维的同时保留关键高频细节，提升模型对微表情的辨别能力。

高频细节捕获特征图对比。如图5(a)所示，基线模型的特征图激活区域分散且模糊，难以聚焦于眼角和嘴角等关键部位。传统卷积神经网络(CNN)[54,55]常忽略高频细节，影响对微表情的敏感度。图5(b)展示了加入空间和通道注意力机制后的特征图，激活区域更加集中。空间注意力使模型聚焦于关键区域（如眼角、嘴角），通道注意力则强化了最具辨识力的特征通道。

图5　基于小波注意力的高频细节捕获特征图对比

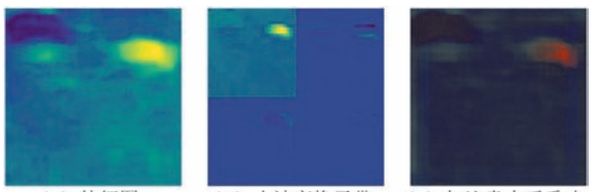
尽管如此，注意力机制在高频细节捕捉上仍显不足。图5(c)显示，仅使用逆小波变换时，关键区域边缘信息清晰，激活区域精准地匹配了微表情的关键特征区域（如眼角和嘴角的细微动作）。小波变换通过分解高频和低频子带，显著提升了对高频信息（如眼角细微运动、嘴角弯曲变化）的捕捉能力，弥补了基线模型频域信息的不足。

图5(d)展示了结合空间与通道注意力机制及逆小波变换后的特征图，其激活区域最清晰、集中，细节更突出。该方法融合了小波变换的高频和低频捕捉能力，并通过注意力机制进一步强化关键区域的聚焦能力。

小波注意力下采样特征处理。小波变换通过将图像分解为不同频率层次，精准捕捉微表情特征，分离高频和低频信息，对识别眼角、嘴角的细微动作尤为重要。多尺度分析使模型能聚焦于细微变化。在微表情识别中，眼睛、嘴角等关键特征的精细分析至关重要。若缺乏注意力机制引导特征聚焦，容易丢失关键微表情细节（如眼部肌肉颤动、嘴角抽动）；同时，通道间交互未能有效建模，导致难以凸显高判别力特征。

结合空间注意力机制，模型可重点关注人脸关键区域，通过动态调整对不同空间位置的关注程度，确保眼睛、嘴角等区域得到更好表征，提高敏感度并减少无关信息干扰。通道注意力机制则通过赋予更具表达力的通道更高权重，进一步优化特征选择。

图6展示了小波注意力下采样模块的流程。图6(a)为原始特征图，包含未分离的低频和高频信息。图6(b)通过小波变换分解为低频和高频子带，低频保留结构信息，高频捕获边缘和纹理细节，其对微表情识别的关键区域（如眼角、嘴角）很重要。图6(c)为加注意力后重建的特征图，其中“重建”是指在小波注意力模块中，首先对输入特征图进行小波分解，提取不同频率的子带信息，随后通过逆小波变换将处理后的频域信息还原至空间域，从而生成与原输入同尺度的特征图。具体地，空间注意力作用于低频分量，增强关键区域分析能力；通道注意力作用于高频分量，突出重要特征。经注意力融合和逆小波变换重建后，使得关键区域细节更加清晰。图6(c)所体现的正是该模块在信息重构和增强方面的优势：通过小波域建模结构细节与全局语义之间的互补关系，并在重建过程中强化关键特征，实现更优的表征能力。



（a）特征图 （b）小波变换子带 （c）加注意力后重建

图6 小波注意力下采样特征处理示意图

3.2.2 方 法

小波注意力下采样模块（如图7所示）的流程如下。

首先，通过小波变换将特征图分解为低频信息（X_{LL}）、高频信息（X_{LH}、X_{HL}、X_{HH}），其中L代表“低频”（Low frequency），H代表“高频”（High frequency），两个字母的组合则表示在水平和垂直两个方向上的频率特性。例如，X_{LL}表示水平和垂直方向都是低频的子带，包含了图像的近似信息；X_{LH}表示水平方向低频、垂直方向高频的子带，主要包含水平边缘特征；X_{HL}表示水平方向高频、垂直方向低频的子带，主要包含垂直边缘特征；X_{HH}表示水平和垂直方向都是高频的子带，主要包含对角线方向的细节和纹理信息。这些不同子带捕获了微表情图像中不同方向和尺度的特征，通过注意力机制进行选择性增强，有助于捕获微表情变化的细微特征。

然后，将空间注意力机制应用于低频信息，以强化关键区域的图像特征；通道注意力机制应用于高频信息，帮助模型提取光流特征中的局部运动梯度和方向变化信息，这些信息直接对应于面部微小肌肉活动产生的速度场变化模式。

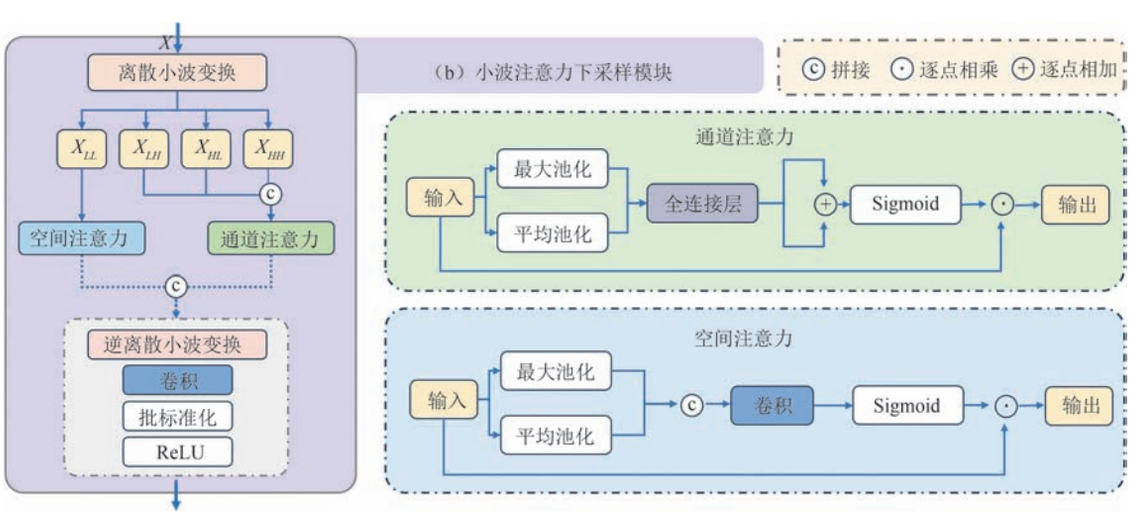


图7 小波注意力下采样模块

具体地，输入X首先通过离散小波变换（DWT）分解为四个子带分量：X_{LL}、X_{LH}、X_{HL}和X_{HH}。这种分解能够隔离出不同的频率信息，从而有助于捕获空间和通道的特征。具体如下所示：

$$X_{LL},X_{LH},X_{HL},X_{HH}=DWT(X) \tag{2}$$

对低频分量X_{LL}进行最大池化和平均池化，将两者结果拼接后通过卷积层处理。卷积输出经过非线性激活函数后与X_{LL}逐元素相乘，得到强化后的低频空间特征F_{spatial}，具体公式如下：

$$\begin{cases} X_{LL}^{\max}=\text{MaxPool}(X_{LL}),X_{LL}^{\text{avg}}=\text{AvgPool}(X_{LL}) \\ F_{\text{spatial}}=\sigma(\text{Conv}(\text{Concat}(X_{LL}^{\max},X_{LL}^{\text{avg}})))\odot X_{LL} \end{cases} \tag{3}$$

将高频分量X_{LH}、X_{HL}和X_{HH}拼接，形成高频特征矩阵X_H，以便模型利用高频信息增强对纹理和细节的捕捉能力，具体公式如下：

$$X_H=\text{Concat}(X_{LH},X_{HL},X_{HH}) \tag{4}$$

对高频特征矩阵X_H的最大池化和平均池化结果分别通过多层感知机（MLP）进行处理，得到通道特征。这些特征经过激活函数σ后与高频特征矩阵X_H逐元素相乘，得到强化后的高频通道特征F_{channel}，具体公式如下：

$$\begin{cases} X_H^{\max}=\text{MaxPool}(X_H),X_H^{\text{avg}}=\text{AvgPool}(X_H) \\ F_{\text{channel}}=\sigma(MLP(X_H^{\max})+MLP(X_H^{\text{avg}}))\odot X_H \end{cases} \tag{5}$$

最后，将提取的空间特征F_{spatial}和通道特征F_{channel}拼接，通过逆离散小波变换（$IDWT$）重建特征图。对重建后的特征图X_{idwt}施加卷积层、批归一化和ReLU激活函数，生成最终输出Y，具体公式如下：

$$\begin{cases} X_{idwt}=IDWT(\text{Concat}(F_{\text{spatial}},F_{\text{channel}})) \\ Y=\text{ReLU}(\text{BatchNorm}(Conv(X_{idwt}))) \end{cases} \tag{6}$$

在训练过程中，采用标准交叉熵损失函数引导模型优化。基于双向光流和小波注意力机制的微表情识别算法总结为算法1。

算法1.　本文算法的训练过程

输入：起始帧I_{onset}，峰值帧I_{apex}，结束帧I_{offset}

输出：

1. FOR 模型训练代数 DO
2. F_{bidir}＝双向光流模块$BOFM(I_{onset},I_{apex},I_{offset})$
3. F＝层标准化LN（卷积$Conv(F_{bidir})$）
4. FOR 阶段数 DO
5. $F\leftarrow$频域变换块$FDTB(F)$
6. END FOR
7. 预测表情类别$Pred$＝分类头$Classifier(F)$
8. $Loss=CrossEntropyLoss(Pred,$ 真实表情类别）
9. 反向传播$Backprop(Loss)$
10. END FOR

4　实验及结果分析

4.1　实验设计与设置

数据集。CASME II数据集[56]包含26名中国志愿者的247个微表情序列，采用200帧/秒、640×480分辨率的高速摄像机拍摄。每个序列标注了起始帧、峰值帧、结束帧、情感类别（如快乐、厌恶、惊讶、压抑等）和动作单元（AUs）。

SAMM数据集[57]由32名不同种族背景的志愿者提供224个微表情序列，平均每位约7条。数据采集使用200帧/秒、2040×1088分辨率的灰度高速摄像机，标注信息包括起始帧、峰值帧、结束帧、情感标签（如生气、悲伤、恐惧等）和动作单元（AUs）。

SMIC数据集[58]由芬兰奥卢大学发布，包含HS、VIS和NIR三个子数据集。本研究选择HS子数据集，其包含164个样本，按情感类别分为积极（51个）、消极（66个）和惊讶（40个），拍摄帧率为100帧/秒，分辨率为1280×720。

CAS（ME）^{3}[59]数据集提供了1109个标注的微表情样本和3490个标注的宏表情样本，每个表情的标注信息包括7种情绪类别以及对应的面部动作单元（FAU），该数据集大约有80个小时的录像，分辨率为1280×720。

DFME数据集[54]是目前数量规模最大、采集帧率最高的动态自发微表情数据集，共有来自671名被试（656名实际有效）的7526个微表情视频样本，包含了七种情感类别：高兴、愤怒、鄙视、厌恶、恐惧、伤心和惊讶。采集帧率覆盖了200、300或500 fps。CCAC2024比赛中使用DFME中259名被试的2629个微表情视频样本。此外还提供了起始帧、峰值帧及结束帧的标注以及AU标签的标注。

预处理。在预处理阶段，我们对图像进行对齐、裁剪，并调整图像大小为240×240像素。为了增强数据多样性，我们应用了随机旋转、水平翻转、色彩调整和随机裁剪等数据增强操作。

评估标准。我们采用与MEGC 2019[55]相同的方法，对由CASME II、SAMM和SMIC数据集构成的复合数据集（MEGC2019数据集）进行分类性能评估。根据MEGC 2019的三类重新标记方案，将原始标签映射到新的标签空间：{积极、消极、惊讶}。具体映射关系为：消极（压抑、生气、鄙视、厌恶等）；积极（快乐）；惊讶（惊讶）。

实验结果通过留一法（LOSO）交叉验证进行报

告，每次保留一个受试者组作为测试集，其余用于训练。核心评估指标为未加权F1分数（UF1）和未加权平均召回率（UAR），用于评估分类器在不平衡数据集上的性能，公式如下：

$$UF1=\frac{1}{C}\sum_{i=1}^{C}\frac{2\times TP_i}{2\times TP_i+FP_i+FN_i} \tag{7}$$

$$UAR=\frac{1}{C}\sum_{i=1}^{C}\frac{TP_i}{N_i} \tag{8}$$

其中，C为样本类别总数，N_i表示第i类的样本总数，TP_i，FP_i，FN_i分别表示第i类样本的预测结果中真正例、假正例和假负例的样本数量。

实验设置。我们采用了Ubuntu 20.04操作系统，并在Python 3.8环境下使用一块NVIDIA Tesla V100显卡训练模型。训练周期设置为100，每个批次处理32个样本。初始学习率设定为0.0001，采用AdamW优化器进行权重更新，其中权重衰减系数设置为0.0001，并采用指数型的学习率衰减策略。此外，鉴于数据集中明显的类别不平衡问题，我们采取了不均衡采样策略，即通过调整不同类别数据的采样频率，来补偿少数类样本的不足，确保模型能够更公正地学习所有类别的特征。起始帧、峰值帧和结束帧的对齐方法与Zhou等人[60]一样，采用dlib工具库检测每一帧图像中的68个面部关键点，并基于关键点定位结果对人脸区域进行统一的裁剪和对齐。

4.2 实验结果与分析

4.2.1 与现有方法性能的比较

根据表1所示，本文模型在MEGC2019数据集上表现优异，UF1和UAR指标分别达到0.9090和0.9103，分别较现有最优方法μ-BERT高出1.87%和2.61%。在SAMM数据集上，本文模型UF1指标为0.8941，位列次优，与最优的MiMaNet方法（0.8960）仅相差0.19%，性能相近。在CASME II数据集上，本文模型UF1和UAR指标分别为0.9861和0.9924，较次优的SRMCL方法分别高出2.26%和2.75%，取得最优性能。总体来看，本文模型在所有数据集上均较大幅度优于基线模型，验证了其在微表情识别任务中的有效性和优势。

表1 与现有方法性能的比较（在MEGC2019数据集上评估）

方法	MEGC2019		SMIC		SAMM		CASME II	
	UF1	UAR	UF1	UAR	UF1	UAR	UF1	UAR
Dual-Inception[60]（FG 2019）	0.7322	0.7278	0.6645	0.6726	0.5868	0.5663	0.8621	0.8560
NMER[63]（FG 2019）	0.7885	0.7824	0.7461	0.7530	0.7754	0.7152	0.8293	0.8209
MTMNet[64]（ACM MM 2020）	0.8640	0.8570	<u>0.8640</u>	<u>0.8610</u>	0.8250	0.8190	0.8700	0.8720
FR[65]（PR 2021）	0.7838	0.7832	0.7011	0.7083	0.7372	0.7155	0.8915	0.8873
GRAPH-AU[66]（CVPR 2021）	0.7914	0.7933	0.7192	0.7215	0.7751	0.7890	0.8798	0.8710
MiMaNet[61]（IJCAI 2021）	0.8830	0.8760	**0.8730**	**0.8670**	**0.8960**	<u>0.8840</u>	0.8810	0.8810
BDCNN[36]（TMM 2022）	0.8509	0.8500	0.7859	0.7869	0.8186	0.7994	0.9501	0.9516
IncepTR[37]（Multim. Syst. 2023）	0.7530	0.7460	0.6550	0.6500	0.6910	0.6940	0.9110	0.8960
FRL-DGT[62]（CVPR 2023）	0.8120	0.8110	0.7430	0.7490	0.7720	0.7580	0.9190	0.9030
μ-BERT[38]（CVPR 2023）	<u>0.8903</u>	0.8842	-	-	-	-	-	-
LAENet[67]（VISUAL COMPUT 2024）	0.7568	0.7405	0.6620	0.6523	0.6814	0.6620	0.9101	0.9119
MFDAN[68]（TCSVT 2024）	0.8453	0.8688	0.6815	0.7043	0.7871	0.8196	0.9134	0.9236
HTNet[39]（Neuro. 2024）	0.8603	0.8475	0.8049	0.7905	0.8131	0.8124	0.9532	0.9516
SRMCL[69]（Affective Comput. 2024）	0.8630	0.8830	0.7946	0.8053	0.8470	**0.8866**	<u>0.9635</u>	<u>0.9649</u>
EDMDBN[40]（Pattern Recognit. Lett. 2025）	0.8821	<u>0.8933</u>	0.7948	0.8085	0.8336	0.8661	0.9484	0.9619
基线模型[70]	0.8580	0.8689	0.8011	0.8095	0.8621	0.8697	0.9212	0.9298
本文模型	**0.9090**	**0.9103**	0.8529	0.8573	<u>0.8941</u>	0.8714	**0.9861**	**0.9924**

注：粗体表示最优数据，下划线对应表现第二数据。

尽管本文方法在MEGC2019和CASME II数据集上取得了优异性能，展现出对标准微表情动态特征的良好建模能力，但在SMIC和SAMM数据集上的表现相对逊色。我们认为，性能差异主要源于以下几点。

（1）样本规模不足：SMIC和SAMM样本量偏少，限制了模型学习能力，影响泛化性能[58]；

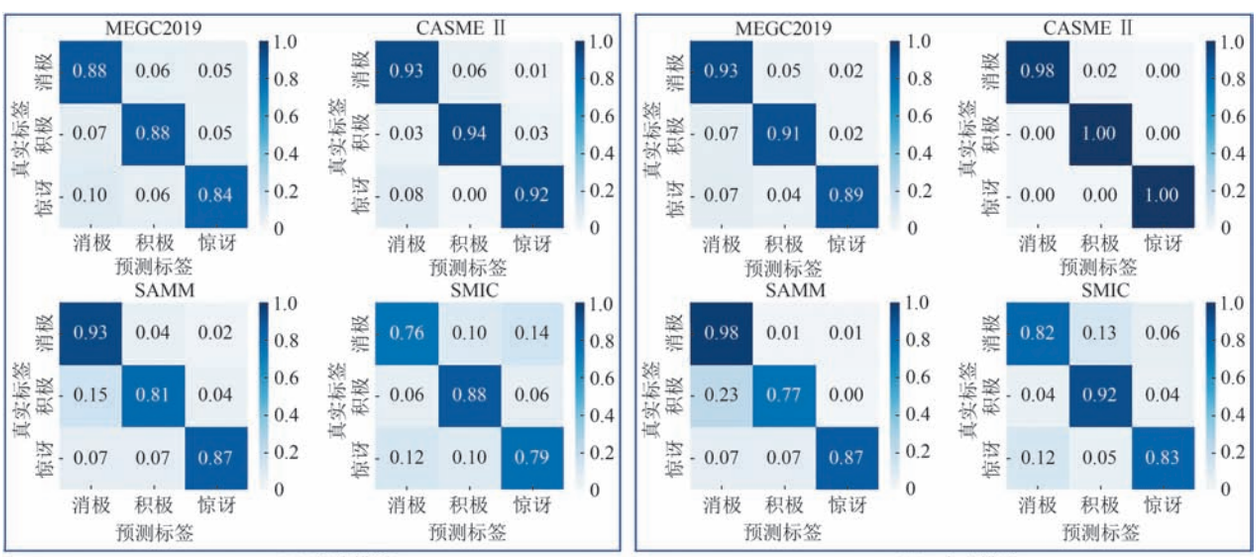
（2）标注一致性不理想：数据中存在较高的主观性与噪声，干扰关键特征提取[57-58]；

（3）场景比较复杂：包含非典型微表情与背景干扰，增加识别难度[55]；

（4）缺乏特定优化：现有在该类数据集上表现优异的方法多采用数据集特定设计[61-62]，而本文更强调通用性建模，未针对极端场景进行额外调整。

在CASME II数据集上，我们获得了超过0.98甚至0.99的高精度结果，这与数据集的固有限制密切相关，如样本量较少、标注一致性高和特征分布集中。这些因素使得模型更容易适应该数据集的模式。尽管如此，所有实验均采用留一法（LOSO）交叉验证，确保测试样本与训练样本完全隔离，有效避免了数据泄露问题。同时，为确保公平性，所有对比方法均采用与Nguyen和Ma等人的工作[38-40]相同的实验设置。因此，我们认为本研究结果在方法有效性验证方面具有重要参考意义。此外，我们的方法在CAS（ME）^{3}和DFME等更具挑战性的数据集上同样展现了优异性能，进一步验证了其稳定性与优良的泛化能力。

基线模型在不同数据集上的混淆矩阵如图8(a)所示，本文模型在不同数据集上的混淆矩阵如图8(b)所示，对比(a)和(b)可知本文模型在大多数数据集和类别上都实现了性能提升。



（a）基线模型　　（b）本文模型

图8　基线与本文模型在MEGC2019、CASME II、SAMM和SMIC数据集上的混淆矩阵

表2展示了在CAS（ME）^{3}数据集上不同方法的微表情识别性能对比。在三分类、四分类和七分类任务中，本文提出的模型均取得了显著优于现有方法的性能。具体而言，本文模型在三分类任务中UF1和UAR分别达到0.8078和0.7998，分别相比现有最优方法HTNet提升了23.11%和25.83%；在四分类任务中分别达到0.6698和0.6781，相比最优方法μ-BERT分别提升了19.80%和18.68%；在七分类任务中分别达到0.5019和0.5031，相比最优方法μ-BERT分别提升了17.55%和17.77%。这些结果表明，本文方法在不同复杂度的任务中均展现出优越的识别能力和鲁棒性，验证了双向光流模块和小波注意力下采样模块在微表情特征提取和表征中的有效性。

表3展示了在DFME数据集（CCAC2024比赛B榜）上，本文提出的微表情识别模型与现有方法的性能对比。其中，第二名方法为本文的早期工作，第一名、第二名和第三名的工作均尚未正式发表。从表中可以看出，本文模型在UF1指标上取得了第二优异的成绩（0.3921），仅比当前第一名方法

表2　与现有方法性能的比较（在CAS(ME)3数据集上评估）

方法	表情种类数	UF1	UAR
FR[65]	3	0.3493	0.3413
μ-BERT[38]	3	0.5604	0.6125
HTNet[39]	3	0.5767	0.5415
本文模型	3	**0.8078**	**0.7998**
AlexNet[59]	4	0.2915	0.2910
AlexNet+Depth[59]	4	0.3001	0.2982
SFAMNet[71]	4	0.4462	0.4797
μ-BERT[38]	4	0.4718	0.4913
本文模型	4	**0.6698**	**0.6781**
AlexNet[59]	7	0.1759	0.1801
AlexNet+Depth[59]	7	0.1773	0.1829
μ-BERT[38]	7	0.3264	0.3254
本文模型	7	**0.5019**	**0.5031**

注：粗体表示最优数据

(0.4016)低0.0095,同时在UAR指标上实现了所有方法中的最优结果(0.4124),体现了更好的类别均衡性能。特别是与传统FR方法相比,本文模型在UF1和UAR上分别提升了0.1046和0.0896,显著增强了微表情识别的准确性和鲁棒性。

表3 与现有方法性能的比较(在DFME数据集(CCAC2024比赛B榜)上评估)

方法	UF1	UAR
FR[65]	0.2875	0.3228
第一名方法	**0.4016**	0.4008
第二名方法	0.3534	0.3661
第三名方法	0.3356	0.3550
本文模型	0.3921	**0.4124**

注:粗体表示最优数据

4.2.2 消融实验

表4展示了在MEGC2019数据集上的消融实验结果,评估了不同组件对模型性能的影响。该实验结果表明本文构建的各组件都能一定程度地带来性能的提升,同时也验证了各个组件的必要性。

表4 消融实验结果

实验	双向光流	小波变换	空间	通道	UF1	UAR
A	→传统光流	×	×	×	0.8580	0.8689
B	√	×	×	×	0.8762	0.8709
C	√	√	×	×	0.8802	0.8873
D	√	√	×	√	0.8966	0.9043
E	√	√	√	×	0.8993	0.9077
F	√	√	√	√	**0.9090**	**0.9103**

注:"→X"表示用X替换相应的组件

光流融合方式对比实验。本文对比了不同光流融合方式的结果,最终选取逐元素加法的原因如下。

(1)减小过拟合风险:逐元素相加无需额外参数,避免了模型复杂度提升和过拟合风险,同时降低了计算成本,适合微表情数据噪声高、样本少的特点。

(2)互补性强:正向与反向光流在空间上对称,加法操作能有效整合互补信息,突出微表情相关特征,同时抵消头部运动噪声。

(3)性能对比:从表5实验结果(在MEGC2019数据集上评估)可以看出,逐元素加法操作的UF1和UAR指标分别为0.9090和0.9103,明显优于其它两种融合方式。相比之下,逐元素乘法操作对噪声敏感且容易抑制信号,导致性能下降;拼接操作虽然能够保留更多信息,但会增加参数量和训练时间,性能提升有限。因此,逐元素加法操作在性能和效率上更具优势,鲁棒性更强。

表5 光流融合方式对比实验结果

融合方式	UF1	UAR
逐元素加法	**0.9090**	**0.9103**
逐元素乘法	0.8646	0.8777
拼接	0.8706	0.8817

小波注意力下采样模块中全连接层设计方式对比实验。对于共享权重方式,从语义建模的角度看,权重共享有助于在不同空间位置间学习一致的通道重要性表示,强化通道间的全局依赖建模,从而提升模型的泛化能力。从模型结构的角度来看,结构上共享权重能够降低参数量和模型复杂度,有助于缓解微表情识别中因数据稀缺导致的过拟合问题,也有助于提升计算效率。同时表6中实验结果(在MEGC2019数据集上评估)表明,使用共享权重的MLP在UF1和UAR指标上均优于不共享权重的设计,性能分别提升了1.62%和1.01%。

表6 小波注意力下采样模块中全连接层设计方式的对比

设计方式	UF1	UAR
共享权重	**0.9090**	**0.9103**
不共享权重	0.8928	0.9002

小波注意力下采样模块中小波基函数对比实验。表7展示了小波注意力下采样模块中小波基函数的对比实验结果(在MEGC2019数据集上评估),其中,Haar小波在UF1和UAR指标上均取得了最优表现,这可能得益于其对边缘信息的良好保持能力[72],有助于特征提取与冗余抑制。相比之下,Daubechies(db4)与Symlet(sym5)等更复杂的小波基虽具备更强的表示能力,但在该模块中可能引入冗余或模糊信息,反而影响整体性能表现。

小波注意力下采样模块中注意力模块超参数的敏感性分析。表8实验结果(在MEGC2019数据集上评估)表明,当通道注意力的压缩比例ratio为

表7 小波注意力下采样模块中小波基函数的对比

小波基函数	UF1	UAR
Haar	**0.9090**	**0.9103**
Daubechies(db4)	0.8954	0.9033
Symlet(sym5)	0.8913	0.9006
Biorthogonal	0.8869	0.8979
Coiflet(coif5)	0.8889	0.8993

16时，模型的UF1和UAR均达到最高，表明该比例在增强关键通道特征表示的同时，避免了对高频细节的过度抑制。

表8　小波注意力下采样模块中通道注意力模块中压缩比例 ratio 的敏感性分析

压缩比例 ratio	UF1	UAR
4	0.8915	0.8979
8	0.8960	0.9047
16	**0.9090**	**0.9103**
32	0.8848	0.8917

表9实验结果（在MEGC2019数据集上评估）表明，对于空间注意力模块，卷积核大小为3时，模型的UF1和UAR指标同样达到最高，这表明较小的卷积核能够更精准地聚焦于微表情的关键局部区域，同时保持较低的计算复杂度。

表9　小波注意力下采样模块中空间注意力模块中卷积核大小的敏感性分析

卷积核大小	UF1	UAR
3	**0.9090**	**0.9103**
5	0.8824	0.8904
7	0.8784	0.8866

表10　频域变换模块中不同阶段数性能和计算量的对比

阶段数	UF1	UAR	参数	FLOPs
3	0.8811	0.8877	13.99 M	3.71 G
4	**0.9090**	**0.9103**	**31.45 M**	**4.47 G**
5	0.8943	0.8998	96.30 M	7.28 G

频域变换模块中不同阶段数性能和计算量的对比实验。表10的实验结果表明，当阶段数为4时，模型在UF1和UAR上均达到最优性能，分别为0.9090和0.9103。然而，随着阶段数的增加，模型的参数量和计算量显著增加。具体而言，阶段数为3时，模型参数量为13.99 M，计算量FLOPs为3.71 G；而阶段数为5时，参数量增加至96.30 M，计算量FLOPs达到7.28 G。这表明阶段数的增加虽然在一定程度上提升了特征提取的复杂性，但并未持续提升模型的识别性能，反而导致计算资源的大幅增加。因此，阶段数为4是性能与计算效率之间的最优平衡点。

4.2.3　对比实验

光流方法对比实验。表11(a)展示了本文提出的双向光流方法与现有光流方法在MEGC2019数据集上的性能对比结果。实验表明，本文的双向光流方法在UF1和UAR两项指标上均明显优于现有方法。与Bi-WOOF和MDMO方法相比，UF1分别提升了3.84%和3.48%，UAR分别提升了2.86%和2.45%。这一性能提升主要源于双向光流方法对微表情动态过程的完整建模能力。Bi-WOOF和MDMO方法仅计算起始帧到峰值帧的单向光流，忽略了微表情从峰值到结束帧的恢复过程，导致时序信息利用不充分，且易受头部姿态变化引入的全局噪声干扰。而本文方法通过融合峰值帧到起始帧和结束帧的双向光流，既强化了与微表情相关的局部肌肉运动特征，又通过方向相反的光流抵消了头部运动引起的噪声。此外，双向光流的对称性设计能够更完整地捕捉微表情的生成与消退动态，从而提升特征表达的鲁棒性和判别力。

表11　对比实验结果

(a)光流方法对比			(b)频率表征方法和频域模块对比		
方法	UF1	UAR	方法	UF1	UAR
Bi-WOOF[4]	0.8706	0.8817	傅里叶变换	0.8677	0.8777
MDMO[8]	0.8742	0.8858	Freq-HD[9]	0.8564	0.8667
本文模型	**0.9090**	**0.9103**	本文模型	**0.9090**	**0.9103**

频率表征方法对比实验。表11(b)展示了频率变换方式对比实验的结果。数据表明，小波变换将UF1和UAR分别提升至0.9090和0.9103，性能分别较傅里叶变换提高了4.13%和3.26%。这种优势主要源于小波变换的两大特性：

(1) 时空分析能力：小波变换能够同时捕捉信号的局部时间和频率信息，适合解析微表情这类短暂的面部动态，而傅里叶变换仅提供全局频域特征，无法定位变化发生的具体时刻。

(2) 非平稳信号适应性：面对微表情中快速波动的频率成分，小波变换通过多尺度分析跟踪瞬时变化，而傅里叶变换因缺乏时间分辨率，难以有效处理此类信号。

频域模块对比实验。表11(b)还展示了频域模块对比实验的结果，即比较了Freq-HD方法与本文模型的性能。具体而言，将本文模型中的小波注意力下采样模块替换为Freq-HD方法中的空间-时间频率分析(STFA)模块后，UF1和UAR分别下降了5.26%和4.36%。这表明本文的小波注意力下采样模块更具优势。原因在于微表情作为一种非平稳信号，其频率成分随时间迅速变化。小波变换通过多尺度分解能够同时捕捉信号的全局特征和局部细节，这种特性使其能够适应微表情中快速变化的频

率成分，更有效地分析非平稳信号。例如，小波变换将信号分解为低频和高频子带，既能保留面部结构的整体信息，又能捕捉到眼角或嘴角细微运动的高频细节，从而为微表情识别提供更丰富的特征表示。相比之下，STFA模块虽然能够从视频中提取片段的空间-时间频率特征，但在处理快速变化的非平稳信号时存在一定局限性。STFA模块主要通过计算片段的动态值来分析频率特征，缺乏对局部时间信息足够精细的捕捉能力。由于微表情的高频变化通常集中在特定的时间段和空间区域（如眼角或嘴角的快速运动），STFA模块难以精确地定位和区分这些细微变化，导致特征表征不够准确，进而影响了模型对微表情的识别能力。

4.2.4 Grad-CAM图可视化分析

为了分析所提出的注意力模型对网络性能的影响，我们使用Grad-CAM[73]探究了双向光流和小波注意力下采样模块在特征提取的过程中对模型的影响，如图9所示。

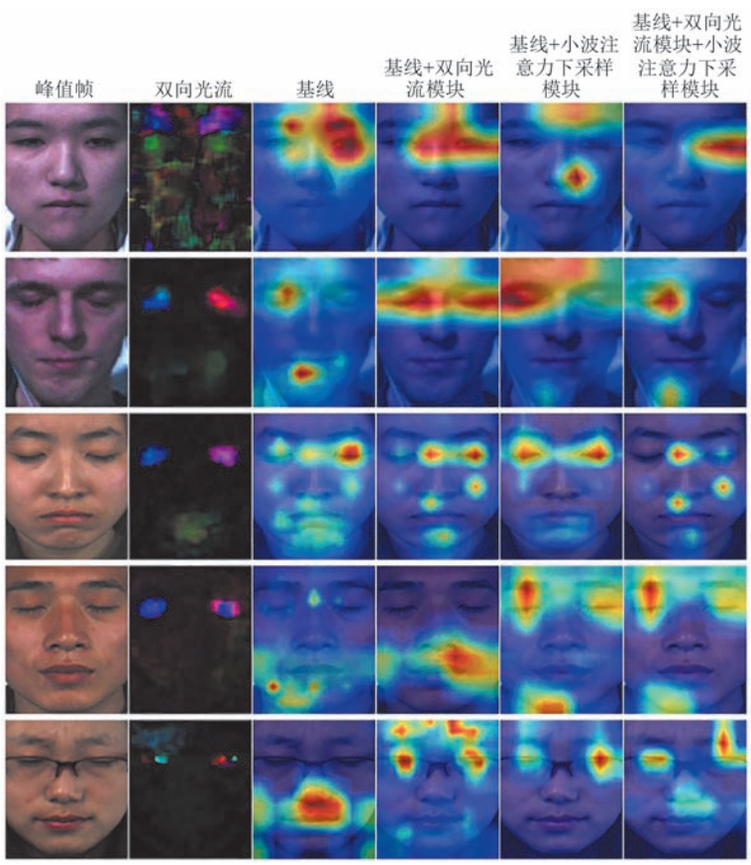


图9 部分微表情样本的Grad-CAM图（每行从左到右依次对应峰值帧、双向光流、输入为传统光流的Grad-CAM图、输入为双向光流的Grad-CAM图、输入为传统光流并使用小波注意力下采样模块的Grad-CAM图、输入为双向光流并使用小波注意力下采样模块的Grad-CAM图）

图9表明，传统光流输入下的Grad-CAM热力图中，注意力区域较为分散，且部分非相关区域（如面部静态区域或背景）也存在较高的激活值。相比之下，双向光流输入下，模型的注意力分布明显更集中于面部的动态区域，例如嘴角、眉毛和眼部周围，这些区域是微表情变化的主要表现部位。这表明双向光流能够提升模型对关键特征区域的捕捉能力。

对于小波注意力下采样模块，Grad-CAM热力图进一步凸显了模型对微表情运动与形变区域的关注能力。例如，在双向光流结合小波注意力下采样模块的输入下，模型在嘴部、鼻翼和眼部等部位的激活区域更加突出，同时对背景和无关区域的激活大幅减少。这种提升主要源于小波注意力下采样模块能够捕捉局部频域特征，并对微小运动和形变区域能进行精细建模。

5 结 论

本文提出了一种基于双向光流和小波变换注意力机制的微表情识别方法。通过分析微表情视频中

峰值帧与起始帧、结束帧之间的光流信息，本文模型能更准确地捕捉面部肌肉的细微运动。此外，引入的小波注意力下采样模块通过在频域中对高低频信息的精细处理，增强了模型对微表情细节的表征识别能力。

尽管取得了一定的进展，我们的工作还存在一些局限性。首先，模型对于表情极端变化的识别能力还有待加强，特别是表情变化较为微妙的情况。未来的工作将集中在提高模型的泛化能力，特别是对于复杂和多变的实际应用场景。此外，我们也将探索更深层次的网络结构和更高效的特征表征策略，以进一步提升识别鲁棒性。最后，我们计划通过大规模学习的多模态数据信息来微调与指导模型，以实现更全面准确的性能提升。

致　谢　感谢国家自然科学基金面上项目(Nos. 62276170, 62576216, 62576076)、国家自然科学基金青年科学基金项目(Nos. 62306061)、广东省科技计划项目(Nos. 2023A1515011549, 2023A1515010688)、人工智能与数字经济广东省实验室(深圳)开放课题(No. GML-KF-24-11)，以及广东省重点实验室(No. 2023B1212060076)对本研究的资助。

参　考　文　献

[1] Zhang Miao-Xuan, Zhang Hong-Gang. A survey on interpretability of facial expression recognition. Chinese Journal of Computers, 2024, 47(12): 2819-2851 (in Chinese)
(张淼萱, 张洪刚. 人脸表情识别可解释性研究综述. 计算机学报, 2024, 47(12):2819-2851)

[2] Ekman P, Friesen W V. Nonverbal leakage and clues to deception. Psychiatry, 1969, 32(1): 88-106

[3] Weinberger S. Airport security: Intent to deceive? Nature, 2010, 465(7297): 412-416

[4] Liong S T, See J, Wong K S, et al. Less is more: Micro-expression recognition from video using apex frame. Signal Processing: Image Communication, 2018, 62: 82-92

[5] Liong S T, Gan Y S, See J, et al. Shallow triple stream three-dimensional CNN (STSTNet) for micro-expression recognition// Proceedings of the 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019). Lille, France, 2019: 1-5

[6] Liu Y J, Zhang J K, Yan W J, et al. A main directional mean optical flow feature for spontaneous micro-expression recognition. IEEE Transactions on Affective Computing, 2015, 7(4): 299-310

[7] Gan Y S, Liong S T, Yau W C, et al. OFF-ApexNet on micro-expression recognition system. Signal Processing: Image Communication, 2019, 74: 129-139

[8] Liu Y J, Li B J, Lai Y K. Sparse mdmo: Learning a discriminative feature for micro-expression recognition. IEEE Transactions on Affective Computing, 2018, 12(1): 254-261

[9] Tao Z, Wang Y, Chen Z, et al. Freq-HD: An interpretable frequency-based high-dynamics affective clip selection method for in-the-wild facial expression recognition in videos// Proceedings of the 31st ACM International Conference on Multimedia (ACM MM 2023. Ottawa, Canada, 2023: 843-852

[10] Deng J, Dong W, Socher R, et al. ImageNet: A large-scale hierarchical image database //Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009). Miami, USA, 2009: 248-255

[11] Ren S, He K, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 39(6): 1137-1149

[12] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation //Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI 2015). Munich, Germany, 2015: 234-241

[13] Qiao Shao-Jie, Xue Qi, Yang Guo-Ping, et al. A multivariate time series forecasting model based on dynamic adaptive spatio-temporal graphs. Chinese Journal of Computers, 2024, 47(12): 2925-2937 (in Chinese)
(乔少杰, 薛骐, 杨国平, 等. 基于动态自适应时空图的多元时序预测模型. 计算机学报, 2024, 47(12):2925-2937)

[14] Kong Qing-Qun, Wu Fu-Chao, Fan Bin. Image matching in deep learning Era: Methods, applications and challenges. Chinese Journal of Computers, 2024, 47(07): 1485-1520 (in Chinese)
(孔庆群, 吴福朝, 樊彬. 基于深度学习的图像匹配:方法、应用与挑战. 计算机学报, 2024, 47(07):1485-1520)

[15] Li H, Sui M, Zhu Z, et al. MMNet: muscle motion-guided network for micro-expression recognition //Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI 2022). Vienna, Austria, 2022

[16] Chen B, Zhang Z, Liu N, et al. Spatiotemporal convolutional neural network with convolutional block attention module for micro-expression recognition. Information, 2020, 11(8): 380

[17] Wang C, Peng M, Bi T, et al. Micro-attention for micro-expression recognition. Neurocomputing, 2020, 410: 354-362

[18] Li Y, Huang X, Zhao G. Micro-expression action unit detection with spatial and channel attention. Neurocomputing, 2021, 436: 221-231

[19] Zhou H, Huang S, Li J, et al. Dual-atme: dual-branch attention network for micro-expression recognition. Entropy, 2023, 25(3): 460

[20] Zafar A, Aamir M, Mohd Nawi N, et al. A comparison of pooling methods for convolutional neural networks. Applied Sciences, 2022, 12(17): 8643

[21] Zhao S, Deng W, Socher R, et al. Affective image content analysis: two decades review and new perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(10): 6729-6751

[22] Pei G, Li H, Lu Y, et al. Affective computing: Recent advances, challenges, and future trends. Intelligent Computing, 2024, 3: 0076

[23] Achlioptas P, Ovsjanikov M, Guibas L, et al. Affection: Learning affective explanations for real-world visual data// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2023). Vancouver, Canada, 2023: 6641-6651

[24] Yang J, Huang Q, Ding T, et al. EmoSet: A large-scale visual emotion dataset with rich attributes //Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV 2023). Paris, France, 2023: 20383-20394

[25] Haydarov K, Shen X, Madasu A, et al. Affective visual dialog: A large-scale benchmark for emotional reasoning based on visually grounded conversations //Proceedings of the European Conference on Computer Vision (ECCV 2024). Milan, Italy, 2024: 18-36

[26] Wu D, Yang D, Zhou Y, et al. Bridging visual affective gap: Borrowing textual knowledge by learning from noisy image-text pairs //Proceedings of the 32nd ACM International Conference on Multimedia (ACM MM 2024). Melbourne, Australia, 2024: 602-611

[27] Xu Q, Yuan S, Wei Y, et al. Multiple feature refining network for visual emotion distribution learning //Proceedings of the AAAI Conference on Artificial Intelligence (AAAI 2025). Philadelphia, USA, 2025: 8924-8932

[28] Wang Y, Guo J, Bai J, et al. InstructAvatar: Text-guided emotion and motion control for avatar generation //Proceedings of the AAAI Conference on Artificial Intelligence (AAAI 2025). Philadelphia, USA, 2025: 8132-8140

[29] Zheng C, Mendieta M, Chen C. POSTER: A pyramid cross-fusion transformer network for facial expression recognition // Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV 2023). Paris, France, 2023: 3146-3155

[30] Wu Y, Wang S, Chang Y. Patch-aware representation learning for facial expression recognition//Proceedings of the 31st ACM International Conference on Multimedia (ACM MM 2023). Ottawa, Canada, 2023: 6143-6151

[31] Xie W, Peng Z, Shen L, et al. Cross-layer contrastive learning of latent semantics for facial expression recognition. IEEE Transactions on Image Processing, 2024, 33: 2514-2529

[32] Zhang Y, Zheng X, Liang C, et al. Generalizable facial expression recognition //Proceedings of the European Conference on Computer Vision (ECCV 2024). Milan, Italy, 2024: 231-248

[33] Wang C, Chen L, Wang L, et al. QCS: Feature refining from quadruplet cross similarity for facial expression recognition // Proceedings of the AAAI Conference on Artificial Intelligence (AAAI 2025). Philadelphia, USA, 2025: 7563-7572

[34] Li M, Zhang X, Liao T, Lin S, Xiao G. PTH-Net: dynamic facial expression recognition without face detection and alignment. IEEE Transactions on Image Processing, 2025, 34: 30-43

[35] Zhang B, Wang X, Wang C, et al. Dynamic stereotype theory induced micro-expression recognition with oriented deformation // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2025). Nashville, USA, 2025: 10701-10711

[36] Chen B, Liu K H, Xu Y, et al. Block division convolutional network with implicit deep features augmentation for micro-expression recognition. IEEE Transactions on Multimedia, 2022, 25: 1345-1358

[37] Zhou H, Huang S, Xu Y. Inceptr: micro-expression recognition integrating inception-CBAM and vision transformer. Multimedia systems, 2023, 29(6): 3863-3876

[38] Nguyen X B, Duong C N, Li X, et al. Micron-BERT: BERT-based facial micro-expression recognition //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2023). Vancouver, Canada, 2023: 1482-1492

[39] Wang Z, Zhang K, Luo W, et al. Htnet for micro-expression recognition. Neurocomputing, 2024, 602: 128196

[40] Ma B, Wang L, Wang Q, et al. Entire-detail motion dual-branch network for micro-expression recognition. Pattern Recognition Letters, 2025, 189: 166-174

[41] Fleet D, Weiss Y. Optical flow estimation//Handbook of mathematical models in computer vision. Boston, USA: Springer US, 2006: 237-257

[42] Lucas B D, Kanade T. An iterative image registration technique with an application to stereo vision //Proceedings of the 7th International Joint Conference on Artificial Intelligence (IJCAI 1981). Vancouver, Canada, 1981: 674-679

[43] Farnebäck G. Two-frame motion estimation based on polynomial expansion //Proceedings of the 13th Scandinavian Conference on Image Analysis (SCIA 2003). Halmstad, Sweden, 2003: 363-370

[44] Dosovitskiy A, Fischer P, Ilg E, et al. FlowNet: Learning optical flow with convolutional networks //Proceedings of the IEEE International Conference on Computer Vision (ICCV 2015). Santiago, Chile, 2015: 2758-2766

[45] Ilg E, Mayer N, Saikia T, et al. FlowNet 2.0: Evolution of optical flow estimation with deep networks //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017). Honolulu, USA, 2017: 2462-2470

[46] Teed Z, Deng J. RAFT: Recurrent all-pairs field transforms for optical flow //Proceedings of the 16th European Conference on Computer Vision (ECCV 2020). Glasgow, UK, 2020: 402-419

[47] Oh Y H, Le Ngo A C, See J, et al. Monogenic Riesz wavelet representation for micro-expression recognition //Proceedings of the IEEE International Conference on Digital Signal Processing (DSP 2015). Singapore, 2015: 1237-1241

[48] Oh Y H, Le Ngo A C, Phari R C W, et al. Intrinsic two-dimensional local structures for micro-expression recognition // Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016). Shanghai, China, 2016: 1851-1855

[49] Wietzke L, Fleischmann O, et al. 2D image analysis by generalized Hilbert transforms in conformal space //Proceedings of the European Conference on Computer Vision (ECCV 2008). Marseille, France, 2008: 638-649.

[50] Zhang P, Ben X, Yan R, et al. Micro-expression recognition

system. Optik, 2016, 127(3): 1395-1400.
[51] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Advances in Neural Information Processing Systems, 2017, 30: 5998-6008.
[52] Hu J, Shen L, Sun G. Squeeze-and-excitation networks // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018). Salt Lake City, USA, 2018: 7132-7141
[53] Woo S, Park J, Lee J Y, et al. CBAM: Convolutional block attention module //Proceedings of the European Conference on Computer Vision (ECCV 2018). Munich, Germany, 2018: 3-19
[54] Zhao S, Tang H, Mao X, et al. Dfme: A new benchmark for dynamic facial micro-expression recognition. IEEE Transactions on Affective Computing, 2023, 15(3): 1371-1386
[55] See J, Yap M H, Li J, et al. MEGC 2019: The second facial micro-expressions grand challenge //Proceedings of the 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019). Lille, France, 2019: 1-5
[56] Yan W J, Li X, Wang S J, et al. CASME II: An improved spontaneous micro-expression database and the baseline evaluation. PloS One, 2014, 9(1): e86041
[57] Davison A K, Lansley C, Costen N, et al. Samm: A spontaneous micro-facial movement dataset. IEEE Transactions on Affective Computing, 2016, 9(1): 116-129
[58] Li X, Pfister T, Huang X, et al. A spontaneous micro-expression database: Inducement, collection and baseline // Proceedings of the 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG 2013). Shanghai, China, 2013: 1-6
[59] Li J, Dong Z, Lu S, et al. CAS (ME) 3: A third generation facial spontaneous micro-expression database with depth information and high ecological validity. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(3): 2782-2800
[60] Zhou L, Mao Q, Xue L. Dual-Inception network for cross-database micro-expression recognition //Proceedings of the 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019). Lille, France, 2019: 1-5
[61] Xia B, Wang S. Micro-expression recognition enhanced by macro-expression from spatial-temporal domain //Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI 2021). Montreal, Canada, 2021: 1186-1193
[62] Zhai Z, Zhao J, Long C, et al. Feature representation learning with adaptive displacement generation and transformer fusion for micro-expression recognition //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2023). Vancouver, Canada, 2023: 22086-22095
[63] Liu Y, Du H, Zheng L, et al. A neural micro-expression recognizer //Proceedings of the 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019). Lille, France, 2019: 1-4
[64] Xia B, Wang W, Wang S, et al. Learning from macro-expression: A micro-expression recognition framework // Proceedings of the 28th ACM International Conference on Multimedia (ACM MM 2020). Seattle, USA, 2020: 2936-2944
[65] Zhou L, Mao Q, Huang X, et al. Feature refinement: An expression-specific feature learning and fusion method for micro-expression recognition. Pattern Recognition, 2022, 122: 108275
[66] Lei L, Chen T, Li S, et al. Micro-expression recognition based on facial graph representation learning and facial action unit fusion //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2021). Nashville, USA, 2021: 1571-1580
[67] Gan Y S, Lien S E, Chiang Y C, et al. LAENet for micro-expression recognition. The Visual Computer, 2024, 40(2): 585-599
[68] Cai W, Zhao J, Yi R, et al. MFDAN: Multi-level flow-driven attention network for micro-expression recognition. IEEE Transactions on Circuits and Systems for Video Technology, 2024, 34(12): 12823-12836
[69] Bao Y, Wu C, Zhang P, et al. Boosting micro-expression recognition via self-expression reconstruction and memory contrastive learning. IEEE Transactions on Affective Computing, 2024, 15(4): 2083-2096
[70] Liu Z, Mao H, Wu C Y, et al. A ConvNet for the 2020s // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2022). New Orleans, USA, 2022: 11976-11986
[71] Liong G B, Liong S T, Chan C S, et al. SFAMNet: A scene flow attention-based micro-expression network. Neurocomputing, 2024, 566: 126998
[72] Shensa M J. The discrete wavelet transform: wedding the a trous and Mallat algorithms. IEEE Transactions on Signal Processing, 2002, 40(10): 2464-2482
[73] Selvaraju R R, Cogswell M, Das A, et al. Grad-CAM: Visual explanations from deep networks via gradient-based localization // Proceedings of the IEEE International Conference on Computer Vision (ICCV 2017). Venice, Italy, 2017: 618-626

XIE Wei-Cheng, Ph. D., associate professor. His main interests include facial expression analysis and robust network design.

XIAO Hang, M. S. candidate. His main interests include deep learning and micro-expression recognition.

FAN Wei-Jia, M. S. candidate. His main interests include visual recognition tasks and multi-modal learning.

WANG Zi-Han, M. S. candidate. His main interests include facial action units recognition and facial expression editing.

YU Zi-Tong, Ph. D., associate professor. His main interests include Human-centered visual computing, multimedia security, multimodal foundation model.

SHEN Lin-Lin, Ph. D., professor. His main interests include deep learning, facial recognition, analysis/synthesis and medical image processing.

Background

Micro-expression recognition is crucial in computer vision and affective computing. These brief, involuntary facial movements reveal hidden emotions and are valuable for psychology, security, and human-computer interaction.

From one aspect, traditional micro-expression recognition methods, particularly those based on optical flow, face challenges in accurately capturing subtle facial movements due to noise sensitivity and insufficient temporal information. For example, conventional optical flow methods often confuse genuine facial muscle movements with head motion-induced pixel changes, resulting in noisy feature representation. Additionally, traditional downsampling techniques like max-pooling tend to discard critical high-frequency details essential for detecting nuanced changes in micro-expressions.

From another aspect, recent research has explored attention mechanisms and frequency-domain features to address these issues, where frequency-domain features capture both high- and low-frequency information for a comprehensive representation. However, existing methods still lack effective frequency-aware feature preservation for the task of micro-expression recognition.

To address these limitations, our method introduces two novel components: the Bidirectional Optical Flow Module (BOFM) and the Wavelet-based Attention Downsampling Module (WADM). The BOFM enhances feature representation by analyzing optical flow from both apex-to-onset and apex-to-offset frames, reducing noise interference and improving muscle movement detection. The WADM integrates wavelet transform with attention mechanisms to preserve high-frequency details typically lost in downsampling. By selectively focusing on key regions and capturing both subtle and significant aspects of micro-expressions, our method provides a more balanced and comprehensive representation. Especially, the employed wavelet transforms offer better spatiotemporal analysis capabilities and adaptability to non-stationary signals, making them more suitable for detecting subtle changes in micro-expressions, compared with the Fourier transforms.

Our method outperforms existing techniques, achieving higher F1 scores and recall rates on the MEGC2019 dataset.