中图法分类号:TP391.4　　**文献标识码**:A　　**文章编号**:1006-8961(XXXX)XX-0001-15

论文引用格式:Li Ce, Wang Kai, Xiao Limei, Wang Ru, Ping Mengmeng, Lu Ming. Cross-fusion multi-level receptive field network for facial expression recognition[J/OL]. Journal of Image and Graphics, XXXX:1-15. DOI: 10.11834/jig.260054.(李策, 王凯, 肖利梅, 王茹, 平梦梦, 卢明. 用于面部表情识别的交叉融合多级感受野网络[J/OL]. 中国图象图形学报, XXXX:1-15. DOI: 10.11834/jig.260054.)[DOI:10.11834/jig.260054]

用于面部表情识别的交叉融合多级感受野网络

李策[2,1],王凯[1],肖利梅[1],王茹[2],平梦梦[1],卢明

1. 兰州理工大学自动化与电气工程学院,兰州　730050　　2　兰州理工大学微电子现代产业学院,兰州　730050; 2. 甘肃省科学院自动化研究所,兰州　730030

摘　要:**目的**　面部表情识别在计算机视觉领域广泛应用,涵盖人机交互、医疗健康和在线行为监测等场景。然而,现有方法往往未能充分建模面部关键区域的局部特征,导致在应对类间相似性高、类内差异性大的复杂表情时性能受限。为了提升模型对面部关键区域的捕捉能力,提出了一种用于面部表情识别的交叉融合多级感受野网络(cross-fusion multi-level receptive field network,CFMRFN)。**方法**　首先,利用Transformer框架融合面部的整体表情特征与局部标志点特征,从而利用标志点引导模型聚焦于面部关键区域。其次,针对Transformer捕捉面部标志点这种局部特征中适应性不足的问题,提出的滑动膨胀窗口注意力机制,在保持全局感知与并行计算优势的同时限制注意力计算范围,实现对面部关键区域的深度建模。最后,为了进一步捕获眼角、嘴角等细微区域,在滑动膨胀窗口注意头中配置多种膨胀率,构建多级感受野检测器,以多尺度感受野捕获面部的关键区域特征,从而加强模型对关键区域的捕捉能力并提升识别性能。**结果**　所提方法在RAF-DB、AffectNet(7cls/8cls)和FERPlus三个公开数据集上进行了实验验证,整体准确率分别达到92.14%、67.35%(7cls)、63.44%(8cls)和91.67%,在41.8M参数量和6.8G FLOPs的计算开销下,与Expllm相比在RAF-DB上提升了1.11%。为进一步检验模型在复杂表情下的判别能力,对三个数据集进行了单类表情准确率评估,并通过模型分析与消融实验验证了所提方法在复杂表情模式下的性能。**结论**　本文所提方法将局部标志点与全局表情特征进行交叉融合,并利用所提的多级感受野准确捕获眼角、嘴角等关键区域的局部特征,使模型在复杂表情模式下依然保持精准的识别表现。具体代码可在此处获取https://www.scidb.cn/anonymous/SmZ1aWFx

关键词:人脸表情识别;Transformer;交叉融合网络;滑动膨胀窗口;多级感受野

Cross-fusion multi-level receptive field network for facial expression recognition

Li Ce[2,1], Wang Kai[1], Xiao Limei[1], Wang Ru[2], Ping Mengmeng[1], Lu Ming[3]

1. School of Automation and Electrical Engineering, Lanzhou University of Technology, Lanzhou 730050, China; 2. School of Microelectronics Industry-Education Integration, Lanzhou University of Technology, Lanzhou 730050, China; 3. Institute of Automation, Gansu Academy of Sciences, Lanzhou 730030, China

Abstract: **Objective** Facial expression recognition (FER) has emerged as a pivotal research topic in the field of computer vision due to its broad applicability in diverse domains such as human-computer interaction, intelligent education systems, healthcare monitoring, mental health assessment, driver fatigue detection, and online behavior analysis. Accurate FER

收稿日期:2026-01-26;**修回日期**:2026-04-22

基金项目:国家自然科学基金(62363025);甘肃省科技厅重点研发计划--社会发展类(23YFFA0064);2025年度甘肃省文化和旅游科学研究项目(2025KZ003)

Supported by: National Natural Science Foundation of China (62363025); Key Research Program of Gansu Provincial Science and Technology Department - Social Development Category (23YFFA0064); 2025th Gansu Provincial Cultural and Tourism Scientific Research Project (2025KZ003)

enables machines to interpret human emotions, thereby enhancing the naturalness and adaptability of interactive systems. Despite recent advances driven by convolutional neural network (CNN) and, more recently, Transformer-based architectures, the performance of FER models remains constrained in challenging scenarios characterized by high inter-class similarity and large intra-class variability. One of the primary limitations of existing methods lies in their insufficient ability to effectively model fine-grained discriminative features from facial key regions—such as the corners of the eyes and mouth—which are crucial for recognizing subtle emotional differences. This deficiency often results in feature representations that are dominated by global semantic patterns while failing to capture subtle but semantically critical local cues, ultimately reducing classification accuracy in complex and ambiguous expression categories. **Methods** To address these challenges, we propose a novel Cross-Fusion Multi-Level Receptive Field Network specifically designed for facial expression recognition. The central idea of our approach is to explicitly integrate local structural priors derived from facial landmarks with global semantic expression features, enabling a more balanced and discriminative feature representation that combines detailed regional cues with holistic facial context. Concretely, we first design a dual-stream feature extraction framework in which one stream encodes the global expression semantics of the entire face, while the other focuses on fine-grained geometric structures extracted from facial landmark points. These two feature streams are then fused using a Transformer-based cross-fusion module, allowing rich bidirectional interactions in the spatial domain. This fusion mechanism effectively guides the model to attend to expression-critical regions while simultaneously leveraging the contextual information provided by the full-face representation. A key technical innovation in our model is the Sliding Dilated Window Attention mechanism, which is designed to overcome the limitations of conventional global self-attention in standard Transformers. While global attention offers strong long-range dependency modeling and parallel computation efficiency, it lacks the local inductive bias inherent to CNNs, making it less effective for capturing small scale discriminative features. To remedy this, our SDWA mechanism restricts attention computation to localized sliding windows, significantly reducing computational complexity while concentrating attention resources on critical facial areas. Furthermore, inspired by the dilated convolution paradigm, we introduce varying dilation rates within these sliding windows to expand the receptive field without sacrificing resolution. Specifically, we assign different dilation rates to different attention heads, thereby constructing a Multi-Level receptive field encoder within the Transformer. This design allows the network to simultaneously perceive small-scale details, such as subtle movements in the eyes or mouth corners, and larger-scale structural variations, enabling a richer and more semantically aligned representation of facial expressions. **Result** We evaluate our method on three widely used benchmark datasets: RAF-DB, AffectNet, and FERPlus. The proposed model achieves top-tier performance, with classification accuracies of 92.14% on RAF-DB (seven-class classification), 67.35% and 63.44% on AffectNet (seven- and eight-class classification settings), and 91.67% on FERPlus (eight-class classification). Our experiments are conducted using a standardized training pipeline with data augmentation strategies such as random cropping, horizontal flipping, and color jittering, and optimization via AdamW with a learning rate warm-up schedule. We also perform a comprehensive ablation study, demonstrating that removing any of the three major components—the cross-fusion framework, the SDWA mechanism, or the multi-dilation receptive field encoder—results in a significant performance drop. Additionally, qualitative analysis using attention heatmaps confirms that the proposed method consistently focuses on semantically meaningful facial regions, such as the eyes and mouth, across different expression categories and datasets. Beyond raw accuracy numbers, we compare our method with representative state-of-the-art FER models, including CNN-based architectures and pure Transformer baselines. Our network consistently outperforms these methods, particularly in categories with subtle local differences, validating the effectiveness of explicitly modeling key facial regions. Notably, the proposed architecture achieves this without incurring excessive computational cost, maintaining an inference speed suitable for real-time applications. **Conclusion** In this work, the proposed Cross-Fusion Multi-Level Receptive Field Network offers a new paradigm for FER by effectively combining the strengths of global and local feature modeling within a unified Transformer-based framework. Its ability to attend to subtle local details while preserving global semantic coherence leads to notable improvements in recognition accuracy on challenging benchmark datasets. Future work will explore adaptive attention windowing strategies and integration with temporal modeling modules to extend the approach to video-based FER and other dynamic facial analysis tasks.

Key words: facial expression recognition; transformer; cross-fusion network; sliding dilated window; multi-level receptive

论文引用格式:DOI:10.11834/jig.260054

0 引　言

面部表情作为人类非语言情感表达的重要方式,在传递情绪、意图和态度方面具有不可替代的作用(Li等,2020)。近年来,深度学习方法逐渐取代传统的手工特征提取技术,在面部表情识别任务(facial expression recognition,FER)中表现出更高的识别精度与泛化能力。面部表情识别的核心目标是将输入人脸图像准确分类为七种基本表情(Ekman和Friesen,1971):愤怒、厌恶、恐惧、高兴、悲伤、惊讶以及中性。在后续的研究中还进一步扩展为八类基本表情,在上述七种表情的基础上新增蔑视,以更全面地刻画人类面部情绪状态,为心理健康监测、智能教育与人机交互等领域提供评价范式。

近年来大量研究围绕特征学习与面部动作单元(Action Unit,AU)的建模展开。一方面Zhong等人(2019)和Song等人(2021)通过构建图结构以建模面部局部区域间的空间关系及其动态变化,增强了模型对面部结构的感知能力。Ruan等人(2020)从特征提取和网络结构优化的角度出发,通过改进网络架构提升了模型的表达能力和识别性能。另一方面Li等人(2019)聚焦于面部动作单元相关建模,强调AU之间的语义依赖关系。Zhang等人(2020)则基于图卷积网络(graph convolutional network,GCN)挖掘面部AU之间的空间关联性。这些方法在处理面部姿态变化、光照干扰和身份差异等挑战时展现出一定的性能,但应对表情中的类间相似性高与类内差异性大的挑战时,仍存在识别有误的情况。

面部表情识别面临的两大挑战如图1所示,(a)是类间相似性高,不同表情中常出现闭眼、张嘴、皱眉等重叠局部动作,造成类别边界模糊;(b)是类内差异性大,同一表情在不同个体下呈现出差异化的局部特征组合。这种“类间相似性高”与“类内差异性大”的并存挑战,提升了FER任务的复杂性。为了应对面部表情识别中的这两大挑战,本文提出了一种用于面部表情识别的交叉融合多级感受野网络,主要贡献如下:

1)针对现有方法未能充分利用面部关键区域的问题,提出双流特征交叉融合Transformer框架,将面部标志点信息作为结构先验引入,与表情语义特征在空间维度进行交互融合,强化模型对面部关键区域的关注能力。

图1　面部表情识别中的两大挑战((a)类间相似性高;(b)类内差异性大)

Figure 1　Two major challenges in facial expression recognition ((a) High similarity among classes; (b) There are significant differences within the class)

2)针对传统Transformer(Vaswani等,2017)缺乏卷积神经网络(convolutional neural networks,CNN)所具备的局部归纳偏置,对于捕捉面部标志点这种局部特征中展现出适应性不足的问题,借鉴膨胀卷积思想(Fisher Yu等,2015),提出了一种新颖的滑动膨胀窗口注意力机制,限制注意力的计算范围,使模型专注于关键区域细节建模,同时保持全局建模能力与高效并行计算。

3)针对Transformer模型捕获眼角、嘴角等关键区域不足的问题,构建多级感受野编码器,以多尺度感受野捕获眼角、嘴角等细微区域,实现局部与全局的协同表达,从而提升表情识别的能力。

4)在RAF-DB(Li等,2017)、AffectNet(Mollahosseini等,2017)和FERPlus(Barsoum等,2016)三个常用表情识别数据集上分别取得92.14%(7分类)、67.35%/63.44%(7/8分类)和91.67%(8分类)的分类准确率,验证了所提方法的有效性与优越性。

1 相关工作

1.1 人脸表情识别

近年来,深度学习在计算机视觉领域取得了显著进展,在解决面部表情识别这一具有挑战性的任

务中发挥着越来越关键的作用(Peng等,2020)。由于不同表情类别之间存在高度相似性,且同一类别内部又存在显著的个体差异与特征变化,许多FER方法聚焦于构建高效的网络结构与优化损失函数,以提取更加具备判别力的表情特征,从而提升识别准确性。例如,Zhou等人(2015)提出了一种将面部表情映射为多种具有不同强度的情绪状态的框架,实现捕捉面部表情的整体语义信息。Li等人(2018a)提出一种基于深度局部保持损失(locality preserving loss)的特征学习方法,通过保持特征的局部相似性,增强特征空间中的类内紧致性与类间分离性,从而提取更具判别力的表情表示。Cai等人(2018)设计了一种岛屿损失函数(island loss),以同时增强不同类别间的区分性并提升同类样本的紧密聚集程度。为进一步增强特征的判别能力,Sang等人(2018)针对类内表情变化问题,在FER任务中引入了密集卷积网络(densely connected convolutional networks, DenseNet)(Huang等,2017),以提升深层特征的表达能力。Xie等人(2019)设计了一种显著区域注意力模块(salient region attention),使网络能够聚焦于面部关键区域,从而有效增强了表情识别的性能。Wang等人(2020a)提出的区域注意力网络(regional attention network, RAN)通过对抗遮挡与姿态变化等干扰因素,精准捕捉面部局部区域的判别性特征,显著提升了FER模型在复杂场景下的性能。此外,Farzaneh等人(2021)提出一种深度注意力中心损失机制(deep attention center loss, DACL),通过为特征分配注意力权重强化其区分性,并结合稀疏中心损失进一步压缩类内特征分布、扩大类间距离,增强整体判别能力。Ruan等人(2021)则致力于建模表情特征间的相似性与差异性,通过细粒度特征的提取捕捉表情类别间的微妙变化,从而提升识别精度。Wang等人(2019)提出的高分辨率网络(high-resolution network, HRNet)通过并行多分辨率分支与多次跨尺度融合,保持整个网络中的高分辨率表示,取得了视觉识别任务的整体性能。

将视觉Transformer(Alexey等,2021)与面部区域信息结合,也已成为提升识别精度的有效策略。专注于面部关键区域有助于Transformer更好地学习辨识性的局部特征。Aouayeb等人(2021)在FER任务中直接应用ViT架构,并在其MLP头部前加入SE块,从而显著提升了性能。Cui等人(2023)提出的Face Transformer(towards high fidelity and accurate face swapping),利用Transformer构建源面孔与目标面孔之间的语义对应关系,并将源面孔的身份特征映射到目标面孔的相应区域。同样,Cui等人(2024)将对比学习融合进ViT的自监督学习取得了显著的效果。Dan等人(2023)提出了从数据中心角度校准Transformer训练(Transface: calibrating transformer training for face recognition from a data-centric perspective),能够聚焦面部关键区域特征,获得了较高的准确率。Xu等人(2023)提出了一种融合全局与局部特征的Transformer模型(HFFT: hierarchical global-local feature fusion transformers),专门应对复杂环境下的表情识别任务。Yang等人(2024)提出了采用掩码重建技术的CL-TransFER框架(collaborative learning based transformer for facial expression recognition with masked reconstruction),通过联合训练CNN和Transformer,能够有效提取面部图像的局部语义特征和全局结构信息。这些研究充分展示了ViT在面部识别和表情分析中的应用潜力,尤其在全局与局部特征建模方面的优势,即使在数据有限的情况下,依然能够取得不错的准确性,展示了ViT在表情识别中的显著优势。

1.2 人脸标志识别

面部标志由眉毛、眼睛、鼻子、嘴巴和下巴等关键区域的稀疏关键点组成(Wu等,2019),这些面部标志点旨可直接补充面部中的表情特征。Khan等人(2018)和Jin等人(2021)引入了面部标志来加强深度学习面部表情识别方法,特别是在关注面部关键区域方面,这有助于解决不同表情类别之间的相似性问题。Jung等人(2015)提出了两个独立的网络结构,一个接收图像输入,另一个接收面部标志输入,二者的输出通过加权求和进行融合。Xiong等人(2024)和Huang等人(2024)将面部标志与特定层的表情特征结合,有效提升了模型对空间特征的感知能力。然而,现有的面部标志利用方法未能充分考虑面部标志特征与表情特征之间的关联性。Chen等人(2020)利用面部标志和动作单元空间的拓扑关系,挖掘更多信息以支持标签分布学习。Thanh-Hung等人(2020)提出了金字塔超分辨率网络架构(pyramid with super resolution for in-thewild facial expression recognition)应对不同环境下表情图像干扰问题,同时通过结合面部标志特征与表情特征的

融合方法，进一步提升了特征提取能力，为FER中的难题提供了有效的解决方案。

2 方 法

针对面部表情识别中类间相似性高、类内差异性大的两大挑战，提出的交叉融合多级感受野网络（Cross-Fusion Multi-Level Receptive Field Network，CFMRFN）如图2所示，包括表情特征和面部标志特征交叉融合、滑动膨胀窗口注意力、多级感受野检测器三个部分组成。

首先，为了同时获取人脸图像中的全局表情语义信息与面部标志点的局部几何结构信息，并行提取面部全局表情特征与标志点特征，为后续特征融合奠定基础。在此基础上，提出了一种Transformer的交叉融合结构，实现两类特征间的交叉融合，从而显著提升整体特征的表达能力。其次，针对传统Transformer在局部特征建模中适应性不足的问题，提出滑动膨胀窗口注意力机制，通过限制注意力计算范围，不仅有效降低计算复杂度，还能更加聚焦于面部标志点的局部细节特征，从而对表情识别过程形成有效引导。最终，在滑动膨胀窗口内引入多膨胀率设计策略，为不同注意力头分配不同尺度的感受野，构建多层次的感知能力，使模型在建模局部特征时具备更丰富的表达能力，有效增强对面部关键区域的捕捉能力。

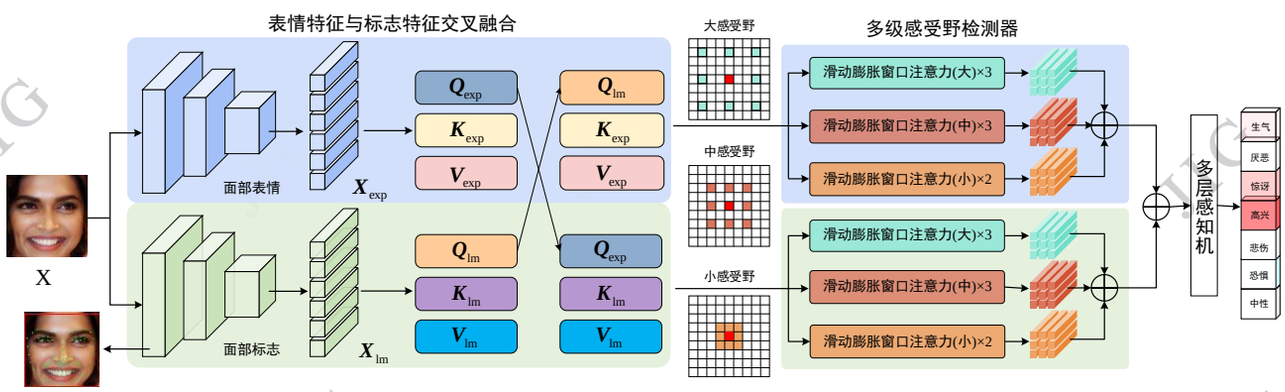


图2 所提用于面部表情识别的交叉融合多级感受野网络框架

Figure 2 The proposed cross-fusion multi-level receptive field network framework for facial expression recognition

2.1 表情特征和面部标志特征交叉融合

针对现有方法未能充分利用面部关键区域，导致在应对类间相似性高、类内差异性大的复杂表情模式时性能受限的问题。利用了在大规模数据集上预训练的骨干网络，分别提取面部的全局表情特征与局部标志点结构特征，该双流特征提取结果为后续的特征融合与注意力建模提供了丰富且互补的信息基础。由于面部标志是一组用于面部区域定位的关键点构成，这些关键点通常与表情变化最显著的区域（如眼角、眉间和嘴角）高度相关，具备良好的局部结构表达能力。然而，面部标志特征在建模全局语义方面存在一定局限，而表情特征则能够提供完整的面部外观与动作信息，作为全局表征的重要补充。为充分挖掘两类特征的互补性，提出交叉融合结构的Transformer编码器，构建跨特征的交叉融合机制，促进局部结构信息与全局表情语义的协同学习。

给定输入图像 $\boldsymbol{X} \in \boldsymbol{R}^{H \times W \times 3}$，分别使用骨干网络提取表情特征 $\boldsymbol{X}_{\text{exp}} \in \boldsymbol{R}^{N \times D}$ 和标志特征 $\boldsymbol{X}_{\text{lm}} \in \boldsymbol{R}^{N \times D}$。其中，exp和lm分别代表表情特征和面部标志特征，H和W是特征图的高度和宽度，$N = H \times W$为展开后的空间token数量，D表示特征维度。在交叉融合阶段，表情特征 $\boldsymbol{X}_{\text{exp}}$ 和面部标志特征 $\boldsymbol{X}_{\text{lm}}$ 分别映射到三个表示矩阵：表情的查询矩阵 $\boldsymbol{Q}_{\text{exp}}$，键矩阵 $\boldsymbol{K}_{\text{exp}}$，值矩阵 $\boldsymbol{V}_{\text{exp}}$，标志的查询矩阵 $\boldsymbol{Q}_{\text{lm}}$、键矩阵 $\boldsymbol{K}_{\text{lm}}$、值矩阵 $\boldsymbol{V}_{\text{lm}}$，它们分别如公式(1-3)所示。

$$\boldsymbol{Q}_{\text{exp}} = \boldsymbol{X}_{\text{exp}}\boldsymbol{W}_{q1}, \boldsymbol{Q}_{\text{lm}} = \boldsymbol{X}_{\text{lm}}\boldsymbol{W}_{q2} \tag{1}$$

$$\boldsymbol{K}_{\text{exp}} = \boldsymbol{X}_{\text{exp}}\boldsymbol{W}_{k1}, \boldsymbol{K}_{\text{lm}} = \boldsymbol{X}_{\text{lm}}\boldsymbol{W}_{k2} \tag{2}$$

$$\boldsymbol{V}_{\text{exp}} = \boldsymbol{X}_{\text{exp}}\boldsymbol{W}_{v1}, \boldsymbol{V}_{\text{lm}} = \boldsymbol{X}_{\text{lm}}\boldsymbol{W}_{v2} \tag{3}$$

式中，$\boldsymbol{W}_{q1}, \boldsymbol{W}_{q2}, \boldsymbol{W}_{k1}, \boldsymbol{W}_{k2}, \boldsymbol{W}_{v1}, \boldsymbol{W}_{v2} \in \mathbf{R}^{D \times D}$ 是映射矩阵。

在特征融合阶段，为实现表情特征与面部标志特征之间的深层交互，利用Transformer的自注意力

计算将表情特征与面部标志特征所构建的查询矩阵 $\boldsymbol{Q}$ 进行互换，以实现不同特征流之间的查询信息进行协同建模。其自注意力的计算方式如公式(4)和(5)所示：

$$\text{Attention}_{\text{exp}} = \text{Softmax}\left(\boldsymbol{Q}_{\text{lm}}\boldsymbol{K}_{\text{exp}}^{\text{T}}/\sqrt{d}\right)\boldsymbol{V}_{\text{exp}} \tag{4}$$

$$\text{Attention}_{\text{lm}} = \text{Softmax}\left(\boldsymbol{Q}_{\text{exp}}\boldsymbol{K}_{\text{lm}}^{\text{T}}/\sqrt{d}\right)\boldsymbol{V}_{\text{lm}} \tag{5}$$

式中，*Attention* 代表不同 token 之间的相关性权重，$1/\sqrt{d}$ 是用于适当归一化以防止极小梯度的缩放因子。

2.2 滑动膨胀窗口注意力

针对 Transformer 模型在捕捉面部标志点这种局部特征中展现出适应性不足的问题，在交叉融合框架的基础上针对面部标志特征提出了滑动膨胀窗口注意力机制(cross-fusion sliding dilated window attention，CFSDWA)如图 3 所示。

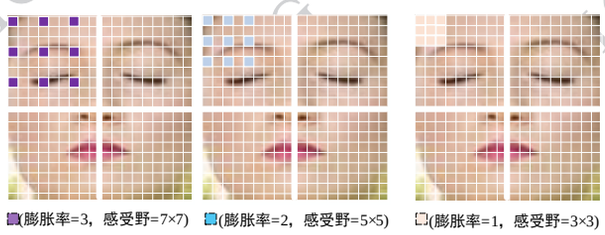


图 3 滑动膨胀窗口注意力机制

Figure 3 Sliding Expanding Window Attention Mechanism

尽管 Transformer 具备优越的全局建模能力，其标准的全局自注意力机制由于缺乏局部空间归纳偏置，在建模面部表情动态过程中，难以捕捉如眼、眉、口、鼻等关键区域的细节差异。为增强模型对面部关键区域的关注能力，在空间维度上对注意力机制的计算范围进行限制，仅在局部滑动膨胀窗口内执行自注意力操作，在 Transformer 全局建模优势的同时，使模型能更有效聚焦于面部的关键区域，从而显著提升对局部结构特征的关注能力。

滑动膨胀窗口机制以查询区域为中心构建局部滑动窗口，在空间维度上稀疏地选取关键区域内的键 $\boldsymbol{K}$ 和值 $\boldsymbol{V}$ 向量，并在这些具有代表性的局部 patch 上执行自注意力操作。通过限制注意力的作用范围，不仅提升了模型对局部细节特征的关注能力，同时有效抑制了来自非显著区域的冗余干扰，为特征融合过程中的结构引导提供了稳定的上下文支持。该机制的具体计算过程描述如公式(6)和(7)所示。

$$\boldsymbol{X}_{\text{exp}} = \text{CFSDWA}(\boldsymbol{Q}_{\text{lm}},\boldsymbol{K}_{\text{exp}},\boldsymbol{V}_{\text{exp}},r) \tag{6}$$

$$\boldsymbol{X}_{\text{lm}} = \text{CFSDWA}(\boldsymbol{Q}_{\text{exp}},\boldsymbol{K}_{\text{lm}},\boldsymbol{V}_{\text{lm}},r) \tag{7}$$

式中，CFSDWA(·)代表交叉融合的滑动膨胀窗口注意力块，r 代表在滑动膨胀窗口中可设置的膨胀率。

对于特征图中位置 (i,j) 处的查询，CFSDWA 以 (i,j) 为中心，大小为 w × w 的滑动膨胀窗口，并在该窗口内稀疏采样相应的 $\boldsymbol{K}$ 和 $\boldsymbol{V}$ 向量以执行自注意力计算。对于位置 (i,j)，CFSDWA 操作中输出 $\boldsymbol{X}$ 按照公式(8)和(9)计算：

$$\begin{aligned}\boldsymbol{X}_{\text{exp}(i,j)} &= \text{Attention}(\boldsymbol{Q}_{\text{lm}(i,j)},\boldsymbol{K}_{\text{exp}(r)},\boldsymbol{V}_{\text{exp}(r)})\\ &= \text{Soft max}\left(\boldsymbol{Q}_{\text{lm}(i,j)},\boldsymbol{K}_{\text{exp}(r)}^{T}/\sqrt{d}\right)\boldsymbol{V}_{\text{exp}(r)},\\ &\quad 1 \leqslant i \leqslant W,1 \leqslant j \leqslant H\end{aligned} \tag{8}$$

$$\begin{aligned}\boldsymbol{X}_{\text{lm}(i,j)} &= \text{Attention}(\boldsymbol{Q}_{\text{exp}(i,j)},\boldsymbol{K}_{\text{lm}(r)},\boldsymbol{V}_{\text{lm}(r)})\\ &= \text{Soft max}\left(\boldsymbol{Q}_{\text{exp}(ij)},\boldsymbol{K}_{\text{lm}(r)}^{T}/\sqrt{d}\right)\boldsymbol{V}_{\text{lm}(r)},\\ &\quad 1 \leqslant i \leqslant W,1 \leqslant j \leqslant H\end{aligned} \tag{9}$$

式中，H 和 W 是特征图的高度和宽度，$\boldsymbol{Q}_{\text{exp}(i,j)}$ 与 $\boldsymbol{Q}_{\text{lm}(i,j)}$ 代表表情特征和面部标志特征对位置 (i,j) 的查询向量，$\boldsymbol{K}_{\text{exp}(r)}$、$\boldsymbol{K}_{\text{lm}(r)}$、$\boldsymbol{V}_{\text{exp}(r)}$ 与 $\boldsymbol{V}_{\text{lm}(r)}$ 分别从表情特征和面部标志特征中提取对应的 $\boldsymbol{K}$ 和 $\boldsymbol{V}$ 向量。

CFSDWA 以查询中心对局部区域的 patch 进行特征聚合，促进表情语义与标志结构的深度协同。在查询操作过程中，若当前查询位置靠近特征图边缘，则采用零填充策略，以保证滑动膨胀窗口尺寸的一致性及全图范围内的可比性。在每个局部窗口内，通过稀疏采样面部标志特征中的键和值，引导模型有效关注标志结构的关键区域。

为了进一步验证 CFSDWA 的有效性，对其可视化效果进行了展示，如图 4 所示。面部标志点信息与整体特征之间的协同作用，能够有效引导模型更加准确地聚焦于表情判别的关键区域，从而增强对复杂表情变化的识别性能。

2.3 多级感受野检测器

为了进一步提升模型捕获眼角、嘴角等关键区域的能力，在滑动膨胀窗口内引入不同的膨胀率，构建了交叉融合多级感受野检测器，其整体结构如图 2 中所示。在经过交叉融合的 $\boldsymbol{Q}$、$\boldsymbol{K}$ 和 $\boldsymbol{V}$ 向量，被划分至多个滑动膨胀窗口注意力头中。在每个注意力头中，结合膨胀卷积策略以实现稀疏感受野调控。通过对每个注意力头设置不同的膨胀率，能够在多层次感受野范围内捕捉语义信息，每个注意力头仅

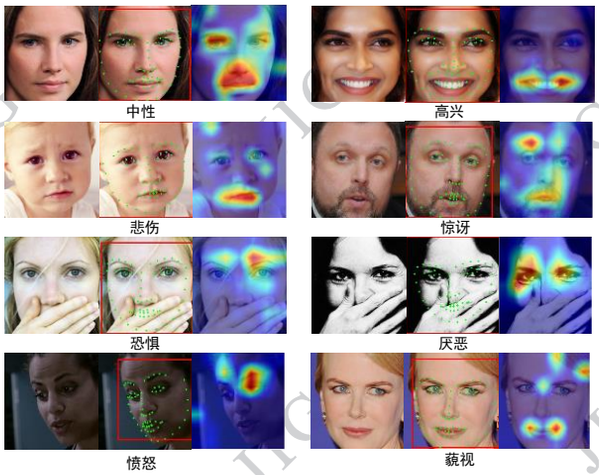


图4　CFSDWA的可视化

Figure 4　Visualization effect of CFSDWA

在其对应的感受野内执行局部注意力计算，从而实现对表情特征的全局上下文建模与面部标志特征的局部细节感知的有效融合，还通过滑动膨胀窗口的稀疏机制降低了冗余计算和参数量，有效避免了额外的计算开销。

多级感受野检测器具体操作是将注意力头分配至不同感受野大小的子空间，如图2中所示以红色查询块为中心的滑动膨胀窗口中，对彩色表示的键和值区域执行局部自注意力操作。各注意力头采用不同的膨胀率以调节其关注范围，分别对应3×3、5×5和7×7的感受野大小，具体通过核大小为3×3，膨胀率$r=1,2,3$实现，采用零填充策略在特征维度上逐位置滑动。最后将所有注意力头的特征进行拼接，并输入至线性投影层进行融合。对于交叉融合之后的$\boldsymbol{X}_{\text{exp}}$和$\boldsymbol{X}_{\text{lm}}$经过多级感受野的处理过程如公式(10-13)所示：

$$\boldsymbol{X}_{\text{exp}(i)}=\text{CFSDWA}(\boldsymbol{Q}_{\text{lm}},\boldsymbol{K}_{\text{exp}},\boldsymbol{V}_{\text{exp}},r_i),\quad 1\leqslant i\leqslant n \tag{10}$$

$$\boldsymbol{X}_{\text{exp}}=\text{Linear}(\text{Concat}[\boldsymbol{X}_{\text{exp}\_1},\cdots,\boldsymbol{X}_{\text{exp}\_n}]) \tag{11}$$

$$\boldsymbol{X}_{\text{lm}(i)}=\text{CFSDWA}(\boldsymbol{Q}_{\text{exp}},\boldsymbol{K}_{\text{lm}},\boldsymbol{V}_{\text{lm}},r_i),\quad 1\leqslant i\leqslant n \tag{12}$$

$$\boldsymbol{X}_{\text{lm}}=\text{Linear}(\text{Concat}[\boldsymbol{X}_{\text{lm}\_1},\cdots,\boldsymbol{X}_{\text{lm}\_n}]) \tag{13}$$

式中，CFSDWA(·)代表交叉融合滑动膨胀窗口注意力块，r_i是第i个头部的膨胀率，n代表头数，输出从1到n拼接在一起，发送到线性层进行特征聚合。膨胀注意力头分配策略：设置总的头数为H，标准注意力头占比为r，膨胀率集合为$\boldsymbol{D}=\{d_1,d_2,\cdots d_m\}$，则标准注意力头为$H_s=rH$，膨胀注意力头$H_d=H-H_s$，每个膨胀率分配的头数$h_d=H_d/m$，其中$m$代表膨胀率种类的个数，每组膨胀头仅在对应的稀疏感受野中执行注意力计算。

通过在滑动膨胀窗口注意力中设置不同的膨胀率构成最终的交叉融合多级感受野编码器(cross-fusion multi-level receptive-field encoder，CFMRFA Encoder)如图5所示。CFMRFA Encoder在表情特征和面部标志特征交叉融合的基础上，在每个滑动膨胀窗口注意力头中采用不同大小的膨胀卷积操作，使模型能够在多个感受野范围内灵活地捕捉不同层次的语义信息。较小的膨胀率有助于建模面部标志特征中的局部关键结构，较大的膨胀率则使模型能够聚焦于表情特征的全局上下文关系。在得到交叉融合的表情特征$\boldsymbol{X}_{\text{exp}}$和面部标志特征$\boldsymbol{X}_{\text{lm}}$的情况下，将$\boldsymbol{X}_{\text{exp}}$和$\boldsymbol{X}_{\text{lm}}$通过CFMRFA Encoder进行整合，最终的输出表示如公式(14-18)所示：

$$\boldsymbol{X}'_{\text{exp}}=\text{CFMRFA}(\boldsymbol{Q}_{\text{lm}},\boldsymbol{K}_{\text{exp}},\boldsymbol{V}_{\text{exp}},r_i)+\boldsymbol{X}_{\text{exp}},1\leqslant i\leqslant n \tag{14}$$

$$\boldsymbol{X}_{\text{exp\_out}}=\text{MLP}(\text{Norm}(\boldsymbol{X}'_{\text{exp}}))+\boldsymbol{X}'_{\text{exp}} \tag{15}$$

$$\boldsymbol{X}'_{\text{lm}}=\text{CFMRFA}(\boldsymbol{Q}_{\text{exp}},\boldsymbol{K}_{\text{lm}},\boldsymbol{V}_{\text{lm}},r_i)+\boldsymbol{X}_{\text{exp}},1\leqslant i\leqslant n \tag{16}$$

$$\boldsymbol{X}_{\text{lm\_out}}=\text{MLP}(\text{Norm}(\boldsymbol{X}'_{\text{lm}}))+\boldsymbol{X}'_{\text{lm}} \tag{17}$$

$$\mathbf{X}_{\text{fuse\_out}}=\mathbf{X}_{\text{exp\_out}}+\mathbf{X}_{\text{lm\_out}} \tag{18}$$

式中，CFMRFA(·)表示交叉融合多级感受野的MSA块，Norm(·)是归一化算子，MLP(·)表示多层感知机。

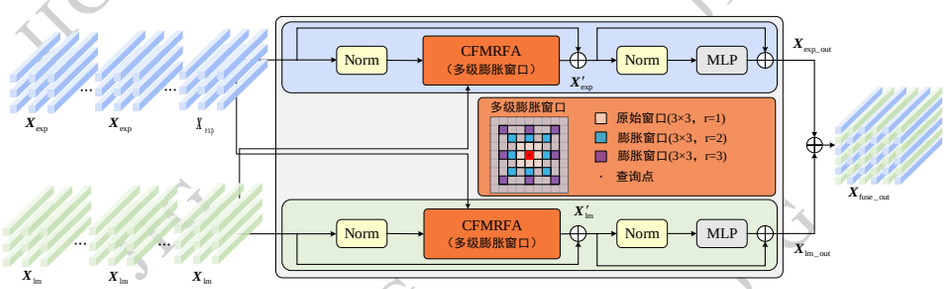


图5　交叉融合多级感受野Transformer编码器(Norm(·)是归一化算子；MLP(·)表示多层感知机)

Figure 5　cross-fused multi-level receptive field transformer encoder(Norm(·)is the normalization operator; MLP(·)B is multi-layer perceptron)

3 实验结果与分析

为了验证所提方法在面部表情识别任务中的有效性。首先介绍所使用的数据集,以明确实验的评测基础。随后给出实验设置与实现细节,以保证方法的可复现性。在此之上,呈现与现有先进方法的对比结果,验证模型的整体性能提升,并进一步通过模型分析挖掘关键模块在复杂表情建模中的作用。最后结合消融实验,定量评估各组件对最终性能的贡献。

3.1 数据集

RAF-DB (Li 等,2017)是一个真实世界的面部表情识别数据集,包含29,672张多样化人脸图像,涵盖年龄、性别、种族、姿态、光照和遮挡等变化。其基本情绪子集包括15,339张图像,分为12,271张训练集和3,068张测试集,标注有七类基本情绪:快乐、惊讶、悲伤、愤怒、厌恶、恐惧和中性。

AffectNet (Ali Mollahossein 等,2017)是目前最大的面部表情数据集,含有超过100万张人脸图像,其中约45万张带有11类情绪的人工标签。在本研究中采用7类和8类分类设置,7类:使用287,651张图像训练,7类为基本情绪;8类:在原7类基础上加入蔑视。训练集存在类别不平衡,测试集为每类500张图像的严格平衡集(共4000张)。

FERPlus (Emad Barsou 等,2016)是对FER2013数据集的增强版本,包含35,887张48×48的灰度人脸图像(28,709训练、3,589验证、3,589测试)。其标签由多个注释者重新标注,涵盖8类情绪:快乐、惊讶、悲伤、愤怒、厌恶、恐惧、中性和蔑视。评估指标为测试集的整体准确率。

3.2 实验细节

所有实验基于PyTorch框架实现,并在NVIDIA RTX 3090 GPU内存为24GB上进行训练与测试。训练阶段batch_size为200,epoch为300,学习率调度采用指数衰减策略,初始学习率设为4×10^{-5},衰减系数gamma=0.98。优化器采用SAM,基优化器为Adam,rho=0.05,通过两次前向和后向传播寻找平坦极小值。

为缓解类别不平衡问题,损失函数采用分类交叉熵与标签平滑交叉熵的加权组合。训练中使用随机水平翻转与随机擦除数据增强方法,输入图像统一调整为224 × 224像素,并使用ImageNet均值与标准差进行归一化。表情分支的主干网络为IR-50 (Deng 等,2019),面部标志分支采用外部检测器MobileFaceNet (Chen 等,2021)用于特征的提取,两者均在MS-Celeb-1M (Guo 等,2016)上预训练。

3.3 实验结果

本文将所提出的方法与近几年17种代表性的方法进行了比较,包括SCN(suppressing uncertainties for large-scale)(Wang 等,2020b)、PSR(pyramid with super resolution)(Thanh-Hung 等,2020)、RAN(Region attention networks)(Wang 等,2020a)、DACL(deep attentive center loss)(Farzaneh 等,2020)、KTN(Adaptively learning facial expression representation via c-f labels and distillation)、DMUE(latent distribution mining and pairwise uncertainty estimation)(She 等,2021)、FDRL(Feature decomposition and reconstruction learning)(Ruan 等,2021)、TransFER(learning relation-aware facial expression representations with transformers)(Xue 等,2017)、AMP-Net(Adaptive multilayer perceptual attention network)(Liu 等,2022)、Face2Exp(combating data biases for facial expression recognition)(Zeng 等,2022)、EAC(erasing attention consistency)(Zhang 等,2022)、EDGL-FLP(Enhanced discriminative global

-local feature learning with priority)(Zhang 等,2023)

FER-former (Multi-modal transformer)(Li 等,2023)、CLIPER(A unified vision-language framework for)(Li 等,2024)、MMATRans(Muscle movement aware representation learning)(Liu 等,2024)、Expllm(towards chain of thought)(Lan 等,2025)。

在RAF-DB数据集上的对比如表1所示,展示了CFMRFN与RAF-DB数据集上其他方法的比较。在整体准确率上均优于其他SOTA方法。其中,CFMRFN实现了92.14%的最高准确率,比次优方法CLIPER(Li 等,2024)提高了0.53%,充分验证了所提出方法在该数据集上的有效性和竞争优势。

在AffectNet数据集上的实验结果如表1所示。由于其类别分布严重平衡,对模型的泛化能力提出了高要求。在七类表情分类任务中,CFMRFN的准确率达到了67.35%,相比CLIPER 提升了1.06%;在八类表情分类任务中,其性能较 Expllm(Lan 等,

2025)提高了0.58%,所提出方法在大规模且类别不平衡的数据场景能够保持相对稳定的识别性能。

在FERPlus数据集上的结果如表1所示,所提方法依然达到了91.67%的准确率,尽管FERPlus中的所有图像均为分辨率48×48的灰度图,比次优方法FER-former(Li等,2023)提高了0.71%。说明模型在低分辨率和信息受限条件下仍具有较强的判别能力。

表1　RAF-DB、AffectNet和FERPlus数据集的比较

Table 1　Comparison of the RAF-DB, AffectNet and FERPlus datasets

方法	期刊	RAF-DB/ Acc %	AffectNet/ Acc(7cls) %	AffectNet/ Acc(8cls) %	FERPlus/ Acc %
SCN(Wang等,2020b)	CVPR	87.03	60.23	60.28	89.39
PSR(Thanh-Hung等,2020)	CVPR	88.98	63.77	60.68	–
RAN(Wang等,2020a)	TIP	86.90	–	–	89.16
DACL(Farzaneh等,2020)	WACV	87.78	65.20	–	–
KTN(Li等,2021)	TIP	88.07	63.97	–	90.49
DMUE(She等,2021)	CVPR	89.42	63.11	–	–
FDRL(Ruan等,2021)	CVPR	89.47	–	–	–
TransFER(Xue等,2017)	ICCV	90.91	66.23	–	90.83
AMP-Net(Liu等,2022)	TCSVT	89.19	61.32	61.74	89.37
Face2Exp(Zeng等,2022)	CVPR	88.54	64.23	–	–
EAC(Zhang等,2022)	ECCV	89.99	65.32	–	89.64
EDGL-FLP(Zhang等,2023)	INS	89.90	–	61.25	–
FER-former(Li等,2023)	arXiv	91.30	–	–	90.96
CLIPER(Li等,2024)	ICME	91.61	66.29	61.98	–
MMATRans(Liu等,2024)	TII	89.67	64.89	–	90.32
Expllm(Lan等,2025)	TMM	91.03	65.93	62.86	–
CFMRFN	–	**92.14**	**67.35**	**63.44**	**91.67**

注:加粗字体表示各列最优结果,"_"表示次优结果,"–"表示无实验数据,"cls"是分类的缩写。

表2　RAF-DB、AffectNet和FERPlus数据集单类准确性

Table 2　Comparison of single-class accuracy of RAF-DB, AffectNet and FERPlus datasets

数据集	方法	中性	高兴	悲伤	惊讶	恐惧	厌恶	愤怒	蔑视	平均准确率
RAF-DB 7cls/%	CFMRFN	**93.23**	**97.13**	90.58	**90.88**	66.21	**71.87**	**90.12**	–	**85.71**
	交叉融合	91.61	96.62	90.58	90.68	**66.31**	65.00	89.50	–	84.32
AffectNet 7cls/%	CFMRFN	**67.33**	**88.17**	67.01	**65.56**	65.00	**56.28**	**63.61**	–	**67.56**
	交叉融合	66.12	87.08	65.41	65.55	**65.12**	56.00	59.46	–	66.39
FERPlus 8cls/%	CFMRFN	**94.19**	96.18	**94.72**	**80.67**	**90.77**	**53.33**	61.53	**46.15**	**77.19**
	交叉融合	93.08	**96.52**	93.46	80.41	89.29	43.66	61.53	40.13	74.76

注:加粗字体表示每类表情在不同数据集上的最优结果,"–"表示无实验数据,"cls"是分类的缩写。

为了进一步展现CFMRFN对于复杂表情的分辨能力,对所提方法进行了单类别分析。在表2中,展示了CFMRFN与单独的交叉融合在RAF-DB、FERPlus和AffectNet数据集上的性能对比。如表2

所示，在RAF-DB数据集中，CFMRFN在中性、快乐、惊讶和愤怒四个类别上的准确率均超过90%。然而，恐惧类别的准确率相对较低，原因在于该类别的训练样本数量较少（仅281张恐惧表情，而其他类别如中性、快乐、悲伤和惊讶则超过1,000张图像）。在FERPlus数据集中，厌恶和蔑视这两个类别的准确率也较低，这同样是由于这两个类别的训练样本稀缺（分别只有116和120张，远少于其他类别）。对于AffectNet数据集，所有样本均来自互联网，并且大部分训练和测试样本来自野外环境。导致AffectNet的性能远低于RAF-DB和FERPlus数据集上的结果。

3.4 模型分析

为了评价所提方法中Transformer模型深度以及注意力头分配策略对模型性能的影响，本文在RAF-DB数据集上对所提方法中的Transformer模块进行了分析。通过设置不同的网络深度和注意力头数量，对模型在表情识别任务中的性能变化进行了对比实验，为Transformer参数配置的合理选择提供依据。

1）对所提出的Transformer模型进行了深度敏感性分析，探讨不同网络深度下的面部表情识别性能。评估集成多级感受野机制的Transformer模型在深度为{2,3,4,5,6,7,8}时的识别效果。如表3所示，随着模型深度的增加性能逐步提升，并在深度为6时达到最优。然而，继续加深网络结构（深度为7和8）时，模型性能反而略有下降。这一现象表明，过深的Transformer网络并不一定带来性能提升，过多的网络层可能引入冗余建模甚至导致一定程度的过拟合，从而影响模型的整体性能。

2）在Transformer深度固定为6的前提下分析对注意力头分配不同膨胀率时对模型性能的影响，结果如表4所示。当8个注意力头全部分配为小感受野{8,0,0}分配策略仅达到91.43%的识别准确率，表明单一尺度的感受野在特征提取方面存在一定局限。相比之下，将注意力头分配为{0,4,4}时，准确率提升至91.79%，体现出中大感受野在建模全局和局部上下文中的优势。然而，在{4,2,2}的配置下，模型达到91.98%的准确率，验证了小感受野对局部细节建模的贡献并不是很大。当注意力头配置为{2,3,3}时模型性能提升至最佳，达到了92.14%的识别准确率，表明融合不同尺度的感受野信息，对于提升面部表情的识别能力具有显著效果。

表3 Transformer模型在深度为{2,3,4,5,6,7,8}的识别效果

Table 3 The recognition effect of the proposed Transformer model at depths of {2,3,4,5,6,7,8}

模型深度	RAF-DB/Acc %	模型总参数/M
8	91.75	61.03
7	91.69	57.37
6	**92.14**	**53.68**
5	91.69	49.99
4	91.56	46.30
3	91.92	42.61
2	91.69	38.92

注：加粗字体表示模型深度最优的结果。

表4 深度为6时CFMRFN注意力头数的分配策略

Table 4 Allocation strategies of attention heads in CFMRFN at depth 6

注意力总头数=8			RAF-DB/Acc %
小感受野×8	中感受野×0	大感受野×0	91.43
小感受野×0	中感受野×4	大感受野×4	91.79
小感受野 2	**中感受野 3**	**大感受野 3**	**92.14**
小感受野×4	中感受野×2	大感受野×2	91.98

注：加粗字体表示感受野分配最优的结果。

3）为验证不同注意力机制对模型性能与计算效率的影响，对全局注意力、滑动窗口注意力及滑动膨胀窗口注意力进行了对比实验，结果如表5所示。从模型复杂度来看，全局注意力需在所有特征token间建立关系，参数量最高52.8M。滑动窗口通过限制注意力范围显著降低复杂度，而滑动膨胀窗口在局部建模基础上引入膨胀策略，使参数量与FLOPs进一步降至41.8M和6.8G。从识别性能来看，滑动膨胀窗口注意力在RAF-DB和AffectNet上均取得最佳结果92.14和67.35，同时优于滑动窗口与全局注意力。表5的结果表明，全局建模易引入冗余背景关系，而纯局部窗口限制跨区域信息交互。滑动膨胀窗口在保持低计算复杂度的同时扩大有效感受野，实现了局部细节与跨区域语义的平衡，从而获得更优识别性能。

表5　不同注意力下的对比结果

Table 5　Comparison results under different attention

注意力机制	参数量(Param)	计算量(FLOPs)	RAF-DB Acc%	AffectNet(7cls) Acc%
全局注意力	51.8M	10.3G	91.24	66.80
滑动窗口	43.7M	8.4G	91.43	67.12
滑动膨胀窗口	**41.8M**	**6.8G**	**92.14**	**67.35**

注：加粗字体表示模型参数最优的结果。

3.5　消融实验

为了验证所提模型中各部分的有效性，在RAF-DB和AffectNet数据集上进行消融分析，所有模型均采用相同的实验设置，结果如表6所示，并在RAF-DB和AffectNet数据集上进行可视化分析。

表6　在RAF-DB、AffectNet数据集上对每个组件进行消融研究

Table 6　Ablation study of each component on RAF-DB and AffectNet datasets

方法	RAF-DB/Acc%	AffectNet/Acc%	
	7cls	7cls	8cls
CFMRFN	**92.14**	**67.35**	**63.44**
w/o多级感受野	91.43	67.12	62.04
w/o 滑动膨胀窗口	91.24	66.80	61.51
w/o 交叉融合	90.26	65.35	61.12
单一面部表情	88.14	65.95	54.70
单一面部标志	80.18	48.78	45.34

注：加粗字体表示模型最优的结果，w/o代表没有。

1）为了验证所提出的交叉融合多级感受野模块在表情识别任务中的有效性，构建了仅具备单一原始感受野的滑动膨胀窗口注意力版本的Transformer模型，并在相同设置下进行对比。根据表6中的实验结果，当移除多级感受野设计后，模型在RAF-DB数据集上的准确率下降了0.71%，在AffectNet的两类任务中分别下降了0.23%和1.40%。该结果表明，多级感受野的引入有助于模型更全面地感知面部关键区域，从而提升识别性能。

2）为了验证提出的滑动膨胀窗口的有效性，将所提出的滑动膨胀窗口机制替换为标准的自注意力，以评估滑动膨胀窗口的贡献。替换后模型在RAF-DB与AffectNet数据集上的性能均有所下降，表明滑动膨胀窗口在提升FER模型对面部关键区域的感知能力方面发挥了显著作用，增强了整体表达的判别性。

3）为了验证提出的交叉融合策略的有效性，将表情特征与面部标志特征在编码器外部融合，通过两个独立的Transformer编码器处理后进行特征求和。实验结果显示，在RAF-DB数据集上，该对照方法的性能比交叉融合策略低0.98%，交叉融合机制通过结构引导增强了模型对面部显著区域的关注能力，在应对类间相似性高与类内差异性大等挑战中发挥了关键作用。

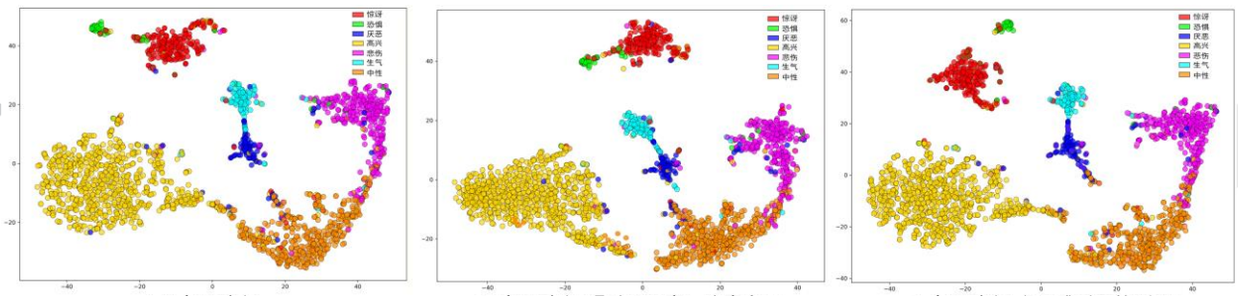
4）为验证表情主干网络IR-50与面部标志主干网络MobileFaceNet的有效性，分别采用Transformer对单一表情特征和单一面部标志特征进行建模与评估。表6中的结果表明，在仅使用单一特征的情况下，模型识别精度均低于两类特征融合时的结果，说明面部表情与面部标志特征具有互补性，二者结合能够显著提升表情识别的性能。

5）为了验证所提方法中各个组件的有效性，采用t-SNE（Maaten等，2008）对RAF-DB数据集中经过不同组件的高维特征可视化如图6所示，通过观察各个组件特征空间中的分布情况，对于区分不同表情之间特征的相似性以及同类表情之间的差异性都得到了提升，其中(b)是将注意力头的感受野全部设置为3×3的可视化结果。

6）为了进一步澄清单个表情的类别信息，展示了面部标志点和CFSDWA的可视化效果。从AffectNet数据集中选取了8个类别的面部图像，并通过图7呈现了这些面部图像在交叉融合阶段的高级特征注意力可视化结果。图像中的面部标志特征成功地帮助捕捉到各类表情的关键面部特征，从而有效地检测到了这些表情中的关键特征。

4　结　论

本文提出了用于面部表情识别的交叉融合多级感受野网络（CFMRFN），解决面部表情类间相似性高和类内差异性大的问题。所提方法通过交叉融合Transformer编码器实现面部标志特征和表情特征的交互学习，利用面部标志特征来引导模型关注面部的关键区域，有效提升了模型对面部关键区域的关



(a)交叉融合　(b)交叉融合+滑动膨胀窗口注意力　(b)交叉融合+多级感受野检测器

图 6　使用t-SNE在RAF-DB数据集对各个组件进行高维特征的可视化((a)交叉融合;(b)交叉融合+滑动膨胀窗口注意力;(c)交叉融合+多级感受野检测器)

Figure 6　t-SNE visualization of high-dimensional feature distributions on the RAF-DB dataset under different component configurations((a) cross-fusion; (b) cross-fusion+sliding dilated window attention; (c) cross-fusion+multi-level receptive field detector)

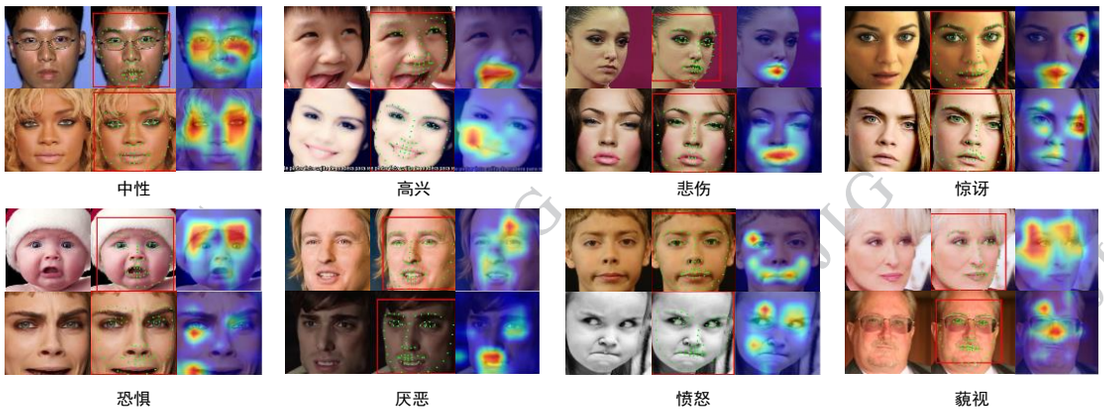


图7　面部标志点和滑动膨胀窗口自注意力可视化

Figure 7　Visualization of facial landmarks and sliding-dilated window attention

注能力和对复杂表情特征的建模能力。其次,构建滑动膨胀窗口来限制Transformer注意力计算在局部范围内进行,使模型更专注于关键区域的细节建模。最后,在注意力模块中设置不同膨胀率,构建多尺度感受野检测器,从而兼顾面部关键结构感知与全局上下文表达。实验结果表明,CFMRFN在RAF-DB、AffectNet和FERPlus等多个公开数据集上均取得了优异性能,验证了所提方法在提高类间区分能力与增强类内表达一致性方面的有效性。

尽管所提方法在识别精度与特征建模上表现突出,但对外部人脸标志检测和预训练模型的依赖降低了野外遮挡场景的性能,且膨胀率人工设定缺乏自适应能力。未来将通过自监督结构搜索与可学习感受野模块,提升系统与样本的自适应性。

参考文献(References)

Li S and Deng W H. 2020. Deep facial expression recognition: A survey. IEEE Transactions on Affective Computing, 13(3): 1195-1215[DOI: 10.1109/TAFFC.2020.2981446]

Ekman P and Friesen W V. 1971. Constants across cultures in the face and emotion. Journal of Personality and Social Psychology, 17(2): 124-129[DOI: 10.1037/h0030377]

Zhong L, Bai C M, Li J F, Chen T, Li S G and Liu Y G. 2019 .A graph structured representation with brnn for static-based facial expression recognition//Proceedings of 2019 IEEE International Conference on Automatic Face and Gesture Recognition. Lille, France: IEEE: 1 - 5 [DOI: 10.1109/FG.2019.8756615]

Song T F, Cui Z J, Wang Y R, Zheng W M, and Ji Q. 2021. Dynamic probabilistic graph convolution for facial action unit intensity estimation//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. California, United States: IEEE: 4845-4854[DOI: 10.1109/CVPR46437.2021.00481]

Ruan D L, Yan Y, Chen S, Xue J H, and Wang H Z. 2020. Deep disturbance-disentangled learning for facial expression recognition//Proceedings of 2020 28th ACM International Conference on Multimedia, Online, ACM: 2833-2841 [DOI: 10.1145/3394171.3413907]

Li G B, Zhu X, Zeng Y R, Wang Q and Lin L. 2019 .Semantic relationships guided representation learning for facial action unit recognition//Proceedings of 2019 AAAI Conference on Artificial Intelligence. Macau, China: AAAI: 8594-8601 [DOI: 10.1609/aaai.v33i01.33018594]

Zhang Z, Wang T Y, and Yin L J. 2020. Region of interest based graph convolution: A heatmap regression approach for action unit detection//Proceedings of 2020 ACM International Conference on Multimedia. Beijing, China: ACM: 2890-2898 [DOI:10.1145/3394171.3413674]

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł and Polosukhin I. 2017. Attention is all you need//Proceedings of the 2017 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc:6000-6010 [DOI:10.48550/arXiv.1706.03762]

Fisher Yu. and Vladlen Koltun. 2015. Multi-scale context aggregation by dilated convolutions. [EB/OL]. [2015-11-07]. https://arxiv.org/pdf/1511.07122.pdf

Peng X J and Qiao Y. 2020. Advances and challenges in facial expression analysis. Journal of Image and Graphics, 25(11):2337-2348 (彭小江,乔宇.2020.面部表情分析进展和挑战.中国图象图形学报, 25(11):2337-2348) [DOI:10.11834/jig.200308]

Zhou Y, Xue H, and Geng X. 2015 .Emotion distribution recognition from facial expressions//Proceedings of 2015 ACM International Conference on Multimedia, Brisbane. Queensland, Australia: ACM :1247-1250[DOI: 10.1145/2733373.2806328]

Li S and Deng W H. 2018a .Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. Transactions on Image Processing, 28(1):356-370 [DOI:10.1109/TIP.2018.2868382]

Cai J, Meng Z B, Ahmed Shehab Khan, Li Z Y, James O'Reilly and Tong Y. 2018. Island loss for learning discriminative features in facial expression recognition//Proceeding of 2018 IEEE International Conference on Automatic Face & Gesture Recognition. Istanbul,Turkey: IEEE: 302-309 [DOI:10.1109/fg. 2018.00051]

Dinh Viet Sang, Le Tran Bao Cuong ,Pham Thai Ha. 2018. Discriminative deep feature learning for facial emotion recognition//Proceeding of 2018 1st International Conference on Multimedia Analysis and Pattern Recognition. Vietnam : IEEE: 1-6 [DOI:10.1109/MAPR.2018.8337514]

Huang G, Liu Z, Laurens Van Der Maaten and Kilian Q Weinberger. 2017 Densely connected convolutional networks//Proceeding of 2017 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington, United States: IEEE: 4700-4708 [DOI:10.1109/CVPR.2017.243]

Xie S Y, Hu H F, and Wu Y B. 2019. Deep multi-path convolutional neural network joint with salient region attention for facial expression recognition. Pattern Recognition, 92 : 177-191 [DOI: 10.1016/j.patcog.2019.03.019]

Wang K, Peng X J, Yang J F, Meng D B and Qiao Y. 2020a. Region attention networks for pose and occlusion robust facial expression recognition. IEEE Transactions on Image Processing, 29: 4057-4069 [DOI:10.1109/TIP.2019.2956143]

Amir Hossein Farzaneh and Qi X J. 2021 .Facial expression recognition in the wild via deep attentive center loss// Proceedings of 2021 IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, HI, USA: IEEE: 2402-2411 [DOI: 10.1109/wacv48630.2021.00245]

Ruan D L, Yan Y, Lai S Q, Chai Z H, Shen C H and Wang H Z. 2021. Feature decomposition and reconstruction learning for effective facial expression recognition// Proceedings of 2021 IEEE /CVF International Conference on Computer Vision and Pattern Recognition. Washington , United States: IEEE: 7660-7669 [DOI: 10.48550/arXiv.2104.05160]

Xue F L, Wang Q C, Guo G. 2021.Transfer: learning relation-aware facial expression representations with transformers//Proceeding of 2017 IEEE/CVF Conference on Computer Vision. Venice, Italy: IEEE: 3601-3610 [DOI:10.48550/ arXiv. 2108.11116]

Wang J D, Sun K, Cheng T H, Jiang B R, Deng C R, Zhao Y, Liu D, Mu Y D, Tan M K, Wang X G, Liu W Y and Xiao B.2019. Deep high-resolution representation learning for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43.10: 3349-3364 [DOI:10.1109/TPAMI.2020.2983686]

Wu Y, Ji Q.2019. Facial landmark detection: A literature survey. International Journal of Computer Vision, 127(2): 115-142. [DOI: 10.48550/arXiv.1805.05563]

Fuzail Khan. 2018. Facial expression recognition using facial landmark detection and feature extraction via neural networks. [EB/OL]. [2018-12-074]. https://arxiv.org/pdf/1812.04510.pdf

Jin H B, Liao S C, and Shao L.2021. Pixel-in-pixel net: Towards efficient facial landmark detection in the wild. International Journal of Computer Vision, 129.12: 3174-3194 [DOI:10.48550/arXiv.2003.03771]

Jung H, Lee S, Yim J, Sunjeong Park and Junmo Kim.2015. Joint fine-tuning in deep neural networks for facial expression recognition// Proceeding of 2015 IEEE/CVF International Conference on Computer Vision. Santiago, Chile: IEEE: 2983-2991 [DOI: 10.1109/iccv.2015.341]

Xiong K H, Qing L B, Li L D, Guo L and Peng Y H. 2024. Facial expression recognition based on local - global information reasoning and spatial distribution of landmark features. The Visual Computer, 41.1: 535-548 [DOI: 10.1007/s00371-024- 03345-y]

Huang Z W, Yu Z, Li H Y and Yang D W.2024. Dynamic facial expression recognition based on spatial key-points optimized region fea-

ture fusion and temporal self-attention. Engineering Applications of Artificial Intelligence, 133: 108535 [DOI: 10.1016/j. engappai. 2024.108535]

Chen S K, Wang J F, Chen Y D, Shi Z C, Geng X and Rui Y. 2020. Label distribution learning on auxiliary label space graphs for facial expression recognition//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Online: IEEE: 13984-13993 [DOI:10.1109/cvpr42600.2020.14 00]

Thanh-Hung Vo, Guee-Sang Lee, Hyung-Jeong Yang and Soo-Hyung Kim. 2020. Pyramid with super resolution for in-thewild facial expression recognition. IEEE Access, 8:131988-132001 [DOI:10.1109/ACCESS.2020.3010018]

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X H, Thomas U, Mostafa D, Matthias M, Georg H, Sylvain G, Jakob U and Neil H. 2021. An image is worth 16x16 words: Transformers for Image Recognition at Scale. [EB/OL]. [2021-06-03]. https://arxiv.org/abs/2010.11929

Mouath Aouayeb, Wassim Hamidouche, Catherine Soladie, Kidiyo Kpalma and Renaud Seguier. 2021.Learning vision transformer with squeeze and excitation for facial expression recognition. [EB/OL]. [2021-07-07]. https://arxiv.org/pdf/2107.03107.pdf

Cui, K W, Wu R L, Zhan F N amd Lu S J. 2023. Face transformer: Towards high fidelity and accurate face swapping// Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 668-677 [DOI: 10.1109/cvprw 59228.2023.00074]

Cui X Y, He C, Zhao H K and Wang M L. 2024. Combining ViT with contrastive learning for facial expression recognition. Journal of Image and Graphics, 29 (01):0123-0133 (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): 0123-0133) [DOI: 10. 11834/jig. 230043]

Dan J, Liu Y, Xie H Y, Deng J K, Xie H R, Xie X S and Sun B G. 2023. Transface: Calibrating transformer training for face recognition from a data-centric perspective//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 20642-20653 [DOI:10.1109/iccv51070. 2023.01887]

Xu R, Huang A B, Hu Y J and Feng Xibo. 2023. HFFT: Hierarchical Global-local Feature fusion transformers for facial expression recognition in the wild. Image and Vision Computing, 139: 104824 [DOI:16/j.imavis.2023.104824]

Yang Y J, Hu L, Zu C, Zhang J J, Hou Y, Chen Y , Zhou J L, Zhou L P and Wang Y. 2024. CL-TransFER: Collaborative learning based transformer for facial expression recognition with masked reconstruction. Pattern Recognition, 156: 110741 [DOI: 16/ j.imavis.2023.104824]

Li S, Deng W H and Du J P. 2017. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild// Proceedings of 2017 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Hawaii, United States: IEEE: 2584-2593 [DOI: 10.1109/CVPR.2017.277]

Ali Mollahosseini, Behzad Hasani and Mohammad H Mahoor. 2017. Affectnet: A database for facial expression, valence, and arousal computing in the wild. IEEE Transactions on Affective Computing, 10(1):18-31 [DOI:10.1109/CVPR.2017.277]

Emad Barsoum, Zhang C, Cristian Canton Ferrer and Zhengyou Zhang. 2016. Training deep networks for facial expression recognition with crowd-sourced label distribution//Proceedings of 2016 ACM International Conference on Multimodal Interaction. Tokyo, Japan: IEEE: 279-283 [DOI:10.1145/299314 8.2993165]

Deng J K, Guo J, Xue N N and Stefanos Zafeiriou. 2019. Arcface: Additive angular margin loss for deep face recognition//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. California, United: IEEE: 4690-4699 [DOI: 10.1109/TPAMI.2021.3087709]

Chen C J. 2021. PyTorch Face Landmark: A fast and accurate facial landmark detector, 2021.

Guo Y D, Zhang L, Hu Y X, He X D and Gao J F. 2016. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition//Proceedings of 2016 European Conference on Computer Vision. Amsterdam, Netherlands: IEEE: 87-102 [DOI: 10.1007/978-3-319-46487-96]

Wang K, Peng X J, Yang J F, Lu S J and Qiao Y. 2020b. Suppressing uncertainties for large-scale facial expression recognition //Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Online: IEEE: 6897-6906 [DOI: 10.1109/CVPR42600.2020.00693]

Li H Y, Wang N N, Ding X P, Yang X and Gao X B.2021. Adaptively learning facial expression representation via c-f labels and distillation. IEEE Transactions on Image Processing, 30: 2016-2028 [DOI:10.1109/TIP.2021.3049955]

She J H, Hu Y B, Shi H L, Wang J, Shen Q and Mei T. 2021. Dive into ambiguity: latent distribution mining and pairwise uncertainty estimation for facial expression recognition //Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Online: IEEE: 6248-6257 [DOI:10.48550/arXiv.2104.00232]

Liu H W, Cai H L, Lin Q C and Li X F. 2022. Adaptive multilayer perceptual attention network for facial expression recognition. IEEE Transactions on Circuits and Systems for Video Technology, 32 (9): 6253-6266 [DOI:10.1109/tcsvt.2022.3165 321]

Zeng D, Lin Z Y, Yan X, Liu Y T, Wang F and Tang B. 2022. Face2exp: Combating data biases for facial expression recognition// Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Online: IEEE: 20291-20300 [DOI: 10.1109/cvpr52688.2022.01965]

Zhang Y , Wang C , Ling X , Deng W.2022.Learn From All: Erasing

Attention Consistency for Noisy Label Facial Expression Recognition//Proceedings of 2022 European Conference on Computer Vision. Tel Aviv, Israel: IEEE: 418-434 [DOI: 10. 48550arXiv.2207.10299]

Zhang Z Y, Tian X, Zhang, Guo K L and Xu X M. 2023. Enhanced discriminative global-local feature learning with priority for facial expression recognition. Information Sciences, 630: 370-384 [DOI: 10.1016/j.ins.2023.02.056]

Li Y D, Wang M J, Gong M L, Lu Y G and Liu L.2023. Fer-former: Multi-modal transformer for facial expression recognition [EB/OL]. [2023-03-23]. https://doi.org/10.48550/arXiv.2303.12997

Li H T, Niu H J, Zhu Z Q and Zhao F. 2024. Cliper: A unified vision-language framework for in-the-wild facial expression recognition. IEEE International Conference on Multimedia and Expo, 15: 1-6 [DOI: 10.48550/arXiv.2303.00193]

Liu H, Zhou Q Y, Zhang C, Liu T T, Zhang Z L and Li Y F.2024 MMATrans: Muscle movement aware representation learning for facial expression recognition via transformers. IEEE Transactions on Industrial Informatics, 20(12): 13753-13764 [DOI: 10.1109/TII.2024.3431640]

Lan X, Xue J, Qi J, Jiang D M, Lu K and Chua T S.2025. Expllm: Towards chain of thought for facial expression recognition. IEEE Transactions on Multimedia, 29: 3069-3081 [DOI: 10.1109/TMM.2025.3557704]

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. Journal of Machine Learning Research, 9: 2579-2605.

作者简介

李策，通信作者，男，教授，博士生导师，主要研究方向为计算机视觉与模式识别、医学图像处理。 E-mail：lice@lut.edu.cn

王凯，男，硕士研究生，主要研究方向为基于视觉的睡眠状态监测与质量评估。 E-mail：232085406051@lut.edu.cn

肖利梅，女，正高级工程师，硕士生导师，主要研究方向为智能信息处理，智能机器人。 E-mail：xlm@lut.edu.cn

王茹，女，硕士研究生，主要研究方向为基于深度学习的睡眠信号分期算法研究。 E-mail：ruwang@lut.edu.cn

平梦梦，女，博士研究生，主要研究方向为智能信息处理，机器人控制。 E-mail：mmping@lut.edu.cn

卢明，男，正高级工程师，主要研究方向空间信息与数字技术、人工智能技术。 E-mail：99168704@qq.com