刘灵儿, 晏峻峰. LEFA-ConvNeXt:基于光照增强与可学习面部先验注意力的表情识别网络[J]. 智能计算机与应用,2026,16(7):10-17. DOI:10.20169/j.issn.2095-2163.26010502

LEFA-ConvNeXt:基于光照增强与可学习面部先验注意力的表情识别网络

刘灵儿, 晏峻峰
(湖南中医药大学 信息科学与工程学院, 长沙 410208)

摘 要: 随着深度学习技术的发展,面部表情识别(Facial Expression Recognition, FER)在情感计算、人机交互、智驾等领域得到了广泛应用。然而,受光照变化、表情细微差异及局部特征模糊等因素影响,现有方法在复杂环境下的识别性能仍存在不足。为此,本文提出了一种基于光照增强与可学习面部先验注意力模块的表情识别网络。该网络以 ConvNeXt 为主干结构,引入了光照增强模块(Light Enhance Block)以自适应调整图像亮度,提升模型在非均匀光照条件下的鲁棒性。同时设计了可学习面部先验注意力模块(Learnable Facial Prior Attention),通过可学习的面部掩模机制引导网络聚焦于关键表情区域(如眼睛、嘴角等),从而增强细粒度特征的辨识能力。为了进一步融合局部上下文信息,模型在高层特征阶段引入了局部上下文自注意力模块(Local Context Self-Attention Block)进行窗口级自注意力计算,实现全局与局部特征的协同建模。在 FER-2013 公开数据集上的实验结果表明,所提出的 LEFA-ConvNeXt 网络在识别准确率和泛化性能方面均优于现有主流方法,最高准确率达到 74.78%,充分验证了该方法在复杂光照条件与多样化表情场景下的有效性与鲁棒性。

关键词: 面部表情识别; 光照增强; 可学习面部先验注意力; ConvNeXt
中图分类号: TP391.41 **文献标志码:** A **文章编号:** 2095-2163(2026)07-0010-08

LEFA-ConvNeXt: A facial expression recognition network based on illumination enhancement and Learnable Facial Prior Attention

LIU Ling′er, YAN Junfeng
(School of Informatics, Hunan University of Chinese Medicine, Changsha 410208, China)

Abstract: With the rapid development of deep learning techniques, Facial Expression Recognition (FER) has been widely applied in affective computing, human-computer interaction, intelligent driving, and related fields. However, due to factors such as illumination variations, subtle expression differences, and blurred local features, existing methods still suffer from limited recognition performance in complex environments. To address these challenges, this paper proposes a facial expression recognition network based on illumination enhancement and Learnable Facial Prior Attention. The proposed network adopts ConvNeXt as the backbone and introduces an illumination enhancement module (Light Enhance Block) to adaptively adjust image brightness, thereby improving robustness under non-uniform illumination conditions. In addition, a Learnable Facial Prior Attention (LFPA) module is designed to guide the network to focus on critical facial regions (e.g., eyes and mouth corners) through a learnable facial mask mechanism, enhancing the discriminability of fine-grained features. To further integrate local contextual information, a Local Context Self-Attention (LCSA) Block is incorporated at the high-level feature stage to perform window-level self-attention, enabling collaborative modeling of global and local features. Experimental results on the public FER-2013 dataset demonstrate that the proposed LEFA-ConvNeXt network outperforms existing mainstream methods in terms of recognition accuracy and generalization ability, achieving a maximum accuracy of 74.78%. These results fully validate the effectiveness and robustness of the proposed method under complex illumination conditions and diverse facial expression scenarios.

Key words: facial expression recognition; illumination enhancement; Learnable Facial Prior Attention; ConvNeXt

基金项目: 2025 年湖南中医药大学交叉学科方向培育项目(2025JC0102)。

作者简介: 刘灵儿(2002—),女,硕士研究生。主要研究方向:深度学习,面部情感识别。
通信作者: 晏峻峰(1965—),女,博士,教授。主要研究方向:中医人工智能。Email:junfengyan@hnucm.edu.cn。

收稿日期: 2026-01-05

0 引　言

面部表情是人类传达情绪和心理状态最直接、最自然的方式之一。在日常人际交往中，细微的表情变化往往能够反映出个体的情绪倾向和认知意图。随着人机交互（Human-Computer Interaction，HCI）[1]和人工智能技术[2]的快速发展，使机器具备情绪感知与理解能力、即面部表情识别（Facial Expression Recognition，FER）[3]，已成为计算机视觉[4]领域的重要研究方向之一。

情绪是人类对环境变化的本能反应[5]，能够通过多种生理与行为特征表现出来，其中最显著的就是面部表情。研究表明，在人类交流中，超过 90% 的有效沟通依赖于情绪的表达，这使得面部表情识别成为情感计算中不可或缺的组成部分。

多年来，不少学者已着手研究以应对这一新兴挑战。Ekman 等学者[6]指出了人类存在的 7 种与出生文化无关的基本情感（愤怒、恐惧、厌恶、快乐、悲伤、惊讶及中性）。Kalsum 等学者[7]提出了在复杂现实环境下实现面部表情识别的挑战，通过构建基于 EfficientNetB0 的卷积神经网络，结合数据增强和早停策略，但模型在最具挑战性的 FER-2013 数据集上性能仍有提升空间。Roy 等学者[8]针对 FER-2013 数据集上面临的表情识别准确率不高和模型泛化能力有限的问题，提出了一种基于 CNN 的模型，通过数据预处理和网络结构优化将准确率提升至约 67%，但模型在“厌恶”、“恐惧”等少数类别上识别率仍然较低，且未解决类别不平衡和复杂场景下的泛化问题。Sahoo 等学者[9]针对 FER2013 数据集中因图像质量差、类别不平衡和复杂真实场景导致的表情识别准确率不高的问题，通过设计 10 层 CNN 模型并结合数据增强技术，将准确率提升至 68.34%，但准确率相比其他先进方法仍有提升空间。Singh 等学者[10]提出了人脸表情识别在复杂场景下准确率低、泛化能力弱的问题，通过采用 EfficientNet 迁移学习的方法有效提升了性能，但其在数据集 FER-2013 上识别率仍偏低，且对少数类别（如厌恶）和小样本场景的识别效果不佳。Niu 等学者[11]提出了在 FER-2013 数据集上，传统 CNN 模型因深度和特征提取能力不足导致表情识别准确率低的问题，通过采用基于迁移学习的预训练模型进行微调显著提升了性能，但模型对少数类别（如厌恶、恐惧）的识别仍不理想，且未引入注意力机制等先进技术来进一步提升性能。Oguine 等学者[12]提出传统 FER 模型在实时性和准确率上存在不足的问题，通过结合 DCNN 与 Haar Cascade 的混合模型，并采用数据增强和超参数优化，在 FER2013 数据集上取得了 70.04% 的准确率，但模型对“厌恶”和“愤怒”等少数类别识别较差，且在真实场景中的泛化能力仍有局限。

要在真实复杂环境中实现高鲁棒性和高精度的面部表情识别仍面临诸多挑战[13-14]。表情的识别过程容易受到光照变化、遮挡、姿态变化及表情差异等多种因素的影响，尤其是光照不均和低光环境会导致人脸特征的失真，使网络难以提取稳定且具有判别力的特征。此外，现有的深度学习模型往往缺乏对关键面部区域的自适应关注能力，导致情绪相关的局部特征未能得到充分挖掘，从而影响识别性能。

针对上述问题，本文提出了一种基于光照增强与可学习面部先验注意的表情识别网络（LEFA-ConvNeXt）。该网络的核心思想是：在复杂光照条件下提升面部图像质量，同时引导模型自适应地聚焦于与情绪密切相关的面部区域。具体而言，本文设计了一个光照增强模块（Light Enhancement Module），用于改善图像对比度并恢复低光条件下的纹理细节，同时构建了可学习面部先验注意力模块，以自适应地捕获表情判别性最强的局部特征，从而增强特征的区分性与鲁棒性。此外，在 ConvNeXt[1]骨干网络的最后阶段引入局部上下文自注意力模块，用于提升模型的全局上下文理解能力与多尺度特征交互能力。

1 方　法

1.1 LEFA-ConvNeXt 总体架构

本文针对复杂光照条件下面部表情识别任务中的关键挑战，提出了一种深度融合卷积与注意力机制的先进神经网络架构，网络结构如图 1 所示。该架构以经典的 ConvNeXt 网络为基础骨架，通过系统性地集成 3 个创新性模块：光照自适应增强模块、可学习面部先验注意力模块以及局部上下文自注意力模块，构建了一个端到端的表情识别网络。这些模块分别针对输入质量优化、局部特征增强和全局关系建模三个关键环节进行了专门设计，形成了一个协同工作的多层次特征处理网络。

本架构在以下几个关键位置引入了创新性改进：

（1）输入预处理层：在网络最前端插入光照自适应增强模块，在特征提取之前对输入图像进行质

量优化,减轻光照变化带来的负面影响。

(2)中层特征增强:在 Stage 3 和 Stage 4 的输出端分别集成可学习面部先验注意力模块,通过自适应的空间权重调整,引导网络关注与表情相关的关键面部区域。

(3)高层特征优化:在 Stage 4 的末端加入轻量级全局上下文建模模块,通过自注意力机制整合全局空间信息,弥补纯卷积操作在长程依赖建模方面的不足。

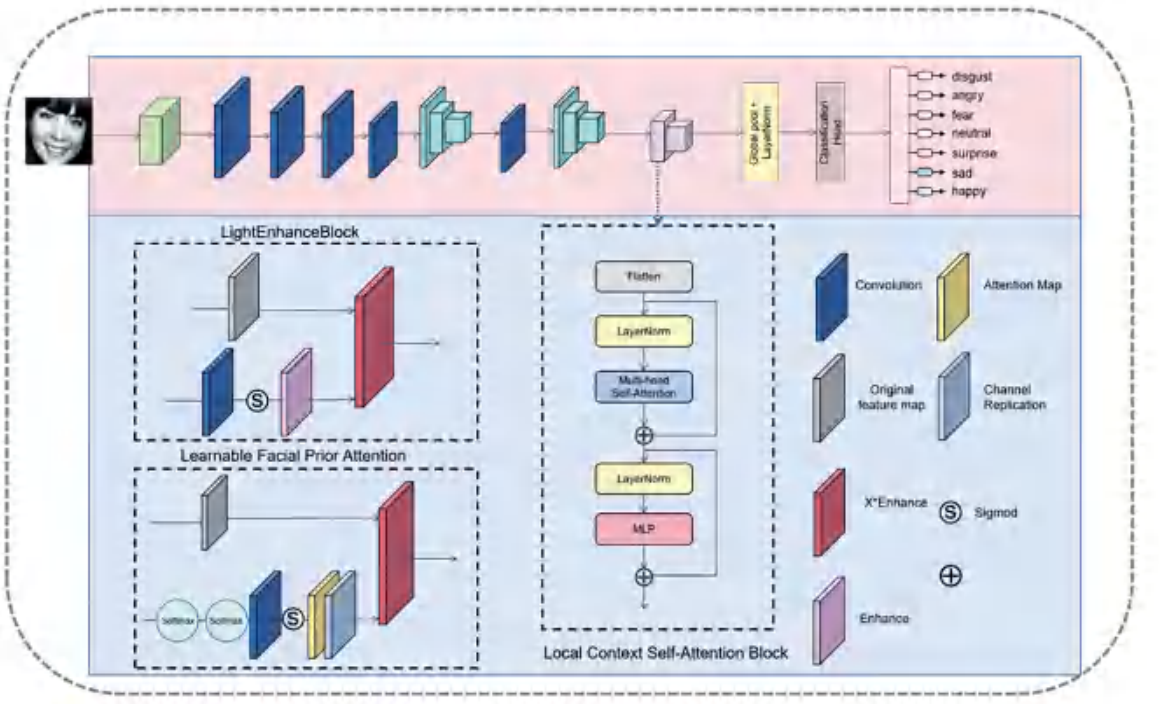


图 1 LEFA-ConvNeXt 总体架构

Fig. 1 Overall architecture of LEFA-ConvNeXt

1.2 光照自适应增强模块

面部表情识别系统在实际部署中经常面临光照条件变化的挑战,不均匀照明、低对比度和过曝光等问题严重影响了特征的可靠提取。传统解决方案如直方图均衡化[15]、同态滤波[16]或基于 Retinex 理论[17]的方法往往依赖于手工设计的参数,缺乏对不同场景的自适应能力。为此,本文提出了一种基于数据驱动的光照自适应增强模块,该模块能够在不引入大量计算开销的前提下,有效提升输入图像的质量,图像增强模块结构如图 2 所示。

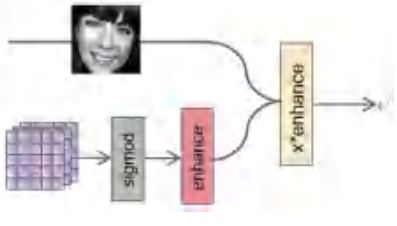


图 2 图像增强模块

Fig. 2 Image enhancement module

该模块通过 3×3 卷积学习输入图像的局部亮度分布,生成亮度调整映射,对暗部区域实施增强而对亮部区域适度抑制。该模块以端到端方式训练,能够根据具体图像内容自适应调整,显著改善低质量输入图像的视觉效果,为后续特征提取奠定基础。模块的前向传播过程如下。

首先,通过卷积操作提取光照分布特征,可表示为:

$$f_{(\mathrm{conv})} = W_{3\times3} * x \tag{1}$$

其中, $W_{3\times3}$ 表示 3×3 卷积核的权重参数;“ * ”表示卷积操作。

该卷积使用同填充(padding = 1)保持空间分辨率不变,且不引入偏置项以减少参数数量。接着,通过 Sigmoid 函数将特征映射转换为亮度调整系数,用到的公式为:

$$\mathrm{light_map} = \sigma(f_{\mathrm{conv}}) = \frac{1}{1 + e^{-f_{\mathrm{conv}}}} \tag{2}$$

Sigmoid 函数将输出值压缩到(0,1)区间,形成与输入同尺寸的亮度分布图。最后,基于亮度分布图生成增强系数并应用于原始输入。具体公式为:

$$x_{\mathrm{enhanced}} = x \odot \mathrm{enhance_factor} \tag{3}$$

其中，“⊙”表示逐元素相乘。

与传统方法中固定方向的亮度调整不同，本文的设计允许网络根据学习到的亮度分布图动态决定每个区域的调整方向：

（1）当某区域的 light_map 值大于 0.5 时，enhance_factor 将小于 1，实现对该区域的亮度抑制，防止过曝光区域的特征丢失。

（2）当某区域的 light_map 值小于 0.5 时，enhance_factor 将大于 1，实现对该区域的亮度增强，改善欠曝光区域的细节可见性。

（3）当 light_map 值等于 0.5 时，enhance_factor 等于 1，保持原始亮度不变。

这种自适应机制使得模块能够同时处理过曝光和欠曝光问题，在各种极端光照条件下均能保持鲁棒的性能。此外，模块的轻量级设计确保了其在计算开销上的高效性，仅增加约 $19C^2$ 个参数（对于 3 通道输入约为 81 个参数），几乎可忽略不计。处理之后的图像如图 3 所示。

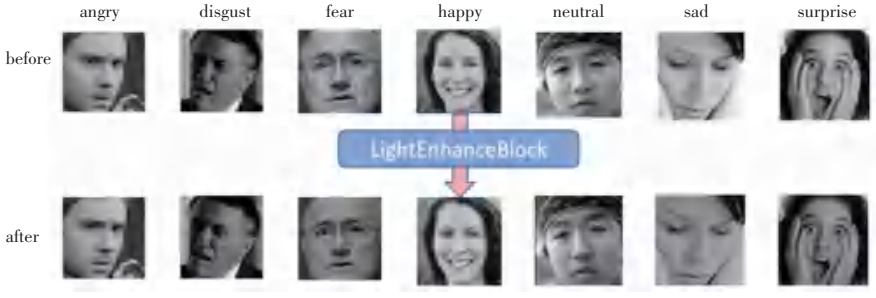


图 3　图像处理前后

Fig. 3　Images before and after processing

1.3　可学习面部先验注意力模块

面部表情的表达主要通过特定面部区域的肌肉运动实现，如眼睛的睁闭、眉毛的扬起等。然而，传统卷积网络在处理面部图像时往往平等对待所有空间位置，缺乏对表情关键区域的针对性关注。为此，本文提出了一种可学习的面部区域注意力模块。该模块能够自动识别并增强与表情相关的面部区域特征，抑制无关背景和不变区域的干扰，LFPA 模块结构如图 4 所示。

该模块通过学习面部关键区域的空间注意力权重，使模型能够聚焦于眉毛、眼睛、嘴巴等与情感表达密切相关的区域。注意力参数通过 Softmax 函数进行空间归一化，确保注意力权重的合理分布。该模块在网络的第 3 和第 4 阶段分别以 14×14 和 7×7 分辨率应用，通过残差连接方式与主干特征融合，增强了对细微表情变化的敏感性。

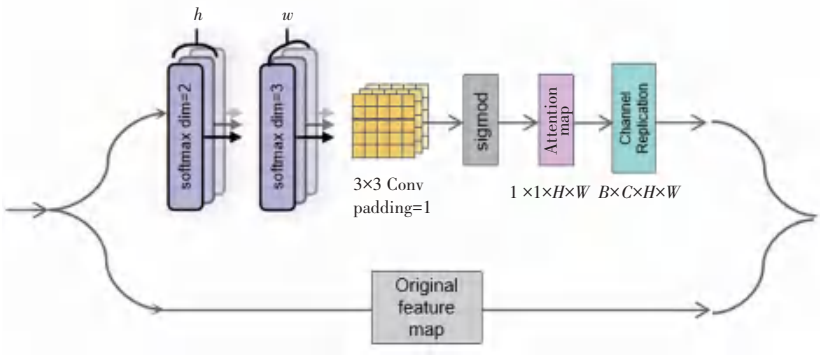


图 4　LFPA 模块

Fig. 4　LFPA module

研究中模块初始化时定义了一组可学习的注意力参数：

$$P_{\text{mask}} \in R(1 \times M \times H_s \times W_s) \tag{4}$$

其中，$M = 2$ 表示注意力通道数（分别对应前景关注和背景抑制），H_s 和 W_s 为目标特征图尺寸（14 或 7）。

参数采用 Xavier 均匀分布初始化[18]：

$$P_{\text{mask}} \sim U\left(-\sqrt{\frac{6}{H_s \times W_s \times M}}, \sqrt{\frac{6}{H_s \times W_s \times M}}\right) \tag{5}$$

为了确保注意力权重在空间维度上的合理分布，采用了独特的双 Softmax 归一化策略：

$$A_{\text{temp}} = \text{Softmax}_{\text{height}}(P_{\text{mask}}) \tag{6}$$

$$A_{\text{final}} = \text{Softmax}_{\text{width}}(A_{\text{temp}}) \tag{7}$$

这种先按高度维度、再按宽度维度的双重归一

化确保了注意力图在 2 个空间维度上均具有合理的概率分布特性,避免了注意力权重过度集中或过度分散的问题。

通过 3×3 卷积将双通道注意力图融合为单通道注意力权重图,计算公式为:

$$A_{\text{fixed}} = W_{\text{spatial}} * A_{\text{final}} \tag{8}$$

其中, $W_{\text{spatial}} \in \mathbb{R}^{1\times M\times 3\times 3}$ 为空间卷积权重,使用同填充保持空间尺寸不变,且不引入偏置项。

最后,通过 Sigmoid 函数生成最终的注意力图,并通过残差连接方式增强原始特征,公式为:

$$A_{\text{output}} = \sigma(A_{\text{fused}}) \tag{9}$$

$$x_{\text{enhanced}} = x \odot (1 + A_{\text{output}}) \tag{10}$$

其中,"⊙"表示逐元素相乘,$(1 + A_{\text{output}})$ 构成了一个大于 1 的增强系数,确保重要区域的特征得到加强。

1.4 局部上下文自注意力模块

卷积操作在局部特征提取方面表现出色,但在捕获长程空间依赖关系方面存在固有局限性。全局上下文信息对于理解复杂表情(如困惑、沉思等涉及多个面部区域协调的表情)至关重要。为了在保持计算效率的前提下引入全局建模能力,研发设计了一个轻量级的视觉变换器模块。该模块基于 Transformer[19] 的核心思想,但针对表情识别任务进行了优化和简化,Local Context Self-Attention 模块结构如图 5 所示。

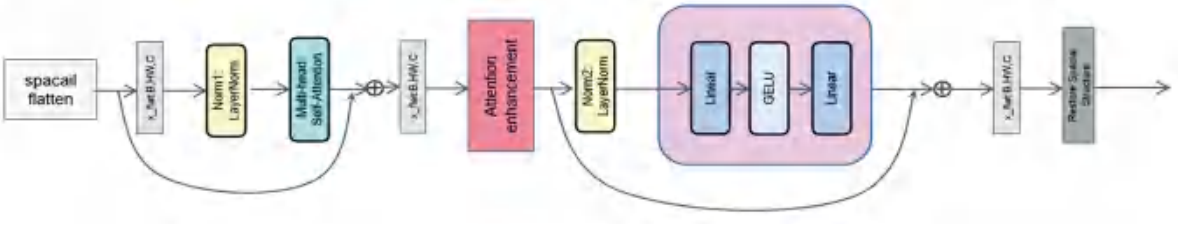


图 5 Local Context Self-Attention 模块

Fig. 5 Local Context Self-Attention module

轻量级全局上下文建模模块包含 4 个主要组件:层归一化、多头自注意力机制、前馈神经网络和残差连接。模块输入为 Stage 4 输出的特征图 $x \in \mathbb{R}^{B\times C\times H\times W}$,其中 $C = 768$,$H = W = 7$。

首先,将空间特征图转换为序列形式,以适应 Transformer 架构,具体公式为:

$$x_{\text{flat}} = \text{Reshape}(x) \in \mathbb{R}^{B\times N\times C} \tag{11}$$

其中, $N = H \times W = 49$ 为序列长度,每个特征向量对应原始特征图中的一个空间位置。采用 Pre-LN 架构,在自注意力操作前应用层归一化,即:

$$x_{\text{norm1}} = \text{LayerNorm}(x_{\text{flat}}) \tag{12}$$

多头自注意力机制通过并行多个注意力头捕获多样化的上下文关系:

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \text{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{13}$$

其中, $\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}$ 分别由输入序列通过线性变换得到,d_k 表示键向量的维度。研究中采用 4 个注意力头,每个头的维度为 $d_k = d_v = C/4 = 192$。

自注意力输出与原始输入通过残差连接合并:

$$x_{\text{attn}} = x_{\text{flat}} + \text{MultiHeadAttention}(x_{\text{norm1}}, x_{\text{norm1}}, x_{\text{norm1}}) \tag{14}$$

前馈网络由 2 个线性变换和 1 个 GELU 激活函数组成,推得公式为:

$$x_{\text{norm2}} = \text{LayerNorm}(x_{\text{attn}}) \tag{15}$$

$$x_{\text{ffn}} = \boldsymbol{W}_2 \cdot \text{GeLU}(\boldsymbol{W}_1 \cdot x_{\text{norm2}} + b_1) + b_2 \tag{16}$$

其中, $\boldsymbol{W}_1 \in \mathbb{R}^{4C\times C}$,$\boldsymbol{W}_2 \in \mathbb{R}^{C\times 4C}$ 表示权重矩阵,实现了先升维后降维的变换策略。

最终输出通过第 2 个残差连接得到:

$$Z_{\text{final}} = \text{Reshape}(Z_{\text{out}}) \in \mathbb{R}^{B\times C\times H\times W} \tag{17}$$

将序列化输出重新转换为空间特征图格式,定义公式为:

$$x_{\text{final}} = \text{Reshape}(x_{\text{out}}) \in \mathbb{R}^{B\times C\times H\times W} \tag{18}$$

该全局建模机制使网络能够有效整合分布于不同面部区域的特征信息,例如眼部与口部动作之间的关联,从而捕获依赖多区域协同的复杂表情模式。同时,在面部局部遮挡[20] 的情况下,模型可借助全局上下文信息对缺失区域进行合理推断,提升表情识别的鲁棒性。

2 实验

2.1 数据集介绍

本研究使用 FER2013 数据集[21] 进行模型训练与评估。FER2013 数据集共包含 35 887 张灰度人脸图像,每张图像的分辨率固定为 48×48 像素。所有图像均通过 Google 搜索引擎的 API 自动收集,并经过人脸检测和对齐预处理,确保了所用数据的标准化

程度。数据集涵盖了人类 7 种基本情感类别，包括愤怒(4 953)、厌恶(547)、恐惧(5 121)、高兴(8 989)、悲伤(6 077)、惊讶(4 002)和中性(6 198)。

2.2　实验环境与参数设置

模型以 ImageNet 预训练的 ConvNeXt-Tiny 为基础架构，训练 50 个 epoch，批次大小设为 16。优化采用 AdamW[22]，初始学习率为 1e-4，并配合 ReduceLROnPlateau 调度策略：若测试集损失连续 5 个 epoch 未下降，则学习率衰减为原来的 0.1，最低学习率设为 1e-6。该设置兼顾训练效率与收敛稳定性。

3　实验结果

3.1　实验结果分析

通过系统实验评估，本研究提出的 LEFA-ConvNeXt 模型在 FER2013 数据集上取得了显著性能提升。不同模型在 FER2013 数据集上的性能对比见表 1。传统模型如 ResNet-50[23]和 VGG-16[24]在该数据集上表现有限，尤其在召回率和 *F*1 分数方面存在不足。相比之下，基线模型 ConvNeXt-Tiny 性能已有明显改善，而本文提出的 LEFA-ConvNeXt 进一步取得了 74.78%的准确率、75.63%的精确率、74.42%的召回率和 74.95%的 *F*1 分数，4 项关键指标均达到最优。与原始 ConvNeXt-Tiny 相比，准确率提升 3.78 个百分点，验证了所引入的光照增强与可学习面部注意力机制的有效性。

表 1　不同模型在 FER2013 数据集上的性能对比

Table 1　Performance comparison of different models on the FER2013 dataset　%

模型	准确率	精确率	召回率	*F*1 分数
ResNet-50	64.28	65.12	63.85	64.21
VGG-16	57.30	58.42	56.78	57.35
ConvNeXt-Tiny	71.00	71.85	70.62	71.10
CNN[8]	66.48	67.10	65.82	66.45
10 层 CNN 模型[9]	68.34	69.02	67.58	68.29
MoEffNet[10]	71.02	71.94	70.31	71.11
DeepFace[11]	70.01	70.86	69.42	70.13
LEFA-ConvNeXt	74.78	75.63	74.42	74.95

定性对比结果(见图 6)显示，针对恐惧、厌恶、悲伤和愤怒等易混淆表情，传统卷积神经网络模型普遍存在误判，而本文方法均能给出正确预测。这表明 LEFA-ConvNeXt 在复杂表情场景下具有更强的判别能力和更好的鲁棒性，能够有效提升细微表情的识别性能。

为进一步验证本文所提出注意力机制的有效性，在相同实验设置下，将所设计的 LEFA 注意力模块与多种典型注意力机制进行了对比实验，实验结果见表 2。可以看出，在未引入注意力机制的 ConvNeXt-Tiny 基线模型中，模型准确率为 71.00%。在此基础上，引入通道注意力(SE)后，模型准确率提升至 73.12%，说明通道维度建模有助于增强判别性特征表达；采用空间注意力机制时，准确率提升至 72.25%，表明空间位置建模在一定程度上能够引导模型关注关键区域，但其提升幅度相对有限。进一步地，ECA-Net[25]和 CBAM[26]等代表性轻量化注意力模块分别取得了 73.48% 和 73.65%的准确率，较基线模型均有不同程度的性能提升。相比之下，本文提出的 LEFA 注意力模块在 FER2013 数据集上取得了最高的准确率 74.78%，较基线模型提升了 3.78 个百分点，优于其他对比注意力方法。这表明 LEFA 模块能够更有效地融合人脸表情中的关键区域信息与细粒度特征，对复杂光照条件和细微表情变化具有更强的建模能力。

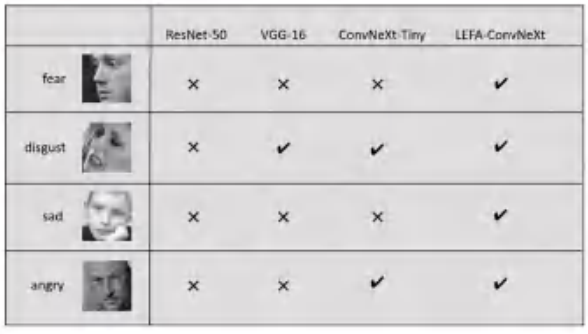


图 6　不同网络面部表情识别结果的定性对比

Fig. 6　Qualitative comparison of facial expression recognition results by different networks

表 2　不同注意力在 LEFA-ConvNeXt 的性能对比

Table 2　Performance comparison of different attention mechanisms in LEFA-ConvNeXt

注意力模块	准确率/%	相比基线提升
无注意力(ConvNeXt-Tiny)	71.00	-
SENet[27]	73.12	+2.12
Spatial Attention[28]	72.25	+1.25
ECA-Net[25]	73.48	+2.48
CBAM[26]	73.65	+2.65
LEFA(本文方法)	74.78	+3.78

3.2　消融实验分析

为深入分析各改进模块对模型性能提升的贡

献,本文在 FER2013 数据集上开展了系统性的消融实验,实验结果见表 3。消融实验以 ConvNeXt-Tiny 作为基线模型,逐步引入不同功能模块,通过对比各阶段性能变化来验证每个模块的有效性。

表 3 消融实验结果

Table 3 Ablation study results

实验配置	准确率/%	相比基线提升
基线模型 (ConvNeXt-Tiny)	71.00	-
+ Light Enhance Block	72.15	+1.15
+Light Enhance Block + LFPA	73.86	+2.86
+Light Enhance Block + LFPA+LCSA Block	74.78	+3.78

基线模型 ConvNeXt-Tiny 在验证集上取得了 71.00% 的分类准确率,作为后续对比的参考标准。在此基础上,引入 Light Enhance Block 后,模型准确率提升至 72.15%,相较基线模型提高了 1.15%。这一结果表明,光照增强模块能够有效缓解面部表情图像中因光照不均或低对比度带来的干扰,从而提升特征提取的稳定性和判别能力。

在引入 Light Enhance Block 的基础上进一步加入可学习面部先验注意力模块,模型准确率显著提升至 73.86%,整体性能相较基线提高 2.86%。该结果说明,LFPA 模块能够自适应地强调面部关键区域如眼部、眉部和口部,有效抑制背景和非判别区域的干扰,从而增强模型对细粒度表情差异的建模能力,尤其有助于区分 Fear、Angry 和 Sad 等易混淆情绪。

在上述结构基础上进一步引入局部上下文自注意力模块,用于增强局部与全局特征建模能力,模型最终准确率达到 74.78%,较基线模型累计提升 3.78%。这一结果表明 Local Context Self-Attention Block 所引入的窗口级自注意力机制[29-30]能够有效捕获面部区域之间的长程依赖关系,与 Light Enhance Block 和 LFPA 模块形成良好的互补,从而进一步提升模型的整体表达能力和泛化性能。

图 7 展示了在 ConvNeXt-Tiny 主干网络基础上逐步引入 Light Enhance Block、LFPA 以及 LCSA 模块后的面部表情识别定性对比结果。可以观察到,原始 ConvNeXt-Tiny 在 fear、disgust、sad 和 angry 等易混淆表情样本上均存在不同程度的误判现象。引入 Light Enhance Block 后,模型在 sad 类别上能够给出正确预测,表明亮度增强模块有助于缓解光照变化对表情判别的干扰。进一步加入 LFPA 模块后,模型在 disgust 与 angry 样本上的识别结果得到明显改善,说明局部特征感知能力的增强有利于捕获关键表情区域。最终,在同时引入 Light Enhance Block、LFPA 与 LCSA 模块后,模型在所有展示样本上均实现了正确分类,验证了各模块之间具有良好的互补性,并能够协同提升细微表情的识别性能。

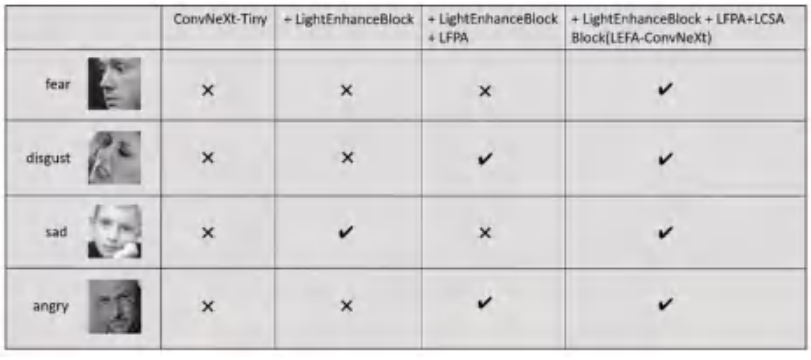


图 7 不同模块组合下面部表情识别结果的定性对比

Fig. 7 Qualitative comparison of facial expression recognition results with different module combinations

4 结束语

本研究提出了一种基于改进 ConvNeXt 架构的人脸表情识别方法,通过集成光照增强、多级注意力机制和全局上下文建模,显著提升了在 FER2013 数据集上的识别性能。实验结果表明,本文方法在准确率、精确率、召回率和 $F1$ 分数四个指标上均优于多个基线模型,消融实验验证了各改进模块的有效性。特别地,本文方法在光照变化和细微表情识别方面表现出明显优势。这些结果为复杂场景下的鲁棒表情识别提供了新的思路和技术途径,具有重要的理论和应用价值。

参考文献

[1] Mackenzie I S. Human-computer interaction: An empirical research perspective[M]. Amsterdam:Morgan Kaufman, 2013.

[2] Sharma L, Garg P K. Artificial intelligence: Technologies, applications, and challenges[M]. New York:Chapman and Hall/CRC, 2021.

[3] Ge H, Zhu Z, Dai Y, et al. Facial expression recognition based on deep learning [J]. Computer Methods and Programs in Biomedicine, 2022, 215: 106621.
[4] Voulodimos A, Doulamis N, Doulamis A, et al. Deep learning for computer vision: A brief review [J]. Computational Intelligence and Neuroscience, 2018, 2018: 7068349.
[5] Kals E, Müller M M. Emotions and environment[M]//Clayton S. The Oxford Handbook of Environmental and Conservation Psychology. Oxford: Oxford University Press, 2012: 128-149.
[6] Ekman P, Friesen W V. Constants across cultures in the face and emotion[J]. Journal of Personality and Social Psychology, 1971, 17 (2): 124-129.
[7] Kalsum T, Mehmood Z. A novel lightweight deep convolutional neural network model for human emotions recognition in diverse environments[J]. Journal of Sensors, 2023,2023: 6987708.
[8] Roy S, Madhurima D, Anupam G, et al. CNN-based model for facial expression recognition using FER 2013 dataset [C]// Proceedings of International Conference on Data Mining And Information Security. Cham: Springer, 2024:387-403.
[9] Sahoo G K, Ponduru J, Das S K, et al. Deep leaning-based facial expression recognition in fer2013 database: An in - vehicle application[C]// Proceedings of 2022 IEEE 19th India Council International Conference. Piscataway,NJ:IEEE, 2022:1-6.
[10] Singh R, Sharama H, Mehta N K, et al. EfficientNet for human FER using transfer learning [J]. ICTACT Journal on Soft Computing, 2022,13(1):2792-2797.
[11] Niu B N, Rashid A R, Ihnaini B,et al. Emotion recognition on FER-2013: A deep learning-based comparative analysis[C]// Proceedings of 2025 10th International Symposium on Advances in Electrical, Electronics and Computer Engineering (ISAEECE). Piscataway,NJ:IEEE, 2025: 529-533.
[12] Oguine O C, Oguine K J, Bisallah H I, et al. Hybrid facial expression recognition (FER2013) model for real-time emotion classification and prediction [J]. arXiv preprint arXiv, 2206. 09509, 2022.
[13] Zhang Z, Fort J M, Giménez M L. Facial expression recognition in virtual reality environments: challenges and opportunities[J]. Frontiers in Psychology, 2023, 14: 1280136.
[14] Liu Z, Mao H, Wu C Y, et al. A convnet for the 2020s[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway,NJ:IEEE, 2022: 11976-11986.
[15] Chowdhury J H, Liu Q, Ramanna S. Simple histogram equalization technique improves performance of VGG models on facial emotion recognition datasets [J]. Algorithms, 2024, 17 (6): 238.
[16] Fan C N, Zhang F Y. Homomorphic filtering based illumination normalization method for face recognition[J]. Pattern Recognition Letters, 2011, 32(10): 1468-1479.
[17] Nam G P, Park K R. New fuzzy-based retinex method for the illumination normalization of face recognition [J]. International Journal of Advanced Robotic Systems, 2012, 9(4): 103.
[18] Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks[C]//Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. ChiaLaguna Resort:PMLR, 2010: 249-256.
[19] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [C]//Proceedings of Advances in Neural Information Processing Systems. Long Beach, CA:NeurIPS, 2017, 30:5998-6008.
[20] 章晓爽, 黎格献, 贺永姣, 等. 基于 PyraMSA-ResNet18 的局部遮挡面部表情识别[J/OL]. 重庆工商大学学报(自然科学版). (2025-04-27). https://link. cnki. net/urlid/50. 1155. N. 20250425. 1806. 04.
[21] Giannopoulos P, Perikos I, Hatzilygeroudis I. Deep learning approaches for facial emotion recognition: A case study on FER-2013 [M]//Proceedings of Advances in Hybridization of Intelligent Methods: Models, Systems and Applications. Cham: Springer, 2017: 1-16.
[22] Zhou P, Xie X, Lin Z, et al. Towards understanding convergence and generalization of AdamW[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(9): 6486-6493.
[23] Koonce B. ResNet 50[M]//Convolutional Neural Networks with Swift for Tensorflow: Image Recognition and Dataset Categorization. Berkeley, CA: Apress, 2021: 63-72.
[24] Tammina S. Transfer learning using VGG - 16 with deep convolutional neural network for classifying images [J]. International Journal of Scientific and Research Publications (IJSRP), 2019, 9(10): 143-150.
[25] Wang Q, Wu B, Zhu P, et al. ECA-Net: Efficient channel attention for deep convolutional neural networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway,NJ:IEEE, 2020: 11534-11542.
[26] Woo S, Park J, Lee J Y, et al. Cbam: Convolutional block attention module[C]//Proceedings of the European Conference on Computer Vision (ECCV). Cham: Springer, 2018: 3-19.
[27] Bastidas A A, Tang H. Channel attention networks [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Piscataway, NJ: IEEE, 2019: 881-888.
[28] Rizzolatti G, Craighero L. Spatial attention: Mechanisms and theories [M]//Advances in Psychological Science. Hove: Psychology Press, 2014: 171-198.
[29] Yao D, Shao Y. A data efficient transformer based on Swin Transformer[J]. The Visual Computer, 2024, 40(4): 2589-2598.
[30] Roshanzadeh F, Barati H, Barati A. Lightweight intrusion detection system using a hybrid CNN and ConvNeXt-Tiny model for Internet of Things networks[J]. arXiv preprint arXiv, 2509. 06202, 2025.