文章编号:1671-6833(2026)03-0126-08

面向学习情感的多模态多尺度面部表情识别分析

姬莉霞, 任晗靓, 王 威, 杜云龙, 周洪鑫, 付元忠

(郑州大学 网络空间安全学院,河南 郑州 450002)

摘 要: 针对学习情感识别中细微特征难以捕捉、数据样本稀缺等问题,提出了一种基于生成扩散模型与多模态多尺度视觉编码的面部情感识别方法。首先,构建融合多尺度全局和局部细节特征的学习情感数据集,并利用生成扩散模型对稀缺情感样本进行扩充,从而缓解少样本学习场景的数据约束;其次,设计了一种多模态多尺度视觉编码机制,通过融合原始人脸图像的全局特征与显著区域的局部细节信息,实现微表情与细粒度情感特征的高精度建模与有效融合。最后,在 CNN、视觉 Transformer 及混合架构等多类模型上进行实验。结果表明,本文方法整体识别准确率达 68.10%,相较于现有基准方法平均提升约 2.98%,最高提升 5.30%。消融实验进一步验证了生成扩散模型与多模态多尺度融合模块在增强模型对微表情细节捕捉及整体识别鲁棒性方面的有效性与协同作用。

关键词:学习情感; 面部表情识别; 人工智能; 多模态; 多尺度特征

中图分类号:TP391;TP181 **文献标志码**:A **doi**:10.13705/j.issn.1671-6833.2025.06.016

学习情感是一种特定场景下的情感表达,相对于基本情绪的短暂强烈,通常表现为较为细微且持续时间较长[1]。传统的学习情感往往以考试结果、调查问卷或主观经验为依据,忽略了在学习过程中直接反映的实时情感。

针对传统方法的局限性,基于深度学习的情感识别技术为解决这一问题提供了新的可能。深度学习技术的兴起显著提升了情感识别的自动化和实时性。Simonyan 等[2]首次将卷积神经网络(convolutional neural network , CNN)应用于面部表情识别,引发了深度学习助力面部情感识别的新篇章。Jain 等[3]通过深度卷积网络改进了特征表示。研究表明,CNN 作为特征提取器与分类器结合可提高情感识别的准确性[4-5]。Singh 等[6]使用 CNN 进行面部情感分类,Sharma 等[7]的研究覆盖了最新趋势下人类情感检测的多种面部表情识别方法及其应用,为实时情感检测提供理论依据。

近年来,随着 Dosovitskiy 等[8]提出了 vision transformer (ViT),基于 Transformer 架构的面部表情识别方法也开始受到关注。Xue 等[9]提出 TransFER 模型,利用 Transformer 学习关系感知面部表情表征。此后,研究者受掩码自编码器的启发,提出 MAE-DFER[10],这是一种自监督方法,利用未标记数据进行大规模预训练。为进一步解决在无约束场景中部分遮挡的面部表情识别问题,提出了基于 Transformer 的面部表情识别方法 TFE[11],它能够自适应地关注最重要和未被遮挡的面部区域。这些方法在一定程度上弥补了传统工程方法在面部表情特征提取上的不足,但在学习情感识别中仍面临诸多挑战。

学习情感的波动通常较为细微,如何更精确地捕捉这些微小的面部表情变化,以真实反映学习过程中的情感,依然是一个难题。此外,情感识别的深度学习方法主要以有监督学习范式为主导,然而,现有的公共数据集大多关注基本情感,且收集深度学习模型所需的数据也面临着诸多障碍,一方面互联网上的人脸图像和视频很少涉及学习环境;另一方面收集人脸信息受到隐私和法律限制。针对以上问

收稿日期:2025-11-02;**修订日期**:2025-12-28

基金项目:河南省重大科技专项资助项目(231100210200)

作者简介:姬莉霞(1979—),女,河南新乡人,郑州大学教授,主要从事信息内容安全研究,E-mail:jilixia@zzu.edu.cn。

通信作者:付元忠(1974—),男,河南郑州人,郑州大学工程师,主要从事信息管理及信息安全研究,E-mail:fuyuanzhong@zzu.edu.cn。

引用本文:姬莉霞,任晗靓,王威,等. 面向学习情感的多模态多尺度面部表情识别分析[J]. 郑州大学学报(工学版),2026, 47(3): 126-133. (JI L X, REN H L, WANG W, et al. Multimodal and multiscale facial expression recognition analysis for learning emotions[J]. Journal of Zhengzhou University (Engineering Science), 2026, 47(3): 126-133.)

题,提出了一种基于生成扩散模型与多模态多尺度视觉编码相结合的学习情感识别方法,通过与基准模型进行对比实验,验证了方法的优越性。本文贡献如下:

1)提出了一种多模态多尺度学习情感识别方法。该方法结合了 RGB 图像与显著特征图两种模态信息,同时利用全局人脸图像与局部关键区域进行特征提取,实现多尺度信息融合。

2)构建融合图像多尺度全局和局部细微特征的学习情感数据集,该数据集结合了实际采集的学习表情与生成扩散模型生成的增强样本,缓解了学习表情数据稀缺问题。

1　学习情感表征建模

为实现对学习情感的有效识别,需构建贴合实际场景的情感表征模型。本研究重点关注在学习情境中发生频率较高的情感。为此,研究通过对来自 10 个真实课堂的 729 名学生进行 8 小时的观察,通过摄像头采集学生在学习任务中的面部表情数据,并采用基于 CNN 的人脸检测算法自动提取图像流中可识别的人脸区域,该过程中剔除了中性或难以判断的表情样本,仅保留具备明确情感特征的图像。统计结果显示,学习过程中最常出现的 10 种情感分别为:专注、思索、愉悦、困惑、无聊、瞌睡、惊讶、焦虑、沮丧和兴奋。

情感心理学研究领域的情感表征建模主要有 2 种路径:分类表征和维度表征。图 1 所示的维度模型,涵盖了学习场景的 4 个类别、2 个层级、8 种高频情感。从分类表征的角度出发,本研究注意到“兴高采烈”和“焦虑”这 2 种强烈情感在学习场景中出现的比例低于 2%,因此它们不属于核心学习情感。基于此,在图 1 的分类表征模型中将学习情感划分为“愉悦”、“专注”、“思索”、“无聊”、“困惑”和“瞌睡”。

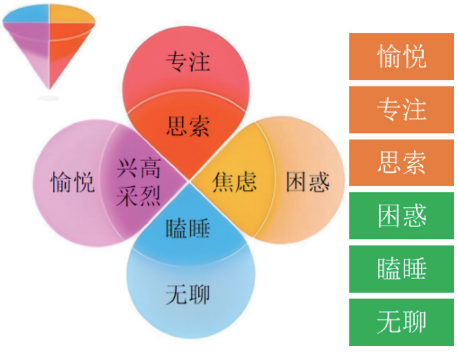


图 1　学习情感表征模型图

Figure 1　Learning emotion representation model diagram

2　学习情感面部识别方法

本研究提出的学习情感全局-局部多尺度多模态识别总体方案如图 2 所示。在数据集构建环节,整合获取的多源原始数据,并采用以生成扩散模型为核心的智能生成技术进行数据增广。在深度情感识别模型构建环节,设计了 1 个双分支编码架构。全局分支基于 Transformer 模型处理人脸 RGB 图像,以捕获整体上下文信息。局部分支则利用局部描述符提取显著特征图,以聚焦细微纹理变化。最终通过 1 个分层融合机制,实现 RGB 图像与特征图在不同尺度上的深度融合,完成全局人脸表征与局部显著特征在模态间的有效对齐。

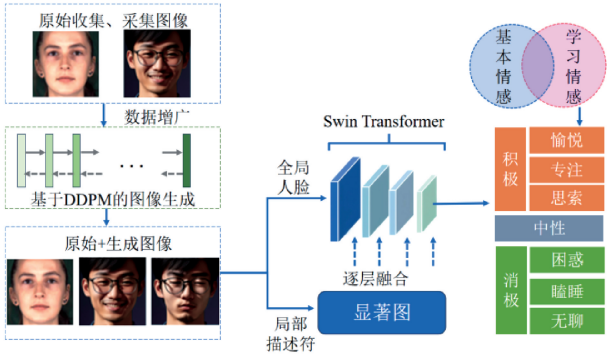


图 2　学习情感面部识别总体方案

Figure 2　Learning emotional facial recognition overall scheme

2.1　数据生成与质量控制

2.1.1　基于扩散模型的数据生成

本研究借助生成模型进行数据增广,利用基于隐空间的去噪扩散模型(denoising diffusion probabilistic models, DDPM)[12]来扩充训练阶段的样本。为此,从现有面部表情数据集、志愿者以及网络收集了体现情感的 5 682 张不同表情,按照如图 3 所示的方法生成数据集。

首先,进行数据收集和预处理。图像的第 1 个来源是现有数据集,主要包括中国科学院微表情数据库 CASME、CAS(ME)[13],以及 RAF-DB[14]、FER[6]等,并对数据集内容进行筛选和重新标注。图像的第 2 个来源是志愿者,在线上线下对 42 名志愿者进行表情采集,得到 576 张表情。此外,从网络上获取了部分图像。为了便于处理,将图像保留脸部并重采样为 256×256 大小,并对低分辨率图像进行盲脸修复[15]。然后,基于扩散模型从初始数据集的部分图像生成新的样本。扩散模型通常包括如图 4 所示的 2 个主要部分:前向加噪过程与反向去噪过程[16]。前向过程在每一个时间步 t 中将较小的

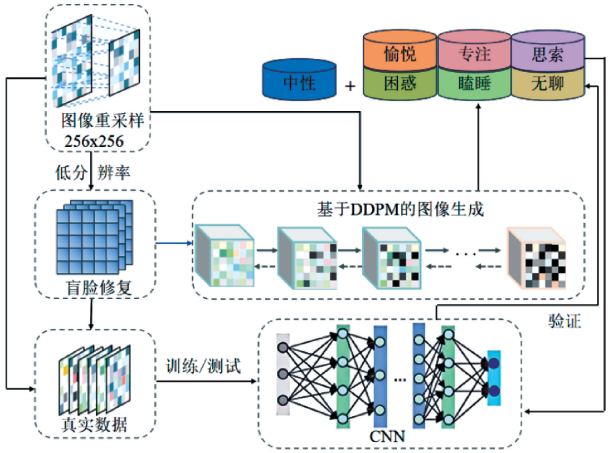


图 3 数据集生成

Figure 3 Dataset generation

高斯噪声逐步加入到原数据中,直到时间 T 时将原始数据完全变为高斯噪声,而反向过程则使用去噪函数在每一次迭代中去除前向过程所添加的噪声,直到原始数据被恢复。在生成数据时,扩散模型会利用在反向过程中学习到的去噪函数,从高斯噪声中生成较为清晰的图像。

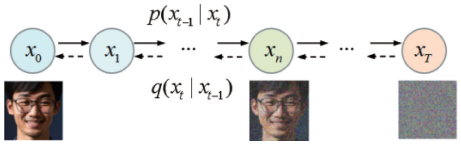


图 4 基于 DDPM 的数据增广

Figure 4 Data augmentation based on DDPM

2.1.2 生成数据的语义与质量控制

在数据生成过程中,使用编码器将原始数据映射到隐空间中实现扩散过程。为在反向去噪过程中约束生成样本类别的一致性,引入类别条件约束,将情感类别标签嵌入噪声预测模块,计算如式(1)所示。

$$\epsilon_\theta(Z_t,y)=\mathrm{CrossAttn}(Z_t,\varepsilon(y);\theta)。\quad(1)$$

式中:ϵ_θ 为基于可学习参数 θ 的噪声预测函数;Z_t 为扩散第 t 步隐空间特征;CrossAttn 为交叉注意力机制;$\varepsilon(y)$ 为学习情感类别标签 y 的嵌入向量。

为提升生成样本与真实样本的特征一致性,引入对比损失最小化两者距离,计算如式(2)所示。

$$\mathcal{L}_{\mathrm{con}}=\sum_i \max[0,m-\mathrm{d}(x_i,x_j)^2+\mathrm{d}(z_i,z_j)^2]。\quad(2)$$

式中:$\mathcal{L}_{\mathrm{con}}$ 为对比损失值;x_i 和 x_j 分别是生成和真实样本的高层特征;z_i,z_j 为生成和真实样本的隐空间低维特征;$\mathrm{d}(\cdot)$ 为欧氏距离函数;m 是距离间隔。

2.2 全局-局部多尺度多模态融合视觉编码

2.2.1 总体架构

多模态多尺度融合方法总体架构如图 5 所示,由 2 个分支组成:全局分支使用 swin transformer 模型[17]作为主干,从人脸 RGB 图像中提取全局表情特征;swin transformer 使用滑动窗口的自注意力机制,相较于 ViT,可以缩短输入序列的长度、降低模型复杂度,更好地兼顾图像局部和全局之间的关联。局部分支使用局部描述符,利用显著特征图来捕捉人脸图像的纹理、随机性和几何分析等局部特征,包括面部表情的细微变化等。

该方法利用层级视觉 Transformer 方式分 4 个阶段融合 2 种模态的特征。具体来说,自下而上对每个阶段 $i\in\{1,2,3,4\}$ 进行编号,每个阶段包含 2 个视觉编码层、1 个多模态特征融合模块和 1 个自适应的局部门控单元(local gate, LG),并通过以下 3 个步骤融合不同模态、不同尺度的视觉特征:

(1) 视觉编码。视觉编码层的 4 个阶段对应于 swin transformer 中的 4 个阶段,分别对人脸图像和显著特征进行编码,得到全局人脸特征 $G_i\in\mathbb{R}^{H_i\times W_i\times C_i}$ 和局部显著特征 $L_i\in\mathbb{R}^{H_i\times W_i\times C_i}$,其中 C_i、H_i

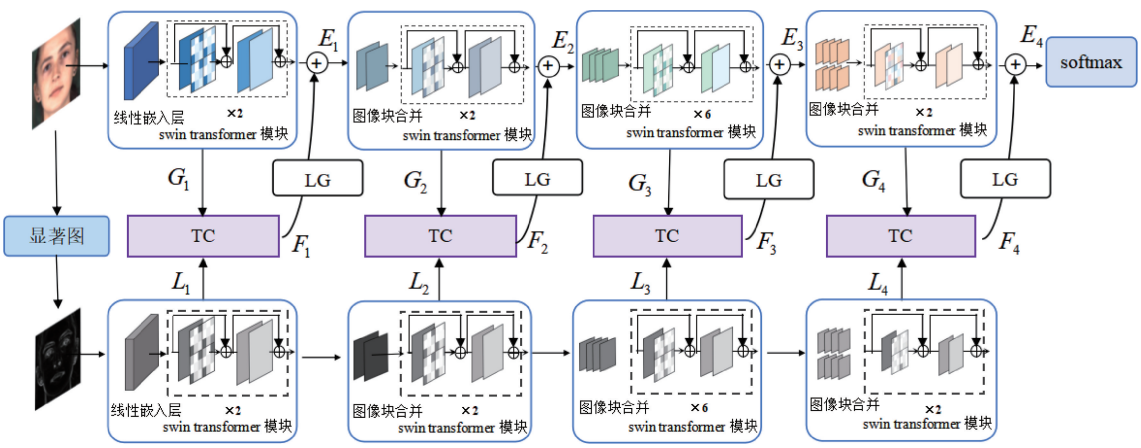


图 5 多尺度多模态视觉编码总体框架

Figure 5 A general framework for multiscale multimodal visual coding

和 W_i 分别表示第 i 阶段的通道数、特征图的高度和宽度。后 3 个阶段通过图像块合并降低输入特征图的分辨率,增加通道数逐层扩大感受野。

(2) 多模态融合。通过平铺拼接融合(tile-and-concatenate, TC)模块将 G_i 与 L_i 相结合,产生一组不同尺度的多模态特征,记为 $F_i \in \mathbb{R}^{H_i \times W_i \times C_i}$。

(3) 局部门控。通过门控单元对 F_i 中的每个元素进行加权,然后将其逐元素添加到 G_i 中,以产生一组嵌入局部描述符信息的增强视觉特征,记为 $E_i \in \mathbb{R}^{H_i \times W_i \times C_i}$。

2.2.2　局部描述符

为更有效地提取局部关键点,采用局部描述符来提取局部特征,如图 6 所示。局部二值图(local binary patterns, LBP)、能量图和显著图更适用于像素级捕获纹理、随机性和几何分析等特征,通过实验发现,显著图在面部表情识别方面表现最佳。本文使用自然统计的模型显著度自下而上计算显著度图[18]。虽然单一的显著图不足以追踪人脸情感分类的全部特征,但在视觉特征中分层添加显著图可以突出图像的局部兴趣点,有利于表情细节的检测。

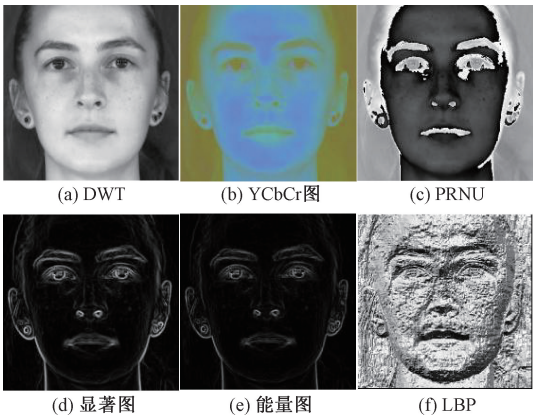


图 6　常用的局部描述特征图

Figure 6　Commonly used local description feature maps

2.2.3　多模态融合

为实现人脸与背景的有效分离,需确保全局人脸特征与局部显著特征的模态对齐。为此,本文采用图 7 所示的平铺拼接融合方法,通过特征对齐与融合,在保持图像空间结构和连续性的同时,有效整合全局信息与局部细节,提升视觉质量与自然度。给定输入全局特征 G_i 和局部特征 L_i,对每个阶段不同模态的不同尺度特征进行拼接,得到 $\widetilde{F}_i \in \mathbb{R}^{H_i \times W_i \times 2C_i}$,再通过卷积得到融合特征 F_i。

$$F_i = \mathrm{Conv}(\mathrm{concat}(G_i, L_i))。\tag{3}$$

式中:F_i 为第 i 阶段多模态融合特征;concat(·)为通道维度拼接操作;Conv(·)表示 3×3 卷积层。

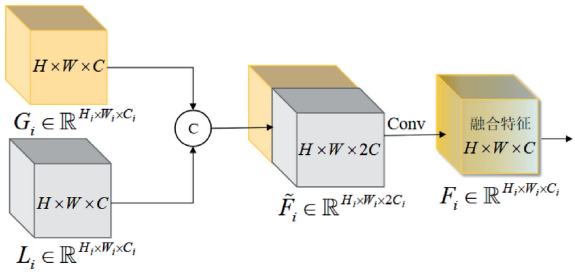


图 7　平铺拼接融合

Figure 7　Tile splice fusion

此外,为提高多个阶段图像之间的关联,使用如图 8 所示的局部门控单元[19],对融合特征进行自适应调节。局部门控是一种可学习的门控单元,能够对融合后的特征 F_i 中的每个元素进行加权,然后将其逐元素地添加到人脸图像的特征 G_i 中,从而产生一组嵌入局部描述符信息的增强视觉特征 E_i。局部门控的作用是防止融合后的特征对人脸图像的特征产生过强影响,从而保持人脸图像的原始信息,同时利用局部描述符的信息增强人脸图像的表情信息。首先,将融合后的特征 F_i,经过卷积和激活函数,得到 $S_i \in \mathbb{R}^{H_i \times W_i \times C_i}$,与原始的人脸图像和显著特征图像的大小相同。接着,将 S_i 和 F_i 进行哈达玛积(逐元素相乘),并将其结果逐元素地添加到 G_i 中,得到 E_i,即

$$S_i = \gamma_i(F_i(x,y))；\tag{4}$$

$$E_i(x,y) = S_i \odot F_i + G_i。\tag{5}$$

式中:S_i 为第 i 阶段门控权重矩阵;$F_i(x,y)$ 为第 i 阶段的融合特征值;γ_i 为激活函数组合;$E_i(x,y)$ 为增强视觉特征值;⊙表示元素乘法;G_i 作为残差项,用于保留原始的图像信息。为进一步提升 E_i 的特征质量,对 E_i 进行 2 层的卷积操作,分别是 1×1 卷积后接 ReLU 激活函数和双曲正切函数。

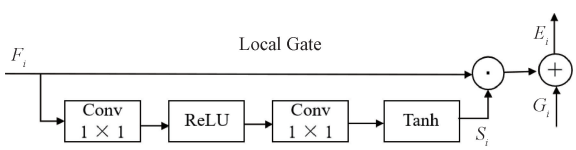


图 8　局部门控单元

Figure 8　Local gate unit

2.2.4　模型训练

面部情感识别是一个多分类问题。本文采用交叉熵损失函数作为代价函数,定义如下:

$$L = -\frac{1}{M}\sum_{i=1}^{M}\sum_{k=1}^{K} y_k^{(i)} \log(\hat{y}_k^{(i)})。\tag{6}$$

式中:L 为交叉熵损失值;M 表示样本数量;K 表示类别数量;$y_k^{(i)}$ 表示第 i 个样本的第 k 个类别的真实标签;$\hat{y}_k^{(i)}$ 表示第 i 个样本的第 k 个类别的模型输出

概率。

2.2.5 时间复杂度分析

本文方法的时间复杂度主要来源于2部分:生成扩散模型的数据扩充和多模态多尺度融合模型。DDPM生成N张图像需T_{step}步扩散,则:

$$O_{\text{DDPM}} = O(T_{\text{step}} \cdot N \cdot d^2) C_{\text{out}}。\tag{7}$$

式中:O_{DDPM}为DDPM数据扩充复杂度;T_{step}为扩散步数;N为生成图像数量;d为隐空间特征维度。

多模态融合模型的复杂度主要由swin transformer分支、平铺拼接融合和局部门控3个模块构成。

$$O_{\text{f}} = O(\sum_{i=1}^{4} H_i W_i (4C_i^2 + 2C_i(C_{\text{out}} + C_i)))\text{;}\tag{8}$$

$$O_{\text{total}} = O_{\text{DDPM}} + O_{\text{f}}。\tag{9}$$

式中:O_{f}为多模态融合模型复杂度;C_{out}为卷积输出通道数;O_{total}为方法总复杂度。

通过独立的生成扩散模型实现数据扩充、训练与多模态模型解耦。如表3所示,通过数据扩充带来的识别精度提升了8.95百分点,验证了本方法在计算效率与识别性能之间的有效平衡。

3 实验结果与分析

3.1 实验设置

3.1.1 数据集与数据扩充

实验采用2.1.1节所述的多源数据集,原始数据集共包含5 682张标注面部表情的样本。为确保数据独立性,采用分层随机划分策略,将数据集分为70%训练集、15%验证集和15%测试集。为改善类别不平衡问题,采用DDPM对训练集进行数据增强,生成约原始训练集规模2倍的高质量样本。其中,对稀缺类别进行重点扩充,使其样本量提升3到5倍,数据扩充前后表情类别占比如图9所示。

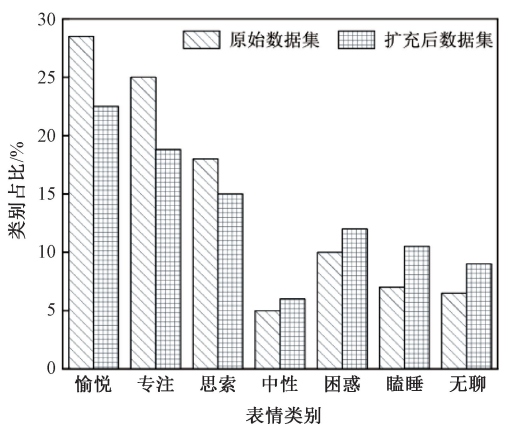


图9 数据扩充前后表情类别占比

Figure 9 The proportion of expression categories before and after data expansion

3.1.2 实验环境与参数设置

实验在Ubuntu 20.04系统上进行,硬件配置为Intel Xeon Gold 6226R 12核CPU(主频2.9 GHz),内存32 GB,GPU为NVIDIA RTX A6000(24 GB显存)。模型采用swin transformer作为主干网络,训练时使用AdamW优化器,实验的初始学习率为1e-5,权重衰减因子为1e-3,批量大小为32,训练过程中采用交叉熵损失函数进行优化。在训练过程中,如果验证集损失不再降低,则采用学习率衰减策略,学习率每5个epoch降低为原来的0.9倍。总迭代次数为200,每轮训练完后,在验证集上进行验证,保存最优模型权重。

3.2 对比实验

3.2.1 性能对比实验

为了评估本文方法,使用了3个不同的性能指标:准确度、平均准确度、宏观$F1$分数。实验选取了3类具有代表性的基准模型进行对比:基于卷积神经网络的ResNet-50[20]、VGG16[2]和EfficientNet[21],基于Transformer架构的ViT[8]和MAE-DFER[10],融合架构的ConvNeXt[22]和MaxViT[23]模型以及多模态Transformer方法POSTER[24]。将本文方法与上述基准模型在测试集上的性能进行比较,结果如表1所示。

表1 3个不同性能指标评估

Table 1 Three different performance metrics evaluated

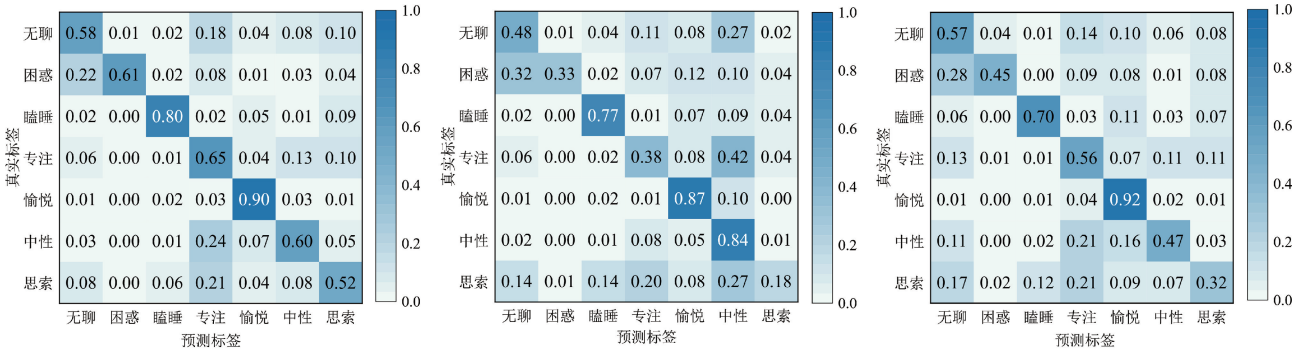
单位:%

模型	*Acc*	*M_Acc*	*F*1
ResNet-50	63.18	56.78	57.96
VGG16	65.39	59.75	60.31
EfficientNet	65.35	63.48	61.74
ViT	62.80	56.83	58.97
MAE-DFER	66.43	62.91	61.83
ConvNeXt	63.00	54.27	54.22
MaxViT	66.94	63.62	63.21
POSTER	67.90	65.20	66.23
本文方法	68.10	66.03	66.86

从表1可以看出,本文方法在准确率和其他指标上均优于基准模型,原因在于多模态多尺度融合机制整合了人脸图像的全局语义信息与局部的细微特征,克服了传统模型因依赖单一模态或单一尺度而导致的表征能力局限。层级视觉Transformer中采用的分阶段融合策略,通过自适应调整不同层次特征间的贡献权重,增强了模型对多源异质特征的表征与融合能力,使模型在面对类间差异小、特征耦合度高的复杂情感场景时,能保持稳定的判别性能,展现出较好的泛化能力与场景鲁棒性。

3.2.2　混淆矩阵分析

为了进一步验证本文方法对学习情感的精确捕捉能力,进行了混淆矩阵分析,如图 10 所示。从图中可以看出,思索情感类别的误分类概率较高。主要原因在于它与专注等情感的面部特征相似,传统模型难以实现有效区分。本文方法通过局部描述符捕捉微表情细节,如眼神、眉毛等的细微变化,结合平铺拼接融合模块增强全局-局部特征对齐,提升细粒度情感的分类精度。实验表明,该方法在“思索-专注”这类易混淆情感对上较 VGG16 等基准模型表现出更优的区分能力,验证了多尺度特征融合对细粒度情感识别的有效性。



(a) 本文模型混淆矩阵　(b) VGG16混淆矩阵　(c) ResNet50混淆矩阵

图 10　混淆矩阵对比

Figure 10　Comparison of confusion matrices

3.3　消融实验

3.3.1　模块间消融实验

为了评估不同模块对本文方法性能的影响,进行了消融实验。首先测试了未经修改的基线模型,然后移除了 DDPM 数据增广,之后分别在基线模型的基础上仅使用全局分支和替代平铺拼接融合方法。结果如表 2 所示:

表 2　消融实验

Table 2　ablation experiment　　单位:%

模型	*Acc*	*F*1
基线模型	68.10	66.86
移除 DDPM 数据增广	63.23	58.53
仅使用全局分支	59.64	53.23
替代融合方法	66.73	62.09

从表 2 可以看出,移除生成扩散模型后,模型的 *Acc* 和 *F*1 分数下降,原因为生成扩散模型的引入显著增强了数据多样性,有效缓解了数据不足的问题。仅依赖 RGB 图像的全局特征不足以捕捉学习表情的细微情感变化,本文方法通过结合显著特征图的局部特征,使模型能够更全面地表达情感细节。平铺拼接融合方式在结合全局与局部特征方面具有显著优势。其保留原始图像空间信息的同时,突出了重点特征区域,使得情感分类更具准确性。

3.3.2　参数敏感性分析

为评估模型对关键超参数的鲁棒性,本文对学习率与批量大小 2 个对模型训练效果影响显著的参数进行了敏感性分析。实验分别测试学习率为 1e-6,1e-5,1e-4 和批量大小为 16,32 和 64 的 9 组实验组合。实验采用控制变量法,固定其他训练参数分别测试了以上参数组合,实验结果如图 11 所示。

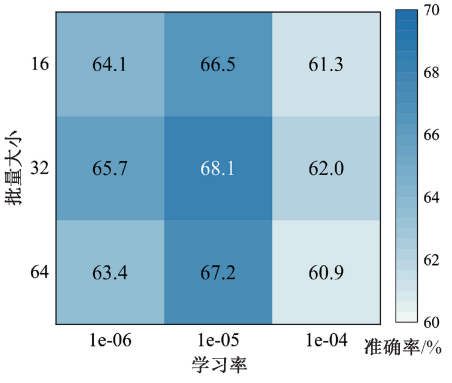


图 11　参数敏感性分析

Figure 11　Parameter sensitivity analysis

从图中可以观察到,模型的性能对不同超参数组合具有一定的敏感性。其中,学习率为 1e-5、批量大小为 32 时模型表现最佳,达到 68.1% 的准确率,显示出该组合在保持训练稳定性和有效性方面的良好平衡。相比之下,当学习率或批量过小或过大时,模型准确率均出现下降,表明超参数的选取对模型性能具有显著影响。

3.4　数据集扩充实验

3.4.1　扩充前后表情识别精度对比

为了评估通过 DDPM 生成的数据对模型性能的影响,进行了数据集扩充质量评价实验。分别在原始和扩充后数据集上训练模型,并对比其精确率 *Precision* 和召回率 *Recall*。结果如表 3 所示:

表3 精度对比实验

Table 3 Precision comparison experiment

单位:%

数据集	*Precision*	*Recall*
原始数据集	61.74	56.64
扩充后数据集	70.69	66.56

从表3可以看出,使用DDPM生成的数据扩充后,模型的精确率和召回率均显著提升。说明扩充数据提高了模型对目标类别的准确预测能力,减少了误分类。召回率的提升进一步表明,DDPM生成的表情样本增强了模型对稀缺类别的检测能力,使模型能够更全面地识别不同表情。

3.4.2 数据多样性分析

为了评估通过DDPM生成的数据扩充是否改善了数据集的多样性,并分析扩充数据与原始数据在质量和覆盖范围上的差异,在*FID*和*IS*指标上进行了数据多样性分析实验。结果如表4所示:

表4 数据多样性分析

Table 4 Data diversity analysis

评价指标	原始数据集	扩充后数据集
FID	—	44.58
IS	2.81	3.42

从表4可以看出,DDPM生成的扩充数据在质量和多样性方面均有显著提升。*FID*值为44.58,说明生成数据与真实数据的分布较为接近。同时,*IS*值从2.81提升至3.42,表明扩充数据不仅类别判别性更强,且多样性得到有效改善。同步优化证实了DDPM在平衡生成质量与多样性方面的优势。

4 结论与展望

本文针对学习情感识别中存在的细微表情特征捕捉困难与标注数据稀缺等核心挑战,提出了1种融合生成扩散模型与多模态多尺度视觉编码的解决方案。通过构建1个融合全局与局部细节的情感数据集,并利用生成扩散模型进行遵循数据分布的高质量样本扩充,有效改善了模型在少样本条件下的泛化能力。然后设计了1种基于层级Transformer的编码架构,通过全局人脸特征与局部显著区域特征的深度融合,实现了对微表情等高阶细粒度特征的鲁棒建模。实验表明,该方法在多个主流骨干网络上的识别性能均优于基线模型,验证了其整体优越性;同时,系统性的模块评估也证实了数据扩充策略与多尺度融合机制对于提升模型鲁棒性与准确性的关键作用。然而,本研究目前主要基于静态图像进行分析,未来工作将致力于引入时序动态信息并融合多模态行为线索,以构建更为完备的学习情感分析框架。

参考文献:

[1] 翟雪松, 许家奇, 王永固. 在线教育中的学习情感计算研究——基于多源数据融合视角[J]. 华东师范大学学报(教育科学版), 2022, 40(9): 32-44.
ZHAI X S, XU J Q, WANG Y G. Research on learning affective computing in online education: from the perspective of multi-source data fusion[J]. Journal of East China Normal University (Educational Sciences), 2022, 40(9): 32-44.

[2] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2014-09-04)[2025-05-25]. https://arxiv.org/abs/1409.1556.

[3] JAIN D K, SHAMSOLMOALI P, SEHDEV P. Extended deep neural network for facial emotion recognition[J]. Pattern Recognition Letters, 2019, 120: 69-74.

[4] SAJJAD M, ULLAH F U M, ULLAH M, et al. A comprehensive survey on deep facial expression recognition: challenges, applications, and future guidelines[J]. Alexandria Engineering Journal, 2023, 68: 817-840.

[5] HU M, YANG C J, ZHENG Y Q, et al. Facial expression recognition based on fusion features of center-symmetric local signal magnitude pattern[J]. IEEE Access, 2019, 7: 118435-118445.

[6] SINGH S, NASOZ F. Facial expression recognition with convolutional neural networks[C]//2020 10th Annual Computing and Communication Workshop and Conference (CCWC). Piscataway: IEEE, 2020: 324-328.

[7] SHARMA A, BAJAJ V, ARORA J. Machine learning techniques for real-time emotion detection from facial expressions[C]//2023 2nd Edition of IEEE Delhi Section Flagship Conference (DELCON). Piscataway: IEEE, 2023: 1-6.

[8] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: transformers for image recognition at scale[EB/OL]. (2020-10-22)[2025-05-25]. https://arxiv.org/abs/2010.11929.

[9] XUE F L, WANG Q C, GUO G D. TransFER: learning relation-aware facial expression representations with transformers[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2022: 3581-3590.

[10] SUN L C, LIAN Z, LIU B, et al. MAE-DFER: efficient masked autoencoder for self-supervised dynamic facial expression recognition[C]//Proceedings of the 31st ACM International Conference on Multimedia. New York: ACM, 2023: 6110-6121.

[11] GAO J X, ZHAO Y Y. TFE: a transformer architecture for occlusion aware facial expression recognition[J]. Frontiers in Neurorobotics, 2021, 15: 763100.

[12] HO J, JAIN A, ABBEEL P. Denoising diffusion probabi-

listic models[EB/OL]. (2020-06-19)[2025-05-25]. https://arxiv.org/abs/2006.11239.

[13] LI J T, DONG Z Z, LU S Y, et al. CAS(ME)3: a third generation facial spontaneous micro-expression database with depth information and high ecological validity[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(3): 2782-2800.

[14] LI S, DENG W H, DU J P. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2017: 2584-2593.

[15] WANG Z X, ZHANG J W, CHEN R J, et al. RestoreFormer: high-quality blind face restoration from undegraded key-value pairs[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 17491-17500.

[16] 姬莉霞, 周洪鑫, 肖士杰, 等. 一种基于邻域注意力的扩散模型训练方法研究[J]. 计算机工程, 2025, 51(8): 262-269.
JI L X, ZHOU H X, XIAO S J, et al. A research on training method for diffusion model based on neighborhood attention[J]. Computer Engineering, 2025, 51(8): 262-269.

[17] LIU Z, LIN Y T, CAO Y, et al. Swin transformer: hierarchical vision transformer using shifted windows[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2022: 9992-10002.

[18] ANWAR M A, TAHIR S F, FAHAD L G, et al. Image forgery detection by transforming local descriptors into deep-derived features[J]. Applied Soft Computing, 2023, 147: 110730.

[19] YANG Z, WANG J Q, TANG Y S, et al. LAVT: language-aware vision transformer for referring image segmentation[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 18134-18144.

[20] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2016: 770-778.

[21] TAN M X, LE Q V. EfficientNet: rethinking model scaling for convolutional neural networks[EB/OL]. (2019-05-28)[2025-05-25]. https://arxiv.org/abs/1905.11946.

[22] LIU Z, MAO H Z, WU C Y, et al. A ConvNet for the 2020s[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 11966-11976.

[23] TU Z, TALEBI H, ZHANG H, et al. MaxViT: multi-axis vision transformer[C]//2022 European Conference on Computer Vision (ECCV). Cham: Springer, 2022: 459-479.

[24] ZHENG C, MENDIETA M, CHEN C. POSTER: a pyramid cross-fusion transformer network for facial expression recognition[C]//2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). Piscataway: IEEE, 2023: 3138-3147.

Multimodal and Multiscale Facial Expression Recognition Analysis for Learning Emotions

JI Lixia, REN Hanliang, WANG Wei, DU Yunlong, ZHOU Hongxin, FU Yuanzhong

(School of Cyber Science and Engineering, Zhengzhou University, Zhengzhou 450002, China)

Abstract: To address the challenges of capturing subtle features and the scarcity of data samples in learning emotion recognition, a facial emotion recognition method based on a generative diffusion model and multimodal multiscale visual encoding was proposed. Firstly, a learning emotion dataset integrating multi-scale global and local detailed features was constructed, and the generative diffusion model was used to augment scarce emotional samples, thereby alleviating data constraints in few-shot learning scenarios. Secondly, a multimodal multiscale visual encoding mechanism was designed, which achieved high-precision modeling and effective fusion of micro-expressions and fine-grained emotional features by combining global features of facial images with local details from salient regions. Finally, the experiments were conducted on various models, including CNNs, Vision Transformers, and hybrid architectures. The results showed that the proposed method achieved an overall recognition accuracy of 68.10%, with an average improvement of 2.98% and a maximum improvement of 5.30% compared with existing baseline methods. The ablation experiments further verified the effectiveness and synergistic contribution of the generative diffusion model and the multimodal multiscale fusion module in enhancing the model's capability to capture micro-expression details and improving overall recognition robustness.

Keywords: learning emotions; facial expression recognition; artificial intelligence; multimodal; multiscale features