林帅男, 张伟, 胡敏, 等. DFNet: 融合多尺度特征与自注意力的表情识别算法[J]. 智能计算机与应用,2024,14(6):64-70.
DOI:10.20169/j.issn.2095-2163.240609

DFNet: 融合多尺度特征与自注意力的表情识别算法

林帅男, 张 伟, 胡 敏, 赵 瑞
(吉林师范大学 数学与计算机学院, 吉林 四平 136000)

摘 要: 为解决人脸表情识别时存在的特征表达能力不足以及识别率不高的问题,提出了一种新的融合多尺度特征与自注意力的表情识别算法-DFNet。进行多尺度特征融合时,通过采用空洞卷积以及通道降维的形式,在扩大感受野的同时,获得多尺度信息;提出一种快自注意力机制,改进了传统的 Transformer block,提升了模型的特征提取能力,进一步提高了模型性能。实验结果表明,所提方法在 RAF-DB 和 KDEF 表情数据集分别取得了 89.31%和 89.05%的准确率,证明了所提网络具有较强的泛化性。

关键词: 人脸表情识别; 残差网络; 自注意力机制; 空洞卷积; 多尺度
中图分类号: TP391.4 **文献标志码**: A **文章编号**: 2095-2163(2024)06-0064-07

DFNet:Expression recognition algorithm based on fusion of multi-scale features and self attention

LIN Shuainan, ZHANG Wei, HU Min, ZHAO Rui
(College of Mathematics and Computer, Jilin Normal University, Siping 136000, Jilin, China)

Abstract: To address the problems of large model size, high computational complexity, and low recognition rate in facial expression recognition, a new expression recognition algorithm based on fusion of multi - scale features and self attention mechanism- DFNet is proposed. When performing multi - scale feature fusion, by using dilated convolution and channel dimensionality reduction, multi-scale information is obtained while expanding the receptive field. A fast self attention mechanism is proposed that improves the traditional Transformer block and reduces the number of parameters, therefore improves the performance of the model. The experimental results show that the proposed method achieves accuracy of 89.31% and 89.05% on the RAF-DB and KDEF expression datasets, respectively, proving that the proposed network has strong generalization ability.

Key words: facial expression recognition; residual network; self attention mechanism; dilated convolution; multi-scale

0 引 言

人脸表情是人们在日常生活中进行交流的非语言交往手段,是人们传达情感和意图的最基本方式。随着人工智能算法的发展,表情识别技术也取得了显著进步。人脸表情识别主要分为人脸检测、图像预处理、特征提取以及表情分类四个部分[1]。其中,特征提取在图像处理领域占据重要地位,是表情识别的前提与基础。

早期的人脸表情识别主要是基于传统手工方法[2]。2006 年,Hinton 等学者[3]开启了深度学习的研究进程。与传统的特征提取方法相比,基于深度学习网络的 FER 方法省去了人工提取特征的步骤,在处理大量复杂数据时可以取得更好的效果,泛化性更强,并且对不同光照条件、遮挡物和多姿态等人脸表情图像的识别更加有效。在图像识别方面,卷积神经网络(Convolutional Neural Networks,CNN)得到了广泛应用,通过采用卷积层、池化层以及全连接层等多种层次相互连接的方式,从输入的图像数据中进行信息过滤,提取出更高层次的特征表示,进而提高模型识别性能。

基金项目: 吉林省科技厅科研项目(20230101243JC)。
作者简介: 林帅男(2000-),女,硕士研究生,主要研究方向:图像识别; 张 伟(1981-),男,博士,副教授,主要研究方向:智能影像处理; 胡敏(1999-),女,硕士研究生,主要研究方向:图像识别。
通讯作者: 赵 瑞(1975-),女,博士,教授,硕士生导师,主要研究方向:数值模拟,智慧教育。Email:lijiatong_zr@163.com
收稿日期: 2023-12-06

Gao 等学者[4]提出了一种新的 FER 方法，在 CNN 中引入光谱和空间注意(SSA)模块以及数据集内持续学习(ICL)模块，有效解决了表情识别过程中空间动作单元的相互作用、光谱表达语义信息的不足以及数据分布不平衡的问题。针对网络参数量较大等问题，李嘉乾等学者[5]基于残差网络，通过引入深度可分离卷积，在一定程度上改善了参数量特征。Hossain 等学者[6]则是将基本 CNN 与双线性 CNN 架构融合，这样在降低网络运算参数的同时，增强了网络中表情特征的权重，提升了网络性能，但上述方法仅在静态的面部动作表情识别上有良好表现。那么，为体现面部表情识别网络的实时性能，Sun 等学者[7]提出了一种多模态层次融合策略，通过将卷积神经网络(CNN)与双边长短时记忆循环神经网络(BiLSTM-RNN)相结合，从视频中学习时空层次特征。Zheng 等学者[8]提出了一种高效的全局和局部感知集成网络，在实时状态下准确、快速提取面部特征的同时，解决了人脸图像特征提取过程中的信息不完全以及大规模数据集鲁棒性较弱的问题。

本文针对人脸表情识别特征提取能力不足、精确度低的问题，进行了 2 个方面的改进：

(1)融合多尺度特征时，利用空洞卷积、通道降维的形式，增大感受野，获得多尺度信息；

(2)提出一种快自注意力机制(Fast Self Attention，FSA)，提高网络中面部表情特征的权重，从而进一步提升 FER 精度。

1　表情识别网络模型

理论上来讲，随着网络层数的加深，网络的性能会越来越好，但研究表明，过深的网络反而会出现梯度爆炸和梯度消失的现象，从而导致网络性能的下降。对此，本文构建了一种基于残差网络(Residual Network，ResNet)的 FER 方法，选用 ResNet50 作为基准框架，提取基础图像特征，增强网络的表达能力，模型结构如图 1 所示。图 1 中，左图为 DF-Net 整体结构；中间图为各模块具体结构；右图为 Bottleneck 具体结构。

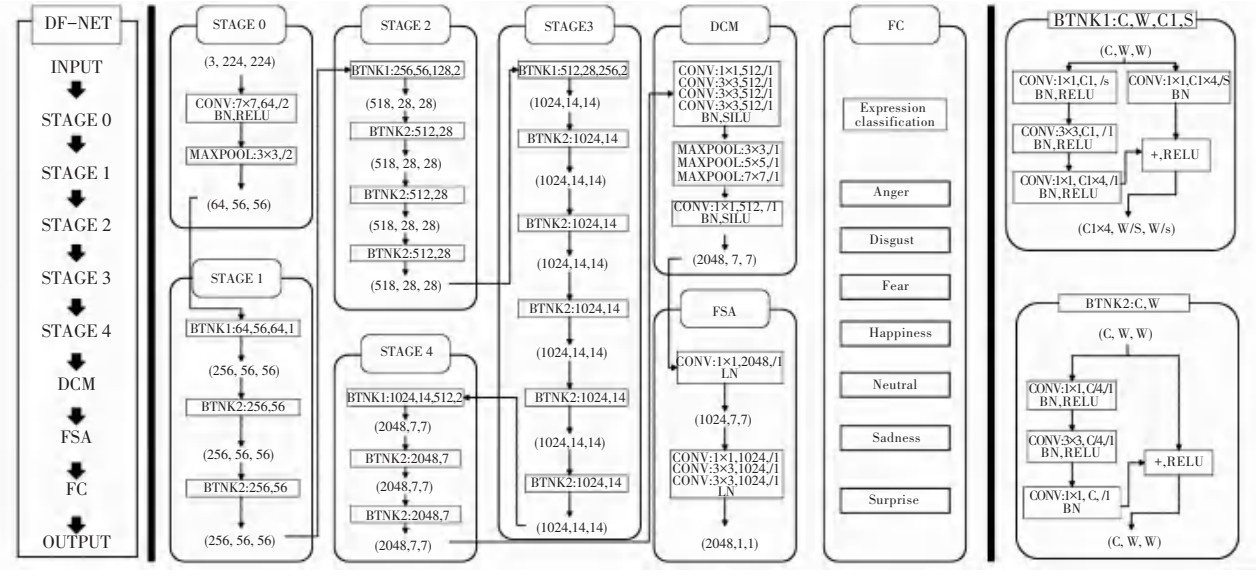

图 1　模型结构图

Fig. 1　Model structure diagram

1.1　DCM 模块

本文提出的空洞卷积-多尺度(Dilated Convolution Multi-scale，DCM)模块，在进行多尺度特征融合时，主要采用空洞卷积的形式，利用不同的膨胀率，使用 3×3 的卷积替代 5×5 和 7×7 的卷积，增大了感受野，同时采用了通道降维的形式，在一定程度上提升模型的检测精度。DCM 结构如图 2 所示。

输入 7×7 的特征图后，将并行经过 1 次 1×1 卷积，1 次 3×3 卷积，2 次 3×3 卷积和 3 次 3×3 卷积的数据进行 concat 操作，也就是通道方面的叠加，通道数由 $C/4$ 变为 C。接着在 BatchNorm2d 和 SILU 层进行激活，后利用 1×1 的卷积实现通道降维。为加速提高网络的收敛速度和稳定性，并提高网络的泛化能力，再次经过 BatchNorm2d 和 SILU 层，激活完成后对所有数据进行整合。

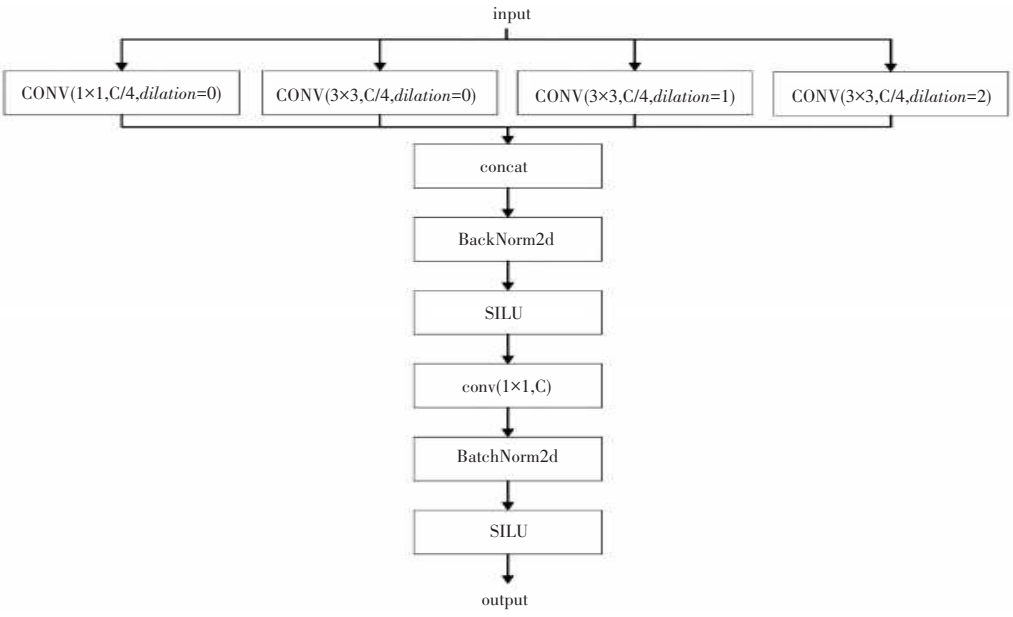

图 2 DCM 结构图

Fig. 2 DCM structure diagram

利用多尺度特征融合进行表情识别,可以在一定程度上改进特征的表达能力。与单一尺度的特征相比,多尺度特征融合能够更加有效地捕捉表情的特征信息,从而提高 FER 的精度与鲁棒性。

空洞卷积(Dilated Convolution)也叫扩张卷积或者膨胀卷积,也就是在卷积核元素之间填充一些空格(0)来扩大卷积核的过程。假设以一个变量 r 来衡量空洞卷积的扩张系数,则加入空洞后的实际卷积核尺寸与原始卷积核尺寸之间的关系满足式(1):

$$K = k + (k - 1)(r - 1) \tag{1}$$

其中,k 为原始卷积核大小;r 为空洞卷积扩张率(dilation rate);K 为经过扩展后实际卷积核大小。

在进行多尺度特征融合时,使用空洞卷积有如下好处:

(1)扩大感受野。相对于池化来说,使用空洞卷积可以在扩大感受野的同时不丢失分辨率,从而获得更加密集的数据;

(2)获取多尺度上下文信息。当多个带有不同空洞卷积扩张率的空洞卷积核叠加时,不同的感受野会带来多尺度信息。

1.2 FSA 模块

针对 Attention 机制,目前核心思路可阐释分述如下:

(1)对每个输入元素计算一个非负对归一化权重;

(2)将这些权重与对应成分的表示相乘;

(3)将得到的结果求和,产生一个固定长度的表示。

那么,为了更好地建立神经网络对于多个相关输入的相关性,主要使用自注意力机制来解决这个问题,自注意力机制实际上是想让机器注意到整个输入中不同部分之间的相关性。

本文提出的快自注意力机制(Fast Self Attention,FSA),通过降维、替换全连接等方式以较低的计算量,提高模型对特征的提取能力,进而提升模型性能。FSA 结构如图 3 所示。图 3 中,左图为 FSA 整体结构图,中间图为多头自注意力机制(Multi-Headed Self-Attention, MHSA)模块,右图为多层感知机(Multi Layer Perceptron, MLP)模块。

为避免因维度过高导致计算量增大、产生过拟合现象,FSA 主要通过 1×1 卷积和维度的一些变换来达到全连接的效果。首先将上一层的输出作为 FSA 模块的输入,输入的 tensor 维度是 $batchsize \times 2\,048 \times 7 \times 7$,输入通道数设为 c。随后,通过 1×1 的卷积实现降维操作,输出通道数是 $c \times r$,r 设为0.25,即把 c 压缩成了原来的四分之一。之后,为了让中间层的数据分布稳定,便于训练,使用 LayerNorm 层进行层归一化。将归一化后的数据输入到 MHSA 多头注意力模块,随后生成一个残差,再次在 LayerNorm 层进行归一化,经过 MLP 多层感知机后,生成一个残差,最后经过 1×1 卷积层升维后输出。

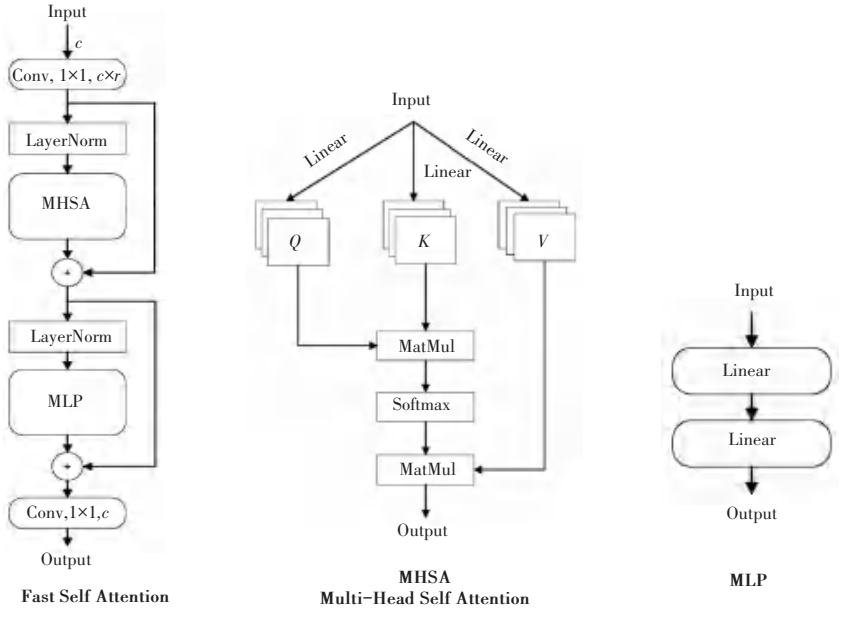

图 3　FSA 结构图

Fig. 3　FSA structure diagram

多头注意力(MHSA)是 Transformer 模型中的一个重要组成部分,这种结构的设计可以均衡同一种注意力机制可能产生的偏差,丰富特征子空间的多样性,从而提升模型的训练效果。Linear 代表线性变换,具体实现时使用卷积+维度变换代替全连接得到多头的 Q(query)、K(key)、V(value),并行整合每个头特征,分别求注意力后再连接起来。

MLP 多层感知机中的 2 层 Linear 也同样使用卷积+维度变换实现语义空间的转换。

1.3　算法流程

本文提出的 FER 方法,算法流程如图 4 所示。这里将对各流程步骤展开剖析论述如下。

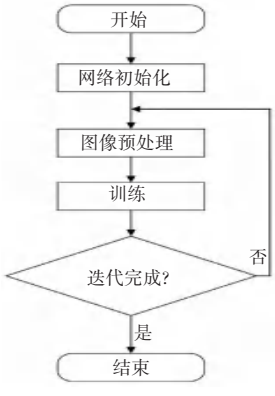

图 4　算法流程图

Fig. 4　Flow chart of the algorithm

(1) 网络初始化:本步骤完成网络的构建及初始化;

(2) 图像预处理:在每次迭代训练前,需要对图像进行预处理,预处理过程包括:

① 图像缩放至 224×224;

② 图像归一化至[0, 1],可由式(2)来确定:

$$y = \frac{\boldsymbol{x}_{\text{img}}}{255} \tag{2}$$

其中,$\boldsymbol{x}_{\text{img}}$ 表示图像矩阵。

③ 白化处理:降低数据间的相关度,即降低不同数据所含信息的重复性,从而提高网络的训练效率,可由式(3)进行描述:

$$y = \frac{\boldsymbol{x}_{\text{img}} - \boldsymbol{mean}}{\boldsymbol{std}} \tag{3}$$

其中,***std*** 为数据集标准差[0.229, 0.224, 0.225];***mean*** 为数据集均值[0.485, 0.456, 0.406]。

④ 随机擦除:随机选择某个区域擦除该区域内的图像信息,使得模型对遮挡更具有鲁棒性。具体是将该区域的像素值设置为 0,执行该操作的概率为 0.5。

(3)训练:将预处理后的图像输入到 Backbone 中,接着输入到 DCM 模块与 FSA 模块,最终输入到全连接层。训练过程中,采用 Adam 优化器,初始学习率为 0.000 1, *epoch* 设为 80;

(4)判断训练是否完成:若完成(训练轮数达到设置的 *epochs*) 则退出;否则,返回至步骤(2)。

2　实验

2.1　实验准备

本文的实验操作系统是 Ubuntu 20.04.3 LTS,实

验环境是采用了 Pytorch1.9.0 框架搭建,硬件平台 CPU 为 Intel(R) Xeon(R) Gold 5218R,GPU 为6 G 的 NVIDA GeForce RTX3080Ti。采用了 Adam 优化器,学习率设为 0.000 1,实 验 *batchsize* 设置为 32,训练 *epoch* 为 80。

2.2 数据集

分别使用 RAF-DB 表情数据集和 KDEF 表情数据集进行实验。

RAF-DB 数据集[9]由29 672张真实世界图像组成,提供了 5 个精确的人脸关键点,包含 7 种基本表情以及 12 种复合表情。为了验证本文提出方法的有效性,本次实验中选取了 7 种基本表情图像中的 12 271张图像作为训练样本,3 068 张图像作为测试样本。图像原始尺寸为 100×100,经过处理后输入到模型的尺寸224×224。图5为RAF-DB数据集样本分布。RAF-DB 数据集中的 7 种表情样例图像如图 6 所示。

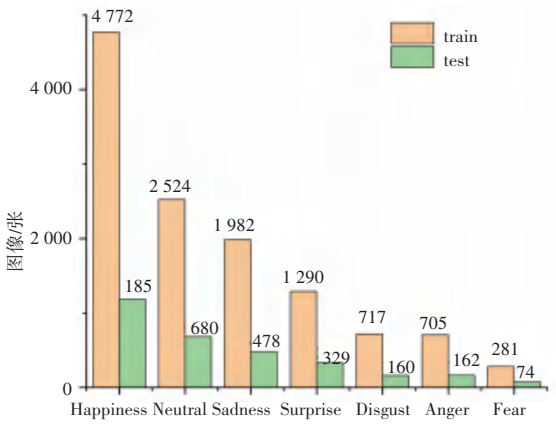

图 5 RAF-DB 数据集各类表情分布

Fig. 5 Distribution of various expressions in the RAF-DB

图 6 RAF-DB 表情库部分图像

Fig. 6 Partial images of RAF-DB emotion library

KDEF 数据集[10]是在光照比较柔和的情况下采集的,被采集者没有胡须、耳环或眼镜,因此不会受到这些因素的遮挡影响。该数据集包括 7 种不同的表情,每个表情有 5 个角度,共有 4 900 张彩色图。在本实验中,选取其中的3 911张图像作为训练样本,977 张图像作为测试样本,图像原始尺寸为 562×762,经过处理后输入到模型的尺寸为 224×224。图 7 为 KDEF 数据集样本分布。KDEF 数据集中的 7 种表情样例图像如图 8 所示。

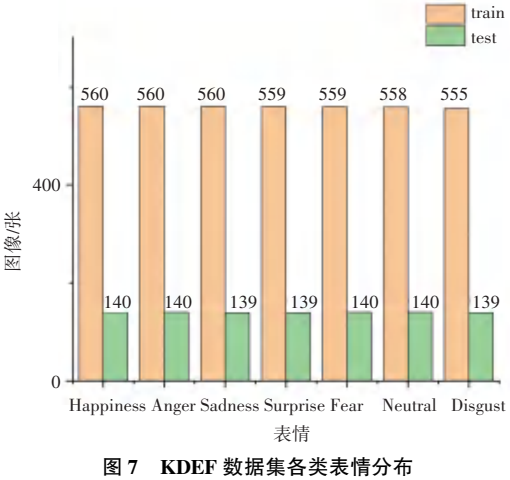

图 7 KDEF 数据集各类表情分布

Fig. 7 Distribution of various expressions in the KDEF

图 8 KDEF 表情库部分图像

Fig. 8 Partial images of KDEF emotion library

2.3 评价指标

为评估网络模型的有效性,使用分类精度(*Accuracy*)来计算模型的平均识别率,即所有预测对的数量/所有样本数量。可由式(4)来求得:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{4}$$

其中,*Accuracy* 表示分类精度;*TP*、*TN*、*FP*、*FN* 分别表示真阳性、真阴性、假阳性以及假阴性。

2.4 实验结果分析

混淆矩阵能够详细描述每种表情识别精度和被误区分为其他表情的比例。对本文的网络模型在 2

个数据集上进行了混淆矩阵测试实验，并通过图 9 和图 10 展示了实验结果。图 9、图 10 中每行都代表真实类别，每列都代表预测类别，对角线项表示每个表情的识别精度。

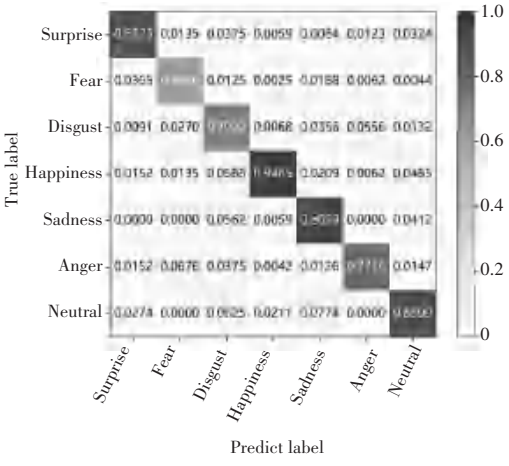

图 9　RAF-DB 数据集混淆矩阵

Fig. 9　RAF-DB dataset confusion matrix

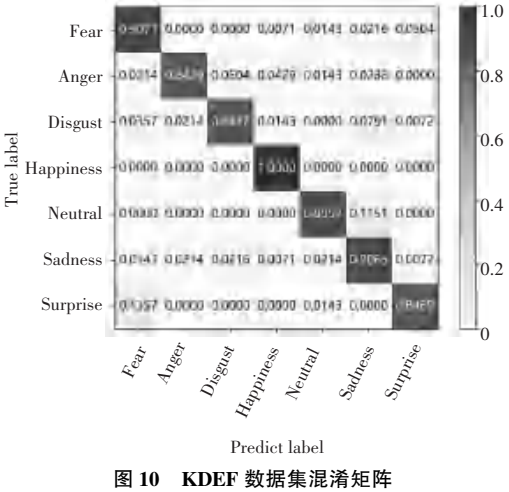

图 10　KDEF 数据集混淆矩阵

Fig. 10　KDEF dataset confusion matrix

实验结果表明，本文的网络模型在 RAF-DB 数据集的实验中，对 Happiness 类识别率较高，对 Disgust 和 Fear 的识别率较低，可能的原因之一是数据集中所选的大部分样本都是 Happiness，而 Fear 与 Disgust 样本相对较少，因此可能会因样本分布不平衡对表情识别产生偏差；除此之外，几类表情之间存在一定程度的混淆，导致提取到的特征相似度较高，因此各类表情识别率会存在一定差距。因 Happiness 表情与其他表情相比较为独特，识别率最高，但 Sadness 和 Disgust 表情与其它表情存在相似特征，容易出现分类错误，因此 Sadness 和 Disgust 表情的识别率相对较低。

相比之下，在 KDEF 数据集的实验中，各类表情的样本分布较均匀，虽然各个表情存在轻微混淆，但其特征较容易区分，因此所提网络模型对各类表情均有较好识别。由此可以说明，通过添加 DCM 模块与 FSA 模块可以有效提高表情识别的鲁棒性。

为验证本文算法对人脸表情特征提取更具优势，在 RAF-DB 和 KDEF 公开表情数据集上，与近年其他先进表情识别算法进行对比，对比实验结果见表 1。由表 1 可知，本文构建的网络模型具有更强的泛化性能。

表 1　与其他方法对比结果

Table 1　Comparison table with other methods

数据集	算法	识别率/%
RAF-DB	eXnet[11]	86.37
	DDL[12]	87.71
	IF-GAN[13]	88.33
	FT-CSAT[14]	88.61
	DFNet	**89.31**
KDEF	DML-Net[15]	88.20
	P-PCANet[16]	84.08
	SAFEPA[17]	84.19
	SSA-Net[18]	88.50
	DFNet	**89.05**

2.5　消融实验

本文研究的主要改进重点在空洞卷积-多尺度与自注意力机制这两部分，为了证明文中所提各模块的有效性与必要性，本文采用了消融实验来进行验证。其中，ResNet 对应 Baseline 结果，ResNet + DCM 是只添加空洞卷积-多尺度的结果，ResNet + FSA 是只添加快自注意力机制的结果，ResNet+DCM+FSA 是添加了空洞卷积-多尺度和快自注意力机制的结果、即本文网络。实验结果见表 2。

表 2　消融实验

Table 2　Ablation experiment

ResNet	DCM	FSA	RAF-DB 识别率/%	KDEF 识别率/%	*Params*/M
√			88.40	87.61	23.52
√	√		88.82	88.02	57.09
√		√	88.66	88.84	32.97
√	√	√	89.31	89.05	66.53

由表 2 可以看出，随着 baseline 添加 DCM 与 FSA 模块，虽然参数量有所增加，但 2 个数据集上表情识别率均有所提升。在二者的共同作用下，相较于基础网络，准确率分别提高了 0.91% 和 1.44%，且相比添加单一模块，模型性能均有明显提升，表明两

者在共同作用下效果更优,验证了本文所提网络的有效性。

3 结束语

本文提出一种新的表情识别方法、即添加 DCM 模块,在增大感受野的同时,有效获得了多尺度信息;融合了 FSA 模块,提高了模型对特征的提取能力,进而提高了表情识别效率。在已公开的人脸表情数据集 RAF-DB 以及 KDEF 上做了对比实验来进行模型评估。实验结果显示,本文提出的表情识别方法较现有模型来说有较高的识别准确率与泛化性。未来的研究中,在考虑降低模型参数的同时,要考虑更多样本的人脸表情识别,实现由单目标到多目标的识别。

参考文献

[1] 魏艳涛,雷芬,胡美佳,等. 学生表情识别研究综述[J]. 中国教育信息化,2020(21):48-55.

[2] LOGIE R H, BADDELEY A D, WOODHEAD M M. Face recognition, pose and ecological validity[J]. Applied Cognitive Psychology, 2015, 1(1): 53-69.

[3] HINTON G E, SALAKHUTDINOV R R. Reducing the dimensionality of data with neural networks[J]. Science, 2006, 313(5786): 504-507.

[4] GAO Hongxiang, WU Min, CHEN Zhenghua, et al. SSA-ICL: Multi-domain adaptive attention with intra-dataset continual learning for Facial expression recognition[J]. Neural Networks, 2023, 158: 228-238.

[5] 李嘉乾,张雷. 基于深度可分离卷积的表情识别改进方法[J]. 智能计算机与应用,2023,13(5):58-63,69.

[6] HOSSAIN S, UMER S, ROUT R K, et al. Fine-grained image analysis for facial expression recognition using deep convolutional neural networks with bilinear pooling [J]. Applied Soft Computing, 2023, 134: 109997.

[7] SUN Bo, CAO Siming, HE Jun, et al. Affect recognition from facial movements and body gestures by hierarchical deep spatio-temporal features and fusion strategy[J]. Neural Networks, 2018, 105: 36-51.

[8] ZHENG He, MENG Bin, WANG Lining, et al. Global and local fusion ensemble network for facial expression recognition [J]. Multimedia Tools and Applications, 2023, 82(4): 5473-5494.

[9] LI Shan, DENG Weilong, DU Junping. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Hawaii, USA :2017: 2852-2861.

[10] CALVO M G, LUNDQVIST D. Facial expressions of emotion (KDEF): Identification under different display - duration conditions[J]. Behavior Research Methods, 2008, 40(1): 109-115.

[11] RIAZ M N, SHEN Yao, SOHAIL M, et al. Exnet: An efficient approach for emotion recognition in the wild[J]. Sensors, 2020, 20(4): 1087.

[12] RUAN Delian, YAN Yan, CHEN Si, et al. Deep disturbance-disentangled learning for facial expression recognition [C]// Proceedings of the 28th ACM International Conference on Multimedia. Seattle, USA :ACM,2020: 2833-2841.

[13] CAI Jie, MENG Zibo, KHAN A S, et al. Identity-free facial expression recognition using conditional generative adversarial network [C]//2021 IEEE International Conference on Image Processing (ICIP). Anchorage, USA: IEEE, 2021: 1344-1348.

[14] YAO H, YANG X, CHEN D, et al. Facial expression recognition based on fine-tuned channel - spatial attention transformer[J]. Sensors, 2023, 23(15): 6799.

[15] LIU Yuanyuan, DAI Wei, FANG Fang, et al. Dynamic multi-channel metric network for joint pose-aware and identity-invariant facial expression recognition [J]. Information Sciences, 2021, 578: 195-213.

[16] SUN Zhe, ZHANG Hehao, MA Suwei, et al. Combining filtered dictionary representation based deep subspace filter learning with a discriminative classification criterion for facial expression recognition[J]. Artificial Intelligence Review, 2022, 55(8): 6547-6566.

[17] ALGHAMDI T, ALAGHBAND G. SAFEPA: An expandable multi-pose facial expressions pain assessment method[J]. Applied Sciences, 2023, 13(12): 7206.

[18] LIU Yuanyuan, PENG Jiyao, DAI Wei, et al. Joint spatial and scale attention network for multi-view facial expression recognition [J]. Pattern Recognition, 2023, 139: 109496.