深度多尺度融合注意力残差人脸表情识别网络

高涛, 杨朝晨, 陈婷, 邵倩, 雷涛

引用本文:
高涛,杨朝晨,陈婷,邵倩,雷涛. 深度多尺度融合注意力残差人脸表情识别网络[J]. 智能系统学报, 2022, 17(2): 393–401.
GAO Tao,YANG Zhaochen,CHEN Ting,SHAO Qian,LEI Tao. Deep multiscale fusion attention residual network for facial expression recognition[J]. *CAAI Transactions on Intelligent Systems*, 2022, 17(2): 393–401.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202107028

DOI: 10.11992/tis.202107028
网络出版地址: https://kns.cnki.net/kcms/detail/23.1538.TP.20211202.1101.004.html

深度多尺度融合注意力残差人脸表情识别网络

高涛[1], 杨朝晨[1], 陈婷[1], 邵倩[1], 雷涛[2]

（1. 长安大学 信息工程学院, 陕西 西安 710000; 2. 陕西科技大学 电子信息与人工智能学院, 陕西 西安 710021）

摘　要: 针对人脸表情呈现方式多样化以及人脸表情识别易受光照、姿势、遮挡等非线性因素影响的问题, 提出了一种深度多尺度融合注意力残差网络 (deep multi-scale fusion attention residual network, DMFA-ResNet)。该模型基于 ResNet-50 残差网络, 设计了新的注意力残差模块, 由 7 个具有三条支路的注意残差学习单元构成, 能够对输入图像进行并行多卷积操作, 以获得多尺度特征, 同时引入注意力机制, 突出重点局部区域, 有利于遮挡图像的特征学习。通过在注意力残差模块之间增加过渡层以去除冗余信息, 简化网络复杂度, 在保证感受野的情况下减少计算量, 实现网络抗过拟合效果。在 3 组数据集上的实验结果表明, 本文提出的算法均优于对比的其他先进方法。

关键词: 人脸表情识别; 残差网络; 多尺度特征; 注意力机制; 遮挡人脸; 卷积神经网络; 特征融合; 深度学习
中图分类号: TP391　　**文献标志码**: A　　**文章编号**: 1673−4785(2022)02−0393−09

中文引用格式: 高涛, 杨朝晨, 陈婷, 等. 深度多尺度融合注意力残差人脸表情识别网络 [J]. 智能系统学报, 2022, 17(2): 393–401.
英文引用格式: GAO Tao, YANG Zhaochen, CHEN Ting, et al. Deep multiscale fusion attention residual network for facial expression recognition[J]. CAAI transactions on intelligent systems, 2022, 17(2): 393–401.

Deep multiscale fusion attention residual network for facial expression recognition

GAO Tao[1], YANG Zhaochen[1], CHEN Ting[1], SHAO Qian[1], LEI Tao[2]

(1. School of Information Engineering, Chang'an University, Xi'an 710000, China; 2. School of Electronic Information and Artificial Intelligence, Shaanxi University of Science and Technology, Xi'an 710021, China)

Abstract: This paper proposes a deep multiscale fusion attention residual network based on the ResNet-50 model to solve the problems of the diversification of facial expression presentation and the susceptibility of facial expression recognition to nonlinear factors, such as illumination, posture, and occlusion. A novel attention residual module consisting of seven attention residual learning units with three branches is designed to perform multiple convolution operations on the input image in parallel and obtain multiscale features. To highlight important local areas, the attention mechanism is introduced simultaneously, which is conducive to the feature learning of the occluded images. Furthermore, a novel transition layer is added between the attention residual modules to remove redundant information, simplify the network complexity, reduce the amount of calculation while ensuring the receptive field, and realize the anti-overfitting effect of the network. Experimental results on three datasets demonstrate that the proposed algorithm is superior to other advanced methods.

Keywords: facial expression recognition; residual network; multiscale features; attention mechanism; occlusion of human faces; convolution neural network; feature fusion; deep learning

情绪包含大量的情感信息, 当人们面对面交流时, 情绪会自动或不自觉地通过面部表情表现出来[1]。随着人工智能技术的飞速发展, 人脸表情识别 (FER) 已成为计算机图像处理中一个重要的研究课题。

人脸表情识别主要包括预处理、特征提取和分类识别 3 个部分[2]。其中, 算法识别精度高低主要由特征提取方法决定。人脸表情特征提取方法主要分为基于传统特征提取的方法和基于深度

收稿日期: 2021−07−16.　**网络出版日期**: 2021−12−05.
基金项目: 国家重点研发计划项目 (2019YFE0108300); 国家自然科学基金项目 (62001058); 陕西省重点研发计划项目 (2019GY-039); 长安大学中央高校基本科研业务费专项资金项目 (300102241201).
通信作者: 陈婷. E-mail: tchenchd@126.com.

学习的方法[3]。传统的特征提取方法主要包括局部二值模式 (LBP)[4]、类 Haar 特征[5]、Gabor 小波变换[6] 和方向梯度直方图 (HOG) 等。Li 等[7] 基于 LBP 方法提出了一种使用三个正交平面的局部二值基线方法 (LBP-TOP), 一定程度上消除了光照变化的影响, 但旋转不变性使得算子对方向信息过于敏感。为了解决这一问题, Rivera 等[8] 学者提出的局部特征描述符 LDN 利用梯度信息使得算子对光照变化和噪声具有较强的鲁棒性。然而, 传统的表情识别算法无法有效处理由于不同姿势、遮挡等引起的非线性面部外观变化, 难以有效提高分类水平。

近年来, 深度学习凭借其优异的特征提取能力逐步应用于人脸表情识别领域。Kim 等[9] 学者对适用于大规模图像识别的 VGG-face 模型进行渐进式微调识别人脸表情, 但大多数人脸表情数据库样本较少导致该网络易出现过拟合问题。An 等[10] 学者提出了一种基于 MMN 线性激活函数的自适应模型参数初始化方法, 可有效克服过拟合问题, 但面对含有大量表情无关因素时算法鲁棒性较差。Xie 等[11] 学者提出了一种多路径变异抑制网络 (MPVS-NET), 但该网络速度较慢且不宜收敛。由于模糊的面部表情、低质的面部图像及注释者的主观性带来的不确定性, 对定性的大规模面部表情数据集进行标注是非常困难的。针对这一问题, Wang 等[12] 学者提出了一种能有效抑制不确定性的自修复网络 (SCN), 防止网络过度拟合不确定的人脸图像。一般来说, 深层网络更易提取到具有丰富语义信息的深层特征。但过深的网络容易出现梯度爆炸或梯度消失现象。针对这一问题, He 等[13] 学者提出了深度残差网络 (ResNet), 利用短路链接使得梯度正常回传, 较好地解决了网络退化问题。但训练参数量仍旧较大, 且残差网络并没有考虑不同尺度特征之间的相互关系对特征识别的影响, 导致大量有效特征丢失。

上述研究均使用完整特征图作为特征输入, 然而在实际分类任务中, 特征的作用程度是不同的。为了突出对特征识别有效的信息, 一些研究引入了注意力机制。Li 等[14] 学者提出了一种具有注意力机制的 CNN 网络结构可识别脸部遮挡区域, 但网络依赖于人脸关键点检测, 遮挡面积较大时, 关键点难以与人脸数据集生成映射。在此基础上, Liu 等[15] 学者提出了一种条件 CNN 增强型随机森林算法 (CoNERF), 从显著引导的人脸区域中提取深层特征, 抑制光照、遮挡和低分辨率带来的影响。然而上述方法仍保留了较多的冗余信息, 且均为完整网络结构, 不易迁移。Hu 等[16] 学者采用全新特征重标定方式提出一种通道注意力网络 (SE-Net), 显示建模特征通道之间的相互依赖关系, 进而提升有用特征并抑制用处不大的特征, 且能够直接集成到现有网络中, 计算代价小, 没有冗余信息。

针对上述问题, 本文提出一种深度多尺度融合注意力残差网络 (deep multi-scale fusion attention residual network, DMFA-ResNet), 主要改进包括以下 3 个方面:

1) 设计了一个由 7 个注意力残差学习单元构成的注意力残差模块, 注意力残差学习单元由 2 条包含卷积层的支路和 1 个短路链接构成, 将融合后的特征经过注意力机制, 对输入图像进行并行多卷积操作, 以获得图像多尺度特征, 突出局部重点区域, 有利于遮挡图像特征学习;

2) 提出多尺度融合模块, 网络整体将各个注意力残差模块的特征输出进行多尺度融合, 以获取更丰富的图像特征;

3) 在网络模型中增加过渡层以去除冗余信息, 在保证感受野的情况下简化网络复杂度。并使用全局平均池化+ Dropout 的设计减少参数运算, 使网络具有更好的抗过拟合性能。

1　DMFA-ResNet 算法

1.1　ResNet 网络结构

ResNet 网络通过引入残差模块, 在算法前向传播过程中使得卷积层之间形成跳跃连接, 实现对输入、输出的恒等映射, 并采用 1×1、3×3 的小卷积核, 在解决网络退化问题的同时进一步加深网络, ResNet-50 的基本残差学习单元如图 1 所示。

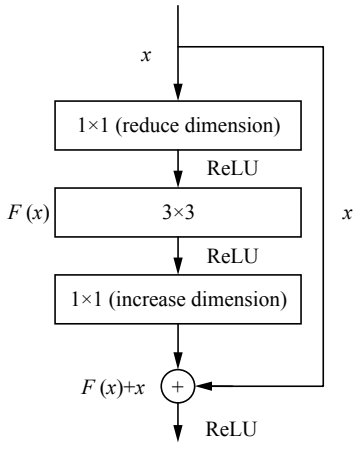


图 1　残差学习单元

Fig. 1　Residual learning unit

图 1 中, x表示输入, $F(x)$表示残差映射, 残差

单元的输出为

$$H(x)=F(x)+x \tag{1}$$

当残差$F(x)=0$, 残差学习单元的功能就是恒等映射; 则深层L的输出为

$$H(x_L)=x_l+\sum_{i=l}^{L-1}F(x_i) \tag{2}$$

其反向梯度为

$$\frac{\partial \mathrm{LOSS}}{\partial x_l}=\frac{\partial \mathrm{LOSS}}{\partial H(x_L)} \tag{3}$$

其中$\frac{\partial \mathrm{LOSS}}{\partial x_l}$为损失函数的梯度下降。

1.2　SE-Net 注意力模块

SE-Net 是 Hu 等[16]学者提出的一种通道注意力网络, 核心为特征压缩操作F_{sq}和特征激励操作F_{ex}。F_{sq}从通道维度将$[H,W,C]$的输入特征图压缩为$[1,1,C]$的输出特征图, 使得每个二维特征通道转换为一个具有全局感受野的实数。F_{ex}通过对每个通道生成权重, 显式建模特征通道间的相关性, 并逐通道加权到原始特征图上, 完成通道维度上的特征重标定, 加强关键特征, 抑制非显著特征, 从而提高网络的整体表征能力。

2　深度多尺度融合注意力残差网络

基于 ResNet-50 残差网络, 本文提出一种深度多尺度融合注意力残差网络 (DMFA-ResNet), 该网络由注意力残差模块 (attention residual module, ARM)、多尺度特征融合模块、过渡层、全局平均池化层、Dropout 和 Softmax 分类层构成, 网络结构如图 2 所示。

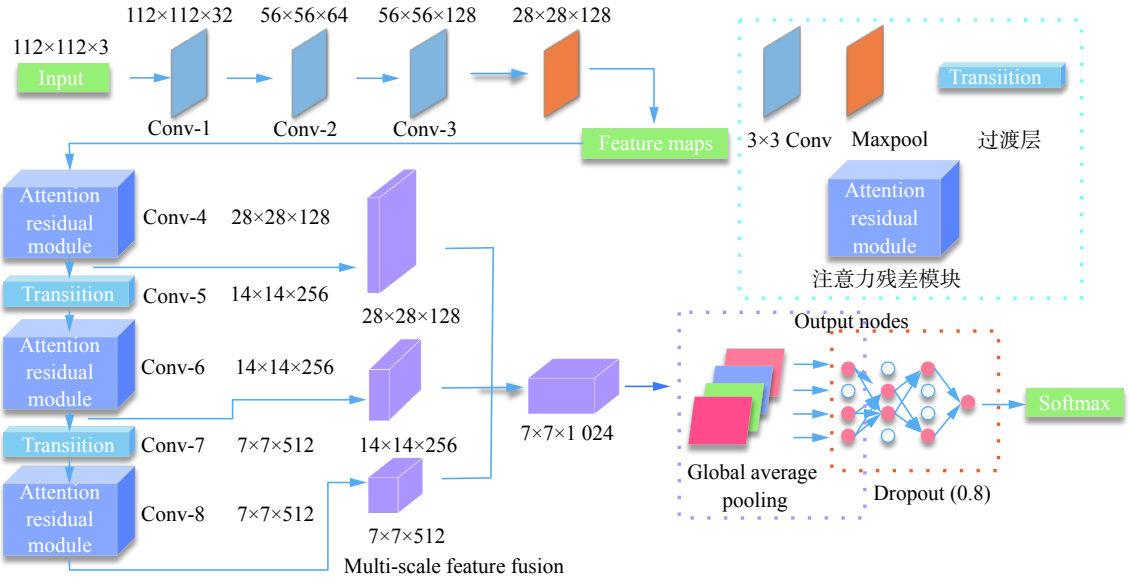


图 2　DMFA-ResNet 结构图

Fig. 2　DMFA-ResNet structure

深度神经网络的输入图片一般较大, 为避免后续计算量爆炸, 需要将输入图片进行下采样后再输入进卷积神经网络。原 ResNet 网络将输入图像经过一个 7×7 大卷积层和最大池化层后, 再输入进后续残差模块。7×7 大卷积层和最大池化层将输入图片的分辨率从 224×224 下采样至 56×56, 在减少计算量的同时最大程度保留了原始图像细节信息。DMFA-ResNet 使用 3 个 3×3 小卷积层代替原 7×7 大卷积层, 在保证与原网络层相同感受野的前提下, 进一步提升了网络深度, 使得网络能够提取到更深层次的语义信息。

2.1　注意力残差模块

注意力残差模块 (ARM) 由 7 个具有 3 条支路的注意力残差学习单元构成。注意力残差学习单元由两条残差学习支路、一条恒等映射支路和 SE-Net 注意力模块构成。为了使输入经过 3×3 卷积层后的特征图维数相同, 通过残差学习支路的第一个 1×1 卷积层对输入进行降维。通过对输入图像进行并行的多卷积操作, 使得网络能够提取到不同深度的多尺度表情图像特征。再将这两条残差学习支路所提取到的特征采用 Concat 方法进行融合, 即将两个需要融合的特征图的通道进行拼接, 将两条残差学习支路输出的特征图融合后的特征通过 1×1 卷积进行升维, 确保输入、输出的维数相等。最后利用注意力机制突出重点局部区域, 获得图像更准确的特征以提高识别准确率, 有利于遮挡图像的特征学习。注意力残差模块和注意力残差单元的结构图分别如图 3、4 所示。

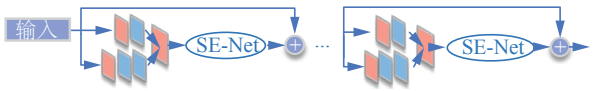


图 3　注意力残差模块

Fig. 3　Attention residual module

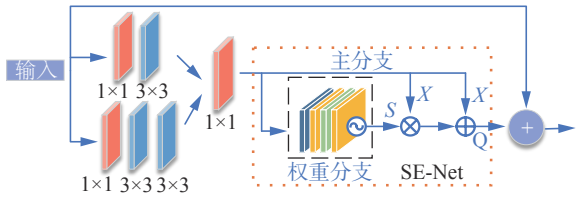


图 4　注意力残差单元

Fig. 4　Attention residual unit

2.2　过渡层

随着网络深度不断加深，运算参数量持续增多，容易使得网络过度学习输入与输出之间的映射关系，将大量干扰信息错认为重点特征。

在注意力残差模块之间引入由一个 3×3 卷积层和最大池化层组成的过渡层以去除冗余信息。3×3 卷积层能够在不改变特征图大小的情况下增大维数，提升网络线性转换能力。最大池化层能够对输入图像进行下采样以减小参数矩阵的尺寸以及卷积层参数误差造成估计均值的偏移，其结构如图 5 所示。

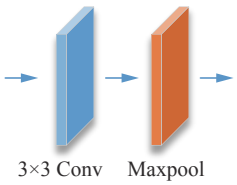


图 5　过渡层结构

Fig. 5　Transition layer structure

2.3　多尺度特征融合模块

经过各个注意力残差模块后，人脸表情图像的多尺度特征具有不同特点：浅层特征图尺寸较大，通道数较少，具有丰富的细节信息；深层特征图尺寸较小，通道数较多，包含丰富的抽象语义信息。因此本文设计了一个多尺度特征融合模块将 3 个注意力残差模块产生的多尺度特征图进行融合。首先将前两个注意力残差模块的输出特征经过最大池化操作下采样至 7×7×128 和 7×7×256；然后通过 Concat 通道融合方法将下采样过后的输出特征图和最后一个注意力残差模块的输出特征图进行融合；再将融合后的特征图使用 1×1 卷积核进行升维，最终得到具有丰富特征信息的 7×7×1 024 输出特征图。

2.4　全局平均池化+随机失活

通常情况下，神经网络都会添加全连接层减少特征位置对分类带来的影响。但人脸基本位于图像中央且占据绝大部分像素，位置信息并不重要。因此采用全局平均池化层代替全连接层加强特征图与类别的一致性，直接对空间信息进行求和实现降维，极大地减少了网络参数。Dropout 原理又名随机失活原理，是指在网络训练过程中随意抛弃某些神经元，破坏特征信息之间密切的交互作用，使得网络不会过于依赖某些局部特征，增强模型泛化性。

本文使用全局平均池化+随机失活设计，简化网络复杂度，减少运算量，避免过拟合现象，进而提高网络泛化性。

3　实验结果与分析

3.1　实验环境与评价指标

实验使用的深度学习框架为 Tensorflow，计算机操作系统为 Windows10，显卡型号为 NVIDIA Quadro P4000，显存为 8BG。

实验使用错误率 (error rate)、准确率 (accuracy rate)、混淆矩阵和 F1-score 作为评价指标。

错误率是指预测值与真实值不相同的样本数占总样本数的比例，准确率是指预测值与真实值相同的样本数占总样本数的比例。将真阳性 (TP)、假阳性 (FP)、真阴性 (TN) 和假阴性 (FN)4 个指标一起呈现在表格中称为混淆矩阵。F1-score 为精准率和召回率的调和平均数，取值范围从 0~1，其计算公式为

$$\text{F1-score} = 2 \cdot \frac{\text{Pre} \cdot \text{Rec}}{\text{Pre} + \text{Rec}} \tag{4}$$

3.2　实验数据集及预处理

3.2.1　实验数据集

实验采取 3 个人脸表情数据库验证算法有效性，分别为 CK+、JAFFE 和 Oulu-CASIA。

CK+数据集共有 123 名实验者，实验共使用 981 张标记图片用于本文实验。JAFFE 数据集共包含 213 个图像、7 类表情，平均每人每种表情有 4 张左右。Oulu-CASIA 数据集由 80 个人的 6 类基本表情构成，实验选取可见光成像系统下的 Strong 强光图像集，在每个序列中选取最后 5 个峰值帧，形成共 2400 幅图像。

3.2.2　数据预处理

由于人脸表情识别数据库样本较少，本文使用裁剪、旋转以及遮挡方法对数据集进行扩充，具体步骤如下：

1）首先对 CK+和 JAFFE 数据集进行裁剪处理，去除多余的背景，将背景对模型的影响降到最低。

2)分别将 JAFFE 数据集图像以顺时针、逆时针旋转 5°后的图像扩充数据集, 扩充完毕共 852 张标记图片用于实验, 其中训练集 680 张, 验证集 172 张, 如表 1 所示。

表 1　JAFFE 扩充数据集样本分布

Table 1　Sample distribution of expanded JAFFE

类别	数量/张	占比
愤怒	120	0.14
厌恶	116	0.13
恐惧	128	0.15
快乐	124	0.15
中性	120	0.14
悲伤	124	0.15
惊讶	120	0.14

3)通过在眼睛、嘴巴位置添加黑色框来模拟现实中存在的遮挡情况, 如由墨镜、口罩等引起。

3.3　实验结果与分析

3.3.1　网络性能实验分析

1) 训练样本对性能影响

为探讨训练样本对网络性能的影响, 设置训练样本数目对比实验。在其余参数量一致的情况下, 在 JAFFE 扩充数据集 (852 张) 上进行训练样本分别为 341、511、680 的对比实验, 实验结果如表 2 所示。

表 2　训练样本对性能影响

Table 2　Effect of training sample number on performance

训练样本/个	验证样本/个	测试集识别率A/%
341	511	94.2
511	341	94.7
680	172	96.3

由表 2 可知, 随着训练样本不断增多, 网络性能逐步增强, 当训练样本为 680 个时, 网络识别率达到最高 96.3%, 因此在网络训练过程中, 应尽可能增大训练样本数目, 保证网络能够学习到足够信息。

2) 网络结构

为验证各个模块的有效性, 设置包含针对不同模块的对比网络进行消融实验。在参数量基本一致的情况下, 以改进的基础残差模块网络 DFR (deep fusion residual network) 为对比基准, 将多尺度特征融合模块添加进网络结构中构成深度多尺度融合残差网络 DMFR (deep multi-scale fusion residual network), 将注意力机制添加进网络结构中构成深度融合注意力残差网络 DFAR (deep fusion attention residual network), 在 Oulu-CASIA 数据集上进行表情识别消融实验, 实验结果如表 3 所示。

表 3　表情识别消融实验

Table 3　Ablation experiment of facial expression recognition

方法	残差单元	多尺度融合	注意力机制	A/%
DFR	√	×	×	91.16
DMFR	√	√	×	91.69
DFAR	√	×	√	91.53
DMFA-ResNet	√	√	√	92.57

由表 3 可知, 改进的基础残差模块网络 DFR 在 Oulu-CASIA 数据集上的识别率为 91.16%。当分别增加多尺度特征模块和注意力机制模块后, Oulu-CASIA 的识别率分别提升到 91.69% 和 91.53%, 表明多尺度特征融合模块对网络的贡献大于注意力机制模块。

为探讨注意残差单元数目对网络性能的影响, 设置注意残差单元数目对比实验。在其余参数量基本一致的情况下, 将注意残差单元数目分别设置为 4、5、6、7、8、9, 并在 JAFFE 数据集上进行实验, 实验结果由图 6 所示。

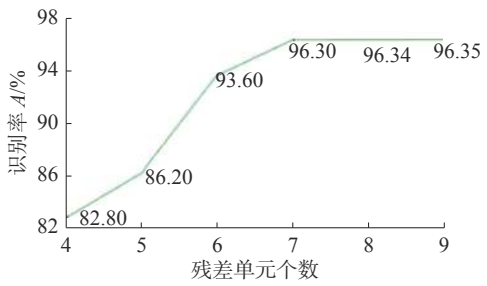


图 6　注意残差单元个数对性能的影响

Fig. 6　Effect of the number of attention residual elements on peraformance

由图 6 可知, 当注意残差单元个数小于 7 时, 算法识别率随残差单元个数的增加增幅明显。当注意残差单元个数为 9 时, 算法识别率达到最高 96.35%。但注意残差单元个数大于 7 时, 识别率增幅缓慢, 考虑到网络复杂度对计算量及网络运行速度带来的影响, 最终选择将 7 个注意残差单元作为一个注意残差模块。

3.3.2　无遮挡表情实验

表 4 是不同方法在 Oulu-CASIA 数据集上的测试结果。结果表明, DFR 算法在 Oulu-CASIA 数据集上的识别率能够达到 91.16%。DMFA-ResNet 的识别率达到 92.57%, 比 LCE 的识别率高出 9.31%, 比 IDFERM 的识别率高出 4.32%。

表 4　不同方法在 Oulu-CASIA 数据集上的测试结果

Table 4　Test results of different methods on Oulu-CASIA data sets

方法	*A*/%
参数重要性正则[17]	88.19
DFLF[18]	87.85
LCE[19]	83.26
IDFERM[20]	88.25
DeRL[21]	88.0
DFR	91.16
DMFA-ResNet	**92.57**

表 5 是不同方法在 CK+和 JAFFE 数据集上的测试结果。结果表明，DFR 算法在 CK+和 JAFFE 数据集上分别能够达到 99.68% 和 96.25% 的识别率。比文献 [22] 在两个数据集中的识别率分别高出 6.22% 和 1.5%，比文献 [23] 在两个数据集中的识别率分别高出 2.92% 和 9.51%。

表 5　不同方法在 CK+和 JAFFE 数据集上的测试结果

Table 5　Test results of different methods on CK+ and JAFFE data sets　%

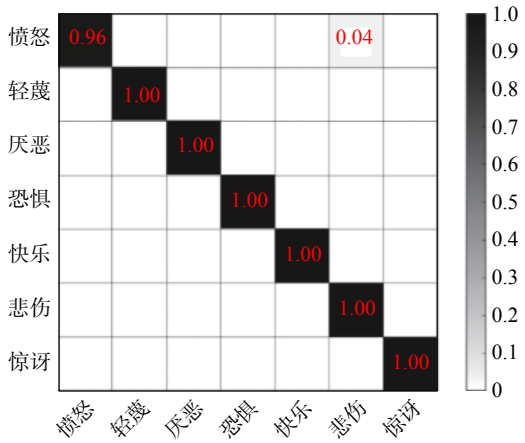
方法	CK+数据集	JAFFE数据集
DCMA-CNNs[22]	93.46	94.75
CNN pre-processing[23]	96.76	86.74
WMDCNN[24]	98.50	92.3
WMDNN[25]	97.02	92.21
HDNNS[26]	96.46	91.27
DFR	99.68	96.25
DMFA-ResNet	99.70	**96.30**

图 7 分别为 DFR 算法在 CK+和 JAFFE 数据集的混淆矩阵，其中 DFR 能够在 CK+数据集上对轻蔑、厌恶、恐惧、快乐、悲伤和惊讶这六种表情达到 100% 识别率；在 JAFFE 数据集上对恐惧及中性表情能够达到 100% 识别率，但惊喜表情容易被误判为中性表情，因此识别精度最低。

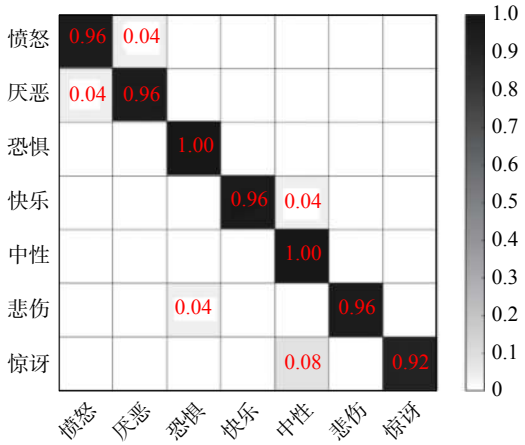
DFR 算法对比其他先进算法在识别率上有很大提升，充分验证了改进的残差模块和过渡层能够提取更加精确的人脸表情特征。DMFA-ResNet 算法在 CK+和 JAFFE 数据集上的识别率分别为 99.7% 和 96.3%，比 DFR 算法在两个数据集中分别提高 0.02% 和 0.05%，证明了引入注意力机制模块和多尺度特征融合模块对提升人脸表情识别率是有利的。

3.3.3　遮挡表情实验

实际生活中，人脸表情图像采集会伴有遮挡情况，一般由墨镜、口罩等引起。若局部区域被遮挡，卷积神经网络就难以抓住重点区域进行特征提取，针对这种情况，本章将在遮挡的扩充数据集上进行实验。表 6 和表 7 分别为各种算法在 CK+和 JAFFE 数据集上的遮挡。



(a) DFR 在 CK+ 上的混淆矩阵



(b) DFR 在 JAFFE 上的混淆矩阵

图 7　DFR 在 CK+和 JAFFE 数据集上的混淆矩阵

Fig. 7　Confusion matrix of DFR on CK+ and JAFFE

表 6　CK+上遮挡表情识别

Table 6　occlusion facial expression recognition on CK+　%

方法	眼睛遮挡	嘴巴遮挡
PCA+SVM	89.04	83.84
CNN	91.69	89.17
DCGAN+CNN	87.83	85.43
end-to-end GAN	92.69	90.57
微调VGGface	89.90	87.34
MU-VGGNet	94.06	91.63
DFR	93.81	92.78
DMFA-ResNet	**95.31**	**93.81**

表 7　JAFFE 上遮挡表情识别

Table 7　Occlusion facial expression recognition on JAFFE　%

方法	眼睛遮挡	嘴巴遮挡
Gabor	85.66	86.97
Gabor+gray-level matrix	88.88	90.90
WLDH	87.23	89.47
DFR	89.47	91.81
DMFA-ResNet	**91.22**	**93.56**

表 8 和表 9 分别为 DMFA-ResNet 算法在 CK+和 JAFFE 数据集上的 F1-score值。图 8 和图 9 分别为 DMFA-ResNet 算法在 CK+和 JAFFE 数据集上的遮挡混淆矩阵。

表 8　CK+上遮挡表情 F1-score 值

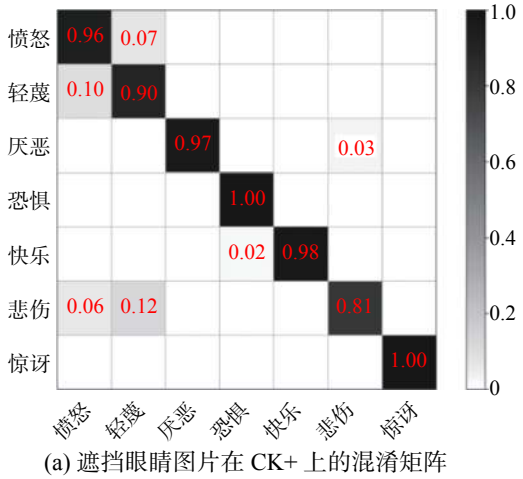
Table 8　F1-score of occlusion facial expression on CK +

表情	眼睛遮挡	嘴巴遮挡
愤怒	0.89	0.96
轻蔑	0.86	0.76
厌恶	0.98	0.90
恐惧	0.99	0.75
快乐	0.99	**1**
悲伤	0.89	0.90
惊讶	**1**	0.90

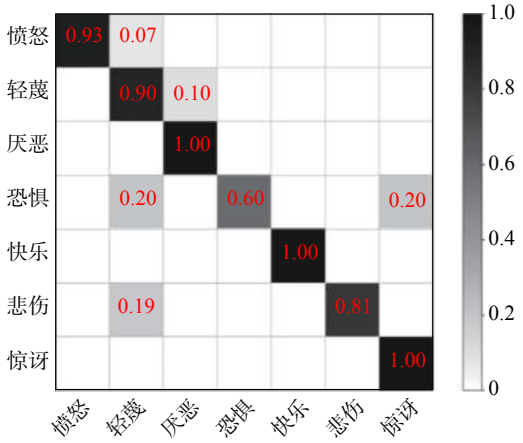
表 9　JAFFE 上遮挡表情 F1-score 值

Table 9　F1-score of occlusion facial expression on JAFFE

表情	眼睛遮挡	嘴巴遮挡
愤怒	**0.96**	0.96
厌恶	0.90	0.92
恐惧	0.88	0.89
快乐	0.94	0.94
中性	0.92	0.96
悲伤	0.82	0.88
惊讶	0.94	**0.98**



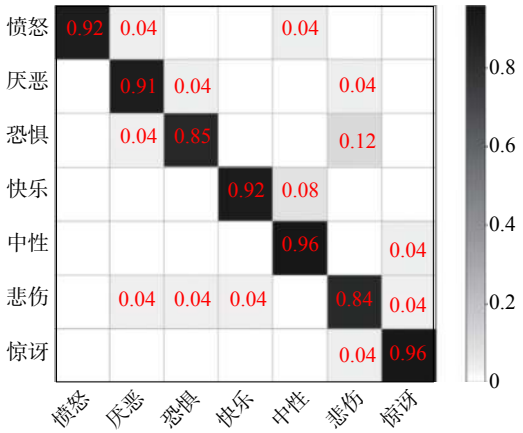
(a) 遮挡眼睛图片在 CK+ 上的混淆矩阵



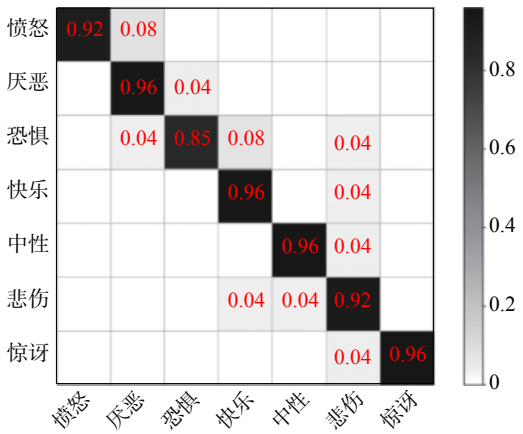
(b) 遮挡嘴巴图片在 CK+ 上的混淆矩阵

图 8　在 CK+数据集上的遮挡混淆矩阵

Fig. 8　Occlusion confusion matrix on the CK+



(a) 遮挡眼睛图片在 JAFFE 上的混淆矩阵



(b) 遮挡嘴巴图片在 JAFFE 上的混淆矩阵

图 9　在 JAFFE 数据集上的遮挡混淆矩阵

Fig. 9　Occlusion confusion matrix on the JAFFE

由表 6、表 7 可知，对于遮挡图像，DMFA-ResNet 比 DFR 算法在 CK+和 JAFFE 数据集上的识别精度分别提升 2.5% 和 1.5%，且 DMFA-ResNet 对遮挡表情的识别在两个数据集上均取得最高识别精度。

由表 8 和图 8 可知，遮挡眼睛后，DMFA-Res-

Net 算法在 CK+数据集上能够对害怕和惊讶两种表情达到 100% 识别率；遮挡嘴巴后，能够对困惑、快乐和惊讶 3 种表情达到 100% 识别率。而轻蔑和恐惧表情的 F1-score 分别只达到 0.76 和 0.75，说明这两种表情的有效特征大部分在于嘴巴部分。

由图 9 和表 9 可知，遮挡眼睛情况下的悲伤表情 F1-score 仅达到 0.82，说明悲伤表情的有效特征大部分在于眼睛部分，虽然该值达到最低，但 DMFA-ResNet 在 JAFFE 数据集上也取得相当不错的效果。由于该数据集样本间的差异较小，导致算法仍出现较多误判情况，无法完全精准识别某一类表情。以上实验结果证明了 DMFA-ResNet 在应对遮挡图像问题上的优越性，更适用于人脸表情识别任务。

4　结束语

本文提出一种多尺度融合注意力残差网络(DMFA-ResNet)。该网络主要提出一种新的注意力残差模块，提高了网络对局部重点部位特征的提取，有利于学习到非遮挡部位的信息；提出多尺度融合模块，将各残差模块的输出进行融合以提取更加丰富的人脸表情特征；为了减少参与网络运算的参数量，在各个残差模块之间添加过渡层，主要进行下采样操作并使用全局平均池化+Dropout 设计防止网络过拟合。在 CK+、JAFFE 和 Oulu-CASIA 数据集上进行实验均取得了不错的效果，注意力残差模块对局部区域的特征能够进行有效提取，实验验证本文算法具有优越性。但所提算法为针对静态图像的表情识别算法，不适用于动态连续的视频识别，在接下来的工作中，可以重点研究基于视频的动态表情识别技术。

参考文献：

[1] BEN Xianye, REN Yi, ZHANG Junping, et al. Video-based Facial micro-expression analysis: a survey of datasets, features and algorithms[EB/OL].(2021-03-19)[2021-05-01].https://arxiv.org/abs/2201.12728v1.

[2] CHEN Boyu, GUAN Wenlong, LI Peixia, et al. Residual multi-task learning for facial landmark localization and expression recognition[EB/OL].(2021-07-01)[2021-07-05].https://www.sciencedirect.com/science/article/pii/S0031320321000807.

[3] LI Shan, DENG Weihong. Deep facial expression recognition: a survey[EB/OL].(2020-03-17)[2021-05-01]. https://ieeexplore.ieee.org/document/9039580.

[4] ZHAO Guoying, PIETIKAINEN M. Dynamic texture recognition using local binary patterns with an application to facial expressions[J]. IEEE transactions on pattern analysis and machine intelligence, 2007, 29(6): 915–928.

[5] WHITEHILL J, OMLIN C W. Haar features for FACS AU recognition[C]//Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition. Southampton, UK, 2006: 5−101.

[6] BARTLETT M S, LITTLEWORT G, FRANK M, et al. Recognizing facial expression: machine learning and application to spontaneous behavior[C]//Proceedings of 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Diego, USA, 2005: 568−573.

[7] LI Xiaobai, PFISTER T, HUANG Xiaohua, et al. A spontaneous micro-expression database: inducement, collection and baseline[C]//2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG). Shanghai, China, 2013: 1−6.

[8] RIVERA A R, CASTILLO J R, CHA E O O. Local directional number pattern for face analysis: face and expression recognition[J]. IEEE transactions on image processing, 2013, 22(5): 1740–1752.

[9] KIM T H, YU C, LEE S W. Facial expression recognition using feature additive pooling and progressive fine-tuning of CNN[J]. Electronics letters, 2018, 54(23): 1326–1328.

[10] AN Fengping, LIU Zhiwen. Facial expression recognition algorithm based on parameter adaptive initialization of CNN and LSTM[J]. The visual computer, 2020, 36(3): 483–498.

[11] XIE Siyue, HU Haifeng, WU Yongbo. Deep multi-path convolutional neural network joint with salient region attention for facial expression recognition[J]. Pattern recognition, 2019, 92: 177–191.

[12] WANG Kai, PENG Xiaojiang, YANG Jianfei, et al. Suppressing uncertainties for large-scale facial expression recognition[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 6897−6906.

[13] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Deep residual learning for image recognition[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 770−778.

[14] LI Yong, ZENG Jiabei, SHAN Shiguang, et al. Occlusion aware facial expression recognition using CNN with attention mechanism[J]. IEEE transactions on image processing, 2019, 28(5): 2439–2450.

[15] LIU Yuanyuan, YUAN Xiaohui, GONG Xi, et al. Conditional convolution neural network enhanced random forest for facial expression recognition[J]. Pattern recognition, 2018, 84: 251–261.

[16] HU Jie, SHEN Li, SUN Gang. Squeeze-and-excitation networks[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 7132−7141.

[17] 江静, 邓伟洪. 持续学习改进的人脸表情识别 [J]. 中国图象图形学报, 2020, 25(11): 2361–2369.
JIANG Jing, DENG Weihong. Facial expression recognition improved by continual learning[J]. Journal of image and graphics, 2020, 25(11): 2361–2369.

[18] 王善敏, 帅惠, 刘青山. 关键点深度特征驱动人脸表情识别 [J]. 中国图象图形学报, 2020, 25(4): 813–823.
WANG Shanmin, SHUAI Hui, LIU Qingshan. Facial expression recognition based on deep facial landmark features[J]. Journal of image and graphics, 2020, 25(4): 813–823.

[19] 张文萍, 贾凯, 王宏玉, 等. 改进的 Island 损失函数在人脸表情识别上的应用 [J]. 计算机辅助设计与图形学学报, 2020, 32(12): 1910–1917.
ZHANG Wenping, JIA Kai, WANG Hongyu, et al. Application of improved Island loss in facial expression recognition[J]. Journal of computer-aided design & computer graphics, 2020, 32(12): 1910–1917.

[20] LIU Xiaofeng, KUMAR B V K V, JIA Ping, et al. Hard negative generation for identity-disentangled facial expression recognition[J]. Pattern recognition, 2019, 88: 1–12.

[21] YANG Huiyuan, CIFTCI U, YIN Lijun. Facial expression recognition by de-expression residue learning[C]// Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018.

[22] XIE Siyue, HU Haifeng. Facial expression recognition using hierarchical features with deep comprehensive multipatches aggregation convolutional neural networks [J]. IEEE transactions on multimedia, 2019, 21(1): 211–220.

[23] LOPES A T, DE AGUIAR E, DE SOUZA A F, et al. Facial expression recognition with convolutional neural networks: coping with few data and the training sample order[J]. Pattern recognition, 2017, 61: 610–628.

[24] ZHANG Hepeng, HUANG Bin, TIAN Guohui. Facial expression recognition based on deep convolution long short-term memory networks of double-channel weighted mixture[J]. Pattern recognition letters, 2020, 131: 128–134.

[25] YANG Biao, CAO Jinmeng, NI Rongrong, et al. Facial expression recognition using weighted mixture deep neural network based on double-channel facial images [J]. IEEE access, 2017, 6: 4630–4640.

[26] KIM J H, KIM B G, ROY P P, et al. Efficient facial expression recognition algorithm based on hierarchical deep neural network structure[J]. IEEE access, 2019, 7: 41273–41285.

作者简介：

高涛，教授，博士，主要研究方向为数字图像处理、模式识别。获得国家专利 9 项。发表学术论文 16 篇。

杨朝晨，硕士研究生，主要研究方向为数字图像处理、深度学习。

陈婷,副教授，博士，主要研究方向为图形图像处理、计算机视觉。