文章编号：1007-2780(2024)04-0522-10

NCA-MobileNet：一种轻量化人脸表情识别方法

左义海[1]，白武尚[2]，何秋生[2*]
（1. 太原工业学院 工程训练中心，山西 太原 030008；
2. 太原科技大学 电子信息工程学院，山西 太原 030024）

摘要：针对目前人脸面部表情识别方法存在参数量多、计算资源消耗大和识别精度低的问题，提出了一种基于条件协调注意力机制的轻量化人脸面部表情识别方法。首先，对MobileNet V3网络层数进行缩减，同时将倒残差结构中间通道数和输出通道数增大至原来的1.5～3.2倍，使用Mish代替Hardswish激活函数，实现特征提取后的非线性化。其次，引入改进的协调注意力机制，在张量信息嵌入中沿水平和竖直方向依次通过最大池化和平均池化进行编码，并通过张量信息集成产生具有全局感受野和精确位置信息特征，提取面部表情在空间和通道位置上的详细信息。最后，在公开数据集FERPlus和RAF-DB上进行实验，结果表明所提方法参数量降低15.91%，准确率分别为88.84%和85.90%，比改进前模型准确率分别提升0.83%和1.39%。该方法具有良好的识别性能，验证了所提方法的有效性。

关　键　词：表情识别；轻量化；注意力机制；特征提取
中图分类号：TP391.4　　**文献标识码**：A　　**doi**：10.37188/CJLCD.2023-0153

NCA-MobileNet: a lightweight facial expression recognition method

ZUO Yihai[1], BAI Wushang[2], HE Qiusheng[2*]

(1. *Engineering Training Center, Taiyuan Institute of Technology, Taiyuan* 030008, *China*;
2. *School of Electronic Information Engineering, Taiyuan University of Science and Technology, Taiyuan* 030024, *China*)

Abstract: At present, facial expression recognition methods have the problems of large number of parameters, large consumption of computing resources and low recognition accuracy. Aiming at the above problems, a lightweight human facial expression recognition method based on conditional coordinated attention mechanism is studied. First, the number of layers of MobileNet V3 network is reduced, while the numbers of intermediate channels and output channels of the inverse residual structure are increased to 1.5～3.2 times of the original number. Mish is used instead of Hardswish activation function to realize the nonlinearization after feature extraction. Secondly, an improved coordinated attention mechanism is introduced to encode the tensor information embedding along horizontal and vertical directions sequentially

收稿日期：2023-04-25；**修订日期**：2023-05-20.
基金项目：山西省自然科学基金(No. 20210302123222)；山西省教学改革项目(No. J20221103)
Supported by Natural Science Foundation of Shanxi Province(No. 20210302123222); Teaching Reform Project of Shanxi Province(No. J20221103)
*通信联系人，E-mail: heqs2008@126.com

by maximum pooling and average pooling. And tensor information integration is used to generate features with global sensory field and precise location information to extract detailed information of facial expressions in space and channel location. Finally, experiments are conducted on the publicly available datasets FERPlus and RAF-DB, and the results show that the proposed method reduces the number of parameters by 15.91%, and the accuracy rates are 88.84% and 85.90%, respectively, which are 0.83% and 1.39% higher than the accuracy rates of the model before improvement. The method has good recognition performance and validate the effectiveness of the proposed method.

Key words: facial expression recognition; lightweight; attention mechanism; feature extraction

1 引　　言

面部表情是人类情绪的最直接表现形式，而赋予机器感知人类情感的能力是实现人机交互的重要目标之一。人脸表情识别（Facial Expression Recogniton，FER）在情感计算、人机交互、驾驶员疲劳检测、教学效果评价等众多领域有着广泛的应用[1]。1971年，著名的心理学家Ekman[2]确定了6种基本的表情类别。通过这些表情，不同种族之间能够互相辨认，即使是远离现代文明的部落文明与普通的哺乳动物也具有类似的表情。此外，Ekman的研究表明，在非人类的哺乳动物中也观察到类似的表情，最终Ekman和Friesen确定了人类的6种基本表情，即快乐（Happy）、悲伤（Sad）、愤怒（Anger）、厌恶（Disgust）、惊奇（Surprise）和恐惧（Fear）。在社会不断发展的过程中，中性（Neutral）表情也被提出，形成了当前主流的7种表情状态。学术界通过对这7类表情进行分类研究，开启了对计算机自动表情识别任务的探索。

随着卷积神经网络在图像识别领域的巨大成功，神经网络逐渐被用于人脸表情识别任务。Li[3]等人提出了一种具有注意力机制（ACNN）的卷积神经网络（CNN），它可以感知人脸的遮挡部分，并专注于非遮挡部分最具区分特征的区域，实现对人脸不同角度和遮挡物情况下的表情识别。人脸表情数据集的质量参差不齐，这种不确定性给深度学习时代的大规模面部表情识别带来重大挑战。为了解决此问题，Kai[4]等人提出了一种简单有效的自愈合网络（SCN），它通过自注意力重要性模块学习每张图片的重要性后对其损失进行加权，不确定的面部表情图像权重较小，反之较大。最大预测概率比给定标签概率高出阈值，则修改其对应的标签值。此模型可以有效地抑制表情的不确定性，阻止深度网络对不确定的人脸图像进行过拟合，有效提高了人脸表情识别的准确率。随着注意力机制的发展与应用，2020年，Wang[5]等人分别在特征层和图像层使用局部块注意力机制，以提高特征学习的能力。针对微表情持续时间短、数据集有限等造成表情特征提取困难的问题，李召峰[6]等人提出一种基于图像预处理技术和双分支网络的识别方法，提升了表情特征提取能力。

在2012年的ImageNet竞赛中，AlexNet网络获得冠军，随后研究人员设计了越来越多的深度神经网络模型，而且层数越来越深，如经典的VGGNet、GoogleNet、ResNet50等网络。与传统算法相比这些算法非常优秀。但深度网络模型对硬件设备的要求也相对较高，带来了巨大的存储压力和计算负担。传统的深度神经网络内存需求较大，计算量也大，在移动设备和嵌入式设备上运行效果较差。2017年，Google公司提出了一种轻量级神经网络MobileNet V1[7]，第一次引入深度可分离卷积来减小模型参数量，使得传统神经网络有了一种轻量化的方法。2018年和2019年，Google先后推出MobileNet V2[8]和MobileNet V3[9]网络，这些网络在ImageNet数据集上能够实现较高的精度，而且模型参数更少，计算速度更快。目前人脸表情识别算法是在经典神经网络基础上进行的变体，存在网络结构复杂、参数量大和识别精度低的特点，不适用于嵌入式移动设备等算力较小的平台。

针对以上问题，本文研究了一种基于改进协调注意力机制的人脸面部表情识别方法（New Coordinate Attention MobileNet，NCA-MobileNet），在轻量化网络MobileNet V3的基础上进行改进。

首先对其卷积层数和通道数进行适当调整，然后引入非线性激活函数 Mish[10]，其对负值拥有更好的梯度流，允许复杂的信息输入神经网络模型。其次引入改进的协调注意力机制，增强特定区域的特征提取能力。本文方法在公开数据集 FER-Plus[11]和 RAF-DB[12]上进行了实验，并对最新的人脸表情识别模型进行对比分析。实验结果表明，所提方法的性能有明显的提升。

2 基于改进协调注意力机制的模型

本文提出一个基于改进协调注意力机制的人脸面部表情识别模型(NCA-MobileNet)，解决了人脸面部表情识别模型复杂和准确率低的问题。首先，对 MobileNet V3 主干网络进行改进，同时引入非线性激活函数 Mish，其对负值拥有更好的梯度流，允许复杂的信息输入神经网络模型。其次，设计改进的协调注意力机制模块(New Coordinate Attention Bneck，NCA Bneck)增强对人脸表情特征提取的能力，提升表情识别准确率。模型结构参数如表 1 所示，整体结构如图 1 所示。表 1 中，Input 表示输入图像大小，Operator 表示构建网络基本计算单元的算子，Exp size 表示算子中点卷积升维后的通道数，Out 表示输出通道数，NL 表示非线性激活函数，s 表示步幅。

表 1 NCA-MobileNet 结构参数表

Tab. 1 NCA-MobileNet structure parameter table

Input	Operator	Exp size	Out	NL	s
$224^2\times3$	Conv2d	—	32	MH	2
$112^2\times32$	NCA Bneck, 3×3	32	32	RE	1
$112^2\times32$	Bneck, 3×3	128	64	RE	2
$56^2\times64$	Bneck, 3×3	128	64	RE	1
$56^2\times64$	NCA Bneck, 5×5	128	64	RE	2
$28^2\times64$	Bneck, 5×5	256	128	RE	1
$28^2\times128$	Bneck, 5×5	256	128	RE	1
$28^2\times128$	NCA Bneck, 3×3	256	128	MH	2
$14^2\times128$	Bneck, 3×3	512	256	MH	1
$14^2\times256$	Bneck, 3×3	512	256	MH	1
$14^2\times256$	NCA Bneck, 3×3	512	256	MH	1
$14^2\times256$	Bneck, 3×3	1 024	512	MH	1
$14^2\times256$	Bneck, 3×3	1 024	512	MH	1
$14^2\times512$	NCA Bneck, 5×5	1 024	512	MH	2
$7^2\times512$	Avg Pool, 7×7	—	512	—	1
$1^2\times512$	FC	—	512	—	—
$1^2\times512$	Softmax	—	7/8	—	—

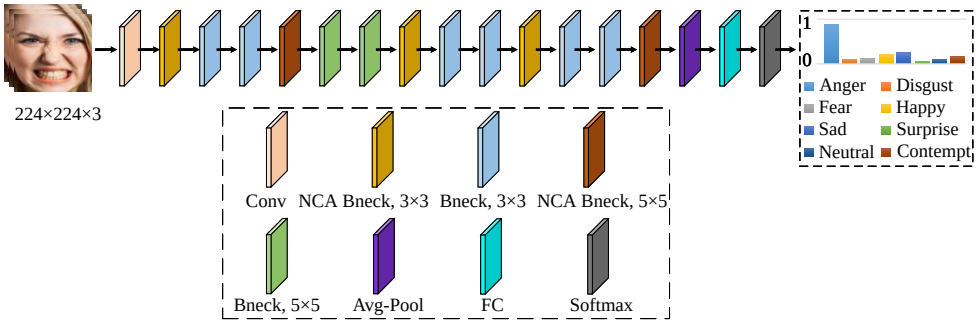

图 1 NCA-MobileNet 基本框架

Fig. 1 NCA-MobileNet basic framework

输入图像是 224 像素×224 像素的 RGB 图像。首先通过普通 3×3 卷积初步提取特征；然后依次通过一个改进的协调注意力机制模块(NCA Bneck)和两个普通卷积模块(Bneck)对数据进行表情特征的详细提取，提取出最具面部表情特性的表情特征；最后通过 FC 层将特征处理为单维度特征，通过 Softmax 输出不同表情的概率值，概率值最大的表情即为输入表情的类别。为了更好的识别特征，使网络快速收敛，在 Bneck 和 NCA Bneck 模块中进行点卷积升维后和深度卷积之后都加入了批标准化和非线性激活函数 Mish(或者 Relu)进行处理。

2.1 改进的 MobileNet V3 主干网络

MobileNet V3 提供了 L 和 S 版本，分别适用于不同硬件资源的情况。基本卷积单元到残差结构(Bneck)采用深度可分离卷积发挥轻量级作

用，首先用逐点卷积对输入特征进行升维操作，然后采用深度卷积提取特征，最后把张量信息映射到低维空间。其倒残差结构见图2。当整体输入与输出相等时，采用残差结构进行连接。

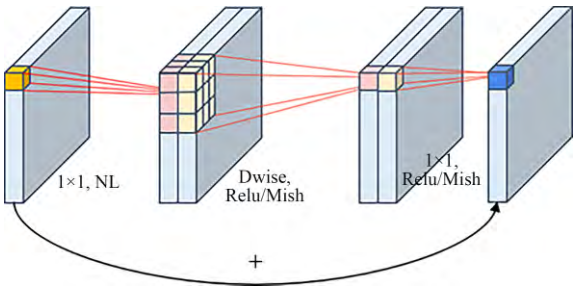

图2　倒残差结构

Fig. 2　Inverted residual structure

假设标准卷积输入张量为 $h_i \times w_i \times d_i$，卷积核大小为 $k \times k \times d_i \times d_j$，输出张量为 $h_i \times w_i \times d_j$，则标准卷积的参数量 Z_C 如式(1)所示。采用深度可分离卷积的参数量 Z_{DSC} 如式(2)所示，与标准卷积相比，深度可分离卷积能减小大约 $(k^2-1)d_j-k^2$ 的参数量。

$$Z_C = h_i \cdot w_i \cdot d_i \cdot d_j \cdot k \cdot k\,, \tag{1}$$

$$Z_{DSC} = h_i \cdot w_i \cdot d_i (k^2 + d_j)\,. \tag{2}$$

MobileNet V3在同级别模型中表现优秀，其网络结构是在ImageNet数据集上应用神经网络架构搜索(NAS)技术获得的最佳模型。而在人脸表情识别任务中，由于人脸的相似性与不同表情之间的区分性相比ImageNet数据集差距巨大，因此原来的网络结构不适用于本文中的人脸表情识别任务。

本文中人脸表情识别任务分为7～8类，输入图片大小统一固定为224×224×3。经过研究与多次实验，最终确定改进后主干网络模型结构参数如表2所示。其中中间通道数表示Benck模块中的点卷积进行升维后的维度，输出通道数表示Bneck输出后的通道数。MobileNet V3的结构中有15个Bneck，本文的网络结构有13个，减少2个Bneck，使网络整体参数量降低。对Bneck中深度卷积后的通道数进行了扩充，点卷积后的升维操作统一使通道数增大为对应Bneck输出通道数的2倍。新网络相比原来的网络整体参数量降低了16.15%。

MobileNet V3中激活函数使用的是Hard-swish函数，它的计算成本较大。本文采用Mish激活函数代替Hardswish激活函数。Hardswish和Mish函数式如式(3)和式(4)所示。图3为对应激活函数图，分析可知，Mish激活函数图像化更加平滑，在非线性激活过程中可以容许更好的信息保留，当深入神经网络时会得到更多的细节特征，提升模型的准确性和泛化性。Mish激活函数主要添加在第一层网络卷积层，以及7个

表2　改进的MobileNet V3主干网络参数表

Tab. 2　Improved MobileNet V3 backbone network parameter table

Input	Operator	Exp size	Out size	SE	NL	s
$224^2\times3$	Conv2d	—	32	—	MH	2
$112^2\times32$	Bneck,3×3	32	32	—	RE	1
$112^2\times32$	Bneck,3×3	64	64	—	RE	2
$56^2\times64$	Bneck,3×3	128	64	—	RE	1
$56^2\times64$	Bneck,5×5	128	64	√	RE	2
$28^2\times64$	Bneck,5×5	128	128	√	RE	1
$28^2\times128$	Bneck,5×5	256	128	√	RE	1
$28^2\times128$	Bneck,3×3	256	128	—	MH	2
$14^2\times128$	Bneck,3×3	256	256	—	MH	1
$14^2\times256$	Bneck,3×3	512	256	—	MH	1
$14^2\times256$	Bneck,3×3	512	256	—	MH	1
$14^2\times256$	Bneck,3×3	512	512	√	MH	1
$14^2\times256$	Bneck,3×3	1 024	512	√	MH	1
$14^2\times512$	Bneck,5×5	1 024	512	√	MH	2
$7^2\times512$	AvgPool,7×7	—	512	—	—	1
$1^2\times512$	FC	—	512	—	—	—
$1^2\times512$	Softmax	—	7/8	—	—	—

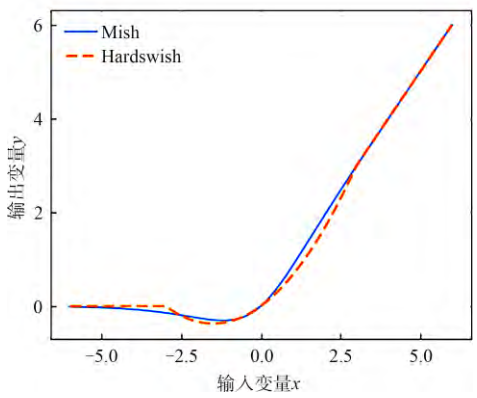

图3　激活函数图

Fig. 3　Activation function diagram

Bneck之中的点卷积和深度卷积之后。

$$\text{Hardswish}(x)=x\frac{\text{ReLU}6(X+3)}{6}, \quad (3)$$

$$\text{Mish}(x)=x\cdot\tanh\left(\ln\left(1+\mathrm{e}^{x}\right)\right). \quad (4)$$

2.2 改进的协调注意力机制模块

在MobileNet V3网络中有SE通道注意力机制,其加入在倒残差结构的深度卷积之后、点卷积降维之前,能够提升模型在通道方面的特征提取能力,但它忽略了空间上的位置信息。针对本文中人脸面部表情识别任务,不同的类别表情在图像整体区域表现不同,空间上的信息差异是不容忽略的。在人脸表情的特征提取阶段,应加强对人脸面部不同区域信息表达的关注。因此本文方法引入协调注意力机制(Coordinate Attention)[13]并对其进行改进,提出改进的协调注意力机制模块(New Coordinate Attention Bneck,NCA Bneck),提升模型在空间上和通道上的特征提取能力,增强感兴趣的对象区域。改进协调注意力机制分为两步:张量信息嵌入和张量信息集成,其结构如图4所示。

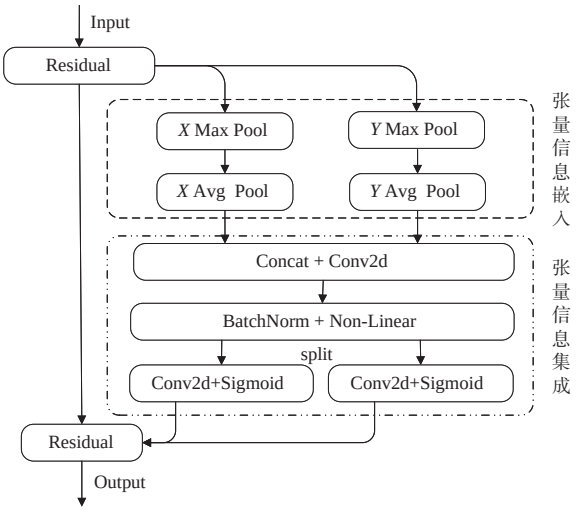

图4 改进的协调注意力机制

Fig. 4 New coordinate attention

(1)张量信息嵌入主要进行特征的聚合,特征聚合通过张量信息编码实现。将一个x的张量作为输入,并使用大小为$(H,1)$和$(1,W)$的池化核分别沿水平和垂直坐标方向对来自各通道的数据进行编码。水平和竖直方向编码可由式(5)、(6)和式(7)、(8)分别表示,式(6)和式(8)表示高度为h的第c个通道的输出和宽度为w的第c个通道的输出。

全局最大池化会提取特征图纹理结构等信息,全局平均池化会捕获图像的全局信息。在水平和竖直两个空间利用自适应最大池化和自适应平均池化进行特征聚合,能够提取图中最具有表现力的区域,返回方向感知的注意力图。经过实验,输入张量先进行全局最大池化,可以提取人脸图像中最突出的部分,重点提取不同类别表情的主要特征。然后再进行全局平均池化,可以增强人脸表情特征的表示能力,还能减少模型过拟合的风险。

$$x_{c^*}=\text{Max}\left\{x_c(h,0)\ldots,x_c(h,W-1)\right\}, \quad (5)$$

$$Z_c^h(h)=\frac{1}{W}\sum_{0\leqslant i<W}x_{c^*}(h,i), \quad (6)$$

$$y_{c^*}=\text{Max}\left\{y_c(0,w)\ldots,y_c(H-1,w)\right\}, \quad (7)$$

$$Z_c^w(w)=\frac{1}{H}\sum_{0\leqslant j<H}y_{c^*}(j,w). \quad (8)$$

(2)张量信息集成是一个可以使模型产生具有全局感受野和精确位置信息的过程。首先,对在张量信息嵌入过程所产生的两个特征图进行合并,并应用逐点卷积进行F1转换。这就产生了一个为$f\in R^{C/r\times(H+W)}$的中间特征图(r为下采样比例),其中包含了水平和垂直方向的空间信息,这个过程在数学上可以表示为式(9)。然后将张量$\boldsymbol{f}$分割成两个独立张量$\boldsymbol{f}^h\in R^{C/r\times H}$和$\boldsymbol{f}^w\in R^{C/r\times W}$,再分别使用逐点卷积操作,将两个特征图转换为与初始输入张量一致的通道数,这一操作可由式(10)、(11)实现。相应通道的注意力权重分别为g^h和g^w,条件协调注意力机制输出结果如式(12)所示。

$$\boldsymbol{f}=\delta\left(F_1\left(\left[z^h,z^w\right]\right)\right), \quad (9)$$

$$\mathrm{g}^h=\sigma\left(F_h\left(f^h\right)\right), \quad (10)$$

$$g^w=\sigma\left(F_w\left(f^w\right)\right), \quad (11)$$

$$y_c(i,j)=x_c(i,j)\times g_c^h(i)\times g_c^w(j). \quad (12)$$

将改进的协调注意力机制加入倒残差结构中的深度卷积之后、点卷积降维之前,条件协调注意力机制的倒残差结构如图5所示。在此处加入条件协调注意力机制能够使特征图在提升通道数的情况下更好地提取位置和通道信息,提高注意力机制的效率。

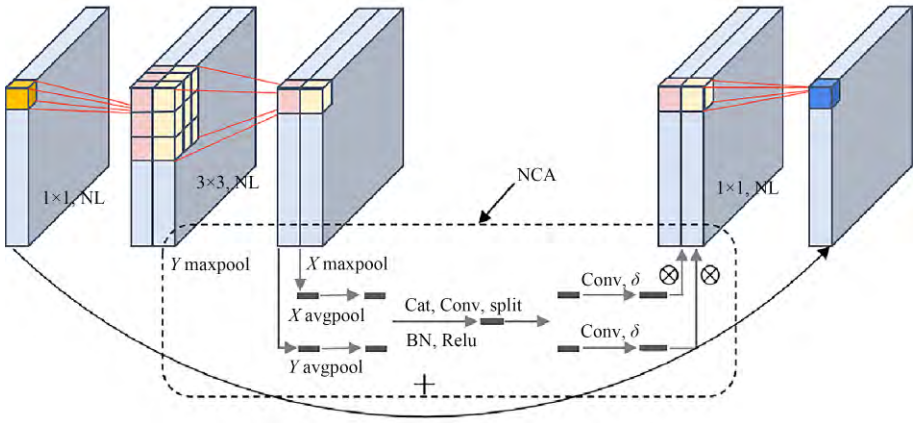

图 5　改进协调注意力机制模块（NCA Bneck）

Fig. 5　New coordinate attention bneck（NCA Bneck）

3　实验结果和分析

3.1　实验环境及配置

本文实验是在 Windows10 操作系统下完成的。实验环境包括计算机基础硬件、图形图像处理单元（GPU）、中央处理器（CPU）、并行计算库 CUDA、Pytorch 深度学习框架、Anaconda、VS-Code、Python 等，具体实验环境配置如表 3 所示。

表 3　实验环境配置

Tab. 3　Experimental environment configuration

名称	配置
CPU	Intel(R) Core(TM) i7－10875H
GPU	NVIDIA GeForce GTX1650(4G)
操作系统	Windows10
并行计算库	CUDA11. 1＋cuDNN8. 2. 1
深度学习框架	Pytorch1. 9. 0
编程语言	Python3. 7

选用 AdamW 优化器，设置衰减策略来调整学习率，损失函数使用交叉熵损失。在网络训练过程中涉及超参数较多，主要包括：Bacth Size、Epochs、初始学习率、学习衰减率、优化器等，具体超参数设置如表 4 所示。

表 4　超参数设置

Tab. 4　Hyperparameter setting

超参数名称	数值
Bacth size	24
Epochs	150
初始学习率	0. 000 5
学习衰减率	0. 01
优化器	AdamW

3.2　表情数据集及数据预处理

在两个公开数据集 FERPlus[11] 和 RAF-DB[12] 上进行实验验证。FERPlus 数据集由 25 045 张训练集、3 191 张私人测试集和 3 137 张公共测试集图像组成，均为 48×48 的灰度图，共分为 8 类基本表情（愤怒、厌恶、恐惧、高兴、悲伤、惊讶、中性、蔑视）。RAF-DB 数据集包含单标签表情图像和多标签表情图像共计 29 672 张，主要是从互联网上下载后经专业标注所得的大小均为 100×100 的 RGB 图像。单标签图像包含 7 类表情（愤怒、厌恶、恐惧、高兴、悲伤、惊讶、中性），共计 15 339 张图片，选择其作为数据集。两个数据集的具体表情数量如表 5 所示，部分样例图如图 6 所示。

表 5　数据集 FERPlus 和 RAF-DB 的表情数量

Tab. 5　Number of images expression from FERPlus and RAF-DB

Data	Anger	Disgust	Fear	Happy	Sad	Surprise	Neutral	Contempt
FER-Plus	2 100	119	532	7 287	3 014	3 149	8 740	119
RAF-DB	867	877	355	5 957	2 460	1 619	3 204	—

图 6 FERPlus 和 RAF-DB 数据集样例图

Fig. 6 Example diagram of FERPlus and RAF-DB dataset

3.3 实验结果与分析

3.3.1 消融实验及其分析

在数据集 RAF-DB 上对不同的改进方案进行实验，可以验证各部分模块对模型整体的作用效果，实验结果见表 6。

分析表 6 可知，方案 1 主干网络 Base 中的激活函数为 Hardswish。方案 2 引入 Mish 替换 Hardswish，不仅降低了模型复杂度，而且提升了识别准确率。方案 4 单独引入条件坐标注意力机制 NCA，虽然模型复杂度变高，但是识别准确率得到提升，说明条件坐标注意力机制能够有效提升模型的特征提取能力。方案 6 同时引入 Mish 和 NCA 后，模型复杂度得到降低，与方案 1 相比准确率提升了 1.76%，说明二者的引入具有相互促进作用。对比方案 5 和方案 6，可以看出方案 6 准确率更高，更适合人脸表情的识别任务。

表 6 NCA-MobileNet 消融实验

Tab. 6 Ablation experiment of NCA-MobileNet

方案	Base	Mish	CA	NCA	Params/M	FLOPs/M	RAF-DB Acc/%
1	√	×	×	×	3.41	826.15	84.14
2	√	√	×	×	3.41	824.14	84.44
3	√	×	√	×	3.54	827.94	84.76
4	√	×	×	√	3.54	827.94	84.78
5	√	√	√	×	3.54	825.93	85.12
6	√	√	×	√	3.54	825.93	85.90

3.3.2 对比实验及其分析

深度学习模型性能优劣的评估不只有准确率等指标，还应该兼顾模型参数量、模型复杂程度、推理时间等指标。为了更好地对比本文方法和最新的人脸表情识别算法的性能，本文对部分表情识别算法进行复现，并提取其关键性指标进行对比实验。在 FERPlus 和 RAF-DB 数据集上进行相关实验，结果见表 7。

分析表 7 可知，本文所提方法在数据集 FERPlus 和 RAF-DB 上准确率最高，分别为 88.84%

表 7 各模型算法性能对比

Tab. 7 Performance comparison of various models

Model	Params/M	FLOPs/M	Inference time/ms	FERPlus Acc /%	RAF-DB Acc /%
LDR[14]	—	—	—	87.60	—
DSAN[15]	—	—	—	—	85.37
MFN[16]	31.00	16 950	—	—	85.39
MFN+[16]	8.26	8 110	—	—	82.43
CERN[17]	1.45	1 780	5.69	87.47	84.08
Ada-CM[18]	11.18	1 819	7.93	88.07	85.02
MobileNet V3S	1.52	58.79	5.45	87.68	83.65
MobileNet V3L	4.21	227.95	7.63	88.01	84.44
本文方法	3.54	825.93	6.19	88.84	85.90

和 85.90%。本文方法与 MobileNet V3S 网络相比，在 FERPlus 和 RAF-DB 数据集上准确率分别提高 1.16% 和 2.25%；与 MobileNet V3L 网络相比，参数量减小 15.91%，推理时间减少 18.87%，准确率分别提高 0.83% 和 1.46%。与最近的表情识别算法 LDR、DSAN、MFN、CERN、Ada-CM 等相比，本文方法的参数量较小，模型推理时间适中，准确率最高。上述试验验证了本文方法对人脸面部表情识别的有效性。

3.3.3　混淆矩阵及其分析

绘制 FERPlus 和 RAD-DB 数据集的混淆矩阵，如图 7 所示。分析图 7 可知，图 7(a)、(b)两个混淆矩阵中高兴表情的识别率最高，都为 95%，这是因为高兴标签的图像数量最多，拥有丰富的表情特征，易于与其他表情区别开。FERPlus 混淆矩阵中，蔑视、厌恶、恐惧表情的误识别率较高，蔑视表情被误识别为中性表情的概率为 17%，厌恶表情被误识别为愤怒表情的概率为 13%，恐惧表情被误识别为惊讶表情的概率为 13%。这是因为这些误识别的表情和正确表情之间拥有高度相似的外观特征，易造成混淆，从而判断失误。悲伤表情被误识别为中性表情的概率为 13%。分析可知，悲伤表情的部分图像表情表现程度较轻，人为判定存在歧义，数据集前期图像标注时存在误标注的现象。RAF-DB 混淆矩阵中，厌恶与恐惧表情的识别率最低，厌恶表情被误识别为生气表情的概率为 6%，恐惧表情被误识别为惊讶表情的概率为 11%。这同样是因为这些表情拥有高度相似的区域特征，易造成混淆。

3.3.4　热力图可视化分析

本文对部分测试图像进行特征区域可视化，使用基于公理的梯度类激活映射（Axiom-based

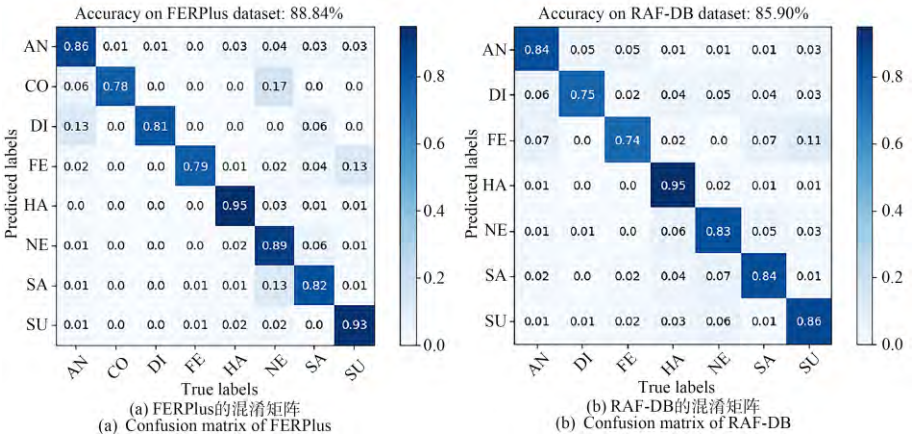

图 7　不同数据集上的混淆矩阵

Fig. 7　Confusion matrix on different datasets

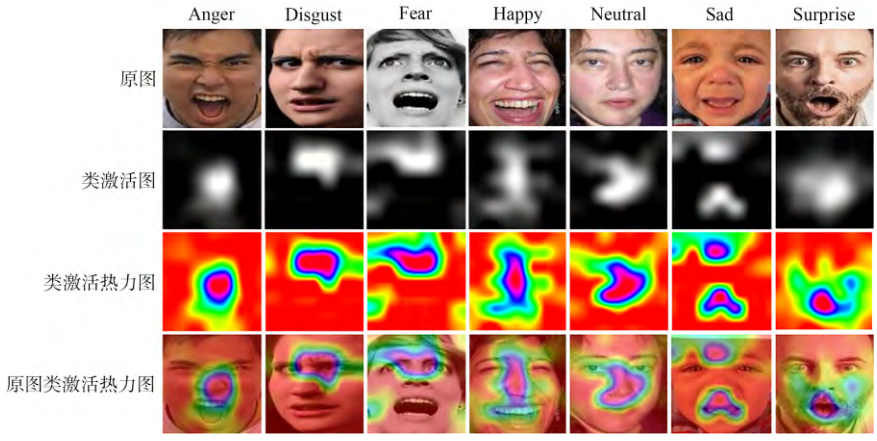

图 8　表情热力图

Fig. 8　Heat maps of different expressions

Grad-CAM, XGradCAM)[19]可以显示出不同类别的识别概率与输入图片中的像素敏感区域,可以直观地解释和分析模型所激活的脸部位置,有助于表情识别的进一步研究。使用RAF-DB数据集中部分图片用于测试,对不同表情进行特征可视化显示,其结果如图8所示。

类激活图中的白色区域为类别表情热力图概率的敏感区域。类激活热力图中的蓝色圈部分为敏感区域,越敏感的区域则温度越高、颜色越红。分析图8可知,由原图上的类激活热力图可以看出,每一表情类别的主要敏感区域都不同,很容易区分。但是厌恶和害怕的敏感区域重合较多,这也反映出二者在类别判断时有许多类似特征区域,容易造成误识别的情况出现,从而使得二者的识别准确率下降。

4 结 论

本文研究了一种结合条件协调注意力机制的轻量化人脸面部表情识别方法。模型构建中,为了获取人脸面部不同表情的深层特征信息,首先增大深度卷积中间通道数,并引入新的激活函数代替原激活函数,降低模型复杂度,提升推理速度;然后加入改进后的条件协调注意力机制,在充分提取表情本身特征基础上,尽可能获取不同表情区域的特征信息进行编码,提升表情识别准确率。在公开数据集FERPlus和RAF-DB进行实验,结果表明,所提方法的参数量为3.54M,准确率分别为88.84%和85.90%,优于目前主流的CERN等轻量级表情识别算法。

参 考 文 献:

[1] YU M, GUO Z Q, YU Y, *et al*. Spatiotemporal feature descriptor for micro-expression recognition using local cube binary pattern [J]. *IEEE Access*, 2019, 7: 159214-159225.

[2] 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.

[3] LI Y, ZENG J B, SHAN S G, *et al*. Occlusion aware facial expression recognition using CNN with attention mechanism [J]. *IEEE Transactions on Image Processing*, 2019, 28(5): 2439-2450.

[4] WANG K, PENG X J, YANG J F, *et al*. Suppressing uncertainties for large-scale facial expression recognition [C]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE, 2020: 6896-6905.

[5] WANG K, PENG X J, YANG J F, *et al*. Region attention networks for pose and occlusion robust facial expression recognition [J]. *IEEE Transactions on Image Processing*, 2020, 29: 4057-4069.

[6] 李召峰,朱明. 基于视频放大和双分支网络的微表情识别[J]. 液晶与显示,2022,37(3):386-394.
LI Z F, ZHU M. Micro-expression recognition based on video magnification and dual-branch network [J]. *Chinese Journal of Liquid Crystals and Displays*, 2022, 37(3): 386-394. (in Chinese)

[7] HOWARD A G, ZHU M L, CHEN B, *et al*. MobileNets: efficient convolutional neural networks for mobile vision applications [J/OL]. *arXiv*, 2017: 1704.04861.

[8] SANDLER M, HOWARD A, ZHU M L, *et al*. MobileNetV2: inverted residuals and linear bottlenecks [C]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018: 4510-4520.

[9] HOWARD A, SANDLER M, CHEN B, *et al*. Searching for MobileNetV3 [C]. 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019: 1314-1324.

[10] DIGANTA M. Mish: a self regularized non-monotonic neural activation function [J/OL]. *arXiv*, 2019: 1908.08681.

[11] BARSOUM E, ZHANG C, FERRER C C, *et al*. Training deep networks for facial expression recognition with crowd-sourced label distribution [C]//*Proceedings of the 18th ACM International Conference on Multimodal Interaction*. Tokyo: ACM, 2016: 279-283.

[12] LI S, DENG W H, DU J P. 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*. Honolulu,

USA：IEEE，2017：2584-2593.

[13] HOU Q B，ZHOU D Q，FENG J S. Coordinate attention for efficient mobile network design [C]. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville，USA：IEEE，2021：13708-13717.

[14] FAN X Y，DENG Z Y，WANG K，*et al*. Learning discriminative representation for facial expression recognition from uncertainties [C]. 2020 IEEE International Conference on Image Processing. Abu Dhabi，United Arab Emirates：IEEE，2020：903-907.

[15] FAN Y R，LI V O K，LAM J C K. Facial expression recognition with deeply-supervised attention network [J]. *IEEE Transactions on Affective Computing*，2022，13(2)：1057-1071.

[16] 唐宏，向俊玲，陈海涛，等. 多区域融合轻量级人脸表情识别网络[J]. 激光与光电子学进展，2023，60(6)：0610006.
TANG H，XIANG J L，CHEN H T，*et al*. Lightweight network based on multiregion fusion for facial expression recognition [J]. *Laser & Optoelectronics Progress*，2023，60(6)：0610006.（in Chinese）

[17] GERA D，BALASUBRAMANIAN S，JAMI A. CERN：compact facial expression recognition net [J]. *Pattern Recognition Letters*，2022，155：9-18.

[18] LI H Y，WANG N N，YANG X，*et al*. Towards semi-supervised deep facial expression recognition with an adaptive confidence margin [C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans：IEEE，2022：4156-4165.

[19] FU R G，HU Q Y，DONG X H，*et al*. Axiom-based grad-CAM：towards accurate visualization and explanation of CNNs [C]. 31st British Machine Vision Conference 2020. BMVC，2020.

作者简介：

左义海，男，学士，高级实验师，2008年于中北大学获得学士学位，主要从事机器人控制及应用方向的研究。E-mail：zuoyh@tit. edu. cn

何秋生，男，博士，教授，2007年于中国矿业大学（北京）获得博士学位，主要从事机器人控制、机器视觉方面的研究。E-mail：heqs2008@126. com