一种基于注意力模型的面部表情识别算法

褚晶辉，汤文豪，张姗，吕卫*
天津大学电气自动化与信息工程学院，天津 300072

摘要 深度卷积网络提取的表情特征易受背景、个体身份等因素影响，其与无用特征混合在一起对表情识别造成干扰。针对此问题，提出一种基于注意力模型的面部表情识别算法，该方法基于一个轻量级的卷积神经网络以避免过拟合，通过通道注意力模块和空间注意力模块对特征图元素进行加强或抑制，应用残差学习单元使注意力模型学习到更丰富的特征并获得更好的梯度流。此外，还提出一种面部表情关键区域截取方案，以解决非表情区域的噪声干扰问题。在两个常用的表情数据集 CK+和 MMI 上对所提方法进行了验证，实验结果证明了该方法的优越性。

关键词 图像处理；表情识别；面部分析；卷积神经网络；注意力模型
中图分类号 TP391.4 **文献标志码** A **doi**：10.3788/LOP57.121015

An Attention Model-Based Facial Expression Recognition Algorithm

Chu Jinghui, Tang Wenhao, Zhang Shan, Lü Wei*
School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, *China*

Abstract Facial features extracted by a deep convolutional network are susceptible to background, individual identity, and other factors, which are mixed with unnecessary features that interfere with facial expression recognition. To solve this problem, an attention model-based facial expression recognition algorithm is proposed in this paper. To avoid overfitting, this method is based on a lightweight convolutional neural network. Moreover, the channel attention model and the spatial attention model are employed to strengthen or suppress the feature map elements. A residual learning unit is used to enable the attention model to learn rich features and obtain an excellent gradient flow. In addition, a key area crop scheme for facial expressions is proposed to solve the problem of noise interference in non-expressive regions. The proposed method is validated on two commonly used expression datasets: CK+ and MMI. Experimental results demonstrate the superiority of the proposed method.

Key words image processing; expression recognition; facial analysis; convolutional neural network; attention model
OCIS codes 100.3008; 100.4996; 100.2000

1 引 言

随着深度学习和人机交互技术的发展，从图像中理解人类情感变得越来越重要。面部表情是人类表达情感的主要方式，是人机界面系统识别人类内心情感的重要途径之一[1]。面部表情识别是计算机视觉领域的经典问题，旨在从人脸图像中预测基本面部表情，如快乐、悲伤、惊讶、愤怒、恐惧、厌恶等。自动面部表情识别在社交机器人、疲劳驾驶监测、广告推荐、教育评价、医学治疗等领域受到广泛关注。

传统面部表情识别方法大多采用人工特征，如 LBP[2]、LPQ[3]、HoG[4]等。这些人工设计的特征在一些简单受控数据集上表现不错，但仍然不足以解决真实场景中的面部表情识别问题。样本的多样性不足，使得训练的模型缺乏对广泛场景的泛化能力，难以满足实际应用需要。这种情形下，研究人员收集了许多不受控的面部表情数据集。这些数据集大多来自互联网，各种角度、光线、遮挡以及身份不同的样本都被收集下来，传统的人工特征已经很难识别与面部表情无关的多种因素。

收稿日期：2019-09-09；**修回日期**：2019-10-29；**录用日期**：2019-11-02
基金项目：国家自然科学基金(61572356)、天津市科技重大专项与工程计划项目(17ZXRGGX00180)
* **E-mail**：luwei@tju.edu.cn

近年来，随着计算机算力的大幅提升，深度学习在计算机视觉领域展现出强大的性能。卷积神经网络(CNN)能够自动从原始数据中提取有效特征，从而避免人工设计特征，有效处理面部表情的多样性。自第一个大规模的深度神经网络 AlexNet[5]诞生以来，各种骨干架构如 VGGNet[6]、GoogLeNet[7]、ResNet[8]、DenseNet[9]等相继被提出。这些骨干架构的性能不断提升，网络规模越来越大。然而，大规模的深层网络不适用于小型的表情数据集，这是因为大量的网络参数使得网络很容易陷入过拟合，导致网络性能下降。此外，表情识别的难点在于，一些表情的类间差异比较微小，而由于人脸身份、背景等非面部表情因素的影响，表情的类内差异比较突出，这使得表情识别更加具有挑战性。

本文提出一种基于注意力模型的面部表情识别网络，能够同时在特征图的通道和空间上寻找表情特征关键区域，并生成一个三维注意力图，通过残差学习单元对特征图元素进行加强或抑制。此外，还提出一种面部表情关键区域截取方案，截取眉毛、眼睛、鼻子、嘴巴这 4 个部位及其周围的区域组成人脸 T 型区，将其作为网络输入，以解决非表情因素的干扰问题。同时，本文采用轻量级的网络结构，在保证模型容量的前提下有效缓解了过拟合问题。本文方法在两个常用的表情数据集 CK+[10]和 MMI[11-12]上进行验证，实验结果证明了本文方法的优越性。

2 相关工作

2.1 面部表情识别

面部表情识别一般包括人脸预处理、特征提取和分类这 3 个阶段。人脸预处理阶段包括：1)人脸检测，从图像中检测人脸区域，去除背景干扰，常见的人脸检测器有 V-J 人脸检测器[13]、MTCNN[14]等；2)人脸关键点检测，从人脸图像中检测面部关键点，定位眼睛、鼻子、嘴巴等部位，常见的关键点检测器有 Dlib[15]人脸关键点检测、MTCNN[13]、FAN[16]等；3)人脸对齐，对检测到的人脸图像关键点坐标进行仿射变换后，将其映射到一个固定位置，达到人脸对齐的目的。特征提取阶段从对齐的人脸图像中提取特征，如几何特征、外观特征、运动特征等。在分类阶段，本文算法将面部表情归类为 6 种或 7 种基本情绪。

2.2 基于深度学习的面部表情识别

Liu 等[17]提出了 BDBN 方法，把特征提取、特征选择和分类器统一构建在一个增强深度信念网络中，可以获得更好的性能。Jung 等[18]提出了 DTGAN 方法，该方法由两个深度网络组成，一个深度网络从图像序列中提取时域外观特征，另一个深度网络从时域人脸关键点中提取时域几何特征，两个网络联合微调，提高了表情识别的准确性。Kuo 等[19]提出了一种小规模的网络结构来识别面部表情，该网络结构在使用更少参数的情况下获得了较优的性能；同时提出一种结合直方图均衡化和线性映射的光照增强方案来解决混合数据源训练深度网络时的过拟合问题，在跨数据集验证上获得较好的性能。何志超等[20]提出一种多分辨率特征融合的卷积神经网络，通过两个深度不同的卷积网络提取图像的不同分辨率特征，多分辨率特征融合提升了模型的识别性能，增强了模型的泛化能力。姚丽莎等[21]提出一种基于卷积神经网络局部特征融合的面部表情识别方法，将卷积神经网络当作特征提取器，提取眼睛、眉毛、嘴巴 3 个局部区域的特征，然后使用 SVM 多分类器进行决策级加权融合，该方法实时性强、识别效果好。

为了减缓人脸身份信息对表情信息的干扰，Meng 等[22]提出了同时利用表情信息和身份信息进行面部表情识别的 IACNN 方法，提出的联合损失函数可以同时考虑表情识别的分类错误以及表情与身份的相似性。Ding 等[23]发现人脸识别与面部表情识别之间存在较大的差异，基于人脸识别网络微调的面部表情识别网络中仍然存在人脸身份信息，这削弱了网络表达不同表情的能力。本文提出一种新的表情识别训练方法 FaceNet2ExpNet，其两阶段的功能主要为：第一阶段，训练表情识别网络的卷积层，由被冻结的人脸识别网络的卷积层进行监督；第二阶段附加上全连接层，共同训练整个网络。

针对类内差异大、类间差异小的问题，Li 等[24]提出了局部保持损失函数(LP loss)，在保持特征局部紧凑的同时增大类间分布，从而增强了深度特征的判别能力，因此 LP loss 更适用于面部表情多模态的情形。考虑到不同类在特征空间的分布上有重叠，Cai 等[25]提出一种岛屿损失函数(island loss)，在相同类表情相互聚集的同时，将不同的表情相互推开。吴慧华等[26]结合岛屿损失函数和 AM-softmax loss 函数，提出了基于余弦距离的损失函数，可用来监督卷积神经网络学习到具有更大的类间距和更小的类内距的人脸表情特征。

2.3 注意力机制

注意力机制[27]在人类感知过程中扮演着重要

的角色。人类视觉系统的一个重要特性是不会试图一次处理整个场景，而是选择性地聚焦于需要关注的目标区域，以获得目标的细节信息，而抑制其他无用信息，这极大地提高了视觉信息处理的效率。Jaderberg 等[28]提出一种 STN 模块，其端到端学习一个仿射变换矩阵，使其能够对特征图进行平移、缩放、旋转等空间变换，从而聚焦于关键目标区域。Hu 等[29]提出一个重要的 SENet 模型结构，通过全局平均池化将三维特征图的空间信息压缩到一维通道特征图中，然后通过两个全连接层融合各通道的信息，学习每个通道的注意力权重。

3 基于注意力模型的面部表情识别算法

3.1 卷积神经网络模型

所提出的基于通道和空间注意力的卷积神经网络(CSACNN)结构如图 1 所示。该模型由 8 个卷积层、4 个池化层和 3 个全连接层组成，在每个卷积层后进行批规范化和 PReLU 激活。本文模型的主干结构是在文献[19]的基础上改进的，主要改进内容包括：1)用 2 个 3×3 的卷积核替代 5×5 的卷积核，在维持较大感受野的前提下，减小了模型的计算量；2)将每个池化层后卷积核的数量增加一倍，从而增强了模型的表示能力；3)在每个池化层前添加注意力模型，使网络加强对表情相关特征的学习。

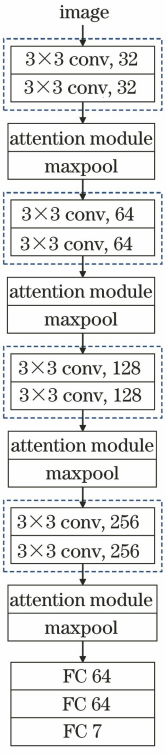

图 1 CSACNN 模型结构

Fig. 1 CSACNN model structure

3.2 注意力模型

面部表情识别的关键问题是寻找表情变化突出的表情特征区域。注意力机制是被广泛认为有助于解决这类问题的方法。受文献[30]的启发，本研究将注意力模型应用于面部表情识别网络，该注意力模型由通道注意力和空间注意力两部分组成。

3.2.1 通道注意力分支

特征图的每个通道都可看作是一个特征检测器，因此通道注意力关注的是哪种特征对输入图像来说更有意义。如图 2 所示，通道注意力首先在特征图 $\boldsymbol{X}\in\mathbf{R}^{c\times h\times w}$ 上取全局平均池化(GAP)，生成一个 $\mathbf{R}^{c\times1\times1}$ 的通道特征图，并经过两个全连接层进行通道信息融合。为了减小计算量，通道特征向量经过第一个完全连接层(FC)降维到 $\mathbf{R}^{\frac{c}{r}\times1\times1}$，其中 r 为降维比例，再经过第二个全连接层恢复至 $\mathbf{R}^{c\times1\times1}$。此外，在第一个全连接层后添加批规范化(BN)和 ReLU 激活。最后将 $\mathbf{R}^{c\times1\times1}$ 的通道特征向量扩展得到通道注意力图 $\boldsymbol{X}_{\mathrm{ch\_att}}\in\mathbf{R}^{c\times h\times w}$。

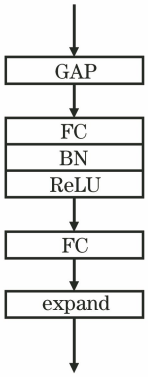

图 2 通道注意力分支

Fig. 2 Channel attention branch

3.2.2 空间注意力分支

不同于通道注意力，空间注意力侧重于寻找输入图像中更需关注的位置。如图 3 所示，空间注意力分支由 4 个卷积层组成。首先，特征图 $\boldsymbol{X}\in\mathbf{R}^{c\times h\times w}$ 经过一个 1×1 的卷积层降维到 $\mathbf{R}^{\frac{c}{r}\times h\times w}$，以节省计算开销；然后，经过两个扩张率为 d 的卷积层进行信息融合，再经过一个 1×1 卷积层后，特征图进一步降维到 $\mathbf{R}^{1\times h\times w}$，并在每个卷积层后添加批规范化和 ReLU 激活。最后，将 $\mathbf{R}^{1\times h\times w}$ 的特征图扩展得到空间注意力图 $\boldsymbol{X}_{\mathrm{sp\_att}}\in\mathbf{R}^{c\times h\times w}$。

3.2.3 注意力模型集成与残差学习单元

如图 4 所示，使用对应元素求和的方式对通道注意力图和空间注意力图进行集成，然后经过

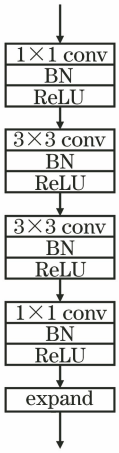

图 3 空间注意力分支

Fig. 3 Spatial attention branch

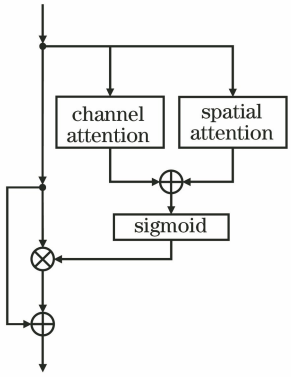

图 4 注意力模型集成与残差学习单元

Fig. 4 Attention model integration and residual learning unit

sigmoid 激活函数得到最终的注意力图，其取值范围为(0,1)。此外，以残差学习的方式学习注意力模型，以获得更丰富的特征和更好的梯度流。集成注意力模型的公式为

$$\boldsymbol{X}_{\mathrm{att}}=\boldsymbol{X}\otimes[\boldsymbol{1}\oplus\mathrm{sigmoid}(\boldsymbol{X}_{\mathrm{ch\_att}}\oplus\boldsymbol{X}_{\mathrm{sp\_att}})], \tag{1}$$

式中：$\boldsymbol{X}_{\mathrm{att}}\in\mathbf{R}^{c\times h\times w}$ 表示集成注意力图；$\boldsymbol{1}$ 表示大小为 $\mathbf{R}^{c\times h\times w}$，元素全部为1的张量；$\boldsymbol{X}_{\mathrm{ch\_att}}\in\mathbf{R}^{c\times h\times w}$ 表示通道注意力图；$\boldsymbol{X}_{\mathrm{sp\_att}}\in\mathbf{R}^{c\times h\times w}$ 表示空间注意力图；⊗表示对应元素相乘；⊕表示对应元素相加。

3.3 面部表情关键区域

在面部动作编码系统(FACS)[31]中，面部表情由与面部肌肉变化有关的动作单元编码而成。眉毛、眼睛、鼻子、嘴巴及其周围区域是面部肌肉的主要分布区。面部表情特征在这 4 个部位周围更加显著。为了减小非表情信息的干扰，提出一种面部表情关键区域截取方案。首先，使用 FAN[16]面部关键点检测方法，检测出面部 68 个关键点，记第 n 个关键点的坐标为 (x_n,y_n)，面部关键点的分布如图 5所示。然后，挑选出覆盖眉毛、眼睛、鼻子、嘴巴的 51 个关键点，使用(2)式计算得到一个二维的面部遮罩。

$$f_{\mathrm{Mask}}(i,j)=\min\left\{\sum_{n=17}^{67}\exp\left[-\frac{(i-x_n)^2+(j-y_n)^2}{2\sigma^2}\right],1\right\}, \tag{2}$$

式中：i、j 分别表示遮罩图的横、纵坐标；σ 为变量，用于控制遮罩范围的大小，本研究中取 $\sigma=10$。最后，使用(3)式截取面部表情关键区域。

$$\boldsymbol{I}_{\mathrm{crop}}=\boldsymbol{I}\otimes\boldsymbol{f}_{\mathrm{Mask}}, \tag{3}$$

式中：$\boldsymbol{I}$ 表示原图；$\boldsymbol{I}_{\mathrm{crop}}$ 表示截取图像。整个过程如图 6 所示。

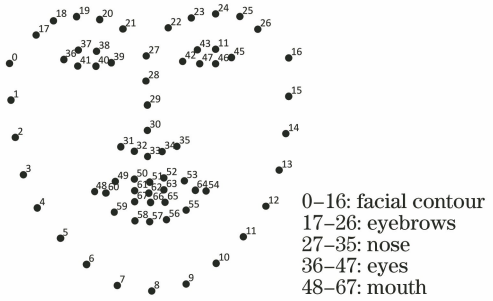

图 5 面部 68 个关键点

Fig. 5 68 face landmarks

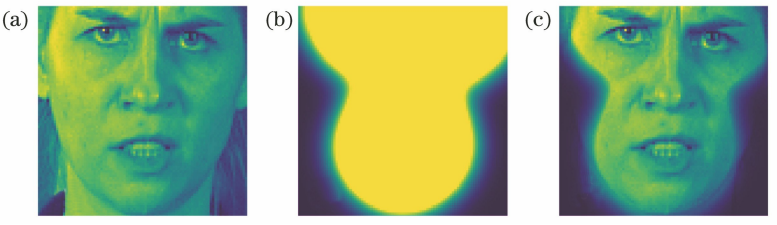

图 6 面部表情关键区域的截取。(a)原图；(b)面部遮罩；(c)截取图像

Fig. 6 Cropping of key areas of facial expression. (a) Original image; (b) facial mask; (c) cropped image

4 实验结果

为了验证所提方法的有效性，在 CK＋、MMI 两个表情数据集上进行实验。

4.1 预处理

使用 3.3 节描述的方法截取面部表情关键区域，得到如图 6(c)所示的截取图像，将截取图像的尺寸缩放到 100 pixel×100 pixel。同时，对数据集使用数据扩充手段，对图像进行翻转、旋转等操作，扩大训练样本的数量，以增加训练样本的多样性。首先，将所有面部区域图像水平翻转得到水平翻转图像集；其次，将每幅图像分别旋转－15°、－12°、－9°、－6°、－3°、3°、6°、9°、12°、15°得到旋转图像集；最后，对旋转图像进行水平翻转得到旋转图像的水平翻转图像。本实验只对训练数据进行扩充，最终得到 22 倍于原数据的实验数据集。

4.2 实验数据集

4.2.1 CK＋数据集

CK＋[10]数据集包含了 118 个不同受试者的 327 个图像序列。每个序列以中性状态开始，以表情峰值结束。图像序列是从正面视角采集的，每个序列被标记为 7 类表情(愤怒、蔑视、厌恶、恐惧、高兴、悲伤和惊喜)。收集了每个序列的最后 3 个峰值帧，建立了包含 981 幅图像的 7 类表情的实验数据集。这些图像按照受试者身份编号的升序分为 10 组，其中 9 组用于训练，1 组用于验证。实验数据集中用于训练的受试者不会用于测试，也就是说实验数据集进行与身份无关的十折交叉验证。

4.2.2 MMI 数据集

MMI[11-12]表情数据集包含了 32 个受试者的 236 个图像序列，每个序列从中性状态到表情峰值再回到中性状态。这些人脸图像是从正面和侧脸视角拍摄的，每个序列被标记为 6 种基本表情(愤怒、厌恶、恐惧、高兴、悲伤和惊喜)中的一种。选取正面视角拍摄的 31 个受试者的 208 个序列，收集每个序列的中间 3 个表情峰值帧，建立了包含 624 幅图像的 6 类表情的实验数据集。实验数据集进行与身份无关的十折交叉验证。

4.3 实验环境和实现细节

本实验是在 Ubuntu 18.04 系统环境下，基于 PyTorch 深度学习框架构建的。中央处理器为 3.20 GHz 的 Intel Core i7-8700 CPU，内存为 64 GB，显卡型号是 NVIDIA GeForce GTX 2080Ti。

所提网络使用预处理后的 90 pixel×90 pixel的人脸图像作为输入，使用批量大小为 440、动量参数为 0.9、权值衰减参数为 0.00005 的随机梯度下降算法进行训练。由于 CK＋数据集与 MMI 数据集包含的表情类数不一致，因此网络的最后一个全连接层要分别构建，对 CK＋数据集进行十折交叉验证时，网络的最后一个全连接层包含 7 个神经元，而对 MMI 来说只用 6 个神经元。

4.4 对比实验

4.4.1 超参数的选择

在 CK＋数据集上研究了注意力模型中的全连接层、1×1 卷积层的降维比 r 和 3×3 卷积层的扩张率 d 对网络性能的影响，实验结果如表 1 所示。结果显示，$r=16$、$d=4$ 时网络性能最好。空间注意力分支需要较大的感受野，以增强对空间信息的感知能力。注意力模型的降维比对网络性能的影响较大。降维比小则模型容量大，在增强了模型表征能力的同时，也会削弱泛化性能；反之，则导致模型表征能力不足。

表 1 超参数对网络性能的影响

Table 1 Effect of hyper-parameters on network performance

Variable	Value	Accuracy /%
d	1	97.35
	4	**97.45**
	8	97.25
r	8	95.72
	16	**97.45**
	32	95.41

4.4.2 注意力模型位置的选择

在网络的三个位置——卷积层后、池化层前，以及池化层后，分别添加注意力模型，研究注意力模型的位置对网络性能的影响，实验结果如表 2 所示。可以看出，将注意力模型加在池化层前对网络性能的提升效果更加显著。注意力模型适用于对相同卷积层堆叠的卷积块进行学习，而不是对每一个卷积

表 2 注意力模型的位置对网络性能的影响

Table 2 Effect of attention model location on network performance

Dataset	Location	Accuracy /%
CK＋	After conv	96.64
	Before pooling	**97.45**
	After pooling	95.72
MMI	After conv	72.69
	Before pooling	**74.73**
	After pooling	72.59

层进行学习。此外，池化层会带来一定的空间信息损失，因此将注意力模型加在池化层前能更有效地处理空间信息。

4.4.3 表情识别性能

本文方法在两个常用的面部表情数据集 CK+和 MMI 上进行了与身份无关的十折交叉验证，实验结果如表 3 所示。

表 3 不同表情识别方法性能对比

Table 3 Performance comparison of different expression recognition methods

Method	Experimental setting	Accuracy /%	
		CK+	MMI
3DCNN[32]	Sequence-based	85.90	53.20
LBP-TOP[33]	Sequence-based	88.99	59.51
HOG 3D[34]	Sequence-based	91.44	60.89
STM-ExpLet[35]	Sequence-based	94.19	**75.12**
DTGAN[18]	Sequence-based	**97.25**	-
Island Loss[25]	Image-based	94.39	**74.68**
IACNN[22]	Image-based	95.37	71.55
DLPCNN[24]	Image-based	95.78	-
DeRL[36]	Image-based	97.30	73.23
Ref.[19]	Image-based	**97.37**	-
PPDN[37]	Image-based	97.30	-
VGG16(ours)	Image-based	91.72	64.13
ResNet5(ours)	Image-based	86.87	57.09
CSACNN(ours)	Image-based	**97.45**	**74.73**

将本文方法在两个经典神经网络结构 VGG16 和 ResNet50 进行表情识别性能的验证，结果表明，直接使用高性能的 CNN 网络结构进行表情识别是不合适的，这是因为较大的模型容易在表情数据集上产生过拟合。本文方法基于静态帧设计，更适合用于单幅图像的识别。相比于图像序列方法，单幅图像方法的计算量更小，性能上没有明显差距。单幅图像方法中 DeRL[36]方法和 PPDN[37]方法使用了中性表情帧作为参考，因此取得了比其他方法更好的性能，Kuo 等[19]采用紧凑网络结构在 CK+数据集的对比中取得了最好的性能。本文方法使用轻量级的网络结构，模型的参数量仅有 1.6×10^6，相比于 Kuo 等[19]研究模型的 2.6×10^6 参数量减少了 38%，但模型准确率并没有显著下降，在 CK+数据集上，本文模型的准确率比文献[19]模型高出了 0.08 个百分点。本文方法取得了相比于其他单幅图像方法更优的性能。

4.5 模块有效性验证

本节验证了通道注意力、空间注意力以及面部表情关键区域截取三个模块的有效性，实验结果如表 4 所示，其中 Base 表示移除了所有注意力模块的基准模型，CA 表示通道注意力模块，SA 表示空间注意力模块，Crop 表示面部关键区域截取。结果显示，通道注意力模块、空间注意力模块以及面部表情关键区域截取对两个数据集的准确率都有一定的提升。通道注意力模块和空间注意力模块的集成则显著提升了识别性能，这说明通道注意力模块和空间注意力模块形成了互补，在通道和空间上共同促进了表情特征的学习。在注意力模型的基础上应用面部关键区域截取方法，进一步提高了识别性能，在 CK+和 MMI 数据集上相较于基准模型的准确率分别提升了 2.72%和 7.12%，证明了 3 个模块的有效性。

表 4 不同模块的性能对比

Table 4 Performance comparison of different modules

Model	Accuracy /%	
	CK+	MMI
Base	94.73	67.61
Base+CA	95.21	71.47
Base+SA	95.62	70.17
Base+Crop	95.82	71.41
Base+CA+SA	96.43	72.98
Base+CA+SA+Crop	97.45	74.73

5 结 论

提出一种基于注意力模型的面部表情识别算法，该方法由通道注意力模块和空间注意力模块两部分组成。通道注意力模块关注不同特征通道的重要性，加强表情特征的作用，减少无用特征的干扰，实现对特征的自适应学习。空间注意力模块对特征图的不同区域赋予不同的重要性权重，实现对显著表情区域的关注。该方法通过通道降维和扩张卷积的方式，以较小的开销实现了对特征图的面部表情识别。此外，提出一种面部关键区域截取方法以避免背景干扰。实验结果表明，本文方法对 CK+和 MMI 表情数据集的面部表情识别准确性明显提高。

参 考 文 献

[1] Li S, Deng W H. Deep facial expression recognition: a survey[EB/OL]. [2019-10-29]. https://arxiv.org/abs/1804.08348.

[2] Shan C F, Gong S G, McOwan P W. Facial expression recognition based on Local Binary Patterns: a comprehensive study[J]. Image and Vision Computing, 2009, 27(6): 803-816.

[3] Ahonen T, Rahtu E, Ojansivu V, et al. Recognition of blurred faces using Local Phase Quantization[C]// 2008 19th International Conference on Pattern Recognition, December 8-11, 2008, Tampa, FL, USA. New York: IEEE, 2008: 1-4.

[4] Dalal N, Triggs B. Histograms of oriented gradients for human detection [C] // 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), June 20-25, 2005, San Diego, CA, USA. New York: IEEE, 2005: 1-8.

[5] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks [J]. Communications of the ACM, 2017, 60(6): 84-90.

[6] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[EB/OL]. [2019-10-29]. https://arxiv.org/abs/1409.1556.

[7] Szegedy C, Liu W, Jia Y Q, et al. Going deeper with convolutions [C] // 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 7-12, 2015, Boston, MA, USA. New York: IEEE, 2015: 1-9.

[8] 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), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE, 2016: 770-778.

[9] Huang G, Liu Z, van der Maaten L, et al. Densely connected convolutional networks [C] // 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI. New York: IEEE, 2017: 4700-4708.

[10] Lucey P, Cohn J F, Kanade T, et al. The Extended Cohn-Kanade Dataset (CK+): a complete dataset for action unit and emotion-specified expression [C] // 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, June 13-18, 2010, San Francisco, CA, USA. New York: IEEE, 2010: 94-101.

[11] Pantic M, Valstar M, Rademaker R, et al. Web-based database for facial expression analysis [C] // 2005 IEEE International Conference on Multimedia and Expo, July 6, 2005, Amsterdam, The Netherlands. New York: IEEE, 2005: 1-5.

[12] Valstar M F, Pantic M. Induced disgust, happiness and surprise: an addition to the mmi facial expression database[C]. Workshop on EMOTION (Satellite of LREC): Corpora for Research on Emotion and Affect, 2010: 65.

[13] Viola P, Jones M J. Robust real-time face detection [J]. International Journal of Computer Vision, 2004, 57(2): 137-154.

[14] Zhang K P, Zhang Z P, Li Z F, et al. Joint face detection and alignment using multitask cascaded convolutional networks [J]. IEEE Signal Processing Letters, 2016, 23(10): 1499-1503.

[15] King D E. Dlib-ml: A machine learning toolkit [J]. Journal of Machine Learning Research, 2009, 10: 1755-1758.

[16] Bulat A, Tzimiropoulos G. How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230000 3D facial landmarks) [C] // 2017 IEEE International Conference on Computer Vision (ICCV), October 22-29, 2017, Venice. New York: IEEE, 2017: 1021-1030.

[17] Liu P, Han S Z, Meng Z B, et al. Facial expression recognition via a boosted deep belief network [C] // 2014 IEEE Conference on Computer Vision and Pattern Recognition, June 23-28, 2014, Columbus, OH, USA. New York: IEEE, 2014: 1805-1812.

[18] Jung H, Lee S, Yim J, et al. Joint fine-tuning in deep neural networks for facial expression recognition [C] // 2015 IEEE International Conference on Computer Vision (ICCV), December 7-13, 2015, Santiago, Chile. New York: IEEE, 2015: 2983-2991.

[19] Kuo C M, Lai S H, Sarkis M. A compact deep learning model for robust facial expression recognition [C] // 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), June 18-22, 2018, Salt Lake City, UT, USA. New York: IEEE, 2018: 2121-2129.

[20] He Z C, Zhao L Z, Chen C. Convolution neural network with multi-resolution feature fusion for facial expression recognition [J]. Laser & Optoelectronics Progress, 2018, 55(7): 071503.
何志超, 赵龙章, 陈闯. 用于人脸表情识别的多分辨率特征融合卷积神经网络[J]. 激光与光电子学进展, 2018, 55(7): 071503.

[21] Yao L S, Xu G M, Zhao F. Facial expression recognition based on local feature fusion of convolutional neural network [J]. Laser & Optoelectronics Progress, 2020, 57(4): 041513.
姚丽莎, 徐国明, 赵凤. 基于卷积神经网络局部特征融合的人脸表情识别[J]. 激光与光电子学进展, 2020, 57(4): 041513.

[22] Meng Z B, Liu P, Cai J, et al. Identity-aware convolutional neural network for facial expression recognition [C] // 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), May 30-June 3, 2017, Washington, DC, USA. New York: IEEE, 2017: 558-565.

[23] Ding H, Zhou S K, Chellappa R. FaceNet2ExpNet: regularizing a deep face recognition net for expression recognition [C] // 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), May 30-June 3, 2017, Washington, DC, USA. New York: IEEE, 2017: 118-126.

[24] 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), July 21-26, 2017, Honolulu, HI. New York: IEEE, 2017: 2852-2861.

[25] Cai J, Meng Z B, Khan A S, et al. Island loss for learning discriminative features in facial expression recognition [C] // 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), May 15-19, 2018, Xi′an. New York: IEEE, 2018: 302-309.

[26] Wu H H, Su H S, Liu G H, et al. Facial expression recognition algorithm based on cosine distance loss function [J]. Laser & Optoelectronics Progress, 2019, 56(24): 241502.
吴慧华, 苏寒松, 刘高华, 等. 基于余弦距离损失函数的人脸表情识别算法[J]. 激光与光电子学进展, 2019, 56(24): 241502.

[27] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [EB/OL]. [2019-10-29]. https://arxiv.org/abs/1706.03762.

[28] Jaderberg M, Simonyan K, Zisserman A, et al. Spatial transformer networks [EB/OL]. [2019-10-29]. https://arxiv.org/abs/1506.02025.

[29] Hu J, Shen L, Sun G. Squeeze-and-excitation networks [C] // 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT. New York: IEEE, 2018: 7132-7141.

[30] Park J, Woo S, Lee J, et al. BAM: bottleneck attention module [EB/OL]. [2019-10-29]. https://arxiv.org/abs/1807.06514.

[31] Ekman P, Friesen, W V. Facial Action Coding System (FACS): a technique for the measurement of facial action[J]. Rivista di Psichiatria, 1978, 47(2): 126-138.

[32] Liu M Y, Li S X, Shan S G, et al. Deeply learning deformable facial action parts model for dynamic expression analysis [M] // Computer Vision-ACCV 2014. Cham: Springer International Publishing, 2015: 143-157.

[33] Zhao G Y, 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.

[34] Klaser A, Marszałek M, Schmid C. A spatio-temporal descriptor based on 3D-gradients [J]. Proceedings of the British Machine Conference, 2008, 99: 1-10.

[35] Liu M Y, Shan S G, Wang R P, et al. Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition [C] // 2014 IEEE Conference on Computer Vision and Pattern Recognition, June 23-28, 2014, Columbus, OH, USA. New York: IEEE, 2014: 1749-1756.

[36] Yang H Y, Ciftci U, Yin L J. Facial expression recognition by de-expression residue learning [C] // 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT. New York: IEEE, 2018: 2168-2177.

[37] Zhao X Y, Liang X D, Liu L Q, et al. Peak-piloted deep network for facial expression recognition [M] // Computer Vision-ECCV 2016. Cham: Springer International Publishing, 2016: 425-442.