激光与光电子学进展 先进成像

多区域融合轻量级人脸表情识别网络

唐宏[1,2]，向俊玲[1,2*]，陈海涛[3]，吕榕城[1]，夏泽昊[3]

[1]重庆邮电大学通信与信息工程学院，重庆 400065；

[2]重庆邮电大学移动通信技术重庆市重点实验室，重庆 400065；

[3]重庆邮电大学国际学院，重庆 400065

摘要 由于人脸表情特有的微妙性和复杂性，对全局面部进行研究时无法突出表情特性。为了增强表情识别在自然环境下的鲁棒性并且优化模型参数，提出一种基于多区域融合的轻量级人脸表情识别方法，融合局部细节特征和全局整体特征，实现粗细粒度结合，增强模型对表情细微变化的判别能力。首先，通过一个分支从人脸子区域提取局部特征，以眼部和嘴部作为细节区域输入，描述面部细节。其次，通过另一个分支从人脸全局自适应地获取面部整体特征，以关键点生成掩模，辅助调节面部注意力图。注意力图作用于全局特征，突出未遮挡部位权重，描述整体高级语义信息。并且，采用剪枝算法对整体模型进行轻量级优化，使用更少的运行内存和计算操作，得到更紧凑的网络。最后，在公开数据集RAF-DB和AffectNet上，所提方法对表情的识别精度分别达85.39%和58.81%。实验结果表明：所提方法的识别精度高于其他先进方法，并显著减少了参数量，有效性和先进性得到证明。

关键词 图像处理；人脸表情识别；注意力图；轻量级网络；剪枝算法；多区域融合

中图分类号 TP391.4 **文献标志码** A **DOI**：10.3788/LOP213204

Lightweight Network Based on Multiregion Fusion for Facial Expression Recognition

Tang Hong[1,2], Xiang Junling[1,2*], Chen Haitao[3], Lü Rongcheng[1], Xia Zehao[3]

[1]*College of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing* 400065, *China*;

[2]*Chongqing Key Laboratory of Mobile Communication Technology, Chongqing University of Posts and Telecommunications, Chongqing* 400065, *China*;

[3]*International College, Chongqing University of Posts and Telecommunications, Chongqing* 400065, *China*

Abstract It is difficult to highlight the features of facial expressions in the study of global faces, due to the unique subtleties and complexity of facial expressions. To improve the robustness of expression recognition in natural environments and optimize model parameters, this paper proposes a lightweight facial expression recognition method based on multiregion fusion, which integrates local details and global features to realize a combination of coarse and fine granularity, thus improving the model's efficacy in discriminating subtle changes in expressions. First, local features are extracted from the human face through a branch, which uses eyes and mouth as input. Second, the facial global features are adaptively acquired by another branch, and a mask is generated by key points to assist in adjusting the facial attention map. The facial attention map acts on the global features to highlight the weight of the unmasked parts and describes the overall high-level semantic information. A pruning algorithm is used to perform lightweight optimization for the overall model, using less memory and few computational operations to obtain a more compact network. The recognition accuracy of the proposed method on RAF-DB and AffectNet datasets is determined to be 85.39% and 58.81%, respectively. The experimental results show that the recognition accuracy of the proposed method is higher than other advanced methods and the proposed method significantly reduces the number of parameters, which proves the effectiveness and progressiveness.

Key words image processing; facial expression recognition; attentional map; lightweight network; pruning algorithm; multi-region fusion

收稿日期：2021-12-13；**修回日期**：2022-01-11；**录用日期**：2022-01-17；**网络首发日期**：2022-01-27

通信作者：*390098758@qq.com

1 引 言

面部表情是人类非语言交流的关键，表情的产生、感知和解读已经得到了广泛的研究。近年来，面部表情识别取得了很大的成就。随着人机交互的兴起，人脸表情识别系统被广泛应用于远程教育、交互游戏和智能交通等领域[1]。然而在真实场景中，人脸图像中存在不相关的背景信息（帽子、头发等），人脸容易受到光照、遮挡的影响，并且表情存在细微变化的特性，使人脸表情识别技术还存在着很大的挑战。

目前，对人脸表情识别（FER）技术的研究主要采用深度学习方法，因为相比传统机器学习方法[2]，卷积神经网络（CNN）可以提取到更深层次、更抽象的特征，可以自适应地调整每一层的卷积核以获得所需的特征，在各种图像分类任务中表现出了优异的性能[3]。Hamester等[4]提出一种基于多通道卷积神经网络的人脸表情识别新体系结构，通过无监督学习的方法，融合两个卷积神经网络通道信息。实验结果表明，该结构融合无监督学习后可以提高准确率，在JAFFE数据集上的精确度达95.8%。Hasani等[5]提出一种三维卷积神经网络，用以解决视频图像中的人脸表情识别问题，在CK+数据集上的识别率达95.53%。

JAFFE数据集和CK+数据集是实验者按照指导要求做出相应表情得到的，诸多学者在这些数据集上进行研究，以提高实验精度，这些实验精度比人类肉眼判断表情的精度还要高，研究也逐渐向更为复杂的方向转变[6-7]。对于人脸表情数据集，逐渐从实验室环境下的摆拍转换到真实场景下，人脸会受到光照、姿势和遮挡等影响，传统神经网络并不能很好地处理这些情况，精度会大幅下降。Happy等[8]和Majumder等[9]认为表情变化通常发生在一些显著的五官区域，如口、鼻、眼附近。对表情的研究方向也从全脸转向为重点关注表情相关区域[10]。目前很多先进的方法是添加注意力机制，这些研究意味着面部局部区域的细节在表情识别中是有区别的，虽然可以聚焦于表情重点区域，但大多数研究仅从整个图像中提取表情特征。这些方法强调面部表情的完整性，而忽略局部细节信息。Wang等[11]根据人脸关键点对人脸面部进行随机剪裁，并将剪裁结果输入网络中，虽然考虑到局部细节，但有很多重复的剪裁区域。Gan等[12]为从关键的面部区域中提取识别特征，提出了一种新的多重注意网络来模拟人类的粗细视觉注意，以提高表情识别性能；定义两个网络，从粗到细粒度级别学习二进制掩码，以定位与突出相关的关键区域。以上方法虽然提高了复杂环境下对表情识别的精度，但在解决复杂情况问题的同时，网络模型的复杂度不可避免会大幅增加，网络计算量也递增，网络运行时也需要很大内存空间和计算资源，不能部署在资源受限的设备上[13]。申毫等[14]提出基于改进残差倒置网络的表情识别模型，筛选浅层特征与深层特征，并将它们融合，实现模型轻量化和多特征融合。尹鹏博等[15]对模型进行降维处理，并嵌入注意力机制，提高了模型的特征提取能力，降低了模型的复杂度。

针对上述问题，本文提出一种基于多区域融合的轻量级人脸表情识别模型，通过两个分支，分别对人脸图像局部细节和全局进行细粒度特征提取；同时在全局分支提出注意力图，以调整特征权重，并用关键点生成掩模，辅助调节注意力图。通过剪枝策略，优化整体模型，实现识别精度和速度的提升。

2 基于多区域融合的轻量级人脸表情识别

提出一种基于局部和全局特征融合细粒度的轻量级人脸表情识别方法，用以解决复杂环境下的人脸表情识别问题。主要模型框架如图1所示，包括两个分支，即局部细节分支和全局自适应关键点分支，以改进的VGG网络为模型主干，将RGB人脸图像作为输入，表情预测值作为输出。对于局部细节分支，选取眼部和嘴部为重点区域，进行图像剪裁，将剪裁结果输入到网络中进行细节特征提取。对于全局自适应关键点分支，选择整张图像输入网络，引入注意力分支，并检测面部关键点，以关键点生成面部掩模，用以调节该分支生成的注意力图。所提模型融合了不同粒度的特征，提高了学习综合判别特征的能力，并将注意力图作用于特征图，使网络更关注表情相关重点区域，抑制非重点区域。引入了剪枝策略，以减小模型的大小，减少运行时的内存和计算量。

2.1 改进VGG网络结构

VGG网络在各项图像分类任务中有优异的表现，突出贡献是：与7×7或5×5的大卷积核相比，3×3卷积核可以在输入数据中表达更强大的特征，并且使用更少的参数量。假设一个卷积层有C个通道，一个7×7卷积层包含$C\times(7\times7\times C)=49C\times C$个参数，而$3\times3$卷积层只包含$3\times\left[C\times(3\times3\times C)\right]=27C\times C$个参数。并且VGG网络对图像数据集具有良好的泛化能力，常用VGG网络来提取图像特征。在该网络中，每个卷积层后使用ReLU激活函数，该函数为不饱和函数，使计算结果稀疏，更能减小反向传播误差，加快网络收敛速度。鉴于所提两个分支需要融合到一个全连接层，会导致网络训练收敛速度变慢，甚至遇到“梯度爆炸”情况，因此提出VGG-BN-16。改进后的VGG网络如图2所示，在每层的第一个卷积后加上batch normalization（BN）[16]，放在激活函数之前。BN层主要解决中间层数据分布的问题，防止梯度消失或爆炸，加快网络收敛速度。

具体的BN操作将隐层中每个神经元的激活值归一化。对于具有d维的输入$x=\left(x^{(1)}\cdots x^{(d)}\right)$，归一化每

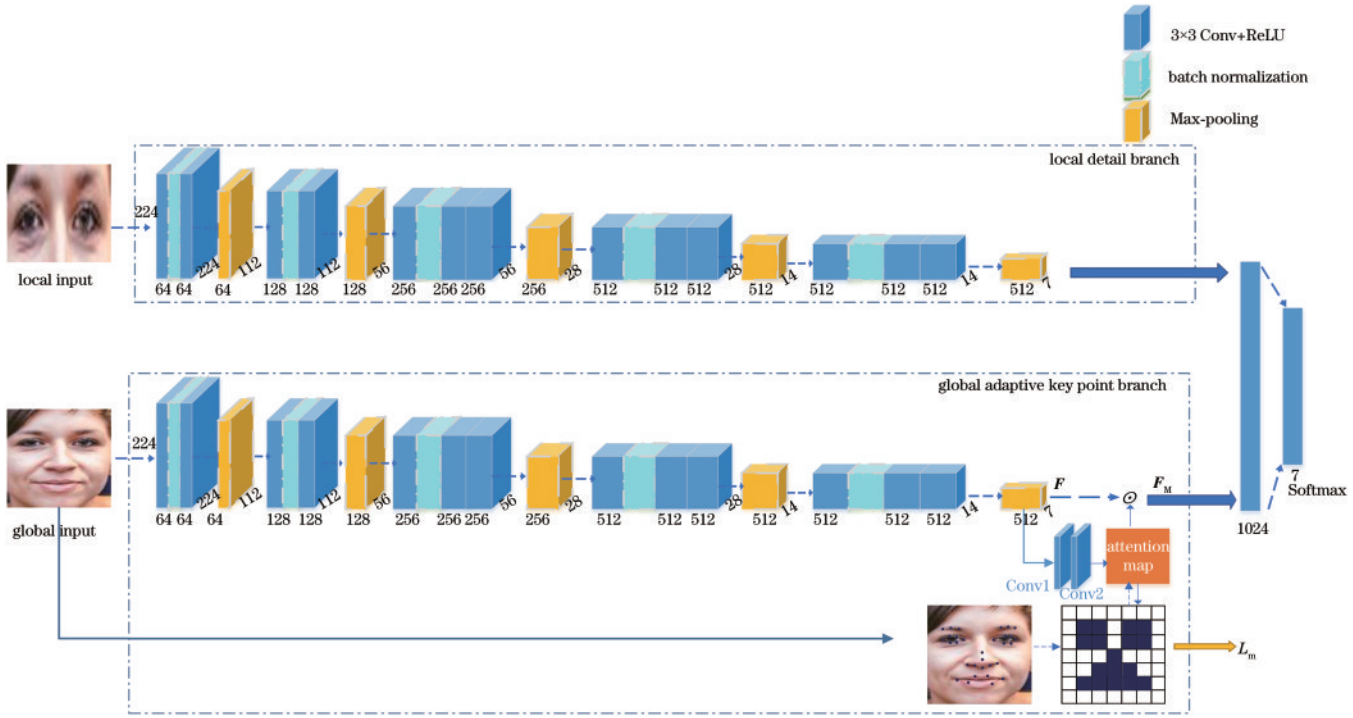

图 1　基于多区域融合的网络

Fig. 1　Multi-area fusion network

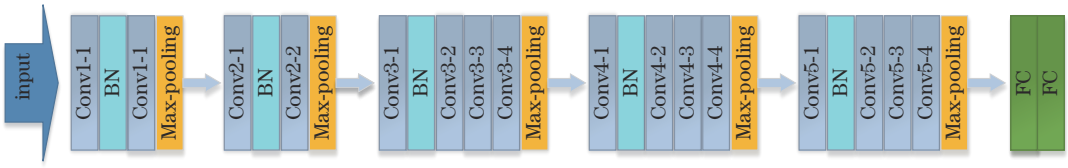

图 2　VGG-BN-16 网络结构

Fig. 2　VGG-BN-16 network structure

个维度下的输入，表示为

$$\hat{x}^{(k)}=\frac{x^{(k)}-E\left[x^{(k)}\right]}{\sqrt{\mathrm{var}\left[x^{(k)}\right]}},\tag{1}$$

式中：$x^{(k)}$表示神经元的激活值；$E\left[x^{(k)}\right]$表示每批训练数据$x^{(k)}$的平均值，$\mathrm{var}\left[x^{(k)}\right]$表示每批训练数据$x^{(k)}$的方差，经过变换后，表示成均值为0、方差为1的正态分布，目的是加快收敛速度，但这将导致网络表达能力下降。为了防止这种情况，对每个神经元都增加了γ和β调节参数，表示为

$$y^{(k)}=\gamma^{(k)}\hat{x}^{(k)}+\beta^{(k)},\tag{2}$$

式中：γ是可学习变换因子，表示为$\gamma^{(k)}=\sqrt{\mathrm{var}\left[x^{(k)}\right]}$；$\beta$是缩放因子，表示为$\beta^{(k)}=E\left[x^{(k)}\right]$。通过训练$\gamma$和$\beta$，可以恢复原始网络，学习特征分布，保持模型的表现力。在所提网络模型中加入BN层，可以提高网络收敛速度和泛化能力，并且也为剪枝提供条件。

2.2　局部细节分支

众所周知，表情是人面部中由五官表达出的情感，若将整张图像直接输入网络进行特征提取，无法聚焦于表情重点区域，如眼部、眉毛和嘴部。现有的基于深度学习的方法大多只专注于提取表情的高级语义概念，而忽略了局部面部区域的细粒度信息。本文提出一个单独的CNN分支来提取局部特征，强调了局部详细信息在表情分析中的重要性。

Fan等[17]提出了一种用于面部表情识别的多区域集成的CNN框架，旨在从多个人脸子区域捕获全局和局部特征来增强CNN模型的学习能力，虽然很大程度上关注表情部分区域，但很多区域被重复计算。考虑到计算量问题，并且Tang等[18]经过实验证明，在遮挡的情况下，鼻子对表情识别的影响最小，因此本文截取面部图像中的眼部和嘴部作为局部分支的输入。如图1所示，局部图像大小调整为224×224像素，每次将特定局部伴随整张面部图片输入到网络中，构成一个子网络。最后对每个子网络的表情预测分数进行加权和运算，得到最终的准确分类结果。

2.3　全局自适应关键点分支

若针对人脸面部有遮挡或头部姿态变化等复杂情况，当整幅图像输入网络时，直接提取全局特征会带入大量与表情无关的信息，因此全局自适应关键点分支

引入注意力图，对面部重要位置增加权重值，重点关注未遮挡五官区域，对其余区域进行抑制。

分支主要结构如图 1 所示，在 VGG-BN-16 网络的最后一个池化层后添加一个分支，用来生成注意力图，池化层后的特征表示为 $\boldsymbol{F}$。注意力图由两个 3×3 卷积生成，输出的注意力图 $\boldsymbol{M}_{\mathrm{H}}$ 可表示为

$$\boldsymbol{M}_{\mathrm{H}}=\mathrm{Sigmoid}\left[\boldsymbol{w}_2\tanh\left(\boldsymbol{w}_1\boldsymbol{F}+\boldsymbol{b}_1\right)+\boldsymbol{b}_2\right], \tag{3}$$

式中：$\boldsymbol{w}_1$ 和 $\boldsymbol{w}_2$ 分别代表图 1 中 Conv1 和 Conv2 的参数矩阵；$\boldsymbol{b}_1$ 和 $\boldsymbol{b}_2$ 分别代表 Conv1 和 Conv2 的偏置。

为了将注意力图聚焦于面部无遮挡的五官区域，利用面部关键点生成掩模，并将其作用于注意力图，使网络根据关键点自适应调节注意力图，从而更关注面部无遮挡关键区域。采取 Dong 等[19]提出的面部关键点检测器（SAN）进行面部关键点检测，面部关键点检测器已在 300W 数据集[20]上进行了预训练。检测器预测坐标和置信度得分，提取 68 个关键点，如图 3(a)到图 3(b)过程所示，只选取其中 31 个主要关键点。为了去除被遮挡的面部区域，设置一个阈值 I 来过滤置信度得分小于 I 的关键点。获取关键点的公式为

$$\left(c_i,c_j\right)=\begin{cases}\left(x_i,x_j\right), & s_{i,j}\geqslant I\\ 0, & s_{i,j}<I\end{cases}, \tag{4}$$

式中：$\left(c_i,c_j\right)$ 表示第 (i,j) 个关键点；x_i 和 x_j 表示关键点坐标；$s_{i,j}$ 表示置信度。根据过滤后的关键点，映射得到一个二维的面部掩模矩阵 $\boldsymbol{f}_{\mathrm{M}}$，表示为

$$f_{\mathrm{M}}(i,j)=\begin{cases}1, & (i,j)\in\left(c_i,c_j\right)\\ 0, & (i,j)\notin\left(c_i,c_j\right)\end{cases}。 \tag{5}$$

需要形成关键点到特征图的映射关系，此时考虑到掩模调节的注意图尺寸为 7×7，因此将掩模划分为 7×7 的网格状，掩模矩阵若值为 1，表示关键点在掩模中被激活，若为 0，表示此区域没有关键点。实现过程如图 3(b)到图 3(c)所示。$\boldsymbol{f}_{\mathrm{M}}$ 显示为一个 7×7 矩阵，有遮挡的关键点 $\left(c_i,c_j\right)$ 如图 4 所示，图 4(a)到图 4(b)表示掩模生成。

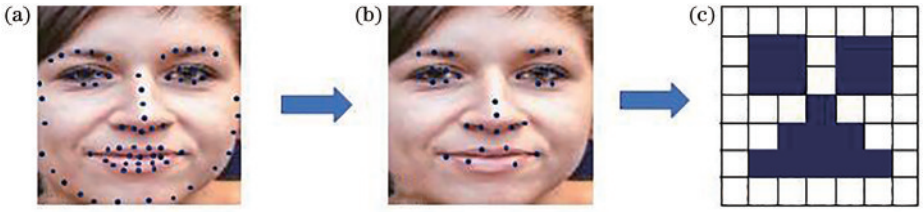

图 3　关键点掩模生成图。(a)68 个关键点；(b)31 个关键点；(c)掩模图

Fig. 3　Key point mask generation map. (a) 68 key points; (b) 31 key points; (c) mask map

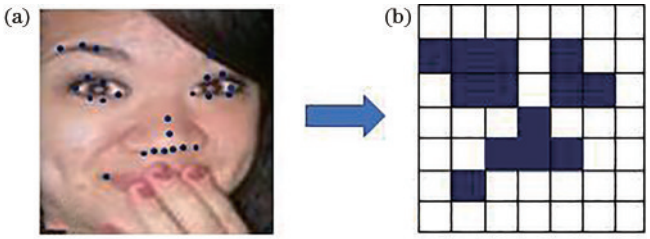

图 4　遮挡下的关键点及掩模生成。(a)关键点；(b)掩模图

Fig. 4　Key points and mask generation under occlusion. (a) Key points; (b) mask map

定义一个回归损失函数来实现调节注意力图的过程，可表示为

$$L_{\mathrm{m}}=\sum_{i=1}^{N}\left(\boldsymbol{M}_{\mathrm{H}}-\boldsymbol{M}_{\mathrm{H}}\odot\boldsymbol{f}_{\mathrm{M}}\right)^2, \tag{6}$$

式中：$\odot$ 表示两个矩阵的哈达玛内积。式(6)中注意力图 $\boldsymbol{M}_{\mathrm{H}}$ 无限接近于 $\boldsymbol{f}_{\mathrm{M}}$，实现对掩模中取值为 0 的位置抑制，也就是非五官区域，但没有增加未遮挡关键部位的权重，需要加大两部分权重差别，因此通过 $\boldsymbol{M}_{\mathrm{H}}\odot\boldsymbol{f}_{\mathrm{M}}$ 来抑制非关键点区域。$\boldsymbol{f}_{\mathrm{M}}$ 矩阵参数只包括 0 和 1，其中 1 表示关键点部位，使关键点部位得到增强，而 $\boldsymbol{M}_{\mathrm{H}}$ 矩阵其余数和 0 相乘结果为 0。然后将注意力图作用于特征 $\boldsymbol{F}$，得到输出特征 $\boldsymbol{F}_{\mathrm{M}}$，从而加强对未遮挡五官的关注，抑制其余区域的影响。

$$\boldsymbol{F}_{\mathrm{M}}=\boldsymbol{M}_{\mathrm{H}}\odot\boldsymbol{F}。 \tag{7}$$

对于一个批次，调节注意力图的总损失为

$$L_{\mathrm{M}}=\frac{1}{r}\sum_{1}^{r}L_{\mathrm{m}}^{r}, \tag{8}$$

式中：r 表示批次；L_{m}^{r} 表示第 r 批次的损失。

2.4　轻量级优化模型

目前 ResNet、VGG、Inception 等深层次网络架构能更好地解决人脸表情识别问题，同时也需要指数级大量数据集。现有的数据集大小有限，无法同时捕捉年龄、性别、种族导致的面部特征变化。因此，训练这些大小有限的数据集的深度架构需要大量的数据增强工作。构建一个小而高效的人脸表情识别模型也是目前研究的挑战。目前遇到主要的问题是模型本身的尺寸太大，运行时产生的参数也占用大量内存，模型中的计算操作量很大。为了达到模型轻量化，Liu 等[21]提出网络瘦身剪枝方法，该方法是一种通道级的稀疏化方法，能够自动识别并修剪掉一些对结果影响不大的通道，即给每个通道加上权重，在训练过程中使用 L1 正则化对通道权重进行稀疏化，然后对最终通道权重低于阈值的通道剪枝，微调后重新训练得到剪枝模型，来实现对网络模型的

轻量化。

剪枝具体方法如下：首先，为每个通道引入一个比例因子 γ，将其乘以对应的通道输出；其次，将这些比例因子与网络权重值相结合，删除权重值最小的通道；最后，剪枝掉 γ 值比较小的通道并且重新训练剪枝后的网络。由于比例因子和加权函数同时优化，网络能够自动识别并去除不重要的信道，对网络的泛化能力影响不大。目标函数定义为

$$L_s=\sum_{(a,b)}l\left[f(a,w),b\right]+\lambda\sum_{\gamma\in\Gamma}g(\gamma),\tag{9}$$

式中：a 和 b 为训练输入和输出；w 为可训练的权重；$g(\cdot)$ 为 L1 正则化，即 $g(s)=|s|$，会将 γ 稀疏到 0 和 1 附近。其中 γ 前向传播可以表示为

$$\gamma^*=\arg\min_\gamma\left\{\sum_{(a,b)}l\left[f(a,w),b\right]+\lambda\sum_{\gamma\in\Gamma}g(\gamma)\right\}。\tag{10}$$

优化式(10)，求梯度的公式为

$$\nabla\gamma^*=\nabla_\gamma\lambda\sum_{\gamma\in\Gamma}g(\gamma)。\tag{11}$$

式(9)的第一项是模型损失函数，第二项中 λ 是用来约束 γ 的超参。可直接对所提 VGG-BN-16 网络中的 BN 层进行权重稀疏化，不需要引入额外参数增加计算量，选取式(2)中 γ 为比例因子进行 L1 稀疏化，即 γ 越小，所对应的通道越不重要，每个比例因子都与特定的 CNN 卷积相关联，可以实现稀疏效应。并且 L1 正则化对精度损失影响很小。但即使有暂时的精度下降，通过微调剪枝，网络可以补偿精度。剪枝后的网络更窄，更紧凑。

2.5　表情分类

局部特征和全局特征一起送入一个全连接层进行特征融合，全连接层后接分类器，它将前一层的输出映射到表达式。选择 Softmax 作为分类器，公式为

$$f(t_j)=\frac{e^{t_j}}{\sum_{k=1}^{K}e^{t_k}},\tag{12}$$

式中：t_j 表示真实分类的值；$\sum_{k=1}^{K}e^{t_k}$ 表示某个样本所有分类值之和。对选取的两个子网络预测值进行加权和运算融合，两个子网络分别为眼部细节和全局网络、嘴部细节和全局网络，模型整体框架如图 5 所示。表情预测结果 $\boldsymbol{P}$ 可表示为

$$\boldsymbol{P}=\delta_1\sum_{m=1}^{M}\frac{1}{\sum_{k=1}^{K}e^{t_k}}\begin{bmatrix}e^{z_1}\\e^{z_2}\\\vdots\\e^{z_M}\end{bmatrix}+\delta_2\sum_{m=1}^{M}\frac{1}{\sum_{k=1}^{K}e^{t_k}}\begin{bmatrix}e^{z_1}\\e^{z_2}\\\vdots\\e^{z_M}\end{bmatrix},\tag{13}$$

式中：M 表示训练数据数量；δ_1 和 δ_2 分别表示不同子网络的权重，δ_1 和 δ_2 都取 0.5。

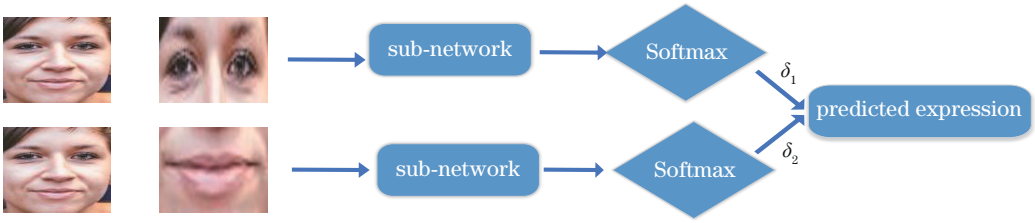

图 5　模型整体框架

Fig. 5　Overall framework of the model

3　实验分析

3.1　实验数据集

为了验证所提模型的有效性，选择公开的 RAF-DB 和 AffectNet 数据集进行实验验证。这两个数据集均是来自互联网的图像或视频片段，更适合在真实环境下对人脸表情识别问题进行研究。

RAF-DB 数据集[22]包括从互联网上下载的 29672 幅多样化人脸图像，是一个真实世界的情感数据集，每张图像大约有 40 个独立的标签，数据库中的面部表情图像在年龄、性别、种族、头部姿势、光照条件等方面都有所不同，符合现实生活中表情图像的特征。选择其中愤怒、厌恶、恐惧、高兴、悲伤、惊讶和自然这 7 类表情标签。本文选取了 16489 幅带有表情分类标签的图像进行实验，其中 13307 幅作为训练样本，3182 幅作为测试样本。

AffectNet 数据集[23]是最大的公开野外人脸表情数据集，由从谷歌网站搜索情感关键词得到的图像组成，总共收录了 100 多万来自互联网的图片，选择标注了 6 种基本表情类别（愤怒、厌恶、恐惧、高兴、悲伤和惊讶）和自然表情。本文选取了其中 235046 幅样本进行实验，其中 191109 幅作为训练样本，43937 幅作为测试样本。

3.2　实验预处理

一张图像不仅包含人脸还包括背景、姿态等干扰因素。要保证人脸尺寸、位置的一致性，首先需要对数据集中的图像进行预处理。预处理的主要步骤包括人脸检测、人脸对齐、图像尺寸归一化。MTCNN[24]用于检测所有图像的 5 个面部地标，通过仿射变换获得对齐后的人脸图像，并将图像大小调整为 224 × 224 像素。为了更准确提取面部未遮挡部分的关键点，选取 SAN 人脸关键点检测方法，首先在 300 W 数据集上进

行预训练，得到面部68个坐标，由此可以得到每个点的置信度。然后只选取其中31个关键点，再对每个点置信度信息的阈值进行取值。选取阈值为0.4、0.5、0.6进行实验，以RAF-DB数据集为例，图6展示了三个不同阈值下的实验精度，结果说明阈值取0.5时能获得更高精度。

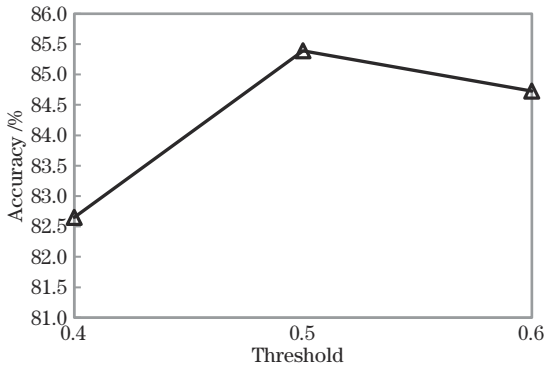

图6　不同置信度阈值下的精度

Fig. 6　Accuracy under different confidence thresholds

3.3　实验配置

所提模型是基于PyTorch深度学习架构的，运行在Windows10操作系统上，使用设备是NVIDIA Quadro RTX 6000 GPU。实验中采用随机梯度下降法(SGD)进行模型整体优化，动量设置为0.9，权值衰减为0.0005。对训练集图像进行随机裁剪和水平翻转，以缓解训练数据不足导致过拟合的问题，并对测试图像进行图像四角和中心切割。在RAF-DB数据集实验中，迭代次数设置为60，初始学习率为0.01，每20次迭代学习率衰减为之前的0.1。在AffectNet数据集实验中，迭代次数设置为20，初始学习率为0.01，每5次迭代学习率衰减为之前的0.2。两个数据集都设置128批次。

3.4　实验结果及分析

3.4.1　主流方法对比实验

表1显示了在两个数据集中所提方法和之前工作的对比结果，其中ResNet-50和VGG-16是基本分类模型。所提方法在RAF-DB数据集上的识别精确度是85.39%，在此数据集中，所提方法优于基本分类模型和其他先进模型；gACNN模型结合注意力机制设置一个门单元，以此权衡每个通道的重要性，并将全局和局部斑块表示相结合，所提方法的识别精确度比gACNN高0.32个百分点，证明所提方法重点选取的两部分的局部特征对细节的关注度是足够的；Wgan模型基于生成对抗网络修复面部被遮挡的图像，可以最大限度地还原图像，但修复受损图像需要了解遮挡物的准确位置和遮挡大小，所提方法的识别精确度比Wgan模型高1.9个百分点，表示对于人脸表情这种微妙的属性，所提细粒度表情识别更适用。

在AffectNet数据集上，所提方法的识别精确度是58.81%，比基本分类模型VGG-16高出7.7个百分点；IPFR模型是基于生成对抗网络的，同时探究姿势对表情的影响，考虑了多种姿态下的表情问题，如果只考虑AffectNet数据集中的样本，不引入额外多姿态样本，所提方法比该模型更具有鲁棒性，识别精确度高出1.41个百分点；SPA-SE模型采用滑动的方式得到面部细节，以此得到每个面部块的边缘信息，该模型为防止漏掉面部细节，提出不需要检测面部关键点信息的方法，但是表情具有一定的区域范围，不需要精确定位到面部各个关键点，因此所提方法重点关注表情相关区域，对面部表情的识别更有效，所提方法的识别精确度比SPA-SE模型高出0.67个百分点；SPWFA-SE在SPA-SE基础上增加了全局分支WFANet、SE和注意力模块，SE模块具有激励作用，自动学习每个特征通道的权值，结合注意力机制后可以减少冗余信息，突出面部特征，最后精度达59.23%，比所提方法高0.42个百分点，但加入SE模块后模型参数大幅增加，不适合后续的模型优化。

表1　不同人脸表情识别方法的性能比较

Table 1　Performance comparison of different FER methods

Method	Accuracy in RAF-DB /%	Accuracy in AffectNet /%
ResNet-50[25]	82.83	54.37
VGG-16[26]	80.96	51.11
pACNN[11]	83.05	55.33
gACNN[11]	85.07	58.78
E2-CapsNet[27]	85.24	
APM[28]	85.17	
DLP-CNN[29]	84.13	54.47
Wgan[30]	83.49	
IPFR[31]		57.40
SPA-SE[32]		58.14
SPWFA-SE[32]		59.23
Proposed method	85.39	58.81

为了更精准探究所提方法在不同表情下的表现能力，表2和表3分别是所提方法在两个数据集上的混淆矩阵。混淆矩阵详细地说明了每个表情识别精度和被误分类为其他表情的比例，其中对角线项表示对每个表情的识别精度。由表2和表3所知：在这两个数据集中，对高兴的识别率最高，都达到了90%以上，是因为高兴样本在两个数据集中个数都最多，其次是高兴表情幅度大，在7个表情中最易辨别；愤怒和惊讶识别率最低，愤怒最容易误识别为自然，惊讶最容易误识别为高兴。对于RAF-DB数据集，厌恶和恐惧样本量最少，识别错误率最高，其中厌恶识别错误率最高为自然表情，恐惧识别错误率最高为惊讶，这是因为这几类表情都有相似的外观特征。相比于RAF-DB数据集，AffectNet数据集中的表情识别率更低，由于AffectNet

表 2 基于 RAF-DB 数据集的人脸表情识别混淆矩阵

Table 2 Facial expression recognition confusion matrix on RAF-DB dataset

Expression	Angry	Disgust	Fear	Happy	Sadness	Surprise	Neutral
Angry	**0.81**	0.02	0.04	0.03	0.03	0.04	0.03
Disgust	0.06	**0.59**	0.02	0.08	0.09	0.03	0.11
Fear	0.02	0.04	**0.57**	0.03	0.14	0.15	0.05
Happy	0.00	0.01	0.00	**0.92**	0.02	0.01	0.04
Sadness	0.03	0.06	0.01	0.03	**0.82**	0.00	0.05
Surprise	0.04	0.01	0.01	0.02	0.01	**0.85**	0.06
Neutral	0.01	0.01	0.01	0.04	0.09	0.02	**0.82**

表 3 基于 AffectNet 数据集的人脸表情识别混淆矩阵

Table 3 Facial expression recognition confusion matrix on AffectNet dataset

Expression	Angry	Disgust	Fear	Happy	Sadness	Surprise	Neutral
Angry	**0.36**	0.24	0.02	0.13	0.07	0.03	0.15
Disgust	0.06	**0.57**	0.03	0.04	0.03	0.02	0.25
Fear	0.01	0.05	**0.49**	0.05	0.10	0.20	0.10
Happy	0.02	0.02	0.00	**0.90**	0.03	0.01	0.03
Sadness	0.03	0.06	0.02	0.04	**0.56**	0.02	0.27
Surprise	0.02	0.05	0.03	0.23	0.03	**0.43**	0.21
Neutral	0.01	0.04	0.02	0.13	0.04	0.05	**0.71**

样本量最大，选择于互联网，包含错误样本也较多，整体识别率下降。

3.4.2 消融实验

所提 VGG-BN-16 一共包括 2 个分支，分别是局部细节分支和全局自适应关键点分支，并改进了基础的 VGG-16 网络。为了明确模型各部分对分类性能的影响，以 RAF-DB 数据集为例进行消融实验，对提出的 7 种基本表情进行分类。

分别保留网络部分分支进行实验，如表 4 所示，包括 Local detail branch、Global branch、Global branch+attention map、Proposed method（VGG-16）、Proposed method。先以两个分支进行单独实验，再以不结合注意力分支的全局自适应关键点分支进行实验，其次结合两个分支选择未改进的 VGG-16 进行实验，最后与所提模型进行实验对比。如表 4 所示：与所提方法相比，单独选择局部细节分支的识别精度大幅下降，识别精确度只有 42.17%，由于局部细节只选择眼部和嘴部，如果样本头部有姿态变化或这两部分被遮挡，那不足以明确判断出表情类别；但单独选择全局自适应关键点分支的实验结果会更稳定，因为全局包含的表情信息要比局部包含的信息多，其次全局自适应关键点分支结合注意力分支的识别精确度比单独全局自适应关键点分支高 2.27 个百分点，证明用注意力图调节网络权重的正确性。全局和局部结合后实验精度提升很多，证明粗细粒度相结合的有效性，对基础分类模型 VGG-16 进行改进，以表情特质为重点，同时关注网络整体运行的性能，经过消融实验证明了所提模型各部分对表情识别的重要性。

表 4 RAF-DB 数据集上的消融实验结果

Table 4 Ablation experimental results on RAF-DB dataset

Method	Accuracy /%	Dataset
Local detail branch	42.17	RAF-DB
Global branch	78.26	
Global branch+attention map	80.53	
Proposed method (VGG-16)	84.59	
Proposed method	85.39	

3.4.3 轻量级实验

剪枝训练中需要确定两个参数，即修剪百分比和平衡因子 λ。修剪百分比决定对网络模型的修剪程度，分别取剪枝通道的 40%，50%，60% 作为剪枝百分比。式(9)中的 λ 作为超参数权衡损失函数与稀疏性之间的比例。λ 越大，比例因子趋于 0 的数量越多；当 $\lambda=0$ 时，相当于没有稀疏正则化，和原始网络结构没有区别；当 $\lambda=10$ 时，几乎所有的比例因子都会非常接近 0。Liu 等[21]在分类数据集上以基础网络进行大量研究，对于 VGGNet 网络，$\lambda=10^{-4}$ 时达到最好效果，对于所提 VGG-BN-16 网络，选择 $\lambda=10^{-4}$，可以保证适当数量标度因子接近 0。

以未进行剪枝的网络作为基准模型，选择在 RAF-DB 数据集上进行剪枝实验。剪枝过程是使模型朝着结构性稀疏的方向调整参数，来创建一个新的更薄、更紧凑的模型。通过微调来提升剪枝后的模型精度。表 5 为所提方法和 VGGNet 模型在不同剪枝百分比下的剪枝结果，其中 accuracy 表示不同剪枝百分比

下的表情识别精度，parameters表示参数量，pruned表示参数修剪率和浮点数修剪率。由于所提方法采用双分支，参数量要比基本分类模型VGGNet多，未剪枝的参数量有31049638，在两个模型中都表现出几乎相同的剪枝效果，只经过40%的剪枝，参数量就少于原模型的50%左右，在RAF-DB数据集的精确度只下降了0.61个百分点，浮点数也明显降低，意味着操作量减少；剪枝百分比为60%时，参数修剪率已达到70%以上，识别精度下降2个百分点～3个百分点，与基础分类模型的精确度不相上下，表明精度下降的范围仍在可接受的范围内，同时参数量和计算操作大幅度减少。

表5 两个方法的剪枝结果

Table 5 Pruned results of two methods

Method	Trim percentage	Accuracy /%	Parameters /10^6	Pruned (parameter) /%	Flops /10^9	Pruned (floating-point) /%
Proposed method	0	85.39	31.00		16.95	
	40%	84.78	13.77	55.58	11.64	31.33
	50%	84.19	11.48	62.97	9.91	41.53
	60%	82.43	8.26	73.35	8.11	52.15
VGGNet	0	80.96	20.48		14.26	
	40%	75.31	10.69	47.80	9.75	31.62
	50%	52.92	8.79	57.08	8.66	39.27
	60%	49.25	5.28	74.22	7.59	46.77

4 结 论

提出一种人脸表情识别方法，目的是提高复杂环境下的识别精度并优化模型参数。模型一共包括局部细节分支和全局自适应关键点分支，结合粗细粒度，更精确地提取表情微妙信息；并在全局自适应关键点分支提出注意力图以调整特征权重，用关键点生成掩模，辅助调节注意力图，目的是增大面部表情相关区域和其他区域的差别。提出改进后的VGG-BN-16，以提高学习率和网络的泛化能力，加快模型收敛速度；并结合剪枝方法，对模型的参数进行优化，得到更紧凑的模型。最后在公开野外环境下的两个数据集上进行实验，与部分主流方法和基本分类模型相比，所提方法取得了更高的识别精度。模型剪枝后，参数修剪率超过70%时，识别精度只下降2个百分点左右，达到了简化模型参数量同时保证识别精度的目的。

参 考 文 献

[1] Li S, Deng W H. Deep facial expression recognition: a survey[EB/OL]. (2018-04-18) [2021-02-05]. https://arxiv.org/abs/1804.06655.

[2] Wu T F, Bartlett M S, Movellan J R. Facial expression recognition using Gabor motion energy filters[C]//2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, June 13-18, 2010, San Francisco, CA, USA. New York: IEEE Press, 2010: 42-47.

[3] Russakovsky O, Deng J, Su H, et al. ImageNet large scale visual recognition challenge[J]. International Journal of Computer Vision, 2015, 115(3): 211-252.

[4] Hamester D, Barros P, Wermter S. Face expression recognition with a 2-channel Convolutional Neural Network[C]//2015 International Joint Conference on Neural Networks (IJCNN), July 12-17, 2015, Killarney, Ireland. New York: IEEE Press, 2015: 1-8.

[5] Hasani B, Mahoor M H. Facial expression recognition using enhanced deep 3D convolutional neural networks [C]//2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 2278-2288.

[6] Ji Y L, Hu Y H, Yang Y, et al. Cross-domain facial expression recognition via an intra-category common feature and inter-category distinction feature fusion network[J]. Neurocomputing, 2019, 333: 231-239.

[7] Baddar W J, Ro Y M. Bilateral hemiface feature representation learning for pose robust facial expression recognition[C]//2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), December 13-16, 2016, Jeju, Korea (South). New York: IEEE Press, 2016.

[8] Happy S L, Routray A. Automatic facial expression recognition using features of salient facial patches[J]. IEEE Transactions on Affective Computing, 2015, 6(1): 14948838.

[9] Majumder A, Behera L, Subramanian V K. Automatic facial expression recognition system using deep network-based data fusion[J]. IEEE Transactions on Cybernetics, 2018, 48(1): 103-114.

[10] Minaee S, Minaei M, Abdolrashidi A. Deep-emotion: facial expression recognition using attentional convolutional network[J]. Sensors, 2021, 21(9): 3046.

[11] 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.

[12] Gan Y L, Chen J Y, Yang Z K, et al. Multiple attention

network for facial expression recognition[J]. IEEE Access, 2020, 8: 7383-7393.

[13] Ma H, Celik T, Li H C. Lightweight attention convolutional neural network through network slimming for robust facial expression recognition[J]. Signal, Image and Video Processing, 2021, 15(7): 1507-1515.

[14] 申毫, 孟庆浩, 刘胤伯. 基于轻量卷积网络多层特征融合的人脸表情识别[J]. 激光与光电子学进展, 2021, 58(6): 0610005.
Shen H, Meng Q H, Liu Y B. Facial expression recognition by merging multilayer features of lightweight convolutional networks[J]. Laser & Optoelectronics Progress, 2021, 58(6): 0610005.

[15] 尹鹏博, 潘伟民, 张海军. 基于卷积注意力的轻量级人脸表情识别方法[J]. 激光与光电子学进展, 2021, 58(12): 1210023.
Yin P B, Pan W M, Zhang H J. Lightweight facial expression recognition method based on convolutional attention[J]. Laser & Optoelectronics Progress, 2021, 58(12): 1210023.

[16] Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift[C]//Proceedings of the 32nd International Conference on International Conference on Machine Learning-Volume 37, July 6-11, 2015, Lille, France. New York: ACM Press, 2015: 448-456.

[17] Fan Y, Lam J C K, Li V O K. Muti-region ensemble convolutional neural network for facial expression recognition[M]//Hammer B, Maglogiannis I. Artificial neural networks and machine learning. Lecture notes in computer science. Cham: Springer, 2018, 11139: 84-94.

[18] Tang H, Xiang J L, Wei H Y. Facial expression recognition based on double-channel facial images with robust occlusion[J]. Journal of Network Intelligence, 2021, 6(3): 606-623.

[19] Dong X Y, Yan Y, Ouyang W L, et al. Style aggregated network for facial landmark detection[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 379-388.

[20] Sagonas C, Tzimiropoulos G, Zafeiriou S, et al. 300 faces in-the-wild challenge: the first facial landmark localization challenge[C]//2013 IEEE International Conference on Computer Vision Workshops, December 2-8, 2013, Sydney, NSW, Australia. New York: IEEE Press, 2013: 397-403.

[21] Liu Z, Li J G, Shen Z Q, et al. Learning efficient convolutional networks through network slimming[C]//2017 IEEE International Conference on Computer Vision, October 22-29, 2017, Venice, Italy. New York: IEEE Press, 2017: 2755-2763.

[22] 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, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 2584-2593.

[23] Mollahosseini A, Hasani B, Mahoor M H. AffectNet: a database for facial expression, valence, and arousal computing in the wild[J]. IEEE Transactions on Affective Computing, 2019, 10(1): 18-31.

[24] 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.

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

[26] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2014-09-04)[2021-11-20]. https://arxiv.org/abs/1409.1556.

[27] Cao S, Yao Y Q, An G Y. E2-capsule neural networks for facial expression recognition using AU-aware attention[J]. IET Image Processing, 2020, 14(11): 2417-2424.

[28] Li Z Y, Han S Z, Khan A S, et al. Pooling map adaptation in convolutional neural network for facial expression recognition[C]//2019 IEEE International Conference on Multimedia and Expo, July 8-12, 2019, Shanghai, China. New York: IEEE Press, 2019: 1108-1113.

[29] Li S, Deng W H. Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition[J]. IEEE Transactions on Image Processing, 2019, 28(1): 356-370.

[30] Lu Y, Wang S G, Zhao W T, et al. WGAN-based robust occluded facial expression recognition[J]. IEEE Access, 2019, 7: 93594-93610.

[31] Wang C, Wang S F, Liang G. Identity- and pose-robust facial expression recognition through adversarial feature learning[C]//Proceedings of the 27th ACM International Conference on Multimedia, October 21-25, 2019, Nice, France. New York: ACM Press, 2019: 238-246.

[32] Li Y J, Lu G M, Li J X, et al. Facial expression recognition in the wild using multi-level features and attention mechanisms[J]. IEEE Transactions on Affective Computing, 2020, 1(1): 1-15.