引用格式:李剑鹏,苏楠. 基于局部空间特征引导的表情识别算法[J]. 信息对抗技术,2024, 3(1):46-56.[LI Jianpeng, SU Nan. Expression recognition algorithm guided by local spatial features[J]. Information Countermeasure Technology, 2024, 3(1): 46-56. (in Chinese)]

基于局部空间特征引导的表情识别算法

李剑鹏,苏 楠*

(清华大学电子工程系,北京 100084)

摘 要 面部表情识别在计算机视觉任务中受到越来越多的关注,由于真实场景中的表情往往包含着大量由姿态、年龄、图像质量、标注等因素带来的噪声,大大增加了类内变化,给表情的分类任务带来了很大的困难。现有的基于此类问题的研究往往聚焦于数据本身,通过对数据进行筛选或者扩大模型接受的数据类型的形式提高识别能力,没有考虑到卷积网络本身对图像特征关注的局限性。针对该问题,提出了一种基于局部空间特征引导的卷积神经网络,对于特征图的某部分像素点进行强调,引导卷积网络的深层特征图能够关注到多个对分类有效的局部面部区域,同时使用对数据重标记的形式抑制由标签错误导致的噪声问题。经过在多个公开的表情识别数据集中测试,并与多个同类方法对比,所提方法具有较好的识别效果。

关键词 面部表情识别;卷积神经网络;特征图可视化;空间特征聚合

中图分类号 TP 391.4 **文章编号** 2097-163X(2024)01-0046-11

文献标志码 A **DOI** 10.12399/j.issn.2097-163x.2024.01.005

Expression recognition algorithm guided by local spatial features

LI Jianpeng, SU Nan*

(Department of Electronic Engineering, Tsinghua University, Beijing 100084, China)

Abstract Facial expression recognition has received increasing attention in computer vision tasks. In real-world scenarios, facial expressions often contain a significant amount of noise introduced by factors such as pose, age, image quality, and annotation, which have greatly increased intra-class variation and have posed significant challenges for facial expression classification tasks. The existing researches addressing this problem often focus on the data itself, improving recognition capabilities by filtering or expanding the types of data accepted by the models, without considering the limitations of the convolutional networks in attending to image features. To address this issue, this paper proposed a convolutional neural network (CNN) based on local spatial feature guidance. It emphasizes certain pixels in the feature maps, enabling deep layers of the convolutional network to attend to multiple local facial regions that are effective for classification. Additionally, a re-labeling approach was employed to suppress noise caused by label errors. The proposed method was tested on multiple publicly available facial expression recognition datasets and has achieved better recognition performance compared to several existing methods.

收稿日期:2023-05-16 **修回日期**:2023-08-28

通信作者:苏楠,E-mail:sunan@tsinghua.edu.cn

Keywords　facial expression recognition; CNN; feature map visualization; spatial feature aggregation

0　引言

人类的情感识别一直是自然人机交互的重要组成部分。使用机器实现对人类情感的精准把握对于智能家居、智能驾驶等智能系统的发展具有巨大的应用价值。人脸作为人的情感最直接也是最重要的表达器官，正是情感认知问题的核心。由于图像信息本身容易采集的特性，基于人脸图像的情感认知研究，即人脸表情识别的研究最为广泛。进行表情识别工作之前，首先要通过情绪表达模型对表情进行编码。目前，常见的情绪表达模型主要有面部动作编码系统（facial action coding system，FACS）[1]、效价和唤醒值表示法（valence and arousal，VA）[2]以及离散情绪表示法[3]。其中，离散情绪表示法的使用最为广泛，对应的数据集种类也最多，例如：FERPlus[4]、ExpW[5]、RAF-DB[6]、Aff-wild[7-8]等。本文也主要基于此类表达模型进行表情识别方法研究。

人脸表情识别方法的工作流程基本可以分为人脸检测、特征提取和表情分类 3 个部分。人脸检测方法能够从图像中获取对应的人脸区域，该类方法已经有了较为成熟的算法实现，例如：多任务级联卷积神经网络（multi-task cascaded convolutional neural networks，MTCNN）[9]、Dilib[10]等，本文对此不进行过多讨论。得到处理好的人脸图像之后，通过特征提取器提取人脸区域与表情分类相关的特征，然后使用分类器对这些提取好的特征进行分类，得到合适的情绪类别输出，即完成了整个表情识别的过程。由于面部表情变化以及表意的复杂性，因此如何提取出合适的面部特征用于分类是一个较为困难的问题，也是本文的研究重点。

对人脸图像特征的部分提取，其本质是对输入的人脸图像进行一定的信息精炼和信息降维。从图像中提取出主要的表情特征，过滤掉非表情特征，从客观上降低图像信息的维度，为后续的分类工作提供方便。特征提取方法按照发展历程可以分为传统手工特征方法、基于机器学习与深度学习的方法 2 大类。传统手工特征方法往往使用一些工程性的特征算子来处理和分析人脸图像，例如：SIFT[11]、HOG[12]、LBP[13]、Gabor 小波系数[14]等，这些方法通常基于人的主观先验知识设计具有特殊结构的局部算子作为滤波器，可以有效地提取出想要的图像信息，而且不需要太多的训练数据，往往具有较小的计算开销和较高的计算效率。但是，传统方法的缺点在于需要大量的专业领域知识与人工设计，例如：对于不同的问题，往往难以复用，需要重新设计和调整；对于比较复杂的图像问题和大规模的图像数据，往往会遇到困难，难以处理。基于深度学习的方法能够自适应地精准提取任务需要的图像特征，相较于传统手工特征方法具有更好的泛化性，对于一些具有挑战性的真实场景具有更好的效果。FASEL[15]发现卷积神经网络对人脸的姿势和尺度变换具有良好的鲁棒性。LIU 等[16]提出了一个基于面部动作单元的卷积网络框架用于人脸识别任务。深度卷积网络能够多尺度地提取特征，对面部特征的表达更加准确。在 2013 年举办的 FER2013 和 Emotiw2013 表情识别挑战中，TANG[17]和 KAHOU[18]分别使用深度卷积神经网络进行特征提取并获得了挑战的最优效果。尔后一些经典的深度网络结构陆续出现，例如：AlexNet[19]、VGG[20]、GoogleNet[21]、ResNet[22]等，这些深度网络结构具有传统方法难以达到的图像特征提取能力，在很多情况下已经足以满足任务的需求。

但是面对人脸表情特征提取这个复杂问题，经典的深度网络往往只能关注到面部某块主要特征区域，因而会丢失很多有效信息，并且由于图像本身的问题（比如遮挡）导致该区域特征不明显，还会产生很明显的错误。从局部特征入手是一个较为直观的思路。LI 等[23]提出了 Multiple CNNs 方法，在多个局部面部区域分别训练 CNN 网络来实现对局部面部特征的关注。WANG 等[24]使用区域关注的注意力网络来解决姿态变化与面部遮挡的问题。然而，这些基于局部人脸区域的方法通常丢失了人脸的全局信息，为了补充全局信息，TBE-CNN[25]方法通过共享底层和中层特征将全局脸和多个局部面部区域的网络整合到一个模型中，实现同时对全局与局部信息的关注。但是这类方法缺乏灵活性，无法

增强可识别性区域的重要性，也无法抑制信息量较小的部分和噪声信息，并且由于涉及人脸分割，非常依赖面部关键点检测的准确率。XUE 等[26]和 PHAN 等[27]通过改进的 Transformer 网络实现了对多个人脸局部区域特征的密切关注，但是此类方法对数据量敏感且参数量较大，与深度卷积网络相比在实现同等性能的情况下需要较高的计算代价。

针对以上这些问题，本文提出了一个基于 ResNet 网络的具有局部空间特征引导的表情识别算法，主要贡献如下：

1）提出了局部空间信息引导的卷积神经网络，通过多个并行的局部空间信息聚合网络引导网络关注不同的面部区域，获得区分度更高的特征图输出。

2）针对“野外”真实人脸表情数据集，同时关注了数据集噪声的影响与模型特征本身的区分度问题，在使用再标记方法抑制数据带来的不确定性的同时，通过局部空间特征强调网络增强模型对面部细节的关注能力，从数据与表情特征的表征 2 个方面增强算法的识别能力。所提出的算法经公共表情识别数据集 RAF-DB 与 ExpW 的测试，展现出了很好的表情识别效果。

1 相关工作

经典的深度卷积网络在处理表情识别问题时往往只能关注到一个最为明显的面部特征区域，因此，由于图像本身的质量因素导致这个代表性区域的特征不够明显时，就会出现误判。DENG 等[28]基于 ResNet-50 网络用知识蒸馏的思路学习采用不同情绪表达模型的情感数据集的数据，通过扩充模型认知的信息类型来提高识别性能。WANG 等[29]提出了自修复网络（self-cure network，SCN），基于 ResNet-18 网络在训练过程中利用网络中预训练获得的先验知识对训练数据进行判别与校正。这 2 种方法在当前的表情识别数据集上拥有不错的识别表现，但是都是从训练方法与数据的层面去弥补经典网络在识别能力上的缺陷，本质上没有对特征提取网络进行改进。为了进一步提高算法的识别能力，需要加强对局部区域特征的关注，使网络关注到更多的信息。一个直观的思路是进行面部分割，然而基于面部分割的方法只是通过主观经验将人脸分块提取特征，并不能强调对于特定分类任务有效的区域，也不能抑制无效信息的影响，并且分割本身较为依赖人脸关键点检测算法的准确性。基于与 SENet 类似的思路，LS-CNN[30]提出空间信息聚合网络（local aggregation network，LANet）强化人脸的局部空间信息，通过串联的 SENet 与 LANet 分别对特征图中有用的空间信息与通道信息进行强调，并使用 Inception 网络学习多尺度特征，在人脸识别任务中取得了良好的效果。单一的 LANet 虽然能在一定程度上起到拓展模型对局部关注的作用，但是受噪声影响较大，即使拓展了特征图的关注区域，有时也会引入错误的信息。

为了尽可能关注到更多的局部特征并保证这种区域强调的鲁棒性，本文采用了将多个 LANet 并行连接，并将它们的输出合并成一张特征图，共同作为最终的空间信息强调特征图的思路。为了保证这些并行的 LANet 能够分别学习到不同的局部特征，使用了多注意力随机丢弃机制（multi-attention dropping，MAD）[1]，在训练过程中随机将某个分支的特征图置为 0，这样就可以引导各个分支探索多样化的面部区域。

2 基于局部空间特征引导的深度卷积神经网络

本文着眼于卷积神经网络特征图的关注点，采用 LANet 强化模型对于局部空间信息的关注，扩大卷积网络深层特征图的关注区域，以增强网络的识别能力。由于 SCN 方法[28]对噪声处理的优秀性能和良好的可迁移性，以及应对大型人脸表情数据集中由标注错误引起的不确定性问题的需求，因此使用其中的再标记形式对主干网络进行训练，以解决大型人脸数据集中普遍存在的图像质量低以及标记错误的问题。

2.1 整体结构

模型的整体结构如图 1 所示。网络整体采用 ResNet-50 作为骨干特征提取模块，在其之后以多个并行的 LANet 作为特征强调模块，用于聚合空间信息；采用 7 类离散情绪模型作为情绪表达的方式，使用全连接网络与 Softmax 作为分类输出层。此外，由于大量的公共人脸表情数据集要参与模型训练，因此使用 SCN[1]方法中的再标记模块来抑制噪声问题。

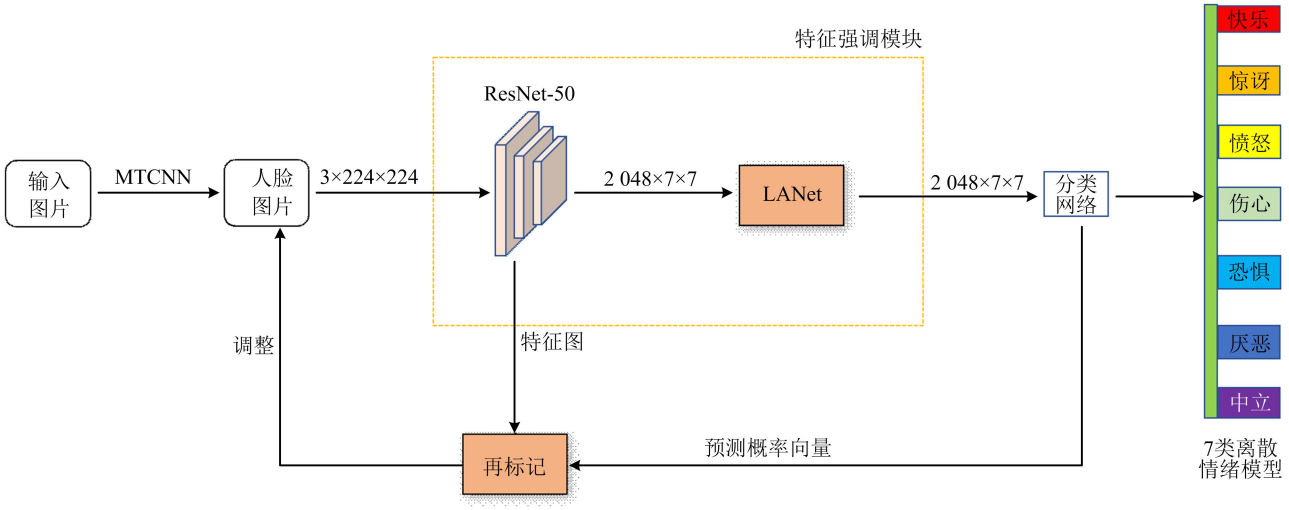


图 1　模型整体结构

Fig. 1　Structure of the model

2.2　特征强调模块

对卷积神经网络而言，图像的面部特征会在识别的过程中被自动捕获，但是如果没有加以引导，那么网络往往就不能够关注到所有的可判别的面部特征，而会把关注点集中在某个区分度最高的区域中。如果这个区域被遮挡或者这个局部区域的收敛并不是一个全局最优的选择，那么网络的识别能力就会受到明显的影响。特征强调模块的主要作用是尽可能地指导网络去学习并提取不同的局部特征，避免将主要关注点集中在一个点，其整体结构如图 2 所示，其中，h，w，c 分别表示特征图的长、宽和通道数。

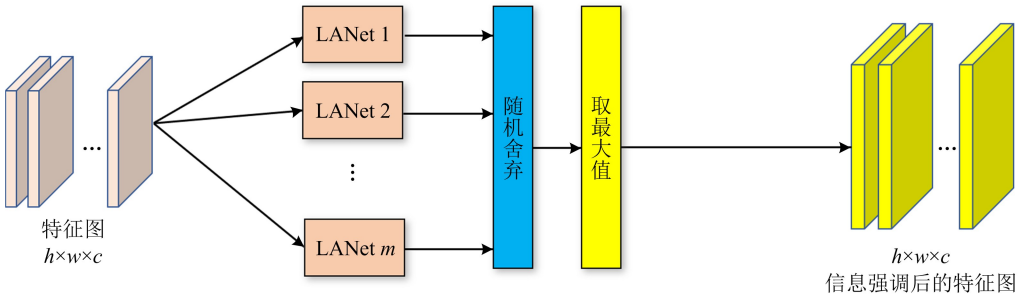


图 2　特征强调模块网络结构

Fig. 2　Network structure of feature emphasis module

特征强调模块是基于对空间信息进行聚集的 LANet 网络实现的，其具体结构如图 3 所示。它采用 2 个连续的 1×1 卷积将各个通道的空间信息分 2 步汇总到一个通道中，得到一个信息富集的单通道特征图，将这个特征图作为权重图赋予原本的输入特征图得到最终的输出。

图 3 中，r 表示第 1 个卷积所带来的通道数减少率，并且在此卷积层后跟随有一个 ReLU 激活层，另一个输出为 1 通道的卷积层的输出激活采用 Sigmoid 函数，输出的特征图为空间注意力权重。由于输入特征图的每个像素单元都对应原本输入图像的一个区域斑块，因此信息量更大的局部区域会获得更高的关注度，即有更大的权重值，而不太重要的区域则会被赋予较低的权重。由图 3 可以看出，LANet 网络的输出与输入特征图的大小相同，因此该模块可以较为容易地插入到不同的网络结构中去。

为了使网络尽可能地关注到更多的局部区域，本文使用了 MAD[26]。该机制的思想类似于 Dropout，以特征强调模块为例，由于使用了多个并行的 LANet，在训练过程中，为了使不同的 LANet 学习到不同的参数，每次训练都会随机舍弃其中的数个特征图，然后对剩下的特征图以按

像素取最大值的形式合并，保留各个分支的强调信息。因此，不同的分支在反向传播的作用下就会开始自主学习并关注不同的局部区域，对卷积网络的特征提取进行引导。定义 LANet 分支的个数为 m，输入的特征图大小为 $h\times w$，每个分支输出的特征图为 M_i，则特征强调网络的输出 M_{out} 可以表示为：

$$M_{out}(x,y)=\max\{\mathrm{MAD}(M_1(x,y),M_2(x,y),\cdots,M_m(x,y))\} \quad (1)$$

式中，$1\leqslant x\leqslant w$，$1\leqslant y\leqslant h$。

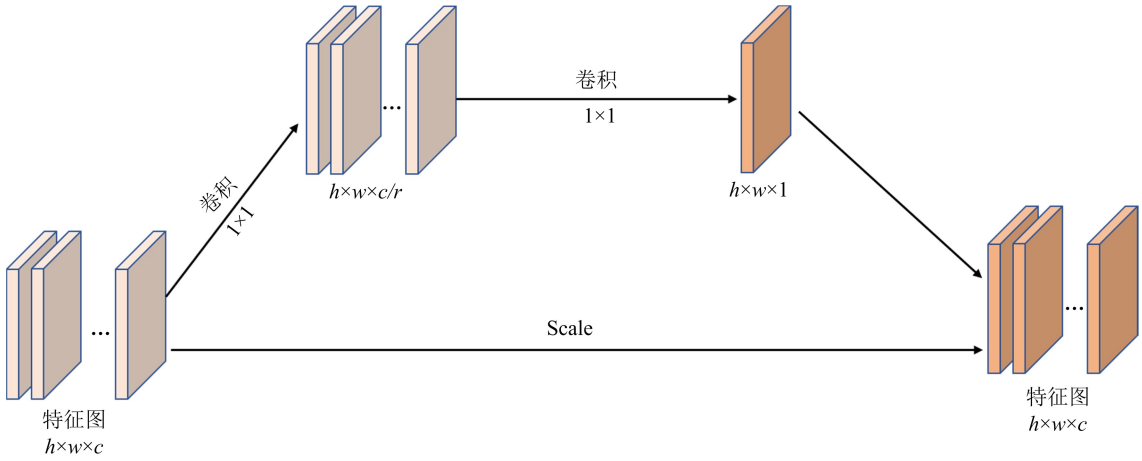


图 3 LANet 网络结构

Fig. 3 Network architecture of LANet

3 实验及结果分析

为了测试模型的识别效果，在多个公开的数据集中进行测试，并与当前的一些表情识别算法进行了比较。

3.1 数据预处理

模型的训练基于目前的公开表情数据集进行，为了尽可能削弱单一数据集有可能存在的标注偏见，本文使用多个数据集混合进行训练，具体如下：

1) Aff-wild/Aff-wild2 数据集[7-8]。Aff-wild 数据集是由 106 个平均包含 507.208 帧的视频数据构成的“野外”型（即非实验室环境中拍摄获得）人脸表情数据集。数据集中包含大约 150 张不同身份的人脸。标注形式为效价与唤醒值的形式，并且数据集中的情绪具有良好的分布。但是由于标注全部为同一人完成，因此对于情绪的判断可能存在一定的主观性。Aff-wild2 数据集在原本数据集的基础上进行了扩充，同时，对于数据的标签形式也在原本的效价和唤醒值模型的基础上扩充了面部活动单元模型与离散情绪模型 2 种不同类型的标签。

2) ExpW 数据集[5]。该数据集是一个包含了 91 793 张用 7 种离散情绪模型标记的人脸表情图像的表情识别数据集。该数据集创建的目的是研究“野外”图像中基于心理学的高级情绪关系的表征和量化。

3) FER2013 与 FER+数据集[4]。FER2013 数据集包含大小限制为 48×48 的大约 30 000 张不同表情的面部灰度图像，图像经过了初步的面部检测和对齐，标签方式采用 7 种离散表情表示法。FER+数据集对 FER2013 数据集进行了重新标注，将原本的 7 种表情的标注扩充为包含轻蔑、未知和非人脸 3 种新标签的 10 种离散情绪表示。

4) 真实世界面部情感数据集(RAF-DB)[6]是一个包含从互联网中下载的约 30 K 面部图像的大型面部表情数据集。数据集包含 2 种不同类型的子集，即含有 7 种离散情绪标签类型的单标签子集和含有 12 类复合情绪标签的双标签子集。数据集的标记采取众包的形式进行，每个图像都由大约 40 名标记者单独注释。数据集中的图像在年龄、性别、种族、头部姿势、光照、遮挡以及人工后期处理（比如滤镜）等方面存在很大差异，具有较大的包容性。

要同时使用不同的数据集首先需要对标签进行统一，虽然本文中选用的几个数据集都是基本情绪类分类，但是其中的情绪种类与编号并不相同，因此本文中统一采用的标签形式为：中立、愤怒、厌恶、恐惧、快乐、伤心、惊讶分别对应 0～6 的数字。将原本数据集的标签形式统一建立到

新标签形式的映射。

完成标签统一工作后，还需要对训练数据的数据类分布进行均衡处理。原本的各个数据集中存在分类不均衡的情况，而且类别分布情况较为相似，若将它们直接混合，则会加重类别分布的不均衡，直接影响到模型的训练结果。将不同数据集按照各个分类的数据量进行采样后混合，得到了情绪类别分布相对均衡的训练数据集。Aff-wild2、ExpW、FER＋和 RAF-DB 4 种数据集的原本类别分布如图 4 所示。

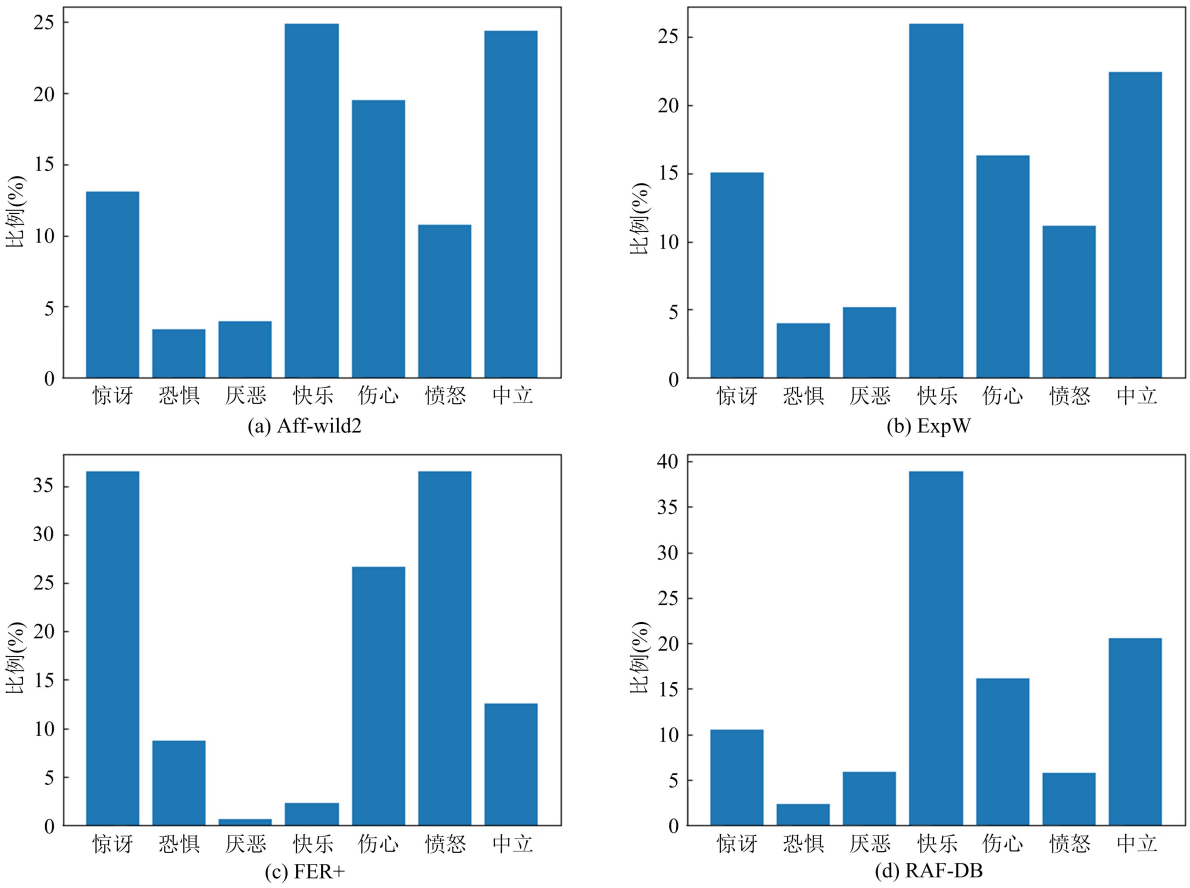


图 4　4 种数据集数据分布对比

Fig. 4　Data distribution comparison of the four dataset

3.2　模型训练

为了应对数据集的噪声，模型在训练过程中采用了 SCN 中再标记的方法对训练数据重新标记以实现噪声抑制。再标记模块是为了处理公开数据集中广泛存在的不确定性，包括模糊图像、遮挡图像等低质量图像带来的噪声以及由标注者的主观性和不专业性引起的标注错误产生的噪声。对于低质量的图像数据，通过人脸检测方法可以进行一定程度的筛选，但是，相对难以解决由标记者的主观认知带来的图像标签与事实相违背的问题，将不同数据集中的数据进行混合可以在一定程度上减轻该问题的影响，但是并没有解决问题。

再标记模块利用网络中预训练学习到的先验知识对输入数据的标签的正确性进行判别，对大概率标注错误的图片重新标记，有效抑制了噪声的影响。再标记模块的整体结构如图 5 所示。

该模块在整体识别模型的输出层增加了一个独立的全连接层用来学习一个自适应权重，该权重可以捕捉训练过程中各个样本的贡献，不确定性较高的样本将会拥有较低的贡献值。由图 5 可以看出，自适应权重由骨干网络中提取的特征通过一个全连接层与 Sigmoid 激活函数构成。自适应权重可以表示为：

$$\alpha_i = \sigma(\boldsymbol{W}_a^{\mathrm{T}} x_i) \tag{2}$$

式中，α_i 表示第 i 个样本的权重，$\boldsymbol{W}_a^{\mathrm{T}}$ 表示对应的全连接层的权重。对于每一批输入图片，先通过骨干网络进行特征提取得到特征 $F_1, F_2, \cdots, F_n$，进而计算得到权重 $\alpha_1, \alpha_2, \cdots, \alpha_n$。自适应权重与

分类层的输出相乘，经过 Softmax 模块得到最终的分类概率向量。因此，自适应权重可以衡量训练过程中样本的重要性。于是通过权重 α_i 可以将每个训练批次的输入数据分为低质量类与高质量类 2 个类别，对于被分类为低质量类的图片，就需要进行数据重标记。再标记的过程是利用网络中学习到的知识逐渐对数据中的噪声进行矫正的过程。重要性低的权重值样本往往具有较高的分类不确定性，这代表其所对应的样本较大可能是噪声样本。在训练过程中，如果能够制定一个策略去抑制这些不确定性样本发挥作用，那么就可以达到抑制样本噪声的效果。因此，权重大小衡量了训练过程中样本的不确定性，并在此基础上，通过赋予它们一个正确率更高的新标签，本文分离出了需要处理的部分低质量数据。

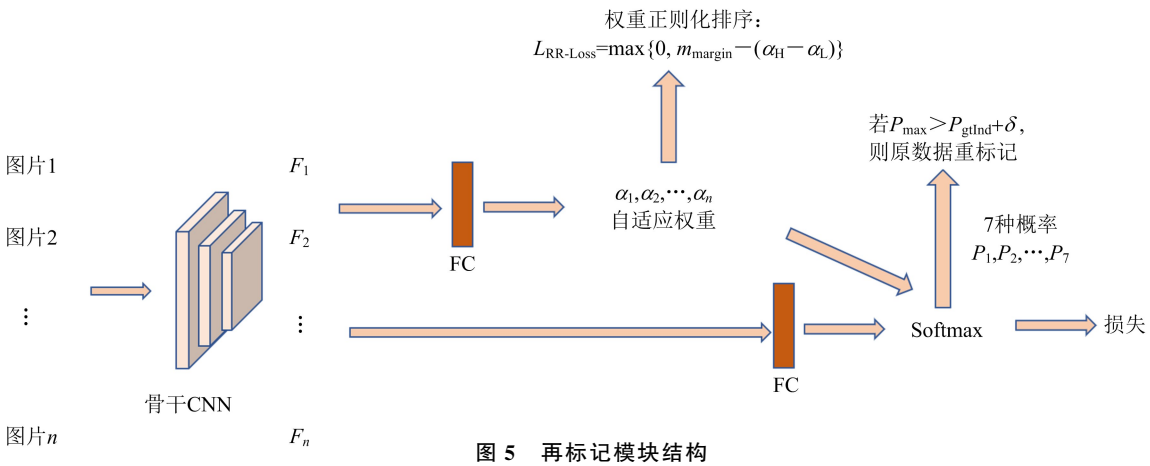


图 5 再标记模块结构

Fig. 5 Network architecture of relabel module

分级正则化损失(rank regularization loss，RR-Loss)确保了高质量权重的数据分组始终比低质量权重的分组大固定的阈值，保证了分组的区分度。正则化损失表示为：

$$L_{\text{RR-Loss}}=\max\{0, m_{\text{margin}}-(\alpha_H-\alpha_L)\} \tag{3}$$

式中，m_{margin} 为设定好的超参数，也可以设置为可学习参数，α_H 与 α_L 分别为高重要性权重的均值与低重要性权重的均值。

对于需要重新标记的低重要性图片，需要去比较估计的概率向量中最大值的类别是否与原本标签类对应的类别的概率差大于某个阈值。若是，则用估计概率最大值对应的类别标签替换原本的类别标签，可表示为：

$$P_{\max}=P_{\text{gtInd}}+\delta \tag{4}$$

式中，$P_{\max}$ 表示模型输出估计的最大类别的概率，P_{gtInd} 表示图像标签对应的表情分类的估计概率，δ 表示设定的阈值。式(4)表明在再标记的过程中使用了网络中学习到的知识来对每个样本的标签进行实时预测，因此本文往往在将网络训练了 15～20 个 epoch 之后再启动再标记模块，以保证整体模型学习到足够多的可用于矫正的知识。

多分类任务的损失函数采用加权的交叉熵损失函数(logit-weighted cross-entropy loss，WCE-Loss)，可表示为：

$$L_{\text{WCE-Loss}}=-\frac{1}{N}\sum_{i=1}^{N}\lg\frac{e^{\alpha_i \boldsymbol{W}_i^{\mathrm{T}} x_i}}{\sum_{j=1}^{C} e^{\alpha_i \boldsymbol{W}_j^{\mathrm{T}} x_i}} \tag{5}$$

式中，$\boldsymbol{W}_j^{\mathrm{T}}$ 是第 j 个分类的权重，C 代表类别数。

于是，模型的整体损失函数可以表示为：

$$L=(1-\lambda)L_{\text{WCE-Loss}}+\lambda L_{\text{RR-Loss}} \tag{6}$$

式中，λ 为权重参数。

3.3 参数设置

模型的输入图像数据由 MTCNN 算法进行人脸检测与配准，并调整为 224×224 的分辨率；骨干网络 ResNet-50 由 Pytorch 工具箱实现。在视觉感知任务中，网络中学习到的某些特征可以在相关的任务之间进行转移。WANG 等[29]证明了使用在人脸识别任务中预训练过的网络可以大大加快表情识别任务的收敛速度并达到更高的识别准确率。因此，特征提取的骨干网络在默认情况下采用经由 MS-Celeb-1M 人脸识别数据集进行预训练过 ResNet-50 网络，面部特征从最后一个池化层中提取。算法的训练在 2 个

RTX2080Ti 显卡上完成，设置 batch size 大小为 256。在每次的迭代过程中，用于噪声抑制的再标记模块会将训练数据分为低重要性和高重要性 2 组，分组的比例为 7∶3。文献[28]中的方法通过实验证明，再标记模块中用来计算正则化排序损失的参数 m_{margin} 为 0.15、再标记阈值参数 δ 为 0.2 时具有最好的识别效果，因此本文采用相同的参数设置。整个网络的优化过程由 RR-Loss 与 WCE-Loss 共同维护，经过实验，损失函数权重 λ 取 0.5，即二者的比例为 1∶1。经测试，将特征强调模块的分支数 m 设定为 3，再标记模块从训练的第 15 个 epoch 处开始进行优化。初始学习率设置为 0.1，在第 20 个和第 40 个 epoch 后分别将被除以 10，在第 60 个 epoch 后停止训练，优化器采用 Adam(adaptive moment estimation)。

3.4　实验对比

为了验证算法的训练效果，分别在公共数据集 ExpW 与 RAF-DB 上进行测试，并与多标签知识蒸馏网络(MER-IL)[29]、自修复网络(SCN)[28]、共享表示集成卷积网络(ESRs)[31]以及基于 Transformer 的多模态融合表情识别(TMIF-FEA)[27]等方法进行对比，结果见表 1 所列。

表 1　不同数据集中识别准确率(%)

Tab. 1　The recognition accuracy of different datasets(%)

方法	数据集	
	RAF-DB	ExpW
ESRs[31]	85.90	65.24
MER-IL[29]	86.15	67.29
SCN[28]	87.03	68.82
TMIF-FEA[27]	88.91	70.46
本文方法	87.74	70.59

可以看出，在 ExpW 数据集上，本文方法比其他方法具有更优秀的识别效果，但是在 RAF-DB 数据集上略逊于 TMIF-FEA 方法。ExpW 数据集中存在着大量噪声数据，而由于再标记模块的存在，本文方法对于这类噪声具有良好的抗性，因此取得了较好的识别效果。而在 RAF-DB 数据集上，本文方法虽然没有取得最优的识别效果，但是事实上，本文方法在参数量远小于 TMIF-FEA 方法的基础上，达到了相近的结果。

参数量的对比见表 2 所列。本文方法主要由骨干网络、特征强调模块以及再标记模块构成，而再标记模块实际上体现在模型中的只是一个额外的全连接层，对参数量的影响很少，主要参数量由骨干网络 ResNet-50 以及特征强调模块提供，特征强调模块主要由 1×1 卷积模块与全连接网络构成，参数量也相对较小。而 TMIF-FEA 方法由于使用了 Transformer 模块，即使骨干网络的参数量较小，总体参数量也远多于本文方法。因此，本文方法更加轻量化，在低参数量的情况下取得了较好的识别效果。

表 2　不同方法参数量对比

Tab. 2　Comparison of parameters in different methods

方法	ESRs[31]	MER-IL[29]	SCN[28]	TMIF-FEA[27]	本文方法
参数量(M)	54.3	30.6	27.4	46.0	28.2

3.5　消融实验

本节将分别测试 ResNet-50 骨干网络、ResNet-50 骨干网络＋噪声抑制模块、ResNet-50 骨干网络＋特征强调模块以及整体网络 4 种情况在不同测试集中的识别结果并进行对比，验证各个模块对最终识别结果的有效性，结果见表 3 所列。

表 3　不同网络结构下的识别准确率(%)

Tab. 3　The recognition accuracy of different network structures(%)

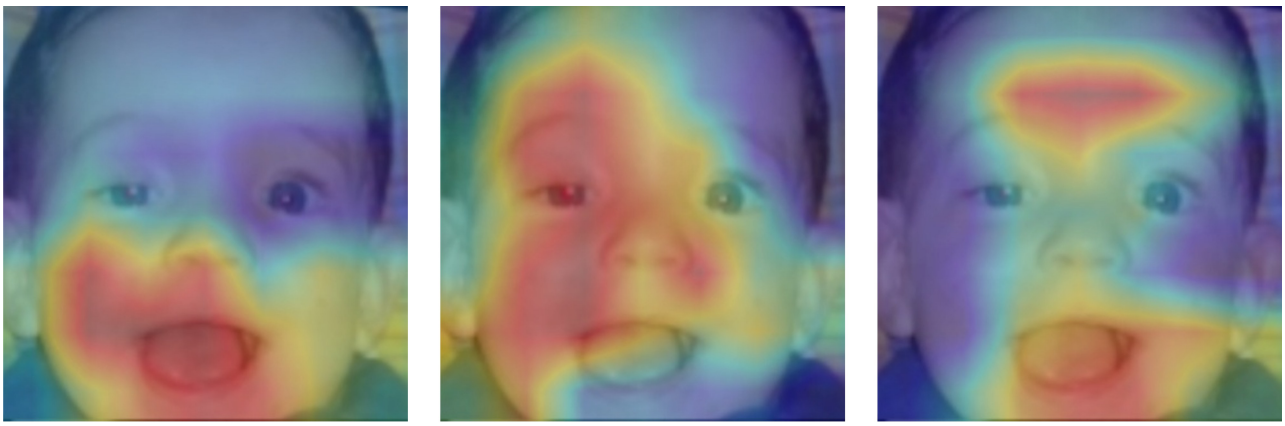
方法	数据集	
	RAF-DB	ExpW
ResNet-50	84.20	65.24
ResNet-50＋噪声抑制	87.03	68.67
ResNet-50＋特征强调	85.14	65.85
ResNet-50＋特征强调＋噪声抑制	87.74	70.59

从结果可以看出，与骨干网络相比，噪声抑制模块确实对于“野外”型数据集中存在的大量噪声起到了抑制作用，对于骨干网络在混合的测试集中的识别效果具有显著提升。而特征强调模块也符合预期地起到了良好的信息聚集与强调作用，二者共同作用下达到了最好的识别效果。

为了更加清楚地展示特征强调模块产生的效果，本文使用 CAM[32]工具对卷积网络最终特征层输出的特征图进行可视化，结果如图 6 所示。

图 6 中将多种情况的特征图做了可视化对比，图 6(a)～(c)分别为不使用特征强调模块、使

用单一 LANet 模块、多个 LANet 并行模块。可以看出，在不使用特征强调模块的情况下，原本骨干网络的特征图的关注点只集中在嘴巴这一个主要区域，而在加入 LANet 模块之后，网络的关注点明显变多了，一些次要的判别区域也被纳入考虑的范围，比如眼睛。但同时也使得一些不重要的面部区域的权重虚高，从而也被纳入关注范围。采用 MAD 训练的多个并行的 LANet 分支，由于在每次训练时都会舍弃随机分支，使得每个分支的 LANet 网络可以自由探索多样化的可辨别的面部区域特征，这在实现了多个局部关注点的同时也一定程度上抑制了单个 LANet 时出现的错误强调的问题，最终呈现的结果就是热力图上出现了一个新的高光区域。而单一通道与多分支的 LANet 的热力图可视化结果看起来似乎有所矛盾，这是因为使用单一分支训练的模型对于复杂的人脸五官变化的处理能力不足，容易出现错误强调的情况，反映在单一样本上就是如图 6(b)中所示的不符合直观的关注区域，而 MAD 引导的多分支训练大大增强了对不同局部特征的关注能力，具有更强的分析能力，因此关注点也更加准确。



(a) 不使用特征强调模块　(b) 使用单一LANet模块　(c) 使用多个LANet并行模块

图 6　热力图可视化对比结果

Fig. 6　Heat map visual comparison results

可视化对比结果证明了特征强调模块对于卷积网络特征图关注点的影响，即确实能够起到引导网络关注到不同的局部面部区域的作用。

此外本文还讨论了信息聚合网络的分支数 m 以及不同的损失函数权重 λ 对于模型识别效果的影响。表 4 列出了在 RAD-DB 数据集上不同的权重系数 λ 下模型识别准确率的变化。

从表 4 中可以看出，在 2 种损失函数取相同比例时，模型有最好的学习效果；当 $\lambda>0.5$ 时，学习效果出现了明显的下降，说明交叉熵损失在学习任务中的重要性相对更大。

表 5 展示了不同的分支数 m 对于识别效果的影响。本文设置了 0～7 这 8 个梯度进行实验。从表中可以看出，在 $m=3$ 时，模型达到了最好的性能；当 $m>5$ 时，性能开始逐渐下降。分支数较小时，模型难以具备较好的鲁棒性，对重要部位的定位能力较差，而较大的分支数也会降低模型的性能，因为在分支数较多时，多个分支会陷入几乎相同的解中。

表 4　损失函数权重 λ 对模型识别准确率的影响

Tab. 4　The effect of the loss function weights λ on model recognition accuracy

λ	0.2	0.3	0.5	0.6	0.8
准确率(%)	86.22	86.35	87.74	86.47	82.26

表 5　分支数 m 对模型识别准确率的影响

Tab. 5　Effect of the number of branches m on model recognition accuracy

m	0	1	2	3	4	5	6	7
准确率(%)	87.04	86.98	87.32	87.74	87.41	87.35	87.30	87.22

4　结束语

本文实现了一个基于 2D 图像的表情识别算法，该算法使用区域特征聚合网络引导网络关注的关注点，使得卷积网络能够关注到更多有效的面部特征并用于分类。同时，为了应对公开数据集中广泛存在的噪声问题，算法使用了再标记模块进行了噪声抑制。该算法在公开数据集中完成训练，并在多个数据集的测试集中与现有算法进行比较，证明了其具有较好的识别效果。

参 考 文 献

[1] EKMAN P, ROSENBERG E L. What the face reveals: basic and applied studies of spontaneous expression using the facial action coding system(facs)[M]. 2nd ed. Oxford: Oxford University Press, 2005.

[2] BARRETT L F. Discrete emotions or dimensions? the role of valence focus and arousal focus[J]. Cognition & Emotion, 1998, 12(4): 579-599.

[3] MARTINEZ A, DU S. A model of the perception of facial expressions of emotion by humans: research overview and perspectives[J]. Journal of Machine Learning Research, 2012, 13(5):1589-1608.

[4] 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. [S. l.: s. n.], 2016:279-283.

[5] ZHANG Z P, LUO P, LOY C C, et al. From facial expression recognition to interpersonal relation prediction[J]. International Journal of Computer Vision, 2018, 126(5): 550-569.

[6] LI S, DENG W H, DU J P. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. [S. l.]: IEEE, 2017: 2584-2593.

[7] ZAFEIRIOU S, KOLLIAS D, NICOLAOU M A, et al. Aff-wild: valence and arousal "in-the-wild" challenge[C]//Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition Workshops. [S. l.]: IEEE, 2017: 1980-1987.

[8] KOLLIAS D, ZAFEIRIOU S. Aff-wild2: extending the Aff-wild database for affect recognition[EB/OL]. (2019-12-13) [2023-08-27]. https://arxiv.org/abs/1811.07770.

[9] ZHANG K P, ZHANG Z P, LI Z F, et al. Joint face detection and alignment using multi-task cascaded convolutional networks[J]. IEEE Signal Processing Letters, 2016, 23(10): 1499-1503.

[10] AMOS B, LUDWICZUK B, SATYANARAYANAN M. OpenFace: a general-purpose face recognition library with mobile applications[EB/OL]. [2023-08-27]. https://elijah.cs.cmu.edu/DOCS/CMU-CS-16-118.pdf.

[11] NG P C, HENIKOFF S. Sift: predicting amino acid changes that affect protein function[J]. Nucleic Acids Research, 2003, 31(13): 3812-3814.

[12] DARAL N, TRIGGS B. Histograms of oriented gradients for human detection[C]//Proceedings of 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. [S. l.: s. n.], 2005: 886-893.

[13] 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.

[14] LIU C J, WECHSLER H. Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition[J]. IEEE Transactions on Image Processing, 2002, 11(4): 467-476.

[15] FASEL B. Robust face analysis using convolutional neural networks[C]//Proceedings of the 16th International Conference on Pattern Recognition. [S. l.]: IEEE, 2002: 40-43.

[16] LIU M Y, LI S X, SHAN S G, et al. AU-inspired deep networks for facial expression feature learning[J]. Neurocomputing, 2015, 159: 126-136.

[17] TANG Y C. Deep learning using linear support vector machines[EB/OL]. (2015-02-21)[2023-08-28]. https://arxiv.org/abs/1306.0239v2.

[18] KAHOU S E, PAL C, BOUTHILLIER X, et al. Combining modality specific deep neural networks for emotion recognition in video[C]//Proceedings of the 15th ACM on International Conference on Multimodal Interaction. [S. l.: s. n.], 2014: 461-466.

[19] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[C]//Proceedings of the 26th Annual Conference on Neural Information Processing Systems. [S. l.: s. n.], 2012: 1097-1105.

[20] SZEGEDY C, LIU W, JIA Y Q, et al. Going deeper with convolutions[C]//Proceedings of the 28th IEEE Conference on Computer Vision and Pattern Recognition. [S. l.: s. n.], 2015: 1-9.

[21] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2015-04-10)[2023-08-28]. https://arxiv.org/abs/1409.1556.

[22] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//Proceedings of the 29th IEEE Conference on Computer Vision and Pattern Recognition. [S. l.: s. n.], 2016: 770-778.

[23] 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.

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

[25] SUN Y, WANG X G, TANG X O. Deep learning face representation from predicting 10000 classes[C]//Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. [S. l.: s. n.], 2014: 1891-1898.

[26] XUE F L, WANG Q C, GUO G D. TransFER: learning relation-aware facial expression representations with transformers[C]//Proceedings of the 18th IEEE/CVF International Conference on Computer Vision. [S. l.: s. n.], 2021: 3601-3610.

[27] PHAN K N, NGUYEN H H, HUYNH V T, et al. Facial expression classification using fusion of deep neural network in video [C]//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. [S. l.]: IEEE, 2022: 2506-2510.

[28] DENG D D, CHEN Z K, SHI B E. Multi-task emotion recognition with incomplete labels[C]//Proceedings of the 15th IEEE International Conference on Automatic Face and Gesture Recognition. [S. l.: s. n.], 2020: 828-835.

[29] WANG K, PENG X J, YANG J F, et al. Suppressing uncertainties for large-scale facial expression recognition[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S. l.: s. n.], 2020: 6896-6905.

[30] WANG Q C, GUO G D. LS-CNN: characterizing local patches at multiple scales for face recognition[J]. IEEE Transactions on Information Forensics and Security, 2020, 15: 1640-1653.

[31] SIQUEIRA H, MAGG S, WERMTER S. Efficient facial feature learning with wide ensemble-based convolutional neural networks[C]//Proceedings of the 34th Conference on Artificial Intelligence. [S. l.: s. n.], 2020: 5800-5809.

[32] ZHOU B L, KHOSLA A, LAPEDRIZA A, et al. Learning deep features for discriminative localization [C]//Proceedings of the 29th IEEE Conference on Computer Vision and Pattern Recognition. [S. l.: s. n.], 2016: 2921-2929.

作 者 简 介

李剑鹏

男,1997 年生,硕士研究生,研究方向为计算机视觉与表情识别

E-mail:ljpastar@foxmail.com

苏 楠

男,1982 年生,工程师,研究方向为计算机视觉、人工智能、多模态目标识别、人脸识别等

E-mail:sunan@tsinghua.edu.cn

责任编辑 董 莉