

基于对比学习的细粒度遮挡人脸表情识别^①



奚 琰

(江苏大学 计算机科学与通信工程学院, 镇江 212013)

通信作者: 奚 琰, E-mail: xiy@stmail.ujs.edu.cn

摘 要: 和实验室环境不同, 现实生活中的人脸表情图像场景复杂, 其中最常见局部遮挡问题会造成面部外观的显著改变, 使得模型提取到的全局特征包含与情感无关的冗余信息从而降低了判别力. 针对此问题, 本文提出了一种结合对比学习和通道-空间注意力机制的人脸表情识别方法, 学习各局部显著情感特征并关注局部特征与全局特征之间的关系. 首先引入对比学习, 通过特定的数据增强方法设计新的正负样本选取策略, 对大量易获得的无标签情感数据进行预训练, 学习具有感知遮挡能力的表征, 再将此表征迁移到下游人脸表情识别任务以提高识别性能. 在下游任务中, 将每张人脸图像的表情分析问题转化为多个局部区域的情感检测问题, 使用通道-空间注意力机制学习人脸不同局部区域的细粒度注意力图, 并对加权特征进行融合, 削弱遮挡内容带来的噪声影响, 最后提出约束损失联合训练, 优化最终用于分类的融合特征. 实验结果表明, 无论是在公开的非遮挡人脸表情数据集 (RAF-DB 和 FER2013) 还是人工合成的遮挡人脸表情数据集上, 所提方法都取得了与现有先进方法可媲美的结果.

关键词: 人脸表情识别; 对比学习; 局部遮挡; 注意力机制; 深度学习

引用格式: 奚琰. 基于对比学习的细粒度遮挡人脸表情识别. 计算机系统应用, 2022, 31(11): 175-183. <http://www.c-s-a.org.cn/1003-3254/8813.html>

Fine-grained Occluded Facial Expression Recognition Based on Contrastive Learning

XI Yan

(School of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang 212013, China)

Abstract: Different from the laboratory environment, the scenes of facial expression images in real life are complex, and local occlusion, the most common problem, will cause a significant change in the facial appearance. As a result, the global feature extracted by a model contains redundant information unrelated to emotions, which reduces the discrimination of the model. Considering this problem, a facial expression recognition method combining contrastive learning and the channel-spatial attention mechanism is proposed in this study, which learns local salient emotion features and pays attention to the relationship between local features and global features. Firstly, contrastive learning is introduced. A new positive and negative sample selection strategy is designed through a specific data augmentation method, and a large amount of easily accessible unlabeled emotion data is pre-trained to learn the representation with occlusion-aware ability. Then, the representation is transferred to the downstream facial expression recognition task to improve recognition performance. In the downstream task, the expression analysis of each face image is transformed into the emotion detection of multiple local regions. The fine-grained attention maps of different local regions of a face are learned using the channel-spatial attention mechanism, and the weighted features are fused to weaken the noise effect caused by the occlusion content. Finally, the constraint loss for joint training is proposed to optimize the final fusion feature for classification. The experimental results indicate that the proposed method achieves comparable results to existing state-of-the-art methods on both public non-occluded facial expression datasets (RAF-DB and FER2013) and synthetic occluded

^① 收稿时间: 2022-03-05; 修改时间: 2022-04-12; 采用时间: 2022-04-22; csa 在线出版时间: 2022-07-14

facial expression datasets.

Key words: facial expression recognition; contrastive learning; local occlusion; attention mechanism; deep learning

1 引言

人脸表情识别 (facial expression recognition, FER) 一直是计算机视觉 (CV) 领域的研究热点. 自动人脸表情分析在社交机器人、医疗、驾驶员疲劳监测等许多人机交互系统中具有重要的实际意义. 最近, 深度学习的发展显著提高了 FER 任务的性能, 研究人员在构建数据集和开发 FER 方法方面取得了很大的进展^[1,2]. 实验室环境和真实场景下的人脸表情数据集, 如 CK+、SFEW、FER2013、AffectNet、EmotioNet 和 RAF-DB 等相继出现. 最近, 随着真实场景下的训练数据越来越多, 对 FER 的研究已经逐渐从实验室环境控制转向真实场景^[3]. 但真实场景下的人脸图像情景复杂, 存在许多情感无关因素, 如姿态各异、光照变化、以及局部遮挡. 其中最常见的是存在眼镜、帽子、手部等局部人脸遮挡. 尤其是在全球新冠肺炎疫情的大环境下, 为了防止新冠病毒的传播, 口罩成了人们出行必不可少的防护工具之一. 这些局部遮挡造成面部外观的显著改变, 使得模型提取到的特征包含与情感无关的冗余信息而降低了判别力, 从而降低 FER 准确率.

近年来, 自监督学习受到了广泛的关注, 其特点是不需要人工标注的类别标签信息, 直接利用数据本身作为监督信号来学习数据的特征表达, 并用于解决特定的下游任务. 自监督学习通过精心设计的前置任务来获取对下游任务有益的信息, 例如: 预测图像中两个分块之间的位置关系, 解决拼图问题或预测灰度图像的颜色信息等. 在计算机视觉中, 自监督学习已经应用至各种类型的下游任务, 比如常见的图像分类^[4], 目标检测^[5], 单目深度估计^[6]等. 这些工作表明了特定的数据增强对表征学习的重要性, 并启发我们探索如何利用自监督学习从大量易获得的无标签情感数据中学习具有感知遮挡能力的信息, 并将其用于遮挡 FER 任务.

除了前置任务, 下游遮挡 FER 任务的设计也十分多样和重要^[7]. 其中最关键的表情特征提取步骤的目的是在人脸图像中寻找与面部情绪相关的语义信息, 并将其作为后续分类任务的输入. 根据特征提取方法的不同, 面部表情特征可以分为手工特征和深度学习型特征. 常用的手工特征有尺度不变特征转换 (SIFT)^[8]、

局部二元模式 (LBP)^[9]、梯度直方图 (HOG)^[10] 和基于各种手工特征的混合特征^[11,12]. 但这些特征被认为是浅层特征, 无法从原始图像中获得深层语义特征. 手工特征提取算法往往无法从原始图像中自动提取特征, 面对大规模数据时, 传统的手工特征方法往往暴露出自身的缺点和困难. 深度学习型特征通常利用深度网络自动学习面部表情特征. 深度学习方法利用多层非线性变换, 因此它们在各种任务中表现出比手工特征更好的性能. 自然地, 如果对面脸表情的训练数据进行更准确的标注, 深度学习技术可以获得更好的收敛性和更鲁棒的模型. 随着 GPU 芯片处理能力的显著提高和算法的快速发展, 人脸表情识别也有了突破性进展^[13]. 从早期的手工特征的提取, 到现在各种各样的深度神经网络自动学习和提取情感特征. 在各类公开的人脸表情数据集上, 人脸表情识别精度被一次次刷新了新的记录. 如实验室环境下的人脸表情数据集 CK+ 已经达到了近百分之百的准确率, 但是在真实场景下, 人脸表情识别存在诸多困难和挑战.

对于遮挡状态下的 FER, 目前主要有以下两种解决方法: 一是从局部未遮挡的区域进行研究. 二是去除遮挡区域, 但这些方法适用于实验室环境, 而真实场景中的遮挡是随机的, 难以检测. 由于人类的视觉处理系统可以快速扫描全局图像, 并获取需要更多关注的一个或多个局部区域, 抑制其他无用信息, 从而能够在遮挡下感知情绪. 受此启发, 考虑到遮挡 FER 比传统分类任务更需要关注细节, 而粗粒度的整体面部的特征学习方式难以挖掘并学好表情局部区域的情感特征, 细粒度的特征学习方式更能学习到表情中局部的细微变化^[14], 从而突显情感特征的显著性与可区分性. Song 等人^[15] 的研究表明, 在存在遮挡的情况下, 不同的特征图通道的响应不同. 而且, 对于给定的输入刺激, 同一个通道在不同空间位置上显示出强度的变化. 因此在下游 FER 任务中, 引入适用于细粒度分类的通道-空间注意力机制来学习局部细节特征.

针对上述研究内容, 本文提出基于对比学习的局部全局关系约束算法. 其解决思路是在前置任务中, 首先利用自监督对比学习, 构建经过特定数据增强的正

样本对学习具有感知遮挡能力的表征,再将学习到的信息迁移到下游 FER 任务,以提高局部遮挡 FER 的性能.下游 FER 任务中使用通道-空间注意力机制学习不同人脸局部区域的细粒度加权特征并融合,削弱遮挡内容带来的噪声影响,并用融合后的全局特征进行人脸表情分类.为了进一步约束局部特征与融合后的全局表征之间的关系,提出了约束损失来捕获人脸局部区域与全局之间的联系,从而优化融合特征,提高模型在遮挡场景下对表情分类的判别力.本文的主要贡献总结如下:1)为了学习具有感知遮挡能力的信息,提出一种新的对比学习策略,通过特定的随机数据增强方式学习来自同源样本的不同局部遮挡状态下的不变性.2)将每张人脸图像的表情分析问题转化为多个局部区域的情感检测问题,采用通道-空间注意力机制学习各个局部区域的细粒度特征并在特征层融合,以达到更为鲁棒的人脸表情识别效果.3)引入约束损失,确保融合后的特征识别正确表情类别的概率大于每个局部区域,并与分类损失联合训练,为融合特征提供进一步的监督与优化.

2 基于对比学习的局部全局关系约束算法

2.1 概述

为了自适应减少或消除遮挡内容和不相关区的影响,学习区域之间的互补特征.提出了基于对比学习的细粒度遮挡人脸表情识别方法,该方法主要包含两个分支:遮挡对比学习前置任务和精心设计的下游人脸表情识别任务.本文方法总体框架如图1所示,主要包括3个模块:1)遮挡对比学习预训练模块(occlusion-aware contrastive learning, OCL):在预训练阶段,设计了一种新的对比学习策略,通过特定的随机数据增强方式学习来自同源样本的不同局部遮挡条件下的相似性,从而为下游任务提供具有感知遮挡能力的信息.2)细粒度加权情感特征学习模块(fine-grained weighted expression feature learning, FGW):学习人脸每个局部区域的细粒度特征,并通过细粒度注意力图对其进行加权融合.3)局部全局深监督模块(local-global deep-supervision, LGDS):通过约束损失进一步监督和优化全局特征,提高最终的表情分类效果.

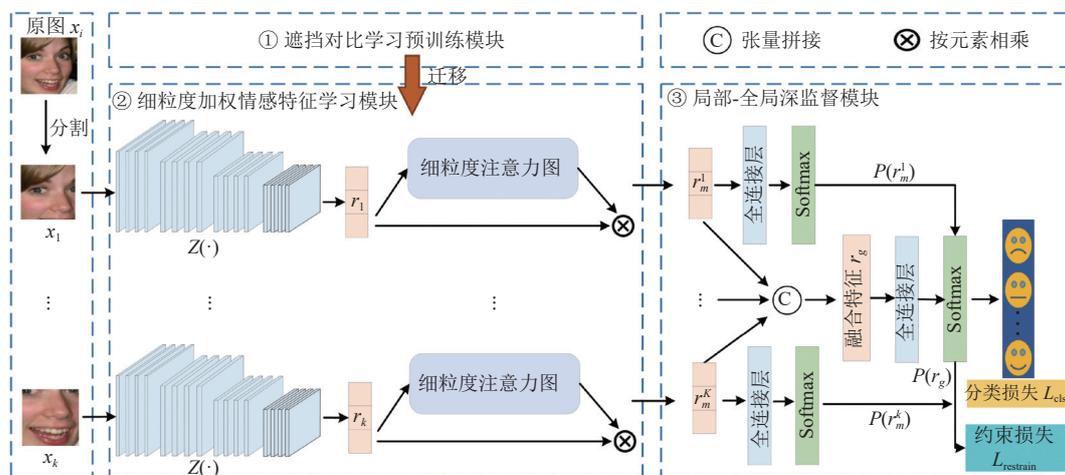


图1 本文方法总体框架图

2.2 遮挡对比学习预训练模块

对比学习方法能够在标注数据少的情况下学习具有判别力的表征,并且将获得的表征迁移至下游任务中,能够加快目标模型的收敛.针对遮挡 FER 任务,提出了一种具有感知遮挡能力的对比学习框架,设计了一种新的选择正负样本的策略.如图2所示,对于任意一个样本(anchor),其正样本对来自在 anchor 上添加了局部遮挡和其他数据增强操作的两个样本,负样本

为该训练批中剩余的其他样本.对比学习的原则是在表征空间中,将正样本对拉近并与负样本分离.

详细来说,假设在一个训练批 D 中有 N 张人脸表情图像,记为 $D = x_n, n = 1, \dots, N$.对于 N 个样本中的每张图像 x_n ,首先进行两次随机数据增强,特别地,在数据增强操作中加入随机局部遮挡,得到属于 x_n 的一对正样本,分别记为 x_i 和 x_j ,即总共 $2N$ 个数据点.对于 $2N$ 个样本中的每个样本 x_i ,接着利用基础编码器网络

$Enc(\cdot)$ 提取特征向量 h_i , 再采用一个浅层投影网络 $Proj(\cdot)$ 将 h_i 映射为向量 z_i . 对于一对正样本 (i, j) 的损失函数 $l_{(i,j)}$ 计算公式如下:

$$l_{(i,j)} = \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{2N-1} \exp(\text{sim}(z_i, z_k)/\tau)} \quad (1)$$

其中, z_k 是 $2N$ 个样本中除样本 x_i 外的所有样本的表征, τ 是温度参数. 最后, 对训练批中的所有正样本对, 包括 (i, j) 和 (j, i) 计算对比损失函数. 遮挡对比学习预训练阶段不仅减少了模型所需要的人工标注数据的数量, 而且通过充分利用来自同源样本的不同局部遮挡条件下的相似性和不同实例 (instance) 间的差异性, 最大化的挖掘和学习遮挡样本中潜在的情感信息. 得到的表征可迁移到下游 FER 任务, 降低带遮挡的噪声数据对人脸表情分类产生的负面影响, 以提高遮挡状态下人脸表情识别的性能.

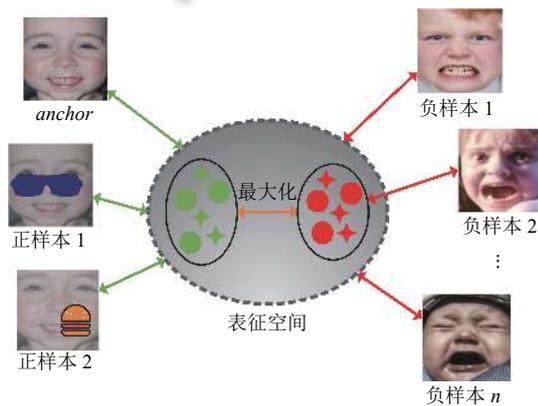


图2 遮挡对比学习预训练模块示意图

2.3 细粒度加权情感特征学习模块

考虑到遮挡人脸表情识别比传统分类任务更需要关注细节, 才能突显情感特征的显著性与可区分性. 因此在下游 FER 任务中, 引入适用于细粒度分类的注意力机制来学习表情中局部的细微变化. 为了获得更有效的注意力, 提出将一张人脸图像的表情分析问题转化为多个局部区域的情感检测问题, 利用细粒度加权情感特征学习模块为人脸不同局部特征计算其细粒度注意力图, 将得到的注意力图与对应的局部特征相乘得到局部加权特征, 获得每个局部区域的显著细粒度特征, 并将获得的每个局部加权特征融合起来作为分类器的输入. 具体来说, 对于每张通过预处理操作获得

的人脸表情图像 x_i , 首先将其按部分区域重叠的方式分割成 K 个局部区域, 其分割区域表示为 $x_1, \dots, x_k, \dots, x_K$. 则整个模型的输入集 X 定义为:

$$X = \{x_1, \dots, x_k, \dots, x_K\} \quad (2)$$

其中, x_k 表示第 k 块人脸局部区域, 并且 $0 \leq k \leq K$.

这 K 个局部区域将被统一调整成相同的大小作为共享参数的主干卷积网络 $Z(\cdot)$ 的输入, 主干卷积网络利用深度学习网络, 独立学习每个区域的特征, 为后续通道和空间注意力图的学习提供基础. 则给定输入集 X , 其输出特征集 R 定义为:

$$R = \{r_1, \dots, r_k, \dots, r_K\} = \{Z(x_1), \dots, Z(x_k), \dots, Z(x_K)\} \quad (3)$$

其中, r_k 是第 k 个局部区域 x_k 的特征.

接着, 通过在每个局部特征上使用通道-空间注意力网络来探索该区域在不同通道和空间下的显著细粒度信息. 具体来说, 对于每个局部特征 r_k , 按串联通道注意力操作 $M_C(\cdot)$ 和空间注意力操作 $M_S(\cdot)$ 的顺序, 自适应学习该局部区域的细粒度注意力图, 并将其与对应的局部特征 r_k 相乘得到第 k 个区域的局部加权特征 r_m^k , 计算公式如下:

$$r_m^k = (M_S(M_C(r_k) \otimes r_k)) \otimes (M_C(r_k) \otimes r_k) \quad (4)$$

其中, \otimes 代表元素相乘操作符. 对于任意三维特征 $F \in R^{C \times W \times H}$, 其对应的一维通道注意力图 $M_C(F) \in R^{C \times 1 \times 1}$ 和二维空间注意力图 $M_S(F) \in R^{1 \times W \times H}$ 的计算操作^[16] 分别如下:

$$M_C(F) = \sigma(MLP(\text{AvgPool}(F)) + MLP(\text{MaxPool}(F))) \quad (5)$$

$$M_S(F) = \sigma(f^{7 \times 7}([\text{AvgPool}(F); \text{MaxPool}(F)])) \quad (6)$$

其中, σ 是 Sigmoid 函数, MLP 是一个共享的多层感知器. $f^{7 \times 7}$ 代表一个过滤器尺寸为 7×7 的卷积操作. AvgPool 和 MaxPool 分别代表平均池化和最大池化操作. 平均池化对特征图上的每个像素点都有反馈, 其关注一张特征图上哪些内容是有意义的, 而最大池化在进行梯度反向传播时, 只有特征图中响应最大的地方有梯度的反馈, 其关注哪里的特征是有意义的, 两者结合更能获得显著情感信息.

为了保证图像信息的完整性, 添加了一个局部信息的融合操作来获取原始图像的高级语义特征. 具体来说, 计算出 K 个区域的局部加权特征后, 将所有的加权特征进行特征层的拼接 (concatenate) 操作, 得到一

个加权全局表征 r_g , 实现信息的融合, 并用该全局表征作为分类器的最终输入进行人脸表情的识别。

2.4 局部全局深监督模块

为了约束局部特征与融合后的全局表征 r_g 之间的关系, 进一步提升模型所学融合特征的判别力, 提高模型在遮挡场景下对表情分类的有效性, 提出了局部全局深监督模块来捕获人脸局部区域与全局之间的联系。具体来说, 在监督训练下, 无论是全局表征 r_g 还是第 k 个区域的加权特征 r_m^k , 将该特征通过一层全连接层 $FC(\cdot)$ 和 $Softmax$ 激活函数后, 即可得到概率向量中正确表情标签 y_i 对应的概率值 P , 用式(7)表示:

$$P(r) = \text{Softmax}_{y_i}(FC(r)) \quad (7)$$

其中, r 代表全局表征 r_g 或第 k 个区域的加权特征 r_m^k 。受到 bag of words 工作^[17]的启发, 融合特征往往会获得比单个区域特征更高的预测概率。因此, 在局部全局深监督模块中加入约束损失 L_{restrain} 用于约束局部特征和全局特征之间的关系。具体来说, 约束损失可以保证最终融合特征 r_g 进行正确分类的预测概率 $P(r_g)$ 大于单个区域特征 r_m^k 的预测概率 $P(r_m^k)$, 从而优化最终的分类器。该损失函数定义如下:

$$L_{\text{restrain}} = \sum (\max\{0, P(r_m^k) - P(r_g)\}) \quad (8)$$

模型的最终目的是正确地将表情样本 x_i 分类到正确的标签 y_i 中, 因此使用标准交叉熵损失以监督学习的方式进行情感预测。其中 N 是人脸表情数据集的图片数量, W 为整个模型的参数, y_i 以 one-hot 编码形式。将该损失记为 L_{cls} :

$$L_{\text{cls}} = - \sum_{n=1}^N y_i \log \frac{\exp(W^T \cdot r_g)}{\sum_n \exp(W^T \cdot r_g)} \quad (9)$$

其中, 细粒度加权情感特征学习模块是局部特征学习阶段, 而局部全局深监督模块是全局特征优化阶段。将这两个阶段结合, 可以同时学习更好的局部特征和更鲁棒的全局融合特征, 以获得更好的性能。最终损失函数定义如下:

$$L = L_{\text{cls}} + \lambda L_{\text{restrain}} \quad (10)$$

其中, 参数 λ 是平衡因子。模型训练过程中, 通过深度神经网络的迭代和反向传播, 正确的人脸表情标签会自适应正确引导注意力权重的学习。在反向传播中, 网络各个模块的梯度将通过 Adam 优化器进行更新, 整个下游模型为一个端到端的模型, 促使其学习到更符

合人脸表情识别需要的局部与全局特征, 从而提升整个网络的人脸表情识别性能。

3 实验结果与分析

3.1 预训练数据和目标数据集

为了验证本文提出的方法的有效性, 在两个具有挑战的人脸表情数据集上对提出的方法进行了评估, 分别为人脸表情数据集 RAF-DB^[18]和 FER2013^[19]。这些数据集涵盖真实世界下遮挡等各种复杂的场景。并通过在互联网上收集到的预训练数据用于对比学习预训练阶段。下面分别介绍数据集的基本信息及划分情况。

(1) 预训练数据: 是从互联网上收集, 通过一组包括大约 20 个相关单词的关键词 (如, 开心、微笑、大笑、有趣、悲伤、哭泣、惊讶、生气、厌恶、恐惧、恐怖等) 和 3 个与身份相关的词 (如, 小孩子、女人、男人等)。此外还添加了来自公开人脸表情数据集的图像进行预训练, 使数据分布更符合人脸表情数据。预训练数据的图像总数约为 12 万张, 所有的图像都经过 MTCNN^[20], 切割出人脸区域并对齐。需要注意的是, 在预训练时未使用任何原始数据集的人工标注。

(2) RAF-DB^[18]: 是一个真实场景下的公开人脸表情数据集, 其 29 672 张极其多样的人脸表情图像源自互联网。基于众包标注, 每幅图片都由约 40 位标注者独立标注。在本文中只使用了其单标签子集, 包括 7 类基本情绪, 即快乐 (happy)、悲伤 (sad)、恐惧 (fear)、厌恶 (disgust)、惊讶 (surprise)、愤怒 (angry) 和中性 (neutral)。

(3) FER2013^[19]: 是谷歌图像搜索 API 自动收集的无约束的公开人脸表情数据集。FER2013 训练集包含 28 709 张图像, 验证集包含 3 589 张验证图像以及测试集包含 3 589 张图像。其同样有 7 个基本表情标签, 该数据集中所有图像大小为 48×48 像素。

3.2 实验设置

对于对比学习预训练, 系统使用 PyTorch 深度学习框架进行训练, 模型在单个 NVIDIA 2080Ti 显卡上训练了 800 个周期 (epochs)。随机数据增强主要采用随机裁剪、随机局部遮挡、随机水平翻转、随机颜色抖动和随机灰度缩放等操作。而在下游 FER 任务中, 对于每张人脸表情图像, 首先将其分割成 5 个区域, 即 K 取 5。一方面, 当 K 过小时, 模型无法学到全部的细

节信息,当 K 过大时,融合特征会因为重叠太多冗余信息而产生噪声造成负面影响.另一方面,考虑到整个模型的大小和参数优化过程,也该将分割区域的数量控制在一定范围内.在这5个区域中,左上、右上、左下、右下4个区域是固定大小为原始图片75%比例的区域,最后一个区域是大小为原始图片85%比例的中心区域.所有的局部区域统一被调整为 112×112 大小作为主干网络的输入.对于主干网络,主要使用ResNet和VGG16这两种流行的网络进行实验.当约束损失和交叉熵损失联合训练时,默认 $\lambda=0.5$.在所有数据集上,网络使用Adam优化器进行训练,超参数 $\beta_1=0.9, \beta_2=0.999$,学习率设置为0.00025.并且设置了Early Stop,避免冗余训练和过拟合.同时使用单一学习速率衰减策略,即在80个训练epochs之后,每隔5个epochs学习率从0.01衰减10%.

3.3 消融实验

为了验证本文所提方法的每个模块对最终人脸表情识别的影响,本小节在数据集RAF-DB和FER2013的测试集上进行了消融实验并进行分析.首先,利用ResNet18(Res18)作为主干网络复现了一些基础方法,从是否使用遮挡对比学习预训练模块(OCL)、是否使用细粒度加权情感特征学习模块(FGW)以及是否使用局部全局深监督模块(LGDS)等3个方面分析了基础方法与所提方法之间的差异.其中,“Pre”表示主干网络经过在ImageNet数据集上的预训练,“Pre-contrast”表示经过遮挡对比学习进行的预训练,“CSA”表示通道-空间注意力机制, $L_{restrain}$ 表示约束损失.模型消融实验采用Top-1准确率作为评估准则,结果如表1所示.

表1 模型消融实验评估结果(%)

方法	Pre	Pre-contrast	CSA	$L_{restrain}$	RAF-DB	FER2013
Res18	√	—	—	—	84.67	68.59
OCL+Res18	—	√	—	—	85.90	70.51
OCL+Res18 +FGW	—	√	√	—	86.35	70.96
OCL+Res18 +FGW+LGDS	—	√	√	√	87.14	71.63

遮挡对比学习预训练模块:当未加上遮挡对比学习的预训练时,“OCL+Res18”比基础方法“Res18”在RAF-DB和FER2013上的识别精度分别提升了1.23%和1.92%.显而易见,相比于其他预训练方法,对比学习预训练能够提高下游任务的性能.分析可知,对比学习预训练步骤通过充分利用来自同源样本的不同局部

遮挡条件下的相似性和不同实例间的差异性,能够学习到具有感知遮挡能力的信息,这对下游人脸表情识别任务是有益的.

细粒度加权情感特征学习模块:“OCL+Res18+FGW”比“OCL+Res18”基本方法在RAF-DB和FER2013上的识别精度有所提高,验证了模型中细粒度加权情感特征学习模块的必要性和合理性.分析可知,细粒化的区域特征学习对于人脸表情识别,尤其是遮挡情况下的人脸表情识别是十分重要的.如果仅学习粗粒度全局特征,则无法对遮挡场景中的遮挡内容以及情感无关区域进行抑制,必然会影响最终的识别性能及模型的泛化能力.

局部全局深监督模块:通道-空间注意力机制在多个研究工作中被证明是有效的,但在所提方法中,单独使用它会忽略局部与全局之间的关系.通过添加局部全局深监督模块,在局部区域和全局区域之间提供了额外的监督.如表1所示,本文所提出的方法“OCL+Res18+FGW+LGDS”添加约束损失联合训练后,在RAF-DB和FER2013的识别性能有所提升.该增益源于约束损失约束了细粒度加权情感特征学习模块得到的融合特征的预测概率大于每个局部区域的预测概率,从而进一步监督和优化了作为分类器最终输入的融合特征.总体来讲,本文所提方法取得了最优的结果,验证了各个模块的必要性.

3.4 与其他方法比较

将本文所提出的方法和其他先进的(state-of-the-art, SOTA)人脸表情识别方法分别在非遮挡数据集RAF-DB和FER2013上进行性能比较,主要包括基于先进网络结构的方法^[21-24]、基于自监督学习的方法^[4]和针对遮挡FER的方法^[25-28]等,实验均来自原始的未添加人工遮挡的数据集,结果如表2所示.

结果显示,本文所提出的方法的识别性能优于目前领先的人脸表情识别方法.所提方法在RAF-DB数据集上的性能提高了0.24%–3.87%,超过了目前领先的方法.在FER2013数据集,所提方法的人脸表情识别精度最高提高了7.22%.BLOCK-FRENET^[21]提出频率神经网络(FreNet)并首次在频域上处理图像.E-FCNN^[23]提出了一种边缘感知反馈卷积神经网络,将图像超分辨率和面部表情识别结合在一起用于解决低分辨率人脸图像的问题.IPA2LT^[24]在RAF-DB的准确

率为 86.77%，然而，它是在几个不同的数据集上训练的。具体来说，该方法在多个人工标注的数据和未标注的数据上进行训练。相反，本文所提出的方法的训练只在目标数据集的人工标记的训练数据上进行，其数量远远小于方法 IPA2LT。先前较好的针对遮挡 FER 的方法 RAN-ResNet18^[28] 使用区域自注意力模型来学习每个区域的权重，但这种方法忽略了局部区域和全局之间的关系来提供额外的监督。本文的方法既未使用结构庞大的网络结构，也未受益于其他自监督学习方法 SimCLR^[4] 所使用的先进 GPU 设备。该提升归功于对比学习预训练步骤能够学习到具有感知遮挡能力的信息，能够帮助下游 FER 任务，而细粒度加权情感特征学习模块分块学习每个局部分块的细节比直接学习整个全局粗粒度的特征能获得更好的情感信息，降低了遮挡噪声对模型的影响。并通过局部全局深监督模块约束局部特征与全局特征之间的关系，深度优化融合特征来获得增强的特征表示，从而提高表情识别的正确率。

表2 本文方法和其他 SOTA 方法比较结果 (%)

类别	方法	RAF-DB	FER2013
先进网络结构	BLOCK-FRENET ^[21]	—	64.41
	ALT ^[22]	84.55	69.85
	E-FCNN ^[23]	84.62	66.17
	IPA2LT ^[24]	86.77	—
自监督学习	SimCLR ^[4]	85.76	69.56
	PG-CNN ^[25]	83.27	—
	gACNN ^[26]	85.07	—
遮挡FER	OAENet ^[27]	85.69	—
	RAN-ResNet18 ^[28]	86.90	—
	本文方法	87.14	71.63

3.5 遮挡人脸表情数据集实验

为了验证所提方法解决遮挡问题的有效性，人工合成了源自 RAF-DB 测试集的遮挡表情数据集，命名为 occlusion-RAF-DB。考虑到真实场景中人脸存在不同位置不同程度的复杂情况的局部遮挡，所以在图像上添加不同比例的随机遮挡块作为噪声 (0%, 5%, 10%, ..., 30%)，并从实际场景出发，以生活中常见的口罩作为遮挡物，该口罩图片收集自互联网，并根据人脸中的关键点固定在特定位置。除遮挡属性外遮挡样本的类别及基本信息与非遮挡样本保持一致。在 Occlusion-RAF-DB 数据集上分别使用基于手工特征的方法 (SIFT^[8]、LBP^[9] 和 HOG^[10]) 和深度学习方法 (在 ImageNet 数据

集上预训练过的 Res18 和 DenseNet^[29] 网络) 与本文所提出的方法进行了对比实验，结果如图 3 所示。

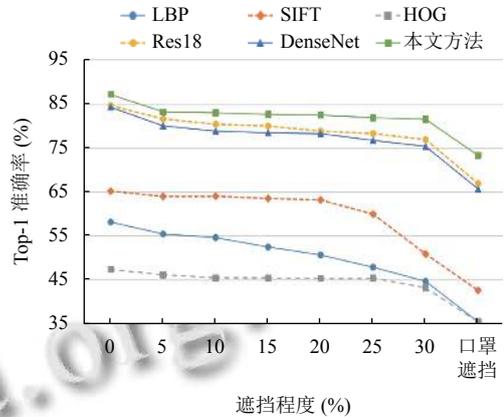


图3 本文方法与其他方法在遮挡 FER 任务的比较结果

实验结果表明，无论是在无遮挡的原始数据集上，还是在不同程度的人工合成的局部遮挡数据集上，本文所提方法的性能大幅度优于基于手工特征的识别方法。所提方法相比于基于手工特征的识别方法，能更好地应对复杂的遮挡场景下的人脸表情识别。该实验结果也表明深度学习学到的特征相比手工特征具有更强的分类能力。此外，本文所提方法性能明显优于 Res18 和 DenseNet 这两种深度学习识别方法。

随着人脸随机遮挡块面积的增加，图中所有方法的平均性能都明显下降，但本文所提出的方法具有较高的稳定性。所提方法性能下降的最大幅度仅为 13.79%，而 Res18 和 DenseNet 这两种深度学习方法性能分别下降 17.65% 和 18.57%。这种稳定性归功于所提方法中的细粒度加权情感特征学习模块降低了遮挡噪声对模型的影响。由于表情产生的脸部形变比较细微，分块学习每个局部的细节比直接学习整个全局粗粒度的特征能获得更好的情感信息。并通过局部全局深监督模块约束局部特征与全局特征之间的联系，从而优化融合后的特征。总的来说，这些结果证明了本文所提出的方法对局部遮挡下的人脸表情数据的有效性。

3.6 可视化分析

为了直观地验证本文所提出的方法的有效性，本节将得到的部分结果进行可视化展示。如图 4 所示，该可视化图主要包括未添加人工遮挡 (0% 遮挡)、5%、10%、15%、20%、25%、30% 比例的人工遮挡以及口罩遮挡等不同遮挡程度的原图以及显著情感定位的效果

图. 其中, (I) 行是原图, (II) 行是效果图. 其中高亮区域代表情感显著区域, 白色部分越亮说明该区域情感越明显.



图4 不同遮挡程度的情感显著区域定位可视化图

从图4中可以看出, 在各种不同程度的遮挡情况下, 白色部分能够有效地定位在显著表达情感的未遮挡区域, 包括嘴巴, 眼睛等部位. 在遮挡程度较重及在“口罩遮挡”组中, 当鼻子、嘴巴等下半边脸被完全遮挡时, 本文所提出的方法能够聚焦于眼睛、眉毛等上半部分人脸情感显著区域, 为 лица表情识别保留有效的局部区域, 从而削弱口罩噪声对人脸表情识别的影响. 以上可视化结果充分说明了本文所提出的方法的有效性, 粗粒度的全局整体面部的特征学习方式难以挖掘并学好表情局部区域的情感特征, 细粒度的特征学习方式更能学习到表情中局部的细微变化, 从而突显情感特征的显著性与可区分性. 通过细粒度加权情感特征学习模块学习每个局部分块的细粒度特征并互相补充, 从而关注面部的情感可判别区域而忽略遮挡噪声. 并且通过局部全局深监督模块深度优化融合后的特征来获得增强的特征表示, 从而提高表情识别的正确率.

4 总结与展望

针对真实场景局部遮挡的情况, 本文提出了基于

对比学习的细粒度遮挡人脸表情识别方法. 通过经特定数据增强的对比学习来获取具有感知遮挡能力的信息, 再采用细粒度加权情感特征学习模块获取每个局部区域的细粒度注意力图, 并将获得的每个局部加权特征进行融合, 削弱遮挡内容带来的噪声影响, 并用融合后的全局特征进行人脸表情分类. 最后引入了约束损失确保融合特征识别正确表情类别的概率大于每个局部区域, 进一步保证了在监督训练下融合特征的分类性能优于各局部特征. 本文所提出的方法在公开的非遮挡人脸表情数据集以及合成的遮挡人脸表情数据集上的实验均获得了很好的结果, 证明了所提方法的有效性和优越性. 在未来的工作中, 将尝试更具挑战性的遮挡人脸表情数据集.

参考文献

- 1 李星燃, 张立言, 姚树婧. 结合特征融合和注意力机制的微表情识别方法. 计算机科学, 2022, 49(2): 4–11.
- 2 龙寒潮, 丁美荣, 林桂锦, 等. 基于视听觉感知系统的多模态情感识别. 计算机系统应用, 2021, 30(12): 218–225. [doi: 10.15888/j.cnki.csa.008235]
- 3 武中华. 基于图卷积多标签学习的复合人脸表情识别. 计算机系统应用, 2022, 31(1): 259–266. [doi: 10.15888/j.cnki.csa.008273]
- 4 Chen T, Kornblith S, Norouzi M, *et al.* A simple framework for contrastive learning of visual representations. Proceedings of the 37th International Conference on Machine Learning. Online: PMLR, 2020. 1597–1607.
- 5 Yang CY, Wu ZR, Zhou BL, *et al.* Instance localization for self-supervised detection pretraining. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 3986–3995. [doi: 10.1109/CVPR46437.2021.00398]
- 6 Tan FT, Zhu H, Cui ZP, *et al.* Self-supervised human depth estimation from monocular videos. Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2020. 647–656. [doi: 10.1109/CVPR42600.2020.00073]
- 7 南亚会, 华庆一. 遮挡人脸表情识别深度学习研究方法研究进展. 计算机应用研究, 2022, 39(2): 321–330. [doi: 10.19734/j.issn.1001-3695.2021.08.0307]
- 8 Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Research, 2003, 31(13): 3812–3814. [doi: 10.1093/nar/gkg509]
- 9 Zhao GY, Pietikainen M. Dynamic texture recognition using

- local binary patterns with an application to facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2007, 29(6): 915–928. [doi: [10.1109/TPAMI.2007.1110](https://doi.org/10.1109/TPAMI.2007.1110)]
- 10 Dalal N, Triggs B. Histograms of oriented gradients for human detection. *Proceedings of 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. San Diego: IEEE, 2005. 886–893. [doi: [10.1109/CVPR.2005.177](https://doi.org/10.1109/CVPR.2005.177)]
- 11 Gupta SK, Agrwal SL, Meena YK, *et al.* A hybrid method of feature extraction for facial expression recognition. *Proceedings of the 7th International Conference on Signal Image Technology & Internet-based Systems*. Dijon: IEEE, 2011. 422–425. [doi: [10.1109/SITIS.2011.64](https://doi.org/10.1109/SITIS.2011.64)]
- 12 Wang XH, Jin C, Liu W, *et al.* Feature fusion of HOG and WLD for facial expression recognition. *IEEE/SICE International Symposium on System Integration*. Kobe: IEEE, 2013. 227–232. [doi: [10.1109/SII.2013.6776664](https://doi.org/10.1109/SII.2013.6776664)]
- 13 吕勇强. 自然场景下的人脸表情识别研究 [硕士学位论文]. 武汉: 华中师范大学, 2021. [doi: [10.27159/d.cnki.ghzsu.2021.002034](https://doi.org/10.27159/d.cnki.ghzsu.2021.002034)]
- 14 苏志明, 王烈, 蓝峥杰. 基于多尺度分层双线性池化网络的细粒度表情识别模型. *计算机工程*, 2021, 47(12): 299–307, 315. [doi: [10.19678/j.issn.1000-3428.0060133](https://doi.org/10.19678/j.issn.1000-3428.0060133)]
- 15 Song LX, Gong DH, Li ZF, *et al.* Occlusion robust face recognition based on mask learning with pairwise differential Siamese network. *Proceedings of the IEEE/CVF International Conference on Computer Vision*. Seoul: IEEE, 2019. 773–782. [doi: [10.1109/ICCV.2019.00086](https://doi.org/10.1109/ICCV.2019.00086)]
- 16 Woo S, Park J, Lee JY, *et al.* CBAM: Convolutional block attention module. *Proceedings of the 15th European Conference on Computer Vision*. Munich: Springer, 2018. 3–19. [doi: [10.1007/978-3-030-01234-2_1](https://doi.org/10.1007/978-3-030-01234-2_1)]
- 17 Csurka G, Dance CR, Fan LX, *et al.* Visual categorization with bags of keypoints. *Proceedings of the 8th European Conference on Computer Vision*. Prague: Springer, 2004. 59–74.
- 18 Zhang KP, Zhang ZP, Li ZF, *et al.* Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 2016, 23(10): 1499–1503. [doi: [10.1109/LSP.2016.2603342](https://doi.org/10.1109/LSP.2016.2603342)]
- 19 Li S, Deng WH, Du JP. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. *Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition*. Honolulu: IEEE, 2017. 2584–2593. [doi: [10.1109/CVPR.2017.277](https://doi.org/10.1109/CVPR.2017.277)]
- 20 Goodfellow IJ, Erhan D, Carrier PL, *et al.* Challenges in representation learning: A report on three machine learning contests. *Proceedings of the 20th International Conference on Neural Information Processing*. Daegu: Springer, 2013. 117–124. [doi: [10.1007/978-3-642-42051-1_16](https://doi.org/10.1007/978-3-642-42051-1_16)]
- 21 Tang Y, Zhang XM, Hu XP, *et al.* Facial expression recognition using frequency neural network. *IEEE Transactions on Image Processing*, 2021, 30: 444–457. [doi: [10.1109/TIP.2020.3037467](https://doi.org/10.1109/TIP.2020.3037467)]
- 22 Florea C, Florea L, Badea MS, *et al.* Annealed label transfer for face expression recognition. *Proceedings of the 30th British Machine Vision Conference*. Cardiff: BMVA Press, 2019. 104.
- 23 Shao J, Cheng QY. E-FCNN for tiny facial expression recognition. *Applied Intelligence*, 2021, 51(1): 549–559. [doi: [10.1007/s10489-020-01855-5](https://doi.org/10.1007/s10489-020-01855-5)]
- 24 Zeng JB, Shan SG, Chen XL. Facial expression recognition with inconsistently annotated datasets. *Proceedings of the 15th European Conference on Computer Vision*. Munich: Springer, 2018. 227–243. [doi: [10.1007/978-3-030-01261-8_14](https://doi.org/10.1007/978-3-030-01261-8_14)]
- 25 Li Y, Zeng JB, Shan SG, *et al.* Patch-gated CNN for occlusion-aware facial expression recognition. *Proceedings of the 2018 24th International Conference on Pattern Recognition*. Beijing: IEEE, 2018. 2209–2214. [doi: [10.1109/ICPR.2018.8545853](https://doi.org/10.1109/ICPR.2018.8545853)]
- 26 Li Y, Zeng JB, Shan SG, *et al.* Occlusion aware facial expression recognition using CNN with attention mechanism. *IEEE Transactions on Image Processing*, 2019, 28(5): 2439–2450. [doi: [10.1109/TIP.2018.2886767](https://doi.org/10.1109/TIP.2018.2886767)]
- 27 Wang ZN, Zeng FW, Liu SC, *et al.* OAENet: Oriented attention ensemble for accurate facial expression recognition. *Pattern Recognition*, 2021, 112: 107694. [doi: [10.1016/j.patcog.2020.107694](https://doi.org/10.1016/j.patcog.2020.107694)]
- 28 Wang K, Peng XJ, Yang JF, *et al.* Region attention networks for pose and occlusion robust facial expression recognition. *IEEE Transactions on Image Processing*, 2020, 29: 4057–4069. [doi: [10.1109/TIP.2019.2956143](https://doi.org/10.1109/TIP.2019.2956143)]
- 29 Huang G, Liu Z, van der Maaten L, *et al.* Densely connected convolutional networks. *Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition*. Honolulu: IEEE, 2017. 2261–2269. [doi: [10.1109/CVPR.2017.243](https://doi.org/10.1109/CVPR.2017.243)]

(校对责编: 孙君艳)