Journal of Computer Applications
计算机应用, 2025, 45(1): 253 - 260
ISSN 1001-9081
CODEN JYIIDU
2025-01-10
http://www.joca.cn

文章编号:1001-9081(2025)01-0253-08
DOI:10.11772/j.issn.1001-9081.2024010098

基于上下文通道注意力机制的人脸属性估计与表情识别

徐　杰[1],钟　勇[2],王　阳[3],张昌福[4],杨观赐[1,3*]

(1. 现代制造技术教育部重点实验室(贵州大学),贵阳 550025;　2. 中国科学院 成都计算机应用研究所,成都 610213;
3. 省部共建公共大数据国家重点实验室(贵州大学),贵阳 550025;　4. 贵州大学 机械工程学院,贵阳 550025)

(*通信作者电子邮箱 gunaci_yang@163.com)

摘　要:人脸特征蕴含诸多信息,在面部属性和情感分析任务中具有重要价值,而面部特征的多样性和复杂性使人脸分析任务变得困难。针对上述难题,从面部细粒度特征角度出发,提出基于上下文通道注意力机制的人脸属性估计和表情识别(FAER)模型。首先,构建基于ConvNext的局部特征编码骨干网络,并运用骨干网络编码局部特征的有效性来充分表征人脸局部特征之间的差异性;其次,提出上下文通道注意力(CC Attention)机制,通过动态自适应调整特征通道上的权重信息,表征深度特征的全局和局部特征,从而弥补骨干网络编码全局特征能力的不足;最后,设计不同分类策略,针对人脸属性估计(FAE)和面部表情识别(FER)任务,分别采用不同损失函数组合,以促使模型学习更多的面部细粒度特征。实验结果表明,所提FAER模型在人脸属性数据集CelebA(CelebFaces Attributes)上取得了91.87%的平均准确率,相较于次优模型SwinFace(Swin transformer for Face)高出0.55个百分点;在面部表情数据集RAF-DB和AffectNet上分别取得了91.75%和66.66%的准确率,相较于次优模型TransFER(Transformers for Facial Expression Recognition)分别高出0.84和0.43个百分点。

关键词:人脸属性估计;面部表情识别;注意力机制;细粒度特征;特征差异

中图分类号:TP18　　**文献标志码**:A

Facial attribute estimation and expression recognition based on contextual channel attention mechanism

XU Jie[1], ZHONG Yong[2], WANG Yang[3], ZHANG Changfu[4], YANG Guanci[1,3*]

(1. *Key Laboratory of Advanced Manufacturing Technology of the Ministry of Education (Guizhou University), Guiyang Guizhou* 550025, *China*;
2. *Chengdu Institute of Computer Application, Chinese Academy of Sciences, Chengdu Sichuan* 610213, *China*;
3. *State Key Laboratory of Public Big Data (Guizhou University), Guiyang Guizhou* 550025, *China*;
4. *School of Mechanical Engineering, Guizhou University, Guiyang Guizhou* 550025, *China*)

Abstract: Facial features contain a lot of information and hold significant value in facial attribute and expression analysis tasks, but the diversity and complexity of facial features make facial analysis tasks difficult. Aiming at the above issue, a model of Facial Attribute estimation and Expression Recognition based on contextual channel attention mechanism (FAER) was proposed from the perspective of fine-grained facial features. Firstly, a local feature encoding backbone network based on ConvNext was constructed, and by utilizing the effectiveness of the backbone network in encoding local features, the differences among facial local features were represented adequately. Secondly, a Contextual Channel Attention (CC Attention) mechanism was introduced. By adjusting the weight information on feature channels dynamically and adaptively, both global and local features of deep features were represented, so as to address the limitations of the backbone network ability in encoding global features. Finally, different classification strategies were designed. For Facial Attribute Estimation (FAE) and Facial Expression Recognition (FER) tasks, different combinations of loss functions were employed to encourage the model to learn more fine-grained facial features. Experimental results show that the proposed model achieves an average accuracy of 91.87% on facial attribute dataset CelebA (CelebFaces Attributes), surpassing the suboptimal model SwinFace (Swin transformer for Face) by 0.55 percentage points, and the proposed model achieves accuracies of 91.75% and 66.66% respectively on facial expression datasets RAF-DB and AffectNet, surpassing the suboptimal model TransFER (Transformers for Facial Expression Recognition) by 0.84 and 0.43 percentage points respectively.

收稿日期:2024-01-26;**修回日期**:2024-03-28;**录用日期**:2024-04-01。　**基金项目**:国家自然科学基金资助项目(62373116, 62163007);贵州省科技计划项目(黔科合支撑[2023]一般118,黔科合平台人才[2020]6007-2)。

作者简介:徐杰(1997—),男,安徽阜阳人,硕士研究生,CCF会员,主要研究方向:自主智能系统;　钟勇(1966—),男,四川岳池人,研究员,博士,主要研究方向:大数据及其智能处理、云计算、软件工程;　王阳(1987—),男,河南鹤壁人,高级工程师,博士,主要研究方向:人工智能、计算机视觉、大数据智能分析;　张昌福(1990—),男,贵州瓮安人,高级工程师,主要研究方向:工业大数据、人工智能;　杨观赐(1983—),男,湖南嘉禾人,CCF高级会员,教授,博士生导师,博士,主要研究方向:自主智能系统与机器人、多模态数据驱动的认知计算。

Key words: Facial Attribute Estimation (FAE); Facial Expression Recognition (FER); attention mechanism; fine-grained feature; feature difference

0 引言

人脸图像分析在生物特征识别和计算机视觉中具有重要意义，人脸属性估计（Facial Attribute Estimation，FAE）和面部表情识别（Facial Expression Recognition，FER）引起了学者的广泛关注，FAE和FER任务要求计算机能识别人脸特征并进行面部特征分析，在安防监控[1]、情感分析[2]和隐私保护[3]等方面有着广泛的应用。

FAE任务的核心挑战是如何有效编码人脸细粒度特征，以更充分地挖掘面部图像的局部特征，如刘海、五官和饰品等特征。由于人脸局部特征的多样性和复杂性等难题，致使FAE仍然是一个具有挑战性的问题[4-5]。随着深度学习在图像分析中的发展，诸多工作重点关注设计基于卷积神经网络（Convolution Neural Network，CNN）的判别特征[6-9]。此外，文献[6]中首先探索了人脸属性之间的相互关系，提出了联合属性表征模型；文献[8]中设计了人脸属性共享特征表征模块，以挖掘更多人脸特征；由于感受野的存在，卷积主要集中较小的局部特征，进而忽略了全局特征的重要性[7]。相反，一些工作将ViT（Vision Transformer）引入到FAE任务上，建模人脸的全局特征[10-13]，并取得了不错的效果；Dosovitskiy等[12]验证了基于ViT网络在人脸属性任务上能获得比ResNet[14]更好的性能；Label2Label（Label sentence to Label sentence）[11]通过引入标签转换器以增强特征学习能力，构建了多个相关联人脸属性之间的依赖关系；SwinFace（Swin transformer for Face）[10]基于共享骨干网络学习多任务特征，并构建注意力机制以表征属性之间的关联性，可以提升人脸分析任务准确率。

尽管上述工作在FAE任务上取得了较好的效果，但基于ViT架构的网络仅在编码全局特征上具有优越性，在编码局部特征时仍有欠缺。考虑到CNN在编码人脸局部细粒度特征的有效性，但无法充分表征全局特征，受ConvNext[15]编码图像特征有效性的启发，本文将ConvNext作为Baseline，提出了上下文通道注意力（Contextual Channel Attention，CC Attention）机制，以编码人脸全局特征；结合Baseline和CC Attention的优点，构建能有效编码全局和局部特征的基于上下文通道注意力机制的人脸属性估计与表情识别（Facial Attribute estimation and Expression Recognition based on contextual channel attention，FAER）模型，以充分表征人脸属性的细粒度特征。

此外，在FER任务中，Ekman等[16]定义了7种典型的基本面部表情类别，包括愤怒、厌恶、恐惧、高兴、自然、悲伤和惊讶，在文献[17-18]工作中，阐述了FER任务的挑战性，主要存在以下两个方面：1）类内差异大。属于同一类别的表情可能具有截然不同的外观，如不同身份、性别、年龄和种族的人可能会表达出相似的面部表情，相同的人在不同的灯光和姿态条件下可能具有不同的面部表情。2）类间差异小。来自不同类别的表情可能具有很大的相似性，由愤怒、恐惧和开心表情的嘴巴部分及开心、自然和惊讶的眼睛部分可以观察到它们之间存在着相似性。为了解决上述难题，一些研究者从损失函数角度出发，惩罚深度特征之间的差异[19-23]；Wen等[19]通过引入Center Loss函数，惩罚样本与对应类别中心之间的距离差异；Farzaneh等[21]设计了深度可分离损失，促使样本在深度可分离空间中更紧凑，以增加类内紧凑性和类间可分离性。与此同时，一些方法通过设计注意力机制，促使模型更关注面部局部特征[24-28]；考虑到面部表情的姿态变化和局部遮挡问题，Wang等[25]通过设计区域注意力促使模型聚焦于局部特征，Ma等[26]构建了注意力选择机制学习面部特征具有判别性的信息，同时抑制噪声或冗余信息。

尽管基于特征损失函数和注意力机制的方法在FER任务中有不错的表现，但未充分考虑面部表情中局部和全局特征之间的内在联系，且面部表情之间的差异体现在多个区域特征，如眉毛、嘴巴和眉心等部位。因此，本文提出了CC Attention，捕获全局和局部特征，构建面部表情上下文之间的依赖关系，以充分表征面部表情之间的相关性和差异性。

结合骨干网络编码图像局部特征的有效性和CC Attention表征上下文之间的关系，本文提出了FAER模型。

本文主要工作如下：

1）构建ConvNext网络编码图像局部特征的骨干网络，以编码人脸特征的细粒度信息，促使模型表征人脸局部特征之间的差异性和关联性。

2）提出了CC Attention机制，通过自适应动态调整通道上不同权重信息，学习人脸图像的全局和局部特征，从而表征深度特征之间的上下文信息。

1 相关工作

FAE任务阐述了人类可直观理解的人脸特征，如性别、眼镜、胡须和鼻子等。FAE任务通常是对人脸是否具有某种特点的二元判断。Liu等[5]发布了（CelebFaces Attributes，CelebA）数据集，包含超过20万张人脸图片，具有40个二元属性标签，加速了FAE任务的研究[10, 29-32]。近年来，为了提高FAE任务的性能，一些新的模型不断被提出。Rudd等[29]设计了一种混合目标优化网络（Mixed Objective Optimization Network，MOON），利用多个目标函数来优化不同层次和人脸属性估计的性能；与此同时，考虑到人脸属性和人脸特征点之间的关系，Hand等[30]基于人脸属性关系，提出了一种改进多任务辅助属性分类方法（Multi-task CNN with an AUXiliary network，MCNN-AUX），利用隐式和显式关系进行人脸属性分类；此外，Zhuang等[31]提出了一种端到端的级联卷积模型（Multi-task learning of Cascaded CNN for Facial Attribute，MCFA），基于动态加权方法为每个人脸属性分配不同权重，促使模型更聚焦于难以预测的属性；Mao等[32]采用了多任务多标签模型（Deep Multi-task Multi-label CNN，DMM-CNN），通过引入特殊层促使不同任务之间共享信息，以提高FAE任务的性能；不同于上述单一分类任务，SwinFace[10]和Savchenko等[33]通过共享骨干网络编码人脸属性特征，构建多尺度特征融合模块，学习多尺度人脸特征，并设计注意力机制以提高模型对人脸属性的鲁棒性。

FER任务因具有类内差异大和类间差异小的特点，仍然是一项具有挑战性的任务。针对上述难题，诸多学者从多角度提出方法，以提高面部表情识别的性能[21, 26-28, 34-36]。Farzaneh等[21]通过引入新的特征损失函数（Deep Attentive Center Loss，DACL）增强表情特征的判别能力；VTFF（Visual Transformers with Feature Fusion）[26]和TransFER（Transformers for Facial Expression Recognition）[27]基于ViT架构捕获面部表

情之间的关系,并设计注意力机制来处理不同面部区域之间的关系,促使模型动态选择最相关和最显著的人脸特征区域;Vo 等[28]基于金字塔结构模型(Pyramid with Super Resolution,PSR)构建多尺度表情特征,可以更好地捕捉面部表情的细节和局部特征,增强模型对多尺度面部表情的适应性和鲁棒性;考虑到面部表情的复杂性和多样性,Zhao 等[34]采用一种鲁棒轻量级面部表情识别网络(Efficient robust Facial expression recognition, EfficientFace),并采用标签分布式训练策略提高复杂场景下面部表情识别的精度;与此同时,Zhao 等[35]通过引入全局多尺度特征和局部注意力机制(Multi-scale and local Attention Network, MA-Net)捕获面部表情的多尺度特征,从而学习到更丰富的特征表示,以提高 FER 任务性能。不同于上述方法,Wen 等[36]从面部表情之间细微差异角度出发,提出了多头交叉注意力机制(Distract Attention Network, DAN)构建局部特征之间的高阶信息,以表征多个面部区域特征;Liu 等[37]提出了一种自适应多层感知注意网络(Adaptive Multilayer Perceptual attention Network, AMP-Net),通过在不同层级上对特征进行加权和选择,进而从不同层次上捕获面部表情关键信息,更好地聚焦具有判别力的面部区域;EAC(Erasing Attention Consistency)[22]和 ARM(Amending Representation Module)[38]通过构建一种随机擦除方法防止模型过分关注噪声标签,更好地聚焦正确的面部表情类别。

2　本文模型 FAER

针对 FAE 和 FER 任务,本文提出了一个简单且高效的基于 CNN 的人脸分析网络。通过共享骨干网络模块,针对不同任务设计不同分类方法,从而在 FAE 和 FER 任务中获得了优异的性能表现。

2.1　人脸属性估计与表情识别模型

如前所述,面部特征之间的差异主要体现在局部特征区域,如五官、饰品等关键部位;与此同时,全局特征对于区分不同类别具有一定作用,如面部表情中愤怒和惊讶的嘴巴部分、恐惧和惊讶的眼睛部分。考虑到 CNN 在编码图像局部特征具有显著优势,且增加注意力机制能弥补 CNN 编码全局特征的不足,受 ConvNext[15]网络深度和编码图像有效性的启发,为促使模型在编码图像局部特征时学习到更多的深度特征,如图 1 所示,提出了以 ConvNext 为骨干的网络,并添加 CC Attention 机制以增强模型编码全局特征能力。基于上述网络,进一步用人脸属性估计方法和面部表情识别方法进行不同任务学习。

ConvNext 骨干网络能有效编码输入图像的局部特征,并表征相同面部特征之间的区域特征及不同面部之间局部特征的差异性。尽管骨干网络在编码人脸图局部特征差异上具有显著效果,但骨干网络编码后的深度特征无法表征图像的全局特征,受文献[39-40]中编码全局特征有效性的启发,为了促使本文模型能充分编码全局特征,提出了 CC Attention 机制,通过学习不同面部之间的特征差异,表征人脸图像全局和局部特征。

基于骨干网络编码后的深度特征,CC Attention 通过动态自适应调整深度特征每个通道上的权重信息,并对每个通道进行全局池化,以学习深度特征的全局信息;与此同时,注意力机制编码图像静态信息,并融合深度特征的动态全局和静态局部信息,表征不同面部类别之间的特征差异,进而弥补骨干网络在编码图像全局特征能力上的不足。

FAE 和 FER 任务需要学习并区分多个面部局部特征,因此,针对不同任务,本文分别采用不同分类方法进行表征学习。由于 FAE 为二元分类任务,考虑到人脸属性样本存在类别不平衡问题,即易分类样本和难分类样本存在严重不平衡问题,而聚焦损失(Focal Loss)[41]可以有效解决类别不平衡问题,并提高模型对困难样本和错误样本的学习能力,因此,在 FAE 任务中采用 Focal Loss 函数。针对 FER,则采用基于 Softmax 损失的多分类任务交叉熵损失(Cross-Entropy Loss);与此同时,考虑到面部表情类内差异大和类间差异小的特点,本文还采用人脸识别任务中常用的中心损失(Center Loss)[19]以惩罚样本与对应类别中心距离。

图 1 为 FAER 的架构图。首先,FAER 通过骨干网络对人脸进行深度特征编码,以获取人脸的局部特征;然后,CC Attention 通过动态自适应调整通道权重,并对深度特征进行全局池化,以获取全局特征,进而弥补骨干网络编码图像全局特征的不足;最后,针对 FAE 和 FER 任务分别采用不同损失函数,以惩罚深度特征与其对应真实标签之间的距离,进而优化整体网络参数。

2.2　融合动态与静态特征的上下文通道注意力机制

不同人脸的某些区域存在相似性,如 FAE 任务中,蕴含多个二元标签的不同人脸可能存在相似的二元标签,如男性和女性图像都具有眼袋、眼镜和吸引人等标签;FER 任务中,如恐惧、开心和惊讶的嘴巴部分。因此,充分编码人脸多个

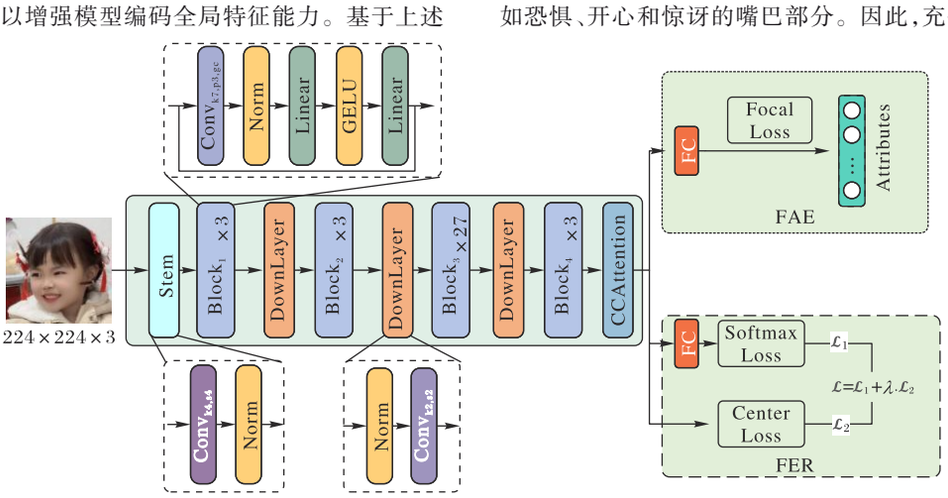


图 1　FAER 的架构

Fig. 1　Architecture of FAER

局部特征，以表征不同类别之间的差异，对提高分类任务的有效性具有重要意义。尽管CNN能根据输入特征本身有效学习多个局部特征信息，但全局特征信息对区分不同类别之间差异仍具有重要意义；考虑到深度特征之间的学习都是相互独立的，且无法充分融合上下文全局和局部信息，为了捕获深度特征之间的相互依赖关系，充分表征输入特征的局部和全局特征，受文献[39-40, 42]的启发，设计了融合动态自适应通道特征信息和静态特征信息的CC Attention，如图2所示。

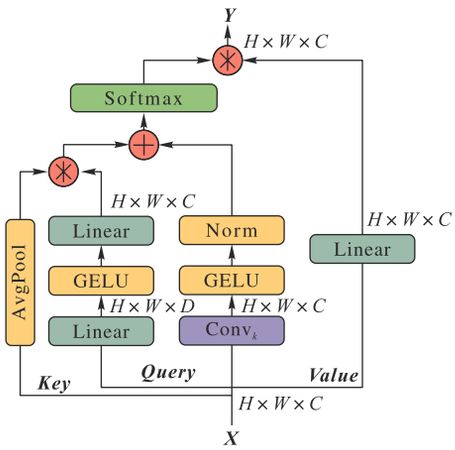


图2 CC Attention机制的结构

Fig. 2 Structure of CC Attention mechanism

假设输入特征 $\boldsymbol{X} \in \mathbb{R}^{H \times W \times C}$，$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 分别定义为如式(1)~(4)所示：

$$\boldsymbol{Q} = \mathrm{LN}\left(\mathrm{GELU}\left(\mathrm{Conv}_{k \times k}(\boldsymbol{X})\right)\right) \in \mathbb{R}^{H \times W \times C} \tag{1}$$

$$\boldsymbol{K} = \mathrm{Linear}\left(\mathrm{GELU}\left(\mathrm{Linear}(\boldsymbol{X})\right)\right) \in \mathbb{R}^{H \times W \times C} \tag{2}$$

$$\boldsymbol{K}' = \boldsymbol{K} \otimes \mathrm{AvgPool}(\boldsymbol{X}) \in \mathbb{R}^{H \times W \times C} \tag{3}$$

$$\boldsymbol{V} = \mathrm{Linear}(\boldsymbol{X}) \in \mathbb{R}^{H \times W \times C} \tag{4}$$

其中：⊗表示哈达玛积操作。不同于自注意力机制通过线性映射获取 $\boldsymbol{K}$ 矩阵，CC Attention 首先压缩通道数，然后由GELU(Gaussian Error Linear Unit)函数激活深度特征，最后再扩展到初始通道数，该方法能动态自适应调整通道上权重信息，以获得动态信息矩阵 $\boldsymbol{K}$，动态信息矩阵 $\boldsymbol{K}$ 与全局池化后的输入特征 $\boldsymbol{X}$ 进行哈达玛积，以编码动态全局信息，进而获得动态全局信息矩阵 $\boldsymbol{K}'$。通过 $k \times k$ 滤波器对相邻 k^2 范围特征进行卷积操作，以GELU和LayerNorm(Layer Normalization)获取特征矩阵 $\boldsymbol{X}$ 中静态局部特征信息矩阵 $\boldsymbol{Q}$。

$$\boldsymbol{R} = \mathrm{Softmax}(\boldsymbol{K}' + \boldsymbol{Q}) \in \mathbb{R}^{H \times W \times C} \tag{5}$$

与此同时，通过静态局部信息矩阵 $\boldsymbol{Q}$ 和动态全局信息矩阵 $\boldsymbol{K}'$ 学习输入特征 $\boldsymbol{X}$ 的上下文重要信息矩阵 $\boldsymbol{R}$，而非基于独立的 $\boldsymbol{Q}$ 和 $\boldsymbol{K}$ 矩阵，该方法在学习输入特征的上下文动态全局和静态局部信息中提供了强有力的指导。聚合注意力矩阵 $\boldsymbol{R}$ 和 $\boldsymbol{V}$ 以计算特征信息矩阵 $\boldsymbol{A}$，形同自注意力机制操作：

$$\boldsymbol{A}(\boldsymbol{X}) = \boldsymbol{R} \otimes \boldsymbol{V} \in \mathbb{R}^{H \times W \times C} \tag{6}$$

CC Attention机制能充分学习输入特征的动态全局和静态局部信息，进而弥补骨干网络编码全局特征能力的不足，促使模型充分关注输入图像的全局和局部特征，经过CC Attention编码后的输出特征为 $\boldsymbol{X}_{\mathrm{CC}} \in \mathbb{R}^{H \times W \times C}$。

2.3 不同分类任务训练策略

针对不同任务，采用不同的学习框架，使模型可以分别解决FAE任务和FER任务，骨干网络和CC Attention用于编码输入图像的局部特征和全局特征。FAE和FER任务的训练和测试时的损失函数如下所示。

FAE任务中，人脸属性估计使用CelebA数据集[5]进行训练和测试，它由40个二分类标签组成。考虑到Focal Loss[41]在二元分类任务能解决类别样本不平衡问题，更关注难以分类的样本，因此本文采用Focal Loss作为损失函数。假设输入人脸图像为 $\boldsymbol{X} \in \mathbb{R}^{H \times W \times C}$，模型预测概率标签为 p，对应真实标签为 q，则总的损失函数 $\mathcal{L}_{\mathrm{A}}$ 为：

$$\mathcal{L}_{\mathrm{A}} = \frac{1}{N}\sum_{j=1}^{B}\sum_{i=1}^{N}\Big[-(1-p_i)^{\gamma} \cdot q_i \cdot \log(p_i) - p_i^{\gamma} \cdot (1-q_i) \cdot \log(1-p_i)\Big] \tag{7}$$

其中：q_i 为第 i 个样本对应的真实标签，若第 j 个属性对应的真实标签存在则 q_i=1，否则 q_i=0；p_i 为第 i 个输入图像所包含第 j 个属性的预测概率；B 为二元分类任务标签数；N 为每训练批次样本数；γ 为可调参数，默认为2。

FER任务中，面部表情识别标签为多分类问题，故采用交叉熵损失(Cross-Entropy Loss, CE Loss)，假设输入人脸图像为 $\boldsymbol{X} \in \mathbb{R}^{H \times W \times C}$，模型预测概率标签为 p，对应的真实标签为 y，则交叉熵损失函数 $\mathcal{L}_{\mathrm{CE}}$ 为：

$$\mathcal{L}_{\mathrm{CE}} = -\frac{1}{N}\sum_{j=1}^{M}\sum_{i=1}^{N} y_{ji} \cdot \log(p_{ji}) \tag{8}$$

其中：y_{ji} 为第 i 个样本的标签为 j 时所对应的真实标签，若第 i 个样本属于表情类别 M 则 y_{ji}=1，否则 y_{ji}=0；p_{ji} 为第 i 个样本属于表情类别 j 的概率；M 为多分类任务标签数；N 为每训练批次样本数。

与此同时，考虑到面部表情具有类内差异大和类间差异小的问题，受Center Loss[19]在惩罚样本与其对应中心距离上优越性的启发，本文同时采用Center Loss作为辅助多标签分类损失函数，基于输入图像 $\boldsymbol{X} \in \mathbb{R}^{H \times W \times C}$，模型在编码后的深度特征为 $\boldsymbol{X}_f \in \mathbb{R}^{H \times W \times C}$，则 $\mathcal{L}_{\mathrm{CL}}$ 可以定义为：

$$\mathcal{L}_{\mathrm{CL}} = \frac{1}{2N}\sum_{i=1}^{N}\left\| \boldsymbol{X}_{f_i} - C_{y_i} \right\|_2^2 \tag{9}$$

其中：$\boldsymbol{X}_{f_i}$ 为第 i 个样本编码后的深度特征；C_{y_i} 为对应中心；N 为每训练批次样本数。本文中FER任务的损失函数由CE Loss和以 λ 为比例的Center Loss构成：

$$\mathcal{L}_{\mathrm{E}} = \mathcal{L}_{\mathrm{CE}} + \lambda \cdot \mathcal{L}_{\mathrm{CL}} \tag{10}$$

针对FAE和FER任务，分别采用 $\mathcal{L}_{\mathrm{A}}$ 和 $\mathcal{L}_{\mathrm{E}}$ 损失函数计算模型预测标签与对应真实标签之间的差异，并更新网络整体参数。

3 实验与结果分析

3.1 数据集与评价指标

为了评估FAER的优越性，针对FAE和FER任务分别选取人脸属性数据集CelebA[5]和面部表情数据集AffectNet[17]、RAF-DB[18]。表1为相关数据集的统计结果。

表1 人脸属性估计和面部表情识别数据集

Tab. 1 Datasets for facial attribute estimation and facial expression recognition

任务	数据集	样本数		特征描述
		训练集	测试集	
FAE	CelebA	162 770	19 962	40种二元人脸属性类别
FER	RAF-DB	12 271	3 068	7种基础表情类别
	AffectNet	283 901	3 500	7种基础表情类别

在人脸属性估计和面部表情识别任务中,准确率和平均准确率指标分别被用来评估模型在FER和FAE任务的性能。类别准确率能体现模型在相应类别上的表现。此外,梯度图可视化能直观展示模型聚焦点。

3.2　实验设置

在Ubuntu18.04系统上使用PyTorch 1.12.0框架实现的代码,并在两张NVIDIA A40 GPU上运行了所有实验。

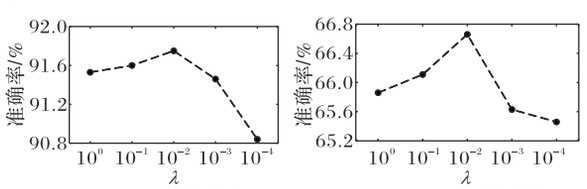
在MS-Celeb-1M数据集[43]上对ConvNext[15]进行预训练,并保存权重文件;基于ConvNext网络,构建如图1所示的FAER模型架构,并在训练时更新整体网络参数。

使用AdamW优化器对模型参数进行优化,批大小设置为256,使用ExponentialLR学习策略并设置衰减因子为0.95,依概率使用随机水平翻转数据增强策略。此外,考虑到FAE和FER任务的不同,在FAE任务中学习率设为10^{-3},并训练60个轮次,在FER任务中学习率设为2.5×10^{-4}。对RAF-DB数据集训练120个轮次,对AffectNet数据集训练60个轮次。

3.3　消融实验

3.3.1　FER超参数λ对模型性能影响

在FER中为了确定不同超参数λ条件下Center Loss对模型性能的影响,即λ在取不同值时,FER任务识别准确率的表现,在$\{10^0,10^{-1},10^{-2},10^{-3},10^{-4}\}$中分别取不同数值并在RAF-DB和AffectNet数据集上进行消融实验,图3为超参数λ的实验统计结果。



(a) RAF-DB数据集　　(b) AffectNet数据集

图3　超参数λ在AffectNet和RAF-DB数据集上对准确率的影响

Fig. 3　Influence of hyperparameter λ on accuracy on AffectNet and RAF-DB datasets

从图3可看出:超参数λ在10^0至10^{-2}上表现出准确率上升趋势,在10^{-2}至10^{-4}上表现出准确率下降趋势,在10^{-2}时准确率最高,因此本文超参数λ选择为10^{-2}。

3.3.2　相关模块对FER任务准确率的影响

为了考察CC Attention和Center Loss对本文模型在FER任务准确率的影响,根据实验设置,分别对CC Attention和Center Loss进行消融实验。实验统计结果如表2所示。

表2　不同模块的消融实验结果　　单位:%

Tab. 2　Ablation experimental results of different modules　　unit: %

模块	准确率	
	AffectNet	RAF-DB
Baseline	64.77	90.74
Baseline+ CC Attention	65.03	90.91
Baseline+ Center Loss	**66.66**	**91.75**

从表2可看出:在AffectNet数据集上,Baseline模型取得64.77%的准确率;增加CC Attention时准确率提升了0.26个百分点;增加Center Loss时,准确率进一步提升了1.63个百分点。与此同时,在RAF-DB数据集上,Baseline取得90.74%的准确率;增加CC Attention时,准确率提升了0.17个百分点;增加Center Loss时,准确率进一步提升了0.84个百分点。

正是因为CC Attention编码图像的全局特征,弥补骨干网络编码图像全局特征不足问题;Center Loss通过惩罚样本与其对应中心距离,有效解决了FER任务类内差异大和类间差异小的问题。

3.3.3　FAE任务中损失函数的选择

考虑到FAE为二元分类任务,为了确定不同损失函数对FAE任务性能影响,本文分别采用二元分类任务常用的Asymmetric Loss[44]、BCE Loss[45]和Focal Loss[41]函数,表3里不同损失函数下FAER模型在CelebA数据集上性能表现。

表3　不同损失函数的平均准确率对比　　单位:%

Tab. 3　Comparison of average accuracies with different loss functions　　unit: %

损失函数	平均准确率	
	Baseline	FAER
Asymmetric Loss	90.69	91.23
BCE Loss	91.03	91.71
Focal Loss	**91.66**	**91.87**

观察表3可知,在二分类任务上,基于Focal Loss函数的Baseline和FAER的平均准确率最高。出现上述差异的主要原因是,Asymmetric Loss函数更集中于正负样本分类错误的问题,并且BCE Loss函数常用于衡量标签和预测之间得到差异,不关注数据分布类型;相反,Focal Loss通过引入调节因子,促使模型更关注难以分类样本。

3.4　实验结果与分析

本文模型与MCNN-AUX[30]、MCFA[31]、DMM-CNN[32]、SwinFace[10]和Baseline在FAE任务中类别准确率和平均准确率对比结果如图4所示,FER任务中的准确率对比统计结果如表4所示。

表4　在RAF-DB和AffectNet数据集上FER任务的准确率表现比较　　单位:%

Tab. 4　Accuracy performance comparison for FER tasks on RAF-DB and AffectNet datasets　　unit: %

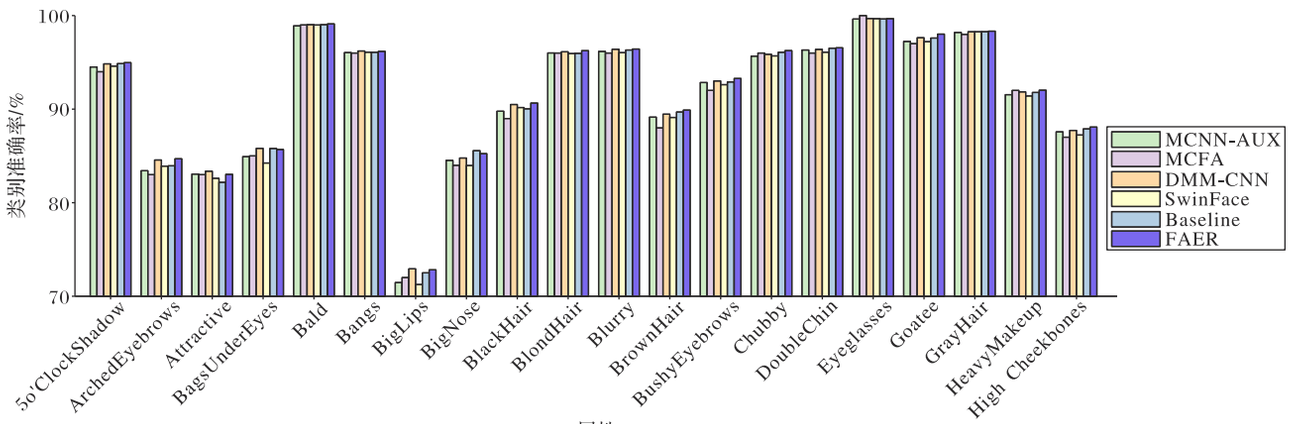
模型	准确率	
	RAF-DB	AffectNet
DACL[21]	87.87	65.20
VTFF[26]	88.14	61.85
EfficientFace[34]	88.36	63.70
MA-Net[35]	88.42	64.53
PSR[28]	88.98	63.77
AMP-Net[37]	89.25	64.54
DAN[36]	89.70	65.69
EAC[22]	90.35	65.32
ARM[38]	90.42	65.20
TransFER[27]	90.91	66.23
Baseline	90.74	64.77
FAER	**91.75**	**66.66**

本文FAER和Baseline模型在RAF-DB和AffectNet两个数据集上的类别准确率统计结果如表5所示,深度特征可视化效果如图5~6所示,直观地展示了模型所关注输入图像的特征。

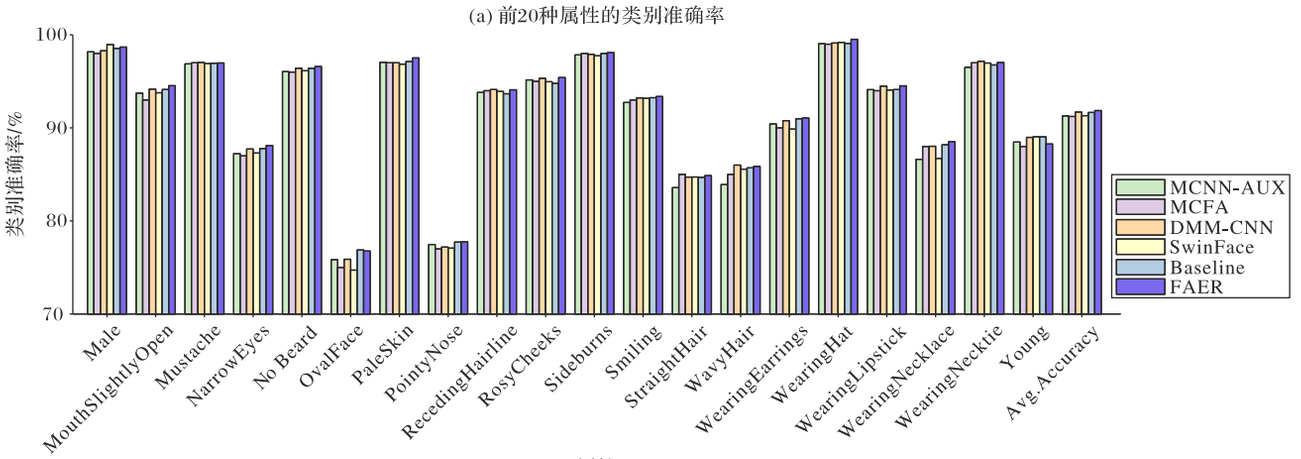
从图4可知:在FAE任务中,FAER在人脸属性CelebA数据集上取得了最高的平均准确率表现,优于MCNN-AUX、MCFA、DMM-CNN、SwinFace和Baseline,并且在绝大部分类别上表现出高于对比模型准确率性能;相较于SwinFace所取得91.32%的平均准确率高出0.55个百分点;Baseline模型

在绝大多数类别准确率以及平均准确率上高于对比方法；因此，FAER 模型在简化 FAE 任务的同时，仍然表现出优越的性能。

从表 4 可以看出：在 FER 任务中，本文 FAER 模型的准确率在 RAF-DB 和 AffectNet 数据集上分别取得了最高的 91.75% 和 66.66%；在 RAF-DB 数据集上，本文模型的准确率比 Baseline、TransFER、ARM 和 EAC 分别高出 1.01、0.84、1.33 和 1.40 个百分点。在 AffectNet 数据集上，本文模型的准确率比 Baseline、TransFER、DAN 和 EAC 分别高出 1.89、0.43、0.97 和 1.34 个百分点。从表 5 可以看出：在 RAF-DB 和 AffectNet 数据集的类别准确率上，本文 FAER 模型在超过半数的类别上表现出类别准确率提升的性能表现。



(a) 前20种属性的类别准确率



(b) 后20种属性的类别准确率和平均准确率

图 4 CelebA 数据集上 FAE 任务的类别准确率和平均准确率

Fig. 4 Class-wise accuracy and average accuracy of FAE tasks on CelebA dataset

表 5 在 RAF-DB 和 AffectNet 数据集上 FER 任务的类别准确率统计结果 单位：%

Tab. 5 Class-wise accuracy statistics for FER tasks on RAF-DB and AffectNet datasets unit: %

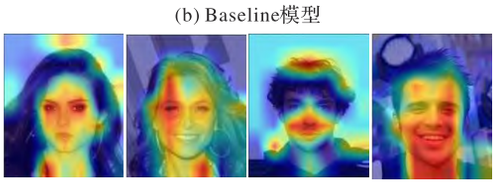
数据集	模型	类别准确率						
		愤怒	厌恶	恐惧	开心	自然	悲伤	惊讶
RAF-DB	Baseline	88.20	72.80	75.39	95.74	90.86	87.65	89.97
	FAER	**90.06**	**79.62**	72.37	**97.11**	86.99	**92.94**	**92.72**
AffectNet	Baseline	61.39	66.18	67.23	77.20	53.05	71.07	59.26
	FAER	**62.25**	**72.86**	**70.22**	**80.97**	**56.57**	66.33	**59.93**

注：加粗数值表示类内差异已经缩小。

从图 5 可以看出：在 CelebA 数据集上，FAER 模型能关注更多的面部细粒度特征，如眼睛、项链和发型等主要属性标签，且关注的人脸属性特征更聚焦；相比之下，Baseline 模型关注的细粒度特征不够充分，它关注的重点区域面积相较于 FAER 模型较少，且无法关注除脸部以外的属性标签，如项链、发型和领带等，在人脸属性特征上不够聚焦。

(a) 输入图像

(b) Baseline模型

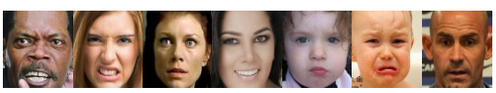


(c) FAER模型

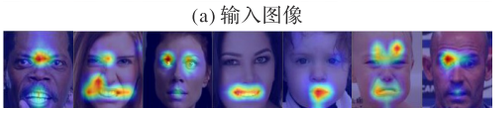
图 5 Grad-CAM[46]在 CelebA 数据集上的可视化

Fig. 5 Grad-CAM visualization on CelebA dataset

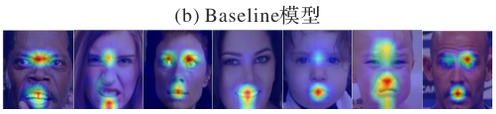
从图6可看出:在AffectNet数据集上,FAER方法能关注面部表情的多个局部特征,如愤怒和厌恶表情的眉心和嘴巴部分,更关注面部精细特征,如愤怒的嘴巴部分和恐惧的下巴部分;相反,尽管Baseline方法也能捕捉面部多细粒度特征,但在人脸多个区域特征上,表现出无法同时关注多个特征,且多特征上表现不佳,如恐惧和惊讶的关注点较为相似,厌恶和悲伤类别的关注点较为分散。



(a) 输入图像



(b) Baseline模型



(c) FAER模型

图6　Grad-CAM[46]在AffectNet数据集上的可视化

Fig. 6　Grad-CAM visualization on AffectNet dataset

综上,FAER在准确率、平均准确率、类别准确率指标上均优于对比模型,相较于对比模型,FAER在多个数据集上均取得了最高的准确率性能;对比基线模型,FAER在绝大多数类别上都表现出准确率上升的性能;可视化结果可以看出,FAER在捕捉多个面部细粒度特征以及不同类别之间差异上具有显著优势。

4　结语

人脸特征蕴含着丰富的信息,基于深度学习的人脸特征识别对于理解人脸属性和面部表情具有重要意义。因此,解决人脸属性估计和表情识别任务中存在的难题,提高模型识别准确率性能,洞察人脸特征之间所隐含的深层丰富信息显得尤为重要。从人脸特征的全局和局部特征角度出发,提出了基于上下文通道注意力机制的人脸属性估计和表情识别(FAER)模型。首先,利用ConvNext编码图像局部特征的有效性,促使骨干网络充分编码图像细粒度特征;其次,提出了上下文通道注意力机制,编码人脸图像的全局特征,弥补骨干网络编码全局特征能力的不足;最后,针对不同任务标签采用不同损失函数,以引导模型学习具有鉴别力的人脸特征。由于人脸蕴含着丰富的信息,未来会继续研究多模态多任务人脸特征分类,以挖掘更多人脸信息。本文的代码链接如下:https://github.com/XUJ1Er/FAER/。

参考文献(References)

[1] 张晓行,田启川,廉露,等. 人脸关键点检测研究综述[J]. 计算机工程与应用, 2024, 60(12): 48-60. (ZHANG X H, TIAN Q C, LIAN L, et al. Review of research on facial landmark detection [J]. Computer Engineering and Applications, 2024, 60(12): 48-60.)

[2] 张波,兰艳亭,鲜浩,等. 基于通道注意力机制的人脸表情识别机器人交互研究[J]. 电子测量技术, 2021, 44(11): 169-174. (ZHANG B, LAN Y T, XIAN H, et al. Research on robot interaction of facial expression recognition based on channel attention mechanism [J]. Electronic Measurement Technology, 2021, 44(11): 169-174.)

[3] LIN J, LI Y, YANG G. FPGAN: face de-identification method with generative adversarial networks for social robots [J]. Neural Networks, 2021, 133: 132-147.

[4] LIU Z, LUO P, WANG X, et al. Large-scale CelebFaces attributes (CelebA) dataset [EB/OL]. [2023-10-22]. https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html.

[5] LIU Z, LUO P, WANG X, et al. Deep learning face attributes in the wild [C]// Proceedings of the 2015 IEEE International Conference on Computer Vision. Piscataway: IEEE, 2015: 3730-3738.

[6] HAN H, JAIN A K, WANG F, et al. Heterogeneous face attribute estimation: a deep multi-task learning approach [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(11): 2597-2609.

[7] HE S, LUO H, WANG P, et al. TransReID: Transformer-based object re-identification [C]// Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2021: 14993-15002.

[8] NGUYEN H M, LY N Q, PHUNG T T T. Large-scale face image retrieval system at attribute level based on facial attribute ontology and deep neuron network [C]// Proceedings of the 2018 Asian Conference on Intelligent Information and Database Systems, LNCS 10752. Cham: Springer, 2018: 539-549.

[9] CAO J, LI Y, ZHANG Z. Partially shared multi-task convolutional neural network with local constraint for face attribute learning [C]// Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 4290-4299.

[10] QIN L, WANG M, DENG C, et al. SwinFace: a multi-task transformer for face recognition, expression recognition, age estimation and attribute estimation [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2024, 34(4): 2223-2234.

[11] LI W, CAO Z, FENG J, et al. Label2Label: a language modeling framework for multi-attribute learning [C]// Proceedings of the 2022 European Conference on Computer Vision, LNCS 13672. Cham: Springer, 2022: 562-579.

[12] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: Transformers for image recognition at scale [EB/OL]. [2023-10-20]. https://arxiv.org/pdf/2010.11929.pdf.

[13] 戴国庆,张晟磊,袁玉波. 老龄面部数据抽取的肤色显著性方法[J]. 计算机应用, 2022, 42(S2): 217-223. (DAI G Q, ZHANG S L, YUAN Y B. Aged facial dada extraction method by using skin color saliency [J]. Journal of Computer Applications, 2022, 42(S2): 217-223.)

[14] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition [C]// Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2016: 770-778.

[15] LIU Z, MAO H, WU C Y, et al. A ConvNet for the 2020s [C]// Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 11966-11976.

[16] EKMAN P, FRIESEN W V. Constants across cultures in the face and emotion [J]. Journal of Personality and Social Psychology, 1971, 17(2): 124-129.

[17] 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.

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

[19] WEN Y, ZHANG K, LI Z, et al. A discriminative feature learning approach for deep face recognition [C]// Proceedings of the 2016 European Conference on Computer Vision, LNCS 9911. Cham: Springer, 2016: 499-515.

[20] WAN W, ZHONG Y, LI T, et al. Rethinking feature distribution for loss functions in image classification [C]// Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 9117-9126.

[21] FARZANEH A H, QI X. Facial expression recognition in the wild via deep attentive center loss [C]// Proceedings of the 2021 IEEE Winter Conference on Applications of Computer Vision. Piscataway: IEEE, 2021: 2401-2410.

[22] ZHANG Y, WANG C, LING X, et al. Learn from all: erasing attention consistency for noisy label facial expression recognition [C]// Proceedings of the 2022 European Conference on Computer Vision, LNCS 13686. Cham: Springer, 2022: 418-434.

[23] 刘希未,宫晓燕,赵红霞,等. 基于混合注意力机制的动态人脸表情识别[J]. 计算机应用, 2023, 43(S1): 1-7. (LIU X W, GONG X Y, ZHAO H X, et al. Dynamic facial expression recognition based on hybrid attention mechanism [J]. Journal of Computer Applications, 2023, 43(S1): 1-7.)

[24] FERNANDEZ P D M, PEÑA F A G, REN T I, et al. FERAtt: facial expression recognition with attention net [C]// Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Piscataway: IEEE, 2019: 837-846.

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

[26] MA F, SUN B, LI S. Facial expression recognition with visual transformers and attentional selective fusion [J]. IEEE Transactions on Affective Computing, 2023, 14(2): 1236-1248.

[27] XUE F, WANG Q, GUO G. TransFER: learning relation-aware facial expression representations with Transformers [C]// Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2021: 3581-3590.

[28] VO T H, LEE G S, YANG H J, et al. Pyramid with super resolution for in-the-wild facial expression recognition [J]. IEEE Access, 2020, 8: 131988-132001.

[29] RUDD E M, GÜNTHER M, BOULT T E. MOON: a mixed objective optimization network for the recognition of facial attributes [C]// Proceedings of the 2016 European Conference on Computer Vision, LNCS 9909. Cham: Springer, 2016: 19-35.

[30] HAND E M, CHELLAPPA R. Attributes for improved attributes: a multi-task network utilizing implicit and explicit relationships for facial attribute classification [C]// Proceedings of the 31st AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2017: 4068-4074.

[31] ZHUANG N, YAN Y, CHEN S, et al. Multi-task learning of cascaded CNN for facial attribute classification [C]// Proceedings of the 24th International Conference on Pattern Recognition. Piscataway: IEEE, 2018: 2069-2074.

[32] MAO L, YAN Y, XUE J H, et al. Deep multi-task multi-label CNN for effective facial attribute classification [J]. IEEE Transactions on Affective Computing, 2022, 13(2): 818-828.

[33] SAVCHENKO A V. Facial expression and attributes recognition based on multi-task learning of lightweight neural networks [C]// Proceedings of the IEEE 19th International Symposium on Intelligent Systems and Informatics. Piscataway: IEEE, 2021: 119-124.

[34] ZHAO Z, LIU Q, ZHOU F. Robust lightweight facial expression recognition network with label distribution training [C]// Proceedings of the 35th AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2021: 3510-3519.

[35] ZHAO Z, LIU Q, WANG S. Learning deep global multi-scale and local attention features for facial expression recognition in the wild [J]. IEEE Transactions on Image Processing, 2021, 30: 6544-6556.

[36] WEN Z, LIN W, WANG T, et al. Distract your attention: multi-head cross attention network for facial expression recognition [J]. Biomimetics, 2023, 8(2): No. 199.

[37] LIU H, CAI H, LIN Q, et al. Adaptive multilayer perceptual attention network for facial expression recognition [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(9): 6253-6266.

[38] SHI J, ZHU S, LIANG Z. Learning to amend facial expression representation via de-albino and affinity [EB/OL]. [2023-10-22]. https://arxiv. org/pdf/2103. 10189. pdf.

[39] HU J, SHEN L, SUN G. Squeeze-and-excitation networks [C]// Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 7132-7141.

[40] LI Y, YAO T, PAN Y, et al. Contextual Transformer networks for visual recognition [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(2): 1489-1500.

[41] LIN T, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection [C]// Proceedings of the 2017 IEEE International Conference on Computer Vision. Piscataway: IEEE, 2017: 2999-3007.

[42] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2017: 6000-6010.

[43] GUO Y, ZHANG L, HU Y, et al. MS-Celeb-1M: a dataset and benchmark for large-scale face recognition [C]// Proceedings of the 2016 European Conference on Computer Vision, LNCS 9907. Cham: Springer, 2016: 87-102.

[44] RIDNIK T, BEN-BARUCH E, ZAMIR N, et al. Asymmetric loss for multi-label classification [C]// Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2021: 82-91.

[45] MAHBUB U, SARKAR S, CHELLAPPA R. Segment-based methods for facial attribute detection from partial faces [J]. IEEE Transactions on Affective Computing, 2020, 11(4): 601-613.

[46] GILDENBLAT J. PyTorch library for CAM methods [EB/OL]. [2023-10-22]. https://github. com/jacobgil/pytorch-grad-cam.

This work is partially supported by National Natural Science Foundation of China (62373116, 62163007); Guizhou Province Science and Technology Program (Qiankehe Zhicheng [2023] Yiban 118, Qiankehe Pingtairencai [2020]6007-2).

XU Jie, born in 1997, M. S. candidate. His research interests include intelligent autonomous system.

ZHONG Yong, born in 1966, Ph. D., research fellow. His research interests include big data and its intelligent processing, cloud computing, software engineering.

WANG Yang, born in 1987, Ph. D., senior engineer. His research interests include artificial intelligence, computer vision, intelligent analysis of big data.

ZHANG Changfu, born in 1990, senior engineer. His research interests include industrial big data, artificial intelligence.

YANG Guanci, born in 1983, Ph. D., professor. His research interests include intelligent autonomous system and robotics, multimodal data-driven cognitive computing.