第3章 表情识别

CHAPTER 3

生活中，人们除了通过语言进行交流以外，表情也是非常重要的交流方式之一。从微笑到皱眉，从愉快到不满，从平静到悲伤，表情能够传达出很多信息。即使在无声的世界里，仅通过解读表情，也可以快速了解人们的心理变化和内在需求。

表情识别技术是一种通过客观分析人脸图像，进而判定人类情绪类别的计算机视觉技术。随着科技的进步，人们希望机器在按部就班完成指定任务的同时，还能主动挖掘用户需求，即表现得更加智能化。例如，医疗机器在完成医疗任务的同时还能够自动记录病人的生理状况和就医感受，在线教育平台在完成教授课程任务的同时还能够自动检测学生的专注度和心理活动等。将表情识别技术配置到机器系统中，可使机器学会“察言观色”，实现更为人性化的服务。

利用表情识别技术，能让机器具有“情绪交互”的功能，不需要使用肢体和语言，仅凭开心、惊讶和悲伤等表情即可自动触发相应指令，使机器做出贴心的反馈。接下来，让我们一起学习表情识别技术，尝试教会机器“读心”吧。

3.1 背景介绍

面部表情是传递人类情绪内涵的最常见、最直观和最重要的媒介之一。表情识别技术(Facial Expression Recognition，FER)能够辅助机器实现更为人性化的人机交互。表情识别系统在智能医疗、智能服务、驾驶员疲劳检测等领域均有实际应用。在医疗看护领域，FER 技术可以通过识别病患的表情来协助医护人员诊断患者的生理状况，提升医务工作者的决策效率和病人的就诊体验；在安全驾驶领域，FER 技术通过识别驾驶者的表情可判断驾驶人员是否有疲劳驾驶行为，提高道路交通的安全性；在虚拟现实领域，FER 技术可以通过识别用户的表情实现人机同步的沉浸式真实感体验；在行为分析领域，FER 技术可以识别城市中智能摄像头所拍摄的目标人员的表情，结合相应的犯罪预测模型，实现对路人异常行为的监测；在服务领域，FER 技术通过识别客户的表情不仅能推断客户对于该服务的满意度，还能估测出客户的消费喜好与习惯。

面部表情识别旨在利用计算机科学方法，首先对采集到的自然图像进行人脸检测和定位裁剪，再对图像进行预处理，例如，人脸对齐、数据增强和标准化等操作，然后使用传统机器学习算法或深度学习方法获取表情特征，最后通过分类器客观分析特征以判定图像中待

测人员的情绪，情绪的判定结果可以是离散型、复合型、连续型或脸部动作单元组合型的。相较于其他类型的表情定义来说，离散型的表情更为典型且易于预测，故大多数现有方法均针对离散型表情进行研究，常见的表情类别有愤怒、厌恶、开心、恐惧、悲伤、惊讶、中性和轻蔑。基于计算机科学技术的面部表情识别系统的一般工作流程如图 3.1 所示。

图 3.1　面部表情识别系统的一般工作流程

早期基于传统算法的表情识别的研究思路，主要是将人工设计特征与常见分类器进行结合，通过建模完成表情识别任务。常用的传统人工特征有局部二值模式、Haar 特征、方向梯度直方图、尺度不变特征变换和 Gabor 变换等；常见的分类模型有支持向量机、K 近邻算法、贝叶斯算法和 Adaboost 算法等。一个有效的识别模型往往需要经过研究人员反复斟酌设计、合理调整结构和重复验证性能才能得到，因而基于特征设计的传统表情识别方法具有较好的逻辑性以及较为清晰的可解释性。但是，随着户外表情数据集规模的不断扩大，面对各种姿态偏移、遮挡和光照变化的表情图像，仅靠人力去获取具有区分性的特征是十分低效的。

随着高性能硬件设备的迭代、深度学习算法的创新与大数据技术的发展，卷积神经网络(Convolutional Neural Networks，CNN)在智能感知与图像理解任务中的优异性能逐渐显现。面对复杂多样的实际应用场景以及庞大的表情数据集，针对不同的理论和数据先验知识，设计并叠加相应的网络模块、训练机制以及损失函数，可以使网络按照各种期望来学习和提取更加显著和精练的表情特征，实现具有较高效率和较高性能的表情识别。

特征提取在表情识别流程中十分关键，它将直接影响到表情识别算法性能的优劣。Xie 等[1]提出了一种结合注意力机制与编解码重构特征对比的方法以获得深层高级泛化特征，抑制不同种族、身份、性别给表情识别任务带来的负面影响。Wang 等[2]提出了一种基于区域注意力网络的表情识别方法，输入裁剪图像获取局部特征，再利用注意力机制计算区域权重，并与全局特征进行加权融合以达到特征增强的目的，主要解决姿态变化和有遮挡情况下的人脸表情识别问题。Xue 等[3]提出了一种结合 Transformer 与卷积神经网络的结构来提取与表情相关的特征，实现了较高的识别准确率。

损失函数作为指导表情识别网络学习和有效提升表情识别精度的重要因素，也受到了研究学者们的关注，各种构思巧妙的损失函数被相继提出和应用。Vo 等[4]提出了一种多尺度特征融合与先验分布标签平滑损失函数相结合的方法，在计算交叉熵损失时，既考虑被划分成正确类别的损失，还以先验分布概率为权重，通过加权来融合被分类至其他错误类别的损失。Farzaneh 等[5]提出了一种深度注意中心损失，该方法采用多头注意力机制，计算每个特征向量与中心特征向量之间对应位置上的元素之间距离的权重，联合加权中心损失与交叉熵损失来训练网络。Fard 等[6]提出了自适应相关损失，分别从个体特征层面和群体特征层面计算各个维度下特征之间的相关性，利用特定的惩罚机制指导网络生成类内样本高相关而类间样本低相关的特征向量。

3.2 算法原理

本实验基于2022年ECCV论文*Learn From All: Erasing Attention Consistency for Noisy Label Facial Expression Recognition*开展，文中提出了Erasing Attention Consistency (EAC)模型，该模型利用不平衡的框架，结合类激活映射与注意力一致性，抑制噪声数据给网络识别性能所带来的负面影响，提升表情识别准确率。

3.2.1 ResNet介绍

ResNet[7]是由何凯明等提出的一种深度学习卷积神经网络模型，其解决了随着卷积神经网络深度的增加，训练集的识别准确率反而下降的退化问题。ResNet在ILSVRC&COCO2015竞赛中斩获了包含ImageNet分类任务、ImageNet检测任务、ImageNet定位任务、COCO检测任务及COCO分割任务在内的五项“第一”，是CNN史上一大里程碑模型。

一般直觉认为深度神经网络的性能与其宽度和深度均成正比，深层网络的表现一般都比浅层网络更优，或二者性能相近，但是实验结果却表明随着网络层数的逐渐增加，模型在训练集上的识别准确率却趋近于饱和甚至下降，这种反直觉的退化问题表明并不是所有模型都是易于优化的。

针对上述退化问题，ResNet引入了深度残差学习框架。当网络模型出现退化问题时，浅层网络性能优于深层网络，那么将深层网络中的某些层仅进行恒等映射，则深层网络也会获得和浅层网络一样好的性能。然而通过具有非线性映射操作的卷积神经网络模块来直接拟合恒等函数是较为困难的，然而，通过非线性层来拟合恒等于0的函数则相对容易一些。

在原始卷积神经网络模型中，设x为输入，$H(x)$为所期望的底层映射，则恒等函数表示为$H(x)=x$，由于非线性层的存在，这种恒等函数的拟合是比较困难的。在深度残差学习框架中，设底层映射$H(x)=F(x)+x$，则残差函数映射$F(x)=H(x)-x$，此时恒等函数依然为$H(x)=x$，当$F(x)=0$时，该恒等函数成立。即使有非线性层的存在，这种恒等于0的函数的拟合也较为容易。为构建满足$F(x)=H(x)-x$形式的残差函数映射，何凯明等设计了如图3.2所示的残差结构。其中，x为残差模块的输入，$F(x)$为残差映射，$H(x)$为所期望的底层映射输出，ReLU为激活函数，权重层为若干卷积层。

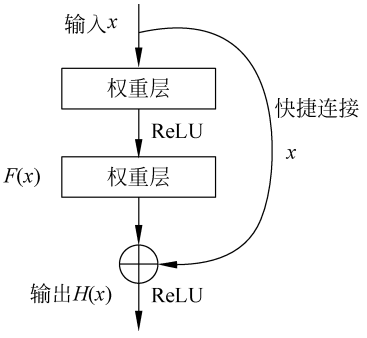


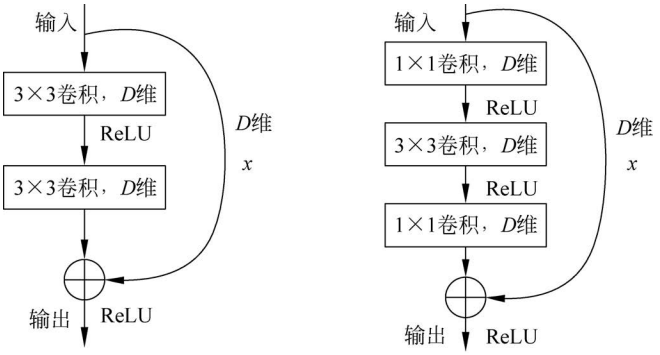
图3.2 残差模块结构图

通常$F(x)$不为0且包含一定特征信息，引入残差结构后网络对于数据的微小变化更加敏感，网络在易于训练的同时性能也会有所提升。使用跳接连接能够简单辅助实现恒等映射，不会产生额外参数且不增加网络的计算复杂度。在残差模块中，x和$F(x)$的维度必须是相等的，否则需要对x进行变换实现维度匹配，此时底层映射$H(x)$可表示为

$$H(x)=F(x)+W_s x \tag{3-1}$$

其中，W_s表示变换参数。

残差函数 $F(x)$是可变的，为了更好地优化不同深度的网络，何凯明等提出了两种不同的基本残差结构：基本模块和瓶颈模块。其中基本模块的输入、输出维度一致，常用于构建相对较浅的网络，而瓶颈模块的输入、输出维度不一，通过 1×1 卷积调整特征维度以降低计算复杂度，常用于构建相对较深的网络。ResNet 的基本模块和瓶颈模块的结构分别如图 3.3(a)和图 3.3(b)所示，其中，D 为大于 1 的常数。



(a) 基本模块结构图　　(b) 瓶颈模块结构图

图 3.3　ResNet 的基本模块和瓶颈模块结构图

ResNet 的设计主要遵循两个原则：一是若输出特征图具有相同尺寸，则此类残差模块具有相同数量的卷积核；二是若输出特征图尺寸减半，则残差模块的卷积核数量加倍，这样便能让每个残差模块的时间复杂度保持一致。当输入和输出的特征图大小不同时，则通过调节步长并进行卷积操作实现特征图尺寸及通道的匹配以便进行跳接连接。ResNet 的细节及其变体结构如表 3.1 所示。

表 3.1　ResNet 的细节及其变体结构

模块	尺寸	18 层	34 层	50 层	101 层	152 层
conv1	112×112	卷积核尺寸为 7，卷积核数量为 64，卷积步长为 2				
conv2	56×56	最大池化层				
		$\begin{bmatrix}3\times3,64\\3\times3,64\end{bmatrix}\times2$	$\begin{bmatrix}3\times3,64\\3\times3,64\end{bmatrix}\times3$	$\begin{bmatrix}1\times1,64\\3\times3,64\\1\times1,256\end{bmatrix}\times3$	$\begin{bmatrix}1\times1,64\\3\times3,64\\1\times1,256\end{bmatrix}\times3$	$\begin{bmatrix}1\times1,64\\3\times3,64\\1\times1,256\end{bmatrix}\times3$
conv3	28×28	$\begin{bmatrix}3\times3,128\\3\times3,128\end{bmatrix}\times2$	$\begin{bmatrix}3\times3,128\\3\times3,128\end{bmatrix}\times4$	$\begin{bmatrix}1\times1,128\\3\times3,128\\1\times1,512\end{bmatrix}\times4$	$\begin{bmatrix}1\times1,128\\3\times3,128\\1\times1,512\end{bmatrix}\times4$	$\begin{bmatrix}1\times1,128\\3\times3,128\\1\times1,512\end{bmatrix}\times8$
conv4	14×14	$\begin{bmatrix}3\times3,256\\3\times3,256\end{bmatrix}\times2$	$\begin{bmatrix}3\times3,256\\3\times3,256\end{bmatrix}\times6$	$\begin{bmatrix}1\times1,256\\3\times3,256\\1\times1,1024\end{bmatrix}\times6$	$\begin{bmatrix}1\times1,256\\3\times3,256\\1\times1,1024\end{bmatrix}\times23$	$\begin{bmatrix}1\times1,256\\3\times3,256\\1\times1,1024\end{bmatrix}\times36$
conv5	7×7	$\begin{bmatrix}3\times3,512\\3\times3,512\end{bmatrix}\times2$	$\begin{bmatrix}3\times3,512\\3\times3,512\end{bmatrix}\times3$	$\begin{bmatrix}1\times1,512\\3\times3,512\\1\times1,2048\end{bmatrix}\times3$	$\begin{bmatrix}1\times1,512\\3\times3,512\\1\times1,2048\end{bmatrix}\times3$	$\begin{bmatrix}1\times1,512\\3\times3,512\\1\times1,2048\end{bmatrix}\times3$
conv6	1×1	全局平均池化层				

3.2.2 类激活映射

类激活映射[8]是一种注意力方法，利用该方法可在给定的图像上将网络检测到的显著特征进行可视化，使用类激活映射，可以清楚地观察到网络关注图像的哪块区域，进而知晓网络主要基于哪部分特征得出表情的判定结果。

对于表情识别任务中的卷积神经网络来说，注意力图来自最后一层卷积层的特征图与全连接层的权重的加权和。注意力图的计算方式可表示为

$$M_j(h,w)=\sum_{c=1}^{C}W(j,c)F_c(h,w) \tag{3-2}$$

其中，C、h 和 w 分别为特征图的通道数、高和宽；M_j 为第 j 类表情预测结果所对应的类激活映射注意力图；$W(j,c)$为第 j 类表情预测结果所对应的全连接层的权重；$F_c(h,w)$为最后一层卷积层的输出特征图。

3.2.3 注意力一致性

注意力一致性由 Guo 等[9]首次提出，模型所学习到的注意力图应当与输入图像遵循相同的变换。例如，将同一张输入图像进行翻转后再次输入同一注意力网络中，重新计算得到的新注意力图应该与翻转之后的旧注意力图保持一致，这被称为注意力一致性。即针对输入图像 I_i，其对应的翻转图像为 I_i'，其对应的注意力图为 I_{att}，I_i'所对应的注意力图为 I'_{att}，则 I_{att} 和 I'_{att}之间的关系应满足：

$$I'_{\text{att}}=\text{flip}(I_{\text{att}}) \tag{3-3}$$

其中，flip()为翻转操作。通过考虑空间变换情况下的视觉注意力一致性，Guo 等实现了更合理的视觉感知与更高精度的多标签图像分类。

在表情数据集中，多数图像均含有一定程度的噪声，噪声掺杂现象在户外表情数据集中尤为常见，普通的注意力机制很有可能将某些噪声当成有用数据来学习，进而让网络发生过拟合。由于对一张图像进行常见的空间变换并不会改变其面部表情的标签属性，若使用注意力一致性对注意力机制加以约束，则网络能够学习到更加泛化和重要的显著信息，并抑制网络趋于过拟合。

3.2.4 EAC 模型介绍

EAC 模型结构如图 3.4 所示。首先，将输入图像依次进行随机擦除与水平翻转，得到翻转擦除图像样本，并与原始擦除图像样本一起构成输入样本对，由 ResNet50 分别提取其表情特征；其次，将原始图像分支的表情特征进行全局平均池化，并利用全连接实现表情识别；之后利用全连接层的参数，结合类激活映射方法分别计算得到原始图像注意力图与翻转图像注意力图；最后将翻转图像注意力图进行水平翻转，并根据注意力一致性计算一致性损失，结合交叉熵损失一同优化整体网络，实现较高性能的表情识别。

应当注意的是，翻转前后的面部图像具有相同的表情标签，在网络训练过程中，随着网络的每一轮迭代，随机擦除的位置与面积均在一定范围内不断变化，这可以防止网络模型记住噪声图像，结合一致性损失的指导，能够抑制网络过拟合，提高识别准确率。一致性损失的计算方式可表示为

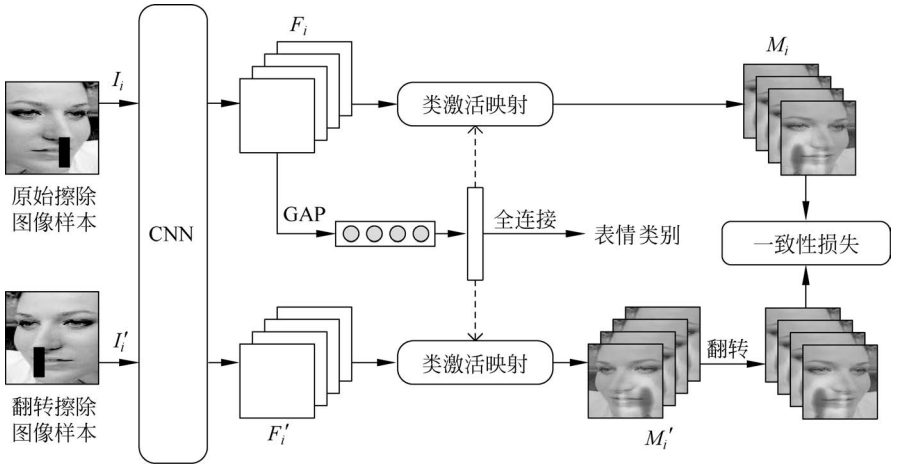


图 3.4　EAC 模型结构

$$L_c = \frac{1}{NCHW}\sum_{i=1}^{N}\sum_{j=1}^{C} \| M_{ij} - \mathrm{flip}(M'_{ij}) \|_2 \tag{3-4}$$

其中，N 为样本总数，C 为通道数，H 为特征图高度，W 为特征图宽度，M_{ij} 为原始注意力图，M'_{ij} 为翻转注意力图。

该网络模型的总损失为

$$L_{\text{total}} = L_{\text{cls}} + \lambda L_c L_{\text{total}} = L_{\text{cls}} + \lambda L_c \tag{3-5}$$

其中，L_{cls} 为交叉熵损失，λ 为超参数。

3.3　实验操作

3.3.1　实验环境

本实验所需要的环境配置如表 3.2 所示。

表 3.2　实验环境配置

条　　件	环　　境
操作系统	Ubuntu 16.04
开发语言	Python 3.8
深度学习框架	Pytorch 1.8.0
相关库	torchvision 0.9.0 pandas 1.2.5 numpy 1.20.2

实验项目文件下载网址可参考书中提供的二维码，下载完毕后解压得到名为 Erasing-Attention-Consistency-main 的实验项目文件夹。文件夹内相关内容如下：

```
Erasing-Attention-Consistency-main ---------------------------------- 工程根目录
├─imgs ------------------------------------------------------------- 论文结果图
├─model ------------------------------------------------ ResNet50 预训练模型参数
├─raf-basic ------------------------------------------------------- RAF-DB 数据集
|   ├─EmoLabel ------------------------------------- 与论文实验相关的标签文件
```

```
├─src
    ├─dataset.py ---------- 数据读取与预处理文件
    ├─loss.py ---------- 一致性损失函数文件
    ├─main.py ---------- 表情识别训练与预测的主程序
    ├─model.py ---------- 表情识别模型构建
    ├─resnet.py ---------- 表情识别主干网络
    ├─train.sh ---------- 运行指令文件
    ├─utils.py ---------- 表情识别工具辅助类文件
```

3.3.2 数据集介绍

本实验所使用的数据集为 Real-world Affective Faces Database(RAF-DB),该数据集的申请下载地址可参考书中提供的二维码。

RAF-DB 数据集[10]是一个含有 29 672 张真实户外面部图像的表情数据集,该数据集中的所有图像均来源于互联网,每张图像中受试者的种族、性别、年龄和外貌等属性不一,图像中的环境背景和光照条件也有很大差异,整个数据集的标注任务由大约 40 名训练有素的注释工作人员完成。根据标注方式的不同,RAF-DB 数据集可分为两个不同的子集,即基本情绪子集和复合情绪子集。基本情绪子集为单标签型标注,共包含 7 个离散型表情类别:中性、愤怒、厌恶、悲伤、开心、恐惧和惊讶。复合情绪子集为双标签型标注,共包含 12 种组合表情类别。目前针对 FER 任务的大部分研究都是基于基本情绪子集所开展,该子集一共包含 15 339 张表情图像,经过人脸对齐后图像尺寸均为 100×100px,其中有 12 271 张图像被选取为训练样本,其他 3068 张图像作为测试样本,各个类别的表情样本数量较为均衡。RAF-DB 数据集中的部分人脸表情图像如图 3.5 所示。

图 3.5 RAF-DB 数据集中部分人脸表情图像

3.3.3　实验操作与结果

(1) 下载 RAF-DB 数据集并解压,将解压后的数据集放入 Erasing-Attention-Consistency-main/raf-basic 目录中,数据存放目录格式如下:

```
├─raf - basic/
|   ├─Image/aligned/
|   |   ├─train_00001_aligned.jpg
|   |   ├─test_0001_aligned.jpg
|   |   ├─ ...
```

(2) 下载 ResNet50 的预训练模型参数并将其放入 Erasing-Attention-Consistency-main/model 目录中,利用该预训练模型参数初始化 EAC 模型中的主干网络,该参数的下载地址可参考书中提供的二维码。

(3) 在终端上将路径切换到 Erasing-Attention-Consistency-main/src 目录下,并开始训练和测试 EAC 模型:

```
$ cd /Erasing - Attention - Consistency - main/src
$ sh train.sh
```

(4) 训练完毕后,将自动生成一个名为 rebuttal_50_noise_list_patition_label.txt 的训练日志文件,但是训练模型参数不会自动保存,如有需要,可在 main.py 文件中的 main()函数中添加 torch.save(网络模型参数,'网络参数命名.pth')命令以保存训练好的整个网络。

(5) 实验结果表明,在 RAF-DB 训练集(主干网络为 ResNet50)上训练 EAC 模型直至收敛,在 RAF-DB 测试集上可达到 90%以上的识别准确率。

3.4　总结与展望

本次实验所使用的算法主要结合类激活映射与一致性损失,通过所设计的双分支不平衡框架抑制噪声所导致的网络过拟合现象,该模型的结构与训练机制并不复杂,但具有较好的识别性能,且该算法模型同样可以很好地推广到具有大量类别的图像分类任务中。

在表情识别任务中,表情数据本身具有一定的类内差异性与类间相似性,类内差异性是指同一类表情数据会因受试者的个人属性或环境背景的不同而不同,类间相似性是指不同类别的表情数据之间会因为受试者之间的个人属性或环境条件之间的相似而相似。同时,解读表情图像是一种较为主观的行为,这导致现存的表情数据集表情标注规则不一,数据集的制作标准也难以统一,表情标签错误混淆的情况难以避免。

标签混淆作为表情识别领域的一大难题,给提升表情识别网络性能带来巨大的挑战,针对标签混淆问题,许多科研工作者提出了新颖的算法以增强对表情特征的判别能力。Wang 等[11]从数据集入手,提出了一种注意力机制和重标记相结合的方法,在训练时学习给标签不确定的表情图像赋予较小的权重,配合重标记操作,达到清洗数据标签,训练得到更为鲁棒的识别网络的目的。She 等[12]从探索表情潜在分布和估计数据对之间的不确定性这两个角度来解决表情标签模糊的问题,其所引入的多分支辅助学习框架和实例语义特征相似模块在一定程度上抑制了标签混淆对网络识别性能带来的影响。Zhang 等[13]提出了一种

相对不确定性学习方法，在建立两个分支同时学习表情特征和图像的不确定性权重，通过累加损失函数，鼓励模型从混合特征中同时识别出两种表情，不确定性学习分支给标签不确定的面部表情图像分配较大的权重，同时给标签确定的表情图像分配较小的权重。

在进行科学研究时，不仅可以从网络模型角度去设计辅助模块、损失函数与训练策略等，还可以从数据本身出发，分析和利用数据特性，结合数据先验知识与相应的算法策略，有时也能获得意想不到的显著效果。

参考文献

[1] Xie S, Hu H, Wu Y. Deep multi-path convolutional neural network joint with salient region attention for facial expression recognition[J]. Pattern Recognition, 2019, 92: 177-191.

[2] 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.

[3] Xue F, Wang Q, Guo G. Transfer: Learning relation-aware facial expression representations with transformers[C]//Proceedings of the IEEE International Conference on Computer Vision, 2021, 3601-3610.

[4] Vo T H, Lee G S, Yang H, et al. Pyramid with super resolution for in-the-wild facial expression recognition[J]. IEEE Access, 2020, 8: 131988-132001.

[5] Farzaneh A H, Qi X. Facial expression recognition in the wild via deep attentive center loss[C]// Proceedings of the IEEE Winter Conference on Applications of Computer Vision, 2021.

[6] Fard A P, Mahoor M H. Ad-corre: Adaptive correlation-based loss for facial expression recognition in the wild[J]. IEEE Access, 2022, 10: 26756-26768.

[7] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[8] Zhou B, Khosla A, Lapedriza A, et al. Learning deep features for discriminative localization[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[9] Guo H, Zheng K, Fan X, et al. Visual attention consistency under image transforms for multi-label image classification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019.

[10] Li S, Deng W, Du J. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[11] Wang K, Peng X, Yang J, et al. Suppressing uncertainties for large-scale facial expression recognition [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020.

[12] She J, Hu Y, Shi H, et al. Dive into ambiguity: Latent distribution mining and pairwise uncertainty estimation for facial expression recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2021.

[13] Zhang Y, Wang C, Deng W. Relative Uncertainty Learning for Facial Expression Recognition[C]// Proceedings of Advances in Neural Information Processing Systems, 2021.