中图法分类号:TP399　　**文献标识码**:A　　**文章编号**:1006-8961(2024)01-0123-11

论文引用格式:Cui X Y, He C, Zhao H K and Wang M L. 2024. Combining ViT with contrastive learning for facial expression recognition. Journal of Image and Graphics, 29(01):0123-0133(崔鑫宇, 何翀, 赵宏珂, 王美丽. 2024. 融合ViT与对比学习的面部表情识别. 中国图象图形学报, 29(01):0123-0133)[DOI:10.11834/jig.230043]

融合ViT与对比学习的面部表情识别

崔鑫宇[1],何翀[1],赵宏珂[1],王美丽[1,2,3*]

1. 西北农林科技大学信息工程学院,杨凌　712100; 2. 农业农村部农业物联网重点实验室(西北农林科技大学), 杨凌　712100;
3. 陕西省农业信息感知与智能服务重点实验室(西北农林科技大学),杨凌　712100

摘　要:**目的**　面部表情识别是计算机视觉领域中的重要任务之一,而真实环境下面部表情识别的准确度较低。针对面部表情识别中存在的遮挡、姿态变化和光照变化等问题导致识别准确度较低的问题,提出一种基于自监督对比学习的面部表情识别方法,可以提高遮挡等变化条件下面部表情识别的准确度。**方法**　该方法包含对比学习预训练和模型微调两个阶段。在对比学习预训练阶段,改进对比学习的数据增强方式及正负样本对对比次数,选取基于Transformer的视觉Transformer(vision Transformer,ViT)网络作为骨干网络,并在ImageNet数据集上训练模型,提高模型的特征提取能力。模型微调阶段,采用训练好的预训练模型,用面部表情识别目标数据集微调模型获得识别结果。**结果**　实验在4类数据集上与13种方法进行了比较,在RAF-DB(real-world affective faces database)数据集中,相比于Face2Exp(combating data biases for facial expression recognition)模型,识别准确度提高了0.48%;在FER-Plus(facial expression recognition plus)数据集中,相比于KTN(knowledgeable teacher network)模型,识别准确度提高了0.35%;在AffectNet-8数据集中,相比于SCN(self-cure network)模型,识别准确度提高了0.40%;在AffectNet-7数据集中,相比于DACL(deep attentive center loss)模型,识别准确度略低0.26%,表明了本文方法的有效性。**结论**　本文所提出的人脸表情识别模型,综合了对比学习模型和ViT模型的优点,提高了面部表情识别模型在遮挡等条件下的鲁棒性,使面部表情识别结果更加准确。

关键词:表情识别;对比学习;自监督学习;Transformer;正负样本对

Combining ViT with contrastive learning for facial expression recognition

Cui Xinyu[1], He Chong[1], Zhao Hongke[1], Wang Meili[1,2,3*]

1. College of Information Engineering, Northwest A&F University, Yangling　712100, China;
2. Key Laboratory of Agricultural Internet of Things, Ministry of Agriculture (Northwest A & F University), Yangling　712100, China;
3. Shaanxi Key Laboratory of Agricultural Information Perception and Intelligent Service (Northwest A & F University), Yangling　712100, China

Abstract: **Objective**　Facial expression is one of the important factors in human communication to help understand the intentions of others. The task of facial expression recognition is to output the category of facial expression corresponding to a given face picture. Facial expression has broad applications in areas such as security monitoring, education, and human-computer interaction. Currently, facial expression recognition under uncontrolled conditions suffers from low accuracy due to factors such as pose variations, occlusions, and lighting differences. Addressing these issues will remarkably advance the development of facial expression recognition in real-world scenarios and hold great relevance in the field of artificial

收稿日期:2023-01-29;**修回日期**:2023-07-03;**预印本日期**:2023-07-10

*通信作者:王美丽　wml@nwsuaf.edu.cn

基金项目:陕西省林业科学院科技创新计划项目(SXLK2021-0214)

Supported by:Science and Technology Innovation Program, Shaanxi Academy of Forestry(SXLK2021-0214)

intelligence. Self-supervised learning is proposed to utilize specific data augmentations on input data and generate pseudo labels for training or pretraining models. Self-supervised learning leverages a large amount of unlabeled data and extracts the prior knowledge distribution of the images themselves to improve the performance of downstream tasks. Contrast learning belongs to self-supervised learning, which can further learn the intrinsic consistent feature information between similar images under the change of posture and light by increasing the difficulty of the task. This paper proposes an unsupervised contrastive learning-based facial expression classification method to address the problem of low accuracy caused by occlusion, pose variation, and lighting changes in facial expression recognition. **Method** To address the issue of occlusions in facial expression recognition datasets under real-world conditions, a method based on negative sample-based self-supervised contrastive learning is employed. The method consists of two stages: contrastive learning pretraining and model fine-tuning. First, in the pretraining stage of contrastive learning, an unsupervised contrastive loss is introduced to reduce the distance between images of the same type and increase the distance between images of different classes to improve the discrimination ability of intraclass diversity and interclass similarity images of facial expression images. This method involves adding positive sample pairs for contrastive learning between the original images and occlusion-augmented images, enhancing the robustness of the model to image occlusion and illumination changes. Additionally, a dictionary mechanism is applied to MoCo v3 to overcome the issue of insufficient memory during training. The recognition model is pretrained on the ImageNet dataset. Next, the model is fine-tuned on the facial expression recognition dataset to improve the classification accuracy for facial expression recognition tasks. This approach effectively enhances the performance of facial expression recognition in the presence of occlusions. Moreover, the Transformer-based vision Transformer (ViT) network is employed as the backbone network to enhance the model's feature extraction capability. **Result** Experiments were conducted on four datasets to evaluate the performance of the proposed method compared with the latest 13 methods. In the RAF-DB dataset, compared with the Face2Exp model, the recognition accuracy increased by 0.48%; in the FERPlus dataset, compared with the knowledgeable teacher network (KTN) model, The recognition accuracy increased by 0.35%; in the AffectNet-8 dataset, compared with the self-cure network (SCN) model, the recognition accuracy increased by 0.40%; in the AffectNet-7 dataset, compared with the deep attentive center loss (DACL) model, the recognition accuracy was slightly lower by 0.26%, which proves the effectiveness of the method in this paper. **Conclusion** A self-supervised contrastive learning-based method for facial expression recognition is proposed to address the challenges of occlusion, pose variation, and illumination changes in uncontrolled conditions. The method consists of two stages: pretraining and fine-tuning. The contributions of this paper lie in the integration of ViT into the contrastive learning framework, which enables the utilization of a large amount of unlabeled, noise-occluded data to learn the distribution characteristics of facial expression data. The proposed method achieves promising accuracy on facial expression recognition datasets, including RAF-DB, FERPlus, AffectNet-7, and AffectNet-8. By leveraging the contrastive learning framework and advanced feature extraction networks, this work enhances the application of deep learning methods in everyday visual tasks.

Key words: facial expression recognition; comparative learning; self-supervised learning; Transformer; positive and negative samples

0 引 言

面部表情是指通过面部肌肉的变化来表现各种情绪状态,在人类沟通中帮助理解他人意图。美国心理学家 Ekman 和 Friesen(1971)定义基本面部表情类型为快乐、悲伤、愤怒、恐惧、惊讶、厌恶。随后 Ekman 和 Friesen(1978)提出了一种面部动作编码系统来研究人的面部动作进而判断表情。目前面部表情分析已经成为计算机视觉和人工智能领域的重要方向(彭小江和乔宇,2020),面部表情识别技术在安全监控、教育教学和人机交互等领域具有广阔应用。

人脸表情识别的任务是对于给定的一幅人脸图像,输出这张人脸对应表情的类别。目前的深度人脸表情识别系统存在真实世界环境下其他与表情无关因素变量带来的干扰问题(李珊和邓伟洪,2020)。在非受控环境下(自然条件)构建的数据集,如 RAF-DB(real-world affective faces database)(Li 等,2017)、AffectNet(Mollahosseini 等,2019)、FERPlus(facial expression recognition plus)(Barsoum 等,2016),相比

于受控条件下数据集更接近真实场景，使其更具有研究意义。目前，非受控条件下的人脸表情识别由于受到姿态变换、遮挡和光照差异等多种因素的影响，导致人脸表情识别的准确性较低。解决好以上问题将会大大推动真实场景下人脸表情识别的发展，对人工智能领域具有重大意义。

自监督学习使用特定的数据增强应用于输入数据，并产生伪标签来训练或预训练模型，利用大量的无标注数据并提取图像本身的先验知识分布，提升下游任务的效果。然而，对比学习属于自监督学习，它可以通过增加任务难度从而进一步学习姿态、光线变化下同类图像之间的内在一致特征信息。受此思路的启发，本文认为基于负例的对比学习模型可以解决现实条件下遮挡、姿态和光线变化对于表情识别影响的问题，提升识别准确度。

具有强大的特征提取网络可以增强对比学习的特征表征能力。基于Transformer（Vaswani等，2017）的网络结构性能超过标准的卷积神经网络（convolutional neutral network，CNN），视觉Transformer（vision Transformer，ViT）（Dosovitskiy等，2021）通过在非重叠图像块上直接应用自然语言处理（natural language processing，NLP）中的标准Transformer编码器，展现了图像分类的强大性能。本文利用ViT强大的特征提取能力来提高对比学习在非受控条件下的表情识别性能。

为了更好地降低真实场景中各类因素对识别效果的影响，提高表情识别准确率，本文工作贡献如下：1）针对现实条件下的人脸表情识别数据集存在较多遮挡问题，采用基于负例的自监督对比学习方法实现人脸表情识别，增加原图像和经过遮挡增强图像的正样本对对比，并将字典机制应用到MoCo v3）（momentum contrast v3）（Chen等，2021）中，解决训练中显存不足的问题，将该识别模型在ImageNet数据集（Deng等，2009）上进行预训练，将预训练模型应用于表情识别数据集进行微调，以提高表情识别任务的分类准确率。2）将基于Transformer的ViT网络作为骨干特征提取网络，结合对比学习预训练，在表情识别公共数据集RAF-DB、FERPlus、AffectNet-7和AffectNet-8上取得89.02%、90.84%、64.94%和60.63%的识别准确度，与一些表情识别流行算法进行比较，表明了算法的有效性。

1 相关工作

1.1 人脸表情识别方法

表情标注具有主观性和差异性，因此数据集中的标签噪声难以避免（姚鸿勋 等，2022）。由于大规模人脸表情数据集具有不确定性，Wang等人（2020a）提出了一个简洁高效的自修复网络（self-cure network，SCN），有效地抑制不确定性，从而提高表情识别算法的鲁棒性。Zhao等人（2021）提出轻量级表情识别算法EfficientFace，引入简单但有效的标签分布学习（label distribution training，LDL）作为训练策略，通过学习样本之间的关系，对每个表情类别建模其可能的分布。Face2Exp（combating data biases for facial expression recognition）（Zeng等，2022）通过元优化框架从数据中提取去偏信息来消除数据偏差。本文增加原图像和经过遮挡增强后图像的正样本对对比，减少遮挡和数据不确定性对识别效果带来的影响。

Li等人（2019b）为了增强模型学习特征的判别力，提出了Separate Loss损失函数，使用该损失函数学习的特征具有类内紧凑和类间分离的特点。Farzaneh和Qi（2020）提出了判别无关分布（discriminant distribution-agnostic，DDA）损失用于野外面部表情识别，该研究通过优化极端类不平衡场景的嵌入空间，在嵌入空间中产生分离良好的深度特征簇。Farzaneh和Qi（2021）提出了一种深度注意力中心损失（deep attentive center loss，DACL）的方法，集成注意力机制，并使用卷积神经网络中的中间空间特征图作为上下文来估计与特征重要性相关的注意力权重。Fard和Mahoor（2022）提出了一种自适应相关性（adaptive correlation，Ad-Corre）损失来引导网络生成对于类内样本具有高相关性、对于类间样本具有较低相关性的嵌入特征向量。以上方法虽然通过使用更好的损失函数来引导模型学习，但并未聚焦特征提取和学习。

因此，一些工作增强模型学习特征的能力，以此来提高表情识别的性能。Li等人（2019a）提出了基于CNN和注意力机制的人脸表情识别方法自注意力卷积神经网络（CNN with attention mechanism，ACNN），能够自动感知遮挡的面部区域，并将注意力集中于未遮挡和信息丰富的区域，在实验室条件

下和真实环境条件下都取得了不错的识别效果，并提出了第一个专注真实环境下面部遮挡的数据集FEF-RO（facial expression dataset with real-world occlusions）。Wang等人（2020b）提出了区域注意力网络（region attention networks，RAN）来捕捉对遮挡和姿态变换图像重要的区域，并提出了区域偏差损失函数（region biased loss，RB-loss）来让更重要的区域得到更多的注意力权重。Li等人（2020）提出了使用SE（squeeze excitation）模块和滑动块的整体面部注意力模型（slide-patch and whole-face attention model with SE blocks，SPWFA-SE），感知面部的判别局部特征和信息丰富的全局特征，利用多级特征提取和注意机制增强所学特征的代表性。Huang等人（2021）提出FER-VT（facial expression recognition with grid-wise attention and visual Transformer）表情识别算法，利用注意力机制从面部图像中捕捉不同区域的依赖关系，利用Transformer注意力机制学习全局表示。Zhang等人（2022）提出了一种擦除注意力一致性（erasing attention consistency，EAC）方法来自动抑制训练过程中的噪声样本。

然而，上述研究使用注意力机制在一定程度上解决了在遮挡等不利因素影响下的表情识别，但忽略了数据本身的特征分布。本文融合ViT和对比学习，不仅更好地提取特征，还能学习表情的特征分布。

此外，还有使用知识蒸馏实现表情识别的工作，如Li等人（2021）提出知识富集的教师网络（knowledgeable teacher network，KTN），解决在各类表情数据集分布不平衡的影响下区分高度相似的表情，提出AdaReg损失函数和粗—细标签策略，引导模型从易到难对高度相似的表示进行分类。

1.2 对比学习概况

对比学习属于自监督学习，无需标注数据，在多个模型上的效果超过有监督模型。基于负例的对比学习方法动量对比（MoCo）（He等，2020）开创自监督视觉表示学习，通过在ImageNet-1k数据集预训练，在7个下游学习任务上超过标准监督算法的性能。

SimCLR（simple framework for contrastive learning of representations）（Chen等，2020a）算法表明，多种数据增强方式的组合、表示和对比损失之间引入非线性转换、更大的批量和更多的训练步骤能有效提高图像分类准确率。但是由于采用了大量的数据增强策略，训练复杂度较高，需要大量的计算资源。

BYOL（bootstrap your own laten）（Grill等，2020）模型基于非对称网络结构的方法，不使用负样本对，训练分为在线网络和目标网络两部分，有效解决了训练坍缩问题。但是训练过程中需要维护在线网络和目标网络，增加了模型训练和存储的复杂度。

SwAV（swapping assignments between multiple views of the same images）（Caron等，2020）算法为基于对比聚类的方法，在对比学习中引入聚类，先对训练样本进行聚类，然后在类间进行对比学习，通过聚类后再做对比操作，减少了对比的数量，可有效降低计算复杂度。然而，聚类的数量对训练速度有较大影响，需要进行合理的聚类数量选择。

Barlow Twins（self-supervised learning via redundancy reduction）（Zbontar等，2021）是一种基于冗余消除损失函数的方法，它提出一个目标函数，通过衡量同一样本两个不同视图输入到相同网络得到嵌入的互相关矩阵，使其接近单位矩阵，该方法依赖于非常高维的输出向量。由于该方法使用高维特征进行计算，需要更多的存储空间和计算资源。

2 提出的面部表情识别方法

本文结合MoCo对比学习框架设计了针对表情识别任务的学习框架，运用了MoCo系列的动量编码机制、队列机制等，同时加入了投影头、预测头等提高模型的提取特征的能力，最后对正负样本对做对比损失计算。

编码器网络f_q包括1个骨干网络ViT、1个投影头、1个预测头。动量编码器网络f_k包括1个骨干网络ViT和1个投影头，没有预测头。f_k通过f_q的滑动平均更新，预测头不更新。投影头是1个3层多层感知机（multilayer perceptron，MLP），预测头是一个2层MLP。

输入的图像经过数据增强后，原图像和数据增强后的图像分别经过编码器网络进行特征提取，编码器提取的特征经过投影层、预测层和动量编码器后，进行正负样本对的相似度计算，结束后编码器网络进行参数更新，动量编码器网络进行动量更新。

MoCo v1(He等,2020)主要使用移动平均更新模型权重和队列方法形成一个字典查询问题,解决难以应用大量负样本进行训练的问题。MoCo v2(Chen等,2020b)改进了图像数据增强的方法,增加使用模糊增强方法,在编码器得到表示信息后添加非线性层等,使模型以更小的批量大小和训练轮数,在ImageNet数据集分类任务上取得了更好的效果。MoCo v3中使用更大的批次大小进行训练,取代了之前改进的Memory Queue,使用基于Transformer的ViT作为编码器,在负例编码器上添加预测头,并解决了对比学习在ViT训练过程中表现出的不稳定问题,图像分类效果得到进一步提升。

由于MoCo v3框架预训练需要大量的显存资源支持,导致其无法广泛使用,本文将MoCo v1和MoCo v2所采用的字典机制运用到MoCo v3中,其中的样本队列采用的是字典机制,在显存资源缺乏的情况下仍能进行训练。改进数据增强和正样本对对比方式,改进后的对比学习框架如图1所示。

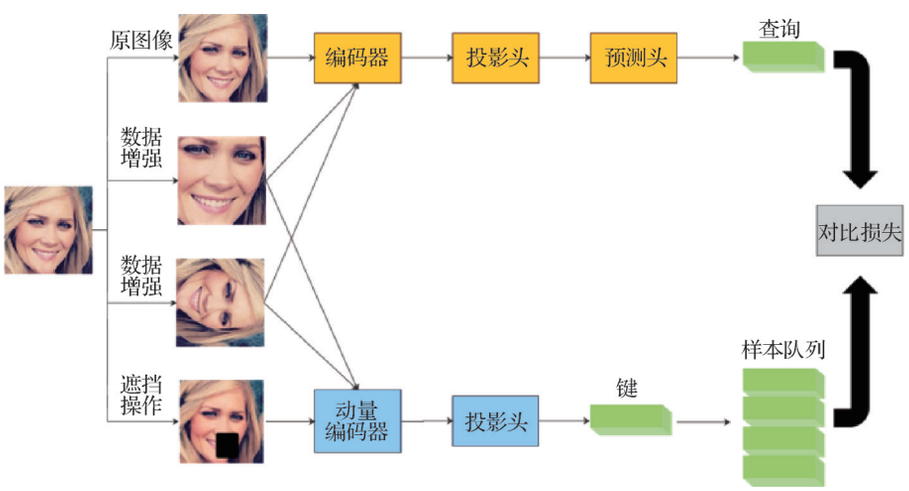


图1　对比学习框架图

Fig. 1　Contrastive learning frame diagram

2.1　数据增强及正负样本对对比

数据增强方式的选择对于对比学习的效果会产生比较大的影响。针对现实情况下图像识别存在遮挡的问题,本文在MoCo框架的基础上,对数据增强方式和正负样本对比方式进行改进,原框架中对输入的图像进行随机数据增强,包括水平翻转、颜色抖动、随机调整大小的裁剪、灰度转换、模糊和曝光等,对数据增强后的图像进行两次裁剪,将裁剪后的两幅图像输入两个编码器中进行编码。本文在此基础上,增加了一组正样本对比,将原图进行数据增强的遮挡操作,再将原图像和经过遮挡后的图像输入到编码器中进行编码。编码器f_q和动量编码器f_k输出向量为$\boldsymbol{q}$和$\boldsymbol{k}$,采用最小化对比损失函数InfoNCE进行相似度计算,分子部分表示正例之间的相似度,分母表示正例与负例之间的相似度。相同类别样本相似度越大,不同类别样本相似度越小,损失函数的值越小,计算为

$$Loss=-\log\frac{\exp\left(\frac{\boldsymbol{q}\times\boldsymbol{k}^{+}}{t}\right)}{\sum_{i=0}^{K}\exp\left(\frac{\boldsymbol{q}\times\boldsymbol{k}^{-}}{t}\right)} \tag{1}$$

式中,$\boldsymbol{k}^{+}$是$\boldsymbol{q}$的同一幅图像经过编码器f_q的输出,作为$\boldsymbol{q}$的正样本;$\boldsymbol{k}^{-}$表示其他图像经过动量编码f_k的处理输出,作为$\boldsymbol{q}$的负样本;t是归一化$\boldsymbol{q}$、$\boldsymbol{k}$的温度超参数;K代表字典中样本数量。本文设计的对比学习正样本对示意图如图2所示。

通过这样的对比方式可以有效地挖掘令网络学习更有效的样本,从而推动网络寻找图像分类的边界线。$\boldsymbol{X}$为原图像,$\boldsymbol{X}_0$为加入随机遮挡后的图像,$\boldsymbol{X}_1$是经过水平翻转后的图像,$\boldsymbol{X}_2$是经过随机裁剪后的图像,$\boldsymbol{X}$,$\boldsymbol{X}_1$,$\boldsymbol{X}_2$通过编码器f_q分别得到特征$\boldsymbol{q}$,$\boldsymbol{q}_1$,$\boldsymbol{q}_2$,$\boldsymbol{X}_0$,$\boldsymbol{X}_1$,$\boldsymbol{X}_2$通过动量编码器f_k分别得到特征$\boldsymbol{k}$,$\boldsymbol{k}_1$,$\boldsymbol{k}_2$,维度为n,特征空间由一个长度为c的向量表示,分别做矩阵相乘,得到3个维度为(n,n)的矩阵,3个矩阵相加得到1个维度为(n,n)

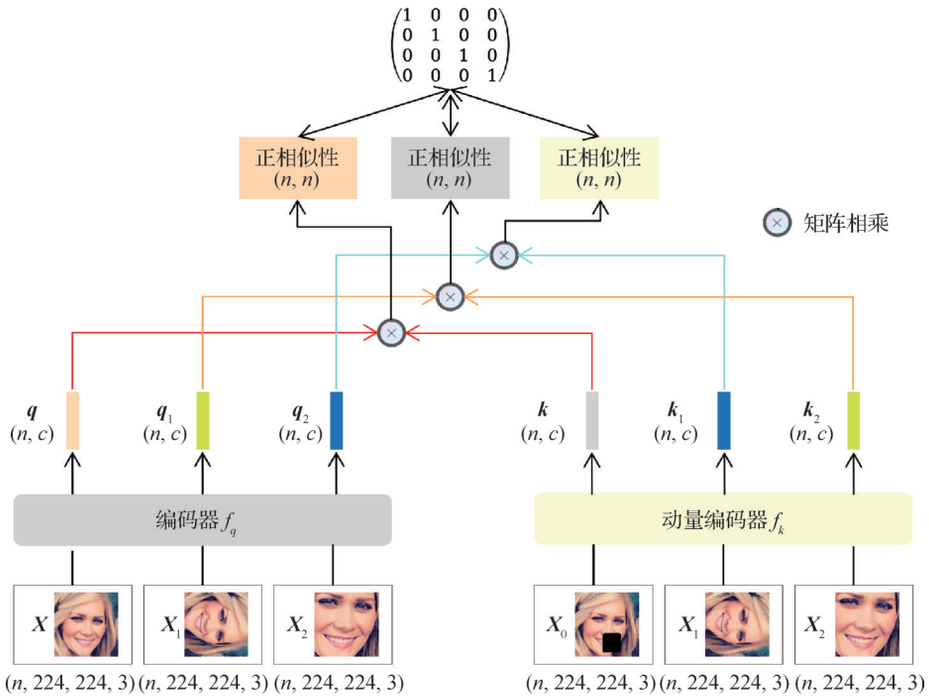


图2　对比学习正样本对示意图

Fig. 2　Schematic diagram of contrastive learning positive sample pairs

的矩阵，矩阵对角线元素代表的是正样本对的相似度，对角线元素越大越好，整个矩阵接近单位矩阵越好。

2.2　编码器

使用ViT-small作为特征提取网络，输入图像尺寸为224 × 224像素，patch大小为16 × 16，嵌入维度大小为384，ViT的核心流程包括图像分块处理、图像块嵌入与位置编码、Transformer编码器和MLP分类处理等4个主要部分，其模型架构如图3所示。

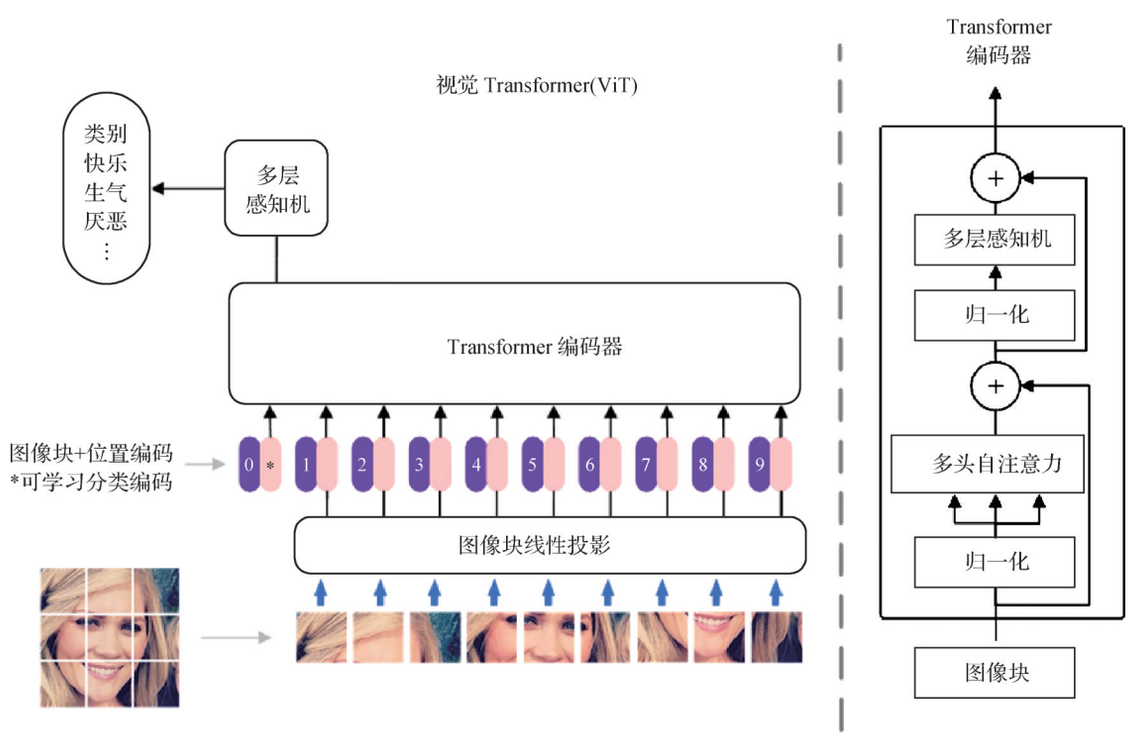


图3　ViT框架图

Fig. 3　ViT frame diagram

ViT的前向过程为：首先将输入图像分为N个图像块$\boldsymbol{X}_p$，表示为向量[$\boldsymbol{X}_p^1$；$\boldsymbol{X}_p^2$；…；$\boldsymbol{X}_p^N$]；然后将图像块变换嵌入为$N \times D$大小的特征序列，并在特征序列头部加入可学习的面部表情分类编码$\boldsymbol{X}_s$，对特征序列加入位置编码得到Transformer第1层输入z_0；然后经过多个Transformer编码器，其中包括层归一化（layer normalization，LN）、多头自注意力模块（multi-head self-attention，MSA）和MLP，最后得到面部特征。具体计算为

$$z_0 = \left[\boldsymbol{X}_s; \boldsymbol{X}_p^1\boldsymbol{E}; \boldsymbol{X}_p^2\boldsymbol{E}; \cdots; \boldsymbol{X}_p^N\boldsymbol{E}\right] + \boldsymbol{E}_p$$
$$\boldsymbol{E} \in \mathbf{R}^{(P^2 \times C) \times D},\ \boldsymbol{E}_p \in \mathbf{R}^{(N+1) \times D} \quad (2)$$
$$z_l' = MSA\left(LN\left(z_{l-1}\right)\right) + z_{l-1},\ l = 1, \cdots, L \quad (3)$$
$$z_l = MLP\left(LN\left(z_l'\right)\right) + z_l',\ l = 1, \cdots, L \quad (4)$$
$$y = LN\left(z_L^0\right) \quad (5)$$

式中，N代表图像分块数，P代表图像分块大小，C代表通道数，$\boldsymbol{E}$代表图像分块线性变换嵌入，D为线性变换嵌入后的特征大小，$\boldsymbol{E}_p$表示图像分块的位置编码，z_l'表示在第l层经过MSA模块和残差连接后得到的面部特征序列，z_l表示在第l层经过MLP模块和残差连接后得到的面部特征序列，z_L^0表示经过L层Transformer编码器后得到的面部特征，最终得到经过多层Transformer编码器处理后的面部特征$\boldsymbol{y}$。

3 数据集介绍

针对实际问题和实际需要采用RAF-DB、FERPlus、AffectNet等自然条件下的公开数据集。

RAF-DB数据集包含29 672幅从网络采集的真实人脸图像，包含单标签子集和多标签子集。在本文实验中使用单标签子集，包括快乐、中性、惊讶、愤怒、悲伤、恐惧、厌恶等7种基本情绪，由12 271幅训练样本和3 068幅测试样本组成。

FERPlus数据集包括28 709幅训练样本、3 589幅验证样本和3 589幅测试样本，是从谷歌搜索引擎收集而来，除7种基本情绪外，蔑视类别也包括在标签中。

AffectNet数据集是一个大型的表情数据集，是从网络中收集而成，由超过1 000 000幅面部图像组成。在实验中，使用与FERPlus数据集相同的8个基本表情以及去除蔑视外的7类表情作为识别任务。

4 实验结果与分析

首先在ImageNet数据集上进行无监督对比学习预训练，预训练完成后在目标表情数据集上进行微调。本文实验均在Ubuntu 20.04系统中完成，环境配置为Intel i7-6700 CPU 3.40 GHz，16 GB内存，GeForce RTX 1080Ti GPU。为了验证所提出方法的有效性，本研究在RAF-DB、FERPlus、AffectNet-7和AffectNet-8数据集上进行实验，并与一些流行的算法进行比较，训练集、验证集按照官网默认划分的结果进行实验。

将InfoNCE（information noise contrastive estimation）作为损失函数，采用网络结构ViT-small，以自适应矩估计（adaptive moment estimation，Adam）为优化器，批量大小设置为128，迭代轮次设置为200，学习率初始化为0.000 3，采用余弦退火学习率衰减策略，学习率衰减周期设置为30。

基于本算法模型在RAF-DB、FERPlus、AffectNet-7和AffectNet-8公开面部表情数据集上进行实验，得到各类表情识别结果，本文将该模型与其他13种方法进行了对比，结果如表1所示。本模型在RAF-DB数据集上取得了89.02%的识别准确度，比Face2Exp算法提升了0.48%，在FERPlus数据集上取得了90.84%的识别准确度，比KTN算法提升了0.35%，说明本模型在小规模数据集上具有良好的性能。本模型在AffectNet数据集（7类）上取得了64.94%的识别准确度，略低于DACL算法的65.20%，在AffectNet数据集（8类）上取得了60.63%的识别准确度，比SCN算法提升了0.4%，说明本模型在较大规模数据集上仍然具有良好的性能。图4展示了本模型在AffectNet-8数据集上的识别效果。

该算法模型在各个面部表情数据集上的混淆矩阵如图5—图8所示，由图中可以看出，快乐、惊讶、中性表情具有较高的识别准确度，蔑视、恐惧、厌恶等表情识别准确度较低，与面部变化不明显有关系。不平衡数据集在分类问题中，当训练集样本数量在类中分布不均匀时，样本较多的类别识别效果较好，如果样本数量不平衡度很高，会影响分类器的性能并导致网络偏向较大的样本。本数据集训练集中高兴表情数量最多，厌恶、蔑视等表情样本数较少，对识别准确度造成了一定程度的影响。

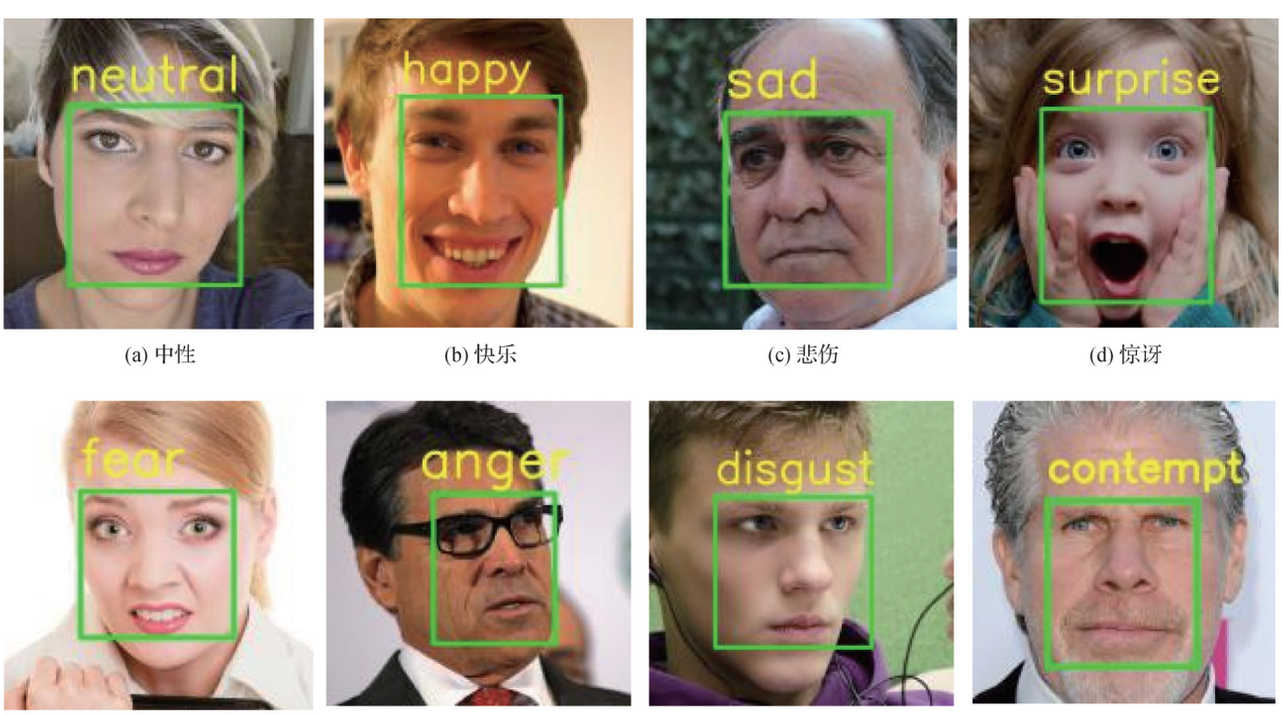
表1　本模型与其他方法结果比较

Table 1　Comparison of the results of this model with other methods

/%

方法	RAF-DB	FERPlus	AffectNet-7	AffectNet-8
ACNN(Li, 2019a)	85.07	–	58.78	–
SPWFA-SE(Li等, 2020)	86.31	–	59.23	–
RAN(Wang等, 2020b)	86.90	89.16	–	59.50
SCN(Wang等, 2020a)	88.14	89.35	–	60.23
DACL(Farzaneh和Qi, 2021)	87.78	–	**65.20**	–
KTN(Li等, 2021)	88.07	90.49	63.97	–
FER-VT(Huang等, 2021)	88.26	90.04		–
EfficientFace(Zhao等, 2021)	88.36	–	63.70	58.89
Separate loss(Li等, 2019b)	86.38	–	58.89	–
DDA loss(Farzaneh和Qi, 2020)	86.90	–	62.34	–
Ad-Corre(Fard和Mahoor, 2022)	86.96	–	63.36	–
EAC(Zhang等, 2022)	88.02	87.03	61.11	–
Face2Exp(Zeng等, 2022)	88.54	–	64.23	–
本文	**89.02**	**90.84**	64.94	**60.63**

注：加粗字体表示各列最优结果，“–”表示无实验数据。



(a) 中性　(b) 快乐　(c) 悲伤　(d) 惊讶

(e) 恐惧　(f) 愤怒　(g) 厌恶　(h) 蔑视

图4　在AffectNet-8数据集的检测识别效果

Fig. 4　Detection and recognition effect in AffectNet-8 dataset

((a) neutral; (b) happy; (c) sad; (d) surprise; (e) fear; (f) anger; (g) disgust; (h) contempt)

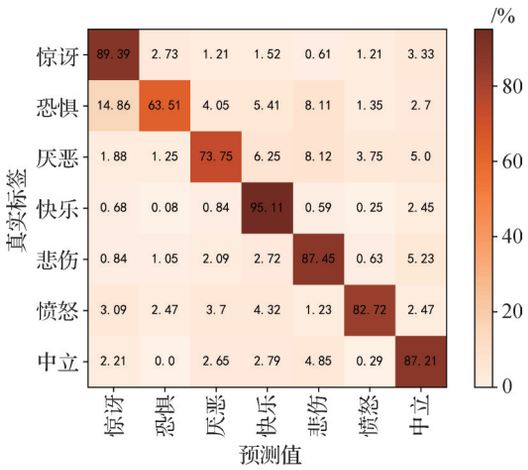


图5　在RAF-DB数据集中测试的混淆矩阵

Fig. 5　Confusion matrix tested on the RAF-DB dataset

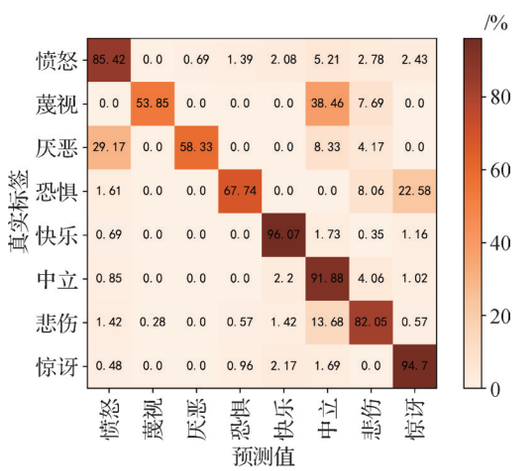


图6　在FERPlus数据集中测试的混淆矩阵

Fig. 6　Confusion matrix tested on the FERPlus dataset

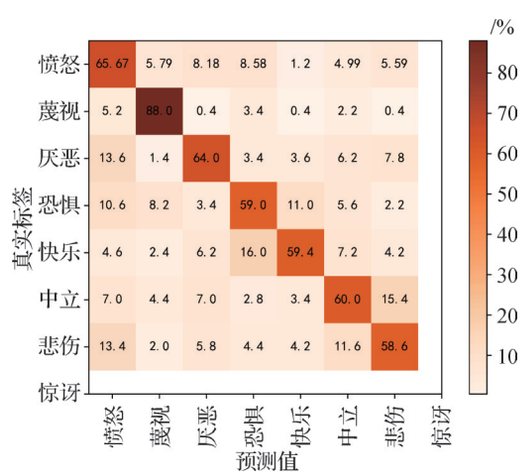


图7　在AffectNet-7数据集中测试的混淆矩阵

Fig. 7　Confusion matrix tested on the AffectNet-7 dataset

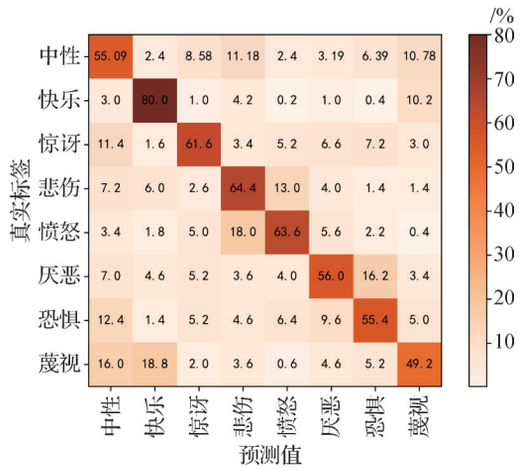


图8　在AffectNet-8数据集中测试的混淆矩阵

Fig. 8　Confusion matrix tested on the AffectNet-8 dataset

5　结　论

针对非受控条件下表情识别中存在的遮挡、姿态变化和光照变化等问题，本文提出一种基于自监督对比学习的人脸面部表情识别算法。该算法包括预训练和微调两个阶段，预训练阶段融合ViT为骨干网络，并增加数据增强方式和正负样本对比次数，强化了模型的特征表征提取能力和对图像遮挡、姿态、光照条件变化的鲁棒性；微调阶段，利用人脸表情识别数据集进行训练获得识别结果。

本文主要创新在于融合ViT到对比学习框架，可以充分利用大量无标签和噪声遮挡数据，充分学习人脸表情数据分布特征，在人脸表情识别数据集RAF-DB、FERPlus、AffectNet-7和AffectNet-8上获得了较好的准确率。利用对比学习框架和先进的特征提取网络，将增强深度学习方法应用于日常视觉任务中。

然而，本文存在以下不足：负面表情的识别准确率有待提高，这与数据集中负面表情样本数量有限有关，网络没有学习到更多的负面表情特征。未来，本文将致力于对负面表情的准确识别，探索更强大的特征提取网络和高效简洁的对比学习方法，进一步提高人脸表情识别的准确率。此外，本文方法基于2D人脸表情数据集，随着未来3D人脸数据的完善，表情识别将有更多的研究空间，本文将继续探索研究3D人脸数据下的表情识别。

参考文献(References)

Barsoum E, Zhang C, Ferrer C C and Zhang Z Y. 2016. Training deep networks for facial expression recognition with crowd-sourced label distribution//Proceedings of the 18th ACM International Conference on Multimodal Interaction. Tokyo, Japan: ACM: 279-283 [DOI: 10.1145/2993148.2993165]

Caron M, Misra I, Mairal J, Goyal P, Bojanowski P and Joulin A. 2020. Unsupervised learning of visual features by contrasting cluster assignments//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: 9912-9924

Chen T, Kornblith S, Norouzi M and Hinton G E. 2020a. A simple framework for contrastive learning of visual representations//Proceedings of the 37th International Conference on Machine Learning. Virtual Event: JMLR.org: 1597-1607

Chen X L, Fan H Q, Girshick R and He K M. 2020b. Improved baselines with momentum contrastive learning [EB/OL]. [2023-01-29]. https://arxiv.org/pdf/2003.04297.pdf

Chen X L, Xie S N and He K M. 2021. An empirical study of training self-supervised vision Transformers//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 9620-9629 [DOI: 10.1109/ICCV48922.2021.00950]

Deng J, Dong W, Socher R, Li L J, Li K and Li F F. 2009. ImageNet: a large-scale hierarchical image database//Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, USA: IEEE: 248-255 [DOI: 10.1109/CVPR.2009.5206848]

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X H, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J and Houlsby N. 2021. An image is worth 16 × 16 words: Transformers for image recognition at scale [EB/OL]. [2023-01-29]. https://arxiv.org/pdf/2010.11929.pdf

Ekman P and Friesen W V. 1971. Constants across cultures in the face and emotion. Journal of Personality and Social Psychology, 17(2): 124-129 [DOI: 10.103//h0030377]

Ekman P and Friesen W V. 1978. Facial Action Coding System: A Technique for the Measurement of Facial Movement. Palo Alto, USA: Consulting Psychologists Press

Fard A P and Mahoor M H. 2022. Ad-Corre: adaptive correlation-based loss for facial expression recognition in the wild. IEEE Access, 10: 26756-26768 [DOI: 10.1109/ACCESS.2022.3156598]

Farzaneh A H and Qi X J. 2020. Discriminant distribution-agnostic loss for facial expression recognition in the wild//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Seattle, USA: IEEE: 1631-1639 [DOI: 10.1109/CVPRW50498.2020.00211]

Farzaneh A H and Qi X J. 2021. Facial expression recognition in the wild via deep attentive center loss//Proceedings of 2021 IEEE Winter Conference on Applications of Computer Vision. Waikoloa, USA: IEEE: 2401-2410 [DOI: 10.1109/WACV48630.2021.00245]

Grill J B, Strub F, Altché F, Tallec C, Richemond P H, Buchatskaya E, Doersch C, Pires B A, Guo Z D, Azaret M G, Piot B, Kavukcuoglu K, Munos R and Valko M. 2020. Bootstrap your own latent a new approach to self-supervised learning//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: 21271-21284

He K M, Fan H Q, Wu Y X, Xie S N and Girshick R. 2020. Momentum contrast for unsupervised visual representation learning//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 9726-9735 [DOI: 10.1109/CVPR42600.2020.00975]

Huang Q H, Huang C Q, Wang X Z and Jiang F. 2021. Facial expression recognition with grid-wise attention and visual Transformer. Information Sciences, 580: 35-54 [DOI: 10.1016/j. ins. 2021.08.043]

Li H Y, Wang N N, Ding X P, Yang X and Gao X B. 2021. Adaptively learning facial expression representation via C-F labels and distillation. IEEE Transactions on Image Processing, 30: 2016-2028 [DOI: 10.1109/tip.2021.3049955]

Li S and Deng W H. 2020. Deep facial expression recognition: a survey. Journal of Image and Graphics, 25(11): 2306-2320 (李珊, 邓伟洪. 2020. 深度人脸表情识别研究进展. 中国图象图形学报, 25(11): 2306-2320) [DOI: 10.11834/jig.200233]

Li S, Deng W H and Du J P. 2017. 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, USA: IEEE: 2584-2593 [DOI: 10.1109/CVPR.2017.277]

Li Y, Zeng J B, Shan S G and Chen X L. 2019a. Occlusion aware facial expression recognition using CNN with attention mechanism. IEEE Transactions on Image Processing, 28(5): 2439-2450 [DOI: 10.1109/TIP.2018.2886767]

Li Y J, Lu G M, Li J X, Zhang Z and Zhang D. 2020. Facial expression recognition in the wild using multi-level features and attention mechanisms. IEEE Transactions on Affective Computing, 14(1): 451-462 [DOI: 10.1109/TAFFC.2020.3031602]

Li Y J, Lu Y, Li J X and Lu G M. 2019b. Separate loss for basic and compound facial expression recognition in the wild//Proceedings of the 11th Asian Conference on Machine Learning. Nagoya, Japan: PMLR: 897-911

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

［DOI：10.1109/TAFFC.2017.2740923］

Peng X J and Qiao Y. 2020. Advances and challenges in facial expression analysis. Journal of Image and Graphics，25(11)：2337-2348（彭小江，乔宇. 2020. 面部表情分析进展和挑战. 中国图象图形学报，25(11)：2337-2348）［DOI：10.11834/jig.200308］

Vaswani A，Shazeer N，Parmar N，Uszkoreit J，Jones L，Gomez A N，Kaiser Ł and Polosukhin I. 2017. Attention is all you need//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach，USA：Curran Associates Inc.：6000-6010

Wang K，Peng X J，Yang J F，Lu S J and Qiao Y. 2020a. Suppressing uncertainties for large-scale facial expression recognition//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle，USA：IEEE：6896-6905［DOI：10.1109/CVPR42600.2020.00693］

Wang K，Peng X J，Yang J F，Meng D B and Qiao Y. 2020b. Region attention networks for pose and occlusion robust facial expression recognition. IEEE Transactions on Image Processing，29：4057-4069［DOI：10.1109/TIP.2019.2956143］

Yao H X，Deng W H，Liu H H，Hong X N，Wang S J，Yang J F and Zhao S C. 2022. An overview of research development of affective computing and understanding. Journal of Image and Graphics，27(6)：2008-2035（姚鸿勋，邓伟洪，刘洪海，洪晓鹏，王甦菁，杨巨峰，赵思成. 2022. 情感计算与理解研究发展概述. 中国图象图形学报，27(6)：2008-2035）［DOI：10.11834/jig.220085］

Zbontar J，Jing L，Misra I，LeCun Y and Deny S. 2021. Barlow twins：self-supervised learning via redundancy reduction//Proceedings of the 38th International Conference on Machine Learning. Virtual Event：PMLR：12310-12320

Zeng D，Lin Z Y，Yan X，Liu Y T，Wang F and Tang B. 2022. Face2Exp：combating data biases for facial expression recognition//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans，USA：IEEE：20259-20268［DOI：10.1109/CVPR52688.2022.01965］

Zhang Y H，Wang C R，Ling X and Deng W H. 2022. Learn from all：erasing attention consistency for noisy label facial expression recognition//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv，Israel：Springer：418-434［DOI：10.1007/978-3-031-19809-0_24］

Zhao Z Q，Liu Q S and Zhou F. 2021. Robust lightweight facial expression recognition network with label distribution training. Proceedings of the AAAI Conference on Artificial Intelligence，35(4)：3510-3519［DOI：10.1609/aaai.v35i4.16465］

作者简介

崔鑫宇，男，硕士研究生，主要研究方向为人脸表情识别。
E-mail：xyc1@nwsuaf.edu.cn
王美丽，通信作者，女，教授，博士生导师，主要研究方向为计算机图形学、三维建模、仿真与可视化。
E-mail：wml@nwsuaf.edu.cn
何翀，男，博士研究生，主要研究方向为计算机视觉。
E-mail：chonghe@nwafu.edu.cn
赵宏珂，男，硕士研究生，主要研究方向为计算机视觉。
E-mail：zhaohk9896@nwafu.edu.cn