融合局部特征与两阶段注意力权重学习的面部表情识别*

郑 剑，郑 炽，刘 豪，于祥春†
（江西理工大学 信息工程学院，江西 赣州 341000）

摘 要：面部的局部细节信息在面部表情识别中扮演重要角色，然而现有的方法大多只关注面部表情的高层语义信息而忽略了局部面部区域的细粒度信息。针对这一问题，提出一种融合局部特征与两阶段注意力权重学习的深度卷积神经网络 FLF-TAWL（deep convolutional neural network fusing local feature and two-stage attention weight learning），它能自适应地捕捉重要的面部区域从而提升面部表情识别的有效性。该 FLF-TAWL 由双分支框架构成，一个分支从图像块中提取局部特征，另一个分支从整个表情图像中提取全局特征。首先提出了两阶段注意力权重学习策略，第一阶段粗略学习全局和局部特征的重要性权重，第二阶段进一步细化注意力权重，并将局部和全局特征进行融合；其次，采用一种区域偏向损失函数鼓励最重要的区域以获得较高的注意力权重。在 FERPlus、Cohn-Kanada（CK + ）以及 JAFFE 三个数据集上进行了广泛实验，分别获得 90.92%、98.90%、97.39% 的准确率，实验结果验证了 FLF-TAWL 模型的有效性和可行性。

关键词：面部表情识别；深度卷积神经网络；局部特征融合；两阶段注意力权重学习；区域偏向损失
中图分类号：TP391.41 **文献标志码**：A **文章编号**：1001-3695(2022)03-043-0889-06
doi:10.19734/j.issn.1001-3695.2021.07.0287

Deep convolutional neural network fusing local feature and two-stage attention weight learning for facial expression recognition

Zheng Jian, Zheng Chi, Liu Hao, Yu Xiangchun†
(*School of Information Engineering, Jiangxi University of Science & Technology, Ganzhou Jiangxi* 341000, *China*)

Abstract: Facial local detail information plays an important role in facial expression recognition(FER). However, most of the existing methods only focus on the high-level semantic information of facial expressions, while ignoring the fine-grained information of local facial regions. To solve this problem, this paper proposed a deep convolutional neural network fusing local feature and two-stage attention weight learning(FLF-TAWL), which could adaptively capture important facial regions to improve the effectiveness of facial expression recognition. The FLF-TAWL model was composed of a dual-branch framework, one branch extracted local features from image blocks, and the other branch extracted global features from the entire expression image. Firstly, this paper proposed a two-stage attention weight learning strategy. In the first stage, it roughly learned the importance weights of global and local features, in the second stage, it further refined the attention weight, and fused the local and global features. Secondly, the model used a region-biased loss function to encourage the most important regions to obtain higher attention weights. Finally, this paper carried out extensive experiments on FERPlus, Cohn-Kanada(CK +) and JAFFE datasets to obtain accuracy rates of 90.92%, 98.90% and 97.39% respectively. The experimental results verify the effectiveness and feasibility of the FLF-TAWL model.

Key words: facial expression recognition; deep convolutional neural network(DCNN); fusing local feature; two-stage attention weight learning; region-biased loss function

0 引言

面部表情识别（facial expression recognition，FER）可辅助计算机理解人类行为从而完成有效的人机交互，其应用极其广泛，如智能教学系统、服务机器人、智能人机交互以及驾驶员疲劳监控等。近年来，基于深度学习的面部表情识别研究已成为国内外学术研究的热点。

一般来说，面部表情可以分为七种基本表情，包括愤怒、厌恶、恐惧、快乐、悲伤、惊讶以及自然表情[1]，表情识别的任务就是对这七类基本表情进行分类。面部表情识别不同于其他图像识别，需要对面部特征进行精细的刻画才能更加精确地完成识别任务。近年来，深度卷积神经网络（DCNN）在计算机视觉领域取得了巨大的成功，DCNN 能够自动从原始数据中提取有效特征，具有自适应学习特征表达的能力，相比手工特征具有更好的高层语义表达和本质映射能力。许多研究利用 DCNN 来改善FER 的性能。最早，Tang[2] 和 Kahou 等人[3] 设计了更深的 DCNN 用于面部特征提取，分别赢得了 FER2013 和 EMotiw2013 表情识别挑战赛的冠军。Ding 等人[4] 提出了一种联合训练 FER 任务和人脸识别任务的 FaceNet2ExpNet 架构。Albanie 等人[5] 利用 VGGFace 2.0 上预训 SeNet50 进行迁移学

收稿日期：2021-07-03；**修回日期**：2021-08-30 **基金项目**：国家自然科学基金资助项目（61563069，61462034）；江西省教育厅科学技术研究项目（GJJ170517，GJJ190468）；江西理工大学研究生创新专项资金资助项目（ZS2020-S049）

作者简介：郑剑（1977-），男，湖北武汉人，副教授，博士，主要研究方向为计算机视觉、大数据隐私保护；郑炽（1996-），男，湖北黄冈人，硕士研究生，主要研究方向为计算机视觉、图像处理；刘豪（1998-），男，江西九江人，硕士研究生，主要研究方向为深度学习、图像处理；于祥春（1989-），男（通信作者），山东泰安人，讲师，博士，主要研究方向为计算机视觉、深度学习（yuxc@jxust.edu.cn）.

习,并使用 softmax 分类函数在 FERPlus 数据集上进行微调。同时,心理学研究表明[6],人类可以有效地利用局部区域和整体区域来感知不完整的面部所传递的语义信息。Majumder 等人[7]研究发现面部表情变化通常与一些特定的面部区域,如嘴巴、眼睛以及鼻子等存在密切关联,这意味着局部面部区域特征对面部表情识别至关重要。姚丽莎等人[8]提出一种基于卷积神经网络局部特征融合的面部表情识别方法,通过构建的 DCNN 模型提取眼睛、眉毛以及嘴巴三个局部区域特征,然后采用 SVM 多分类器进行决策级加权融合,取得了较好的识别结果。Wang 等人[9]设计了一种基于局部区域的注意力网络,用来解决 FER 问题中姿势和遮挡的干扰问题。Xie 等人[10]提出双分支的 DCNN 将面部全局特征和局部特征简单地融合在一起,丰富了面部表情特征,但是不能自动抑制不相关的局部区域,在一定程度上限制了该方法的性能。Li 等人[11]提出了一种抗面部遮挡的表情识别方法,利用注意力机制使网络关注未遮挡的部分从而提高识别效果,但是该方法所获得的关键注意力权重还不够精细。最近,Ben 等人[12]不仅对微表情识别进行了全面的综述,还对宏观表情识别的基本技术、最新进展和主要挑战进行了系统的阐述和讨论。

综上所述,在心理学研究以及上述工作基础上,本文对基于 DCNN 的 FER 方法进行相应改进,提出了融合局部特征与两阶段注意力权重学习的深度卷积神经网络模型 FLF-TAWL。该模型更加关注局部特征的重要性,能够提取更加精细的面部局部细节信息,更全面地表征表情信息。本文的主要工作如下:a)设计了一个包含两分支的特征网络融合框架,即全局面部特征提取模块和局部特征提取模块,该融合框架同时融合面部表情全局特征和局部子块特征,实现两个尺度信息的相互补充,更全面地表示表情图像;b)提出了一种两阶段注意力权重计算策略,在第一阶段通过自注意力权重模块粗略计算局部子块的注意力权重,在第二阶段通过关系注意力模块对拼接后的特征进一步细化注意力权重,完成注意力权重由粗到细的计算,自动感知具有判别性的局部图像子块和抑制非重要的局部图像子块;c)有机整合对数加权交叉熵损失(WCE-loss)和面部局部图像子块区域排名正则化损失(RR-loss),目的是完成目标任务的联合优化,从而使得本文模型能够获得更优的注意力权重参数和更具判别性的识别效果。

1 相关理论

1.1 特征融合网络

许多研究工作通过设计相应的深度卷积网络来完成不同类型特征的融合,这通常比使用单一类型特征的网络能获得更好的识别效果。例如,Majumder 等人[7]从表情图像中提取 LBP 特征和面部几何特征,这两种类型特征最终通过两层自动编码器进行融合,获得了可观的效果;彭玉青等人[13]提出将一种将卷积神经网络与 DenseSIFT 特征进行融合的混合模型,从输入信息中提取出了更为细微的特征,从而有效地提升了表情识别率;Sun 等人[14]提出了一种多通道深度时空特征融合神经网络(MDSTFN)来执行静态图像的深度时空特征提取和融合,该网络同时捕获了时空特征从而取得了满意的效果。然而值得注意的是,现有的基于深度学习的方法大多只关注面部表情的高层语义信息而忽略了局部面部区域的细粒度信息,与已有的工作不同,本文提出了一种可以同时有效融合全局和局部面部特征的方法,同时本文方法也致力于挖掘局部细节信息在表情识别中的重要性。

1.2 注意力网络

注意力机制起初是在强化学习的基础上发展而来。Mnih 等人[15]使用带有注意力机制的 RNN 模型进行图像分类,并成功地应用到了机器翻译任务。之后,越来越多的研究者针对不同的研究任务提出了不同的自注意力模型。Wang 等人[16]提出了一种用于人脸检测的注意力网络,其在多选框生成步骤中突出显示面部区域。Yang 等人[17]提出了一种神经聚合网络(neural aggregation network,NAN),NAN 使用级联注意力机制来融合视频的面部特征或将其设置为紧凑的视频表示。在 NAN 模型的启发下,本文将注意力机制引入所设计的模型中。

2 本文方法

2.1 FLF-TAWL 网络

本文所提 FLF-TAWL 模型如图 1 所示,它通过两个独立分支有效地融合全局和局部面部的深层特征信息。全局面部特征提取分支从整幅面部图像中提取整体特征。局部特征提取分支从带重叠面部图像裁剪子块中提取局部特征,将局部面部区域按照第一阶段注意力权重系数进行加权聚合后得到局部聚合特征,接着在第二阶段注意力权重计算将这两个分支得到的输出特征进行聚合,目的是有效覆盖面部表情图像的全局和局部尺度,同时有效实现两个尺度信息的相互补充。这两个分支的有机融合不仅丰富了特征提取尺度,而且在一定程度上降低了 FER 识别中干扰因素的影响并提升了模型的表示能力,从而增强了模型的泛化能力。整幅面部图像表示为 I,面部图像的副本表示为 $\boldsymbol{x}_0$,均匀裁剪的带重叠的局部区域依次为 $\boldsymbol{x}_1,\cdots,\boldsymbol{x}_L$,$L$ 为每幅面部图像所裁剪的局部图像子块数。当输入图像为 I 时,网络的输入数据集用 X 表示为

$$X=[\boldsymbol{x}_0,\boldsymbol{x}_1,\cdots,\boldsymbol{x}_L] \tag{1}$$

其中:$\boldsymbol{x}_0,\boldsymbol{x}_1,\cdots,\boldsymbol{x}_L$ 是各部分图像的矩阵表示。将 X 分别输入两分支的主干网络进行特征提取,分别得到全局特征和局部特征 $\boldsymbol{F}$,具体表示如下:

$$\begin{aligned}\boldsymbol{F}=&(\boldsymbol{v}_h^0;\boldsymbol{v}_l^1;\boldsymbol{v}_l^2;\cdots;\boldsymbol{v}_l^k;\cdots;\boldsymbol{v}_l^L)=\\&(r(\boldsymbol{x}_0;\theta);r(\boldsymbol{x}_1;\theta);\cdots;r(\boldsymbol{x}_L;\theta))\end{aligned} \tag{2}$$

其中:$\boldsymbol{v}_h^0$ 为全局面部特征 CNN 提取模块所提取的全局特征;$\boldsymbol{v}_l^k$ 为局部面部特征 CNN 提取模块所提取的第 k 个局部特征,$k=1,2,3,\cdots,L$;$r(.;\theta)$ 表示特征提取网络 CNN,θ 是特征提取网络 CNN 中的参数。局部特征被输入到自注意力权重模块进行第一阶段注意力权重计算,全局特征和局部特征在关系注意力模块进行第二阶段注意力权重计算,获得最终的聚合特征后以全连接形式输入到 softmax 的分类器中,softmax 函数如式(3)所示,其中 C 为表情类别数。

$$f(z_j)=\frac{\mathrm{e}^{z_j}}{\sum_{c=1}^{C}\mathrm{e}^{z_c}} \tag{3}$$

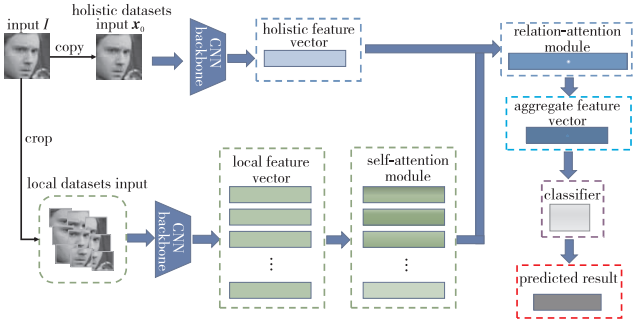

图 1 FLF-TAWL 模型结构

Fig. 1 FLF-TAWL model structure

2.2 两阶段注意力权重计算模式

不同的面部局部图像子块将在 FER 任务中扮演不同的角色。为了在网络的训练中自动感知具有判别性的局部图像子

块和抑制非重要的局部图像子块,本文设计了一种两阶段注意力权重计算模式:a)引入了自注意力权重加权模块和排名正则化来对面部图像子块的贡献度进行排名,具有较高判别性的局部子块被赋予较高的重要性权重,同时判别性较弱的局部子块被赋予较低的重要性权重; b)在获得粗略计算的局部子块注意力权重后,该模式又引入关系注意力模块对局部子块特征以及全局面部特征分别与来自第一阶段融合后的表征进行关系建模以寻求细化的注意力权重。两阶段注意力权重计算模式的具体设计如图 2 所示。

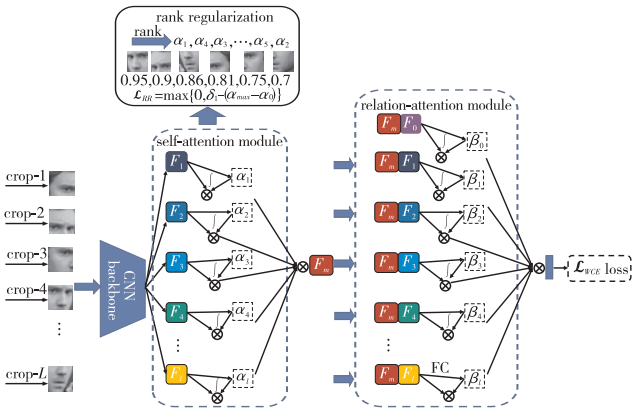

图 2　两阶段注意力权重计算模式的详细架构

Fig. 2　Detailed architecture of two-stage attention weight learning strategy

2.2.1　第一阶段注意力权重计算

1)自注意力权重模块　由式(2)可知第 k 个局部特征向量为 $\boldsymbol{v}_l^k$,按通道融合得到的 $\boldsymbol{F}\in\mathbb{R}^{D\times L}$,其中 D 为特征维度,L 为局部图像子块数。自注意力权重加权模块输入为每幅图像中所有局部子块的特征,输出为每个局部子块的粗略注意力权重。具体地,自注意力权重加权模块计算公式为

$$\alpha_k=\sigma(\boldsymbol{W}_a\boldsymbol{v}_l^k) \tag{4}$$

其中:$\boldsymbol{W}_a$ 为自注意力权重加权模块全连接 FC 层的权重,该权重与局部子块特征 $\boldsymbol{v}_l^k$ 进行向量相乘,通过 sigmoid 函数 σ 过滤后得到第 k 个局部子块注意力权重 α_k。本模块得到的粗略计算后的局部子块注意力权重,将会用于后续第二阶段注意力计算模块,从而得到更加细化的注意力权重。

2)面部局部图像子块排名正则化　不同类别面部表情的形成由面部不同的子区域所影响[4]。为了深度挖掘不同面部局部子块的重要性,本文采用面部局部图像子块区域排名正则化来提升具有判别性的局部图像子块的权重和抑制非重要的局部图像子块权重。如图 2 所示,在排名正则化约束模块中,首先对局部子块特征按照自注意力权重模块所得到的注意力权重 $\alpha_k\in[0,1]$ 的大小进行降序排列;然后要求局部子块中最大注意力权重应大于全局面部图像的注意力权重,两者之间的边距阈值由超参数 δ_1 来控制。本文使用下面的局部子块排名正则化损失函数 RR-loss 来实现面部局部图像子块排名正则化约束,即

$$\mathcal{L}_{RR}=\max\{0,\delta_1-(\alpha_{max}-\alpha_0)\} \tag{5}$$

其中:δ_1 为边距阈值超参数;α_0 为原始图像副本(即整体面部图像)的注意力权重;α_{max} 为局部子块中注意力权重最大值。排名正则化损失函数所施加约束的目的是突出某些局部子块特征(如嘴巴、眼睛以及鼻子等),同时抑制非重要局部子块特征(如额头等)。排名正则化约束策略可以让模型深度挖掘判别性更强的面部局部表情特征。在自注意力权重模块得到粗略计算的 α_k 后,本文将所有局部特征 $\boldsymbol{v}_l^k$ 及其注意力权重进行有机整合从而得到第一阶段的聚合特征 $\boldsymbol{F}_m$。

$$\boldsymbol{F}_m=\frac{1}{\sum_{k=1}^{L}\alpha_k}\sum_{k=1}^{L}\alpha_k F_k \tag{6}$$

其中:$\boldsymbol{F}_k$ 为局部特征;$\boldsymbol{F}_m$ 为第一阶段所得到的自注意力权重集合的特征。

2.2.2　第二阶段注意力权重计算

第一阶段获得的注意力权重在一定程度上是粗糙的,例如所得到的针对于每一个局部子块的注意力权重并不具备感知剩余其他局部子块的信息,从而缺乏全局判别能力。本文提出通过第二阶段的关系注意力策略来进一步细化逐个局部子块的注意力权重。具体来讲,首先将第一阶段所获得的自注意权重聚合特征 $\boldsymbol{F}_m$ 分别与每个局部子块特征 $\boldsymbol{v}_l^k$ 以及全局特征 $\boldsymbol{v}_h^0$ 进行拼接;然后通过全连接层来分别自动学习各个局部子块特征 $\boldsymbol{v}_l^k$ 以及全局特征 $\boldsymbol{v}_h^0$ 与该自注意权重聚合特征 $\boldsymbol{F}_m$ 之间的关系;最后继续按照式(4)所描述的方式得到进一步细化的注意力权重。第二阶段的关系注意力模块中第 k 个区域的细化注意权重表示为

$$\beta_k=\sigma(\boldsymbol{W}_\beta[\boldsymbol{F}_k:\boldsymbol{F}_m]^{\mathrm{T}}) \tag{7}$$

其中:$\boldsymbol{W}_\beta$ 是关系注意力模块全连接 FC 层的权重,该权重与局部子块特征 $\boldsymbol{F}_k$ 和 $\boldsymbol{F}_m$ 的拼接特征进行向量点乘,通过 sigmoid 函数 σ 过滤后得到第 k 个细化的局部子块注意力权重 β_k。最后将两阶段注意力权重计算进行整合,得到最终的聚合特征为

$$P_{FLF\text{-}TAWL}=\frac{1}{\sum_{k=0}^{n}\beta_k}\sum_{k=0}^{n}\beta_k[\boldsymbol{F}_k:\boldsymbol{F}_m] \tag{8}$$

其中:$\boldsymbol{F}_m$ 为第一阶段的聚合特征,具体如式(6)所示;对于 $\boldsymbol{F}_k$,当 $k=0$ 时,$\boldsymbol{F}_0$ 为 $\boldsymbol{v}_h^0$;当 $k>0$ 时,$\boldsymbol{F}_k$ 为 $\boldsymbol{v}_l^k$。$P_{FLF\text{-}TAWL}$ 将作为 FLF-TAWL 网络最终的特征表征。

2.2.3　WCE-loss 与 RR-loss 联合优化

通过上述设计的两阶段注意力权重计算模块得到了最终的聚合特征 $P_{FLF\text{-}TAWL}$,注意力权重在上述特征提取过程中扮演了重要角色,受文献[15,18]的启发,本文将所得到的注意力权重用于损失加权,目的是从目标损失函数角度引导注意力权重参数的学习,从而进一步完成具有判别性局部子块特征的提升和非重要局部子块特征的抑制。本文设计了对数加权交叉熵损失(WCE-loss)来完成目标优化任务,具体表示为

$$\mathcal{L}_{WCE}=-\frac{1}{N}\sum_{i=1}^{N}\log\frac{\mathrm{e}^{\beta_i\boldsymbol{W}_{y_i}^{\mathrm{T}}\boldsymbol{x}_i}}{\sum_{j=1}^{C}\mathrm{e}^{\beta_i\boldsymbol{W}_j^{\mathrm{T}}\boldsymbol{x}_i}} \tag{9}$$

其中:$\boldsymbol{W}_j$ 是第 j 个类别的网络分类参数,由上述公式可推断最终的损失函数 $\mathcal{L}_{WCE}$ 与注意力权重 β_i 成正相关。同时,RR-loss 损失可以直接衡量面部表情局部子块的注意力权重偏向于提升或偏向于抑制,因此,本文将 WCE-loss 和 RR-loss 进行整合,完成目标任务的联合优化。FLF-TAWL 网络的整体损失函数为

$$\mathcal{L}_{all}=\gamma\,\mathcal{L}_{RR}+(1-\gamma)\mathcal{L}_{WCE} \tag{10}$$

其中:γ 为平衡 $\mathcal{L}_{RR}$ 和 $\mathcal{L}_{WCE}$ 两种损失系数,在训练过程中通过上述的联合优化方式鼓励 FLF-TAWL 网络获得更优的注意力权重参数和更具判别性的识别性能。

2.3　裁剪方式

将面部图像裁剪出多个局部子块是 FLF-TAWL 中的一个基本任务。裁剪区域过大将导致特征的多样性降低,裁剪区域过小将导致区域特征的区分能力不足。本文重点研究三种形式的局部子块裁剪方案,即固定位置裁剪、随机裁剪以及基于关键点位置裁剪,如图 3 所示。

a)固定位置剪裁。以固定的比例在固定的位置进行局部子块裁剪。具体地,使用该方法裁剪五个区域,其中三个是左上、右上和中下的面部区域,其大小固定为原始人脸的 0.75 比例;另外两个区域类似于微笑分类任务中使用的区域,裁剪原始面部图像大小为 0.9 和 0.85 比例的中心区域[19]。

b)随机剪裁。在基于深度学习面部识别任务中,

DeepID[20] 在每幅面部表情图像进行 200 次随机裁剪，得到更多的局部子块来提高其性能。本文在随机裁剪过程中随机裁剪 N 个区域，其中随机区域的尺寸比例为原始人脸的 0.7 ~ 0.95 不等。

c）基于关键点位置剪裁。给定面部表情关键点，在关键点周围的区域进行裁剪。本文使用 MTCNN[21] 来检测五个典型的面部标志点（即左眼、右眼、鼻子、左嘴角和右嘴角），并根据这些标志点为中心点得到半径为 r 的裁剪区域。最后将所有裁剪下来的局部子块进行缩放至 64 × 64 的统一大小。

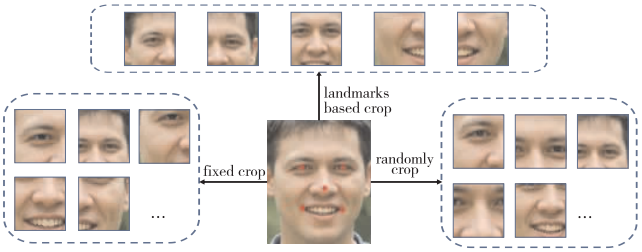

图 3 人脸图像的三种裁剪方法

Fig3 Three clipping methods for the face image

3 实验验证与结果分析

为了验证本文提出的 FLF-TAWL 模型的有效性，在三个公开的面部表情数据集上进行了大量的实验，分别是 FERPlus、CK + 以及 JAFFE 数据集，这三个数据集的部分样本示例如图 4 所示。本实验是在 Ubuntu 18.04.5 LTS 环境下，基于 TensorFlow 实现完成的。实验硬件平台为 Intel ® Core™ i5-6500 CPU，主频为 3.2 GHz，内存为 8 GB，同时借助显存 12 GB 的 NVIDIA GeForce RTX 2080Ti GPU 进行加速处理。

3.1 数据集与数据预处理

FERPlus、CK + 以及 JAFFE 数据集在面部表情识别领域应用广泛，许多面部表情识别方法均在该数据集上进行验证。FERPlus 和 CK + 都包含八种基本表情，即自然、开心、惊讶、悲伤、生气、厌恶、恐惧以及轻蔑。FERPlus 数据集由 ICML2013 挑战赛中的 FER2013 数据集扩展而来，该数据集是通过谷歌搜索引擎从互联网上收集的大规模数据集，由 28 709 张训练图像、3 589 张验证图像以及 3 589 张测试图像组成。相比于 FER2013，扩展后的 FERPlus 数据集的标签精度更高，在图 4 中第 1 行显示了该数据集的一些样本。CK + 数据集是一个动态表情数据集，它包含来自 123 个人共 593 例的动态表情图像序列，每一个序列都包含表情从平静到表情峰值的所有帧，但是仅有 327 个图像序列带表情标签。图 4 中第 2 行显示了 CK + 数据集的部分样本示例，本文选取了 327 个共有八种基本表情类别的序列进行实验，对每个序列收集最后 3 帧峰值的表情帧作为表情图像。JAFFE 数据集是最常用的静态图像数据集，它包含 10 名日本女性共 213 张大小为 256 × 256 的面部正面静态图像，每人都有除轻蔑类别之外的七种基本表情，其中每种表情有 2 ~ 4 幅图像，该数据集标签比较标准，图 4 第 3 行显示了 JAFFE 数据集的部分样本示例。

在训练模型前先对数据进行预处理。在实验中针对 FERPlus 和 CK + 数据集中样本数据的不平衡问题，采用数据增强来提高样本数量的均衡分布，从而尽可能避免因样本数量不均衡所造成的面部表情识别率下降的影响。例如执行图像水平翻转，每个图像顺时针和逆时针旋转 5°，此外还可以通过随机添加具有零均值和 0.01 方差的高斯噪声等方式获得更多的样本。针对人脸表情识别易受人脸光照和姿态的影响，采用如图 5 所示的 MTCNN 人脸检测器[23] 检测所有选定面部图像中的人脸并进行面部对齐，对齐之后再通过直方图均衡化将图像的直方图分布变成近似均匀分布以增加图像对比度、增强图像细节。因此，经过数据预处理后的实验样本集得到了很大的扩展和丰富。表 1 显示了经过数据预处理后实验中选取样本的数量分布情况。

图 4 实验中使用的 FERPlus、CK + 、JAFFE 数据集部分图像

Fig. 4 Some examples of FERPlus、CK + 、JAFFE datasets

表 1 数据预处理后 FERPlus、CK + 和 JAFFE 数据集数量分布

Tab. 1 Sample size of FERPlus、CK + and JAFFE datasets after data preprocessing

表情	数量			表情	数量		
	FERPlus	CK +	JAFFE		FERPlus	CK +	JAFFE
angry	19 812	564	120	happiness	26 943	912	124
neutral	17 721	1 176	116	sadness	18 231	420	120
disgust	2 192	672	128	surprise	13 386	936	120
fear	15 363	348	124	contempt	1 664	216	0

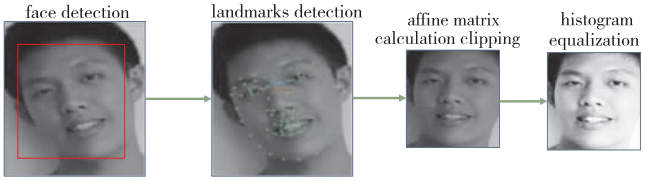

图 5 表情图像预处理

Fig. 5 Facial expression image preprocessing

3.2 实验设置与实现细节

本文利用迁移学习思想，分别选择 VGG16 以及 ResNet50 作为主干网络，其中 VGG16 和 ResNet50 分别在 VGG-Face 2.0、MSCeleb-1M 人脸识别数据集上进行了预训练。

为了与五点关键点裁剪出的局部图像数量相等，在固定裁剪的训练阶段，本文使用所有五个区域以及每个原始人脸图像的副本（即图 2 中的 $L = 5$）作为网络输入；对于随机裁剪的训练，本文采用随机裁剪的区域替换固定裁剪的五个区域。当使用 RB-loss 和 WCE-loss 进行联合训练时，默认的权重比为 1 ∶ 1，它们之间的占比对表情识别的影响将在随后的消融实验中进行研究。在所有数据集上，学习率初始化为 0.01，每隔 15 个 epoch 学习率减少 10 倍，$epoch = 100$，RR-loss 中的超参数 δ_1 默认设置为 0.02。为了评估该方法的性能，所有实验均采用 10 折交叉验证（即图像被随机分成 10 个等大小的子集，9 个子集用于训练，剩余的 1 个子集用于测试）。最后的结果通过平均识别精度得出。

3.3 实验结果分析

为了进一步验证本文提出的 FLF-TAWL 模型的有效性，首先采用以 ResNet50 作为 FLF-TAWL 的主干网络，按照图 3 中三种裁剪方式得到的局部和全局图像数据作为网络输入，其中随机裁剪中分别随机 9、30、60 次，随机取五个局部图像输入模型，分别得到三种裁剪方式的平均识别准确率；另外本文复现了文献[10]的 DCMA-CNN 算法作为对比方法；同时还将原始的人脸图像作为输入，对传统方法 VGG16 + SVM 进行微调作为基线对比模型，实验对比结果如表 2 ~ 4 所示，同时在三个数据集上的可视化结果如图 6 所示。

表 2　FLF-TAWL 模型上三种裁剪方式识别准确率的对比

Tab. 2　Comparison of recognition accuracy of three clipping methods on FLF-TAW model　/%

数据集	裁剪方式				
	固定裁剪	关键点位置裁剪	随机裁剪(9)	随机裁剪(30)	随机裁剪(60)
FERPlus	90.41	**90.92**	90.11	90.32	89.71
CK +	98.61	**98.90**	98.30	98.52	98.11
JAFFE	97.09	**97.39**	96.77	96.99	96.52

表 3　DCMA-CNN 模型上三种裁剪方式识别准确率的对比

Tab. 3　Comparison of recognition accuracy of three clipping methods on DCMA-CNN model　/%

数据集	裁剪方式				
	固定裁剪	关键点位置裁剪	随机裁剪(9)	随机裁剪(30)	随机裁剪(60)
FERPlus	85.98	88.67	84.15	86.01	82.33
CK +	91.31	93.46	90.47	94.44	91.23
JAFFE	92.01	94.75	91.95	93.10	89.12

表 4　基线模型的识别准确率的对比

Tab. 4　Comparison of recognition accuracy on baseline model　/%

方法	FERPlus	CK +	JAFFE
VGG16 + SVM	81.24	91.67	89.70

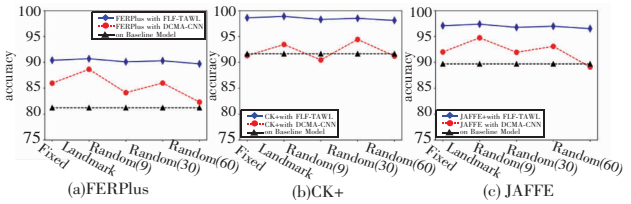

图 6　对比结果可视化展示

Fig. 6　Visual presentation of comparison results

通过表 2 ~ 4 及图 6 可以看出,本文提出的方法取得了最优结果。与传统的基线模型比较,输入单一特征的人脸表情图像只能从整个表情图像中提取特征,仅强调面部表情的完整性,从而忽略了局部细节信息,因此识别精度不高。与 DCMA-CNN 相比,本文的 FLF-TAWL 在三种裁剪方式上表现更稳定,说明本文方法更易学习到局部面部区域的细粒度信息,从而充分利用编码在表情图像中的有效识别信息达到较优的识别效果。另外,本文发现基于人脸关键点的裁剪方式产生的识别效果始终优于随机裁剪方式和固定裁剪方式,甚至使用多倍随机裁剪策略的情况下,网络模型也不会对识别精度提高很多。该实验结果表明,人类面部表情的变化通常发生在面部的一些显著区域,如嘴巴、嘴角、眼和鼻子周围区域。更重要的是,本文的 FLF-TAWL 模型在固定位置裁剪和基于关键点位置剪裁的识别率差异小于 DCMA-CNN 模型,这表明了 FLF-TAWL 模型可以有效突出某些局部子块特征(如嘴巴、眼睛以及鼻子等),同时抑制非重要局部子块特征,从而提升表情识别任务的区分性,后续实验最终选择用人脸关键点裁剪方式的数据输入 FLF-TAWL。图 7 中给出了本文 FLF-TAWL 模型在三个数据集下每个表情类别的混淆矩阵。

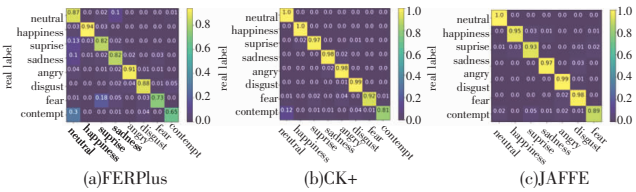

图 7　在不同数据集上的混淆矩阵

Fig. 7　Confusion matrixes on different datasets

从图 7 中可以看出,本文方法在中性、开心、生气、悲伤这四类表情上识别率最高,其中在 JAFFE 和 CK + 数据集上中性的表情识别率达到 100%,主要原因是 JAFFE 和 CK + 数据集中表情数据较为规范标准,中性类别的表情数量也最丰富。同样的方法在 JAFFE 和 CK + 数据集上的表现要优于 FERPlus 数据集,造成这种结果的原因是 FERPlus 数据集是一个从互联网上收集的大规模数据集,它更加符合大规模真实世界环境下的表情数据集,在光照、头部姿态以及面部遮挡等方面具有多样性,这也从侧面说明本文方法对光照等外界因素具有鲁棒性。

为了进一步对 FLF-TAWL 中的三个模块进行评估,本文设计了一项消融实验,研究 WCE-loss、自我注意力模块和关系注意力模块在三个数据集上的性能影响,结果如表 5 所示。对应表 5 中三个模块有效性的评估结果,在图 8 中展示了具体样本案例的实验结果。其中,最下方显示图片的原始标签;样本上的识别标签中绿色代表识别正确,红色代表识别错误(见电子版)。

表 5　FLF-TAWL 中的三个模块在各数据集上的性能评估

Tab. 5　Performance evaluation of three modules of FLF-TAWL on each dataset

WCE-loss	自我注意力模块	关系注意力模块	识别准确率/%		
			FERPlus	CK +	JAFFE
×	×	×	88.82	97.00	93.46
√	×	×	89.83	97.77	94.75
√	√	×	90.10	98.01	95.22
√	√	√	**90.92**	**98.90**	**97.39**

图 8　具体样本案例的实验结果

Fig. 8　Experimental results of specific sample cases

表 5 中第一行为基础模型,它使用传统的 softmax 损失函数替换 WCE-loss,并且去掉了所有注意力模块,选取的具体样本实验结果对应图 8 中的第一行。

对于这种训练方案,在基础模型上添加 WCE-loss 之后,在三个数据集上的识别精度都有所提升,这是因为该改进后的 WCE-loss 能很好地调整人脸特征的最大类内距离小于最小类间距离。通过图 8 第二行样本案例可以看出进一步拉近了惊讶和开心类别之间的距离,这也是本文模型性能体现的原因之一。

当再在表 5 第二行的基础上增加自我注意力模块,表情识别结果相比第二行的精度在三个数据集上精度提升为 0.27%、0.24%、0.47%。图 8 中第三行实验结果可以看出自我注意力模块能够明显提高识别性能,这种提升得益于自我注意力模块中粗略的权重值以及权重正则化带来的效果增益。在表 5 第四行加上关系注意力权重模块后,从图 8 中第四行实验结果可以看到,让原本眉毛内侧和上眼皮有着相似动作的惊讶和恐惧表情也能够正确区分开,使模型整体识别精度进一步提高,由此看出注意力机制对于分类准确率的提升有突出贡献,同时也证明了三个模块的有效性。

在同一数据集上将本文模型与其他的识别效果对比,如表 6 ~ 8 所示,可以看出,本文提出的 FLF-TAWL 模型在识别准确率上具有优势。基于深度学习的方法(如 Rest18 + VGG16、Em-AlexNet 和 C-LetNet5)采用单分支结构来提取图像特征,而本文方法通过增加一个分支来提取特征,从而更全面地表示表情。结果表明,局部特征提取的分支确实有利于表情分类。在 JAFFE 数据集上的实验涉及到旋转和噪声等变化的图像,但对比方法中大多是采用手工特征的方法,从实验结果来看 FLF-

TAWL 仍然可以正确地对大多数表达式进行分类,这说明了本文方法对表情图像的微小变化具有一定的鲁棒性。

另外对本文方法中的主干网络采用除 ResNet50 之外的另一个经典神经网络 VGG16 结构进行表情识别性能的对比验证。结果表明,本文方法中使用网络层数更深的 ResNet50 作为主干网络提取特征能力加强,从而使得识别率有所提升。

表 6 在 FERPlus 数据集上准确率对比

Tab. 6 Comparison of accuracy on FERPlus dataset /%

对比方法	准确率	对比方法	准确率
Rest18 + VGG16[9]	87.40	STSN[23]	89.66
SENet-50[4]	88.80	KTN[23]	90.49
SCN-(ResNet-18)[9]	89.35	FLF-TAWL(VGG16)	87.89
GCN + DCN *[24]	89.39	FLF-TAWL(ResNet50)	90.92

表 7 在 CK + 数据集上准确率对比

Tab. 7 Comparison of accuracy on CK + dataset /%

对比方法	准确率	对比方法	准确率
Khor-Net[27]	91.25	C-LetNet5[26]	83.74
Em-AlexNet[22]	94.25	MDSTFN[14]	98.38
CNN + DBN[25]	95.73	FLF-TAWL(VGG16)	94.51
FN2EN[4]	98.6	FLF-TAWL(ResNet50)	98.90

表 8 在 JAFFE 数据集上准确率对比

Tab. 8 Comparison of accuracy on JAFFE dataset /%

对比方法	准确率	对比方法	准确率
C-LetNet5[26]	94.37	Em-AlexNet[22]	93.02
LogGabor + HOG + DBN[25]	96.30	FLF-TAWL(VGG16)	93.10
Mob_Inc + DenseSIFT[21]	96.51	FLF-TAWL(ResNet50)	97.39

实验最后,本文在图 9 中评估了分类损失 WCE-loss 与自注意力模块中的排序正则化损失 RR-loss 之间不同比率 γ 对表情分类结果的影响。由图 9 可以发现,对这两个损失函数平均分配相等的权重可以获得最佳分类结果。将 RR-loss 的权重从 0.5 增加到 0.8,导致模型的识别性能显著降低。

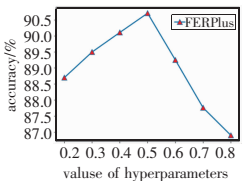

图 9 RR 损失与 WCE 损失之间的比率 γ 对结果的影响

Fig. 9 Effect of ratio γ between RR-loss and WCE-loss on expression classification results

4 结束语

本文提出了一种新的面部表情识别网络模型 FLF-TAWL。首先,该模型由两个独立的 CNN 分支机构组成,其中一个分支用于整幅面部表情图像特征提取,另一个分支对裁剪后的面部表情图像块进行局部特征提取。全局特征与局部特征的融合既丰富了面部表情特征又确保提取到的特征更具区分性。然后,在训练阶段提出了一种两阶段注意力权重计算策略,通过该注意力权重策略使得模型自动感知具有判别性的局部图像子块和抑制非重要的局部图像子块;将 WCE-loss 和 RR-loss 联合优化,加快了模型迅速收敛。最后,在三个公开的面部表情数据集上的大量实验验证了 FLF-TAWL 模型在提高识别精度、泛化能力的同时,也提高了识别算法的鲁棒性,在三个数据集上的分类结果优于其他许多有竞争力的工作。虽然本文的 FLF-TAWL 模型表现出了较好的性能,但仍存在一些不足,例如,本文的面部表情识别是基于静态图像的,而现实生活中的情感变化是有一定时间的,静态图像只能反映一个人在某个时间的表情状态。接下来的工作将研究动态人脸表情识别,致力于设计出更加精准的用于动态面部表情识别的网络模型。

参考文献:

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

[2] Tang Yichuan. Deep learning using linear support vector machines [EB/OL]. (2015-02-21). https://arxiv.org/pdf/1306.0239.pdf.

[3] Kahou S E, Pal C, Bouthillier X, *et al*. Combining modality specific deep neural networks for emotion recognition in video[C]//Proc of the 15th ACM on International Conference on Multimodal Interaction. New York:ACM Press,2013:543-550.

[4] Ding Hui, Zhou S K, Chellappa R. FaceNet2ExpNet: regularizing a deep face recognition net for expression recognition[C]//Proc of the 12th IEEE International Conference on Automatic Face & Gesture Recognition. Piscataway, NJ:IEEE Press,2017:118-126.

[5] Albanie S, Nagrani A, Vedaldi A, *et al*. Emotion recognition in speech using cross-modal transfer in the wild[C]//Proc of the 26th ACM International Conference on Multimedia. New York:ACM Press,2018:292-301.

[6] Yovel G, Duchaine B. Specialized face perception mechanisms extract both part and spacing information: evidence from developmental prosopagnosia[J]. Journal of Cognitive Neuroscience, 2006, 18(4):580-593.

[7] Majumder A, Behera L, Subramanian V K. Automatic facial expression recognition system using deep network-based data fusion[J]. IEEE Trans on Cybernetics,2016,48(1):103-114.

[8] 姚丽莎,徐国明,赵凤. 基于卷积神经网络局部特征融合的人脸表情识别[J]. 激光与光电子学进展,2020,57(4):041513. (Yao Lisha, Xu Guoming, Zhao Feng. Facial expression recognition based on local feature fusion of convolutional neural network[J]. Laser & Optoelectronics Progress,2020,57(4):041513.)

[9] Wang Kai, Peng Xiaojiang, Yang Jianfei, *et al*. Suppressing uncertainties for large-scale facial expression recognition[C]//Proc of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ:IEEE Press,2020:6897-6906.

[10] Xie Siyue, Hu Haifeng. Facial expression recognition using hierarchical features with deep comprehensive multipatches aggregation convolutional neural networks[J]. IEEE Trans on Multimedia, 2019, 21(1):211-220.

[11] Li Yong, Zeng Jiabei, Shan Shiguang, *et al*. Occlusion aware facial expression recognition using CNN with attention mechanism[J]. IEEE Trans on Image Processing, 2019, 28(5):2439-2450.

[12] Ben Xianye, Ren Yi, Zhang Junping, *et al*. Video-based facial micro-expression analysis: a survey of datasets, features and algorithms[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2021, DOI:10.1109/TPAMI.2021.3067464.

[13] 彭玉青,王纬华,刘璇,等. 基于深度学习与 Dense SIFT 融合的人脸表情识别[J]. 中国科学技术大学学报,2019,49(2):105-111. (Peng Yuqing, Wang Weihua, Liu Xuan, *et al*. Facial expression recognition based on fusion of deep learning and Dense SIFT[J]. Journal of University of Science & Technology of China, 2019, 49(2):105-111.)

[14] Sun Ning, Li Qi, Huan Ruizhi, *et al*. Deep spatial-temporal feature fusion for facial expression recognition in static images[J]. Pattern Recognition Letters, 2019, 119(3):49-61.

[15] Mnih V, Heess N, Graves A. Recurrent models of visual attention [C]//Proc of the 27th International Conference on Neural Information Processing Systems. Cambridge, MA:MIT Press,2014:2204-2212.

[16] Wang Jianfeng, Yuan Ye, Yu Gang. Face attention network: an effective face detector for the occluded faces[EB/OL]. (2017-11-22). https://arxiv.org/abs/1711.07246.

[17] Yang Jiaolong, Ren Peiran, Zhang Dongqing, *et al*. Neural aggregation network for video face recognition[C]//Proc of IEEE Conference on Computer Vision and Pattern Recognition. Washington DC: IEEE Computer Society,2017:5216-5225.

(下转第 918 页)

org/abs/1409.0473.
[13] Wu Qi, Shen Chunhua, Liu Lingqiao, *et al.* What value do explicit high-level concepts have in vision to language problems[C]//Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2016: 203-212.
[14] Gu Jiuxiang, Wang Zhenhua, Kuen J, *et al.* Recent advances in convolutional neural networks[J]. Pattern Recognition, 2018, 77: 354-377.
[15] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[C]//Proc of the 25th Conference and Workshop on Neural Information Processing Systems. 2012: 1097-1105.
[16] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.
[17] Donahue J, Hendricks L, Guadarrama S. Long-term recurrent convolutional networks for visual recognition and description[C]//Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2015: 2626-2634.
[18] Vinyals O, Toshev A, Bengio S, *et al.* Show and tell: a neural image caption generator[C]//Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2015: 3156-3164.
[19] Chen Long, Zhang Hanwang, Xiao Jun, *et al.* SCA-CNN: spatial and channel-wise attention in convolutional networks for image captioning [C]//Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2017: 6298-6306.
[20] Wang Qingzhong, Chan A B. CNN + CNN: convolutional decoders for image captioning[EB/OL]. (2018). https://arxiv.org/abs/1805.09019.
[21] Yao Ting, Pan Yingwei, Li Yehao, *et al.* Boosting image captioning with attributes[C]//Proc of IEEE International Conference on Computer Vision. Piscataway, NJ: IEEE Press, 2017: 4894-4902.
[22] You Quanzeng, Jin Hailin, Wang Zhaowen, *et al.* Image captioning with semantic attention[C]//Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2016: 4651-4659.
[23] Wang Yufei, Lin Zhe, Shen Xiaohui, *et al.* Skeleton key: image captioning by skeleton-attribute decomposition[C]//Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2017: 7378-7387.
[24] Kumar A, Barman D, Sarkar R, *et al.* Overlapping community detection using multi objective genetic algorithm[J]. IEEE Trans on Computational Social Systems, 2020, 7(3): 802-817.
[25] Elkelesh A, Ebada M, Cammerer S, *et al.* Decoder-tailored polar code design using the genetic algorithm[J]. IEEE Trans on Communications, 2019, 67(7): 4521-4534.
[26] Yan Jiaqi, Gou Yang, Zhang Siyu, *et al.* Output current optimization for multi brick parallel discharge drivers based on genetic algorithm [J]. IEEE Trans on Plasma Science, 2019, 47(6): 3015-3025.
[27] Alavi M, Henderson J C. An evolutionary strategy for implementing a decision support system[J]. Management Science, 1981, 27(11): 1309-1323.
[28] Lara A, Sanchez G, Coello A, *et al.* HCS: a new local search strategy for memetic multi objective evolutionary algorithms[J]. IEEE Trans on Evolutionary Computation, 2010, 14(1): 112-132.
[29] Huang Han, Su Junpeng, Zhang Yushan, *et al.* An experimental method to estimate running time of evolutionary algorithms for continuous optimization[J]. IEEE Trans on Evolutionary Computation, 2020, 24(2): 275-289.
[30] He Kaiming, Zhang Xiangyu, Ren Shaoqing, *et al.* Deep residual learning for image recognition[C]//Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2016: 770-778.
[31] Papineni K, Roukos S, Ward T, *et al.* BLEU: a method for automatic evaluation of machine translation[C]//Proc of the 40th Annual Meeting of the Association for Computational Linguistics. New York: ACM Press, 2002: 311-318.
[32] Lin T Y, Maire M, Belongie S, *et al.* Microsoft COCO: common objects in context[C]//Proc of the 12th European Conference on Computer Vision. Berlin: Springer, 2014: 740-755.
[33] Denkowski M, Lavie A. Meteor universal: language specific translation evaluation for any target language[C]//Proc of the 9th EACL Workshop on Statistical Machine Translation. New York: ACM Press, 2014: 376-380.
[34] Lin C Y. ROUGE: a package for automatic evaluation of summaries [C]//Proc of the 42nd Annual Meeting of the Association for Computational Linguistics. New York: ACM Press, 2004: 74-81.
[35] Vedantam R, Zitnick C L, Parikh D. CIDER: consensus based image description evaluation[C]//Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2015: 4566-4575.
[36] Anderson P, Fernando B, Johnson M, *et al.* SPICE: semantic propositional image caption evaluation[C]//Proc of the 14th European Conference on Computer Vision. Cham: Springer, 2016: 382-398.
[37] Wang Hanzhang, Wang Hanli, Xu Kaisheng. Evolutionary recurrent neural network for image captioning[J]. Neurocomputing, 2020, 401: 249-256.

(上接第894页)

[18] Hu Wei, Huang Yangyu, Zhang Fan, *et al.* Noise-tolerant paradigm for training face recognition CNNs[C]//Proc of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2019: 11879-11888.
[19] Zhang Kaipeng, Tan Lianzhi, Li Zhifeng, *et al.* Gender and smile classification using deep convolutional neural networks[C] //Proc of IEEE Conference on Computer Vision and Pattern Recognition. Washington DC: IEEE Computer Society, 2016: 34-38.
[20] Sun Yi, Wang Xiaogang, Tang Xiaoou. Deep learning face representation from predicting 10 000 classes[C]//Proc of IEEE Conference on Computer Vision and Pattern Recognition. Washington DC: IEEE Computer Society, 2014: 1891-1898.
[21] Zhang Kaipeng, Zhang Zhanpeng, Li Zhifeng, *et al.* Joint face detection and alignment using multitask cascaded convolutional networks[J]. IEEE Signal Processing Letters, 2016, 23(10): 1499-1503.
[22] 杨旭,尚振宏.基于改进 AlexNet 的人脸表情识别[J].激光与光电子学进展,2020,57(14):141026. (Yang Xu, Shang Zhenhong. Facial expression recognition based on improved AlexNet[J]. Laser & Optoelectronics Progress, 2020, 57(14): 141026.)
[23] Li Hangyu, Wang Nannan, Ding Xinpeng, *et al.* Adaptively learning facial expression representation via CF labels and distillation[J]. IEEE Trans on Image Processing, 2021, 30: 2016-2028.
[24] Jiang Ping, Wan Bo, Wang Quan, *et al.* Fast and efficient facial expression recognition using a Gabor convolutional network[J]. IEEE Signal Processing Letters, 2020, 27: 1954-1958.
[25] 王琳琳,刘敬浩,付晓梅.融合局部特征与深度置信网络的人脸表情识别[J].激光与光电子学进展,2018,55(1):011002. (Wang Linlin, Liu Jinghao, Fu Xiaomei. Facial expression recognition based on fusion of local features and deep belief network[J]. Laser & Optoelectronics Progress, 2018, 55(1): 011002.)
[26] 李勇,林小竹,蒋梦莹.基于跨连接 LeNet-5 网络的面部表情识别[J].自动化学报,2018,44(1):176-182. (Li Yong, Lin Xiaozhu, Jiang Mengying. Facial expression recognition with cross-connect LeNet-5 network[J]. Acta Automatica Sinica, 2018, 44(1): 176-182.)
[27] Khorrami P, Paine T, Huang T. Do deep neural networks learn facial action units when doing expression recognition? [C] //Proc of IEEE International Conference on Computer Vision. Washington DC: IEEE Computer Society, 2015: 19-27.