中图法分类号:TP391　　**文献标识码**:A　　**文章编号**:1006-8961(2024)08-2377-11

论文引用格式:Lu L D, Xia H Y, Tan Y M and Song S X. 2024. Attention-guided local feature joint learning for facial expression recognition. Journal of Image and Graphics, 29(08):2377-2387(卢莉丹, 夏海英, 谭玉枚, 宋树祥. 2024. 注意力引导局部特征联合学习的人脸表情识别. 中国图象图形学报, 29(08):2377-2387)[DOI:10.11834/jig.230410]

注意力引导局部特征联合学习的人脸表情识别

卢莉丹[1,2],夏海英[1*],谭玉枚[3],宋树祥[1]

1. 广西类脑计算与智能芯片重点实验室,广西师范大学电子与信息工程学院,桂林　541004;
2. 南宁理工学院大数据与人工智能学院,南宁　530105; 3. 广西师范大学计算机科学与工程学院,桂林　541004

摘　要:**目的**　在复杂的自然场景下,人脸表情识别存在着眼镜、手部动作和发型等局部遮挡的问题,这些遮挡区域会降低模型的情感判别能力。因此,本文提出了一种注意力引导局部特征联合学习的人脸表情识别方法。**方法**　该方法由全局特征提取模块、全局特征增强模块和局部特征联合学习模块组成。全局特征提取模块用于提取中间层全局特征;全局特征增强模块用于抑制人脸识别预训练模型带来的冗余特征,并增强全局人脸图像中与情感最相关的特征图语义信息;局部特征联合学习模块利用混合注意力机制来学习不同人脸局部区域的细粒度显著特征并使用联合损失进行约束。**结果**　在2个自然场景数据集RAF-DB(real-world affective faces database)和FERPlus上进行了相关实验验证。在RAF-DB数据集中,识别准确率为89.24%,与MA-Net(global multi-scale and local attention network)相比有0.84%的性能提升;在FERPlus数据集中,识别准确率为90.04%,与FER-VT(FER framework with two attention mechanisms)的性能相当。实验结果表明该方法具有良好的鲁棒性。**结论**　本文方法通过先全局增强后局部细化的学习顺序,有效地减少了局部遮挡问题的干扰。

关键词:人脸表情识别;注意力机制;局部遮挡;局部显著特征;联合学习

Attention-guided local feature joint learning for facial expression recognition

Lu Lidan[1,2], Xia Haiying[1*], Tan Yumei[3], Song Shuxiang[1]

1. Guangxi Key Laboratory of Brain-inspired Computing and Intelliyent Chips, School of Electronic and Information Engineering, Guangxi Normal University, Guilin　541004, China; 2. College of Big Data and Artificial Intelligence, Nanning College of Technology, Nanning　530105, China; 3. School of Computer Science and Engineering, Guangxi Normal University, Guilin　541004, China

Abstract: **Objective** When communicating face to face, people use various methods to convey their inner emotions, such as conversational tone, body movements, and facial expressions. Among these methods, facial expression is the most direct means of observing human emotions. People can convey their thoughts and feelings through facial expression, and they can also use it to recognize others' attitudes and inner world. Therefore, facial expression recognition belongs to one of the research directions in the field of affective computing. It can obviously be applied to many fields, such as fatigue

收稿日期:2023-06-30;**修回日期**:2023-11-07;**预印本日期**:2023-11-14

*通信作者:夏海英　xhy22@mailbox.gxnu.edu.cn

基金项目:国家自然科学基金项目(62106054);广西创新驱动重大专项项目(AA20302003);广西师范大学(自然科学类)科研项目(2021JC012);桂林市科技开发项目(20222C243986)

Supported by: National Natural Science Foundation of China (62106054); Major Special Projects of Guangxi Science and Technology (AA20302003); The Research Projects of Guangxi Normal University (Natural Sciences) (2021JC012)

driving detection, human-computer interaction, students' listening state analysis, and intelligent medical services. However, in complex natural situations, facial expression recognition suffers from direct occlusion issues such as masks, sunglasses, gestures, hairstyles, or beards, as well as indirect occlusion issues such as different lighting, complex backgrounds, and pose variation. All these concerns can pose great challenges to facial expression recognition in natural scenes, where extracting discriminative features is difficult. Thus, the final recognition results are poor. Therefore, we propose an attention-guided joint learning method for local features in facial expression recognition to reduce the interference of occlusion and pose variation problems. **Method** Our method is composed of a global feature extraction module, a global feature enhancement module, and a joint learning module for local features. First, we use ResNet-50 as the backbone network and initialize the network parameters using the MS-Celeb-1M face recognition dataset. We think that the rich information available in the face recognition model can be used to complement the contextual information needed for facial expression recognition, especially the middle layer features such as eyes, nose, and mouth. Thus, the global feature extraction module is used to extract the global features of the middle layer, which consists of a 2D convolutional layer and three bottleneck residual convolutional blocks. Second, most of the facial expression features are concentrated in localized key regions such as eyes, nose, and mouth. Accordingly, the overall face information can be ignored and the expression categories can be directly recognized correctly with the help of local key information. Given that face recognition requires overall facial information, the face recognition pretraining model introduces some unimportant features for expression recognition. Therefore, we utilize a global feature enhancement module to suppress the redundant features (e. g. , features in the nose region) brought by the pretrained model for face recognition and enhance the semantic information of global face image that is most relevant to the emotion. This module is implemented by the effcient channel attention(ECA) attention mechanism, which strengthens the channel features that contribute to the classification and weakens the weights of the channel features that are detrimental to the classification through cross-channel interactions between high-level semantic channel features. Finally, we divide the output features of the global feature enhancement module into four non-overlapping local regions uniformly in terms of spatial dimensions. This method exactly distributes the eye and mouth regions in most of the face images in four sub-image blocks. The global facial expression analysis problem is split into multiple local regions for calculation. Then, the fine-grained salient features of different localized regions of the face are learned through the mixed-attention mechanism. The local feature joint learning module learns information from complementary contexts, which reduces the negative effects of occlusion and pose variations. Considering that our method integrates four classifiers for local feature learning, a decision-level fusion strategy is used for the final prediction. That is, after summing the output probability results of the four classifiers, the category corresponding to the maximum probability is the model prediction category. **Result** Relevant experimental validation was performed on two in-the-wild expression datasets, namely, real-world affective faces database (RAF-DB) and face expression recognition plus (FERPlus) datasets. The results of the ablation experiments show that the gains of our method compared with the base model on the two datasets are 1. 89% and 2. 47%, respectively. In the RAF-DB dataset, the recognition accuracy is 89. 24%, which has a performance improvement of 0. 84% compared with global multi-scale and local attention network (MA-Net). In the FERPlus dataset, the recognition accuracy is 90. 04%, which is comparable to the performance of FER framework with two attention mechanisms (FER-VT). Therefore, our method has good robustness. We test the model trained on the RAF-DB dataset by incorporating it with the FED-RO dataset with real occlusion and achieved an accuracy of 67. 60%. We also use Grad-CAM++ to visualize the attention heatmap of the proposed model for demonstrating the effectiveness of the proposed method more intuitively. The visualization of the joint learning module for local features illustrates that the module can direct the overall model to focus on the features in each individual local image block that are useful for classification. **Conclusion** In general, the proposed method is guided by the attention mechanism, which enhances the global features first and then learns the salient features in the local region. This approach effectively reduces the interference of the local occlusion problem through the learning sequence of global enhancement followed by local refinement. Experiments on two natural scene datasets and the occlusion test set prove that the model is simple, effective, and robust.

Key words: facial expression recognition; attention mechanism; partial occlusion; local salient feature; joint learning

0 引 言

在进行面对面沟通时，人们会用各种不同的方法来传达自己的内心情感，比如对话语气、肢体动作和面部表情等。在这些方法中，面部表情是最直接的一种观察人类情感的手段，人们可以通过面部表情来传达自己的思想和感情，也可以通过面部表情识别他人的态度和内心世界。因此人脸表情识别属于情感计算（affective computing）领域的一个研究方向，其具体任务是将离散的人类基本表情分类为7类基本情感（Ekman 和 Friesen，1971）（中立、快乐、悲伤、惊讶、害怕、厌恶以及生气）。显然这可以应用于众多领域，比如疲劳驾驶检测（Pratama 等，2017）、人机交互（Corneanu 等，2016）、学生听课状态分析（Sawyer 等，2017）和智能医疗服务（Rehman 等，2013）等。

人们在实验室受控场景数据集 CK+（the extended Cohn-Kanade dataset）（Lucey 等，2010）、OULU-CASIA（Oulu-CASIA NIR&VIS facial expression database）（Zhao 等，2011）和 MMI（MMI facial expression database）（Pantic 等，2005）等）的人脸表情识别研究中已经取得了很大的进步（Zhang 等，2017；Yang 等，2018；陈拓 等，2022），但是在现实世界中，人脸图像的背景会受到光照、遮挡和姿态变化等因素的影响，这给人脸表情识别工作带来了很大的挑战。而人脸表情是由人脸局部区域的形变体现的，所以现有的大多数人脸表情识别系统都聚焦于具有辨别信息的区域，例如眉毛、眼睛和嘴巴等。

本文针对自然场景下的局部遮挡人脸表情识别提出了一种基于局部特征联合学习的方法，主要工作如下：1）提出全局特征提取和全局特征增强模块用于从人脸识别预训练模型中学习关键表征并丢弃一些冗余信息。2）提出局部特征联合学习模块将全局面部表情分析问题拆分为多个局部区域来解决，通过混合注意力机制和多分类器决策级特征融合的方式来学习互补的上下文的信息，从而减少遮挡和姿态变化带来的不利影响。3）在自然场景数据集 RAF-DB（real-word affective faces database）、FERPlus 和具有真实遮挡的 FED-RO（facial expression dataset with real occlusions）数据集上的实验结果表明了本文方法的有效性。

1 相关工作

本节将从基于局部子区域的方法和基于注意力机制的方法两方面来探讨近年来自然场景下的人脸表情识别方法。

1.1 基于局部子区域的方法

当人类在对存在遮挡和姿态变化情况的图像进行情感类别判断时，通常只需要观察具有五官形变的局部区域就可以进行分类。因此，近年来人脸表情识别的研究热点是将输入人脸图像划分为多个局部子区域，并引导网络关注局部显著特征，以此提升表情识别的性能。Fan 等人（2018）提出了一种新的多区域集成网络，通过将不同的子区域输入到不同的子网络中，捕捉人脸面部多个子区域的全局和局部特征，从而探究不同局部子区域对面部表情识别的影响。Hazourli 等人（2020）根据人脸关键点将人脸图像裁剪为7个块，再将它们分别输入到7个子网络中学习深度特征，最后将所有子网络的输出聚集到2个全连接层中进行分类。Wang 等人（2020a）采用固定位置裁剪、随机裁剪和基于关键点裁剪3种图像块生成方法来生成不同的人脸子图像，从而减少遮挡和姿态变化带来的不利影响。Li 等人（2019）根据人脸关键点提取了24个感兴趣的区域，再使用 PG-Unit（patch-gated unit ）和 GG-Unit（global-gated unit）分别对区域图和全局人脸图计算自适应权重。然而，这些方法都需要依赖人脸关键点来划分子区域，使得整体网络变得复杂，且遮挡和姿态变化情况会导致可用的人脸关键点数量变少。

1.2 基于注意力机制的方法

注意力机制是当前的一种主流方法，它使得卷积神经网络能够更加关注输入图像中最有效和最具分辨力的特征并抑制重要程度较低的特征。因此，在自然场景的人脸表情识别中，许多研究者会将注意力机制与卷积神经网络相结合，使整体模型更聚焦于眼睛、眉毛和嘴巴等表情信息丰富的区域，忽略鼻子、额头和未正面展示出来的对最终分类作用不大的区域。Ding 等人（2020）提出的 OADN（occlusion-adaptive deep network ）通过基于人脸关键点的注意力分支对骨干网络提取到的特征进行调制，引导模型丢弃遮挡区域并更关注非遮挡区域。DAN（distract your attention network）（Wen 等，2023）

通过特征聚类网络、多头交叉注意力网络和注意力网络3部分来捕捉表情之间细微的和局部变化。MA-Net (global multi-scale and local attention network)(Zhao等,2021)中的局部注意力模块通过2个3 × 3卷积块和CBAM(convolutional block attention module)注意力机制(Woo等,2018)来引导网络学习关键区域中的显著特征。毛君宇等人(2022)通过全局注意力模块对原图像特征图和子图像特征图进行重新分配权重,从而得到有重要特征信息的图像。梁华刚等人(2022)使用标准卷积、深度可分离卷积和通道加权的组合,在减少网络参数的同时提高表情识别准确率。FER-VT(FER framework with two attention mechanisms)(Huang等,2021)在低层特征学习中使用一种基于网格的注意力机制,以从面部表情图像中捕获不同区域的相关性,在高层特征学习中使用visual Transformer注意力机制来学习全局特征。受上述注意力机制的启发,本文网络无需借助人脸关键点,也无需复杂的子网络,即可加强对重要关键区域的关注,抑制无表情信息区域的表达。

2 本文模型

本文方法由全局特征提取模块、全局特征增强模块和局部特征联合学习模块3部分组成,模型结构如图1所示。首先全局特征提取模块用来提取中级人脸特征,它由1个二维卷积层和3个瓶颈残差卷积块组成。接着将全局特征提取模块中的Conv4_x的输出作为全局特征增强模块的输入,消除从人脸识别预训练模型带来的冗余信息。最后将全局特征增强模块的输出特征图沿空间轴方向划分为4个无重叠的局部区域特征图,并利用4个并行的局部注意力网络来学习局部显著性特征,再将4个分支所提取到的局部注意力特征图分别送入全局平均池化层(global average pooling, GAP)和全连接层(fully-connected, FC)中得到最终分类结果。

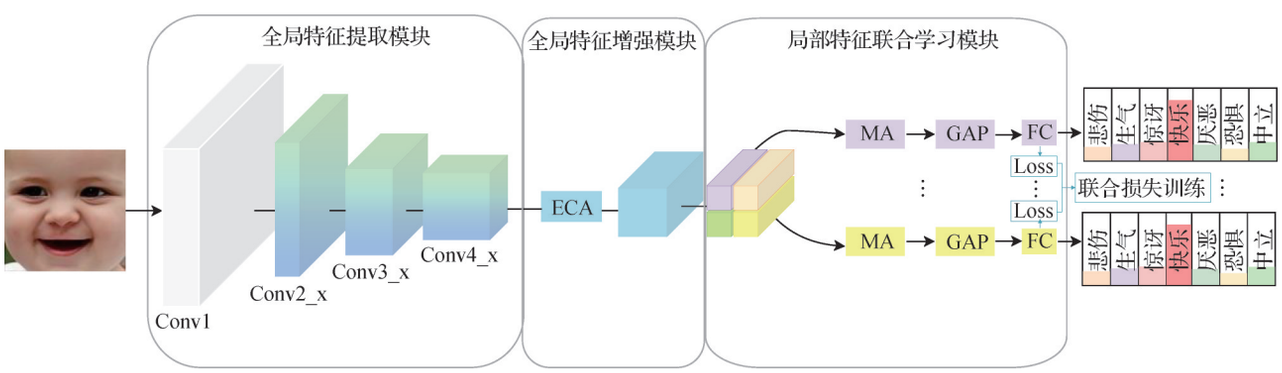

图1 整体网络结构

Fig. 1 The proposed network structure

2.1 全局特征提取模块

本文采用在人脸识别数据集MS-Celeb-1M(Guo等,2016)上预训练好的ResNet50(residual network 50)(He等,2016)作为特征提取的骨干网络,具体结构如表1所示,其中Conv1表示单个卷积层,Pool表示最大池化层,Conv2_x、Conv3_x和Conv4_x表示ResNet50中的瓶颈残差卷积块,它们分别需要堆叠3次、4次和6次。每一个卷积块都由卷积层、批量归一化层(batch normalization, BN)以及ReLU层构成。假设输入图像尺寸为224 × 224 × 3,经过整个全局特征提取模块后,输出特征图为14 × 14 × 1 024。

2.2 全局特征增强模块

人脸识别预训练模型中可用的丰富信息可以用来补充表情识别所需要的上下文信息,特别是中间层的眉毛、眼睛和嘴巴等特征。因此,为了强调中间层的这些重要特征并消除不必要的冗余特征(例如对表情分类作用较小的鼻子区域的特征),本节将全局特征提取模块中Conv4_x的输出特征送入到全局特征增强ECA(efficient channel attention)模块(Wang等,2020b)进行处理,ECA模块结构如图2所示。假设输入特征图$\boldsymbol{X} \in \mathbf{R}^{H \times W \times C}$,这里$H$,$W$,$C$表示特征图的高度、宽度和通道数,首先在空间维度上使用全局平均池化GAP对空间特征进行压缩,得到1 ×

表 1　全局特征提取模块的结构

Table 1　Structure of the global feature extraction module

层名称	层操作	输出特征图尺寸
Conv1	7 × 7,64, stride 2	112 × 112 × 64
Pool	3 × 3 max pool, stride 2	56 × 56 × 64
Conv2_x	$\begin{bmatrix}1\times1,64\\3\times3,64\\1\times1,256\end{bmatrix}\times3$	56 × 56 × 256
Conv3_x	$\begin{bmatrix}1\times1,128\\3\times3,128\\1\times1,512\end{bmatrix}\times4$	28 × 28 × 512
Conv4_x	$\begin{bmatrix}1\times1,256\\3\times3,256\\1\times1,1\,024\end{bmatrix}\times6$	14 × 14 × 1 024

$1\times C$的特征图，接着使用1×1卷积对压缩后的特征图进行通道特征学习，增加相邻通道之间的交互信息；再经过一个sigmoid激活函数输出得到新的特征通道加权值$\boldsymbol{\omega}\in\mathbf{R}^{1\times1\times C}$，$\boldsymbol{\omega}$计算为

$$\boldsymbol{\omega}=\sigma\left(Conv1\left(GAP\left(\boldsymbol{X}\right)\right)\right) \tag{1}$$

最后通道注意力模块输出的加权特征$\boldsymbol{X}'\in\mathbf{R}^{H\times W\times C}$可以表示为

$$\boldsymbol{X}'=\boldsymbol{\omega}\times\boldsymbol{X} \tag{2}$$

即用每个通道的加权值按元素与对应通道进行相乘，在不改变输入特征图大小的情况下，实现对高层语义通道特征间的跨通道交互，并对有助于分类的通道特征进行加强，削弱对分类不利的通道特征权值。

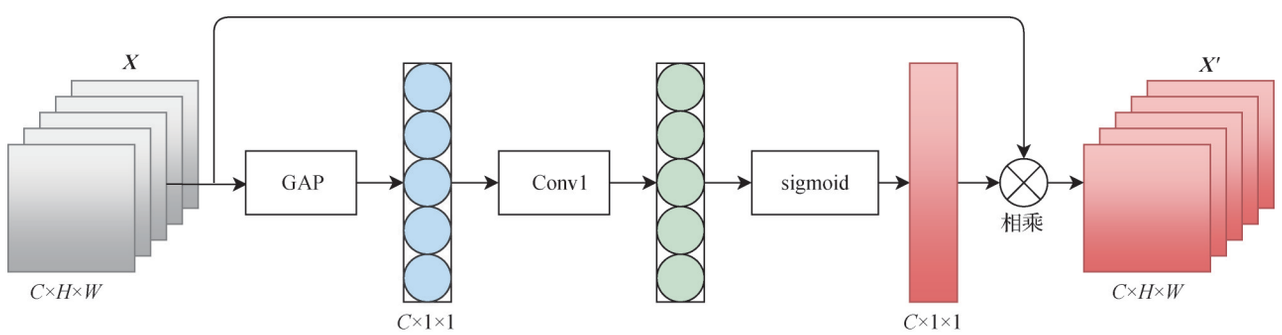

图 2　全局特征增强模块的内部结构

Fig. 2　Internal structure of the global feature enhancement module

2.3　局部特征联合学习模块

对于开放场景下的遮挡和姿态变化情况，人类在进行表情判断时会重点去关注未被遮挡的、能够传递情感信息的局部区域。而卷积神经网络为了有效地学习到局部人脸特征，在以往的工作中主要是采用根据人脸关键点裁剪或随机裁剪的方式将人脸图像分割，然而裁剪位置的不准确会导致信息丢失，且根据人脸关键点的方式需要先进行关键点定位，比较烦琐。为了让模型学到更细粒度的局部形变特征，本文直接将全局特征增强模块的输出特征映射$\boldsymbol{X}'\in\mathbf{R}^{14\times14\times1\,024}$均匀地划分为 4 个不重叠的局部特征映射$\boldsymbol{X}_i'$，其中$i\in\{1,2,3,4\}$。这种方法正好能将大部分人脸图像中的眼睛和嘴巴区域分布在 4 个子图像块中。

局部特征联合学习模块的输入是 4 个 7 × 7 × 1 024 的局部特征图，每个局部特征图经过如图 3 所示的混合注意力模块（mixed attention，MA）处理后，再送入 GAP 层和 FC 层进行分类。每个子网络进行独立的预测，训练过程中采用联合交叉熵损失函数来进行整体优化，以此学习互补的上下文信息，从而增强鲁棒性。每个局部注意力网络的损失L_n计算为

$$L_n=-\sum_{c=1}^{M}y_{ic}\log\left(p_{ic}\right) \tag{3}$$

式中，M表示类别数量，如果样本i的真实类别为c则y_{ic}取 1，否则y_{ic}取 0，p_{ic}表示待测样本i属于类别c的概率。最后整体网络的总损失为

$$L=\sum_{n=1}^{4}L_n \tag{4}$$

由于本文联合了 4 个分类器来进行局部特征学习，因此最终预测时使用决策级融合策略，即，将 4 个分支的识别结果直接相加。最终预测结果P表示为

$$P=\operatorname{argmax}\left(p_1+p_2+p_3+p_4\right) \tag{5}$$

即 4 个分类器的输出概率结果相加后，最大概率对

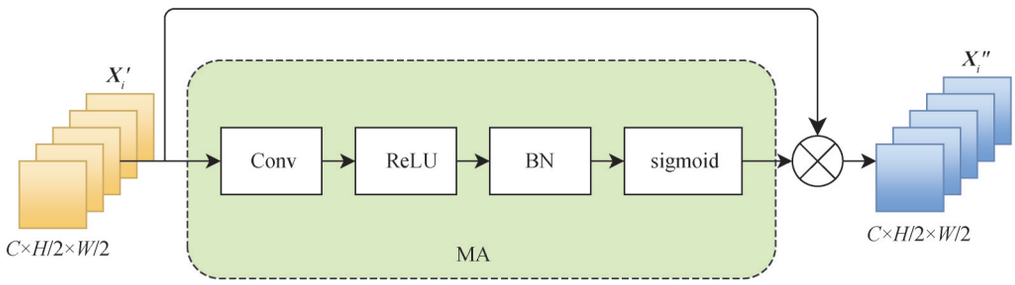

图3　局部特征联合学习模块中的混合注意力(MA)结构

Fig. 3　Mixed attention (MA) structure in the local feature joint learning module

应的类别即为模型预测类别。

3　实验结果与分析

3.1　实验数据集

为了验证提出方法的有效性,本文在3个人脸表情数据集上进行实验,分别是自然场景数据集RAF-DB、FERPlus和遮挡测试集FED-RO。

RAF-DB(Li等,2017)数据集中包含7类基本表情的图像15 339幅,其中训练集图像12 271幅,测试集图像3 068幅。

FERPlus(Barsoum等,2016)数据集是对FER2013(Goodfellow等,2013)数据集的扩展,并对其重新标签而得,由28 709幅训练图像、3 589幅验证图像和3 589幅测试图像组成。标签类别为7类基本表情加上一个“蔑视”类。

FED-RO是Li等人(2019)为了解决遮挡问题而收集的自然场景下具有真实遮挡的面部表情数据集,共400幅图像,分为7种基本表情类别,每幅图像的遮挡模式在颜色、形状、位置和遮挡率上是不同的。

3.2　实验设置与实验环境

本文设计的网络在WIN 10系统下使用PyTorch深度学习框架完成搭建,实验所用硬件配置:CPU为Intel(R) Core(TM) i7-12700K,主频为3.61 GHz,内存32 GB,GPU为NVIDIA GeForce RTX 3090,显存为24 GB。

由于不同数据集的图像尺寸各不相同,因此在模型训练之前先进行图像预处理,首先将所有图像的大小调整为224 × 224 × 3,为了避免训练数据的过拟合,数据增强时采用了随机水平翻转和随机擦除。本文采用ResNet-50作为骨干网络,并使用它在MS-Celeb-1M人脸识别数据集上的预训练模型初始化网络参数。在设置实验参数时,批处理大小为32,初始学习率为0.01,动量为0.9,权重衰减为0.000 1,优化器采用随机梯度下降(stochastic gradient descent,SGD),训练周期总数为100次。

3.3　消融实验

为了验证本文方法中全局特征增强模块和局部特征联合学习模块(即图像分块模块与混合注意力模块的组合)的有效性,在RAF-DB和FERPlus数据集上进行消融分析,所有模型均采用相同的实验设置,结果如表2所示,具体分析如下:

1)对没有添加任何模块的基础网络ResNet-50进行实验,在RAF-DB和FERPlus数据集上的识别准确率分别为87.35%和87.57%。

2)在基础网络的基础上加入图像分块模块,即将ResNet-50的Conv4_x的输出特征图按空间维度划分为4个不重叠的局部特征图,再分别将局部特征图送入4个并行的分类模块(包括GAP层和FC层)进行分类,最终在两个数据集上分别得到87.48%和88.27%的准确率,相比于只使用基础网络,性能提升了0.13%和0.7%。说明人脸识别预训练模型给网络带来了丰富的中间层可用信息,如眼睛、鼻子和嘴巴等,而分类模块能够辅助网络更关注局部显著特征。

3)考虑到面部表情特征大多集中在眼睛、鼻子和嘴巴等局部关键区域,使得人们可以忽略整体人脸信息,直接借助局部关键信息来正确地识别出表情类别。而人脸识别需要的是整体的面部信息,所以人脸识别预训练模型对于表情识别来说会引入一些不重要的特征。因此,本文在分块模块之前加入全局特征增强模块,通过ECA注意力机制的全局平均池化层对输入特征图进行空间特征压缩,抛弃一些冗余信息并强调重要特征。结果显示,在两个数据集上,使用基础网络 + 全局特征增强模块 + 图像

表 2　消融实验结果

Table 2　Results of ablation experiments

全局特征增强模块	局部特征联合学习模块		识别准确率/%	
	图像分块模块	混合注意力模块	RAF-DB	FERPlus
–	–	–	87.35	87.57
–	√	–	87.48	88.27
√	–	√	88.40	88.87
√	√	–	89.02	89.79
√	√	√	**89.24**	**90.04**

注：加粗字体表示各列最优结果，“√”和“–”分别为使用和未使用该模块。

分块模块的组合比不加入全局特征增强模块时有 1.54% 和 1.52% 的性能提升，识别准确率达到了 89.02% 和 89.79%。

4）在开放场景数据集中存在大量的遮挡和姿态变化情况，注意力机制会反映出特征图的通道响应程度随着空间位置的不同而不同。为了有效减少遮挡和姿态变化问题的干扰，在图像分块模块之后加入混合注意力模块，使得每个局部子区域网络能够自主地关注局部显著特征，并将整体网络的关注点集中在与表情最相关的动作单元区域。由表 2 的结果可知，使用基础网络 + 全局特征增强模块 + 局部特征联合学习模块的组合性能达到了最优，在 RAF-DB 和 FERPlus 数据集上的识别准确率分别为 89.24% 和 90.04%。与只使用全局特征增强模块 + 混合注意力模块相比，性能分别提升了 0.84% 和 1.17%，说明图像分块模块能够有效地引导模型关注判别性显著特征。

3.4　对比实验与分析

本节在 RAF-DB 和 FERPlus 数据集上将提出的方法与几种基于注意力机制的先进方法进行比较，并在 FED-RO 数据集上进行可视化分析。

3.4.1　RAF-DB 和 FERPlus 实验结果

在 RAF-DB 数据集上的对比如表 3 所示。RAN（region attention network）（Wang 等，2020a）是利用区域级注意力机制来减轻遮挡或姿态变化对表情识别的影响。MA-Net（Zhao 等，2021）采用全局多尺度特征分支和局部注意力特征分支相结合，并使用决策级融合策略来弥补两个分支在特征层次上具有较弱互补性的缺点。VTFF（visual Transformers with feature fusion）（Ma 等，2023）通过一个注意力选择融合器将局部二值模式（local binary patterns，LBP）特征与卷积神经网络（convolutional neural network，CNN）特征结合起来，以获得全局—局部注意力和全局自注意力。HPMI（height performance module implement）（Yao 等，2022）提出一种高性能模块（即基于空间和通道的混合注意力机制）来增强有用信息并抑制无用信息。与这些复杂的多分支、多模块融合方法相比，本文方法只需要按照注意力引导的、先全局增强后局部细化的顺序学习，即可获得 89.24% 的最优准确率，比 MA-Net、FER-VT 和 HPMI 的准确率分别提高了 0.84%、0.98% 和 0.8%。

表 3　在 RAF-DB 数据集上与其他方法的比较

Table 3　Comparison with other methods on the RAF-DB dataset

方法	准确率/%
gACNN	85.07
RAN	86.90
MA-Net	88.40
PyConv-Attention	87.34
VTFF	88.14
FER-VT	88.26
HPMI	88.44
本文	**89.24**

注：加粗字体表示最优结果。

在 FERPlus 数据集上的实验对比如表 4 所示。LSGB（Su 等，2023）使用局部关系模块（local relation，LR）、自注意力模块和全局注意力模块（self- and global-attention，SG）与批量归一化模块（batch normalization，BN）的组合，捕获样本重要区域并编码全局注意力特征。本文方法获得了与 FER-VT 相当的性能，准确率为 90.04%。与 RAN、VTFF 和 LSGB 相比获得了 1.49%、1.23% 和 1.24% 的性能提升。

图 4 为最优模型在 RAF-DB 和 FERPlus 数据集上的混淆矩阵，可以看出，该方法对“快乐”、“中立”、和“惊讶”这几种特征明显的表情识别率较高，尤其是对“快乐”的表情识别准确率达到了 95% 以上；而对于“恐惧”、“厌恶”、“悲伤”和“蔑视”的表情识别率较低，因为同属于消极表情类别，容易混淆，有时

表4　在FERPlus数据集上与其他方法的比较

Table 4　Comparison with other methods on the FERPlus dataset

方法	准确率/%
RAN	88.55
VTFF	88.81
FER-VT	**90.04**
LSGB	88.80
本文	**90.04**

注：加粗字体表示最优结果。

人类也难以区分这几类表情。另外一个原因是在大多数数据集中属于快乐类别的图像数量最多，模型可以充分学习到此类表情的特征，而厌恶和蔑视等消极情绪较难激发和获取，导致样本数量缺少，因此识别率较低。

3.4.2　在FED-RO数据集上的测试

在RAF-DB数据集上训练好的模型放到具有真实遮挡的FED-RO数据集上进行测试，实验结果如表5所示，本文方法取得了67.60%的准确率。其中gACNN和RAN使用的是RAF-DB和AffectNet（Mollahosseini等，2019）的合并训练集来进行训练，再在FED-RO数据集上进行测试。而本文方法在训练数据不够丰富的情况下取得了与它们相当的性能。说明本文方法简单且有效，在图像分块之后依然可以有效关注无遮挡区域的重要表情特征，同时抑制遮挡区域的特征在网络中的表达。

为了更直观地展示本文方法的有效性，使用梯度加权类激活映射（Grad-CAM++）（Chattopadhay等，2018）来可视化本文模型在遮挡测试集上的注意力热图，深红色表示高度关注的区域。如图5所示，局部特征联合学习模块的可视化效果说明该模块可以引导整体模型去集中关注每个单独的局部图像块中对分类有用的区域，例如，第1幅属于“惊讶”类别的图在有手部动作遮挡住嘴巴时，模型能够侧重关注睁大的眼睛区域；对于第2幅“恐惧”和第3幅“厌恶”，模型能够关注扭曲变形的眼睛和嘴巴区域；第4幅属于“快乐”类别的图在有墨镜遮挡的情况下，模型只关注嘴角上扬区域。这些结果说明了本文方法能够在复杂背景下从人脸区域中有效提取出关键特征，增强对面部遮挡的鲁棒性。

表5　在FED-RO数据集上与其他方法的比较

Table 5　Comparison with other methods on the FED-RO dataset

方法	准确率/%
VGG-16	60.15
ResNet-18	64.25
gACNN	66.50
RAN	67.89
MA-Net	**70.00**
本文	67.60

注：加粗字体表示最优结果。

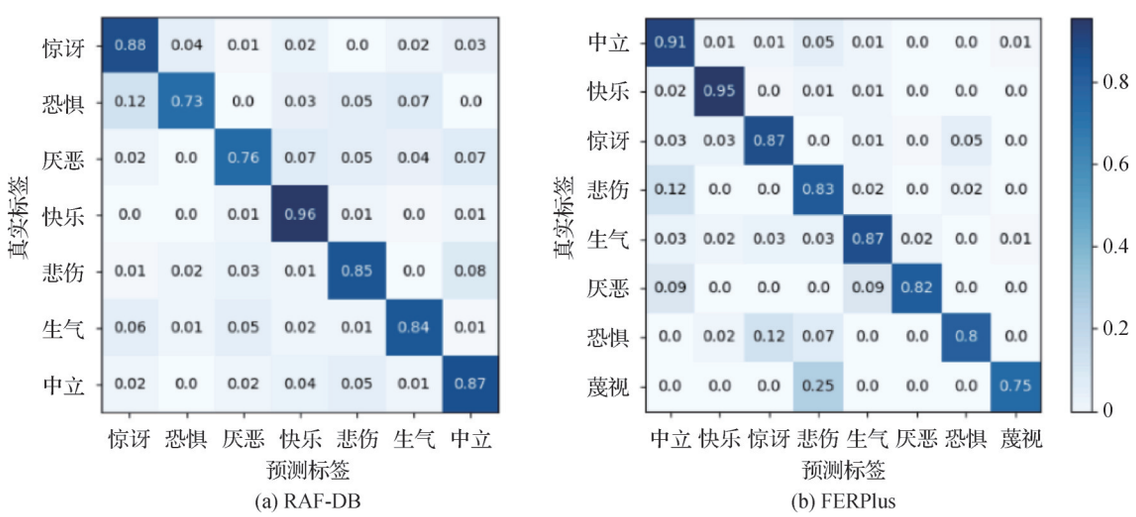

图4　最优模型在RAF-DB和FERPlus数据集上的混淆矩阵

Fig. 4　Confusion matrixs of the best model on the RAF-DB and FERPlus datasets ((a) RAF-DB; (b) FERPlus)

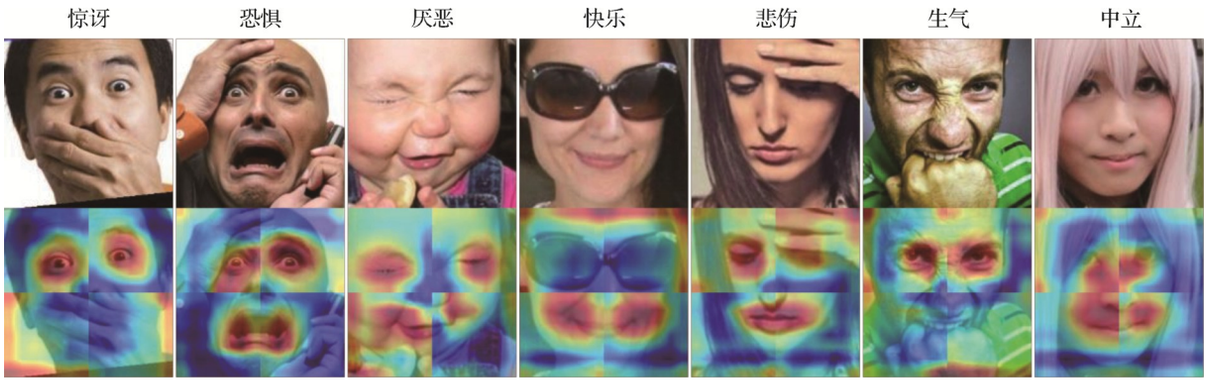

图5　本文方法在FED-RO数据集上的可视化效果展示

Fig. 5　Visualization of our method on the FED-RO dataset

4　结　论

本文提出了一种注意力引导局部特征联合学习的网络，可以有效解决自然场景下人脸表情识别中的局部遮挡问题。首先通过全局特征提取和增强模块来获取人脸识别预训练模型中对表情分类有用的全局信息，其次采用局部特征联合学习模块来获取每个局部子区域的细粒度关键特征，通过决策级特征融合实现局部特征间的互补，从而削弱局部遮挡带来的不利影响。在两个自然场景数据集和遮挡测试集上的实验证明，该模型简单有效，且具有较强的鲁棒性。

但是本文也存在一定不足，在进行子区域划分时，采用均匀划分的方法会使该模型缺乏对不同头部姿态偏转的适应性，比如，头部偏转角度较大时，可能会出现一只眼睛被分为两个部分的情况。因此，为了在遮挡和姿态变化的情况下提高表情识别准确率，如何更有效地划分局部子区域是后续工作的重点。

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]

Chattopadhay A, Sarkar A, Howlader P and Balasubramanian V N. 2018. Grad-CAM++: generalized gradient-based visual explanations for deep convolutional networks//Proceedings of 2018 IEEE winter Conference on Applications of Computer Vision (WACV). Lake Tahoe, USA: IEEE: 839-847 [DOI: 10.1109/WACV.2018.0009]

Chen T, Xing S, Yang W W and Jin J Q. 2022. Spatio-temporal features based human facial expression recognition. Journal of Image and Graphics, 27(7): 2185-2198 (陈拓, 邢帅, 杨文武, 金剑秋. 2022. 融合时空域特征的人脸表情识别. 中国图象图形学报, 27(7): 2185-2198) [DOI: 10.11834/jig.200782]

Corneanu C A, Simón M O, Cohn J F and Guerrero S E. 2016. Survey on RGB, 3D, thermal, and multimodal approaches for facial expression recognition: history, trends, and affect-related applications. IEEE Transactions on Pattern Analysis and Machine Intelligence, 38(8): 1548-1568 [DOI: 10.1109/TPAMI.2016.251560]

Ding H, Zhou P and Chellappa R. 2020. Occlusion-adaptive deep network for robust facial expression recognition//Proceedings of 2020 IEEE International Joint Conference on Biometrics (IJCB). Houston, USA: IEEE: 1-9 [DOI: 10.1109/IJCB48548.2020.9304923]

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: 10.1037/h0030377]

Fan Y R, Lam J C K and Li V O K. 2018. Multi-region ensemble convolutional neural network for facial expression recognition//Proceedings of the 27th International Conference on Artificial Neural Networks. Rhodes, Greece: Springer: 84-94 [DOI: 10.1007/978-3-030-01418-6_9]

Goodfellow I J, Erhan D, Carrier P L, Courville A, Mirza M, Hamner B, Cukierski W, Tang Y C, Thaler D, Lee D H, Zhou Y B, Ramaiah C, Feng F X, Li R F, Wang X J, Athanasakis D, Shawe-Taylor J, Milakov M, Park J, Ionescu R, Popescu M, Grozea C, Bergstra J, Xie J J, Romaszko L, Xu B, Chuang Z and Bengio Y. 2013. Challenges in representation learning: a report on three machine learning contests//Proceedings of the 20th International Conference on Neural Information Processing. Daegu, Korea (South): Springer: 117-124 [DOI: 10.1007/978-3-642-42051-1_16]

Guo Y D, Zhang L, Hu Y X, He X D and Gao J F. 2016. MS-Celeb-

1M: a dataset and benchmark for large-scale face recognition//Proceedings of the 15th European Conference on Computer Vision-ECCV 2016. Amsterdam, the Netherlands: Springer: 87-102 [DOI: 10.1007/978-3-319-46487-9_6]

Hazourli A R, Djeghri A, Salam H and Othmani A. 2020. Deep multi-facial patches aggregation network for facial expression recognition [EB/OL]. [2023-06-15]. https://arxiv.org/pdf/2002.09298.pdf

He K M, Zhang X Y, Ren S Q and Sun J. 2016. Deep residual learning for image recognition//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE: 770-778 [DOI: 10.1109/CVPR.2016.90]

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 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 (CVPR). Hawaii, USA: IEEE: 2852-2861 [DOI: 10.1109/CVPR.2017.277]

Li Y, Zeng J B, Shan S G and Chen X L. 2019. 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]

Liang H G, Bo Y, Lei Y X, Yu Z X and Liu L H. 2022. A CNN-improved and channel-weighted lightweight human facial expression recognition method. Journal of Image and Graphics, 27(12): 3491-3502 (梁华刚, 薄颖, 雷毅雄, 喻子鑫, 刘丽华. 2022. 结合改进卷积神经网络与通道加权的轻量级表情识别. 中国图象图形学报, 27(12): 3491-3502) [DOI: 10.11834/ jig.210945]

Lucey P, Cohn J F, Kanade T, Saragih J, Ambadar Z and Matthews I. 2010. The extended Cohn-Kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression//Proceedings of 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops. San Francisco, USA: IEEE: 94-101 [DOI: 10.1109/CVPRW.2010.5543262]

Ma F Y, Sun B and Li S T. 2023. Facial expression recognition with visual Transformers and attentional selective fusion. IEEE Transactions on Affective Computing, 14(2): 1236-1248 [DOI: 10.1109/TAFFC.2021.3122146]

Mao J Y, He T N, Guo Y and Li A B. 2022. Expression recognition based on global attention and pyramidal convolution network. Computer Engineering and Applications, 58(23): 214-220 (毛君宇, 何廷年, 郭艺, 李爱斌. 2022. 基于全局注意力及金字塔卷积网络的表情识别. 计算机工程与应用, 58(23): 214-220) [DOI: 10.3778/j.issn.1002-8331.2105-0422]

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]

Pantic M, Valstar M, Rademaker R and Maat L. 2005. Web-based database for facial expression analysis//Proceedings of 2005 IEEE International Conference on Multimedia and Expo. Amsterdam, the Netherlands: IEEE: #5 [DOI: 10.1109/ICME.2005.1521424]

Pratama B G, Ardiyanto I and Adji T B. 2017. A review on driver drowsiness based on image, bio-signal, and driver behavior//Proceedings of the 3rd International Conference on Science and Technology-Computer (ICST). Yogyakarta, Indonesia: IEEE: 70-75 [DOI: 10.1109/ICSTC.2017.8011855]

Rehman S, Raza S J, Stegemann A P, Zeeck K, Din R, Llewellyn A, Dio L, Trznadel M, Seo Y W, Chowriappa A J, Kesavadas T, Ahmed K and Guru K A. 2013. Simulation-based robot-assisted surgical training: a health economic evaluation. International Journal of Surgery, 11(9): 841-846 [DOI: 10.1016/j.ijsu.2013.08.006]

Sawyer R, Smith A, Rowe J, Azevedo R and Lester J. 2017. Enhancing student models in game-based learning with facial expression recognition//Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization. Bratislava, Slovakia: ACM: 192-201 [DOI: 10.1145/3079628.3079686]

Su C, Wei J G, Lin D Y and Kong L H. 2023. Using attention LSGB network for facial expression recognition. Pattern Analysis and Applications, 26(2): 543-553 [DOI: 10.1007/s10044-022-01124-w]

Wang K, Peng X J, Yang J F, Meng D B and Qiao Y. 2020a. 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]

Wang Q L, Wu B G, Zhu P F, Li P H, Zuo W M and Hu Q H. 2020b. ECA-Net: efficient channel attention for deep convolutional neural networks//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: 11534-11542 [DOI: 10.1109/CVPR42600.2020.01155]

Wen Z Y, Lin W Z, Wang T and Xu G. 2023. Distract your attention: multi-head cross attention network for facial expression recognition. Biomimetics, 8(2): #199 [DOI: 10.3390/biomimetics8020199]

Woo S, Park J, Lee J Y and Kweon I S. 2018. CBAM: convolutional block attention module//Proceedings of the 15th European Conference on Computer Vision-ECCV 2018. Munich, Germany: Springer: 3-19 [DOI: 10.1007/978-3-030-01234-2_1]

Yang H Y, Ciftci U and Yin L J. 2018. Facial expression recognition by de-expression residue learning//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 2168-2177 [DOI: 10.1109/CVPR.2018.00231]

Yao L S, He S X, Su K and Shao Q T. 2022. Facial expression recognition based on spatial and channel attention mechanisms. Wireless Personal Communications, 125(2): 1483-1500 [DOI: 10.1007/s11277-022-09616-y]

Zhang K H, Huang Y Z, Du Y and Wang L. 2017. Facial expression recognition based on deep evolutional spatial-temporal networks. IEEE

Transactions on Image Processing, 26(9): 4193-4203 [DOI: 10.1109/TIP.2017.2689999]

Zhao G Y, Huang X H, Taini M, Li S Z and Pietikäinen M. 2011. Facial expression recognition from near-infrared videos. Image and Vision Computing, 29(9): 607-619 [DOI: 10.1016/j.imavis.2011.07.002]

Zhao Z Q, Liu Q S and Wang S M. 2021. Learning deep global multi-scale and local attention features for facial expression recognition in the wild. IEEE Transactions on Image Processing, 30: 6544-6556 [DOI: 10.1109/TIP.2021.3093397]

作者简介

卢莉丹，女，硕士研究生，主要研究方向为机器视觉和模式识别。E-mail：lulidan2023@163.com

夏海英，通信作者，女，教授，主要研究方向为模式识别、医学图像处理和人脸表情识别。

E-mail：xhy22@mailbox.gxnu.edu.cn

谭玉枚，女，博士研究生，主要研究方向为情感计算和人脸表情识别。E-mail：tanyumei@stu.gxnu.edu.cn

宋树祥，男，教授，主要研究方向为智能检测、自动控制、信号与图像处理。E-mail：songshuxiang@mailbox.gxnu.edu.cn