基于边缘辅助和多尺度Transformer的无参考屏幕内容图像质量评估

陈羽中[1,2,3],陈友昆[1,2],林闽沪[1,2],牛玉贞[1,2,3*]

(1. 福州大学计算机与大数据学院,福建福州 350108;2. 福建省网络计算与智能信息处理重点实验室(福州大学),福建福州 350108;3. 大数据智能教育部工程研究中心,福建福州 350108)

摘 要: 与从现实场景中拍摄的自然图像不同,屏幕内容图像是一种合成图像,通常由计算机生成的文本、图形和动画等各种多媒体形式组合而成. 现有评估方法通常未能充分考虑图像边缘结构信息和全局上下文信息对屏幕内容图像质量感知的影响. 为解决上述问题,本文提出一种基于边缘辅助和多尺度Transformer的无参考屏幕内容图像质量评估模型. 首先,使用高斯拉普拉斯算子构造由失真屏幕内容图像高频信息组成的边缘结构图,然后通过卷积神经网络(Convolutional Neural Network,CNN)对输入的失真屏幕内容图像和相应的边缘结构图进行多尺度的特征提取与融合,以图像的边缘结构信息为模型训练提供额外的信息增益. 此外,本文进一步构建了基于Transformer的多尺度特征编码模块,从而在CNN获得的局部特征基础上更好地建模不同尺度图像和边缘特征的全局上下文信息. 实验结果表明,本文提出的方法在指标上优于其他现有的无参考和全参考屏幕内容图像质量评估方法,能够取得更高的主客观视觉感知一致性.

关键词: 无参考屏幕内容图像质量评估;高斯拉普拉斯算子;卷积神经网络;Transformer;多尺度特征

基金项目: 国家自然科学基金(No.U21A20472,No.61972097);国家重点研发计划(No.2021YFB3600503);福建省科技重大专项(No.2021HZ022007);福建省自然科学基金(No.2021J01612,No.2020J01494);福建省科技厅高校产学合作项目(No.2021H6022)

中图分类号: TN911.73;TP391 **文献标识码:** A **文章编号:** 0372-2112(2024)07-2242-15

电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20230607

No-Reference Screen Content Image Quality Assessment Based on Edge Assistance and Multi-Scale Transformer

CHEN Yu-zhong[1,2,3], CHEN You-kun[1,2], LIN Min-hu[1,2], NIU Yu-zhen[1,2,3*]

(1. *College of Computer and Data Science*, *Fuzhou University*, *Fuzhou*, *Fujian* 350108, *China*;
2. *Fujian Key Laboratory of Network Computing and Intelligent Information Processing*, *Fuzhou University*, *Fuzhou*, *Fujian* 350108, *China*;
3. *Big Data Intelligence Engineering Research Center of the Ministry of Education*, *Fuzhou*, *Fujian* 350108, *China*)

Abstract: Different from the natural images captured from real-world scenes, screen content images (SCI) are synthetic images typically composed of various multimedia contents, such as computer-generated text, graphics, and animations. Existing SCI quality assessment methods usually fail to fully consider the impacts of image edge and global context on the perceived quality of screen content images. To address the above issues, this paper proposed a no-reference screen content image quality assessment model based on edge assistance and multi-scale Transformer. Firstly, an edge structure map consisting of the high-frequency information in a distorted SCI is constructed using Gaussian Laplace operators. Then a convolutional neural network (CNN) is used to extract and fuse the multi-scale features from the input distorted SCI and the corresponding edge structure map, thus providing additional edge information gain for model training. In addition, this paper further proposed a multi-scale feature encoding module based on Transformer to better model the global context information of different scale images and edge features on the basis of the local features obtained by CNN. The experimental results show that the model proposed in this paper outperforms the state-of-the-art no-reference and full-reference SCI quality

收稿日期:2023-07-03;修回日期:2024-03-04;责任编辑:王天慧
*通讯作者:牛玉贞

assessment methods, and achieves higher consistency with the subjective visual perception.

Key words: no-reference screen content image quality assessment; laplacian of gaussian; convolutional neural network; Transformer; multi-scale features

Foundation Item(s): National Natural Science Foundation of China (No.U21A20472, No.61972097); National Key Research and Development Program of China (No.2021YFB3600503); Major Science and Technology Project of Fujian Province (No.2021HZ022007); Natural Science Foundation of Fujian Province (No.2021J01612, No.2020J01494); Industry-Academy Cooperation Project of Fujian Province (No.2021H6022)

1 引言

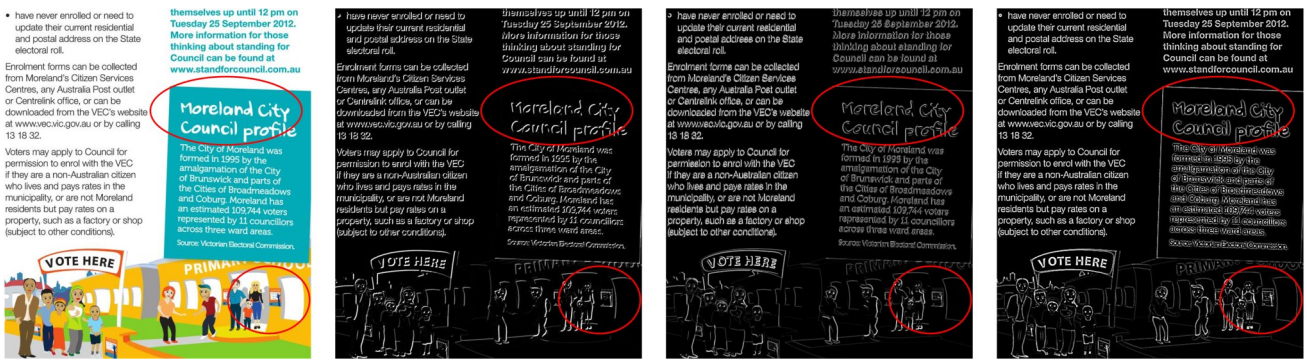
近年来,随着多媒体技术和社交网络的迅速发展,由计算机等终端设备产生的屏幕内容图像(Screen Content Image,SCI)日益频繁地出现在人们的日常生活中,在屏幕共享、在线会议、云游戏和云计算等场景中得到了普遍应用[1]. 然而,由于技术限制或硬件局限性等原因,屏幕内容图像在采集[2]、压缩[3]、处理、传输和显示[4]等过程中不可避免地会引入各种类型的失真,从而导致图像的视觉质量出现不同程度的下降,最终影响到用户视觉体验和系统交互性能. 因此,如何准确有效地评估失真屏幕内容图像的质量在电子与信息科学、数字图像处理等领域具有重要的研究意义.

根据图像质量评估(Image Quality Assessment,IQA)过程中所需原始参考图像信息量的差异,客观屏幕内容图像质量评估方法可以进一步细分为全参考(Full Reference,FR)、半参考(Reduced Reference,RR)和无参考(No Reference,NR)三种类型. 但由于现实场景中无失真的参考图像往往难以获取,因而其中的无参考方法具有更强的实用性与发展前景. 目前,针对屏幕内容图像提出的无参考质量评估方法可分为基于传统机器学习的方法和基于深度学习的方法两大类.

基于传统机器学习的方法[5~7]大多根据人类视觉系统(Human Visual System,HVS)的感知特性,从失真屏幕内容图像中手工设计提取不同的质量感知特征,如颜色、纹理、梯度、形状等,再结合有效的机器学习算法将图像特征映射为相应的质量评估分数. 这一类方法的性能高度依赖手工设计特征的合理性和代表性,在图像信息的利用率和性能提升上相对有限.

近年来,随着深度学习技术的不断发展,出现了许多基于卷积神经网络(Convolutional Neural Network,CNN)的屏幕内容图像质量评估模型[8~10],使得图像质量评估的研究重点从手工设计特征转向图像深度特征的自动学习,该类方法端到端地学习输入图像到输出质量分数之间的映射,不仅取得了较大的性能提升,还具有更好的泛化能力[11].

由于人类视觉系统对屏幕内容图像中文本和图形等边缘显著区域的畸变高度敏感,已有少量研究工作[5~7,12]利用边缘信息辅助进行失真屏幕内容图像质量分数的预测. 如图1所示,现有方法基于Sobel和Gabor滤波器构建的边缘结构图存在模糊以及边缘细节缺失等问题,限制了利用边缘信息辅助进行失真屏幕内容图像质量评估的有效性. 此外,人类视觉系统在进行主观质量感知的过程中不仅关注失真图像的局部信息,同时也会受到图像全局上下文信息的影响. 现有的基于卷积神经网络的屏幕内容图像质量评估模型大多使用卷积操作来提取失真图像的感知质量特征,但卷积操作的局部特性导致其难以对失真图像的全局上下文信息进行有效建模,因此在一定程度上会影响质量感知过程的准确性[13]. 而结合卷积神经网络和Transformer的质量评估工作[14,15]集中在自然图像质量评估领域,并且这类方法未充分考虑如何有效地结合边缘特征和图像特征.



(a) 失真屏幕内容图像　(b) Sobel滤波器　(c) Gabor滤波器　(d) LoG滤波器

图1　对比度失真下不同边缘滤波器输出的边缘结构图示例

为解决上述问题,本文提出了一种基于边缘辅助和多尺度Transformer的无参考屏幕内容图像质量评估模型.已有研究表明[16],经过高斯拉普拉斯滤波器(Laplacian of Gaussian,LoG)输出的屏幕内容图像更符合HVS中视觉皮层的感知,其构建的边缘结构图具有更加清晰和完整的边缘结构信息(如图1所示),从而能够为SCI质量评估模型的训练提供额外有效的信息增益.因此,本文首先通过高斯拉普拉斯算子构建失真屏幕内容图像对应的边缘结构图,然后将失真屏幕内容图像及其边缘结构图分别输入到基于CNN的多尺度特征提取和融合模块,以CNN不同网络层级输出的具有不同尺度大小的图像特征和边缘特征作为其多尺度特征表示,同时利用提取的边缘特征来增强失真屏幕内容图像中的高频边缘结构信息.此外,考虑到质量感知过程还受到失真图像全局上下文信息的影响,本文进一步提出了基于Transformer的多尺度特征编码模块,基于CNN网络提取的同时蕴含失真图像信息和边缘结构信息的多源特征,在多个尺度上对其进行长距离依赖关系的有效建模.实验结果表明,本文所提方法能够更充分地利用屏幕内容图像的全局上下文信息及其边缘结构信息,提高主客观质量评估的一致性.

综上所述,本文的主要贡献如下:

(1)基于HVS对图像边缘显著区域高度敏感的特性,本文提出利用高斯拉普拉斯滤波器构建更加清晰和具有完整边缘的边缘结构图,继而辅助失真屏幕内容图像进行质量评估,为其提供额外有效的信息增益.

(2)提出了基于Transformer的多尺度特征编码模块,通过自注意力机制捕获各个尺度上图像特征和边缘特征之间的长距离依赖关系,从而在CNN获得的局部特征的基础上提供有效的全局上下文信息.据我们所知,本文是首个将Transformer应用于屏幕内容图像质量评估任务的工作.

(3)在两个主流的屏幕内容图像数据集SIQAD和SCID上的实验结果表明,本文所提方法在指标上优于其他现有的无参考和全参考屏幕内容图像质量评估方法,其评估结果更符合人类的主观感知.

2 相关工作

2.1 无参考屏幕内容图像质量评估

图像质量评估任务主要通过提取失真图像的感知质量特征对其客观质量分数进行评估.而屏幕内容图像与自然图像在内容组成、结构布局等方面存在较大差异[17].如图2所示,屏幕内容图像通常包含大量的文本、线条等边缘显著区域,且其中包含的图像尺寸和颜色变化相对有限,而自然图像则一般含有相对较少的边缘、复杂的纹理细节和丰富的颜色信息[18].基于两者间的特性差异,以往针对自然图像的质量评估模型往往并不能直接适用于屏幕内容图像,因此需要开展针对屏幕内容图像的质量评估研究.

(a) 自然图像 (b) 屏幕内容图像

图2 自然图像和屏幕内容图像示例

已有研究表明,基于深度学习的图像质量评估模型通常比基于传统机器学习的方法性能更好.因此,深度学习已成为图像质量评估领域的主流方法.Yue等人[19]首先将输入图像分解为预测部分和残差部分,然后分别通过CNN进行特征提取,同时使用全参考SCI质量度量来生成训练标签,以解决数据短缺问题.为解决单一图像块在模型训练过程中所产生的误差,Jiang等人[20]提出了一种基于多区域特征学习的NR-IQA模型,通过利用多区域局部特征生成的伪全局特征进行质量评估,同时将失真类型分类以及图像质量排名作为辅助任务,以提高模型表征能力.受视觉边缘模型可以有效捕获SCI感知质量变化这一理论的启发,Yang等人[12]提出了一种结合人类视觉边缘模型和AdaBoosting反向传播神经网络的方法.考虑到多任务学习的有效性,Yang等人[21]设计了一个基于多任务学习框架的NR-IQA模型,首先分别提取SCI的失真类型和失真等级特征,然后联合这两类特征进行质量分数预测.Zhang等人[22]首先将SCI分割成文本块和图像块,然后通过双路卷积神经网络分别预测文本块和图像块的质量,最后通过质量分数聚合自适应加权策略得到整个失真图像的质量分数.

已有方法主要依赖卷积神经网络实现图像特征的自动提取,但通常缺乏对图像边缘结构信息以及全局上下文信息的有效利用,因而在性能上具有一定的局限性.本文根据屏幕内容图像具有强边缘性的特点,采用边缘检测算法提取失真图像的边缘结构特征,并通过卷积神经网络对图像特征和边缘特征进行深度融合,而后利用Transformer模型建立多个尺度上图像特征的长距离依赖关系,以获得信息更加全面的图像特征,进一步提高模型的表征能力.

2.2 基于Transformer的图像质量评估

Transformer模型[23]是一种基于自注意力机制的深

度神经网络,它起初被应用于自然语言处理任务,近年来在图像分类、目标检测、图像生成等计算机视觉任务中得到了广泛应用,并取得了优异的性能表现. 与传统的卷积神经网络不同,视觉Transformer模型可以对特征图中的所有像素进行注意力计算,从而有效捕捉图像的全局语义信息. 基于这一特性,You等人[14]设计了首个将Transformer应用于图像质量评估的模型TRIQ,该模型在CNN提取的特征图上使用浅层的Transformer编码器,并使用自适应位置嵌入以处理具有不同分辨率的图像. Ke等人[24]设计了一个多尺度Transformer模型MUSIQ以处理具有不同分辨率的图像,通过多尺度的图像表示以捕获不同粒度的图像感知质量. Cheon等人[15]将Transformer架构应用于全参考图像质量评估任务中,并在NTIRE 2021挑战赛中获得了第一名. Wang等人[25]提出了一种基于Swin Transformer[26]的多级特征融合模型,该模型聚合局部和全局特征以更好地预测图像质量,同时引入了相对排名和回归损失作为辅助任务. Zhu等人[27]提出了一种结合局部特征嵌入的显著性引导网络,通过将Transformer与显著性预测相结合,引导模型在聚合全局信息时更加关注显著区域.

与已有工作不同,为了更好地模拟人类视觉系统对屏幕内容图像的质量感知过程,本文结合卷积神经网络和Transformer有效地利用边缘特征和图像特征进行屏幕内容图像质量评估. 具体地,本文提出了基于Transformer的多尺度特征编码模块,首先在每个尺度上,该模块通过建模图像特征和边缘特征之间的长距离依赖关系,以提供图像全局上下文信息;其次在多个尺度上,能够同时表征细节特征和语义特征.

3　本文方法

为了充分利用失真屏幕内容图像中的边缘结构信息,并考虑全局上下文信息对图像质量感知的影响,本文提出了一种基于边缘辅助和多尺度Transformer的无参考屏幕内容图像质量评估模型,该模型的网络结构图如图3所示. 首先,使用高斯拉普拉斯算子构建失真屏幕内容图像对应的边缘结构图,并将其与失真图像分别输入到ResNet50网络中进行多尺度特征提取与融合,以加强网络对失真图像高频信息的感知学习. 然后,将融合边缘结构信息后的多尺度图像特征输入到基于Transformer的多尺度特征编码模块中,从而获得不同尺度图像和边缘特征的全局信息表示. 最后,对编码得到的两个蕴含不同尺度信息的特征向量进行拼接,并输入到质量分数回归模块中,同时预测失真屏幕内容图像的质量分数和失真类型. 以下小节将分别对边缘结构图的构建、多尺度特征提取与融合模块、多尺度特征编码模块以及质量分数回归模块进行详细介绍.

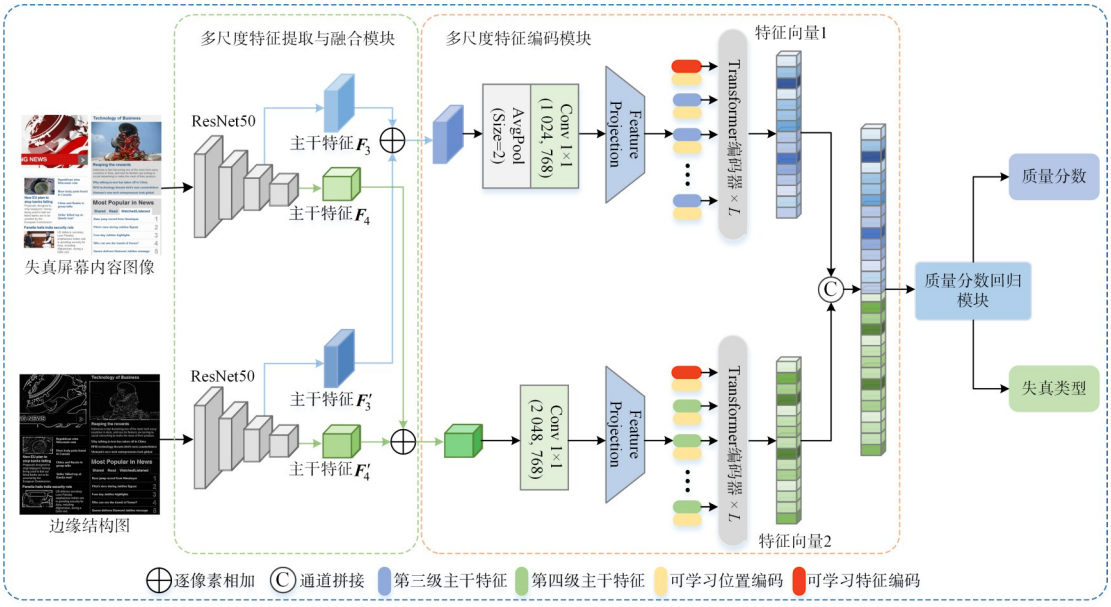


图3　基于边缘辅助和多尺度Transformer的SCI质量评估模型结构图

3.1　边缘结构图的构建

鉴于边缘显著区域对视觉质量感知的重要影响,本文利用屏幕内容图像中的边缘结构信息辅助表征图像失真. 研究表明,相比于Canny、Sobel和Gabor等其他边缘检测算子,高斯拉普拉斯算子不但适用于灰度渐变和噪声较多的图像,而且具有各向同性,得到的边缘图像具有更好的连贯性和清晰度. 因此,本文采用高斯拉普拉斯算子进行边缘结构图的构建.

具体地,首先将输入的失真图像转为灰度图,然后对其进行高斯平滑处理,以提高算子对于噪声的鲁棒

性,之后将平滑后的图像与高斯拉普拉斯算子卷积核进行卷积运算,得到输入图像对应的边缘结构图.本文中设定二维高斯拉普拉斯卷积核的窗口大小为13×13、标准差为1.图4给出了不同失真类型下的屏幕内容图像以及通过高斯拉普拉斯算子构建的边缘结构图示例.

(a) 无失真参考图像和边缘结构图

(b) 高斯噪声失真图像和边缘结构图

(c) 运动模糊失真图像和边缘结构图

(d) 对比度变化失真图像和边缘结构图

图4 不同类型失真屏幕内容图像以及对应的边缘结构图示例

从图4中可以直观地看出,通过高斯拉普拉斯算子提取的边缘结构图可以有效捕获失真屏幕图像中文本、图形、线条等区域的高频结构信息,并且能够反映不同失真类型下屏幕内容图像感知质量的不同表现.因此,高斯拉普拉斯算子构建的边缘结构图可以有效表征失真图像中高频的边缘结构信息.

3.2 多尺度特征提取与融合模块

基于卷积神经网络出色的特征提取能力,与传统手工设计的屏幕内容图像特征相比,通过深层神经网络提取的图像特征往往更加准确和全面.如图3所示,本文采用在ImageNet数据集上预训练的ResNet50模型[28]作为图像特征提取的主干网络,并将失真屏幕内容图像及其对应的边缘结构图作为网络的输入.

具体地,对于给定的失真屏幕内容图像$\boldsymbol{I}_{\text{img}}\in\mathbb{R}^{H\times W\times 3}$以及对应的边缘结构图$\boldsymbol{I}_{\text{edge}}\in\mathbb{R}^{H\times W\times 3}$,我们将其分别输入到两个ResNet50分支子网络中进行多尺度特征提取.由于本文所使用的ResNet50模型具有四个不同阶段的输出,且浅层网络提取到的图像特征通常含有丰富的细节信息,而深层网络提取到的图像特征包含更多的语义信息[29],为充分利用所提取图像特征的不同尺度信息,本文使用ResNet50第三级和第四级输出的主干特征作为失真图像以及边缘结构图的多尺度特征表示.同时记失真屏幕内容图像$\boldsymbol{I}_{\text{img}}$以及对应的边缘结构图$\boldsymbol{I}_{\text{edge}}$经过ResNet50第三级输出的主干特征为$\boldsymbol{F}_3\in\mathbb{R}^{\frac{H}{16}\times\frac{W}{16}\times 1\,024}$和$\boldsymbol{F}_3'\in\mathbb{R}^{\frac{H}{16}\times\frac{W}{16}\times 1\,024}$,经过ResNet50第四级输出的主干特征为$\boldsymbol{F}_4\in\mathbb{R}^{\frac{H}{32}\times\frac{W}{32}\times 2\,048}$和$\boldsymbol{F}_4'\in\mathbb{R}^{\frac{H}{32}\times\frac{W}{32}\times 2\,048}$.

为了增强失真屏幕内容图像的边缘信息表示,需要对来自不同尺度的失真图像特征以及对应的边缘结构特征进行融合,使得网络更加注重失真屏幕内容图像中的高频信息的特征学习.特征融合的计算公式为

$$\boldsymbol{I}_3=\boldsymbol{F}_3\oplus\boldsymbol{F}_3' \tag{1}$$

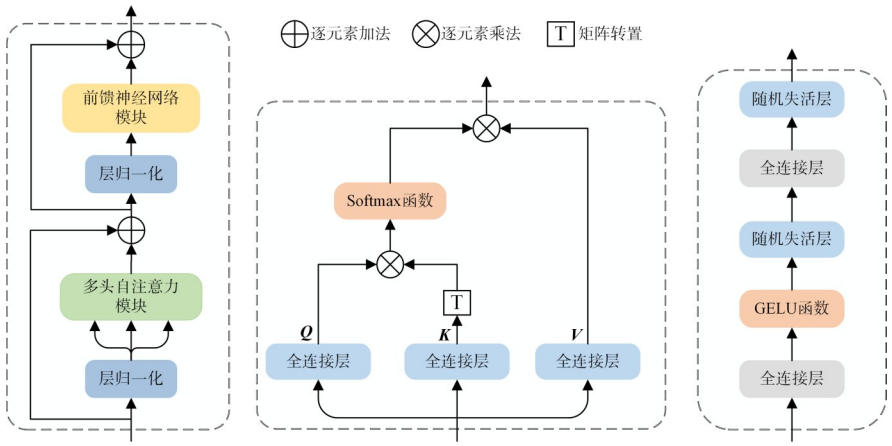
$$\boldsymbol{I}_4=\boldsymbol{F}_4\oplus\boldsymbol{F}_4' \tag{2}$$

其中,"⊕"表示逐像素相加运算,$\boldsymbol{I}_3\in\mathbb{R}^{\frac{H}{16}\times\frac{W}{16}\times 1\,024}$表示第三级失真图像特征$\boldsymbol{F}_3$和边缘结构特征$\boldsymbol{F}_3'$融合后输出的特征图,$\boldsymbol{I}_4\in\mathbb{R}^{\frac{H}{32}\times\frac{W}{32}\times 2\,048}$表示第四级失真图像特征$\boldsymbol{F}_4$以及边缘结构特征$\boldsymbol{F}_4'$融合后输出的特征图,$H$和$W$分别表示输入图像的高度和宽度.鉴于人类视觉系统对边缘显著区域的畸变高度敏感,本文通过逐像素相加的方式将同一尺度的图像特征和边缘特征进行融合,可以将边缘特征直接补偿到图像特征中,从而让模型充分学习图像边缘信息对质量评估的影响.

3.3 多尺度特征编码模块

传统的卷积神经网络模型在处理图像时主要进行空间局部特征的提取,导致其难以对图像的全局上下文特征进行有效建模. 因此,本文使用Transformer的编码器部分来执行多尺度特征间的注意力操作,建模输入特征序列之间的长距离依赖关系,从而更准确地评估图像质量. 如图3所示,首先把CNN提取的不同尺度的三维图像特征展平为二维信息序列,并附加一个可学习的特征编码充当输入序列的全局信息表示,然后将可学习的位置编码添加到特征序列中以表示图像的空间位置信息,最后将其输入到多个Transformer编码器中计算输入特征序列的全局信息表示.

如图5所示,每个Transformer编码器由多头自注意力模块、前馈神经网络模块、层归一化以及残差连接组成. 多头自注意力模块通过利用自注意力机制计算序列化后的各个局部特征之间的相关性,并利用这些相关性作为注意力权重整合所有局部特征,从而获得全局上下文信息,而前馈神经网络对每个位置的上下文信息进行非线性变换.



(a) Transformer编码器　(b) 多头自注意力模块　(c) 前馈神经网络模块

图5　Transformer编码器结构图

具体地,对于多尺度特征提取与融合模块输出的多尺度主干特征$\boldsymbol{I}_3$和$\boldsymbol{I}_4$,首先对主干特征$\boldsymbol{I}_3$进行平均池化下采样操作,使得特征图$\boldsymbol{I}_3$与$\boldsymbol{I}_4$具有相同的空间分辨率,然后将其分别输入到多个结构相同的Transformer编码器中提取全局上下文信息. 对于输入特征图$\boldsymbol{F}\in\mathbb{R}^{H'\times W'\times C}$(即主干特征$\boldsymbol{I}_3$或$\boldsymbol{I}_4$),首先通过一个$1\times 1$卷积进行线性特征投影,然后将其展平为二维特征序列$\boldsymbol{F}'=[\boldsymbol{F}'_1,\boldsymbol{F}'_2,\cdots,\boldsymbol{F}'_N]\in\mathbb{R}^{N\times D}$(其中,$N=H'\times W'$,$D$表示每个特征的维数),接着添加可学习的特征编码以及位置编码后输入到L个相互堆叠而成的Transformer编码器中进行非局部特征建模,取附加的可学习特征编码经过Transformer编码器后的输出作为最终的特征表示. 该过程可以表示为

$$\boldsymbol{F}'=[\boldsymbol{F}'_1,\boldsymbol{F}'_2,\cdots,\boldsymbol{F}'_N]=\mathrm{RELU}\left(\mathrm{Conv}\left(\boldsymbol{F}\right)\right) \tag{3}$$

$$\boldsymbol{Y}_0=[\boldsymbol{F}'_0+\boldsymbol{P}_0,\boldsymbol{F}'_1+\boldsymbol{P}_1,\cdots,\boldsymbol{F}'_N+\boldsymbol{P}_N] \tag{4}$$

$$\boldsymbol{Y}_l'=\mathrm{MHSA}\left(\mathrm{LN}\left(\boldsymbol{Y}_{l-1}\right)\right)+\boldsymbol{Y}_{l-1},\ l=1,2,\cdots,L \tag{5}$$

$$\boldsymbol{Y}_l=\mathrm{FFN}\left(\mathrm{LN}\left(\boldsymbol{Y}_l'\right)\right)+\boldsymbol{Y}_l' \tag{6}$$

其中,$\mathrm{Conv}(\cdot)$表示一个卷积核大小为1×1的卷积层,$\mathrm{RELU}(\cdot)$表示ReLU激活函数,$\boldsymbol{F}'_0\in\mathbb{R}^{1\times D}$表示额外添加的可学习特征编码,$[\boldsymbol{P}_0,\boldsymbol{P}_1,\cdots,\boldsymbol{P}_N]\in\mathbb{R}^{(N+1)\times D}$表示额外添加的定长位置编码,$\mathrm{MHSA}(\cdot)$表示多头自注意力模块,$\mathrm{FFN}(\cdot)$表示前馈神经网络模块,$\mathrm{LN}(\cdot)$表示层归一化操作,$\boldsymbol{Y}_{l-1}\in\mathbb{R}^{(N+1)\times D}$与$\boldsymbol{Y}_l\in\mathbb{R}^{(N+1)\times D}$分别表示第$l-1$个和第$l$个Transformer编码器的输出,$\boldsymbol{Y}'_l\in\mathbb{R}^{(N+1)\times D}$表示第$l$个Transformer编码器中经过第一个残差连接后得到的输出,L表示Transformer编码器的层数. 最后取可学习特征编码$\boldsymbol{F}'_0$经过Transformer编码器的输出,即$\boldsymbol{Y}_l[0]\in\mathbb{R}^{1\times D}$作为图像的全局上下文特征表示.

3.4 质量分数回归模块和损失函数

研究表明[20,21],将屏幕内容图像的失真类型预测作为辅助任务有助于提高图像质量评估模型的准确性. 因此,为充分利用失真屏幕内容图像的各类信息对模型训练进行约束,进而得到更符合人类视觉感知的屏幕内容图像质量评估模型,本文采用多任务学习策略,设计了两个不同的任务学习分支,其中一支进行质量分数预测,另一支进行失真类型预测,并且质量分数预测分支结合失真类型预测分支中学习到的失真类型特征综合评估屏幕内容图像的感知质量. 质量分数回归模块的结构如图6所示,其中失真类型预测分支主要由两个全连接层及一个Softmax层组成. 质量分数预测分支主要由三个全连接层组成,并且将第一个全连接层的输出特征与失真类型预测分支第一个全连接层的

输出特征进行拼接,从而将两个不同的任务进行关联,提高模型的特征表征能力.

因此,本文网络模型的损失函数定义如下:

$$L_{\text{score}}=\frac{1}{m}\sum_{i=1}^{m}\left(\hat{y}_i-y_i\right) \tag{7}$$

$$L_{\text{style}}=-\sum_{i=1}^{m}\sum_{j=1}^{c}p_j^{(i)}\log\hat{p}_j^{(i)} \tag{8}$$

$$L=\lambda\times L_{\text{score}}+L_{\text{style}} \tag{9}$$

其中,L表示模型的总损失函数,L_{score}和L_{style}分别表示质量分数预测损失和失真类型预测损失,m表示一个训练批次中的样本数量,c表示失真类型的种类数,y_i表示第i个图像样本的真实质量分数,$\hat{y}_i$表示第i个图像样本经过网络预测的质量分数,$p_j^{(i)}$表示真实标签中第i个图像样本对应第j类失真类型的概率值,$\hat{p}_j^{(i)}$表示第i个图像样本经过网络预测的第j类失真类型的概率值,λ表示权重.

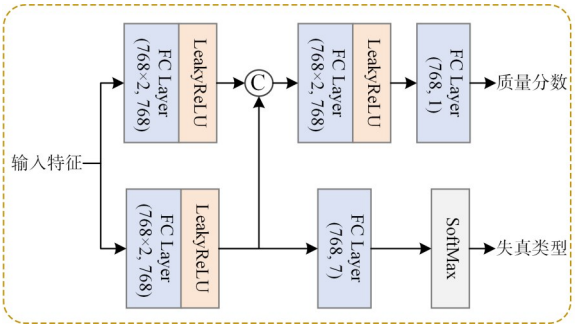


图6 质量分数回归模块结构图

4 实验

4.1 数据集和实现细节

4.1.1 数据集

为了验证所提方法的有效性,本文在两个主流的屏幕内容图像数据集SIQAD[30]和SCID[31]上进行了实验.其中SIQAD数据集由20幅参考图像以及980幅失真屏幕内容图像组成,数据集中的图像分辨率范围从626×612到832×728不等,涵盖了7种不同的失真类型,包括高斯噪声(GN)、高斯模糊(GB)、运动模糊(MB)、对比度变化(CC)、JPEG压缩(JC)、JPEG2000压缩(J2K)以及基于层划分的压缩(LSC),且每种失真类型有7种不同的失真等级.SCID数据集中包含40幅参考图像以及1 800幅失真屏幕内容图像,数据集中所有图像的分辨率均为1 280×720,涵盖了9种失真类型,包括GN、GB、MB、CC、JC、J2K、色彩饱和度变换(CSC)、高效视频编码标准(HEVC-SCC)以及带抖动的色彩量化(CQD),每种失真类型具有5种不同的失真等级.这两个数据集中的每幅失真图像都具有人工标注的主观质量分数、失真类型以及失真等级这三种标签信息.

4.1.2 评价指标

本文使用三种常见的性能指标来衡量图像质量评估算法的性能,包括皮尔森线性相关系数(PLCC)、斯皮尔曼秩序相关系数(SROCC)以及均方根误差(RMSE).PLCC、SROCC和RMSE分别用来衡量质量分数预测结果的准确性、单调性以及一致性.其中,PLCC和SROCC的值越接近于1,表明模型预测的质量分数与主观评分的相关性越高,算法的性能越好;相反,较小的RMSE值表示较低的预测误差,其值越低表明主客观评分之间的相关性越好.同时,由于不同图像质量评估模型得到的客观质量分数可能处于不同的分数区间,因此需要将模型预测的质量分数映射到一个共同的范围.本文使用视频质量专家组(Video Quality Experts Group,VQEG)在HDTV测试中建议的性能评估标准程序[32],即带五个参数的非线性逻辑回归函数对质量评估分数进行映射,公式如下所示:

$$Q(x)=\beta_1\left(\frac{1}{2}-\frac{1}{1+\mathrm{e}^{\beta_2\times(x-\beta_3)}}\right)+\beta_4 x+\beta_5 \tag{10}$$

其中,β_1、β_2、β_3、β_4和β_5表示拟合过程需要计算的五个参数,x表示模型预测的质量分数,$Q(x)$表示拟合后的质量分数.非线性逻辑回归函数中的五个拟合参数通过使用迭代最小二乘法进行计算,即通过最小化映射得到的客观质量分数$Q(x)$与主观质量分数之间的误差平方和来确定.

4.1.3 实验细节

在数据划分上,对于每个数据集随机选取80%的失真图像作为训练数据,其余20%的失真图像作为测试数据.为保证训练集和测试集之间不会出现样本重叠,数据划分均以参考图像为基准进行.对于SIQAD数据集,将784幅失真图像(对应16幅参考图像)作为训练数据,196幅失真图像(对应4幅参考图像)作为测试数据;对于SCID数据集,将1 440幅失真图像(对应32幅参考图像)作为训练数据,360幅失真图像(对应8幅参考图像)作为测试数据.每个数据集的训练和测试数据被随机划分5次,取5次重复实验的中值作为模型的整体性能.

在实验细节方面,本文方法采用Pytorch深度学习框架,并使用两张NVIDIA Tesla K80 GPU进行训练和测试.我们使用了在ImageNet数据集上预训练的ResNet50与ViT模型来对网络的部分参数进行初始化设置.根据经验,实验设置Transformer编码器的层数L为12,编码的特征维度D为768,损失函数中的权重λ设置为0.1.

考虑到现有屏幕内容图像数据集通常只含有数量较少的失真图像,所以在训练阶段本文采用图像分块

的方式来实现数据量的扩充. 但由于单一图像块并不能完全表征整个失真图像的质量,因此本文使用较大的图像块作为模型输入,以减少使用单一图像块带来的训练误差. 具体地,我们首先将训练集中的失真图像随机裁剪成448×448大小的图像块,并使用随机水平翻转进行数据增强,训练批次大小设置为6,采用Adam算法优化损失函数,将初始学习率设置为0.001,训练轮数设置为80轮,在前20轮的训练过程中冻结载入的预训练模型参数,而后以较小的学习率对网络的所有参数进行微调,从第21轮开始每经过10轮将学习率衰减为原来的0.1倍. 在测试阶段,从测试集的每张失真图像中以224像素大小的间距裁剪多个448×448大小的图像块,取各个图像块经过模型输出的质量分数的平均值作为该失真图像的质量评估分数.

4.2　方法对比

为了评估所提方法的有效性,本文选择了以下16个具有代表性的全参考或无参考质量评估算法进行比较,包括2个经典的自然图像质量评估算法SSIM[33]、GMSD[34]、2个基于Transformer的自然图像质量评估算法TRIQ[14]、MUSIQ[24]以及12个现有的屏幕内容图像质量评估算法ESIM[31]、GFM[5]、SQMS[35]、DDEGSM[36]、BES[6]、PQSC[37]、Lin[7]、Gao[10]、ABPNN[12]、RIQA[20]、Yang[38]、Zhang[22]. 其中ESIM、GFM、SQMS、DDEGSM、BES、PQSC、Lin为传统的屏幕内容图像质量评估方法,而Gao、ABPNN、RIQA、Yang、Zhang为基于深度学习的屏幕内容图像质量评估方法.

此外,为了更全面地评估各个模型对不同失真类型导致的图像质量退化的感知能力,本文在SIQAD数据集以及SCID数据集上对每种失真类型以及模型的整体性能进行了测试,实验对比结果如表1和表2所示. 对于对比的六种全参考方法以及两种基于Transformer的自然图像质量评估方法,我们首先运行作者开源的代码获得初步的客观质量分数,然后采用非线性逻辑回归函数将初步预测得到的客观质量分数映射至相同的分数区间,取拟合后的质量分数进行PLCC、SROCC以及RMSE三个指标的计算;而对比的八种无参考屏幕内容图像质量评估方法由于其代码未开源,因此实验结果数据均来源于原论文.

表1　在SIQAD数据集上的实验对比结果

指标	失真类型	全参考方法						无参考方法										
		SSIM	GMSD	ESIM	GFM	SQMS	DDEGSM	BES	TRIQ	MUSIQ	PQSC	Lin[7]	Gao[10]	ABPNN	RIQA	Yang[38]	Zhang[22]	Proposed
PLCC	GN	0.865 6	0.899 4	0.895 2	0.894 1	0.901 2	0.908 4	0.909 0	0.922 9	0.926 0	0.949 8	0.926 4	0.917 7	0.913 9	0.920	0.924 9	**0.950 7**	0.932 9
	GB	0.885 8	0.909 5	0.923 4	0.914 5	0.912 6	0.919 3	0.922 4	0.882 4	0.855 8	0.956 6	0.923 0	0.953 5	0.922 5	0.930	0.922 9	0.961 9	**0.964 9**
	MB	0.788 9	0.843 7	0.889 4	0.865 7	0.867 3	0.854 3	0.869 4	0.878 8	0.887 6	0.910 2	0.867 0	0.928 2	0.894 8	0.905	0.898 0	**0.959 2**	0.925 4
	CC	0.718 3	0.782 8	0.767 8	0.812 3	0.802 8	0.826 3	0.822 4	0.724 4	0.894 3	0.818 9	0.835 8	0.704 9	0.777 2	0.857	0.782 3	0.884 9	**0.906 4**
	JC	0.819 8	0.774 7	0.799 8	0.840 2	0.789 6	0.822 6	0.757 6	0.903 7	0.803 2	0.921 5	0.831 2	0.913 8	0.801 4	0.872	0.811 5	0.893 4	**0.927 6**
	J2K	0.789 8	0.850 9	0.794 0	0.849 0	0.826 3	0.870 7	0.818 2	0.791 0	0.749 6	0.923 8	0.831 1	0.936 8	0.798 4	0.891	0.873 4	**0.940 7**	0.889 3
	LSC	0.747 3	0.856 1	0.792 1	0.830 1	0.815 2	0.864 6	0.757 5	0.619 1	0.876 3	0.819 9	0.807 4	0.812 8	0.790 7	0.857	0.746 0	0.882 7	**0.908 8**
	ALL	0.639 6	0.739 1	0.879 1	0.883 1	0.887 2	0.900 1	0.870 5	0.895 4	0.903 6	0.916 4	0.880 3	0.900 0	0.852 9	0.910 7	0.873 8	0.926 0	**0.933 9**
SROCC	GN	0.849 7	0.885 6	0.875 6	0.879 5	0.886 1	0.894 5	0.885 1	0.912 4	0.880 7	0.940 9	0.900 1	0.869 2	0.910 2	0.901	0.918 0	**0.943 9**	0.930 5
	GB	0.885 9	0.912 2	0.923 9	0.912 2	0.914 9	0.918 5	0.904 5	0.865 4	0.777 2	0.948 5	0.902 0	0.936 5	0.922 3	0.936	0.916 3	0.939 4	**0.959 5**
	MB	0.789 1	0.843 9	0.893 8	0.863 3	0.869 5	0.851 9	0.857 1	0.859 9	0.879 0	0.918 9	0.859 9	0.918 4	0.886 7	0.893	0.893 1	**0.928 4**	0.901 5
	CC	0.490 0	0.544 3	0.610 7	0.707 8	0.694 9	0.745 7	0.658 2	0.559 9	0.770 7	**0.844 6**	0.642 0	0.583 5	0.747 1	0.714	0.778 5	0.809 2	0.811 2
	JC	0.811 0	0.771 3	0.795 9	0.843 6	0.789 4	0.818 6	0.707 2	0.861 5	0.798 0	**0.917 0**	0.817 5	0.892 2	0.776 8	0.859	0.808 1	0.849 3	0.916 8
	J2K	0.760 5	0.843 2	0.782 5	0.844 4	0.819 2	0.870 1	0.793 9	0.783 8	0.708 3	0.919 9	0.807 9	**0.956 2**	0.778 3	0.890	0.866 9	0.924 5	0.882 9
	LSC	0.758 9	0.859 6	0.796 1	0.842 6	0.829 2	0.878 9	0.729 1	0.621 2	0.877 9	0.784 2	0.777 5	0.789 8	0.758 5	0.883	0.735 5	0.840 9	**0.887 2**
	ALL	0.635 0	0.730 5	0.863 2	0.873 5	0.880 4	0.896 7	0.848 8	0.875 9	0.902 3	0.906 9	0.861 3	0.896 2	0.833 6	0.900 2	0.854 3	0.924 2	**0.930 2**
RMSE	GN	7.469 2	6.520 3	6.648 9	6.681 7	6.463 2	6.237 2	6.106 2	5.751 3	5.458 4	4.802 6	5.525 3	5.919 5	5.974 5	6.172	6.362 5	**4.737 8**	5.422 4
	GB	7.043 3	6.309 5	5.825 1	6.141 1	6.203 8	5.972 4	5.597 4	6.767 7	7.441 3	4.365 3	5.634 8	4.433 6	5.731 9	5.712	5.776 7	4.348 5	**3.640 7**
	MB	7.990 0	6.979 7	5.943 6	6.508 6	6.471 0	6.758 1	6.432 3	5.696 3	5.497 4	5.429 5	6.287 6	**4.699 7**	6.714 4	5.283	5.684 6	5.321 6	4.768 1
	CC	8.750 9	7.827 5	8.059 3	7.336 3	7.499 2	7.085 4	6.713 5	8.040 7	5.218 4	5.425 8	6.762 7	9.935 1	8.068 4	6.617	7.598 5	7.690 0	**4.886 2**
	JC	5.380 6	5.941 5	5.640 9	5.095 9	5.765 7	5.343 3	5.988 0	4.154 3	5.779 9	**3.701 4**	5.001 4	3.702 7	6.800 6	4.767	5.383 5	5.662 6	3.760 7
	J2K	6.375 3	5.459 3	6.318 8	5.492 5	5.854 4	5.112 1	5.704 8	6.386 2	6.908 8	3.497 4	5.434 8	**3.118 9**	6.553 8	4.597	4.896 7	4.901 9	4.376 4
	LSC	5.669 3	4.410 1	5.208 0	4.758 2	4.941 5	4.287 4	5.390 8	6.122 5	3.906 7	4.588 9	4.953 1	5.389 2	5.455 6	4.476	5.859 3	5.653 8	**3.793 4**
	ALL	11.002 8	9.642 2	6.822 1	6.715 6	6.603 6	6.237 7	6.914 7	6.663 7	6.412 3	5.708 0	6.700 5	6.353 5	7.281 7	5.880 3	6.933 5	5.701 7	**5.117 3**

注:加粗/下划线/波浪线分别表示性能第一/第二/第三的数据.

从表1可以看出，在SIQAD数据集上，本文所提方法在PLCC、SROCC以及RMSE三个评价指标上均实现了整体性能最优（对应表1中失真类型为ALL的结果），并且除了GN、MB、J2K等部分失真类型外，本文所提方法在单个失真类型的评估上均达到了最优或者次优性能，且取得最优性能的次数远多于其他的无参考和全参考方法．虽然在SIQAD数据集上针对GN、MB、J2K三种失真类型本文方法没有取得最优的性能，但在对比的16种方法中，针对GN失真类型，本文方法在PLCC、SROCC和RMSE上的性能表现位列第三；针对MB失真类型，本文方法在PLCC、SROCC和RMSE上的性能表现分别为第三、第四和第二；而针对J2K失真类型，本文方法在PLCC、SROCC和RMSE上的性能表现分别为第五、第五和第三．此外，如表2所示，本文方法对于SCID数据集上的GN、MB两种失真类型取得了第一、第二的性能表现．由此可见，虽然本文方法在两个数据集上的整体性能均为最优，但是在单个失真类型上的性能表现因受数据规模和数据分布的影响而略有波动．

从表2可以看出，在SCID数据集上，相较于对比的15种方法，本文方法在包含所有失真类型上的整体性能（对应表2中失真类型为ALL的结果）达到了最优的性能表现；其次，在单独的九种失真类型对应的三个评价指标上（即表2中除了失真类型为ALL的27个性能数据），本文方法在10个、3个及5个性能数据上分别取得了第一、第二和第三的性能表现．因此，本文方法在SCID数据集上的整体性能以及在单个失真类型上性能

表2 在SCID数据集上的实验对比结果

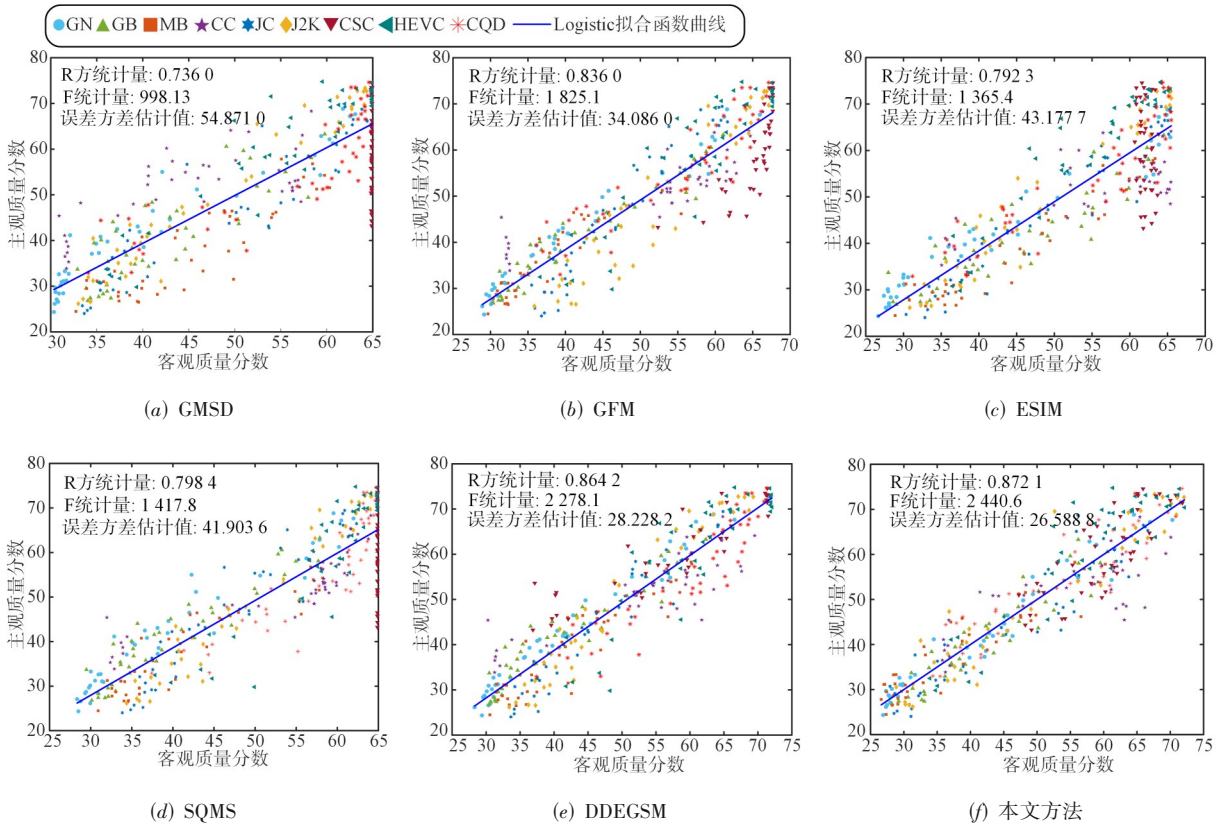
指标	失真类型	全参考方法						无参考方法									
		SSIM	GMSD	ESIM	GFM	SQMS	DDEGSM	BES	TRIQ	MUSIQ	PQSC	Lin[7]	Gao[10]	ABPNN	Yang[38]	Zhang[22]	Proposed
PLCC	GN	0.842 2	0.954 1	0.959 2	0.949 9	0.931 4	0.958 3	0.939 8	0.973 3	0.956 6	0.951 6	0.969 4	0.972 2	—	—	0.971 0	**0.978 4**
	GB	0.842 6	0.791 7	0.870 6	0.915 9	0.909 8	0.914 7	0.855 3	0.914 5	0.858 6	0.924 0	0.851 9	0.857 9	—	—	**0.959 7**	0.918 9
	MB	0.844 6	0.833 0	0.885 9	0.893 5	0.897 3	0.879 4	0.931 4	0.926 2	0.933 7	0.908 3	0.796 8	0.898 3	—	—	0.913 5	**0.940 6**
	CC	0.691 5	0.810 9	0.790 8	**0.881 2**	0.849 7	0.827 8	0.812 2	0.762 9	0.740 5	0.753 7	0.805 4	0.758 4	—	—	0.783 7	0.835 5
	JC	0.802 1	0.935 3	0.941 9	0.933 2	0.930 2	**0.964 4**	0.820 5	0.931 3	0.931 4	0.958 9	0.862 0	0.955 8	—	—	0.959 6	0.943 8
	J2K	0.819 6	0.942 6	0.945 8	0.922 6	0.946 8	**0.957 4**	0.795 4	0.948 0	0.947 4	0.951 6	0.811 5	0.941 0	—	—	0.941 0	0.952 8
	CSC	0.906 7	0.096 8	0.131 0	0.877 2	0.046 9	**0.916 4**	0.311 0	0.638 3	0.765 7	0.722 8	0.316 2	0.616 7	—	—	0.837 1	0.806 8
	HEVC	0.798 2	0.904 7	0.910 6	0.874 2	0.852 5	0.883 0	0.587 5	0.816 8	0.869 9	0.891 2	0.603 9	0.837 9	—	—	0.895 2	**0.928 6**
	CQD	0.806 2	**0.917 9**	0.900 5	0.893 1	0.899 2	0.915 1	0.830 3	0.829 4	0.817 3	0.862 7	0.871 1	0.784 2	—	—	0.916 1	0.905 3
	ALL	0.749 2	0.833 7	0.863 6	0.876 0	0.856 3	0.914 0	0.785 2	0.872 3	0.883 6	0.917 9	0.804 2	0.861 3	0.714 7	0.786 7	0.913 3	**0.926 7**
SROCC	GN	0.835 3	0.934 1	0.943 1	0.937 0	0.915 5	0.950 5	0.933 3	0.974 1	0.966 2	0.941 2	0.958 5	0.964 1	—	—	0.937 6	**0.978 6**
	GB	0.839 6	0.793 0	0.870 3	0.908 2	0.910 0	0.907 8	0.847 9	0.903 9	0.879 2	0.928 8	0.845 6	0.840 1	—	—	**0.960 8**	0.932 8
	MB	0.823 4	0.814 7	0.861 1	0.882 8	0.881 4	0.861 6	0.788 3	0.885 7	0.898 4	0.906 8	0.764 4	**0.910 7**	—	—	0.882 2	0.898 7
	CC	0.628 1	0.577 3	0.617 8	**0.822 5**	0.802 6	0.755 7	0.477 3	0.507 1	0.599 5	0.601 6	0.511 6	0.530 0	—	—	0.590 9	0.670 0
	JC	0.790 9	0.934 4	0.933 9	0.928 1	0.923 7	0.959 4	0.795 7	0.917 1	0.916 5	**0.961 0**	0.849 6	0.955 8	—	—	0.957 5	0.922 7
	J2K	0.812 3	0.928 0	0.930 6	0.908 4	0.932 0	0.942 7	0.763 4	0.924 6	0.912 6	**0.952 7**	0.773 2	0.900 6	—	—	0.858 3	0.923 5
	CSC	0.905 9	0.123 0	0.103 7	0.874 1	0.063 5	**0.908 7**	0.109 8	0.592 5	0.624 0	0.694 5	0.096 6	0.535 0	—	—	0.724 3	0.788 7
	HEVC	0.823 9	0.896 0	**0.904 2**	0.871 1	0.866 8	0.879 4	0.465 4	0.689 5	0.700 9	0.879 1	0.454 3	0.856 9	—	—	0.883 1	0.813 1
	CQD	0.804 1	0.904 8	0.886 8	0.890 5	0.891 4	0.906 9	0.754 8	0.820 3	0.803 9	0.856 0	0.787 9	0.750 5	—	—	**0.908 2**	0.876 4
	ALL	0.729 6	0.813 9	0.829 6	0.875 9	0.831 9	0.914 7	0.761 3	0.860 5	0.871 3	0.914 7	0.782 8	0.856 9	0.692 0	0.756 2	0.905 0	**0.923 5**
RMSE	GN	6.776 9	3.762 4	3.554 2	3.926 7	4.575 4	3.592 8	4.516 2	3.005 2	3.812 1	3.345 5	3.101 8	2.768 1	—	—	3.252 1	**2.702 4**
	GB	5.702 3	6.469 1	5.208 6	4.251 0	4.395 9	4.279 4	5.436 5	4.187 5	5.306 1	3.843 9	5.372 2	5.302 6	—	—	**3.268 4**	4.082 1
	MB	5.851 8	6.047 1	5.070 3	4.908 3	4.825 8	5.204 3	6.425 3	3.868 2	3.675 4	4.327 5	6.498 5	4.619 7	—	—	5.631 5	**3.428 4**
	CC	6.466 0	5.238 0	5.475 0	**4.231 8**	4.719 4	5.022 6	5.183 6	5.142 4	5.346 0	6.180 8	5.256 4	5.378 3	—	—	8.127 3	4.426 5
	JC	8.975 5	5.319 6	5.049 9	5.400 4	5.517 4	3.973 8	8.634 2	5.186 4	5.178 1	**3.922 1**	7.527 1	4.498 7	—	—	4.452 5	4.840 0
	J2K	9.115 7	5.310 5	5.166 4	6.138 4	5.119 4	4.595 1	9.254 0	4.874 4	5.317 5	**4.262 9**	9.122 5	4.803 7	—	—	5.409 7	4.955 1
	CSC	4.149 7	9.793 1	9.754 5	4.723 4	9.828 5	**3.938 1**	9.265 4	7.087 9	7.854 9	5.676 1	9.335 0	7.144 0	—	—	7.408 9	5.565 9
	HEVC	8.379 4	5.926 7	5.749 0	6.754 4	7.273 1	6.530 5	11.138 2	7.565 0	6.462 5	4.972 8	10.870 2	6.284 4	—	—	5.163 7	**4.864 1**
	CQD	7.564 2	5.073 5	5.560 8	5.752 2	5.593 4	5.155 8	7.101 5	6.829 0	7.044 1	4.863 0	6.100 9	8.732 9	—	—	**4.822 2**	5.059 4
	ALL	9.379 7	7.820 5	7.139 9	6.829 6	7.313 5	5.745 9	8.831 9	6.950 9	6.656 9	5.479 3	8.435 3	6.799 1	10.398 8	8.594 9	6.253 5	**5.298 7**

注：加粗/下划线/波浪线分别表示性能第一/第二/第三的数据，符号“—”表示对比方法的论文中无相应的数据．

表现达到第一、第二和第三的总次数均优于其他对比方法.

整体上,相比于基于Transformer的图像质量评估模型(如TRIQ、MUSIQ)以及Gao[10]、RIQA、Zhang[22]等基于卷积神经网络的无参考屏幕内容图像质量评估模型,本文方法通过联合失真图像的边缘结构图对模型进行训练,进一步加强了对图像边缘和纹理等高频特征的感知学习,并引入Transformer编码器来提取失真图像的全局上下文信息,充分考虑了不同视觉内容的感知质量差异,因而能取得更高的主客观视觉感知一致性.

如图7所示,为了更加直观地可视化模型的总体性能,我们在SCID数据集上绘制了不同方法依据失真类型划分的质量分数散点图,并使用R方统计量、F统计量以及拟合误差方差估计值三个定量指标对散点图中线性回归模型的拟合效果进行度量. 其中R方统计量是衡量线性回归模型拟合程度的常用统计量,取值范围为0到1,R方统计量越接近1,表示模型对数据的拟合越好;而F统计量通过假设检验来确定模型是否具有统计显著性,F统计量越大,一定程度上说明模型的拟合效果越好;拟合误差方差估计值是用于衡量模型预测误差的统计量,误差方差估计值越小,表示模型的预测精度越高. 从图7的定量指标计算中可以看出,本文方法在SCID数据集上具有最大的R方统计量、F统计量以及最小的拟合误差方差估计值,因此可以说明本文模型预测的客观质量分数与主观质量分数之间具有更强的相关性,相比于其他方法可以更精准地预测失真屏幕内容图像的质量分数,这与表2中的实验结果相一致.



(a) GMSD　(b) GFM　(c) ESIM

(d) SQMS　(e) DDEGSM　(f) 本文方法

图7　不同方法在SCID数据集上的散点图对比

4.3　消融实验

本节将对网络模型中各个模块的有效性进行消融研究,具体地,我们在SIQAD数据集上进行了有关实验,并使用相同的训练方法和相同的训练测试数据. 本文从以下五个方面来进行消融验证,分别为:(1)边缘结构图的有效性;(2)多尺度特征提取的有效性;(3)基于Transformer的多尺度特征编码模块的有效性;(4)多任务学习的有效性及关联性;(5)测试阶段不同裁剪间距对模型性能的影响.

4.3.1　边缘结构图的有效性

为了促进网络模型对失真屏幕内容图像中畸变效果明显的高频区域的特征学习,本文联合了失真屏幕内容图像及其对应的边缘结构图来对网络进行训练,从而为模型的质量感知学习过程提供额外的信息增益. 在表3中,我们列出了不同输入组合下所提方法的性能. 从表3中可以看出,相比于组合1仅使用原始失

表3 不同输入组合下的模型性能表现对比

组合	原始失真图像	边缘结构图	PLCC	SROCC	RMSE
组合1	√	×	0.920 0	0.918 1	5.865 2
组合2	×	√	0.901 4	0.891 3	6.480 6
组合3	√	√	**0.934 6**	**0.931 4**	**5.323 1**

真图像的情况，组合3中加入边缘结构图后，模型在PLCC、SROCC以及RMSE三个指标上都有显著提升．而组合2仅使用边缘结构图进行训练则取得了较低的性能表现，其主要原因在于边缘结构图只关注失真图像中的高频信息，对平滑图像区域的提取效果不理想，从而造成网络难以学习到完整的失真图像信息表示．虽然组合2的性能低于组合1和组合3，但是也取得了较高的PLCC和SROCC及较低的RMSE，充分证明了高频信息对于失真屏幕内容图像质量评估的重要性和有效性．

此外，为了验证本文使用高斯拉普拉斯算子构建的边缘结构图的有效性，我们使用Sobel、Canny以及Gabor等几种常见的边缘检测算子构建了不同的边缘结构图，并使用这些边缘结构图进行了相关的对比实验，实验结果如表4所示．从表4中可以看出，相比于使用其他边缘检测算子，本文通过使用高斯拉普拉斯算子构建的边缘结构图辅助网络进行训练，取得了更加优异的性能表现．并且，我们在图8中给出了针对同一失真类型（高斯噪声）下的屏幕内容图像，不同边缘检测算子构建的边缘结构图实例．从图8中可以看出，本文使用高斯拉普拉斯算子构建的边缘结构图具有更加清晰和完整的边缘结构信息，能够为网络训练提供额外有效的信息增益．

表4 不同边缘检测算子下构建的边缘结构图的消融实验结果

指标	Sobel	Canny	Gabor	Scharr	Laplacian	LoG
PLCC	0.910 8	0.885 0	0.923 2	0.895 2	0.912 9	**0.934 6**
SROCC	0.907 4	0.892 1	0.921 9	0.886 7	0.911 0	**0.931 4**
RMSE	6.180 8	6.969 0	5.752 6	6.672 6	6.109 4	**5.323 1**

(a) 无失真参考图像 (b) 高斯噪声失真图像 (c) Sobel算子 (d) Canny算子

(e) Gabor算子 (f) Scharr算子 (g) Laplacian算子 (h) LoG算子

图8 使用不同的边缘检测算子构建的边缘结构图示例

4.3.2 多尺度特征提取的有效性

经过CNN中不同深度的卷积输出的特征图通常携带不同的视觉信息，因此综合不同尺度的特征对失真图像的质量感知进行建模往往能取得更好的效果．在表5中给出了本文模型使用ResNet50单尺度特征以及多尺度特征下的性能表现．可以看到相比于仅使用单尺度特征的情况（组合1和2），使用多尺度特征（组合3）可以获得更高的指标．并且对比组合1和组合2可以看出，不同尺度的图像特征对失真图像的质量感知具有不同的贡献，更高尺度的图像特征取得了更好的效果．

表5 主干网络ResNet50中多尺度特征提取模块的消融实验结果

组合	C3	C4	PLCC	SROCC	RMSE
组合1	√	×	0.909 2	0.907 2	6.233 1
组合2	×	√	0.915 6	0.912 1	6.018 7
组合3	√	√	**0.934 6**	**0.931 4**	**5.323 1**

注：C3和C4分别表示主干网络ResNet50中输出的第三级和第四级的主干特征．

进一步地,我们对主干网络ResNet50中不同尺度主干特征的组合进行了消融实验,实验结果如表6所示. 可以看出在不同尺度特征的组合下,使用ResNet50中输出的第三级和第四级的主干特征可以取得更加优异的性能表现.

表6　ResNet50中不同尺度特征组合下的模型性能表现对比

指标	C1+C2	C1+C3	C1+C4	C2+C3	C2+C4	C3+C4
PLCC	0.900 1	0.910 9	0.918 5	0.917 7	0.910 8	**0.934 6**
SROCC	0.888 8	0.900 5	0.913 5	0.910 5	0.900 9	**0.931 4**
RMSE	6.522 4	6.174 7	5.917 8	5.946 5	6.180 2	**5.323 1**

注:C1、C2、C3和C4分别表示主干网络ResNet50中输出的第一级、第二级、第三级和第四级的主干特征.

4.3.3　基于Transformer的多尺度特征编码模块的有效性

只使用传统的卷积神经网络模型通常难以有效建模失真图像的全局上下文信息,因此本文在CNN特征提取模块的基础上进一步构建了基于Transformer的多尺度特征编码模块,以学习失真图像不同尺度特征的全局信息表示. 为了验证Transformer编码器的有效性,我们在原始网络模型的基础上去除基于Transformer的多尺度特征编码模块,直接使用沿通道方向的大小为1×1的全局平均池化层来对多尺度特征提取与融合模块输出的特征图进行降维,其实验结果如表7所示. 可以看出去除基于Transformer的多尺度特征编码模块后,模型在三个评价指标上分别下降了3.9%、4.0%和23.5%,这表明失真图像的全局上下文信息对质量感知过程有着重要的影响,进一步验证了该模块的有效性.

表7　基于Transformer的多尺度特征编码模块的消融实验结果

方法	PLCC	SROCC	RMSE
w/o	0.898 5	0.893 8	6.571 6
w	**0.934 6**	**0.931 4**	**5.323 1**

由于多尺度特征编码模块中Transformer编码器的个数也会影响模型性能,所以本文对此也进行了消融实验. 对于两个不同尺度的特征提取分支来说,我们将Transformer编码器的个数保持一致,对不同Transformer编码器个数进行消融实验,结果如图9所示. 可以看出起初随着编码器个数的不断增加,模型的整体性能不断上升,当编码器的数量为12时,模型的整体性能达到最大值. 因此综合考虑模型的参数量以及整体的性能指标后,本文将Transformer编码器的个数设置为12.

4.3.4　多任务学习的有效性及关联性

考虑到图像质量与失真类型之间存在的高度相关性,本文采用了同时预测失真图像质量分数以及失真类型的多任务学习方式. 为了验证质量分数回归模块中多任务学习策略的有效性,本文进行了去除失真类型预测分支的实验,实验结果如表8所示. 可以看出去除失真类型预测这一子任务后,模型在三个性能指标上都出现了一定幅度的下降,这说明采用多任务学习策略对于提高模型整体性能是有益的,质量分数预测任务可以通过从失真类型预测任务中获得的失真类型特征来更准确地评估失真屏幕内容图像的感知质量.

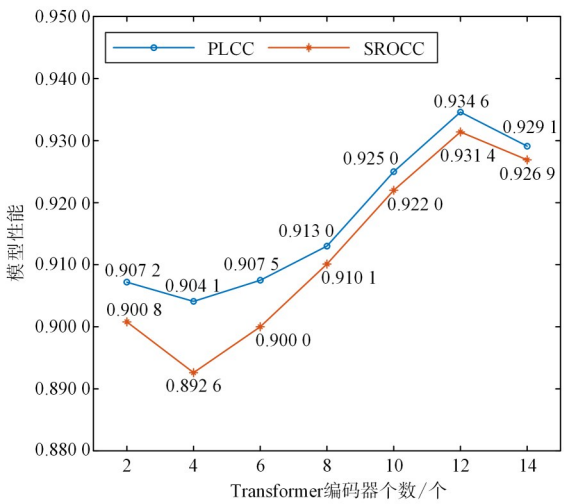


图9　不同Transformer编码器个数对模型性能的影响

表8　有无失真类型预测子任务的消融实验结果

方法	PLCC	SROCC	RMSE
w/o	0.920 5	0.914 3	5.848 2
w	**0.934 6**	**0.931 4**	**5.323 1**

此外,由于式(9)中不同大小的损失函数权重值会对不同子任务的性能表现产生影响,因此我们对不同损失函数权重值下质量分数预测任务以及失真类型预测任务的性能表现进行了消融实验,结果如表9所示. 由表9的实验结果可以看出,在一定程度上减小质量分数预测任务的损失权重值,即相对的赋予失真类型预测子任务更大的学习权重后,网络模型对于失真类型预测子任务的分类准确率取得了较大的提高. 当损失函数权重值λ为0.005的情况下,网络模型对于失真类型预测子任务的分类准确率达到了92.35%,但对于质量分数预测任务的精度却出现了一定程度的下降. 同时,从表9第一行的数据中可以看出,在只进行失真类型预测任务的情况下,模型的分类准确率达到了95.41%. 因此,综合考虑本文的主要任务是对失真屏幕内容图像进行质量分数评估,所以本文更加倾向于将损失权重值设置为0.1,以提高质量分数的预测精度.

4.3.5　测试阶段不同裁剪间距对模型性能的影响

在测试阶段,本文以224像素大小的间距裁剪多个448×448大小的图像块,取各个图像块经过模型输出

表9 SIQAD数据集上不同损失函数权重值下不同子任务取得的性能表现

损失函数权重值	PLCC	SROCC	RMSE	Accuracy
只进行失真类型预测	—	—	—	**95.41%**
λ=0.005	0.906 1	0.905 0	6.332 8	92.35%
λ=0.01	0.916 2	0.908 3	5.998 8	87.24%
λ=0.05	0.914 1	0.907 1	6.069 8	77.55%
λ=0.1	**0.934 6**	**0.931 4**	**5.323 1**	62.24%
λ=0.5	0.930 9	0.929 5	5.468 0	42.86%

注:Accuracy表示失真类型预测任务的分类准确率.

的质量分数的平均值作为该失真图像的质量评估分数,通过综合评估失真图像不同区域的视觉质量差异,以提高质量评估模型的准确性. 为此,我们补充了不同裁剪间距大小对模型性能影响的消融实验,其结果如表10所示,可以看到在裁剪步距为112、168和224的情况下,模型的性能表现差距不大,但当裁剪步距为224时模型的计算效率更高,而当裁剪步距为336和448时,模型的性能略有下降.

表10 测试阶段不同裁剪间距大小对模型性能的实验结果

裁剪间距	PLCC	SROCC	RMSE
Patch_Size=448, Stride=112	0.933 1	0.931 2	5.381 4
Patch_Size=448, Stride=168	0.933 4	0.930 7	5.331 6
Patch_Size=448, Stride=224	**0.934 6**	**0.931 4**	**5.323 1**
Patch_Size=448, Stride=336	0.932 8	0.930 3	5.353 8
Patch_Size=448, Stride=448	0.931 1	0.929 2	5.461 4

5 总结

为解决现有无参考屏幕内容图像质量评估方法未充分考虑图像边缘信息和全局上下文信息对SCI质量感知的影响,本文提出了一种基于边缘辅助和多尺度Transformer的无参考屏幕内容图像质量评估模型. 首先使用高斯拉普拉斯算子构建屏幕内容图像的边缘结构图,并采用CNN网络对失真图像和相应的边缘结构图进行多尺度的特征提取和融合,从而显著增强模型对边缘信息的表征能力. 此外,本文进一步提出了基于Transformer的多尺度特征编码模块,以更好地建模失真图像不同尺度特征的全局上下文信息. 实验结果表明,本文所提模型能够取得更高的主客观视觉感知一致性,并且在指标上优于其他现有的无参考和全参考屏幕内容图像质量评估方法. 未来工作将进一步探索失真屏幕内容图像的边缘特征与原始图像特征的融合策略,以进一步提高对边缘特征的有效利用.

参考文献

[1] PENG W H, WALLS F G, COHEN R A, et al. Overview of screen content video coding: Technologies, standards, and beyond[J]. IEEE Journal on Emerging and Selected Topics in Circuits and Systems, 2016, 6(4): 393-408.

[2] 贾旭, 曹玉东, 孙福明, 等. 基于无参考质量评价模型的静脉图像采集方法[J]. 电子学报, 2015, 43(2): 236-241.
JIA X, CAO Y D, SUN F M, et al. Vein image acquisition method based on quality assessment model without reference[J]. Acta Electronica Sinica, 2015, 43(2): 236-241. (in Chinese)

[3] 陈文俊, 杨春玲. 图像压缩感知的特征域优化及自注意力增强神经网络重构算法[J]. 电子学报, 2022, 50(11): 2629-2637.
CHEN W J, YANG C L. Feature-space optimization-inspired and self-attention enhanced neural network reconstruction algorithm for image compressive sensing[J]. Acta Electronica Sinica, 2022, 50(11): 2629-2637. (in Chinese)

[4] 李群迎, 张晓林. 基于多描述和不等差错保护的航空遥感图像传输方法[J]. 电子学报, 2010, 38(11): 2655-2659.
LI Q Y, ZHANG X L. Aerial remote sensing image transmission using multiple description coding and unequal error protection[J]. Acta Electronica Sinica, 2010, 38(11): 2655-2659. (in Chinese)

[5] NI Z K, ZENG H Q, MA L, et al. A Gabor feature-based quality assessment model for the screen content images[J]. IEEE Transactions on Image Processing, 2018, 27(9): 4516-4528.

[6] 魏乐松, 陈俊豪, 牛玉贞. 基于边缘和结构的无参考屏幕内容图像质量评估[J]. 北京航空航天大学学报, 2019, 45(12): 2449-2455.
WEI L S, CHEN J H, NIU Y Z. Blind quality assessment for screen content images based on edge and structure[J]. Journal of Beijing University of Aeronautics and Astronautics, 2019, 45(12): 2449-2455. (in Chinese)

[7] 林冠妙, 魏乐松, 牛玉贞. 基于多尺度特征的无参考屏幕内容图像质量评估[J]. 小型微型计算机系统, 2022, 43(2): 372-380.
LIN G M, WEI L S, NIU Y Z. No reference image quality assessment for screen content image based on multi-scale features[J]. Journal of Chinese Computer Systems, 2022,

43(2): 372-380. (in Chinese)

[8] ZHANG Y, CHANDLER D M, MOU X Q. Quality assessment of screen content images via convolutional-neural-network-based synthetic/natural segmentation[J]. IEEE Transactions on Image Processing, 2018, 27(10): 5113-5128.

[9] JIANG X H, SHEN L Q, DING Q, et al. Screen content image quality assessment based on convolutional neural networks[J]. Journal of Visual Communication and Image Representation, 2020, 67: 102745.

[10] GAO R, HUANG Z Q, LIU S G. Multi-task deep learning for no-reference screen content image quality assessment[C]//International Conference on Multimedia Modeling. Cham: Springer, 2021: 213-226.

[11] 徐少平, 李芬, 陈孝国, 等. 一种利用改进深度图像先验构建的图像降噪模型[J]. 电子学报, 2022, 50(7): 1573-1578.
XU S P, LI F, CHEN X G, et al. An image denoising model using the improved deep image prior[J]. Acta Electronica Sinica, 2022, 50(7): 1573-1578. (in Chinese)

[12] YANG J C, BIAN Z L, LIU J C, et al. No-reference quality assessment for screen content images using visual edge model and Adaboosting neural network[J]. IEEE Transactions on Image Processing, 2021, 30: 6801-6814.

[13] ZENG C, KWONG S. Learning transformer features for image quality assessment[EB/OL]. (2021-12-01) [2023-07-01]. http://arxiv.org/abs/2112.00485.

[14] YOU J Y, KORHONEN J. Transformer for image quality assessment[C]//2021 IEEE International Conference on Image Processing (ICIP). Piscataway: IEEE, 2021: 1389-1393.

[15] CHEON M, YOON S J, KANG B, et al. Perceptual image quality assessment with transformers[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Piscataway: IEEE, 2021: 433-442.

[16] GUO H, MA K K, ZENG H Q. A Log-Gabor feature-based quality assessment model for screen content images[C]//2019 IEEE International Conference on Image Processing (ICIP). Piscataway: IEEE, 2019: 4499-4503.

[17] KUANG W, CHAN Y L, TSANG S H, et al. Online-learning-based Bayesian decision rule for fast intra mode and CU partitioning algorithm in HEVC screen content coding[J]. IEEE Transactions on Image Processing, 2020, 29: 170-185.

[18] MIN X K, GU K, ZHAI G T, et al. Screen content quality assessment: Overview, benchmark, and beyond[J]. ACM Computing Surveys, 2021, 54(9): 1-36.

[19] YUE G H, HOU C P, YAN W Q, et al. Blind quality assessment for screen content images via convolutional neural network[J]. Digital Signal Processing, 2019, 91: 21-30.

[20] JIANG X H, SHEN L Q, YU L W, et al. No-reference screen content image quality assessment based on multi-region features[J]. Neurocomputing, 2020, 386: 30-41.

[21] YANG J C, BIAN Z L, ZHAO Y, et al. Staged-learning: Assessing the quality of screen content images from distortion information[J]. IEEE Signal Processing Letters, 2021, 28: 1480-1484.

[22] ZHANG C F, HUANG Z Q, LIU S G, et al. Dual-channel multi-task CNN for no-reference screen content image quality assessment[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(8): 5011-5025.

[23] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: ACM, 2017: 6000-6010.

[24] KE J J, WANG Q F, WANG Y L, et al. MUSIQ: Multi-scale image quality transformer[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2021: 5128-5137.

[25] WANG J, FAN H T, HOU X X, et al. MSTRIQ: No reference image quality assessment based on swin transformer with multi-stage fusion[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Piscataway: IEEE, 2022: 1268-1277.

[26] LIU Z, LIN Y T, CAO Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2021: 9992-10002.

[27] ZHU M M, HOU G Q, CHEN X J, et al. Saliency-guided transformer network combined with local embedding for no-reference image quality assessment[C]//2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). Piscataway: IEEE, 2021: 1953-1962.

[28] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2016: 770-778.

[29] ZEILER M D, FERGUS R. Visualizing and understanding convolutional networks[C]//European Conference on Computer Vision. Cham: Springer, 2014: 818-833.

[30] YANG H, FANG Y M, LIN W S. Perceptual quality as-

sessment of screen content images[J]. IEEE Transactions on Image Processing, 2015, 24(11): 4408-4421.
[31] NI Z K, MA L, ZENG H Q, et al. ESIM: Edge similarity for screen content image quality assessment[J]. IEEE Transactions on Image Processing, 2017, 26(10): 4818-4831.
[32] VQEG. Final Report From the Video Quality Experts Group on the Validation of Objective Models of Video[R/OL]. (1997-10-14) [2023-06-27]. https://vqeg. org/vqeg-home.aspx.
[33] WANG Z, BOVIK A C, SHEIKH H R, et al. Image quality assessment: From error visibility to structural similarity[J]. IEEE Transactions on Image Processing, 2004, 13(4): 600-612.
[34] XUE W F, ZHANG L, MOU X Q, et al. Gradient magnitude similarity deviation: A highly efficient perceptual image quality index[J]. IEEE Transactions on Image Processing, 2014, 23(2): 684-695.
[35] GU K, WANG S Q, YANG H, et al. Saliency-guided quality assessment of screen content images[J]. IEEE Transactions on Multimedia, 2016, 18(6): 1098-1110.
[36] TOLIE H F, FARAJI M R. Screen content image quality assessment using distortion-based directional edge and gradient similarity maps[J]. Signal Processing, 2022, 101: 116562.
[37] FANG Y M, DU R G, ZUO Y F, et al. Perceptual quality assessment for screen content images by spatial continuity[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 30(11): 4050-4063.
[38] YANG J C, ZHAO Y, LIU J C, et al. No reference quality assessment for screen content images using stacked autoencoders in pictorial and textual regions[J]. IEEE Transactions on Cybernetics, 2022, 52(5): 2798-2810.

作者简介

陈羽中 男,1979 年 4 月出生于福建省福州市,博士. 现为福州大学计算机与大数据学院教授、博士生导师. 主要研究方向为计算机视觉和机器学习.
E-mail: yzchen@fzu.edu.cn

陈友昆 男,1999 年 8 月出生于福建省泉州市. 现为福州大学计算机与大数据学院硕士研究生. 主要研究方向为图像质量评价、计算机视觉和机器学习.
E-mail: 1490708728@qq.com

林闽沪 男,1999 年 6 月出生于福建省福州市. 现为福州大学计算机与大数据学院硕士研究生. 主要研究方向为计算机视觉和图像恢复.
E-mail: 211020026@fzu.edu.cn

牛玉贞 女,1982 年 7 月出生于山东省济南市,博士. 现为福州大学计算机与大数据学院教授、博士生导师. 主要研究方向为计算机视觉和机器学习.
E-mail: yuzhenniu@gmail.com