文章编号:1674-2974(2024)12-0087-11 DOI:10.16339/j.cnki.hdxbzkb.2024286

EMMA注意力与多尺度融合下的图像修复

魏赟,王璐璐,邬开俊†,单宏全,田彬
(兰州交通大学 电子与信息工程学院,甘肃 兰州 730070)

摘 要:针对图像缺失区域与其周围的纹理、结构密切相关而无法准确推断缺失区域内容的问题,提出一种单阶段图像修复模型.通过卷积层和FastStage模块对特征进行压缩、重建和增强,结合自注意力和多层感知机来捕捉特征之间的上下文关系.在模型中引入EMMA机制.以增强生成器对特征的注意力和重要性感知,避免模型参数的更新出现抖动和振荡现象,从而提高生成器的性能和生成结果的质量.通过判别器对修复后的图像与原始图像的一致性进行评估.针对CelebA、Places2以及Paris StreetView数据集进行的端到端实验结果表明,相较于现有的经典方法,该模型的修复结果更符合视觉语义,能够精细地修复图像的细节纹理和局部特征.

关键词:图像修复;注意力机制;膨胀卷积;深度学习
中图分类号:TP751.1 **文献标志码**:A

Image Inpainting with EMMA Attention and Multi-scale Fusion

WEI Yun,WANG Lulu,WU Kaijun†,SHAN Hongquan,TIAN Bin
(School of Electronic and Information Engineering, Lanzhou Jiaotong University, Lanzhou 730070, China)

Abstract: To address the problem of accurately inferring the content of missing regions in an image when they are closely related to the surrounding textures and structures, we propose a single-stage image inpainting model. The model first compresses, reconstructs, and enhances features through convolutional layers and the FastStage module, while self-attention and multi-layer perceptron are incorporated to capture contextual relationships among features. Furthermore, in order to enhance the attention and importance perception on features, we propose EMMA in the models, which avoids the shaking and oscillation during updating the model parameters, thereby improving the performance of the generator and the quality of the generated results. Lastly, we introduce a discriminator to evaluate the consistency between the inpainted image and the original image. The end-to-end experimental results conducted on CelebA, Places2, and Paris StreetView datasets demonstrate that, compared with classical methods, the inpainting results of this model exhibit better visual semantics, and it is capable of finely inpainting details, textures, and local features of images.

Key words: image inpainting; attention mechanism; dilated convolution; deep learning

* **收稿日期**:2023-11-01
基金项目:甘肃省自然科学基金项目(23JRRA913), Natural Science Foundation of Gansu Province(23JRRA913);全国高等院校计算机基础教育研究会项目(2023-AFCEC-039), Association of Fundamental Computing Education in Chinese Universities(2023-AFCEC-039)
作者简介:魏赟(1970—),女,河南鹤壁人,兰州交通大学副教授,博士
† 通信联系人,E-mail: wkj@mail.lzjtu.cn

图像修复旨在对受损图像进行最大限度的还原,使填充像素与未缺失像素融合在一起,修复结果接近于原图[1].随着技术的发展,图像修复的应用领域也越来越广泛.如恢复老照片、去除水印、医学影像等.

传统的图像修复方法分为两类:基于补丁的方法和基于扩散的方法,基于补丁的方法[2-4]是选择图像中合适的补丁或邻域来修复受损区域,通常根据周围的像素信息进行补丁的选择和合成.基于扩散的方法[5-8]通过在图像中扩散或传播受损像素来修复受损区域,常见的基于扩散的方法有扩散滤波、偏微分方程等.然而,传统的图像修复方法缺乏高级语义特征,无法修复高度模式化的图像.

近年来,深度学习方法在图像修复领域得到了广泛的应用,它能够从大规模数据中学习图像的特征表示,并通过生成器网络生成高质量的修复结果.生成器网络通过学习修复缺失或损坏的图像,并能够处理各种复杂情况.基于注意力机制的修复算法将传统像素匹配概念引导到图像特征空间中,利用注意力机制在缺失区域中搜索最相似的特征块进行特征匹配.与传统方法相比,这种方法能够更好地捕捉图像特征的相关性,并实现更准确的特征匹配.You 等[5]引入了上下文注意力机制和编码器-解码器结构的神经网络模型,以生成逼真且上下文一致的修复图像.该方法利用图像中其他已知区域的上下文信息,有效地生成高质量的修复图像.但其忽略了区域内图像特征的相关性,导致修复结果出现不连续或不自然的现象.为了解决这个问题,Liu 等[6]提出一个新的连贯语义注意层,通过建立缺失部分特征之间的语义关联模型,有效解决了修复图像的色彩缺陷和边界畸变问题.Zeng 等[7]引入了一种可学习的金字塔上下文网络和多尺度解码网络模型.在图像和特征层面进行修复,旨在保证修复结果在纹理和语义上的一致性.Li 等[8]提出了一种具有区域特征的增量图像修复网络,以保证图像修复过程中不同方向的一致性.当图像中的缺失区域与其周围的纹理和结构密切相关时,可能无法准确推断缺失区域的内容,由于需要对图像进行逐像素的修复,并考虑到图像的局部和全局上下文,这些方法可能需要更长的时间来生成高质量的修复结果.

针对以上问题,本文提出一种 EMMA 注意力与多尺度融合下的图像修复方法,该方法是一个单阶段修复网络,它主要由三部分构成:生成网络、局部判别器和全局判别器.生成网络是编码器-解码器的架构,用于完成图像修复任务.其通过学习图像的上下文信息和纹理特征,尝试填充和恢复缺失的像素.通过引入全局和局部判别器,以提高修复网络的性能和稳定性.

在人脸、街道和建筑数据集上进行的实验表明,本文提出的方法能更好地修复受损图片.其创新性主要体现在:

1)提出一个 EMMA 注意力模块,通过综合利用平滑的特征表示和注意力感知的特征表示,增强网络对输入特征的表达能力,使模型在修复过程中更加关注缺失区域周围的重要结构和纹理细节.

2)在编码和解码阶段引入 FastStage 模块,该模块提供高效的特征提取和表示学习能力,有助于网络更好地理解和恢复图像的缺失区域,从而在与周围联系紧密的缺失区域中较为合理地填充内容.

3)通过改进全局-局部判别器,利用 VGG 网络提取特征并计算损失,对细节网络的输出进行判别优化,从而增强修复结果的整体和局部一致性.这种修复方法旨在改善修复质量,使修复结果更加真实.

1 相关工作

1.1 图像修复

Pathak 等[9]提出了 Context Encoder 模型,通过对抗性损失训练来修复图像中的缺失内容.Iizuka 等[10]提出了局部判别器和泊松融合技术,用于保证修复结果在局部区域的一致性,并通过细化处理获取更加透明的修复效果.然而,这两种方法都是基于一般卷积的深度生成模型,无法有效修复不规则损伤图像.为了解决这个问题,Liu 等[6]提出了利用部分卷积来修复不规则区域的掩膜,但这种方法可能导致修复结果的语义模糊性.因此,一些学者提出了渐进式修复方法,先修复边缘再逐步修复中心区域,以更好地保留图像的轮廓信息,例如 Yi 等[11]提出利用上下文残差聚合模块和上下文注意力机制从多层次上下文中获取特征信息,以填充缺失的内容,保证高分辨率的修复结果.此外,还有一些研究人员提出了多尺度结构用于图像修复,例如 Wang 等[12]提出了生成式多列卷积神经网络,利用不同大小的卷积核(如 3×3、5×5、7×7)来获得各种感受野,从而提高了修复性能.另外,Li 等[8]提出了递归特征推理(RFR)模块,利用循环特征推理机制填充图像中缺失部

分,以实现更加自然连贯的图像修复效果.

1.2　生成对抗网络

生成对抗网络(GAN)是一种参数模型,主要用于生成高质量的图像.它由生成器(G)和判别器(D)两部分组成.生成器基于随机向量(通常服从某种先验分布)生成图像,判别器评估输入的图像属于真实图像的概率.生成器的目标是生成逼真的图像,而判别器则在训练过程中起"监督"作用,通过评估图像的真实性来"鼓励"生成器学习真实图像的分布.然而,原始GAN模型难以训练,且容易遭受崩溃模式的影响.因此,研究者提出了一些方法来改进GAN的性能.部分方法通过修改损失函数以获得更好的性能,如Wasserstein生成对抗网络(WGANs)[13]和最小二乘生成对抗网络(LSGANs)[14].还有一些方法提出了基于原始GAN的新框架,如深度卷积生成对抗网络(DCGAN)[15].相较于原始GAN,DCGAN在内部结构上进行了调整,如在生成器和判别器中使用批标准化(batch normalization)并移除了全连接的隐藏层,为GAN的训练提供了实用的技术,并确保了生成照片级别真实图像的能力.

然而,鉴于生成对抗网络可能产生完全不相关的图像,直接将其应用于图像修复任务是不可行的.图像修复任务要求更精确和准确地恢复丢失或损坏的图像内容.因此,针对图像修复,需要在生成对抗网络的基础上进行进一步的改进和定制,以满足图像修复任务的特定需求.

2　本文算法

2.1　整体模型

本文模型如图1所示,该模型为一个单阶段修复模型,由生成器、局部判别器和全局判别器组成.其中,生成器用于产生合理的修复结果,判别器用于对抗性训练.

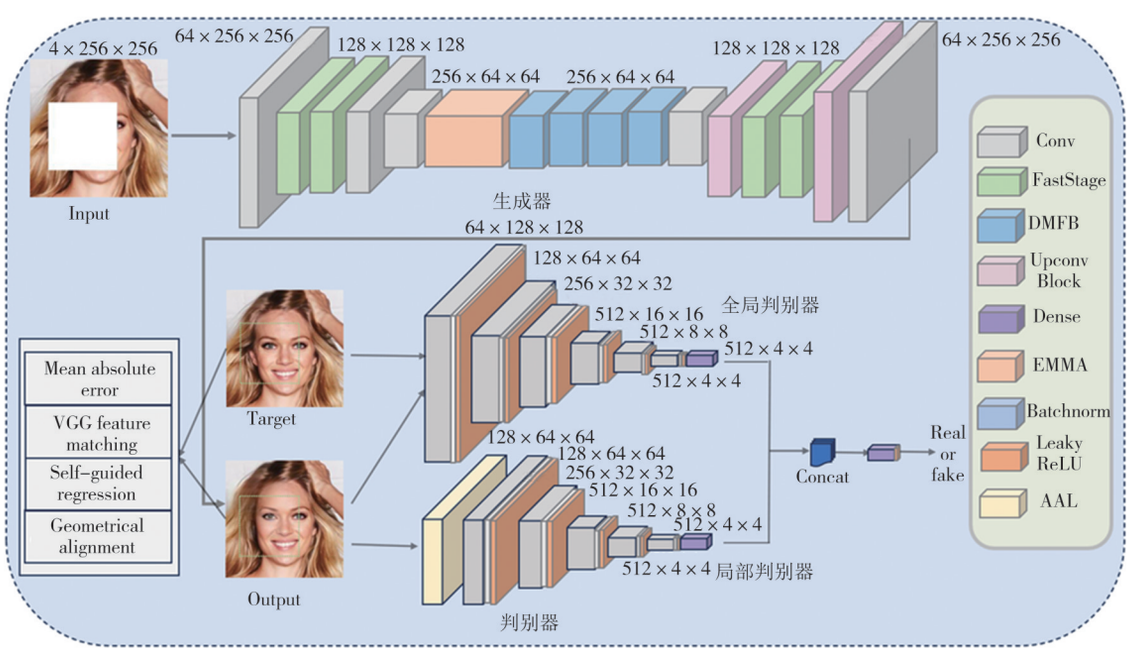


图1　整体结构

Fig.1　Overall structure

生成器采用编码器-解码器结构.编码器通过卷积层和FastVit[16]中的FastStage模块对输入图片进行特征压缩、重建和增强.接着通过EMMA注意力模块和密集多尺度融合模块(dense multi-scale fusion block,DMFB)[17]强化特征提取和融合能力,并对生成的特征图进行反噪和稳定.解码器通过卷积层处理特征图,接着通过上采样、FastStage模块和另一级上采样恢复图像尺寸.最后通过卷积层将特征图通道数转换为输出图像所需通道数,并通过tanh激活映射输出.

判别器通过使用全局判别器和局部判别器进行训练,以区分真实图像和修复图像.全局判别器对整个图像进行评估,用于判断整体一致性;局部判别器则专注于补全小区域,以确保生成的补丁在局部上

保持一致性.

2.2 生成器

生成器采用编码器-解码器结构.

编码器首先通过卷积层对输入图片进行压缩;然后,通过两个FastStage模块,对特征进行重建和增强.其中FastStage模块的结构如图2所示.FastStage模块由RepMixer block和Conv FFN构成.RepMixer block对输入图像进行卷积和特征映射,使用结构重参数化来减少网络中的跳跃连接,从而降低内存访问成本.Conv FFN模块对输入图像进行局部特征提取、通道数调整和非线性变换,从而增强模型的特征表达能力和非线性拟合能力,之后两层卷积用于提取和压缩输入特征图的信息,进一步提取和处理特征.

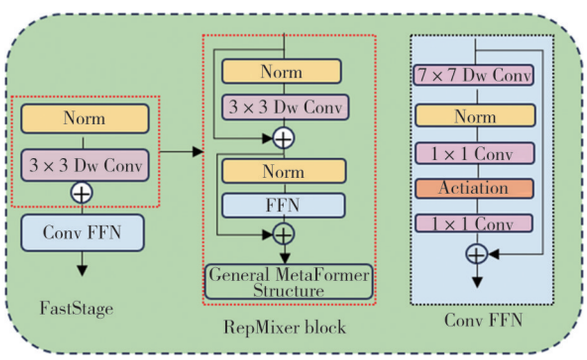


图2 FastStage结构

Fig.2 FastStage structure

其中,FastStage模块能够提供更准确的特征表示,从而在与周围联系紧密的缺失区域中较为合理地填充内容.

在解码器和编码器中间,利用EMMA注意力模块以及密集多尺度融合模块进行数据处理.其中,密集多尺度融合模块如图3所示,在前向传播过程中,输入经过一个卷积层和规范化层,然后将输出分别输入四个具有不同膨胀率的卷积层Conv-3-1、Conv-3-2、Conv-3-4和Conv-3-8.将得到的特征图相加、与Conv-3进行拼接形成combine特征图.再经过一个卷积层和规范化层得到第二个输出.将第二个输出与输入相加得到最终的输出.

DMFB主要作用是提取和加强输入特征,实现特征提取、特征融合和残差连接.其通过一系列的卷积操作和残差连接来执行特征处理,用于提取输入特征的高级表示.通过多尺度的卷积操作(包括不同的空洞率)和残差连接来引入非线性变换和上下文信息,从而增强输入特征的表达能力.

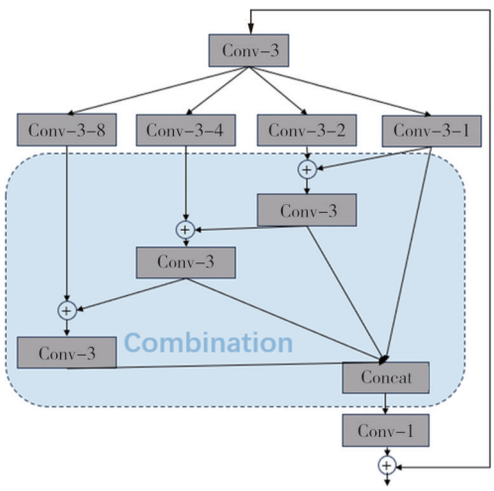


图3 密集多尺度融合模块

Fig.3 Dense multi-scale fusion block

解码器通过卷积层对特征进行处理并减少通道数,然后通过上采样-FastStage模块-上采样结构进行处理.其中,第一个上采样模块将输入特征图从4×nf维度上采样到2×nf维度,即从尺寸为64×64的特征图恢复到尺寸为128×128.第二个上采样模块将尺寸为128×128的特征图恢复到尺寸为256×256.FastStage模块在上采样层之间用于增强特征图的表示能力.最后通过卷积将特征图通道数转换为输出图像所需的通道数,并通过tanh激活进行输出结果的映射.

2.3 EMMA

EMMA结构如图4所示,灵感来自于文献[17]和文献[18],其中左边部分[17]先对输入特征进行分组,通过捕捉多尺度的空间信息来学习不同的语义信息.为了同时考虑不同大小的感受野,采用三条平行路径来提取分组特征图的注意力权重.两条路径在1×1分支上,另一条在3×3分支上.在1×1卷积的路径中,模块通过两个全局平均池化操作来捕获通道间的依赖关系,这样做可以在减少计算开销的同时,对通道进行有效的编码.这两个池化操作分别在图像的高度和宽度方向上进行,然后将得到的特征在高度方向上拼接,并共享一个1×1的卷积核,以保持特征图的维度不变.接下来,1×1卷积的输出被分解为两个向量,并通过两个非线性的Sigmoid函数来拟合卷积后的数据,这样可以实现不同的跨通道特征交互.3×3分支堆叠了一个3×3的卷积核,以捕获多尺度的特征表示,同时捕捉局部的跨通道交互,以扩大特征空间.这样既能编码通道间的信息,调整不同通道的重要性,同时还能保留精确的空间结构信息.

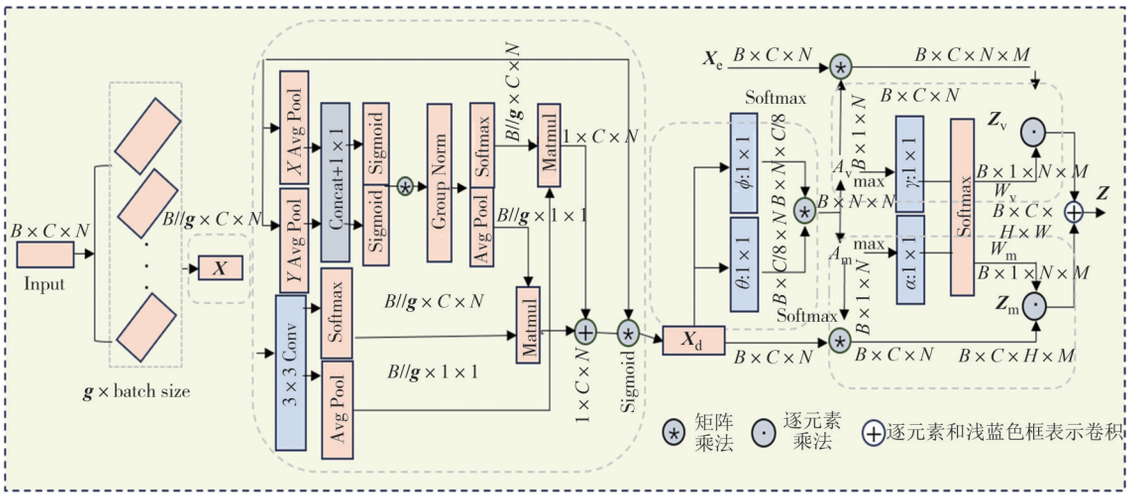


图 4　EMMA 结构

Fig.4　EMMA structure

综上所述，左侧部分同时保留每个通道上的信息并降低计算开销.其核心思想是将部分通道重塑为批量维度，并将通道维度分组成多个子特征组，以确保空间语义特征在每个特征组中均匀分布.

右侧部分[18]针对左侧部分输出的特征 $\boldsymbol{X}_{\mathrm{d}}$，首先,计算其注意力分数，接着基于掩码进行区分，并将注意力分数划分为两部分:A_{v}(可见区域相似度)和 A_{m}(生成区域相似度).然后,将 A_{v} 和 A_{m} 通过 Softmax 函数进行归一化,以确保每个空间位置上的权重和为 1.接下来,通过使用 1×1 滤波器 γ 和 α 来调节两个特征的权重比值.最后,经过注意力分数计算、权重调节和归一化之后,将得到的权重与编码特征进行乘法运算,并将结果进行加权求和,得到最终的输出特征 $\boldsymbol{Z}$.

可见,右侧部分通过注意力分数和权重调节，根据可见区域和生成区域的相似度，在每个空间位置上对编码特征和解码特征进行加权.这样既可以保留重要的可见区域信息，也可以结合生成区域的相似度,生成最终的加权特征表示.

综上所述，左侧部分能够捕捉不同层级的特征信息,提供了更准确的纹理和结构表示.它可以理解图像中不同尺度的细节信息,使模型对于缺失区域与周围区域的关系有更深入的认知.而右侧部分使模型能够更加集中地关注重要的图像区域.这样可以使模型在修复过程中更加关注缺失区域周围的重要结构和纹理细节.

2.4　判别器

判别器由全局判别器和局部判别器组成.

全局判别器由一系列卷积块组成，通过卷积操作、激活函数和批归一化来提取和处理输入图像的全局特征，这些卷积块逐渐减小特征图的尺寸并增加通道数，将输出的特征输入全局分类器进行分类.

局部特征提取器首先初始化一个注意力感知模块(AAL)[18],使其关注输入特征中的重要信息部分,提高模型对重要信息的敏感性.局部特征提取器由 local_fea1、local_fea2、local_fea3、local_fea4、local_fea5 组成,也是通过卷积操作、激活函数和批归一化来提取图像的局部特征.这些卷积模块逐渐减小特征图的尺寸并增加通道数.将输出特征经展平后输入局部分类器进行分类.

将局部特征和全局特征在通道维度上拼接在一起，形成最终的特征向量.将该特征向量输入到 LeakyReLU 激活函数和线性层,得到一个表示对输入图像的判别结果的标量.

2.5　损失函数

2.5.1　自引导回归损失(self-guided regression loss)

自引导回归损失是一种用于解决图像语义结构保持问题的方法.它将输入图像通过 VGG19 网络进行前向传递,生成一个五级特征图金字塔,空间分辨率逐渐降低.该方法旨在提高完整图像的细节保真度,尤其是针对建筑图像和人脸图像,通过使用自引导回归约束,可以优化生成图像与真实图像之间的差异图,从而更好地保留图像的细节信息.

为了更好地区分掩码区域和已知区域,我们根据生成图像和原图之间的误差来生成引导图,其可以在训练过程中用作额外的约束,以指导生成器生成更精确、质量更高的图像.为此,采用式(1)来计算平均误差:

$$M_{\mathrm{AEM}}=\frac{1}{3}\sum_{n\in N}(I_{\mathrm{out},n}-I_{\mathrm{GT},n}) \tag{1}$$

其中:N代表三个颜色通道;n表示输出图像的第n个通道;$I_{\mathrm{out},n}$表示输出图像的第n个通道;$I_{\mathrm{GT},n}$表示输入图像的第n个通道.

归一化掩码通过式(2)计算:

$$M_{\mathrm{GUI},p}=\frac{M_{\mathrm{AEM},p}-\min(M_{\mathrm{AEM}})}{\max(M_{\mathrm{AEM}})-\min(M_{\mathrm{AEM}})} \tag{2}$$

其中:$M_{\mathrm{AEM},p}$表示在位置p处的错误映射值;$\max(M_{\mathrm{AEM}})$是该层特征图中最大的误差值;$\min(M_{\mathrm{AEM}})$则是最小的误差值.

$\boldsymbol{M}^{l}_{\mathrm{GUI}}$对应第$l$层特征图,表示为

$$\boldsymbol{M}^{l+1}_{\mathrm{GUI}}=\mathrm{AP}(\boldsymbol{M}^{l}_{\mathrm{GUI}}) \tag{3}$$

AP表示平均池化操作,内核大小为2、步长为2.将其表示为:$\boldsymbol{M}^{1}_{\mathrm{GUI}}=\boldsymbol{M}_{\mathrm{GUI}}$,$\boldsymbol{M}^{l}_{\mathrm{GUI}}$的值范围为0~1.

由于低层特征图包含更为详细的信息,因此,本文选择第一层和第二层这两个层级描述图像语义结构的特征张量.这些层级通常具有更多关于图像语义结构的细节.自引导回归损失定义为

$$L_{\mathrm{SG}}=\sum_{l=1}^{2}m^{l}\frac{\left\|\boldsymbol{M}^{l}_{\mathrm{GUI}}\odot\left(\boldsymbol{\Phi}^{l}_{I_{\mathrm{gt}}}-\boldsymbol{\Phi}^{l}_{I_{\mathrm{output}}}\right)\right\|_{1}}{G_{\boldsymbol{\Phi}^{l}_{I_{\mathrm{gt}}}}} \tag{4}$$

其中:$\boldsymbol{\Phi}^{l}_{I_*}$是给定输入$I_*$的第$l$层的激活图;$G_{\boldsymbol{\Phi}^{l}_{I_{\mathrm{gt}}}}$是$\boldsymbol{\Phi}^{l}_{I_{\mathrm{gt}}}$的第$l$层的元素数量;$\odot$是元素乘积算子;$m^{l}=\frac{1\mathrm{e}3}{(N_{\boldsymbol{\Phi}^{l}_{I_{\mathrm{gt}}}})^{2}}$引用了文献[19],$N$是特征图$\boldsymbol{\Phi}^{l}_{I_*}$的通道大小,这种正则化的一个明显好处是抑制不确定性较高的区域.$\boldsymbol{M}_{\mathrm{GUI}}$被视为一种空间注意力图,针对难以处理的区域进行优化操作时表现最佳.

与传统的像素空间相比,本文的自引导回归损失更加注重较低层级的语义空间.这种方法的优势在于它能够在感知图像合成中保留具有结构信息的特点.这种方式能够更好地优化具有复杂语义结构的图像区域,提高图像合成的质量.

2.5.2 几何对齐约束(geometrical alignment constraint)

传统解决方案通常使用基于像素的损失来评估高级特征空间的度量,如L_1或L_2损失.然而,这种方法没有考虑到每个高级特征图的语义中心是否对齐.

为了更准确地衡量预测特征和真实目标之间的距离,本文引入了几何对齐约束,应用于第四层的响应图中.这个约束有助于生成器创建一个在位置上与目标图像对齐的合理图像.具体而言,几何对齐约束使生成图像的输出特征图的中心与目标特征图的中心在空间上靠近.本文通过计算在轴u上第k个特征图的几何中心来实现这一约束.计算公式为

$$c^{k}_{u}=\sum_{u,v}u\cdot(V(k,u,v)/\sum_{u,v}V(k,u,v)) \tag{5}$$

其中:响应图$V(\cdot)=\mathrm{VGG}(I;\theta_{vgg})\in\mathbb{R}^{K\times H\times W}$;$V(k,u,v)/\sum_{u,v}V(k,u,v)$表示空间概率分布函数;$c^{k}_{u}$表示沿轴$u$的坐标期望.

本文将生成的图像$\boldsymbol{I}_{\mathrm{Output}}$和真实图像$\boldsymbol{I}_{\mathrm{gt}}$通过VGG网络传递,得到相应的响应图$\boldsymbol{R}'$和$\boldsymbol{R}$.在给定这些响应图的情况下,我们使用式(5)计算中心点$<n^{k'}_{u},n^{k'}_{v}>$和$<n^{k}_{u},n^{k}_{v}>$.几何对齐损失表述为:

$$L_{\mathrm{GAC}}=\left\|<n^{k'}_{u},n^{k'}_{v}>-<n^{k}_{u},n^{k}_{v}>\right\|^{2}_{2} \tag{6}$$

2.5.3 特征匹配损失(feature matching loss)

VGG特征匹配损失L_{FMV}比较了训练良好的VGG19模型的中间层的激活图,可以写成

$$L_{\mathrm{FMV}}=\sum_{l=1}^{5}m^{l}\frac{\left\|\boldsymbol{\Phi}^{l}_{I_{\mathrm{gt}}}-\boldsymbol{\Phi}^{l}_{I_{\mathrm{output}}}\right\|_{1}}{G_{\boldsymbol{\Phi}^{l}_{I_{\mathrm{gt}}}}} \tag{7}$$

其中:$G_{\boldsymbol{\Phi}^{l}_{I_{\mathrm{gt}}}}$是$\boldsymbol{\Phi}^{l}_{I_{\mathrm{gt}}}$中的元素数;$\boldsymbol{\Phi}^{l}_{I_{\mathrm{gt}}}$代表第$l$层输入图像的激活图;$\boldsymbol{\Phi}^{l}_{I_{\mathrm{output}}}$是第$l$层输出图像的激活图.

本文在判别器特征匹配损失中引入了局部分支.这种局部分支的作用是合理地假设生成的图像与任何以高维空间度量的真实地面图像是一致的.判别器特征匹配损失的定义如下:

$$L_{\mathrm{FMD}}=\sum_{l=1}^{5}m^{l}\frac{\left\|A^{l}_{\mathrm{local}}(I_{\mathrm{gt}})-A^{l}_{\mathrm{local}}(I_{\mathrm{output}})\right\|_{1}}{G_{\boldsymbol{\Phi}^{l}_{\mathrm{local}}}(I_{\mathrm{gt}})} \tag{8}$$

其中:$A^{l}_{\mathrm{local}}(I_*)$是给定输入$I_*$的判别器的第$l$个选定层中的激活.需要注意的是,判别器的隐藏层是可训练的,与在ImageNet数据集上预训练的VGG19网络有所不同.本文判别器可以根据特定的训练数据自适应地更新其隐藏层参数.

2.5.4 对抗性损失(adversarial loss)

为了提高内部绘图结果的视觉质量,本文采用了ESRGAN中的相对论平均判别器.对于生成器部分,本文引入对抗性损失,定义如下:

$$\begin{aligned}L_{\mathrm{adv}}=&-Z_{x_{\mathrm{r}}}[\log(1-A_{R_a}(x_{\mathrm{r}},x_{\mathrm{f}}))]-\\&Z_{x_{\mathrm{f}}}[\log(A_{R_a}(x_{\mathrm{r}},x_{\mathrm{f}}))]\end{aligned} \tag{9}$$

其中:$A_{R_a}(x_{\mathrm{r}},x_{\mathrm{f}})=\mathrm{Sigmoid}\left[O(x_{\mathrm{r}})-Z_{x_{\mathrm{f}}}\left(O(x_{\mathrm{f}})\right)\right]$,$O(\cdot)$表示不包含最后Sigmoid函数的判别器网络;$(x_{\mathrm{r}},x_{\mathrm{f}})$是从真实图像和输出图像中采样的.

2.5.5 总体损失(overall loss)

本文结合自引导回归损失、几何对齐约束、VGG

特征匹配损失、判别器特征匹配损失、对抗性损失和平均绝对误差(MAE)损失，得到整体损失函数：

$$L_{\text{total}} = L_{\text{MAE}} + \lambda(L_{\text{SG}} + L_{\text{FMV}}) + \beta L_{\text{FMD}} + \delta L_{\text{adv}} + \varepsilon L_{\text{GAC}} \tag{10}$$

其中：λ、β、ε、δ用于平衡上述损失的影响；L_{MAE}为平均绝对误差损失.

3　实验

3.1　实验环境

实验环境是基于Windows 11系统，Python3.7和PyTorch环境，硬件环境包括了Intel Core i9-13900HX处理器，16 G内存，采用NVIDIA GeForce RTX 2060 SUPER图形处理器.

在模型训练初始阶段，将学习率设置为0.000 2，批次大小设置为4，迭代次数设置为100 000. 将输入图像的大小均设置为256×256，评价指标采用峰值信噪比（peak signal-to-noise ratio，PSNR）、结构相似性（structural similarity，SSIM）和L_2损失函数3种客观评价指标和人眼主观视觉对修复结果进行分析，并与Context Encoder[9]、DMFN[20]、ShiftNet[21]、CSA[22]这4种方法进行对比.

本文采用3个公共图像数据集来验证模型的性能. 这些公共图像数据集在图像修复任务中被广泛应用，旨在提供多样化的图像内容和具有挑战性的修复场景，从而评估模型在真实世界中的通用性和准确性. 对于不规则掩膜数据集，它提供了具有不同形状和位置的不规则遮挡掩膜，以模拟真实世界中出现的破损或遮挡情况.

CelebA：一个大规模的人脸数据集，包含超过20万张图像. 该数据集被广泛用于人脸识别、人脸表情识别和属性分析等任务.

Places2：一个用于场景识别的数据集，包含超过1 000万张图像，覆盖了场景、建筑、自然景观等多个类别. 该数据集常被用于图像分类、图像生成和场景理解等领域的研究.

Paris StreetView：该数据集包含了约14 900张用于训练的图像以及100张用于测试的图像，这些图像均来自巴黎真实的街景景观.

3.2　实验结果与分析

将本文所提出的方法与3.1节中提到的4种对比算法在结构相似性(SSIM)、峰值信噪比(PSNR)及L_2损失进行对比，本文算法在各项指标上表现出明显的优势.

图5~图7为本文方法与4种具有代表性的图像修复方法在3个公共数据集上，用中心掩码在人眼视觉上的修复效果对比，表1~表3为修复效果对比. 图8为本文方法在不规则掩码上的修复效果.

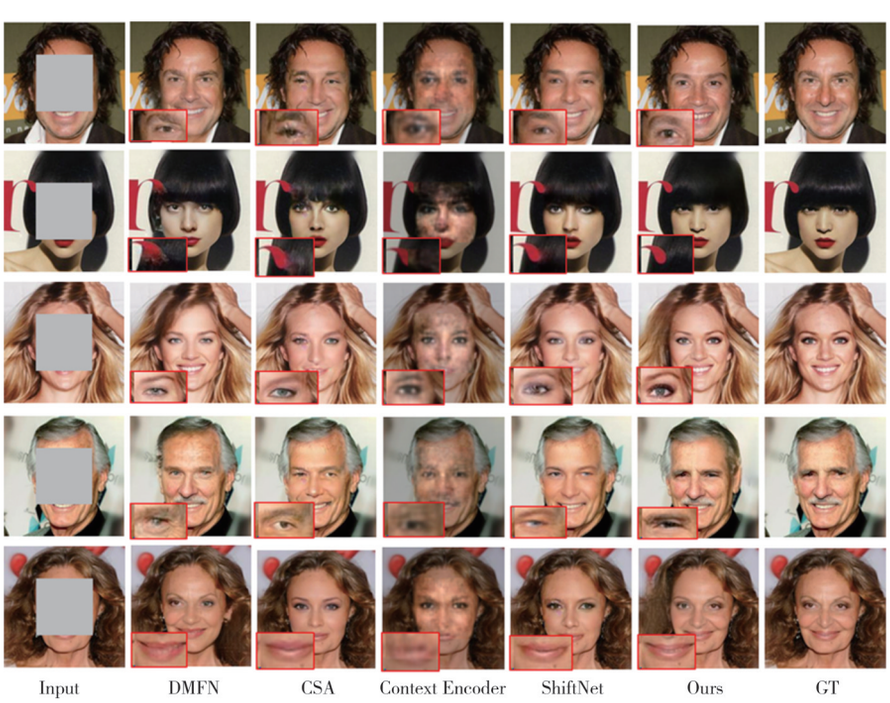


图5　CelebA数据集修复对比

Fig.5　Comparison of repair based on CelebA dataset

图5是在CelebA数据集上叠加规则掩码所得到的修复效果，可以看出掩码几乎覆盖了人脸的五官. DMFN算法在修复过程中，会出现局部修复模糊(第一行、第四行)，额头和头发之间的像素边界不清晰，导致出现伪影(第三行)，以及图像的边缘修复效果较差(第二行). CSA算法主要注重图像的语义信息，对于细节的修复有待提升，在修复的图像中局部的区域会出现纹理失真(第一、二、三、四行眼睛部分)、模糊和失真(第三行嘴巴). Context Encoder算法通常假设修复的区域相对较小，对于大面积缺失的图像部分，算法的修复效果不够理想. 这会导致修复后的图像在整体上出现不协调或突兀的现象. ShiftNet算法在修复过程中对于眼睛、眉毛以及嘴巴等的修复都比较模糊，如第二行中对脸颊左侧修复存在失真现象. 相比前面的4种算法，本文的算法在细节处理和整体结构上体现了优越性，对眼睛、嘴巴、边缘的修复都比较好.

表1为在数据集CelebA上的定量分析结果，其中最优结果用粗体表示，可以看出本文的方法在3个评价指标上的结果都优于其他算法. 由于Context Enocder算法只利用了输入数据的自身特征进行修复，所以修复效果较差.

表1 CelebA数据集定量分析

Tab.1 Quantitative analysis of the CelebA dataset

修复算法	CelebA		
	PSNR/dB	SSIM	L_2
DMFN	24.31	0.923 6	0.106 2
CSA	26.03	0.923 7	0.091 2
Context Encoder	22.14	0.872 4	0.173 1
ShiftNet	26.12	0.921 7	0.079 6
Ours	**27.21**	**0.937 1**	**0.073 4**

图6展示了在Paris StreetView数据集上应用不同算法修复后的效果. CSA算法可以勉强修复掩码区域的大致结构，但效果不理想，无法完整保留结构信息. 尤其是第五行中出现了大量与图像无关的内容，导致结构混乱. Context Encoder和ShiftNet算法因人物背景复杂而无法完全修复图像. 然而，CSA算法能利用已知区域的信息，完成整体结构修复. DMFN算法能够保留物体的整体结构，但在细节处理上仍有待改进，结构纹理不够清晰. 相比之下，本文算法修复结果结构清晰，无冗余内容，与原图相似度高.



图6 Paris StreetView数据集修复对比

Fig.6 Comparison of repair based on Paris StreetView dataset

表2为在数据集Paris StreetView上定量分析结果，其中最优结果用粗体表示，可以看出，ShiftNet算法的修复效果较CelebA数据集来说有所下降. 这是由于ShiftNet算法在执行具有复杂纹理和结构的图像修复任务时可能效果较差，在处理细节丰富的纹理或复杂的物体结构时，算法无法准确恢复缺失或损坏的区域.

表2　Paris StreetView数据集定量分析

Tab.2　Quantitative analysis of the Paris StreetView dataset

修复算法	Paris StreetView		
	PSNR/dB	SSIM	L_2
DMFN	21.46	0.845 5	0.134 7
CSA	23.24	0.832 6	0.125 6
Context Encoder	22.99	0.753 5	0.136 0
ShiftNet	21.35	0.835 6	0.115 9
Ours	**24.31**	**0.862 5**	**0.102 0**

图7展示了在Places2数据集上应用不同算法修复后的效果.根据其修复结果可看出DMFN、Context Encoder、ShiftNet由于背景复杂,不能修复出图像的整体结构.其中DMFN和ShiftNet算法可以将颜色与原图对应,但缺乏纹理结构的修复.CSA算法对于复杂背景的修复相对更好,它可以对整体结构进行填充,但填充细节较差,比如第一行的石头的边缘产生冗余内容,第二行中的格子未修复出来,第三行的置物架杆子下面出现伪影和第四行的棚子底下白色管道的边界模糊,而本文算法填充的内容虽然与原图有所差异,但符合视觉语义.

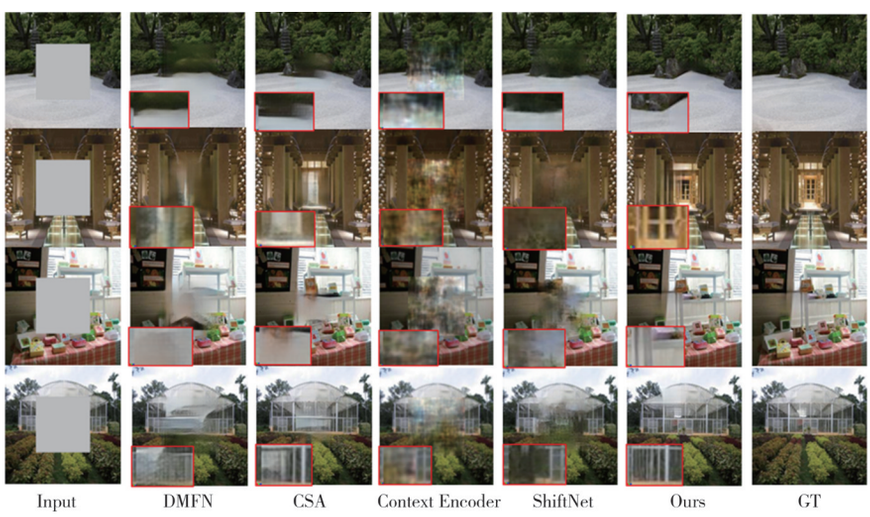


图7　Places2数据集修复对比

Fig.7　Comparison of repair based on Places2 dataset

表3为在Places2数据集上的定量分析结果,其中最优结果用粗体表示,可以看出由于这组数据集的内容以及背景的复杂性,修复效果相较于前面的数据集有所下降.但本文算法的修复效果优于其他算法.

表3　Places2数据集定量分析

Tab.3　Quantitative analysis of the Places2 dataset

修复算法	Place2		
	PSNR/dB	SSIM	L_2
DMFN	21.58	0.857 0	0.160 7
CSA	22.47	0.869 5	0.124 7
Context Encoder	19.43	0.752 5	0.166 2
ShiftNet	20.71	0.864 1	0.175 3
Ours	**23.31**	**0.878 5**	**0.102 6**

图8展示了本算法在处理不规则掩码下的图像修复效果.结果表明,该算法在修复图像时表现出色,即便是对于具有复杂背景的图像,也能够实现高质量的修复效果.

3.3　消融实验

为了验证本文方法的有效性,本文在中心正方形掩码下进行测试.在同等条件下,本文分别移除了相应的模块,并对3个数据集进行了定性和定量对比.这样的对比可以更好地揭示本文方法与其他方法之间的差异和优势.通过这些实验数据,可以更具说服力地证明本文方法的有效性.

如图9所示,第二列为加入FastStage模块的效果,可以看到加入之后的整体结构较为合理,但缺失区域与周围区域密切相关时,并不能完全准确地推断其内容,修复清晰度、纹理细腻度和轮廓锐利度都有待提高.第三列在第二列基础上加入了AAL,可以看出它在细微纹理和局部特征上的修复效果更好,鼻孔、窗户以及柱子都比之前有了更清晰的轮廓.第四列为本文提出的方法,在第三列的基础上加入了EMMA,加入之后对于细节的处理更加精准,有较好的语义连续性和合理性,当修复区域与周围密切相关时,可以较为清晰地修复缺失区域的内容.表4展示了不同模块在3个评价指标上的效果,其中最优结果用粗体表示.

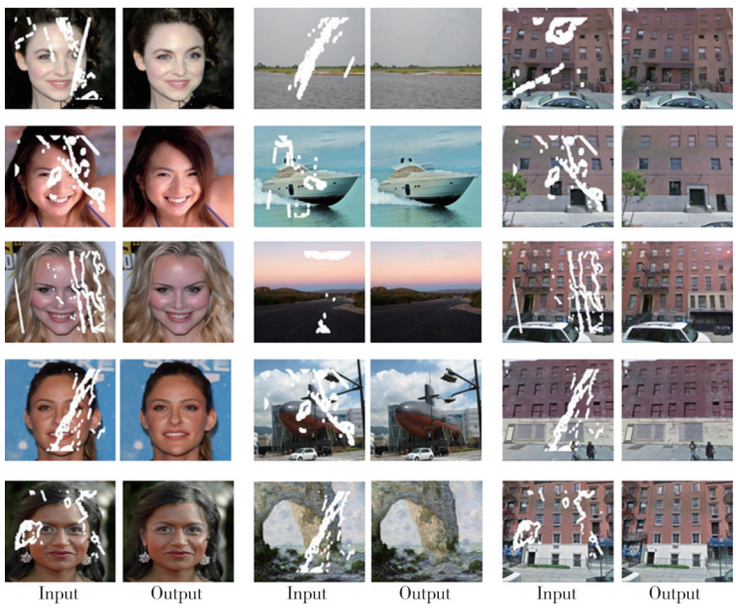


图 8 在 3 个数据集上不规则掩码修复对比

Fig.8 Comparison of irregular mask repair on three datasets

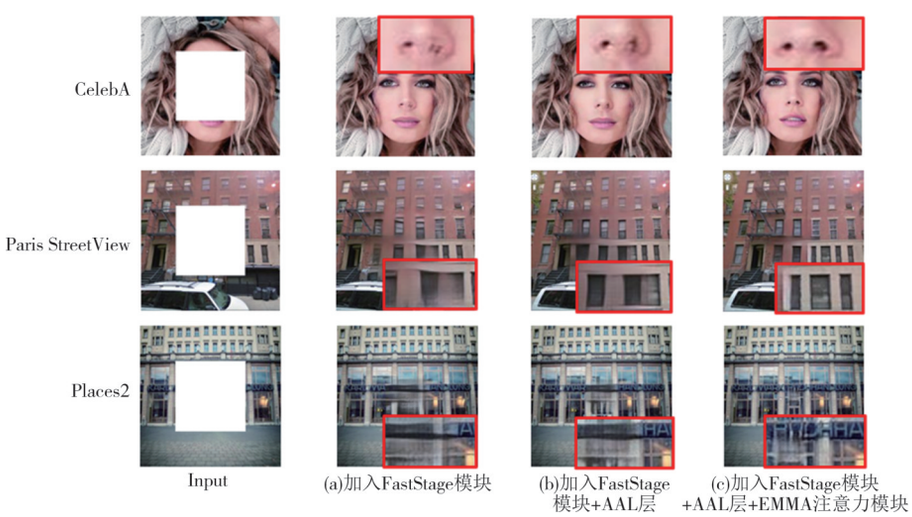


图 9 消融实验结果对比

Fig.9 Comparison of ablation experimental result

表 4 模块消融实验对比

Tab.4 Comparison of module ablation experiments

模块	PSNR/dB	SSIM	L_2
+FastStage(a)	24.31	0.924 1	0.102 2
+FastStage+AAL(b)	25.46	0.930 5	0.078 9
+FastStage+AAL+EMMA(c)	**28.64**	**0.937 6**	**0.071 2**

4 总 结

本文提出的单阶段修复模型旨在解决图像中缺失区域与周围纹理和结构密切相关时，传统方法可能无法准确推断缺失区域内容的问题，从而使修复结果更符合实际需求. 该模型充分利用卷积、自注意力和多层感知机的优点，以实现对特征的提取和重建. 此外，本文通过引入 EMMA 注意力模块，进一步提升了特征融合和降噪的能力，确保修复结果能够满足实际需求. 经过实验评估，针对人脸、街景以及建筑数据集进行的端到端实验结果表明，该模型的修复结果更符合视觉语义，能够精细地修复图像的细节纹理和局部特征.

参考文献

[1] 张红英，彭启琮. 数字图像修复技术综述[J]. 中国图象图形学

报,2007(01):1-10.
ZHANG H Y, PENG Q C. A survey on digital image inpainting [J]. Journal of Image and Graphics, 2007, 12(1): 1-10. (in Chinese)

[2] CRIMINISI A, PEREZ P, TOYAMA K. Region filling and object removal by exemplar-based image inpainting [J]. IEEE Transactions on Image Processing, 2004, 13(9): 1200-1212.

[3] BARNES C, SHECHTMAN E, FINKELSTEIN A, et al. PatchMatch: a randomized correspondence algorithm for structural image editing [C]//Seminal Graphics Papers: Pushing the Boundaries. August 3-7, 2009, New York, NY, USA: ACM, 2023: 619-629.

[4] SUN J, JIA J, TANG C K. Efficient patch-based inpainting for large-scale image editing [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 34(6): 1038-1050.

[5] YOU Y L, XU W Y, TANNENBAUM A, et al. Behavioral analysis of anisotropic diffusion in image processing [J]. IEEE Transactions on Image Processing, 1996, 5(11): 1539-1553.

[6] LIU G L, REDA F A, SHIH K J, et al. Image inpainting for irregular holes using partial convolutions [M]//Lecture Notes in Computer Science. Cham: Springer International Publishing, 2018: 89-105.

[7] ZENG Y H, FU J L, CHAO H Y, et al. Learning pyramid-context encoder network for high-quality image inpainting [C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). June 15-20, 2019, Long Beach, CA, USA: IEEE, 2019: 1486-1494.

[8] LI J Y, WANG N, ZHANG L F, et al. Recurrent feature reasoning for image inpainting [C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). June 13-19, 2020, Seattle, WA, USA. IEEE, 2020: 7757-7765.

[9] PATHAK D, KRÄHENBÜHL P, DONAHUE J, et al. Context encoders: feature learning by inpainting [C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). June 27-30, 2016, Las Vegas, NV, USA: IEEE, 2016: 2536-2544.

[10] IIZUKA S, SIMO-SERRA E, ISHIKAWA H. Globally and locally consistent image completion [J]. ACM Transactions on Graphics, 2017, 36(4): 1-14.

[11] YI Z L, TANG Q, AZIZI S, et al. Contextual residual aggregation for ultra high-resolution image inpainting [C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). June 13-19, 2020, Seattle, WA, USA. IEEE, 2020: 7505-7514.

[12] WANG Y, TAO X, QI X J, et al. Image inpainting via generative multi-column convolutional neural networks [EB/OL]. 2018: 1810.08771. https://arxiv.org/abs/1810.08771v1.

[13] ARJOVSKY M, CHINTALA S, BOTTOU L. Wasserstein generative adversarial networks [C]//International conference on machine learning. PMLR, 2017: 214-223.

[14] MAO X D, LI Q, XIE H R, et al. Least squares generative adversarial networks [C]//2017 IEEE International Conference on Computer Vision (ICCV). October 22-29, 2017, Venice, Italy. IEEE, 2017: 2813-2821.

[15] MEHRALIAN M, KARASFI B. RDCGAN: unsupervised representation learning with regularized deep convolutional generative adversarial networks [C]//2018 9th Conference on Artificial Intelligence and Robotics and 2nd Asia-Pacific International Symposium. December 10-10, 2018, Kish Island, Iran. IEEE, 2018: 31-38.

[16] HATAMIZADEH A, HEINRICH G, YIN H X, et al. FasterViT: fast vision transformers with hierarchical attention [EB/OL]. 2023: 2306.06189. https://arxiv.org/abs/2306.06189v2

[17] OUYANG D, HE S, ZHAN J, et al. Efficient multi-scale attention module with cross-spatial learning [EB/OL]. ArXiv, 2023, abs/2305.13563. DOI: 10.1109/ICASSP49357.2023.10096516.

[18] ZHENG C X, CHAM T J, CAI J F, et al. Bridging global context interactions for high-fidelity image completion [C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). June 18-24, 2022, New Orleans, LA, USA: IEEE, 2022: 11502-11512.

[19] ZHOU Y, ZHU Z, BAI X, et al. Non-stationary texture synthesis by adversarial expansion [J]. ACM Transactions on Graphics, 2018, 37(4): 1-13.

[20] OUYANG D L, HE S, ZHANG G Z, et al. Efficient multi-scale attention module with cross-spatial learning [C]//ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). June 4-10, 2023, Rhodes Island, Greece: IEEE, 2023: 1-5.

[21] YAN Z Y, LI X M, LI M, et al. Shift-net: image inpainting via deep feature rearrangement [M]//Lecture Notes in Computer Science. Cham: Springer International Publishing, 2018: 3-19.

[22] LIU H Y, JIANG B, XIAO Y, et al. Coherent semantic attention for image inpainting [C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). October 27-November 2, 2019, Seoul, Korea (South). IEEE, 2019: 4169-4178.