DOI:10.19651/j.cnki.emt.2417416

融合多特征与全局-局部 Transformer 的图像修复算法*

滕诗宇 何丽君
(大连民族大学计算机科学与工程学院 大连 116600)

摘 要:针对当前图像修复领域所面临的高计算复杂度以及在生成结构合理且细节丰富的图像方面的局限,提出了一种融合多尺度分层特征与全局-局部协同 Transformer 的图像修复模型。首先提出多尺度分层特征融合模块,以实现深层特征与浅层特征细节上的有效融合,在扩大感受野的同时减少关键信息丢失情况。其次提出用于全局推理的全局-局部协同 Transformer 模块,它通过集成矩形窗口注意力机制和局部前馈神经网络,在降低计算复杂度的同时,提高模型对全局上下文信息的宏观理解和对局部细节特征的微观捕捉能力,增强图像的整体一致性。实验在 CelebA-HQ 和 Places2 数据集上进行了验证,在处理 40%~50%掩码时,所提方法与常用的修复方法对比,PSNR 平均提高了 0.26~6.25 dB,SSIM 平均提升了 1.4%~19%,L1 平均下降了 0.2%~5.66%。实验证明,所提方法修复后的图像在视觉上具有更加真实和自然的效果,进一步验证了该方法的有效性。

关键词:深度学习;图像修复;多尺度分层特征融合;全局-局部协同 Transformer;矩形窗口注意力机制;局部前馈神经网络

中图分类号:TP391.4;TN919.8 **文献标识码**:A **国家标准学科分类代码**:510.1050

Fusion of multi-features and global-local Transformer for image inpainting

Teng Shiyu He Lijun
(College of Computer Science and Engineering, Dalian Minzu University, Dalian 116600, China)

Abstract: Addressing the challenges in the domain of image inpainting, such as the high computational complexity, loss of information during feature extraction, and the blurring of textures in the inpainting images, this study proposed a image inpainting model that integrates multiscale hierarchical feature fusion with synergetic global-local Transformer. Initially, the multi-scale hierarchical feature fusion block was proposed as a means of effectively fusing deep and shallow features in detail, thereby reducing the loss of key information while expanding the sensory field. Subsequently, synergetic global-local Transformer blocks for global reasoning was proposed, featuring an integrated rectangle-window self-attention mechanism and local feed-forward neural networks. This design reduced computational complexity while enhancing the model′s macroscopic understanding of global context and microscopic grasp of local detail characteristics. The proposed method was validated on the CelebA-HQ and Places2 datasets, and the results demonstrated that it yielded improvements in PSNR by an average of 0.26~6.25 dB, SSIM by an average of 1.4%~19%, and L1 decreased by an average of 0.2%~5.66% compared to commonly used inpainting methods when dealing with 40%~50% masks. The experiments show that the inpainted images resulting from the proposed method exhibit a more realistic and natural visual effect, thereby providing further validation of the method′s effectiveness.

Keywords: deep learning; image inpainting; multi-scale hierarchical feature fusion; synergetic global-local Transformer; rectangle-window self-attention; local feed-forward network

0 引 言

图像修复是图像处理领域的重要课题之一,旨在完善缺失的内容和纹理细节,恢复受损或不完整的图像[1]。图像修复技术在许多应用中发挥重要作用,包括图像编辑[2]、对象移除[3]以及文物修复[4]等。通过运用先进的图像修

收稿日期:2024-11-21
*基金项目:辽宁省应用基础研究计划项目(2023JH2/101300191)、辽宁省自然科学基金(2023-MS-133)、辽宁省教育厅科研项目(LJKZ0024)资助

复技术，可以显著提升图像的整体质量，有效地修复受损区域，让修复后的图像在视觉上和结构上更加自然和完整。

图像修复技术的早期发展主要基于两类算法：基于补丁的修复方法和基于扩散的修复方法。虽然这类方法在应对小面积缺失或少量像素缺失的任务上有不错的成效，然而，当面对图像中有复杂的语义内容或较大缺失区域时，它们往往难以生成与周围内容一致的纹理和结构，导致修复效果不佳。

随着深度学习的发展，基于深度学习的方法在图像修复领域取得显著的成果。尤其是基于生成对抗网络(generative adversarial networks，GAN)的方法[5]，它通过利用生成器和判别器的相互对抗学习机制，使得模型在理解和重建图像中的深层语义信息及细节方面展现出了卓越的能力，显著提升了修复的质量。Pathak 等[6]提出的上下文编码模型(context encoders，CE)，将 GAN 的思想引入图像修复领域，通过设计的编码器-解码器架构，有效地利用了图像中的上下文信息来指导缺失区域的重建。Yu 等[7]提出的两阶段修复策略，先通过卷积层大致恢复缺失图像的轮廓，紧接着利用上下文注意力机制获取远距离的信息特征，从而精细化破损区域的细节信息。Li 等[8]结合部分卷积[9]提出了循环推理网络，该网络利用循环特征推理模块和知识一致性注意力模块，有效修复了包含复杂语义信息的图像。Zeng 等[10]则结合门控卷积[11]引入辅助上下文重建机制和改进的 GAN 架构对图像进行多样化修补。高杰等[12]借助门控卷积对已知区域与掩码区域的关系展开动态学习，从而提升图像的整体修复质量。陈婷等[13]提出的边缘先验融合动态门控特征的方法，能高效提取动态特征及边缘信息，用以修复人脸图像。Zeng 等[14]通过构建多层聚合上下文转换模型(aggregated contextual-transformation gan，AOT-GAN)，从图像的不同尺度捕获信息，并整合这些信息来理解和填补图像中的缺失部分。Suvorov 等[15]提出的 LaMa 模型采用快速傅里叶卷积在频域内进行卷积操作，使得模型能够生成具有真实感和细节丰富的图像。这些方法在理解和重建复杂的语义信息方面取得了显著进步，但在面对大面积损坏时，往往会因为难以充分捕捉上下文信息而导致修复后的图像出现结构扭曲、纹理模糊现象。为了应对这些挑战，最近的研究探索了更具灵活性的 Transformer 模型[16]，它通过自注意力机制和并行处理，能够更有效地捕获长距离依赖关系和复杂特征。Wan 等[17]受到 Transformer 的启发，提出了一种具有双向注意力机制的修复模型(image completion with Transformer，ICT)，该机制可以有效捕捉全局信息，显著提升了修复图像的自然度。Zheng 等[18]提出的内容推理 Transformer，通过在所有层中，显式地以同等重要性建立长距离的可见上下文关系，使得模型理解图像全局信息的能力得到了增强。向泽林等[19]充分借助 Swin-Transformer 强大的特征提取效能来捕获全局特征信息，解决了卷积难以整合长距离信息的问题。为了解决 Transformer 因侧重全局信息整合而忽略局部细节，导致修复图像完整性欠佳的问题，杨红菊等[20]利用局部增强滑动窗口 Transformer，在充分捕获全局信息的同时，有效提升模型对于图像局部细节的感知精度。Liu 等[21]通过双向特征交互模块强化 Transformer 对局部特征的提取能力，有效弥补了 Transformer 在局部信息获取方面存在的短板，提高了修复的整体质量。徐志刚等[22]提出一种结合 CSWin-Transformer 和门卷积的修复方法，在全局层利用动态条纹窗口提取特征，在局部层借助残差连接融合输入特征与残差块输出权值，使得生成的图像结构和纹理更加协调一致。上述基于 Transformer 的修复方法虽然取得了良好的修复效果，但是在处理具有复杂纹理和结构的图像时有时也难达成自然融合的效果，同时它们还带来了高计算复杂度、参数需求大和训练周期过长等问题，增加了对内存和计算资源的需求。

针对上述问题，本研究提出了一种融合多尺度分层特征与全局-局部协同 Transformer 的轻量化两阶段修复模型，它能有效地为图像缺失区域生成合理的内容，同时保持修复图像的真实感和协调性。主要工作为：1)提出了全局-局部协同 Transformer(synergetic global-local Transformer，SGT)，SGT 利用矩形窗口注意力机制，捕获图像中的复杂共现全局特征，并基于这些特征生成更精细的修复结果。此外，为了解决 Transformer 在训练过程中缺乏归纳偏差，易忽视细粒度局部细节的问题，本研究在 SGT 中设计了局部前馈神经网络(local feed-forward network，LFFN)，以增强对局部特征的建模能力，使生成的图像在自然性和真实感上有显著提升，进而增强整体一致性。2)提出了多尺度分层特征模块(multi-scale hierarchical feature fusion block，MHFFB)，该模块在训练过程中通过融合不同层次和不同尺度的特征，能够自适应地提取图像的关键信息，同时抑制非关键信息对模型的干扰。这一设计有助于在粗修复阶段保留更多的有效信息，使得生成的图像具有更丰富的细节信息。3)在广泛的数据集上进行训练和测试。与现有方法相比，本文方法在处理具有复杂结构和纹理以及大面积缺失区域的图像时，展现出更高的修复质量和更好的视觉效果。

1　本文方法

1.1　模型架构

本文的修复框架如图 1 所示。第一阶段是粗修复阶段，旨在快速填补损坏的区域，为损坏的部分提供一个大致的轮廓。具体地，破损的图像经过由门控卷积组成的编码器进行下采样操作，提取图像中的特征。接着，通过 4 个 MHFFB，以获取不同层次的之间的特征并将它们有效融合。最后，这些融合后的特征被输入解码器，生成粗修复的图像结果。

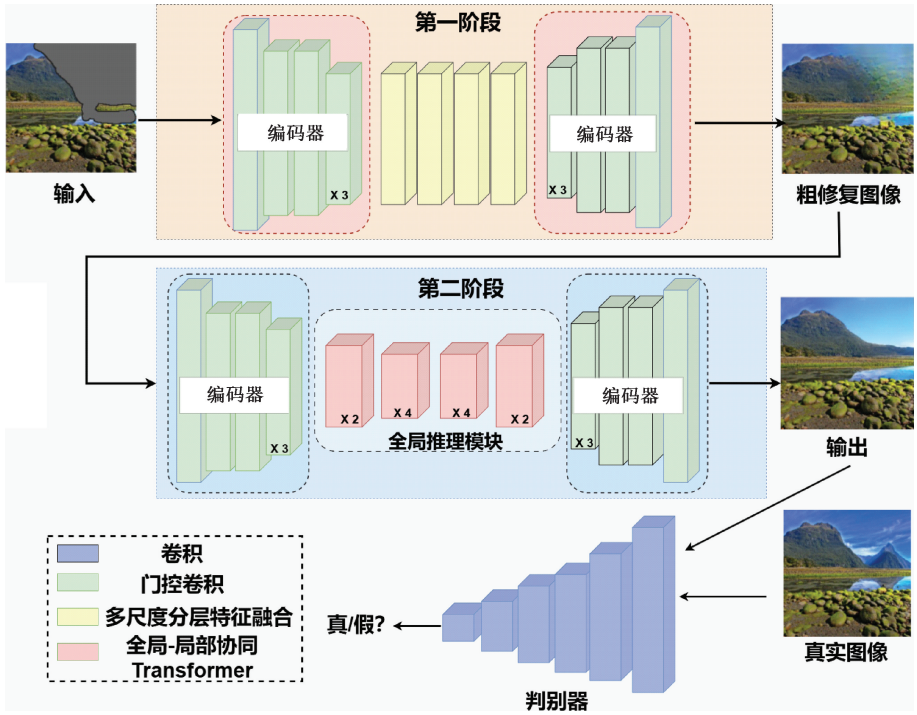


图 1 融合多特征和全局-局部协同 Transformer 的图像修复框架

Fig. 1 The framework of fusion of multi-features and global-local Transformer for image inpainting

第二阶段是精修复阶段,旨在对图像的细节、纹理进行更深入地学习和处理。具体地,粗修复图像经过编码器处理后,被送进全局推理模块中,该模块的作用是全面考虑缺失区域与未受损区域之间的依赖关系,它由 N 个 SGT 块组成,每个 SGT 块用于同时处理图像的全局特征和局部细节特征,通过这种设计,使得网络能够更全面地理解图像的内容。最后,将生成最终的修复图像结果并送入马尔可夫鉴别器中进行对抗学习。

1.2 多尺度分层特征融合模块

现有的图像修复网络大多采用编码器-解码器的体系结构,该结构在反复进行上采样和下采样操作时容易出现信息丢失的情况。在这种情况下,感受野的大小变得尤为关键,合适的感受野能够更有效地捕获图像中的上下文信息,有助于保留深层特征的同时,减少浅层特征丢失的风险,提升修复结果的质量。先前的研究通常采用膨胀卷积技术来扩展神经网络的感受野,在不引入额外的参数量的情况下增加图像的处理区域,有效扩大了感受野的范围,进而提升网络对特征信息的提取能力。然而,膨胀卷积的内核结构相对稀疏,可能会限制网络对特征中纹理细节和边缘信息的充分利用,对修复图像的细节和纹理效果产生负面影响。

为了解决这一问题,本文提出了 MHFFB,以应对传统的膨胀卷积在特征信息利用方面的局限性。MHFFB 通过结合不同扩张率的策略和分层特征融合技术,在扩大感受野范围的同时实现对不同尺度和不同层级特征的有效整合。与膨胀卷积相比,MHFFB 有着更密集的内核结构,使得网络在处理图像时能够动态地提取关键特征,同时有效抑制无关信息对模型的影响。MHFFB 的结构如图 2 所示。

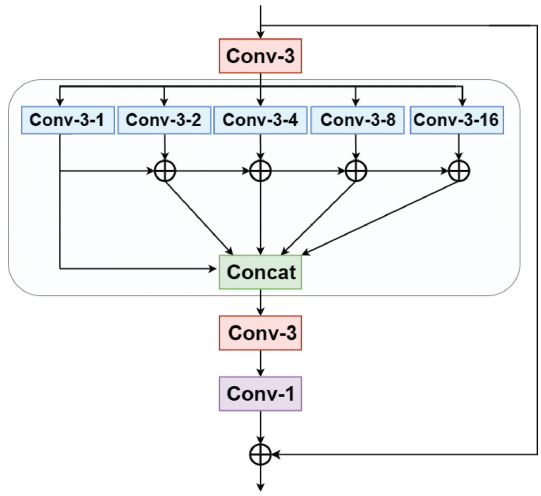


图 2 多尺度分层特征融合模块

Fig. 2 Multi-scale hierarchical feature fusion block

输入特征经过一个 3×3 卷积运算将其的通道数降低,再通过分层添加不同扩张率的方式,将不同尺度的特征信息融合起来。本文通过设置了 5 个 3×3 膨胀卷积核,这 5 个膨胀卷积核的扩张率分别是 1、2、4、8、16,通过使用具有不同扩张率的膨胀卷积来提取多尺度特征,得到的特

征可表示为 $X_i(i=1,2,3,4,5)$。之后将不同尺度的特征以逐元素相加的方式结合起来，这不仅保留了每个尺度特征图中的关键信息，还通过叠加的方式增强了这些信息的密度和丰富性。本文用 Y_i 表示的输出：

$$Y_i=\begin{cases}X_i, & i=1\\ X_{i-1}+X_i, & i=2\\ Y_{i-1}+X_i, & 2<i\leqslant 5\end{cases} \tag{1}$$

经过上述操作，具有不同感受野范围的特征和不同层级的密集特征被整合在一起，有效地扩大了感受野范围也实现了从浅层特征到深层特征的关键信息的全面提取，增强了修复网络重建细节丰富的图像的能力。

1.3　全局-局部协同 Transformer 模块

对图像的纹理和结构进行更加精细地修复，关键在于充分挖掘图像中未破损区域的上下文信息及捕获长距离的依赖关系。为了提升修复质量，增强图像的真实感，本文提出一种用于全局推理的 SGT 结构，该结构专注于捕获从全局到局部的纹理结构信息，以此引导神经网络实现对受损图像的高精度修复。

SGT 结构如图 3 所示，本文引入注意力机制——矩形窗口自注意力机制(rectangle-window self-attention，Rwin-SA)[23]，该机制采用矩形滑动窗口的形式，通过不同头并行处理水平和垂直方向的矩形窗口注意力，扩展了注意力区域并聚合来自不同窗口的特征，与文献[19]中使用的方形窗口注意力机制相比，Rwin-SA 不仅增强了相邻窗口之间的信息交流，还提升了模型对全局特征的捕捉和聚合能力。Rwin-SA 的结构图如图 3 右下方所示。

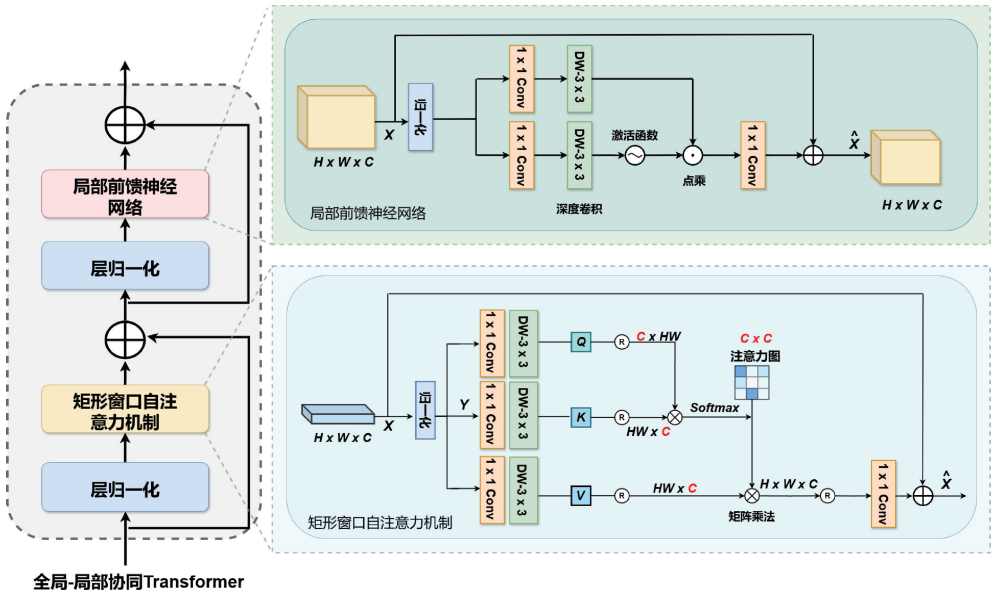


图 3　全局-局部协同 Transformer 结构

Fig. 3　The structure synergetic global-local Transformer

具体而言，对给定的 $X\in R^{C\times H\times W}$ 执行归一化操作得到 $Y\in R^{C\times H\times W}$，然后将 Y 拆分为不重叠的 $s_h\times s_w$ 矩形窗口，并将第 i 个矩形窗口表示为 $Y_i\in R^{(s_h\times s_w)\times C}$。

其中，H、W、C 分别代表图片高度、宽度和输入通道数，s_h，s_w 分别代表矩形窗口的高度和宽度，$i=1,\cdots,\frac{H\times W}{s_h\times s_w}$。

为了更好地计算通道间的交叉协方差，在 Rwin-SA 中应用 1×1 卷积来聚合像素级跨通道特征，随后使用 3×3 深度卷积来处理通道级的空间信息，产生了 $\boldsymbol{Q}_i^m=W_d^QW_p^QY$，$\boldsymbol{K}_i^m=W_d^KW_p^KY$ 以及 $\boldsymbol{V}_i^m=W_d^VW_p^VY$。

其中，$\boldsymbol{Q}_i^m,\boldsymbol{K}_i^m,\boldsymbol{V}_i^m\in R^{C\times d}$ 表示第 m 个头的查询、键和值的投影矩阵，$W_p^{(\cdot)}$ 表示 1×1 卷积，$W_d^{(\cdot)}$ 表示 3×3 深度卷积，Rwin-SA 可计算为：

$$Attention(\boldsymbol{Q}_i^m,\boldsymbol{K}_i^m,\boldsymbol{V}_i^m)=SoftMax\left(\frac{\boldsymbol{Q}_i^m(\boldsymbol{K}_i^m)^{\mathrm{T}}}{\sqrt{d}}+B\right)\boldsymbol{V}_i^m \tag{2}$$

$$\hat{X}=W_pAttention(\boldsymbol{Q}_i^m,\boldsymbol{K}_i^m,\boldsymbol{V}_i^m)+X \tag{3}$$

其中，$d=\frac{C}{M}$ 是每个头的信道维度，M 表示执行注意力操作次数，B 是动态相对位置编码。

尽管 Transformer 在捕获长距离依赖关系方面具有显著优势，但在处理图像的线条、纹理和形状等细节信息方面仍存在局限，这些细节信息对提升模型的修复精度至关重要。为解决这一问题，提出 LFFN，如图 3 右上方所示。受到先前研究[24]的启发，本文在前馈神经网络中结合门控机制和深度卷积。门控层通过调节通道中的信息流，使每层能够专注于与其他层互补的细节特征，从而提升细节捕捉的精度。而深度卷积通过聚合每个输入通道的特征，有

效编码空间相邻像素位置的信息,增强模型对局部区域特征的敏感度,使其能够更好地捕捉图像的细微变化与结构信息,从而提升整体修复性能。LFFN 的公式为:

$$Gating(X)=\phi(W_d^1W_p^1(LN(X)))\odot W_d^2W_p^2(LN(X)) \tag{4}$$

$$\hat{X}=W_p^0Gating(X)+X \tag{5}$$

其中,$\odot$ 表示逐元素乘法,ϕ 表示 GELU 非线性激活函数,$LN(\cdot)$ 是 LayerNorm 运算。SGT 由 Rwin-SA 和 LFFN 组成,它的主要优势:在增强 Transformer 对长程依赖关系捕获能力的同时,显著提升了对图像局部细节的处理能力。SGT 的计算过程的公式为:

$$\hat{x}^l_f=Rwin-SA(LN(x^{l-1}_f))+x^{l-1}_f \tag{6}$$

$$x^l_f=LFFN(LN(\hat{x}^l_f))+\hat{x}^l_f \tag{7}$$

其中,x^{l-1}_f 和 x^l_f 分别代表 SGT 的输入和输出特征图。

1.4 损失函数

令真实图像表示为 I_{g_t},对应的二进制掩模表示为 M(已知像素为 0,缺失像素为 1),第 i 阶段的生成器表示为 G_i。第 i 阶段的图像修复结果可以表示为:

$$I_{out_i}=(1-M)\odot I_{g_t}+M\odot G_i((1-M)\odot I_{g_t},M) \tag{8}$$

重建损失是用于衡量生成的修复图像与真实图像之间的差异。它包含第一阶段和第二阶段的像素级重构损失,其定义为:

$$L_{rec}=\sum_{i=1}^{2}|I_{g_t}-I_{out_i}| \tag{9}$$

感知损失不仅考虑了像素级别的差异,还考虑了更高级别的特征差异。其定义为:

$$L_{per}=\sum_i\frac{\|\phi_i-\phi_i(I_{out_2})\|}{N_i} \tag{10}$$

其中,ϕ_i 表示修复网络第 i 层的特征映射,N_i 表示 ϕ_i 中元素的数量的元素数量。

对抗损失本文采用的是谱归一化马尔可夫鉴别器。其定义为:

$$L_{adv}=E_{I_{g_t}}[\log D(I_{g_t})]+E_{I_{out}}[\log(1-D(I_{out}))] \tag{11}$$

用于训练修复模型的总损失 L 可表示为:

$$L=\lambda_rL_{rec}+\lambda_pL_{per}+\lambda_aL_{adv} \tag{12}$$

其中,L_{rec},L_{per} 和 L_{adv} 分别代表重建损失、感知损失和对抗损失。通过实验,本文设置 $\lambda_r=1$,$\lambda_p=0.1$,$\lambda_a=0.01$。

2 实验及分析

2.1 实验数据集

在两个广泛认可的数据集 CelebA-HQ[25] 和 Places2[26] 上对所提出的方法进行评估。CelebA-HQ 数据集包含 30 000 张高质量人脸图像,选择其中 5 000 张被用作测试集,其余作为训练集。Places2 数据集总共包含超过 1 000 万张场景图片,400 多种类别的大规模数据集。选择其中 10 个类别,每个类别随机选择 3 000 张图像,其中 25 000 张作为训练,5 000 张作为测试。在实验中,所有图像均被调整至 256×256 像素的统一分辨率。

2.2 实验设置与评价指标

本文使用 Pytorch 框架实现所提出的方法,并在一块显存为 24 GB 的 NVIDIA 3090 GPU 上完成所有的实验。经过多次实验,最终将 batch size 为 16,并且根据全局推理块中 SGT 层级不同,将 SGT 块的数量依次定义为 2、4、4、2。在训练阶段,使用不规则 mask 数据集,全面地评估本文提出的模型,选用的 Adam 优化器参数配置为 $\beta_1=0.5$,$\beta_2=0.999$,生成器初始学习率为 0.000 1,判别器初始学习率为 0.000 4。其次选择 L1 损失、峰值信噪比(peak signal to noise ratio,PSNR)和结构相似性(structural similarity index measure,SSIM)作为图像修复任务的客观评估指标,其中,L1 损失越低代表性能越好,PSNR 和 SSIM 指标越高代表性能越好。最后,通过消融实验分析模型中各个模块对提升修复效果的贡献。

2.3 实验结果与分析

将提出方法与近年来的图像修复主流算法做对比分析,即:PC[9]、GC[11]、RFR[8]、AOT-GAN[14]、ICT[17] 和 LaMa[15]。不同方法之间的计算性能,本文选择参数数量和推理时间作为评估指标,其中推理时间是模型处理单张图片所需的时间,如表 1 所示。从结果来看,在与不同方法进行对比时,本文方法凭借在模型参数数量和推理时间之间达成的有效平衡,展示出了良好的的综合性能。

表 1 不同方法之间参数与推理时间对比

Table 1 Comparison of parameters and inference time between methods

模型	参数×10^6	推理时间/ms
PC	49	22
GC	4	10
RFR	30	20
AOT-GAN	15	19
ICT	150	113
LaMa	51	32
本文方法	29	25

本文方法与其他方法在 CelebA-HQ 和 Places2 数据集上的定性比较结果如图 4 和 5 所示。可以观察到:PC 能恢复图像基本的形状与结构,然而生成的图像往往伴随着较为明显的伪影。例如,在图 4(c)第 3 张图中,修复出来的人脸面部特征有些模糊不清,缺乏清晰度。GC 在基本语义填充方面表现尚可,但生成的图像在纹理清晰度上有

所欠缺。例如，在图 4(d)的 5 张图中红框标注的区域，眼睛、鼻子、嘴和牙齿部分的修复效果显得不够自然。RFR 大部分能够生成结构合理的图像，但在纹理清晰度方面仍有提升空间。在图 4(e)第 2 张图和第 5 张图中，在处理嘴唇细节上显得较为粗糙，缺乏真实感。AOT-GAN 在处理大面积缺失时能生成语义合理的图像，但在图像局部细节修复方面不是很好。例如，图 5(f)第 2 张图的红框标注区域，其结构也修复得较为杂乱。ICT 在处理大面积损坏时能够生成较为合理的结构，但在细节处理上仍有不足。在图 5(g)的第 1 张图中未能成功修复柜子的结构和细节纹理。LaMa 能够生成语义正确且相对自然的修复结果，但在细节纹理的生成上仍有改进空间，在图 5(h)的 5 张图中方框标注的区域，它对纹理和结构的修复效果也稍显不足。

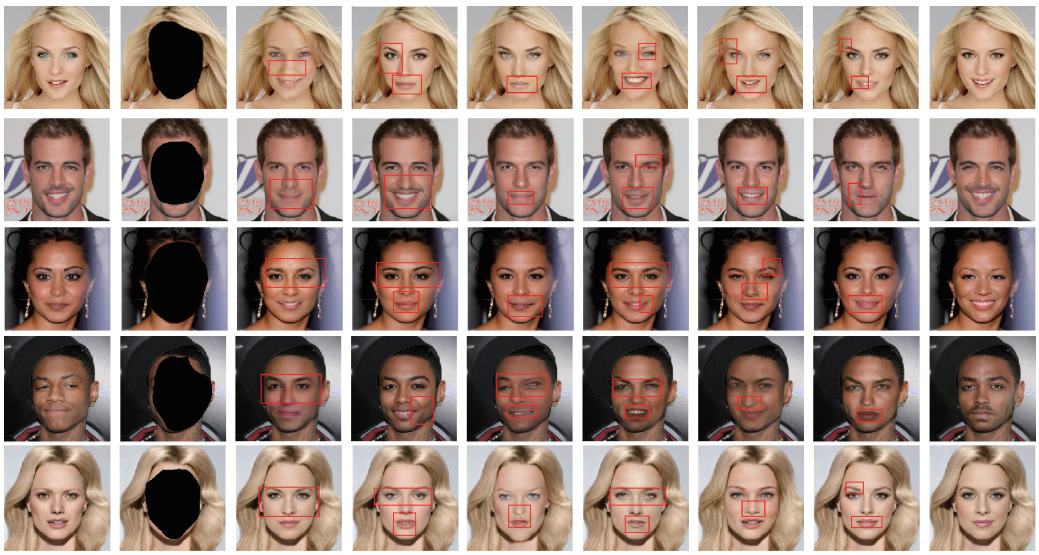


图 4　不同方法 CelebA-HQ 数据集上的定性对比

Fig. 4　Qualitative effect comparison of different methods on CelebA-HQ dataset

图 5　不同方法 Places2 数据集上的定性对比

Fig. 5　Qualitative effect comparison of different methods on Places2 dataset

与其他方法相比,本文的方法在大多数情况下提供了更为合理和逼真的修复结果,而且在结构和纹理的修复上都展现出较高的准确性和自然性。

为了客观评估本文方法的有效性,对结果图选用 L1、PSNR 和 SSIM 3 个客观指标进行评估,如表 2 所示。在掩码率达到 40%～50%的情况下,相比于其他算法,本文提出的方法在 CelebA-HQ 数据集上 PSNR 提高了 0.31～7.63 dB,SSIM 提升了 0.54%～21%,L1 损失降低了 0.21%～5.65%。在 Places2 数据集上,PSNR 平均提高了 0.21～4.87 dB,SSIM 提高了 2.3%～17%,L1 损失降低了 0.19%～5.68%,证明了本研究方法的有效性。

表 2 不同方法在 CelebA-HQ 和 Places2 数据集上定量对比

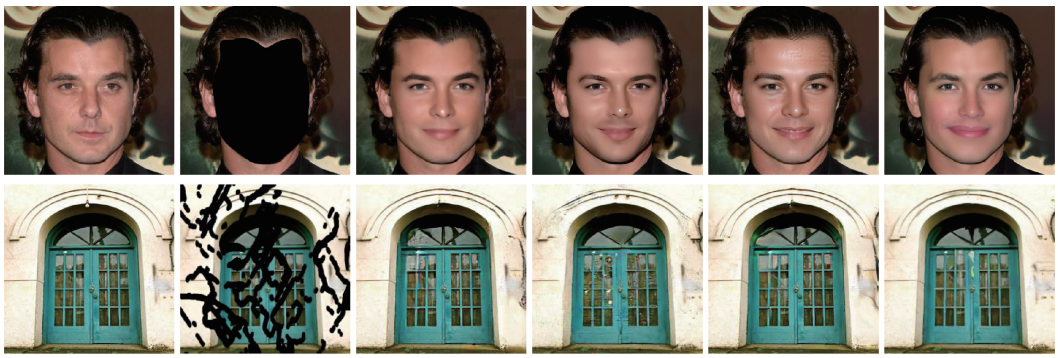
Table 2 Quantitative comparison of different methods on the CelebA-HQ and Places2 datasets

指标	模型	掩码率 10%～20%		掩码率 20%～30%		掩码率 30%～40%		掩码率 40%～50%	
		CelebA-HQ	Places2	CelebA-HQ	Places2	CelebA-HQ	Places2	CelebA-HQ	Places2
L1 ↓	PC	2.13	2.86	3.91	5.10	6.00	7.54	8.48	10.41
	GC	0.84	1.88	1.50	3.36	2.26	5.05	3.16	7.12
	RFR	0.90	1.92	1.83	2.83	**1.99**	3.90	3.16	5.18
	AOT-GAN	0.86	1.42	1.38	2.69	2.63	3.91	3.35	5.15
	ICT	0.89	1.42	1.79	2.66	2.63	3.72	3.12	5.21
	LaMa	0.81	1.46	1.56	2.53	2.31	3.61	3.04	4.92
	本研究	**0.75**	**1.32**	**1.37**	**2.27**	2.33	**3.46**	**2.83**	**4.73**
PSNR ↑	PC	25.50	23.16	22.21	20.19	19.78	18.23	17.87	16.65
	GC	32.28	26.11	29.12	23.11	26.93	20.92	24.80	19.20
	RFR	30.93	26.05	27.31	24.07	**27.11**	22.58	24.78	20.56
	AOT-GAN	30.15	27.20	28.94	24.53	26.72	22.75	24.67	20.53
	ICT	31.19	27.36	28.52	24.64	26.62	23.13	24.88	20.72
	LaMa	32.77	27.44	29.23	24.83	26.89	22.84	25.19	21.31
	本研究	**33.51**	**28.16**	**29.87**	**25.45**	26.97	**23.38**	**25.50**	**21.52**
SSIM ↑	PC	0.925	0.880	0.860	0.761	0.785	0.700	0.699	0.600
	GC	0.979	0.928	0.959	0.864	0.931	0.779	0.896	0.681
	RFR	0.969	0.925	0.939	0.879	**0.938**	0.820	0.893	0.720
	AOT-GAN	0.968	0.932	0.958	0.886	0.923	0.826	0.884	0.724
	ICT	0.969	0.930	0.944	0.888	0.923	**0.852**	0.899	0.733
	LaMa	0.976	0.932	0.950	0.901	0.926	0.832	0.905	0.753
	本研究	**0.980**	**0.943**	**0.964**	**0.914**	0.930	**0.852**	**0.910**	**0.776**

2.4 消融实验

为了验证所提出的 MHFFB 与 SGT 模块的有效性,本文在掩码率为 40%～50%的数据集中进行消融实验。模型 1 是 Baseline,它由门控卷积、普通的膨胀卷积和普通的 Transformer 构成的;在模型 1 基础上添加 MHFFB 作为模型 2;在模型 1 基础上添加 SGT 作为模型 3,实验结果如图 6 所示。图 6(a)为原图,图 6(b)为输入,图 6(c)、(d)、(e)、(f)分别为模型 1、模型 2、模型 3、本文方法。观察结果:模型 1 修复的图像纹理不够清晰且部分结构没有得到完整地修复;模型 2 在加入 MHFFB 模块后,其中的伪影有明显减少,可以看出在第一阶段使用 MHFFB 可以使得模型生成有丰富细节特征的图像,但它对图像的整体结构恢复得不是很完整;模型 3 在加入 SGT 模块后,修复的图像内容变得更加真实和合理,尽管如此,部分边缘和纹理的处理仍有待改进;最后,当所有模块都加入后,生成的图像在细节表现和结构连贯性上都有了显著改善,纹理和边缘的处理也更为精确,整体修复效果更为自然和真实。

消融实验的结果在客观指标上的表现如表 3 所示。模型 2 在加入 MHFFB 后,PSNR 和 SSIM 分别提升了 0.27 dB、1.6%,L1 降低了 0.24%;模型 3 加入 SGT 后 PSNR 和 SSIM 分别提升了 0.37 dB、1.9%,L1 降低了 0.27%;而在本文方法中,PSNR 提升了 0.78 dB,SSIM 提升了 2.6%,L1 降低 0.48%,进一步验证了本文算法的有效性。



(a) 原图 (a) Origin image　(b) 输入 (b) Input　(c) 模型1 (c) Model 1　(d) 模型2 (d) Model 2　(e) 模型3 (e) Model 3　(f) 本文方法 (f) Ours

图 6　消融对比实验

Fig. 6　Comparison results of ablation experiments

表 3　MHFFB 和 SGT 的消融实验

Table 3　Ablation experiments with MHFFB and SGT

模型	PSNR/dB	SSIM	L1
模型 1	24.72	0.884	3.31
模型 2	24.99	0.900	3.07
模型 3	25.09	0.903	2.95
本文方法	25.50	0.910	2.83

3　结　　论

本研究提出了一个融合多尺度分层特征和全局-局部协同 Transformer 的图像修复模型。多尺度分层特征融合模块的引入，不仅扩大感受野范围，还促进了网络不同层次的特征融合，有效提取图像的关键信息使得生成的图像细节丰富，伪影减少。另一方面，通过使用全局-局部协同 Transformer 模块，在降低了计算的复杂度同时，有效地增强了网络对图像全局信息和局部细节信息的捕获能力，提高了修复的质量。通过定量和定性的实验分析，与现有的修复方法相比，本研究的方法不仅在定量指标上取得了较好的结果，在视觉效果方面，修复后的图像也展现了更清晰的纹理以及更合理的结构，极大地提升了图像的真实感和协调性。不过，该模型在生成多样化的图像方面存在一定局限，对于相同的缺失区域，模型倾向于产生相似的修复结果，缺乏多样性输出。未来的工作将着重于引入动态损失函数或是结合多模态图像修复方法，以此增强模型的灵活性，丰富输出结果的多样性。

参考文献

[1] 童俊毅，张银胜，张培琰，等. 基于双阶段多尺度生成对抗网络的图像复原方法[J]. 国外电子测量技术，2024，43(6)：50-58.
TONG J Y，ZHANG Y SH，ZHANG P Y，et al. Image restorationmethod based on two-stage multiscale generative adversarial network[J]. Foreign Electronic Measurement Technology，2024，43(6)：50-58.

[2] JO Y，PARK J. SC-FEGAN：Face editing generative adversarial network with user′s sketch and color[C]. IEEE/CVF International Conference on Computer Vision (ICCV)，2019：1745-1753.

[3] YI Z L，TANG Q，AZIZI S，et al. Contextual residual aggregation for ultra high-resolution image inpainting[C]. IEEE/CVF Conferenceon Computer Vision and Pattern Recognition，2020：7508-7517.

[4] 张双，杨帆. 改进的双阶段生成对抗数字壁画修复算法[J]. 电子测量技术，2023，46(11)：123-129.
ZHANG SH，YANG F. Digital mural inpainting model based on improved two-stage generative adversarial network [J]. Electronic Measurement Technology，2023，46(11)：123-129.

[5] GOODFELLOW I，POUGET A J，MIRZA M，et al. Generative adversarial networks[J]. Communications of the ACM，2020，63(11)：139-144.

[6] PATHAK D，KRAHENBUHL P，DONAHUE J，et al. Context encoders：Feature learning by inpainting[C]. IEEE Conference on Computer Vision and Pattern Recognition，2016：2536-2544.

[7] YU J H，LIN ZH，YANG J M，et al. Generative image inpainting with contextual attention[C]. IEEE Conferenceon Computer Vision and Pattern Recognition，2018：5505-5514.

[8] LI J Y，WANG N，ZHANG L F，et al. Recurrent feature reasoning for image inpainting[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition，2020：7760-7768.

[9] LIU G L，REDA F A，SHIH K J，et al. Image inpaintingfor irregular holes using partial convolutions[C]. European Conference on Computer Vision (ECCV)，

2018: 85-100.

[10] ZENG Y, LIN ZH, LU H C, et al. CR-Fill: Generative image inpainting with auxiliary contextual reconstruction [C]. IEEE/CVF International Conference on Computer Vision, Montreal, 2021: 14164-14173.

[11] YU J H, LIN ZH, YANG J M, et al. Free-form image inpainting with gated convolution [C]. IEEE/CVF International Conference on Computer Vision, 2019: 4471-4480.

[12] 高杰, 霍智勇. 一种门控卷积生成对抗网络的图像修复算法[J]. 西安电子科技大学学报, 2022, 49(1): 216-224.
GAO J, HUO ZH Y. A lgorithm for image inpainting in generative adversarial networks based on gated convolution[J]. Journal of Xidian University, 2022, 49(1): 216-224.

[13] 陈婷, 王通, 张冀武, 等. 基于边缘先验融合动态门控特征的人脸图像修复[J]. 计算机应用研究, 2023, 40(11): 3478-3484.
CHEN T, WANG T, ZHANG J W, et al. Face image inpainting algorithm based on edge prior fusion dynamic gating features[J]. Application Research of Computers, 2023, 40(11): 3478-3484.

[14] ZENG Y H, FU J L, CHAO H Y, et al. Aggregated contextual transformations for high-resolution image inpainting [J]. IEEE Trans on Visualization and Computer Graphics, 2022, 29(7): 3266-3280.

[15] SUVOROV R, LOGACHEVA E, MASHIKHIN A, et al. Resolution-robust large mask inpainting with fourier convolutions [C]. IEEE/CVF Winter Conference on Applications of Computer Vision, 2022: 2149-2159.

[16] 刘华咏, 黄聪, 金汉均. 注意力增强的视觉 Transformer 图像检索算法[J]. 电子测量技术, 2023, 46(23): 50-55.
LIU H Y, HUANG C, JIN H J. Image retrieval method with attention-enhanced visual transformer. [J]. Electronic Measurement Technology. 2023, 46(23): 50-55.

[17] WAN Z Y, ZHANG J B, CHEN D D, et al. High-fidelity pluralistic image completion with transformers[C]. IEEE/CVF International Conference on Computer Vision, 2021: 4692-4701.

[18] ZHENG CH X, CHAM T J, CAI J F, et al. Bridging global context interactions for high-fidelity image completion[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 11512-11522.

[19] 向泽林, 楼旭东, 李旭伟. 基于 Swin Transformer 和 Style-based Generator 的盲人脸修复[J]. 四川大学学报(自然科学版), 2023, 60(3): 65-73.
XIANG Z L, LOU X D, LI X W. Blind face restoration based on swin transformer and style-based generator[J]. Journal of Sichuan University (Natural Science Edition), 2023, 60(3): 65-73.

[20] 杨红菊, 高敏, 张常有, 等. 一种面向图像修复的局部优化生成模型[J]. 图学学报, 2023, 44(5): 955-965.
YANG H J, GAO M, ZHANG CH Y, et al. A local optimization generation model for image inpainting[J]. Journal of Graphics, 2023, 44(5): 955-965.

[21] LIU J L, GONG M G, GAO Y, et al. Bidirectional interacton of cnn and transformer for image inpainting[J]. Knowledge Based Systems, 2024, 299: 112046.

[22] 徐志刚, 杨欣宇. 结合 CSWin-Transformer 和门卷积的壁画图像修复方法[J]. 计算机工程与应用, 2024, 60(21): 215-224.
XU ZH G, YANG X Y. Mural image restoration method based on cswin-transformer and gate convolution [J]. Computer Engineering and Applications, 2024, 60(21): 215-224.

[23] CHEN ZH, ZHANG Y L, GU J J, et al. Cross aggregation transformer for image restoration [J]. Advances in Neural Information Processing Systems, 2022, 35: 25478-25490.

[24] HUA W ZH, DAI Z H, LIU H X, et al. Transformer quality in linear time[C]. International Conference Onmachine Learning, PMLR, 2022: 9099-9117.

[25] KARRAS T, AILA T, LAINE S, et al. Progressive growing of gans for improved quality, stability, and variation[J]. ArXiv preprint arXiv: 1710.10196, 2017.

[26] ZHOU B L, LAPEDRIZA A, KHOSLA A, et al. Places: A 10 million image database for scene recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 40(6): 1452-1464.

作者简介

滕诗宇,硕士研究生,主要研究方向为计算机视觉和图像处理图像修复。

E-mail:283673217@qq.com

何丽君(通信作者),硕士,副教授,硕士生导师,主要研究方向为计算机图形学、数字图像处理和数据处理。

E-mail:51932929@qq.com