内外特征交互与融合的双流注意力图像修复方法

黄光远[1],黄 荣[1,2*],周树波[1,2],蒋学芹[1,2]
(1. 东华大学信息科学与技术学院,上海 201620;2. 东华大学数字化纺织服装技术教育部工程研究中心,上海 201620)

摘 要: 注意力机制及其变体已广泛应用于基于深度学习的图像修复领域,它们将破损图像内部分为完好区域和缺失区域,捕获完好区域的远距离上下文信息以填充缺失区域. 随着缺失区域增大,完好区域特征减少,限制了注意力机制的性能,从而导致修复效果不佳. 为拓展注意力机制捕获上下文的范围,本文通过矢量量化码本学习视觉原子. 这些视觉原子刻画了图像块的结构、纹理等特征,组成用于图像修复的外部特征,以弥补图像内部完好区域特征的不足. 在此基础上,本文提出一种内外特征交互与融合的双流注意力图像修复方法. 该方法结合内部和外部两个信息源,设计了内部掩码注意力和内外交叉注意力,组成双流注意力以实现内部特征之间以及内部和外部特征之间的交互,生成内外源修复特征. 内部掩码注意力通过掩码屏蔽缺失区域特征的干扰,仅在完好区域捕获上下文信息,生成内源修复特征. 内外交叉注意力通过计算内部特征与由视觉原子组成的外部特征之间的相似度关系,实现内外特征之间的交互,生成外源修复特征. 此外,本文设计了可控特征融合模块,利用内外源修复特征之间的相关性生成空间权重图,为每个空间位置精确地筛选内外源修复特征,从而实现内部与外部特征的融合. 在 Places2、FFHQ 和 Paris StreetView 三个公开的数据集上的实验结果表明本文方法在 PSNR、SSIM、$L1$、LPIPS 和 FID 指标上比其他先进方法平均提高了 3.45%、1.34%、13.91%、13.64% 和 16.92%. 消融实验结果和可视化实验结果表明图像内部特征与由视觉原子组成的外部特征均有益于修复破损图像.

关键词: 图像修复;矢量量化码本;视觉原子;掩码注意力;交叉注意力;特征融合
基金项目: 国家自然科学基金(No.62001099);中央高校基本科研业务费专项资金(No.2232023D-30)
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112(2025)04-1293-15
电子学报 URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20240780

Dual-Stream Attention Image Inpainting Method Based on Interacting and Fusing Internal-External Features

HUANG Guang-yuan[1], HUANG Rong[1,2*], ZHOU Shu-bo[1,2], JIANG Xue-qin[1,2]
(1. *College of Information Science and Technology*, *Donghua University*, *Shanghai* 201620, *China*;
2. *Engineering Research Center of Digitized Textile & Apparel Technology*, *Ministry of Education*, *Donghua University*, *Shanghai* 201620, *China*)

Abstract: The attention mechanism and its variants have been widely applied in the field of image inpainting. They divide corrupted images into complete and missing regions, and capture long-range contextual information only within the complete regions to fill in the missing regions. As the area of missing regions increases, the features of complete regions decrease, which limits the performance of the attention mechanisms and leads to suboptimal inpainting results. In order to extend the context range of the attention mechanism, we employ a vector-quantized codebook to learn visual atoms. These visual atoms, which describe the structural and textural of image patches, constitute external features for image inpainting and thus compensate for the internal features of the image. On this basis, we propose a dual-stream attention image inpainting method based on interacting and fusing internal-external features. Based on internal and external information sources, we design an internal mask attention module and an internal-external cross attention module. These two attention modules form a dual-stream attention to facilitate interaction within internal features and between internal and external features, thereby generating internal and external source inpainting features. The internal mask attention shields the interference of missing region features with a mask. It captures contextual information exclusively within the complete regions, thereby generating in-

收稿日期:2024-08-28;修回日期:2025-01-09;责任编辑:朱梅玉
*通讯作者:黄荣

ternal-source inpainting features. The internal-external cross attention interacts with internal and external features by calculating the similarity relationship between internal features and external features composed of visual atoms, thereby generating external-source inpainting features. In addition, we design a controllable feature fusion module that generates spatial weight maps based on the correlation between internal and external source inpainting features. These spatial weight maps fuse internal and external features by element-wise weighting of internal and external source inpainting features. Extensive experimental results on Places2, FFHQ and Paris StreetView datasets demonstrate that the proposed method achieves average improvements of 3.45%, 1.34%, 13.91%, 13.64%, and 16.92% for PSNR, SSIM, $L1$, LPIPS, and FID metrics respectively, compared with the state-of-the-art methods. Visualization experimental results demonstrate that both internal features and external features composed of visual atoms are beneficial for repairing corrupted images.

Key words: image inpainting; vector-quantized codebook; visual atoms; masked attention; cross attention; feature fusion

Foundation Item(s): National Natural Science Foundation of China (No.62001099); Fundamental Research Funds for the Central Universities (No.2232023D-30)

1 引言

图像修复(image inpainting)是指利用破损图像完好区域信息推断并补全缺失区域的过程,也称为图像补全(image completion). 图像修复属于不适定(ill-posed problem)任务[1,2],其难点在于如何为缺失区域生成结构完整、语义一致、纹理逼真的内容. 目前,图像修复已成为计算机视觉领域中重要的研究方向,广泛应用于照片编辑[3]、文物修复[4,5]以及医学成像[6]等领域.

现有的图像修复方法可分为传统方法和基于深度学习的方法. 传统图像修复方法通常基于偏微分方程或图像块匹配机制填充缺失区域. 前者将完好区域的像素沿等照度线的方向传播到缺失区域内,再经过各向异性扩散,得到修复结果[7]. 后者通常基于马尔可夫随机场[8]、低秩矩阵近似[9]等从完好区域搜索最相似的图像块以合成缺失区域的内容. 这些传统方法机械地复制、搬运像素或图像块,缺乏语义推导的能力. 因此当破损图像的缺失区域较大或较复杂时,传统方法难以生成语义合理的内容,导致修复内容失真.

以卷积神经网络(Convolutional Neural Network, CNN)[10]和生成对抗网络(Generative Adversarial Network, GAN)[11]为代表的深度学习技术的兴起,推动了图像修复的发展. 基于深度学习的修复方法[1~5,12~29]从大规模数据中学习图像语义信息,能够为缺失区域生成结构完整、语义一致、纹理逼真的内容. 上下文编码器(context encoders)[12]是首个基于CNN的图像修复方法. 为了捕获完好区域的远距离上下文信息,后续研究在上下文编码器的基础上相继提出了上下文注意力(contextual attention)[25]、多尺度上下文注意力[29]、多模态注意力[20]和连贯语义注意力[26]等. 这些注意力机制及其变体将破损图像特征分为完好与缺失区域,并按照注意力分数加权完好区域的特征块填充缺失区域. 上述注意力机制捕获上下文信息的范围仅局限于图像内部特征,信息源单一. 因此,上述注意力机制的性能很大程度上取决于缺失区域的大小. 当缺失区域较大时,图像内部完好区域特征减少,可捕获的上下文信息范围收缩,注意力机制的性能下降,从而导致修复效果不佳.

为了突破现有注意力机制信息源单一的局限性,本文以VQ-GAN(Vector Quantized-Generative Adversarial Networks)[30]为基本框架,利用矢量量化码本学习视觉原子(visual atoms)[31]. 这些视觉原子组成外部特征,作为修复破损图像的外部信息源,弥补了内部特征的不足. 在此基础上,本文提出一种内外特征交互与融合的双流注意力图像修复方法. 该方法包含内部和外部两个信息源,设计了内部掩码注意力(Internal Masked Attention, IMA)和内外交叉注意力(Internal-External Cross Attention, IECA)模块,组成双流注意力以实现内部特征之间以及内部与外部特征之间的交互,从而拓展了注意力机制捕获上下文的范围. 具体地,IMA通过掩码屏蔽缺失区域特征块的干扰,仅在完好区域捕获上下文信息,生成内源修复特征. IECA根据内部与外部特征之间的相似度关系,实现内部与外部特征之间的信息交互,生成外源修复特征. 此外,本文设计了可控特征融合(Controllable Feature Fusion, CFF)模块,利用内源和外源修复特征之间的相关性生成空间权重图,自适应地对每个空间位置进行特征筛选,实现内部与外部特征融合. 本文的贡献在于两个方面:

(1)通过矢量量化码本所学习的视觉原子扩充了图像修复的信息来源,拓展了注意力机制捕获上下文的范围,解决了信息源单一的问题.

(2)提出了内外特征交互与融合的双流注意力,形成了一套内部特征和外部特征"互为补充、相辅相成、共同修复"的图像修复新框架.

2 相关工作

传统的图像修复方法通常可分为两类:基于偏微分方程的方法[7,32,33]和基于图像块匹配的方法[8,9,34].

Bertalmio等人[7]基于缺失区域边界的完好像素,估计出等照度线的方向,然后沿该方向传播完好像素以填充缺失区域. 在此基础上,Shen等人[32]提出了全变分(total variational)模型,以加快像素扩散的收敛速度. Criminisi算法[34]根据缺失区域边界等照度线与法线的角度,确定待填充像素的优先级并在完好区域搜索相似的图像块进行填充. Ružić等人[8]根据分割算法自上而下将图像分成大小可变的图像块,并将图像块的搜索范围限制在经匹配筛选的上下文区域,以提高修复的速度. 这些传统方法只是机械地复制、搬运像素或图像块,语义推导能力差,因而仅适用于修复缺失区域面积较小或纹理、结构简单的破损图像. 当破损图像的缺失区域面积较大时,修复结果往往存在纹理模糊和语义不一致等问题.

近年来,基于深度学习的图像修复方法[1~5,12~29]从大规模数据中学习图像的语义信息,从而生成结构完整、语义一致、纹理逼真的修复结果. Pathak等[12]提出了基于CNN的上下文编码器(context encoders). 它采用编码器-解码器的结构,通过最大限度地降低重构损失和对抗损失[11],为缺失区域生成语义合理的内容. 在此基础上,Iizuka等[13]引入了空洞卷积[35]以增大卷积核的感受野,并通过全局和局部双鉴别器来保证修复结果的整体语义连贯性和局部细节真实性. 为了排除缺失像素的干扰,Liu等[22]提出了一种基于部分卷积的图像修复方法,通过一个自动更新的二值掩码保留特征图中完好区域特征块. 在此基础上,Yu等[27]提出了可学习软掩码的门控卷积. 与二值掩码不同,软掩码的取值介于0到1之间,规避了二值掩码在网络深层的失效问题.

受制于卷积核的局部感受野,CNN难以捕获远距离上下文信息,导致生成的内容与周围区域结构不一致. Yu等[25]提出了上下文注意力(contextual attention)机制,从完好区域搜索与缺失区域相似的特征块,并通过特征块加权融合实现远距离上下文信息的捕获. Liu等[26]提出了连贯语义注意力,额外考虑缺失区域内部相邻特征块之间的相关性,以确保修复结果在语义上的连贯性. Zeng等[1]提出了注意力转移网络,将金字塔结构中高层特征学习到的注意力分数复制给低层特征图,从而实现注意力跨尺度的转移. Li等[23]提出了掩码注意力机制,在Swin transformer[36]窗口注意力的基础上,引入二值掩码排除缺失区域特征块的干扰. Deng等[21]提出了一种跨尺度上下文注意力机制,通过两个并行的分支分别捕获结构和纹理特征. 然而,上述注意力机制捕获上下文的范围仅局限于破损图像本身,依赖单一的信息源. 因此,上述注意力机制的性能受到缺失区域大小的影响. 当缺失区域扩大时,图像内部完好区域特征减少,注意力机制可捕获的上下文信息范围收缩,致使性能下降,从而导致修复效果不佳.

矢量量化作为一种数据压缩技术已应用于图像恢复领域. 矢量量化以近似原始数据的分布为目标,通过码本的学习将图像的连续特征量化为视觉原子的离散化索引,从而实现数据压缩. VQ-VAE(Vector Quantized-Variational Auto Encoder)[37]将矢量量化技术引入基于自编码器的生成模型中,通过对码本中视觉原子的分布自回归建模,从而缓解“模式坍塌”(mode collapse)问题. VQ-GAN[30]进一步采用对抗损失[11]和感知损失[38]增强监督信号,不但增大了压缩率还提高了视觉原子的感知质量. 在图像恢复领域,相关研究[31,39~41]通过在高质量图像上训练矢量量化码本,为恢复低质量图像提供细节特征. 与本文相似,Liu等[24]将矢量量化码本引入至图像修复领域. 他们将图像修复任务视为码本中视觉原子索引的回归任务. 然而,该方法仅依靠单一的外部特征进行图像修复,直接忽略了内部完好区域特征. 相比于外部特征,图像内部的完好区域特征与缺失区域特征来源于同一张图像,往往能提供与缺失区域相似的结构和纹理信息. 与该方法[24]不同,本文方法同时考虑了内部和外部特征,并通过双流注意力和可控特征融合进行特征的交互与融合,实现了内外部特征的互为补充、相辅相成,共同对缺失区域进行修复.

3　方法设计

为解决现有注意力机制信息源单一的问题,本文提出了一种内外特征交互与融合的双流注意力图像修复方法,整体框架分为预训练阶段和修复阶段,如图1所示. 在预训练阶段,矢量量化码本中的向量通过吸收图像特征实现更新,最终形成能够刻画图像块的结构、纹理等特征的视觉原子. 这些视觉原子为后续修复阶段提供外部特征,弥补了内部特征的不足. 此外,矢量量化码本的训练与特定数据集解耦,即码本可在任意的通用数据集上进行训练,有利于提高修复模型面对多样化场景的适应性和鲁棒性. 修复阶段通过内部掩码注意力和内外交叉注意力组成双流注意力,实现内部特征之间以及内部与外部特征之间的交互,分别生成内源和外源修复特征. 可控特征融合模块根据内源和外源修复特征之间的相关性矩阵生成空间权重图,为每个空间位置自适应地筛选内源和外源修复特征,从而实现内部与外部特征的融合. IMA、IECA和CFF组成一个复合模块,共级联T次.

3.1　视觉原子学习

预训练阶段旨在通过训练矢量量化码本学习视觉原子,为修复破损图像提供丰富的外部特征. 如图1的

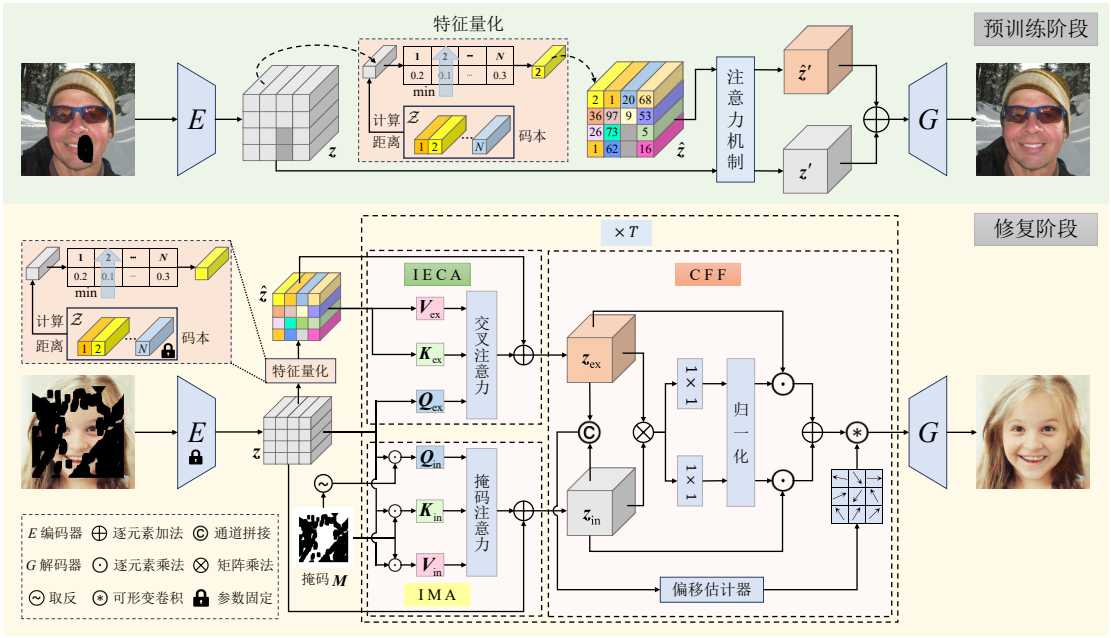


图1 内外特征交互与融合的双流注意力整体框图

上半栏所示，预训练阶段的网络模型由码本 $\mathcal{Z}$、注意力机制、编码器 E 和解码器 G 组成. 码本 $\mathcal{Z}=\{\boldsymbol{z}_n\}(n=1,2,\cdots,N)\in\mathbb{R}^{N\times c}$ 由 N 个可学习的 c 维向量 $\boldsymbol{z}_n$ 构成. 在训练过程中，编码器 E 学习图像特征的提取，$\boldsymbol{z}_n$ 吸收编码器 E 提取的图像特征并进行更新，最终形成视觉原子. 经训练后，$\boldsymbol{z}_n$ 能够刻画图像块的结构、纹理，对应于稀疏的视觉原子. 具体地，随机初始化码本 $\mathcal{Z}$；将编码器 E 提取的特征图记为 $\boldsymbol{z}\in\mathbb{R}^{h\times w\times c}$，其中 h、w 和 c 分别为特征图的高、宽和通道数. 根据二值的缺损掩码 $\boldsymbol{M}$（1 为完好区域，0 为缺失区域），将特征图 $\boldsymbol{z}$ 划分为完好区域和缺失区域. 为了排除缺失区域特征块对修复结果的干扰，预训练阶段仅对完好区域的特征块进行量化，选取码本中最相近的向量 $\boldsymbol{z}_n$ 替换 $\boldsymbol{z}$，缺失区域则保持不变. 特征图 $\hat{\boldsymbol{z}}$ 可由下式得到：

$$\hat{\boldsymbol{z}}_i=\begin{cases}\arg\min\limits_{\boldsymbol{z}_n\in\mathcal{Z}}\|\boldsymbol{z}_i-\boldsymbol{z}_n\|_2^2, & \boldsymbol{M}_i=1\\ \boldsymbol{z}_i, & \boldsymbol{M}_i=0\end{cases}\tag{1}$$

其中，i 为 $\boldsymbol{M}$ 每个空间位置特征块的索引. 预训练阶段采用注意力机制[25]分别填充特征图 $\boldsymbol{z}$ 和 $\hat{\boldsymbol{z}}$ 的缺失区域. 对于图1中注意力机制下输入分支，特征图 $\boldsymbol{z}$ 完好区域的特征块 $\boldsymbol{z}_i$ 与缺失区域特征块 $\boldsymbol{z}_j$ 之间的相似度分数通过下式计算得到：

$$s_{ij}=\cos\left(\frac{\boldsymbol{z}_i}{\|\boldsymbol{z}_i\|},\frac{\boldsymbol{z}_j}{\|\boldsymbol{z}_j\|}\right)\tag{2}$$

其中，$\cos(\cdot,\cdot)$ 为余弦相似度，$i\in\{1,2,\cdots,I\}$，$j\in\{1,2,\cdots,J\}$. 注意力分数可表示为：$a_{ij}=\exp(s_{ij})\Big/\sum_i^I\exp(s_{ij})$. 按照 $\sum_i^I a_{ij}\boldsymbol{z}_i, j\in\{1,2,\cdots,J\}$ 更新缺失区域的特征块，从而生成特征图 $\boldsymbol{z}'$. 同理，注意力机制上输入分支通过加权 $\hat{\boldsymbol{z}}$ 完好区域特征块填充缺失区域，生成特征图 $\hat{\boldsymbol{z}}'$. 最后，将 $\boldsymbol{z}'$ 和 $\hat{\boldsymbol{z}}'$ 逐元素相加后送入解码器 G 生成预训练阶段的修复结果.

本文采用量化损失 L_{vq} 指导码本中视觉原子的学习. 由于式(1)的量化过程不可微，本文采用直通梯度估计器(straight-through gradient estimator)[30,37]将梯度从解码器 G 复制到编码器 E，L_{vq} 如下式所示：

$$L_{\mathrm{vq}}=\left\|\mathrm{sg}[\boldsymbol{z}]-\hat{\boldsymbol{z}}\right\|_2^2+\eta\left\|\boldsymbol{z}-\mathrm{sg}[\hat{\boldsymbol{z}}]\right\|_2^2\tag{3}$$

其中，$\mathrm{sg}[\cdot]$ 表示停止梯度，即不计算也不反传梯度. 量化损失的第一项用于优化码本，第二项用于优化编码器. 为促进码本训练，本文将 η 设为 0.25[24,31,37,39].

预训练结束后，将编码器 E 的参数与矢量量化码本 $\mathcal{Z}$ 固定；码本中的向量成为能够刻画图像块结构、纹理等特征的视觉原子. 这些视觉原子为后续修复阶段提供外部特征，形成对于图像内部特征的补充.

3.2 内部特征交互

完好区域特征与缺失区域特征来源于同一张图像，往往能够提供与缺失区域相似的结构和纹理信息. 因此，本文设计了内部掩码注意力模块，通过挖掘图像内部完好区域的特征，与缺失区域的特征进行交互，从而利用完好区域特征修复破损图像. IMA 借助二值掩码 $\boldsymbol{M}$ 将特征图 $\boldsymbol{z}$ 分离成缺失区域和完好区域. 为了实现缺失区域与完好区域之间的特征交互，IMA 分别将缺失区域映射为查询向量 $\boldsymbol{Q}_{\mathrm{in}}$，将完好区域映射为键向

量$\boldsymbol{K}_{\text{in}}$和值向量$\boldsymbol{V}_{\text{in}}$,如式(4)所示:

$$\begin{aligned}\boldsymbol{Q}_{\text{in}}&=\left[z\odot(\boldsymbol{1}-\boldsymbol{M})\right]\boldsymbol{W}_{\text{Q}}\\ \boldsymbol{K}_{\text{in}}&=\left[z\odot\boldsymbol{M}\right]\boldsymbol{W}_{\text{K}}\\ \boldsymbol{V}_{\text{in}}&=\left[z\odot\boldsymbol{M}\right]\boldsymbol{W}_{\text{V}}\end{aligned}\tag{4}$$

其中,$\boldsymbol{Q}_{\text{in}},\boldsymbol{K}_{\text{in}},\boldsymbol{V}_{\text{in}}\in\mathbb{R}^{h\times w\times c}$,$\boldsymbol{W}_{\text{Q/K/V}}\in\mathbb{R}^{c\times c}$为可学习的参数,$\odot$为逐元素相乘.$z\odot\boldsymbol{M}$表示完好区域,$z\odot(\boldsymbol{1}-\boldsymbol{M})$表示缺失区域,$\boldsymbol{1}$是值全为1的矩阵.IMA通过查询向量$\boldsymbol{Q}_{\text{in}}$与键向量$\boldsymbol{K}_{\text{in}}$之间的点积运算,实现缺失区域与完好区域之间特征交互,进而根据注意力分数加权值向量$\boldsymbol{V}_{\text{in}}$以填充缺失区域.最终,IMA生成的内源修复特征图$z_{\text{in}}$可表示为

$$z_{\text{in}}=z+\text{softmax}\left(\frac{\boldsymbol{Q}_{\text{in}}\boldsymbol{K}_{\text{in}}^{\top}}{\sqrt{c}}\oplus\boldsymbol{m}\right)\boldsymbol{V}_{\text{in}}\tag{5}$$

其中,"+"代表残差连接,旨在缓解梯度消失的问题.式(5)中,加号后的一项表示内部完好区域特征修复后的特征图,其中$\oplus$为逐元素相加.为了防止$\boldsymbol{Q}_{\text{in}}$与$\boldsymbol{K}_{\text{in}}$点积结果过大,导致梯度消失,将点积结果与比例系数$\sqrt{c}$相除[42].IMA在计算注意力分数时,通过掩码$\boldsymbol{m}$屏蔽缺失区域特征块,仅在完好区域捕获上下文信息.掩码$\boldsymbol{m}$的表达式如下:

$$\boldsymbol{m}_i=\begin{cases}0, & \boldsymbol{M}_i=1\\ -\infty, & \boldsymbol{M}_i=0\end{cases}\tag{6}$$

其中,i为二值缺损掩码$\boldsymbol{M}$每个空间位置特征块的索引.经过softmax归一化后,缺失区域特征块的注意力分数接近于0,起到了掩码屏蔽的作用.

3.3　内部与外部特征交互

本文设计了内外交叉注意力模块,通过内部特征与外部特征之间的交互,拓展了注意力捕获上下文的范围.在修复阶段,按照式(1)对内部特征图z的所有特征块进行量化,得到外部特征图$\hat{z}$.与IMA的单一输入不同,IECA的输入包括编码器E提取的内部特征图z和量化后的外部特征图$\hat{z}$.具体地,IECA将内部特征图z映射为查询向量$\boldsymbol{Q}_{\text{ex}}$,将外部特征图$\hat{z}$映射为键向量$\boldsymbol{K}_{\text{ex}}$和值向量$\boldsymbol{V}_{\text{ex}}$:

$$\boldsymbol{Q}_{\text{ex}}=z\boldsymbol{W}_{\text{Q}}',\ \boldsymbol{K}_{\text{ex}}=\hat{z}\boldsymbol{W}_{\text{K}}',\ \boldsymbol{V}_{\text{ex}}=\hat{z}\boldsymbol{W}_{\text{V}}'\tag{7}$$

其中,$\boldsymbol{Q}_{\text{ex}},\boldsymbol{K}_{\text{ex}},\boldsymbol{V}_{\text{ex}}\in\mathbb{R}^{h\times w\times c}$,$\boldsymbol{W}_{\text{Q/K/V}}'\in\mathbb{R}^{c\times c}$为可学习的参数.IECA通过内外特征交叉匹配的方式,根据内部特征的语义信息,在外部特征中搜寻语义匹配的特征块,以弥补内部特征的不足.IECA生成的外源修复特征图z_{ex}可表示为

$$z_{\text{ex}}=\hat{z}+\text{softmax}\left(\frac{\boldsymbol{Q}_{\text{ex}}\boldsymbol{K}_{\text{ex}}^{\top}}{\sqrt{c}}\right)\boldsymbol{V}_{\text{ex}}\tag{8}$$

IECA也引入了残差连接以缓解梯度消失的问题.式(8)中,加号后的一项表示IECA利用外部特征修复后的特征图.此外,IECA与IMA将$\boldsymbol{Q}_{\text{ex/in}}$、$\boldsymbol{K}_{\text{ex/in}}$和$\boldsymbol{V}_{\text{ex/in}}$划分为$N_h$个部分,形成多头注意力,旨在从多个角度捕捉缺失区域与完好区域之间的特征关系,提高模型的泛化能力.

3.4　内部与外部特征融合

由于来源不同,内部特征和外部特征在分布上通常具有差异性.为了融合内部与外部特征,本文设计了可控特征融合模块.CFF通过构建内源修复特征图z_{in}和外源修复特征图z_{ex}之间的相关性矩阵,预测每个空间位置的融合权重,自适应地进行特征(z_{in}和z_{ex})筛选,从而实现内部与外部特征的融合.

具体地,CFF先对特征图z_{in}和z_{ex}进行张量尺寸变换,得到$z_{\text{in}}\in\mathbb{R}^{c\times h\times w}$和$z_{\text{ex}}\in\mathbb{R}^{h\times w\times c}$,则$z_{\text{in}}$和$z_{\text{ex}}$之间的相关性矩阵$\boldsymbol{A}$可通过下式计算得到:

$$\boldsymbol{A}=z_{\text{ex}}\otimes z_{\text{in}}\tag{9}$$

其中,$\otimes$代表矩阵乘法.随后,通过1×1卷积对$\boldsymbol{A}$进行通道压缩,并运用归一化操作确保每个空间位置的权重和为1,即

$$[\boldsymbol{w}_{\text{ex}},\boldsymbol{w}_{\text{in}}]=\text{softmax}\left(\text{Concat}\left(\alpha(\boldsymbol{A}),\gamma(\boldsymbol{A})\right)\right)\tag{10}$$

其中,Concat(·)为按通道拼接,α和γ为1×1卷积,$\boldsymbol{w}_{\text{ex}},\boldsymbol{w}_{\text{in}}\in\mathbb{R}^{h\times w}$为空间权重图.筛选后的特征图为$\left(\boldsymbol{w}_{\text{ex}}\odot z_{\text{ex}}\right)\oplus\left(\boldsymbol{w}_{\text{in}}\odot z_{\text{in}}\right)$.随后,CFF通过可形变卷积[43]校正筛选后特征存在的不对齐问题.可形变卷积根据可学习的偏移量$\boldsymbol{Y}$扭曲感受野,以调整筛选后的特征图,从而生成融合后的特征图z_{out},具体过程如式(11)所示:

$$\begin{aligned}\boldsymbol{Y}&=\text{Conv}\left(\text{Concat}(z_{\text{ex}},z_{\text{in}})\right)\\ z_{\text{out}}&=\text{DeformConv}\left[\left(\boldsymbol{w}_{\text{ex}}\odot z_{\text{ex}}\right)\oplus\left(\boldsymbol{w}_{\text{in}}\odot z_{\text{in}}\right),\boldsymbol{Y}\right]\end{aligned}\tag{11}$$

其中,Conv(·)为3×3卷积,DeformConv[·]为可形变卷积.z_{out}为由IMA、IECA和CFF组成的复合模块的输出.

IMA、IECA和CFF组成的复合模块共级联T次.每一级复合模块输出的z_{out},作为下一级复合模块的输入.最后一级复合模块的输出z_{out}作为解码器G的输入,生成修复结果$\boldsymbol{I}_{\text{out}}$.$T$为超参数,本文在4.4节消融实验中对其设置进行探讨.

3.5　损失函数

本文采用联合损失分别在预训练阶段和修复阶段训练模型,包括重构损失、梯度损失、感知损失[38]、风格损失[44]和对抗损失[11].其中,重构损失和梯度损失作用在像素层面,感知损失、风格损失和对抗损失作用在特征层面.

重构损失旨在像素层面最小化修复结果$\boldsymbol{I}_{\text{out}}$与真实图像$\boldsymbol{I}_{\text{gt}}$的差异.本文采用$L1$损失作为重构损失$L_{\text{rec}}$,具体如式(12)所示:

$$L_{\text{rec}}=\left\|\boldsymbol{I}_{\text{out}}-\boldsymbol{I}_{\text{gt}}\right\|_1\tag{12}$$

图像梯度通常反映局部邻域内像素值变化的速

度,其在图像边缘区域呈现出较强响应. 本文引入梯度损失L_{grad},以衡量修复结果$\boldsymbol{I}_{\text{out}}$在边缘处与真实图像$\boldsymbol{I}_{\text{gt}}$的一致性,具体损失如下式所示:

$$L_{\text{grad}}=\frac{1}{2}\left[\left\|\nabla_x\left(\boldsymbol{I}_{\text{out}}\right)-\nabla_x\left(\boldsymbol{I}_{\text{gt}}\right)\right\|_1+\left\|\nabla_y\left(\boldsymbol{I}_{\text{out}}\right)-\nabla_y\left(\boldsymbol{I}_{\text{gt}}\right)\right\|_1\right] \tag{13}$$

重构损失和梯度损失仅在像素层面最小化修复结果$\boldsymbol{I}_{\text{out}}$与真实图像$\boldsymbol{I}_{\text{gt}}$的差异,难以保证二者的语义一致性. 为此,本文引入特征层面的感知损失[38]和风格损失[44],以衡量$\boldsymbol{I}_{\text{out}}$与$\boldsymbol{I}_{\text{gt}}$之间的语义一致性. 本文采用在ImageNet数据集[45]上预训练的VGG-19(Visual Geometry Group-19)网络[46]来提取图像的语义特征. 感知损失L_{perc}定义为

$$L_{\text{perc}}=E\left[\sum_i\left\|\phi_i(\boldsymbol{I}_{\text{out}})-\phi_i(\boldsymbol{I}_{\text{gt}})\right\|_1\right] \tag{14}$$

其中,$\phi_i(\cdot)$为VGG-19网络中第i层的特征图,i为ReLu1_1、ReLu2_1、ReLu3_1、ReLu4_1和ReLu5_1的索引.

本文引入风格损失L_{style},用于缓解解码器中转置卷积导致的棋盘伪影[47]. L_{style}通过计算特征图的格拉姆矩阵来衡量修复结果$\boldsymbol{I}_{\text{out}}$和真实图像$\boldsymbol{I}_{\text{gt}}$之间的风格相似性:

$$L_{\text{style}}=E\left[\sum_i\left\|\psi_i(\boldsymbol{I}_{\text{out}})-\psi_i(\boldsymbol{I}_{\text{gt}})\right\|_1\right] \tag{15}$$

其中,$\psi_i(\cdot)=\phi_i(\cdot)^{\mathrm{T}}\phi_i(\cdot)$为格拉姆矩阵,是特征图在通道维度上的自相关矩阵.

为了确保修复修复结果$\boldsymbol{I}_{\text{out}}$的全局一致性,本文采用对抗损失$L_{\text{adv}}$训练模型,定义如下:

$$L_{\text{adv}}=E_{\boldsymbol{I}_{\text{gt}}}\left[\log D(\boldsymbol{I}_{\text{gt}})\right]+E_{\boldsymbol{I}_{\text{out}}}\left[1-\log D(\boldsymbol{I}_{\text{out}})\right] \tag{16}$$

其中,D是采用谱归一化[48]训练的PatchGAN[49]鉴别器.

在预训练阶段,除上述损失外,本文还采用式(3)的量化损失L_{vq}约束码本中视觉原子的学习. 预训练阶段和修复阶段的总体损失L和L'是上述损失的加权和,如下式所示:

$$L=L_{\text{rec}}+\lambda_{\text{g}}L_{\text{grad}}+\lambda_{\text{p}}L_{\text{perc}}+\lambda_{\text{s}}L_{\text{style}}+\lambda_{\text{a}}L_{\text{adv}}+L_{\text{vq}} \tag{17}$$

$$L'=L_{\text{rec}}+\lambda_{\text{g}}L_{\text{grad}}+\lambda_{\text{p}}L_{\text{perc}}+\lambda_{\text{s}}L_{\text{style}}+\lambda_{\text{a}}L_{\text{adv}} \tag{18}$$

其中,λ_{g}、λ_{p}、λ_{s}和λ_{a}是用于平衡各项损失的权重超参数.

4 实验

4.1 实验设置

4.1.1 实验设置及运行环境/配置

两阶段均使用Adam(Adaptive moment estimation)优化器[50],一阶动量$\beta_1=0$,二阶动量$\beta_2=0.9$,批次大小为16. 前5 000次迭代中,学习率从0上升到2×10^{-4},随后采用余弦退火策略对学习率进行衰减. 矢量量化码本中视觉原子的个数N设为512,多头注意力中将N_{h}设为8. 若无特别说明,模块级联次数T设为3. 将损失函数的权重超参数设置为$\lambda_{\text{g}}=5$、$\lambda_{\text{p}}=0.1$、$\lambda_{\text{s}}=250$和$\lambda_{\text{a}}=0.1$[18,24,51]. 本文在Ubuntu 20.04系统基于PyTorch 1.12.0框架进行实验,硬件配置为单张NVIDIA RTX 3090 GPU(24 GB). 实验中所有训练和测试图像的尺寸均被缩放至256×256.

4.1.2 实验数据集及基准方法

本文采用公开的Places2[52]、FFHQ[53]和Paris StreetView[54]图像数据集进行训练和测试. 其中,Place2数据集包含超过180万张训练图像,328 500张测试图像,涵盖机场、美术馆、游乐园等365种不同类型的场景. 这些图像的背景复杂多样,能够验证修复模型在真实世界应用中的性能. FFHQ数据集包含70 000张涵盖不同年龄段和性别的人脸图像. 本文将前69 000张图像作为训练图像,后1 000张作为测试图像. Paris StreetView数据集源于巴黎的真实街景,图像内容覆盖巴黎街道的建筑、树木和雕塑等景观,包含14 900张训练图像和100张测试图像.

本文与近三年提出的五种主流图像修复方法[14,16,18,21,24]进行性能对比. 其中,CTSDG[14]和MISF[18]是图像修复领域基于CNN的经典方法. AOT-GAN[16]与CANet[21]分别提出了多尺度扩张卷积和跨尺度上下文注意力机制,以捕获远距离的上下文信息,但范围仅局限于图像内部. 与本文方法相似,PUT(Patch-based Unquantized Transformer)[24]也引入了矢量量化码本学习视觉原子. 但该方法仅通过Transformer[42]为缺失区域预测视觉原子的索引,忽略了内部特征的作用.

4.2 定量比较

本文采用峰值信噪比(Peak Signal-to-Noise Ratio, PSNR)、结构相似性(Structural SIMilarity, SSIM)、平均L1损失、图像块感知相似度(Learned Perceptual Image Patch Similarity, LPIPS)[38]和弗雷切特起始距离(Fréchet Inception Distance, FID)[55]五个指标来定量评价图像修复的性能. 前三者从像素层面、后两者分别从特征层面和分布层面衡量图像修复的性能. 表1展示了各图像修复方法在Places2、FFHQ和Paris StreetView数据集上在不同掩码率下的修复结果.

从表1可以看出,本文方法在三个数据集上的大多数指标上取得了最优的性能. 特别是对于真实街景Paris StreetView数据集,在三种不同掩码率下本文方法的五个指标平均为29.126 dB、0.894 1、0.017 3、0.123 4和38.703,这表明本文方法具有实际应用的潜力. 与CTSDG[14]、MISF[18]、AOT-GAN[16]和CANet[21]相比,本文

表1　在Places2、FFHQ和Paris StreetView数据集上不同方法性能比较

数据集		Places2			FFHQ			Paris StreetView		
掩码率		0%~20%	20%~40%	40%~60%	0%~20%	20%~40%	40%~60%	0%~20%	20%~40%	40%~60%
PSNR /dB ↑	CTSDG[14]	32.090	24.909	20.412	36.637	28.355	23.050	34.973	27.699	22.738
	MISF[18]	32.254	24.927	20.536	35.932	28.139	23.255	34.512	27.827	23.084
	AOT-GAN[16]	30.761	24.656	20.202	33.675	26.476	21.822	34.552	27.566	23.010
	CANet[21]	30.583	24.474	20.254	34.527	27.876	23.084	33.566	27.789	23.344
	PUT[24]	31.546	25.226	20.565	35.087	27.962	23.168	34.521	27.960	23.423
	本文方法	**33.406**	**25.669**	**20.915**	**36.824**	**28.773**	**23.600**	**35.382**	**28.354**	**23.641**
SSIM ↑	CTSDG[14]	0.971 6	0.889 8	0.711 6	0.988 9	0.950 9	0.848 2	0.979 4	0.911 9	0.751 5
	MISF[18]	0.971 6	0.891 7	0.726 5	0.987 9	0.950 0	0.853 1	0.978 3	0.913 8	0.766 1
	AOT-GAN[16]	0.966 0	0.888 3	0.719 3	0.982 1	0.932 5	0.813 8	0.977 6	0.910 1	0.764 4
	CANet[21]	0.966 1	0.885 8	0.719 1	0.985 1	0.946 9	0.847 6	0.975 4	0.915 0	0.774 3
	PUT[24]	0.970 1	0.888 5	0.728 9	0.986 7	0.951 4	**0.861 8**	0.977 6	0.915 7	0.779 6
	本文方法	**0.977 0**	**0.905 6**	**0.743 7**	**0.989 6**	**0.955 2**	**0.861 8**	**0.981 1**	**0.921 3**	**0.780 0**
*L*1 ↓	CTSDG[14]	0.006 5	0.021 7	0.048 3	0.003 6	0.013 2	0.040 8	0.004 4	0.015 2	0.037 5
	MISF[18]	0.006 5	0.021 2	0.046 4	0.004 1	0.013 9	0.032 6	0.004 8	0.015 7	0.036 2
	AOT-GAN[16]	0.008 3	0.022 7	0.048 7	0.004 7	0.014 6	0.033 8	0.004 8	0.016 4	0.036 6
	CANet[21]	0.007 5	0.022 5	0.047 8	0.005 2	0.016 9	0.038 6	0.005 3	0.016 0	0.035 5
	PUT[24]	0.007 0	0.022 3	0.047 9	0.004 1	0.013 6	0.031 6	0.004 8	0.015 4	0.033 9
	本文方法	**0.005 4**	**0.018 8**	**0.043 5**	**0.003 4**	**0.012 4**	**0.030 5**	**0.004 1**	**0.014 3**	**0.033 5**
LPIPS ↓	CTSDG[14]	0.067 5	0.182 6	0.316 0	0.032 0	0.100 8	0.202 1	0.046 3	0.162 0	0.270 5
	MISF[18]	0.043 0	0.129 7	0.249 9	0.028 6	0.085 3	0.168 9	0.039 2	0.115 2	0.223 6
	AOT-GAN[16]	0.057 1	0.142 9	0.266 0	0.031 6	0.093 1	0.185 3	0.041 2	0.120 0	0.224 4
	CANet[21]	0.053 9	0.132 1	0.246 1	0.035 7	0.086 7	0.166 3	0.044 3	0.121 7	0.222 2
	PUT[24]	0.044 3	0.133 0	0.255 7	0.029 9	0.088 0	0.178 0	0.041 9	0.123 3	0.235 6
	本文方法	**0.038 8**	**0.121 2**	**0.240 9**	**0.025 3**	**0.078 2**	**0.159 2**	**0.036 7**	**0.112 2**	**0.221 3**
FID ↓	CTSDG[14]	2.611	14.907	50.397	4.023	12.871	27.535	15.010	48.153	109.652
	MISF[18]	1.126	4.545	**13.969**	3.667	10.224	19.073	**11.007**	34.852	71.009
	AOT-GAN[16]	1.779	5.228	15.537	4.026	10.890	19.922	25.657	41.474	73.192
	CANet[21]	1.558	4.788	15.110	4.654	10.652	19.427	13.294	41.123	72.743
	PUT[24]	1.209	4.904	15.136	3.709	10.329	21.248	12.334	36.544	81.057
	本文方法	**1.066**	**4.475**	14.438	**3.349**	**9.620**	**18.447**	11.433	**33.849**	**70.826**

注:↑表示数值越大,性能越优;↓表示数值越小,性能越优.粗体表示性能最优.

方法在三个数据集40%~60%掩码率下的五个指标性能分别提高了2.98%、2.69%、10.88%、8.77%和14.15%.这得益于本文方法融合了内部和外部两个信息源.在缺失区域较大时,相比于仅利用内部特征的方法[14,16,18,21],本文方法通过外部特征弥补了内部特征的不足,拓展了注意力机制捕获上下文的范围.与仅依赖于外部特征的PUT[24]相比,本文方法在三个数据集40%~60%掩码率下在像素层面的三个指标上分别提升了1.5%、6.94%和4.62%,而在0%~20%掩码率下,提升幅度分别达到了4.45%、4.54%和18.17%.这表明本文方法在不同缺失区域面积下的性能均有所提升,特别是在缺失区域面积较小时,性能提升更为显著.这是由于本文方法结合了内部特征和外部特征两个信息源,能够提供与缺失区域相似的结构和纹理信息.

4.3　定性比较

图2所示是本文方法与基准方法在三个数据集上定性比较结果.每两行图像为一组,从上至下分别是在Places2、FFHQ和Paris StreetView数据集上的修复结果.图中红色虚线框表示基准方法与本文方法修复结果差异较大的区域.CTSDG和MISF的修复结果通常存在伪影等问题,如图2(*c*)和图2(*d*)第二行修复结果中的天空区域.AOT-GAN和CANet的修复结果存在模糊,如图2(*e*)和图2(*f*)第一行的窗户与第五行的墙壁.PUT的修复结果存在扭曲,如图2(*g*)第四行中虚线框所示.

相比之下,本文方法的修复结果具有清晰的纹理细节,缺失区域与完好区域颜色一致,更接近于真实图像. 例如,本文方法修复街景的建筑结构完整,无明显模糊与扭曲. 特别地,图2(a)中第四行的缺失区域已完全覆盖住面部,本文方法修复后的人脸图像五官分明,无明显的色差与扭曲. 这得益于内外交叉注意力可根据内部特征的语义信息从外部特征中搜寻语义相近的特征块填充缺失区域,弥补了内部特征的不足.

(a) 破损图像 (b) 真实图像 (c) CTSDG (d) MISF (e) AOT-GAN (f) CANet (g) PUT (h) 本文方法

图2 在Places2、FFHQ和Paris StreetView数据集上不同方法修复结果比较

4.4 消融实验

本节探讨不同消融条件对图像修复性能的影响,主要分为三个部分:第一,探究复合模块级联次数T的增减对修复性能的影响;第二,通过模块的消融,揭示各个模块对整体性能的贡献;第三,探究视觉原子的预训练数据集对修复性能的影响. 若无特殊说明,所有消融实验均在Paris StreetView数据集上进行.

4.4.1 关于级联次数T的消融实验

本文对复合模块级联次数T进行消融实验,分别将T设为1、2、3、4、5和10,实验结果如图3所示. 其中,图3(a)至图3(e)分别对比了PSNR、SSIM、$L1$、LPIPS和FID五个指标,前两个指标数值越大,性能越优,后三个指标数值越小,性能越优. 从图中可以看出,在0%~20%和20%~40%掩码率下,$T=3$在五个指标上均取得了最优性能. 具体地,在0%~20%掩码率下,$T=3$的性能比其他五个实验设置平均提高了0.61%、0.1%、2.8%、4.38%和3.03%;在20%~40%掩码率下,平均提高了0.39%、0.21%、1.85%、2.86%和5.27%. 当掩码率为40%~60%时,$T=3$在$L1$和FID指标上取得了最优性能. 随着T的增大,性能逐渐下降. 当$T=10$时,性能下降显著,这是因为模型的参数量随T的增大而增加,$T=10$的参数量相比于$T=1$的增加了26.82%,模型过度拟合训练数据,导致泛化能力下降. 上述结果表明随着模块级联次数T增大,模型的性能先提升后降低. 基于上述比较,本文将复合模块级联次数T设为3.

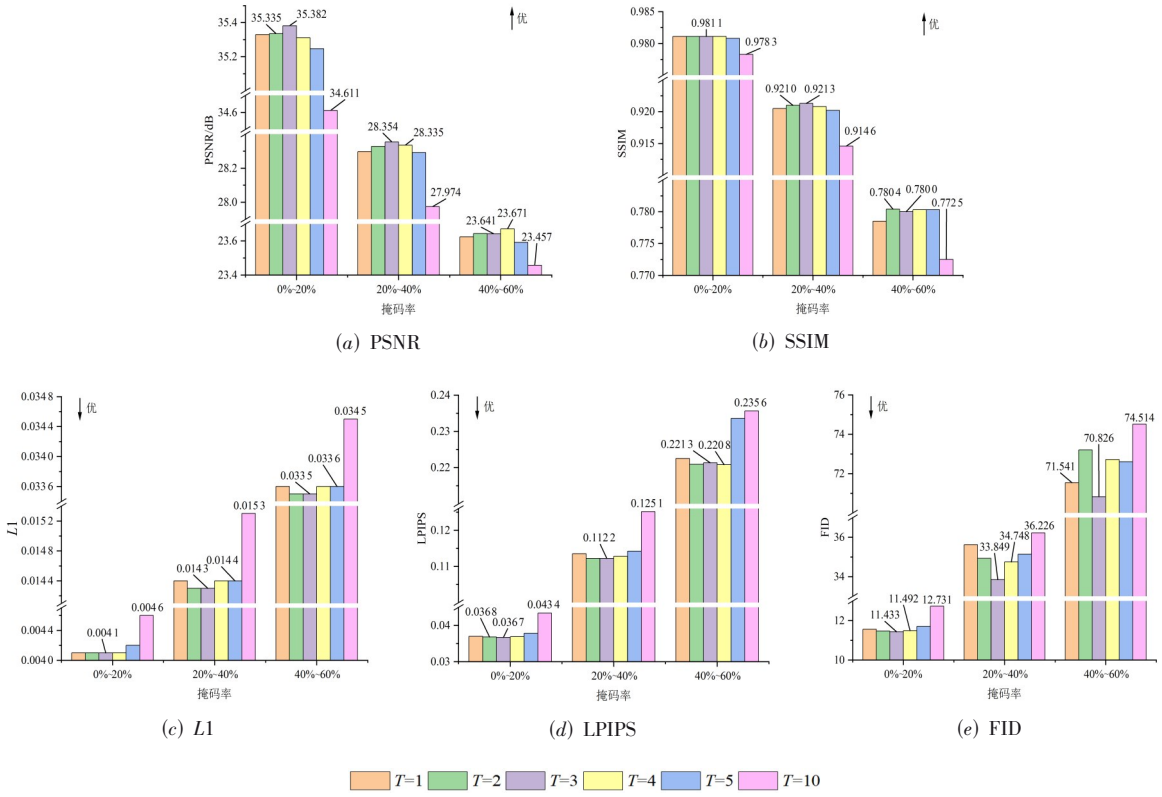


图3　在Paris StreetView数据集上对模块级联次数T的消融实验结果

4.4.2　模块的消融实验

为了验证本文提出的模块在图像修复模型中的有效性,本文对内部掩码注意力、内外交叉注意力和可控特征融合模块进行了消融实验,定量实验结果如表2所示,定性结果如图4所示. 表2中括号内的数值表示相较于本文方法性能下降的百分比. 图4从上至下依次为破损图像、真实图像、本文修复结果和其他实验设置的修复结果. 图中红色虚线框表示不同实验设置修复结果与本文修复结果差异较大的区域,实线框则为对应虚线框区域的放大图. 本文方法与三个实验设置进行对比,即“w/IMA”、“w/IECA”和“w/IMA&IECA”. 其中“w/IMA”去除了IECA和CFF模块,IMA的输出通过解码器G直接生成修复结果. “w/IECA”去除了IMA和CFF模块,IECA的输出直接作为解码器G的输入. “w/IMA&IECA”去除了CFF模块,仅通过一个可学习的标量加权内源修复特征和外源修复特征.

从表2中可以看出,“w/IMA”的五个指标的性能在0%~20%掩码率下平均比本文方法下降了0.37%、0.02%、2.44%、2.18%和1.27%. 当掩码率增至40%~60%时,“w/IMA”在五个指标上比本文方法平均下降了1.26%、0.85%、1.79%、2.94%和8.79%. 上述结果表明,随着掩码率的增大,“w/IMA”的性能下降越显著. 这是因为“w/IMA”仅依靠图像内部特征修复缺失区域. 随着缺失区域面积增大,IMA模块可捕获的上下文信息范围收缩,从而导致修复结果存在模糊与伪影. 如图4(*a*)第四行虚线框所示,“w/IMA”未能修复窗户的主体结构,以及在窗户与墙壁之间产生伪影.

从表2括号中的数值可知,“w/IECA”在所有实验设置中性能下降最为显著. 具体而言,在0%~20%掩码率时,在$L1$和LPIPS指标上分别下降了12.2%和23.43%. 在40%~60%掩码率下,这两个指标性能的降幅分别为4.78%和8.5%. 由于内部特征的缺失,w/IECA仅依赖于外部特征捕获结构与纹理,导致性能下降幅度扩大. 这是因为,相比于外部特征,内部特征往往能够提供与缺失区域相似的结构与纹理. 因此,在修复较小的缺失区域时,“w/IECA”容易产生伪影、色差等问题. 如图4(*b*)第四行虚线框所示,“w/IECA”修复的天空区域存在伪影,以及将墙壁错误地修复成玻璃.

“w/IMA&IECA”在五个指标上比本文方法分别下降0.63%、0.43%、2.54%、4.07%和6.9%. 虽然“w/IMA&IECA”的信息源包含内部和外部特征,但仅通过一个可学习的标量加权内源和外源修复特征,难以适应缺失区域与完好区域特征之间的差异. 在测试时,该标量无法根据缺失区域的变化动态调整,从而导致修复结果存在模糊与伪影等问题. 如图4(*c*)第四行虚线框所示,“w/IMA&IECA”错误地将树干的颜色修复成白色,以

表 2 在 Paris StreetView 数据集上对不同模块的消融实验结果 单位:%

指标	掩码率	实验设置			本文方法
		w/IMA	w/IECA	w/IMA&IECA	
PSNR ↑	0%~20%	35.250(−0.37)	34.536(−2.39)	35.191(−0.54)	35.382
	20%~40%	28.229(−0.44)	27.916(−1.54)	28.172(−0.64)	28.354
	40%~60%	23.342(−1.26)	23.301(−1.44)	23.475(−0.70)	23.641
SSIM ↑	0%~20%	0.980 9(−0.02)	0.978 1(−0.31)	0.980 4(−0.07)	0.981 1
	20%~40%	0.919 2(−0.23)	0.914 1(−0.78)	0.918 4(−0.31)	0.921 3
	40%~60%	0.773 4(−0.85)	0.766 0(−1.79)	0.773 0(−0.90)	0.780 0
*L*1 ↓	0%~20%	0.004 2(−2.44)	0.004 6(−12.20)	0.004 2(−2.44)	0.004 1
	20%~40%	0.014 5(−1.40)	0.015 3(−6.99)	0.014 7(−2.80)	0.014 3
	40%~60%	0.034 1(−1.79)	0.035 1(−4.78)	0.034 3(−2.39)	0.033 5
LPIPS ↓	0%~20%	0.037 5(−2.18)	0.045 3(−23.43)	0.038 6(−5.18)	0.036 7
	20%~40%	0.115 4(−2.85)	0.126 8(−13.01)	0.116 6(−3.92)	0.112 2
	40%~60%	0.227 8(−2.94)	0.240 1(−8.50)	0.228 2(−3.12)	0.221 3
FID ↓	0%~20%	11.578(−1.27)	13.070(−14.32)	11.868(−3.80)	11.433
	20%~40%	35.454(−4.74)	39.751(−17.44)	36.679(−8.36)	33.849
	40%~60%	77.051(−8.79)	81.321(−14.82)	76.875(−8.54)	70.826

注:↑表示数值越大,性能越优;↓表示数值越小,性能越优.

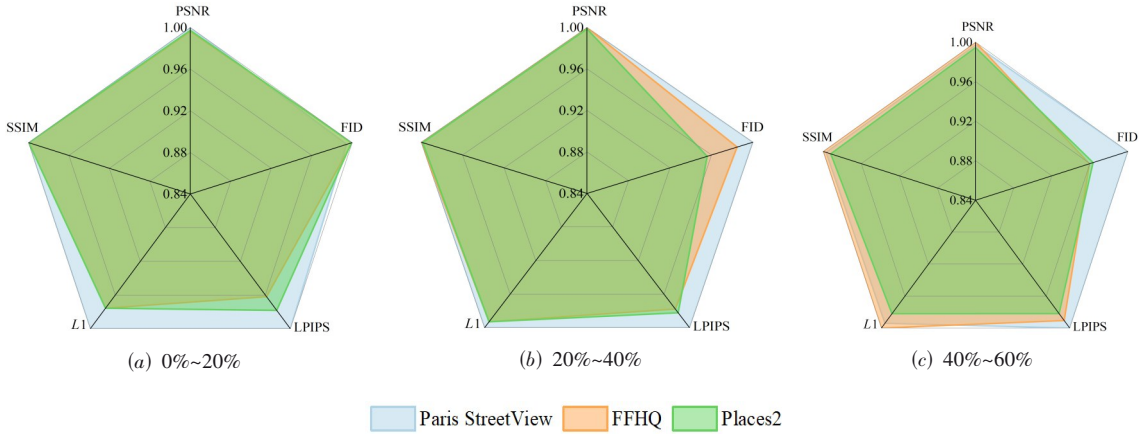
(*a*) w/IMA (*b*) w/IECA (*c*) w/IMA&IECA

图 4 在 Paris StreetView 数据集上不同实验设置修复结果比较

及修复的墙壁存在模糊. 相比之下,本文提出的CFF模块生成空间权重图,能够根据缺失区域的变化对各空间位置的权重进行动态调整,具有自适应性.

4.4.3　视觉原子的预训练数据集消融实验

为探究视觉原子的预训练数据集对修复性能的影响,本文对预训练数据集进行消融实验. 实验中,修复阶段均在Paris StreetView[54]数据集上进行训练和测试. 本实验设计了两组预训练方案:一组在Paris StreetView数据集上预训练视觉原子,确保视觉原子的学习与修复任务在相同数据集上进行;另一组分别在Places2[52]和FFHQ[53]数据集上进行预训练,以探索跨数据集预训练的视觉原子对修复性能的影响. 为了保证公平性,在预训练阶段分别从FFHQ和Places2中随机抽取14 900张图像(与Paris StreetView训练集图像数量保持一致)训练码本学习视觉原子. 实验结果如图5所示,图5(*a*)~图5(*c*)分别表示在掩码率为0%~20%、20%~40%和40%~60%的修复性能. 蓝色、橙色、绿色区域分别表示在Paris StreetView、FFHQ和Places2数据集上预训练视觉原子.



(*a*) 0%~20%　(*b*) 20%~40%　(*c*) 40%~60%

图5　不同数据集下预训练视觉原子的修复性能对比

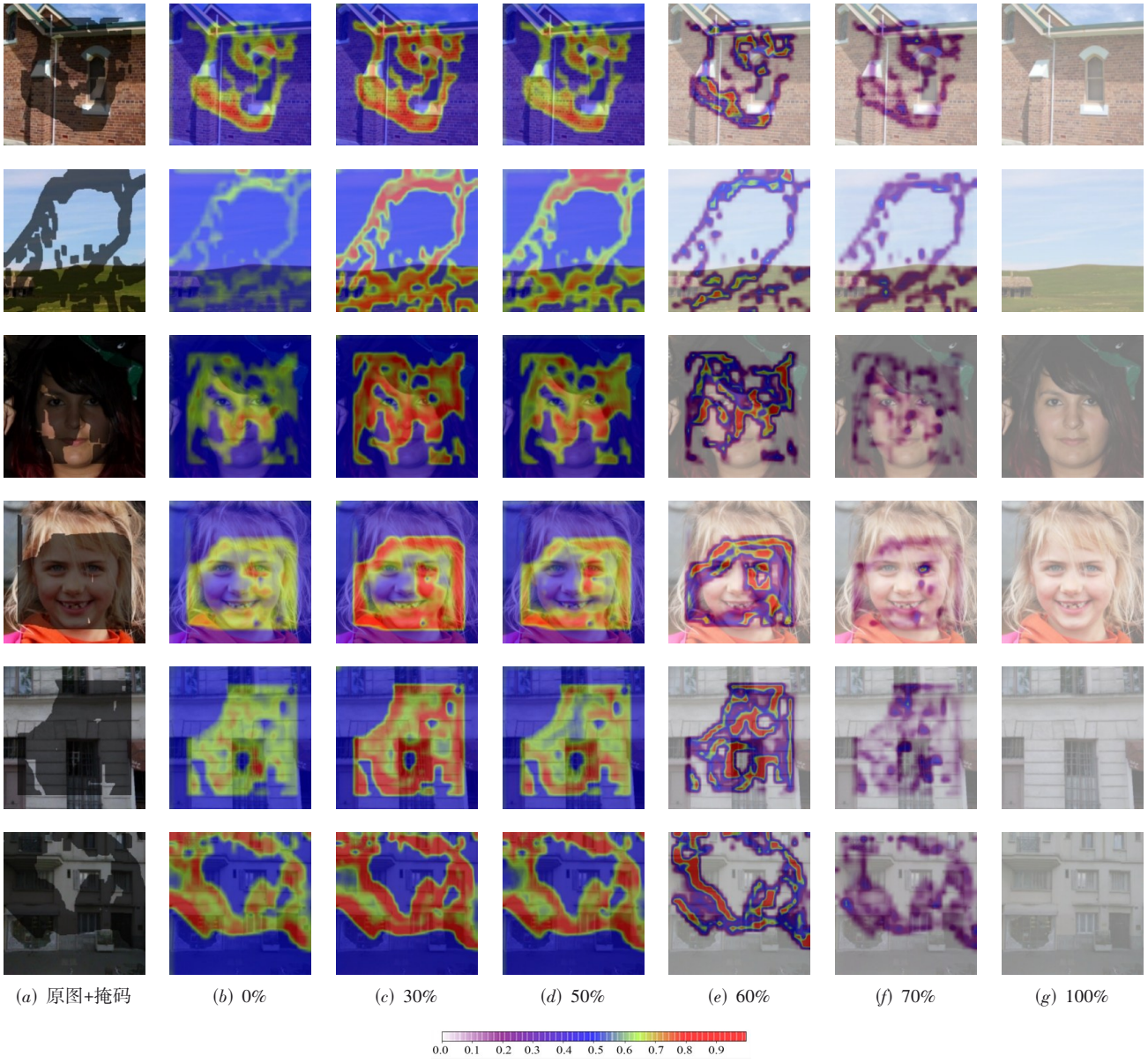
从图5中可以看出,跨数据集预训练视觉原子的整体修复性能与相同数据集的大体一致,性能保有率均超过95%. 特别地,在一些条件下跨数据集的性能指标实现了超越. 例如,当掩码率为0%~20%时,与Paris StreetView数据集相比,FFHQ和Places2数据集预训练的视觉原子在FID指标上的修复性能分别提升了0.55%和0.48%. 此外,在40%~60%的掩码率下,在FFHQ数据集上预训练的视觉原子在PSNR、SSIM和$L1$指标上分别提升了0.29%、0.31%和0.6%. 上述实验结果说明,视觉原子的预训练与特定数据集解耦,可在任意的通用数据集上进行,有利于提高修复模型面对多样化场景的适应性和鲁棒性.

4.5　可视化实验

为了探究内部特征和外部特征对修复破损图像的贡献,本文对空间权重图$\boldsymbol{w}_{ex}$进行可视化. 本实验对式(1)中的量化过程进行干预,按照一定概率错误地选择码本中距离最远(而非最近)的视觉原子替换内部特征的特征块,生成含有干扰的外部特征,以此来反映内外特征对于图像修复的作用. 本实验在Places2、FFHQ和Paris StreetView数据集上进行,将错误选择概率设置为0%、30%、50%、60%、70%和100%,$\boldsymbol{w}_{ex}$的可视化结果如图6所示. 图6(*a*)为叠加缺损掩码的真实图像,图6(*b*)至图6(*g*)为不同错误选择概率下获得的权重图$\boldsymbol{w}_{ex}$. 其中,$\boldsymbol{w}_{ex}$在每个空间位置的取值范围为0到1之间,当取值为1(红色)或0(白色)时,意味着CFF仅依赖外源或内源修复特征进行特征融合.

图6(*b*)为错误选择概率为0%条件下的权重图,即不引入任何错误,对应本文式(1)量化过程的做法. 从图6(*b*)可以看出,完好区域的权重值均在0.5左右,缺失区域的权重值则明显大于0.5. 这表明在修复缺失区域时,外部特征是重要的信息源,与内部特征互为补充,相辅相成,共同修复破损图像. 图6(*b*)~图6(*g*)的可视化结果表明,随着错误选择概率的增大,$\boldsymbol{w}_{ex}$完好区域的权重值最初保持不变(在0.5左右),随后减小至0. 这表明当错误选择概率小于50%时,内部和外部特征对重构完好区域的贡献各占一半. 当概率大于50%,则完全依赖内部特征进行重构. 这一实验结果说明,CFF能够自适应地进行特征筛选,从而实现内部与外部特征的融合. 特别地,当概率增大至60%时,从图6(*d*)与图6(*e*)可以看出,$\boldsymbol{w}_{ex}$完好区域的权重值骤减至0,这表明当外部特征含有的干扰超过60%时,CFF将完全依赖于内源修复特征重构完好区域. 这是由于当错误选择概率超过50%,干扰占据外部特征中的主导地位,迫使CFF仅依赖于内源修复特

征重构完好区域. 上述分析表明,在重构完好区域时,内部特征和外部特征的贡献大致相同;在修复缺失区域时,外部特征发挥主导作用,弥补了内部特征的不足.



(a) 原图+掩码 (b) 0% (c) 30% (d) 50% (e) 60% (f) 70% (g) 100%

图6 在Places2、FFHQ和Paris StreetView数据集上对权重图$\boldsymbol{w}_{ex}$的可视化

5 结论

本文通过训练矢量量化码本学习视觉原子,为修复破损图像提供外部特征,作为内部特征的补充,拓展了注意力机制捕获上下文的范围. 在此基础上,本文提出一种内外特征交互与融合的双流注意力图像修复方法. 该方法通过内部掩码注意力和内外交叉注意力组成的双流注意力,结合可控特征融合模块,实现了内部和外部特征的交互与融合. 两种特征互为补充,相辅相成,在特征层面共同修复破损图像. 此外,视觉原子的学习与数据集解耦,可在任意通用数据集上学习,不依赖于特定的数据集. 在Places2、FFHQ和Paris StreetView三个公开的数据集上的实验结果表明本文方法优于其他先进方法,在五个指标上分别平均提高了3.45%、1.34%、13.91%、13.64%和16.92%. 通过消融和可视化实验,证明了内部特征和外部特征通过双流注意力和可控特征融合进行交互与融合,二者互为补充,相辅相成,共同修复破损图像特征.

参考文献

[1] ZENG Y, FU J, CHAO H, et al. Learning pyramid-context encoder network for high-quality image inpainting[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2019: 1486-1494.

[2] ZUO Z, ZHAO L, LI A, et al. Generative image inpainting with segmentation confusion adversarial training and contrastive learning[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI, 2023: 3888-3896.

[3] JO Y, PARK J. Sc-fegan: Face editing generative adversarial network with user's sketch and color[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE, 2019: 1745-1753.

[4] 陈善雄, 朱世宇, 熊海灵, 等. 一种双判别器GAN的古彝文字符修复方法[J]. 自动化学报, 2022, 48(3): 853-864.
CHEN S X, ZHU S Y, XIONG H L, et al. A method of inpainting ancient yi characters based on dual discriminator generative adversarial networks[J]. Acta Automatica Sinica, 2022, 48(3): 853-864. (in Chinese)

[5] 赵磊, 吉柏言, 邢卫, 等. 基于多路编码器和双重注意力的古画修复算法[J]. 计算机研究与发展, 2023, 60(12): 2814-2831.
ZHAO L, JI B Y, XING W, et al. Ancient painting inpainting algorithm based on multi-channel encoder and dual attention[J]. Journal of Computer Research and Development, 2023, 60(12): 2814-2831. (in Chinese)

[6] 李建锋, 廖胜辉, 梅楚璇. 基于Mean Shift和插值图像修复算法的CT图像金属伪影消除方法[J]. 电子学报, 2017, 45(8): 1919-1924.
LI J F, LIAO S H, MEI C X. A mean shift algorithm and interpolation image restoration algorithm based method for metal artifact reduction[J]. Acta Electronica Sinica, 2017, 45(8): 1919-1924. (in Chinese)

[7] BERTALMIO M, SAPIRO G, CASELLES V, et al. Image inpainting[C]//Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques. New York: ACM, 2000: 417-424.

[8] RUŽIĆ T, PIŽURICA A. Context-aware patch-based image inpainting using Markov random field modeling[J]. IEEE Transactions on Image Processing, 2015, 24(1): 444-456.

[9] GUO Q, GAO S, ZHANG X, et al. Patch-based image inpainting via two-stage low rank approximation[J]. IEEE Transactions on Visualization and Computer Graphics, 2018, 24(6): 2023-2036.

[10] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

[11] GOODFELLOW I J, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[C]//Proceedings of the 27th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates, 2014: 2672-2680.

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

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

[14] GUO X, YANG H, HUANG D. Image inpainting via conditional texture and structure dual generation[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE, 2021: 14134-14143.

[15] 刘微容, 米彦春, 杨帆, 等. 基于多级解码网络的图像修复[J]. 电子学报, 2022, 50(3): 625-636.
LIU W R, MI Y C, YANG F, et al. Generative image inpainting with multi-stage decoding network[J]. Acta Electronica Sinica, 2022, 50(3): 625-636. (in Chinese)

[16] ZENG Y H, FU J L, CHAO H Y, et al. Aggregated contextual transformations for high-resolution image inpainting[J]. IEEE Transactions on Visualization and Computer Graphics, 2023, 29(7): 3266-3280.

[17] 罗会兰, 敖阳, 袁璞. 一种生成对抗网络用于图像修复的方法[J]. 电子学报, 2020, 48(10): 1891-1898.
LUO H L, AO Y, YUAN P. Image inpainting using generative adversarial networks[J]. Acta Electronica Sinica, 2020, 48(10): 1891-1898. (in Chinese)

[18] LI X, GUO Q, LIN D, et al. MISF: Multi-level interactive siamese filtering for high-fidelity image inpainting[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 1869-1878.

[19] 曹承瑞, 刘微容, 史长宏, 等. 多级注意力传播驱动的生成式图像修复方法[J]. 自动化学报, 2022, 48(5): 1343-1352.
CAO C R, LIU W R, SHI C H, et al. Generative image inpainting with attention propagation[J]. Acta Automatica Sinica, 2022, 48(5): 1343-1352. (in Chinese)

[20] 王山豹, 梁栋, 沈玲. 利用多模态注意力机制生成网络的图像修复[J]. 计算机辅助设计与图形学学报, 2023, 35(7): 1109-1121.
WANG S B, LIANG D, SHEN L. Image inpainting with multi-modal attention mechanism generative networks[J]. Journal of Computer-Aided Design & Computer Graph-

ics, 2023, 35(7): 1109-1121. (in Chinese)
[21] DENG Y, HUI S, ZHOU S, et al. Context adaptive network for image inpainting[J]. IEEE Transactions on Image Processing, 2023, 32: 6332-6345.
[22] LIU G, REDA F A, SHIH K J, et al. Image inpainting for irregular holes using partial convolutions[C]//Proceedings of the European Conference on Computer Vision. Cham: Springer, 2018: 85-100.
[23] LI W, LIN Z, ZHOU K, et al. MAT: Mask-aware transformer for large hole image inpainting[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 10758-10768.
[24] LIU Q, TAN Z, CHEN D, et al. Reduce information loss in transformers for pluralistic image inpainting[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 11347-11357.
[25] YU J, LIN Z, YANG J, et al. Generative image inpainting with contextual attention[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 5505-5514.
[26] LIU H, JIANG B, XIAO Y, et al. Coherent semantic attention for image inpainting[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE, 2019: 4170-4179.
[27] YU J, LIN Z, YANG J, et al. Free-form image inpainting with gated convolution[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE, 2019: 4471-4480.
[28] MA Y, LIU X, BAI S, et al. Coarse-to-fine image inpainting via region-wise convolutions and non-local correlation[C]//Proceedings of the International Joint Conference on Artificial Intelligence. San Francisco: Morgan Kaufmann, 2019: 3123-3129.
[29] WANG N, LI J, ZHANG L, et al. MUSICAL: Multi-scale image contextual attention learning for inpainting[C]//Proceedings of the International Joint Conference on Artificial Intelligence. San Francisco: Morgan Kaufmann, 2019: 3748-3754.
[30] ESSER P, ROMBACH R, OMMER B. Taming transformers for high-resolution image synthesis[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2021: 12873-12883.
[31] ZHOU S, CHAN K, LI C, et al. Towards robust blind face restoration with codebook lookup transformer[C]//Proceedings of the 36st International Conference on Neural Information Processing Systems. Red Hook: Curran Associates, 2022: 30599-30611.
[32] SHEN J, CHAN T F. Mathematical models for local nontexture inpaintings[J]. SIAM Journal on Applied Mathematics, 2002, 62(3): 1019-1043.
[33] CHAN T F, SHEN J. Nontexture inpainting by curvature-driven diffusions[J]. Journal of Visual Communication and Image Representation, 2001, 12(4): 436-449.
[34] CRIMINISI A, PÉREZ P, TOYAMA K. Region filling and object removal by exemplar-based image inpainting[J]. IEEE Transactions on Image Processing 2004, 13(9): 1200-1212.
[35] YU F, KOLTUN V. Multi-scale context aggregation by dilated convolutions[C]//Proceedings of the International Conference on Learning Representations. Washington: ICLR, 2016: 1-16.
[36] LIU Z, LIN Y, CAO Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE, 2021: 10012-10022.
[37] VAN DEN OORD A, VINYALS O. Neural discrete representation learning[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook: Curran Associates, 2017: 6309-6318.
[38] JOHNSON J, ALAHI A, FEI-FEI L. Perceptual losses for real-time style transfer and super-resolution[C]//Proceedings of the European Conference on Computer Vision. Cham: Springer, 2016: 694-711.
[39] GU Y, WANG X, XIE L, et al. VQFR: Blind face restoration with vector-quantized dictionary and parallel decoder[C]//Proceedings of the European Conference on Computer Vision Cham: Springer, 2022: 126-143.
[40] WANG Z, ZHANG J, CHEN R, et al. Restoreformer: High-quality blind face restoration from undegraded key-value pairs[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 17512-17521.
[41] ZOU W, GAO H, YE T, et al. VQCNIR: Clearer night image restoration with vector-quantized codebook[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI, 2024: 7873-7881.
[42] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st Conference on Neural Information Processing Systems. New York: ACM, 2017: 6000-6010.
[43] ZHU X, HU H, LIN S, et al. Deformable convnets v2: More deformable, better results[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2019: 9308-9316.

[44] GATYS L A, ECKER A S, BETHGE M. Image style transfer using convolutional neural networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2016: 2414-2423.
[45] DENG J, DONG W, SOCHER R, et al. ImageNet: A large-scale hierarchical image database[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2009: 248-255.
[46] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[C]//Proceedings of the International Conference on Learning Representations. Washington: ICLR, 2015: 1-13.
[47] SAJJADI M S, SCHOLKOPF B, HIRSCH M. Enhancenet: Single image super-resolution through automated texture synthesis[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE, 2017: 4491-4500.
[48] MIYATO T, KATAOKA T, KOYAMA M, et al. Spectral normalization for generative adversarial networks[C]//Proceedings of the International Conference on Learning Representations. Washington: ICLR, 2018: 1-13.
[49] ZHU J Y, PARK T, ISOLA P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks [C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE, 2017: 2223-2232.
[50] KINGMA D P, BA J. Adam: A method for stochastic optimization[C]//Proceedings of the International Conference on Learning Representations. Washington: ICLR, 2014: 58-64.
[51] NAZERI K, NG E, JOSEPH T, et al. EdgeConnect: Structure Guided Image Inpainting using Edge Prediction[C]//Proceedings of the IEEE International Conference on Computer Vision Workshops. Piscataway: IEEE, 2019: 3265-3274.
[52] ZHOU B, 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.
[53] KARRAS T, LAINE S, AILA T. A style-based generator architecture for generative adversarial networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2019: 4401-4410.
[54] DOERSCH C, SINGH S, GUPTA A, et al. What makes Paris look like Paris?[J]. ACM Transactions on Graphics, 2012, 31(4): 1-9.
[55] HEUSEL M, RAMSAUER H, UNTERTHINER T, et al. Gans trained by a two time-scale update rule converge to a local nash equilibrium[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook: Curran Associates, 2017: 6629-6640.

作者简介

黄光远　男,2000年生,江苏徐州人. 东华大学信息科学与技术学院硕士研究生. 主要研究方向为图像修复.
E-mail: gyhuang@mail.dhu.edu.cn

周树波　男,1988年生,浙江绍兴人. 东华大学信息科学与技术学院助理研究员. 主要研究方向为计算成像、工业视觉检测等.
E-mail: zhoushubo@dhu.edu.cn

黄　荣　男,1985年生,浙江绍兴人. 东华大学信息科学与技术学院副教授. 主要研究方向图像修复、语义分割等.
E-mail: rong.huang@dhu.edu.cn

蒋学芹　男,1981年生,江苏苏州人. 东华大学信息科学与技术学院教授. 主要研究方向为图信号处理、工业视觉检测等.
E-mail: xqjiang@dhu.edu.cn