结构与纹理双生成的二阶段网络图像修复

石计亮[1,2],张 乾[1,2]
1. 贵州民族大学 数据科学与信息工程学院, 贵阳 550025
2. 贵州民族大学 贵州省模式识别与智能系统重点实验室,贵阳 550025

摘 要:目的 针对现有图像修复方法不能很好地实现结构和纹理信息之间的双向交互,在修复缺失面积较大或纹理复杂的图像时存在纹理模糊、结构失真等问题。**方法** 提出了一种基于双向坐标注意融合模块和傅里叶特征聚合模块的二阶段网络图像修复方法。首先,使用结构编-解码器和纹理编-解码器对受损图像进行结构重建和纹理合成,产生初步的修复结果;然后,将粗修复结果输入到细化修复网络,利用双向坐标注意融合模块和傅里叶特征聚合模块对图像内部纹理细节进行修复;为增强全局一致性,设计了双向坐标注意融合模块来实现结构和纹理信息之间的双向交互,并设计了傅里叶特征聚合模块,用于捕获全局上下文信息,增强图像局部特征之间的相关性,以获得精细的修复结果;此外,还利用双流判别器来估计结构和纹理的特征统计量,以区分原始图像和生成图像。**结果** 在 CelebA-HQ 数据集上进行实验,与 4 种图像修复方法进行比较,定性结果表明方法生成的人脸图像更加清晰自然;定量结果表明方法在峰值信噪比、结构相似性指数和弗雷歇距离上均优于对比算法;对模型中各模块的消融实验结果也验证了所提出创新点的有效性。**结论** 因此,所提出的方法能够有效地修复受损的人脸图像,特别是在大面积遮挡下也能生成具有结构合理、纹理清晰的图像。

关键词:图像修复;二阶段网络;生成对抗网络;深度学习
中图分类号:TP391.41 **文献标识码:**A doi:10.16055/j.issn.1672-058X.2025.0001.002

Two-stage Network Image Inpainting with Dual Generation of Structure and Texture

SHI Jiliang[1,2], ZHANG Qian[1,2]
1. School of Data Science and Information Engineering, Guizhou Minzu University, Guiyang 550025, China
2. Key Laboratory of Pattern Recognition and Intelligent Systems of Guizhou Province, Guizhou Minzu University, Guiyang 550025, China

Abstract: Objective Existing image inpainting methods fail to achieve effective bidirectional interaction between structure and texture information, resulting in issues like texture blur and structural distortion when repairing images with large missing areas or complex textures. **Methods** A two-stage network image inpainting method was proposed, employing a bidirectional coordinate attention fusion module and a Fourier feature aggregation module. Firstly, the damaged image was subjected to structure reconstruction and texture synthesis using structure encoder-decoder and texture encoder-decoder, generating preliminary inpainting results. Subsequently, the coarse inpainting result was input to a refinement inpainting network, where the bidirectional coordinate attention fusion module and the Fourier feature aggregation module were

收稿日期:2023-09-10 **修回日期:**2023-10-25 **文章编号:**1672-058X(2025)01-0009-11
基金项目:贵州民族大学校级科研项目(GZMUZK[2021]YB23).
作者简介:石计亮(1999—),男,贵州榕江人,硕士研究生,从事图像修复研究.
通讯作者:张乾(1984—),男,贵州贵定人,教授,博士,硕士生导师,从事人工智能、计算机视觉和模式识别研究. Email:gzmuzq@gzmu.edu.cn.
引用格式:石计亮,张乾. 结构与纹理双生成的二阶段网络图像修复[J]. 重庆工商大学学报(自然科学版),2025,42(1):9-19.
SHI Jiliang, ZHANG Qian. Two-stage network image inpainting with dual generation of structure and texture[J]. Journal of Chongqing Technology and Business University (Natural Science Edition), 2025, 42(1): 9-19.

utilized to repair the internal texture details of the image. To enhance global consistency, the bidirectional coordinate attention fusion module was designed to facilitate bidirectional interaction between structure and texture information. Additionally, the Fourier feature aggregation module was designed to capture global contextual information, enhancing the correlation between local image features to obtain fine inpainting results. Moreover, dual-stream discriminators were employed to estimate the feature statistics of structure and texture, distinguishing between original and generated images. **Results** In experiments conducted on the CelebA-HQ dataset, compared with four image inpainting methods, qualitative results indicated that face images generated by this method were clearer and more natural; the quantitative results showed that this method outperformed the contrastive algorithms in peak signal-to-noise ratio, structural similarity index, and Fréchet distance. Ablation experiments on various modules of the model also validated the effectiveness of the proposed innovations. **Conclusion** Therefore, the proposed method effectively restores damaged face images, especially generating images with reasonable structure and clear texture even under large occlusions.

Keywords: image inpainting; two-stage network; generative adversarial network; deep learning

1 引 言

图像修复[1]是利用合理的内容填充图像缺失区域,以使得修复后的图像在视觉感知上看起来更自然、连续。图像修复已经成为计算机视觉一个重要的研究领域,其广泛应用于人们日常生活中,例如图像编辑[2]、图像生成[3]、文化遗产保护和艺术修复[4-5]、航空航天遥感图像处理[6]、医学影像恢复[7]等领域。

图像修复方法主要包括传统图像修复方法和基于深度学习的图像修复方法。传统图像修复方法大多使用数学计算提取图像的完整信息,并填充图像的缺失区域。但方法未能提取图像的深层信息,导致生成的图像缺乏语义信息,修复效果与人类视觉感知不相符。特别是在修复遮挡面积较大的图像时,传统图像修复方法未能生成复杂且非重复的图像结构和纹理。近年来,随着卷积神经网络的发展,深度学习方法在图像修复任务中达到最先进的水平,应用卷积神经网络可以缓解这些问题。通过卷积计算,可以更好地提取图像的深层信息,使得生成图像的信息更加丰富。Yan 等[8]提出了一种新的 CNN 模型,即 Shift-Net,它通过将移位连接层与 U-Net 相结合,能够清晰地填充任何形状的缺失区域,并保留精细的纹理。虽然该方法能够产生清晰、精细和视觉上可信的结果,但无法解决单阶段网络模型模糊的问题,导致生成的图像中丢失了一些纹理细节。在随后的工作中,Liu 等[9]提出了一种互编解码器 CNN,以平衡结构和纹理特征。该方法有效地恢复了结构和纹理特征,并在修复结果中很好地保留了两者的信息。然而,方法没有充分利用图像的上下文信息,从而使得局部特征与整体的一致性连接有限。大多数这类单阶段图像修复方法都无法处理图像细节,导致合成图像过于平滑,出现与周围区域不一致的边界伪影、扭曲结构和模糊纹理。这些方法可能是因为卷积神经网络在建模远距离上下文信息时存在一些不足之处。

为解决单阶段图像修复方法的不足,Yu 等[10]提出了一种基于上下文注意力机制的二阶图像修复模型。方法在第一阶段使用扩张卷积进行初始预测,在生成的信息和已知信息之间进行匹配,并经过情景关注通道和卷积通道进行处理,最终将它们生成的结果合并以得到更精细的修复结果。虽然方法解决了单阶段图像修复方法在建模远距离上下文信息的不足。然而,在修复大面积缺失的图像时仍然存在伪影的问题。Nazeri 等[11]提出了 Edge Connect (EC) 模型,模型将图像修复分为两个阶段:边缘生成和图像补全。边缘生成仅关注图像缺失区域的结构边缘信息,然后使用结构边缘信息来重建最终的图像。方法确保了结构边缘和原始图像在视觉上的一致性。然而,方法仅用单一的结构边缘作为先验信息,无法正确处理图像修复任务。Xiong 等[12]也提出了类似的模型,采用 Sobel 算子从分割图中检测出图像的不完整轮廓作为结构先验,而不是边缘。然后,将不完整轮廓输入轮廓补全模块以预测缺失的轮廓,最后将完整的轮廓与输入图像一起送入图像补全模块,进一步恢复受损图像。尽管方法在缺失面积较小的图像修复任务中能够预测出合理的对象轮廓。然而,对于缺失面积较大的图像进行修复任务时,方法从受损图像中获取合理的对象轮廓有限。这些方法的成功表明边缘、轮廓等结构知识在生成合理内容的图像修复中起着重要作用。然而,由于使用串联耦合架构,在推理过程中很容易受到不合理

结构前提条件下的负面影响。为了生成有意义的结构,Ren 等[13]提出了 StructureFlow 模型。模型由两部分组成:结构重构器和纹理生成器。结构重构器用于预测缺失的结构,生成全局结构图像;纹理生成器则利用重构的结构来恢复纹理细节。方法通过使用边缘保留的平滑图像,更好地表示图像的全局结构,并专注于恢复全局结构,避免受到相关纹理信息的干扰。然而,方法在恢复合格的局部结构较为困难,并对结构(如边缘和轮廓)的准确性要求较高,难以保证生成合理的结构。为了解决这一难题,一些研究人员尝试对图像的结构和纹理进行建模。Li 等[14]设计了一种视觉结构重建(VSR)层,采用部分卷积和瓶颈块来恢复缺失区域的部分边缘。然后,将重建的边缘与带孔的图像结合,通过填充语义上有意义的内容来逐步恢复受损图像。方法结合了转置卷积和部分卷积的优点,提高了图像修复的效果;Yang 等[15]提出了一种多任务学习框架,框架结合结构知识来辅助图像修复,训练一个共享生成器来同时修复损坏的图像和相应的结构信息。这些方法对结构和纹理进行建模,虽然产生了视觉上令人满意的修复结果。然而,在单一的共享架构中同时对结构和纹理进行建模,并使它们的信息充分互补,是相当困难的。以上这些方法并未完全考虑结构和纹理之间的相互影响关系,使得结构和纹理难以获取互补的信息。

认为现有的图像修复方法不能很好地实现结构和纹理信息之间的双向交互,导致在修复缺失面积较大或纹理复杂的图像时出现纹理模糊和结构失真等缺陷。为了解决这些问题,提出了一种由粗到细的二阶段网络图像修复方法。首先,在第一阶段分别运用结构编码器-解码器和纹理编码器-解码器进行结构重建和纹理合成,以实现结构引导的纹理合成和纹理促进的结构重建模式,并得到粗糙的修复结果。其次,第二阶段采用双向坐标注意融合(Bi-directional Coordinate Attention Fusion, Bi-CAF)模块,能够很好地实现结构和纹理信息之间的双向交互,相互促进,从而增强一致性。此外,通过设计的傅里叶特征聚合(Fourier Feature Aggregation, FFA)模块,捕获全局上下文信息,增强图像局部特征之间的相关性,并学习合理的全局结构,以获得更好的修复结果。同一数据集在相同环境下的对比实验表明,方法在主观感受和定量评价指标上均优于其他对比算法。生成的人脸图像纹理清晰、结构合理且符合图像语义。此外,还对各个模块进行了消融实验,以验证所提出模块的有效性。

2　相关工作

2.1　基于传统的图像修复

传统的图像修复方法主要包括:基于扩散[16]的方法和基于样本块[17]的方法。前者通过从缺失区域周围的边界,将已知信息填充到边缘内部。Bertalmio 等[1]提出了一种基于偏微分方程的扩散方法,将艺术修复的思想与之结合,方法对具有少量缺失区域或重复图案的图像非常有效,但对于破损区域较大或结构复杂的图像,往往无法生成合理的图像。基于样本块的方法则从周围图像区域中提取样本块,来填补或替代缺失区域的像素值。Drori 等[18]采用平滑插值方法迭代处理图像中缺失区域的内容,寻找最相似的样本块填充到缺失区域。Criminisi 等[19]通过采用优先级来确定修复图像的顺序,并在整张图像范围内寻找相似区域进行填充,以优先保持图像结构和纹理的传播方向,算法适用于恢复简单纹理结构的图像,但在纹理结构复杂的图像上仍存在局限。Wilczkowiak 等[20]通过搜索相似的补丁区域,并与用户进行交互引导修复,尽管方法在遮挡修复方面优于基于扩散的方法,但是方法仍依赖图像中的相似块。这些传统的图像修复方法在固定纹理上表现良好。然而,它们通常无法为复杂场景合成令人满意的结果,因为它们无法理解图像中的高级语义信息。

2.2　基于深度学习的图像修复

对于大面积缺失或复杂纹理的图像修复,基于深度学习的方法能更好地捕捉图像中的高级语义信息。Goodfellow 等[21]提出生成对抗网络(Generative Adversarial Nets, GAN)后,Pathak 等[22]首次将 GAN 引入图像修复领域,提出了 Context Encoders 模型。模型采用编码器-解码器的结构,可修复图像的语义信息和大面积遮挡。然而,模型的编码器和解码器之间采用全连接网络,导致图像过于平滑和模糊。随后,lizuka 等[23]使用全卷积生成器,并添加局部判别器,形成全局与局部图像修复模型,以增强结构和纹理细节。然而,这种双判别器仍无法很好地处理图像细节。为了提高修复结果的质量和准确性,Yu 等[10]设计了一个由粗到细的生成器架构,并引入一个上下文注意力机制,以确保修复区域与完整区域之间的一致性。此外,他们采用改进的 Wasserstein GAN[24]来稳定网络训练。与此类似,Nazeri 等[11]也采用由粗到细的网络结构。在第一阶段,网络

恢复图像的结构信息,如边缘和轮廓;而在第二阶段,则进行了精细的填充。然而,在第一阶段一次性重建图像结构信息时,容易生成错误的图像轮廓。在深度修复模型中,单阶段修复模型生成的图像结构模糊且纹理细节不足。因此,许多研究人员致力于改进从粗到细的图像修复模型,后者能够产生更真实的纹理。此外,由粗到细的图像修复模型可以一次性恢复受损区域。然而,如果第一阶段产生不准确的图像结构信息,会直接影响第二阶段的生成结果,最终导致生成图像与原始图像之间的不一致。

3 方 法

提出了一种基于双向坐标注意融合(Bi-directional Coordinate Attention Fusion, Bi-CAF)模块和傅里叶特征聚合(Fourier Feature Aggregation, FFA)模块的二阶段网络图像修复方法。生成器包括粗糙生成器 G_1 和细化生成器 G_2,通过粗糙到细化的填充方式来修复缺失区域。网络的整体结构如图 1 所示。为了提高生成部分与原始图像在结构和纹理上的相似性,首先将输入图像经过粗糙生成器 G_1 处理,然后再次经过细化生成器 G_2 进行进一步修复。在第一阶段,粗糙生成器 G_1 分别使用纹理编-解码器和结构编-解码器进行初始预测,得到粗糙的结构和纹理特征。在第二阶段中,使用细化生成器 G_2 来处理粗糙的结构和纹理特征。生成器通过本文设计的核心模块进行处理,包括双向坐标注意融合(Bi-directional Coordinate Attention Fusion, Bi-CAF)模块和傅里叶特征聚合(Fourier Feature Aggregation, FFA)模块。使用这些模块,本文方法能够得到精细的修复结果。为了区分本文方法生成图像与原始图像,借助双流判别器[25]来估算图像的结构和纹理特征统计量。

3.1 粗糙生成阶段

设 I^{gt},E^{gt} 和 Y^{gt} 分别表示原始图像、完整的边缘图和灰度图,M 表示掩码。其中,缺失区域标记为 1,否则标记为 0。在粗糙生成器 G_1 中输入的数据被描述为

$$\begin{cases} I^{in} = I^{gt} \odot (1-M) \\ E^{in} = E^{gt} \odot (1-M) \\ Y^{in} = Y^{gt} \odot (1-M) \end{cases}$$

其中,I^{in},E^{in} 和 Y^{in} 分别表示受损图像、受损边缘图和灰度图。

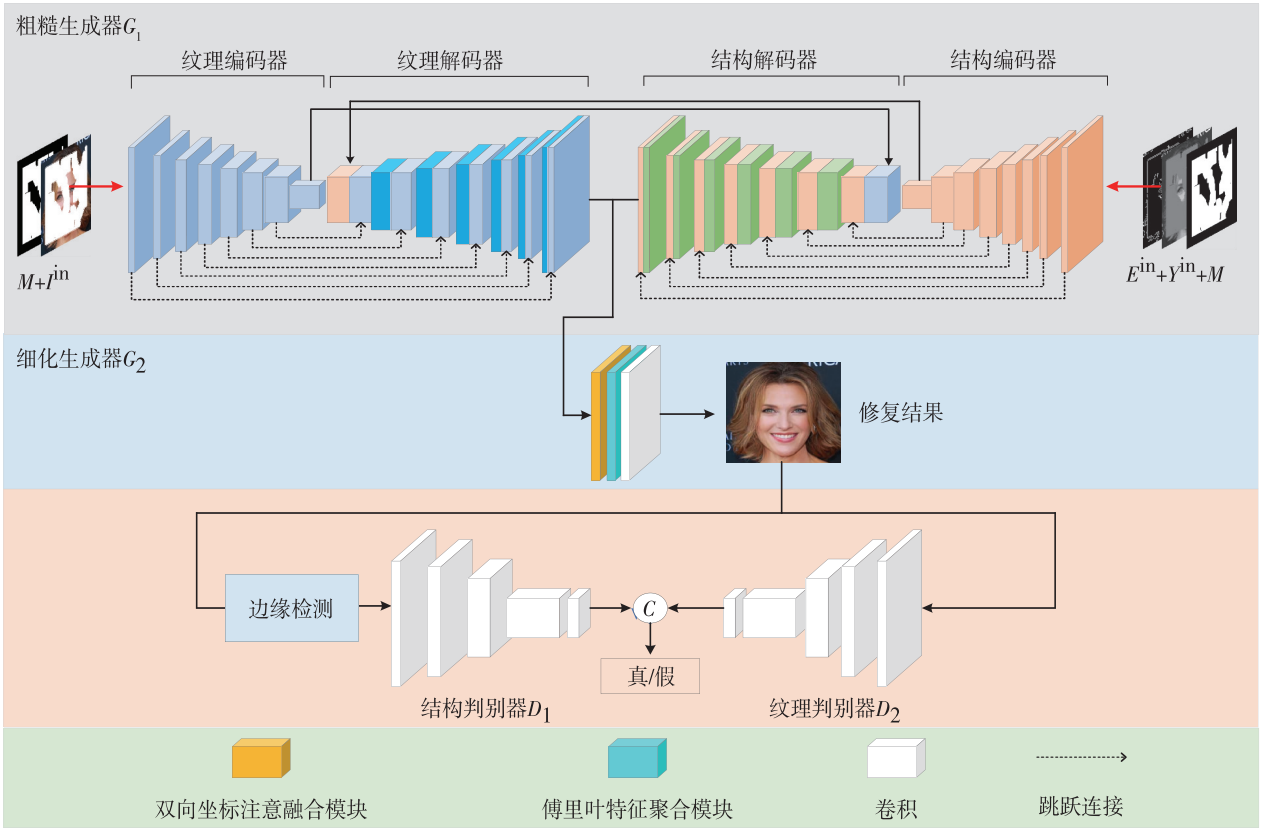


图 1 总体网络结构图

Fig. 1 The diagram of the overall network structure

在粗糙生成器 G_1[25]中,左边为纹理编码器-解码器,右边为结构编码器-解码器。在编码阶段,将受损

图像 I^{in} 和掩码 M 作为纹理编码器的输入,将受损的边缘图 E^{in}、灰度图像 Y^{in} 和掩码 M 作为结构编码器的输入。而在解码阶段中,两个解码器相互利用对方编码器最后一层输出的特征,以实现纹理引导的结构重建和结构约束的纹理合成,这样的操作使得结构和纹理信息之间相互利用,从而在编码阶段生成更合理的内容。为了更好地从受损图像中捕获不规则边界的信息,在编码器-解码器中,使用部分卷积层来替代普通卷积,因为部分卷积只依赖于可见的信息,不会受到遮挡的影响。采用跳跃连接[26]将低级和高级特征在多个尺度上进行融合,以产生更复杂的预测结果。通过粗糙生成器对输入图像进行初步预测,得到粗糙的结构和纹理特征。其过程可表示为

$$F_{st},F_{te}=G_1(I^{in},E^{in},Y^{in},M)$$

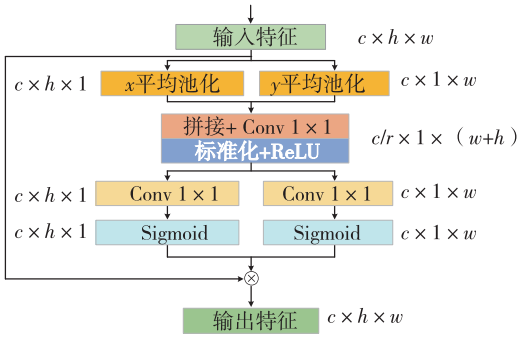
其中,F_{st} 和 F_{te} 分别是粗糙生成器所生成的结构特征图和纹理特征图。

3.2　细化生成阶段

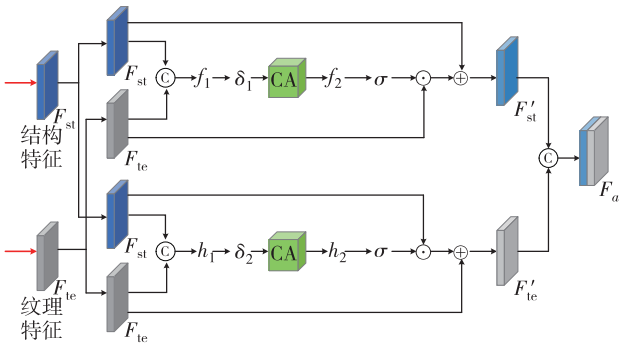
细化生成阶段使用粗糙生成阶段生成的结构和纹理特征作为输入,先后经过本文设计的 Bi-CAF 模块和 FFA 模块进行精细化。其中,Bi-CAF 模块旨在自适应融合粗糙生成阶段生成的结构和纹理特征,以增强它们的一致性。FFA 模块则旨在捕获全局上下文信息,增强图像局部特征之间的相关性,并学习合理的整体结构。接下来,本节内容将详细描述 Bi-CAF 模块和 FFA 模块。

3.2.1　双向坐标注意融合

在图像修复任务中,确保图像缺失区域和未缺失区域之间特征信息的一致性至关重要。通过引入注意力机制,网络能够有针对性地从未缺失区域中提取准确的特征信息,并将其用于重建缺失区域中的特征。Hou 等[27]提出了坐标注意力(Coordinate Attention, CA)机制,如图 2(a)所示。该机制可以预测高级特征信息的残差,精确地突出网络感兴趣的对象,增强有用特征、削弱无用特征,有助于模型准确提取有用的特征信息,并修复图像中的缺失区域。在当前的图像修复领域,虽然已经有方法尝试同时处理图像的结构和纹理,如 MED[9],它通过互编码器-解码器分别学习结构和纹理。其中,深层 CNN 特征用于结构学习,浅层 CNN 特征用于纹理学习。方法设计了两个分支,分别对结构和纹理特征进行空洞填充,并在特征融合过程中提高了结构和纹理之间的一致性。然而,方法没有充分考虑图像结构和纹理之间的关系,很难传递全面的互补信息,这可能会限制结构和纹理之间的一致性。为了解决这个问题,在提出的 Bi-CAF 模块中嵌入了 CA 机制,Bi-CAF 模块结构如图 2(b)所示。



(a) Coordinate Attention 结构图



(b) Bi-CAF 模块结构图

图 2　Bi-CAF 模块和 CA 机制结构图

Fig. 2　Architecture diagrams of Bi-CAF module and CA mechanism

模块能够自适应地融合粗糙生成阶段输出的结构和纹理特征,允许结构和纹理信息在网络中双向交互,并通过门控机制控制信息的传递,以确保它们之间的空间一致性。粗糙生成阶段输出的纹理特征记为 F_{te},结构特征记为 F_{st}。首先,在通道维度上将 F_{te} 与 F_{st} 级联,以构建纹理感知的结构特征 F'_{st},为了控制纹理信息被整合的程度,定义了一个软门控 P_{te},表示为

$$P_{te}=\sigma(f_2(\mathrm{CA}(\delta_1(f_1(\mathrm{Concat}(F_{st},F_{te}))))))$$

其中,Concat(·)表示对特征进行通道的级联操作,$f_1(\cdot)$ 与 $f_2(\cdot)$ 均为 3×3 的卷积操作,$\delta_1(\cdot)$ 为 LeakyReLU 激活函数,CA(·)表示特征经过坐标注意力机制,$\sigma(\cdot)$ 为 Sigmoid 函数。对于 P_{te},采用自适应机制,将粗糙生成阶段输出的纹理特征 F_{te} 融合为 F_{st},可表示为

$$F'_{st}=\alpha(P_{te}\odot F_{te})\oplus F_{st}$$

其中,α 是初始化为零的训练参数,$\odot$ 和 $\oplus$ 分别表示元素相乘操作和相加操作。

而对于构建结构感知纹理特征 F'_{te},定义一个软门

控 P_{st},可表示为

$$P_{st}=\sigma(h_2(\mathrm{CA}(\delta_2(h_1(\mathrm{Concat}(F_{st},F_{te}))))))$$

$$F'_{te}=\beta(P_{st}\odot F_{st})\oplus F_{te}$$

其中,$h_1(\cdot)$与$h_2(\cdot)$均为3×3的卷积操作,$\delta_2(\cdot)$与$\delta_1(\cdot)$相同,β是初始化为零的训练参数。

最后,在通道维度上将F'_{st}与F'_{te}进行级联操作,从而得到融合后的特征图F_a,过程表示为

$$F_a=\mathrm{Concat}(F'_{st},F'_{te})$$

3.2.2 傅里叶特征聚合

在不规则大掩码修复中,需要充分考虑全局的上下文信息。因此,较大的有效感受野[28]。对于理解图像的全局结构以及解决修复问题至关重要。Suvorov等[29]提出使用傅里叶卷积进行频域学习,图3中频域特征融合(Frequency Domain Feature Fusion, FDFF)模块具有较大有效感受野,可以有效解决图像全局上下文信息的问题。FDFF具有几种卷积,用于下采样和上采样图像特征。模块最关键的部分是快速傅里叶卷积(Fast Fourier Convolution, FFC)[30]块,FFC块基于信道快速傅里叶变换(Fast Fourier Transform, FFT)[31],它由两个分支组成:(1)局部分支使用传统的卷积;(2)全局分支使用FFT来考虑全局结构并捕获长距离的上下文信息。最后,局部分支和全局分支的输出被堆叠在一起。为了解决单一尺度难以获取图像全局与局部信息的问题,提出傅里叶特征聚合(Fourier Feature Aggregation, FFA)模块,如图3所示。模块能捕获全局上下文信息,以增强图像局部特征之间的相关性,并学习合理的全局结构。

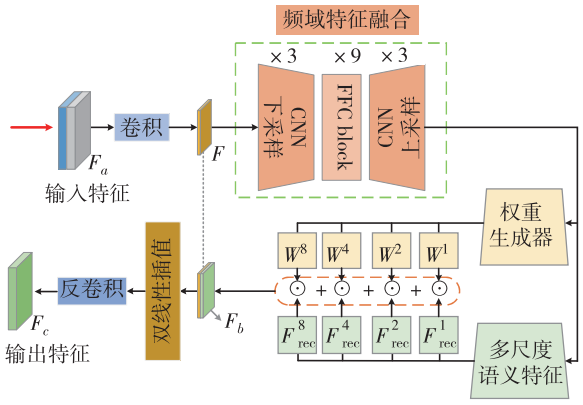


图3 FFA模块结构图

Fig. 3 Structure diagram FFA module

为了更好地融合图像多尺度特征空间信息,并对多尺度下丰富的语义特征进行编码,以处理更具挑战性的情况。使用了4组带有不同扩张率的卷积层来捕获多尺度语义特征,过程表示为

$$F^k_{rec}=\mathrm{Conv}_k(F_{rec})$$

其中,$\mathrm{Conv}_k(\cdot)$为不同扩张率的扩张卷积层k,F^k_{rec}为经过扩张卷积层后的特征图,$k\in\{1,2,4,8\}$,F_{rec}为频域特征融合后的特征图。

为了更好地聚合多尺度语义特征,设计了一个像素级权重生成器G_w,以更好地处理不同像素点的重要性。生成器由一系列卷积和激活函数组成,具体包括两个卷积层。第一个卷积层的卷积核大小为3,第二个卷积层的卷积核大小为1。在每个卷积层后面接ReLU非线性激活函数,并使用Softmax函数在通道维度上进行归一化处理。第二个卷积层的输出通道数为4。过程表示为

$$W^1,W^2,W^4,W^8=\mathrm{Slice}(\mathrm{Softmax}(G_w(F_{rec})))$$

其中,$\mathrm{Softmax}(\cdot)$是使用Softmax函数在通道维度上的归一化操作,$Slice(\cdot)$为对通道进行切片。通过元素加权和聚合多尺度语义特征,可以获得更精细化的特征图,过程表示为

$$F_b=(F^1_{rec}\odot W^1)\oplus(F^2_{rec}\odot W^2)\oplus(F^4_{rec}\odot W^4)\oplus(F^8_{rec}\odot W^8)$$

其中,F_b为经过元素加权和多尺度语义操作后的特征图,$\odot$和$\oplus$分别表示各元素相乘和相加操作。

其次,将特征图F和F_b在通道维度上进行拼接,并使用双线性插值方法将拼接后的特征图恢复到与原始特征图相同的尺寸大小。最后,通过反卷积得到修复后的图像,过程表示为

$$F_c=f(B(\mathrm{Concat}(F,F_b)))$$

其中,$B(\cdot)$为双线性插值操作,$f(\cdot)$为反卷积操作,F_c为经过FFA模块融合后的特征图。

3.3 判别器

判别器由结构判别器D_1和纹理判别器D_2组成,用于估计原始图像和生成图像之间的结构和纹理特征,以实现二者的区分。判别器参数如表1所示,两个判别器结构相同。判别器由5个卷积层组成,前3个卷积层的核大小为4,步长为2,后面2个卷积层的核大小为4,步长为1。前4个卷积层使用斜率为0.2的LeakyReLU作为激活函数,本文使用Sigmoid函数作为最后一个卷积层的激活函数。最后,将2个判别器的输出在通道维度上进行级联,并基于此计算对抗损失。D_2的输入通道数为3,D_1的输入通道数为4。与D_2不同,D_1采用灰度图像作为附加条件,并将其与检测到的边缘图作为输入以优化D_1的对抗损失。因此,D_1不仅

可以判断生成结构的真实性,还可以保证与真实图像的一致性。为了解决生成对抗网络中训练不稳定的问题,使用谱归一化的方法来稳定生成对抗网络的训练,使判别网络满足 Lipschitz 连续连。

表 1　判别器网络结构

Table 1　Network structure of the discriminator

层	网络类型	输入通道数	输出通道数	卷积核数	步　长
1	Conv2d	3(4)	64	4	2
2	Conv2d	64	128	4	2
3	Conv2d	128	256	4	2
4	Conv2d	256	512	4	1
5	Conv2d	512	1	4	1

3.4　损失函数

联合训练损失函数主要包括像素重建损失、感知损失、风格损失、对抗损失和中间监督损失,定义为

$$L_{\text{total}}=\lambda_{\text{rec}}L_{\text{rec}}+\lambda_{\text{perc}}L_{\text{perc}}+\lambda_{\text{style}}L_{\text{style}}+\lambda_{\text{adv}}L_{\text{adv}}+\lambda_{\text{inter}}L_{\text{inter}}$$

其中,L_{rec} 为像素重建损失,L_{perc} 为感知损失,L_{style} 为风格损失,L_{adv} 为对抗损失,L_{inter} 为监督粗糙生成器中两个解码器准确获取纹理特征和结构特征的损失,λ_{rec}、λ_{perc}、λ_{style}、λ_{adv} 和 λ_{inter} 分别为各损失函数的权重值。

使用 L_1 正则化作为输出图像和真实图像之间的像素重构损失,定义为

$$L_{\text{rec}}=\| I^{\text{out}}-I^{\text{gt}} \|_1$$

其中,I^{out} 为模型输出图像;I^{gt} 为真实图像。

在许多图像修复模型中,已经验证了在图像修复领域应用感知损失[32]和风格损失[33]。使用在 ImageNet[34]数据集上预训练的 VGG-16 网络来提取图像的高层语义信息,感知损失定义为

$$L_{\text{perc}}=\sum_i \frac{\| \phi_i(I^{\text{out}})-\phi_i(I^{\text{gt}}) \|_1}{N_i}$$

其中,$\boldsymbol{\phi}_i(\cdot)$ 是在 VGG-16 网络的第 i 个特征层中提取到的特征,使用前 3 个池化层的特征信息,N_i 为 $\boldsymbol{\phi}_i(\cdot)$ 中的元素个数。

通过在特征空间中引入风格损失约束,以进一步增强修复图像的真实性,定义为

$$L_{\text{style}}=E\Big[\sum_i \| (\boldsymbol{\phi}_i(I^{\text{out}})-\boldsymbol{\phi}_i(I^{\text{gt}})) \|_1\Big]$$

$$\boldsymbol{\phi}_i(I^{\text{out}})=\boldsymbol{\phi}_i(I^{\text{out}})^{\text{T}}\boldsymbol{\phi}_i(I^{\text{out}})$$

$$\boldsymbol{\phi}_i(I^{\text{gt}})=\boldsymbol{\phi}_i(I^{\text{gt}})^{\text{T}}\boldsymbol{\phi}_i(I^{\text{gt}})$$

其中,$\boldsymbol{\phi}_i(\cdot)$ 表示由 $\boldsymbol{\phi}_i$ 激活映射构造的克拉姆矩阵。

对抗损失是用于 GANs 训练的一种损失约束。为确保方法生成图像的真实性,并保持图像结构和纹理信息的一致性,利用对抗损失来监督修复模型的训练过程,定义为

$$L_{\text{adv}}=\min_G \max_D E_{I^{\text{gt}},E^{\text{gt}}}[\log D(I^{\text{gt}},E^{\text{gt}})]+E_{I^{\text{out}},E^{\text{out}}}[\log(1-D(I^{\text{out}},E^{\text{out}}))]$$

其中,E^{gt} 为完整边缘图,E^{out} 为生成的边缘图。

为确保粗糙生成器中的纹理解码器和结构解码器准确捕获纹理和结构特征,引入了以下损失对输出特征图 F_{st} 和 F_{te} 进行中间监督[25],定义为

$$L_{\text{inter}}=L_{\text{structure}}+L_{\text{texture}}=\text{BCE}(E^{\text{gt}},G_{\text{st}}(F_{\text{st}}))+\ell_1(I^{\text{gt}},G_{\text{te}}(F_{\text{te}}))$$

其中,BCE(·)是二元交叉熵损失,G_{st} 和 G_{te} 表示由残差块和卷积层实现的投影函数,它们分别将 F_{st} 和 F_{te} 映射到边缘图和 RGB 图像。

4　实验与结果分析

方法基于 Pytorch 框架的 1.8.1 版本进行开发,并在 24 GB 显存的 RTX 3090 GPU 上进行训练和测试。在训练阶段,使用 Adam 优化器对模型进行优化,与文献[25]相似,模型训练初始阶段的学习率设置为 2×10^{-4},然后以 5×10^{-5} 的学习率对模型进行微调,批量大小设置为 6。损失函数的各部分权重设置与文献[25]类似,分别设置为 $\lambda_{\text{rec}}=10$,$\lambda_{\text{perc}}=0.1$,$\lambda_{\text{style}}=250$,$\lambda_{\text{adv}}=0.1$,$\lambda_{\text{inter}}=1$。实验使用了公开人脸数据集 CelebA-HQ[35],数据集包含了 30 000 张高清人脸图像。将前 3 000 张图像作为测试集,其余 27 000 张图像作为训练集,在 CelebA-HQ 数据集上进行了 50 万次迭代的训练,并进行了 10 万次迭代的微调。对于不规则掩码,从文献[36]中采用了不规则掩码进行模型的训练和测试,所有输入图像和掩码的大小均被调整为 256×256 像素。

4.1　定性评价

为了客观比较本文方法与其他对比方法在 CelebA-HQ 数据集上的人脸图像修复效果,选择的对比方法有 RFR[37],PRVS[38],CTSDG[25]和 HAN[39]。5 种图像修复方法生成的人脸图像效果如图 4 所示。由图 4 可以看出,RFR[37]和 PRVS[38]所生成的人脸图像整体效果并不理想,如图 4 第 3 列和第 4 列。HAN[39]方法生成的人脸图像效果与原图明显有些差异,如图 4 第 6 列。CTSDG[25]方法生成的人脸图像在耳朵和眼睛等部位修复效果较差,如图 4 第 5 列所示。相比之下,方法生成的人脸图像更加合理和自然,特别是在修复大面积遮挡的人脸图像时,方法生成的人脸图像结构合理、纹理清晰,修复效果均优于其他对比方法,如图 4 最后一列所示。

图 4 不同模型在 CelebA-HQ 数据集上的定性对比

Fig. 4 Qualitative comparison of different models on the CelebA-HQ dataset

4.2 定量评价

与文献[39]的工作类似,选择了如下 3 个常用的评价指标。峰值信噪比(peak signal-to-noise ratio, PSNR)是一个客观评价指标,用于评估生成图像的质量。结构相似性(structural similarity, SSIM)指数[40]是比较生成图像与原始图像之间的差异。弗雷歇距离(Fréchet distance, FID)[41]是评估生成图像的准确性和多样性,用来表示修复结果的感知质量。用 P_{PSNR}、S_{SSIM}、F_{FID} 来分别表示它们值的大小(同下文)。并与 4 个对比方法 RFR[37]、PRVS[38]、CTSDG[25]和 HAN[39]进行定量比较。P_{PSNR} 与 S_{SSIM} 越高,说明修复效果越好;相反,F_{FID} 越低,说明图像在高级特征上越相似。使用文献[36]的不同比例不规则掩码在 CelebA-HQ 数据集上进行测试,5 种图像修方法定量结果如表 2 所示。由表 2 可以看出,方法在 F_{FID}、P_{PSNR} 和 S_{SSIM} 这 3 个指标上均取得了最高值,随着掩码比例的增加,方法的优势更加明显。与其他 4 个对比方法相比较,方法在定量评价中取得了优异的结果。

表 2 不同模型在 CelebA-HQ 数据集上的定量对比

Table 2 Quantitative comparison of different models on the CelebA-HQ dataset

指 标	模 型	1%~10%	10%~20%	20%~30%	30%~40%	40%~50%	50%~60%
F_{FID}	RFR[37]	17.326 0	29.720 1	38.960 3	45.408 6	61.965 7	91.272 0
	PRVS[38]	4.779 2	16.983 9	20.146 9	28.183 4	40.102 7	66.960 3
	CTSDG[25]	0.280 1	1.276 7	2.984 9	3.978 3	5.882 4	9.014 1
	HAN[39]	1.686 3	1.778 1	3.680 5	4.020 4	5.747 4	10.130 3
	本文方法	**0.246 4**	**1.060 2**	**2.384 5**	**3.237 4**	**4.948 5**	**7.570 1**
P_{PSNR}	RFR[37]	37.353 8	29.654 7	25.683 5	24.762 5	22.128 9	20.775 2
	PRVS[38]	37.459 0	28.817 2	25.260 0	23.810 8	21.260 8	19.386 2
	CTSDG[25]	44.598 3	35.588 4	30.142 9	28.653 4	25.795 5	23.921 9
	HAN[39]	38.879 3	33.764 5	29.053 8	27.907 8	25.325 9	23.361 3
	本文方法	**45.333 5**	**36.380 1**	**30.918 7**	**29.424 1**	**26.617 2**	**24.660 4**
S_{SSIM}	RFR[37]	0.980 0	0.909 1	0.825 3	0.767 5	0.681 9	0.610 0
	PRVS[38]	0.981 9	0.903 6	0.809 1	0.749 3	0.654 3	0.547 4
	CTSDG[25]	0.995 8	0.973 5	0.924 7	0.892 7	0.839 3	0.786 6
	HAN[39]	0.985 3	0.959 4	0.907 0	0.880 6	0.830 1	0.773 8
	本文方法	**0.996 2**	**0.976 0**	**0.931 5**	**0.901 8**	**0.854 6**	**0.804 4**

4.3　消融实验

为了验证方法中的双向坐标注意融合(Bi-CAF)、傅里叶特征聚合(FFA)和频域特征融合(FDFF)模块对图像修复结果的影响,进行了定性和定量的消融实验。为保证实验结果的可靠性,所有消融实验均在同一环境下进行训练和测试。

4.3.1　定性消融实验

为了验证方法不同模块对图像修复结果的影响,实验中采用相同的掩码进行定性消融实验,定性结果如图 5 所示。其中,图 5(a)为本文的基准模型修复结果,即未添加 Bi-CAF 模块与 FFA 模块的情况。可以看出帽子边缘、眼镜、眼睛和耳饰等部位修复效果极差,生成的人脸图像纹理模糊。添加 Bi-CAF 模块后,眼镜、帽子边缘和耳饰等部位的修复效果进一步提升,如图 5(b)所示。在基准模型上添加了傅里叶特征聚合(FFA)模块后,修复质量显著提高,特别是在修复图像的整体结构和纹理细节方面,相比基准模型有很大提升,如图 5(c)所示。在基准模型上添加了 Bi-CAF 模块和 FFA 模块(去除了 FDFF)后,修复效果优于基准模型,如图 5(d)所示。方法将 Bi-CAF 模块和 FFA 模块同时加到基准模型中,修复结果如图 5(e)所示。可以看出,在修复细节方面有所提升,如图 5(e)中眼镜、帽子边缘、眼睛、头发和耳饰等部分修复效果与原图更接近。方法能够很好地还原受损图像中缺失的部分,使得修复结果与原始图像更接近,也说明了方法所提出模块的有效性。

4.3.2　定量消融实验

表 3 展示了在 CelebA-HQ 数据集上使用(40%~50%掩码率)掩码进行的模块消融实验定量对比结果。为了直观比较各个模块在方法的作用,从基准模型开始逐个模块添加,并对修复结果进行对比,与之相对应的定性比较结果如图 5 所示。由表 3 可知,在未添加任何模块时的各评价指标值均很低。进一步,通过在基准模型上分别添加 Bi-CAF 和 FFA 模块,网络修复性能逐渐提高,且均优于基准模型。三者具体的定量对比结果如表 3 中图 5 序号(a)、(b)、(c)所示。而在 CelebA-HQ 数据集上,方法在 P_{PSNR}、S_{SSIM} 和 F_{FID}3 个指标上都取得了最高值,分别为 26.617 2 dB、0.854 6 和 4.948 5,与之对应的可视化图也说明本文方法修复效果。相较于基准模型,方法的性能提高了 0.989 5 dB、0.021 4 和 1.577 7。实验结果表明,在面对大面积缺失修复任务时,本文的 FFA 和 Bi-CAF 模块能够稳定地提升模型的性能。进一步也验证了方法所提出的各模块对修复性能的影响。

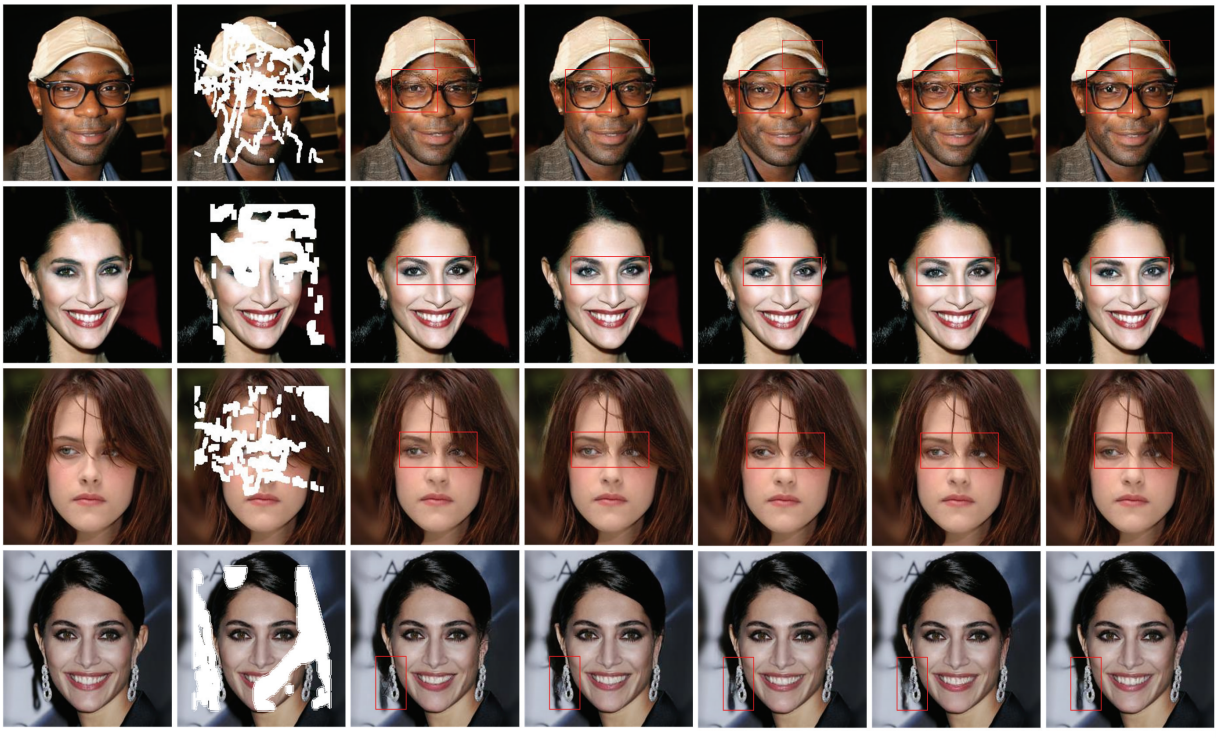


图 5　消融实验定性结果

Fig. 5　Qualitative results of ablation experiments

表 3 消融实验定量结果

Table 3 Quantitative results of ablation experiments

图 5 序号	模块组合	P_{PSNR}	S_{SSIM}	F_{FID}
(a)	基准	25.627 7	0.833 2	6.526 2
(b)	基准+Bi-CAF	25.924 3	0.843 8	5.692 9
(c)	基准+FFA	26.447 4	0.850 1	5.217 1
(d)	基准+Bi-CAF+FFA(去除 FDFF)	26.041 9	0.846 0	5.523 1
(e)	基准+Bi-CAF+FFA(本文方法)	**26.617 2**	**0.854 6**	**4.948 5**

5 结 论

提出了一种基于双向坐标注意融合模块和傅里叶特征聚合模块的二阶图像修复方法。方法由粗糙修复网络和细化修复网络构成。首先,粗糙修复阶段使用双编-解码器进行结构重建和纹理合成,得到粗糙的结构特征和纹理特征;其次,细化修复阶段则采用双向坐标注意融合模块来促进结构和纹理信息之间的双向交互,并利用傅里叶特征聚合模块来捕获全局上下文信息和增强图像局部特征之间的相关性,以学习合理的全局结构,以获得精细的修复结果。

在定性和定量对比实验中,与其他图像修复方法相比,方法表现出了有效性。但在处理佩戴眼镜的人脸图像时,尤其是遭遇较大的遮挡面积,方法生成的修复效果与原始图像存在一定差异。因此,未来的工作中,需要进一步解决大面积遮挡的人脸图像修复方面的缺陷。

参考文献(References):

[1] BERTALMIO M, SAPIRO G, CASELLES V, et al. Image inpainting[C]//Proceedings of the 27th annual conference on Computer graphics and interactive techniques, 2000: 417-424.

[2] BROOKS T, HOLYNSKI A, EFROS A A. InstructPix2Pix: learning to follow image editing instructions[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 18392-18402.

[3] WANG L, WU Z, ZHONG Y, et al. Snapshot spectral compressive imaging reconstruction using convolution and contextual Transformer[J]. Photonics Research, 2022, 10(8): 1848-1858.

[4] ZHOU Z, LIU X, SHANG J, et al. Inpainting digital Dunhuang murals with structure-guided deep network[J]. Journal on Computing and Cultural Heritage, 2022, 15(4): 77.

[5] WANG H, LI Q, ZOU Q. Inpainting of Dunhuang murals by sparsely modeling the texture similarity and structure continuity[J]. Journal on Computing and Cultural Heritage, 2019, 12(3): 1710-1720.

[6] WEI Q, ZUO Z, NIE J, et al. Inpainting of remote sensing sea surface temperature image with multi-scale physical constraints[C]//Proceedings of the IEEE International Conference on Multimedia and Expo. 2023: 492-497.

[7] ZHANG R, LU W, WEI X, et al. A progressive generative adversarial method for structurally inadequate medical image data augmentation[J]. IEEE Journal of Biomedical and Health Informatics, 2022, 26(1): 7-16.

[8] YAN Z, LI X, LI M, et al. Shift-net: image inpainting via deep feature rearrangement[M]//Computer Vision-ECCV 2018. Cham: Springer International Publishing, 2018: 3-19.

[9] LIU H, JIANG B, SONG Y, et al. Rethinking image inpainting via a mutual encoder-decoder with feature equalizations[C]//European Conference on Computer Vision. Cham: Springer, 2020: 725-741.

[10] YU J, LIN Z, YANG J, et al. Generative image inpainting with contextual attention[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018: 5505-5514.

[11] NAZERI K, NG E, JOSEPH T, et al. EdgeConnect: structure Guided Image Inpainting using Edge Prediction[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision Workshop. 2019: 3265-3274.

[12] XIONG W, YU J, LIN Z, et al. Foreground-aware image inpainting[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 5840-5848.

[13] REN Y, YU X, ZHANG R, et al. StructureFlow: image inpainting via structure-aware appearance flow[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 181-190.

[14] LI J, HE F, ZHANG L, et al. Progressive reconstruction of visual structure for image inpainting[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 5961-5970.

[15] YANG J, QI Z, SHI Y. Learning to incorporate structure knowledge for image inpainting[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 12605-12612.

[16] BALLESTER C, BERTALMIO M, CASELLES V, et al. Filling-in by joint interpolation of vector fields and gray levels[J]. IEEE Transactions on Image Processing, 2001, 10(8): 1200-1211.

[17] XU Z, SUN J. Image inpainting by patch propagation using patch sparsity[J]. IEEE Transactions on Image Processing, 2010, 19(5): 1153-1165.

[18] DRORI I, COHEN-OR D, YESHURUN H. Fragment-based image completion[J]. ACM Transactions on Graphics, 2003, 22(3): 303-312.

[19] 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.

[20] WILCZKOWIAK M M, BROSTOW G J, TORDOFF B, et al. Hole filling through photomontage[C]//Proceedings of the British Machine Vision Conference. 2005: 492-501.

[21] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[C]//Proceedings of the 27th International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2014: 2672-2680.

[22] 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. 2016: 2536-2544.

[23] IIZUKA S, SIMO-SERRA E, ISHIKAWA H. Globally and locally consistent image completion[J]. ACM Transactions on Graphics, 2017, 36(4): 107-114.

[24] ARJOVSKY M, CHINTALA S, BOTTOU L. Wasserstein generative adversarial networks[C]//Proceedings of the 34th International Conference on Machine Learning. 2017: 214-223.

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

[26] RONNEBERGER O, FISCHER P, BROX T. U-net: convolutional networks for biomedical image segmentation[C]//Lecture Notes in Computer Science. Cham: Springer International Publishing, 2015: 234-241.

[27] HOU Q, ZHOU D, FENG J. Coordinate attention for efficient mobile network design[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 13713-13722.

[28] LUO W, LI Y, URTASUN R, et al. Understanding the effective receptive field in deep convolutional neural networks[C]//Proceedings of the 30th International Conference on Neural Information Processing Systems. 2016: 4905-4913.

[29] SUVOROV R, LOGACHEVA E, MASHIKHIN A, et al. Resolution-robust large mask inpainting with Fourier convolutions[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022: 3172-3182.

[30] CHI L, JIANG B, MU Y. Fast Fourier convolution[C]//Proceedings of the 34th International Conference on Neural Information Processing Systems. 2020: 4479-4488.

[31] BRIGHAM E O, MORROW R E. The fast Fourier transform[J]. IEEE Spectrum, 1967, 4(12): 63-70.

[32] JOHNSON J, ALAHI A, LI F F. Perceptual losses for real-time style transfer and super-resolution[C]//European Conference on Computer Vision. Cham: Springer, 2016: 694-711.

[33] 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. 2016: 2414-2423.

[34] RUSSAKOVSKY O, DENG J, SU H, et al. ImageNet large scale visual recognition challenge[J]. International Journal of Computer Vision, 2015, 115(3): 211-252.

[35] KARRAS T, AILA T, LAINE S, et al. Progressive growing of GANs for improved quality, stability, and variation[J]. Arxiv Preprint Arxiv, 2017, 79: 1710-1720.

[36] LIU G, REDA F A, SHIH K J, et al. Image inpainting for irregular holes using partial convolutions[C]//Proceedings of the Computer Vision-ECCV 2018: 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XI. 2018: 89-105.

[37] LI J, WANG N, ZHANG L, et al. Recurrent feature reasoning for image inpainting[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 7757-7765.

[38] LI J, HE F, ZHANG L, et al. Progressive reconstruction of visual structure for image inpainting[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 5961-5970.

[39] DENG Y, HUI S, MENG R, et al. Hourglass attention network for Image inpainting[C]//Lecture Notes in Computer Science. Cham: Springer Nature Switzerland, 2022: 483-501.

[40] 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.

[41] 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. 2017: 6629-6640.

责任编辑:田　静