doi:10.12068/j.issn.1005－3026.2024.09.002

基于去噪扩散概率模型的人脸图像修复模型

刘纪红，黄熙雄
（东北大学 信息科学与工程学院，辽宁 沈阳 110819）

摘　　要：针对使用主流人脸图像修复模型在修复图像后，出现图像质量欠佳、修复边缘模糊，且模型复杂、训练困难的问题，提出了一种基于去噪扩散概率模型的人脸图像修复模型．通过使用Guided-diffusion中的U-Net网络结构，并在网络中引入快速傅里叶卷积来改进去噪扩散概率模型，最后在CelebA-HQ高清人脸图像数据集上进行模型的训练与结果评估．实验结果表明，改进后的去噪扩散概率模型在修复随机掩码的人脸图像时，修复结果与原图的PSNR（峰值信噪比）可以达到25.01，SSIM（结构相似性）可以达到0.886，优于改进前的去噪扩散概率模型与现有的基于生成对抗网络的人脸图像修复模型．

关　键　词：深度学习；人脸图像修复；去噪扩散概率模型；快速傅里叶卷积；U-Net网络
中图分类号：TP 391.41　　**文献标志码**：A　　**文章编号**：1005-3026(2024)09-1227-08

Face Inpainting Model Based on Denoising Diffusion Probability Models

LIU Ji-hong, *HUANG Xi-xiong*
(School of Information Science & Engineering, Northeastern University, Shenyang 110819, China. Corresponding author: LIU Ji-hong, E-mail: liujihong@ise.neu.edu.cn)

Abstract: A face inpainting model based on the denoising diffusion probability model is proposed aiming at the problems of poor image quality, blurred repair edges, complex model, and difficult training of the mainstream face inpainting model after image inpainting. By improving the denoising diffusion probability model, the U-Net network structure in Guided-diffusion is adopted. The fast Fourier convolution is introduced into the network, and then the model is trained and tested on the CelebA-HQ high-definition face image dataset. The experimental results show that the improved denoising diffusion probability model can achieve a PSNR of 25.01 and a SSIM of 0.886 compared to the original image, when inpainting face images with random mask, both of which are better than the model before improvement and the existing face image inpainting model based on generative adversarial networks.

Key words: deep learning; face inpainting; denoising diffusion probability models; fast Fourier convolution; U-Net network

面部识别作为一种安全便捷地利用生物特征进行认证和识别的技术，在安全验证、刑侦工作等任务上有重大的应用．然而在进行面部识别时，往往会因为采集到的人脸图像上有遮挡或者模糊而导致识别失败．人脸图像修复是在这种情况下保证识别率的一种重要技术手段．

传统的图像修复算法，如基于目标周围信息的图像修复和补丁匹配算法[1]，虽然可以根据整张图像的语义信息，在待修复区域生成修补补丁，但由于算法表达能力有限，难以完整且精确地捕捉到特征空间中的高级语义和低维特征．因此在修复包含复杂非重复结构的图像（如人脸）时，修复的边缘往往会出现明显的模糊或色彩冲突．

近些年来基于深度学习的方法被广泛应用于图像修复任务中，其修复精度对比传统的算法

收稿日期：2023-05-08
作者简介：刘纪红（1969-），女，辽宁沈阳人，东北大学副教授．

提高了10%以上.在基于深度学习的人脸图像修复算法中,生成对抗网络(generative adversarial networks,GANs)[2]是被运用最多的一种算法.GANs凭借其对抗性的训练过程,使生成器获得了优异的图像生成能力.Li等[3]使用GANs成功对有遮挡人脸图像进行了有效的修复.Yu等[4]提出了一种新型的注意力机制来捕获全局和局部特征,大大提升了在规则性遮挡和随机性遮挡下的人脸图像修复效果.Nazeri等[5]提出了一种两阶段的对抗式边缘连接模型,获得了更好的修复效果.最近的GANs,如AOT-GAN[6]通过将原始卷积核进行拆分,使用不同的子核分别捕捉全局语义和局部特征,进一步提升了生成器的图像修复能力.最近,基于自注意力机制的Transformer模型[7]在各种自然语言处理任务上大放异彩,为了利用该模型中自注意力机制的强大全局建模能力,相关学者将Transformer模型应用于计算机视觉的图像修复任务中,如ViTGAN[8],并取得了较好的效果.但是基于GANs的算法往往在训练时更容易遇到模型不收敛以及模型崩溃的问题.

最近,Ho等[9]通过改进扩散概率模型,提出去噪扩散概率模型(denoising diffusion probability model,DDPM),通过使用神经网络参数化的逆向去噪过程,使得DDPM在图像生成和修复任务中取得了巨大的成功.而且DDPM利用U-Net网络[10]的归纳偏置,在获得更好的特征提取效果的同时,避免了在训练时出现模型崩溃的问题.相比GANs,不但在修复质量上有所提升,而且由于去噪扩散概率模型仅需要使用神经网络进行噪声预测,所以其在训练时无须复杂的损失设计,更容易训练.最近的工作如Guided-diffusion[11],Palette[12]展现了去噪扩散概率模型的优异图像生成能力.

为了进一步提高人脸图像修复效果,本文基于Guided-diffusion中提出的改进的U-Net网络结构,通过对去噪扩散概率模型中的超参数进行调节,强化模型在人脸图像修复任务中的能力.并在U-Net网络浅层引入快速傅里叶卷积[13],利用其能在浅层网络快速捕获图像高级语义信息的优势,增加人脸图像修复模型浅层神经网络的感受野.最后通过消融实验,证明使用快速傅里叶卷积后的去噪扩散概率模型,在人脸图像修复能力上得到了增强.

1 方法描述

本文采用去噪扩散概率模型,对人脸图像数据进行建模并利用模型学习到的数据分布修复被遮挡的人脸图像.为了提升模型的修复质量,引入快速傅里叶卷积.实验分为3个部分:

1)人脸图像数据的获取及人脸图像数据预处理.本文所使用的人脸图像数据集是分辨率为256像素×256像素的CelebA-HQ[14]高清人脸图像开源数据集.在获得数据集后,进行图像增强操作,以提升数据集数据的多样性;

2)人脸图像修复模型的搭建与训练.搭建去噪扩散概率模型,并引入快速傅里叶卷积,然后将经过预处理后的人脸图像数据送入模型进行训练;

3)图像修复指标评估与定性结果分析.模型训练完成后,使用未参与训练的人脸图像数据测试模型人脸图像修复效果,并进行指标评估.

实验流程示意图如图1所示.

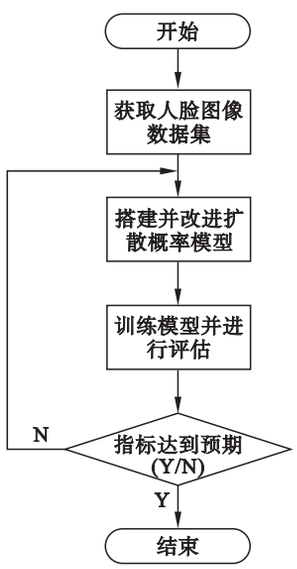

图1 实验流程图

Fig. 1 Experimental flowchart

2 人脸图像修复模型

2.1 去噪扩散概率模型

作为图像生成领域中的新秀,去噪扩散概率模型在各种图像生成任务中都展现了出色的生成能力,同时去噪扩散概率模型还具有稳定的训练过程和更加多样化的图像生成结果.去噪扩散概率模型主要包括前向加噪和反向去噪过程[15].

如图2所示,模型的前向过程是对1张图像逐步添加高斯噪声直至原图变为随机噪声的过程;反向过程是将1幅与原图大小相同的随机噪声进行逐步去噪,直至重新生成为1张图像的过程;反向过程是需要使用神经网络进行噪声预测的部分.

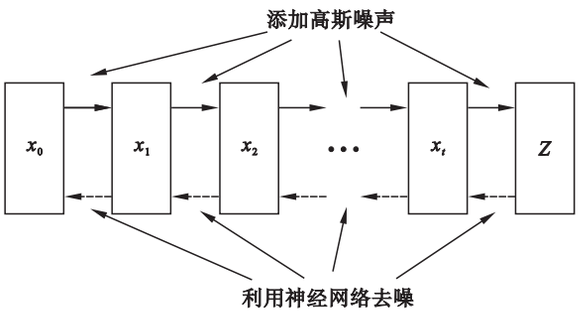

图2　去噪扩散概率模型

Fig. 2　Denoising diffusion probability model

前向的扩散过程共包含t步,每一步都是对上一步得到的数据x_{t-1}按式(1)的方式增加高斯噪声.

$$q\left(x_t|x_{t-1}\right)=N\left(x_t;\sqrt{1-\beta_t}\,x_{t-1},\beta_t I\right). \tag{1}$$

式中:q代表前向过程获得的先验概率分布;N代表高斯噪声;I代表常数;β_t代表前向过程每一步增加噪声的方差,方差由专门的方差方案(variance schedule)确定,通常随着扩散步数的增加,方差也随之增加,一般将方差限定在0~1之间.均值往往设为0,这样,通过确定的方差和均值,就确定了前向扩散中每步对上一步所添加的高斯噪声.这个过程有1个重要的特性,即可以通过式(2)~(3)直接基于原始数据x_0来确定任意t步的x_t.

$$x_t=\sqrt{\alpha_t}\,x_{t-1}+\sqrt{1-\alpha_t}\,\epsilon_{t-1}. \tag{2}$$

式中:α_t等于$1-\beta_t$;ϵ表示随机噪声,通过数学归纳法,式(2)可转化为式(3):

$$x_t=\sqrt{\overline{\alpha}_t}\,x_0+\sqrt{1-\overline{\alpha}_t}\,\epsilon. \tag{3}$$

式中的$\overline{\alpha}_t$代表$\prod_{i=1}^{t}\alpha_i$,由式(3)可知,x_t就是原始数据x_0和随机噪声ϵ的线性组合,组合系数为$\sqrt{\overline{\alpha}_t}$,$\sqrt{1-\overline{\alpha}_t}$,两者的平方和为1.

前向过程是将数据噪声化的过程,反向过程则是去噪的过程.去噪扩散概率模型的反向过程是1个马尔科夫链,由一系列神经网络参数化的高斯分布组成.这个高斯分布的均值和方差由训练的神经网络给出.神经网络的优化目标L可以简化为1个预测噪声的过程[9],如式(4)所示:

$$L_{t-1}=\left[\left\|\epsilon-\epsilon_\theta\left(\sqrt{\overline{\alpha}_t}\,x_0+\sqrt{1-\overline{\alpha}_t}\,\epsilon,t\right)\right\|^2\right]. \tag{4}$$

式中,ϵ_θ表示神经网络预测的噪声,神经网络用来指导梯度下降的损失,即为随机噪声ϵ与ϵ_θ的均方误差.在神经网络进行推理与前向传播时,通过式(5)利用神经网络预测的噪声ϵ_θ以及前向过程获得的参数,逐步将随机噪声恢复为目标图像.

$$x_{t-1}=\frac{1}{\sqrt{\alpha_t}}\left(x_t-\frac{1-a_t}{\sqrt{1-\bar{a}_t}}\epsilon_\theta\left(x_t,t\right)\right)+\sigma_t\epsilon. \tag{5}$$

式中σ_t代表超参数,一般取随采样步数变化的常数.

2.2　快速傅里叶卷积

传统卷积神经网络在搭建时常采用小卷积核尺寸以及多卷积网络层数的方式来构建模型.这样所获得的模型虽然能在更深的卷积层获得较大的感受野,捕捉到更加抽象的高维特征,但是在一些对上下文语义比较敏感的任务(比如人脸图像修复)中,在模型浅层就学习到特征空间中更抽象与更高级的特征是非常必要的.

根据傅里叶变换,在频域中进行卷积处理可以影响到输入特征图中所有的点.快速傅里叶卷积利用频域中的特殊性质,将输入网络的张量,按通道分为两部分,第一部分进行2次相同的普通卷积(卷积核大小为3)计算,以获得局部特征信息,另一部分则使用1次普通卷积,以及1次快速傅里叶卷积来获得全局语义.然后将获得的4张特征图交叉组合,再相加为2个部分,分别代表捕获局部特征信息的局部模块,与获取全局语义信息的全局模块,最后将2个模块的特征图在通道维度拼接起来,构成最后的输出.快速傅里叶卷积示意图如图3所示.

快速傅里叶卷积在频域中进行卷积操作,使得深度学习模型能在靠前的神经网络层中即可学习到全局上下文语义信息.在人脸图像修复任务中,为了使得修复的图像在修复区域与背景更协调,避免结构混乱以及修复边缘处出现伪影,快速傅里叶卷积提供的大感受野能很好地解决这些问题.

2.3　U-Net网络结构

在去噪扩散概率模型中,神经网络具有预测噪声的功能,利用神经网络预测噪声,去噪扩散概率模型逐步对人脸图像中待修复的区域进行填充,最后获得高质量的修复后人脸图像.因此,模型中神经网络的设计将在极大程度上决定去噪扩散概率模型的修复质量与效率.

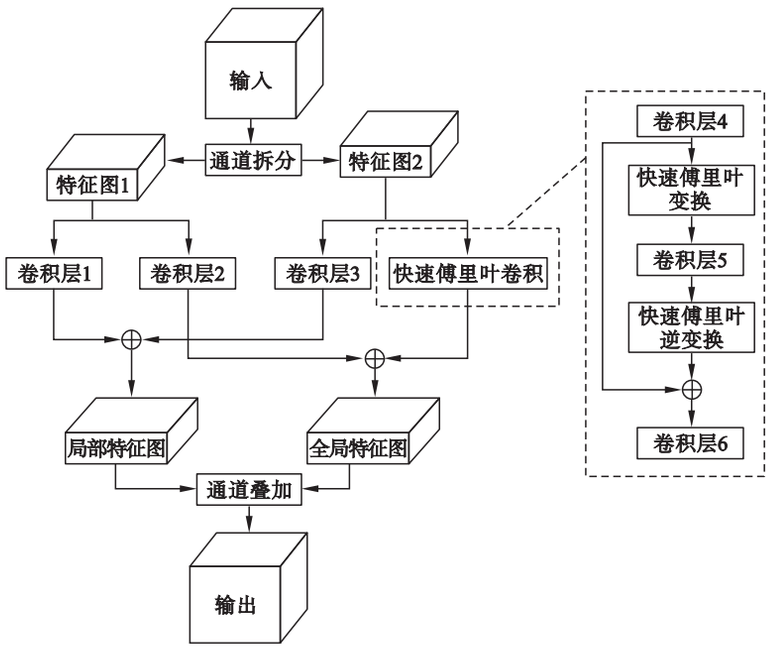

图3 快速傅里叶卷积

Fig. 3 Fast Fourier convolution

在生成式模型中,自编码器模型的网络结构具有最佳的归纳偏置.其往往包含进行下采样操作的编码器以及进行上采样操作的解码器.

U-Net网络一开始被广泛应用于语义分割等任务中,不过由于其同样拥有编码器与解码器,故可以尝试将其作为1种生成式模型使用.相较于其他的生成式网络中的自编码器结构,如图4所示,U-Net在编码器和解码器之间添加了不改变特征图尺寸的中间层.

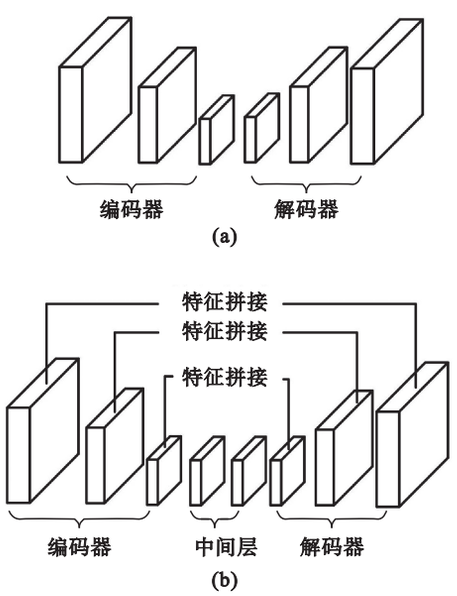

图4 自编码器与U-Net网络结构对比

Fig. 4 Comparison of autoencoder and U-Net structures

(a)—自编码器结构;(b)—U-Net网络结构.

相较其他的自编码器结构,U-Net网络结构被认为更加适合作为去噪扩散概率模型的神经网络结构.在去噪扩散概率模型中,利用深度U-Net网络,对人脸图像进行像素级的建模,获得更好的噪声预测结果,使得修复的图像更加精细,结构更协调.

本文所使用的U-Net网络参考Guided-diffusion中的网络结构,还引入了自注意力模块来提升模型的全局建模能力.参考Suvorov等[16]在图像修复模型的开始将残差模块替换为快速傅里叶卷积的方式,使得模型在浅层的神经网络中即可获得特征空间中的高级全局语义.总体网络结构如图5所示.

人脸图像数据输入U-Net网络后,先对数据进行1次卷积操作,初步提取图像特征并扩充特征图通道数,将数据映射至高维特征空间.然后使用快速傅里叶卷积,在模型浅层即获得特征空间中的高级语义,接着在网络中使用残差网络[17]块,其中卷积层的卷积核大小为3.

自注意力机制关注特征空间中特征的相互影响,其利用相隔较远距离特征之间的相关度来获取全局语义,所以相比普通的卷积网络,自注意力机制具有更大的感受野,这对于人脸图像修复任务很有帮助.但是自注意力层往往需要巨大的计算量,为了减少计算开销,加快模型的推理速度,仅在中间层的低分辨率特征图中运用多头自注意力,其中$\boldsymbol{K}$,$\boldsymbol{Q}$,$\boldsymbol{V}$分别为3个特征矩阵.

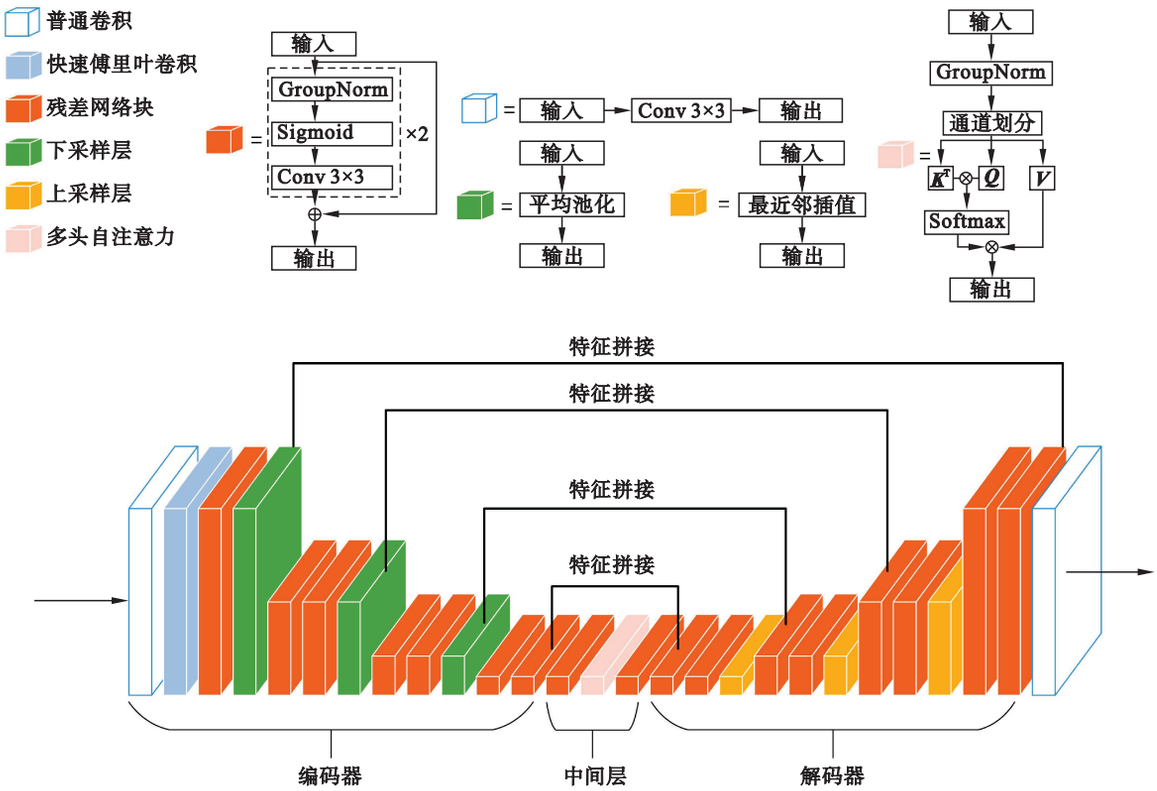

图 5　改进的 U-Net 网络结构

Fig. 5　Improved U-Net network structure

3　实验与分析

3.1　数据集及数据的预处理

本文所使用的数据集为高清人脸图像开源数据集 CelebA-HQ. 该数据集是由 NVIDIA 基于 CelebA 数据集[18]通过训练深度学习模型生成的新的高分辨率人脸图像数据集,一共包含 30 000 张分辨率为 1 024 像素×1 024 像素的高清人脸图像. 由于现实任务中摄像头捕捉到的人脸图像宽度与高度通常在 300 像素以内,故本实验采用经过下采样、分辨率大小为 256 像素×256 像素的 CelebA-HQ 数据集. 实验采用留去法,将数据集分为包含 27 982 张图像的训练集,包含 18 张图像的验证集和包含 2 000 张图像的测试集.

在将图像送入模型训练之前,对人脸图像进行随机的数据增强操作,例如随机裁剪、随机水平翻转和随机色彩变化,以此提高模型的泛化能力.

3.2　模型训练策略

1) 扩散模型前向过程共包含 2 000 步,添加的高斯噪声的方差满足从 10^{-6} 到 10^{-2} 的线性增长.

2) 采用余弦函数衰减策略,使学习率从高到低按余弦曲线随着训练轮数逐渐降低. 这样做的好处是加快模型前期训练速度,并防止模型在后期出现损失震荡的情况.

3) 对式(3)中的 $\overline{\alpha}_t$ 编码并参与神经网络的训练,具体做法为将 $\overline{\alpha}_t$ 由 1 维张量(每次迭代对应 1 个)通过线性层变换为维度与网络每层输出通道数一致的张量,并与网络每层输出相加. 最后作为超参数参与模型推理.

4) Saharia 等[19]的研究表明,使用 L1 范数(即绝对平均值误差)作为损失函数可以在一些图像修复任务中减少修复区域中的伪影,但是原始的去噪扩散概率模型采用的是 L2 范数(即均方误差)来作为损失函数. Palette 中分别使用这两种损失在去噪扩散概率模型上训练,评估结果表明,使用均方误差(mean square error, MSE)作为目标函数能提高模型的样本生成多样性,改善泛化能力. 故本文采用 MSE 作为损失函数,给定分辨率都为 $m\times n$,通道为 R, G, B 的真实图像 I 与含噪图像 K,式(6)表示两者的 MSE:

$$\mathrm{MSE}=\frac{1}{3mn}\sum_{\mathrm{R,G,B}}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}[I_{\mathrm{color}}(i,j)-K_{\mathrm{color}}(i,j)]^2. \tag{6}$$

5) 模型训练时,对数据集中的人脸图像数据添加噪声掩码,将加噪图像输入模型,使用原图与模型输出的修复图像进行损失计算,只计算图像修复区域与原图该区域的均方误差[4],以此节约计算资源.模型前向传播过程如图6所示.

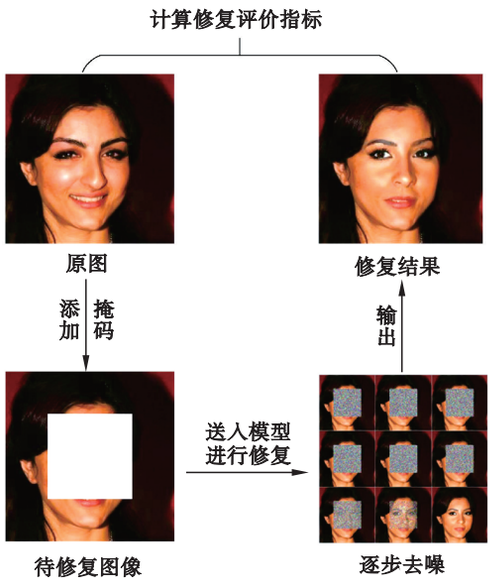

图6 模型前向传播过程

Fig. 6 Model forward propagation process

3.3 评价指标

人脸图像修复的质量在评估时可能会受到非客观因素的影响,为了更加客观地定量评估模型修复质量,需要综合考虑模型在2个评价指标下测定的结果.本文采用图像生成与修复领域最常用的2个评价指标,分别是峰值信噪比[20](peak signal-to-noise ratio, PSNR)与结构相似性[21](structural similarity, SSIM).

PSNR表示最有可能的真实信号分布与干扰该分布的噪声的比值. PSNR利用均方误差(MSE)来定义.则PSNR计算公式为

$$\mathrm{PSNR}=10\cdot\lg\left(\frac{\mathrm{MAX}_{\mathrm{I}}^{2}}{\mathrm{MSE}}\right). \tag{7}$$

其中,$\mathrm{MAX}_{\mathrm{I}}$表示原图中像素灰度值的最大值,由于数据集图像均为8 b图像,故该值取$2^{8}-1$. PSNR值越大,说明模型输出的图像中含有的噪声对图像的影响越小,修复质量越佳.

SSIM比较2个样本之间的明亮度、对比度和结构,其更吻合人类视觉对图像的判断.对真实图像I与含噪图像K,SSIM定义为

$$\mathrm{SSIM}=\frac{(2\mu_{\mathrm{I}}\mu_{\mathrm{K}}+C_{1})(2\sigma_{\mathrm{IK}}+C_{2})}{(\mu_{\mathrm{I}}^{2}+\mu_{\mathrm{K}}^{2}+C_{1})(\sigma_{\mathrm{I}}^{2}+\sigma_{\mathrm{K}}^{2}+C_{2})}. \tag{8}$$

其中:μ_{I}及μ_{K},σ_{I}及σ_{K}分别代表图像I与图像K的平均值和标准差;σ_{IK}代表图像I与图像K的协方差;C_1,C_2取常数. SSIM值越大,代表模型修复的图像越接近原始图像,SSIM为1时,代表图像I与图像K为相同图像.

3.4 实验结果分析

本文在去噪扩散概率模型的基础上,引入了快速傅里叶卷积,提高了模型的人脸图像修复能力.快速傅里叶卷积在神经网络的浅层中即可提取到高级语义,而普通卷积则含有更多的细节信息,需要通过多层卷积神经网络的堆叠,才能获得全局语义.2种卷积输出特征对比如图7所示.

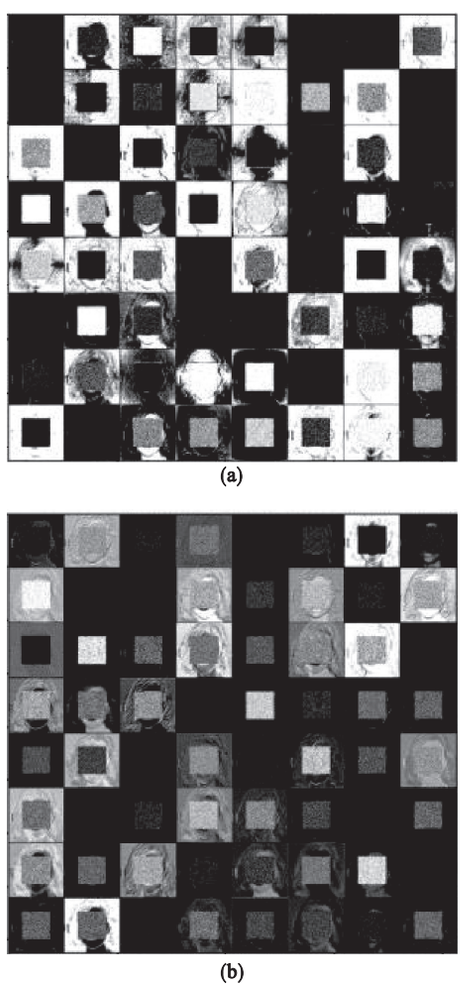

图7 两种卷积输出特征图对比

Fig. 7 Comparison of two convolution outputs

(a)—快速傅里叶卷积; (b)—普通卷积.

在测试集上分别使用改进前和改进后的模型对含有原图25%面积的中心掩码的待修复图像进行修复,并计算其评价指标.结果如表1所示,证明引入快速傅里叶卷积对人脸图像修复任务是有益的.

本文的人脸图像修复模型与其他经典的图

像修复模型,例如 GMCnn[22]、Edge Connect[5]和 AOT-GAN 等,使用相同的数据集进行训练及测试,评估结果如表 2 与表 3 所示.前者采用占原图面积 25% 的中心掩码、后者采用占原图面积 20%~30% 的随机掩码来模拟图像被遮挡的情况.实验结果表明,相比之前的修复模型,本文提出的方法实现了更好的图像修复效果.同时,去噪扩散概率模型避免了训练不稳定的问题,相比传统生成模型成本更低.

表 1　模型改进前后评估指标对比

Table 1　Comparison of evaluation indicators before and after model improvement

模型	PSNR	SSIM
不使用快速傅里叶卷积	24.28	0.873
使用快速傅里叶卷积	24.93	0.882

表 2　在中心掩码下与其他修复模型评估结果对比

Table 2　Comparison of evaluation results with other inpainting models under the center mask

模型	PSNR	SSIM
GMCnn	23.73	0.779
Gated Conv[4]	23.70	0.780
Edge Connect	23.28	0.776
AOT-GAN	24.47	0.849
本文方法	24.93	0.882

表 3　在随机掩码下与其他修复模型评估结果对比

Table 3　Comparison of evaluation results with other inpainting models under the random mask

模型	PSNR	SSIM
GMCnn	23.98	0.781
Gated Conv[4]	24.10	0.790
Edge Connect	23.57	0.794
AOT-GAN	24.83	0.852
本文方法	25.01	0.886

本文模型的局限之处在于推理时需要进行多步去噪操作来获得最终生成的图像,虽然能采用 Song 等[23]的方法大大减少去噪步数,但相比之前的方法,在推理时间上不具有优势.

为了定性地评估本文模型图像修复效果,对模型修复后的人脸图像进行人眼视觉评估,图 8 展示了本文方法分别在中心掩码、随机掩码和大面积随机掩码下的分辨率为 256 像素×256 像素的人脸图像修复效果.

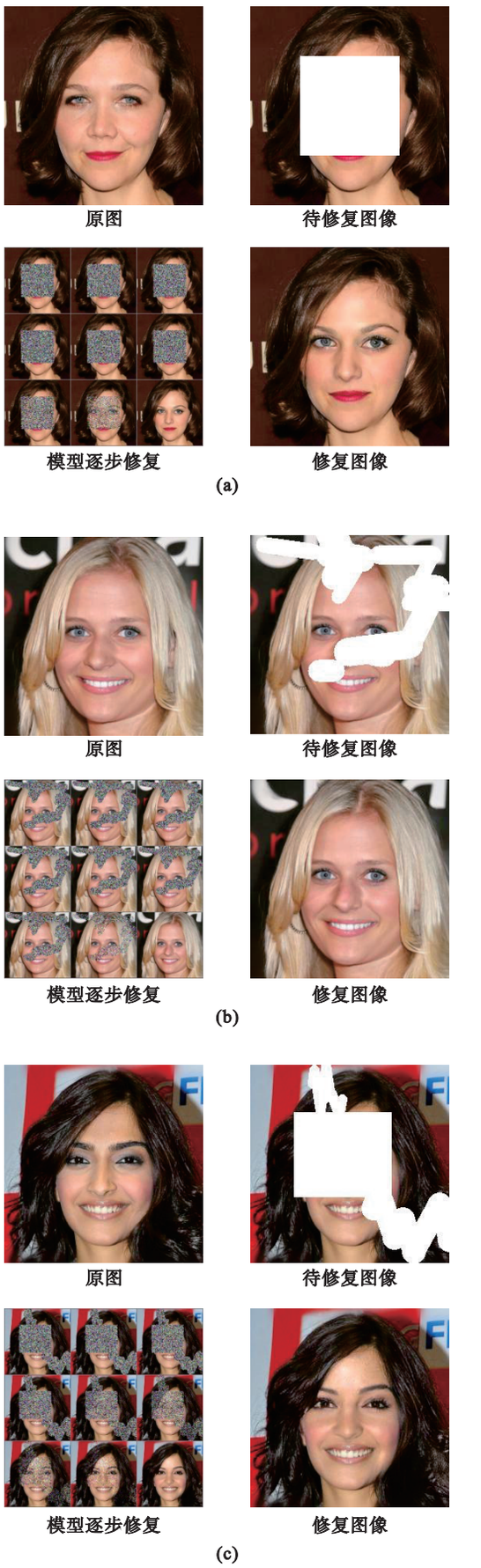

图 8　人脸图像修复结果

Fig. 8　Face inpainting results

(a)—中心掩码;(b)—随机掩码;(c)—大面积随机掩码.

4 结 语

本文聚焦于人脸识别中重要的人脸图像修复技术,为提高获得的人脸图像修复质量,针对修复后图像容易出现伪影和结构混乱的问题,提出了一种基于去噪扩散概率模型的人脸图像修复深度学习模型.为了解决修复后的图像在修复区域出现与全局语义不符的问题,引入快速傅里叶卷积,最后在图像大小为256像素×256像素的高清人脸图像数据集CelebA-HQ上进行模型训练与测试.通过定量评价指标对比,证明本文方法优于改进前的模型以及传统的人脸图像修复模型.结果表明,本文方法在各种掩码下均能实现符合人类视觉感官的人脸图像修复效果,且在细节处没有出现结构混乱和边缘模糊不清的问题,修复图像与原图像的人脸整体形状和五官没有大的出入,因而具有实际应用意义.

参考文献:

[1] 刘颖,佘建初,公衍超,等.基于深度学习的面部修复技术综述[J].计算机应用研究,2021,38(1):9-14.
(Liu Ying, She Jian-chu, Gong Yan-chao, et al. Survey of facial completion techniques based on deep learning [J]. *Application Research of Computers*,2021,38(1):9-14.)

[2] 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. Montreal, 2014: 2672-2680.

[3] Li Y J, Liu S F, Yang J M, et al. Generative face completion [C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, 2017: 5892-5900.

[4] Yu J H, Lin Z, Yang J M, et al. Generative image inpainting with contextual attention [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, 2018: 5505-5514.

[5] Nazeri K, Eric N, Joseph T, et al. Edge connect: generative image inpainting with adversarial edge learning [J]. *arXiv Preprint arXiv*, 2019: 1901.00212.

[6] 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.

[7] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, 2017: 6000-6010.

[8] Lee K, Chang H W, Jiang L, et al. ViTGAN: training GANs with vision transformers [J]. *arXiv Preprint arXiv*, 2021: 2107.04589.

[9] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models [C]//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, 2020: 6840-6851.

[10] Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation [C]//Medical Image Computing and Computer-Assisted Intervention-MICCAI 2015. Munich, German: Springer International Publishing, 2015: 234-241.

[11] Dhariwal P, Nichol A. Diffusion models beat gans on image synthesis [J]. *Advances in Neural Information Processing Systems*, 2021, 34: 8780-8794.

[12] Saharia C, Chan W, Chang H W, et al. Palette: image-to-image diffusion models [C]//SIGGRAPH '22: ACM SIGGRAPH 2022 Conference Proceedings. Vancouver, 2022: 1-10.

[13] Chi L, Jiang B, Mu Y. Fast Fourier convolution [J]. *Advances in Neural Information Processing Systems*, 2020, 33: 4479-4488.

[14] Karras T, Aila T, Laine S, et al. Progressive growing of gans for improved quality, stability, and variation [J]. *arXiv Preprint arXiv*, 2017: 1710.10196.

[15] Yang L, Zhang Z L, Song Y, et al. Diffusion models: a comprehensive survey of methods and applications [J]. *arXiv Preprint arXiv*, 2022: 2209.00796.

[16] Suvorov R, Logacheva E, Mashikhin A, et al. Resolution-robust large mask inpainting with Fourier convolutions [C]//2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Waikoloa, 2022: 3172-3182.

[17] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition [C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, 2016: 770-778.

[18] Liu Z W, Luo P, Wang X G, et al. Large-scale celebfaces attributes (CelebA) dataset [DS/OL]. [2022-11-01]. https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html.

[19] Saharia C, Ho J, Chan W, et al. Image super-resolution via iterative refinement [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023, 45(4): 4713-4726.

[20] Voronin V V, Sizyakin R A, Marchuk V I, et al. Video inpainting of complex scenes based on local statistical model [J]. *Electronic Imaging*, 2016, 28(15): 1-6.

[21] 何凯,牛俊慧,沈成南,等.基于SSIM的自适应样本块图像修复算法[J].天津大学学报(自然科学与工程技术版),2018,51(7):763-767.
(He Kai, Niu Jun-hui, Shen Cheng-nan, et al. Image inpainting algorithm with adaptive patch using SSIM [J]. *Journal of Tianjin University (Science and Technology)*, 2018, 51(7): 763-767.)

[22] Wang Y, Tao X, Qi X J, et al. Image inpainting via generative multi-column convolutional neural networks [C]// Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montréal, 2018: 329-338.

[23] Song J M, Meng C L, Ermon S. Denoising diffusion implicit models [J]. *arXiv Preprint arXiv*, 2020: 2010.02502.