DOI:10.19651/j.cnki.emt.2517904

# 基于扩散模型的二阶段细化图像修复模型*

张 绪 胡峻峰
(东北林业大学计算机与控制工程学院 哈尔滨 150040)

**摘 要**:近年来,图像生成领域的技术取得了显著进展,但图像修复任务中修复区域与未修改区域之间的一致性仍是一个普遍存在的挑战。本文旨在提出一种基于扩散模型的两阶段图像修复模型(Diff-2sIR),以提升修复区域与未修复区域的一致性,进一步提高图像修复质量。本文以扩散模型理论为基础,设计了一种两阶段修复框架。通过改进 U-Net 架构和扩散模型采样算法,对初步修复结果进行二次细化修复,缓解了修复区域与未修复区域之间的不一致性问题。在 CelebA-HQ 数据集的人脸修复任务中,Diff-2sIR 模型取得了最优 FID 分数(2.92),显著提升了修复质量。实验结果表明,该模型在指导模块修复结果的基础上进一步细化修复效果,展现了卓越的性能。本文提出的 Diff-2sIR 模型有效解决了修复区域与未修复区域之间的不一致性问题,为图像修复任务提供了一种新的解决方案,具有重要的理论意义和应用价值。

**关键词**:图像修复;图像生成;扩散模型;二阶段模型
**中图分类号**:TP391.41;TN0 **文献标识码**:A **国家标准学科分类代码**:520.2060

# Diff-2sIR: Diffusion-based refinement two-stage image restoration model

Zhang Xu Hu Junfeng
(College of Computer Science and Control Engineering, Northeast Forestry University, Harbin 150040, China)

**Abstract**: In recent years, significant progress has been made in the field of image generation, but the consistency between the repaired and unmodified regions remains a common challenge in image inpainting tasks. This paper proposes a two-stage image inpainting model based on diffusion models (Diff-2sIR) to enhance the consistency between the repaired and unmodified regions, thereby improving the overall quality of image inpainting. Based on the theory of diffusion models, a two-stage inpainting framework is designed. By improving the U-Net architecture and the diffusion model sampling algorithm, the initial inpainting results are further refined in a second stage, alleviating the inconsistency between the repaired and unmodified regions. In the face inpainting task on the CelebA-HQ dataset, the Diff-2sIR model achieves the best FID score (2.92), significantly improving the inpainting quality. Experimental results show that the model further refines the inpainting results based on the guidance module, demonstrating exceptional performance. The Diff-2sIR model effectively addresses the inconsistency between the repaired and unmodified regions, providing a new solution for image inpainting tasks, with significant theoretical and practical implications.

**Keywords**: image inpainting;image generation;diffusion model;two-stage model

## 0 引 言

图像修复(image inpainting)旨在根据图像中的上下文信息重建缺失或受损区域,生成视觉自然且语义一致的完整图像。传统的图像修复方法主要依赖于低级视觉线索,例如文献[1]提出的偏微分方程(partial differential equations, PDEs)和文献[2]提出的纹理合成技术。PDEs 方法利用数学模型对缺失区域进行像素填充,保持局部结构的平滑性;纹理合成利用样本特征进行像素生成,有着较好的局部细节修复性能。然而,这些方法在处理复杂语义信息和大范围缺损时,难以对全局信息进行建模,导致大量语义错误,这些语义错误是限制修复质量的重要因素之一。

与传统方法相比,生成模型能够更好地建模图像中的

收稿日期:2025-01-14
* 基金项目:国家自然科学基金(32202147)、中央高校基本科研专项资金项目(2572019BF09)资助

像素细节与高层语义信息,从而在复杂场景下展现出更强的修复能力。

早期基于卷积神经网络(CNN)的方法,例如文献[3]提出的扩张卷积(dilated convolution)、文献[4]提出的部分卷积(partial convolution)和文献[5]提出的门控卷积(gated convolution)等技术提高了对上下文信息的捕捉能力,但在处理具有复杂纹理和语义结构的图像时仍显得不足。文献[6]中提出的 WavePaint 设计了基于 WaveMix 的全卷积架构,进一步减少了训练和修复计算开销。文献[7]提出了生成对抗网络(generative adversarial network, GAN),GAN 的引入,推动了图像修复领域的发展。例如文献[8]提出的 Context Encoder 是最早将 GAN 应用于图像修复的研究之一,文献[9]提出的 EdgeConnect 通过加强边缘约束,显著提升了修复性能。而文献[10]提出的 LBAM 提出的双向注意力图进一步提高了修复质量。文献[11]提出的 LaMa 利用频域操作提升了模型恢复高频细节的能力。

此外,文献[12]提出的 Transformer 架构也被引入到图像修复任务中,相关工作包括:文献[13]提出基于 Transformer 的结构信息重建的 ICT 方法,能够有效恢复图像语义结构;文献[14]提出的 BAT-Fill 构建了双向 Transformer 架构,通过双向上下文感知提升图像修复的语义一致性;文献[15]提出的 MAT 则创新性地引入多头上下文注意力机制,在提升图像修复质量的同时,显著增强了生成结果的多样性特征。

近年来,扩散模型(diffusion model, DM)在图像生成任务中的优势逐渐显现,并成功应用于图像修复。其通过逐步去噪生成高质量图像。文献[16]提出了 DDPM 模型,该模型使用神经网络对噪声进行预测,并对扩散理论进行了完善。后续研究如文献[17]提出的 Score-based SDE 和文献[18]提出的 DDIM 均对采样过程进行了优化,显著提升了效率。同时,DM 的可控生成能力也引起了广泛关注,文献[19]提出 Guided Diffusion 和文献[20]提出 Stable Diffusion 引入条件控制和注意力机制,控制生成过程进一步改善了生成效果。

在图像修复任务中,DM 展现出了卓越的性能。例如,文献[21]提出的 SDEdit 通过优化采样过程,修复质量取得了显著提升;文献[22]提出的 DDNM 在反向过程中细化零空间内容,减少了修复中的语义错误;文献[23]提出的 DiffIR 通过引入 IR 先验提取网络和动态 IR 变换器,通过双阶段训练流程实现高效的图像修复;文献[24]提出的 M2S 提出从粗至细的采样策略,显著提升修复效率;文献[25]提出的 Palette 方法在图像修复、图像超分等多个修复任务中表现出色。尽管如此,语义一致性问题仍然是 DM 在图像修复应用中的一大挑战。

虽然基于深度学习的图像修复技术取得了诸多成就,修复区域与未修复区域之间在细节和语义上的一致性问题依然突出,且有些模型需要一定的计算成本。针对这些问题,本文提出了一种基于扩散模型的图像修复方法,同时探讨了采样策略与噪声表设计对修复效果的影响。通过这些改进,有效缓解了语义不一致问题,并提升了修复质量,推动图像修复技术的进一步发展。

## 1　方法与实验

在本节中,首先回顾经典扩散理论,然后详细介绍本文提出的方法。

### 1.1　预备知识:扩散模型

扩散模型在图像生成、音频合成、图像超分、图像翻译及图像编辑等任务中取得了显著进展。扩散模型主要包括前向加噪过程和反向去噪过程。DDPM 将前向过程定义为一个马尔科夫链加噪过程:

$$q(x_t \mid x_{t-1}) = \mathcal{N}(x_t; \sqrt{\alpha_t}x_{t-1}, (1-\alpha_t)\varepsilon) \tag{1}$$

其中,$\varepsilon \sim \mathcal{N}(0, I)$。通过不断向样本中添加高斯噪声:

$$q(x_{1:T} \mid x_0) = \prod_{t=1}^{T} q(x_t \mid x_{t-1}) \tag{2}$$

图像的原始语义信息被逐步抹除,最终加噪后的样本近似服从于高斯分布 $x_t \sim \mathcal{N}(0, I)$,借助高斯分布的可加性,边缘化前向过程可以得到:

$$q(x_t \mid x_0) = \mathcal{N}\left(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\varepsilon\right) \tag{3}$$

其中,$\bar{\alpha}_t = \prod_{s=0}^{t} \alpha_s$。$q(x_t \mid x_0)$ 中包含了前向过程的所有信息,在给定样本 $x_0$、前向噪声调度参数 $\alpha_t$ 以及时间步 $t$ 的情况下,通过重参数化边缘分布:

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{(1-\bar{\alpha}_t)}\,\varepsilon, \varepsilon \sim \mathcal{N}(0, I) \tag{4}$$

可以采样出前向过程中任意时间步的 $x_t$。进一步,在反向过程中,依据当前步 $x_t$ 来预测 $x_{t-1}$,结合式(3)和贝叶斯公式,此时可以从式(5)得到后验分布。

$$q(x_{t-1} \mid x_t, x_0) = q(x_t \mid x_{t-1}) \frac{q(x_{t-1} \mid x_0)}{q(x_t \mid x_0)} \tag{5}$$

简化为:

$$x_{t-1} \sim \mathcal{N}(x_{t-1}; \mu, \sigma^2 I) \tag{6}$$

其中,均值 $\mu = \dfrac{\sqrt{\bar{\alpha}_t}(1-\alpha_t)}{1-\bar{\alpha}_t}x_0 + \dfrac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t$,以及方差 $\sigma^2 = \dfrac{(1-\bar{\alpha}_{t-1})(1-\alpha_t)}{1-\bar{\alpha}_t}$。显然,$x_{t-1}$ 的分布依赖 $x_0$ 以及 $x_t$,然而 $x_0$ 是未知的。实际上 DDPM 中首先通过优化目标式(7)来缩小近似去噪转移分布 $p_\theta(x_{t-1} \mid x_t)$ 和真实去噪转移分布 $q(x_{t-1} \mid x_t, x_0)$ 的偏差。

$$\arg\min_{\theta} \mathcal{D}_{KL}(q(x_{t-1} \mid x_t, x_0) \,||\, p_\theta(x_{t-1} \mid x_t)) \tag{7}$$

直至:

$$p_\theta(x_{t-1} \mid x_t) \approx q(x_{t-1} \mid x_t, x_0) \tag{8}$$

由于 $x_0$ 未知,因此 DDPM 假设在反向过程中,$x_t$ 与 $x_0$ 依旧遵循与式(4)近似的映射关系。最终简化式(7)为预

测噪声：

$$L_{simple}=\mathbb{E}_{x_0,t,\varepsilon}\varepsilon-\varepsilon_\theta(x_t,t)^2 \tag{9}$$

在进行推理时，给定含噪图像 $x_t$ 和时间索引 $t$，模型预测噪声 $\varepsilon_t$，然后从式(10)中采样 $x_0$，最终由式(8)得到 $x_{t-1}$。

$$p_\theta(x_0\mid x_t)=\mathcal{N}\left(x_0;\frac{1}{\sqrt{\bar{\alpha}_t}}x_t,\left(\sqrt{\alpha_t}-\frac{1}{\sqrt{\bar{\alpha}_t}}\right)\varepsilon_t\right) \tag{10}$$

从一个标准高斯分布开始，执行 $T$ 步，直至采样得到 $x_0$，推理结束。除了预测噪声 $\varepsilon$，DDPM 中也提到了直接预测 $x_0$，而 progressive distillation[26] 则通过数学变换将预测噪声变成预测间接目标 $v$，一定程度上提高了模型的训练稳定性和生成效果，具体如何实施这里不再赘述。

### 1.2　Diff-2sIR

相较于 PDEs 等方法，Diff-2sIR 的优势如图 1 所示。

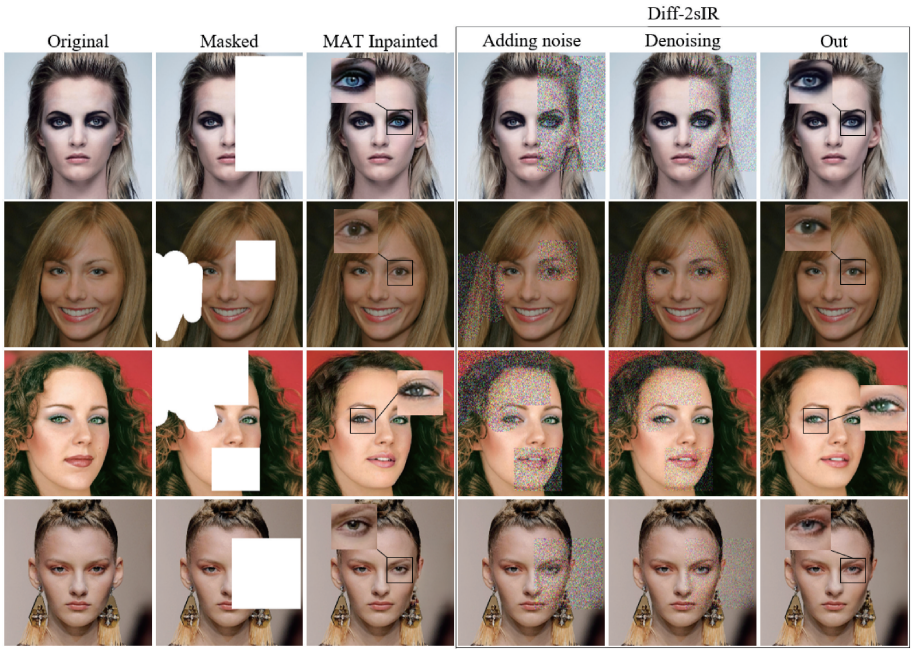



图 1　Diff-2sIR 效果展示

Fig. 1　Diff-2sIR results demonstration

在说明 Diff-2sIR 之前，首先给出下文中将会使用到的一些定义：观测数据 $x$，掩码 $m$，$x_{unkown}=m\odot x$ 和 $x_{kown}=(1-m)\odot x$ 分别表示未知像素和已知像素。将退化观测数据 $x_{kown}$ 由高斯噪声填充后表示为：

$$x_m=m\odot\varepsilon_m+(1-m)\odot x_0 \tag{11}$$

其中，$\varepsilon_m\sim\mathcal{N}(0,I)$。

图像修复的本质是一类逆向求解问题。而其中的关键是解决语义一致性问题，为了缓解这个问题，本文提出了：Diff-2sIR，一种基于扩散模型细化的二阶段图像修复模型(diffusion-based refinement two-stage image restoration model，Diff-2sIR)。Diff-2sIR 由一个指导模块和一个细化模块组成。其中细化模块基于扩散模型实现，而指导模块则较为灵活，可以由任何具备修复能力的模型构成。

细化模块的训练与一般扩散修复模型类似。如算法 1 所示，给定观测样本 $x_0$，在经过式(4)加噪后得到对应的噪声样本 $x_t$。模型训练的目标是根据已知像素 $x_{kown}$ 推测未知像素 $x_{unkown}$，使损毁样本恢复到接近原始图像 $x_0$。与一般扩散修复模型将掩码作为输入不同的是，本文直接使用 $x_m$ 作为掩码信息的输入。在训练过程中，使用式(12)作为优化目标来预测噪声向量 $\boldsymbol{\varepsilon}$，其中 $m$ 为二进制掩码，这样将预测目标经过掩码处理，可以有效减少模型反向传播时的计算量。

$$\mathbb{E}_{x_0,\varepsilon,\bar{\alpha}}\parallel m\odot f_\theta(x_m,\bar{x},\bar{\alpha})-m\odot\boldsymbol{\varepsilon}\parallel \tag{12}$$

**Algorithm 1** Training process of Diff-2sIR

**repeat**
　$x_0\sim p(x_0)$
　$m\sim p(m)$
　$x_t=\sqrt{\bar{\alpha}_t}x_0+\sqrt{1-\bar{\alpha}_t}\varepsilon,\varepsilon\sim\mathcal{N}(\mathbf{0},\boldsymbol{I})$
　$x_m\leftarrow m\odot\varepsilon_m+(1-m)\odot x_0,\varepsilon_m\sim\mathcal{N}(\mathbf{0},\boldsymbol{I})$
　$\bar{x}\leftarrow m\odot x_t+(1-m)\odot x_0$
　Take a gradient descent step on
　　$\nabla_\theta\parallel m\odot f_\theta(x_m,\bar{x},\bar{\alpha})-m\odot\varepsilon\parallel$
**until** converged

在推理时，DDPM 通常从纯高斯噪声开始采样，即 $x_t\sim\mathcal{N}(0,I)$。采样前期，大尺度图像特征就已基本确定，采样中后期模型主要倾向于围绕这些特征生成像素细节[27]。Diff-2sIR 通过在潜空间中对整个采样过程进行控制，细化修复不仅能够保证指导模块初步修复的质量，还能进一步提升修复质量，在一定程度上缓解语义不一致问题。

如图 2 所示，Diff-2sIR 主要包括两部分：指导模块和

细化模块。其中前者作为串联框架的第一部分,其本质上是对待修复样本 $x_{known}$ 进行初步修复。在算法 2 中,如果使用扩散修复模型作为指导模块,首先由式(8)在采样 $T_g$ 步后输出初步修复结果 $\bar{x}_0$,然后再由式(4)将 $\bar{x}_0$ 加噪到 $T$ 步得到 $x_T$,并将 $x_T$ 初始化为细化模块的采样先验分布。指导模块非扩散修复模型的情况下(Diff-2sIR$^+$),如算法 3 所示,在获得 $\bar{x}_0$ 后,推理时设置中间步 $T_k \in [1,T]$,然后再由式(13)采样得到 $x_{T_k-1}$。重复操作直至修复完成。在实验中本文对中间步采样与起始位置采样策略进行了对比。

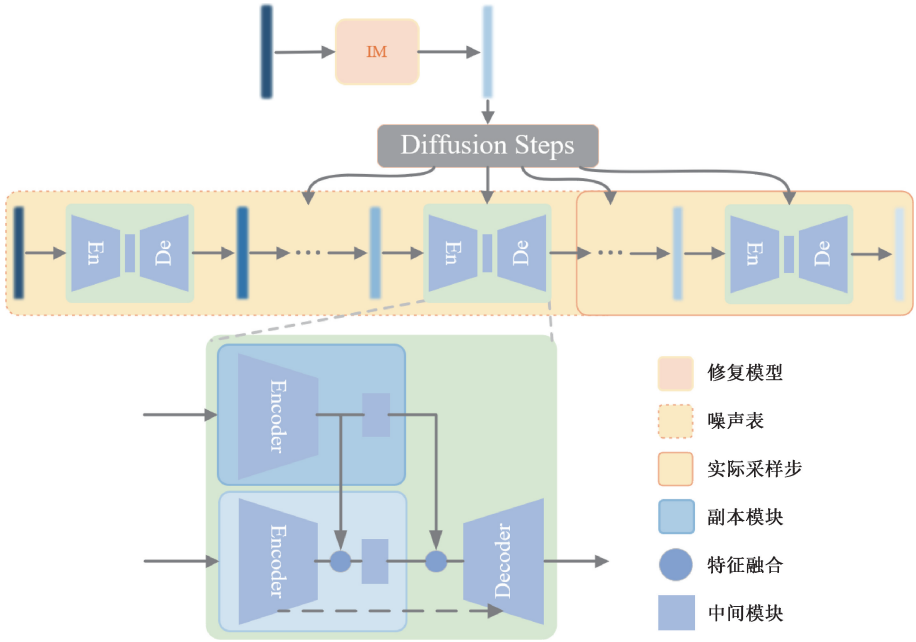



图 2　Diff-2sIR 原理介绍图

Fig. 2　The illustration of the principles of Diff-2sIR

$$x_{t-1}=\frac{1}{\sqrt{\alpha_t}}\left(x_t-\frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}f_\theta(x_m,x_t,\bar{\alpha}_t)\right)+\sqrt{1-\alpha_t}\varepsilon \tag{13}$$

**Algorithm 2** Sampling process of Diff-2sIR

**Stage 1:**
**if** Utilize diffusion models **then**
  $x_{T_g}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I})$
  **for** $t=T_g,\dots,1$ **do**
    $\varepsilon'\sim\mathcal{N}(\mathbf{0},\boldsymbol{I})$ if $t>1$,else $\varepsilon'=0$
    $x'_{t-1}=\frac{1}{\sqrt{\alpha'_t}}(x'_t-\frac{1-\alpha'_t}{\sqrt{1-\bar{\alpha}'_t}}f'_\theta(x_m,x'_t,\bar{\alpha}'_t))+\sqrt{1-\alpha'_t}\varepsilon'$
    $x'_{t-1}=m\odot x'_{t-1}+(1-m)\odot x_0$
  **end for**
**else**
  Perform restoration using other models
**end if**
**return** $\bar{x}_0$
**Stage 2:**
$x_T=\sqrt{\bar{\alpha}_T}\bar{x}_0+\sqrt{1-\bar{\alpha}_T}\varepsilon_T,\varepsilon_T\sim\mathcal{N}(\mathbf{0},\boldsymbol{I})$
**for** $t=T,\dots,1$ **do**
  $\varepsilon\sim\mathcal{N}(\mathbf{0},\boldsymbol{I})$ if $t>1$,else $\varepsilon=0$
  $x_{t-1}=\frac{1}{\sqrt{\alpha_t}}(x_t-\frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}f_\theta(x_m,x_t,\bar{\alpha}_t))+\sqrt{1-\alpha_t}\varepsilon$
  $x_{t-1}=m\odot x_{t-1}+(1-m)\odot x_0$
**end for**
**return** $x_0$

**Algorithm 3** Sampling process of Diff-2sIR+

**Stage 1:**
Same as **Algorithm 2**
**Stage 2:**
$x_T=\sqrt{\bar{\alpha}_t}\bar{x}_0+\sqrt{1-\bar{\alpha}_t}\varepsilon_T,\varepsilon_T\sim\mathcal{N}(\mathbf{0},\boldsymbol{I})$
Let $T_k\in\{1,2,\dots,T\}$
**for** $t=T_k,\dots,1$ **do**
  $\varepsilon\sim\mathcal{N}(\mathbf{0},\boldsymbol{I})$ if $t>1$,else $\varepsilon=0$
  $x_{t-1}=\frac{1}{\sqrt{\alpha_t}}(x_t-\frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}f_\theta(x_m,x_t,\bar{\alpha}_t))+\sqrt{1-\alpha_t}\varepsilon$
  $x_{t-1}=m\odot x_{t-1}+(1-m)\odot x_0$
**end for**
**return** $x_0$

第二部分则是 Diff-2sIR 的核心:细化模块。受 ControlNet[28] 启发,在本文的设想中,细化模块不需要再重新训练一个扩散修复模型,或者再从头设计并训练一个控制接口模块,本文仅仅利用作为指导模块的扩散修复模型本身,再对其进行简单的修改,这样需要消耗的成本相对来说大大降低。然而,目前还没有类似的工作这样做过,幸运的是,本文利用扩散模型中 Unet 本身的一些特性,经过简单的设计修改,得到了一个可以供外部控制的改进 Unet 架构,其相比于原始 Unet 多了一个接口模块,这个模块不是从头设计一个模型,也不需要重新训练或者微调,而是利用其本身原有的模块在架构上做了一些调整。具体来说,细化模块作为被控制的一方,其推理不再从 $\mathcal{N}(0,\mathrm{I})$ 开始采样,而是变成从分布开始采样:

$$q(x_T \mid \bar{x}_0) = \mathcal{N}\left(\sqrt{\bar{\alpha}_T}\bar{x}_0, \sqrt{1-\bar{\alpha}_T}\varepsilon_T\right) \tag{14}$$

其中，$\varepsilon_T \sim \mathcal{N}(0, \boldsymbol{I})$。除此之外，还额外接收了指导模块的输出 $\bar{x}_0$ 作为输入，同样，本文对 $\bar{x}_0$ 进行了加噪处理，将其加噪到当前时间步得到 $\bar{x}_t$，然后再输入到细化模块中。Asyrp[29] 在 $h$ 空间中引入一个 $f_t$ 模块接收最深层特征 $h_t$，制造 $\triangle h_t$ 偏差，进而实现对图像的编辑，受此启发，在 Diff－2sIR 细化模块内部，本文使用 Unet 编码器 $\mathcal{F}_\varepsilon(\cdot)$ 和中间模块 $\mathcal{F}_M(\cdot)$ 的副本 $\mathcal{F}_{\varepsilon'}(\cdot)$ 及 $\mathcal{F}_{M'}(\cdot)$ 来接收 $\bar{x}_0$，定义 $\mathcal{F}_\varepsilon(\cdot)$ 的输出空间为 $h_{-1}$ 空间，为了实现对采样过程的控制，将指导模块的输出信息融入到当前采样过程中，本文直接将副本的输出在对原始编码器以及中间模块的输出在 $h_{-1}$ 以及 $h$ 空间进行通道的替换：

$$Output = \mathcal{F}_M(Cat\left[\mathcal{F}_\varepsilon(\mathrm{x_m}, \mathrm{x_t})^1, \mathcal{F}_{\varepsilon'}(\mathrm{x_m}, \bar{\mathrm{x}}_\mathrm{t})^2\right]) \tag{15}$$

$$Output_M = Cat\left[Output^1, \mathcal{F}_{M'}(\mathcal{F}_{\varepsilon'}(\bar{x}_t))^2\right] \tag{16}$$

$$\varepsilon_{pred} = \mathcal{F}_D(Output_M) \tag{17}$$

其中，上标 1 和 2 分别表示通道的前半部分和后半部分。进行两次拼接后最终得到中间模块的输出 $Output_M$，再输入到模型的解码器 $\mathcal{F}_{\mathcal{D}}(\cdot)$ 中得到预测噪声 $\varepsilon_{pred}$，进一步由式(3)及式(10)采样得到 $x_0$ 和 $x_t$，最终由式(6)得到 $x_{t-1}^{unkown}$。采样时本文保证了模型的输入和训练时一致，因此：

$$x_{t-1} = m \odot x_{t-1}^{unkown} + (1-m) \odot x_0 \tag{18}$$

注意，与 ControlNet 和 Asyrp 不同，在整个过程中，没有重新再训练或者微调一个模型，也没有添加任何模块，只是对扩散修复模型的架构进行了上述调整，利用原始模型 Unet 的编码以及瓶颈模块的副本实现了这个过程，而关于原始模型的训练则与上文中的一致，没有任何不同。并且，实际上指导模块的训练以及推理相对于细化模块是解耦的，这赋予了指导模块极大的自由度，本文的实现中不仅使用了原始模型作为指导模块，事实上在实验中还尝试了一些其他主流修复模型作为指导模块，这极大增加了 Diff-2sIR 的延续性。在实际修复任务中，如果对修复结果不满意，Diff-2sIR 保留了细化模块的输入接口，将修复图像输入至细化模块，即可对初步修复的结果做进一步的细化修复。

### 1.3 实　　验

本文在 CelebA-HQ[30] 数据集上验证了 Diff-2sIR，实验中采用分辨率为 256 的版本，实验使用的数据集划分与 MAT 基本一致，其中 24 181 张图像作为训练集，2 张图像作为验证集，选取 2 993 张图像作为测试集。为保证模型对掩码的鲁棒性和泛化能力以及方便验证本模型的性能，Diff-2sIR 的训练采用与 MAT 一样的掩码策略，即使用大掩码在图像上随机涂抹，测试使用 MAT 提供的掩码进行测试。同样，本文选择以下指标对 Diff-2sIR 的性能进行评估：

1）FID(frechet inception distance)[31]：FID 通过计算生成图像与真实图像在 Inception-v3 特征空间中的分布距离，综合评估生成样本的质量和多样性，其值越低表明生成效果越接近真实数据分布；

2）P-IDS/U-IDS(paired/unpaired inception discriminative s-core)[32]：P-IDS/U-IDS 是一种基于预训练特征空间中线性可分性的评估指标，旨在衡量生成图像与真实图像之间的区分难度。该方法利用预训练的 Inception v3 模型提取图像特征，并通过线性支持向量机(SVM)拟合特征，以反映特征空间中的线性可分性；

3）IS(inception score)[33]：IS 通过 Inception 分类器预测结果的熵值，平衡生成样本的质量与多样性，其值越高表明模型同时具备良好的类间多样性和类内一致性。

关于实验的具体实施细节，首先，对于模型训练，本文采用 UNet 基础超参数如下：输入分辨率为 256，输入通道数为 6，输出分辨率为 256，输出通道数为 3，UNet 总共进行 3 次下采样，从分辨率 256 到 32，且仅在瓶颈处使用注意力机制。前向过程的噪声表使用线性表，即 β 从 0.000 001 到 0.01，总共扩散 2 000 步。模型训练时的 batch size 设置为 3，优化器使用 Adam，初始学习率为 $5\times10^{-5}$。训练实验环境如下：1）硬件环境：一张 NVIDIA RTX4090 GPU，内存 32 GB；2）软件环境：操作系统 Ubuntu22.04 LTS，Python3.10，torch 2.0.1(cu118)。

其次，在推理阶段，本文并未采用与原始 Palette 模型一致的噪声表配置(从 0.000 1～0.09 的 1 000 步线性噪声表)，而是使用与训练时一致的噪声表(从 0.000 001 到 0.01 的线性噪声表)。实验中还分别设置了 $T$ = 2 000，1 500，1 000，500 不同步数进行修复。对于 Diff-2sIR$^+$，设置 $T_k$ = 100，150，…，500 进行实验。推理的 batch size 为 8，测试实验环境如下：1）硬件环境：一张 NVIDIA RTX4080 GPU，其余与训练实验一致。2）软件环境：与训练实验一致。

在指导模块和细化模块均使用 Palette 基础模型的情况下，指导模块采用 500 步的线性噪声表(从 0.000 1～0.09)，细化模块的噪声表分别使用从 100 至 500 步，中间间隔为 50 步的设置。当指导模块使用其他模型时，细化模块的噪声表保持不变。在推理实验中，采用 MAT 官方代码库提供的两种分辨率为 256×256 的掩码(即小掩码和大掩码)进行测试。

## 2 结　　果

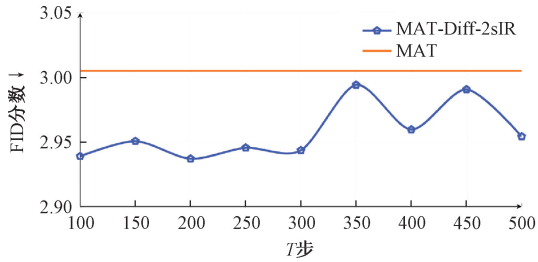
本节将会探索模型中的一些参数对修复效果的影响。在 2.1 节探索初步修复结果优劣对细化模块修复效果的影响；2.2 节主要探讨两种噪声表策略对修复效果的影响。

### 2.1 指导模块对修复效果的影响

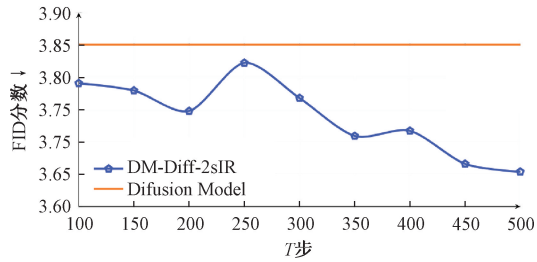
在图 3 中，分别比较了两种指导模块的性能对细化模

块修复效果的影响。实验基于 CelebA-HQ 数据集,两者细化模块参数一致,且均使用小掩码进行测试,扩散模型作为指导模块时设置 $T_g = 500$ 步,MAT 使用其论文中配置,两者细化模块参数保持一致,使用小掩码进行测试。FID 分数如图 3(a)和(b)所示。

当 MAT 作为指导模块时,初步修复的 FID 分数(图 3(a)MAT 曲线)优于扩散指导模块(图 3(b)Diffusion Model 曲线)。尽管两者性能有所差距,Diff-2sIR 的细化模块依然在两种情况下展现了强大的再修复能力:对于 MAT,$T = 100$ 时 FID 从 3.00 降至 2.93(如图 3(a)所示);对于扩散指导模块,FID 从 3.85 降至 3.79(如图 3(b)所示)。这一现象表明,即使初步修复效果较差,Diff-2sIR 依然能够显著提升修复质量,展现出极强的鲁棒性和适应性。此外,这也证明了 Diff-2sIR 不仅适配高质量初步修复输入,还能从低质量输入中提取信息并优化结果,从而在面对不同指导模块时具有更高的灵活性。在图 3(a)中,MAT-Diff-2sIR 曲线随着 $T$ 的提升,以 $T = 300$ 为分界线呈现总体向上的趋势,而图 3(b)中则是呈现总体向下的趋势。导致这一现象的主要原因在于 $T$ 的不同取值对细化模块修复过程的影响。




(a) MAT单独修复与Diff-2sIR修复性能比较
(a) Comparison of MAT alone and Diff-2sIR for restoration performance




(b) DM单独修复与Diff-2sIR修复性能比较
(b) Comparison of DM alone and Diff-2sIR for restoration performanc

图 3　指导模块单独修复和使用 Diff-2sIR 效果对比

Fig. 3　Comparison of the results of single module restoration and Diff-2sIR

当 $T < 300$ 时,由于细化模块与扩散指导模块有着一致的解空间,先验结果对采样过程影响较大。此时,指导模块的修复质量较低,语义错误较多,导致细化模块即便引入随机性纠正错误,其结果仍难以完全摆脱初步修复的限制。而 MAT 的修复语义错误较少,为细化模块提供了更为准确的指导,因此表现优于扩散指导模块。

当 $T > 300$ 时,MAT 的先验分布在细化模块的采样过程中出现偏移,由于细化模块性能的局限性,这种偏移导致结果质量下降。而扩散指导模块通过更多的采样步数,能够缓解部分语义错误,逐步优化修复效果,从而表现出更大的提升空间。

从结果来看,无论选择扩散模型还是 MAT 作为指导模块,Diff-2sIR 均能进一步提升修复质量,尤其是在较高质量初步修复结果的基础上仍能取得显著改善。这表明 Diff-2sIR 具有良好的适应性,可适配不同质量的初步修复输入。通过上述分析,验证了指导模块对修复效果的影响,进一步证明了 Diff-2sIR 在不同指导模块输入下的优越性能和灵活性。

### 2.2　中间步采样与起始位置采样的影响

本实验中主要探讨了扩散模型两种采样策略对修复效果的影响:从噪声表起始位置采样和从中间步采样。两者均使用从 0.000 001～0.01 的噪声表,采样步数设置为 $T = 100$～500,间隔为 50 步。实验结果如图 4 所示。

从图 4 可以观察到,中间步采样在各个采样步数下的

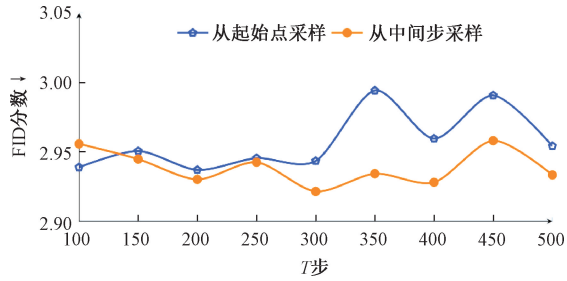



图 4　起始位置采样与中间步采样比较($T$=1 500)

Fig. 4　Comparison of sampling from the start and from the middle($T$=1 500)

FID 分数普遍低于从起始位置采样。这表明,相比于将初步修复图像再次加噪到接近完全破坏的状态,通过减少对初步修复图像的加噪可以保留更多的图像细节,从而显著提升修复效果。具体而言,中间采样策略由于直接从较低噪声的起点开始,保留了较多的先验信息,模型基于这些信息进行修复,避免了从高噪甚至纯噪的状态下重建图像(如图 5 所示,图中 FS 与 FM 分别代表从起始位置开始采样以及从中间步开始采样)。

此外,中间采样在不同采样步数下的性能更加稳定,这一稳定性进一步表明减少过度加噪对修复任务的重要性。

相比之下,从起始位置采样由于遵循完整噪声表进行加噪,先验图像被破坏得较为严重,残留信息较少。模型需

要花费更多的采样步数重建图像细节，这不仅对模型的能力提出了更高的要求，也使得修复结果对于采样步数的依赖性增强，导致修复效果波动较大(如图 4 后半段所示)。

综上，中间采样策略在图像修复任务中能够更加高效地利用已有的图像信息，从而显著提升修复质量和稳定性。进一步地，为了探讨噪声表细粒度对修复性能的影响，在实验中设置了不同细粒度的噪声表($T$ = 2 000，1 500，1 000，500)，实验结果如图 6 所示。

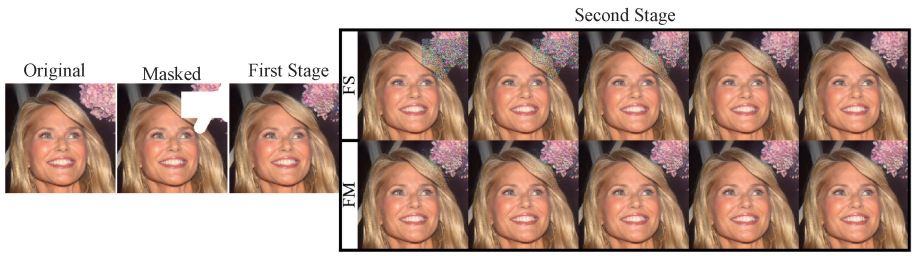



图 5　不同噪声表策略采样对比

Fig. 5　Comparison of sampling with different noise schedule strategies

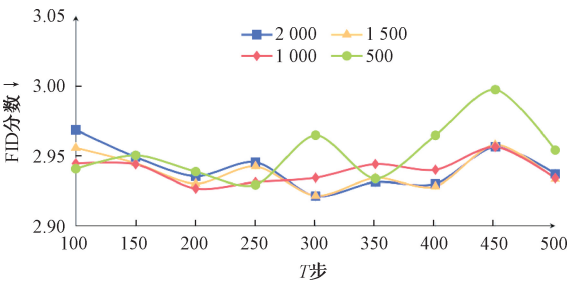



图 6　不同噪声表设置不同采样步对比

Fig. 6　Comparison of different noise table settings with varying intermediate steps

从图 6 中可以看出，随着细粒度的提高(即 $T$ 增大)，修复效果整体呈现出更加平稳的趋势。这种现象的主要原因在于：细粒度更高的噪声表在相同采样步数下对图像的破坏相对更小，能够保留更多的图像信息。

例如 $T$ = 2 000 和 $T$ = 1 000 的噪声表，后者第 250 步 $\beta$ 的取值与前者第 500 步非常接近，在采样步数相同的情况下，更高细粒度的噪声表在每一步的加噪强度更小(如图 7 所示)。


图 7　不同细粒度噪声表( $T_k$ = 200)对比

Fig. 7　Comparison of different fine-grained noise schedules( $T_k$ = 200)

因而在采样过程中，相比于低细粒度噪声表来说，细化模块拥有较多的先验信息。最后，结合由图 4 分析得出的结论，这进一步表明减少加噪强度和保留更多细节信息对于提升模型性能的重要性。

此外，从图 6 的细节可以观察到，在较低采样步数(例如 $T_k$ = 100，150)的情况下，细粒度更高的噪声表对修复效果的提升比细粒度更低的噪声表平均较差。这说明，对于高效修复任务，采用细粒度更低的噪声表能够显著提升模型的性能。另一方面，当采样步数较多时(例如 $T_k$ = 500)，不同细粒度噪声表的性能差距逐渐缩小。这主要因为在较多的采样步数下，模型有足够的时间从噪声中逐步恢复图像细节，噪声表的细粒度差异对最终修复效果的影响被进一步弱化，而 $T_k$ = 500 的噪声表，随着 $T_k$ 的增大，对于先验图像的破坏更大，在 $T_k$ = 500 时，先验图像几乎被完全破坏，模型能够接收的有效参考信息最少，因此总体效果相比于其他噪声表表现的更差。综上所述，本实验表明：

首先，中间采样策略优势显著。通过减少过度加噪，中间采样更有效保留初步修复细节，提高修复质量并增强稳定性，显著减少采样步数对性能的依赖。

其次，噪声表细粒度影响修复效果。低细粒度噪声表在低采样步数下表现更优，适用于时间受限场景；高细粒度噪声表在高采样步数下效果更平稳，适合追求精度的任务。

最后，噪声表策略与细粒度需根据任务需求选择。需要一定修复效率的场景优先使用中间采样和低细粒度噪声表；任务需要较高精度时，应结合高采样步数与高细粒度噪声表以发挥模型性能。遗憾的是。尽管本文提出了中间步采样策略，一定程度上提高了修复效率并保证了修复质量，但是由于模型参数量以及扩散模型本身采样速度的限制，Diff-2sIR$^+$ 依然不能做到实时修复。

本文将 Diff-2sIR$^+$ 与 CelebA-HQ 数据集上的先进模型进行了对比，如表 1 所示，最优结果由下划线标注，使用 Diff-2sIR$^+$ 前后分数较高者被加粗标注。表中 Diff-2sIR$^+$ 使用了 $T$ = 1 000，$T_k$ = 200 的配置。

表 1　与其他先进修复模型的对比

Table 1　Comparison with other advanced inpainting models

| 模型 | Para. $\times 10^6$ | CelebA-HQ (256 × 256) | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Small Mask | | | | Large Mask | | | |
| | | FID ↓ | IS ↑ | P-IDS/% ↑ | U-IDS/% ↑ | FID ↓ | IS ↑ | P-IDS/% ↑ | U-IDS/% ↑ |
| MAT | 62 | 3.00 | 3.70 | **20.25** | **31.32** | 5.20 | 3.52 | **11.96** | **22.84** |
| +Diff-2sIR$^+$ | 62 | **2.92** | **3.71** | 17.91 | 30.20 | **5.15** | **3.55** | 8.45 | 21.58 |
| LaMa | 27 | 3.96 | 3.55 | 10.12 | 22.39 | 8.50 | 3.37 | 2.10 | 9.20 |
| +Diff-2sIR$^+$ | 62 | **3.17** | **3.59** | **14.27** | **26.78** | **6.24** | **3.46** | **4.81** | **15.39** |
| DDNM | 113 | 3.60 | 3.59 | 10.89 | 23.91 | 7.67 | 3.45 | 3.01 | 10.57 |
| +Diff-2sIR$^+$ | 62 | **3.07** | **3.63** | **16.20** | **29.69** | **5.84** | **3.49** | **7.22** | **17.89** |
| DiffIR$_{S2}$ | 26 | 3.74 | 3.61 | 9.42 | 23.24 | 8.06 | 3.39 | 2.61 | 9.61 |
| +Diff-2sIR$^+$ | 62 | **3.02** | **3.64** | **14.53** | **27.13** | **5.86** | **3.50** | **5.21** | **16.37** |
| EdgeConnect | 22 | 5.13 | 3.45 | 6.05 | 15.85 | 12.10 | 3.21 | 0.94 | 2.92 |
| +Diff-2sIR$^+$ | 62 | **3.87** | **3.59** | **11.66** | **24.17** | **8.33** | **3.42** | **3.44** | **11.01** |
| M2S | 68 | 3.79 | 3.66 | 12.50 | 24.64 | 7.25 | 3.50 | 4.04 | 13.03 |
| +Diff-2sIR$^+$ | 62 | **3.21** | 3.65 | **16.40** | **29.62** | **5.77** | **3.57** | **8.05** | **19.55** |
| WavePaint | 13 | 4.14 | 3.53 | 7.35 | 16.56 | 8.89 | 3.20 | 1.34 | 3.91 |
| +Diff-2sIR$^+$ | 62 | **3.17** | **3.59** | **14.27** | **26.78** | **6.42** | **3.43** | **4.94** | **14.62** |

## 3　结　　论

针对图像修复中的语义不一致问题,本文利用 Unet 本身以及扩散模型采样过程所具有的一些特性,提出了一种新颖的解决方案:Diff-2sIR,Diff-2sIR 在 CelebA-HQ 数据集上得到了最佳的结果(FID=2.92),并且还可作为一个再修复插件来使用,事实上,本文提出的这个二阶段修复框架仍然有许多可以扩展的地方,比如集成先进的采样算法,探索更强性能的细化模型的影响等,最终通过广泛的实验及定性比较证明了 Diff-2sIR 相对于基线模型的优越性。

## 参考文献


[1] BERTALMIO M, SAPIRO G, CASELLES V, et al. Image inpainting[C]. 27th Annual Conference on Computer Graphics and Interactive Techniques, 2000.

[2] EFROS A A, LEUNG T K. Texture synthesis by non-parametric sampling[C]. 7th IEEE International Conference on Computer Vision, 1999.

[3] YU F, KOLTUN V. Multi-scale context aggregation by dilated convolutions[J]. ArXiv preprint arXiv: 1511.07122, 2015.

[4] LIU G, REDA F A, SHIH K J, et al. Image inpainting for irregular holes using partial convolutions[C]. European Conference on Computer Vision, 2018.

[5] 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.

[6] JEEVAN P, KUMAR D S, SETHI A. WavePaint: Resource-efficient token-mixer for self-supervised inpainting[J]. ArXiv preprint arXiv: 2307.00407, 2023.

[7] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[J]. Advances in Neural Information Processing Systems, 2014, 27: 2672-2680.

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

[9] NAZERI K, NG E, JOSEPH T, et al. EdgeConnect: Structure guided image inpainting using edge prediction[C]. IEEE/CVF International Conference on Computer Vision Workshop, 2019:3265-3274.

[10] XIE CH H, LIU SH H, LI CH, et al. Image inpainting with learnable bidirectional attention maps[C]. IEEE/CVF International Conference on Computer Vision, 2019: 8857-8866.

[11] 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.

[12] VASWANI A. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017, 30: 5998-6008.

[13] WAN Z Y, ZHANG J B, CHEN D D, et al. High-fidelity pluralistic image completion with trans formers[C]. IEEE/CVF International Conference on Computer Vision, 2021.

[14] YU Y CH, ZHAN F N, WU R L, et al. Diverse image inpainting with bidirectional and autoregressive transformers[C]. 29th ACM International Conference on Multimedia, 2021.

[15] LI W B, LIN ZH, ZHOU K, et al. Mat: Mask-aware transformer for large hole image inpainting [C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022.

[16] HO J, JAIN A, ABBEEL P. Denoising diffusion probabilistic models [J]. Advances in Neural Information Processing Systems, 2020, 33: 6840-6851.

[17] SONG Y, SOHL-DICKSTEIN J, KINGMA D P, et al. Score-based generative modeling through stochastic differential equations [C]. International Conference on Learning Representations, 2020.

[18] SONG J M, MENG CH L, ERMON S. Denoising diffusion implicit models[C]. International Conference on Learning Representations, 2020.

[19] DHARIWAL P, NICHOL A. Diffusion models beat gans on image synthesis [J]. Advances in Neural information Processing Systems, 2021, 34: 8780-8794.

[20] ROMBACH R, BLATTMANN A, LORENZ D, et al. High-resolution image synthesis with latent diffusion models [C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022.

[21] MENG CH L, HE Y T, SONG Y, et al. Sdedit: Guided image synthesis and editing with stochastic differential equations[C]. International Conference on Learning Representations, 2022.

[22] WANG Y H, YU J W, ZHANG J. Zero-shot image restoration using denoising diffusion null-space model[C]. International Conference on Learning Representations, 2022.

[23] XIA B, ZHANG Y L, WANG SH Y, et al. DiffIR: Efficient diffusion model for image restoration[C]. IEEE/CVF International Conference on Computer Vision, 2023.

[24] ZHANG L T, DU X CH, TOMYENRIQUE L, et al. Minutes to seconds: Speeded-up ddpm-based image inpainting with coarse-to-fine sampling[C]. 2024 IEEE International Conference on Multimedia and Expo, 2024.

[25] SAHARIA C, CHAN W, CHANG H, et al. Palette: Image-to-image diffusion models [C]. ACM SIGGRAPH 2022 Conference Proceedings, 2022.

[26] SALIMANS T, HO J. Progressive distillation for fast sampling of diffusion models [C]. International Conference on Learning Representations, 2022.

[27] HERTZ A, MOKADY R, TENENBAUM J, et al. Prompt-to-prompt image editing with cross attention control[J]. ArXiv preprint arXiv:2208.01626, 2022.

[28] ZHANG L M, RAO A Y, AGRAWALA M. Adding conditional control to text-to-image diffusion models[C]. IEEE/CVF International Conference on Computer Vision, 2023.

[29] KWON M, JEONG J, UH Y. Diffusion models already have a semantic latent space[C]. International Conference on Learning Representations, 2023.

[30] KARRAS T. Progressive growing of GANs for improved quality, stability, and variation [C]. International Conference on Learning Representations, 2018.

[31] HEUSEL M, RAMSAUER H, UNTERTHINER T, et al. Gans trained by a two time-scale update rule converge to a local nash equilibrium[J]. Advances in Neural Information Processing Systems, 2017, 30: 6626-6637.

[32] ZHAO SH Y, CUI J, SHENG Y L, et al. Large scale image completion via co-modulated generative adversarial networks[C]. International Conference on Learning Representations, 2021.

[33] SALIMANS T, GOODFELLOW I, ZAREMBA W, et al. Improved techniques for training gans [J]. Advances in Neural Information Processing Systems, 2016, 29: 2234-2242.

## 作者简介

**张绪**，硕士研究生，主要研究方向为计算机视觉与图像生成与处理。

E-mail：daxu@nefu.edu.cn

**胡峻峰**（通信作者），博士，副教授，硕士生导师，主要研究方向为机器视觉、图形处理、模式识别与智能控制。

E-mail：nefuhjf@126.com