RePaint: Inpainting using Denoising Diffusion Probabilistic Models

Andreas Lugmayr Martin Danelljan Andres Romero Fisher Yu Radu Timofte Luc Van Gool

Computer Vision Lab

ETH Zürich, Switzerland

Figure 1. **We use Denoising Diffusion Probabilistic Models (DDPM) for inpainting.** The process is conditioned on the masked input (*left*). It starts from a random Gaussian noise sample that is iteratively denoised until it produces a high-quality output. Since this process is stochastic, we can sample multiple diverse outputs. The DDPM prior forces a harmonized image, is able to reproduce texture from other regions, and inpaint semantically meaningful content.

Abstract

Free-form inpainting is the task of adding new content to an image in the regions specified by an arbitrary binary mask. Most existing approaches train for a certain distribution of masks, which limits their generalization capabilities to unseen mask types. Furthermore, training with pixel-wise and perceptual losses often leads to simple textural extensions towards the missing areas instead of semantically meaningful generation. In this work, we propose RePaint: A Denoising Diffusion Probabilistic Model (DDPM) based inpainting approach that is applicable to even extreme masks. We employ a pretrained unconditional *DDPM as the generative prior. To condition the generation process, we only alter the reverse diffusion iterations by sampling the unmasked regions using the given image information. Since this technique does not modify or condition the original DDPM network itself, the model produces high-quality and diverse output images for any inpainting form. We validate our method for both faces and general-purpose image inpainting using standard and extreme masks. RePaint outperforms state-of-the-art Autoregressive, and GAN approaches for at least five out of six mask distributions. Github Repository: git.io/RePaint*

1. Introduction

Image Inpainting, also known as Image Completion, aims at filling missing regions within an image. Such inpainted regions need to harmonize with the rest of the im-

RePaint：使用去噪扩散概率模型进行修复

安德烈亚斯·卢格迈尔 马丁·丹内尔扬 安德烈斯·罗梅罗 费舍尔·余 拉杜·蒂莫夫特 卢克·范·古尔

计算机视觉实验室

瑞士苏黎世联邦理工学院

图1. 我们使用去噪扩散概率模型（DDPM）进行图像修复。该过程以掩码输入（{v*}）为条件，从随机高斯噪声样本开始，经过迭代去噪直至生成高质量输出。由于该过程具有随机性，我们可以采样得到多种不同的输出结果。DDPM先验能够强制生成协调的图像，可复现其他区域的纹理，并修复具有语义意义的内容。

摘要

Free-form inpainting is the task of adding new content to an image in the regions specified by an arbitrary binary mask. Most existing approaches train for a certain distribution of masks, which limits their generalization capabilities to unseen mask types. Furthermore, training with pixel-wise and perceptual losses often leads to simple textural extensions towards the missing areas instead of semantically meaningful generation. In this work, we propose RePaint: A Denoising Diffusion Probabilistic Model (DDPM) based inpainting approach that is applicable to even extreme masks. We employ a pretrained 无条件 *DDPM as the generative prior. To condition the generation process, we only alter the reverse diffusion iterations by sampling the unmasked regions using the given image information. Since this technique does not modify or condition the original DDPM network itself, the model produces high-quality and diverse output images for any inpainting form. We validate our method for both faces and general-purpose image inpainting using standard and extreme masks. RePaint outperforms state-of-the-art Autoregressive, and GAN approaches for at least five out of six mask distributions. Github Repository: git.io/RePaint*

1. 引言

图像修复，亦称图像补全，旨在填充图像中的缺失区域。此类修复区域需与图像的其余部分保持协调。

age and be semantically reasonable. Inpainting approaches thus require strong generative capabilities. To this end, current State-of-the-Art approaches [20, 40, 48, 51] rely on GANs [8] or Autoregressive Modeling [33, 42, 49]. Moreover, inpainting methods need to handle various forms of masks such as thin or thick brushes, squares, or even extreme masks where the vast majority of the image is missing. This is highly challenging since existing approaches train with a certain mask distribution, which can lead to poor generalization to novel mask types. In this work, we investigate an alternative generative approach for inpainting, aiming to design an approach that requires no mask-specific training.

Denoising Diffusion Probabilistic Models (DDPM) is an emerging alternative paradigm for generative modelling [12, 38]. Recently, Dhariwal and Nichol [7] demonstrated that DDPM can even outperform the state-of-the-art GAN-based method [4] for image synthesis. In essence, the DDPM is trained to iteratively denoise the image by reversing a diffusion process. Starting from randomly sampled noise, the DDPM is then iteratively applied for a certain number of steps, which yields the final image sample. While founded in principled probabilistic modeling, DDPMs have been shown to generate diverse and high-quality images [7, 12, 28].

We propose RePaint: an inpainting method that solely leverages an off-the-shelf unconditionally trained DDPM. Specifically, instead of learning a mask-conditional generative model, we condition the generation process by sampling from the given pixels during the reverse diffusion iterations. Remarkably, our model is therefore not trained for the inpainting task itself. This has two important advantages. First, it allows our network to generalize to any mask during inference. Second, it enables our network to learn more semantic generation capabilities since it has a powerful DDPM image synthesis prior (Figure 1).

Although the standard DDPM sampling strategy produces matching textures, the inpainting is often semantically incorrect. Therefore, we introduce an improved denoising strategy that *resamples* (RePaint) iterations to better condition the image. Notably, instead of slowing down the diffusion process [7], our approach goes forward and backward in diffusion time, producing remarkable semantically meaningful images. Our approach allows the network to effectively harmonize the generated image information during the entire inference process, leading to a more effective conditioning on the given image information.

We perform experiments on CelebA-HQ [21] and ImageNet [36], and compare with other State-of-the-Art inpainting approaches. Our approach generalizes better and has overall more semantically meaningful inpainted regions.

2. Related Work

Early attempts on Image Inpainting or Image Completion exploited low-level cues within the input image [1–3], or within the neighbor of a large image dataset [10] to fill the missing region.

Deterministic Image Inpainting: Since the introduction of GANs [8], most of the existing methods follow a standard configuration, first proposed by Pathak *et al*. [32], that is, using an encoder-decoder architecture as the main inpainting generator, adversarial training, and tailored losses that aim at photo-realism. Follow-up works have produced impressive results in recent years [15, 20, 30, 34, 50].

As image inpainting requires a high-level semantic context, and to explicitly include it in the generation pipeline, there exist hand-crafted architectural designs such as Dilated Convolutions [16, 45] to increase the receptive field, Partial Convolutions [19] and Gated Convolutions [48] to guide the convolution kernel according to the inpainted mask, Contextual Attention [46] to leverage on global information, Edges maps [9, 27, 43, 44] or Semantic Segmentation maps [14, 31] to further guide the generation, and Fourier Convolutions [40] to include both global and local information efficiently. Although recent works produce photo-realistic results, GANs are well known for textural synthesis, so these methods shine on background completion or removing objects, which require repetitive structural synthesis, and struggle with semantic synthesis (Figure 5).

Diverse Image Inpainting: Most GAN-based Image Inpainting methods are prone to deterministic transformations due to the lack of control during the image synthesis. To overcome this issue, Zheng *et al*. [55] and Zhao *et al*. [53] propose a VAE-based network that trade-offs between diversity and reconstruction. Zhao *et al*. [54], inspired by the StyleGAN2 [18] modulated convolutions, introduces a co-modulation layer for the inpainting task in order to improve both diversity and reconstruction. A new family of autoregressive methods [33, 42, 49], which can handle irregular masks, has recently emerged as a powerful alternative for free-form image inpainting.

Usage of Image Prior: In a different direction closer to ours Richardson *et al*. [35] exploits the StyleGAN [17] prior to successfully inpaint missing regions. However, similar to super-resolution methods [5, 26] that leverage the StyleGAN latent space, it is to limited specific scenarios like faces. Noteworthy, a Ulyanov *et al*. [41] showed that the structure of a non-trained generator network contains an inherent prior that can be used for inpaining and other applications. In contrast to these methods, we are leveraging on the high expressiveness of a pretrained Denoising Diffusion Probabilistic Model [12] (DDPM) and therefore use it as a prior for generic image inpainting. Our method generates very detailed, high-quality images for both seman-

年龄且在语义上合理。因此，修复方法需要强大的生成能力。为此，当前最先进的方法[20, 40, 48, 51]依赖于生成对抗网络[8]或自回归建模[33, 42, 49]。此外，修复方法需要处理各种形式的掩码，例如细或粗的笔刷、方块，甚至是图像大部分缺失的极端掩码。这极具挑战性，因为现有方法使用特定的掩码分布进行训练，这可能导致对新掩码类型的泛化能力较差。在这项工作中，我们研究了一种用于修复的替代生成方法，旨在设计一种无需针对特定掩码进行训练的方法。

去噪扩散概率模型（DDPM）是生成建模领域一种新兴的替代范式[12, 38]。最近，Dhariwal和Nichol[7]证明，在图像合成任务中，DDPM甚至能超越最先进的基于GAN的方法[4]。本质上，DDPM通过逆转扩散过程来训练模型迭代地对图像进行去噪。从随机采样的噪声开始，DDPM被迭代应用一定步数，从而生成最终的图像样本。尽管基于概率建模原理，DDPM已被证明能够生成多样且高质量的图像[7, 12, 28]。

我们提出了RePaint：一种仅利用现成的无条件训练DDPM进行修复的方法。具体而言，我们并非学习一个掩码条件生成模型，而是通过在反向扩散迭代过程中从给定像素采样来条件化生成过程。值得注意的是，因此我们的模型并未针对修复任务本身进行训练。这带来了两个重要优势：首先，它使我们的网络能够在推理阶段泛化到任意掩码；其次，由于它具备强大的DDPM图像合成先验（图1），这使得网络能够学习更具语义的生成能力。

尽管标准DDPM采样策略能生成匹配的纹理，但修复结果常存在语义错误。因此，我们引入一种改进的去噪策略，通过*resamples* (RePaint)迭代来更好地对图像进行条件约束。值得注意的是，我们的方法并非如[7]所述减缓扩散过程，而是在扩散时间轴上双向推进，从而生成语义信息显著的图像。该方法使网络能在整个推理过程中有效协调生成图像信息，从而更高效地对给定图像信息进行条件化处理。

我们在CelebA-HQ [21]和ImageNet [36]上进行了实验，并与其他最先进的图像修复方法进行了比较。我们的方法泛化能力更强，修复区域在整体上具有更丰富的语义信息。

2. 相关工作

早期关于图像修复或图像补全的尝试利用了输入图像[1–3]或大型图像数据集邻近区域[10]中的低级线索来填充缺失区域。

确定性图像修复：自从GANs[8]被提出以来，大多数现有方法遵循由Pathak *et al*率先提出的标准配置[32]，即采用编码器-解码器架构作为主要修复生成器，结合对抗训练及旨在实现照片级真实感的定制损失函数。近年来，后续研究取得了令人瞩目的成果[15, 20, 30, 34, 50]。

由于图像修复需要高层次的语义上下文，并需将其明确纳入生成流程中，现有一些手工设计的架构方案，例如通过扩张卷积[16, 45]来增大感受野，采用部分卷积[19]和门控卷积[48]根据修复掩码引导卷积核，利用上下文注意力[46]获取全局信息，借助边缘图[9, 27, 43, 44]或语义分割图[14, 31]进一步引导生成，以及使用傅里叶卷积[40]高效融合全局与局部信息。尽管近期研究能生成逼真的结果，但GAN以纹理合成见长，因此这些方法在背景补全或物体移除这类需要重复结构合成的任务上表现出色，而在语义合成方面仍存在困难（图5）。

多样图像修复：大多数基于GAN的图像修复方法由于在图像合成过程中缺乏控制，容易产生确定性变换。为克服这一问题，Zheng *et al* [55] 与 Zhao *et al* [53] 提出了一种基于VAE的网络，在多样性与重建质量之间进行权衡。Zhao *et al* [54] 受StyleGAN2 [18] 调制卷积的启发，为修复任务引入协同调制层，以同时提升多样性与重建效果。近期出现的新型自回归方法族 [33, 42, 49] 能够处理不规则掩码，已成为自由形式图像修复的强大替代方案。

图像先验的使用：在另一个更接近我们Richardson *et al* 的方向上，[35]利用StyleGAN [17]先验成功修复了缺失区域。然而，与利用StyleGAN潜在空间的超分辨率方法[5, 26]类似，它局限于如人脸等特定场景。值得注意的是，Ulyanov *et al* [41]表明，未经训练的生成器网络结构包含一种可用于修复和其他应用的内在先验。与这些方法不同，我们利用预训练去噪扩散概率模型[12]（DDPM）的高表达能力，因此将其作为通用图像修复的先验。我们的方法能够为语义...

tically meaningful generation and texture synthesis. Moreover, our method is not trained for the image inpainting task, and instead, we take full advantage of the prior DDPM, so each image is optimized independently.

Image Conditional Diffusion Models: The work by Sohl-Dickstein *et al*. [38] applied early diffusion models to inpainting. More recently, Song *et al*. [39] develop a score-based formulation using stochastic differential equations for unconditional image generation, with an additional application to inpainting. However, both these works only show qualitative results, and do not compare with other inpainting approaches. In contrast, we aim to advance the state-of-the-art in image inpainting, and provide comprehensive comparisons with the top competing methods in literature.

A different line of research is guided image synthesis with DDPM-based approaches [6, 25]. In the case of ILVR [6], a trained diffusion model is guided using the low-frequency information from a conditional image. However, this conditioning strategy cannot be adopted for inpainting, since both high and low-frequency information is absent in the masked-out regions. Another approach for image-conditional synthesis is developed by [25]. Guided generation is performed by initializing the reverse diffusion process from the guiding image at some intermediate diffusion time. An iterative strategy, repeating the reverse process several times, is further adopted to improve harmonization. Since a guiding image is required to start the reverse process at an intermediate time step, this approach is not applicable to inpainting, where new image content needs to be generated solely conditioned on the non-masked pixels. Furthermore, the resampling strategy proposed in this work differs from the concurrent [25]. We proceed through the full reverse diffusion process, starting at the end time, at each step jumping back and forth a fixed number of time steps to progressively improve generation quality.

While we propose a method that conditions an unconditionally trained model, the concurrent work [29] is based on classifier-free guidance [13] for training an image-conditional diffusion model. Another direction for image manipulation is image-to-image translation using diffusion models as explored in the concurrent work [37]. It trains an image-conditional DDPM, and shows an application to inpainting. Unlike both these concurrent works, we leverage an unconditional DDPM and only condition through the reverse diffusion process itself. It allows our approach to effortlessly generalize to any mask shape for free-form inpainting. Moreover, we propose a sampling schedule for the reverse process, which greatly improves image quality.

3. Preliminaries: Denoising Diffusion Probabilistic Models

In this paper, we use diffusion models [38] as a generative method. As other generative models, the DDPM learns a distribution of images given a training set. The inference process works by sampling a random noise vector x_T and gradually denoising it until it reaches a high-quality output image x_0. During training, DDPM methods define a diffusion process that transforms an image x_0 to white Gaussian noise $x_T \sim \mathcal{N}(0,1)$ in T time steps. Each step in the forward direction is given by,

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

The sample x_t is obtained by adding *i.i.d.* Gaussian noise with variance β_t at timestep t and scaling the previous sample x_{t-1} with $\sqrt{1-\beta_t}$ according to a variance schedule.

The DDPM is trained to reverse the process in (1). The reverse process is modeled by a neural network that predicts the parameters $\mu_\theta(x_t, t)$ and $\Sigma_\theta(x_t, t)$ of a Gaussian distribution,

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \tag{2}$$

The learning objective for the model (2) is derived by considering the variational lower bound,

$$\begin{aligned} \mathbb{E}\left[-\log p_\theta(\mathbf{x}_0)\right] &\leq \mathbb{E}_q\left[-\log \frac{p_\theta(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}\right] \\ &= \mathbb{E}_q\left[-\log p(\mathbf{x}_T) - \sum_{t\geq 1} \log \frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q(\mathbf{x}_t|\mathbf{x}_{t-1})}\right] = L \end{aligned} \tag{3}$$

As extended by Ho *et al*. [12], this loss can be further decomposed as,

$$\begin{aligned} &\mathbb{E}_q\Bigg[\underbrace{D_{\mathrm{KL}}(q(\mathbf{x}_T|\mathbf{x}_0) \parallel p(\mathbf{x}_T))}_{L_T} \\ &+ \sum_{t>1} \underbrace{D_{\mathrm{KL}}(q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) \parallel p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t))}_{L_{t-1}} \underbrace{-\log p_\theta(\mathbf{x}_0|\mathbf{x}_1)}_{L_0}\Bigg] \end{aligned} \tag{4}$$

Importantly the term L_{t-1} trains the network (2) to perform one reverse diffusion step. Furthermore, it allows for a closed from expression of the objective since $q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)$ is also Gaussian [12].

As reported by Ho *et al*. [12], the best way to parametrize the model is to predict the cumulative noise ϵ_0 that is added to the current intermediate image x_t. Thus, we obtain the following parametrization of the predicted mean $\mu_\theta(x_t, t)$,

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t)\right) \tag{5}$$

具有语义意义的生成和纹理合成。此外，我们的方法并非针对图像修复任务进行训练，而是充分利用先验DDPM，因此每张图像都是独立优化的。

图像条件扩散模型：Sohl-Dickstein *et al* [38] 的工作将早期扩散模型应用于图像修复。最近，Song *et al* [39] 基于随机微分方程提出了一种用于无条件图像生成的分数形式化方法，并额外应用于图像修复。然而，这两项研究仅展示了定性结果，并未与其他图像修复方法进行比较。相比之下，我们的目标是推动图像修复领域的前沿发展，并提供与文献中顶尖竞争方法的全面比较。

另一条研究路线是基于DDPM的方法进行引导图像合成[6, 25]。在ILVR[6]中，使用条件图像的低频信息来引导训练好的扩散模型。然而，这种条件策略无法应用于修复任务，因为被遮蔽区域同时缺失高频和低频信息。[25]提出了另一种图像条件合成方法，通过在某个中间扩散时间步从引导图像初始化反向扩散过程来执行引导生成。为进一步提升协调性，该方法还采用了重复多次反向过程的迭代策略。由于需要在中间时间步使用引导图像启动反向过程，该方法不适用于修复任务——修复任务需要仅基于非遮蔽像素生成全新图像内容。此外，本文提出的重采样策略与同期研究[25]不同：我们执行完整的反向扩散过程（从终止时间开始），通过在每一步固定时间步数内往返跳跃来逐步提升生成质量。

我们提出了一种对无条件训练模型进行条件化的方法，而同期研究[29]则基于分类器无关引导[13]来训练图像条件扩散模型。图像处理的另一个方向是使用扩散模型进行图像到图像的转换，如同期研究[37]所探索的那样。该研究训练了一个图像条件DDPM，并展示了其在修复中的应用。与这两项同期研究不同，我们利用无条件DDPM，仅通过反向扩散过程本身进行条件化。这使得我们的方法能够轻松泛化到任意掩码形状，实现自由形式的修复。此外，我们为反向过程提出了一种采样调度方案，显著提升了图像质量。

3. 预备知识：去噪扩散概率模型

在本文中，我们使用扩散模型[38]作为一种生成方法。与其他生成模型类似，DDPM通过学习训练集中的图像分布进行训练。推理过程通过采样一个随机噪声向量x_T并逐步去噪，直至得到高质量的输出图像x_0。在训练过程中，DDPM方法定义了一个扩散过程，将图像x_0在T个时间步内转化为高斯白噪声$x_T \sim \mathcal{N}(0, 1)$。前向过程的每一步由以下公式给出：

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

样本 x_t 是通过在时间步 t 添加方差为 β_t 的高斯噪声，并根据方差表按 $\sqrt{1-\beta_t}$ 缩放前一个样本 x_{t-1} 得到的。

DDPM被训练来逆转(1)中的过程。反向过程通过一个神经网络建模，该网络预测高斯分布的参数$\mu_\theta(x_t, t)$和$\Sigma_\theta(x_t, t)$。

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \tag{2}$$

模型（2）的学习目标是通过考虑变分下界推导得出的，

$$\begin{aligned}\mathbb{E}\left[-\log p_\theta(\mathbf{x}_0)\right] &\leq \mathbb{E}_q\left[-\log \frac{p_\theta(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}\right] \\ &= \mathbb{E}_q\left[-\log p(\mathbf{x}_T) - \sum_{t\geq 1}\log\frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q(\mathbf{x}_t|\mathbf{x}_{t-1})}\right] = L\end{aligned} \tag{3}$$

如Ho *et al* [12]所扩展，该损失可进一步分解为：

$$\begin{aligned}&\mathbb{E}_q\Bigg[\underbrace{D_{\mathrm{KL}}(q(\mathbf{x}_T|\mathbf{x}_0)\parallel p(\mathbf{x}_T))}_{L_T} \\ &+\sum_{t>1}\underbrace{D_{\mathrm{KL}}(q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)\parallel p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t))}_{L_{t-1}}\underbrace{-\log p_\theta(\mathbf{x}_0|\mathbf{x}_1)}_{L_0}\Bigg]\end{aligned} \tag{4}$$

重要的是，术语L_{t-1}训练网络(2)执行一个反向扩散步骤。此外，由于$q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)$也是高斯分布[12]，这使得目标的闭式表达成为可能。

如Ho *et al*等人[12]所述，对模型进行参数化的最佳方法是预测累积噪声ϵ_0，该噪声被添加到当前中间图像x_t中。因此，我们得到预测均值$\mu_\theta(x_t, t)$的以下参数化形式：

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t)\right) \tag{5}$$

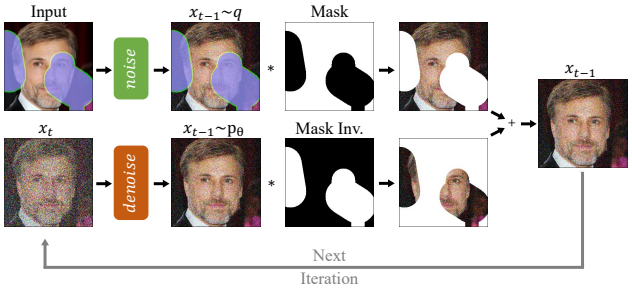


Figure 2. **Overview of our approach.** RePaint modifies the standard denoising process in order to condition on the given image content. In each step, we sample the known region (*top*) from the input and the inpainted part from the DDPM output (*bottom*).

From L_{t-1} in (4), the following simplified training objective is derived by Ho *et al*. [12],

$$L_{\text{simple}} = E_{t,x_0,\epsilon}\left[||\epsilon - \epsilon_\theta(x_t, t)||^2\right] \tag{6}$$

As introduced by Nichol and Dhariwal [28], learning the variance $\Sigma_\theta(x_t, t)$ in (2) of the reverse process helps to reduce the number of sampling steps by an order of magnitude. They, therefore, add the variational lower bound loss. Specifically, we base our training and inference approach on the recent work [7], which further reduced the inference time by factor four.

To train the DDPM, we need a sample x_t and corresponding noise that is used to transform x_0 to x_t. By using the independence property of the noise added at each step (1), we can calculate the total noise variance as $\bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$. We can thus rewrite (1), as a single step,

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\mathbf{I}) \tag{7}$$

It allows us to efficiently sample pairs of training data to train a reverse transition step.

4. Method

In this section, we first present our approach for conditioning the reverse diffusion process of an unconditional DDPM for image inpainting in Section 4.1. Then, we introduce an approach to improve the reverse process itself for inpainting in Section 4.2.

4.1. Conditioning on the known Region

The goal of inpainting is to predict missing pixels of an image using a mask region as a condition. In the remaining of the paper, we consider a trained unconditional denoising diffusion probabilistic model (2). We denote the ground truth image as x, the unknown pixels as $m \odot x$ and the known pixels as $(1-m) \odot x$.

Since every reverse step (2) from x_t to x_{t-1} depends solely on x_t, we can alter the known regions $(1-m) \odot x_t$ as long as we keep the correct properties of the corresponding distribution. Since the forward process is defined by a Markov Chain (1) of added Gaussian noise, we can sample the intermediate image x_t at any point in time using (7). This allows us to sample the know regions $m \odot x_t$ at any time step t. Thus, using (2) for the unknown region and (7) for the known regions, we achieve the following expression for one reverse step in our approach,

$$x_{t-1}^{\text{known}} \sim \mathcal{N}(\sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\mathbf{I}) \tag{8a}$$

$$x_{t-1}^{\text{unknown}} \sim \mathcal{N}(\mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \tag{8b}$$

$$x_{t-1} = m \odot x_{t-1}^{\text{known}} + (1-m) \odot x_{t-1}^{\text{unknown}} \tag{8c}$$

Thus, x_{t-1}^{known} is sampled using the known pixels in the given image $m \odot x_0$, while x_{t-1}^{unknown} is sampled from the model, given the previous iteration x_t. These are then combined to the new sample x_{t-1} using the mask. Our approach is illustrated in Figure 2.

4.2. Resampling

When directly applying the method described in Section 4.1, we observe that only the content type matches with the known regions. For example, in Figure 3 $n = 1$, the inpainted area is a furry texture matching the hair of the dog. Although the inpainted region matches the texture of the neighboring region, it is semantically incorrect. Therefore, the DDPM is leveraging on the context of the known region, yet it is not harmonizing it well with the rest of the image. Next, we discuss possible reasons for this behavior.

From Figure 2, we analyze how the method is conditioning the known regions. As shown in (8), the model predicts x_{t-1} using x_t, which comprises the output of the DDPM (2) and the sample from the known region. However, the sampling of the known pixels using (7) is performed without considering the generated parts of the image, which introduces disharmony. Although the model tries to harmonize the image again in every step, it can never fully converge because the same issue occurs in the next step. Moreover, in each reverse step, the maximum change to an image declines due to the variance schedule of β_t. Thus, the method cannot correct mistakes that lead to disharmonious boundaries in the subsequent steps due to restricted flexibility. As a consequence, the model needs more time to harmonize the conditional information x_{t-1}^{known} with the generated information x_{t-1}^{unknown} in one step before advancing to the next denoising step.

Since the DDPM is trained to generate an image that lies within a data distribution, it naturally aims at producing consistent structures. In our resampling approach, we use this DDPM property to harmonize the input of the model. Consequently, we diffuse the output x_{t-1} back to x_t by sampling from (1) as $x_t \sim \mathcal{N}(\sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I})$. Although this operation scales back the output and adds

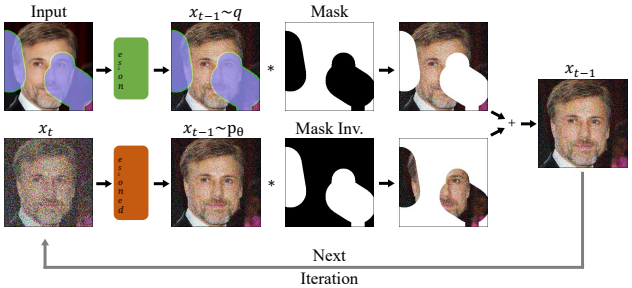


图2. 我们方法的概览。RePaint修改了标准的去噪过程，以便以给定的图像内容为条件。在每一步中，我们从输入中采样已知区域（*top*），并从DDPM输出中采样修复部分（*bottom*）。

从(4)中的L_{t-1}出发，Ho等人[12]推导出以下简化训练目标*et al*。

$$L_{\text{simple}} = E_{t,x_0,\epsilon}\left[||\epsilon - \epsilon_\theta(x_t, t)||^2\right] \tag{6}$$

如Nichol和Dhariwal [28]所介绍，学习反向过程中(2)式的方差$\Sigma_\theta(x_t, t)$有助于将采样步骤数量减少一个数量级。因此，他们引入了变分下界损失。具体而言，我们的训练和推理方法基于近期研究[7]，该研究进一步将推理时间缩短了四倍。

为了训练DDPM，我们需要一个样本x_t以及用于将x_0转换为x_t的对应噪声。利用每一步所添加噪声的独立性特性(1)，我们可以将总噪声方差计算为$\bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$。因此，我们可以将(1)式改写为单步形式，

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\mathbf{I}) \tag{7}$$

它使我们能够高效地采样训练数据对，以训练反向转换步骤。

4. 方法

在本节中，我们首先在第4.1节中介绍了如何为无条件DDPM的反向扩散过程添加条件，以实现图像修复。随后，我们在第4.2节中引入了一种改进反向过程本身以进行修复的方法。

4.1. 基于已知区域的条件化

图像修复的目标是利用掩码区域作为条件来预测图像的缺失像素。在本文的后续部分，我们考虑一个训练好的无条件去噪扩散概率模型（2）。我们将真实图像表示为x，未知像素表示为$m \odot x$，已知像素表示为$(1-m) \odot x$。

由于从x_t到x_{t-1}的每个反向步骤（2）仅依赖于x_t，我们可以改变已知区域$(1-m) \odot x_t$只要我们保持对应分布的正确属性。由于前向过程是通过添加高斯噪声的马尔可夫链（1）定义的，我们可以使用（7）在任意时间点对中间图像x_t进行采样。这使我们能够在任意时间步t对已知区域$m \odot x_t$进行采样。因此，对未知区域使用（2）而对已知区域使用（7），我们得到了方法中单次反向步骤的以下表达式：

$$x_{t-1}^{\text{known}} \sim \mathcal{N}(\sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\mathbf{I}) \tag{8a}$$

$$x_{t-1}^{\text{unknown}} \sim \mathcal{N}(\mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \tag{8b}$$

$$x_{t-1} = m \odot x_{t-1}^{\text{known}} + (1-m) \odot x_{t-1}^{\text{unknown}} \tag{8c}$$

因此，x_{t-1}^{known} 是利用给定图像 $m \odot x_0$ 中的已知像素进行采样的，而 x_{t-1}^{unknown} 则是根据先前迭代 x_t 从模型中采样得到的。随后，通过掩码将这些采样结果组合成新的样本 x_{t-1}。我们的方法如图2所示。

4.2. 重采样

直接应用第4.1节所述方法时，我们观察到仅有内容类型与已知区域相匹配。例如在图3{v*}1中，修复区域呈现的毛绒纹理虽与狗的毛发匹配，且修复区域的纹理与相邻区域一致，但语义层面并不正确。这表明DDPM虽能利用已知区域的上下文信息，却未能使其与图像其余部分良好融合。接下来我们将探讨导致这种现象的可能原因。

从图2中，我们分析了该方法如何对已知区域进行条件化处理。如(8)所示，模型使用x_t预测x_{t-1}，其中x_t包含DDPM（2）的输出和已知区域的样本。然而，使用（7）对已知像素进行采样时并未考虑图像已生成的部分，这导致了不协调。尽管模型在每一步都试图重新协调图像，但由于下一步会出现相同问题，它永远无法完全收敛。此外，在每个反向步骤中，由于β_t的方差调度，图像的最大变化逐渐减小。因此，由于灵活性受限，该方法无法纠正后续步骤中导致边界不协调的错误。因此，模型在进入下一个去噪步骤之前，需要更多时间在单一步骤内协调条件信息x_{t-1}^{known}与生成信息x_{t-1}^{unknown}。

由于DDPM被训练用于生成位于数据分布内的图像，它自然致力于生成一致的结构。在我们的重采样方法中，我们利用DDPM的这一特性来协调模型的输入。因此，我们通过从(1)中采样$x_t \sim \mathcal{N}(\sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I})$，将输出$x_{t-1}$反向扩散回$x_t$。尽管这一操作会缩小输出规模并引入

Figure 3. **The effect of applying n sampling steps.** The first example with $n = 1$ is the DDPM baseline, the second with $n = 2$ is with one resample step. More resampling steps lead to more harmonized images. The benefit saturates at about $n = 10$ resamplings.

noise, some information incorporated in the generated region x_{t-1}^{unknown} is still preserved in x_t^{unknown}. It leads to a new x_t^{unknown} which is both more harmonized with x_t^{known} and contains conditional information from it.

Since this operation can only harmonize one step, it might not be able to incorporate the semantic information over the entire denoising process. To overcome this problem, we denote the time horizon of this operation as jump length, which is $j = 1$ for the previous case. Similar to the standard change in diffusion speed [7] (*a.k.a.* slowing down), the resampling also increases the runtime of the reverse diffusion. Slowing down applies smaller but more resampling steps by reducing the added variance in each denoising step. However, that is a fundamentally different approach because slowing down the diffusion still has the problem of not harmonizing the image, as described in our resampling strategy. We empirically demonstrate this advantage of our approach in Sec. 5.6.

5. Experiments

We perform extensive experiments for face and generic inpainting, compare to the state-of-the-art solutions, and conduct an ablative analysis. In Section 5.3 and 5.4, we report a detailed discussion of mask robustness and diversity, respectively. We also report with additional results, analysis, and visuals in the appendix.

Algorithm 1 Inpainting using our RePaint approach.

$x_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
for $t = T, \dots, 1$ **do**
  for $u = 1, \dots, U$ **do**
    $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $\epsilon = \mathbf{0}$
    $x_{t-1}^{\text{known}} = \sqrt{\bar{\alpha}_t} x_0 + (1 - \bar{\alpha}_t)\epsilon$
    $z \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $\mathbf{z} = \mathbf{0}$
    $x_{t-1}^{\text{unknown}} = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \boldsymbol{\epsilon}_\theta(x_t, t) \right) + \sigma_t z$
    $x_{t-1} = m \odot x_{t-1}^{\text{known}} + (1 - m) \odot x_{t-1}^{\text{unknown}}$
    if $u < U$ and $t > 1$ **then**
      $x_t \sim \mathcal{N}(\sqrt{1 - \beta_{t-1}} x_{t-1}, \beta_{t-1}\mathbf{I})$
    end if
  end for
end for
return x_0

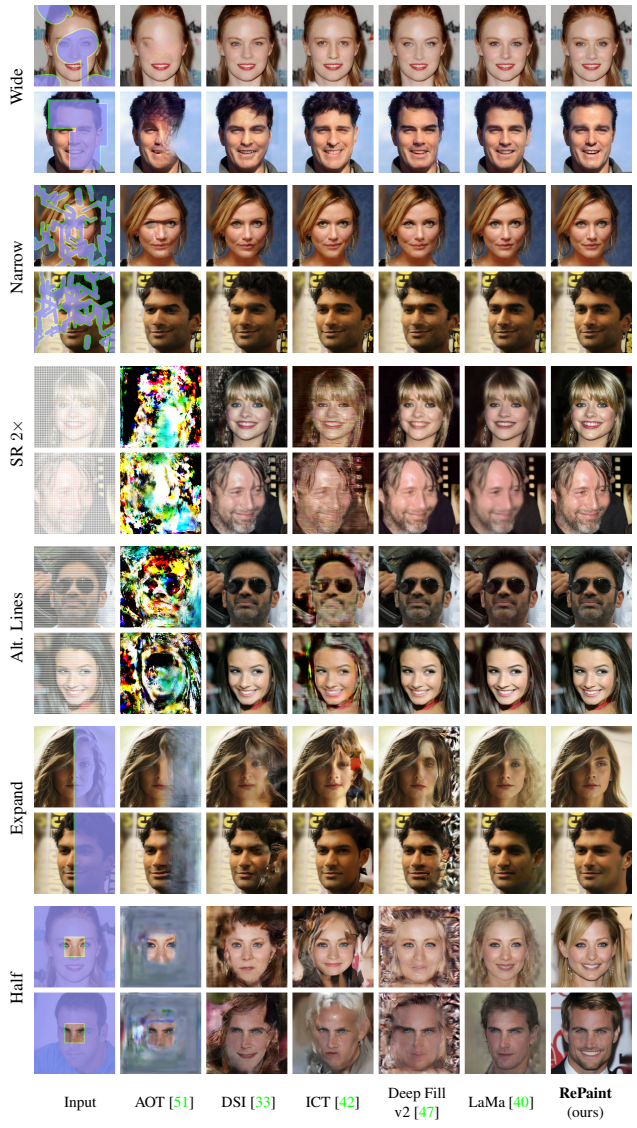


Figure 4. **CelebA-HQ Qualitative Results.** Comparison against the state-of-the-art methods for Face Inpainting over several mask settings. Zoom-in for better details.

图3. 应用n采样步骤的效果。第一个示例采用n =1步，即DDPM基线；第二个示例采用n =2步，即包含一次重采样步骤。更多重采样步骤会使图像更协调。当重采样次数达到约n =10次时，效果趋于饱和。

噪声中，生成区域x_{t-1}^{unknown}所包含的部分信息在x_t^{unknown}中仍得以保留。这催生了一个新的x_t^{unknown}，它既与x_t^{known}更加协调，又包含了来自x_t^{known}的条件信息。

由于此操作仅能协调一个步骤，可能无法整合整个去噪过程中的语义信息。为克服这一问题，我们将此操作的时间跨度定义为跳跃长度，前述案例中该值为$j = 1$。与扩散速度的标准调整方式[7]（*a.k.a.* 减速）类似，重采样同样会增加反向扩散的运行时长。减速策略通过减少每个去噪步骤中增加的方差，采用更小但更频繁的重采样步骤。然而，这是一种根本不同的方法，因为如我们的重采样策略所述，单纯降低扩散速度仍存在无法协调图像的问题。我们将在第5.6节通过实验证明本方法的这一优势。

5. 实验

我们对人脸和通用图像修复进行了广泛的实验，与最先进的解决方案进行了比较，并进行了消融分析。在第5.3节和5.4节中，我们

Algorithm 1 Inpainting using our RePaint approach.

$x_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
for $t = T, \dots, 1$ **do**
  for $u = 1, \dots, U$ **do**
    $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $\epsilon = \mathbf{0}$
    $x_{t-1}^{\text{known}} = \sqrt{\bar{\alpha}_t} x_0 + (1 - \bar{\alpha}_t)\epsilon$
    $z \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $\mathbf{z} = \mathbf{0}$
    $x_{t-1}^{\text{unknown}} = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \boldsymbol{\epsilon}_\theta(x_t, t) \right) + \sigma_t z$
    $x_{t-1} = m \odot x_{t-1}^{\text{known}} + (1 - m) \odot x_{t-1}^{\text{unknown}}$
    if $u < U$ and $t > 1$ **then**
      $x_t \sim \mathcal{N}(\sqrt{1 - \beta_{t-1}} x_{t-1}, \beta_{t-1} \mathbf{I})$
    end if
  end for
end for
return x_0

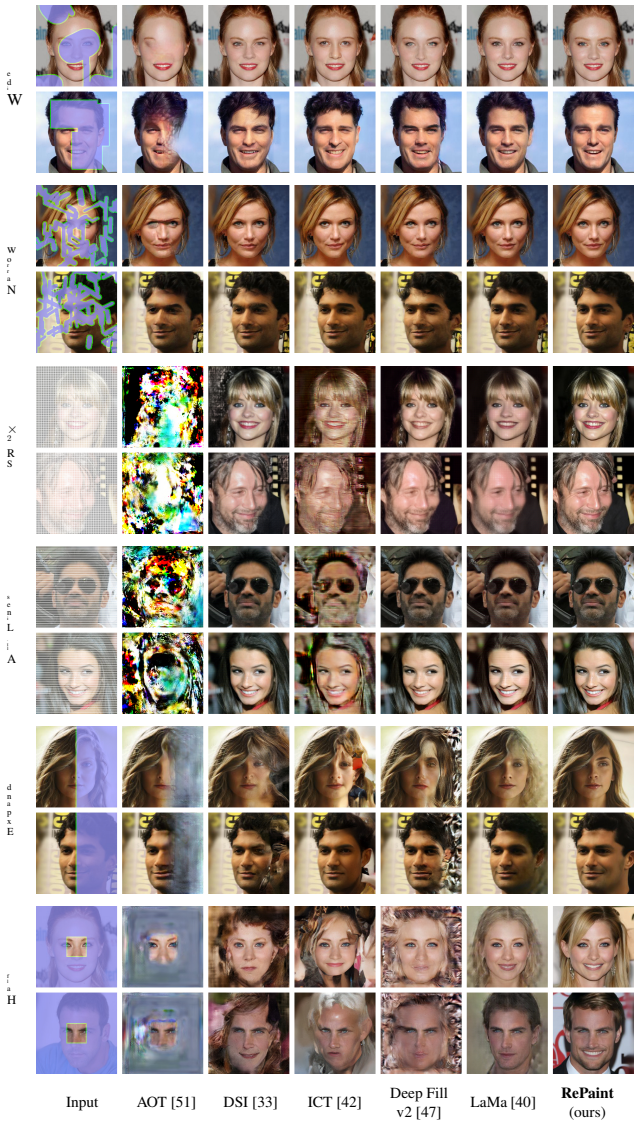


图4. CelebA-HQ定性结果。在多种掩码设置下，与最先进的人脸修复方法进行对比。放大可查看更佳细节。

分别报告了关于掩码鲁棒性和多样性的详细讨论。我们还在附录中提供了额外的结果、分析和可视化材料。

5.1. Implementation Details

We validate our solution over the CelebA-HQ [21], and Imagenet [36] datasets. As our method relies on a pretrained guided diffusion model [7], we use the provided ImageNet model. For CelebA-HQ, we follow the same training hyper-parameters as for ImageNet. We use 256×256 crops in three batches on $4\times$V100 GPUs each. In contrast to the pretrained ImageNet model, the CelebA-HQ one is only trained for 250,000 iterations during roughly five days. Note that all our qualitative and quantitative results in the main paper are for 256 image size.

For our final approach, we use $T = 250$ timesteps, and applied $r = 10$ times resampling with jumpy size $j = 10$.

5.2. Metrics

We compare our RePaint with the baseline methods in a user study described as follows. The user is shown the input image with the blanked missing regions. Next to this image, we display two different inpainting solutions. The user is asked to select "Which image looks more realistic?". The user thus evaluates the realism of our RePaint to the result of a baseline. To avoid biasing the user towards an approach, the methods were anonymized shown in a different random order for each image. Moreover, each user was asked every question twice and could only submit their answer if they were consistent with themselves in at least 75% of their answer. A self-consistency in 100% of the cases is often not possible since, for example, the LaMa method can have a very similar quality to RePaint on the mask settings they provide. Our user study evaluates all 100 test images of the test datasets CelebA-HQ and ImageNet for the following masks: Wide, Narrow, Every Second Line, Half Image, Expand, and Super-Resolve. We use the answers of five different humans for every image query, resulting in 1000 votes per method-to-method comparison in each dataset and mask setting, and show the 95% confidence interval next to the mean votes. In addition to the user study, we report the commonly reported perceptual metric LPIPS [52], which is a learned distance metric based on the deep feature space of AlexNet. We compute the LPIPS over the same 100 test images used in the user study. The results are shown in Table 1. Furthermore, please refer to the appendix for additional quantitative results.

5.3. Comparison with State-of-the-Art

In this section, we first compare our approach against state-of-the-art on standard mask distributions, commonly employed for benchmarking. We then analyze the generalization capabilities of our method against other approaches. To this end, we evaluate their robustness under four challenging mask settings. Firstly, two different masks that probe if the methods can incorporate information from thin structures. Secondly, two masks that require to inpaint a large connected area of the image. All quantitative results are reported in Table 1 and visual results in Figure 4 and 5.

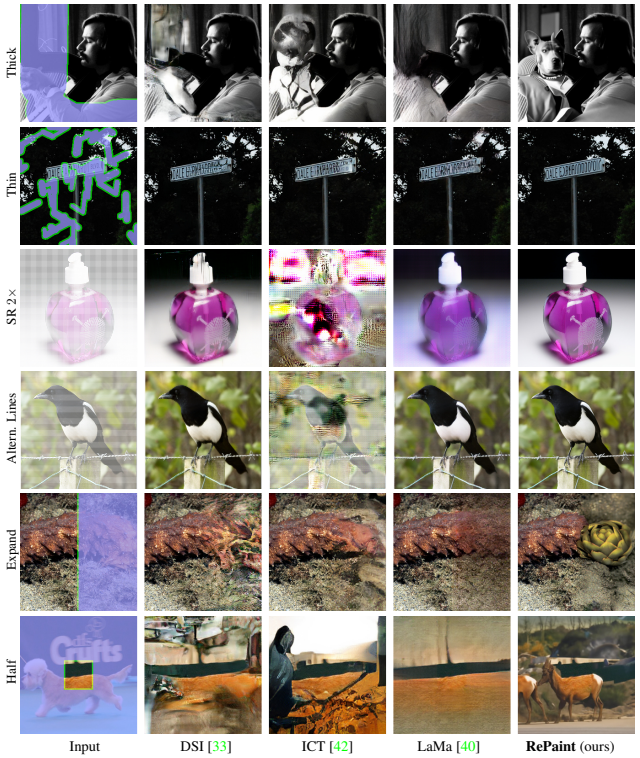


Figure 5. **ImageNet Qualitative Results.** Comparison against the state-of-the-art methods for pluralistic inpainting methods over different mask settings. Zoom-in for better details.

Methods: We compare our approach against several state-of-the-art autoregressive-based or GAN-based approaches. The autoregressive methods are DSI [33] and ICT [42], and the GAN methods are DeepFillv2 [47], AOT [51], and LaMa [40]. We use their publicly available pretrained models. We used the existing FFHQ [17] pretrained model of ICT for our CelebA-HQ testing. As LaMa does not provide ImageNet models, we trained their system for 300,000 iterations of batch size five using the original implementation.

Settings: We use 100 images of size 256×256 from CelebA-HQ [21] and ImageNet test sets. The resulting LPIPS and the average votes of the user study are shown in Table 1. Additionally, refer to the appendix for qualitative and quantitative results over the Places2 [56] dataset.

Wide and Narrow masks: To validate our method on the standard image inpainting scenario, we use the LaMa [40] settings for Wide and Narrow masks. RePaint outperforms all other methods with a significance margin of 95% in both CelebA-HQ and ImageNet, for both Wide and Narrow settings. See qualitative results in Figure 4 and 5 and quantitative in Table 1. The best autoregressive method ICT seems to have less global consistency as observed in Figure 4 in the second row, where the eyes do not to match well. In general,

5.1. 实现细节

我们在CelebA-HQ [21]和Imagenet [36]数据集上验证了我们的解决方案。由于我们的方法依赖于预训练的引导扩散模型[7]，我们使用了提供的ImageNet模型。对于CelebA-HQ，我们采用了与ImageNet相同的训练超参数。我们在4×V100 GPU上分三个批次使用256×256的裁剪图像。与预训练的ImageNet模型相比，CelebA-HQ模型仅训练了约五天，共250,000次迭代。请注意，我们主论文中的所有定性和定量结果均基于256的图像尺寸。

在我们的最终方法中，我们使用了$T = 250$个时间步长，并应用了$r = 10$次重采样，跳跃步长为$j = 10$。

5.2. 指标

我们在用户研究中将我们的RePaint与基线方法进行比较，具体流程如下：向用户展示带有空白缺失区域的输入图像，并在其旁边显示两种不同的修复方案。用户需回答“哪张图像看起来更真实？”，从而评估我们的RePaint与基线结果在真实感上的差异。为避免用户对特定方法产生倾向性，所有方法均匿名处理，且每张图像的展示顺序随机调整。此外，每位用户需对每个问题回答两次，仅当他们的答案在至少75%的情况下保持一致时才能提交。由于某些情况下（例如LaMa方法在其提供的掩码设置中可能与RePaint质量极为接近），100%的一致性往往难以实现。我们的用户研究评估了CelebA-HQ和ImageNet测试数据集中全部100张测试图像，涵盖以下掩码类型：宽掩码、窄掩码、隔行掩码、半图像掩码、扩展掩码和超分辨率掩码。每张图像由五位不同用户进行评价，使得每个数据集和掩码设置下的方法对比共获得1000次投票，我们在平均投票数旁标注了95%置信区间。除用户研究外，我们还报告了常用的感知度量指标LPIPS[52]——这是一种基于AlexNet深度特征空间的学习型距离度量。我们使用用户研究中相同的100张测试图像计算LPIPS，结果如表1所示。更多量化结果请参阅附录。

5.3. 与最先进技术的比较

在本节中，我们首先将我们的方法与最先进的技术在标准掩码分布上进行比较，这些分布通常用于基准测试。然后，我们分析了我们的方法相对于其他方法的泛化能力。为此，我们在四种具有挑战性的掩码设置下评估了它们的鲁棒性。首先，使用两种不同的掩码来探究这些方法是否能从细薄结构中整合信息。其次，使用两种需要修复的掩码来

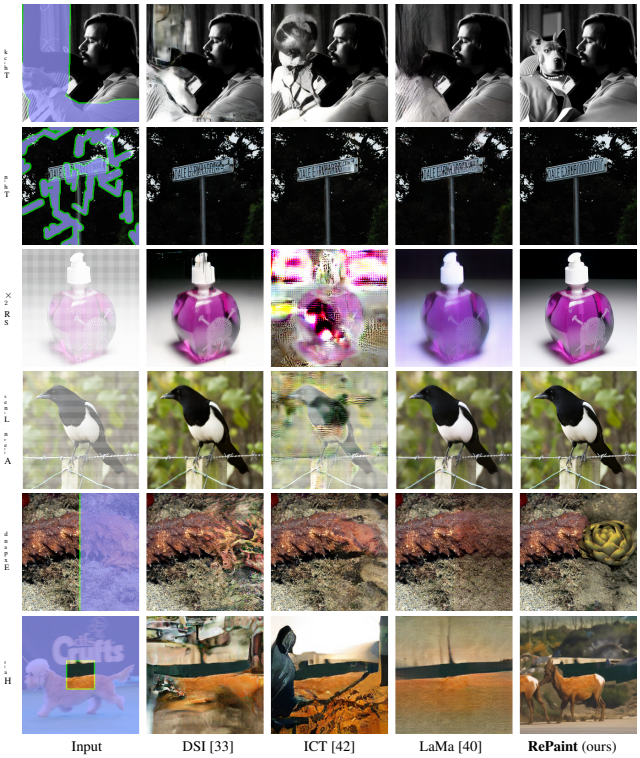


图5. ImageNet定性结果。与不同掩码设置下的最先进多元修复方法进行比较。放大以查看更佳细节。

图像的大面积连通区域。所有定量结果均记录在表1中，视觉结果见图4和图5。

方法：我们将我们的方法与几种基于自回归或基于GAN的先进方法进行比较。自回归方法包括DSI [33]和ICT [42]，GAN方法包括DeepFillv2 [47]、AOT [51]和LaMa [40]。我们使用它们公开可用的预训练模型。对于CelebA-HQ测试，我们使用了ICT现有的FFHQ [17]预训练模型。由于LaMa未提供ImageNet模型，我们使用其原始实现，以批次大小为五训练了300,000次迭代。

设置：我们使用了来自CelebA-HQ [21] 和ImageNet测试集的100张尺寸为256×256的图像。所得LPIPS及用户研究的平均投票结果如表1所示。此外，关于Places2 [56] 数据集的定性与定量结果可参阅附录。

宽与窄掩码：为了在标准图像修复场景中验证我们的方法，我们采用LaMa[40]中宽窄掩码的设置。在CelebA-HQ和ImageNet数据集上，无论是宽掩码还是窄掩码设置，RePaint均以95%的显著优势超越所有其他方法。定性结果见图4与图5，定量结果见表1。最佳自回归方法ICT似乎缺乏全局一致性，如图4第二行所示，其中双眼的匹配度欠佳。总体而言，

CelebA-HQ	Wide		Narrow		Super-Resolve 2×		Altern. Lines		Half		Expand	
Methods	LPIPS↓	Votes [%]	LPIPS↓	Votes [%]	LPIPS↓	Votes [%]	LPIPS↓	Votes [%]	LPIPS↓	Votes [%]	LPIPS↓	Votes [%]
AOT [51]	0.104	11.6 ± 2.0	0.047	12.8 ± 2.1	0.714	1.1 ± 0.6	0.667	2.4 ± 1.0	0.287	9.0 ± 1.8	0.604	8.3 ± 1.7
DSI [33]	0.067	16.0 ± 2.3	0.038	22.3 ± 2.6	0.128	5.5 ± 1.4	0.049	5.1 ± 1.4	0.211	4.5 ± 1.3	0.487	4.7 ± 1.3
ICT [42]	0.063	27.6 ± 2.8	0.036	30.9 ± 2.9	0.483	4.2 ± 1.2	0.353	0.7 ± 0.5	0.166	12.7 ± 2.1	0.432	8.8 ± 1.8
DeepFillv2 [47]	0.066	23.9 ± 2.6	0.049	21.0 ± 2.5	0.119	9.8 ± 1.8	0.049	10.6 ± 1.9	0.209	4.1 ± 1.2	0.467	13.1 ± 2.1
LaMa [40]	0.045	41.8 ± 3.1	0.028	33.8 ± 3.0	0.177	5.5 ± 1.4	0.083	20.6 ± 2.5	0.138	35.6 ± 3.0	0.342	24.7 ± 2.7
RePaint	0.059	*Reference*	0.028	*Reference*	0.029	*Reference*	0.009	*Reference*	0.165	*Reference*	0.435	*Reference*

ImageNet	Wide		Narrow		Super-Resolve 2×		Altern. Lines		Half		Expand	
Methods	LPIPS↓	Votes [%]	LPIPS↓	Votes [%]	LPIPS↓	Votes [%]	LPIPS↓	Votes [%]	LPIPS↓	Votes [%]	LPIPS↓	Votes [%]
DSI [33]	0.117	31.7 ± 2.9	0.072	28.6 ± 2.8	0.153	26.9 ± 2.8	0.069	23.6 ± 2.6	0.283	31.4 ± 2.9	0.583	9.2 ± 1.8
ICT [42]	0.107	42.9 ± 3.1	0.073	33.0 ± 2.9	0.708	1.1 ± 0.6	0.620	6.6 ± 1.5	0.255	51.5 ± 3.1	0.544	25.6 ± 2.7
LaMa [40]	0.105	42.4 ± 3.1	0.061	33.6 ± 2.9	0.272	13.0 ± 2.1	0.121	9.6 ± 1.8	0.254	41.1 ± 3.1	0.534	20.3 ± 2.5
RePaint	0.134	*Reference*	0.064	*Reference*	0.183	*Reference*	0.089	*Reference*	0.304	*Reference*	0.629	*Reference*

Table 1. **CelebA-HQ (*top*) and ImageNet (*bottom*) Quantitative Results.** Comparison against the state-of-the-art methods. We compute the LPIPS (lower is better) and *Votes* for six different mask settings. *Votes* refers to the ratio of votes with respect to ours.

the best GAN approach LaMa [40] has better global consistency, yet it produces notorious checkerboard artifacts. Those observations might have influenced the users to vote for RePaint for the majority of images, in which our method generates more realistic images.

Thin Masks: Similar to a Nearest-Neighbor Super Resolution problem, the "Super-Resolution 2×" mask only leaves pixels with a stride of 2 in height and width dimension, and the "Alternating Lines" mask removes the pixels every second row of an image. As seen in Figure 4 and 5, AOT [51] fails completely, while the others either produce blurry images or generate visible artifacts, or both. These observations are also confirmed by the user study, where RePaint achieves between 73.1% and 99.3% of the user votes.

Thick Masks: The "Expand" mask only leaves a center crop of 64×64 from a 256×256 image, and "Half" mask, which provides the left half of the image as input. As there is less contextual information, most of the methods struggle (see Figure 4 and 5). Qualitatively, LaMa comes closer to ours, yet our generated images are sharper and have overall more semantic hallucination. Noteworthy, LaMa outperforms RePaint in therms of LPIPS on "Expand" and "Half" for both CelebA and ImageNet (Tab. 1). We argue that this behavior is due to our method being more flexible and diverse in the generation. By generating a semantically different image than that in the Ground-Truth, it makes the LPIPS an unsuitable metric for this particular solution.

The artifacts produced by the baselines can be explained by strong overfitting to the training masks. In contrast, as our method does not involve mask training, our RePaint can handle any type of mask. In the case of large-area inpainting, RePaint produces a semantically meaningful filling, while others generate artifacts or copy texture. Finally, RePaint is preferred by the users with confidence 95% except for the inconclusive result of ICT with "Half" masks as shown in Table 1.

Figure 6. Visual results for class guided generation on ImageNet.

5.4. Analysis of Diversity

As shown in (2), every reverse diffusion step is inherently stochastic since it incorporates new noise from a Gaussian Distribution. Moreover, as we do not directly guide the inpainted area with any loss, the model is, therefore, free to inpaint anything that semantically aligns with the training set. Figure 1 illustrates the diversity and flexibility of our model.

5.5. Class conditional Experiment

The pretrained ImageNet DDPM is capable of class-conditional generation sampling. In Figure 6 we show examples for the "Expand" mask for the "Granny Smith" class, as well as other classes.

5.6. Ablation Study

Comparison to slowing down: To analyze if the increased computational budget causes the improved performance of resampling, we compare it with the commonly used technique of slowing down the diffusion process as described in Section 4.2. Therefore, in Figure 7 and Table 2, we show a comparison resampling and the slow down in diffusion using the same computational budget for each setting. We observe that the resampling uses the extra computational budget for harmonizing the image, whereas there is no visible improvement at slowing down the diffusion process.

Jumps Length: Moreover, to ablate the jump lengths j and the number of resampling r, we study nine different settings in Table 3. We obtain better performance at applying the larger jump $j = 10$ length than smaller step length

CelebA-HQ	Wide		Narrow		Super-Resolve 2×		Altern. Lines		Half		Expand	
Methods	LPIPS↓	Votes [%]	LPIPS↓	Votes [%]	LPIPS↓	Votes [%]	LPIPS↓	Votes [%]	LPIPS↓	Votes [%]	LPIPS↓	Votes [%]
AOT [51]	0.104	11.6 ± 2.0	0.047	12.8 ± 2.1	0.714	1.1 ± 0.6	0.667	2.4 ± 1.0	0.287	9.0 ± 1.8	0.604	8.3 ± 1.7
DSI [33]	0.067	16.0 ± 2.3	0.038	22.3 ± 2.6	0.128	5.5 ± 1.4	0.049	5.1 ± 1.4	0.211	4.5 ± 1.3	0.487	4.7 ± 1.3
ICT [42]	0.063	27.6 ± 2.8	0.036	30.9 ± 2.9	0.483	4.2 ± 1.2	0.353	0.7 ± 0.5	0.166	12.7 ± 2.1	0.432	8.8 ± 1.8
DeepFillv2 [47]	0.066	23.9 ± 2.6	0.049	21.0 ± 2.5	0.119	9.8 ± 1.8	0.049	10.6 ± 1.9	0.209	4.1 ± 1.2	0.467	13.1 ± 2.1
LaMa [40]	0.045	41.8 ± 3.1	0.028	33.8 ± 3.0	0.177	5.5 ± 1.4	0.083	20.6 ± 2.5	0.138	35.6 ± 3.0	0.342	24.7 ± 2.7
RePaint	0.059	*Reference*	0.028	*Reference*	0.029	*Reference*	0.009	*Reference*	0.165	*Reference*	0.435	*Reference*

ImageNet	Wide		Narrow		Super-Resolve 2×		Altern. Lines		Half		Expand	
Methods	LPIPS↓	Votes [%]	LPIPS↓	Votes [%]	LPIPS↓	Votes [%]	LPIPS↓	Votes [%]	LPIPS↓	Votes [%]	LPIPS↓	Votes [%]
DSI [33]	0.117	31.7 ± 2.9	0.072	28.6 ± 2.8	0.153	26.9 ± 2.8	0.069	23.6 ± 2.6	0.283	31.4 ± 2.9	0.583	9.2 ± 1.8
ICT [42]	0.107	42.9 ± 3.1	0.073	33.0 ± 2.9	0.708	1.1 ± 0.6	0.620	6.6 ± 1.5	0.255	51.5 ± 3.1	0.544	25.6 ± 2.7
LaMa [40]	0.105	42.4 ± 3.1	0.061	33.6 ± 2.9	0.272	13.0 ± 2.1	0.121	9.6 ± 1.8	0.254	41.1 ± 3.1	0.534	20.3 ± 2.5
RePaint	0.134	*Reference*	0.064	*Reference*	0.183	*Reference*	0.089	*Reference*	0.304	*Reference*	0.629	*Reference*

表1. CelebA-HQ (***top***) 与 ImageNet (***bottom***) 定量结果。与最先进方法的比较。我们针对六种不同掩码设置计算了 LPIPS（越低越好）和 *Votes*。*Votes* 表示相对于我们方法的投票比例。

最好的GAN方法LaMa [40]具有更好的全局一致性，但它会产生臭名昭著的棋盘格伪影。这些观察结果可能影响了用户为大多数图像投票给RePaint，而我们的方法在这些图像中生成了更逼真的图像。

薄掩码：类似于最近邻超分辨率问题，“超分辨率2×”掩码仅在高度和宽度维度上保留步长为2的像素，而“交替行”掩码则移除图像中每隔一行的像素。如图4和图5所示，AOT[51]完全失效，而其他方法要么生成模糊图像，要么产生可见伪影，或两者兼有。这些观察结果也在用户研究中得到证实，其中RePaint获得了73.1%至99.3%的用户投票。

厚掩码："扩展"掩码仅从256×256图像中保留64×64的中心裁剪区域，而"半幅"掩码则提供图像的左半部分作为输入。由于上下文信息较少，大多数方法都表现不佳（见图4和图5）。从质量上看，LaMa的结果更接近我们的方法，但我们生成的图像更清晰，且整体具有更强的语义幻觉。值得注意的是，在CelebA和ImageNet数据集上，LaMa在"扩展"和"半幅"掩码的LPIPS指标上均优于RePaint（见表1）。我们认为这一现象源于我们的方法在生成过程中具有更高的灵活性和多样性。由于生成的图像在语义上与真实图像存在差异，这使得LPIPS成为不适合评估此特定解决方案的指标。

基线方法产生的伪影可以解释为对训练掩码的强烈过拟合。相比之下，由于我们的方法不涉及掩码训练，我们的RePaint能够处理任何类型的掩码。在大面积修复的情况下，RePaint能生成语义上有意义的填充，而其他方法则会产生伪影或复制纹理。最后，如表1所示，除了ICT在“Half”掩码上的不确定结果外，用户以95%的置信度更倾向于选择RePaint。

图6. ImageNet上类别引导生成的视觉结果。

5.4. 多样性分析

如（2）所示，每个反向扩散步骤本质上都是随机的，因为它从高斯分布中引入了新的噪声。此外，由于我们没有通过任何损失函数直接引导修复区域，因此模型可以自由地修复任何在语义上与训练集对齐的内容。图1展示了我们模型的多样性和灵活性。

5.5. 类别条件实验

预训练的ImageNet DDPM能够进行类别条件生成采样。在图6中，我们展示了“Granny Smith”类别及其他类别的“Expand”掩码示例。

5.6. 消融研究

与减速对比：为了分析增加的计算预算是否提升了重采样的性能，我们将其与第4.2节中描述的常用扩散过程减速技术进行比较。因此，在图7和表2中，我们展示了在每种设置下使用相同计算预算时，重采样与扩散减速的对比。我们观察到，重采样将额外的计算预算用于协调图像，而扩散过程减速则未见明显改善。

跳跃长度：此外，为了消融跳跃长度j和重采样次数r的影响，我们在表3中研究了九种不同设置。我们发现应用较大的跳跃长度j =10比使用较小步长能获得更好的性能。

	T	r	LPIPS	T	r	LPIPS	T	r	LPIPS	T	r	LPIPS
Slowing down	250	1	0.168	500	1	0.167	750	1	0.179	1000	1	0.161
Resampling	250	1	0.168	250	2	0.148	250	3	0.142	250	4	0.134

Table 2. **Analysis of the use of computational budget**. We compare slowing down the diffusion process and resampling. We use the ImageNet validation set with 32 images over the LaMa [40] Wide mask setting. The number of diffusion steps is denoted by T, and the number of resamplings by r.

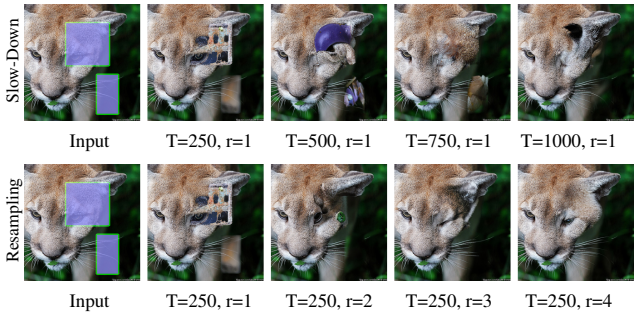


Figure 7. **Qualitative Analysis of the use of computational budget.** RePaint produces higher visual quality with the same computational budget by resampling (*bottom*) compared to slowing down the diffusion process (*top*). The number of diffusion steps is denoted by T and resamplings by r.

steps. We observe that for jump length $j = 1$, the DDPM is more likely to output a blurry image. Furthermore, this observation is stable across different numbers of resampling. Furthermore, the number of resamplings increases the performance.

Comparison to alternative sampling strategy: To compare our resampling approach to SDEdit [25], we first perform reverse diffusion from $t = T$ to $t = T/2$ to obtain the required initial inpainting at $t = T/2$. We then apply the resampling method from SDEdit, which repeats the reverse process from $t = T/2$ to $t = 0$ several times. The results are shown in Table 4. Our approach achieves significantly better performance across all mask types except for one "Expand" case, where LPIPS > 0.6 is outside a meaningful range for comparisons. In case of 'super-resolution masks', our approach reduces the LPIPS by over 53% on all datasets, clearly demonstrating the advantage of our resampling strategy.

	$j = 1$		$j = 5$		$j = 10$	
r	LPIPS	Votes [%]	LPIPS	Votes [%]	LPIPS	Votes [%]
5	0.075	42.50±7.7	0.072	46.88±7.8	0.073	53.12±7.8
10	0.088	42.50±7.7	0.073	45.62±7.8	0.068	56.25±7.8
15	0.065	46.25±7.8	0.063	53.12±5.5	0.065	53.75±7.8

Table 3. **Ablation Study.** Analysis of length of the jumps j and number of resamplings r. We report LPIPS and the average user-study votes with respect to LaMa [40]. We use 32 images from the CelebA validation set, and use the LaMa Wide mask setting.

Dataset	Method	Wide	Narrow	Super-Res.	Alt. Lin.	Half	Expand
ImageNet	SDEdit [25]	0.1532	0.0952	0.3902	0.1852	0.3272	**0.6281**
	RePaint (Ours)	**0.1341**	**0.0641**	**0.1831**	**0.0891**	**0.3041**	0.6292
Places2	SDEdit [25]	0.1302	0.0622	0.2712	0.1302	0.3042	0.6202
	RePaint (Ours)	**0.1051**	**0.0441**	**0.0991**	**0.0511**	**0.2861**	**0.6151**
CelebA-HQ	SDEdit [25]	0.0762	0.0462	0.1132	0.0302	0.1892	0.4492
	RePaint (Ours)	**0.0591**	**0.0281**	**0.0291**	**0.0091**	**0.1651**	**0.4351**

Table 4. Comparison with the resampling schedule proposed in [25] in terms of LPIPS. The resampling method proposed in our RePaint (Sec. 4.2) achieves substantially better results, in particular for the Super-Resolution masks.

6. Limitations

Our method produces sharp, highly detailed, and semantically meaningful images. We believe that our work opens interesting research directions for addressing the current limitations of the method. Two directions are of particular interest. First, naturally, the per-image DDPM optimization process is significantly slower than the GAN-based and Autoregressive-based counterparts. That makes it currently difficult to apply it for real-time applications. Nonetheless, DDPM is gaining in popularity, and recent publications are working on improving the efficiency [23, 24]. Secondly, for the extreme mask cases, RePaint can produce realistic images completions that are very different from the Ground Truth image. That makes the quantitative evaluation challenging for those conditions. An alternative solution is to employ the FID score [11] over a test set. However, a reliable FID for inpainting is usually computed with more than 1,000 images. For current DDPM, this would result in a runtime that is not feasible for most research institutes.

7. Potential Negative Societal Impact

On the one hand, RePaint is an inpainting method that relies on an unconditional pretrained DDPM. Therefore, the algorithm might be biased towards the dataset on which it was trained. Since the model aims to generate images of the same distribution as the training set, it might reflect the same biases, such as gender, age, and ethnicity. On the other hand, RePaint could be used for the anonymization of faces. For example, one could remove the information about the identity of people shown at public events and hallucinate artificial faces for data protection.

8. Conclusions

We presented a novel denoising diffusion probabilistic model solution for the image inpainting task. In detail, we developed a mask-agnostic approach that widely increases the degree of freedom of masks for the free-form inpainting. Since the novel conditioning approach of RePaint complies with the model assumptions of a DDPM, it produces a photo-realistic image regardless of the type of the mask.

Acknowledgements: This work was supported by the ETH Zürich Fund (OK), a Huawei Technologies Oy (Finland) project, and an Nvidia GPU grant.

	T	r	LPIPS	T	r	LPIPS	T	r	LPIPS	T	r	LPIPS
Slowing down	250	1	0.168	500	1	0.167	750	1	0.179	1000	1	0.161
Resampling	250	1	0.168	250	2	0.148	250	3	0.142	250	4	0.134

表2：计算预算使用情况分析。我们比较了减缓扩散过程和重采样两种方法。我们在LaMa [40] 宽掩码设置下，使用包含32张图像的ImageNet验证集进行实验。扩散步数记为T，重采样次数记为r。

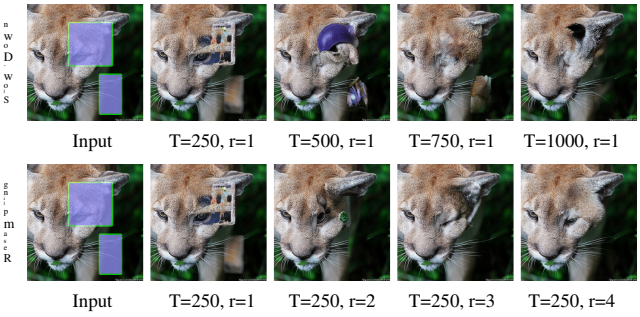


图7. 计算预算使用的定性分析。在相同计算预算下，通过重采样(*bottom*)相比减缓扩散过程(*top*)，RePaint能生成更高的视觉质量。扩散步数由T表示，重采样次数由r表示。

步骤。我们观察到，对于跳跃长度j =为1的情况，DDPM更可能输出模糊图像。此外，这一观察在不同重采样次数下均保持稳定。同时，重采样次数的增加会提升性能。

与替代采样策略的比较：为了将我们的重采样方法与SDEdit [25]进行比较，我们首先执行从$t = T$到$t = T/2$的反向扩散，以获得在$t = T/2$处所需的初始修复。然后，我们应用SDEdit中的重采样方法，该方法多次重复从$t = T/2$到t =0的反向过程。结果如表4所示。除了一种“扩展”情况外，我们的方法在所有掩码类型上都取得了显著更好的性能，在该情况下，LPIPS >0.6超出了有意义的比较范围。对于“超分辨率掩码”，我们的方法在所有数据集上将LPIPS降低了超过53%，这清楚地证明了我们重采样策略的优势。

	$j = 1$		$j = 5$		$j = 10$	
r	LPIPS	Votes [%]	LPIPS	Votes [%]	LPIPS	Votes [%]
5	0.075	42.50±7.7	0.072	46.88±7.8	0.073	53.12±7.8
10	0.088	42.50±7.7	0.073	45.62±7.8	0.068	56.25±7.8
15	0.065	46.25±7.8	0.063	53.12±5.5	0.065	53.75±7.8

表3. 消融研究。关于跳跃长度j和重采样次数r的分析。我们报告了相对于LaMa[40]的LPIPS和平均用户研究投票数。我们使用了CelebA验证集中的32张图像，并采用了LaMa的宽掩码设置。

Dataset	Method	Wide	Narrow	Super-Res.	Alt. Lin.	Half	Expand
ImageNet	SDEdit [25]	0.1532	0.0952	0.3902	0.1852	0.3272	**0.6281**
	RePaint (Ours)	**0.1341**	**0.0641**	**0.1831**	**0.0891**	**0.3041**	0.6292
Places2	SDEdit [25]	0.1302	0.0622	0.2712	0.1302	0.3042	0.6202
	RePaint (Ours)	**0.1051**	**0.0441**	**0.0991**	**0.0511**	**0.2861**	**0.6151**
CelebA-HQ	SDEdit [25]	0.0762	0.0462	0.1132	0.0302	0.1892	0.4492
	RePaint (Ours)	**0.0591**	**0.0281**	**0.0291**	**0.0091**	**0.1651**	**0.4351**

表4. 与文献[25]提出的重采样方案在LPIPS指标上的对比。我们RePaint方法（第4.2节）提出的重采样方案取得了显著更好的结果，尤其在超分辨率掩码任务上表现突出。

6. 局限性

我们的方法能够生成清晰、细节丰富且语义明确的图像。我们相信，这项工作为解决当前方法的局限性开辟了有趣的研究方向。其中两个方向尤其值得关注。首先，自然，基于每张图像的DDPM优化过程明显慢于基于GAN和自回归模型的方法，这使得目前难以将其应用于实时场景。尽管如此，DDPM正日益受到关注，近期的研究也在致力于提升其效率[23, 24]。其次，在极端掩码情况下，RePaint可以生成与真实图像差异巨大但视觉效果逼真的补全结果，这为定量评估这些条件下的性能带来了挑战。一种替代方案是在测试集上采用FID分数[11]进行评估，但可靠的修复任务FID通常需要超过1000张图像进行计算。对于当前的DDPM而言，这将导致运行时间过长，对大多数研究机构来说难以实现。

7. 潜在负面社会影响

一方面，RePaint是一种依赖于无条件预训练DDPM的图像修复方法。因此，该算法可能偏向于其训练所用的数据集。由于模型旨在生成与训练集相同分布的图像，它可能会反映相同的偏见，例如性别、年龄和种族。另一方面，RePaint可用于人脸匿名化。例如，可以移除公共活动中人物身份信息，并通过生成人工面部以实现数据保护。

8. 结论

我们提出了一种新颖的去噪扩散概率模型解决方案，用于图像修复任务。具体而言，我们开发了一种与掩码无关的方法，广泛提高了自由形式修复中掩码的自由度。由于RePaint的新颖条件方法符合DDPM的模型假设，无论掩码类型如何，它都能生成逼真的图像。

致谢：本工作得到了ETH Zürich基金（OK）、华为技术有限公司（芬兰）项目以及英伟达GPU资助的支持。

References

[1] Coloma Ballester, Marcelo Bertalmio, Vicent Caselles, Guillermo Sapiro, and Joan Verdera. Filling-in by joint interpolation of vector fields and gray levels. *IEEE transactions on image processing*, 10(8):1200–1211, 2001. 2

[2] Marcelo Bertalmio, Guillermo Sapiro, Vincent Caselles, and Coloma Ballester. Image inpainting. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 417–424, 2000. 2

[3] Marcelo Bertalmio, Luminita Vese, Guillermo Sapiro, and Stanley Osher. Simultaneous structure and texture image inpainting. *IEEE transactions on image processing*, 12(8):882–889, 2003. 2

[4] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 2

[5] Kelvin CK Chan, Xintao Wang, Xiangyu Xu, Jinwei Gu, and Chen Change Loy. Glean: Generative latent bank for large-factor image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14245–14254, 2021. 2

[6] Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. Ilvr: Conditioning method for denoising diffusion probabilistic models. *arXiv preprint arXiv:2108.02938*, 2021. 3

[7] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. *arXiv preprint arXiv:2105.05233*, 2021. 2, 4, 5, 6, 12, 15

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 2

[9] Xiefan Guo, Hongyu Yang, and Di Huang. Image inpainting via conditional texture and structure dual generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14134–14143, 2021. 2

[10] James Hays and Alexei A Efros. Scene completion using millions of photographs. *ACM Transactions on Graphics (ToG)*, 26(3):4–es, 2007. 2

[11] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in Neural Information Processing Systems 30 (NIPS 2017), 2017. 8

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. 2, 3, 4

[13] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 3

[14] Seunghoon Hong, Xinchen Yan, Thomas Huang, and Honglak Lee. Learning hierarchical semantic image manipulation through structured representations. *arXiv preprint arXiv:1808.07535*, 2018. 2

[15] Zheng Hui, Jie Li, Xiumei Wang, and Xinbo Gao. Image fine-grained inpainting. *arXiv preprint arXiv:2002.02609*, 2020. 2

[16] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. *ACM Transactions on Graphics (ToG)*, 36(4):1–14, 2017. 2

[17] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. 2, 6

[18] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020. 2

[19] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 85–100, 2018. 2

[20] Hongyu Liu, Bin Jiang, Yibing Song, Wei Huang, and Chao Yang. Rethinking image inpainting via a mutual encoder-decoder with feature equalizations. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 725–741. Springer, 2020. 2

[21] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015. 2, 6

[22] Andreas Lugmayr, Martin Danelljan, and Radu Timofte. Ntire 2021 learning the super-resolution space challenge. In *CVPRW*, 2021. 15

[23] Eric Luhman and Troy Luhman. Denoising synthesis: A module for fast image synthesis using denoising-based models. *Software Impacts*, 9:100076, 2021. 8

[24] Eric Luhman and Troy Luhman. Knowledge distillation in iterative generative models for improved sampling speed. *arXiv preprint arXiv:2101.02388*, 2021. 8

[25] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 3, 8

[26] Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 2437–2445, 2020. 2

[27] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Qureshi, and Mehran Ebrahimi. Edgeconnect: Structure guided image inpainting using edge prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 2

[28] Alex Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. *arXiv preprint arXiv:2102.09672*, 2021. 2, 4

[29] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation

参考文献

[1] Coloma Ballester, Marcelo Bertalmio, Vicent Caselles, Guillermo Sapiro, 和 Joan Verdera。通过矢量场和灰度级的联合插值进行填充。*IEEE transactions on image processing*, 10(8):1200−1211, 2001. 2[2] Marcelo Bertalmio, Guillermo Sapiro, Vincent Caselles, 和 Coloma Ballester。图像修复。收录于 *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, 页 417−424, 2000. 2[3] Marcelo Bertalmio, Luminita Vese, Guillermo Sapiro, 和 Stanley Osher。同时进行结构和纹理的图像修复。*IEEE transactions on image processing*, 12(8):882−889, 2003. 2[4] Andrew Brock, Jeff Donahue, 和 Karen Simonyan。用于高保真自然图像合成的大规模 GAN 训练。*arXiv preprint arXiv:1809.11096*, 2018. 2[5] Kelvin CK Chan, Xintao Wang, Xiangyu Xu, Jinwei Gu, 和 Chen Change Loy。GLEAN：用于大因子图像超分辨率的生成式潜在库。收录于 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 页 14245−14254, 2021. 2[6] Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, 和 Sungroh Yoon。ILVR：去噪扩散概率模型的调节方法。*arXiv preprint arXiv:2108.02938*, 2021. 3[7] Prafulla Dhariwal 和 Alex Nichol。扩散模型在图像合成上击败 GAN。*arXiv preprint arXiv:2105.05233*, 2021. 2, 4, 5, 6, 12, 15[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, 和 Yoshua Bengio。生成对抗网络。*Advances in neural information processing systems*, 27, 2014. 2[9] Xiefan Guo, Hongyu Yang, 和 Di Huang。通过条件纹理和结构双重生成的图像修复。收录于 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 页 14134−14143, 2021. 2[10] James Hays 和 Alexei A Efros。使用数百万张照片进行场景补全。*ACM Transactions on Graphics (ToG)*, 26(3):4−es, 2007. 2[11] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, 和 Sepp Hochreiter。通过双时间尺度更新规则训练的 GAN 收敛到局部纳什均衡。神经信息处理系统进展 30 (NIPS 2017), 2017. 8[12] Jonathan Ho, Ajay Jain, 和 Pieter Abbeel。去噪扩散概率模型, 2020. 2, 3, 4[13] Jonathan Ho 和 Tim Salimans。无分类器扩散引导。收录于 *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 3[14] Seunghoon Hong, Xinchen Yan, Thomas Huang, 和 Honglak Lee。通过结构化表示学习分层语义图像操作。*arXiv preprint arXiv:1808.07535*, 2018. 2[15] Zheng Hui, Jie Li, Xiumei Wang, 和 Xinbo Gao。图像细粒度修复。*arXiv preprint arXiv:2002.02609*, 2020. 2[16] Satoshi Iizuka, Edgar Simo-Serra, 和 Hiroshi Ishikawa。全局与局部一致的图像补全。*ACM Transactions on Graphics (ToG)*, 36(4):1−14, 2017. 2[17] Tero Karras, Samuli Laine, 和 Timo Aila。一种用于生成对抗网络的基于风格的生成器架构。发表于 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 页 4401−4410, 2019. 2, 6[18] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, 和 Timo Aila。分析与改进StyleGAN的图像质量。发表于 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 页 8110−8119, 2020. 2[19] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, 和 Bryan Catanzaro。使用部分卷积进行不规则孔洞的图像修复。发表于 *Proceedings of the European Conference on Computer Vision (ECCV)*, 页 85−100, 2018. 2[20] Hongyu Liu, Bin Jiang, Yibing Song, Wei Huang, 和 Chao Yang。通过具有特征均衡化的互编码器-解码器重新思考图像修复。发表于 *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, 页 725−741. Springer, 2020. 2[21] Ziwei Liu, Ping Luo, Xiaogang Wang, 和 Xiaoou Tang。在野外深度学习人脸属性。发表于 *Proceedings of International Conference on Computer Vision (ICCV)*, 2015年12月. 2, 6[22] Andreas Lugmayr, Martin Danelljan, 和 Radu Timofte。NTIRE 2021学习超分辨率空间挑战赛。发表于 *CVPRW*, 2021. 15[23] Eric Luhman 和 Troy Luhman。去噪合成：一个使用基于去噪模型进行快速图像合成的模块。*Software Impacts*, 9:100076, 2021. 8[24] Eric Luhman 和 Troy Luhman。迭代生成模型中的知识蒸馏以提高采样速度。*arXiv preprint arXiv:2101.02388*, 2021. 8[25] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, 和 Stefano Ermon。SDEdit：使用随机微分方程进行引导图像合成与编辑。*arXiv preprint arXiv:2108.01073*, 2021. 3, 8[26] Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, 和 Cynthia Rudin。PULSE：通过生成模型潜在空间探索进行自监督照片上采样。发表于 *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, 页 2437−2445, 2020. 2[27] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Qureshi, 和 Mehran Ebrahimi。EdgeConnect：使用边缘预测进行结构引导的图像修复。发表于 *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 页 0−0, 2019. 2[28] Alex Nichol 和 Prafulla Dhariwal。改进的去噪扩散概率模型。*arXiv preprint arXiv:2102.09672*, 2021. 2, 4[29] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, 和 Mark Chen。GLIDE：迈向逼真的图像生成

and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 3

[30] Evangelos Ntavelis, Andrés Romero, Siavash Bigdeli, Radu Timofte, Zheng Hui, Xiumei Wang, Xinbo Gao, Chajin Shin, Taeoh Kim, Hanbin Son, et al. Aim 2020 challenge on image extreme inpainting. In *European Conference on Computer Vision*, pages 716–741. Springer, 2020. 2

[31] Evangelos Ntavelis, Andrés Romero, Iason Kastanis, Luc Van Gool, and Radu Timofte. Sesame: semantic editing of scenes by adding, manipulating or erasing objects. In *European Conference on Computer Vision*, pages 394–411. Springer, 2020. 2

[32] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016. 2

[33] Jialun Peng, Dong Liu, Songcen Xu, and Houqiang Li. Generating diverse structure for image inpainting with hierarchical vq-vae. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10775–10784, 2021. 2, 5, 6, 7, 13, 15, 17, 18, 19, 20, 21, 22, 23, 24, 25

[34] Yurui Ren, Xiaoming Yu, Ruonan Zhang, Thomas H Li, Shan Liu, and Ge Li. Structureflow: Image inpainting via structure-aware appearance flow. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 181–190, 2019. 2

[35] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2287–2296, 2021. 2

[36] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 2, 6

[37] Chitwan Saharia, William Chan, Huiwen Chang, Chris A Lee, Jonathan Ho, Tim Salimans, David J Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. *arXiv preprint arXiv:2111.05826*, 2021. 3

[38] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics, 2015. 2, 3

[39] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations, 2020. 3

[40] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. *arXiv preprint arXiv:2109.07161*, 2021. 2, 5, 6, 7, 8, 12, 13, 14, 15, 17, 18, 19, 20, 21, 22, 23, 24, 25

[41] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9446–9454, 2018. 2

[42] Ziyu Wan, Jingbo Zhang, Dongdong Chen, and Jing Liao. High-fidelity pluralistic image completion with transformers. *arXiv preprint arXiv:2103.14031*, 2021. 2, 5, 6, 7, 13, 15, 17, 18, 19, 20, 21, 22, 23, 24, 25

[43] Wei Xiong, Jiahui Yu, Zhe Lin, Jimei Yang, Xin Lu, Connelly Barnes, and Jiebo Luo. Foreground-aware image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5840–5848, 2019. 2

[44] Shunxin Xu, Dong Liu, and Zhiwei Xiong. E2i: Generative inpainting from edge to image. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(4):1308–1322, 2020. 2

[45] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015. 2

[46] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5505–5514, 2018. 2

[47] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. *arXiv preprint arXiv:1801.07892*, 2018. 5, 6, 7, 13, 15, 17, 18, 19, 23, 24, 25

[48] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4471–4480, 2019. 2

[49] Yingchen Yu, Fangneng Zhan, Rongliang Wu, Jianxiong Pan, Kaiwen Cui, Shijian Lu, Feiying Ma, Xuansong Xie, and Chunyan Miao. Diverse image inpainting with bidirectional and autoregressive transformers. In *Proceedings of the 29th ACM International Conference on Multimedia*, 2021. 2

[50] Yanhong Zeng, Jianlong Fu, Hongyang Chao, and Baining Guo. Learning pyramid-context encoder network for high-quality image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1486–1494, 2019. 2

[51] Yanhong Zeng, Jianlong Fu, Hongyang Chao, and Baining Guo. Aggregated contextual transformations for high-resolution image inpainting. *arXiv preprint arXiv:2104.01431*, 2021. 2, 5, 6, 7, 13, 15, 17, 18, 19, 23, 24, 25

[52] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. *CVPR*, 2018. 6

[53] Lei Zhao, Qihang Mo, Sihuan Lin, Zhizhong Wang, Zhiwen Zuo, Haibo Chen, Wei Xing, and Dongming Lu. Uctgan: Diverse image inpainting based on unsupervised cross-space translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5741–5750, 2020. 2

以及使用文本引导扩散模型进行编辑。*arXiv preprint arXiv:2112.10741*，2021年。3 [30] Evangelos Ntavelis, Andrés Romero, Siavash Bigdeli, Radu Timofte, Zheng Hui, Xiumei Wang, Xinbo Gao, Chajin Shin, Taeoh Kim, Hanbin Son, 等。AIM 2020图像极限修复挑战赛。收录于 *European Conference on Computer Vision*，第716–741页。Springer，2020年。2 [31] Evangelos Ntavelis, Andrés Romero, Iason Kastanis, Luc Van Gool, 和 Radu Timofte。Sesame：通过添加、操纵或擦除对象进行场景语义编辑。收录于 *European Conference on Computer Vision*，第394–411页。Springer，2020年。2 [32] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, 和 Alexei A Efros。上下文编码器：通过修复进行特征学习。收录于 *Proceedings of the IEEE conference on computer vision and pattern recognition*，第2536–2544页，2016年。2 [33] Jialun Peng, Dong Liu, Songcen Xu, 和 Houqiang Li。通过分层VQ-VAE为图像修复生成多样化结构。收录于*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*，第10775–10784页，2021年。2, 5, 6, 7, 13, 15, 17, 18, 19, 20, 21, 22, 23, 24, 25 [34] Yurui Ren, Xiaoming Yu, Ruonan Zhang, Thomas H Li, Shan Liu, 和 Ge Li。StructureFlow：通过结构感知外观流进行图像修复。收录于*Proceedings of the IEEE/CVF International Conference on Computer Vision*，第181–190页，2019年。2 [35] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, 和 Daniel Cohen-Or。风格编码：用于图像到图像转换的StyleGAN编码器。收录于 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*，第2287–2296页，2021年。2 [36] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, 和 Li Fei-Fei。ImageNet大规模视觉识别挑战赛。 *International Journal of Computer Vision (IJCV)*，115(3):211–252，2015年。2, 6 [37] Chitwan Saharia, William Chan, Huiwen Chang, Chris A Lee, Jonathan Ho, Tim Salimans, David J Fleet, 和 Mohammad Norouzi。Palette：图像到图像扩散模型。 *arXiv preprint arXiv:2111.05826*，2021年。3 [38] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, 和 Surya Ganguli。使用非平衡热力学的深度无监督学习，2015年。2, 3 [39] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, 和 Ben Poole。通过随机微分方程进行基于分数的生成建模，2020年。3 [40] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, 和 Victor Lempitsky。具有傅里叶卷积的、分辨率鲁棒的大掩码修复。*arXiv preprint arXiv:2109.07161*，2021年。2, 5, 6, 7, 8, 12, 13, 14, 15, 17, 18, 19, 20, 21, 22, 23, 24, 25 [41] Dmitry Ulyanov, Andrea Vedaldi, Victor Lempitsky. 深度图像先验。收录于 *Proceedings of the IEEE conference on computer vision and pattern recognition*, 第9446–9454页, 2018年。2[42] Ziyu Wan, Jingbo Zhang, Dongdong Chen, Jing Liao. 基于Transformer的高保真多元化图像补全。 *arXiv preprint arXiv:2103.14031*, 2021年。2, 5, 6, 7, 13, 15, 17, 18, 19, 20, 21, 22, 23, 24, 25[43] Wei Xiong, Jiahui Yu, Zhe Lin, Jimei Yang, Xin Lu, Connelly Barnes, Jiebo Luo. 前景感知的图像修复。收录于 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 第5840–5848页, 2019年。2[44] Shunxin Xu, Dong Liu, Zhiwei Xiong. E2i：从边缘到图像的生成式修复。*IEEE Transactions on Circuits and Systems for Video Technology*, 31(4):1308–1322, 2020年。2[45] Fisher Yu, Vladlen Koltun. 通过空洞卷积进行多尺度上下文聚合。*arXiv preprint arXiv:1511.07122*, 2015年。2[46] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, Thomas S Huang. 基于上下文注意力的生成式图像修复。收录于 *Proceedings of the IEEE conference on computer vision and pattern recognition*, 第5505–5514页, 2018年。2[47] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, Thomas S Huang. 基于上下文注意力的生成式图像修复。 *arXiv preprint arXiv:1801.07892*, 2018年。5, 6, 7, 13, 15, 17, 18, 19, 23, 24, 25[48] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, Thomas S Huang. 基于门控卷积的自由形式图像修复。收录于 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 第4471–4480页, 2019年。2[49] Yingchen Yu, Fangneng Zhan, Rongliang Wu, Jianxiong Pan, Kaiwen Cui, Shijian Lu, Feiying Ma, Xuansong Xie, Chunyan Miao. 基于双向自回归Transformer的多样化图像修复。收录于 *Proceedings of the 29th ACM International Conference on Multimedia*, 2021年。2[50] Yanhong Zeng, Jianlong Fu, Hongyang Chao, Baining Guo. 学习金字塔上下文编码器网络以实现高质量图像修复。收录于 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 第1486–1494页, 2019年。2[51] Yanhong Zeng, Jianlong Fu, Hongyang Chao, Baining Guo. 聚合上下文变换以实现高分辨率图像修复。*arXiv preprint arXiv:2104.01431*, 2021年。2, 5, 6, 7, 13, 15, 17, 18, 19, 23, 24, 25[52] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, Oliver Wang. 深度特征作为感知度量的惊人有效性。*CVPR*, 2018年。6[53] Lei Zhao, Qihang Mo, Sihuan Lin, Zhizhong Wang, Zhiwen Zuo, Haibo Chen, Wei Xing, Dongming Lu. UCTGAN：基于无监督跨空间转换的多样化图像修复。收录于 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 第5741–5750页, 2020年。2

[54] Shengyu Zhao, Jonathan Cui, Yilun Sheng, Yue Dong, Xiao Liang, Eric I Chang, and Yan Xu. Large scale image completion via co-modulated generative adversarial networks. *arXiv preprint arXiv:2103.10428*, 2021. 2

[55] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. Pluralistic image completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1438–1447, 2019. 2

[56] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 6, 12

[54] 赵盛宇、崔乔纳森、盛一伦、董悦、梁潇、Eric I Chang、徐岩。基于共调制生成对抗网络的大规模图像补全。*arXiv preprint arXiv:2103.10428*，2021年。2[55] 郑传侠、詹达仁、蔡剑飞。多元图像补全。见 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*，第1438–1447页，2019年。2[56] 周博磊、Agata Lapedriza、Aditya Khosla、Aude Oliva、Antonio Torralba。Places：一个用于场景识别的千万级图像数据库。*IEEE Transactions on Pattern Analysis and Machine Intelligence*，2017年。6, 12

Appendix

In this appendix, we provide additional details and analysis of our approach. We give more explanation on our user study in Section A. Further, we present additional details on how we implemented the diffusion time schedule for jumps in Section B. Visual results for our ablation for jump size and the number of resamplings are provided in Section C. The evaluation on the second part of the LaMa Benchmark on Places2 is presented in Section D. Furthermore, to compare the diversity of the inpaintings for RePaint compared with state-of-the-art, we provide a quantitative analysis in Section E. Details on failure cases and data bias on the ImageNet dataset are provided in Section F. For gaining a better intuitive understanding of the evolution of the latent space, we provide a video of the inference in Section G. And finally, we show additional visual examples in Section I.

A. User Study

As described in Section 5.2 in the main paper, we conduct a user study to determine which method is best perceived to the human eye. In Figure 8, we depict the user interface, where the user selects the most realistic solution from an input reference. To reduce bias, we show the two candidate images in random order. Additionally, to improve the consistency of the user decision and prevent answers with low effort, we show every example twice. The users that agree in less than 75% of their own votes are discarded.

B. Algorithm for jump size larger than one

In addition to the resampling introduced in Algorithm 1 in the main paper, we use jumps in diffusion time as described in Section 4.2 in the main paper. Figure 9 shows a

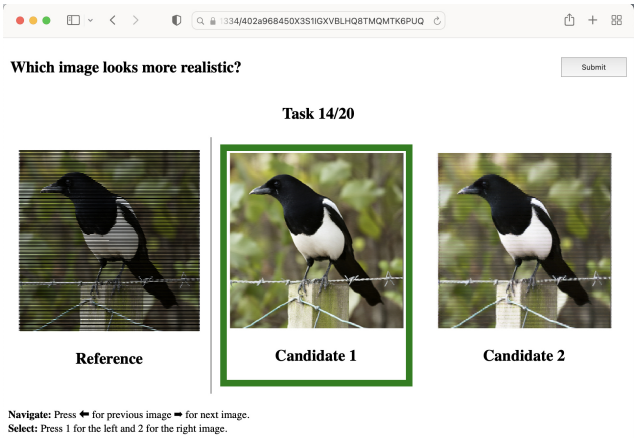


Figure 8. **User Study Interface.** Example of the user-study interface. Based on the reference image on the Left, the user selects the image that looks more realistic.

```
t_T = 250
jump_len = 10
jump_n_sample = 10

jumps = {}
for j in range(0, t_T - jump_len, jump_len):
    jumps[j] = jump_n_sample - 1

t = t_T
ts = []

while t >= 1:
    t = t-1
    ts.append(t)

    if jumps.get(t, 0) > 0:
        jumps[t] = jumps[t] - 1
        for _ in range(jump_len):
            t = t + 1
            ts.append(t)

ts.append(-1)
```

Figure 9. **Diffusion Time Schedule.** Pseudo code to generate diffusion time steps for jump length $j = 10$ and resample $r = 10$.

pseudo-code to further clarify the generation of state transitions. Note that each transition increases or decreases the diffusion time t by one. For example, for a chosen jump length of $j = 10$ shown in Figure 11, we apply ten forward transitions before applying ten reverse transitions. The diffusion time t for the latent vector x_t is plotted in Figure 10.

C. Ablation

In addition to the quantitative analysis in Table 3 in the main paper, this section shows visual examples for different jump lengths j and number or resamplings r. As discussed in Section 5.5 in the main paper, smaller jump lengths j tend to produce blurrier images as shown in Figure 12, and an increased number or resamplings r improves the overall image consistency.

D. Evaluation on Places2

For a more comprehensive experimental framework, in this section, we provide the second part of the benchmark proposed in LaMa [40], which is over the Places2 [56] dataset. The experiments on Places2 were conducted using an unconditional model that we trained for 300k iterations with batch size four on four V100, taking about six days in total. All other training settings were kept as originally [7] used for ImageNet. The model checkpoint will be published. We will clarify these aspects and add further details in the paper. We use the same mask generation procedure and settings described in the main paper. The results shown in Table 5 are in line with those on CelebA and ImageNet

附录

在本附录中，我们将提供关于我们方法的额外细节与分析。关于用户研究，我们将在A节中给出更多解释。此外，我们将在B节中详细介绍如何为跳跃步骤实现扩散时间调度。关于跳跃步长和重采样次数的消融实验视觉结果将在C节中展示。在D节中，我们将展示在Places2数据集上对LaMa基准测试第二部分的评估结果。此外，为了比较RePaint与现有先进方法在修复结果多样性方面的表现，我们将在E节中提供定量分析。关于ImageNet数据集上的失败案例和数据偏差的细节将在F节中说明。为了更直观地理解潜在空间的演化过程，我们将在G节中提供推理过程的视频。最后，我们将在I节中展示更多的视觉示例。

A. 用户研究

如主论文第5.2节所述，我们进行了一项用户研究，以确定哪种方法在人眼感知中效果最佳。图8展示了用户界面，用户需从输入参考中选择最真实的解决方案。为减少偏差，我们以随机顺序呈现两张候选图像。此外，为提高用户决策的一致性并避免低质量回答，每个示例会展示两次。若用户自身投票的一致性低于75%，其数据将被剔除。

B. 跳跃步长大于一的算法

除了主论文中算法1引入的重采样方法外，我们还使用了主论文第4.2节所述的扩散时间跳跃策略。图9展示了

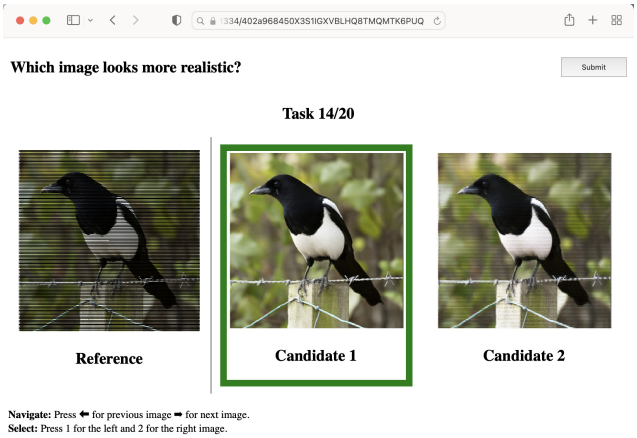


图8. 用户研究界面。用户研究界面的示例。基于左侧的参考图像，用户选择看起来更真实的图像。

```
t_T = 250
jump_len = 10
jump_n_sample = 10

jumps = {}
for j in range(0, t_T - jump_len, jump_len):
    jumps[j] = jump_n_sample - 1

t = t_T
ts = []

while t >= 1:
    t = t-1
    ts.append(t)

    if jumps.get(t, 0) > 0:
        jumps[t] = jumps[t] - 1
        for _ in range(jump_len):
            t = t + 1
            ts.append(t)

ts.append(-1)
```

图9. 扩散时间调度表。为跳跃长度$j = 10$和重采样$r = 10$生成扩散时间步长的伪代码。

伪代码进一步阐明了状态转移的生成过程。注意每次转移会使扩散时间t增加或减少一单位。例如，针对图11中选取的跳跃长度j =为10的情况，我们先进行十次前向转移，再进行十次反向转移。潜在向量x_t对应的扩散时间t变化曲线绘制于图10中。

C. 消融实验

除了主论文中表3的定量分析外，本节展示了不同跳跃长度j和重采样次数r的视觉示例。如主论文第5.5节所述，较小的跳跃长度j往往会产生更模糊的图像（如图12所示），而增加重采样次数r则能提升图像的整体一致性。

D. 在Places2数据集上的评估

为了构建更全面的实验框架，本节我们提供了LaMa [40]中提出的基准测试的第二部分，该部分基于Places2 [56]数据集。在Places2上进行的实验使用了我们在四块V100显卡上以批次大小四训练30万轮的无条件模型，总计耗时约六天。所有其他训练设置均保持与原始ImageNet实验[7]一致。模型检查点将公开发布。我们将在论文中阐明这些细节并补充更多信息。我们采用了与主论文相同的掩码生成流程和设置。表5所示结果与CelebA和ImageNet上的结果保持一致。

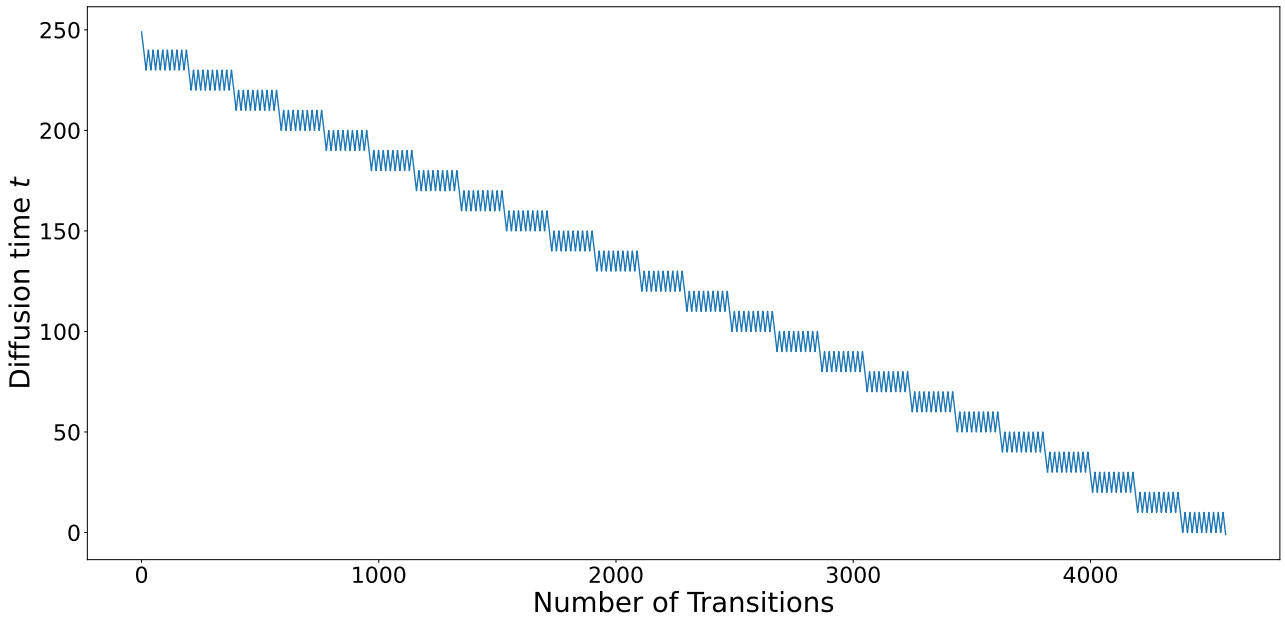


Figure 10. **Diffusion time during inference.** The diffusion time t that a sample x_t is transiting during the inference process with jump length $j = 10$ and resampling $r = 10$.

Datasets	Wide		Narrow		Super-Resolve 2×		Altern. Lines		Half		Expand	
Methods	LPIPS	Votes [%]	LPIPS	Votes [%]	LPIPS	Votes [%]	LPIPS	Votes [%]	LPIPS	Votes [%]	LPIPS	Votes [%]
AOT [51]	0.112	35.4 ± 3.0	0.062	36.0 ± 3.0	0.560	2.2 ± 0.9	0.399	0.8 ± 0.6	0.263	34.0 ± 2.9	0.686	0.7 ± 0.5
DSI [33]	0.101	27.4 ± 2.8	0.054	33.1 ± 2.9	0.157	8.4 ± 1.7	0.083	6.9 ± 1.6	0.265	33.7 ± 2.9	0.565	13.8 ± 2.1
ICT [42]	0.101	35.7 ± 3.0	0.057	33.7 ± 2.9	0.776	0.9 ± 0.6	0.672	1.3 ± 0.7	0.256	26.0 ± 2.7	0.554	26.6 ± 2.7
Deep Fill v2 [47]	0.097	29.7 ± 2.8	0.051	33.0 ± 2.9	0.120	15.8 ± 2.3	0.070	15.4 ± 2.2	0.254	32.8 ± 2.9	0.550	12.9 ± 2.1
LaMa [40]	0.078	47.7 ± 3.1	0.039	43.3 ± 3.1	0.369	7.5 ± 1.6	0.138	21.5 ± 2.6	0.233	34.0 ± 2.9	0.512	39.4 ± 3.0
RePaint	0.105	*Reference*	0.044	*Reference*	0.099	*Reference*	0.051	*Reference*	0.286	*Reference*	0.615	*Reference*

Table 5. **Places2 Quantitative Results.** We compute the LPIPS (lower is better) and *votes* for five different mask settings. *Votes* refers to the ratio of votes in favor our RePaint.

in Table 1 of the main paper. RePaint outperforms all other methods for all masks with significance 95% except for one inconclusive case. This case is when comparing RePaint to LaMa on Wide Masks, where the users vote in 52.4%

```
times = get_schedule()
x = random_noise()

for t_last, t_cur in zip(times[:-1], times[1:]):
    if t_cur < t_last:
        # Apply Equation 8 (Main Paper)
        x = reverse_diffusion(x, t, x_known)
    else:
        # Apply Equation 1 (Main Paper)
        x = forward_diffusion(x, t)
```

Figure 11. **Inference Process.** Pseudo code of RePaint inference process using a precalculated time schedule.

Mask	Wide		Narrow		SR 2x		Alter. Lines		Half		Expand	
Measure	LPIPS	DS	LPIPS	DS	LPIPS	DS	LPIPS	DS	LPIPS	DS	LPIPS	DS
DSI [33]	0.0639	16.68	0.0454	18.74	0.1404	12.38	0.0591	4.78	0.2348	15.30	0.5458	14.33
ICT [42]	0.0596	15.77	0.0402	18.65	0.5427	8.70	0.3916	8.16	0.1817	16.40	0.4779	17.25
RePaint	0.0552	16.40	0.0337	23.79	0.0327	19.84	0.0106	23.00	0.1839	17.31	0.4832	17.11

Table 6. **Diversity Score.** The Diversity Score (DS) and LPIPS calculated on CelebA-HQ on various masks for 32 images.

for RePaint, but the significance interval overlaps with the 50% border. The visual comparison on the and Wide and Narrow mask is shown in Figure 22. Moreover, the visual results further confirm the robustness against sparse masks as shown in Figure 23. The mask pattern is clearly visible in all competing methods, while RePaint shows better harmonization. Regarding large masks, RePaint is able to inpaint semantically meaningful content such as the companion in the Bar in the same age, and overall lightning conditions as shown in the second row of Figure 24.

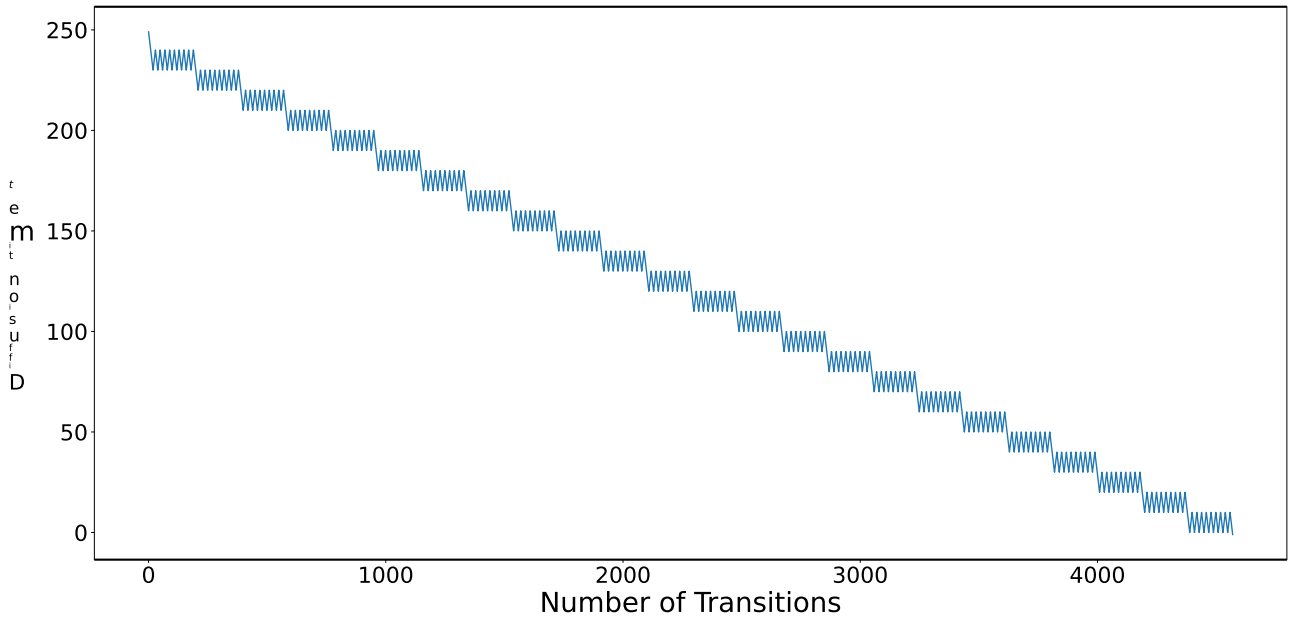


图10. 推理过程中的扩散时间。在跳跃长度j =为10、重采样r =为10的推理过程中，样本x_t所经历的扩散时间为t。

Datasets	Wide		Narrow		Super-Resolve 2×		Altern. Lines		Half		Expand	
Methods	LPIPS	Votes [%]	LPIPS	Votes [%]	LPIPS	Votes [%]	LPIPS	Votes [%]	LPIPS	Votes [%]	LPIPS	Votes [%]
AOT [51]	0.112	35.4 ± 3.0	0.062	36.0 ± 3.0	0.560	2.2 ± 0.9	0.399	0.8 ± 0.6	0.263	34.0 ± 2.9	0.686	0.7 ± 0.5
DSI [33]	0.101	27.4 ± 2.8	0.054	33.1 ± 2.9	0.157	8.4 ± 1.7	0.083	6.9 ± 1.6	0.265	33.7 ± 2.9	0.565	13.8 ± 2.1
ICT [42]	0.101	35.7 ± 3.0	0.057	33.7 ± 2.9	0.776	0.9 ± 0.6	0.672	1.3 ± 0.7	0.256	26.0 ± 2.7	0.554	26.6 ± 2.7
Deep Fill v2 [47]	0.097	29.7 ± 2.8	0.051	33.0 ± 2.9	0.120	15.8 ± 2.3	0.070	15.4 ± 2.2	0.254	32.8 ± 2.9	0.550	12.9 ± 2.1
LaMa [40]	0.078	47.7 ± 3.1	0.039	43.3 ± 3.1	0.369	7.5 ± 1.6	0.138	21.5 ± 2.6	0.233	34.0 ± 2.9	0.512	39.4 ± 3.0
RePaint	0.105	*Reference*	0.044	*Reference*	0.099	*Reference*	0.051	*Reference*	0.286	*Reference*	0.615	*Reference*

表5. Places2定量结果。我们针对五种不同掩码设置计算了LPIPS（数值越低越好）和*votes*。*Votes*指代支持我们]RePaint方法的投票比例。

在主论文的表1中。RePaint在除一个不确定情况外的所有掩码上均以95%的显著性优于其他所有方法。该不确定情况是在宽掩码上比较RePaint与LaMa时，用户投票率为52.4%

```
times = get_schedule()
x = random_noise()

for t_last, t_cur in zip(times[:-1], times[1:]):
    if t_cur < t_last:
        # Apply Equation 8 (Main Paper)
        x = reverse_diffusion(x, t, x_known)
    else:
        # Apply Equation 1 (Main Paper)
        x = forward_diffusion(x, t)
```

图11. 推理过程。使用预计算时间调度的RePaint推理过程的伪代码。

Mask	Wide		Narrow		SR 2x		Alter. Lines		Half		Expand	
Measure	LPIPS	DS	LPIPS	DS	LPIPS	DS	LPIPS	DS	LPIPS	DS	LPIPS	DS
DSI [33]	0.0639	16.68	0.0454	18.74	0.1404	12.38	0.0591	4.78	0.2348	15.30	0.5458	14.33
ICT [42]	0.0596	15.77	0.0402	18.65	0.5427	8.70	0.3916	8.16	0.1817	16.40	0.4779	17.25
RePaint	0.0552	16.40	0.0337	23.79	0.0327	19.84	0.0106	23.00	0.1839	17.31	0.4832	17.11

表6. 多样性得分。在CelebA-HQ数据集上，针对32张图像在不同掩码下计算的多样性得分（DS）和LPIPS。

对于RePaint，但显著性区间与50%边界重叠。在宽窄掩码上的视觉比较如图22所示。此外，视觉结果进一步证实了对稀疏掩码的鲁棒性，如图23所示。所有竞争方法中掩码图案均清晰可见，而RePaint展现出更好的协调性。对于大面积掩码，RePaint能够修复具有语义意义的内容，例如同龄酒吧中的同伴，以及整体光照条件，如图24第二行所示。

Figure 12. **Ablation Study.** Analysis of length of the jumps j and number of resamplings r on ImageNet validation set with LaMa [40] Benchmark mask setting Wide.

图12. 消融研究。在ImageNet验证集上使用LaMa [40]基准掩码设置Wide，分析跳跃长度j和重采样次数r的影响。

Figure 13. **Failure Cases on ImageNet.** When applying RePaint trained on ImageNet for inpainting it is more likely to inpaint dogs, due to the data bias. Zoom-in for better details.

E. Diversity

For our quantitative evaluation in the main paper, we sample a single image per input. However, since our method is stochastic, we can sample from it. To compare the diversity among the stochastic methods, we use the Diversity Score as described in [22] (higher is better). In contrast to the standard diversity metric [33,42] that only computes the mean LPIPS across pair of outputs, this score is designed to describe meaningful diversity yet also weighting the overall performance in LPIPS. It aims at measuring the diversity of the generations inside the manifold of plausible predictions. In detail, too extreme predictions or failures are therefore penalized. As shown in Table 6, for "Wide" and "Half", there is no method with both best LPIPS and Diversity Score and for "Expand" ICT beats RePaint by 0.81% in Diversity Score and 1.1% in LPIPS. RePaint is superior by a large margin in both LPIPS and Diversity Score for the thin structured masks "Narrow", "Super-Resolution $2\times$", and "Alternating Lines" to both ICT [42] and DSI [33].

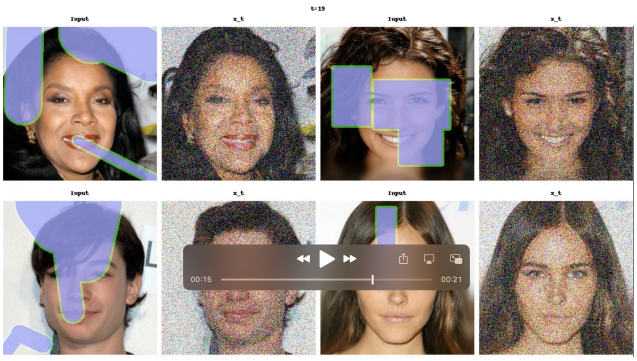


Figure 14. **Video of Diffusion Process.** In the attachment we show the video of the denoising diffusion process on the CelebA-HQ validation set.

Figure 15. Visual results on ImageNet 512×512 for thin mask.

F. Failure Cases

As depicted in Figure 13, RePaint sometimes confuses the semantic context and mixes non-matching objects. Our model on ImageNet seems to be biased towards inpainting dogs more frequently than expected. Since ImageNet has many different breeds of dogs for classification tasks, dogs are over-represented in the training set, hence our model bias.

G. Attached Video

To inspect the latent space of the diffusion space, we provide a video in the attachment as shown in the screenshot in Figure 14. There we show the Ground Truth and the latent space x_t after every transition in the diffusion process. Note that the diffusion time t, shown on top, jumps up and down according to the following schedule: The jump length is $j = 5$, and the number of resamplings is $r = 9$. To focus more on the visually interesting part of the diffusion process we set the number of diffusion steps to $T = 100$ and start resampling below $t = 50$.

H. Experiment on larger resolution

As shown in Figure 15, our inpainting method also works on pretrained model from [7] for 512×512. However, we were not able to conduct our full analysis on that resolution due to limited computational resources.

图13. ImageNet上的失败案例。由于数据偏差，在ImageNet上训练的RePaint进行修复时更容易生成狗的图案。放大以查看更佳细节。

E. 多样性

在主论文的定量评估中，我们对每个输入采样单张图像。然而，由于我们的方法是随机性的，我们可以从中进行多次采样。为了比较随机方法之间的多样性，我们采用[22]中所述的多样性评分（分值越高越好）。与仅计算输出对间平均LPIPS的标准多样性指标[33,42]不同，该评分旨在描述有意义的多样性，同时兼顾LPIPS整体性能。其目标是衡量生成结果在合理预测流形内的多样性。具体而言，过于极端的预测或失败案例会受到惩罚。如表6所示，在“拓宽”和“半幅”任务中，没有方法能同时取得最佳LPIPS和多样性评分；而在“扩展”任务中，ICT在多样性评分上以0.81%、在LPIPS上以1.1%的优势超越RePaint。对于细结构掩码“窄幅”“超分辨率2×”和“交替线条”，RePaint在LPIPS和多样性评分上均大幅领先ICT[42]和DSI[33]。

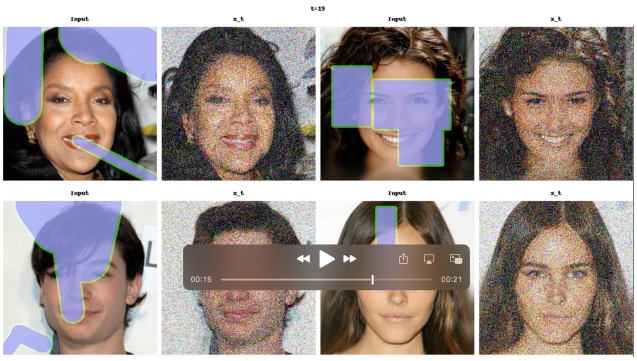


图14. 扩散过程视频。附件中展示了CelebA-HQ验证集上去噪扩散过程的视频。

图15. 在ImageNet 512 × 512上针对细长掩模的视觉结果。

F. 失败案例

如图13所示，RePaint有时会混淆语义上下文并混合不匹配的对象。我们的模型在ImageNet上似乎更频繁地倾向于修复狗的图像，这超出了预期。由于ImageNet包含许多不同品种的狗用于分类任务，狗在训练集中占比过高，因此导致了我们模型的偏差。

G. 附带的视频

为了观察扩散空间的潜在空间，我们在附件中提供了一段视频，如图14的截图所示。其中展示了真实数据（Ground Truth）以及扩散过程中每次转换后的潜在空间x_t。请注意，顶部显示的扩散时间t会根据以下计划上下跳跃：跳跃长度为$j = 5$，重采样次数为$r = 9$。为了更聚焦于扩散过程中视觉上最有趣的部分，我们将扩散步数设置为$T = 100$，并在低于$t = 50$时开始重采样。

H. 更大分辨率实验

如图15所示，我们的修复方法同样适用于[7]中预训练的512×512模型。但由于计算资源有限，我们未能对该分辨率进行完整的分析。

I. Additional Visual Results

We also provide additional visual examples for CelebA-HQ and ImageNet, comparing our approach to the same state-of-the-art methods as in the main paper. We show the results for Wide and Narrow masks in Figures 16 and 19, respectively, for the sparse masks "Super-Resolve 2×" and "Alternating Lines" in Figures 17 and 20 and for "Half" and "Expand" in Figures 18 and 21.

I. 额外视觉结果

我们还为CelebA-HQ和ImageNet提供了额外的视觉示例，将我们的方法与主论文中相同的先进方法进行比较。针对宽掩膜和窄掩膜，我们分别在图表16和19中展示了结果；针对稀疏掩膜“超分辨率2×”和“交替线条”，结果见图表17和20；而针对“半幅”和“扩展”掩膜，结果则展示在图表18和21中。

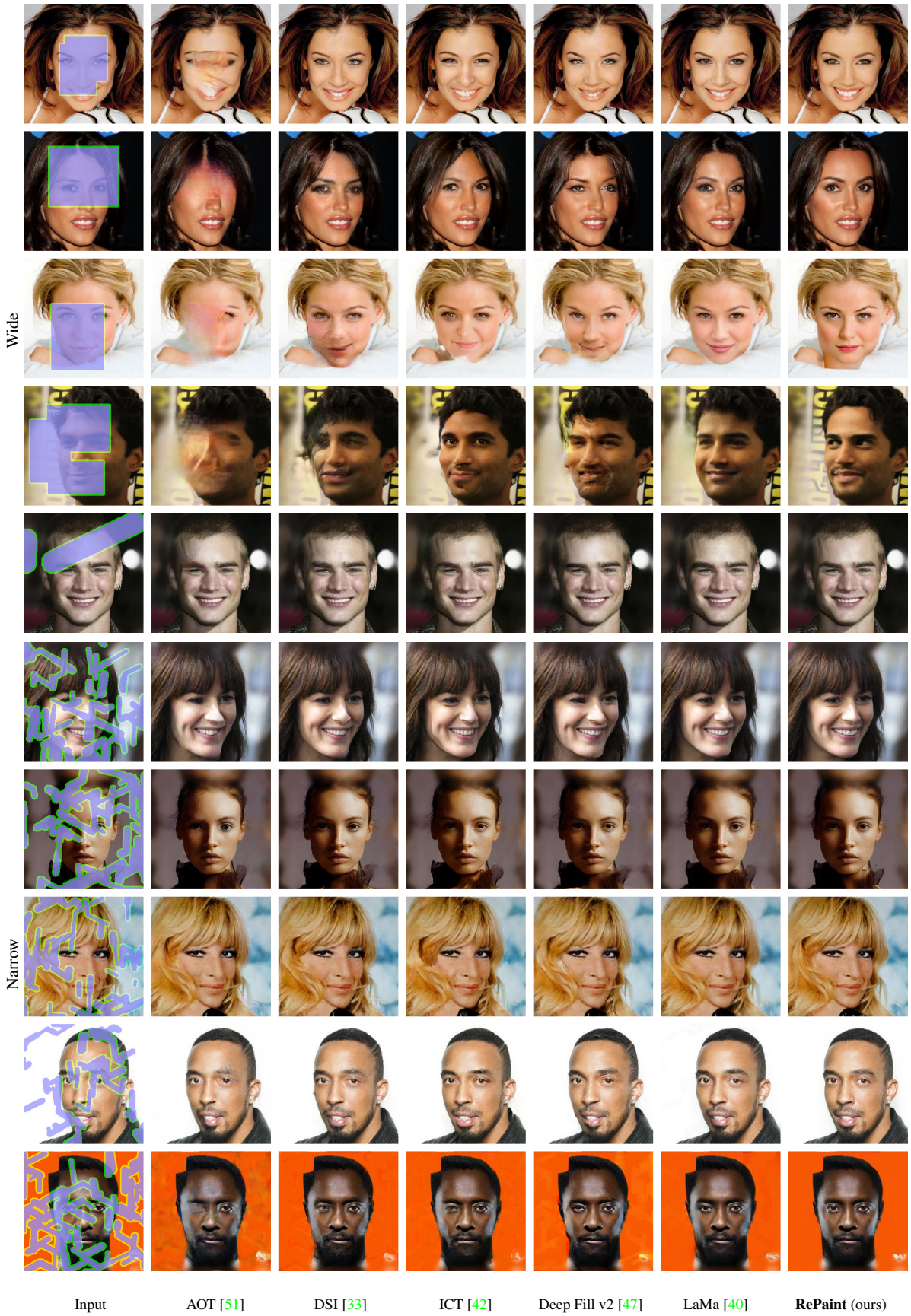
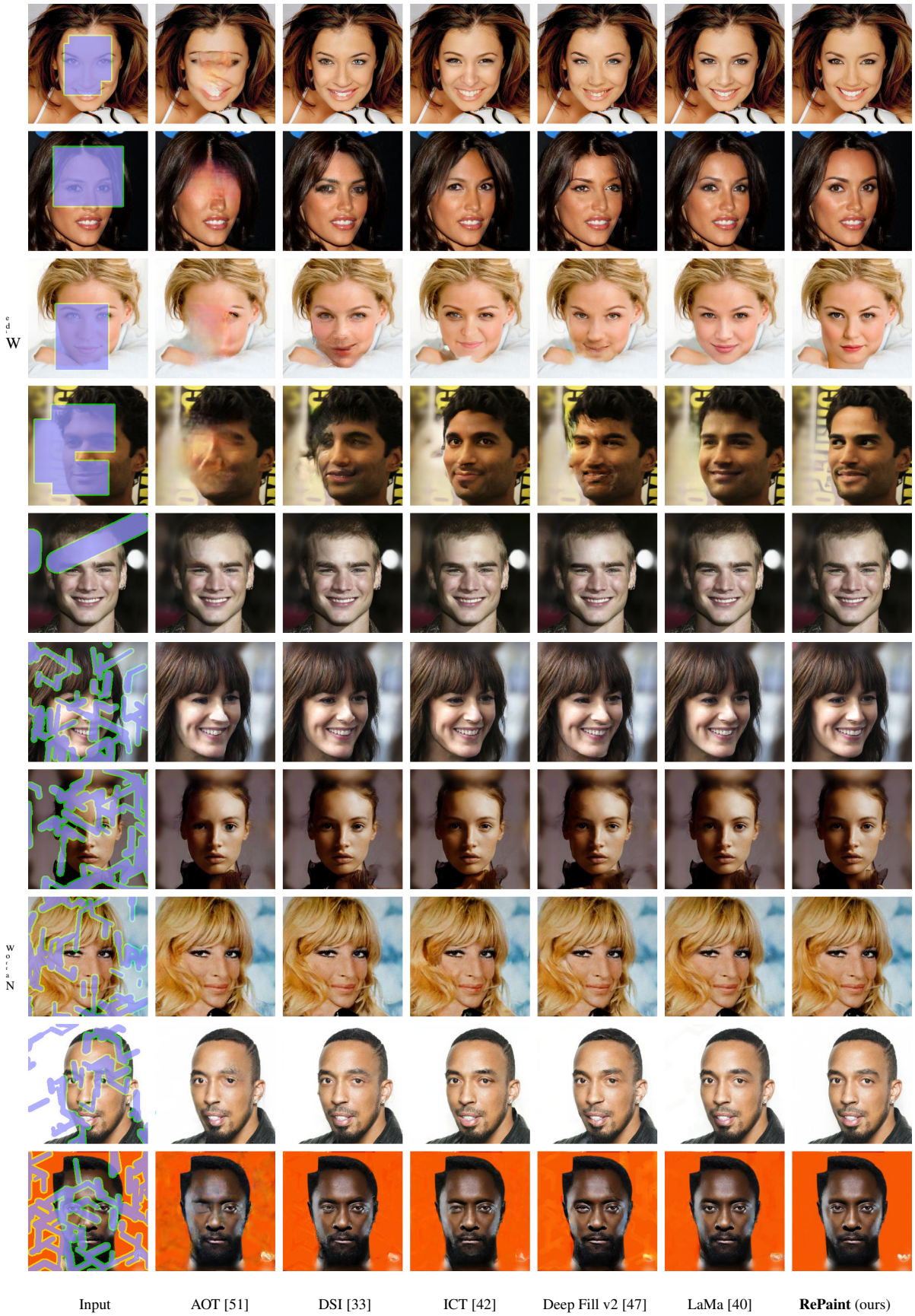


Figure 16. **CelebA-HQ Qualitative Results.** Comparison against the state-of-the-art methods for face inpainting. Zoom for better details.



F图 16. CelebA-HQ 定性结果。与最先进的人脸修复方法进行对比。放大查看更佳细节。

Figure 17. **CelebA-HQ Qualitative Results.** Comparison against the state-of-the-art methods for face inpainting. Zoom for better details.

图17. CelebA-HQ定性结果。与最先进的人脸修复方法对比。放大查看更佳细节。

Figure 18. **CelebA-HQ Qualitative Results.** Comparison against the state-of-the-art methods for face inpainting. Zoom for better details.

图18. CelebA-HQ定性结果。与最先进的人脸修复方法对比。放大查看更佳细节。

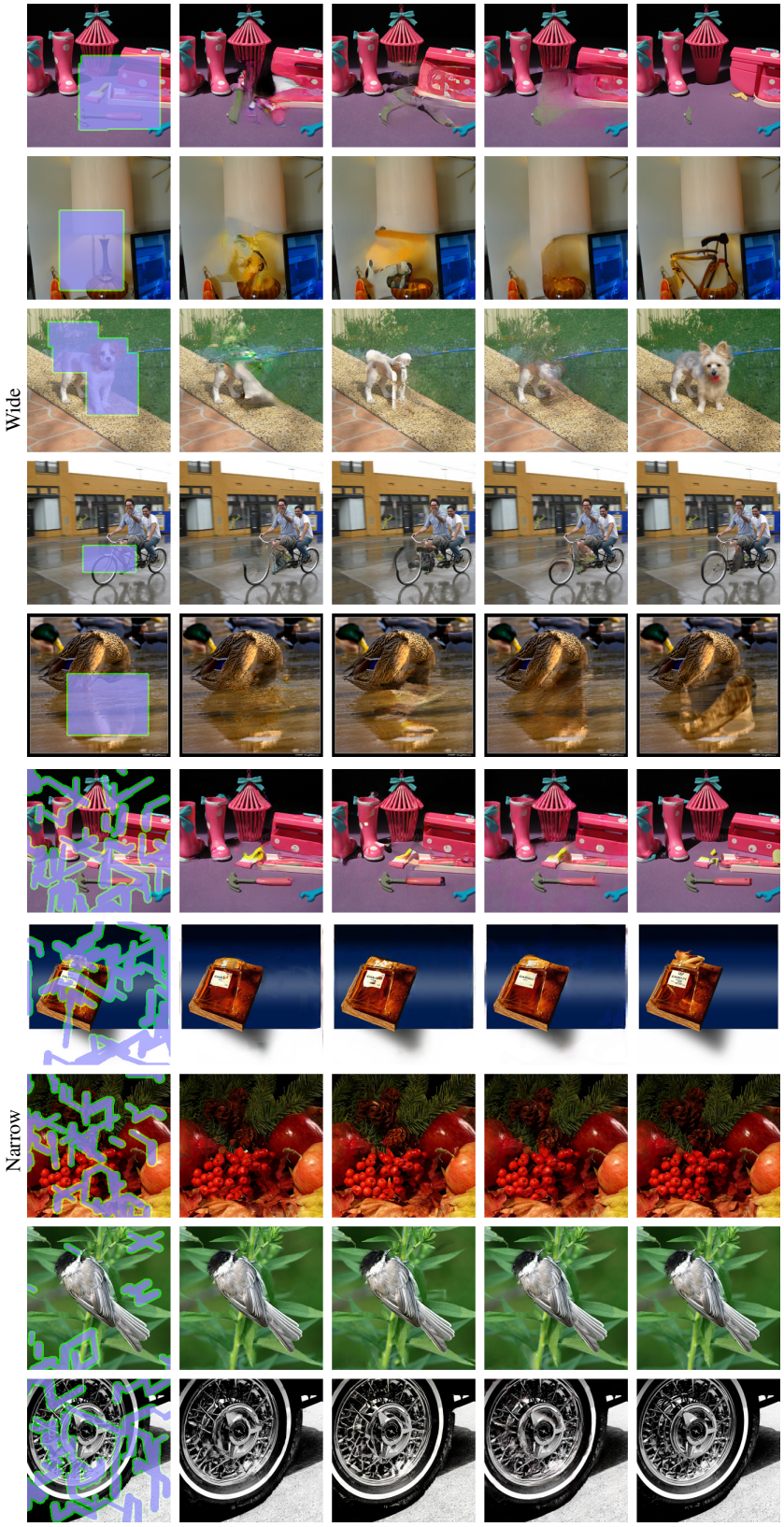
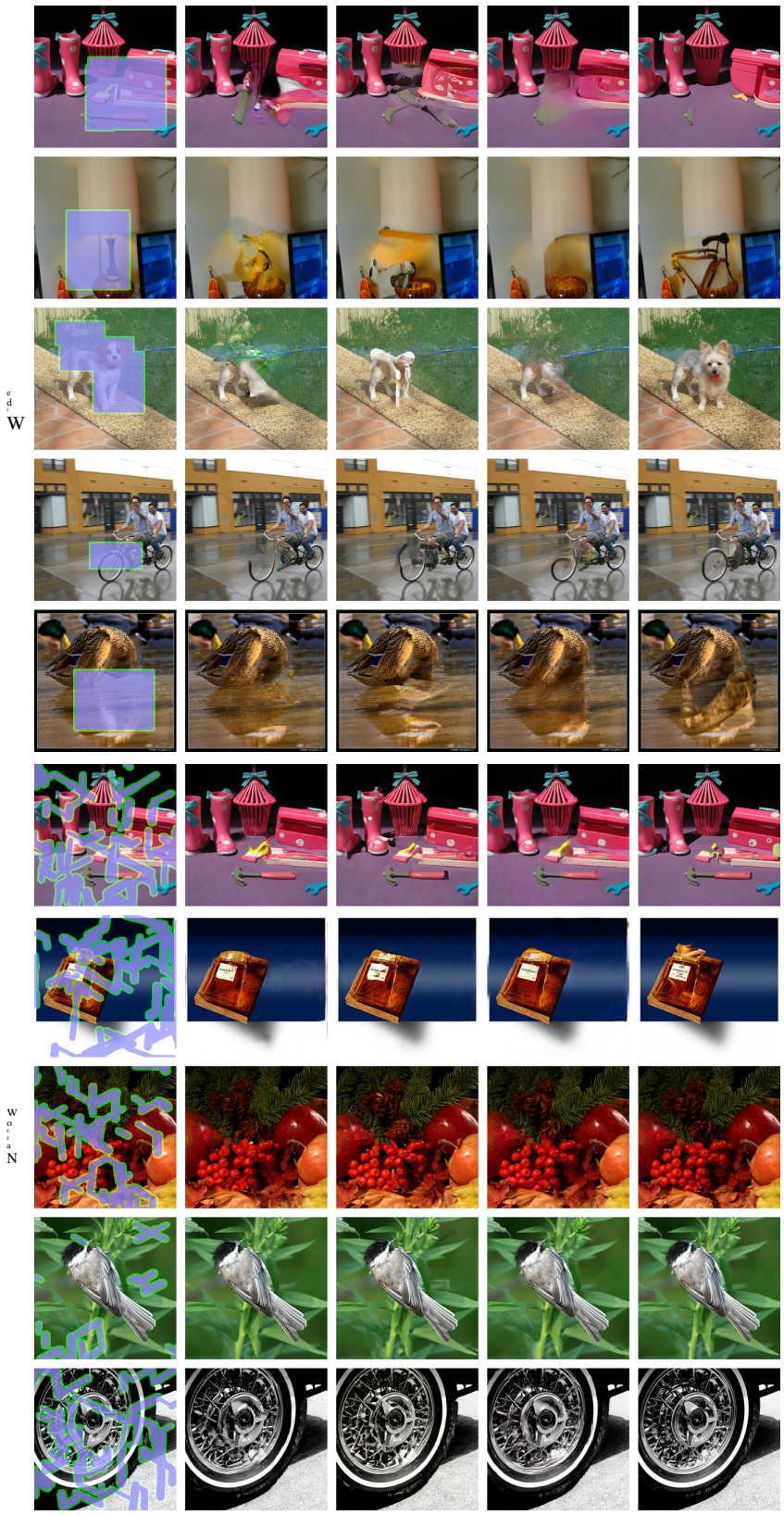


Figure 19. **ImageNet Qualitative Results.** Comparison against the state-of-the-art methods for diverse inpainting. Zoom for better details.



F图19. ImageNet定性结果。与最先进的多样化修复方法对比。放大以查看更佳细节。

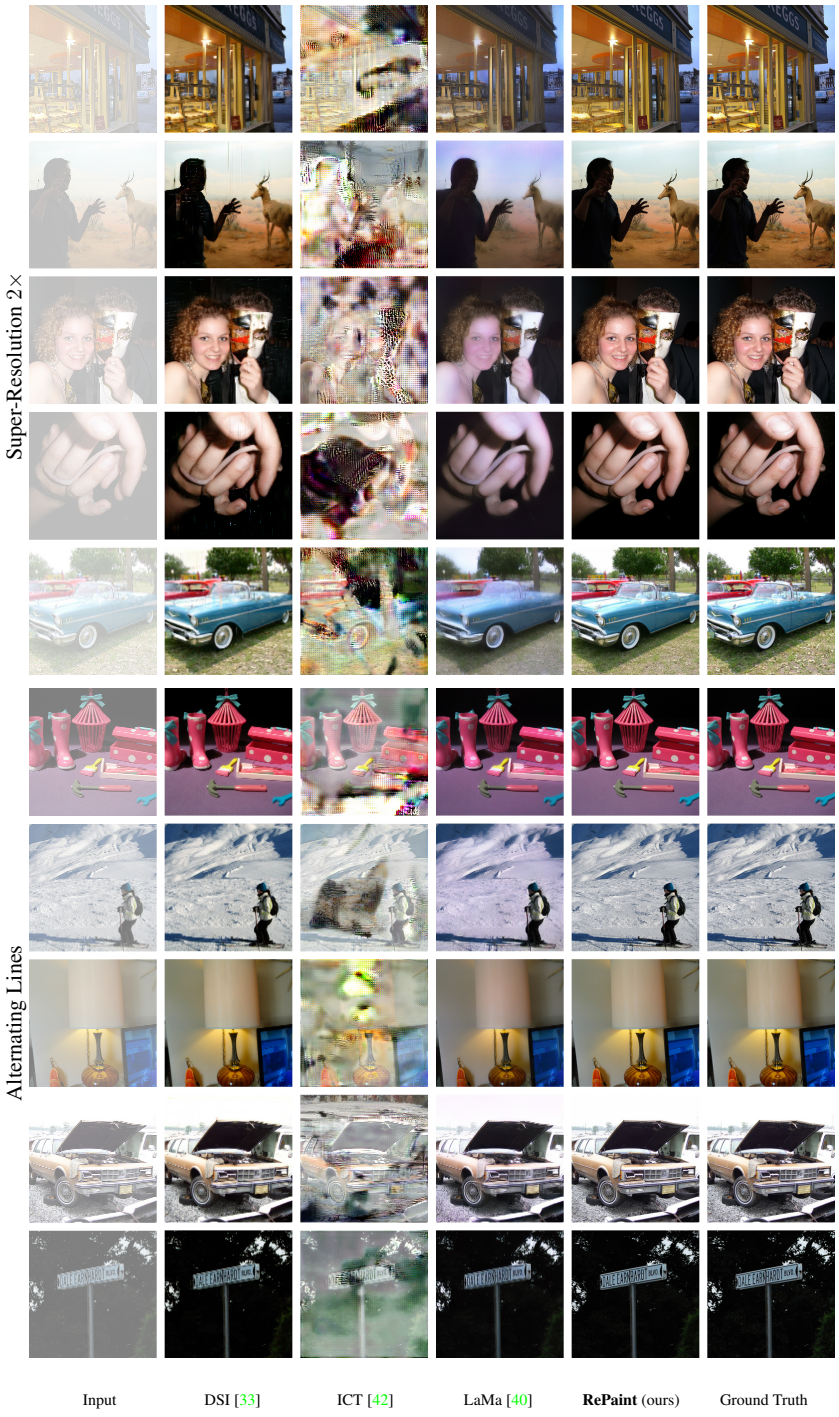


Figure 20. **ImageNet Qualitative Results.** Comparison against the state-of-the-art methods for diverse inpainting. Zoom for better details.

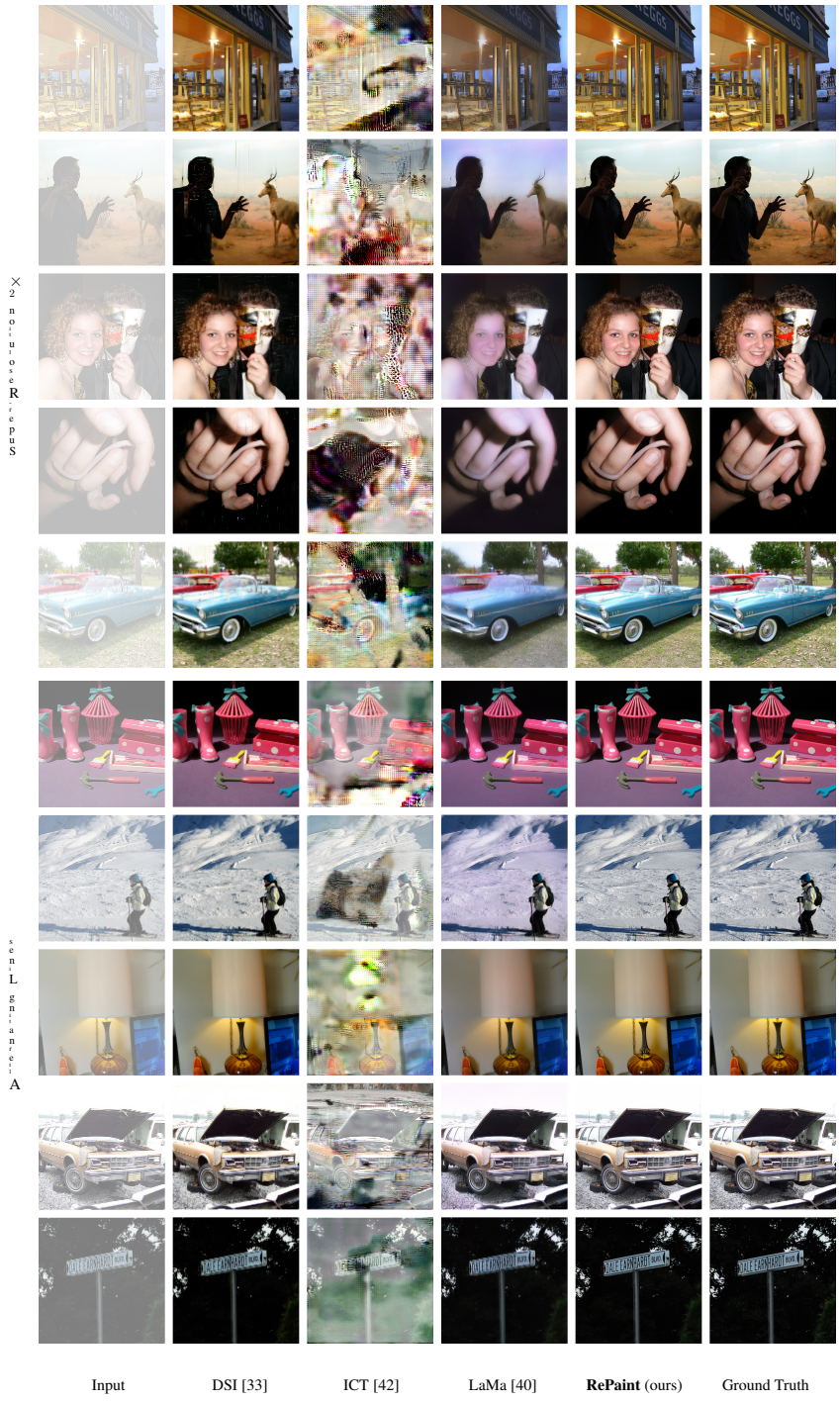


图20. ImageNet定性结果。与最先进方法在多样化修复方面的对比。放大以查看更佳细节。

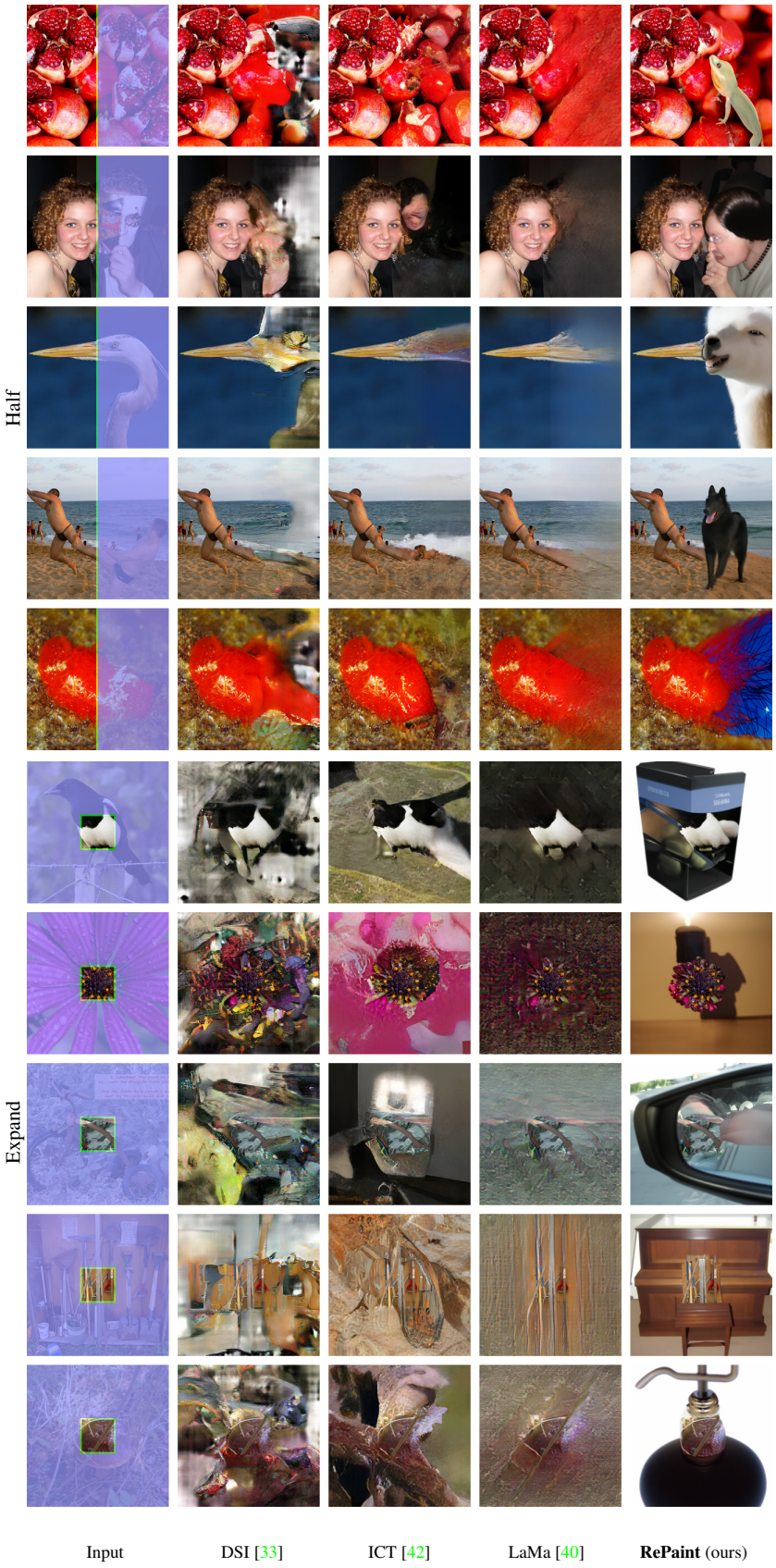


Figure 21. **ImageNet Qualitative Results.** Comparison against the state-of-the-art methods for diverse inpainting. Zoom for better details.

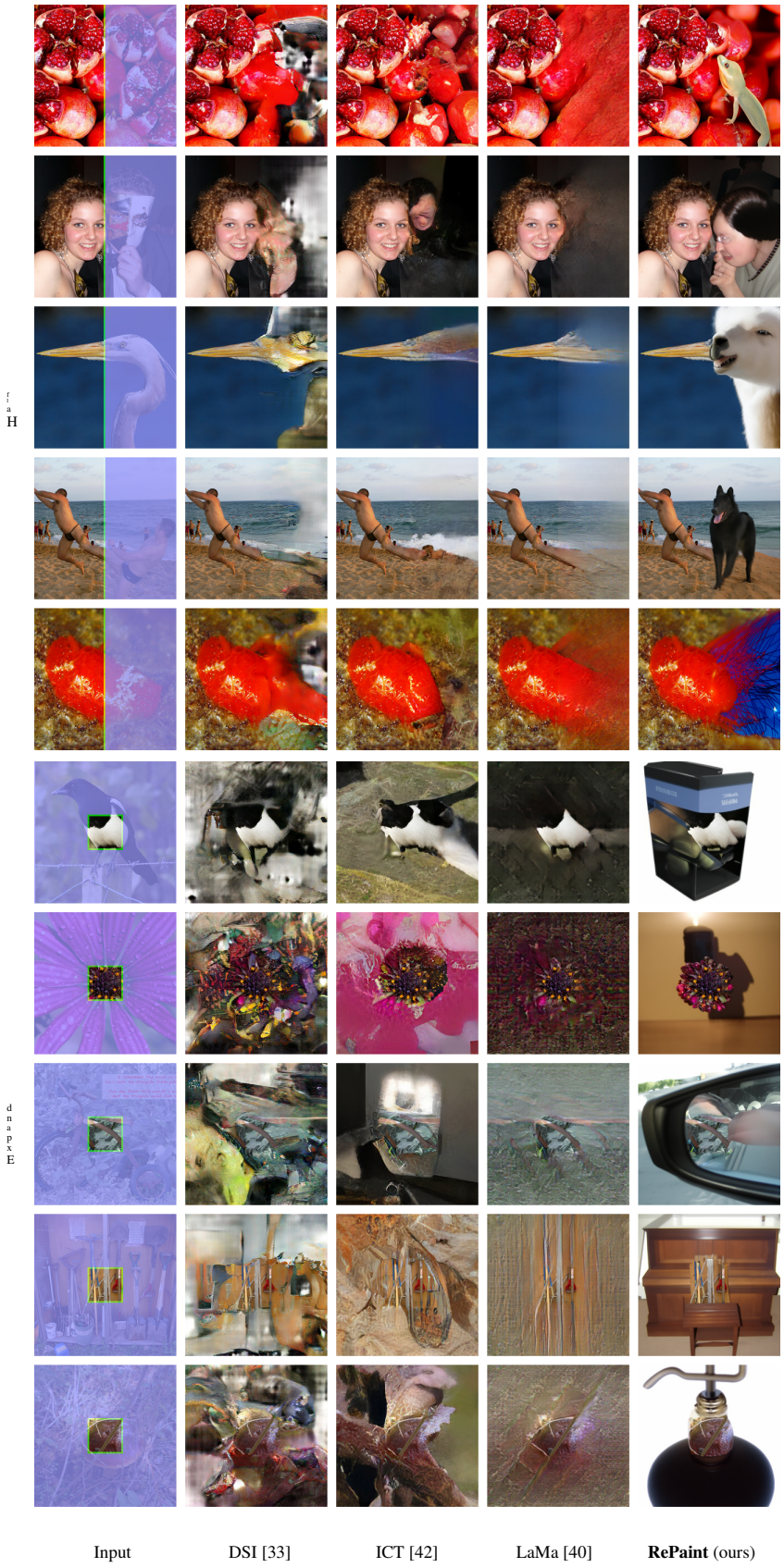


图21. ImageNet定性结果。与最先进的多样化修复方法对比。放大以查看更佳细节。

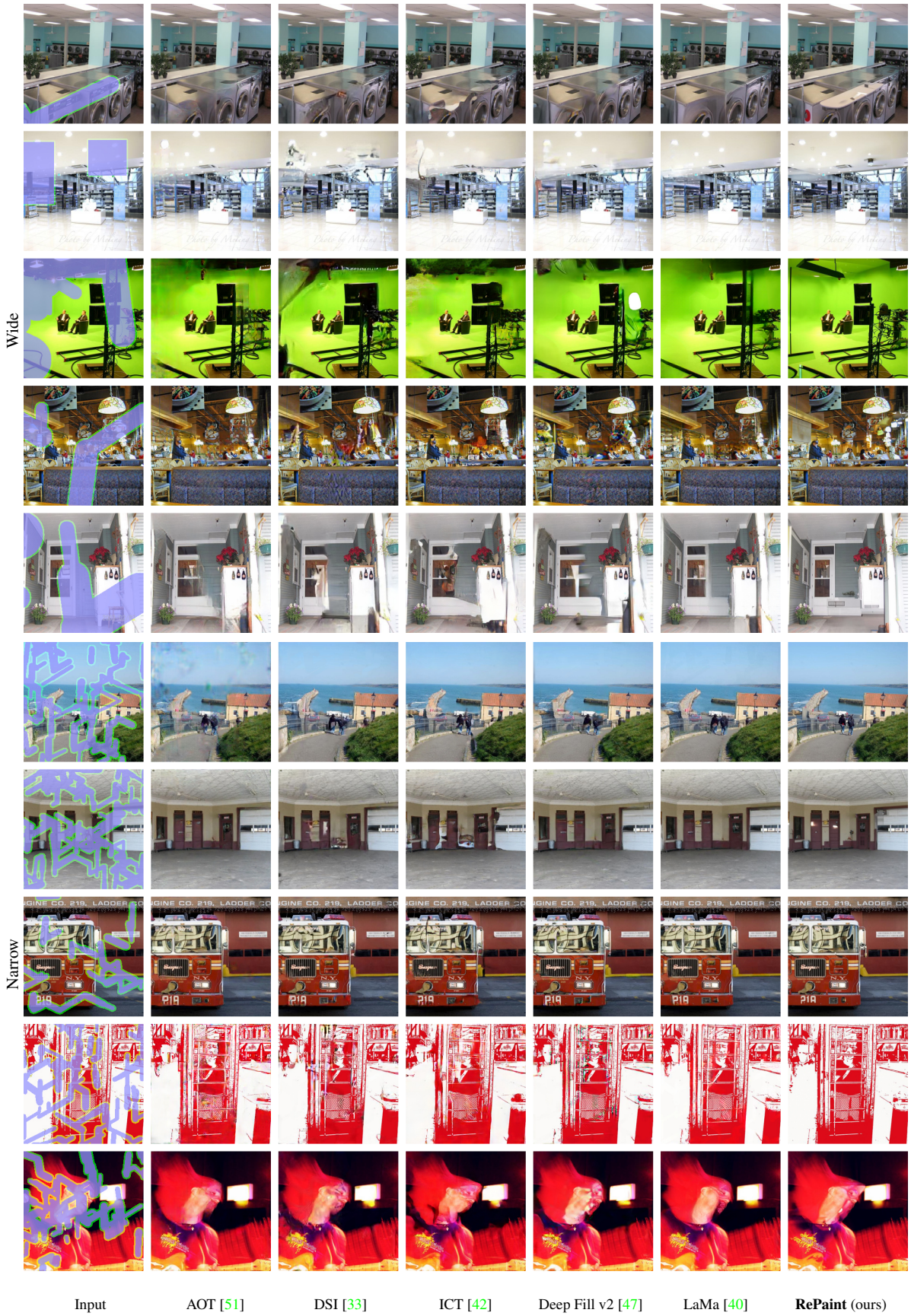


Figure 22. **Places2 Qualitative Results.** Comparison against the state-of-the-art methods for diverse inpainting. Zoom for better details.

图22. Places2定性结果。与最先进方法在多样化修复方面的对比。放大以查看更佳细节。

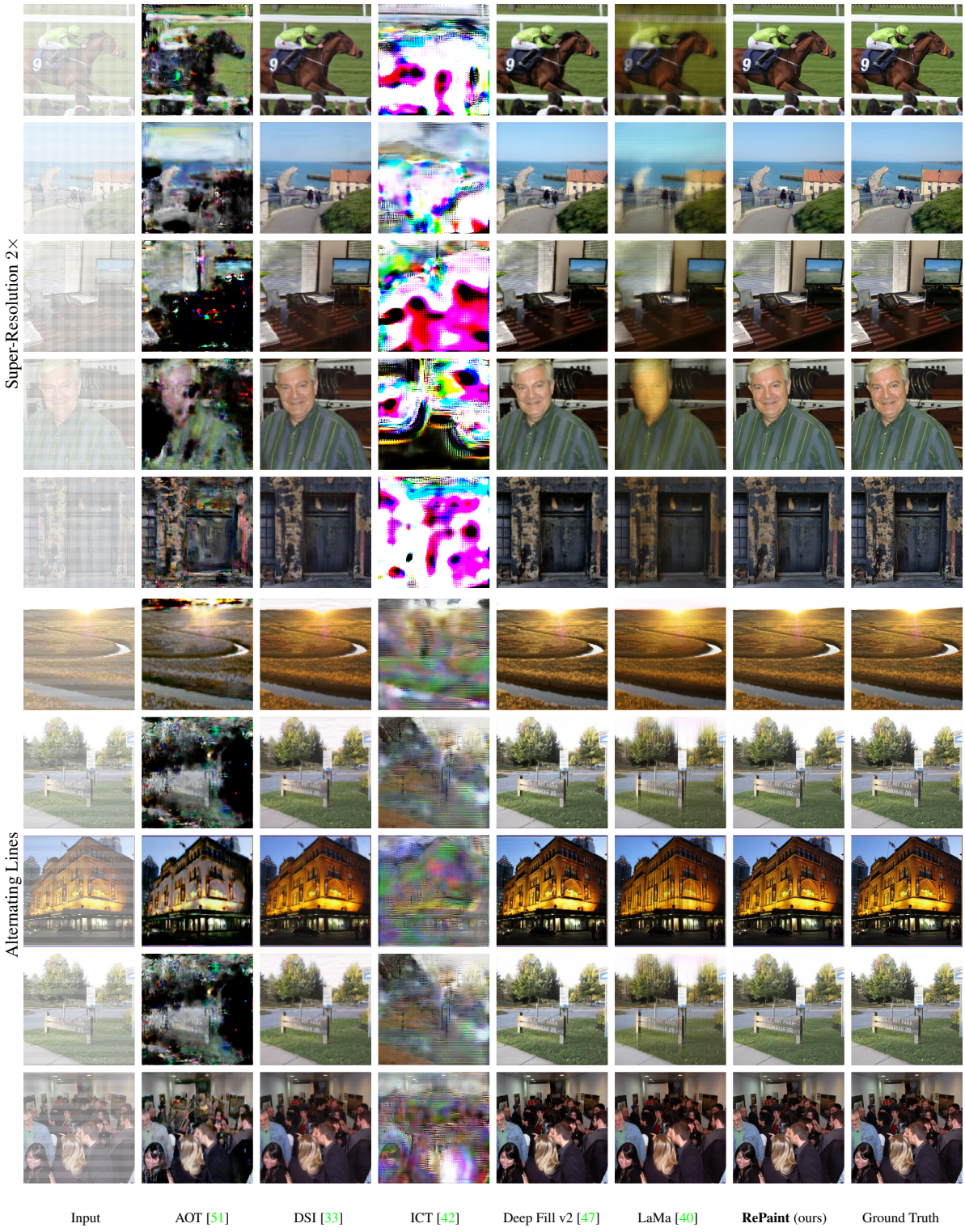


Figure 23. **Places2 Qualitative Results.** Comparison against the state-of-the-art methods for diverse inpainting. Zoom for better details.

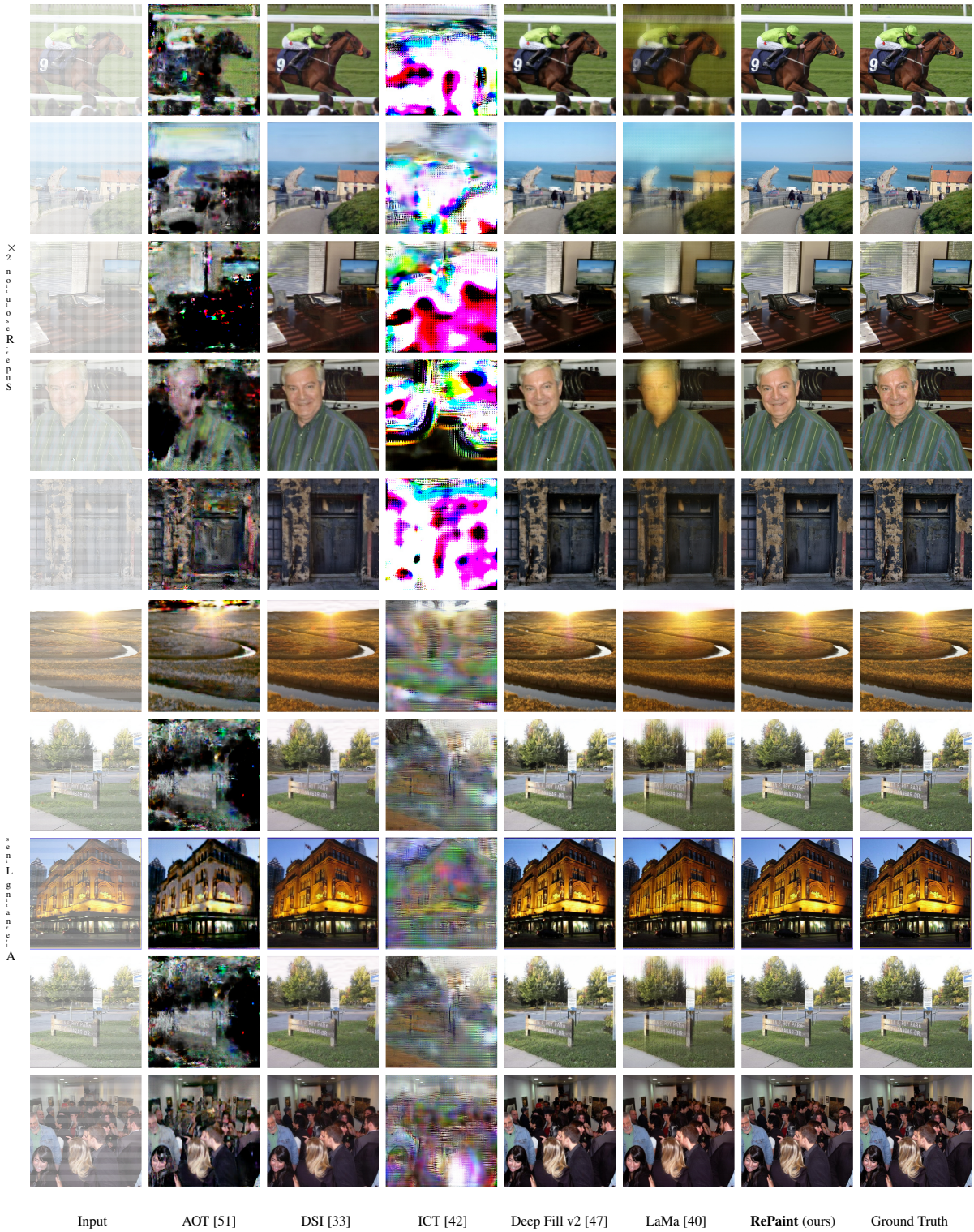


图23. Places2定性结果。与最先进方法的多样化修复效果对比。放大以查看更佳细节。

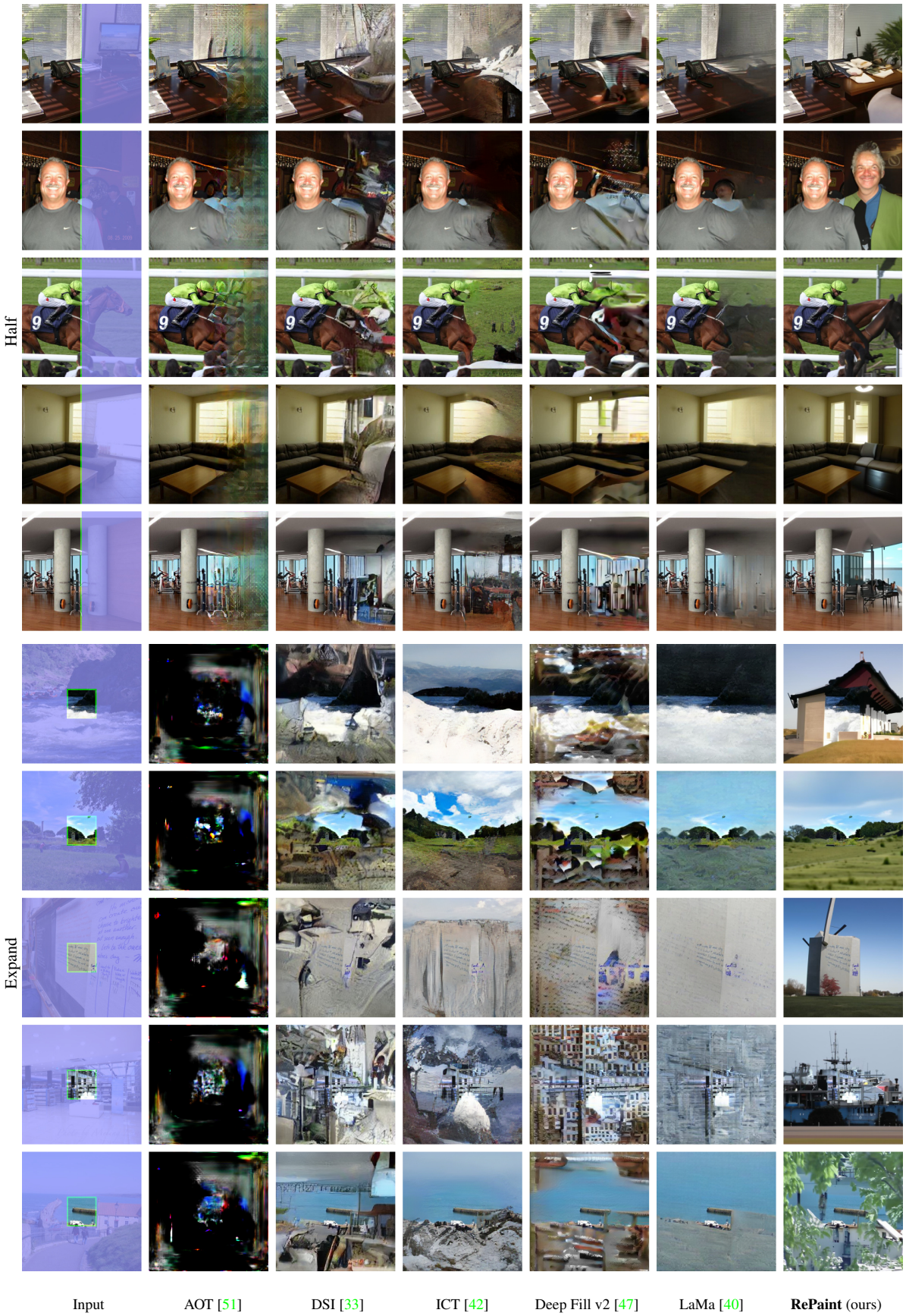


Figure 24. **Places2 Qualitative Results.** Comparison against the state-of-the-art methods for diverse inpainting. Zoom for better details.

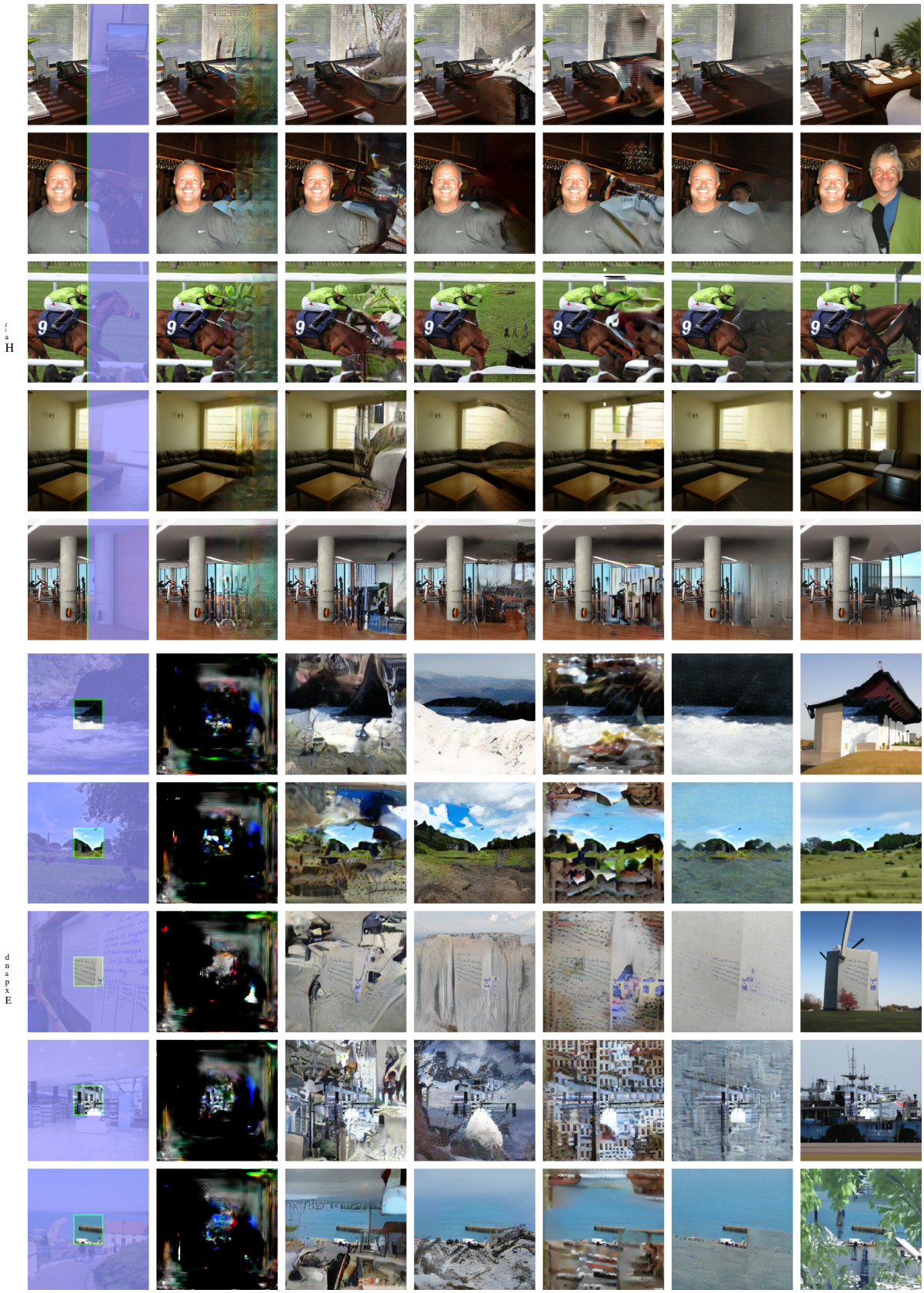


图24. Places2定性结果。与最先进方法在多样化修复方面的比较。放大以查看更佳细节。