RePaint: Inpainting using Denoising Diffusion Probabilistic Models

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

Computer Vision Lab
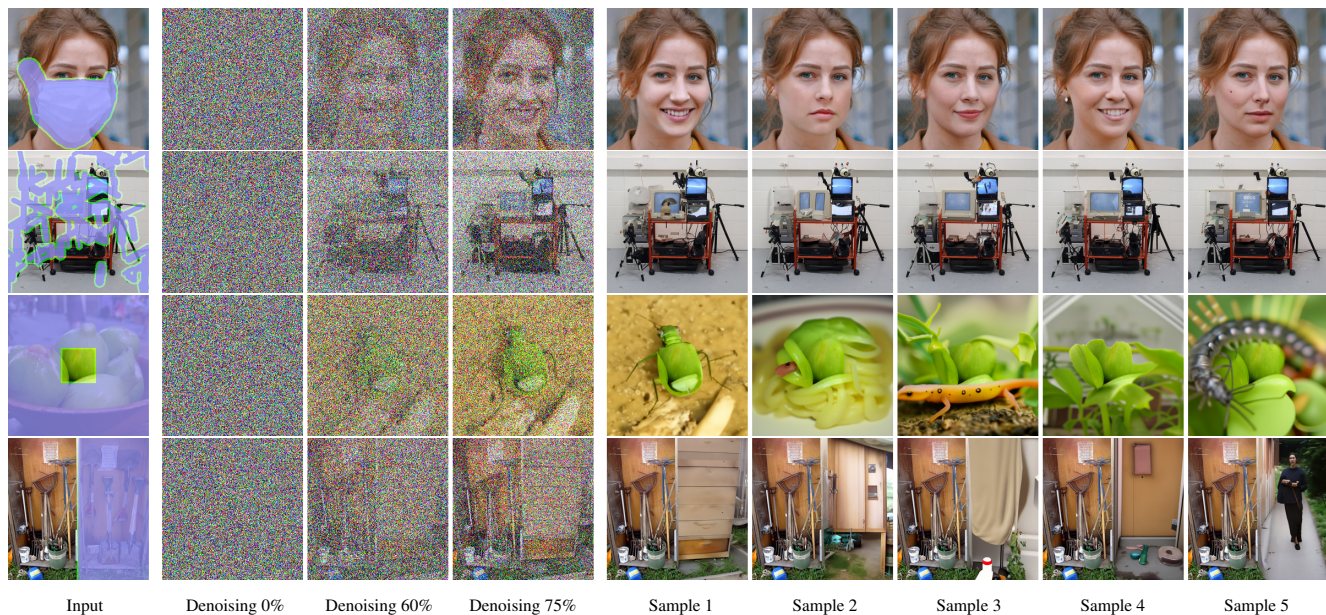
ETH Zürich, Switzerland

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-

age and be semantically reasonable. Inpainting approaches thus require strong generative capabilities. To this end, current State-of-the-Art approaches [20, 39, 47, 50] rely on GANs [8] or Autoregressive Modeling [32, 41, 48]. 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, 37]. 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, 27].

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 [35], 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.* [31], 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, 29, 33, 49].

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, 44] to increase the receptive field, Partial Convolutions [19] and Gated Convolutions [47] to guide the convolution kernel according to the inpainted mask, Contextual Attention [45] to leverage on global information, Edges maps [9, 26, 42, 43] or Semantic Segmentation maps [14, 30] to further guide the generation, and Fourier Convolutions [39] 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.* [54] and Zhao *et al.* [52] propose a VAE-based network that trade-offs between diversity and reconstruction. Zhao *et al.* [53], 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 [32, 41, 48], 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.* [34] exploits the StyleGAN [17] prior to successfully inpaint missing regions. However, similar to super-resolution methods [5, 25] that leverage the StyleGAN latent space, it is to limited specific scenarios like faces. Noteworthy, a Ulyanov *et al.* [40] 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-

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.* [37] applied early diffusion models to inpainting. More recently, Song *et al.* [38] 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, 24]. 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 [24]. 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 [24]. 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 [28] 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 [36]. 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 [37] 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,

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

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,

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

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

$$\mathbb{E}\left[-\log p_\theta(\mathbf{x}_0)\right] \leq \mathbb{E}_q\left[-\log \frac{p_\theta(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T} | \mathbf{x}_0)}\right] \qquad (3)$$

$$= \mathbb{E}_q\left[-\log p(\mathbf{x}_T) - \sum_{t \geq 1} \log \frac{p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t)}{q(\mathbf{x}_t | \mathbf{x}_{t-1})}\right] = L$$

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

$$\mathbb{E}_q\left[\underbrace{D_{\mathrm{KL}}(q(\mathbf{x}_T | \mathbf{x}_0) \| p(\mathbf{x}_T))}_{L_T}\right. \qquad (4)$$

$$\left. + \sum_{t > 1} \underbrace{D_{\mathrm{KL}}(q(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{x}_0) \| p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t))}_{L_{t-1}} \underbrace{- \log p_\theta(\mathbf{x}_0 | \mathbf{x}_1)}_{L_0}\right]$$

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 $\mu_\theta(x_t, t)$,

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_\theta(x_t, t)\right) \qquad (5)$$
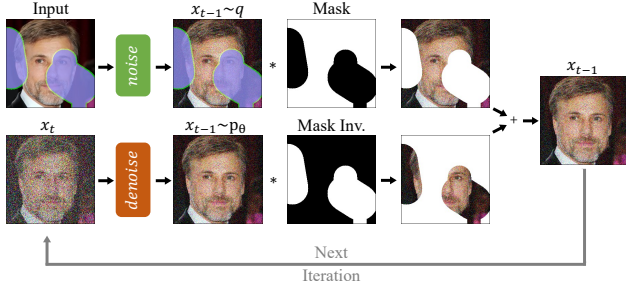
Figure 2. **Overview of our approach.** RePaint modifies the standard denoising process in order to condition on the given image content. In each step, we sample the known region (*top*) from the input and the inpainted part from the DDPM output (*bottom*).

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

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

As introduced by Nichol and Dhariwal [27], 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,

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

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,

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

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

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

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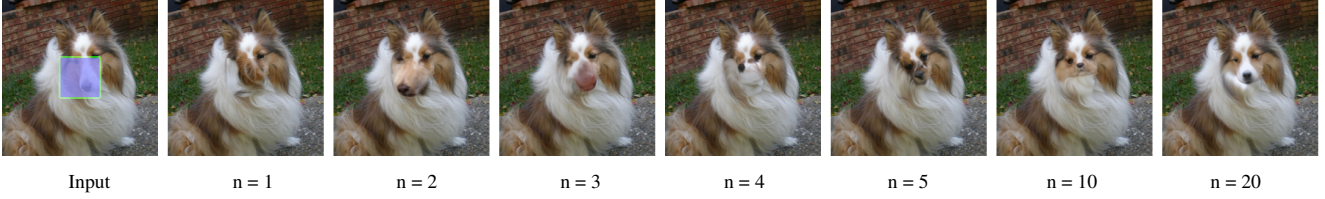
Figure 3. **The effect of applying n sampling steps.** The first example with $n = 1$ is the DDPM baseline, the second with $n = 2$ is with one resample step. More resampling steps lead to more harmonized images. The benefit saturates at about $n = 10$ resamplings.

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

---

Algorithm 1 Inpainting using our RePaint approach.

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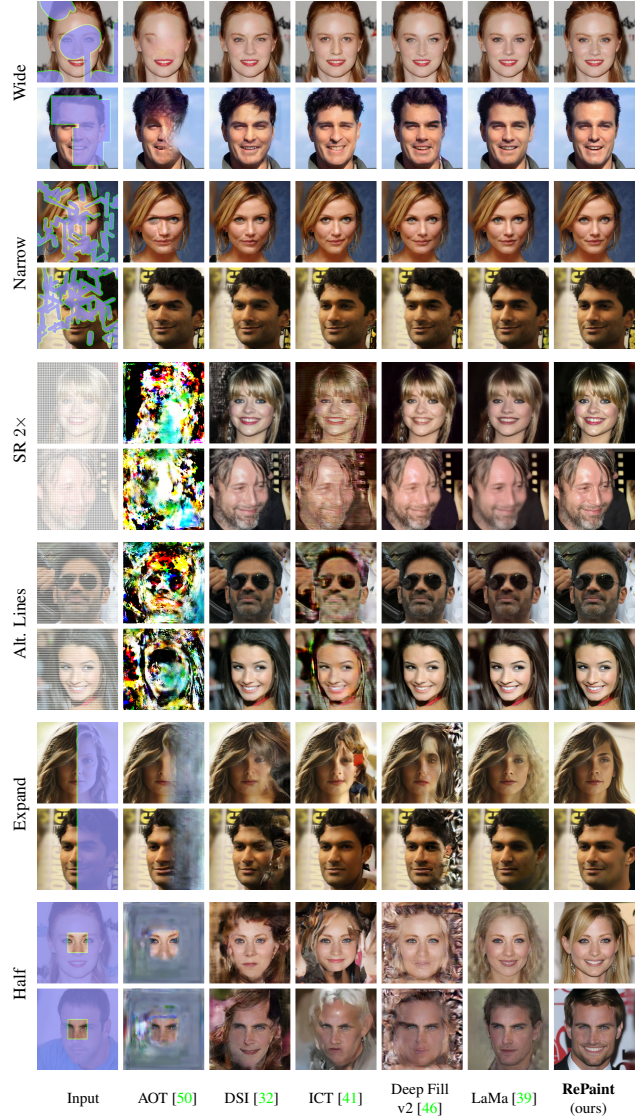12:   **end for**
13: **end for**
14: **return** x_0

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.

report a detailed discussion of mask robustness and diversity, respectively. We also report with additional results, analysis, and visuals in the supplement.

5.1. Implementation Details

We validate our solution over the CelebA-HQ [21], and Imagenet [35] 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×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 [51], 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 supplement 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
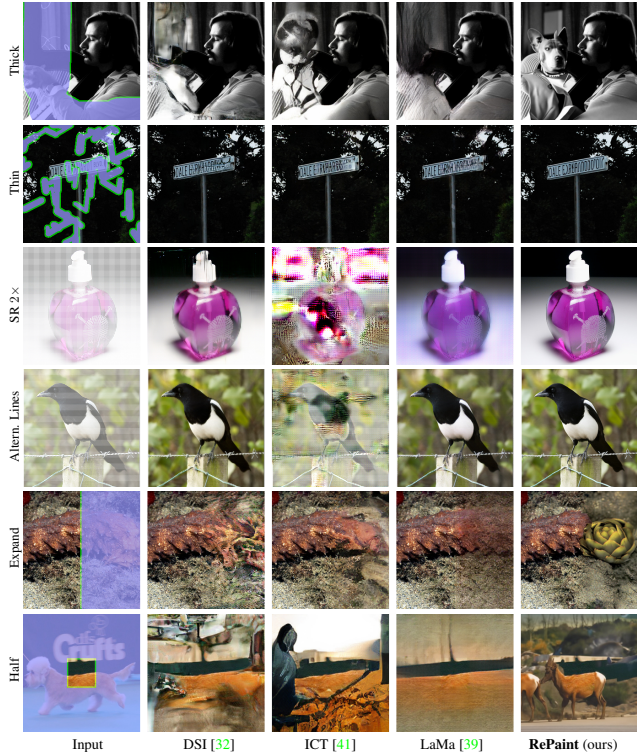

Figure 5. **ImageNet Qualitative Results.** Comparison against the state-of-the-art methods for pluralistic inpainting methods over different mask settings. Zoom-in for better details.

large connected area of the image. All quantitative results are reported in Table 1 and visual results in Figure 4 and 5.

Methods: We compare our approach against several state-of-the-art autoregressive-based or GAN-based approaches. The autoregressive methods are DSI [32] and ICT [41], and the GAN methods are DeepFillv2 [46], AOT [50], and LaMa [39]. 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 supplement for qualitative and quantitative results over the Places2 [55] dataset.

Wide and Narrow masks: To validate our method on the standard image inpainting scenario, we use the LaMa [39] 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,

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 [50]	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 [32]	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 [41]	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 [46]	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 [39]	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 [32]	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 [41]	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 [39]	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 [39] 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 [50] 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

	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 [39] Wide mask setting. The number of diffusion steps is denoted by T, and the number of resamplings by r.
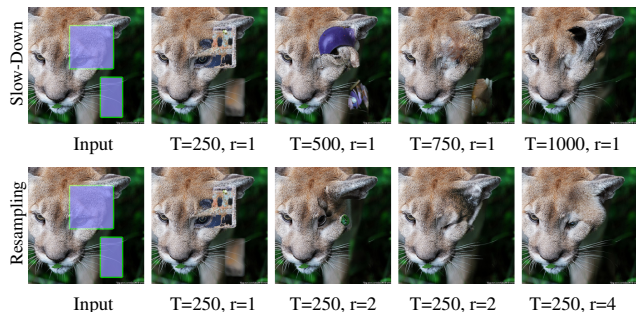

Figure 7. **Qualitative Analysis of the use of computational budget.** RePaint produces higher visual quality with the same computational budget by resampling (*bottom*) compared to slowing down the diffusion process (*top*). The number of diffusion steps is denoted by T and resamplings by r.

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 [24], 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 [39]. 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 [24]	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 [24]	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 [24]	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 [24] 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 [22, 23]. 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.

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] 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.

[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.

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

[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.

[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.

[7] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. *arXiv preprint arXiv:2105.05233*, 2021.

[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.

[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.

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

[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.

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

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

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

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

[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.

[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.

[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.

[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.

[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.

[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.

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

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

[24] 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.

[25] 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.

[26] 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.

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

[28] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

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

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

[39] 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.

[40] 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.

[41] Ziyu Wan, Jingbo Zhang, Dongdong Chen, and Jing Liao. High-fidelity pluralistic image completion with transformers. *arXiv preprint arXiv:2103.14031*, 2021.

[42] 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.

[43] 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.

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

[45] 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.

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[50] Yanhong Zeng, Jianlong Fu, Hongyang Chao, and Baining Guo. Aggregated contextual transformations for high-resolution image inpainting. *arXiv preprint arXiv:2104.01431*, 2021.

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

[52] 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.

[53] 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.

[54] 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.

[55] 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.