RePaint: Inpainting using Denoising Diffusion Probabilistic Models

Kate Liang Huining Liu Hongrui Tang Jennie Wu Mengjia Zhao
https://github.com/HongruiTang/RePaint-reimplementation

1. Introduction

Image inpainting is the problem of filling in an image's masked regions in a semantically consistent way with the surrounding pixels. Traditional methods often struggle to generate semantically coherent results and tend to overfit to specific mask distributions, which severely limit their generalizability, especially when faced with unseen irregular or extreme mask shapes.

To overcome these limitations, the paper proposes an inpainting method that can handle arbitrary mask patterns without requiring mask-specific training while generating high-quality and semantically harmonious image completions. It uses pretrained unconditional Denoising Diffusion Probabilistic Models (DDPM) and introduces a resampling strategy that jumps forward and backward in the reverse denoising process, conditioning on the known area of the image to improve semantic harmony. The key contributions of the paper include: (1) a conditioning method that does not require retraining the DDPM, (2) a resampling schedule that improves semantic coherence, and (3) strong empirical results that outperform GAN and autoregressive baselines on multiple mask types.

2. Chosen Result

In the original paper, the authors trained a DDPM on the CelebA-HQ dataset for 250,000 iterations, which takes 5 days even on 4×V100 GPUs. Due to resource constraints and the fact that the RePaint algorithm is adaptable to various DDPMs, we utilize pretrained models and focus on the CelebA-HQ dataset to reproduce the following: (1) The visualization results using different masks presented in Figure 1; (2) The LPIPS results in Table 1; (3) Ablation results on the effect of resampling steps and jump length demonstrated in Figures 3.

This includes the reproduction of the paper's main contribution, the Repaint method, as well as the evaluation results that will serve as proof of the validity of our implementation.

3. Methodology

In this project, we reimplemented the RePaint method for image inpainting, which modifies the unconditional DDPM's reverse process to include two additional stages: conditioning on the known region and resampling, which involves jumping forward in the reverse process. Then, we evaluated our method by visualizing the generated images, computing the LPIPS scores, and comparing them with the paper results.

3.1. Dataset and Model

To reproduce LPIPS scores from the original paper, we used the pre-trained *ddpm-celebahq-256* model from Google that was trained on the CelebA-HQ-256 dataset. Since no retraining and model architecture redesign are needed, we focused on reimplementing Algorithm 1.

Algorithm 1 Inpainting using our RePaint approach

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

3.2. Diffusion Inference Sampling Algorithm

Prior inpainting methods typically train models to be explicitly conditioned on the mask and known pixels. In contrast, RePaint requires no training or fine-tuning, which saves computation resources while producing promising results.

As illustrated in Figure 2, at each denoising step $t - 1$, random noise is added to the known (unmasked) region of the original image (top row), preserving the original content. This noisy known area is combined with the denoised masked region generated at step $t - 1$ (bottom row). Thus, the denoiser operates on an image where the known pixels guide the model to generate harmonious results between the

Mask Type	Wide		Narrow		Super Res2x		Alt. Lines		Half		Expand	
Input												
RePaint [1]												
Ours												

Figure 1. We compare the images inpainted by our re-implementation and the one produced by the paper.

Table 1. **CelebA-HQ Quantitative Results.** Comparison against state-of-the-art methods. We compute the LPIPS (lower is better) for 6 different mask settings.

CelebA-HQ	Wide LPIPS↓	Narrow LPIPS↓	Super-Res 2x LPIPS↓	Altern. Lines LPIPS↓	Half LPIPS↓	Expand LPIPS↓
RePaint [1]	0.059	0.028	0.029	0.009	0.165	0.435
Ours	0.075	0.043	0.057	0.048	0.193	0.534

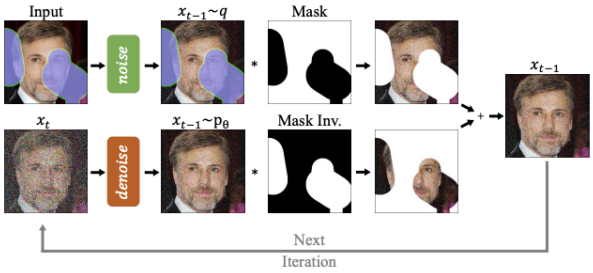

Figure 2. Overview of the RePaint algorithm.

masked and unmasked regions.

In our reimplementation, noise sampled from the original DDPM scheduler is added to the known area in the forward diffusion process (Line 5). Meanwhile, to denoise x_t to x_{t-1}, we use the pretrained UNet to predict the noise added to the masked region at timestep t. Combining the noised unmasked region with the denoised masked region, we obtain the inpainted image x_{t-1}.

However, this process alone produces semantically inconsistent inpainting results, as demonstrated in our ablation study (Figure 3 the image generated with $r = 1$ aligns with the DDPM baseline). In this case, the generated masked region primarily aligns with the immediate surrounding pixels (e.g., forming a face) but fails to ensure coherence with the entire known region. For instance, this inconsistency results in generating a person with inconsistent hairstyles and facial features, as observed in Figure 3. Therefore, the authors do a certain number of forward steps, labeled as jump length, adding noise again to inpainted x_{t-1} to move it forward in timestep, and repeat the denoising process on x_t. The intuition behind this bidirectional resampling process is that the forward diffusion steps inject noise again, erasing part of what was generated incorrectly, while the reverse steps refill details into now more aligned pixels. This resampling process is represented as U in Algorithm 1.

Due to computational constraints, we modified Algorithm 1 in 2 places. First, instead of resampling for every single timestep, we resample every 10 reverse steps to reduce inference time. In addition, instead of going forward for only 1 step, we tried different numbers of forward steps to explore the impact of jump length on generation quality.

3.3. Evaluation metric

As in the original paper, we use the LPIPS metric for quantitative evaluation. It computes the semantic distance between image patches to measure perceptual similarity. Moreover, we eyeball the visual results to observe the effect of varying hyperparameters, i.e. number of resampling and jump length, particularly in the ablation experiments.

4. Result and Analysis

We achieved results comparable to those of the original paper on the CelebA-HQ dataset. In this section, we present our results and discuss the discrepancies we observed.

4.1. Visual Results

The original paper conducted experiments over a wide range of masks with different test images and compared their

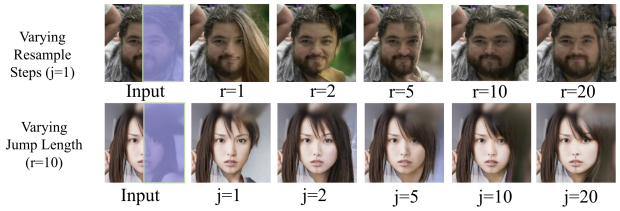

Figure 3. The effect of applying r sampling steps and j jump length. More resampling steps and jump length lead to more harmonized images. The benefit saturates at about $r = 10$ resamplings and $j = 10$ jump length.

results against several other state-of-the-art methods for CelebA-HQ inpainting as shown in Figure 4. We also generated images using the same masks and test images, and visually compared our results with the original paper's in Figure 1. As can be seen from the comparison, our output is similar to the output from the original paper in terms of the level of detail and semantic correctness. This similar outcome confirms that RePaint is robust and generalizes well on different input masks, as well as the effectiveness of our reimplementation.

4.2. Evaluation Result

In order to evaluate the performance of the model, we computed the LPIPS score of our model on different masks—lower LPIPS score are desirable as they indicate that image patches are perceptually similar. Table 1 shows the score from the original paper on the first row, and our result on the last row.

Our results have a slightly higher LPIPS score than the original paper, but match the overall pattern. This discrepancy may come from two different aspects:

1. We used a pre-trained diffusion model from Hugging Face, whereas the original paper trained its own model on the CelebA-HQ dataset. The model we used may have a different performance from the model used in the original paper.

2. The original paper calculated the LPIPS score for each mask using 100 different test images, while we used only 20 because of time and computational limitations. The paper also didn't share which images were used in their testing experiment, so we randomly picked 20 images from the dataset.

While most of our LPIPS scores results match the original paper's result, we observe one mismatch. The original paper reports a score of 0.009 on the Altern. Lines mask, while we obtained 0.048. We are unsure how the authors were able to obtain such a score when the images we reproduced appear extremely similar to theirs. Although even minor discrepancies in implementation can significantly impact performance, we do not feel that this explains the 5x discrepancy well.

4.3. Ablation Study

We conducted an ablation study to observe the effect of using different jump lengths and resample steps on the resulting image. We arrived at a similar conclusion as the original paper—increasing the resampling steps and jump length generates more harmonized images, but the benefits saturate at approximately $r = 10$ and $j = 10$ as shown in Figure 3.

5. Reflections

5.1. Key takeaway

RePaint presents a novel conditioning method that enables pretrained unconditional diffusion models to perform inpainting across a wide variety of mask types. Its unified inference schedule ensures harmonization between known and generated regions, resulting in semantically coherent image completions. Our implementation successfully reproduced the original results on CelebA-HQ, achieving comparable LPIPS metrics across multiple mask types. Moreover, our ablation experiments revealed similar performance trends to those reported in the original paper, reinforcing the validity of RePaint's design.

5.2. Lessons Learned

One of the key lessons was the importance of assessing the computational feasibility of a project early on. While the RePaint method is elegant in theory, its iterative sampling procedure is resource-intensive, requiring thoughtful planning and resource management even when we are simply utilizing the base model at inference time instead of training deep models from scratch. Additionally, it was interesting to note that the intuition behind the RePaint method, including the effects of changing various hyperparameters, was actually quite straightforward, which is perhaps further testament to how simple ideas often work best, even in deep learning.

5.3. Future work

We see promising directions for extending RePaint's capabilities, such as incorporating language-based conditioning to enable instruction-driven inpainting. Another exciting avenue is the adaptation of RePaint from static image inpainting to temporal tasks like video inpainting, where coherence across frames presents a unique challenge.

References

[1] Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., and Van Gool, L. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11461–11471, 2022.

A. Figures From Original Paper

This section includes the figures and results from the original paper that we choose to reproduce. This includes Figure 4, Figure 5 and Figure 6.

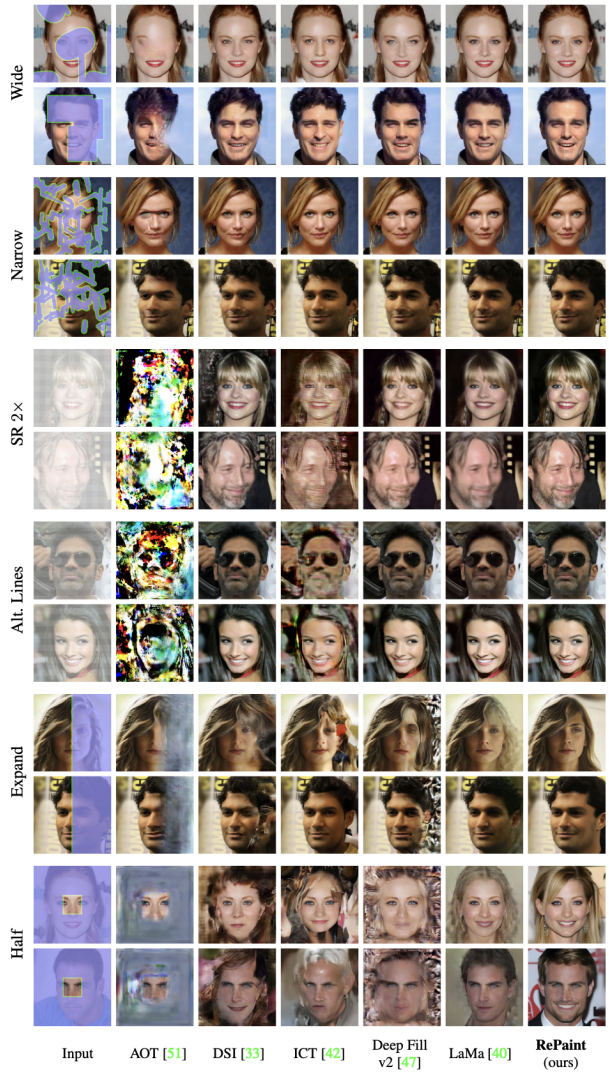

Figure 4. Qualitative Results from the original paper comparing against the state-of-the-art methods for celebrity Face Inpainting over several mask settings.

Figure 5. Ablation study example from the original paper showing the effect of applying n sampling steps. The original paper found more resampling steps lead to more harmonized images. The benefit saturates at about $n = 10$ resamplings.

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*

Figure 6. Quantitative Results from the original paper comparing against the state-of-the-art methods. They computed the LPIPS (lower is better) and Votes for six different mask settings.