Image Inpaiting with Denoising Diffusion Probabilistic Models

Ian Li
Harvey Mudd College
Claremont, CA 91711

Abstract

In this paper, we explore the capabilities of free-form image inpainting through the application of Denoising Diffusion Probabilistic Models (DDPM). Traditional inpainting approaches often struggle with the flexibility and diversity required for complex inpainting tasks, particularly under extreme mask conditions. By employing a pretrained unconditional DDPM, we explore the generative potential of these models without the need for mask-specific adaptations. Our methodology leverages the stochastic nature of the DDPM, where the inpainting process is initiated from a noise distribution and iteratively refined to produce high-quality, contextually appropriate image completions. This project does not alter the original DDPM architecture; instead, it uniquely conditions the generative process on the available image data, which allows for superior generalization across diverse inpainting scenarios.

1 Introduction

1.1 Overview of Image Inpainting

Image inpainting is a sophisticated technique aimed at filling in missing or damaged areas of digital images. The core purpose of image inpainting includes enhancing photo quality, restoring historical images, and editing content within images. The need for advanced inpainting methods has grown with the digital age, as more sophisticated and seamless restoration becomes possible and desirable.

1.2 Traditional Image Inpainting Methods

Traditionally, image inpainting techniques have been employing patch-based techniques that sample and copy patches from the available parts of the image to the occluded regions [3]. While these methods can effectively handle larger areas with repetitive patterns, they struggle with generating semantically meaningful content when the context or the structure in the missing region is not clearly defined.

The limitations of traditional inpainting methods in handling diverse and dynamic inpainting challenges have led to the exploration of more sophisticated generative models. Among these, Deep Learning approaches, particularly Generative Adversarial Networks (GANs) and Convolutional Neural Networks (CNNs), have shown promising results by learning to synthesize new content that is contextually coherent with the existing image data [4][5]. However, even these advanced models can produce artifacts and often require extensive training data tailored to specific inpainting tasks.

Denoising Diffusion Probabilistic Models (DDPMs) present a novel paradigm by framing the inpainting process as a stochastic denoising task. Originating from the field of statistical physics, DDPMs gradually transform a distribution of pure noise into a structured image through a series of learned reverse diffusion steps. This approach naturally integrates the generation of diverse and

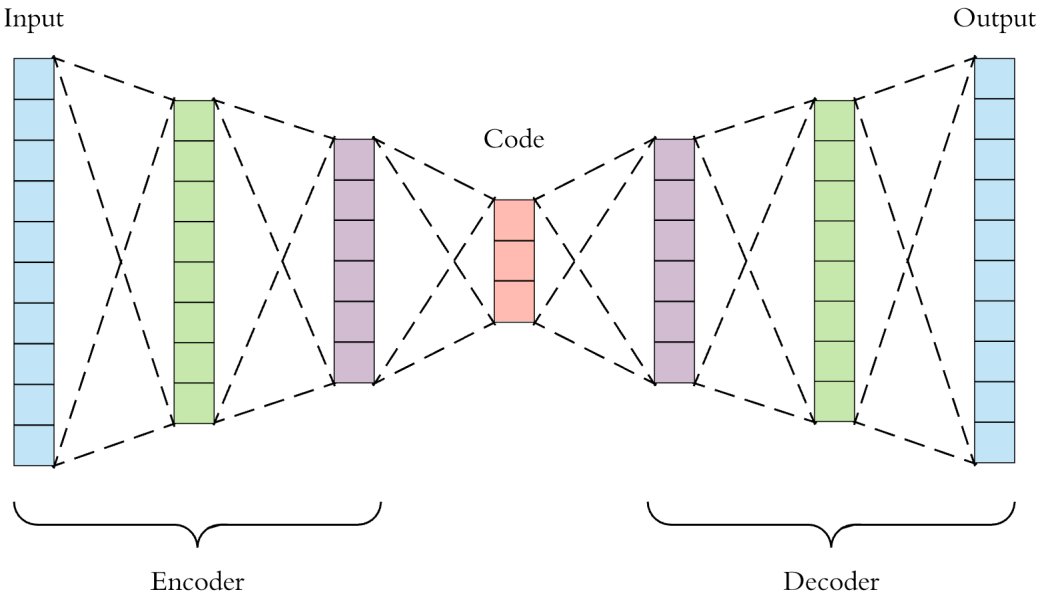
Preprint. Under review.

high-quality images from noisy starting points, making it inherently suited for the complexities of free-form inpainting. The use of DDPMs in inpainting is motivated by their ability to generate varied outputs that are both high in quality and rich in detail, outperforming other state-of-the-art methods, particularly in scenarios involving extreme or irregular mask shapes [6][7].

2 Related Works

2.1 Convolutional Encoders for Image Inpainting

Convolutional encoders have become a staple in the design of deep learning-based inpainting models due to their ability to effectively encode spatial hierarchies of features. Pathak et al. (2016) introduced Context Encoders, which employ convolutional neural networks (CNNs) in an encoder-decoder framework as shown below. The encoder compresses the image into a lower-dimensional feature space, and the decoder then reconstructs the missing parts from this compressed representation. This method was among the first to use adversarial training to refine the inpainted outputs, which significantly improved the sharpness and coherence of the inpainted regions compared to previous methods [4].



2.2 Partial Convolution

While convolutional encoders improved the feasibility of inpainting large areas, the challenge of dealing with irregular masks where the area of missing information is not uniformly distributed remained. Liu et al. (2018) addressed this challenge by introducing partial convolutions, where the convolution is masked and renormalized to be conditioned only on valid pixels. This approach allows the network to dynamically adapt its filtering process depending on the extent of image corruption or missing regions, thereby effectively managing varying shapes and sizes of masks. The network trained with partial convolutions automatically learns to inpaint damaged regions by understanding the context of the surrounding pixels, significantly reducing artifacts commonly seen in earlier techniques [8].

These convolution-based methods have been foundational in illustrating the potential of neural networks in automating the task of image restoration. They leverage the spatial hierarchies learned through deep learning to synthesize plausible image content that blends seamlessly with the intact regions of the image. This not only enhances the visual quality of the inpainted images but also expands the practical applications of inpainting in various domains such as digital art restoration, photo editing, and content creation.

Despite the advancements brought about by convolutional encoders and partial convolutions, these methods still face limitations in terms of handling diverse inpainting scenarios with high semantic complexity. This has led to the exploration of more flexible and powerful generative models, such as Denoising Diffusion Probabilistic Models (DDPMs), which offer promising new directions for further enhancing the state-of-the-art in image inpainting.

3 Preliminaries

3.1 Diffusion Models

Diffusion models are a class of generative models that transform a simple noise distribution into a complex data distribution through a gradual, iterative process. The concept is inspired by the physical process of diffusion, where particles spread from regions of high concentration to low concentration until they reach equilibrium.

Mathematically, the diffusion process is modeled in reverse, starting from a distribution of noise and progressively 'denoising' it to form structured data. This process is divided into a sequence of steps $t = 1, 2, \ldots, T$, where T is the total number of diffusion steps. At each step, a slightly less noisy version of the data is generated by applying a conditional probability model:

$$q(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu(\mathbf{x}_t, t), \sigma_t^2 \mathbf{I})$$

where $\mathcal{N}$ denotes the normal distribution, $\mu(\mathbf{x}_t, t)$ is the mean predicted by the model for the previous timestep given the current state $\mathbf{x}_t$, and σ_t^2 is the variance at time t.

3.2 Denoising Diffusion Probabilistic Models (DDPMs)

DDPMs refine the concept of diffusion models by specifically designing the reverse process as a denoising task. They iteratively convert a noise distribution into a data distribution by learning to reverse a Markov chain that gradually adds Gaussian noise to the data.

The forward process is defined as:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t \mathbf{I})$$

Here, β_t are pre-defined variance schedules spread over T timesteps, gradually transforming data into noise. The reverse process, modeled by the neural network, aims to estimate the parameters of the Gaussian distribution at each step, essentially 'denoising' the input. It is given by:

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \sigma_\theta(\mathbf{x}_t, t))$$

where μ_θ and σ_θ are functions parameterized by the neural network, using the noisy data $\mathbf{x}_t$ at step t to predict the distribution parameters of the previous step $\mathbf{x}_{t-1}$.

The objective of training a DDPM is to minimize the difference between the generated samples and the real data distribution, typically using a variant of the mean squared error (MSE) across the diffusion steps:

$$\mathcal{L}(\theta) = \mathbb{E}_{t,\mathbf{x}_0,\epsilon}\left[\|\epsilon - \epsilon_\theta(\mathbf{x}_t, t)\|^2\right]$$

where ϵ is the noise used in the forward process, and ϵ_θ is the noise predicted by the model.

4 Methods

4.1 Pretrained Model

We utilize a pre-trained DDPM developed by OpenAI, specifically the 256x256 guided diffusion model. This model is unconditionally trained, which means it learns to generate images from a noise distribution without specific conditions apart from the input noise itself. It is trained on the ImageNet Dataset which encompasses a diverse range of images, enabling the model to handle various content types and styles effectively.

The model operates by reversing a diffusion process that initially converts an image into Gaussian noise over a series of steps. During the reverse process, the model iteratively denoises this input

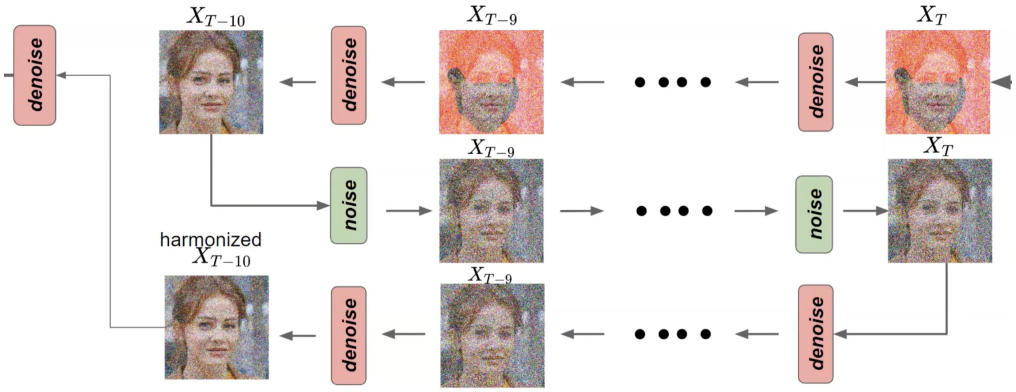
to reconstruct the image. The ability of the model to generate high-quality, diverse outputs makes it an excellent candidate for tasks that require filling in missing parts of images, as it can infer the missing content by effectively 'imagining' the absent parts based on the noise pattern and its training on complete images.

4.2 Resampling

The motivation behind the resampling technique is to address the sometimes semantically inconsistent outputs produced during the reverse diffusion process. While DDPMs are proficient at textural and structural accuracy, they can struggle with semantic coherence, particularly in complex inpainting tasks where the context must be inferred from limited available data.

Resampling involves selectively repeating the diffusion steps multiple times during the reverse process. By doing so, the model can revisit and revise its earlier predictions, refining the output iteratively. This is particularly useful for correcting errors or inconsistencies that may have been introduced in earlier stages of the generation process.

Applying the resampling technique increases the computational overhead because certain stages are processed multiple times, but it significantly enhances the quality of the inpainting. It allows the model to better integrate the inpainted regions with the surrounding image, leading to outputs that are not only visually pleasing but also more contextually appropriate.



4.3 Jumping

The jumping technique is inspired by the need to improve the flexibility of the diffusion process in handling varying inpainting challenges. This technique allows for dynamic adjustments in the generation pathway, enabling the model to 'jump back' to earlier diffusion states and potentially take a different path that might lead to a better outcome.

In the jumping technique, the model can move back and forth within the diffusion timeline during the reverse phase. At any given step, based on the quality of the output, the model can decide to revert to an earlier state and reprocess the subsequent steps. This approach is akin to having a 'second chance' to correct or improve upon the areas that were not ideally inpainted in the first pass.

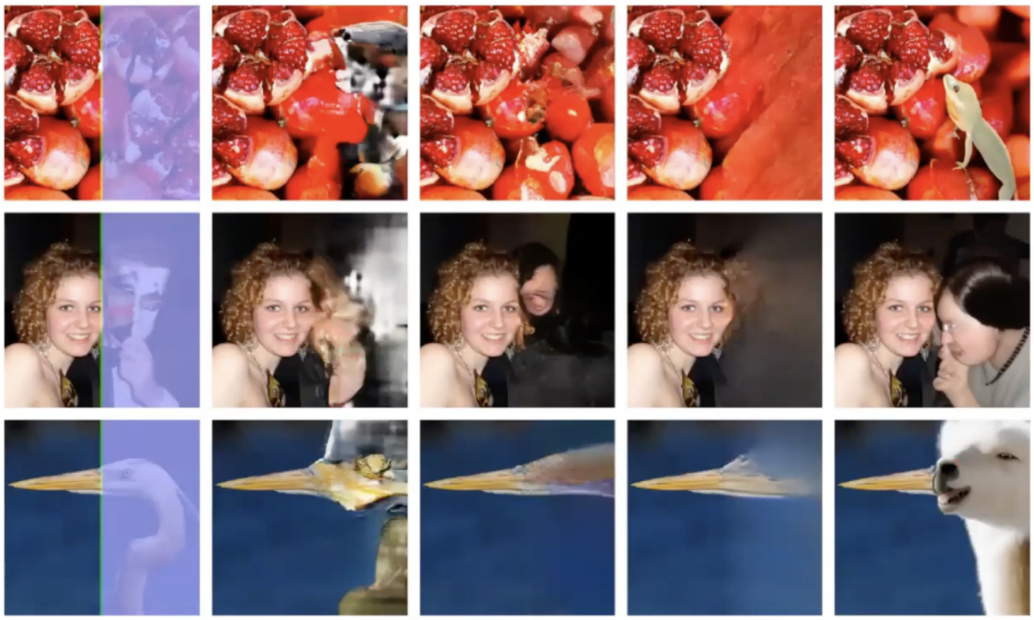
Like resampling, jumping increases the time required for inference due to the additional computations involved in revisiting previous states. However, it provides a significant advantage in terms of output quality. By allowing the model to reconsider and adjust its course during the reverse diffusion, jumping helps in producing more accurate and coherent inpaintings, especially in scenarios with complex or unusual masks.

Together, these techniques leverage the inherent capabilities of the pre-trained DDPM to produce inpaintings that are not only high in resolution and detail but also contextually and semantically coherent. These improvements make the DDPM-based approach particularly robust for free-form inpainting tasks across a variety of image types and conditions.

5 Results

Our evaluation of the Denoising Diffusion Probabilistic Model (DDPM) for image inpainting tasks across a variety of scenes and objects demonstrates the model's proficiency in generating realistic and contextually appropriate image completions. The analysis encompasses various types of images, including food items, human faces, animals, and environmental scenes, providing a comprehensive overview of the model's capabilities and areas for potential enhancement.

The DDPM showed exceptional ability in maintaining textural details, as evidenced by the inpainting results of the strawberry images. The model successfully replicated the complex textures of both the strawberries and the accompanying cream, seamlessly blending the inpainted sections with the original image areas. This highlights the model's effectiveness in preserving and synthesizing fine details, which is crucial for the realism of the inpainted images.



In the portraits, the model demonstrated a robust capacity to handle human features, reconstructing facial elements with a high degree of realism. Despite the challenges associated with accurate skin tone rendering and the preservation of expression, the model managed to maintain the integrity of the facial structures. However, some instances of blurring and minor distortions were observed, particularly around complex areas such as hair. These artifacts, while relatively minor, suggest areas where the model could be further refined.

6 Discussion

6.1 Limitations

While the model excelled in many respects, certain challenges remain. In the provided images, hallucinations can be observed particularly in areas with complex details, such as faces or intricate backgrounds. For instance, in the human portraits, while the general shape and position of facial features are preserved, specific details such as the exact eye placement, hairstyle, or expression may not exactly match the original or expected appearance. This issue is often exacerbated in scenarios where the model needs to inpaint significant portions of an image based on minimal available information.

Also, bias in generative models like DDPMs tends to stem from the data on which they are trained. If the model is predominantly trained on images of certain environments, such as urban scenes, it might struggle with natural landscapes or vice versa. Similarly, if the training data includes predominantly young individuals, the model may not accurately reconstruct the faces of older adults. This bias was evident in the handling of specific textures and structures, such as the smoke and fine details in facial features, where the model defaulted to simpler, more generic textures that may not truly represent the unique characteristics of the original subjects.

6.2 Possible Mitigations

To combat hallucinations and bias, one effective strategy is to diversify the training dataset. By incorporating a wider array of images that cover a broader spectrum of ages, ethnicities, environments, and objects, the model can learn a more balanced representation of the world, which in turn can help minimize bias and reduce the likelihood of inappropriate hallucinations.

In addition, feedback mechanisms that allow the model to learn from its mistakes can be beneficial. By applying corrections based on discrepancies between the model's outputs and ground truth data, and then reintegrating these corrected examples into the training cycle, the model can gradually learn to avoid past mistakes and refine its generative capabilities.

7 Conclusion

In this study, we have explored the capabilities of a Denoising Diffusion Probabilistic Model (DDPM) for the task of image inpainting, assessing its efficacy across a range of scenarios including the inpainting of food, faces, animals, and urban landscapes. Our findings affirm that DDPMs are capable of producing high-quality inpaintings that blend seamlessly with original images, preserving texture, color, and structural integrity. The model particularly excelled in handling complex textures and structured environments, suggesting its potential utility in digital art restoration and urban planning applications.

However, the model also displayed limitations, such as a tendency for hallucination and output bias, which were most evident in the rendering of human features and dynamic elements. These issues underscore the need for broader and more diverse training datasets, as well as enhanced model architectures that can effectively address complex inpainting challenges. Future research should focus on refining DDPMs through advanced training protocols and exploring hybrid generative approaches to mitigate these limitations and fully harness the capabilities of DDPMs in automated image inpainting.

We envision that future works should focus on expanding the diversity of training datasets to ensure broader representational coverage and reduce model bias. Moreover, incorporating feedback loops and iterative refinement techniques, such as resampling and jumping, could further enhance the model's performance, particularly in complex inpainting scenarios. Lastly, exploring hybrid approaches that combine DDPMs with other generative models might offer new pathways to overcome current limitations.

References

[1] Bertalmio, M., Sapiro, G., Caselles, V., Ballester, C. (2000). Image inpainting. In Proceedings of the 27th annual conference on Computer graphics and interactive techniques.

[2] Ballester, C., Bertalmio, M., Caselles, V., Sapiro, G., Verdera, J. (2001). Filling-in by joint interpolation of vector fields and gray levels. IEEE transactions on image processing, 10(8), 1200-1211.

[3] Criminisi, A., Perez, P., Toyama, K. (2004). Region filling and object removal by exemplar-based image inpainting. IEEE Transactions on image processing, 13(9), 1200-1212.

[4] Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A. A. (2016). Context encoders: Feature learning by inpainting. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2536-2544).

[5] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... Bengio, Y. (2014). Generative adversarial nets. In Advances in neural information processing systems (pp. 2672-2680).

[6] Ho, J., Jain, A., Abbeel, P. (2020). Denoising diffusion probabilistic models. In Advances in Neural Information Processing Systems.

[7] Nichol, A., Dhariwal, P. (2021). Improved denoising diffusion probabilistic models. arXiv preprint arXiv:2102.09672.

[8] Liu, G., Reda, F. A., Shih, K. J., Wang, T. C., Tao, A., Catanzaro, B. (2018). Image inpainting for irregular holes using partial convolutions. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 85-100).

[9] Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., Van Gool, L. (2022). RePaint: Inpainting using Denoising Diffusion Probabilistic Models. CVPR 2022.