Diffusion Models for Image Inpainting

Rohan Nafde
Indian Institute of Technology, Bombay
210070068@iitb.ac.in

Abstract

Lately, diffusion models for image inpainting have been a topic of growing interest. However, generating the missing regions consistent with the original image (especially for complex images) has always been challenging. In this report, we plan to first glance over three diffusion models for image restoration and compare their applications and advantages. The first model involves a Partial Differential Equation (PDE) based image restoration [1]. The second model involves the Fast Fourier Transform (FFT) [2]. The third model, called SmartBrush [4], uses text and shape guidance to complete the missing region. Finally, we will look at a survey on Image Inpainting using Diffusion Models [3] to generate a set of diverse images for the same deteriorated image. This report will compare the various advantages and disadvantages of each model so that it helps us decide which model can be used in a specific situation

Index: Diffusion model, image inpainting, partial differential equations, fast fourier transform.

1. Introduction

Image reconstruction, also known as image inpainting, represents the computer vision process of restoring the missing areas of a damaged image as plausibly as possible from the known zones around them [1]. The applications of this field have been increasing with time, and hence, we have developed several models for the same. If we have an old image (hard or soft copy) that is torn in specific regions or has missing patches of pixels, we use image inpainting to reconstruct the image. Something as simple as denoising also involves inpainting. However, one of the most important applications is the removal of objects or entities from an image, where the missing patch created by the removal is filled up using inpainting. In this report, we will understand, analyze and compare different diffusion models for image inpainting

We use diffusion models over generative adversarial networks (GANs) since GANs are less reliable. A diffusion model utilises the concept of 'reverse diffusion,' essentially reversing the diffusion process (homogeneous mixing) by a series of iterative steps. The dataset they are trained on is large and more particular, thus making it more reliable. Diffusion models also allow us to generate diverse images for large areas of missing regions, which GANs cannot do.

The general framework of a diffusion model used for image inpainting is as follows. The main tasks involve denoising and reconstruction. In each case, we develop and train a model that iteratively refines the input image, which consists of noise and masks. Here, the mask is a terminology used for the missing regions. The model either removes noise from the image first and then reconstructs it or roughly reconstructs it and then removes the noise to refine the estimated patch. The various models differ in their implementation methods and the data they are trained on and can be autonomous or user-guided.

In the second section, we will examine the diffusion models listed in the papers and the survey, the mathematics involved and a diagrammatic representation of each. In the third section, we will look at the examples of image inpainting done by each model discussed in the previous section and compare the results. The fourth and final section deals with the concluding remarks about the functioning, advantages and disadvantages of each model,

2. Methodology

2.1. PDE Based Image Inpainting

2.1.1 Variational Image Noise Removal

The general variational framework used for image denoising utilises a cost functional $J(u)$. The task is to minimise this functional by finding u_{min} minimising functional $J(u)$. This minimisation process involves solving the associated Euler-Lagrange Equation leading to the following PDE to get u_{min}:

$$\frac{\partial u}{\partial x} = div(\psi'(||\nabla u||^2)\nabla u) - \lambda(u - u_0) \qquad (1)$$

This helps us in the numerical approximation of the model. It involves the use of a 4-NN discretization of the Laplacian

operator. It leads to the iterative approximation:

$$u^{t+1} \approx u^t + \alpha \sum_{q \in N(p)} \psi'(||\nabla u_{p,q}(t)||^2) \nabla u_{p,q}(t)) - \lambda(u - u_0) \tag{2}$$

where $N(p) = (x-1, y)(x+1, y)(x, y-1)(x, y+1)$, $\nabla u_{p,q}(t) = u(q,t) - u(p,t)$, $p = (x, y)$, $\alpha \in (0, 1)$, $t = 1, 2, ..., N$. This iterative model converges fast to the solution $u^N \approx u_{min}$

2.1.2 Variational Image Inpainting

The method utilised is a robust variational PDE technique. Another functional is to be minimised here with some set boundary conditions. We get our equation for U_{min} (steady state solution. The discrete form of our solution satisfies the following steepest descent algorithm:

$$2u_{k+1} - \beta(u_{k+1}) = u + 0 + u_k, k = 0, 1, ... \tag{3}$$

The degraded image $u_0 = u^0$ is transformed into the restored image u^N. We take $\nabla B(u) = -Au$ to finally get the implicit finite difference scheme:

$$u_{k+1}(hA + 1) = u_k, k = 0, 1, ... \tag{4}$$

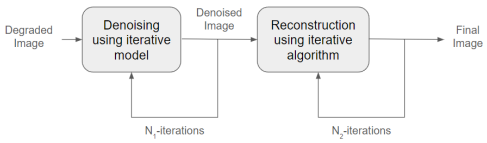

Figure 1. The entire flow of the PDE-Based algorithm [1]

2.2. Diffusion Model for Image Inplanting using FFT - Proposed Methods

A FFT-DM (Fast Fourier Transform Diffusion Model) is proposed in this section. It contains two components: a Denoising Diffusion Probabilistic Model (DDPM) and a Convolutional Neural Network (CNN) [2]. The DDPM is used to generate images prior, while the CNN is used to capture finer details.

The diffusion model proposed is shown in Figure 2, which is trained as an unconditional DDPM. Subsequently, we use a U-Net to predict the Gaussian Distribution parameter for reverse diffusion

We then calculate the loss function of the FFT-DM by making use of the mean and variance of the images at different time steps (as we can see, this is an iterative algorithm). The loss function turns out to be:

$$L = E_{t,x_0,\epsilon}||\epsilon - \epsilon_\theta(x_t, t)||^2 + \lambda(L_0 + L_1 + ...) \tag{5}$$

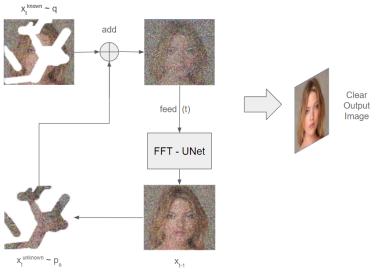

Figure 2. Training and functioning of the FFT-based model [2]

To define our diffusion model, we sample the known and unknown regions of image x_t as x_t^{known} and $x_t^{unknown}$, respectively. The equations for one reverse step are:

$$\begin{aligned} x_t^{known} &= \mathcal{N}(\sqrt{\alpha_t}x_0, \sqrt{1-\alpha_t}\epsilon_0) \\ x_t^{unknown} &= \mathcal{N}(mean(t), variance(t)) \\ \alpha_t &= \textstyle\prod_{s=0}^{t}(1 - variance(s)) \end{aligned} \tag{6}$$

To predict the parameters of the Gaussians, we use a neural network to model it. The UNet is used since it has strong representation capabilities that can capture local changes and texture variations in images. Incorporating the FFT mechanism within CNN improves the generalisation in image inpainting and speeds up the computational speed of convolution operations, improving CNN efficiency. Hence, FFT-UNet is used, whose structure is depicted below:

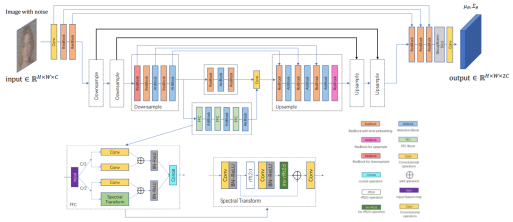

Figure 3. The architecture of the FFT-UNet [2]

2.3. SmartBrush: Text and Shape Guided Image Inpainting

In this section, we consider the task of multi-modal object inpainting conditioned on a text description and the object's shape to be inpainted. We explore diffusion models for this task inspired by their superior performance in modelling complex image distributions and generating high-quality images [4]. The preliminary stuff involves making a diffusion model, which we have discussed in the previous sections. The further sections discuss the approach used in SmartBrush.

In the previous section, we used a similar forward process strategy:

$$\tilde{x}_t = \mathcal{N}(\sqrt{\alpha_t}x_0, \sqrt{1-\alpha_t}\epsilon_0) \tag{7}$$

The inputted mask m generates $\{m_s\}_{s=1}^{N}$ where m_s denotes masks of varying shape-precision. For each mask, we also have a class label c_s. With this, we can write:

$$x_t = \tilde{x}_t \odot m + x_0 \odot (1 - m) \tag{8}$$

This ensures that the generated objects in the foreground instead of the mask (m) are consistent with the background

Now to predict noise ϵ from the noisy x_t, we train a network ϵ_θ:

$$\mathcal{L}_{DM} = \mathbb{E}_{t\ [1,T],x_0,\epsilon_t} ||\epsilon_t - \epsilon_\theta(x_t, t)||^2 \tag{9}$$

To describe more about the shape-precision control (m_s, $s\ [0, S]$)

$$m_s = \text{GaussianBlur}(m, k_s, \sigma_s) \tag{10}$$

where k_s is Gaussian kernal size and σ_s is standard deviation of the kernel.

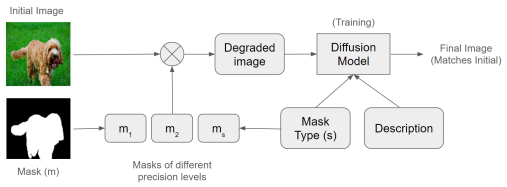

Figure 4. Block diagram for text and shape guided object inpainting [4]

For background preservation, we train our diffusion model to predict an accurate mask m_s:

$$\mathcal{L}_{prediction} = H(\epsilon_\theta(m_s), m) \tag{11}$$

This helps us predict where the object goes in the foreground of our image, and we can thus generate the background to preserve it.

Final Training Strategy:

$$\mathcal{L}_{total} = \mathcal{L}_{seg-DM} + \lambda \mathcal{L}_{prediction} \tag{12}$$

2.4. Survey: Diffusion Models for Image Inpainting

This survey describes different diffusion model-based methods successfully performing high-quality, diverse image inpainting.

2.4.1 RePaint: Inpainting using DDPM

Step 1: A deep neural network roughly estimates the missing pixels. The neural network is trained on a large dataset. This produces a noisy image with the gaps filled roughly.
Step 2: The DDPM takes a rough estimate from the deep neural network and iteratively refines it [3]

2.4.2 Spatial Diffusion Model (SDM)

The proposed method makes use of a Markov random field. It uses a deep neural network to roughly generate missing pixels, and then the SDM refines this image using the Markov random field [3]. This model can handle large missing regions, preserving the original image's texture and structure by considering the relationship between pixels to generate the missing ones. It is versatile and also computationally very efficient.

2.4.3 Structure Guided Diffusion Model (SGDM)

The proposed method utilises a UNet (as seen earlier). Thus, it can generate multiple variations of the same image when the mask area is large. The diffusion models used are called the structure generator and the texture generator. The structure generator generates the edge image and guides the texture generator to refine the inpaint further efficiently [3]. However, there are two disadvantages. The first one is that the inpaint may be faulty, given the mask area the model deals with is large. The second disadvantage is that it is computationally heavy, and specific processes require even days to complete.

2.4.4 Latent Diffusion Models (LDM)

Unlike previous works, this diffusion model takes advantage of image-specific inductive biases, which rely on autoregressive attention-based transformers for operating in discrete latent space. This method uses compressional models along with UNet. The high-resolution images are generated on a low-dimensional latent space with low complexity [3]. Hence, compression is not required, and the quality of the image is preserved. Thus, this model outperforms the previous model significantly for the same task, as shown in Table 2.

3. Results

We see an example of an image reconstructed using our first PDE model in section 2.1 in Figure 5. The input image shown in Figure 5 (b) is noisy and missing a region of pixels compared to the original image in Figure 5 (a). The image in Figure 5 (c) has the noise removed, and the image in Figure 5 (d) is reconstructed.

For our second model (FFT-UNet), we see two examples. Figure 6 shows us the reconstruction when the mask region is small. The model can recover a very similar image to the original one. However, Figure 8 shows us that when the mask region is large, the recovered image differs from the original in some details.

An example of the third model (SmartBrush) is shown in Figure 8. We have given the mask shown in the figure and the text to be 'teddy bear'. We observe that the inpainted image from the SmartBrush model is much better than the others.

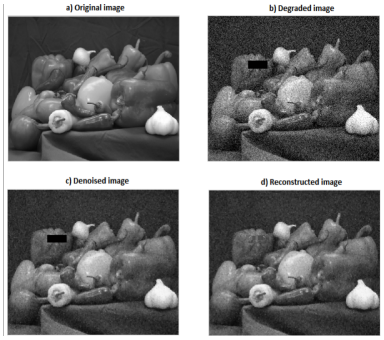

Figure 5. Example of image inpainting using PDE model [1]

Figure 6. Example of image inpainting using FFT-UNet model with a small mask [2]

Figure 7. Example of image inpainting using FFT-UNet model with a large mask [2]

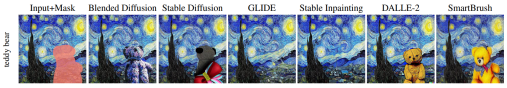

Figure 8. Comparison of inpainted images using SmartBrushand other models [2]

We will now see in Table 1 that SmartBrush outperforms other models in every aspect. The FID score (lower the better) denotes the similarity between the output and the original image. The Clip score (higher the better) denotes the correlation between the output and the original image.

Model	Local FID	Clip Score	FID
Blended Diffusion	29.16	0.265	11.05
Blended Diffusion	22.45	0.252	9.70
Blended Diffusion	15.28	0.265	9.10
Blended Diffusion	12.57	0.264	7.07
SmartBrush	**9.71**	**0.266**	**6.80**

Table 1. Comparing the outputs of various models with SmartBrush [4]

In the survey, we have seen four diffusion models for image inpainting. On comparing the performances and outputs of each model, as shown in Table 2, we see that LDMs perform significantly better, as stated earlier.

Model	FID score	LPIPS score
RePaint	6.98	0.060
SDM	4.05	0.052
SGDM	4.68	0.057
LDM (big, w/ ft)	1.50	0.137
LDM (big, w/o ft)	2.40	0.142
LDM (w/ attn)	2.15	0.144
LDM (w/o attn)	2.37	0.146

Table 2. Comparison of various diffusion models in the survey [3]

4. Conclusions

Several kinds of diffusion models can be used for image inpainting. The PDE-based one utilises that the denoising and reconstruction utilise partial differential equations-based mathematical models, which are easy to solve and discretise. The FFT-UNet model utilises FFT as it speeds up the convolutional operations of the CNN involved. This model, however, has the limitation of the large missing patch, where the inpainted image might not be generated as well as the original one, lacking some details, as seen in the figure. The SmartBrush is a much more improvised version of the text and shape-guided image inpainting by creating different levels of masks. Finally, the survey on various diffusion models analysed various models and concluded that the LDM methods outperform others, as seen in Table 2. From the above results, one can easily decide the best model as per their requirement (speed, efficiency, diversity or similarity).

References

[1] Tudor Barbu, Adrian Ciobanu, and Mihaela Luca. Pde-based image restoration using variational denoising and inpainting models. In *2014 18th International Conference on System Theory, Control and Computing (ICSTCC)*, pages 688–691, 2014. 1, 2, 4

[2] Yuxuan Hu, Hanting Wang, Cong Jin, Bo Li, and Chunwei Tian. A diffusion model with a fft for image inpainting. *JOURNAL OF Cyber-Physical-Social Intelligence*, 1:60–69, 2022. 1, 2, 4

[3] Sibam Parida, Vignesh Srinivas, Bhavishya Jain, Rajesh Naik, and Neeraj Rao. Survey on diverse image inpainting using diffusion models. In *2023 2nd International Conference on Paradigm Shifts in Communications Embedded Systems, Machine Learning and Signal Processing (PCEMS)*, pages 1–5, 2023. 1, 3, 4

[4] Shaoan Xie, Zhifei Zhang, Zhe Lin, Tobias Hinz, and Kun Zhang. Smartbrush: Text and shape guided object inpainting with diffusion model. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22428–22437, 2023. 1, 2, 3, 4