2023 2nd International Conference on Paradigm Shifts in Communications Embedded Systems, Machine Learning and Signal Processing (PCEMS) | 979-8-3503-1071-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/PCEMS58491.2023.10136091

Survey on Diverse Image Inpainting using Diffusion Models

Sibam Parida
VNIT
Nagpur, India
paridasibam@gmail.com

Vignesh Srinivas
VNIT
Nagpur, India
vigneshmadhipatla2001@gmail.com

Bhavishya Jain
VNIT
Nagpur, India

Rajesh Naik
VNIT
Nagpur, India

Neeraj Rao
Prof. ECE,
VNIT
Nagpur, India
neerajrao@ece.vnit.ac.in

***Abstract*—Image inpainting (or Image completion) is the process of reconstructing lost or corrupted parts of images. It can be used to fill in missing or corrupted parts of an image, such as removing an object from an image, removing image noise, or restoring an old photograph. The goal is to generate new pixels that are consistent with the surrounding area and make the image look as if the missing or corrupted parts were never there. Image inpainting can be done using various techniques such as texture synthesis, patch-based methods, and deep learning models. Deep learning-based Image inpainting typically involves using a neural network to generate new pixels to fill the missing parts of an image. Different network architectures can be used for this purpose, including Convolutional Neural Networks (CNNs), Generative Adversarial Networks(GANs), Transformer-based models, Flow-based models, and Diffusion models. In this work, we focus on Image Inpainting using Diffusion models whose task is to provide a set of diverse and realistic inpainted images for a given deteriorated image. Diffusion models use a diffusion process to fill in missing pixels, where the missing pixels are iteratively updated based on the surrounding context. The diffusion process is controlled by a set of parameters, which can be learned from data. The advantage of diffusion models is that they can handle large missing regions, while still producing visually plausible results. The challenges involved in the training of these models will be discussed.**

***Index Terms*—Inpainting, Diverse Inpainting, Deep Learning, Diffusion Models.**

I. Introduction

Image inpainting is an active area for research in many fields such as computer vision and applied mathematics. Image inpainting refers to the virtual reconstruction of missing content in images imperceptible to the observer. It is an inverse problem which has multiple plausible solutions. The non-uniqueness nature of this problem can be understood mathematically and also because the reconstruction is judged independently by humans. There are multiple applications of inpainting in real life including, artefact removal [1], image compression [2], preservation of old vintage film and video [3] [4], art conservation [5] and manuscripts [6] and medical imaging [7] [8].

State-of-the-art methods for image inpainting have produced astounding results when it comes to the difficult task of filling in significant missing sections in an image. However, most approaches for image inpainting use specific masks and do not generalise well on different types of masks. In this paper, we focus on analysing recent advancements in image inpainting with a focus on diffusion model-based methods. Diffusion models generalise better when compared to other methods. This is because diffusion models are probabilistic generative models which do not require mask-specific training and produce high-fidelity outputs.

979-8-3503-1071-9/23/$31.00 ©2023 IEEE

The paper is structured as follows. Section 2 provides a brief overview of the diffusion model. Section 3 presents the underlying theory of the several approaches and the review of the state-of-the-art proposals using those particular strategies. Section 4 presents the evaluation of those approaches using particular metrics. Finally, Section 5 concludes the presented analysis.

II. A walk through Denoising Diffusion Probabilistic Models

Diffusion models are a class of likelihood-based models which have recently been shown to produce high-quality images while offering desirable properties such as distribution coverage, a stationary training objective, and easy scalability. These models generate samples by gradually removing noise from a signal, and their training objective can be expressed as a re-weighted variational lower bound. This class of models already holds the state-of-the-art on several datasets. [9] proposes improvements to the original diffusion model [10] which is used in our model.

Diffusion models sample from a distribution by reversing the gradual noising process. In particular sampling starts with noise X_T produces less noisy samples gradually X_{T-1},X_{T-2}.. till X_0. Each timestep t corresponds to the noise level at that step and X_t can be considered as a mixture of signal X_0 and some noise.

A diffusion model learns to produce a slightly more denoised sample from the previous one. [10] parameterised this model as a noise predictor function, which predicts the noise component from a sample X_t. To train the model each sample in a mini-batch is produced by randomly drawing a data sample X_0, time step t and noise e to form noised sample X_0. The training objective is simple MSE loss between true

and predicted noise.

Since sampling directly from the noise predictor is difficult. Under suitable assumptions the distribution of X_{t-1} given X_t can be modelled as a diagonal Gaussian. The mean of Gaussian is modelled as a function of the noise predictor. [10] considers the variance as a known constant, which yields results only when the number of diffusion steps T is large enough. [9] model the variance as interpolation learnt by neural networks, which uses the reverse diffusion process's upper and lower bound variances. The training objective used is a weighted sum of the simple MSE loss and The lower bound variance loss. This allows sampling from fewer steps without a drop in quality.

We have studied various generative models for image inpainting, diffusion models having lower chances of mode collapse and with the improved fidelity for generation works favourably. The diffusion model is trained to sample distribution from noise, the assumption made here is since the diffusion model can generate samples from noise, given an image with missing information it would generalise well and produce several realistic outputs regardless of the mask used.

III. Multiple and Diverse Inpainting using Diffusion Models

In this section, we will describe different diffusion model-based methods that successfully perform high-quality diverse image inpainting.

RePaint: Inpainting using Denoising Diffusion Probabilistic Models [11]

The authors [11] propose a method to perform high-quality diverse image inpainting that is applicable to even extreme masks. The model employs a pre-trained unconditional Denoising Diffusion Probabilistic Model [9] as the generative prior. Figure 1 displays an overview of the network architecture.

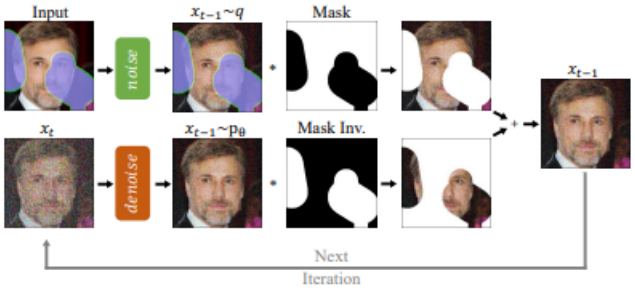

Fig. 1. RePaint architecture

The Authors proposed a two-step Approach. First, it uses a deep neural network to generate a rough estimate of the missing pixels. The network is trained on a large dataset of similar images and uses a combination of convolutional and deconvolutional layers to learn the relationship between the surrounding pixels and the missing pixels. This initial estimate provides a starting point for the second step of the algorithm, which uses the denoising diffusion probabilistic model to refine the prediction.

In the second step, the denoising diffusion probabilistic model takes the rough estimate from the deep neural network and iteratively refines it. The model uses a Markov random field to model the dependencies between the pixels in the image and the surrounding context. The model predicts the most likely values for the missing pixels based on the surrounding pixels and then updates its prediction based on the results of the previous iteration. The algorithm repeats this process until the prediction converges into a final solution.

One of the key benefits of RePaint is its ability to handle a wide range of inpainting scenarios. The algorithm can handle images with missing pixels, cracks, holes, and other types of damage. It can also handle textured surfaces, such as hair, fur, and grass, and preserve the texture in the inpainted region. RePaint has also been shown to outperform other state-of-the-art inpainting algorithms in terms of both accuracy and speed.

SDM: Spatial Diffusion Model for Large Hole Image Inpainting [12]

The authors [12] propose a method for inpainting large missing regions while preserving the texture and structure of the image. The proposed method makes use of a Markov random field for the inpainting of the region. Figure 3 displays an overview of the network architecture.

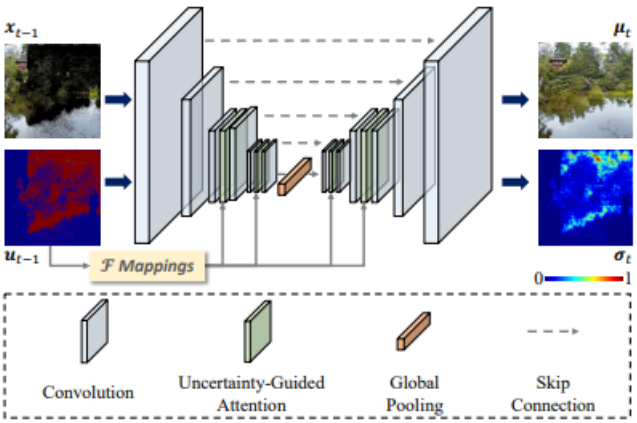

Fig. 2. SDM architecture

It uses a deep neural network to generate a rough estimate of the missing pixels. This is then refined using a spatial diffusion model, which takes into account the surrounding context and structure of the image. The spatial diffusion

model uses a Markov random field to model the dependencies between the pixels and the surrounding context. This means that it takes into account the relationships between the pixels, such as the intensity and colour, and uses this information to predict the most likely values for the missing pixels.

One of the key advantages of SDM is its ability to handle large missing regions in an image. Unlike some other inpainting algorithms, which struggle to handle large missing regions and produce blurry results, SDM can fill in large missing regions while preserving the texture and structure of the original image. This makes it a useful tool for repairing damaged images or removing objects from images.

Another advantage of SDM is its computational efficiency. The algorithm can fill in large missing regions in real time, making it suitable for use in interactive applications, such as photo editing software. Additionally, it has been shown to outperform other state-of-the-art inpainting algorithms in terms of accuracy and speed.

SDM is also versatile and can handle a wide range of image textures, including hair, fur, and grass. This is important, as these textures can be challenging for inpainting algorithms to preserve, and a failure to do so can result in an unnatural-looking inpainted region.

One of the key challenges in implementing SDM is training the deep neural network used to generate the initial estimate of the missing pixels. The network needs to be trained on a large dataset of similar images in order to learn the relationship between the surrounding pixels and the missing pixels. This requires significant computational resources and can take several days to complete.

There are also some limitations to the SDM algorithm. For example, it may not perform as well on images with highly complex textures, such as abstract images, or images with many small, fine details. In these cases, the algorithm may struggle to preserve the fine details of the original image.

Structure-Guided Diffusion Model for Large-Hole Diverse Image Completion [13]

The authors [13] propose a method for generating multiple variations of an image with large missing content. The proposed method utilises a mask for edges along with training two diffusion models jointly for structure and texture preservation. Figure 3 displays an overview of the network architecture.

The proposed method utilises the input image and its corresponding edge image, both images are masked. The diffusion models are called structure generator and texture generator both utilising a modified UNet [21] [14] [15] architecture. The structure generator is used to generate the edge image which guides the texture generator for more

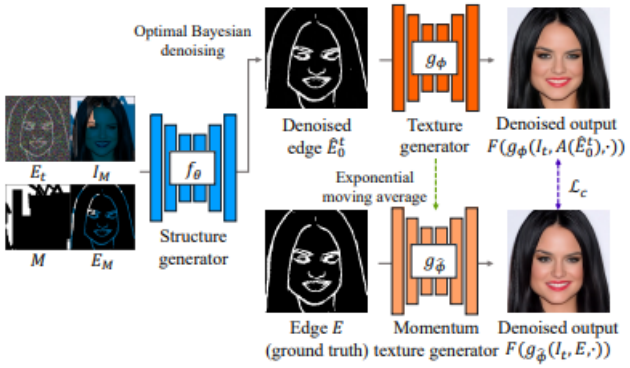

Fig. 3. Structure guided diffusion model

semantically efficient inpainting.

A major drawback is depending on the output of the structure generator faulty inpainting variations may be generated. To mitigate this problem a joint training method is utilised which applies Bayesian denoising and a momentum framework [16] [17]. The momentum framework comprises two networks: one is the regularly trained network and the other is the momentum network where weight is updated as the estimated moving average of the first network.

The proposed method uses an algorithm that erases regions randomly from augmentations to prevent corruption and improve generalisation. Further, they also utilise several noises to make it generalise well over large datasets and distributions.

The disadvantages of the proposed method are it fails to generate images that have adequately closed edges. The computational cost required is large as it uses two diffusion models which are an iterative process as compared to GANs [18] [19] which is a single-step process. Acceleration sampling methods can be used to reduce the amount of time taken. Overall the method produces images that have better generalisation with the structure, texture and colour gradient.

High-Resolution Image Synthesis with Latent Diffusion Models [20]

The authors [20] propose a method that improvises on the current diffusion model for several tasks one of which is image inpainting. The main idea behind the proposed methodology is that the diffusion model works on an image on the pixel space and the proposed method will shift it to vector space to improve computational costs and time required. Figure 4 displays an overview of the network architecture.

The proposed method utilises compressional models along with UNet [21] architecture to perform high-resolution image synthesis on a low dimensional latent space which has low complexity. This makes the diffusion model particularly

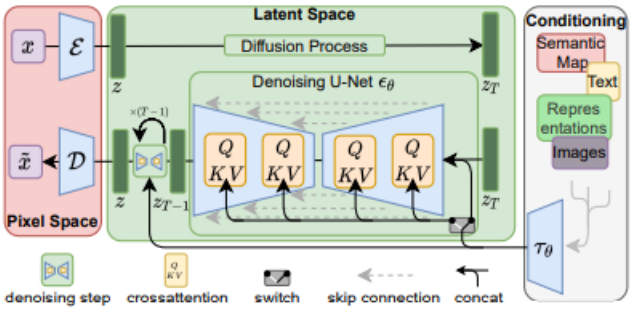

Fig. 4. Latent diffusion model

effective with spatial structure and reduces the need for quality reducing high compression levels unlike previous methods. [22] [23]

The compression model is based on an autoencoder which is trained on a combination of perceptual loss [22] [24] and patch-based [25] adversarial objective [22] [26] [27]. This enforces local realism and avoids blurriness which helps in realistic reconstruction. To avoid high variance latent spaces they test two different kinds of regularisations, Kullback-Liebler [28] [29] and Vector Quantisation [30].

The diffusion model takes advantage of image-specific inductive biases unlike the previous works which relied on autoregressive attention-based transformers for operating in discrete latent space [22] [23] [27]. The underlying UNet architecture is built using 2D convolutional layers using the encoder and decoder of the perceptual compression autoencoder. This model is conditioned for generalising well beyond class labels using cross-attention mechanism [31].

For the purpose of inpainting, the model uses a specialised architecture relying on Fast Fourier Convolutions [32] based on LaMa [33]. The model outperforms all known methods in terms of image generation. While the latent diffusion models reduce computational requirements the sampling process is much slower than methods such as GANs.

A. Qualitative Evaluation

All the above models have been evaluated for inpainting on the CelebA-HQ [34] dataset (512x512) using two evaluation metrics and the results are shown in the table below. The Latent diffusion model was evaluated with four configurations using Fourier convolutions and attention mechanisms. The evaluation metrics used are the Frechet inception distance (FID) [35] score and the Learned Perceptual Image Patch Similarity (LPIPS score) [36]. The lower the value of each of these scores, the better the quality of the results generated by the model.

CelebaHQ		
Methodology	***FID score***	***LPIPS score***
RePaint	6.98	0.060
SDM	4.05	0.052
SDGM	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 I
FID AND **LPIPS** SCORES OF THE ABOVE MODELS. THE LOWER THE SCORE, THE BETTER THE MODEL. THE LATENT DIFFUSION MODEL OUTPERFORMS ALL THE OTHER MODELS.

CONCLUSION

In this paper, we have seen various diffusion models for multiple and diverse image inpainting. We have compared the methods on a public dataset across two metrics. We have shown that the latent diffusion model provides better inpainting quality qualitatively. The analysis highlights it due to the strategy of working in the low dimensional latent space which results in better performance. Moreover, we also see that all the methods suffer from a few drawbacks and the inpainting problem is not solved yet. Therefore, we argue more efforts should be made on improving and exploring new strategies to enhance the quality and diversity of this ill-posed inverse problem.

REFERENCES

[1] Vitoria and Ballester, 2019. Vitoria, P. and Ballester, C. (2019). Automatic flare spot artefact detection and removal in photographs. Journal of Mathematical Imaging and Vision, 61(4):515– 533.
[2] Peter and Weickert, 2015. Peter, P. and Weickert, J. (2015). Compressing images with diffusion and exemplar-based inpainting. In Lecture Notes in Computer Science, pages 154–165. Springer International Publishing
[3] Grossauer, 2006. Grossauer, H. (2006). Inpainting of movies using optical flow. In Mathematics in industry, pages 151–162. Springer Berlin Heidelberg.
[4] Newson et al., 2014. Newson, A., Almansa, A., Fradet, M., Gousseau, Y., and Pérez, P. (2014). Video inpainting of complex scenes. SIAM Journal on Imaging Sciences, 7(4):1993–2019.
[5] Baatz et al., 2008. Baatz, W., Fornasier, M., Markowich, P. A., and bibiane Schönlieb, C. (2008). Inpainting of ancient Austrian frescoes. In Conference Proceedings of Bridges, pages 150–156.
[6] Calatroni et al., 2018. Calatroni, L., d'Autume, M., Hocking, R., Panayotova, S., Parisotto, S., Ricciardi, P., and Schönlieb, C.-B. (2018). Unveiling the invisible: mathematical methods for restoring and interpreting illuminated manuscripts. Heritage Science
[7] Tovey et al., 2019. Tovey, R., Benning, M., Brune, C., Lagerwerf, M. J., Collins, S. M., Leary, R. K., Midgley, P. A., and Schönlieb, C.-B. (2019). Directional sinogram inpainting for limited angle tomography. Inverse Problems
[8] Chen et al., 2012. Chen, Y., Li, Y., Guo, H., Hu, Y., Luo, L., Yin, X., Gu, J., and Toumoulin, C. (2012). CT metal artefact reduction method based on improved image segmentation and sinogram in-painting. Mathematical Problems in Engineering, 2012:1–18.
[9] A. Nichol and P. Dhariwal," Diffusion Models Beat GANs on Image Synthesis," in NIPS,2021
[10] J. Ho, A. Jain and P. Abbeel," Denoising Diffusion Probabilistic Models," in NIPS,2020.
[11] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. RePaint: Inpainting Using Denoising Diffusion Probabilistic Models. In CVPR, 2022
[12] Li, W., Yu, X., Zhou, K., Song, Y., Lin, Z., Jia, J. (2022). SDM: Spatial Diffusion Model for Large Hole Image Inpainting. In CVPR,2022

[13] Horita, D., Yang, J., Chen, D., Koyama, Y., Aizawa, K. (2022). A Structure-Guided Diffusion Model for Large-Hole Diverse Image Completion.In CVPR,2022

[14] 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

[15] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Image Super-Resolution via Iterative Refinement. arXiv preprint arXiv:2104.07636, 2021.

[16] Jean-Bastien Grill, Florian Strub, Florent Altche, Corentin ´ Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Moham- ´ mad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Remi ´ Munos, and Michal Valko. Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning. In NeurIPS, 2020

[17] Antti Tarvainen and Harri Valpola. Weight-averaged consistency targets improve semi-supervised deep learning results. In NeurIPS, 2017.

[18] Qiaole Dong, Chenjie Cao, and Yanwei Fu. Incremental Transformer Structure Enhanced Image Inpainting with Masking Positional Encoding. In CVPR, 2022

[19] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Qureshi, and Mehran Ebrahimi. EdgeConnect: Structure-Guided Image Inpainting using Edge Prediction. In ICCVW, 2019

[20] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-resolution image syn- ¨ thesis with latent diffusion models. In CVPR, pages 10684– 10695, 2022.

[21] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. Unet: Convolutional networks for biomedical image segmentation. In MICCAI (3), volume 9351 of Lecture Notes in Computer Science pages 234–241. Springer, 2015.

[22] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming ¨ transformers for high-resolution image synthesis. CoRR, abs/2012.09841, 2020

[23] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. CoRR, abs/2102.12092, 2021.

[24] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018.

[25] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In CVPR, pages 5967–5976. IEEE Computer Society, 2017.

[26] Alexey Dosovitskiy and Thomas Brox. Generating images with perceptual similarity metrics based on deep networks. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett, editors, Adv. Neural Inform. Process. Syst., pages 658–666, 2016.

[27] Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. Vector-quantized image modelling with improved vqgan, 2021.

[28] Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In 2nd International Conference on Learning Representations, ICLR, 2014.

[29] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In Proceedings of the 31st International Conference on International Conference on Machine Learning, ICML, 2014.

[30] Aaron van den Oord, Oriol Vinyals, and Koray ¨ Kavukcuoglu. Neural discrete representation learning. In NIPS, pages 6306–6315, 2017.

[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In NIPS, pages 5998–6008, 2017.

[32] Lu Chi, Borui Jiang, and Yadong Mu. Fast Fourier convolution. In NeurIPS, 2020.

[33] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor S. Lempitsky. Resolution-robust large mask inpainting with Fourier convolutions. ArXiv, abs/2109.07161, 2021.

[34] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In ICLR, 2018.

[35] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, Gunter Klambauer, and Sepp Hochreiter. GANs ¨ Trained by a Two Time-Scale Update Rule Converge to a Nash Equilibrium. In NeurIPS, 2017.

[36] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In CVPR, 2018.