# Dense Pixel Transformer GAN for Image Inpainting

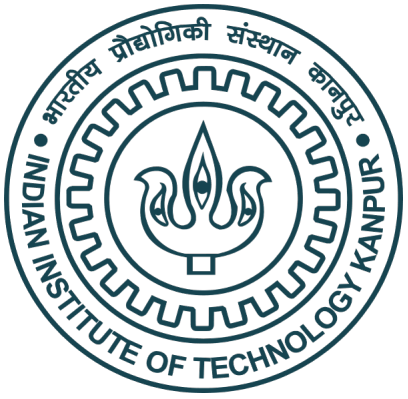
*A thesis submitted in partial fulfilment of the requirements*

*for the degree of Master of Science (By Research)*

*by*

Harshvardhan Pratap Singh

20111410




DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY KANPUR

November 2023

**November 2023**

# Certificate

We certify that the work contained in this thesis on **"Dense Pixel Transformer GAN for Image Inpainting"** by **Harshvardhan Pratap Singh** has been carried out under our supervision and that it has not been submitted elsewhere for a degree.

Gaurav Sharma
Visiting Assistant Professor
CSE Department
Indian Institute of Technology Kanpur

Manindra Agrawal
Professor
CSE Department
Indian Institute of Technology Kanpur

November 2023

# Declaration

This is to certify that the thesis titled **Dense Pixel Transformer GAN for Image Inpainting** has been authored by me. It presents the research conducted by me under the supervision of **Dr. Gaurav Sharma** and **Prof. Manindra Agrawal**.

To the best of my knowledge, it is an original work, both in terms of research content and narrative, and has not been submitted elsewhere, in part or in full, for a degree. Further, due credit has been attributed to the relevant state-of-the-art and collaborations (if any) with appropriate citations and acknowledgments.

Harshvardhan Pratap Singh
Roll No. 20111410
MS-Research, CSE Department
Indian Institute of Technology Kanpur

# *Abstract*

---

Name of the student: **Harshvardhan Pratap Singh** Roll No: **20111410**
Degree for which submitted: **MS by Research** Department: **CSE**
Thesis title: **Dense Pixel Transformer GAN for Image Inpainting**
Thesis supervisors: **Dr. Gaurav Sharma** and **Prof. Manindra Agrawal**
Month and year of thesis submission: **November 2023**

---

Image inpainting, also called image completion or hole-filling, is a fundamental task in computer vision that is utilized in real-world image editing scenarios. Imagine having a picture where someone is standing in front of the Taj Mahal, but there is a part of the image occupied by people or objects, and you want to erase them seamlessly as if they never existed there. When elements are removed from an image, they leave gaps or holes. The challenge lies in determining what should go in those missing pixels so that the result appears authentic. Algorithms need the ability to comprehend the surrounding context to introduce content that blends naturally into the scene intelligently.

This thesis introduces a novel deep-learning method for image inpainting. Our approach incorporates vital components, including Generative Adversarial Networks (GANs), the Vision Transformer, and a U-Net-based convolutional network. The procedure consists of sending a masked image as input to a generator and processing it through a transformer and a U-Net. This results in the generation of an inpainted image as output. Subsequently, this output is fed into a discriminator to receive feedback.

Our generator employs a dual-stage method. The initial step involves using the Dense Vision Transformer to establish the images' foundational structure and texture. The subsequent stage employs a U-Net-based refinement network to enhance pixel-level details, increasing the realism of generated images. To ensure smooth communication between different feature map levels, we combine U-Net's convolutional layers and the vision transformer within the generator. This integration is achieved through skip connections that link the two stages.

The refinement network utilizes high-level information from the Transformer's output feature maps for image enhancement. This process is guided by integrating skip connections from various resolutions originating from the Transformer's fusion blocks.

Earlier approaches encountered challenges in maintaining smooth transitions in generating local pixels, occasionally resulting in blurred textures. This occurred due to feature discontinuity and a limited understanding of semantic information. We opt for the Dense Vision Transformer over the standard Vision Transformer due to its capability to process overlapping patches continuously and densely. This attribute leads to more realistic outcomes than the Vision Transformer's method of handling non-overlapping patches. The latter can result in less coherent results, potentially failing to capture context smoothly. Conversely, the Dense Vision Transformer benefits from shared information among neighbouring patches, facilitating a seamless context integration.

We conduct a thorough empirical evaluation of our approach, utilizing well-known datasets like Paris StreetView (street level images taken in various neighborhoods throughout Paris), CelebA (images of celebrities' faces, showcasing a wide range of facial features and attributes), and FFHQ (variety of high-resolution human faces). This evaluation occurs in real-world conditions, enabling meaningful comparisons with existing methods. Our approach consistently achieves competitive performance in contrast to these methods, producing visually coherent images.

# *Acknowledgements*

My mentor, Dr. Gaurav Sharma from the Department of Computer Science and Engineering, has my highest appreciation. My thesis path would not have been possible without his constant advice. My initial ideas were carefully guided to fruition by Dr. Gaurav, who also inspired me to have the patience and drive to keep pushing my limits.

My sincere thanks go out to Prof. Manindra Agrawal as well, whose steadfast support and guidance have served as a continual source of motivation. I am incredibly grateful for the blessings from God and my parent's continuous support during my journey.

I wish to extend my special thanks to Professor Mainak, the former Head of the Department of Computer Science and Engineering at IIT Kanpur, for providing me with the invaluable support to make a meaningful contribution to this thesis.

**Harshvardhan Pratap Singh**

# Contents

# List of Figures

# List of Tables

*Dedicated to family, friends and teachers.*

# Chapter 1

# Introduction

Image inpainting is a fundamental problem in computer vision that involves filling or painting in missing or damaged regions of an image with plausible content. Image inpainting is challenging, specifically when dealing with large missing parts and high-resolution images. The goal is to complete the picture so the result appears natural and consistent with the surrounding pixels. It is akin to training a virtual artist capable of painting the missing parts realistically, ensuring that the final output looks genuine and coherent. Image inpainting is a challenging task that requires the network to comprehend the image's context and structure to generate a realistic and coherent result. With deep learning techniques, researchers are making remarkable advancements in this field, which has crucial applications in image editing, restoration and synthesis. Inpainting has diverse real-world applications, from restoring old destroyed images to taking off unwanted objects and inpainting with natural-looking scenes. Image inpainting can enhance images' visual and semantic quality, improving image realism. Applications of image inpainting are convenient in diverse fields, such as art entertainment.

Earlier or traditional approaches, such as patch based and exemplar based methods, have been proposed for this task. Patch based approaches in the inpainting process fail when working with large missing areas and complex structures because they use basic methods based on finding similar-looking patches in the surrounding area. These techniques could be more efficient for handling sizeable missing area images.

Deep learning-based approaches show huge success in image inpainting and video inpainting tasks. Convolutional Neural Networks, Generative Adversarial Networks, Transformers, and Diffusion are some methods researchers use extensively in this field. However, these methods have limitations based on use cases, such as GANs suffering from mode collapse, producing blurry outputs and exhibiting unstable training. The earliest deep learning-based approach was proposed by Pathak et al. (2016) [14] for image inpainting using a context-based encoder that utilizes the

surrounding context of the image to generate image patches. Promising results were shown in face images, but this method had limitations for complex structures and large holes. Recently, GAN-based approaches have demonstrated high-quality results. The generator learns to generate realistic samples, while the discriminator learns to distinguish between real and fake samples. Yu et al. (2018) [7] proposes an end-to-end GAN-based approach for image inpainting that shows promising results on various datasets, including the CelebA dataset.

Image inpainting is problematic because it demands the network replace the absent pixels in a visually convincing and contextually coherent manner with the neighbouring areas. To tackle this challenge, one possible solution is presented by Liu et al. (2018) [10] to utilize partial convolutional layers. These layers facilitate the network's ability to update only the relevant pixels in the missing region and retain the information in the surrounding pixels.

Attention-based approaches, such as those based on the Transformer model, have also been proposed for image inpainting. Wang et al. (2020) [15] have proposed a deep cascaded refinement network with attention, in which a transformer-based attention module followed a series of U-Net based refinement stages. Li et al. [16] have proposed an image inpainting method that used a transformer to model the long-range dependencies in the image and achieved state-of-the-art results on several benchmark datasets.

Vision Transformer(ViT) based models have recently shown impressive results in various computer vision tasks, including image classification and segmentation. Yuan et al. (2021) [17] have proposed a transformer-based generative adversarial network for image inpainting that leverages ViT and convolutions in the generator backbone to generate high-resolution images with large holes. However, their model required post-processing steps to refine the generated images.

Image inpainting is the task of predicting dense pixel values. Recently, Image inpainting methods have utilized end-to-end learning of deep neural networks to acquire semantics and useful hidden representations. These methods aim to learn a mapping from the visible parts of the image to the missing parts, ensuring that the inpainted regions are consistent with the rest of the image and appear natural. Initially, encoder-decoder-based models were proposed, but one popular class of deep learning models that transformed image painting were generative adversarial networks (GANs)[18, 19, 20, 10, 7]. Two essential components are at play: the generator and the discriminator networks. These two entities work harmoniously, undergoing simultaneous training through an adversarial process. The primary aim of the generator is to craft images that are so realistic that they could easily be mistaken for genuine ones. In contrast, the discriminator faces the challenge of distinguishing between these authentic images and those generated by the GAN.

Through an iterative training process, GANs can simulate a game-like setting to learn how to generate various realistic images. Recent work has explored various modifications and extensions of the GAN architecture for image inpainting.

To address the limitations of GANs and previous transformer-based models, this thesis proposes a novel image inpainting method that leverages a dense pixel transformer-based generative adversarial network and a U-Net-based refinement network. The model consists of two stages: first, a dense pixel transformer generates the missing regions of the image using dense pixel transformers, and second, a U-Net-based refinement network is used to improve image quality. The skip connections between both stages allow the refinement network to use high-level features from the dense pixel transformer output to generate visually plausible and semantically correct images. During training, a patch-based discriminator is used to distinguish between real and fake images.

Our proposed method achieves competitive performance on benchmark datasets such as Paris StreetView [4], CelebA [3], and FFHQ [5], demonstrating its promising potential in the field of image painting. Our approach uses a dense vision transformer (DenseViT) as the generator in a GAN. DenseViT is a type of transformer-based architecture that has recently attained popularity in the computer vision community, particularly for tasks involving large images or video sequences. Unlike traditional convolutions, which operate on local patches of an image, transformers capture global dependencies and are better suited to modelling long-range relationships in images.

Some approaches to image inpainting often take advantage of post-processing techniques such as poisson blending to improve output quality. However, transformer our proposed method incorporates a refinement network within the inpainting process, utilizing a U-Net-based convolutional network and guiding skip connections from the vision transformer fusion step. Doing so allows us to generate refined output directly from our method without post-processing techniques.

In this thesis, we have also discussed training strategy and implementation details in the Experiments chapter. Qualitative and Quantitative results are shown for the method's effectiveness. The approach's potential for image inpainting can be seen visually and quantitatively.

## 1.1 Contribution

The main contributions of this thesis are as follows:

- We propose a novel architecture, a Dense Pixel Transformer Generative adversarial network and Conv-U-Net based refinement network.

- We show that skip connection between Transformer generator and U-Net-based refinement network helps feature sharing.
- We perform experimentation with different loss functions, and show that our method performs well with limited resources and even on low-resolution images.

## 1.2 Thesis Outline

The structure of the thesis is as follows:

**Chapter 2** We start by exploring image inpainting, covering both traditional and modern deep learning methods.

**Chapter 3** This chapter introduces our approach and architecture for image inpainting. We explain the different parts we use to tackle this challenge.

**Chapter 4** In this chapter, we share the results of our experiments. We have tested our methods on popular datasets like CelebA, Paris StreetView, and FFHQ. We talk about how we used our techniques and our unique architecture. We also dive into the technical details of how we trained our model and analyze different settings and what we discovered.

**Chapter 5** We wrap up the thesis by summarizing what we have learned from our research and discuss future directions for this field.

# Chapter 2

# Related work

## 2.1 History of Image inpainting

Researchers have been working on the image inpainting problem before the Deep Learning era. Image inpainting has shown significant progress from traditional techniques [21, 22, 23, 24] to modern deep learning-based approaches like Transformers [1, 2, 25, 16] or Diffusion [26, 27, 28, 29] based techniques, with representational improvement on quality and accuracy of image inpainting and video inpainting tasks. Various techniques are proposed that have been researched for decades; here, we will discuss subset. Recent methods can generate nearly natural and compatible with the surrounding final image, which can work as a backbone for multiple tasks. The initial methods for image inpainting relied on simple techniques like interpolation, nearest neighbour, and texture creation. These techniques frequently produced unrealistic or aesthetically undesirable results and had limitations when handling sophisticated inpainting assignments.

## 2.2 Non-learning based Approaches

Early works have introduced patch based procedures, texture synthesis based methods and exemplar based techniques. While dealing with complex holes, these techniques fail. The capacity to synthesise material from related patches in the images led to the adoption of patch based techniques, which first appeared in the middle of the 2000s. A texture creation technique based on locating related patches in the image and blending them to fill in the missing areas was proposed by [23]. The patch based synthesis technique is used to fill in the missing region using patches selected from relevant areas of the image. However, this method may produce output with apparent seams due to the patch selection problem. Various strategies have

been developed to address this, including those based on exemplar based inpainting, like those in the work of [22], which chose patches using the nearest neighbour method. The fact that patch selection could have been better and the outcomes were frequently unsightly remained a drawback of these methods.

## 2.3 Modern Deep Learning based Approaches

Repair and completion of missing or damaged parts in images is now feasible thanks to advancements achieved in recent years in deep learning-based methods [25, 16, 27, 28, 29] for image inpainting. Below is a summary of popular deep learning-based image inpainting methods. Deep learning has emerged as a potent method for image inpainting in recent years. In one of the early attempts in this discipline, convolutional neural networks (CNNs) were employed to predict the missing pixels in the image. The technique provided incredible results, but the outcomes could have been more precise. Later studies explored the use of generative adversarial networks (GANs) for image inpainting to address this issue. Recent works in computer vision have extended the Transformer model, incorporating attention techniques into inpainting tasks. Initially designed for natural language processing, transformers have proven highly effective in image inpainting and other computer vision-related tasks.

### 2.3.1 CNN based Approaches

Image inpainting has switched to data-driven machine learning techniques with deep learning. Because they could learn to provide convincing information depending on the surrounding context, convolutional neural networks (CNNs) have become very popular for image inpainting applications. In recent years, convolutional neural networks (CNNs) and generative adversarial networks (GANs) have revolutionised image inpainting. Image Inpainting effectiveness with deep generative models employing CNNs and GANs for inpainting irregular holes in images was demonstrated in [10]. In the work of Pathak et al., [14], they introduced context encoders, one of the pioneering deep learning methods for image inpainting. These context encoders leverage Convolutional Neural Networks (CNNs) to understand how to transform an image with missing parts into a complete image. The context encoder consists of two key components: a generator and a discriminator. The generator is a straightforward encoder-decoder network. The remarkable capability of the context encoder lies in its ability to generate structurally coherent content. It achieves this by learning to paint image patches based on their surrounding context. This approach, rooted in deep learning techniques, produces high-quality image patches. This method produced good results but needed help to deal with large holes and complex architecture.

Recent developments in deep learning have shown encouraging results when tackling the challenging task of filling in blank spaces in images. However, due to the fragmented nature of local pixels, present approaches usually produce results with fuzzy textures and warped structures. The leading cause of this local pixel inconsistency is the disregard for feature coherence and semantic relevance within the regions that need to be filled. To solve this issue, Liu et al. [12] examine how people repair images and present a better deep generative model-based technique that is strengthened by a special coherent semantic attention (CSA) layer. By adding semantic significance to the attributes of the hole areas, this layer improves the accuracy of detecting missing components while maintaining contextual structure.

Rough estimation and refining are the two stages in that this method divides the task. An embedded neural network within the U-Net framework addresses each stage. Within the refining phase's encoder, the CSA layer is located. They incorporate a consistency loss to ensure network stability during training and encourage acquiring more effective parameters inside the CSA layer. This loss concurrently puts the VGG feature layer of a ground truth image close to the CSA layer and the equivalent CSA layer within the decoder.

In conclusion, CNN-based models for image inpainting have shown promising outcomes, particularly when combined with GANs and attentional processes. These techniques could revolutionise the field of image inpainting and are crucial for image editing, restoration, and synthesis.

### 2.3.2 GAN based Approaches

Goodfellow et al.(2014) [18] have first introduced Generative Adversarial Networks (GANs), which swiftly gained popularity as a method. By competitively training a generator and discriminator, GANs could produce realistic and cohesive images. Standalone CNN-based methods were facing issues such as fuzzy output. To solve this problem, the use of generative adversarial networks (GANs) for image inpainting was explored in later research, and found to have the potential of providing sharper and more realistic output. To employ GANs for image inpainting, two neural networks, a generator and a discriminator, must be trained. While the generator network inputs an image with missing portions and produces a final image, the discriminator network aims to discern between fake and real photos. The generator is trained to create images that the discriminator cannot tell apart from real photos, while the discriminator is trained to discern between real and produced images accurately. This process results in the generator producing good-quality photos that bear remarkable similarity to the real images.

One of the first attempts in this area was by Yu et al. [7]. An end to end GAN based image inpainting technique showed promising results on several datasets, including the CelebA dataset. Their method managed the inconsistent masks and generated high-quality images using a partial convolutional neural network (PCNN). The PCNN learns which areas of the image to focus on and which to ignore during the training phase by adaptively changing the convolutional mask.

Iizuka et al. (2017) [19] proposed a GAN-based method that utilised a free-form mask to fill in the image's blank spaces. Their method included a contextual encoder to predict missing regions and a local discriminator to ensure the visual plausibility of the generated images, producing high-quality images. The authors also recommended a post-processing method that could improve the final images.

Yang et al. (2017) [20] proposed a GAN-based method that utilised a dilated convolutional neural network to produce realistic images. They used a dilated convolutional neural network to encode the image and build the missing part. Additionally, the researchers proposed a new reconstruction loss that generates high-quality images while accommodating varying sizes of the missing region.

A partial convolution layer is employed in the first stage of Yu et al. (2018) [7], a two stage GAN design for image inpainting to selectively update the missing region, and the output is enhanced in the second stage. Similarly, Liu et al. (2018) [10] proposed a GAN based method that improved the output's visual quality by incorporating perceptual loss and adversarial loss.

Liu et al. (2018) [10] also proposed a GAN-based method that integrated perceptual and adversarial loss to improve the output's visual quality. They proposed a multi-scale contextual attention technique that can generate high-quality image patches and record the long-range dependencies of the image. The researchers also revealed a promising training technique that can simultaneously handle numerous masks to reduce training time.

Xie et al. (2019) [30] introduced a potent attention-guided technique to image inpainting, utilising principles for learnable spatial attention. By accurately capturing the links between context and target regions, this unique method pushed the limits of image inpainting techniques and improves the quality and realism of inpainted images.

These GAN-based methods have revolutionised the field of image inpainting and synthesis. There are still problems to be overcome in managing several masks, producing high-quality images with detailed textures and structures, and improving the efficiency of the training process. However, with more study and development, GAN-based approaches could provide an effective tool for image inpainting and other comparable applications.

### 2.3.3 Transformer based Approaches

Transformers, originally designed for NLP applications, have now found utility in computer vision tasks as well, where they have been adapted to excel in image recognition. Dosovitskiy et al. [1] use the input image's fixed-size, non-overlapping patches as "words" in the Transformer's context by breaking it into such units. With this modification's help, the Transformer can now manage images as collections of patches, providing global context awareness. A vital element of the Transformer architecture, multi-head self-attention, is used by the Vision Transformer(ViT) [1] to capture interactions between various patches. Thanks to this attention technique, the model may concentrate on pertinent patches while processing the image. They are further extended for dense pixel prediction problems like monocular depth estimation, and Semantic Segmentation named Dense Prediction Transformer(DPT) [2]. DPT introduces an architecture which combines the power of CNN and ViT for dense pixel prediction tasks. CNNs excel at managing local details, while ViTs are skilled at gathering global context and semantic information. The researchers hope to improve dense prediction task performance by integrating these two designs. Transformer models have integrated attention techniques in the image inpainting. Image inpainting is an intricate, poorly understood inverse problem that naturally leaves space for various content to fill gaps or damaged areas convincingly. Convolutional neural networks (CNNs) are used in standard approaches to create visually appealing content. Nevertheless, because of their constrained sensory fields, CNNs struggle with a limited ability to understand global aspects. Convolutional neural networks (CNNs), recognised for their process in texture modelling, have significantly accelerated the development of image completion. CNNs encounter difficulties in grasping overall structures and supporting pluralistic completion, mainly because of their inherent properties like local inductive priors and spatially invariant kernels. CNNs have limitations in comprehending global structures in data and often result in result in multiple possible interpretations due to their reliance on local patterns and fixed spatial features.

Transformers can generate various content through autoregressive modelling of pixel-sequence distributions thanks to their ability to model significant dependencies across an image at the picture level. However, the autoregressive transformers [31] unidirectional attention method has drawbacks, especially when dealing with arbitrarily shaped distorted image regions with contexts coming from any direction. Transformers have recently shown off their talent for grasping distant relationships and producing various results. Though their computing cost increases quadratically with input length making managing high-resolution photos difficult. This study fills the gap between CNN and Transformer strengths in image completion.

BAT-Fill [31] is a novel framework for image inpainting. This method's innovative Bidirectional Autoregressive Transformer (BAT), explicitly created for image inpainting, is at its core. BAT uses a transformer architecture to learn autoregressive distributions, which naturally facilitates the creation of a variety of content to fill in the gaps. A masked language model similar to BERT is also incorporated, allowing for bidirectional contextual information modelling of the blank sections for better image completion.

Wan et al. [25] propose a dual strategy that uses CNNs for texture replenishment and transformers for appearance before reconstruction. The former uses transformers to recover coarse textures and cohesive structures pluralistically. The latter employs CNNs to improve local texture features of the coarse priors while being led by high-resolution masked images. The suggested approach is exceptional in several ways. It dramatically improves image fidelity and even outperforms deterministic completion techniques. Additionally, it improves diversity and faithfulness for the completeness of pluralism. Furthermore, its performance on large masks and generic datasets like ImageNet demonstrates its outstanding generalisation ability.

Existing strategies use solo attention techniques or transformers to tackle the problem of enormous hole painting. Due to computational constraints, MAT [16] provides a brand-new transformer-based model explicitly designed for large-hole inpainting in the current work. This novel method efficiently processes high-resolution images by combining the advantages of transformers and convolutions. The framework parts are carefully planned to guarantee the authenticity and variety of the recovered images. MAT mainly offers an inpainting-focused transformer block where the attention module only accumulates non-local information from tokens that are only partially valid, as indicated by a dynamic mask.

Recently, transformer-based techniques have grown in popularity because they can get beyond these limitations and deliver good results on benchmark datasets. These methods use transformers to characterise the image's long-range relationships and produce high-quality inpainting outputs. These methods have shown excellent performance on benchmark datasets and can advance the area of image inpainting. They can create long-range dependencies and fully capture the context of the image, leading to more realistic and aesthetically pleasing results. It will be interesting to see how these techniques advance and how they might be adjusted to deal with practical applications for image inpainting.

### 2.3.4 Diffusion Based Approaches

Modelling the image as a diffusion process and filling in the missing pixels by solving a partial differential equation are the two steps involved in diffusion-based approaches for image inpainting. These techniques have been used with remarkable effectiveness to fill in missing areas of photographs. Bertalmio et al. (2000) [21] proposal was one of the first efforts in this field, and it utilised a diffusion-based method to fill in image gaps. A partial differential equation that represented the diffusion process in the image had to be solved as part of the procedure. Additionally, the authors suggested a patch based restriction to guarantee that the resulting image matched the pixels in its surroundings.

The Navier-Stokes equation-based strategy put forth by Bertalmio et al. (2001) [24] was another effort in this field. The process entailed simulating the image as a fluid and using the Navier-Stokes equation to fill in the omitted areas. The authors also suggested a texture restriction that ensured the generated image's texture matched its surroundings' texture. While diffusion based methods have occasionally proved successful, they frequently need help with substantial image holes and complex structures. Results from these techniques typically lack more delicate details and appear fuzzy. Additionally, they demand manual parameter selection, which can be laborious and difficult for non-experts. A diffusion based strategy that polishes the outcomes using a deep neural network. The distribution of the missing image patches is learned by the generative adversarial network (GAN), which is then used to produce accurate findings.

A novel method for generative modelling utilising diffusion processes is presented in the paper "Denoising Diffusion Probabilistic Models" [26]. This study suggests a novel probabilistic model that models the diffusion of noise applied to the data to learn the data distribution. The essential concept is to learn several conditional distributions that progressively change the noisy observations into the desired data distribution. The shift from noise-corrupted data to clean data is modelled using a diffusion process by the Denoising Diffusion Probabilistic Model (DDPM) [26]. A sequence of phases is used to simulate this process, with each step involving transforming the existing noisy data in the direction of the desired data distribution. The model optimises the parameters that specify these transformations. The capacity of DDPM to denoise data throughout the creative process is one of its noteworthy features. The model produces denoised samples by starting with noisy observations and gradually improving them through diffusion. This contrasts with conventional generative models, which frequently produce hazy images since they directly sample from a latent space. The model's features are theoretically explained in the paper, and testing on diverse datasets shows the model to be effective. It demonstrates how

DDPM outperforms numerous cutting-edge generative models on tasks like image generation and inpainting and produces high-quality samples with fine details.

In conclusion, "Denoising Diffusion Probabilistic Models" [26] proposes a novel method for generative modelling by utilising diffusion processes to convert skewed observations into accurate samples of data gradually. With this method, high-quality photographs with enhanced details and realism are produced.

Diffusion models (DMs) have achieved outstanding achievements in image synthesis and related disciplines by decomposing the image production process into a sequential application of denoising autoencoders. Without the requirement for retraining, these models provide a guiding mechanism for directing picture production. However, because powerful DMs operate directly in pixel space, training them is computationally heavy and requires a lot of GPU resources. Additionally, it might be show because sequential evaluations are required during the inference process.

A new strategy is presented [27] to solve these issues and enable DM training with constrained computational resources while maintaining quality and flexibility. In this method, DMs are applied to the latent space of already trained autoencoders. This innovative approach compromises between the preservation of detail and complexity reduction, significantly improving visual authenticity. Diffusion models become powerful and flexible generators thanks to the addition of cross-attention layers to the model design. These generators support a variety of conditioning inputs, such as text or bounding boxes, and enable convolutional high-resolution synthesis.

Remarkable results have been obtained using this strategy, known as latent diffusion models (LDMs) [27]. These models demonstrated competitive performance in various tasks, including super-resolution, text-to-image synthesis, and unconditional picture production, while also setting new standards in image inpainting and class-conditional image synthesis. Notably, this progress is made compared to pixel-based DMs while significantly lowering the processing requirements.

Free-form inpainting entails inputting fresh information into a picture when a binary mask has already been applied. However, a lot of existing algorithms are trained for particular mask distributions, which restricts their capacity to handle various and unknown mask types properly. Additionally, when these approaches are trained using pixel-wise and perceptual loss functions, the missing regions are frequently extended with rudimentary textural information and no meaningful semantic generation.

RePaint [28] is a unique inpainting method based on the Denoising Diffusion Probabilistic Model (DDPM). RePaint offers strong generalisation capabilities and is adaptable to extreme masks, unlike earlier techniques. A trained unconditional

DDPM is used as a generative prior in the method. Only the reverse diffusion iterations are altered by sampling unmasked patches using the available picture data to condition the inpainting process. Notably, this method leaves the underlying DDPM network unaltered and unconditioned, allowing the model to provide various high-quality output images for varied inpainting circumstances.

The face and general picture inpainting tasks, and regular and extreme mask circumstances are used to validate RePaint's efficacy. It demonstrates its superiority in inpainting various mask types by outperforming state-of-the-art Autoregressive and GAN-based approaches on at least five of six mask distributions.

Despite their drawbacks, diffusion-based techniques have contributed significantly to the advancement of image inpainting technologies. They have acted as a foundation for more modern approaches and stimulated the creation of fresh solutions that surpass their drawbacks.

# Chapter 3

# Approach and Methods

## 3.1 Model Architecture

This chapter outlines our suggested architecture and methodology. Our architecture design can be seen in Figure 3.3 and is built on Generative Adversarial Networks and Dense Vision Transformer. The generator in our architecture is a two-stage procedure that combines a U-Net based refining network and the Dense Prediction Transformer. The discriminator is patch-based, whereas the generator's two stages are connected via skip connections. The transformer treats the image as a set of tokens, first embedded into feature space and then processed by a cascade of transformer layers to produce updated representations of the tokens. Transformer layers use multi-headed self-attention as their fundamental operation. Multi-headed self-attention is an operation that relates each token to every other in the image and consequently has a global receptive field. The vision transformer does not use down-sampling operations in its intermediate stages and thus supports fine-grained feature maps in deeper layers of the network. To leverage the tokens produced by the vision transformer, we reassemble the tokens at various stages of the backbone into image-like representations. In the Reassemble block, we integrate and then spatially concatenate the tokens and resample the resulting feature maps before passing them to the decoder. In the Fusion block, the decoder iteratively fuses and upsamples the reassembled feature maps from different stages and produces the final dense prediction using an output head. The feature maps from consecutive stages are fused using residual convolutional units followed by bilinear upsampling. Here, we will discuss refinement, transformer, and generative adversarial networks (GANs).

### 3.1.1 Generative Adversarial Network

In 2014, Goodfellow et al. [18] proposed Generative Adversarial Networks (GANs), which derive inspiration from the mini-max algorithm in game theory. GANs are deep learning models that generate data that closely resemble accurate training data. GAN is an implicit generative latent variable model, unlike VAE [32], which has no explicit parametric likelihood model. GANs consist of two two-stage architecture generators and a discriminator. The generator network learns the training data distribution to produce realistic looking data points that follow the actual data distribution. In contrast, the discriminator network is generally a classification network which distinguishes between real and fake data points. During training, the generator and discriminator are set up in a competitive scenario, similar to a game of strategy, where they work against each other to improve their performance and compete with each other in a zero-sum game where the generator tries to generate a data sample identical to accurate data and fool the discriminator by producing realistic images. The discriminator tries to distinguish between the real and generated images. Both network weights get updated in a supervised manner through backpropagation. The generator tries to be better at generating realistic data points close to training data, and the discriminator tries to be a better investigator. Throughout this process, the aim is to reach a point where the discriminator needs help distinguishing between real and fake data.

In our case, the generator is a two-stage Dense Prediction transformer followed by a U-Net [33] based convolutional network, and the discriminator is patch-based. Instead of classifying entire images, a patch-based discriminator [34] is trained to classify small image patches as either natural or fake. This method allows the discriminator to focus on local details and identify subtle differences between real and fake patches, resulting in more realistic-looking generated images. The discriminator network takes natural and generated images as input and produces a probability score for each image, indicating its likelihood of being honest. The generator network is trained to maximize the probability of classifying its generated images as accurate, while the discriminator network is trained to classify the authentic and generated images accurately. This GAN-based method has proven effective in generating diverse and high-quality images across various domains. They have also been used in image inpainting, which involves generating missing parts of an image based on its surrounding pixels. Utilizing a discriminator that classifies image patches as real or fake, these models can produce realistic-looking patches that blend well with the original image.

GANs have proven they can produce realistic and cohesive images by competing in the training of a generator and a discriminator. Traditional CNN-based techniques needed help producing results with precise details in particular. In order to address

this issue, the following research led to the application of generative adversarial networks (GANs) for image inpainting, which has the potential to provide results with improved sharpness and realism.

To use GANs for image inpainting, we train two networks: a generator that generates images and a discriminator that judges their authenticity. The discriminator's role is to tell apart fake and real photos, whereas the generator takes an input image with gaps and creates a finished image. The key to this method is training the generator to produce images that the discriminator finds impossible to separate from actual photographs. Ultimately, due to this interplay, the generator creates images of good quality that nearly resemble the qualities of the source images.

With the introduction of GAN-based approaches, the field of image inpainting has undergone a radical change with ramifications for image editing, restoration, and synthesis. Although these techniques have shown to be transformative, difficulties still exist in handling several masks properly, producing high-quality images with rich textures and structures, and improving the effectiveness of the training process. Nevertheless, GAN-based techniques have the potential to be an effective tool for image inpainting and related applications with further study and improvement.

### 3.1.2 Transformer

Since their launch in 2017, Transformers has significantly contributed to Natural Language Processing (NLP) by handling sequential data. Transformers were first introduced in the 2017 study "Attention is All You Need" by Vaswani et al. [35]. Transformers have become state-of-the-art in numerous disciplines, including NLP, speech synthesis, and recognition.

Transformer is another deep neural Network architecture based on the attention mechanism. The attention mechanism is first introduced by Bahdanau et al. [36]. Attention is developed for long-phrase memorization in Neural machine translation(NMT). Vaswani et al. [35] have introduced multi-headed self-attention. Self-attention helped the transformer model semantic features to produce meaningful generations and concentrate on relevant things while making predictions. Transformer can also consume data at once and not sequentially, creating embedding or representation for the data.

Image inpainting is a dense pixel prediction task which requires high-quality image generation for each pixel. Our method uses a Dense Prediction Transformer, which uses a multiple-vision transformer as its backbone. Vision Transformer (ViT) was proposed by Dosovitskiy et al. in 2021 [1].

DenseViT [2] and ViT [1] differ in their approach in processing input images. ViT divides the image into non-overlapping patches individually processed by the transformer encoder. In contrast, DenseViT processes overlapping patches continuously and densely. In our method, we have used ResNet50 to flatten image patches to be processed by a dense vision transformer. This strategy captures more spatial information and minimizes artefacts that commonly appear in the output of ViT.

#### 3.1.2.1 Self Attention

Self Attention allows transformer to establish relationships between different input parts and capture long-range dependencies, which in our case is the global context of the image, which was a shortcoming of previous convolutional approaches. They are only able to get a local receptive field and to get a better receptive field, the network has to be very deep, which is computationally expensive and prone to overfit because of the number of parameters. Attention defines the score between the relationship of different parts of the image and helps the model generate meaningful generation semantically.

#### 3.1.2.2 Vision Transformer

Several researchers have proposed transformer variants for computer vision tasks in the wake of the enormous success of NLP. The most well-known of these is the Vision Transformer (ViT), which is shown in Figure 3.1, a variant of NLP transformer that is first presented in the paper by Dosovitskiy et al. [1]. By segmenting the image into patches and flattening it to convert it into 1D vectors similar to sequence data, the Vision Transformer performs a similar function to the NLP transformer so that self-attention can create relationships between various patches.

The classification head is placed behind a stack of transformer encoder layers in the vision transformer. The transformer layer has a feedforward neural network, residual connections, and a multi-head self-attention mechanism similar to an NLP transformer. ViT multi-head self-attention may consider various picture components at various scales and resolutions. The performance of vision transformers in classification tasks is excellent.

Overall, ViT has demonstrated encouraging results in computer vision tasks and has the potential to enhance the functionality of numerous computer vision applications.

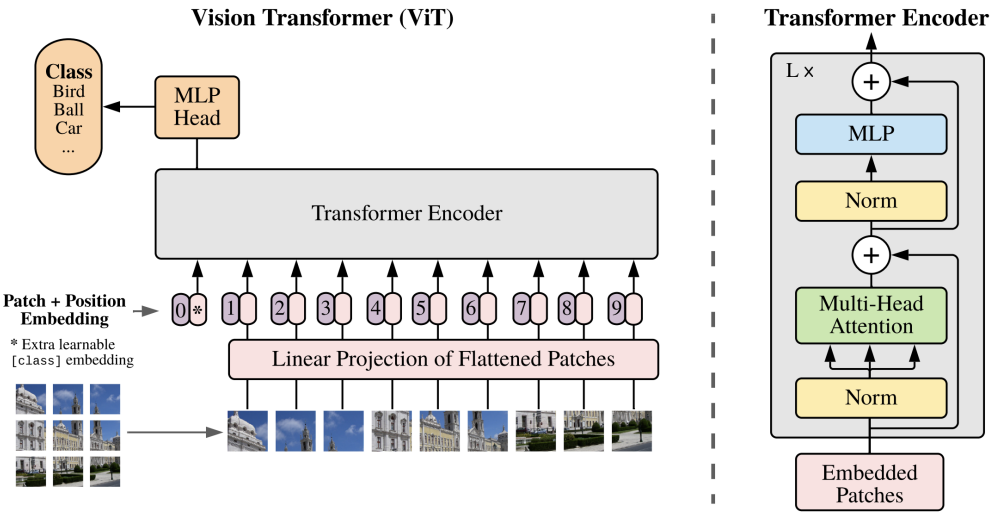



FIGURE 3.1: Network architecture of Vision Transformer(ViT). Figure taken from [1]

## 3.2 Dense Prediction Transformer

Ranftl et al. [2] introduces a new method capable of solving tasks related to assigning pixels or patches in an image, for example, semantic segmentation, image synthesis and generation. A dense Prediction Transformer uses a Vision Transformer(ViT) as a backbone in several stages to extract the features. Different components of DPT are shown in Figure 3.2. ViT based backbone expects embeddings of input from ResNet50 output, and then the output of ViT is a sequence of feature vectors, which are then fed into the Transformer decoder. The Transformer decoder is a convolutional network that combines tokens of different resolutions into image-like representations to generate full-resolution predictions from different stages of the Vision Transformer. The transformer backbone's constant different resolution processing and global receptive field enable the Dense ViT to provide finer-grained and globally coherent predictions compared to fully convolutional networks. The Transformer decoder uses a fusion module to combine different-resolution images into the final image. From this fusion module output at different resolutions, skip connections are connected to the refinement network to help it generate high-quality images.

In transformer-based methods in computer vision, image patches are treated as being equivalent to "words" in the context of the transformer architecture when using transformers for image recognition.

It enables the processing of images as collections of patches in order to understand the overall context. In order to capture interactions between these patches and

enable focused processing of pertinent information, the Vision Transformer (ViT) makes use of multi-head self-attention.

The Dense Vision Transformer was developed by extending the transformer concept to dense pixel prediction applications like monocular depth estimation and semantic segmentation. The advantages of ViTs in capturing global context and CNNs in handling local details are combined in this architecture. Performance in dense prediction jobs is anticipated to increase with the integration of these design components. Additionally, image inpainting has been incorporated into transformer model attention mechanisms.

Transformers are becoming more popular because they can work around the limitations of convolutional methods and do very well empirically. When we use transformers for inpainting, they can create images that focus on fine details, improving the results. Their good performance on standard datasets suggests they can make inpainting better. These methods can understand the whole picture and produce accurate results.

Especially when working with big masks, enabling efficient interactions across distant contexts is essential for completing high-quality images. Deep or widely receptive field (RF) convolutions have been used. However, they frequently find themselves confined by neighbouring interactions, which may not produce the intended results. It sees image completion as a challenge of directionless sequence-to-sequence prediction. Our research presents a new methodology. In the initial phase, it uses a transformer to directly understand long-range dependencies within the encoder. Transformers accurately consider distant connections in all its layers without unintentionally blending nearby elements. Using larger receptive fields can sometimes lead to this unintended mixing of nearby elements.

In the following step, we use an U-Net based architecture, which we change according to our needs and made more profound, which is included in the generator to improve the coherence between visible and generated regions. This layer lessens the isolating impact of conventional attention mechanisms and more effectively utilises distantly related features. In conclusion, extensive experimental findings highlight our method works competitively to cutting-edge techniques across various datasets.

## 3.3 Refinement Network

We propose a refinement network to enhance our image quality on top of the output generated by a Dense Prediction Transformer Generator, which produces blurry outputs with some artefacts. Our refinement network architecture is inspired by U-Net architecture, which is a famous and widely used network for restoration, image

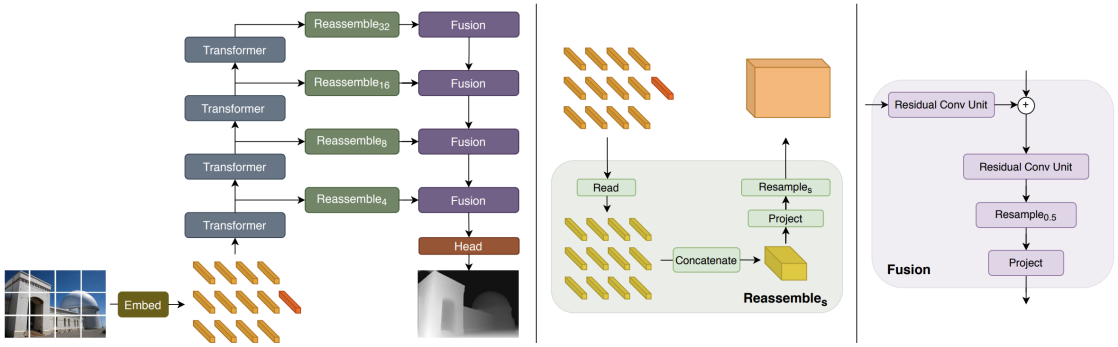



FIGURE 3.2: Architecture overview of Dense Prediction Transformer(DPT). Figure taken from [2]

segmentation and other dense pixel prediction tasks. U-Net is a convolution-based network which consists of a U shaped network in Figure 3.3. On the left is the contracting path, and on the right is the expanding path; the contracting path contains convolutional layers and, going forward, increases channels in the feature map and decreases the spatial resolution of the feature maps. The convolutional layer contain ReLU, followed by the max pool layer, which helps preserve important features and reduce spatial dimensionality. Expanding path works similarly to the Contracting path and contains up-convolutional layers that perform upsampling, and both the paths are connected with skip connections for feature sharing; this setup of the contracting path and expanding path can be seen as an encoder and decoder network that is just expanding path of each up-convolutional layer followed by corresponding feature map concatenation from the contracting path followed by convolutional layer and ReLU activation. The expanding path is tasked with restoring the spatial details of the feature maps without compromising on the high-level features. Also, these skip connections between both paths help do this feature concatenation and enable the network to utilize the high-resolution features obtained from the contracting path, thereby enhancing the quality of the output produced by the expanding path. Our architecture is a slight variation of U-Net, which has more depth in this path, allowing it to take advantage of high-level and low-level features to generate high-quality images. Skip connections ensure the network can recover fine-grained details that may have been lost during the down-sampling process in the contracting path.

The Vision Transformer is highly effective in capturing the semantic details of an image by leveraging global context. Our refinement network, which operates at different levels of the contracting path, incorporates skip connections from the output of the Transformer's fusion blocks at different resolutions. This allows the refinement network to use the high-level features captured by the Transformer's output feature

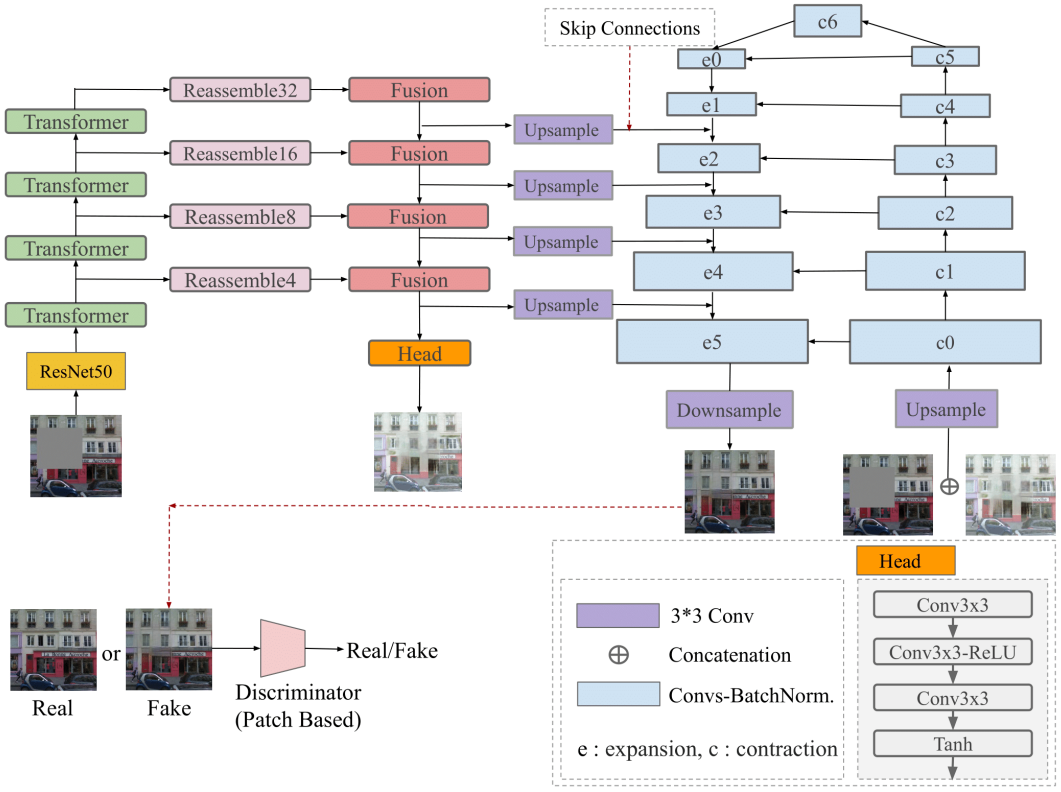



FIGURE 3.3: The proposed architecture of Dense Pixel Transformer GAN for image inpainting, network consists of Transformer, Conv-U-Net, Convolutional upsampler/downsampler and a Convolutional head.

maps to guide the refinement process, resulting in images that are both semantically accurate and visually plausible.

## 3.4 Loss Functions

We use a regression based method during the training phase to match the content of the ground truth image with the generated image. However, we integrate a tripled joint loss function to account for contextual continuity and the potential of several acceptable outputs due to the inherent ambiguity in filling such regions with contextually consistent content.

The perceptual loss crucially preserves the structural coherence of the inpainted image [37, 38] component, ensuring that the created information appropriately matches the surrounding context. The adversarial loss component [18], meanwhile, enhances the inpainted image's visual realism. It directs the generator to produce images that comply with genuine images' structural and statistical properties.

In our model, the role of the reconstruction loss (L1 loss) is to ensure the integrity and cohesiveness of the missing region's structure in its contextual surroundings. Nonetheless, it can sometimes blur distinct but reasonable predictions together. In contrast, the adversarial loss isolates a single convincing outcome from a diverse range of potential results. To produce images with a natural appearance in each iteration, we adjust the loss term by adding both adversarial and perceptual losses.

In both rounds of optimisation inside our system, we apply perceptual loss to improve the quality and diversity of our generation. This contrasts with just using pixel-wise MAE, which frequently results in averaged and fuzzy results. We purposefully use a low coefficient to incorporate the perceptual loss because we have seen that smoother optimisation procedures are made possible by doing so. Here, $x$ is the ground truth image and $\hat{x}$ is generated image

### 3.4.1 Reconstruction Loss

We conducted experiments with pixel-wise L1 and L2 loss functions and found their effects were similar, especially in producing blurry results. We know that L1 loss encourages solution sparsity. This quality becomes especially useful when used in image inpainting, where the goal is to fill in blank spaces realistically. L1 loss generally produces results that are crisper and more visually realistic.

In addition, compared to L2 loss, L1 loss shows stronger resilience to outliers. When outliers pixels with significantly different values occur in the missing region, L1 loss penalises these outliers less severely than L2 loss. This attribute aids in more consistent training and results. The ability of L1 loss to better retain edges than L2 loss is another advantage. L1 loss helps the model capture these edges more accurately because edges are essential for preserving the image's structure and visual coherence.

We chose L1 loss as the reconstruction loss in our method after considering these factors and experimentation.

$$\mathcal{L}_{\text{rec}}(\hat{x}, x) = \sum_{i} |(x_i - \hat{x}_i)| \tag{3.1}$$

### 3.4.2 Adversarial Loss

In our image inpainting methodology, we incorporate adversarial loss as a critical component of our training process. Generative Adversarial Networks (GANs) [18], which consist of a generator and a discriminator network engaged in a competitive

training interplay, are the source of the essence of adversarial loss. We incorporate adversarial loss into our image inpainting technique to enhance the consistency and authenticity of the inpainted results.

In the context of image inpainting, the adversarial loss directs the generator to fill the voids with content that captures both the statistical characteristics of authentic images as well as the structural characteristics of the surrounding context. This forces the generator to create pictures with precise details, finely improved textures, and a general sense of visual authenticity.

Our image inpainting technology generates images of good quality that seamlessly fill in the gaps between the absent regions and the surrounding context while maintaining realism and seamlessness through the combination of adversarial loss. This method has proven remarkably effective in addressing the complexities involved in inpainting situations involving irregular or significant voids within images, leading to results that seamlessly integrate with the more prominent visual narrative.

$$\mathcal{L}_{\mathrm{G}} = -\mathbb{E}_{\hat{x}} \left[\log D(\hat{x})\right] \tag{3.2}$$

$$\mathcal{L}_{\mathrm{D}} = -\mathbb{E}_{x} \left[\log D(x)\right] - \mathbb{E}_{\hat{x}} \left[\log(1 - D(\hat{x}))\right] \tag{3.3}$$

Here, $D$ represents the discriminator, a neural network that assesses and distinguishes between real and generated data in a Generative Adversarial Network (GAN).

### 3.4.3 Perceptual Loss

We make use of the pretrained combination VGG19 [39] network for perceptual loss [37]. The concept behind perceptual loss is that assessing an image's quality is better by looking at its high-level features rather than just comparing individual pixels. To do this, it uses a deep convolutional neural network VGG19, trained on a large dataset of images. This network captures complex textures, patterns, and structural clues in images from several domains.

Perceptual loss measures the difference between the feature profiles of the inpainted image and the corresponding regions in the ground truth image, both collected by VGG19, throughout the training phase. The optimisation of this loss seeks to match the enhanced structural features present in natural images and ensure contextual closure of voids.

We use a pre-trained VGG19, to gather visual cues from simple to complex. It makes images look more visually appealing by adding detailed features and natural textures during the inpainting process.

Perceptual loss and VGG19 help the method in creating inpainted images beyond pixel-level uniformity and capturing the subtle visual themes essential to accurate imaging. The perceptual quality and general authenticity of the results we generate are significantly improved by this method, making them more believable, in tune with, and consistent with the surrounding context.

$$\mathcal{L}_{\text{perceptual}} = \frac{1}{N} \sum_{i=1}^{N} \|\phi_{\text{VGG19}}(x_i) - \phi_{\text{VGG19}}(\hat{x}_i)\|^2 \tag{3.4}$$

where $\phi_{\text{VGG19}}$ denotes feature representations of the image obtained from a VGG19 network.

### 3.4.4 Overall Loss

The overall loss of the model is

$$\mathcal{L}_{\text{total}} = \lambda_{\text{rec}} \cdot \mathcal{L}_{\text{rec}} + \lambda_{\text{adv}} \cdot \mathcal{L}_{\text{adv}} + \lambda_{\text{per}} \cdot \mathcal{L}_{\text{per}} \tag{3.5}$$

In this equation, $\lambda$'s are the hyperparameters that control the influence of the corresponding loss terms.

## 3.5 Training the network

Our network utilizes the Adam optimization algorithm and different learning rates. We also investigated alternative optimization algorithms, including SGD. The reasons for our choices, and the experimental results that influenced our decisions, are discussed in Section 4.2. Additionally, our approach incorporates Generative Adversarial Networks (GANs), which consist of a generator and a discriminator. The generator aims to create images that closely resemble real data, while the discriminator distinguishes between real and fake images in a competitive game like training scenario. Our training process was conducted in multiple stages, comprising four specific phases. A comprehensive outline of this approach can be found in Section 4.2, providing detailed insights into our methodology. During each training iteration, we calculate various loss functions, including L1 loss (for content consistency), adversarial loss (for visual realism), and perceptual loss (for high-level features) to

guide the network's learning. The total loss is a weighted combination of these losses, and the network's parameters are updated accordingly.

The backpropagation of gradients is used to adjust the network's parameters during training, ultimately improving its inpainting capabilities. Our methodology results in high-quality inpainted images that are visually realistic and contextually coherent with their surroundings, even when dealing with challenging inpainting scenarios.

Overall, our training process combines the power of GANs with well-defined loss functions and the efficiency of the Adam optimization algorithm to produce high-quality inpainted images that are visually realistic and contextually consistent with the surrounding regions. This methodology effectively addresses the challenges of inpainting irregular or significant image voids, resulting in seamless and realistic results.

# Chapter 4

# Results and Experiments

We have many ways to measure how well image inpainting works, but we follow previous methods on the respective datasets. These methods have been proven effective and are consistent with what others have done.

## 4.1 Datasets and Metrics

We performed our experiments on three datasets, Paris StreetView [4], CelebA [3], and FFHQ [5] at 256×256 resolution. The Paris StreetView [4] dataset contains images of urban scenes, streets, buildings, and landmarks in Paris, providing a resource for urban and architectural visual data tasks. In contrast, the CelebA [3] dataset primarily comprises facial images of celebrities and individuals, serving as a valuable resource for facial analysis, attribute prediction, and recognition tasks. Lastly, the FFHQ (Flickr-Faces-HQ) [5] dataset is known for its high-quality and high-resolution facial images, offering a diverse range of facial expressions and poses, making it particularly suitable for tasks involving facial image generation and style transfer in computer vision. We adhered to the conventional split of the data into training and testing sets for both Paris StreetView [4] and CelebA [3] datasets, as well as utilizing the metrics outlined in the work of [10, 11, 12], for FFHQ we organized all 70000 images with a split of 65000 train and 5000 test images. In terms of masks, we also chose three different settings to compare with the most relevant papers. One is the irregular mask, which we adopted from the work of [10], which contains a different ratio of masks which is random, another is a centre mask of 128×128, and the third one is 128×128 moving mask across the image to capture the majority of the context through training.

## 4.2 Training Methodology

We compare our method with the following methods:

-CA: Generative Contextual Attention, proposed by Yu et al. [7]

-SH: ShiftNet, proposed by Yan et al. [13]

-PC: Partial Convolution, proposed by Liu et al. [10]

-GC: Gated Convolution, proposed by Yu el al. [11]

-PIC: Pluralistic image completion, proposed by Zheng et al. [8]

-CSA: Coherent Semantic Attention, proposed by Liu et al. [12]

-TFill: High-Fidelity Completion, proposed by Zheng et al.[9]

We conducted our training process in stages, utilizing four distinct phases. In the initial phase, we trained our generator using the entire image and applied pixel-wise loss to the complete image. The subsequent phase involved concentrating on the masked region and applying pixel-wise loss exclusively to that area. In the third stage, our focus shifted to training the discriminator, encompassing both local and global aspects through a patch-based approach. Finally, the last phase encompassed joint training of the entire network within the masked region. This phase incorporated pixel-wise, adversarial, and perceptual loss to enhance the refinement of the output. Our experimentation involved prominent image datasets, such as Paris StreetView, CelebA, and FFHQ, widely recognized in this domain.

We have experimented with different optimization algorithms like SGD [40] and Adam [41] with different learning rates. Adam algorithm is being used for optimization for final models. The reason is that the first and second gradient moments are used by Adam (Adaptive Moment Estimation) to customize learning rates for each parameter. Compared to SGD, this quickens model convergence and minimizes hyperparameter modifications. GANs and Transformers frequently meet sparse gradients, which Adam efficiently manages for consistent updates. SGD, on the other hand, needs to be carefully tuned for sparse gradients. In order to smooth learning and enable automatic momentum adjustment, Adam incorporates exponential moving averages for historical and squared gradients. For the adversarial training of GANs, this stability is essential. Adam's rapid early-stage convergence helps Transformers and GANs, where short gradients are essential. Squared gradients are incorporated into learning rate computation as regularisation to prevent overfitting in complex models.

We have implemented all our models using the PyTorch framework [42] and trained on a single 11GB memory NVIDIA 1080TI GPU On a server Ubuntu 22.04 operating

system. The version of Python used was 3.8.5. The batch size for our experiments was 4, and the resolution of the training and testing images was 256×256. For the training dataset, We set weights learning rate 0.0002 and $\beta_1 = 0.5$

A binary mask $m$ is constructed for the dropped image region for each actual picture $x$. While 0 designates the input pixels, a value of 1 denotes the location of a dropped pixel. These masks are automatically created for each image during training sessions. Irregular masks we used provided by [10].

### 4.2.1 Quantitative Comparisons

TABLE 4.1: Comparison outcomes on Paris StreetView dataset using irregular masks among Partial Conv. [10], Gated Conv. [11], Coherent Semantic Attention [12], and Ours. $^-$ Smaller is better. $^+$ Larger is better

| | **Mask** | **PC [10]** | **GC [11]** | **CSA [12]** | **Ours** |
|---|---|---|---|---|---|
| $L_1^-(\%)$ | 10-20% | 1.47 | 1.14 | 1.05 | 0.94 |
| | 20-30% | 2.12 | 1.71 | 1.41 | 1.22 |
| | 30-40% | 3.49 | 3.19 | 2.69 | 2.17 |
| | 40-50% | 4.58 | 4.49 | 3.70 | 3.23 |
| $L_2^-(\%)$ | 10-20% | 0.17 | 0.14 | 0.08 | 0.05 |
| | 20-30% | 0.28 | 0.22 | 0.13 | 0.11 |
| | 30-40% | 0.60 | 0.57 | 0.45 | 0.38 |
| | 40-50% | 0.86 | 0.90 | 0.68 | 0.59 |
| $PSNR^+$ | 10-20% | 28.91 | 29.58 | 32.67 | 34.19 |
| | 20-30% | 26.78 | 27.43 | 30.32 | 31.00 |
| | 30-40% | 23.27 | 23.19 | 24.85 | 28.65 |
| | 40-50% | 21.67 | 21.33 | 23.10 | 26.70 |
| $SSIM^+$ | 10-20% | 0.937 | 0.945 | 0.972 | 0.984 |
| | 20-30% | 0.894 | 0.920 | 0.951 | 0.964 |
| | 30-40% | 0.815 | 0.846 | 0.873 | 0.892 |
| | 40-50% | 0.678 | 0.731 | 0.768 | 0.840 |

We have thoroughly compared with existing research papers using relevant datasets, highlighting our method's performance. By analyzing various metrics, our method performs competitively, we employ standard evaluation metrics, including L1 and L2 distances, PSNR (Peak Signal-to-Noise Ratio), and SSIM (Structural Similarity Index), to rigorously assess the models' performance. These metrics serve as objective measures, ensuring a comprehensive analysis of the results. Additionally, detailed comparison tables are provided, offering a concise overview of the models' performance across various criteria, facilitating a thorough understanding of their effectiveness, as demonstrated in Table 4.1. In the case of the Paris StreetView dataset [4], we employed the irregular masks technique, resulting in substantial gains, including a 1.52 increase in PSNR. We also tested our method using centre masks to

TABLE 4.2: Comparison outcomes on CelebA dataset using irregular masks among Partial Conv. [10], Gated Conv. [11], Coherent Semantic Attention [12], and Ours. $^-$ Smaller is better. $^+$ Larger is better

| | **Mask** | **PC [10]** | **GC [11]** | **CSA[12]** | **Ours** |
|---|---|---|---|---|---|
| $L_1^-(\%)$ | 10-20% | 1.00 | 1.00 | 0.72 | 0.66 |
| | 20-30% | 1.46 | 1.40 | 0.94 | 0.82 |
| | 30-40% | 2.97 | 2.62 | 2.18 | 1.92 |
| | 40-50% | 4.01 | 3.26 | 2.85 | 2.58 |
| $L_2^-(\%)$ | 10-20% | 0.12 | 0.08 | 0.04 | 0.03 |
| | 20-30% | 0.19 | 0.12 | 0.07 | 0.06 |
| | 30-40% | 0.58 | 0.44 | 0.37 | 0.32 |
| | 40-50% | 0.76 | 0.50 | 0.44 | 0.39 |
| $PSNR^+$ | 10-20% | 31.13 | 31.67 | 34.69 | 35.28 |
| | 20-30% | 29.10 | 29.83 | 32.58 | 33.63 |
| | 30-40% | 23.46 | 24.48 | 25.32 | 29.45 |
| | 40-50% | 22.11 | 23.36 | 24.14 | 27.06 |
| $SSIM^+$ | 10-20% | 0.970 | 0.977 | 0.989 | 0.996 |
| | 20-30% | 0.956 | 0.964 | 0.982 | 0.988 |
| | 30-40% | 0.897 | 0.910 | 0.926 | 0.941 |
| | 40-50% | 0.839 | 0.860 | 0.883 | 0.917 |

TABLE 4.3: Comparison outcomes on CelebA dataset using centering hole among Contextual Attention [7], ShiftNet [13], Coherent Semantic Attention [12], and Ours. $^-$ Smaller is better. $^+$ Larger is better

| | $\mathbf{L_1^-(\%)}$ | $\mathbf{L_2^-(\%)}$ | $\mathbf{SSIM^+}$ | $\mathbf{PSNR^+}$ |
|---|---|---|---|---|
| **CA [7]** | 2.64 | 0.47 | 0.882 | 23.93 |
| **SH [13]** | 1.97 | 0.28 | 0.926 | 26.38 |
| **CSA [12]** | 1.83 | 0.27 | 0.931 | 26.54 |
| **Ours** | 1.77 | 0.27 | 0.942 | 27.12 |

TABLE 4.4: Comparison outcomes on FFHQ dataset using moving square mask among Contextual Attention [7], Pluralistic Completion [8], High-Fidelity Completion(TFill) [9], and Ours. $^-$ Smaller is better. $^+$ Larger is better

| | $\mathbf{L_1^-}$ | $\mathbf{SSIM^+}$ | $\mathbf{PSNR^+}$ |
|---|---|---|---|
| **CA [7]** | 0.0337 | 0.8099 | 22.7745 |
| **PIC[8]** | 0.0241 | 0.8547 | 24.3430 |
| **TFill [9]** | 0.0184 | 0.8778 | 25.2061 |
| **Ours** | 0.0182 | 0.8854 | 25.3647 |

ensure robustness, showcasing its adaptability to different scenarios. While we did not provide a quantitative comparison for the Paris centre setting (since others only explored irregular masks). On CelebA [3], we showcase outcomes for both irregular and central masks in Table 4.2 and Table 4.3 respectively, revealing a significant

performance improvement. Despite resource constraints, we attempted different approaches for the FFHQ dataset, such as bilinear upscaling, which yielded reasonable results even though high-resolution training was not feasible. Our training strategy's inclusion of various mask types ensures a comprehensive exploration of settings, facilitating visual and empirical comparisons and can be seen in Table 4.4.

### 4.2.2 Qualitative Analysis

The results we generated are visually impressive and seamlessly follow the context, as demonstrated by the test images. For instance, in the Paris StreetView dataset, in Figure 4.3 examples, the image in the first row and the second column show how the missing parts of the windows blend well with the surrounding context and structure. Similarly, the gates and balcony structure are realistically generated in the third row and second column, capturing the contextual details. In face dataset FFHQ, as shown in Figure 4.5, various scenarios of missing features are covered, such as one side of the face, one or both eyes and the nose. Our approach generates intricate features that align smoothly with the texture and structure. For example, in the third row and second column, even though both eyes and the nose were almost missing, the generated image includes eyebrows and maintains a natural appearance. Figure 4.6 showcases our approach for post-processing images generated by our model and converting them into high-resolution images. In this context, there are two distinct approaches, the "upscaled" and "restored" settings. In the "upscaled" setting, the process is relatively straightforward. After the model generates an image, it performs a basic upscaling procedure using bilinear interpolation. This method increases the image's dimensions by interpolating between neighbouring pixels, resulting in a larger image. However, it may lack fine details, potentially appearing slightly blurry or pixelated. The primary aim is to boost image resolution without significantly enhancing its quality or realism.

On the other hand, in the "restored" setting, the process is more intricate and involves additional steps to refine the quality of the upscaled image. Following the upscaling of the generated image, a pre-trained GAN (Generative Adversarial Network) model, known as the GFPGAN [6] is deployed for restoration. The GFPGAN [6] is designed to focus on enhancing facial features in upscaled images. This GAN model excels at adding fine details, refining textures, and achieving a more natural and realistic appearance. It can also mitigate any artefacts that may have emerged during the upscaling process. The "restored" setting aims to provide a visually appealing and higher quality output, especially for facial features, surpassing the basic upscaling approach. This method leverages the capabilities of GANs to make the upscaled image resemble one captured at a higher original resolution.

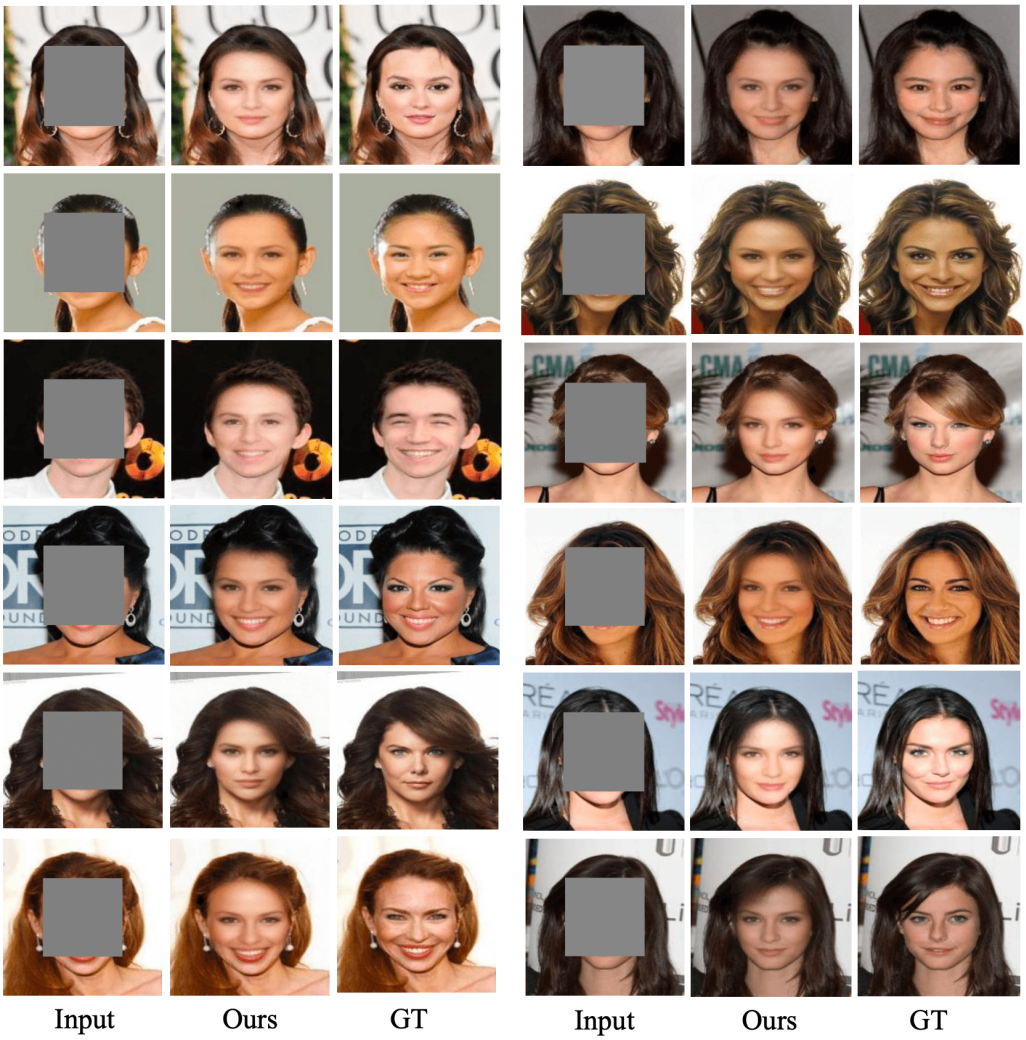



FIGURE 4.1: Qualitative comparisons in centre mask cases on CelebA [3]

The "upscaled" setting primarily enlarges the image size without significant quality enhancement. At the same time, the "restored" setting, in addition to upscaling, uses a specialized GAN model to enhance quality, particularly for facial features, leading to a more visually pleasing and realistic result. This approach proves valuable in situations that demand high-resolution, detailed images, such as applications in facial recognition or image enhancement. GFPGAN [6], a pre-trained GAN model, has been trained on different datasets from CelebA for enhanced image quality.

### 4.2.3 Ablation Study

Our model effectively captures intricate texture details while ensuring pixel consistency with the background. A sole reliance on a transformer based approach tends


FIGURE 4.2: Qualitative comparisons in irregular mask cases on CelebA [3]

to yield blurry outputs. However, integrating skip connections between the Transformer and U-Net at various resolutions significantly enhances texture sharpness, resulting in more realistic details. Strategically placing skip connections from low to high resolutions proves crucial in preventing the loss of intricate information. Low resolution feature maps adeptly capture broad, global context, aiding the model in comprehending overall structures within the inpainting region. Conversely, high-resolution feature maps preserve finer details and local structures, contributing to the faithful reproduction of intricate patterns.

We conducted an ablation study on the FFHQ dataset, systematically removing skip connections. The presented results in Table 4.5 highlight a noticeable drop in performance metrics with each removed skip connection, underscoring their integral role in achieving optimal inpainting results. Despite these findings, it is crucial to note that relying on single skip connections occasionally leads to slightly blurry

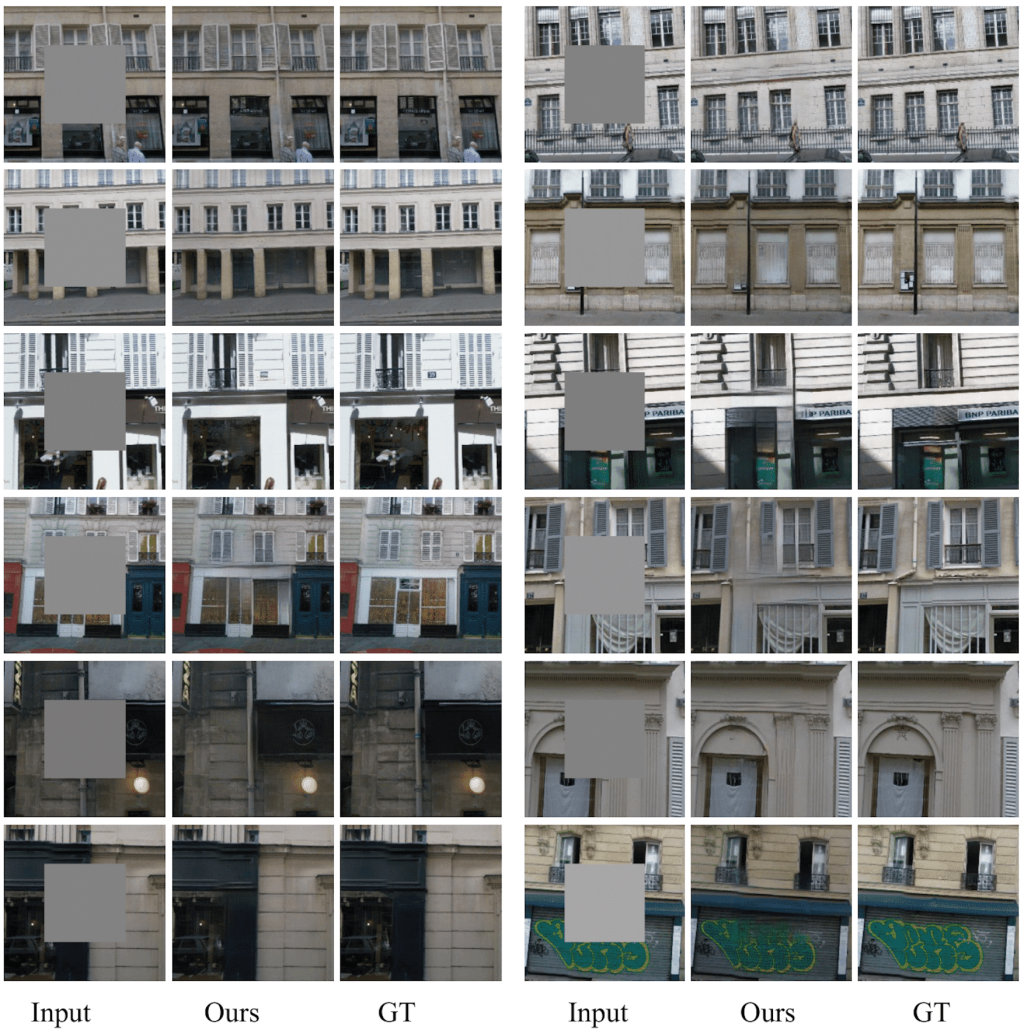



FIGURE 4.3: Qualitative comparisons in centre mask cases on Paris StreetView [4]

outcomes. This underscores the importance of careful design considerations in constructing skip connection architectures for optimal inpainting results. In our case, leveraging four skip connections and a slightly deeper U-Net architecture proved most effective.

In our experiments, we assessed the impact of Perceptual loss on the inpainting model. The addition of Perceptual loss provided enhanced semantic supervision, significantly improving numerical metrics. Excluding Perceptual loss resulted in the generation of box-shaped textures, emphasizing the critical role of semantic supervision in texture synthesis.


FIGURE 4.4: Qualitative comparisons in irregular mask cases on Paris StreetView [4]

TABLE 4.5: Ablation Study on FFHQ dataset on Skip Connections between Transformer and UNet. The base model includes skip connections at different resolutions between Transformer and UNet. Ablation experiments involve removing one by one skip connections individually to assess their impact.. $^{-}$ Smaller is better. $^{+}$ Larger is better

| **Ablations** | $\mathbf{L_1^-}$ | $\mathbf{SSIM^+}$ | $\mathbf{PSNR^+}$ |
|---|---|---|---|
| **0 Skip Connection** | 0.0410 | 0.7223 | 20.8379 |
| **1 Skip Connection** | 0.0329 | 0.8138 | 22.2780 |
| **2 Skip Connection** | 0.0263 | 0.8716 | 24.5674 |
| **3 Skip Connection** | 0.0219 | 0.8744 | 25.0347 |
| **Ours(full)** | 0.0182 | 0.8854 | 25.3647 |

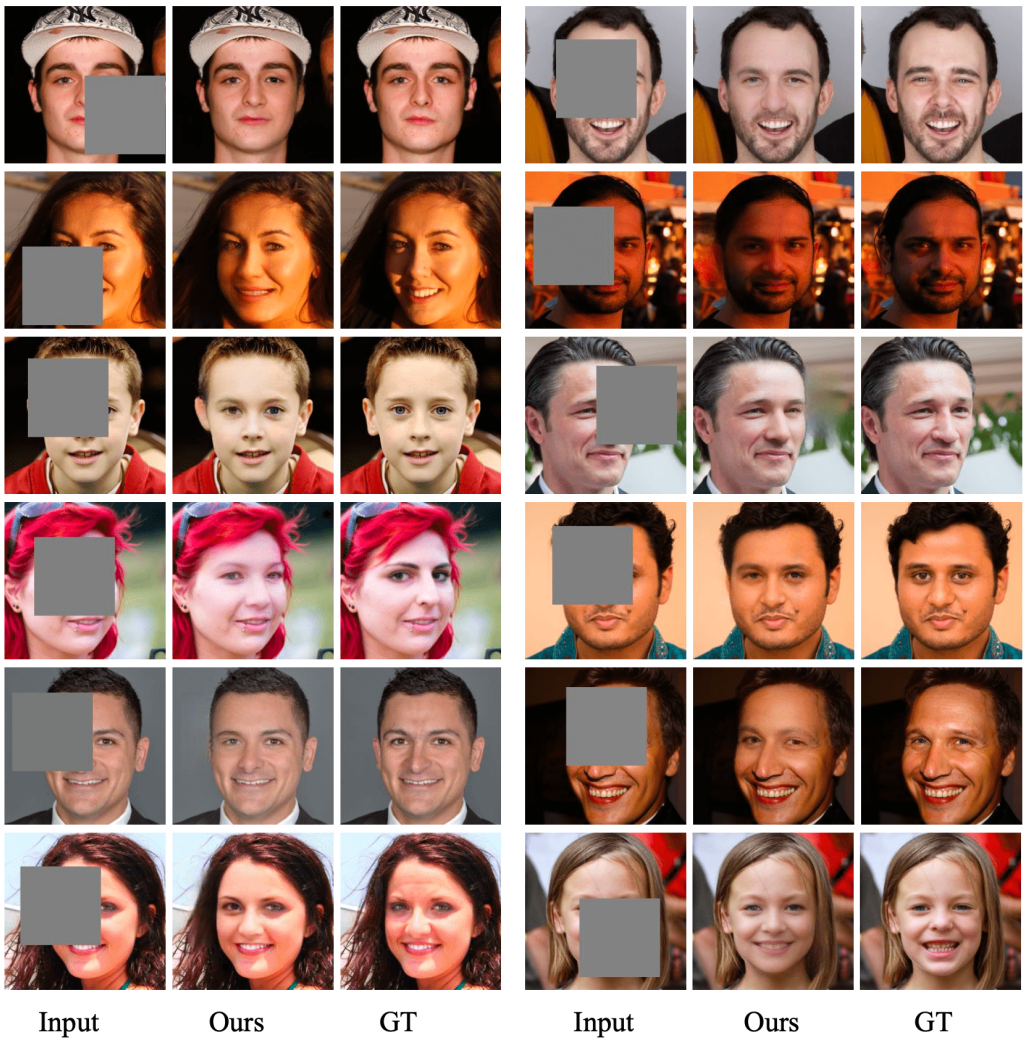



FIGURE 4.5: Qualitative comparisons in square mask cases on FFHQ [5]

### 4.2.4 User Study

We conducted a comprehensive user study to assess the visual quality of our proposed method. Specifically, we generated evaluation samples using the FFHQ test set on square mask cases. Twenty images were randomly selected from the test set, each containing a region to be inpainted. Inpainting results were generated using the methods proposed by Yu et al. [7], Zheng et al. [8], Liu et al. [12], and our proposed approach. For the user study, we engaged fifteen participants who were asked to evaluate and vote for the most visually plausible inpainting result for each input image. The images, along with the inpainting results from different methods, were presented to online users. To mitigate bias, the methods were anonymized and presented in a different randomized order for each image, as illustrated in Figure 4.7. Each user provided their preference for every image query, resulting in a total of 300

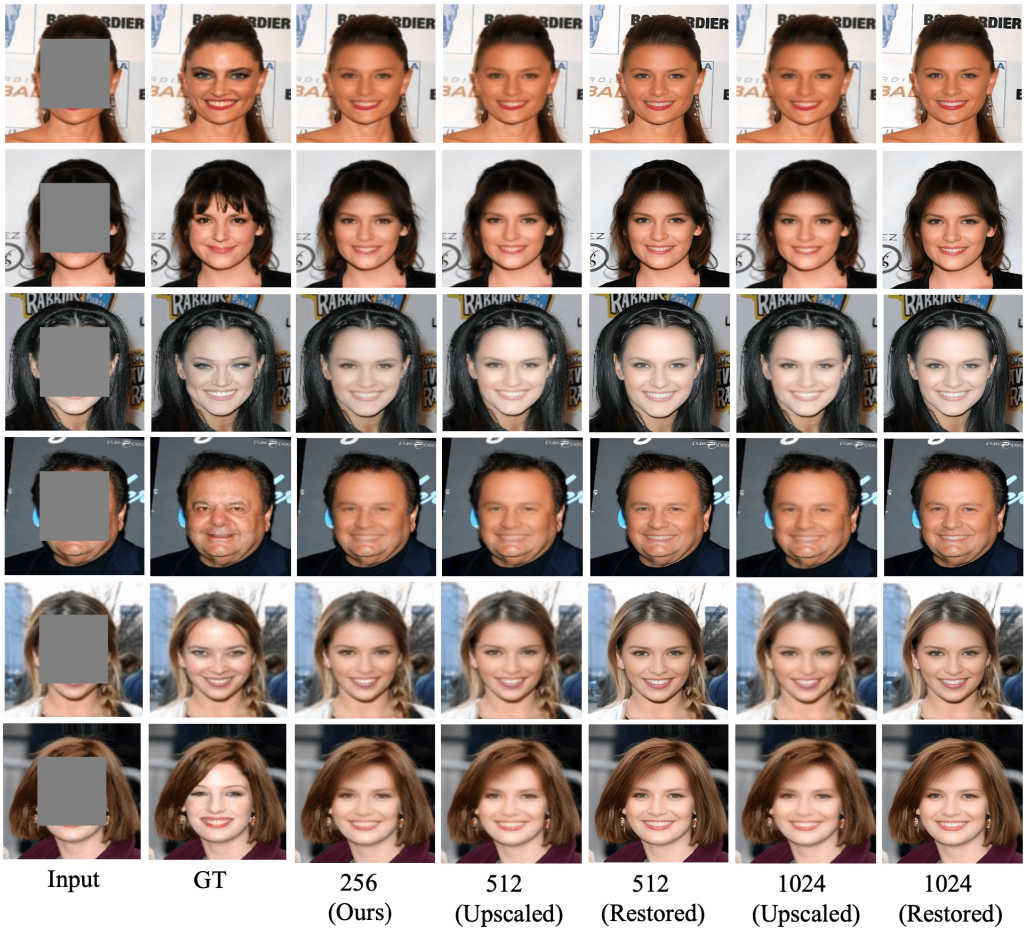



FIGURE 4.6: Qualitative comparisons in centre mask cases on CelebA [3]. Upscaled: bilinear upscaled our output. Restored: restored images with GFPGAN [6]

TABLE 4.6: User Study preferences in Image Inpainting: Analysis of 300 Votes from 15 Participants Evaluating 20 Images using a Moving Square Mask on FFHQ Dataset among Contextual Attention [7], Pluralistic Completion [8], High-Fidelity Completion(TFill) [9], and Ours.

| **Method** | **Votes** | **Votes(%)** |
|---|---|---|
| **CA [7]** | 42 | 14 |
| **PIC[8]** | 70 | 23.3 |
| **TFill [9]** | 71 | 23.6 |
| **Ours** | 117 | 39.0 |

votes in the dataset and mask setting. To quantify the preferences obtained from the user study, we express the outcomes as votes and percentage in Table 4.6. These findings align with our observations, reinforcing the efficacy of our proposed method in achieving visually compelling inpainting results.

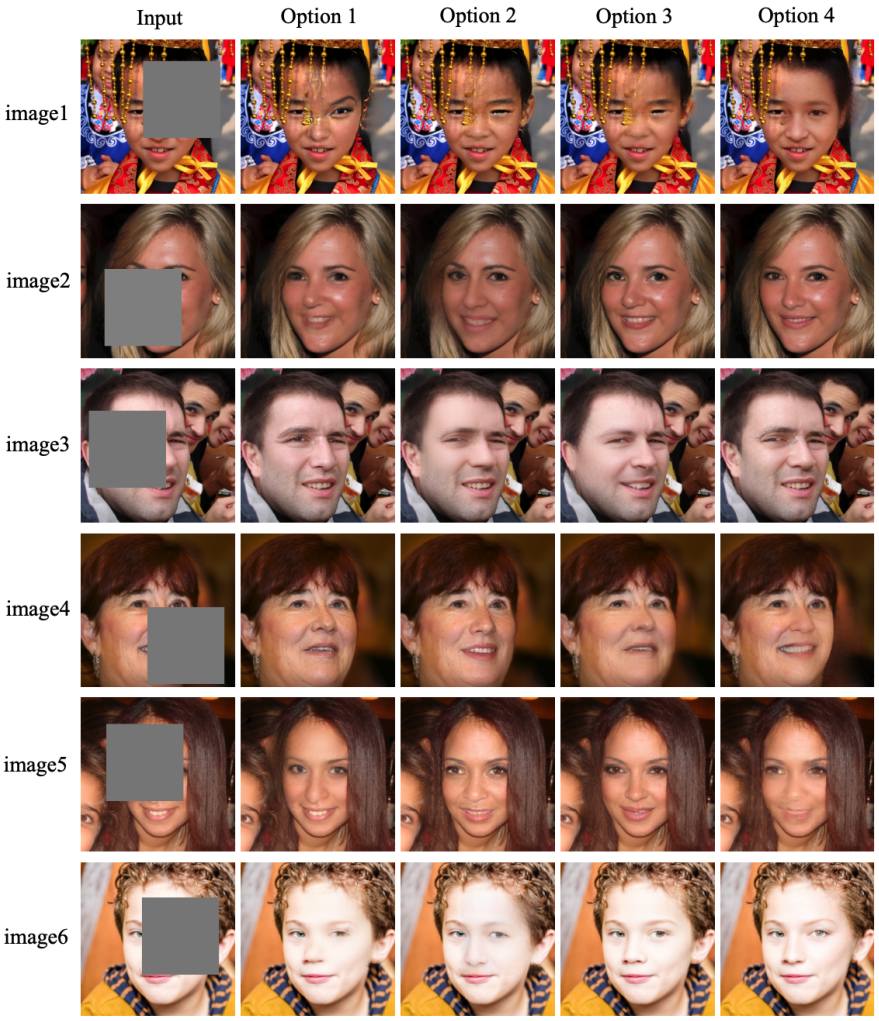



FIGURE 4.7: Comparative Evaluation of Inpainting Methods in User Study. The visual results of inpainting methods by Contextual Attention [7], Pluralistic Completion [8], High-Fidelity Completion(TFill) [9], and ours proposed approach are presented in random order for each image, as anonymized in the user study. Participants provided their preferences based on visual plausibility, resulting in 300 votes across 20 images. Refer to Table 4.6 for a quantitative breakdown of user preferences expressed as percentages. The randomization aimed to mitigate bias and ensure an unbiased assessment of inpainting quality.

# Chapter 5

# Conclusion, Challenges and Future Work

This thesis proposes a novel method for image inpainting, using a Dense Vision Transformer as the core of a generative adversarial network. It discoveres that the DenseViT methodology produced a more realistic result than the Vision Transformer's non-overlapping patch method. The proposed generating network is a two-stage architecture whose first stage uses DenseViT to fill the missing region. Then, the second stage uses a U-Net based refinement network to improve the output image quality. The refinement network creates visually realistic and semantically accurate images using high-level characteristics from the DenseViT output, thanks to the skip connections between the two stages. The suggested technique has produced visually promising results on image inpainting dense prediction tasks. To leverage the high-level features acquired by the Transformer's output feature maps to direct the refinement process, the refinement network combines skip connections from the output of the fusion blocks of the Transformer at various resolutions. The effectiveness of the suggested method was assessed and contrasted with other cutting-edge image inpainting methods. Overall, the suggested approach offers a practical framework for producing accurate predictions for each pixel in an input image. We achieved visually realistic outcomes at 256×256 resolution with our model. Due to limited resources, training for higher resolutions was not feasible.

In our research, we compared our DenseViT based image inpainting approach with existing methods on datasets like Paris StreetView, CelebA, and FFHQ. Our training process involved four stages and employed the Adam optimization algorithm. Quantitative comparisons demonstrated competitive performance, with significant gains in PSNR and SSIM. Qualitatively, our approach seamlessly integrated inpainted regions into the context, achieving natural and detailed results. We also introduced two post-processing settings: "upscaled" and "restored" with the latter

significantly improving image quality, particularly for facial features. In summary, our experiments showcased the effectiveness of our approach, making it a valuable tool for applications in facial recognition, image enhancement, and beyond.

Future work can focus on expanding the approach to handle higher resolutions while effectively managing the challenges associated with scaling up the model and optimizing the resource intensive training processes. Furthermore, there should be a focus on developing resource efficient training methods, thus expanding the model's usability for practical, real world applications. Exploring semi-supervised or weakly-supervised training methods can significantly reduce the dependency on large amounts of fully annotated data, making the model more practical. Given the computational expense associated with training large scale models like DenseViT, finding ways to overcome resource constraints for practical deployment is essential. Further enhancements in the realism of inpainted images are crucial, and this can be achieved by refining the refinement network, experimenting with novel architectures, and exploring new loss functions to deliver even more convincing results. The feasibility of deploying the model in real-time applications, such as video inpainting and live image editing, should be investigated. Adapting the architecture for video inpainting, which requires maintaining temporal continuity and consistency across frames, can open up video editing and restoration applications. Moreover, developing techniques to make the inpainting process more interpretable would enable users to understand and interpret the model's decision-making, enhancing its usability and transparency.

# Bibliography


[1] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. El-Khamy, M. Minderer, G. Heigold, S. Gelly, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations (ICLR)*, 2021.

[2] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[3] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015.

[4] C. Doersch, S. Singh, A. Gupta, J. Sivic, and A. Efros, "What makes paris look like paris?," *ACM Transactions on Graphics*, vol. 31, no. 4, pp. 103–110, 2012.

[5] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[6] X. Wang, Y. Li, H. Zhang, and Y. Shan, "Towards real-world blind face restoration with generative facial prior," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[7] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[8] C. Zheng, T.-J. Cham, and J. Cai, "Pluralistic image completion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[9] C. Zheng, T.-J. Cham, J. Cai, and D. Phung, "Bridging global context interactions for high-fidelity image completion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[10] G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro, "Image inpainting for irregular holes using partial convolutions," in *Proceedings of the European conference on computer vision (ECCV)*, 2018.

[11] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. Huang, "Free-form image inpainting with gated convolution," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[12] R. Liu, O. M. Aodha, and R. Cipolla, "Coherent semantic attention for image inpainting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[13] J. Huang, S. B. Kang, and N. Ahuja, "Shift-net: Image inpainting via deep feature rearrangement," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[14] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[15] X. Wang, F. Yu, and T. Darrell, "Deep network interpolation for continuous image representation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[16] W. Li, Z. Lin, K. Zhou, L. Qi, Y. Wang, and J. Jia, "Mat: Mask-aware transformer for large hole image inpainting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[17] Z. Deng, Y. Cai, L. Chen, Z. Gong, Q. Bao, X. Yao, D. Fang, W. Yang, S. Zhang, and L. Ma, "Rformer: Transformer-based generative adversarial network for real fundus image restoration on a new clinical benchmark," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 9, pp. 4645–4655, 2022.

[18] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 27, pp. 2672–2680, 2014.

[19] S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Globally and locally consistent image completion," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, pp. 1–14, 2017.

[20] C. Yang, X. Lu, Z. Lin, E. Shechtman, O. Wang, and H. Li, "High-resolution image inpainting using multi-scale neural patch synthesis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

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

[22] A. Criminisi, P. Perez, and K. Toyama, "Region filling and object removal by exemplar-based image inpainting," *IEEE Transactions on Image Processing (TIP)*, vol. 13, no. 9, pp. 1200–1212, 2004.

[23] A. A. Efros and W. T. Freeman, "Image quilting for texture synthesis and transfer," in *Proceedings of the ACM SIGGRAPH*, 2001.

[24] M. Bertalmio, A. Bertozzi, and G. Sapiro, "Navier-stokes, fluid dynamics, and image and video inpainting," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR*, 2001.

[25] Z. Wan, J. Zhang, D. Chen, and J. Liao, "High-fidelity pluralistic image completion with transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

[26] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in neural information processing systems*, vol. 34, pp. 8780–8794, 2021.

[27] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[28] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, and L. Van Gool, "Repaint: Inpainting using denoising diffusion probabilistic models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[29] C. Saharia, W. Chan, H. Chang, C. Lee, J. Ho, T. Salimans, D. Fleet, and M. Norouzi, "Palette: Image-to-image diffusion models," in *ACM SIGGRAPH 2022 Conference Proceedings*, 2022.

[30] C. Xie, S. Liu, C. Li, M.-M. Cheng, W. Zuo, X. Liu, S. Wen, and E. Ding, "Image inpainting with learnable bidirectional attention maps," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019.

[31] Y. Yu, F. Zhan, R. Wu, J. Pan, K. Cui, S. Lu, F. Ma, X. Xie, and C. Miao, "Diverse image inpainting with bidirectional and autoregressive transformers," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021.

[32] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *International Conference on Learning Representations (ICLR)*, 2013.

[33] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, vol. 9351, pp. 234–241, 2015.

[34] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017.

[35] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017.

[36] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014.

[37] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European Conference on Computer Vision (ECCV)*, 2016.

[38] R. Suvorov, E. Logacheva, A. Mashikhin, A. Remizova, A. Ashukha, A. Silvestrov, N. Kong, H. Goka, K. Park, and V. Lempitsky, "Resolution-robust large mask inpainting with fourier convolutions," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2022.

[39] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *International Conference on Learning Representations (ICLR)*, 2015.

[40] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proceedings of COMPSTAT'2010: 19th International Conference on Computational StatisticsParis France, August 22-27, 2010 Keynote, Invited and Contributed Papers*, 2010.

[41] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations (ICLR)*, 2014.

[42] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, 2019.

[43] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester, "Simultaneous structure and texture image inpainting," in *Proceedings of the International Conference on Computer Vision (ICCV)*, 2003.

[44] R. A. Yeh, C. Chen, T. Yian Lim, A. G. Schwing, M. Hasegawa-Johnson, and M. N. Do, "Semantic image inpainting with deep generative models," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017.

[45] A. A. Efros and T. K. Leung, "Texture synthesis by non-parametric sampling," *Proceedings of the International Conference on Computer Vision (ICCV)*, vol. 2, pp. 1033–1038, 1999.

[46] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[47] X. Wang, M. H. Ang, and G. Lee, "Cascaded refinement network for point cloud completion with self-supervision," *IEEE Transactions on Pattern Analysis amp; Machine Intelligence*, vol. 44, no. 11, pp. 8139–8150, 2022.

[48] X. Wang, F. Yu, and T. Darrell, "Deep network interpolation for continuous image representation," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[49] W. Li, X. Wang, S. Qi, Q. Zhao, Z. Liu, J. Shi, and J. Wang, "Learning hierarchical semantic image inpainting with contextual attention," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 42, no. 7, pp. 1637–1653, 2019.

[50] H. Tang, S. Liu, J. Zhang, J. Chen, N. Sebe, and J. Chen, "Realistic image inpainting by filling masked regions with context-aware patches," *IEEE Transactions on Image Processing (TIP)*, vol. 29, pp. 5982–5993, 2020.

[51] Y. Wang, J. Yu, and J. Zhang, "Zero-shot image restoration using denoising diffusion null-space model," *The Eleventh International Conference on Learning Representations*, 2023.

[52] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.