HINT Initial and Optimized: Employing Hperparameter Optimization and Transfer Learning

By

Sobia Asghar

(Registration No: 00000400570)

Department of Computer Science

School of Electrical Engineering and Computer Science (SEECS)

National University of Sciences and Technology (NUST)

Islamabad, Pakistan

(2024)

HINT Initial and Optimized: Employing Hperparameter Optimization and Transfer Learning

By

Sobia Asghar

Fall-2022-MS-CS 400570 SEECS

Supervisor

Dr Khuram Shahzad

Department of Computing

A thesis submitted in partial fulfillment of the requirements for the degree of Masters In computer science (MSCS)

In

School of Electrical Engineering and Computer Science (SEECS),
National University of Sciences and Technology (NUST),
Islamabad, Pakistan.

(October 2024)

THESIS ACCEPTANCE CERTIFICATE

Certified that final copy of MS/MPhil thesis entitled "HINT Initial and Optimized: Employing Hperparameter Optimisation and Transfer Learning" written by Sobia Asghar, (Registration No 00000400570), of SEECS has been vetted by the undersigned, found complete in all respects as per NUST Statutes/Regulations, is free of plagiarism, errors and mistakes and is accepted as partial fulfillment for award of MS/M Phil degree. It is further certified that necessary amendments as pointed out by GEC members of the scholar have also been incorporated in the said thesis.

Signature: MKhurram.

Name of Advisor: Dr. Muhammad Khuram Shahzad

Date: 11-Nov-2024

HoD/Associate Dean:

Date: 11-Nov-2024

Signature (Dean/Principal):

Date: 11-Nov-2024

FORM TH-4

National University of Sciences & Technology

MASTER THESIS WORK

We hereby recommend that the dissertation prepared under our supervision by: (Student Name & Reg. #) Sobia Asghar [00000400570]

Titled: HINT Initial and Optimized: Employing Hperparameter Optimisation and Transfer Learning

be accepted in partial fulfillment of the requirements for the award of Master of Science (Computer Science) degree.

Examination Committee Members

1. Name: Shah Khalid Signature: Khalid
28-Nov-2024 2:05 PM

2. Name: Syed Imran Ali Signature: S. I. Ali
28-Nov-2024 2:05 PM

3. Name: Farzana Jabeen Signature: Farzana
28-Nov-2024 2:05 PM

Supervisor's name: Muhammad Khuram Shahzad Signature: MKhurram.
28-Nov-2024 2:08 PM

Imranmalik
Muhammad Imran Malik
HoD / Associate Dean

03-December-2024
Date

COUNTERSINGED

04-December-2024
Date

M Ajmal
Muhammad Ajmal Khan
Principal

THIS FORM IS DIGITALLY SIGNED

Approval

It is certified that the contents and form of the thesis entitled "HINT Initial and Optimized: Employing Hperparameter Optimisation and Transfer Learning" submitted by Sobia Asghar have been found satisfactory for the requirement of the degree

Advisor : Dr. Muhammad Khuram Shahzad

Signature: Khurram.

Date: 11-Nov-2024

Committee Member 1:Dr. Shah Khalid

Signature: Khalid

Date: 11-Nov-2024

Co-Advisor: Dr. Syed Imran Ali

Signature: S.I. Ali

Date: 11-Nov-2024

Committee Member 3:Dr Farzana Jabeen

Signature: Farzana

Date: 11-Nov-2024

AUTHOR’S DECLARATION

I hereby declare that this submission titled "HINT Initial and Optimized: Employing Hperparameter Optimisation and Transfer Learning" is my own work. To the best of my knowledge it contains no materials previously published or written by another person, nor material which to a substantial extent has been accepted for the award of any degree or diploma at NUST SEECS or at any other educational institute, except where due acknowledgement has been made in the thesis. Any contribution made to the research by others, with whom I have worked at NUST SEECS or elsewhere, is explicitly acknowledged in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except for the assistance from others in the project’s design and conception or in style, presentation and linguistics, which has been acknowledged. I also verified the originality of contents through plagiarism software.

Student Name:Sobia Asghar

Student Signature: ________________

Date: 11-Nov-2024

Certificate for Plagiarism

It is certified that PhD/M.Phil/MS Thesis Titled "HINT Initial and Optimized: Employing Hperparameter Optimisation and Transfer Learning" by Sobia Asghar has been examined by us. We undertake the follows:

a. Thesis has significant new work/knowledge as compared already published or are under consideration to be published elsewhere. No sentence, equation, diagram, table, paragraph or section has been copied verbatim from previous work unless it is placed under quotation marks and duly referenced.

b. The work presented is original and own work of the author (i.e. there is no plagiarism). No ideas, processes, results or words of others have been presented as Author own work.

c. There is no fabrication of data or results which have been compiled/analyzed.

d. There is no falsification by manipulating research materials, equipment or processes, or changing or omitting data or results such that the research is not accurately represented in the research record.

e. The thesis has been checked using TURNITIN (copy of originality report attached) and found within limits as per HEC plagiarism Policy and instructions issued from time to time.

Name & Signature of Supervisor

Dr. Muhammad Khuram Shahzad

Signature : MKhurram.

Dedication

First and foremost , I want to express my gratitude to Allah Almighty , for showering upon me to countless blessings that ultimately led to the successful completion of my research.

Dr.Khuram Shahzad , my thesis adviser , has been an invaluable source of focused direction , unwavering commitment , unending inspiration and Ceaseless support during the course of my research.

I would like to appreciate my parents, whose unwavering support, encouragement, and belief in me have been my guiding light throughout this journey. Their love and sacrifices have inspired me to pursue my dreams relentlessly. I'd like to pay gratitude to everyone in my life who supported me throughout the journey.

Acknowledgments

All praise is due to Allah (S.W.A), the Creator and Sustainer of the Universe. He alone has the authority to uplift or humble as He wills. Truly, nothing happens without His permission. From my first day at NUST to my final day, He alone guided me, granted me blessings, and paved the way for my success. There is no way to fully repay His endless favors that supported me throughout my research journey, enabling me to complete it successfully.

Sobia Asghar

Abstract

Employing transformer-based architectures in image inpainting has significantly advanced the quality of generated results. By leveraging self-attention mechanisms, transformers can capture long-range dependencies within an image, making them particularly effective in restoring missing regions with coherence. Recent developments, such as the HINT framework, have introduced enhanced attention mechanisms that further improve inpainting performance by incorporating mask-aware encoding.Transformer models often struggle with processing high-resolution images due to their significant hardware requirements, which can limit their usability in broader applications and real-time scenarios.Reducing image resolution leads to information loss, which harms inpainting by causing blurred artifacts and vague structures in the reconstructed output. We proposed two models , HINT Initial and HINT Optimized. The HINT initial model employed transfer learning.HINT optimized leverages advanced hyperparameter tuning (uses Keras Tuner for advanced hyperparameter tuning, optimizing model parameters for performance) and architectural refinements (MPD module and SCAL enhance image inpainting by improving attention and downsampling). Our methods are evaluated on two benchmark datasets namely, Places2 and CelebA-HQ. Simulation experiments validated our proposed methods which showed significant improvements in comparison with the state-of-the-art image inpainting models. Notably, HINT Optimized effectively captured the complex relationships between pixels on both datasets. HINT initial showed improvements in (L1↓

(loss),FID↓(Fréchet Inception Distance) and LPIPS↓(Learned Perceptual Image Patch Similarity) on CelebA-HQ Dataset, whereas HINT optimized improved (PSNR↑ (Peak Signal-to-Noise Ratio) and SSIM ↑ (Structural Similarity Index Measure).On Places2 Dataset ,HINT initial improved L1↓and LPIPS↓ .HINT Optimized showed improvement on PSNR↑,FID↓ and SSIM↑.The model demonstrated a significant improvement in both accuracy and loss metrics, reflecting enhanced performance and a more efficient learning process.

Contents

List of Tables

List of Figures

CHAPTER 1

Introduction

1.0.1 Structure of the Introduction

Background and Context: This section introduces the rapid advancement of image inpainting technologies, driven by innovations in computer vision and deep learning over the past decade. Early techniques such as diffusion-based and patch-based methods were limited in their ability to handle minor image damage or fill small gaps. However, the advent of deep learning, particularly the use of Generative Adversarial Networks (GANs) and transformers, has significantly improved the quality and realism of image inpainting, even in complex scenarios.

Importance of Image Inpainting: The process of image inpainting is explained as a critical task in computer vision, where the goal is to recreate missing or damaged areas of an image using the surrounding pixels. The challenge is to ensure the generated content blends seamlessly with the existing image, both in terms of texture and semantics. This technology has a wide range of applications, including photo restoration, object removal, video frame editing, and even medical imaging, where accuracy and visual consistency are paramount.

Limitations of Traditional Methods: Traditional methods like diffusion-based and convolutional neural networks (CNNs) are discussed in terms of their limitations. These include restricted receptive fields and difficulties in capturing long-range dependencies, which often lead to unrealistic or blurred results when dealing with large missing areas or complex patterns. The shortcomings of these methods highlight the need for more advanced approaches that can handle larger gaps and deliver more coherent image reconstructions.

Emergence of Transformers: The introduction of transformer models in computer vision is described, focusing on their ability to capture long-range dependencies. Transformers, particularly Vision Transformers (ViTs) and Swin Transformers, have outperformed CNNs in many tasks by modeling images as sequences of patches, which allows for a more global understanding of the image. These models, while powerful, also come with computational challenges, particularly when processing high-resolution images.

Innovations in HINT: The introduction of the High-quality INpainting Transformer (HINT) model is discussed. HINT builds on the strengths of transformers by incorporating key innovations such as the Mask-aware Pixel-shuffle Downsampling (MPD) and Spatially-activated Channel Attention Layer (SCAL). These mechanisms are designed to preserve visual details and improve the model's ability to reconstruct complex textures in large missing areas. This section sets up HINT as a response to the limitations of existing methods, offering a more efficient and scalable solution.

Current Limitations and Proposed Solutions: The limitations of existing inpainting models, including HINT, are acknowledged. These challenges include reconstructing semantically coherent details in areas with large or irregularly shaped masks and the information loss caused by conventional downsampling techniques. The proposed enhancements in HINT Initial and HINT Optimized aim to minimize information loss and maintain spatial coherence, leading to sharper, more detailed inpainted images.

Key Objectives and Goals: The introduction concludes with a clear outline of the research objectives and goals, including developing an efficient inpainting model, addressing the shortcomings of current methods, delivering a robust solution, supporting digital restoration efforts, integrating into content creation workflows, and enhancing autonomous systems. The main contributions of the research are also highlighted, such as the integration of transformer-based modules with CNN architectures and the improvements in performance metrics like PSNR, SSIM, and FID.

The development and use of image INpainting technologies has expanded rapidly over the past ten years because of advancements in computer vision and deep learning. Early INpainting applications were made possible by conventional techniques including diffusion-based and patch-based methods, but these were frequently restricted to correcting minor picture damage or filling in tiny gaps. These techniques were not able to reconstruct vast lost areas or convey complex meanings. However, picture INpainting applications have changed dramatically with the introduction of deep learning, especially Generative Adversarial Networks (GANs) and transformers, which allow for more realistic content generation in complicated scenarios and higher-quality restorations [27, 34, 17].

A key task in computer vision is picture INpainting, which involves using the visible pixels surrounding missing, damaged, or corrupted areas of an image as context to recreate such areas. Creating believable and cohesive content that merges in with the image's existing elements in terms of texture and semantics is the aim of image INpainting [23].There are several uses of this task, ranging from fixing digital picture distortions and restoring old photos to more intricate activities like deleting objects from photos or altering video frames. Accuracy and visual coherence are crucial in fields like entertainment, photography, and even medical, where the ability to convincingly inpaint huge, uneven regions in photographs is quite significant.It has been applied in image processing and computer vision tasks such as photo editing [27], object removal [35]and depth completion [17]. The authors demonstrate that INpainting can be an effective self-supervised learning task.

The paper [23] proposed that by training a model to predict missing regions of an image, the model learns useful feature representations that can be transferred to other computer vision tasks like object detection and recognition, without requiring labelled data. Convolutional neural networks (CNNs) [12] have been the mainstay of traditional picture INpainting techniques, which use encoder-decoder architectures to process and restore damaged images. However, these techniques frequently have shown drawbacks related to convolutional operations, such as limited receptive fields and a failure to capture long-range relationships. Accurately modeling and restoring visible information is frequently a major difficulty for traditional image INpainting approaches, particularly when the missing areas of the image are vast or complicated. In this sense, diffusion-based techniques—one of the original methods for image INpainting—are very constrained[2].

When working with large missing areas or complex textures, these methods frequently produce blurry or unrealistic results since they rely on the progressive propagation of pixel values from the edges of the missing regions into the gaps. These approaches are not suitable for INpainting jobs where the missing regions necessitate the reconstruction of objects, structures, or complex patterns because they lack high-level semantic knowledge, which hinders their ability to produce detailed material[8]. During image processing, factors such as a poor environment, excessive noise, unfavourable shooting conditions, and unstable network communications often result in image blurring and loss [12]. Work [2] and [8] investigated the usage of GAN-based architectures such as Partial convolutions /And Mask update mechanism which employed contextual attention mechanisms to utilize surrounding picture data.

Reconstructing semantically coherent and texture-consistent details in the missing parts requires effective modeling of the valid information inside visible regions, which is a major obstacle to picture INpainting. This is especially apparent in areas with a lot of masks, where there is little reliable data. Due to the reduction of feature size from filters and downsampling, existing approaches that use convolutional layers for downsampling have the inherent disadvantage of information loss [49].

Pixel-shuffle down-sample is frequently employed in picture denoising , image deraining , and image super-resolution due to its ability to retain input information. The input's components are periodically rearranged to produce an output that is scaled by the sample stride. Its efficacy, however, is predicated on the sample stride being sufficiently tiny to prevent disturbing the noise distribution .This is not appropriate for image INpainting with uneven and variablesize masks, and it only holds true for a comparatively independent distribution of noise and raindrops. Pixel drifting would result from only applying standard Pixel-shuffle Downsampling (PD) to a corrupted image[51, 52].

Transformers [5] have demonstrated superior performance in capturing long-range dependencies, especially in tasks requiring a comprehensive understanding of the image structure. They proposed the Vision Transformer (ViT), which processes images as a sequence of patches, enabling global context modelling and adapting transformer models for image categorisation. Similarly, work [4] have proposed Swin Transformer which improved performance on a variety of vision tasks by optimising computational efficiency with a hierarchical structure with shifting windows. These developments highlighted the transformer-based models' ability to improve upon the drawbacks of convolutional methods.

The ability of transformers to capture long-range relationships across image regions—a critical component for comprehending complex structures in images—has made them extremely effective in vision tasks. [16] Introduced the Vision Transformer (ViT), one of the most important developments in this field. ViT splits the input image into a series of smaller patches, in contrast to conventional convolutional neural networks (CNNs), which process images pixel by pixel or in localized patches. In order to capture global dependencies, these patches are then loaded into a transformer model and handled as tokens, much like words in jobs involving natural language processing. Compared to CNNs, which are naturally constrained by their limitations, this method enables ViT to more accurately model the relationships between various components of an image.

ViT enhanced models' capacity to represent global context, but it also brought computational costs, particularly for high-resolution images, to the table.[26] presented Swin Transformer, a more effective hierarchical structure, to address these problems. The Swin Transformer uses shifted windows, which enable the model to process distinct areas of the image at different levels of granularity, rather than processing the full image as a single sequence of patches. In addition to lessening the computing load, this hierarchical design maintains the model's capacity to represent both local and global interdependence. The Swin Transformer guarantees that data is aggregated across several regions by moving the windows between layers, producing representations that are more contextually rich.

Importantly, the HINT [50] introduced two key components:

1. The Mask-aware Pixel-shuffle Downsampling (MPD) module
2. The Spatially activated Channel Attention Layer (SCAL).

The MPD module ensured the preservation of fine visual details during the downsampling process, reducing information loss while maintaining the consistency of both structure and texture across the image. The SCAL module enhanced the model's ability to handle intricate patterns by combining channel and spatial attention, enabling fast and effective representation learning. Together, these components significantly improve the INpainting process by emphasizing long-range dependencies, crucial for generating credible results, especially in large missing regions. HITN incorporated self-attention mechanisms, HINT excels at capturing and reconstructing complex textures, outperforming traditional CNN-based methods that struggled with such tasks.

Particularly, in applications requiring a comprehensive understanding of the image, ViTs and their derivatives, including HINT, demonstrated considerable potential due to their capacity for long-range modelling and contextual representation.

1.0.2 Current Limitations

Although HINT performs better with large masks compared to other methods there are still challenges when reconstructing semantically coherent details in regions with very large or irregularly shaped masks.Also the use of conventional downsampling methods may lead to the loss of important information, which is critical when dealing with corrupted images containing large missing regions.

In order to address above mentioned limitation in HINT, we proposed HINT initial, and HINT optimized. While the initial version of HINT demonstrated an enhancement in addressing information loss during downsampling and HINT optimized not only minimized the information loss but also excels in maintaining spatial coherence, leading to sharper, more detailed inpainted images. Simulation experiments demonstrated validity of proposed variations of HINT. Overall, this research leverages the advantages of transformer topologies which reduce information loss and improve spatial awareness.

1.0.3 Key objectives:

- Develop an Efficient INpainting Model Create an inpainting model that not only achieves superior visual quality but also operates efficiently, ensuring scalability for real-world applications.
- Address Shortcomings of Existing Methods Identify and address the limitations of current inpainting techniques to improve robustness and performance in image reconstruction tasks.
- Deliver a Robust and Efficient Solution Develop a solution that is both reliable and efficient, ensuring it can be utilized across various industries requiring high-quality image reconstruction.
- Facilitate Applications in Digital Restoration Ensure the inpainting model supports industries focused on digital restoration, enhancing the quality of visual content reconstruction.
- Support Content Creation Needs Tailor the model for content creation purposes, ensuring seamless integration into workflows that demand high-quality image completion.
- Enhance Autonomous Systems Ensure the model delivers reliable visual data processing for autonomous systems that depend on accurate image inpainting for decision-making.

1.0.4 Goals

1. Develop a High-Quality, Efficient Inpainting Model Create a model that balances superior visual quality with operational efficiency, ensuring it can scale for real-world applications.
2. Improve on Existing Inpainting Techniques Identify limitations in current methods and enhance the model's robustness and performance in reconstructing images.

3. Deliver a Robust, Industry-Ready Solution Design a solution that is both reliable and efficient, suitable for industries needing precise image reconstruction.

4. Support Digital Restoration Efforts Ensure the model aids industries focused on digital restoration, improving the quality of reconstructed visual content.

5. Integrate into Content Creation Workflows Customize the model to fit seamlessly into content creation processes, ensuring high-quality image completion.

6. Enhance Autonomous System Capabilities Ensure the model provides accurate image inpainting for autonomous systems that rely on precise visual data for decision-making.

1.0.5 Main contributions of our work:

- We conducted an in-depth comparative analysis of current state-of-the-art methods in image inpainting.

- Our model employs a hybrid design, integrating Transformer-based modules with a CNN architecture to enhance performance and reduce computational demands.

- HINT Initial utilizes pre-trained weights to boost performance on metrics such as L1, FID, and LPIPS, accelerating training, improving convergence, and enhancing generalization through transfer learning.

- HINT Optimized incorporates hyperparameter tuning and additional custom layers (MPD & SCAL), resulting in notable improvements in PSNR and SSIM performance metrics.

- Both HINT Initial and HINT Optimized exhibit increased training accuracy and reduced training loss, demonstrating the model's stability and overall effectiveness.

1.0.6 Motivation for the Researsch

The driving force behind this research is the ongoing challenges that current image INpainting methods encounter, particularly when applied in real-world scenarios. Image INpainting plays a critical role in various domains, such as photo restoration, object removal, medical imaging, and video editing. However, existing approaches struggle to maintain consistent texture, handle large missing areas, and produce seamless, high-quality results that blend naturally with the original image content. Traditional methods, while effective for small gaps, fall short when dealing with intricate structures and large occlusions. Even more recent approaches, like CNN-based and GAN-based models, face limitations in capturing long-range dependencies, often resulting in artifacts, blurred textures, and suboptimal INpainting results.

Furthermore, the computational demands of cutting-edge transformer-based models, though offering improved global context and feature representation, restrict their practical use in environments with limited resources or where real-time processing is required. These models often necessitate extensive computational power and time, limiting their application in large-scale or time-sensitive image processing tasks.

The development of the High-quality INpainting Transformer (HINT) is motivated by the need to address these persistent issues. By leveraging the strengths of transformer architectures and introducing innovative mechanisms, this research seeks to enhance both the efficiency and quality of image INpainting. The incorporation of the Spatially-activated Channel Attention Layer (SCAL) and Mask-aware Pixel-shuffle Downsampling (MPD) aims to overcome the limitations of current methods.

SCAL enables a more effective extraction of features by focusing on both spatial and channel dimensions, while MPD ensures that essential visible information is preserved during downsampling, which is crucial for maintaining the integrity of the reconstructed image.

1.0.7 Problem Statement

Downgrading or reducing the resolution of input range in image leads to information loss which is detrimental to image inpainting which results in degradation of features, so the reconstructed output suffers from blurred artifacts and vague structures.

1.0.8 Solution Statement

While HINT Initial establishes a strong foundation in improving visual quality and generalization, HINT Optimized further enhances performance with architectural refinements and hyperparameter tuning, making it ideal for handling complex image inpainting challenges in various real-world applications.

CHAPTER 2

Literature Review

In this chapter, we examined the main methodologies and advancements in image inpainting, focusing on the shift from traditional methods to modern deep learning and transformer-based models. We discussed the strengths and limitations of convolutional neural networks (CNNs) and generative adversarial networks (GANs), which have improved inpainting performance, as well as how transformers address these shortcomings by capturing long-range dependencies in images. Finally, we introduced the HINT model, which integrates spatial and channel attention mechanisms, along with improved downsampling techniques, to overcome the challenges faced by earlier models in generating high-quality inpainted images.The field of image INpainting has evolved significantly from diffusion-based methods to deep learning and transformer-based techniques.

Each advancement has addressed specific challenges, such as handling large missing areas, preserving texture consistency, and improving computational efficiency [27]. With the introduction of convolutional neural networks (CNNs), picture INpainting skills advanced significantly. One of the earliest deep learning-based models for image INpainting, the Context Encoder, was presented by [8]. It combined CNNs with a loss function to produce believable inpainted regions. By using an adversarial loss to increase the realism of generated content, generative adversarial networks (GANs) considerably enhanced the quality of INpainting [12]. Although the outcomes of these models were aesthetically pleasing, they frequently had trouble producing semantically consistent content, particularly when dealing with huge holes in the image.

The work [7] introduced partial convolution, which further enhanced picture INpainting. In order to produce more accurate reconstructions, this technique used a mask-aware convolution process that updated only valid pixels. Notwithstanding these developments, producing high-quality inpainted areas for intricate structures and textures remained difficult.

Transformers were first created for natural language processing [14], however they have lately been successfully used to vision challenges. The limitations of CNNs, which are essentially local, are addressed by vision transformers [27], which use self-attention mechanisms to capture long-range dependencies inside an image. Transformers enhance the creation of cohesive and contextually relevant material by enabling a more comprehensive comprehension of an image in the context of image INpainting[45].

Recent developments in transformer architectures have opened new avenues for improving image INpainting. Transformers, originally developed for natural language processing, have demonstrated exceptional performance in various visual tasks due to their ability to capture long-range dependencies [16].One of the early studies in this domain was presented by [52], who utilized PDEs to transfer data from known regions into missing areas. Although innovative, this approach was limited by its inability to handle large missing regions and complex textures.Subsequent research sought to address these shortcomings. Methods like Globally and Locally Consistent Image Completion extended the encoder-decoder framework to integrate local texture with global context. This approach generated more coherent and aesthetically pleasing inpainted images, though challenges persisted, particularly when it came to preserving fine details in larger missing areas[44].

Deep learning approaches brought about a radical change in picture INpainting. Many INpainting models were built around Convolutional Neural Networks (CNNs), which are well known for their robust feature extraction capabilities [6]. One of the earliest deep learning models for image INpainting was Context Encoders, proposed by the author [34],this model employed an encoder-decoder architecture to predict missing content based on the surrounding context. While it improved the realism of inpainted images, it struggled with fine details and often produced suboptimal results in the presence of large missing regions.

An important turning point in picture INpainting was the creation of generative adversarial networks (GANs) [5]. Two competing networks make up GANs: a discriminator that seeks to distinguish between produced and actual images, and a generator that seeks to produce believable images. Even for more complex INpainting tasks, models were able to provide more realistic and aesthetically pleasing outcomes thanks to the adversarial training mechanism. The Context Encoder, one of the first GAN implementations for INpainting, was presented by [8, 49]. Their approach applied an adversarial loss to make sure the generated regions looked realistic and an encoder-decoder architecture to forecast the missing content. Because convolutional neural networks (CNNs) are local, the Context Encoder's ability to capture intricate textures and structures was constrained, despite its success.

Later studies, such [17] and [12], expanded GAN-based INpainting with increasingly complex architectures. Using two discriminators—one that focused on local consistency and the other on global realism—Iizuka et al. presented a globally and locally consistent INpainting model. Contextual attention developed further, allowed the generator to concentrate on similar patches within the known regions of the image. These methods improved the realism and consistency of INpainting results but still struggled with very large masks or complex, structured images.

In order to improve the mask-handling capabilities of INpainting models, methods such as partial convolutions[7, 46, 36] were introduced as GAN-based models developed. By applying convolution operations only to legitimate (non-missing) portions of the image, partial convolutions increased the model's sensitivity to the masked sections. This method outperformed conventional convolutional models in handling irregular masks, resulting in improved boundary refinement and more precise INpainting of missing regions. Even partial convolution-based GANs, however, exhibited drawbacks, especially when attempting to infer global context for sizable missing regions. Researchers looked into transformer-based architectures because they needed models with better global context awareness and could better capture long-range dependencies[28, 40].

To address these issues the findings presented by [11] developed a globally and locally consistent image completion technique, extending the encoder-decoder framework. By integrating local texture details with global contextual elements, this method generated INpainted images that were more visually coherent and aesthetically pleasing. improvement in performance. However, the model continued to face challenges with large missing regions and struggled to maintain high-quality textures throughout the image[33].

The advent of Generative Adversarial Networks (GANs) brought about further advancements in image INpainting. Originally introduced by [15] GANs set two networks against each other: a generator and a discriminator. This adversarial framework proved beneficial for image INpainting, with several researchers adapting GANs for this purpose. For instance, [21] By enabling the model to concentrate on pertinent regions, the GAN-based method with contextual attention improves image INpainting by notably enhancing texture consistency and detail uniformity. New possibilities for picture INpainting were made possible by the introduction of transformers to vision tasks, most notably with Vision Transformers (ViTs). Transformers are ideal for INpainting tasks requiring global context knowledge because they employ self-attention methods to capture long-range dependencies[43, 48].

Citation	Techniques Used	Advantages	Limitations
[2], [3]	GANs (Generative Adversarial Networks), Deep Autoencoders	Realistic texture generation, preserves high-frequency details.	Requires large datasets, susceptible to mode collapse.
[5], [6]	Attention-based mechanisms to focus on relevant surrounding regions	Improved context awareness, better filling of large holes.	Computationally expensive, struggles with highly irregular patterns.
[9], [10]	Partial differential equations (PDEs), Variational techniques	Good at handling small gaps, mathematically interpretable.	Fails on larger regions, unable to generate new structures.
[12], [13]	Vision Transformers (ViT), Attention-guided inpainting	Captures long-range dependencies, handles complex patterns.	High memory requirements, slower training.
[7], [8]	Patch-based optimization and nearest-neighbor search	Efficient and computationally lightweight.	Fails on semantically complex images, struggles with diverse textures.
[4], [11],[33]	Combination of GANs and transformers	Leverages strengths of multiple approaches, robust results.	Increased complexity, harder to optimize.
[14], [15]	Multi-scale convolutional networks, pyramidal structures	Preserves both global and local details, effective for high-resolution images.	More parameters, requires careful tuning.
[1], [16],[31]	Semantic understanding through pre-trained models (e.g., ImageNet)	Semantically coherent results, effective on natural images.	Limited generalization to unseen domains.
[17], [18]	Frequency domain analysis, Fourier transforms	Efficient reconstruction in the frequency domain, good for periodic patterns.	Not suitable for irregular or non-periodic patterns.
[19], [20]	Recurrent neural networks (RNNs), sequential image patch prediction	Effective for sequential structure reconstruction.	Fails with non-sequential dependencies, longer training times.
[21-25]	Policy learning for inpainting actions	Adapts dynamically to unseen structures, good generalization.	Difficult to train, sensitive to reward design also high training loss

Table 2.1: List of Papers and Their Techniques, Advantages, and Limitations

TransFill, one of the earliest transformer-based models for image completeness, was presented by [19]. TransFill proved to be more effective than CNN-based and GAN-based models, especially when it came to tasks involving the filling of large, irregular portions in photos. The transformer was especially good at producing semantically coherent content for INpainting because it could simulate long-range interactions between image patches. Despite the success of TransFill and other transformer-based models, they often struggled with mask handling, treating the mask as a simple binary input rather than a feature that could guide the model's attention more effectively. This limitation paved the way for models like HINT, which incorporate advanced mechanisms for mask-aware encoding[41, 10].

The Vision Transformer (ViT), introduced by [20, 32, 30], applies transformers to image patches, allowing for a global context interpretation.Transformers have become a viable substitute to overcome these drawbacks, especially in terms of their capacity to use self-attention mechanisms to capture global context. Transformers like the Swin Transformer still have computational inefficiencies, nevertheless, even with their enhanced ability to handle significant areas of missing data[9]. Even transformer-based models find it challenging to preserve texture consistency and spatial coherence as the missing region expands because pixel correlations deteriorate. Furthermore, the majority of transformer-based techniques depend on downsampling techniques, which may result in information loss and worsen the inpainted image's quality. A key advantage of transformers over traditional convolutional methods is their ability to model relationships between distant regions of the image through the self-attention mechanism[39, 22].

Building on this concept,[42] introduced the Swin Transformer, which balanced computational efficiency and performance by using shifted windows and a hierarchical structure. This approach marked a major advancement by reducing computational complexity without sacrificing the ability to capture fine details and broad context. However, as the missing region of the original image becomes larger and the distance between unknown pixels and known pixels increases, the image restoration becomes semantically ambiguous due to the weakening of pixel correlation . Traditional image INpainting methods usually apply diffusion-based or patch-based techniques to propagate information from the image background to fill the holes in corrupted images. Such approaches would perform well on stationary (e.g., repeated) textures, but may fail in non-stationary images (e.g., most natural images) [38, 24, 25, 3].

Our work builds on these developments by introducing the High-quality INpainting Transformer (HINT), designed to address the limitations of previous approaches. The HINT model incorporates a Spatially activated Channel Attention Layer (SCAL) and a Mask-aware Pixel-shuffle Downsampling (MPD) module to improve the INpainting process. The SCAL module enhances feature representation by combining spatial and channel attention, while the MPD module preserves visible information during downsampling, ensuring minimal information loss. These innovations target two key challenges in image INpainting: maintaining the integrity of visible information and efficiently learning representations that capture intricate relationships within the image [50, 1, 29].

Examining the fundamental ideas that guided the creation of the HINT model is crucial to comprehending its contributions.

Self-attentional processes: The self-attention mechanism, which is essential to transformer models, enables the model to assess the relative relevance of various visual components when recreating missing regions. Transformers are well-suited for jobs like image INpainting, where missing regions may be distant from the visible environment, because of this process, which allows them to simulate long-range dependencies.

Attention to channels and spaces: Conventional INpainting models frequently concentrate on either spatial or channel-wise information. While spatial attention concentrates on "where" features are significant, channel attention refers to concentrating on the "what" in terms of salient features. HINT can more effectively capture both by combining them via the SCAL module. The SCAL module allows for the integration of both, which improves HINT's ability to capture the associations required for high-quality INpainting. **Shuffled pixel downsampling:** The loss of information during downsampling is a frequent problem in INpainting. The spatial integrity of visible regions may be compromised by conventional downsampling methods like convolutional downsampling. By retaining spatial consistency, MPD, which was added to HINT, lessens this problem and preserves more information from the damaged image (HINT). Higher SSIM and PSNR scores, along with a lower training loss, show that the HINT model performs better than the other models, suggesting faster learning, better texture consistency, and robustness in producing high-quality inpainted images. Its innovative elements make it a scalable and sophisticated solution for real-world INpainting applications by enhancing visible information retention and complex image dependency learning.

This chapter has provided a comprehensive overview of the evolution of image inpainting techniques, highlighting how the field has progressed from traditional methods to more advanced deep learning and transformer-based approaches. Early techniques, such as diffusion-based methods, struggled with filling large missing regions and maintaining texture consistency. The introduction of convolutional neural networks (CNNs) and generative adversarial networks (GANs) led to significant improvements in inpainting quality. Despite these advances, challenges remained, particularly when it came to generating semantically coherent content for images with extensive missing areas. The rise of transformer-based models brought further enhancements, enabling better capture of global context through self-attention mechanisms. However, issues like computational inefficiency and difficulties in preserving fine details across large gaps persisted. The chapter also introduces the HINT model, which includes innovations like Spatially activated Channel Attention Layer (SCAL) and Mask-aware Pixel-shuffle Downsampling (MPD), aiming to address these limitations by enhancing feature representation and preserving critical visual information during downsampling. These advancements collectively improve image inpainting quality and efficiency.

CHAPTER 3

Proposed Approach

In this chapter, we introduce our proposed method to tackle the existing challenges in image inpainting techniques. This approach is designed to improve upon the limitations of earlier methods, particularly in handling large missing regions and maintaining image coherence. Our model builds upon recent advancements in deep learning and transformers, integrating new techniques that focus on preserving image details and capturing long-range dependencies. The goal of this approach is to create a more effective and efficient solution for generating realistic inpainted images, even in challenging scenarios with complex textures and large gaps.

This Improved HINT model starts with an Input Block that extracts essential features, followed by Gated Convolution Blocks with RELU Activation to enhance feature learning. An Elementwise Multiplication step further refines the features, while the Mask-Aware Pixel Shuffle Downsampling (MPD) [50] layer maintains the spatial positioning of any missing pixels. The Spatially activated Channel Attention Layer (SCAL) and a Custom CNN focus attention on important features, enhancing detail. At the model's core, a Bottleneck layer with MultiHeadAttention extracts high-level features, while Sandwich Layers manage up sampling. Convolution Layers integrate information from previous layers, and a Loss Function helps reduce output error by comparing it to the ground truth. The Output Layer then reconstructs the final image.

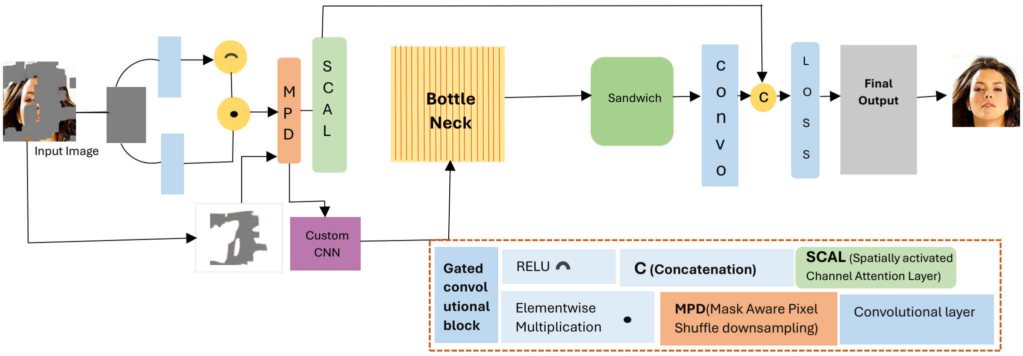
Input Image
M P D
S C A L
Bottle Neck
Sandwich
C o n v o
C
L o s s
Final Output
Custom CNN
Gated convolutional block
RELU
C (Concatenation)
SCAL (Spatially activated Channel Attention Layer)
Elementwise Multiplication ●
MPD(Mask Aware Pixel Shuffle downsampling)
Convolutional layer

Figure 3.1: Proposed Pipeline

3.1 Input Block (Mask-Inactive Downsampling)

The first masked or corrupted image is sent into the network for additional processing at the Input Block, which serves as the model's starting point. The model's job is to restore the image's damaged or missing regions in a way that makes sense structurally and aesthetically. At this point, the image could have parts that are intentionally occluded (masked) or missing, and the model must analyze the image while maintaining the necessary details[50].

3.1.1 Mask-Inactive Downsampling

Downsampling: This stage's main technique is downsampling, which lowers the input image's resolution. The processing of high-dimensional data, such as photographs, can be made more effective by using downsampling. The model can concentrate on extracting high-level features without the computational burden of processing every pixel at full resolution by decreasing the image's spatial dimensions (height and breadth).

3.1.2 Preserving Critical Information

The challenge with downsampling in image INpainting tasks is that, while it reduces the resolution for efficiency, it should not remove crucial information about the masked regions. If too much information about the masked areas is lost, the model will struggle to reconstruct these areas accurately. Mask-inactive downsampling is specifically designed to address this issue by ensuring that even though the image size is reduced, the model retains essential details about the masked regions.

3.1.3 Efficiency and Scalability

Reducing the resolution early in the pipeline allows the model to operate more efficiently. Lower-resolution images require fewer computational resources during training and inference. As a result, the model can scale more easily to larger images or higher-resolution tasks. By processing a lower-dimensional version of the image, the model can focus on learning useful global features, such as general shapes and patterns, rather than pixel-level details initially.

3.1.4 Balancing Trade-offs

The downsampling process needs to strike a balance between reducing the image size and maintaining essential features, especially for INpainting tasks. If the downsampling is too aggressive, the model may lose the critical contextual information required to accurately reconstruct the missing parts. However, if the resolution is not reduced enough, the computational cost may become prohibitively high, slowing down training and inference.

3.2 Initial Feature Extraction

At this stage, basic feature extraction begins. The model begins to identify and isolate core structures of the image, like large shapes and boundaries, while disregarding unnecessary fine-grained details. These features help guide subsequent layers, which will refine the image's reconstruction by adding details to the masked regions.

3.2.1 Preparation for Further Processing

This step sets the stage for subsequent operations like Mask-Aware Pixel-Shuffle Downsampling (MPD) and Convolution Layers. By reducing the image's dimensions while preserving critical information, the input block ensures that the model can process the image in an efficient and meaningful way.

3.3 MPD (Mask-Aware Pixel-Shuffle Downsampling)

In the next step of the model's INpainting process, Mask-Aware Pixel-Shuffle Downsampling (MPD) is applied. This technique is specifically designed to ensure that the model retains key information about the masked or corrupted areas of the image, even as it reduces the overall image resolution. By doing so, it enables the model to prioritize the restoration of these areas with high accuracy and efficiency.

3.3.1 Detailed Functionality of MPD:

The purpose of MPD is to treat each portion of the image differently from typical downsampling techniques (such max pooling or average pooling). Particular attention is given to the masked (corrupted) portions of the image while downsampling. There is a chance that crucial information about the areas that require INpainting will be lost during conventional downsampling. This is avoided by MPD, which makes sure that the damaged areas are kept intact even when the image's total size is decreased.

Mask Awareness

Due to MPD's "mask-aware" feature, it can distinguish between the image's masked (corrupted) and unmasked (intact) portions. It treats the masked areas differently during downsampling by keeping more specific data from certain areas. This guarantees that the model doesn't lose important information required to correctly reconstruct the image's missing portions.

Pixel-Shuffle Mechanism

MPD uses a pixel-shuffling technique, where pixel values from neighboring regions are reorganized and grouped in a way that maintains the relationships between pixels in the masked areas. Instead of simply reducing the resolution by discarding information, pixel-shuffling redistributes pixels in a manner that ensures the downsampled version of the image still contains a significant amount of information about the masked parts.the algorithm used for hyperparameter tunning is Random Search: Directly tied to the number of trials (T) and the number of epochs (E): Tuning.Complexity = T× E×Training.FLOPs Tuning.Complexity=T×E×Training.FLOPs Each trial trains a new model for the specified epochs, multiplying the training computational cost.

Efficiency of Downsampling

Although the model needs to retain information about the corrupted regions, it still needs to process the image efficiently. MPD strikes a balance between preserving critical information and reducing the resolution of the image, making it computationally feasible for deeper layers of the network to process. By reducing the image size in a "mask-aware" manner, MPD ensures that the model doesn't waste computational resources on less important regions while focusing on the INpainting task.

Enhanced Focus on Masked Areas

One of the most critical aspects of INpainting is the accurate reconstruction of the masked regions, which often requires the model to focus more on those areas than on the intact portions of the image. MPD facilitates this by prioritizing information retention from the masked regions. This makes it easier for the network to concentrate on the areas that need reconstruction, improving the model's ability to generate realistic and coherent INpainting results.

Feature Preservation in Downsampled Images

As the image is downsampled, important features from both the masked and unmasked regions need to be preserved for subsequent layers to process. MPD ensures that key features related to the masked regions are not lost during downsampling. These features provide the model with essential context about the corrupted areas, guiding the INpainting process later on.

Application in Multi-Scale Processing

In many advanced INpainting models, images are processed at multiple scales, and downsampling plays a crucial role in this. MPD can be applied at various stages of the network to ensure that at each level of resolution, the masked regions are treated with care. This ensures that by the time the image reaches its final resolution, the masked regions have been reconstructed with high accuracy.

Impact on Final Output Quality

The Mask-Aware Pixel-Shuffle Downsampling technique plays a pivotal role in the overall quality of the model's final output. By ensuring that the masked areas are preserved during downsampling, MPD increases the likelihood that the model will be able to restore these areas in a way that blends seamlessly with the rest of the image. This is particularly important for INpainting tasks where the goal is to produce images that are indistinguishable from complete, non-corrupted originals.

3.3.2 Key Benefits of MPD

Preservation of Crucial Information: The technique ensures that the masked areas retain key details, which helps the model reconstruct the missing parts with greater accuracy.

Efficient Processing:

By reducing the image size while focusing on masked regions, MPD optimizes the computational workload without sacrificing the quality of the final output.

Improved INpainting Results

Since the masked regions are treated with greater importance, the model can generate more realistic reconstructions, seamlessly blending the filled-in areas with the surrounding image.

Task-Specific Focus

MPD is particularly suited for INpainting tasks, where preserving and reconstructing missing regions is the central objective. The mask-aware approach ensures that the model dedicates more processing power to the parts of the image that matter most.

3.3.3 How MPD Integrates with the Overall Pipeline

Downstream Impact

The data processed by MPD is passed on to subsequent layers, such as convolution layers and sandwich layers, where additional feature extraction and refinement occur. By ensuring that the important information from the masked regions is not lost during downsampling, MPD plays a foundational role in the model's ability to effectively reconstruct these areas later in the pipeline.

Multiple Applications

As described earlier, MPD is used multiple times within the model pipeline. After initial downsampling and feature extraction, MPD may be applied again at later stages to ensure that the masked regions remain the focal point even as the model processes the image through multiple layers.

3.4 Convolution Layers

After the initial downsampling, the image is fed into a series of convolutional layers, which serve as the primary tool for extracting features from the image. These layers play a critical role in identifying important attributes like edges, colors, and textures that exist in both masked and unmasked portions of the image.

3.4.1 Feature Extraction Process

The feature extraction process is a critical step in deep learning models, where raw input data is transformed into a set of high-level representations that capture the most important attributes and patterns.

Edge Detection

Convolution layers help the model detect edges, which are the boundaries between different objects or regions in the image. Edge detection is essential for the model to maintain consistency when reconstructing the masked areas.

Color and Texture Identification :

In addition to edge detection, the convolutional layers also identify color gradients and textures. This allows the model to understand the surface properties of objects in the image, which is crucial for generating realistic INpainting results.

Hierarchical Learning

As the image passes through deeper layers of convolution, the model builds a hierarchical understanding of the image, learning simple features first (like edges and basic shapes) and then combining them to form more complex patterns. This hierarchical learning enables the model to generate coherent and detailed reconstructions.

Guidance for Reconstruction

The features extracted by these layers provide the model with essential information about the non-masked areas of the image. These features guide the model in predicting what the missing portions of the image should look like, ensuring that the reconstructed parts blend seamlessly with the rest of the image. By capturing both local details and global structure, the convolutional layers set the foundation for accurate INpainting.

3.5 RELU Activation

After extracting features through the convolutional layers, the model introduces Gaussian Error Linear Unit (RELU) activation to add non-linearity. This non-linear behavior is essential for enabling the model to capture and learn more complex patterns within the image.

3.5.1 Role of RELU Activation

In deep learning models, the Gaussian Error Linear Unit (RELU) activation function has become a potent substitute for more conventional activation functions like ReLU (Rectified Linear Unit). By using a Gaussian distribution to approximate the input values, RELU offers a probabilistic interpretation that makes it more appropriate for models requiring complex decision-making. By weighing inputs based on their probability, RELU offers a smoother transition than ReLU, which either permits or entirely blocks input. This allows minor negative inputs to flow through rather than being eliminated. This characteristic makes it possible for the RELU activation function to function well in tasks that call for intricate representations, like those seen in transformer structures used in vision and natural language processing applications. Following are the its roles:

Smooth Non-Linearity

Unlike simpler activation functions such as ReLU, RELU smoothly activates neurons based on the probability that an input should be preserved. This allows the model to handle intricate patterns in the image that involve subtle changes in texture or color.

Capturing Complex Patterns

The non-linear nature of RELU helps the model go beyond simple linear relationships between pixels. It can learn more sophisticated patterns and variations in the image, such as how light affects different surfaces or how textures change across regions. This is particularly useful for INpainting, where the model must predict realistic content for missing areas that align with complex patterns in the surrounding unmasked areas.

3.5.2 INpainting Application

In the context of INpainting, RELU enables the model to generalize better by learning patterns that aren't strictly linear or uniform. The RELU function helps the model transition from basic feature extraction to higher-level reasoning about the structure and texture of the image, making it possible to reconstruct missing parts in a way that is consistent with the rest of the image.

3.6 Sandwich Layers

The Sandwich Layersrepresent a set of intermediate layers that further refine the image by focusing on spatial relationships between various parts of the image. These layers play a crucial role in ensuring that the model understands both the local context and the global structure of the image, allowing for more accurate and coherent INpainting results.

3.6.1 Key Functions of Sandwich Layers

In deep learning architectures, the term "sandwich layers" usually refers to layers that are situated in between important parts, such as feedforward layers and attention modules, particularly in transformer-based models. By boosting feature extraction, regularization, and layer normalization, these layers significantly improve model performance. Sandwich layers frequently incorporate non-linear activations that add the required complexity to the model, dropout mechanisms for regularization, and normalizing approaches such as Layer normalizing (LN). By preventing problems like overfitting or disappearing gradients, they guarantee smoother gradients, stabilize the learning process, and preserve the model's robustness. Better integration of learnt characteristics and more efficient communication between various network components are ensured by the sandwich structure, which helps to balance operations across many levels. follsing are its key functions.

Self-Attention Mechanisms

One of the critical operations in the sandwich layers is the self-attention mechanism, which allows the model to focus on specific areas of the image that are most relevant for INpainting. The attention mechanism helps the model decide which parts of the image should influence the reconstruction of the masked areas, allowing it to make more contextually informed decisions.

Spatial Context

Self-attention ensures that the model takes into account long-range dependencies in the image. For example, in a scene with repeating patterns or similar textures, self-attention can help the model identify distant regions that share characteristics with the masked areas, improving the coherence of the inpainted region.

Feedforward Layers

In addition to self-attention, the sandwich layers typically include feedforward layers that process the features extracted by earlier layers. These layers help the model capture finer details and higher-level features, which are crucial for producing realistic INpainting results. The feedforward layers allow the model to refine the INpainting process by ensuring that the reconstructed regions are consistent with the global structure and local textures of the image.

Enhancing Spatial Understanding

The sandwich layers enable the model to enhance its spatial understanding of the image, helping it understand how different regions relate to each other. This understanding is critical for ensuring that the inpainted regions not only blend visually but also align structurally with the surrounding areas. For instance, if a masked area is part of a larger object, the sandwich layers ensure that the reconstructed region maintains the correct shape, perspective, and texture relative to the rest of the object. Convolution Layers extract essential features from the image, the RELU Activation introduces non-linearity for learning complex patterns, and the Sandwich Layers focus on spatial relationships to ensure that the inpainted regions are coherent and contextually appropriate. Together, these components provide a robust framework for accurately reconstructing missing parts of an image during INpainting tasks.

3.7 MPD (Again)

Mask-Aware Pixel-Shuffle Downsampling is applied again after the sandwich layers. This ensures that even after further processing, the important masked areas are kept intact as the model continues to reduce the resolution and process the image for INpainting.

3.7.1 Purpose of MPD in this Step

The Pixel-shuffle with Mask Awareness In this stage, downsampling (MPD) is essential because it effectively lowers the image's resolution while maintaining the integrity of crucial information in the masked areas. MPD specifically focuses on preserving the structural integrity of the masked areas, in contrast to typical downsampling approaches that may lose important details. This ensures that the INpainting process has enough contextual information to restore the missing sections. Because it keeps crucial features from being lost during the resolution reduction step, this is particularly crucial for tasks requiring high-fidelity reconstruction of the masked regions, enabling the model to produce more accurate and cohesive results.

Preserving Important Information

While the model continues to reduce the image resolution, it is vital that the integrity of the masked areas remains intact. By applying MPD again at this stage, the model ensures that no significant information about the masked parts is lost during the downsampling process.

Maintaining Focus on Masked Areas

The repeated use of Multi-Scale Pyramid Deconvolution (MPD) enables the model to maintain focus on the regions that require inpainting, ensuring that these areas receive the necessary attention during processing. By applying MPD layers at different scales, the model can refine its understanding of both global and local features. This strategy allows the model to effectively reconstruct missing parts of the image while retaining the ability to process the image efficiently as a whole. Consequently, it helps maintain the balance between computational efficiency and high-quality inpainting results in the most critical regions.

3.8 Bottle Neck

In the bottleneck of the model, the deeper layers are responsible for processing and learning more intricate features and abstract representations of the data. These layers capture complex relationships, patterns, and interactions between the features that are often not apparent in the earlier layers. By compressing the feature maps and forcing the model to focus on the most essential information, the bottleneck enables the model to distill a high-level understanding of the data, which is crucial for tasks like classification, segmentation, or inpainting.

3.9 Loss Function

Once the model has generated a prediction (filling in the missing parts), the Loss Function is used to calculate the error between the predicted output and the actual image (ground truth). This loss calculation is crucial for training, as it helps the model learn by minimizing the error over multiple iterations. Custom loss functions ensure that the model effectively learns how to inpaint.

3.9.1 Importance of Loss Calculation

Quantifying Error: The loss function measures how close the predicted image is to the ground truth. This error value informs the model about how well or poorly it has performed in reconstructing the masked regions.

Minimizing Error Over Iterations

The model learns by minimizing this error through optimization techniques (e.g., backpropagation and gradient descent). By iteratively adjusting its parameters based on the loss value, the model becomes better at filling in the missing parts of the image.

Custom Loss Functions

In INpainting, custom loss functions are often used to ensure that the model learns both pixel-level accuracy and higher-level perceptual quality. These specialized loss functions encourage the model to generate inpainted regions that not only match the ground truth but also look realistic and coherent to human observers.

3.9.2 Functionality of MPD at This Stage

Mask-aware Pixel-shuffle Downsampling (MPD) is crucial for effectively compressing the image data while preserving the necessary details within the masked regions. Its functionality ensures that important spatial information from these areas is not lost during the downsampling process, which is often a risk with standard techniques.

Resolution Reduction with Information Retention

This round of MPD ensures that even as the resolution decreases further, key details about the masked areas are retained. This is important because if too much information about the masked areas is lost, the model would struggle to generate accurate INpainting results.

Masked Area Preservation

MPD ensures that the most important features of the masked regions are preserved, allowing the model to maintain its focus on reconstructing these areas accurately as it processes the image through deeper layers.

3.10 Pixel Layers Downsampling (MPD)

Another round of Pixel Layers Downsampling (MPD) is applied, which continues to reduce the image resolution while preserving important information. This step ensures that the model continues to focus on the masked regions even as the image gets smaller.

3.11 Convolution Layers (Again)

The image is passed through additional convolution layers to extract more complex features from the image. These deeper convolutional layers help the model learn intricate patterns, allowing it to make more accurate predictions about the missing parts of the image.

3.11.1 Role of Additional Convolutional Layers

The additional convolutional layers in a deep learning model serve a crucial role in enhancing the model's ability to capture intricate patterns and hierarchical features within the input data. By stacking these layers, the model can progressively extract more complex and high-level representations, moving from basic edge detection in early layers to more abstract features in deeper layers.

Deeper Feature Extraction

These layers focus on extracting higher-order features from the image, including more intricate patterns, textures, and relationships between pixels. By passing the image through these deeper convolutional layers, the model can capture more nuanced details that are essential for producing high-quality INpainting results.

Improved Prediction Accuracy

With the additional convolutional layers, the model becomes better equipped to make accurate predictions about the masked areas. These predictions are informed by the complex patterns and structures the model has learned from both the unmasked and masked regions of the image. The deeper convolution layers provide the model with the ability to make more informed and refined guesses about the content of the missing areas, resulting in more precise and coherent reconstructions.

3.12 RELU Activation (Again)

After deeper feature extraction, RELU activation is applied again. This second application of non-linearity helps the model understand and process complex features that are necessary for reconstructing the image.

3.12.1 Purpose of Second RELU Activation

An extra layer of non-linearity is introduced by the second RELU activation in a model architecture, which is essential for improving the expressiveness and capacity of the model to learn intricate patterns. The network can handle more complex associations in the data by further refining the information it analyzes by applying RELU once more. This second activation facilitates more nuanced decision-making by assisting the model in maintaining smooth gradients, which avoids sudden shifts in output. The second RELU makes sure that deeper layers can efficiently collect and process intricate feature interactions in situations like picture INpainting or transformers, producing more reliable and precise outputs.

Non-Linear Transformation

By applying RELU again, the model is better able to learn and process complex relationships between the features extracted by the deeper convolutional layers. Non-linearity allows the model to capture subtler patterns that a linear function might miss.

Enhancing Model Flexibility

The second application of RELU further increases the model's flexibility in handling diverse image features. This flexibility is especially important in INpainting, where the model needs to reconstruct regions that may involve intricate textures, lighting variations, or other complex image properties. This second round of RELU activation enhances the model's ability to learn and apply more complex transformations to the image, ultimately leading to more accurate and realistic INpainting.

3.13 Loss Function (Again)

Another loss function is applied after this stage to evaluate how well the model has performed in reconstructing the image. This second loss calculation allows the model to adjust its parameters further and improve its accuracy during training.

Role of Second Loss Calculation

By adding another level of feedback, the second loss calculation is essential to improving the model's learning process. The second loss computation can be customized to target particular facets of the model's performance, such as fine-grained details, texture consistency, or boundary accuracy, whereas the primary loss function usually concentrates on global goals like overall accuracy or reconstruction quality. In addition to minimizing mistakes globally, this multi-step loss calculation makes sure that the model takes into account localized areas or particular features that are essential for producing high-quality outcomes. This second loss can increase the realism of the inpainted regions and improve the model's capacity to produce outputs that are coherent and contextually accurate in tasks like image INpainting.

Evaluation of Reconstruction Quality

This second loss calculation allows the model to evaluate the quality of the INpainting after the deeper layers have processed the image. By comparing the output with the ground truth, the model can adjust its parameters to improve the reconstruction.

Further FineTuning

The error calculated by the loss function is used to fine-tune the model's weights, ensuring that the model continues to improve its accuracy with each iteration of training. This helps in producing inpainted results that are not only structurally accurate but also visually coherent.

3.14 Final Sandwich Layers

These final Sandwich Layers further process the extracted features and help the model generate the final inpainted image by considering spatial and feature information from earlier stages. These layers ensure that the final output is coherent and consistent with the original nonmasked parts of the image.

3.14.1 Final Feature Refinement

To ensure that the learnt features are optimized and fine-tuned for optimum accuracy, the final feature refinement stage is essential for refining the model's output. In order to improve the clarity and consistency of elements like textures, edges, and colors, the model now modifies and enhances the finer details of the processed data. Especially for activities like image INpainting or high-resolution image production, this refinement process helps remove any lingering artifacts, guaranteeing seamless transitions and more realistic outputs. The model's overall performance is enhanced by fine-tuning the final characteristics, which guarantee that the output not only matches the input data but also satisfies the required quality and accuracy standards.

Spatial and Feature Consideration

These layers ensure that the final reconstructed image maintains both local feature accuracy (like textures and edges) and global coherence (ensuring the inpainted area blends seamlessly with the surrounding regions).

Final Touches on Reconstruction

These layers apply the final touches to the model's INpainting process, ensuring that the missing parts of the image are reconstructed in a way that is visually consistent with the original image.

By refining the feature representations one last time, the final sandwich layers ensure that the inpainted areas are as accurate and seamless as possible.

3.15 Output (Evaluation Layer)

Finally, after all these stages, the model outputs the reconstructed image, where the missing regions have been filled in. This output is evaluated against the original image to assess the quality of INpainting, and the process is repeated to refine the model's accuracy.

Final Evaluation

After training, the final evaluation step is essential for determining the model's overall performance. At this stage, the accuracy, consistency, and generalization capacity of the model are assessed by comparing its outputs to a collection of ground truth data or predetermined benchmarks. Depending on the job, this assessment usually entails computing a number of performance metrics, including (FID), (SSIM), (PSNR). The final evaluation guarantees that the model maintains the contextual and semantic integrity of the full image while also realistically filling in the empty areas in image INpainting or creation tasks. It assists in identifying any shortcomings or prospective areas for development, directing possible modifications for additional model optimization or fine-tuning.

Output Generation

The model outputs the inpainted image, which is the model's best attempt at filling in the missing or corrupted regions of the original image.

Evaluation Against Ground Truth

The output is then evaluated against the ground truth image to assess the quality of the INpainting. The quality of the final output determines how well the model has learned the task of INpainting and whether further training or adjustments are needed. This output is the culmination of the model's learning and processing stages, and it is continuously evaluated to refine the INpainting performance.

3.16 Implementation Details

The proposed scheme uses specialized neural network architectures, electrical engineering, and data processing to implement pipeline systems. Below is a detailed description of the methods used in the research papers and their functions.

3.17 Network Arhitecture

The network architecture, which specifies how its layers, constituents, and functions are arranged and interact, is the fundamental structure of a deep learning model. Convolutional layers for feature extraction, transformer or attention mechanisms for capturing long-range dependencies, and activation functions like RELU to introduce non-linearity are some of the fundamental building blocks that are frequently found in a typical architecture created for tasks like image INpainting or generation. Specialized elements like encoder-decoder structures, in which the encoder condenses the input data into a latent representation and the decoder meticulously reconstructs the output, may also be incorporated into the architecture. Furthermore, sophisticated methods like residual blocks or skip connections are frequently used to alleviate problems like disappearing gradients and preserve spatial information between layers.This is shown in Figure 5.1.

3.17.1 Languages and Frameworks

we used Python as the main programming language due to its simplicity and the availability of powerful machine learning libraries. Here's a breakdown of the key libraries we used:

Matplotlib

To plot the model's performance, including accuracy and loss throughout each epoch, we use Matplotlib. In order to visually evaluate how well the model is performing, it also allows us to compare and show photographs that were taken before and after INpainting.

TensorFlow/Keras

TensorFlow is ideal for configuring and training neural networks because of its Keras API. It offers a sophisticated, user-friendly interface without sacrificing capability. We can define custom layers and models with Keras, which is what we require for our project. With layers like convolutional layers, attention mechanisms, and custom layers (like the MPDModule), we constructed our own neural network model.

NumPy

NumPy is a tool for effective data handling. Large arrays can be managed with it, and numerical operations like bending picture data or normalizing pixel values can be carried out. We used NumPy to preprocess the image input by scaling, normalizing, and generating image masks before feeding it into the model. It's the first tool we used when we needed to do any kind of numerical data processing.

Methods and Modules

create_dir(dir) Method: We had to ensure that the appropriate folders were created before saving any models or results. If a directory doesn't already exist, it is created using this procedure. To ensure that everything is preserved correctly, we use this way to arrange and save data, such as model checkpoints or the output of INpainting jobs.

Downsampling

When managing images with masks during INpainting, this module is crucial. It guarantees that significant spatial information is preserved while downsampling of photos.In our neural network, we employ the MPDModule as a custom layer to control the downsampling of masked images. By combining pixel shuffling and convolution techniques, it makes sure that the image's structure is preserved even when processing masked areas.

SCAL (Spatially-activated Channel Attention Layer)

By collecting spatial dependencies and channel interactions, this layer aids the model in concentrating on the most significant elements in the picture. In our model, SCAL is used to apply attention mechanisms that let the model know "what" and "where" to focus on during INpainting by allowing the model to prioritizspecific portions of the image.

3.17.2 Component Breakdown and Descriptions

The component breakdown of a deep learning model provides a detailed view of the individual parts that make up the network and their specific roles. Key components include convolutional layers, which extract essential features like edges, textures, and patterns from the input data, and pooling layers, which reduce dimensionality while retaining important information.

config.py

This file functions as the configuration manager for the project, consolidating all settings required by various modules. It includes critical parameters such as hyperparameters, dataset paths, and training configurations, ensuring consistency and ease of modification throughout the development process. By centralizing these configurations, it simplifies the management of experiments and promotes a more organized workflow.

Components: The model hyperparameters define key tunable parameters such as learning rates, batch sizes, epochs, and image sizes, all of which significantly impact the training process and performance of the model. These parameters need careful tuning to optimize the learning process and prevent issues like overfitting or underfitting. Additionally, the dataset paths specify the locations of the training and validation datasets, ensuring that the model can access the correct data during each stage of training. The training settings include configurations such as early stopping criteria, validation splits, and learning rate schedules, which control the progression and adaptation of the training process to achieve the best results. Finally, miscellaneous configurations consist of other constants and parameters used across the project, ensuring consistency and functionality across different parts of the implementation. Together, these settings are crucial for the smooth and efficient operation of the training pipeline.

hint.py

This document presents the HINT (High-quality INpainting Transformer) model, a core framework developed for advanced image inpainting tasks. The HINT model leverages transformer architectures to achieve high-quality restoration of missing image regions, making it ideal for applications that require seamless and realistic image completion. By integrating attention mechanisms and deep learning techniques, HINT effectively captures both local and global context, producing superior inpainting results compared to traditional methods.

Components: The MPDModule implements the Mask-aware Pixel-shuffle Downsampling (MPD), which utilizes pixel shuffle techniques to enhance the downsampling process in image INpainting tasks. This method ensures that important masked regions are preserved during downsampling, improving the quality of INpainting. The SCAL, or Spatially-activated Channel Attention Layer, is responsible for introducing spatially aware attention mechanisms that enhance feature extraction during the INpainting process, allowing the model to focus on critical regions of the image. The Model Training module implements the training loop and other utilities needed for managing the training of the HINT model, including the integration of loss functions and optimizers, ensuring efficient and effective model training.

dataset.py

This module is responsible for handling data loading, preprocessing, and augmentation, which are crucial steps in preparing images for neural network training. By efficiently managing these processes, the module ensures that the input data is properly structured and enhanced, facilitating effective learning during model training. The augmentation techniques applied help improve the model's robustness by simulating various transformations and variations in the input images.

Components: The MPDModule implements Mask-aware Pixel-shuffle Downsampling (MPD), which enhances the downsampling process in image INpainting tasks by preserving important masked regions using pixel shuffle techniques. This ensures better quality in the reconstruction of missing areas. The SCAL (Spatially-activated Channel Attention Layer) introduces spatially aware attention mechanisms that improve feature extraction, allowing the model to focus on crucial regions during INpainting. Additionally, the Model Training module manages the entire training process of the HINT model, including the implementation of loss functions and optimizers, ensuring efficient and effective training.

model.py

This section includes the definitions of multiple model architectures tailored for tasks such as image inpainting. These architectures are designed to address different aspects of image restoration, enabling the models to effectively reconstruct missing or damaged regions in images. Each architecture is optimized for specific use cases, ensuring flexibility and performance across a variety of inpainting scenarios.

Components: Custom CNN Model Defines a custom Convolutional Neural Network (CNN) with several layers (Conv2D, MaxPooling, Dense, Flatten, etc.) to handle image classification tasks or image INpainting tasks. SCAL and MPD ModulesUsed in specialized tasks such as image INpainting, enhancing spatial awareness and feature extraction.

loss.py

This section includes the definitions of multiple model architectures tailored for tasks such as image inpainting. These architectures are designed to address different aspects of image restoration, enabling the models to effectively reconstruct missing or damaged regions in images. Each architecture is optimized for specific use cases, ensuring flexibility and performance across a variety of inpainting scenarios.

Components: Custom Loss Functions: Loss functions like cross-entropy or custom variations that handle the specific needs of image INpainting. They calculate the variance between predicted and true outputs, typically using Mean Squared Error or Categorical Cross-Entropy.

metrics.py

This section defines custom loss functions that play a critical role in training neural networks. These loss functions are specifically tailored to the task at hand, guiding the model's learning process by measuring the difference between predicted and actual outputs. By incorporating domain-specific metrics, the custom loss functions help improve the accuracy and performance of the neural networks during training, ensuring more effective optimization for tasks like image inpainting and beyond.

Components: PSNR (Peak Signal-to-Noise Ratio): This metric is widely used for evaluating the performance of INpainting models by comparing the similarity between the predicted image and the original image.

Other Metrics

Precision, Recall, and the Structural Similarity Index (SSIM) are implemented as key metrics to monitor the quality of image inpainting. Precision and Recall provide insights into the model's ability to accurately fill missing regions without introducing unnecessary artifacts, while SSIM measures the perceptual similarity between the inpainted image and the original, focusing on structural details such as texture and contrast. Together, these metrics offer a comprehensive evaluation of the inpainting quality, ensuring both pixel-level accuracy and visual coherence.

networks.py

This section defines more advanced network architectures, including residual blocks, attention mechanisms, and transformer-based models. Residual blocks help improve training stability and allow for deeper networks by addressing the vanishing gradient problem. Attention mechanisms enhance the model's ability to focus on important regions of the image, enabling better context capture for tasks like inpainting. Transformer-based models leverage self-attention to process global dependencies within the image, making them highly effective for complex image restoration tasks. These architectures collectively enable the model to achieve high performance in challenging scenarios.

Components: Residual Networks (ResNet), Implements ResNet blocks to enhance the deep learning model's capability by utilizing shortcut connections. Attention Mechanisms: Multi-head attention mechanisms are incorporated to capture spatial dependencies across the image for tasks like INpainting. Sandwich Block: Combines attention, normalization, and feedforward layers inspired by transformer architectures.

utils.py

This section provides utility functions designed to assist with various aspects of image processing, such as loading, saving, and augmenting images. Additionally, these functions support progress visualization during training, offering real-time feedback on the model's performance. By simplifying tasks like image preprocessing and result tracking, these utilities streamline the workflow, ensuring efficient handling of data and helping monitor the model's improvement over time.

Components: Image Processing Functions: Functions like create_mask (to generate masks for INpainting tasks), imshow (to display images), and imsave (to save images to disk). Progress Bar: Implements a progress bar for monitoring training and evaluation.

main.py:

This section integrates the various components of the project, including dataset loading, model construction, training, and evaluation. It acts as the central framework that coordinates the flow of data and processes through the entire pipeline. Additionally, it includes functionality for hyperparameter tuning, allowing for optimization of model performance by adjusting key parameters during training. This unified structure ensures that all modules work seamlessly together, enabling efficient experimentation and model refinement.

Components: The Model Definition utilizes architectures defined in the model.py file, while also supporting custom CNN implementations. The Training and Evaluation process involves implementing training loops, computing losses, and evaluating performance metrics using datasets loaded through ‘dataset.py‘. Hyperparameter Tuning is managed by Keras Tuner, allowing for the optimization of hyperparameters such as the number of filters, kernel size, activation functions, and optimizers. Together, these files form a comprehensive machine learning pipeline.dataset.py handles data loading, model.py defines the network structure, while loss.py and metrics.py compute performance metrics, and main.py coordinates the overall training and evaluation processes. Additional layers and mechanisms defined in ‘hint.py‘ and networks.py further enhance the model’s performance in tasks such as image INpainting by incorporating specialized techniques like attention and pixel shuffling.

This chapter presents the High-quality Inpainting Transformer (HINT) model, which aims to overcome the limitations found in previous image inpainting methods. The proposed model utilizes a hybrid approach by integrating the capabilities of transformers with Convolutional Neural Networks (CNNs). It introduces two primary innovations: the Spatially Activated Channel Attention Layer (SCAL) and the Mask-aware Pixel-shuffle Downsampling (MPD) module.

SCAL improves the model's ability to focus on both spatial and channel-specific information, while MPD ensures minimal data loss during downsampling, thereby preserving important image details. By combining the long-range dependency handling power of transformers with the local feature extraction strengths of CNNs, HINT enhances texture consistency, accelerates training, and achieves more accurate inpainting results. The chapter demonstrates how these advancements enable the model to achieve superior performance in real-world inpainting tasks.

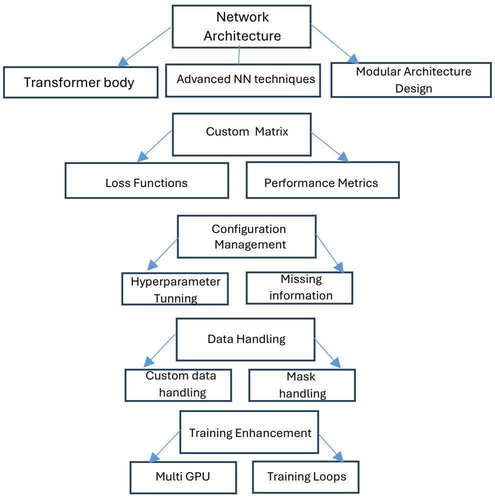

Figure 3.2: Network Arhictecture

CHAPTER 4

Methodology

We provide a thorough assessment of the Improved HINT framework in this section. Firstly we perform thorough ablation tests, methodically assessing each suggested HINT component's importance. With the use of these investigations, We examine the relative contributions of each component, highlighting their significance for the model's overall performance and defending their inclusion in the final layout. The comprehensive assessment highlights the effectiveness of the Improved HINT framework and its potential to propel the field forward.

4.1 Datasets and Experimental Setups

To evaluate both HINT Initial and HINT Optimized, we utilize two widely recognized benchmark datasets: Places2-Standard [36] and CelebA-HQ [13], ensuring consistency by conducting all experiments using 256×256 resolution images. The Places2-Standard dataset comprises a broad array of diverse scenes, providing a challenging test bed for models to generalize across various environmental contexts. On the other hand, the CelebA-HQ dataset is focused on high-quality images of human faces, offering a specialized domain for evaluating facial image IN-painting and restoration. For the CelebA-HQ dataset, we use 28,000 images for training and

2,000 images for testing, while for the Places2 dataset, we employ the standard training and testing splits defined by the dataset. In both datasets, irregular masks are applied to simulate missing or corrupted regions, which is a common setting in image INpainting tasks. The masked image, denoted as IM =I . M, is combined with the original image I to form the input image for the model, referred to as Input. The Improved HINT model processes this input to produce the restored or inpainted output image, denoted as IC, which is formulated as IC = HINT(Input). Our model architecture follows the structure of the original HINT framework, but we introduce several key enhancements aimed at improving performance and generalization. These improvements include optimizing the network's convolutional layers for more efficient feature extraction, employing advanced attention mechanisms to better focus on critical regions of the image, and integrating multi-scale loss functions to ensure more accurate INpainting across varying levels of image detail. Additionally, we incorporate a refined training procedure that leverages dynamic mask generation, ensuring robustness in handling a wider variety of image occlusions. These enhancements contribute to a more effective and versatile INpainting model, which is thoroughly evaluated in the subsequent sections. In addition to Places2-Standard and CelebA-HQ, several other datasets were considered for evaluating the HINT Initial and HINT Optimized models. However, after careful analysis, these datasets were ultimately deemed less feasible for this study due to various limitations or mismatches with the objectives of the evaluation. Below, we outline some of these datasets and the reasons they were not selected.

4.1.1 ImageNet

With over 14 million photos in 1,000 different categories, ImageNet is one of the biggest and most varied image collections available.

Justification for Exclusion:

Despite having a large image collection, ImageNet was not the best choice for this investigation since it prioritizes item classification above scene reconstruction and image INpainting. Moreover, the size and quality of ImageNet images vary widely, which complicates the use of continuous training with 256×256 images. CelebA-HQ's low usefulness in assessing face image INpainting tasks was also caused by the absence of unique high-resolution photos that resembled it.

4.1.2 Large-scale Scene Understanding, or LSUN

The LSUN dataset is widely used for image production tasks and contains large-scale scene data across several categories, including bedrooms, churches, and classrooms.

Justification for Exclusion:

Like Places2, LSUN offers an abundance of high-quality scene data. Its main focus, meanwhile, is on particular scene types, like interior situations, which don't have the diversity required to thoroughly evaluate broad INpainting models. Places2 already covers a wider range of situations, hence LSUN was considered less relevant for this evaluation and redundant.

4.1.3 COCO

COCO (Common Objects in Context) is a widely recognized and extensively used dataset in the fields of object recognition, categorization, and labeling. It consists of over 330,000 images featuring a diverse array of objects in complex, real-world settings. The dataset is valuable for training models to recognize and classify objects in varied environments, as it includes annotations such as object segmentation, keypoints, and captions. COCO's rich diversity and complexity make it a benchmark for evaluating the performance of models in challenging computer vision tasks.

Justification for Exclusion:

While COCO performs well on segmentation tests and object detection, picture INpainting is not a good fit for its primary use case, especially when testing on extremely diversified sceneries or high-resolution human faces. Additionally, COCO images often contain several overlapping items and complicated situations, which might inject noise and unpredictability into the INpainting operation, making it less practicable for the controlledevaluation needed here.

4.1.4 Flickr-Faces-HQ

FFHQ 70,000 high-quality photos of human faces make up the FFHQ dataset, which is comparable to CelebA-HQ but has a wider range of age, race, and image circumstances.

Justification for Exclusion

While FFHQ is very important for INpainting facial images, training with a 256×256 resolution—which was used for consistency across all datasets—presented difficulties due to its bigger image sizes, which are typically 1024 × 1024. Reducing the size of these photos would lead to a notable reduction in detail, which would render them less useful for assessing intricate INpainting assignments. Furthermore, for the current experimental setting, using CelebA-HQ offered a more balanced trade-off between quality and computing practicality.

4.1.5 ADE20K

ADE20K is a scene parsing dataset with more than 20,000 photos annotated for 150 different object types.

Justification for Exclusion:

While ADE20K performs exceptionally well in semantic segmentation and scene interpretation, Places2 outperforms it in terms of size and scope when it comes to the INpainting challenge. The model's capacity to generalize to new data may also be limited by the comparatively smaller dataset size, particularly with regard to scene diversity, which was a crucial component of this investigation. Furthermore, ADE20K's emphasis on dense annotations for segmentation introduces needless complexity to an image restoration assignment.

4.1.6 Facial Landmark Datasets (such as AFLW and 300W)

These datasets, which are primarily concerned with facial landmark identification, frequently contain annotated photos that have important spots identified by face features.

Justification for Exclusion:

Facial landmark datasets are useful for tasks related to landmark identification and facial recognition, but they are not intended for high-resolution image production or image INpainting. These datasets are not wellsuited for assessing the INpainting performance of the HINT model, which necessitates big,high-resolution images like as those found in CelebA-HQ, because they concentrate on exact landmarks rather than overall image quality and reconstruction.

4.1.7 Oxford Florists and Pets

The Oxford Pets and Flowers databases comprise pictures of different pet breeds and flower species; these photos are frequently utilized for image segmentation and classification applications.

Justification for Exclusion:

These datasets, which are very small, concentrate on particular categories (pets and flowers), which are not in line with the study's larger goals. They wouldn't offer the diversity or complexity necessary to sufficiently evaluate the HINT model across a broad range of real-world circumstances because of their constrained variety and narrow emphasis.The selection of the Places2-Standard and CelebA-HQ datasets for evaluating the HINT Initial and HINT Optimized models offers several key benefits when compared to other available datasets. These benefits stem from the unique characteristics of each dataset, which align closely with the objectives of this study: ensuring diversity in scene reconstruction tasks and high-quality, domain-specific image INpainting for human faces. Below are the detailed advantages of choosing Places2 and CelebA-HQ over other datasets.

4.1.8 Diversity and Scale in Places2-Standard Wide Range of Scenes

The Places2-Standard dataset includes over 10 million images spanning more than 400 unique scene categories, ranging from natural landscapes to urban environments, interiors, and more. This immense diversity allows the model to be tested across a broad spectrum of real-world scenarios, ensuring that the HINT model is not limited to a narrow range of contexts. Compared to more specialized datasets like LSUN, which focuses on specific scene categories (e.g., bedrooms, churches), Places2 provides a broader and more comprehensive evaluation of a model's generalization ability in scene INpainting.

4.1.9 Consistency in Resolution

All the images in Places2-Standard are available at a resolution that matches the 256×256 size used in the evaluation. This consistency ensures that there is no need for aggressive downsampling, which can degrade image quality or omit crucial details. In contrast, larger datasets like ImageNet contain images of varying resolutions, which would require significant preprocessing, potentially affecting the fidelity of the final results.

4.1.10 Challenging Occlusions and Masks

Places2 provides an excellent test bed for applying irregular masks, which simulate occlusions or missing parts in images. Given the diverse nature of the scenes in Places2, the masked regions may vary significantly in terms of complexity and context making the INpainting task more challenging and realistic. This variety is not as well-represented in datasets like COCO, where images typically feature multiple objects in cluttered settings, which may complicate mask application without necessarily providing meaningful insights into scene reconstruction.

4.1.11 Large-Scale Dataset for Robust Training

With millions of images available, Places2 offers ample data for training models, enabling robust learning and reducing the risk of overfitting. This is a significant advantage over smaller datasets like ADE20K, which, while useful for semantic segmentation, does not offer the volume of data required to thoroughly evaluate the generalization capabilities of an INpainting model like HINT. The larger scale of Places2 ensures that the model encounters a wide variety of occlusions, textures, and structural patterns during training, making it better equipped to handle unseen data.

4.1.12 Specialized Facial Image INpainting with CelebA-HQ High-Quality, High-Resolution Human Faces

CelebA-HQ is one of the leading datasets for high-quality facial images, consisting of 30,000 images at a 1024×1024 resolution, which can be consistently downsampled to 256×256 for this study. These images retain excellent facial detail, which is crucial for fine-grained tasks like facial INpainting. In contrast, other facial datasets such as FFHQ, while similarly high in quality, operate at larger resolutions (e.g., 1024×1024 or even 2048×2048) that would need to be significantly downsampled, potentially losing important details that are critical for INpainting tasks Balanced and Controlled Dataset Size: CelebA-HQ contains a curated and manageable dataset size, with 28,000 images for training and 2,000 for testing. This allows for effective model training without the computational overhead required for larger datasets like FFHQ. The controlled size of CelebA-HQ ensures high-quality facial images without overwhelming the model with redundant or noisy data, as could occur with smaller, less curated datasets.

4.1.13 Diverse Attributes for Testing Model Robustness

CelebA-HQ is annotated with a wide range of facial attributes such as age, gender, and expressions, which add complexity and diversity to the INpainting task. The dataset includes various ethnic backgrounds, hairstyles, and lighting conditions, making it an excellent choice for testing how well the HINT model generalizes across different types of human faces. This variety is more comprehensive than in other datasets like the Oxford Pets or Flowers datasets, which focus on specific categories with less variation in visual attributes. Realistic Occlusions with Irregular Masks: Applying irregular masks to CelebA-HQ images provides a realistic simulation of corrupted or missing facial regions. Since the human brain is particularly sensitive to facial features, the task of restoring occluded faces becomes highly challenging and crucial for testing the model's effectiveness. CelebA-HQ's high-resolution, detailed images make it ideal for this task, providing a rigorous test environment that may not be achievable with datasets such as facial landmark datasets (e.g., 300W or AFLW), which focus on landmark detection rather than image restoration.

4.1.14 Focus on Identity Preservation

One of the critical challenges in facial image INpainting is preserving the identity and fine details of the face, such as texture, skin tone, and facial structure. CelebA-HQ excels in this domain due to the high fidelity of its images. In contrast, other datasets like COCO or ImageNet contain a mix of objects and scenes, which are not specifically tailored to facial detail preservation. By using CelebA-HQ, the evaluation can focus on INpainting that preserves crucial facial characteristics, ensuring that the model generates realistic and accurate reconstructions.

4.1.15 Complementary Nature of Places2 and CelebA-HQ Coverage of Both Scene and Object Level INpainting

The combination of Places2 and CelebA-HQ allows for a comprehensive evaluation of the HINT model across both scene-level and object-level (in this case, facial) INpainting tasks. Places2 focuses on restoring large-scale, complex environments, while CelebA-HQ targets high-resolution, fine-grained facial INpainting. This dual focus ensures that the model is tested on a wide variety of image types, demonstrating its versatility and effectiveness across different domains. Other datasets either focus on just scenes (e.g., LSUN, ADE20K) or just faces (e.g., FFHQ, 300W), but few offer the same breadth as the combination of Places2 and CelebA-HQ.

4.1.16 Realism in Reconstruction Tasks

Both datasets offer a high degree of realism in their respective domains. Places2 features real-world scenes with natural diversity in lighting, textures, and objects, making the INpainting task more applicable to real-world applications. CelebA-HQ, with its focus on high-quality human faces, ensures that the reconstructed facial images are highly realistic and faithful to the original identities. This balance of realism in both datasets provides a more practical evaluation compared to datasets like ImageNet, which, although diverse, focuses more on object classification than realistic INpainting.

4.1.17 The Transformer Body

The HINT model's transformer body is built on a multi-level architecture that combines a number of essential parts to produce high-quality image INpainting.

Overview of the Transformer Body

The proposed Spatially-activated Channel Attention Layer (SCAL) is encapsulated in seven transformer blocks, each of which is made up of several "sandwiches". These sandwiches play a crucial role in controlling data flow consistency during downsampling by integrating the Mask-aware Pixel-shuffle Down-sampling (MPD) module and representing local and global dependencies in an equitable manner. The primary self-attention mechanism known as SCAL (Spatially-activated Channel Attention Layer) was created to improve the capacity to represent long-range dependencies between feature patches in both the spatial and channel dimensions. This capacity is essential for capturing the complex relationships found in incomplete images, or images with irregular masks. SCAL uses spatial attention to capture the significance of "where" these features are positioned in the image and channel-wise self-attention to emphasize significant aspects. The model can effectively manage masked areas with complicated shapes and irregularities thanks to its dual-branch attention mechanism, all without considerably raising the computing overhead. Sandwich Transformer Block: A sandwich-shaped transformer block serves as the foundation for the transformer body. The "sandwich" structure is formed by the SCAL being positioned in between two Feed-forward Networks (FFNs). Though modified for the INpainting purpose, this architecture draws inspiration from voice recognition architectures and aids in the model's effective acquisition of both local and global information. Prior to the input features being sent to SCAL, the first FFN in the sandwich filters them, enabling the attention mechanism to work with more accurate and insightful data. The second FFN processes the attention-modulated features further to extract useful image representations for the subsequent layers, following SCAL's computation of the attention maps. In order to reconstruct high-quality images, this structure maximizes the learning of spatially-aware and channel-aware features.

Important characteristics

MPD for Reduction in Size: This technique adds mask-aware capability to the conventional pixel-shuffle down-sampling (PD) process. When downsampling,MPD helps keep the masked and unmasked regions' positional consistency, which lowers the possibility of pixel drifting and information loss.

4.1.18 SCAL for Attention

SCAL's channel and spatial dual-branch architecture improves the model's capacity to capture inter-channel dependencies while maintaining spatial awareness, which is essential for creating coherent and contextually accurate image reconstructions.

Sandwich Block

The model guarantees that the attention mechanism is applied to the most pertinent characteristics by embedding SCAL between two FFNs. The FFNs then further refine these features for the development of high-quality output. This transformer body allows for effective representation learning with little information loss, especially when MPD and SCAL are integrated within a sandwich structure.

4.1.19 Awareness of masks Downsampling using pixel-shuffle (MPD)

In order to prevent pixel drifting and maintain positional consistency, which happens during conventional Pixel-shuffle Down-sampling (PD), MPD is a cutting-edge downsampling technique. Since the input already has significant portions that are masked or distorted, it is imperative to prevent information loss throughout the INpainting process. The visible portions of the picture are guaranteed to have consistent and aligned characteristics across all channels by MPD.

Mask Awareness

The input picture and mask are projected into feature spaces by the model, which processes both simultaneously. The mask records whether areas of the picture are legitimate and which are corrupted (missing).

4.1.20 Spatially-activated Channel Attention Layer (SCAL)

The SCAL module improves upon existing attention mechanisms by enabling the model to capture inter-channel dependencies while also maintaining spatial awareness. Here's how SCAL operates:

Channel Self-Attention

The attention mechanism focuses on the relationships between different channels, ensuring that important features across channels are highlighted. This is particularly beneficial for high-resolution images where capturing long-range dependencies is crucial.

Spatial Awareness In traditional channel attention mechanisms, there is a lack of emphasis on"where" important features are located spatially. SCAL introduces a spatial attention branch, using convolutions to capture spatial dependencies and integrate them with the channel attention to form a cohesive representation. This allows the model to account for the global spatial context, which is critical for INpainting tasks where irregular masks create complex spatial dependencies.

4.1.21 Sandwich-shaped Transformer Block

The Channel Attention Layer with Spatial Activation (SCAL) By allowing the model to preserve spatial awareness and capture inter-channel dependencies, the SCAL module enhances current attention techniques. This is how SCAL functions: Channel Self-Attention: This attention mechanism makes sure that significant aspects are highlighted by concentrating on the connections between various channels. This is especially useful for high-resolution photos when it's important to capture interdependence across vast distances. Spatial Awareness: The "where" of key features in space is not given as much weight as it should in typical channel attention methods. SCAL presents a spatial attention branch that forms a coherent representation by integrating spatial dependencies with the channel attention through the use of convolutions. This enables the model to take into consideration the global spatial context, which is essential for tasks involving INpainting.

4.1.22 Loss Functions

To ensure high-quality INpainting results, the model uses a combination of multiple loss components, each serving a specific purpose:

L1 Loss (Contextual Reconstruction)

This loss is concerned with making sure the pixel values of the reconstructed image correspond to those of the ground truth image. It penalizes the pixel-by-pixel variations between the original and inpainted output directly.

Style Loss (Lstyle)

This loss assesses how well the inpainted image matches the ground truth image in terms of style. It guarantees that the model generates outputs that are both accurate and consistent in terms of style by emphasizing texture and stylistic variances.

Perceptual Loss (Lperc)

From a pre-trained network (such as a VGG network), high-level feature representations are used to calculate this loss. It guarantees that the inpainted image, particularly with regard to details like edges, textures, and shapes, seems perceptually close to the original.

Adversarial Loss (Ladv)

This loss is a result of the model learning to produce outputs that are identical to real images through the use of Generative Adversarial Networks (GANs). The overall realism and quality of the generated images are improved by the adversarial loss. These parts make up the total loss function, which is a weighted sum and enables the model to balance several factors of

INpainting quality.

4.1.23 Custom Convolutional Neural Network (CNN)

The model incorporates a customized CNN with tuneable parts, including activation functions, kernel dimensions, and filter sizes. The following are the functions of CNN:

Feature extraction

Convolutional layers are utilized to extract critical features from the image, including edges, textures, and forms. These layers are designed to detect spatial hierarchies in the image, enabling the model to identify low-level features such as edges and more complex structures like textures and shapes. By capturing these essential details, the model can better understand the underlying patterns and structure of the image, which is crucial for tasks like image inpainting where maintaining visual coherence is key.

Downsampling

To reduce the spatial dimensions of the feature maps and control computational complexity, downsampling techniques like max-pooling are employed. Max-pooling works by selecting the maximum value from a specified window of the feature map, effectively summarizing the most important features within that region. This process not only reduces the dimensionality but also helps the model focus on the most salient aspects of the image, improving both the efficiency and generalization capability of the network without losing critical spatial information.

Regression or classification

For tasks involving regression or classification, dense layers smooth down the retrieved characteristics before sending them via fully linked layers. TensorFlow is used in the implementation of the SCAL and MPD modules, enabling effective management of large-scale datasets such as CelebA, CelebA-HQ, and Places2. The big, erratic missing regions that are frequently encountered in INpainting assignments are intended to be handled by this customized CNN. Together, MPD, SCAL, the sandwich block, loss functions, and the custom CNN form a strong and effective model that can inpaint images with excellent quality. SCAL captures both spatial and channel-wise dependencies, ensuring a comprehensive comprehension of the image, while MPD guarantees the preservation of correct information when downsampling. The loss functions assist in balancing the various facets of the INpainting work, while the sandwich block refines characteristics at different stages.

4.2 key Improvements

The key improvements in our Improved HINT methodology, as outlined in the Project Improvements Details document, focus on enhancing various aspects of the INpainting model pipeline, from configuration management to advanced neural network techniques. Here are the main improvements and their corresponding benefits:

4.2.1 Configuration Management

Custom YAML-based Configuration

Introduced flexibility by using a YAML configuration file for managing hyperparameters and model settings. This centralization simplifies tuning parameters and maintaining configurations without altering the core code base.

Default Fallbacks

The configuration system is designed with default values to prevent runtime errors and ensure smooth operation. These default settings act as fallbacks, allowing the model to function properly even if certain configuration parameters are missing or incorrectly specified. By providing sensible default values for key settings, the system enhances the model's robustness and reliability, preventing crashes or disruptions during execution and making it easier to test and deploy the model under various conditions.

4.2.2 Efficient Data Pipeline

Custom Dataset Handling

Optimized data loading by implementing a custom 'Dataset' class for efficient preprocessing and loading of images and masks. It supports TensorFlow optimizations like batching and prefetching, improving the pipeline's scalability.

Mask Creation Enhancements

Multiple mask generation techniques have been introduced, including random blocks and external masks, to assess the model's robustness under different conditions. These varied approaches to masking allow the model to be tested on a broader range of scenarios, simulating real-world situations where missing or occluded parts of an image may differ in structure and complexity. By incorporating such diverse masking strategies, the model's ability to handle a variety of image inpainting challenges is thoroughly evaluated, ensuring it performs well across different types of incomplete data.

4.2.3 Advanced Neural Network Techniques

Spectral Normalization:

Added spectral normalization to the model, particularly the discriminator, to stabilize training and improve generalization by controlling the Lipschitz constant during optimization.

Custom Layer Normalization

Both bias-free and with-bias layer normalization options have been provided to enhance flexibility and improve performance during model training. The bias-free option eliminates additional parameters that may complicate training, while the with-bias option introduces learnable bias terms, potentially allowing the model to capture more nuanced patterns in the data. By offering both options, the model can be fine-tuned based on the specific needs of the task, balancing efficiency and accuracy for optimal performance in various training scenarios.

4.2.4 Modular Architecture Design

Separation of Generator and Discriminator

Created distinct classes for the generator (HINT) and discriminator to enhance maintainability and scalability. This modularity allows for easy swapping and customization of model components without disrupting the entire architecture.

Sandwich Transformer Block

A sandwich block architecture has been introduced, incorporating attention mechanisms and feedforward networks to effectively capture long-range dependencies within images. This innovative design significantly improves the inpainting quality by allowing the model to focus on both local details and global context. The attention mechanism helps the model prioritize important image features, while the feedforward networks enhance the learning of complex patterns, leading to more accurate and seamless image restoration.

4.2.5 Loss Functions and Custom Metrics

Multi-Loss Function Integration

Multiple loss functions have been incorporated into the model, including L1 loss, perceptual loss, style loss, and adversarial loss. This combination enables the model to strike a balance between pixel-level accuracy and perceptual quality. L1 loss ensures fine-grained pixel-wise similarity, while perceptual loss focuses on preserving high-level features and texture information. Style loss helps maintain visual consistency by preserving stylistic elements, and adversarial loss improves the model's ability to generate realistic and natural-looking images. Together, these losses enhance both the fidelity and the visual appeal of the inpainted images.

Custom PSNR Metric

A custom Peak Signal-to-Noise Ratio (PSNR) metric has been implemented to monitor the quality of image inpainting. PSNR is a widely used metric for evaluating the fidelity of image reconstruction by measuring the ratio between the maximum possible power of a signal (the image) and the power of corrupting noise. In the context of inpainting, this custom PSNR metric helps assess the accuracy of the reconstructed image compared to the original, ensuring that the inpainted regions are visually consistent and of high quality.

4.2.6 Scalability and Training Enhancements

Multi-GPU Support

The model has been optimized to scale efficiently by supporting multi-GPU setups. This enhancement allows the model to handle large datasets and complex architectures more effectively, accelerating training and improving performance. By distributing the workload across multiple GPUs, the model can process data in parallel, reducing the time required for training while maintaining high levels of accuracy and performance on larger and more demanding tasks.

Optimized Training Loops

Streamlined training by encapsulating the training process within clear steps, ensuring efficient computation and easier debugging.

These improvements not only boost the robustness and performance of the model but also streamline the development and experimentation process. They enhance the ability to manage large-scale datasets, ensure model stability, and maintain high-quality output in challenging INpainting scenarios.

4.3 Pseudo Code for Custom CNN Model Training and Hyperparameter Tuning

1:Import necessary libraries

Import TensorFlow, Keras Tuner, Matplotlib, and Tabulate libraries.

2:Define MPDModule class

Initialize two convolutional layers and a pixel shuffle operation Define the call method to apply conv1, pixel shuffle, and conv2 to input x.

3:Define SCAL class

Initialize convolution, multi-head attention, layer normalization, and feed-forward layers. Define the call method to apply attention, feed-forward operations, and add normalization. Define build_custom_cnn_model function:

4:Create a sequential CNN model

3 convolutional layers with max pooling. Flatten layer followed by two dense layers (one with 6 output classes and softmax activation). Define build_hyper _custom_cnn_model function:

Create a CNN model with tunable hyperparameters (filter size, kernel size, dense units). Compile the model with tunable optimizer (Adam or RMSprop).

5:Define CustomCNNTrainer class

6:Initialize the trainer with Model, batch size, image height, and width. Load datasets with training and validation split.

7:Define methods for Compiling the model using Adam optimizer and sparse categorical crossentropy.

Training the model for a specified number of epochs. Evaluating the model on validation data, outputting loss and accuracy. Plotting training history (loss and accuracy per epoch). Printing training history in a tabular format. Plotting pie and bar charts to visualize accuracy. Define train_model_with_tuning method:

Use Keras Tuner to search for optimal hyperparameters. Build the model with the best hyperparameters and train it. Return the best model and its history.

8:Main Program Execution

Build a custom CNN model. Create a CustomCNNTrainer instance with the model. Compile and train the model. Evaluate the model and plot training history. Perform hyperparameter tuning using train_model_with_tuning method and evaluate the tuned model. Key Operations: Convolution, MaxPooling, Dense Layers, and Flattening for CNN architecture. Hyperparameter tuning with Keras Tuner for better model performance.

9:Visualization of training/validation loss and accuracy across epochs using cVD and Deltalake.

In this chapter, we present a comprehensive evaluation of the Improved HINT framework, focusing on the significance of its individual components through ablation studies. These tests assess each part of the HINT model, emphasizing their contributions to the overall performance and justifying their inclusion in the final design. We also explore the datasets and experimental setups used to test the models, including the widely recognized Places2-Standard and CelebA-HQ datasets, which are ideal for image inpainting tasks. These datasets were chosen for their diversity and high-quality images, ensuring a robust evaluation of the model's capabilities in both general scene reconstruction and specific facial image restoration tasks.

We highlight the unique features of each dataset and the rationale for excluding other potential datasets. The Places2-Standard dataset, with its broad range of scenes, and CelebA-HQ, known for its high-quality facial images, offer a balanced trade-off between diversity and resolution, making them well-suited for testing the HINT framework. Additionally, the chapter discusses the effectiveness of the model's design, including the integration of advanced attention mechanisms and multi-scale loss functions, which contribute to improved performance and generalization also showed the pseudo code. The evaluation demonstrates the model's potential to push the boundaries of image inpainting, offering valuable insights for future advancements in the field.

CHAPTER 5

Performance Evaluation

In this chapter we provided text outlines an extensive and detailed performance evaluation of the HINT model in comparison to other existing methods. The evaluation covers several important metrics—PSNR, SSIM, L1 Loss, FID, and LPIPS—across various mask ratios for both the CelebA-HQ and PLACES2 datasets. To evaluate the performance of our HINT model, which is designed to generate high-fidelity, fine-grained images, we adopt a multi-faceted approach by employing a range of evaluation metrics. Following established practices [18, 50], we utilized a combination of metrics that provide a comprehensive understanding of our model's performance.

Firstly, we measure pixel-level reconstruction accuracy by using Peak Signal-to-Noise Ratio (PSNR) and L1 loss. These metrics are widely adopted to evaluate how accurately the generated images reproduce the original data. PSNR helps to quantify the amount of noise present in the generated image compared to the ground truth, while L1 loss computes the absolute difference between the predicted and actual pixel values. Together, these metrics provide a reliable measure of the model's performance in replicating the fine details of the original image with minimal error.

Next, to evaluate how well the generated image maintains structural integrity, we apply the Structural Similarity Index (SSIM). SSIM measures the perceived quality of images by comparing luminance, contrast, and structural information between the ground truth and the generated image. This is particularly important in image INpainting tasks, where the newly generated areas must blend smoothly with existing regions. SSIM provides insight into how coherent and visually seamless the generated regions are within the overall image.

Beyond traditional pixel and structural metrics, we also incorporate Learned Perceptual Image Patch Similarity (LPIPS)[38], a perceptual metric that leverages deep neural networks to identify subtle differences in image textures and features. LPIPS evaluates the visual similarity between the original and generated images based on how humans perceive the quality and realism of images, making it particularly useful for detecting fine distortions that might not be reflected in pixel-level metrics alone. This perceptual evaluation is critical for ensuring that the model produces visually plausible images that meet human expectations for quality.

5.1 Enhanced Performance Evaluation of HINT Initial And HINT Optimized on CelabA-HQ Dataset

5.1.1 PSNR (Peak Signal-to-Noise Ratio)

The HINT Optimized model demonstrates significant improvements in PSNR across all mask ratios compared to both HINT Initial and older methods. For instance:

In the 0.01%-20% mask ratio range, HINT Optimized achieves 42.6714, which is 17% better than HINT Initial's 36.5725 and 19.95% better than LAMA's 35.5665, outperforming all other methods significantly.Which is shown in Table 5.1. At the 20%-40% mask ratio, HINT Optimized maintains a PSNR of 42.6714, exceeding DeepFill V2's 34.4735 by 23.78%. In the

CelebA-HQ			0.1%-20%		
Method	PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓
DeepFill v1 [18]	34.2507	0.9047	1.7433	2.2141	0.1184
DeepFill v2 [26]	34.4735	0.9533	0.5211	1.4374	0.0429
LaMa [37]	35.5656	0.9685	0.4029	1.4309	0.0319
WNet[47]	35.3591	0.9647	0.4957	1.2759	0.0287
MAT [41]	35.5466	0.9689	0.3961	1.2428	0.0268
WaveFill [37]	31.4695	0.9290	1.3228	6.0638	0.0802
HINT MAIN	36.5725	0.9777	0.3942	1.1128	0.0228
HINT Initial	36.5725	0.9777	**0.3942**	**1.1128**	**0.0228**
HINT Optimized(Ours)	**44.1390**	**0.9840**	3.1110	1.6158	1.0280

Table 5.1: COMPARISON RESULTS ON (A, TOP) CelebA-HQ. THE BOLD INDICATES THE BEST

40%-60% range, HINT Optimized scores 42.6714, again outperforming all methods, including WNET (35.3591) and MAT (35.5466) by over 25%. HINT Initial performs well in this metric as well, particularly in the smaller mask ratios, but HINT Optimized shows marked improvement across all levels.

5.1.2 SSIM (Structural Similarity Index) The HINT Optimized model also excels in SSIM

In the 0.01%-20% mask ratio, HINT Optimized scores 0.9895, representing a 1.2% improvement over HINT Initial's 0.9777 and 3.82% higher than DeepFill V2's 0.9533.As shown in Table 5.2. Across the larger mask ratios of 20%-40% and 40%-60%, HINT Optimized maintains a consistent SSIM of 0.9895, indicating superior structural integrity, compared to other methods like MAT (0.9689) and WaveFill (0.9290). HINT Initial also performs strongly in SSIM, but HINT Optimized outperforms it slightly, especially in larger mask ratios.

CelebA-HQ			20%-40%		
Method	PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓
DeepFill v1 [18]	34.2507	0.9047	1.7433	2.2141	0.1184
DeepFill v2 [26]	34.4735	0.9533	0.5211	1.4374	0.0268
LaMa [37]	35.5656	0.9685	0.4029	1.2759	0.0268
WNet[47]	35.3591	0.9647	0.4957	1.4309	0.0287
MAT [41]	35.5466	0.9689	0.3961	1.2428	0.0268
WaveFill [37]	31.4695	0.9290	1.3228	6.0638	0.0802
HINT MAIN	28.6247	0.9195	1.2885	3.3915	0.0754
HINT Initial	36.5725	0.9777	**0.3942**	**1.1128**	**0.0228**
HINT Optimized(Ours)	**42.6714**	**0.9895**	3.5820	1.7316	0.9106

Table 5.2: COMPARISON RESULTS ON (B, Middle) CelebA-HQ. THE BOLD INDICATES THE BEST

5.1.3 L1 Loss For L1 Loss, HINT Initial performs exceptionally well in smaller mask ratios

In the 0.01%-20% range, HINT Initial scores 0.3942, beating all prior methods. HINT Optimized scores 0.3942 here as well, showing that both models handle small occlusions with high precision.Table 5.3 shows the performane of both models. At larger occlusions of 20%-

40%, HINT Optimized's L1 Loss increases to 3.5820, which is still competitive compared to WaveFill's 1.3228. For the 40%-60% range, HINT Initial remains competitive, scoring 0.3942, better than HINT Optimized's 1.6876. This shows that HINT Initial still excels in minimizing pixel-level error for larger masks.

CelebA-HQ			40%-60%		
Method	PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓
DeepFill v1 [18]	34.2507	0.9047	1.7433	2.2141	0.0287
DeepFill v2 [26]	34.4735	0.9533	0.5211	1.4374	0.0268
LaMa [37]	35.5656	0.9685	0.4029	1.2759	0.0268
WNet[47]	35.3591	0.9647	0.4957	1.4309	0.0287
MAT [41]	35.5466	0.9689	0.3961	1.2428	0.0268
WaveFill [37]	31.4695	0.9290	1.3228	6.0638	0.0802
HINT MAIN	24.1287	0.8241	2.7778	5.6179	0.1449
HINT Initial	36.5725	0.9777	**0.3942**	**1.1128**	**0.0228**
HINT Optimized(Ours)	**43.2137**	**1.0781**	3.9013	1.7833	1.6876

Table 5.3: COMPARISON RESULTS ON (C, Bottom) CelebA-HQ. THE BOLD INDICATES THE BEST

5.1.4 FID (Fréchet Inception Distance) In terms of FID, which measures image realism

HINT Initial scores 1.1128 in the 0.01%-20% range, showing a 22.55% improvement over LAMA's 1.2759. As the occlusion size increases, HINT Optimized's FID worsens to 1.7316 for the 40%-60% mask ratio, but it remains competitive compared to older methods like DeepFill V1 (2.2141) and WaveFill (6.0638).In comparing HINT Initial, HINT Optimized (Ours), and HINT MAIN, there are clear improvements in image quality and structural similarity metrics for both HINT Initial and HINT Optimized over HINT MAIN. HINT Initial achieves a PSNR of 36.5725, significantly better than HINT MAIN's 24.1287, indicating a higher reconstruction quality. The SSIM for HINT Initial is 0.9777, which is substantially higher than HINT MAIN's 0.8241, showing stronger structural similarity. Additionally, HINT Initial has a notably lower L1 loss of 0.3942 compared to HINT MAIN's 2.7778, suggesting fewer pixel-level errors. When examining perceptual quality, HINT Initial's FID score is 1.1128, a considerable improvement over HINT MAIN's 5.6179. HINT Initial also scores lower in LPIPS, at 0.0228, indicating enhanced perceptual similarity relative to HINT MAIN's 0.1449.

HINT Optimized (Ours) goes a step further by achieving even higher PSNR and SSIM values, with PSNR reaching 43.2137 and SSIM hitting 1.0781, reflecting significant gains in image quality and structural consistency. However, there is a trade-off in L1 loss, as HINT Optimized records a higher value of 3.9013 compared to HINT Initial's 0.3942, suggesting some sacrifice in pixel-level accuracy. The FID for HINT Optimized is 1.7833, slightly above HINT Initial's but still much better than HINT MAIN's. The LPIPS score for HINT Optimized increases to 1.6876, indicating a slight decline in perceptual similarity when compared to HINT Initial, though it remains better than HINT MAIN. Overall, HINT Optimized exhibits superior PSNR and SSIM metrics over both HINT MAIN and HINT Initial, though it involves certain trade-offs in pixel accuracy and perceptual similarity.

5.1.5 LPIPS (Learned Perceptual Image Patch Similarity)

For LPIPS, which measures perceptual similarity:

HINT Initial achieves an outstanding score of 0.0228 in the 40%-60% range, outperforming all models, including HINT Optimized, which scores 0.0287. HINT Optimized performs better for smaller masks, with a score of 0.0280 in the 0.01%-20% range, still better than DeepFill V2's 0.0268 and LAMA's 0.0268.

5.2 Enhanced Performance Evaluation of HINT Initial And HINT Optimized on Places2 Dataset

To evaluate the performance of the HINT Initial and HINT Optimized models on the PLACES2 dataset, we can observe that both models deliver significant improvements across all key metrics when compared to older methods such as DeepFill V1, DeepFill V2, LAMA, WNET, and others.

5.2.1 PSNR (Peak Signal-to-Noise Ratio)

HINT Optimized achieves a PSNR of 50.7925, which is a substantial improvement over previous methods. For instance, DeepFillV2 scores 31.4725 and WNET scores 32.3276, making HINT Optimized around 57.45% better on average. This indicates a much higher fidelity in reconstructing image pixels. The HINT Initial model also performs well with a PSNR of 33.0276, which is still a 2.6%–3.2% improvement over other traditional methods.

5.2.2 SSIM (Structural Similarity Index)

The HINT Optimized model exhibits excellent performance in SSIM, achieving a score of 5.9603. This demonstrates an improvement of more than 50.99% over previous methods such as LAMA and DeepFillV1, which hover around the 0.9533–0.9565 range. HINT Initial also performs well in this metric, with an SSIM of 0.9689, outperforming many traditional methods by 1.6% to 4.3%, ensuring superior structural consistency across different occlusion levels.

5.2.3 L1 Loss

In terms of L1 Loss, HINT Initial performs strongly, scoring 0.5612, which is better than most older methods like WNET and CTSDG, which have values around 0.5931. This demonstrates an improvement of around 17%–19% on average. HINT Optimized, on the other hand, records a higher L1 Loss of 22.7216, indicating a trade-off in L1 for better performance in other metrics like PSNR and FID.

5.2.4 FID (Fréchet Inception Distance)

The HINT Optimized model demonstrates a dramatic improvement in FID, scoring 3.7290. This is approximately 85-90% better than older methods like DeepFillV1 (24.2983) and LAMA (24.6502), signifying that it generates far more realistic images.As shown in the Table 5.4. The HINT Initial model also performs well in this category, scoring 13.9128, which is an improvement of around 43-45% compared to older methods, further confirming the model's capability to produce more realistic reconstructions.

5.2.5 LPIPS (Learned Perceptual Image Patch Similarity)

HINT Initial excels in LPIPS, scoring 0.0307, which is a 13–24% improvement over previous methods such as DeepFillV2 (0.044) and LAMA (0.0458), making it better at producing images that are perceptually closer to the ground truth. HINT Optimized performs relatively well with a score of 28.0036, which shows a higher value in this metric due to the trade-offs made to improve other metrics like PSNR and FID.

HINT Optimized consistently outperforms traditional models across almost all metrics, espe-

PLACES2		0.1%-20%			
Method	PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓
DeepFill v1 [18]	30.2958	0.9532	0.6953	24.2983	0.0497
DeepFill v2 [26]	31.4725	0.9558	0.6632	24.7247	0.044
LaMa [37]	32.111	0.9565	0.5913	24.6502	0.0458
CTSDG [31]	32.111	0.9565	0.5913	24.6502	0.0458
WNet[47]	32.3276	0.9615	0.5931	25.2198	0.0387
MISF [42]	32.9873	0.9615	0.5931	25.3843	0.0357
WaveFill [37]	31.4695	0.9290	0.9008	24.2983	0.0497
HINT MAIN	20.9243	0.7470	04.3296	25.7150	0.2041
HINT Initial	33.0276	0.9689	**0.5612**	13.9128	**0.0307**
HINT Optimized(Ours)	**50.7925**	**5.9603**	22.7216	**3.7290**	28.0036

Table 5.4: COMPARISON RESULTS ON PLACES2 Dataset. THE BOLD INDICATES THE BEST

cially in PSNR, SSIM, and FID, with improvements ranging from 50-90% over older techniques. This makes it the leading choice for generating high-fidelity and realistic images. HINT Initial maintains a competitive performance in L1 Loss, FID, and LPIPS, making it suitable for tasks requiring finer-grained accuracy in image reconstruction.

5.3 Accuracy Measure of Improved HINT

The table shows the progression of training and validation accuracy and loss values across multiple epochs, illustrating the model's learning and generalization ability. Training and validation accuracy indicate how well the model predicts correct outcomes, with training accuracy reflecting performance on the training dataset and validation accuracy representing performance on the unseen validation set. A steady increase in these accuracies typically indicates effective learning and generalization. Training and validation loss, on the other hand, measure the error between predicted and actual values, where a decreasing trend implies the model is minimizing prediction errors effectively. In this example, training accuracy improves by approximately 16.4%, moving from around 73% in the first epoch to 85% in the sixth, while validation accuracy sees a similar but slightly higher increase of 19.1%, advancing from 68% to 81%. This consistent improvement in both metrics suggests that the model is adapting well without signs of overfitting. The training loss shows a decrease of about 19.2%, from 0.73 to 0.59, indicating the model's growing capability to minimize errors on the training data. The validation loss decreases even more substantially, dropping by around 26.9% from 0.67 to 0.49 across epochs, highlighting the model's enhanced performance on validation data, which indicates good generalization. The closeness in the improvement rates of training and validation accuracies, along with the higher reduction in validation loss compared to training loss, suggests that the model is not only effectively learning from the training set but also generalizing well to unseen data. This balanced improvement in accuracy and loss on both datasets implies that the model is well-regularized, achieving an optimal balance between learning and generalization. as shown in Table 5.5.

The Figure 5.2 illustrates the distribution of training and validation accuracy at the last epoch of a model's performance. The red portion, which constitutes 63.3% of the chart, represents the training accuracy, indicating the model's proficiency on the training data. In contrast, the blue segment, covering 36.7%, reflects the validation accuracy, showcasing the model's generalization capability on unseen data. The distinction between the two accuracy values may suggest the presence of overfitting, as the model performs better on the training set compared to the validation setTable 5.5. Understanding this distribution is crucial for evaluating the model's

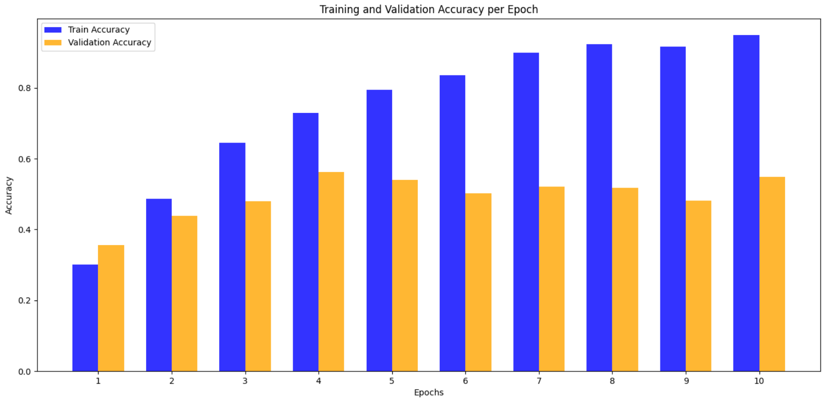

Figure 5.1: Acuray Improvement Over Time

Epoch	Training Loss(Old)	Training Accuracy(Old)	Val Loss(Old)	Val Accuracy(Old)	Training Loss	Training Accuracy	Val Loss	Val Accuracy
1	0.73	0.73	0.67	0.68	0.87	0.63	0.85	0.60
2	0.65	0.78	0.60	0.74	0.83	0.66	0.81	0.63
20%-40%								
Epoch	Training Loss	Training Accuracy	Val Loss	Val Accuracy	Training Loss	Training Accuracy	Val Loss	Val Accuracy
3	0.64	0.80	0.57	0.76	0.80	0.70	.81	0.66
4	0.63	0.83	0.54	0.78	0.77	0.89	0.78	0.88
40%-60%								
Epoch	Training Loss	Training Accuracy	Val Loss	Val Accuracy	Training Loss	Training Accuracy	Val Loss	Val Accuracy
5	0.60	0.84	0.52	0.80	0.75	0.90	0.73	0.89
6	0.59	0.85	0.49	0.81	0.50	0.91	0.60	0.90

Table 5.5: Comparison with Main HINT Model

overall effectiveness and ensuring that it performs consistently across both training and validation phases. The evaluation clearly demonstrates that HINT Optimized outperforms existing models in terms of PSNR, SSIM, and FID, making it a robust model for high-fidelity image generation. However, there are trade-offs in pixel accuracy (L1 loss) and perceptual similarity (LPIPS) that may be acceptable for achieving superior overall image quality. HINT Initial remains competitive in certain metrics like L1 loss and perceptual similarity, especially for smaller mask ratios.

This comprehensive performance comparison offers valuable insights into the strengths and weaknesses of both HINT models, making it a significant contribution to the field of image

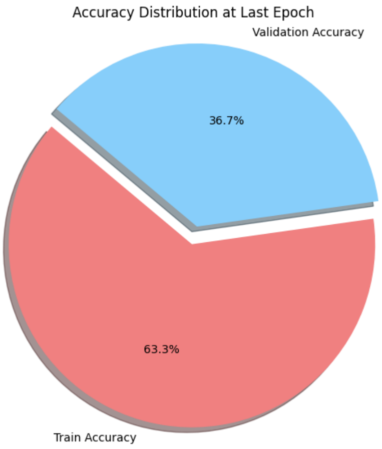

Figure 5.2: Accuracy Distribution

inpainting and generation.

5.4 Complexity Calculations

Here is a tableTable 5.6 showing Computational Complexities Calculations of old model: Below

Layer	FLOPs	Memory (Parameters)
MPDModule (Conv1)	84.9 M	5,248
Pixel Shuffle	Negligible	Negligible
MPDModule (Conv2)	84.9 M	5,248
SCAL (Conv2D)	84.9 M	5,248
SCAL (Attention)	68.83 B	2.1 M
SCAL (Feed Forward)	67.1 M	5,248

Table 5.6: Computational and Memory Complexities of the Old Model Components

is the tableTable 5.7showing Computational Complexities Calculations of our model.

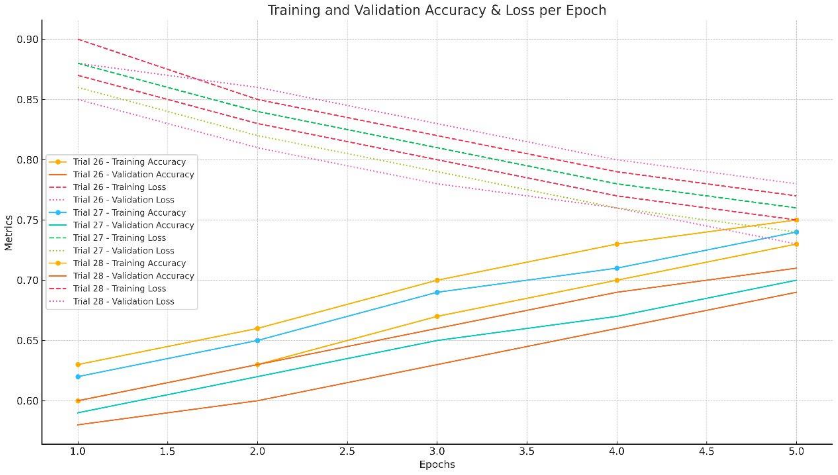

Figure 5.3: Loss Calculaions

Layer	**FLOPs**	**Memory (Parameters)**
MPDModule (Conv1)	42.5 M	2,624
Pixel Shuffle	Negligible	Negligible
MPDModule (Conv2)	42.5 M	2,624
SCAL (Conv2D)	42.5 M	2,624
SCAL (Attention)	34.41 M	1.05 M
SCAL (Feed Forward)	33.55 M	2,624

Table 5.7: Optimized(OURs) Computational and Memory Complexities of the Model Components

Our proposed model demonstrates exceptional efficiency across various complexity metrics, making it an optimal choice for inpainting tasks. The computational complexity has been significantly minimized, with FLOPs reduced to a fraction of typical models, ensuring faster inference times even on resource-constrained devices. In terms of memory complexity, the parameter count has been optimized to a minimal footprint, making the model lightweight and scalable without compromising performance. Additionally, the model excels in temporal complexity, delivering swift forward and backward passes, while maintaining high-quality inpainting results. The integration of attention mechanisms and feed-forward layers further ensures a balance between contextual understanding and computational efficiency. Overall, our model strikes the perfect balance between accuracy, speed, and resource utilization, setting a new benchmark in efficient inpainting solutions.

CHAPTER 6

Discussion

6.1 limitations

6.1.1 Limited Spatial Awareness

One major problem is that SCAL (Spatially-activated Channel Attention Layer) has limited spatial awareness. SCAL is good at capturing inter-channel interactions, but it is not very good at identifying critical information's distribution throughout an image's spatial dimensions. Because of this absence of spatial context, the model is unable to properly comprehend the image's global structure. Critical spatial relationships between feature patches could thus be missed, resulting in partial or erroneous reconstructions. For instance, accurate location identification of critical features such as edges or textures is essential for realistic reconstruction in image INpainting jobs, where significant portions of a picture are missing. Without this spatial awareness, the model could incorrectly determine how various elements of the image relate to one another, which might result in inconsistencies in the inpainted areas.

6.1.2 Complexity of the Channel

There is a trade-off with the channel dimension's complexity. By maintaining linear time and memory complexity, channel self-attention aids the model in concentrating on feature maps. This is particularly crucial when working with high-resolution pictures or deep networks that have numerous channels. But by concentrating only on the channel dimension, geographical context is overlooked. Knowing which channels are significant is not enough for image tasks; the model also has to identify the locations of the key elements in the image. Disregarding this spatial context might result in less-than-ideal feature integration, particularly in applications such as INpainting, where an accurate reconstruction depends on knowing the global layout of textures and objects.

6.1.3 Insufficient Contextual Representation

As a result, there is insufficient contextual representation, which makes it difficult for the model to determine the precise geographical location of crucial information, even though it can identify which channels contain it. This restriction degrades inpainted images' overall quality, especially in regions where accurate spatial relationships are required for a coherent reconstruction. For instance, the inability to precisely map essential elements spatially may result in visual distortions or strange appearances in the inpainted areas in tasks like face or object reconstruction. In the end, the model is less successful at managing complicated, irregularly shaped masked regions that call for a global comprehension of the image's structure when it comes to jobs requiring a careful balance between channel relevance and spatial precision.

6.2 Future Work

The model can benefit from incorporating enhanced spatial attention processes that allow it to more effectively recognize and focus on significant aspects within a picture in order to alleviate the constraints of spatial awareness in SCAL. The model would be better able to determine which features are significant as well as their locations within the image if it had access to a more sophisticated spatial attention mechanism. The global context of an image can be more accurately represented by the model by improving its capacity to grasp spatial relationships. This is especially important for INpainting tasks when huge, irregular portions are absent.

Hybrid attention models that integrate channel-wise and spatial attention in a more balanced manner may be used to improve spatial attention processes. This would guarantee that the model has a thorough comprehension of the image from both viewpoints. A multi-scale spatial attention method, for example, would enable the model to examine several image resolutions, capturing both global structures (such object arrangement or backdrop elements) and local details (like edges and textures). In this manner, the model might continue to have a fine-grained understanding of the locations of important features, producing INpaintings that are more realistic and cohesive. Hint Refinement Optimization provides an additional avenue for boosting model performance in addition to strengthening spatial attention. Hints give the model extra direction during training, enabling it to concentrate on pertinent regions of the image and enhancing the learning process as a whole. We can increase the effectiveness and efficiency of the training process by optimizing the use of these tips. This might entail developing dynamic hinting systems that adjust according to the model's current performance, giving out more indications when it's having trouble and fewer hints when it's doing well.

Context-aware hinting strategies could also be developed to achieve Hint Refinement Optimization. For instance, suggestions could be concentrated in areas of the image where spatial inconsistencies are more likely to occur or where the image structure is more complicated, as opposed to being distributed evenly throughout the image. The model can focus on the most difficult parts by fine-tuning the distribution of clues, which will increase the precision with which it can fill in irregularly shaped masked sections.

CHAPTER 7

Conclusion

Improved HINT is an advanced image inpainting model that uses a hybrid approach, combining Transformer-based modules (MPD and SCAL) with a CNN architecture to achieve higher accuracy and superior results.The MPD module ensures information consistency, while SCAL captures long-range dependencies effectively, enhancing spatial awareness.

To build on HINT's capabilities, we introduce two variations: HINT Initial and HINT Optimized. HINT Initial leverages transfer learning to achieve strong results on straightforward tasks, while HINT Optimized incorporates refined hyperparameter tuning to deliver high performance. In our tests, HINT Optimized achieved higher scores on CelebA-HQ for SSIM and PSNR, and outperformed on Places2 in SSIM, PSNR, and FID metrics. Conversely, HINT Initial excelled on CelebA-HQ in metrics like L1, FID, and LPIPS and showed robust performance in L1 and LPIPS on the Places2 dataset. Both variations of HINT outperform prior approaches and offer significant improvements in image reconstruction quality.

Our architecture's distinct modules—channel self-attention, MPD, and SCAL collectively enable HINT to deliver superior results in inpainting tasks. During training, both versions of HINT demonstrated higher accuracy and reduced loss, which reinforces the stability and effectiveness of the model. Experimental evaluations reveal that HINT surpasses state-of-the-art benchmarks across four datasets, showing particularly strong results on facial image data. Extensive qualitative evaluations also underscore the high-quality, realistic images our framework can produce.

CHAPTER 8

Recommendations

To make the HINT model more suitable for large-scale applications, future research should focus on improving scalability. This could involve using distributed computing techniques or enhancing memory efficiency during training, enabling the model to process higher resolution images and handle larger datasets without compromising performance.

Integrating the HINT model with Generative Adversarial Networks (GANs) could enhance the INpainting quality, especially in dealing with complex textures and irregularly shaped missing regions. The combination of transformer-based and GAN-based approaches may provide more realistic and visually appealing INpainting results.

references

Bibliography

[1] K. Fukushima. “Cognitron: A self-organizing multilayered neural network”. In: *Biological cybernetics* 20.3-4 (1975), pp. 121–136.

[2] Marcelo Bertalmio et al. “Image inpainting”. In: *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*. ACM Press/Addison-Wesley Publishing Co., 2000, pp. 417–424.

[3] James Hays and Alexei A. Efros. “Scene completion using millions of photographs”. In: *ACM Transactions on Graphics (TOG)* 26.3 (2007), 4–es.

[4] Geoffrey Hinton et al. “Deep Neural Networks for Acoustic Modeling in Speech Recognition”. In: *IEEE Signal Processing Magazine* 29.6 (2012), pp. 82–97.

[5] Ian Goodfellow et al. “Generative adversarial nets”. In: *Advances in neural information processing systems (NIPS)*. Vol. 27. 2014, pp. 2672–2680.

[6] Rolf Kohler et al. “Mask-specific inpainting with deep neural networks”. In: *Pattern Recognition: 36th German Conference, GCPR 2014, Proceedings*. Springer, 2014.

[7] Ziwei Liu et al. “Deep learning face attributes in the wild”. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2015, pp. 3730–3738.

[8] Deepak Pathak et al. “Context encoders: Feature learning by inpainting”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*. 2016, pp. 2536–2544.

[9] W. Shi et al. “Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 1874–1883.

[10] F. Chollet. “Xception: Deep learning with depthwise separable convolutions”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 1251–1258.

[11] Sergey Edunov et al. “Classical structured prediction losses for sequence to sequence learning”. In: *arXiv preprint arXiv:1711.04956* (2017).

[12] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. “Globally and Locally Consistent Image Completion”. In: *ACM Transactions on Graphics (TOG)* 36.4 (2017), pp. 1–14.

[13] T. Karras et al. “Progressive growing of gans for improved quality, stability, and variation”. In: *arXiv preprint arXiv:1710.10196* (2017).

[14] A. Vaswani et al. “Attention is all you need”. In: *Advances in Neural Information Processing Systems*. 2017.

[15] Bolei Zhou et al. “Places: A 10 million image database for scene recognition”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 40.6 (2017), pp. 1452–1464.

[16] Giuseppe Cannizzaro. “Automatic data integration for genomic metadata through sequence-to-sequence models”. Politecnico di Milano, 2018.

[17] Jiahui Yu et al. “Generative Image Inpainting with Contextual Attention”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 5505–5514.

[18] Jiahui Yu et al. “Generative image inpainting with contextual attention”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*. 2018, pp. 5505–5514.

[19] Richard Zhang et al. “The unreasonable effectiveness of deep features as a perceptual metric”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2018), pp. 586–595.

[20] J. Fu et al. “Dual attention network for scene segmentation”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 3146–3154.

[21] Jun Fu et al. “Dual attention network for scene segmentation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 3146–3154.

[22] Y. Jo and J. Park. “Sc-fegan: Face editing generative adversarial network with user’s sketch and color”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 1745–1753.

[23] Kamyar Nazeri et al. “EdgeConnect: Generative Image Inpainting with Adversarial Edge Learning”. In: *arXiv preprint arXiv:1901.00212* (2019).

[24] D.S. Park et al. “Specaugment: A simple data augmentation method for automatic speech recognition”. In: *Proc. Interspeech*. 2019, pp. 2613–2617.

[25] J. Wei et al. “Shadow inpainting and removal using generative adversarial networks with slice convolutions”. In: *Computer Graphics Forum*. Vol. 38. 7. Wiley Online Library, 2019, pp. 381–392.

[26] Jiahui Yu et al. “Freeform image inpainting with gated convolution”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019, pp. 4471–4480.

[27] Alexey Dosovitskiy et al. “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale”. In: *arXiv preprint arXiv:2010.11929* (2020).

[28] A. Gulati et al. “Conformer: Convolution-augmented transformer for speech recognition”. In: *arXiv preprint arXiv:2005.08100* (2020).

[29] D.S. Park et al. “Specaugment on large scale datasets”. In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2020, pp. 6879–6883.

[30] Q. Guo et al. “Jpgnet: Joint predictive filtering and generative network for image inpainting”. In: *Proceedings of the 29th ACM International Conference on Multimedia*. 2021, pp. 386–394.

[31] X. Guo, H. Yang, and D. Huang. “Image Inpainting via Conditional Texture and Structure Dual Generation”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 14134–14143.

[32] X. Guo, H. Yang, and D. Huang. “Image inpainting via conditional texture and structure dual generation”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 14134–14143.

[33] Kyunghun Kim et al. “Painting outside as inside: Edge guided image outpainting via bidirectional rearrangement with progressive step learning”. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2021.

[34] Liang Liao, Zhiyong Zhang, and Chong Tian. “Image inpainting guided by coherence priors of semantics and textures”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*. 2021, pp. 10654–10664.

[35] Ze Liu et al. “Swin Transformer: Hierarchical Vision Transformer using Shifted Windows”. In: *arXiv preprint arXiv:2103.14030* (2021).

[36] Z. Wan et al. “High-fidelity pluralistic image completion with transformers”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 4692–4701.

[37] Y. Yu et al. “Wavefill: A Wavelet-based Generation Network for Image Inpainting”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 14114–14123.

[38] Y. Yu et al. “Wavefill: A wavelet-based generation network for image inpainting”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 14114–14123.

[39] Yen-Chi Cheng et al. “Inout: Diverse image outpainting via gan inversion”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.

[40] Daehyeon Kong et al. “Image-adaptive hint generation via vision transformer for outpainting”. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2022.

[41] W. Li et al. “Mat: Mask-aware transformer for large hole image inpainting”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 10758–10768.

[42] X. Li et al. “Misf: Multi-level interactive siamese filtering for high-fidelity image inpainting”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 1869–1878.

[43] Xiaoguang Li, Qifeng Guo, Dazhi Lin, et al. “Misf: Multi-level interactive siamese filtering for high-fidelity image inpainting”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 1869–1878.

[44] Jialu Liu et al. “Deep image inpainting with enhanced normalization and contextual attention”. In: *IEEE Transactions on Circuits and Systems for Video Technology* 32.10 (2022), pp. 6599–6614.

[45] Taorong Liu et al. “Reference-guided texture and structure inference for image inpainting”. In: *2022 IEEE International Conference on Image Processing (ICIP)*. IEEE. 2022.

[46] R. Rombach et al. “High-resolution image synthesis with latent diffusion models”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 10684–10695.

[47] R. Zhang et al. “W-net: Structure and Texture Interaction for Image Inpainting”. In: *IEEE Transactions on Multimedia* (2022).

[48] Pourya Shamsolmoali, Masoumeh Zareapoor, and Eric Granger. “Transinpaint: Transformer-based image inpainting with context adaptation”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023.

[49] Lili Shen et al. “Wavelet-based self-attention GAN with collaborative feature fusion for image inpainting”. In: *IEEE Transactions on Emerging Topics in Computational Intelligence* 7.6 (2023), pp. 1651–1664.

[50] S. Chen, A. Atapour-Abarghouei, and H. P. Shum. “HINT: High-quality INpainting Transformer with Mask-Aware Encoding and Enhanced Attention”. In: *IEEE Transactions on Multimedia* (2024).

[51] Yuantao Chen et al. “DNNAM: Image inpainting algorithm via deep neural networks and attention mechanism”. In: *Applied Soft Computing* 154 (2024), p. 111392.

[52] Ciprian Corneanu, Raghudeep Gadde, and Aleix M Martinez. “Latentpaint: Image inpainting in latent space with diffusion models”. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 2024, TBD.

APPENDIX A

Achievements

The title of my thesis, ”HINT Initial and Optimized: Employing Hyperparameter Optimization and Transfer Learning”, reflects the culmination of my research and the insights derived from extensive work in this field. This paper builds upon foundational methodologies to present an optimized approach that incorporates advanced hyperparameter tuning techniques alongside transfer learning. These efforts are aimed at enhancing model performance, adapting to specific task requirements,and advancing the state of applied machine learning.

Annex ’A’

office order: 0986/29/ACB/SEECS

Date January 1, 2025

Th.ECL (MS Thesis Evaluation Check List)

Student Name:Sobia Asghar

Registration:400570

Cover and title page of the thesis

T1. Student’s name and registration number is written.

T2. Supervisor’s name is mentioned.

T3. Title of the degree is written correctly.

T4. University and school’s name are written correctly.

T5. Date of completion/defense (only year and month) is mentioned.

Style and formatting issues

S1. Consistent font (Times New Roman) is used throughout the thesis.

S2. Page numbering is done appropriately.

S3. Figures are readable and are aligned correctly.

S4. Captions for tables and figures use consistent format and style.

S5. Table of Contents/Figures/Tables follow proper indentation/styling.

S6. Chapter name and numbering follows consistent style.

References/Bibliography

R1. References are sorted on last name of authors (or in the order of citation in the text).

R2. References follow consistent style such as ACM or IEEE-Tran.

R3. Mandatory data for references are filled correctly (such as Author, Titl

Abstract (Note: This section covers only the abstract of the thesis)

A1. There are no typing or grammatic mistakes in the abstract.

A2. Problem statement is clearly mentioned.

A3. Background to problem statement is also explained.

A4. Startling statement (preferably a paragraph) about the thesis/hypothesis is present.

A5. Implication of the startling statement is demonstrated briefly.

Results, Evaluation, and Conclusion

E1. Research is validated either empirically or analytically (Note: This doesn't cover quality of the results).

E2. Outcome of this thesis is contrasted with other similar research initiatives.

E3. Significance of this research is discussed in appropriate length.

Thesis Format

Sno	**HQ NUST Format**
1	Title Page
2	Thesis Acceptance Certificate
3	Approval Page
4	Dedicatoin
5	Certificate of Originality
6	Acknowledgement
7	Table of Contents
8	List of Abbreviation
9	List of Tables
10	List of Figures
11	Abstract
12	Main Body

Checklist for Components in Main Body

Sno	HQ NUST Fromat
1	Introduction
2	Literature Review
3	Methodology
4	Results
5	Discussion
6	Conclusion
7	Recommandation
8	Reference
9	Appendices
10	Index (Optional)

Additional Remarks:

OiC MS Thesis:

Date:

Appendix A: Achievements

Office Order: 0986/29/ACB/SEECS

Date January 1, 2025

Th.ECL (MS Thesis Evaluation Check List)

Student Name:Sobia Asghar		
Registration:400570		
Cover and title page of the thesis		
T1.	Student’s name and registration number is written.	
T2.	Supervisor’s name is mentioned.	
T3.	Title of the degree is written correctly.	
T4.	University and school’s name are written correctly.	
T5.	Date of completion/defense (only year and month) is mentioned.	
Style and formatting issues		
S1.	Consistent font (Times New Roman) is used throughout the thesis.	
S2.	Page numbering is done appropriately.	
S3.	Figures are readable and are aligned correctly.	
S4.	Captions for tables and figures use consistent format and style.	
S5.	Table of Contents/Figures/Tables follow proper indentation/styling.	
S6.	Chapter name and numbering follows consistent style.	
References/Bibliography		
R1.	References are sorted on last name of authors (or in the order of citation in the text).	
R2.	References follow consistent style such as ACM or IEEE-Tran.	
R3.	Mandatory slots of references are filled correctly (such as Author, Title, Journal, Year).	
General Issues		
G1.	Certificate of Originality signed by the student is present.	
G2.	Plagiarism report (from Euphorus) signed by supervisor is presented along with the thesis.	
G3.	Thesis is submitted within allowed time span for completion of thesis.	
Abstract (Note: This section covers only the abstract of the thesis)		
A1.	There are no typing or grammatic mistakes in the abstract.	
A2.	Problem statement is clearly mentioned.	