Advancing Representation Learning and Generative Models for Deep Learning-Based Image Inpainting

SHUANG CHEN

Advancing Representation Learning and Generative Models for Deep Learning-Based Image Inpainting

Shuang Chen

A thesis presented for the degree of
Doctor of Philosophy at Durham University

Department of Computer Science
Durham University
United Kingdom
2025-11-04

Abstract

Image inpainting is a computer vision task that aims to reconstruct an image based on the visible pixels of a damaged or corrupted image with missing regions. Its applications span across image processing and computer vision tasks such as photo editing, objective removal and depth completion. However, significant challenges remain, particularly in remaining the insufficient information and capturing long-range dependencies for improving the overall fidelity and visual quality.

We first explore the use of geometric features to guide the insufficient visual feature for improving the fidelity of facial image inpainting, and the feasibility of using image inpainting for cleft lip surgery to support surgeons as adjuncts to adjust surgical technique and improve surgical results. To achieve this idea, we collect two real-world cleft lip datasets to conduct experiments with a proposed single-stage multi-task image inpainting framework that is capable of covering a cleft lip and generating a lip and nose without a cleft. The results are assessed by expert cleft lip surgeons to demonstrate the feasibility of the proposed methods. Additionally, we embed this framework as software and released it on CodeOcean and Github, to make it convenient and equal to use for both patients and surgeons.

Although insufficient information can be provided and supplemented by landmark points, such approach only works for facial image inpainting and cannot be transferred to natural or architectural scenes, which are more common in the real-world senario. To more effectively use information while more adaptively maintaining fidelity, we propose an end-to-end High-quality INpainting Transformer, abbreviated as HINT, which consists of a novel Mask-aware Pixel-shuffle Downsampling (MPD) module to preserve the visible information extracted from the corrupted image while maintaining the integrity of the information available for high-level inference made within the model. Moreover, we propose a Spatially-activated Channel Attention Layer (SCAL), an efficient self-attention mechanism interpreting spatial awareness to model the corrupted image at multiple scales. To further enhance the effectiveness of SCAL, motivated by recent advanced in speech recognition, we introduce a sandwich structure that places feed-forward networks before and after the SCAL module. We demonstrate the superior performance of HINT compared to contemporary state-of-the-art models on four datasets, CelebA, CelebA-HQ, Places2, and Dunhuang.

Furthermore, capturing global contextual understanding is a crucial challenge to restore missing regions of images with semantically coherent content. Recent advancements have incorporated transformers, leveraging their ability to understand global interactions. However, these methods face computational inefficiencies and struggle to maintain fine-grained details. To overcome these challenges, we introduce $M \times T$ composed of the proposed Hybrid Module (HM), which combines Mamba with the transformer in a synergistic manner. Selective State Space

Model (SSM), known as Mamba, is adept at efficiently processing long sequences with linear computational costs, making it an ideal complement to the transformer for handling long-scale data interactions. However, such method method of directly adopting the vanilla SSM does not solve the inherent limitation of SSM unidirectional scanning the data, making it lack 2D spatial awareness. This insight introduces two key challenges: (i) how to maintain the continuity and consistency of pixel adjacency for pixel-level dependencies learning while processing the SSM recurrence; and (ii) how to effectively integrate 2D spatial awareness to the predominant linear recurrent-based SSMs. To solve these challenges, we spatially enhance the SSM to propose SEM-Net with efficient pixel modelling for image inpainting, involving the Snake Mamba Block (SMB) and Spatially-Enhanced Feedforward Network. These innovations enable SEM-Net to outperform state-of-the-art inpainting methods in capturing long-range dependency and enhancement in spatial consistency.

We validate the effectiveness of our methods through extensive experiments and qualitative analysis. Our approaches surpass the state-of-the-art (SoTA) in Peak Signal-to-Noise Ratio (PSNR), Structural Similarity (SSIM), Perceptual Similarity (LPIPS), L_1, and Fréchet Inception Distance (FID). All contributions have been accepted by peer-reviewed conferences or journals.

Declaration

The work in this thesis is based on research carried out at the Department of Computer Science, Durham University, United Kingdom. No part of this thesis has been submitted elsewhere for any other degree or qualification and it is all my own work unless referenced to the contrary in the text.

Shuang Chen is the author's legal name, *i.e.*, the transliteration of the Chinese name. Additionally, the individual is also known under the English name Chris, fully rendered as Shuang (Chris) Chen. This designation is employed across various professional contexts, encompassing, but not limited to, signatures in source code, emails, and documentations, among others.

Acknowledgements

"When you want something, all the universe conspires in helping you to achieve it."

Like Santiago in The Alchemist, I began my journey believing that I had to prove myself by managing everything alone—seeking acknowledgment solely through my own effort. Yet, just as Santiago learns that the pursuit of one's Personal Legend is never a solitary path, I quickly realized during my PhD that true progress is only possible with the support, guidance, and companionship of others.

I would first like to express my deepest gratitude to Prof. Hubert P.H. Shum, who showed genuine trust in me when I was searching for a PhD position. He was the one who brought me into academia, encouraged me to explore my potential, and acknowledged me as a scholar.

My heartfelt thanks also go to Dr. Amir Atapour-Abarghouei, who first showed me that research is not mere struggle but a source of genuine curiosity and joy. He helped me see that pursuing research for its own sake can be a meaningful reason to devote oneself to academia.

To my friends in the Vivid group—Ruochen Li, Haozheng Zhang, Tanqiu Qiao, Luis Li, Ziyi Chang, and Xiaotang Zhang—thank you for the valuable discussions, encouragement, and shared happiness throughout this journey.

My deepest acknowledgment is reserved for my fiancée, Ludwig (Mingze) Hou. You have stood beside me through every stage of my PhD life, encouraging me, sharing both joy and hardship, and making me a better person. You saved me from my darkest moments and completed me in ways words cannot fully express.

At the foundation of everything are my parents, whose unconditional love, care, and sacrifices have shaped who I am. Even from afar, while I studied in the UK, their guidance and encouragement continued to sustain me. For this, I remain forever grateful.

Finally, to everyone who has been part of this journey—thank you for helping me achieve and become Shuang (Chris) Chen.

Dedication

To my parents.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

Image inpainting, also called image completion, has been greatly benefited by modern learning-based techniques [2, 13–20], even though it has been a focus of study with traditional methods long before the rise of deep learning [21–23]. Unlike traditional methods, which rely on replicating nearby patches or filling minor scratches, learning-based approaches leverage deep networks to understand global semantics and structure, allowing them to generate visually coherent and contextually accurate completions even for complex and large missing regions. While these methods represent a significant improvement in terms of flexibility and performance, limitations remain, particularly in "effectively utilising insufficient contextual information" and "capturing long-range dependencies". To address these limitations, we explore tailor-made fundamental techniques specifically designed for image inpainting, focusing on enhancing representation learning and spatial awareness. Improved representation learning allows the model to better interpret and utilise insufficient contextual information, while heightened spatial awareness ensures greater spatial consistency, enabling the generation of structurally coherent results with less artifacts. By combining these advancements, our approach aims to significantly improve the overall quality of inpainted images, producing outputs that are both semantically accurate and visually seamless.

1.1 Motivations

Image inpainting aims to reconstruct a complete image from a corrupted one, which inherently contains insufficient information. As the masked region becomes larger, the consistency of structure and context is increasingly lost, making it more difficult to produce coherent results. Additionally, the missing regions disrupt both local and global relationships within the image, hindering the model's ability to perceive long-distance dependencies. This often leads to inconsistencies in patterns or features across different parts of the image.

Our motivation primarily comes from two perspectives: **effectively using insufficient information** and **capturing long-range dependency**. Enhancements in these two perspectives can significantly improve the overall performance of image inpainting tasks.

1.1.1 Motivations for Effectively Using Insufficient Information

A significant challenge hindering image inpainting is effectively modeling the valid information within visible regions, which is crucial for reconstructing semantically coherent and texture-consistent details in the missing regions. This is particularly noticeable in large masked regions, where the valid information is limited. Such challenge often results in suboptimal outcomes, such as implausible facial attributes in facial image inpainting or inconsistent structural details in building scenarios, highlighting the need for more robust solutions. Existing methods that utilise convolutional layers for downsampling come with the inherent drawback of information loss [24], attributed to the reduction of feature size from filters and downsampling. Given its capability to preserve input information, pixel-shuffle down-sample is widely used in image denoising [25], image deraining [26] and image super-resolution [27]. It periodically rearranges the elements of the input into an output scaled by the sample stride. However, its effectiveness depends on the assumption that the sample stride is small enough to avoid disrupting the noise distribution [28]. This holds only for a relatively independent distribution of raindrops and noise, and is not suitable for image inpainting with irregular and variable-size masks. Implying that directly applying conventional Pixel-shuffle Down-sampling (PD) [25–27]

to a corrupted image induces a Pixel Drifting effect, as illustrated in Fig. 4.2 (upper branch). In our observation, after PD the features within the hole regions become misaligned with their pre-downsampling counterparts, manifesting as spatial discontinuities across the downsampled feature maps. This drift stems from the periodic spatial-to-channel rearrangement, which mixes valid and corrupted neighborhoods and breaks local continuity around mask boundaries. After the feature is downsampled, the position of the masked regions becomes inconsistent across channels, causing the visible area to be misaligned in the channel, disrupting subsequent feature extraction processes within the model, thus affecting the accurate modelling of the valid information from the visible regions of the input image.

To effectively utilise the limited information from such corrupted images, first, we propose a specific facial image inpainting architecture which is also able to predict the landmark points to identify the facial attributes. We involve landmark prediction to partially guide masked images, resulting in a more precise geometry indicator for repairing facial attributes. Furthermore, to solve this problem in a more generalised manner, we propose a novel High-quality INpainting Transformer (HINT), with a tailor-made mask-aware pixel-shuffle down-sampling strategy, specifically designed for image inpainting, enabling efficient multiscale modelling of the global context while minimising the loss of valid information.

1.1.2 Motivations for Capturing Long-range Dependency

Convolutional Neural Networks (CNNs) are widely used as backbone networks in image inpainting due to their strong performance in learning generalisable representations from images and their effective mining of short-range dependencies through convolution operations [29–32]. However, their slow-grown receptive field constrains the perception of the global context and hampers the ability to capture long-range dependencies within the image. This limitation is particularly problematic for low-level vision tasks like image inpainting, where single-pixel reconstruction must preserve pixel consistency while accounting for dependencies over larger distances. To address this limitation, researchers have shifted towards transformer-based architectures [8, 33] to better capture the long-range dependencies (LRDs) and global structure. However, transformer-based methods

suffer from quadratic computational complexity, which restricts their ability to learn spatial LRDs only at the patch level rather than the pixel level. [10] attempts to model images at the pixel level with transformer, but it focuses on semantic features rather than spatial relations, which means it still lacks the ability to effectively capture spatial LRDs.

Mamba [34], emerging from the domain of long-sequence modelling, offers promising advantages for handling long sequential data and capturing long-range dependency efficiently, all at a linear computational cost. This capability makes Mamba particularly suitable for globally learning interactions at the pixel level, thus complementing transformers by adding detailed context.

We observe that, Mamba and transformer exhibit complementary strengths: Mamba is good at learning long-range pixel-wise dependency, which is computationally expensive for the transformer. Conversely, transformer is good at capturing global interactions between localized patches, such spatial awareness is an area that Mamba lacks due to it being designed for sequence modelling. Based on this observation, we propose $M \times T$, an architecture of mixture of Mamba and Transformer, consisting of proposed Hybrid Modules that synergistically combine the strengths of both transformer and Mamba. This novel approach allows for dual-level interaction learning from the patch level and pixel level.

However, as the vanilla SSM scans the data as a sequence with a single fixed direction, it lacks 2D spatial awareness, making the way to model pixels in SSM crucial. As illustrated in Sample II of Fig. 5.1, a vanilla SSM model [9] shows positional drifting of the inpainted left eye (upper than the right eye). This insight introduces two key challenges: (i) how to maintain the continuity and consistency of pixel adjacency for pixel-level dependencies learning while processing the SSM recurrence; and (ii) how to effectively integrate 2D spatial awareness to the predominant linear recurrent-based SSMs.

To further solve that, we propose **SEM-Net**: **S**patially-**E**nhanced SS**M** **Net**work for image inpainting, which is a simple yet effective encoder-decoder architecture. Both $M \times T$ and SEM-Net are evaluated on the widely-used CelebA-HQ and Places2-standard datasets, and overall outperform than existing state-of-the-art methods.

1.2 Problem Definitions

Formally, the problem is formulated as follows: given the original image I and Mask M, the input image I_{input}, is obtained by concatenating masked image $I_M = I \odot M$, and the mask M, where $\odot$ is element-else multiplication. The input image, I_{input}, is then processed by the inpainting model and a semantically accurate output image, I_C, will be generated. The whole formulation is denoted as: $I_C = f(I_{input})$, f is the inpainting model.

1.3 Research Aims

As mentioned above, how to **effectively use insufficient information from corrupted data** and **how to capture long-range dependency** are two critical perspectives to improve the quality of image inpainting, such that the inpainted images can provide more accurate and comprehensive context. Our research is driven by the following objectives:

1. **Effectively using Insufficient Data**:

    - **Geometric Feature Adjunction**: To more effectively use the information that remains in the corrupted images, we aim to develop an advanced multi-task framework to integrate the related geometric feature for image inpainting to guide the feature understanding, and explicitly complement the information that potentially missing in the representation learning.
    - **Avoiding Information Loss**: Most existing image-inpainting methods perform down-sampling via strided convolutions. However, convolutional down-sampling inherently filters each neighbourhood with a limited kernel and then sparsely samples the result, effectively a low-pass-plus-decimation operation that discards or aliases high-frequency details and disrupts dependencies across windows. To solve the information loss that happens in the convolutional-based down-sampling in existing methods. By exploring the optimal down-sampling method, and tailor-made tuning it to be adaptive to image inpainting, we aim to develop an advanced downsampling strategy,

enabling whole information to remain while reducing the resolution in image inpainting task.

2. **Long-range Dependency**: There are multiple ways to improve the long-range dependency, the critical challenge is the way to improve the long-range dependency while with a low computational cost.

    - **Long-range Pixel-wise Dependency**: We observe that Mamba is good at learning long-range pixel-wise dependency, which is computationally expensive for the transformer. By combining Mamba with the transformer, we aim to build a model for dual-level interaction learning from the patch level and pixel level to capture the long-range dependency.
    - **Enhancing Mamba for Image Inpainting**: To overcome the limitations of vanilla Mamba, which lacks 2D spatial awareness, we plan to extend it to capture both short- and long-range spatial dependencies efficiently, with linear computational cost.

Overall, our research advances image inpainting by addressing **insufficient information** and **long-range dependency**. To tackle insufficient data, we propose integrating geometric features to complement missing information and developing an adaptive downsampling strategy to prevent information loss. For long-range dependency, we introduce two methods: combining Mamba with transformers to leverage dual-level interactions at pixel and patch levels, and enhancing Mamba alone to efficiently capture both short- and long-range dependencies with improved spatial awareness. These innovations contribute to the more effective image inpainting models.

1.4 Contributions

The main contributions of this thesis are summarised as follows:

- A novel facial image inpainting architecture, to produce a non-cleft lip image from patients with cleft lips. We design adaptive feature fusion and landmark indicators to boost parameter sharing and utilise the second task more efficiently.

The landmark prediction is guided by both masked image and partial inpainted information, resulting in a more precise geometry indicator for repairing facial attributes.

- A novel end-to-end transformer-based architecture HINT for image inpainting, taking advantage of multi-scale feature- and spatial-level representations as well as pixel-level visual information. We propose a plug-and-play down-sampling module to preserve useful information while keeping irregular masks consistent during downsampling. Comparative experiments show that HINT outperforms SOTA image inpainting approaches across four datasets, CelebA [35], CelebA-HQ [4], Places2 [5] and Dunhuang challenge [36].

- Two novel Mamba-based architecture, $M \times T$ and SEM are proposed to learn the global and local representation while keeping a relatively low computational cost. $M \times T$ enables dual-level interaction learning, capturing both pixel-level and patch-level dependencies for enhanced global and local context modelling. SEM-Net further innovates with Snake Mamba Blocks, incorporating a Snake Bi-Directional Modelling module for spatial consistency and a Spatially-Enhanced Feedforward Network to refine local spatial dependencies. These architectures effectively balance computational efficiency and spatial awareness, enabling them to handle high-resolution image inpainting.

1.5 Publications

The research related to this thesis has been previously published in the following peer-reviewed publications:

- **Shuang Chen**, Amir Atapour-Abarghouei, Jane Kerby, Edmond S. L. Ho, David C. G. Sainsbury, Sophie Butterworth, Hubert P. H. Shum, "A Feasibility Study on Image Inpainting for Non-cleft Lip Generation from Patients with Cleft Lip." In *International Conference on Biomedical and Health Informatics* (BHI). IEEE, 2022. (Chapter 3)

- **Shuang Chen**, Amir Atapour-Abarghouei, Edmond S. L. Ho, Hubert P. H. Shum, "INCLG: Inpainting for Non-Cleft Lip Generation with a Multi-Task Image Processing Network." In *Software Impacts* (SIMPAC). ELSEVIER, 2023.
  (Chapter 3)

- **Shuang Chen**, Amir Atapour-Abarghouei, Hubert P. H. Shum, "HINT: High-quality INpainting Transformer with Mask-Aware Encoding and Enhanced Attention." In *IEEE Transactions on Multimedia* (TMM). IEEE, 2024.
  (Chapter 4)

- **Shuang Chen**, Amir Atapour-Abarghouei, Haozheng Zhang, Hubert P. H. Shum, "MxT: Mamba x Transformer for Image Inpainting." In *the 2024 British Machine Vision Conference* (BMVC), 2024.
  (Chapter 5)

- **Shuang Chen**, Haozheng Zhang, Amir Atapour-Abarghouei, Hubert P. H. Shum, "SEM-Net: Efficient Pixel Modelling for Image Inpainting with Spatially Enhanced SSM." In *the Winter Conference on Applications of Computer Vision* (WACV), 2025.
  (Chapter 5)

1.6 Thesis Structure

This thesis is structured to systematically explore and present the advancements in deep learning for image inpainting using advancing representation learning and generative model. The organisation of the chapters is designed to take the reader through the motivation, literature, methodology, and findings of the research coherently and logically.

In Chapter 1, we first introduce the our motivation for "**effectively using insufficient information**" and "**capturing long-range dependency**" in image inpainting, reveal the significant challenge that existing methods struggling. These challenges are particularly evident in scenarios with large masked regions, where local and global relations are disrupted, leading to inconsistencies in structure and texture. To address these challenges, the chapter outlines the research aims, which include developing methodologies to preserve valid information, enhance geometric feature adjunction, and prevent

information loss during down-sampling. Additionally, it emphasises the need for modelling long-range dependencies efficiently through dual-level interaction learning and extending Mamba to capture spatial dependencies effectively. This chapter also defines the image inpainting problem, presenting a formal mathematical formulation. After that, we propose the aims that we are targeting to solve in this research, followed by the contributions bullet points and a list of peer-reviewed publications.

Chapter 2 provides an overview of the evolution of image inpainting, starting with traditional approaches, which focus on neighbouring pixel information or external patches to complete missing regions. Then we discuss the learning-based methods, highlighting their success in generating contextually coherent content for large missing regions. This section also explores advances in convolutional and GAN-based architectures, including innovations like partial and gated convolutions, contextual attention mechanisms, and multi-scale patch matching strategies. These developments underscore the ongoing efforts to extract valid information from known regions, while addressing challenges such as enlarging receptive field while saving the computational cost.

In Chapter 3, we validate the feasibility of applying a novel generative image inpainting model to generate non-cleft lip images from cleft lip images, while presenting the system design and implementation details of the proposed framework. This section outlines a multi-task architecture, which integrates facial landmark prediction and image inpainting to enhance parameter sharing and feature interaction. It describes the use of gated convolution layers for efficient inpainting and adaptive fusion mechanisms to refine feature maps. Additionally, the section highlights the data preparation process, including the open facial datasets and the ethical considerations for the steps taken to collect and validate real patient data for testing.

After focusing on facial inpainting, where strong domain priors (e.g., landmarks, facial symmetry, identity cues) can be explicitly leveraged for structure and identity consistency—these assumptions do not generalize to open-world scenes, we move on to the General-scene inpainting, which must cope with diverse layouts, noncanonical objects, and high-frequency textures without reliable semantic anchors. Building on the insights from the facial setting, we retain the core principles of information-preserving down-sampling, mask-aware feature routing, and long-range context aggregation, but we

recast them into category-agnostic modules. Differences are shown in Tab. 1.1. Chapter 4 therefore relaxes identity-specific conditioning and replaces face-centric supervision with self-consistent structural and textural cues, enabling our method to scale beyond faces and deliver robust completion on unconstrained images.

In Chapter 4, we present the High-quality INpainting Transformer (HINT), a novel transformer-based architecture designed to address the challenges of image inpainting. This section provides a detailed explanation of the mask-aware pixel-shuffle down-sampling strategy employed to minimise information loss and preserve the consistency of visible and masked regions. Comprehensive evaluations are conducted on multiple benchmarks, including CelebA-HQ and Places2, to highlight HINT's effectiveness

In Chapter 5, the focus shifts to the integration of Mamba-based architectures for improving long-range dependency modelling in image inpainting. We introduce two Mamba-based architecture, $M \times T$ and SEM, $M \times T$ combines the strengths of Mamba and transformers for dual-level interaction learning, capturing both pixel-level and patch-level dependencies. Additionally, we propose SEM-Net, a spatially-enhanced architecture that incorporates Snake Mamba Blocks (SMBs) to enhance spatial consistency and long-range dependency modelling. Detailed experimental setups, evaluations, and ablation studies are presented to demonstrate the superior performance and efficiency of these methods.

In Chapter 6, we conclude by reviewing the main contributions of this thesis, highlighting the innovative solutions proposed to address the challenges of "**effectively using insufficient information**" and "**capturing long-range dependency**". These include the development of the advanced frameworks for non-cleft lip generation, transformer-based architecture and Mamba-based architecture. Additionally, we outline the potential future directions, emphasising key areas for further advancements and innovations to enhance the field of image inpainting.

Table 1.1: Facial vs. general-scene inpainting: typical priors, representative methods, strengths, and limitations across domains.

Domain	Typical Priors / Assumptions	Representative Methods	Strengths	Limitations
Facial images	Structured anatomy; availability of identity cues (landmarks, parsing maps); strong local symmetry; datasets like CelebA / CelebA-HQ	LaFIn [3]; Identity-preserving face completion [37].	High fidelity on facial components; identity preservation and realism within distribution; works well with semantic/landmark guidance	Domain-specific—can overfit to facial statistics; degraded generalization to non-face content or extreme poses/occlusions; reliance on identity/landmark supervision
General scenes	Broad textures and layouts; often no explicit semantic priors; free-form, irregular masks; large holes	DeepFill / Gated Convolution [18]; EdgeConnect (edge-guided) [38]; LaMa (Fourier conv.) [39]	Robust to irregular masks; strong performance across diverse categories; good structure completion via edges or global receptive fields; scales to large holes/resolutions	May blur fine semantics (e.g., faces) without domain priors; struggles with identity-specific consistency; artifacts on periodic or high-frequency patterns without specialized design

CHAPTER 2

Literature Review

This chapter surveys the literature that underpins our subsequent contributions, moving from foundations to the most recent advances. We begin by outlining the two principal families of modern generative models—GANs and diffusion models—and how their objectives, strengths, and weaknesses relate to inpainting (Section 2.1). We then revisit pre–deep learning inpainting to clarify the historical assumptions and limitations of exemplar- and diffusion-based techniques (Section 2.2). Next, we review convolutional approaches, including partial/gated convolutions and attention-augmented CNNs that improve local texture synthesis while struggling with long-range dependencies (Section 2.3). Building on this, we introduce visual transformers, covering tokenisation and self-attention in vision (Section 2.4), and summarize transformer-based inpainting methods that address global context at varying computational costs (Section 2.5). Finally, we discuss State Space Models—especially Mamba—and their emerging role in efficient long-sequence modelling for vision (Section 2.6). We close with datasets and metrics commonly used to evaluate inpainting quality (Section 2.7), establishing consistent ground for the experiments in later chapters.

2.1 Generative Models

Generative models are a class of machine learning models that learn the underlying probability distribution of a dataset in order to generate new samples that resemble the original data [40, 41]. In contrast to discriminative models that predict labels or targets, generative models aim to model the data distribution itself, capturing the essential structure and variability in an unsupervised manner. By learning a good approximation of the data-generating process, these models can synthesize novel examples to produce realistic images, plausible text, or coherent audio. Over the past decade, rapid advances in deep learning have led to powerful generative modeling frameworks, including variational autoencoders (VAEs) [42], generative adversarial networks (GANs) [43], autoregressive models [44, 45], and diffusion models [46], among others. Each framework embodies different theoretical principles and trade-offs in how it learns data distributions, reflecting the evolution of the field and the pursuit of more expressive, stable, and high-fidelity generative systems. In this research, we only focus on GANs and diffusion-based methods, which are widely used in image inpainting task. These approaches have demonstrated superior performance in generating visually coherent and semantically consistent content, making them particularly well-suited and empirically validated as state-of-the-art methods for image inpainting compared to other generative frameworks, such as VAEs or autoregressive models.

Generative models can be broadly categorised by how they learn and represent the data distribution. One useful distinction is between explicit density models and implicit density models. Explicit models define a probability distribution for data (often enabling likelihood estimation) and are typically trained by maximising likelihood or a surrogate objective. Diffusion models represent a unique class of Explicit models, as they explicitly define a probability distribution over the data. Implicit models, on the other hand, do not directly specify a probability density. Instead, they learn to generate samples that match the data distribution without computing explicit likelihoods. GANs are the quintessential implicit generative models: rather than maximise a likelihood, GANs train a generator network to produce data that a discriminator network cannot distinguish from real data, thereby implicitly learning the target distribution.

2.1.1 Generative Adversarial Networks (GANs)

[43] introduces Generative Adversarial Networks (GANs), which take a very different approach to generative modelling by casting it as a two-player minimax game. In a GAN, there are two neural networks contesting with each other: a generator G that tries to produce realistic data (e.g., images) from random noise, and a discriminator D that examines data and attempts to distinguish real samples (from the training set) from fake samples produced by G. The generator and discriminator are trained simultaneously with opposing objectives: D is optimised to correctly classify inputs as real or fake, while G is optimised to fool D. Such that, GANs are to maximise the probability that D misclassifies the generated output as real. Formally, this setup corresponds to the following minimax game:

$$\min_G \max_D V(G,D) = \mathbb{E}_{x\sim p_{\text{data}}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}\Big[\log\Big(1 - D(G(z))\Big)\Big]. \quad (2.1)$$

At the theoretical equilibrium of this game, the generator G would exactly reproduce the training data distribution and the discriminator would be unable to tell real from fake. In practice, although perfect equilibrium is seldom achieved, this adversarial learning process has shown remarkable ability for the generator to learn complex data distributions implicitly, without ever explicitly computing a likelihood or distance measure between distributions.

The fundamental principle of GANs is this adversarial training dynamic, which differs from likelihood-based training in that it uses a learned discriminator to drive the improvements of the generator. One intuition is that the discriminator is learning an adaptive loss function: rather than a fixed pixel-wise loss, it learns which differences between real and generated data are important and guides the generator to fix those. This often leads GANs to produce visually sharper and more realistic outputs than VAEs [47,48], since G is trained to focus on realism and fine details that fool D, rather than to average over possible outputs. Indeed, GANs have been credited with redefining the state-of-the-art in realistic image generation, producing photographs of people, objects, and scenes that are often indistinguishable from real images to human observers. Over the past ten years, GANs and their variants (DCGAN [49], WGAN [50], StyleGAN [51], BigGAN [52],

etc.) have dominated many image synthesis benchmarks, demonstrating the ability to capture high-frequency details and complex textures in a way earlier likelihood-based models struggled to match.

However, GANs come with their own set of challenges and characteristics. Training GANs can be unstable due to the delicate balance required between the generator and discriminator. If one network overpowers the other, the adversarial loss can converge very soon. Considerable research has gone into techniques to stabilise GAN training, such as using alternative loss functions or architectural constraints and regularisation. For example Wasserstein distance with gradient penalty in WGAN to address gradient vanishing. Even with these improvements, mode collapse is a well-documented issue, which means the generator may find it easier to produce a limited variety of outputs that consistently fool the discriminator, resulting in lack of diversity. This happens because nothing in the standard GAN objective explicitly forces G to cover all modes; it only needs to generate something in each minibatch that can fool D.

Despite being implicit models, GANs demonstrate that adversarial training can yield latent representations that capture meaningful factors of variation. The discriminator learns to encode high-level representations of the data, which are usually related to the semantic content, to tell if an image looks real. Additionally, some GAN variants introduce an encoder or use techniques like bi-directional GANs to map data into the latent space of the generator, effectively adding an inference mechanism to GANs for representation learning. In terms of practical downstream use, GAN-based models have been applied to tasks such as image inpainting, where the generator is tasked with filling in missing regions of an image in a realistic way. By training on large image datasets, a GAN can learn to plausibly imagine the missing content consistent with the surrounding context. For example, if part of an image of a face is masked out, a GAN trained on faces can generate a plausible guess of the missing part (eyes, nose, etc.) that looks coherent. Early works showed that incorporating adversarial loss in inpainting (besides a reconstruction loss) yields much sharper and more realistic filled regions. This underscores how generative modeling advances have enabled new capabilities in image restoration and reconstruction.

2.1.2 Diffusion Models

The fundamental idea of diffusion models is to train a model to reverse a gradual noising process. Diffusion models represent the latest wave of innovation in deep generative modelling. The concept of diffusion-based generation originates from nonequilibrium thermodynamics and was first explored in the context of deep learning [53], but only recently have diffusion models demonstrated state-of-the-art results in generating high-quality images and other data [54]. In a typical diffusion model, a forward diffusion process is defined wherein data x_0 is progressively corrupted by noise through a sequence of T small steps, eventually yielding a sample of pure noise x_T, which is usually from Gaussian noise. This forward process is a fixed Markov chain that ensures by the final step the original structure in the data is almost entirely destroyed. The generative model works by learning a reverse diffusion process. In this process, a neural network is trained to denoise. It predicts the distribution of x_{t-1} from a noisy x_t at each step. The model starts with random noise x_T and gradually refines it. By applying these denoising steps in sequence, it produces a coherent sample x_0 that resembles real data [55, 56].

The training of diffusion models typically uses variational inference or score matching. One common approach trains the denoising network to predict either the original data x_0 or the added noise from a partially noised input x_t at various noise levels [55]. This minimizes a weighted sum of reconstruction losses over all timesteps and can be shown to maximize a variational lower bound on the data likelihood. An alternative perspective is that the model learns the score function of the noisy data distribution at each noise level [56]. It then uses Langevin-like sampling to generate data, linking diffusion models to energy-based models and score matching. Since each step involves only a small denoising task rather than generating an image from scratch, the training is more stable. There is no adversarial objective, and each denoising step is trained with a simple loss (typically an L_2 loss on the noise prediction). This stability helps diffusion models avoid issues like mode collapse that can affect GANs [57].

However, a key drawback of diffusion models is their computational complexity [58, 59]. Generating a single sample requires iterating through hundreds or even thousand denoising steps, each of which involves a forward pass through a large neural network. As a result, producing a batch of samples is considerably more computationally intensive

compared to one-shot generative models [59]. To address these challenges, researchers are exploring several acceleration techniques. These include reducing the number of timesteps with noise-aware sampling [60], and employing distillation methods to skip intermediate steps [61]. Another issue is that diffusion models typically operate in pixel space or in a high-dimensional latent space, which can be inefficient. Recent advances such as latent diffusion [62] mitigate this by compressing images into a lower-dimensional latent space using an autoencoder before applying diffusion. In terms of evaluation, diffusion models do permit likelihood estimation but computing the exact likelihood is not as straightforward as in autoregressive or flow models. Nonetheless, the field has largely embraced diffusion models for their empirical performance. Nevertheless, the inference speed of the diffusion model is still a shortcoming that limits its widespread application in downstream tasks [58], such as large-area image restoration tasks [63].

2.2 Non-Deep Learning Image Inpainting

Image inpainting predates learning-based techniques and the literature on image completion based on conventional strategies is extensive. Traditional approaches (non-deep learning methods) complete minor and narrow stretches using neighbouring visible pixels [21], these methods fill in missing or damaged regions by copying pixels from other parts of the same image or from a database of external images. Exemplar-based methods infer missing regions with plausible edge information based on other patches from background or external data [22, 23], these methods involve searching for the most similar patches to the missing region's boundary and using them to fill in the missing regions. This process often involves iteratively searching for best-matching patches and blending them seamlessly into the missing regions. Such methods are generally effective for reconstructing images with small and constrained missing regions as they are able to produce visually plausible results by leveraging existing textures and patterns. However, a key limitation is that, these methods are unable to generate novel features or content that is not present in the source images or database. Therefore, they struggle to produce satisfactory results while dealing with large missing regions requiring imaginative reconstruction.

2.3 Convolution-based Image Inpainting

Compared with traditional methods, learning-based methods have achieved great success in inpainting, especially when it comes to generating new contextually sound content for large missing regions. [13] proposed a parametric framework for image inpainting based on an encoder-decoder architecture taking advantage of a Generative Adversarial Network (GAN) [43]. Subsequently, numerous GAN-based methods emerged to offer improved inpainting quality [2, 15–17, 32, 38, 64–66] using better training strategies. [14] use two discriminators to calculate both global and local adversarial losses. [29] propose region-wise normalisation for missing and visible areas. Partial [15] and gated convolutions [16, 17] are introduced to handle the irregular masks by improving the convolution operation [32, 64] to efficiently extract valid information for inpainting. [18] propose contextual attention to facilitate the matching of feature patches across distant spatial locations. Building on this, [2, 19] extended [18] by incorporating a multi-scale patch size to further improve its efficiency. [67] introduces fourier convolution-based encoder for image inpainting to avoid generating invalid features inside the missing regions.

Despite the advancements, a persistent challenge in learning-based inpainting methods is the information loss caused by convolutional downsampling, particularly for those methods that rely on convolutional networks. This information loss can suppress the generation of fine-grained details and texture in the inpainted regions. In addition, these convolution-based methods struggle with the slow expansion of the receptive field, which limites their ability to model complex long-range dependencies.

2.4 Visual Transformer

Transformers [68] have revolutionised natural language processing with their self-attention mechanism and have recently gained prominence in computer vision applications [69, 70]. Unlike convolutional networks, transformers excel in capturing long-range dependencies due to their global receptive field. This property makes them highly suitable for tasks requiring the modelling of extensive contextual relationships, such as image inpainting [8, 71–75].

2.4.1 Embedding an Image into Tokens

To process an image with transformers, it must first be converted into a sequence of tokens. This can be achieved through two main approaches: extracting patches or using overlap convolution.

Patch Extraction and Projection

Given an image $I \in \mathbb{R}^{H \times W \times C}$, where H, W and C represent the height, width and number of channels, respectively, is divided into non-overlapping patches of size $P \times P$. Each patch is flattened into a 1D vector of size $P^2 \cdot C$ and then projected into a d-dimensional token embedding using a linear layer:

$$T_i = \text{Linear}(x_i), \quad \forall i \in [1, N], \tag{2.2}$$

where $x_i \in \mathbb{R}^{P^2 \cdot C}$ represents the i-th patch, and $N = \frac{H \times W}{P^2}$ is the total number of patches. Since this approach disrupts the spatial relations between patches, positional encodings are added to the tokens to retain the spatial structure:

$$Z_i = T_i + \text{PE}_i, \quad \forall i \in [1, N], \tag{2.3}$$

where $PE_i \in \mathbb{R}^d$ represents the positional encoding for the i-th patch.

Overlap Convolution and Projection

Another approach embeds the image using overlapped convolutions, which preserve spatial relationships inherently. For an image I, overlapped convolution generates feature maps that embed local spatial information directly. These feature maps are then projected into token embeddings:

$$T_i = \text{Linear}(Conv(x_i)), \quad \forall i \in [1, N], \tag{2.4}$$

where $Conv$ represents the overlapped convolution operation. Unlike the patch-based approach, positional encoding is unnecessary, as the convolution operation already incorporates positional information into the embeddings.

Both methods result in a sequence of tokens, which serve as input to the transformer. The choice between these methods depends on the specific requirements of the application. In our research, we use overlapped convolution to embed the tokens to provide more comprehensive spatial information.

2.4.2 The Self-Attention Mechanism in Vision

The core of the transformer is the self-attention mechanism, which computes the relationship between input tokens. Given an input sequence of features $X \in \mathbb{R}^{N \times d}$, the the scaled dot-product attention is calculated as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V, \tag{2.5}$$

where: $Q = XW_Q$ is the query matrix, $K = XW_K$ is the key matrix, $V = XW_V$ is the value matrix, d is the dimension of the query and key. Here,W_Q, W_K, and W_V are learnable projection matrices. This formulation allows the model to attend to all positions in the input simultaneously, enabling the extraction of global contextual information.

While transformer is able to handle global relations, the quadratic complexity $O(N^2d)$ of self-attention poses a computational bottleneck for high-resolution images.

2.5 Visual Transformers for Image Inpainting

The notable success of Transformers [68] in natural language processing has recently prompted research into their applicability in computer vision [69, 70]. Driven by this, efforts were focused towards applying transformers to image inpainting [8, 71–75]. However, spatial-based self-attention incurs an expensive computational cost. To reduce computation, [71, 72] down-sample the input image into a lower resolution. [8, 73, 75] calculate the spatial self-attention after encoding the input image into low-resolution features. Nonetheless, these approaches fail to change the quadratic complexity of spatial self-attention, which restricts its applicability to high-frequency features.

The Swin Transformer [70] mitigates this issue by introducing a hierarchical structure and shifted window-based attention. The computational complexity is reduced to linearity,

as attention is calculated within non-overlapping windows. For an input image of size $H \times W$ with a window size of $M \times M$, the complexity is reduced to:

$$O\left(\frac{HW}{M^2} \cdot M^2 d\right) = O(HWd), \tag{2.6}$$

where H, W represent the height and width of the image, respectively. However, the shifted-window design splits the local neighbourhood context of the visible and missing area, and thus is not ideally suited for inpainting. [10] propose utilising channel-wise self-attention in multi-scale representation with linear complexity for image reconstruction. Its variant [6] demonstrates the applicability in image inpainting. Nevertheless, both of these models omit spatial attention that is vital in delivering high-quality and contextually sound results. In contrast, our model integrates multiscale channel and spatial attention in an efficient manner, thus resolving the issue that prior work has struggled with [8, 72, 73].

2.6 State Space Models and Mamba

2.6.1 Preliminary

State Space Models (SSMs) are a foundational framework in systems theory, widely used for modelling linear time-invariant systems. In their essence, SSMs describe the evolution of a system's hidden state over time and its relationship with input and output sequences. Specifically, an input sequence $x(t) \in \mathbb{R}$ is mapped to an output response $y(t) \in \mathbb{R}$ via a hidden latent state $h(t) \in \mathbb{R}^N$. The state-space representation is given by:

$$h'(t) = \boldsymbol{A}h(t) + \boldsymbol{B}x(t) \quad , \qquad y(t) = \boldsymbol{C}h(t) \quad ,, \tag{2.7}$$

where $A \in \mathbb{R}^{N \times N}$ represent the system dynamics, $B \in \mathbb{R}^{N \times 1}$ defines how the input sequence influences the state, and $C \in \mathbb{R}^{1 \times N}$ maps the hidden state to the output sequence.

2.6.2 Zero-Order Hold Discretisation

For modern deep learning applications, it is essential to implement SSMs in discrete time, which requires converting the continuous-time parameters $(\Delta, \boldsymbol{A}, \boldsymbol{B})$into discrete-time equivalents $(\overline{\boldsymbol{A}}, \overline{\boldsymbol{B}})$, A widely used method for discretization is the zero-order hold rule, which approximates the continuous system by assuming the input remains constant within each time step. The discrete parameters are computed as:

$$\overline{\boldsymbol{A}} = \exp(\Delta \boldsymbol{A}), \quad \overline{\boldsymbol{B}} = (\Delta \boldsymbol{A})^{-1}(\exp(\Delta \boldsymbol{A}) - \boldsymbol{I}) \cdot \Delta \boldsymbol{B}, \tag{2.8}$$

where Δ represents the time scale, the $\exp \Delta A$ is the matrix exponential of ΔA.

The resulting discrete state-space equations are expressed as:

$$h_t = \overline{\boldsymbol{A}} h_{t-1} + \overline{\boldsymbol{B}} x_t \quad , \qquad y_t = \boldsymbol{C} h_t \quad . \tag{2.9}$$

2.6.3 Selective State Space Models and Mamba

The structured nature of traditional SSMs makes them computationally efficient but often limits their ability to adapt to dynamic and complex input-output relationships. To address this limitation, recent advancements have introduced selective mechanisms into SSMs, enabling them to reason more effectively about content and context.

Mamba [34], one of the most recent selective SSMs, innovates upon traditional models by introducing a gated selective mechanism. This mechanism allows the model to dynamically propagate or eliminate specific information based on the current system state. The key innovation of Mamba lies in its shift from time-invariant to time-varying parameters, enabling it to adapt to varying inputs.

Specifically, Mamba redefines the parameters (Δ, B, C) as input-dependent functions. Let x_t denote the input at time t. Then, Mamba introduces functions $\Delta(x_t)$, $B(x_t)$, and $C(x_t)$ that dynamically adjust the model's behaviour based on the current input:

$$\Delta(x_t) = \sigma(W_\Delta x_t + b_\Delta), \quad \boldsymbol{B}(x_t) = \sigma(W_B x_t + b_B), \quad \boldsymbol{C}(x_t) = \sigma(W_C x_t + b_C), \tag{2.10}$$

where $\sigma()$ is a nonlinear activation function, W_Δ, W_B, W_C are learnable weight matri-

ces. This dynamic adjustment improves the model's ability to handle complex temporal dependencies and enhances its performance in content-reasoning tasks. Mamba has demonstrated significant improvements in tasks requiring precise, context-aware temporal modelling, making it a powerful tool for sequential data analysis.

2.6.4 SSMs in Computer Vision

Recently, State Space Models (SSMs) have demonstrated promising advantages of long sequence modelling and linear-time complexity in Natural Language Processing (NLP) [76]. This work specifically tackles the problem of vanishing gradients in SSMs when solving the exponential function by the linear first-order Ordinary Differential Equations [77]. Building on the rigorous theoretical proofs of the HiPPO framework that enables SSMs to capture long-range dependencies, Gu et al. [34] further introduce a data-dependent selective structure SSM (i.e., Mamba) to significantly improve the computational efficiencies in conventional SSMs. Inspired by pioneering SSMs, vision-specific adaptations of the Mamba architecture, such as Vision Mamba [78] and V-Mamba [79], propose visual SSMs designs for computer vision tasks including image classification and object detection [78, 79]. SSMs are appealing for vision because they aggregate long-range context in linear time with bounded memory, which scales favourably to high-resolution inputs compared with quadratic-cost attention; moreover, their continuous-time formulation and content-selective gating (as in Mamba) enable adaptive information propagation that is robust to occlusions, large masks, and heterogeneous textures often seen in dense prediction. However, their performance is still behind of the state-of-the-art transformer-based models like SpectFormer [80], SVT [81], and WaveViT [82]. U-Mamba [9] effectively extends the capabilities of Mamba for biomedical image segmentation by proposing a hybrid CNN-SSM block. However, these studies ineffectively leverage the capabilities of Mamba in image long-range pixel-level dependency learning and overlook the critical spatial awareness during model designs.

2.6.5 Mamba-based Image Inpainting

Selective State Space Models (SSMs), especially Mamba, have recently been explored for inpainting due to their linear-time long-range modeling and favorable memory footprint. Xiang et al. propose a Mamba-GAN pipeline that embeds visual state-space operators (e.g., VSS with SS2D scanning) into the generator/discriminator, targeting high-resolution hole filling and reporting quality–efficiency gains on natural images [83]. Beyond purely adversarial designs, Wen et al. introduce a hybrid restoration framework (MatIR) that alternates Transformer and Mamba layers and includes an Image Inpainting State Space module with multi-directional scans; while framed as general restoration, the inpainting-oriented state-space block is directly evaluated for completion quality [84]. In a more U-Net–style architecture, Sandooghdar and Yaghmaee integrate "U-Net mamba" blocks with parameter-assisted and edge-guided priors, showing that lightweight SSM components can maintain structural integrity and perceptual quality in inpainted regions [85].

2.6.6 Limitations for GAN, Diffusion-based model, Transformer and Mamba

Despite rapid progress, the three dominant generative families used for inpainting exhibit characteristic limitations. GAN-based methods can suffer from training instability and mode collapse, often producing sharp but semantically inconsistent fills and visible seams near mask boundaries. In our work, we mitigate this by employing a composite training objective that balances adversarial, reconstruction, and perceptual/style losses to stabilize optimization and maintain diversity). Diffusion models offer stable optimization but incur high sampling cost—multi-step sampling frequently leads to inference times of several seconds to minutes, whereas our feed-forward design produces results in sub-second latency (on the order of tenths of a second) at comparable resolutions. Transformer backbones, while capturing global dependencies, typically incur quadratic attention cost in the number of pixels. Restormer [10] reduces this to linear complexity via an efficient attention formulation, yet it lacks explicit pixel-aware spatial attention for precise boundary harmonization. Building on a linear-complexity backbone, our method

introduces mask-aware, pixel-level spatial attention to refine fine structures without sacrificing efficiency. SSM variants reduce complexity with linear-time scans but may be sensitive to scan order and anisotropy, leading to drifting textures or misaligned geometry. To alleviate this, our approach introduces 2D-consistent, mask-aware positional indexing with multi-directional (cross-scan) alignment and boundary-aware feature routing, which enforces spatial correspondence across scans and preserves local neighborhoods near hole edges. These issues highlight the need for mask-aware conditioning, information-preserving down-sampling, multi-scale context aggregation, and lightweight global modules that reconcile structure and detail under diverse, irregular missing regions.

2.7 Datasets and Metrics

In this section, we review the literature related to image inpainting, focusing on both the datasets commonly employed in this domain and the evaluation metrics used to quantify reconstruction quality. We first introduce several benchmark datasets—CelebA, CelebA-HQ, Places2, and the Dunhuang dataset—that provide diverse visual content ranging from high-resolution facial images and natural scenes to culturally significant artworks. These datasets serve as the foundation for developing and testing state-of-the-art inpainting methods. We then discuss a suite of metrics—PSNR, SSIM, LPIPS, L1 loss, and FID—that have become standard for assessing the performance of generative models in restoring missing image regions. Each metric offers unique insights into the fidelity, structural similarity, perceptual quality, and overall realism of the inpainted outputs, thus enabling a comprehensive evaluation of the underlying algorithms.

2.7.1 Datasets

Table 2.1: Comparison of datasets used for image inpainting.

Dataset	Number of Images	Resolution	Domain / Characteristics
CelebA	202,599 (10,177 identities)	$\sim 178 \times 218$	Faces, diverse poses and backgrounds
CelebA-HQ	30,000	1024×1024	High-quality faces, photorealistic details
Places2	$>$10 million (400+ categories)	Varies ($\sim 256 \times 256$ common)	Large-scale natural scenes, indoor/outdoor variety
Places365-Standard	$\sim$1.8 million (365 classes)	$\sim 256 \times 256$	Subset of Places2, scene classification benchmark
Dunhuang	600 (500 train / 100 test)	$\sim 500 \times 800$	Ancient murals, cultural heritage restoration

CelebA

The CelebFaces Attributes (CelebA) dataset is a large-scale face image dataset with 202,599 images of 10,177 celebrities with around 178×218. The images cover diverse facial poses and backgrounds, making CelebA useful for tasks like face recognition, attribute prediction, and generative modelling. In image inpainting research, CelebA often serves as a benchmark for face completion tasks, given its rich annotations and variety.

CelebA-HQ

CelebA-HQ is a high-quality subset of CelebA introduced to facilitate photorealistic image generation and editing. It consists of 30,000 human face images at 1024×1024 resolution, derived and enhanced from the original CelebA dataset. This dataset retains the diversity of CelebA but with higher fidelity images, making it become a standard for evaluating facial inpainting due to its quality and resolution.

Places2

The Places2 dataset is a large-scale scene image database designed for high-level visual understanding and frequently used in inpainting tasks. It contains over 10 million images spanning 400+ scene categories, with each category (e.g., different types of indoor and outdoor environments) having 5,000 to 30,000 training images. A widely used subset is Places365-Standard, with 1.8 million training images across 365 scene classes. Places2 provides diverse and complex backgrounds (city streets, landscapes, rooms, etc.), which helps train and benchmark image inpainting models on general scenes. In fact, many state-of-the-art inpainting methods evaluate on Places2 because of its challenging variety of real-world images. In this research, unless otherwise stated, the Places2-standard dataset is implemented.

Dunhuang

This is a specialised dataset released for an e-Heritage restoration challenge (ICCV 2019) focusing on ancient mural inpainting. It comprises 600 digital photographs of the Mogao

Grotto cave paintings in Dunhuang, with image resolutions around 500×800 pixels. These images capture both well-preserved and deteriorated regions of the murals, covering themes like Buddha figures, architecture, and decorative patterns. The dataset is split into 500 training images and 100 testing images. Researchers use this dataset to develop and evaluate inpainting algorithms for cultural heritage restoration, simulating the filling-in of damaged parts of the ancient artwork. The Dunhuang dataset thus provides a unique testbed for inpainting models on non-photographic, high-historical-value images.

2.7.2 Metrics

Peak Signal-to-Noise Ratio (PSNR)

PSNR is a classic full-reference metric for quantifying image inpainting quality. It is defined in logarithmic decibel scale based on the mean squared error (MSE) between a reconstructed image and the ground truth. For an 8-bit image with pixel values in [0,255], PSNR is given by $PSNR = 10 \log_{10} \left(\frac{255^2}{MSE} \right)$. A higher PSNR indicates that the inpainted image is closer to the original, with less error. In image inpainting, PSNR is used to measure how faithfully missing regions are filled – e.g., a model achieving higher PSNR on a test set produces more accurate, less noisy restorations.

Structural Similarity Index Measure (SSIM)

SSIM [86] is a perceptual metric that evaluates image similarity in terms of structure, luminance, and contrast, rather than absolute pixel differences. It computes similarity by comparing local patterns of pixel intensities between an output and reference image, combining measures of luminance, contrast, and structural correlation (often via mean μ, variance σ^2, and covariance σ_{xy}). The SSIM index ranges from −1 to 1 (with 1 indicating identical images). A higher SSIM means the inpainted image retains more structural information of the ground truth. In image inpainting studies, SSIM is a common metric to report because it reflects the perceptual integrity of filled regions. A model with higher SSIM produces outputs that are structurally more similar to the originals, indicating more

plausible textures and object shapes.

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}, \tag{2.11}$$

where μ_x and μ_y are the means of images x and y, σ_x^2 and σ_y^2 are the variances, σ_{xy} is the covariance, the C_1 and C_2 are small constants to avoid division by zero.

Learned Perceptual Image Patch Similarity (LPIPS)

LPIPS is a learned perceptual metric proposed by Zhang et al. (2018) that measures similarity between images using deep network feature representations. Instead of comparing pixels, LPIPS passes image patches through a pretrained convolutional neural network (e.g. AlexNet or VGG) and computes the distance between the two images in the feature level. These feature distances are then weighted by learned parameters calibrated to human judgments, producing a score that correlates with human perceptual similarity. A lower LPIPS score indicates two images look more alike perceptually. In the context of image inpainting, LPIPS is often used to evaluate the visual realism of the filled-in regions, such as whether textures and structures appear natural to a human observer. Since LPIPS captures high-level differences that may not show in PSNR/SSIM, it complements those metrics.

$$\text{LPIPS}(x, y) = \sqrt{\sum_l \frac{1}{H_l W_l} \sum_{h=1}^{H_l} \sum_{w=1}^{W_l} \left\| w_l \odot \left(\hat{f}_l(x)_{hw} - \hat{f}_l(y)_{hw} \right) \right\|_2^2}. \tag{2.12}$$

In this equation, $\hat{f}_l(x)$ and $\hat{f}_l(y)$ are the normalised feature maps from layer l of a pretrained network for images x and y, respectively. H_l and W_l denote the height and width of the feature map at layer l. w_l are learned scalar weights applied channel-wise. The symbol $\odot$ denotes element-wise multiplication.

Mean Absolute Error (L1)

L1 loss refers to the mean absolute difference between the predicted image and the ground truth image, computed pixel-wise. In formula, $L_1 = \frac{1}{N} \sum_{i=1}^{N} |P_i - G_i|$, where P_i and G_i are the pixel values of the predicted and ground truth images respectively. This metric

(also known as MAE) measures the average magnitude of errors without regard to their direction. A lower L1 value implies the inpainted image is closer to the original on average, indicating more accurate pixel reconstruction. In image inpainting literature, L1 is often used as a loss function during training, due to its robustness against outliers and tendency to produce sharper results than L2 loss. L1 can also serve as an evaluation metric for reconstruction error, for example, reporting the mean L1 error on a masked region quantifies how well a model restored the missing content. Generally, a smaller L1 is associated with better inpainting quality.

Fréchet Inception Distance (FID)

FID [87] is a distribution-based metric widely used to assess the realism of generated images, including inpainted results. It compares the statistics of deep features (typically from an InceptionV3 network [88]) between a set of generated images and real images. The inpainted images are passed through the Inception network to obtain feature vectors (usually from a late layer), and assuming these features follow a multivariate Gaussian, FID computes the Fréchet distance between the two Gaussians – one for the model's outputs and one for the ground-truth data. The formula for FID between the real image distribution $\mathcal{N}(\mu_r, \Sigma_r)$ and generated image distribution $\mathcal{N}(\mu_g, \Sigma_g)$ is:

$$\text{FID} = \|\mu_r - \mu_g\|^2 + \text{Tr}\left(\sigma_r + \sigma_g - 2\left(\sigma_r\sigma_g\right)^{\frac{1}{2}}\right), \tag{2.13}$$

where μ and σ are the feature-wise mean and covariance of the two sets. Lower FID values indicate that the feature distribution of inpainted images is closer to that of real images, meaning higher fidelity and diversity in a perceptual sense. In image inpainting evaluations, FID is very important for judging visual realism: a model that produces inpainting results with unnatural artifacts or mode collapses will have a higher FID, whereas a model that generates plausible, varied completions yields a low FID. FID has become a standard metric [87] for comparing generative models and is often reported to demonstrate the improvements in producing more photo-realistic inpainted results.

CHAPTER 3

Image Inpainting for Non-Cleft Lip Generation

Portions of this chapter have previously been published in the following peer-reviewed publication [66, 89]:

- **Shuang Chen**, Amir Atapour-Abarghouei, Jane Kerby, Edmond S. L. Ho, David C. G. Sainsbury, Sophie Butterworth, Hubert P. H. Shum, "A Feasibility Study on Image Inpainting for Non-cleft Lip Generation from Patients with Cleft Lip." In *International Conference on Biomedical and Health Informatics* (BHI). IEEE, 2022.
- **Shuang Chen**, Amir Atapour-Abarghouei, Edmond S. L. Ho, Hubert P. H. Shum, "INCLG: Inpainting for Non-Cleft Lip Generation with a Multi-Task Image Processing Network." In *Software Impacts* (SIMPAC). ELSEVIER, 2023.

In this work, we present a software that predicts non-cleft facial images for patients with cleft lip, thereby facilitating the understanding, awareness and discussion of cleft lip surgeries. To protect patients' privacy, we design a software framework using image inpainting, which does not require cleft lip images for training, thereby mitigating the risk of model leakage. We implement a novel multi-task architecture that predicts both the non-cleft facial image and facial landmarks, resulting in better performance as evaluated by surgeons. The software is implemented by PyTorch and is usable with consumer-level colour images with a fast prediction speed, enabling effective deployment.

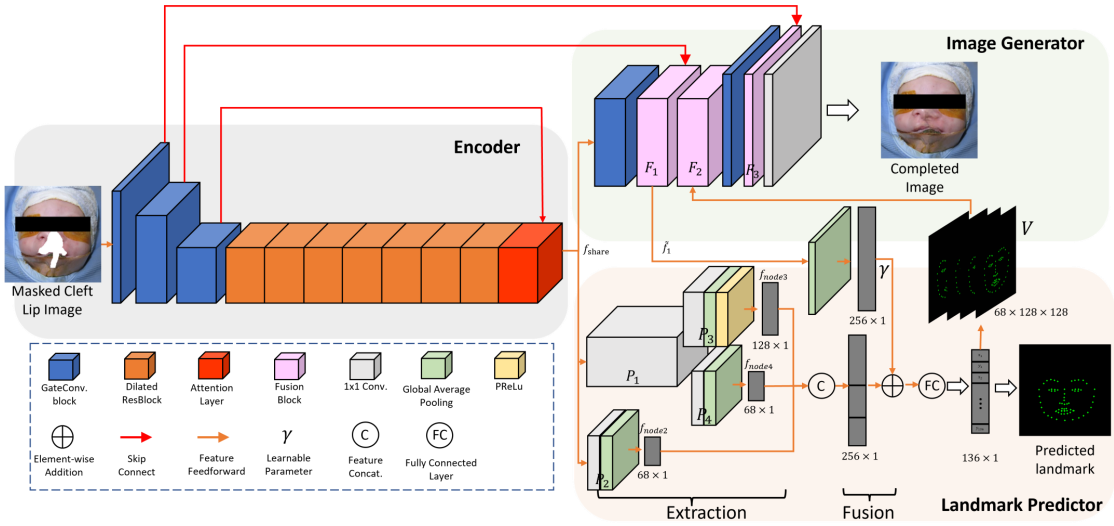

Figure 3.1: Overview of the proposed method.

3.1 Introduction

Our work aims to generate a non-cleft lip from an image of a baby with a cleft lip while protecting patient privacy. While StyleGAN [90] is a powerful option for style transfer, its data demands and vulnerability to model-inversion attacks mean that, if trained on limited clinical images of individuals with cleft lip, it could reveal patient identity [91]. We therefore adopt an image-inpainting paradigm trained exclusively on open, non-clinical datasets, reserving clinical photographs only for held-out validation. This design minimizes memorization risk by construction: because no patient images are used for training, the network has no opportunity to memorize or leak identifiable clinical content.

Federated learning and privacy-preserving training (e.g., differentially private optimization) are viable complementary strategies, but they cannot entirely eliminate residual risks under all configurations. For the present project, the most conservative and reliable safeguard is to avoid training on patient data altogether. We further rely on synthetic occlusions and defect simulations applied to open-source faces to enrich training variability without introducing identifiable clinical information, thereby maintaining utility while preserving privacy.

In the facial inpainting task, the surgical evaluation criteria correspond to the semantic plausibility of the face shape and the image quality. Existing advances in facial inpainting typically ensure the accuracy of generated facial attributes by supplying supplementary

facial geometry information. EdgeConnect [1] first generates a structure map, then combines the corrupted image to perform image inpainting in the second stage. However, the correlation between structural information and texture is frequently redundant and unreliable [39]. Human face can be modelled using landmarks and their geometrical features [92]. Lafin [3] uses landmarks as indicators to more precisely described facial attributes. Nonetheless, both of EdgeConnect and Lafin have a multi-stage limitation: the final image quality is highly dependent on how well the indicator generation in the first stage works.

This work, on the other hand, proposes a single-stage end-to-end multi-task image inpainting framework to generate non-cleft lip from patients with cleft lip.

We propose a single-stage, end-to-end multi-task image inpainting model for synthesizing non-cleft lips from patients with cleft lip. A shared encoder produces features that feed two branches: (i) an image-generation branch and (ii) a landmark-prediction branch. The landmark branch predicts facial keypoints from both the masked input and intermediate inpainted context. These keypoints are then adaptively fused into the generation branch as a geometry prior, yielding more coherent and anatomically plausible facial attributes. Notably, our model is trained without images of individuals with cleft lips, reducing privacy risks for patients. To evaluate our model and assess the feasibility of the proposed method, we curated two clinical test sets, *CleftLip10* and *CleftLip24*, from real patients. We generate facial landmarks using a pretrained FAN detector [93] and employ the resulting keypoints to guide structure-aware inpainting. For each image, segmentation masks are applied to cover the cleft region and any medical equipment; our model then automatically synthesizes a continuous lip and nose without a cleft. *CleftLip10* contains 10 pairs of pre- and post-operative images, enabling paired visual comparison, while *CleftLip24* consists of 24 pre-operative images to evaluate performance in a more challenging unpaired setting. We compare our results with EdgeConnect [1], LaFIn [3], and CTSDG [2], and ask three professional cleft-lip surgeons to rank the outputs. In addition, we report quantitative results on CelebA [35] to highlight the advantages of our design.

The main contributions are summarised as follows:

- We propose an image inpainting approach to produce an non-cleft lip image from

patients with cleft lip.

- We propose a multi-task network in which branches cooperate with each other through parameter sharing between tasks, which can achieve both landmark prediction and image inpainting at the same time.

- The code is available on: https://github.com/ChrisChen1023/ICLG, and validated on CodeOcean: https://codeocean.com/capsule/4388343/tree/v1.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the host organisation, the Research Ethics Committee, the Health Research Authority, and Health and Care Research Wales, under Approval Nos. 19/LO/1690 and under IRAS Project ID: 240451.

3.2 Related Work

3.2.1 Cleft Lip and Palate

A cleft lip/palate is a medical condition where the lip/palate of a patient does not join completely before birth, which usually occurs in the early stages of pregnancy. In the UK, cleft lips are the most common facial birth defect, with one out of every 700 children suffering from cleft lip and palate every year [94]. This explains the importance of cleft lip and palate surgeries, which are usually performed on orofacial cleft patients at an average age of three months [95]. Although the surgical treatment for cleft lip and palate varies, their common objective is to achieve symmetry and enhance a nasolabial look [96].

Achieving symmetry and improving nasolabial appearance is a fundamental goal of cleft lip surgery [96]. There are various surgical approaches to repairing a cleft lip. The most commonly used worldwide at present are a Millard repair or a Fisher repair [97]. Repairing cleft lips is a specialist skill and training in the UK requires an extended period of subspeciality training. Evaluating the outcome of cleft lip and palate surgery is an essential part of being able to improve surgical technique. The current gold standard for assessing outcomes is the Asher-McDade rating scale by using a 5-point standard scale to assess nasolabial profile, nasal symmetry, nasal form, and vermilion border [98].

Recently, with the rapid advancement of AI, technologies based on deep learning have emerged to locate cleft lip surgical annotation and incisions to facilitate surgery [99]. This may help junior surgeons in the early stages of their career and also surgeons who may not be as familiar with repairing cleft lips. Other clinical applications include being able to predict the outcome of a cleft lip repair which would enable surgeons to adjust their surgical procedure to provide the best outcome possible.

3.2.2 Facial Image Inpainting

Facial image inpainting is a crucial area of computer vision focused on reconstructing missing or corrupted regions of a face image while ensuring that the restored areas are semantically consistent and visually natural. This task plays an essential role in medical applications such as cleft lip reconstruction, where it can simulate post-surgical outcomes and assist surgeons in preoperative planning. Compared to generic image inpainting, facial inpainting poses unique challenges due to the complex geometry and texture of human faces. Achieving facial symmetry, maintaining proportionality, and preserving fine-grained details are fundamental requirements for this task.

Recent state-of-the-art methods use auxiliary information, such as structural features or facial landmarks, to guide the inpainting process. For example, EdgeConnect [1] is a notable approach that employs a two-stage pipeline. In the first stage, a structure map is generated to capture the essential edges and contours of the face. In the second stage, the structure map is combined with the corrupted image to produce the final inpainted output. While effective in many cases, this approach often struggles with inconsistencies between structural and textural information. The correlation between the predicted edges and the texture in the missing regions can be redundant or unreliable [39], leading to suboptimal results.

Another method is Lafin [3], which leverages facial landmarks to guide the inpainting process. Landmarks provide precise geometric descriptions of facial attributes [92], allowing the model to focus on restoring critical features with greater accuracy. However, Lafin and similar methods face challenges due to their multi-stage nature. In such pipelines, the final output heavily depends on the quality of the intermediate results, such as the accuracy of the predicted landmarks or structure maps. Inaccurate results in the

early stages propagate to the later stages, compromising the quality of the reconstructed image. Furthermore, the multi-stage design increases computational complexity and processing time, making these methods less suitable for real-time applications. Our work solves this problem by designing an end-to-end multi-task framework, which is able to predict landmark points and regenerate the completed image.

3.2.3 Landmark Detection

Landmark detection is an important task in computer vision and medical imaging, involving the identification of key points or features on objects of interest. In the context of facial analysis, landmark detection plays an essential role in various applications, including facial recognition, expression analysis, and image inpainting. The precise localisation of landmarks enables models to capture geometric relationships, which are pivotal for tasks requiring detailed spatial understanding.

Traditional methods for landmark detection relied heavily on handcrafted features and classical statistical models. Techniques such as Active Shape Models (ASM) [100] were among the earliest approaches. These methods modelled the shape and appearance of facial structures using predefined templates and were effective for controlled scenarios. However, their reliance on handcrafted features and sensitivity to variations in pose, lighting, and occlusions limited their robustness in real-world settings.

Deep learning revolutionise the landmark detection. Dlib library introduced a CNN-based facial landmark detector that demonstrated significant improvements over traditional methods [101]. Similarly, methods such as stacked hourglass networks [102] and deep regression forests [103] further improved accuracy by capturing multi-scale contextual information.

In medical applications, landmark detection has been evident by facilitating precise surgical planning and assessment. For cleft lip and palate surgeries, landmarks are used to evaluate symmetry and guide incisions. Studies like [104] introduced deep learning-based frameworks tailored for medical landmark detection, combining CNNs with attention mechanisms to focus on relevant anatomical features.

Recent approaches have started to explore the application which integrates landmark detection with other related tasks such as segmentation or inpainting. For instance,

Lafin [3] demonstrated the utility of facial landmarks for guiding image inpainting, highlighting the potential for synergistic learning between tasks. Such frameworks aim to improve the robustness of landmark detection while simultaneously enhancing downstream tasks.

3.3 System Description

To protect the privacy of patients' data, we decide to implement the non-cleft facial image prediction system as an image inpainting framework. One key software engineering decision in this research is the framework we use to implement the solution. Existing style transfer-based frameworks [105] allow effective facial image generation with different features. However, they require training data from both the source (i.e., cleft lip images in our case) and target (i.e., non-cleft lip images) domains, which may lead to model leakage where the trained model memorizes the training images. Conditional image translation frameworks using GAN [105] or VAEs [106] may resolve the issue, but those methods mainly focus on the synthesis of new colour patterns instead of geometric structures. Our investigation led us to the image inpainting framework [3] as a suitable solution, as it does not necessitate using cleft facial data for training. Additionally, the binary mask effectively defines the lip area for synthesis with the rest of the face, serving as conditions, making it well-suited to our requirements.

In particular, to implement an image inpainting framework, we utilise the image generation network in [3] as the backbone, which is ameliorated from [1], given its good performance in image inpainting. We also re-implemented the gated convolution algorithm proposed in [16] to dynamically select features for each channel and location, resulting in better inpainting quality.

On top of the backbone, we implement a multi-task system that predicts both the non-cleft facial image and facial landmarks. Facial landmark has shown to be effective in assisting facial image inpainting [1, 3], and is used extensively for cleft lip analysis [99]. Our work differs from existing approaches in that we employ a multi-task model, where two tasks share a part of a common network and facilitate each other.

To prepare the training data, we employ an open facial dataset and a tailor-made

masking algorithm. In particular, we use the CelebA dataset [35], which consists of 202,599 face images of over 10,000 celebrities. To prepare the data for training our inpainting network, we apply an irregular mask algorithm following [15], such that our network can learn to inpaint any masked regions of the face.

To test the system, we work with the NHS to collect a dataset of cleft lip images. Due to the sensitive nature of the data, ethical approvals are obtained from the Research Ethics Committee, the Health Research Authority, and Health and Care Research Wales, under Approval Nos. 19/LO/1690 and under IRAS Project ID: 240451. Given a cleft lip image, we manually draw a mask that covers the mouth area. The masked image is fed into our multi-task network to create the non-cleft facial counterpart, with the facial landmark as a side-product. Since cleft lip images are only used in testing, we mitigate any risk of model leakage [91].

3.4 Methodology

We propose an end-to-end multi-task model, the framework is shown in Fig. 3.1. We first train our model on CelebA, then perform inference step on real patient images with cleft lip to generate non-cleft lip with semantic plausible facial attributes. Our model can simultaneously perform image inpainting and facial landmark prediction. The parameters in two tasks are shared through image-to-landmark and landmark-to-image feature fusion operations. Formally, the whole pipeline could be denoted as:

$$(\hat{I}, \hat{L}) = G(I \odot (1 - M)), \tag{3.1}$$

where G is our multi-task model, I is the real image and M denotes the segmentation mask that occludes the cleft lip and medical equipment. $\hat{I}$ and $\hat{L}$ are completed image and predicted landmarks respectively.

3.4.1 Dataset Collection

To verify that our model operates reliably on real clinical photographs, we collected two datasets—*CleftLip10* and *CleftLip24*—comprising frontal face images from infants (all

under one year of age) who underwent cleft lip repair at the Royal Victoria Infirmary (RVI), Newcastle upon Tyne, during outpatient clinics. *CleftLip10* contains paired images from 10 patients with 20 images, each with a pre-operative and an immediate post-operative photograph (Canon PowerShot G1 X Mark II, 3072×2048), enabling direct paired evaluation. *CleftLip24* consists solely of pre-operative photographs from 24 patients with 24 images (Canon EOS 20D or 5D Mark II with a 105 mm lens, 2574×3861). In total, the collection includes 44 images. For structure guidance, facial landmarks are detected using a pretrained FAN model [93], and the resulting keypoints are used in our inpainting pipeline. All images were included in our experiments.

3.4.2 Encoder and Image Generator

The encoder and the image generator jointly perform the inpainting task. The masked cleft lip image is downsampled three times and fed into the dilated convolutional residual blocks used to improve the receptive field, followed by a short-long attention layer to match feature more efficiently. We use gated convolutions instead of vanilla convolutions only in the image downsampling and downsampling stages. This is because 1) using gated convolution is more efficient for irregular masks [16], 2) its sensitivity to valid and missing pixels seems to be significant only for encoder and decoder [17] and 3) extensive use of gated convolution lead to a significant increase in parameter count. The shared feature is extracted at the end of the encoder:

$$f_{\text{share}} = E(I \odot (1 - M), \tag{3.2}$$

where E is the encoder and f_{share} is the deep feature from the attention layer (See Fig.3.1 (Encoder)).

The image generator is designed to up-sample f_{share} and reconstruct a non-cleft lip and nose. We employ three feature fusion blocks to facilitate parameter sharing, which are denoted by F_1, F_2, F_3 respectively. F_1 and F_3 aim to fuse the uncompleted image features from encoder by skip connections to generate more exquisite results by combining low-level and high-level feature.

$$\widetilde{f}_l = \begin{cases} F_i\left(\text{Concat}\left(f_{ei}, f_{di}\right)\right), \text{ if } (i = 1, 3) \\ F_i\left(\text{Concat}\left(\widetilde{f}_1, V\right)\right), \text{ if } (i = 2) \end{cases}, \tag{3.3}$$

where $\widetilde{f}_i$ is the result from fusion block F_i $(i = 1, 2, 3)$. f_{ei} and f_{di} is the feature map from corresponding encoder and decoder layer. After F_3 followed by a vanilla convolution layer, completed image is generated. F_2 is designed to fuse the Landmark map V from the landmark predictor, which will be detailed in the next subsection.

3.4.3 Landmark Predictor

The landmark predictor involves extraction, fusion block and a fully-connected layer, aims to predict facial landmarks and inform the generator for assisting image inpainting.

The extraction step is designed to collect the landmark information from the encoded image feature. Specifically, f_{share} is fed to a 1×1 convolutional layer P_1 to increase dimensionality, then we conduct dimensionality reduction followed by global average pooling to extract the feature into two vectors with different lengths (P_4 and P_3). Particularly, there is a $PReLu$ layer at the end of P_3 for non-linear projection. Simultaneously, P_2 also returns a vector after dimensionality reduction and global pooling directly acting on f_{share}, then we concatenate them:

$$f_{lmk} = \text{Concat}\left(f_{\text{node2}}, f_{\text{node3}}, f_{\text{node4}}\right), \tag{3.4}$$

where f_{nodei} is the corresponding vector from P_i. In existing multi-stage networks [1, 3], generated indicators are assumed as perfect and are used in final inpainting stage directly. A faulty indicator may mislead image inpainting. To involve both corrupted and regenerated information in landmark predictor, and strengthen the parameter sharing between two tasks, we adaptively borrow f_1 from inpainting task, followed by a global average pooling, we merge it with the concatenated landmark feature vector:

$$f'_{lmk} = \gamma * \widetilde{f}_1 \oplus f_{lmk}, \tag{3.5}$$

where γ is a trainable weight with zero initialization and $\oplus$ is element-wise addition.

Finally, we apply a fully-connected layer to predict facial landmark points.

To strengthen the parameter interaction between the two tasks and improve the completed image quality, we further map the landmark points into a binary feature map V, which is integrated with texture information in F_2. Formally, let v_{pq} be the value in V at position (p, q):

$$v_{pq} = \begin{cases} 1, \text{ if } (p = [\alpha x_i], q = [\alpha y_i]) \\ 0, \text{ otherwise} \end{cases}, \tag{3.6}$$

where α is a scale factor corresponding the size of the feature map in F_2, $[.]$ means integer operation. We create a $68 \times 128 \times 128$ tensor with landmark annotations, and transfer it to F_2 to provide facial geometry indicators.

3.4.4 Loss Function

We follow Yang et al. (2019) to design our loss function. Given the predicted landmark $\hat{L}$ and corresponding landmark ground truth L_{gt}, the landmark loss is:

$$\mathcal{L}_{lmk} = \left\| \hat{L} - L_{gt} \right\|_2^2. \tag{3.7}$$

We also consider L_1 loss, adversarial loss, style loss, perceptual loss, total variation loss. Given a masked image I and the ground truth image I_{gt}:

$$\mathcal{L}_{rec} = \mathbb{E}\left[\left\| \boldsymbol{I}_{out} - \boldsymbol{I}_{gt} \right\|_1 \right], \tag{3.8}$$

$$\mathcal{L}_{perc} = \mathbb{E}\left[\sum_i \left\| \phi_i\left(\boldsymbol{I}_{out}\right) - \phi_i\left(\boldsymbol{I}_{gt}\right) \right\|_1 \right], \tag{3.9}$$

$$\mathcal{L}_{style} = \mathbb{E}\left[\sum_i \left\| \left(\psi_i\left(\boldsymbol{I}_{out}\right) - \psi_i\left(\boldsymbol{I}_{gt}\right)\right) \right\|_1 \right], \tag{3.10}$$

$$\mathcal{L}_{adv} = \min_G \max_D \mathbb{E}_{\boldsymbol{I}_{gt}} \left[\log D\left(\boldsymbol{I}_{gt}\right) \right] + \mathbb{E}_{\boldsymbol{I}_{out}} \log\left[1 - D\left(\boldsymbol{I}_{out}\right) \right], \tag{3.11}$$

where $\phi i(\cdot)$ indicates the activation map from the i-th pooling layer of VGG-16. $\psi_i(\cdot) = \phi_i(\cdot)^T \phi_i(\cdot)$ denotes the Gram matrix. The loss combination is:

$$\begin{aligned} \mathcal{L}_{\text{total}}(I, I_{gt}, L_{gt}) = \mathcal{L}_{\text{pixel}} &+ \lambda_{\text{perc}} \mathcal{L}_{\text{perc}} + \lambda_{sty} \mathcal{L}_{\text{style}} \\ &+ \lambda_{tv} \mathcal{L}_{tv} + \lambda_{adv} \mathcal{L}_{adv} \quad , \\ &+ \lambda_{lmk} \mathcal{L}_{lmk}, \end{aligned} \tag{3.12}$$

where $\lambda_{\text{perc}} = \lambda_{sty} = \lambda_{tv} = 0.1$, $\lambda_{adv} = 0.01$, $\lambda_{lmk} = 0.00046$. The weights of each loss are tuned by using Optuna [107].

Algorithm 1: GAN-based Training for Proposed Multi-task Model MT_θ (Revised)

Input: Generated image x, image ground truth X, predicted landmark k, landmark ground truth K, irregular mask M, maximum iterations T, batch size 4

Output: Trained Multi-task Model MT_θ

Initialise dataloader;

Initialise the multi-task network MT_θ;

for $t \leftarrow 0$ **to** T **do**

  Sample 4 images and corresponding landmarks from dataloader;

  Sample 4 irregular masks from dataloader;

  Compute losses: L_{pixel}, L_{landmark}, L_{tv}, L_{style}, $L_{\text{perceptual}}$, L_g, L_d using $MT_\theta(X, M, k)$;

  Aggregate generator loss:

$$L_G \leftarrow L_{\text{pixel}} + L_{\text{landmark}} + L_{\text{tv}} + L_{\text{style}} + L_{\text{perceptual}} + L_g$$

  ;

  Define discriminator loss:

$$L_D \leftarrow L_d$$

  ;

  Freeze generator parameters θ_G and update discriminator using adversarial loss L_D;

  Freeze discriminator parameters θ_D and update generator using adversarial loss L_G;

Save the trained Multi-task Model MT_θ;

Table 3.1: Valid Possibility on Cleft Lip dataset.

Method	EC [1]	Lafin [3]	CSTDG [2]	Ours
CleftLip10	0.233	0.233	0.233	**0.5**
CleftLip24	0.319	0.222	0.264	**0.333**

Table 3.2: Average Ranking on the Cleft Lip dataset.

Method	EC [1]	Lafin [3]	CSTDG [2]	Ours
CleftLip10	1.857	1.714	2.429	**1.267**
CleftLip24	1.696	1.813	1.947	**1.208**

3.5 Experiments

3.5.1 Training Details

We train our model with a GAN-Based training flow (as shown in Algorithm 1) on CelebA [35], which is a popular human face dataset containing over 160 thousands training face images (Sec.2.7.1). For CelebA, we remove a few images which can not be obtained landmark ground truth. We adopt [93] on CelebA to get the landmark ground truth. During training, the images are resized to 256×256 and we use irregular masks as in [15]. Although adult faces and infant images differ in overall distribution, the structural priors learned from adult lips and noses are sufficiently stable to transfer. We use Adam optimiser and follow [1] to set $\beta_1 = 0$ and $\beta_2 = 0.9$. The learning rate = 2.92×10^{-4} and 2.92×10^{-5} for discriminator, with a learning rate decay ratio of 0.78. Batch size = 4.

3.5.2 Experimental Validation

Cleft Lips Repair

We use the CleftLip10 and CleftLip24 as test sets to compare our model with the current state-of-the-art facial inpainting methods [1–3]. The visualisation results are shown in Fig. 3.2. We crop and resize them to 256×256, then design a mask to cover the cleft lip, as

Figure 3.2: Visual comparison of different facial inpainting methods on real Cleft Lip dataset: (a) input masked image, (b) EdgeConnect [1], (c) CTSDG [2], (d) Lafin [3], (e) Ours, and (f) Before Surgery

Table 3.3: Quantitative Comparison on CelebA.

Mask Ratio	Model	PSNR	SSIM	FID
0-20%	EC [1]	36.1340	0.9880	0.4706
	Lafin [3]	35.9544	0.9870	0.5845
	CTSDG [2]	37.9275	0.9908	**0.3420**
	Ours	**38.1083**	**0.9911**	0.3421
20-40%	EC [1]	28.3684	0.9486	3.1275
	Lafin [3]	28.2797	0.9476	3.3880
	CTSDG [2]	29.3860	0.9570	2.8436
	Ours	**29.6678**	**0.9595**	**2.8327**
40-60%	EC [1]	23.4513	0.8561	6.1253
	Lafin [3]	23.5109	0.8614	6.5367
	CTSDG [2]	**24.3130**	**0.8762**	8.7051
	Ours	24.2076	0.8726	**4.3419**

well as the medical equipment used during surgery, according to the type (unilateral and bilateral) and the severity of cleft lip for each patient. To better evaluate the feasibility of the proposed method, we invited NHS specialist cleft lip surgeons to assess the results based on the quality, consistency and validity. For each patient, results from four models are presented together. To avoid bias, the results are mixed and unlabelled. Images are deemed invalid if it is excessively blurry or illogical, e.g., flying lip or three nostril (see Fig. 3.2(d)). The valid probability represents the the success rate of models in repairing cleft lips images (see Table 3.1), and the average ranking represents the performance of each models in the valid repaired results (see Table 3.2).

From our observation, each of the four models is capable for repairing small cleft

lip areas. However, our method performs best for relatively complex situation, such as Fig. 3.2(f) with severe cleft lips and large medical equipment. The result from EC [1] is too blurry and CTSDG [2] leads obvious artifacts in regenerated region. Lafin [3] seems to be suffering from model collapse and was seriously misled by the input indicator, generating a full nose at the right nostril. From the surgeons assessment, our model generates more natural and semantically plausible images with a higher valid possibility. Additionally, our model is able to generate textures similar to post-surgical scars while we leave certain intimation to the model (see Fig. 3.2(e)).

Facial Inpainting

We compare our model with current state-of-the-art facial inpainting models on CelebA. The evaluation metrics involve peak signal-to-noise ratio (PSNR), structural similarity index (SSIM) [86] and Frechet Inception Distance (FID) [87], which is shown in Table 3.3. Higher PSNR, SSIM, and lower FID, indicate better generated image quality. We further visualize the results in Fig 3.3 for qualitative comparison. We observe that our model overall suppresses state-of-the-art inpainting models in terms of small and medium masked ratio. The latest CTSDG outperforms ours by a small margin in large missing regions case in terms of PSNR and SSIM, but it is much lower than ours in FID.

Table 3.4: Ablation Study on CelebA.

	Mask	Irregular Mask			Regular
	Mask Ratio	0-20%	20-40%	40-60%	Mask
PSNR	Baseline	37.1742	29.1189	23.7541	25.9412
	Base+Lmk	37.3932	29.2115	23.7948	26.074
	Ours	**38.1083**	**29.6678**	**24.2076**	**26.685**
SSIM	Baseline	0.9895	0.9545	0.8605	0.9113
	Base+Lmk	0.9897	0.9554	0.8626	0.9144
	Ours	**0.9911**	**0.9595**	**0.8726**	**0.9231**
FID	Baseline	0.5330	3.3655	5.7635	3.6037
	Base+Lmk	0.4614	2.9762	5.1124	3.410
	Ours	**0.3421**	**2.8327**	**4.3419**	**3.274**

To validate the effectiveness of our multi-task architecture, we remove the parameter sharing between two tasks and take encoder followed by image generator as the baseline.

Then, we implement the landmark predictor (Base+Lmk) and gated convolution (Ours) progressively. As shown in Table 3.4, the integration of both the multi-task model and gated convolutions improve the performance on both irregular and regular masks.

3.6 How to Use the System

To retrieve the training dataset for this image inpainting application, users are required to download the CelebA Dataset [35] and the irregular mask dataset [15] from the respective official websites. The CelebA dataset should then be divided into a standard training set and a validation set, according to the official instruction. Additionally, the corresponding landmark points should be generated with FAN [93]. Furthermore, the irregular mask dataset should be divided into three groups according to the mask ratios (0-20%, 20-40%, 40-60%). 3,300 masks are randomly selected from each group, resulting in a total of 9,900 mask images for training. Another 200 masks are selected from each group, resulting in a total of 600 mask images for verification. For the inference step, all cleft facial images and their corresponding masks serve as the image test set and the mask test set, respectively. The user should then run the provided "./scripts/filst.py" script to generate training, test and validation set file lists, and update the information in the "config.yml" file accordingly to set the model configuration. Once the python environment has been set up using the released "requirements.txt" file, the user may proceed to run the "train.py" script for training and the "test.py" script for testing. For the inference process, although we recommend using our system with GPUs for better speed, the system is fully runnable with only CPUs. Due to the sensitivity of patient privacy, we are not allowed to upload the cleft lip data for an online demonstration. Therefore, we show the reproducibility of our system with the images from CelebA and the irregular masks.

3.7 Impact Overview

While our method primarily focuses on cleft lips, the uses of the implemented source code can be extended to other applications. The key idea of this software is to mask out a particular region of a face, and to employ inpainting techniques for predicting the masked

area. The versatility of our system allows for the implementation of extended facial applications, such as makeup and plastic surgery prediction. To utilize these capabilities, a customized dataset is required for training, such as the Facial Beauty Database [108] or a plastic surgery facial dataset [109]. The users then need to retrain our model according to section 2.3. The resulting model can then be tested using a corresponding mask that covers specific facial components, such as nose or eyebrows, to generate the image of the subject after makeup or plastic surgery. Therefore, it can also be used for supporting plastic surgeries and makeup prediction [110] on specific facial components. This would facilitate the understanding and discussion of those operations and applications among stakeholders.

We put a particular effort in selecting a software framework that is robust against model leakage and attack [91, 111]. In particular, we propose the idea of excluding patient data in training deep learning models if possible, mitigating any privacy concerns and risk of data loss. The high-level concept of training with open data and testing with sensitive data can be employed in other machine learning applications to protect data privacy, particularly those in the healthcare domain or involving people of vulnerable groups.

In theory, our system is also capable of synthesising cleft facial images from non-cleft lip ones. In practice, due to the wide variety of cleft lip conditions, training such a system would require a large dataset of cleft images, which is currently not available. Should there be enough data (and we only need the lip area to protect patients' privacy), this system can be used to generate synthetic cleft lip facial images, which enable the training of machine learning algorithms. As the data is artificially created, there is no privacy or model leakage concern, and an unlimited amount of samples can be created. This aligns with the recent trend of using computer graphics techniques to mock up real-world data [112], facilitating the training of machine learning systems for patient-related applications [113]. Since the beginning of this research, there is raising awareness from both UK universities and hospitals in collecting cleft lip data for research purposes. We believe our vision will be made possible in the future.

3.8 Conclusion and Discussions

This work implements a multi-task image inpainting model to predict non-cleft lip facial images from cleft lip ones. We make an important software engineering decision to implement the system under an inpainting framework, which does not require patient data for training and mitigates model leakage risks. We design and develop a multi-task neural network that co-predicts a facial image and the corresponding facial landmarks, and we find that the two tasks support each other. We collected two real-world patient datasets to demonstrate the feasibility of proposed approach. Three expert cleft lip surgeons assessed that our design outperforms state-of-the-art methods in both valid possibility and image quality, while the performance of our model on CelebA also suppresses the state-of-the-art facial inpainting counterparts. Apart from detailing the design and implementation details of our software, we also discuss its impact within and beyond cleft lip applications. The source code is now publicly released on CodeOcean and GitHub.

This work lays the foundation for advancing general-purpose image inpainting techniques. While our current work demonstrates the clinical and practical value of a multi-task inpainting framework for cleft lip repair, the challenges encountered in preserving sparse yet critical visual cues highlight broader limitations of existing inpainting architectures. These insights motivate the subsequent research, a transformer-based inpainting model $HINT$, that employs a mask-aware downsampling strategy and efficient attention mechanisms to better preserve visible structures and exploit long-range dependencies.

Figure 3.3: Comparisons with visualisations (256×256) on CelebA dataset [4] , showing that our results are more semantically plausible with more clear facial attributes.

CHAPTER 4

High Quality Image Inpainting with Enhanced Transformer

Portions of this chapter have previously been published in the following peer-reviewed publication [58]:

- **Shuang Chen**, Amir Atapour-Abarghouei, Hubert P. H. Shum, "HINT: High-quality INpainting Transformer with Mask-Aware Encoding and Enhanced Attention", In *IEEE Transactions on Multimedia* (TMM). IEEE, 2024.

Existing image inpainting methods leverage convolution-based downsampling approaches to reduce spatial dimensions. This may result in information loss from corrupted images where the available information is inherently sparse, especially for the scenario of large missing regions. Recent advances in self-attention mechanisms within transformers have led to significant improvements in many computer vision tasks including inpainting. However, limited by the computational costs, existing methods cannot fully exploit the efficacy of long-range modelling capabilities of such models. In this paper, we propose an end-to-end High-quality INpainting Transformer, abbreviated as HINT, which consists of a novel mask-aware pixel-shuffle downsampling module (MPD) to preserve the visible information extracted from the corrupted image while maintaining the integrity of the information available for high-level inferences made within the model. Moreover, we propose a Spatially-activated Channel Attention Layer (SCAL), an efficient self-attention

mechanism interpreting spatial awareness to model the corrupted image at multiple scales. To further enhance the effectiveness of SCAL, motivated by recent advanced in speech recognition, we introduce a sandwich structure that places feed-forward networks before and after the SCAL module. We demonstrate the superior performance of HINT compared to contemporary state-of-the-art models on four datasets, CelebA [35], CelebA-HQ [4], Places2 [5], and Dunhuang [36].

4.1 Introduction

A significant challenge hindering image inpainting is effectively modelling the valid information within visible regions, which is crucial for reconstructing semantically coherent and texture-consistent details in the missing regions. This is particularly noticeable in large masked regions, where the valid information is limited. Existing methods that utilise convolutional layers for downsampling come with the inherent drawback of information loss [24], attributed to the reduction of feature size from filters and downsampling. Given its capability to preserve input information, pixel-shuffle down-sample is widely used in image denoising [25], image deraining [26] and image super-resolution [27]. It periodically rearranges the elements of the input into an output scaled by the sample stride. However, its effectiveness depends on the assumption that the sample stride is small enough to avoid disrupting the noise distribution [28]. This holds only for a relatively independent distribution of raindrops and noise, and is not suitable for image inpainting with irregular and variable-size masks. Simply using conventional Pixel-shuffle Downsampling (PD) [25–27] for corrupted image would lead to the problem of pixel drifting, which is shown in Fig. 4.2 (upper branch). The pixel drifting happens in $\hat{X}$. After the feature X' is downsampled, the position of the masked regions (white elements) becomes inconsistent across channels, causing the visible area to be misaligned in the channel, disrupting subsequent feature extraction processes within the model, thus affecting the accurate modelling of the valid information from the visible regions of the input image.

Another challenge in applying spatial self-attention in CNN-based models is its significant computational expense. Considering this, spatial self-attention is typically only employed on low-resolution representations [73, 114]. While transformer-based meth-

ods [8,72] employ multiple spatial self-attention blocks to model long-range dependencies. However, the quadratic computational complexity limits their wider applicability. To address this, the prevalent compromise involves down-sampling [8] or reducing the resolution [72] of the input image prior to being passed through the transformer. However, this strategy leads to information loss from the input images through the model, which is detrimental to image inpainting where visible information is already limited. This loss subsequently results in the degradation of fine-grained features. As long-range dependencies are modelled over these degraded features, the reconstructed output suffer from blurring artefacts and vague structures. [72,73] introduce extra refinement networks to improve image quality after getting coarse completed images, rather than recovering high-quality results directly. The method in [10] replaces spatial self-attention with channel self-attention to reduce computational complexity. Although channel self-attention gains linear computational complexity, it completely loses spatial awareness. This makes it possible to highlight "what" the salient features are but cannot discern "where" the spatially important regions are, which is essential as visible regions often exhibit complex and irregular shapes, especially with large irregular masks. Some existing works [115–117] attempt to address the spatial awareness loss by incorporating spatial self-attention back to the channel self-attention, but at a cost of significant increases in computation.

To address these common challenges currently restricting progress in the existing literature, we present a novel High-quality INpainting Transformer (HINT) for image inpainting, which enables efficient multiscale modelling of the global context while minimising the loss of valid information. Specifically, we propose a tailor-made pixel-shuffle down-sampling (MPD) module for image inpainting to reduce information loss and maintain the consistency of data. To enhance the representation learning capabilities of our model, we develop a Spatially-activated Channel Attention Layer (SCAL) to blend information in both the channel and spatial dimensions. Unlike these existing methods [115–117], the innovation of SCAL lies in its minimalistic and efficient design, only utilising convolutional layers to retrain spatial awareness, thereby mitigating the significant computational cost, which is a major issue in the field. This enhanced self-attention module plays the predominant role in HINT and build HINT as a transformer-based model. To further improve the effectiveness of SCAL with limited parameters, we

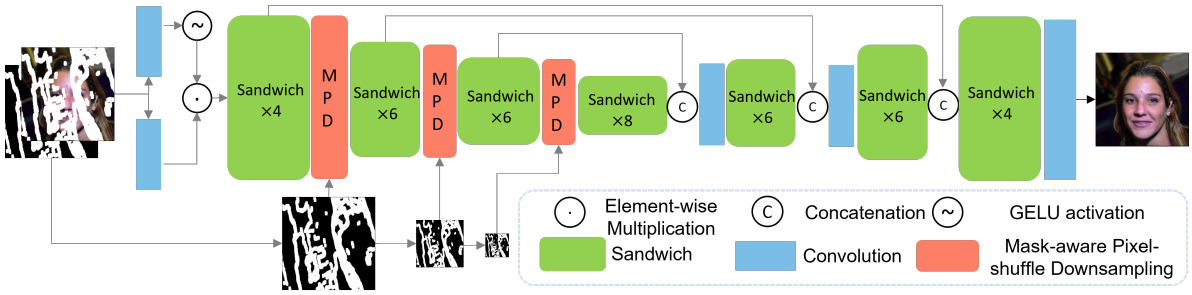

Figure 4.1: The overview of the proposed framework, which is built with a gated embedding block, with multiple stacked "sandwiches" in different levels. The "sandwich" is described in Sec. 4.2.3, the MPD is described in Sec. 4.2.2

propose a module known as the "Sandwich", sandwiching the proposed SCAL between two feed-forward networks (FFNs) for each transformer block. This structure results in better performance compared to alternative designs with the same number of network parameters.

Comparative experiments show that HINT outperforms state-of-the-art image inpainting approaches (Fig. 4.5 and Fig. 4.4) across four datasets, i.e., CelebA [35], CelebA-HQ [4], Places2 [5] and Dunhuang challenge [36]. We also perform ablation experiments to demonstrate the contribution of proposed components in HINT.

Our source code is openly released at `https://github.com/ChrisChen1023/HINT`.

Our major contributions are as follows:

- We propose HINT, an end-to-end transformer-based architecture for image inpainting that takes advantage of multi-scale feature- and spatial-level representations as well as pixel-level visual information.
- We propose a plug-and-play mask-aware pixel-shuffle down-sampling (MPD) module to preserve useful information while keeping irregular masks consistent during downsampling.
- We propose a Spatially-activated Channel Attention Layer (SCAL) using self-attention and convolutional attention to sequentially refine features at the channel and spatial dimensions. We further design an improved sandwich-shaped transformer block to boost the efficacy of the proposed SCAL.

4.2 HINT: High-quality INpainting Transformer

Formally, the problem is formulated as follows: the input image, I_{input}, is obtained by concatenating masked image, $I_M = I \odot M$, and the mask, M. The input image, I_{input}, is then processed by our proposed HINT model and a semantically accurate output image, I_C, will be generated. The whole formulation is denoted as: $I_C = HINT(I_{input})$.

We present our transformer-based HINT approach to image inpainting, which takes advantage of our novel Mask-aware Pixel-shuffle Down-sampling (MPD) to solve the information loss issue during downsampling and further enhance the use of valid information from known areas. Within the architecture, we propose a Spatially-activated Channel Attention Layer (SCAL), which aims to handle spatial awareness while maintaining efficiency within the transformer block. The SCAL is encapsulated between two feed-forward networks, forming a sandwich-shaped transformer block, henceforth referred to as "*Sandwich*". This design enables the effective extraction of long-range dependencies while preserving the smooth and coherent flow of valid information through the model.

4.2.1 The Overall Pipeline

Overall, as seen in Fig. 4.1, HINT consists of an end-to-end network with a gated embedding layer to selectively extract features, followed by a transformer body for modelling long-range correlations, and a projection layer to generate the output. Specifically, we insert a gating mechanism [16] into the embedding layer serving as a feature extractor, achieved by using two parallel paths of vanilla convolutions with one path activated by a GELU non-linearity [118] to dynamically embed the finer-grained features, leading to stronger representation learning and better optimisation [119]. The transformer body is an encoder-decoder architecture comprising multiple transformer blocks. The encoder consists of the first three blocks, each followed by an MPD layer to mitigate incoherence in invalid locations, while the final three blocks with conventional pixel shuffle upsampling form the decoder. Mirrored blocks are connected via skip connections to preserve shared features learned within the encoder. At the end, a convolutional layer is used to project the decoded features to the final output.

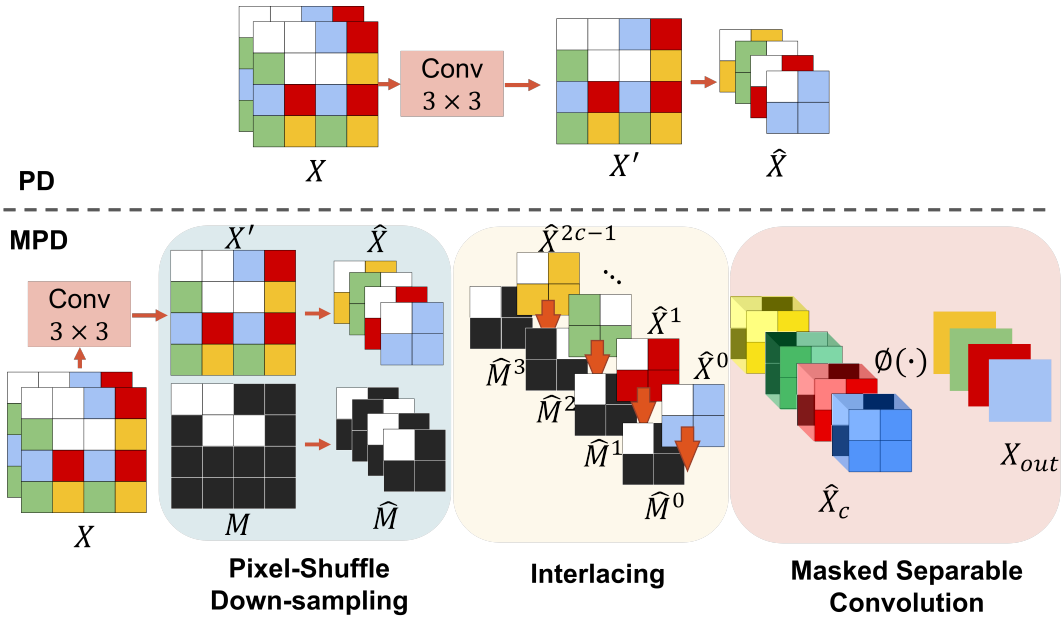

Figure 4.2: The comparison of Pixel-shuffle Down-sampling (PD, upper) and the proposed Mask-aware Pixel-shuffle Down-sampling (MPD, lower). Ours proposed MPD, with one 3×3 convolution, a conventional PD, interlacing (concatenation of feature and mask slices), and a masked-separable convolution. Invalid pixel drifting happens in $\hat{X}$. After the feature X' is downsampled, the masked position becomes inconsistent across channels.

4.2.2 Mask-aware Pixel-shuffle Down-sampling

Conventional Pixel-shuffling Down-sampling (PD) is the inverse operation of Pixel-shuffle [120]. It periodically rearranges the input $T_{in} \in \mathbb{R}^{H \times W \times C}$ into $T_{out} \in \mathbb{R}^{\frac{H}{s} \times \frac{W}{s} \times s^2 C}$ for downsampling with s being the scale factor to denote the sample stride. PD can effectively preserve the input information, which is desirable for inpainting, particularly for reconstructing high-quality images. However, as PD uses non-overlapping sampling with stride s to generate mosaics from the image [120], the consistency of missing pixel locations can be disrupted during the down-sampling, as shown in Fig. 4.2, making it unsuitable for image inpainting.

We propose a Mask-aware Pixel-shuffle Down-sampling (MPD) module, which is a novel down-sampling approach specifically tailored for image inpainting. It resolves the issue of positional drift of masked pixels that occurs during the process of conventional PD. Furthermore, in contrast to convolution-downsampling, MPD preserves all valid information, thereby minimising information loss. Apart from inpainting, this module can be plugged into any other problem that involves masking, such as any that might use image segmentation labels masks as their input.

Given the features $X \in \mathbb{R}^{H \times W \times C}$ and mask $M \in \mathbb{R}^{H \times W \times 1}$, we first project X into X' with half the channels but the same size [120], utilising a 3×3 convolution operator $h(\cdot)$, and perform PD on both X' and M:

$$\hat{M} = PD(M), \hat{X} = PD(h(X)). \tag{4.1}$$

As shown in Fig. 4.2, the positions of the missing pixels in $\hat{X}$ drift and are discontinuous across channels while each channel of $\hat{M}$ sequentially indicates the positions of valid and invalid pixels in $\hat{X}$. To enforce $\hat{M}$ to act on the corresponding channel accurately, we intersperse and concatenate the sliced $\hat{X}$ and $\hat{M}$ across the channel, obtaining $\hat{X}_c \in \mathbb{R}^{\frac{H}{2} \times \frac{W}{2} \times 2C}$.

$$\begin{aligned} \hat{M}^0, \hat{M}^1, \hat{M}^2, \hat{M}^3 &= Slice(\hat{M}), \\ \hat{X}^0, \hat{X}^1, \hat{X}^2, \ldots, \hat{X}^{2C-1} &= Slice(\hat{X}), \end{aligned} \tag{4.2}$$

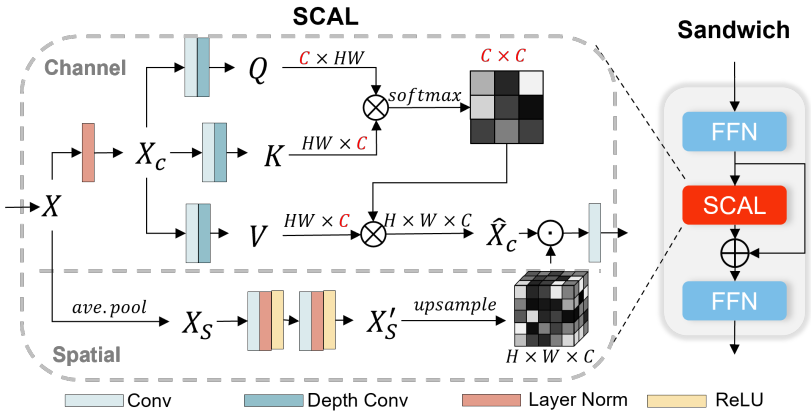

Figure 4.3: "Sandwich" (right) and "Spatially-activated Channel Attention Layer" (left). "$\oplus$","$\otimes$", and "$\odot$" denote the element-wise sum, matrix multiplication, and element-wise multiplication, respectively.

$$\begin{aligned}\hat{X}_c =&(\hat{X}^0||\hat{M}^0)||\dots||(\hat{X}^i||\hat{M}^{(i+4)\%4})||\dots \\ &||(\hat{X}^{2C-1}||\hat{M}^3),\end{aligned} \tag{4.3}$$

where $Slice(\cdot)$ is a channel-wise slice, $||$ is channel-wise concatenation, and $\%$ denotes the modulo operator. Thus, each feature has a paired mask as an indicator. In the end, we exploit a separable convolutional layer [121], denoted as $\phi(\cdot)$, to encode pairs of features and masks, aiming to learn the correct local priors from the features indicated by the shuffled mask, and forcing the encoder to accurately model the valid information within the visible regions. The output is formulated as:

$$X_{out} = \phi(\hat{X}_c). \tag{4.4}$$

4.2.3 The Transformer Body

Each of the seven transformer blocks stacks multiple *sandwiches* encapsulating the proposed SCAL for local-global representation learning, working with MPD to down-sample the features and control data flow consistency (Fig. 4.1).

Spatially-activated Channel Attention Layer

We propose a Spatially-activated Channel Attention Layer (SCAL) to strengthen the model to capture inter-channel dependencies while preserving spatial awareness. Channel self-attention [122] is computationally viable for high-resolution features due to its linear time and memory complexity growth with channel depth. However, it fails to account for "where" the important information is across the entire spatial position, thus ignoring the relationship between feature patches. This is very important for image inpainting as the global context in the valid regions within each image can be distinct and irregularly shaped, as defined by the irregular mask M.

To alleviate this issue, we improve the concept of transposed attention [10] by introducing a convolution-attention branch to capture the attention matrix of spatial locations. This enables HINT to effectively model long-range dependencies in the channel dimension, while attending to spatial locations where features should be emphasised. Unlike alternative approaches [8, 69, 70, 72, 73], SCAL does not increase the computational cost quadratically with input resolution, making it feasible for multi-scale context modelling.

As shown in Fig. 4.3, SCAL contains two branches. Given input feature X, the channel self-attention branch is:

$$\begin{aligned} X_c &= LN(X), \\ \hat{X}_c &= \left(W_{d3}^V W_1^V X_c\right) \cdot \mathrm{Attc}(X_c), \\ \mathrm{Attc}(X_c) &= \varphi\left(\frac{W_{d3}^Q W_1^Q X_c \cdot \left(W_{d3}^K W_1^K X_c\right)^T}{\gamma}\right), \end{aligned} \tag{4.5}$$

where *LN* denotes layer normalisation, γ is a learnable parameter to scale the dot product of key and query, W_1 is the linear projection and W_{d3} is the 3×3 depth-wise convolution, $\mathrm{Attc}(\cdot)$ represents the function to calculate the channel attention map, and φ is a softmax layer. In the spatial branch, we first downsample the input features X but not fully squeeze, via average pooling to preserve global spatial information. Subsequently, two 3×3 convolutions serve as attention descriptors followed by an upsampling process, generating a soft global attention matrix, $\alpha = \mathrm{Atts}(X)$, which is used to reweight the

output obtained through channel attention:

$$\mathrm{Atts}(X) = \mathrm{Up}(f(g(\mathrm{AP}(X)))), \tag{4.6}$$

where AP is an average pooling layer, Up is upsampling. $f(\cdot)$ and $g(\cdot)$ are two similar convolution blocks, one of which contains a 3×3 convolutional layer, a normalisation layer, and a ReLU layer [123]. $\mathrm{Atts}(\cdot)$ represents the function to calculate the spatial attention map. As depicted in Fig. 4.3, the attention matrix α modulates the output of the channel branch $\hat{X}_c$ through point-wise multiplication. Subsequently, the mapping function $\theta(\cdot)$ is a projection layer performed via of a 1×1 convolution. The complete representation of the SCAL is:

$$SCAL(X) = \theta(\hat{X}_c \odot \mathrm{Atts}(X)). \tag{4.7}$$

Algorithm 2: Sandwich Block (FFN–SCAL–FFN)

Require: Feature map $x \in \mathbb{R}^{B \times C \times H \times W}$

Ensure: $y \in \mathbb{R}^{B \times C \times H \times W}$

1: $y_1 \leftarrow x + \mathrm{FFN}(x)$

2: $y_2 \leftarrow y_1 + \mathrm{SCAL}(y_1)$

3: $y \leftarrow y_2 + \mathrm{FFN}(y_2)$

4: **return** y

Algorithm 3: SCAL: Spatial–Channel Adaptive Layer

Require: $z \in \mathbb{R}^{B\times C\times H\times W}$

Ensure: $\text{out} \in \mathbb{R}^{B\times C\times H\times W}$

$c \leftarrow \text{GAP}(z)$

$\hat{c} \leftarrow \sigma(\text{MLP}(c))$

$z_c \leftarrow z \odot \text{Broadcast}(\hat{c})$

$s_{\text{avg}} \leftarrow \text{Mean}(z;\ \text{channel})$

$s_{\max} \leftarrow \text{Max}(z;\ \text{channel})$

$s_{\text{in}} \leftarrow \text{Concat}(s_{\text{avg}}, s_{\max})$

$\hat{s} \leftarrow \sigma(\text{Conv}_{3\times 3}(s_{\text{in}}))$

$z_s \leftarrow z \odot \hat{s}$

$g_{\text{in}} \leftarrow \text{Concat}(z_c, z_s)$

$g \leftarrow \sigma(\text{Conv}_{1\times 1}(g_{\text{in}}))$

$\text{out} \leftarrow g \odot z_c + (1-g) \odot z_s$

return out

Sandwich-shaped Transformer Block

Image inpainting presents a significant challenge: the network must effectively learn from limited context to reconstruct complete images. This task is particularly daunting when faced with irregularly shaped masks, which complicate feature extraction, especially in areas with extensive missing information. This process of masking in image inpainting bears a notable resemblance to the masking of audio spectrograms in speech recognition for data augmentation purposes, as seen in techniques like SpecAugment [124, 125]. The Conformer [126], with its innovative "FFN-Attention-Conv-FFN" architecture, demonstrates remarkable efficiency in speech recognition by using augmented, masked spectrograms as inputs. We hypothesise that such structures are equally effective for image inpainting, since their inputs are also incomplete and insufficient, highlighting a common challenge in both fields that may benefit from similar architectural solutions.

Therefore, to boost the effectiveness of our attention layer, we propose a sandwich-shaped transformer block with an FFN-Attention-FFN structure. This first FFN serves

as a filter, extracting more essential features for the following attention layer to capture long-distance dependencies (see Section 4.3.4 for validations). Unlike [126], we remove the convolutional layer in the middle, and enhance the two FFNs with depth-wise convolutions with a gate mechanism [10]. This is because FFN integrating depth-wise convolution captures local information from every channel, which helps the model learn a more comprehensive and informative feature representation with fewer parameters [121]. Also, the gating strategy selectively filters and modulates the information flow according to the importance of each feature to the final high-quality output, thereby reducing irrelevant information and highlighting the most salient input features for representation learning. Given an input $X \in \mathbb{R}^{H\times W\times C}$, our sandwich is formulated as:

$$X_{\text{out}} = FFN(SCAL(FFN(X))). \tag{4.8}$$

4.2.4 Loss Functions

To obtain high-quality inpainting results, we follow the established literature [38, 127] to develop multiple loss components, including an $\mathcal{L}_1$ loss to enforce a contextually sound reconstruction, style loss $\mathcal{L}_{\text{style}}$ to measure the difference in style, perceptual loss $\mathcal{L}_{\text{perc}}$ to compare the high-level perceptual features extracted from a pre-trained network, and an adversarial loss $\mathcal{L}_{\text{adv}}$ to improve overall output quality.

$$\mathcal{L}_1 = \mathbb{E}\left[\left\|\boldsymbol{I}_{out} - \boldsymbol{I}_{gt}\right\|_1\right], \tag{4.9}$$

$$\mathcal{L}_{perc} = \mathbb{E}\left[\sum_i \left\|\phi_i\left(\boldsymbol{I}_{out}\right) - \phi_i\left(\boldsymbol{I}_{gt}\right)\right\|_1\right], \tag{4.10}$$

$$\mathcal{L}_{style} = \mathbb{E}\left[\sum_i \left\|\left(\psi_i\left(\boldsymbol{I}_{out}\right) - \psi_i\left(\boldsymbol{I}_{gt}\right)\right)\right\|_1\right], \tag{4.11}$$

$$\mathcal{L}_{adv} = \min_G \max_D \mathbb{E}_{\boldsymbol{I}_{gt}}\left[\log D\left(\boldsymbol{I}_{gt}\right)\right] + \mathbb{E}_{\boldsymbol{I}_{out}} \log\left[1 - D\left(\boldsymbol{I}_{out}\right)\right], \tag{4.12}$$

where $\phi i(\cdot)$ indicates the activation map from the i-th pooling layer of VGG-16. $\psi_i(\cdot) = \phi_i(\cdot)^T \phi_i(\cdot)$ denotes the Gram matrix. The final loss function is thus denoted as:

$$\begin{aligned}\mathcal{L}_{total}(\hat{\mathbf{I}}, \mathbf{I_{gt}}) =& \lambda_1 \mathcal{L}_1 + \lambda_2 \mathcal{L}_{\text{style}} + \lambda_3 \mathcal{L}_{\text{perc}} \\ &+ \lambda_4 \mathcal{L}_{\text{adv}},\end{aligned} \tag{4.13}$$

where the weighting coefficients λ_1 = 1, λ_2 = 250, λ_3 = 0.1, λ_4 = 0.001 were chosen based on the parameter analysis (see Section 4.3.4).

4.3 Experiments

In this section, we present a comprehensive evaluation of the proposed HINT. First, we describe the datasets employed and delve into the specifics of the implementation. Then, we compare HINT with state-of-the-art methods to showcase its superior performance, with both quantitative and qualitative results. Finally, we conduct thorough ablation studies to evaluate the significance of each proposed component.

4.3.1 Datasets

To assess the efficacy of our proposed method, we employ CelebA [35], CelebA-HQ [4], Places2-Standard [5] and Dunhuang Challenge [36] datasets. All experiments are conducted with 256×256 images, providing a comprehensive evaluation of our approach in a consistent and well-defined setting. The CelebA [35] and CelebA-HQ [4] are two human face datasets with different qualities, while the Places2-Standard dataset is a subset of the Places2 [5] dataset offering a diverse collection of scenes, such as indoor and outdoor environments, natural landscapes, and man-made structures and constructions. These three datasets are commonly used within the existing literature on inpainting [8, 72, 73], making them ideal for evaluating our approach. The Dunhuang Challenge [36] dataset represents a practical application of image inpainting in real-world scenarios.

For CelebA and Dunhuang, we follow the standard configuration to split the data for training and testing. In the case of the CelebA-HQ dataset, to ensure reproducibility, we use the first 28,000 images for training and the remaining 2,000 images for testing. For the Places2-Standard dataset, we use the standard training set and validation set

Table 4.1: Comparisons on the Dunhuang Challenge dataset.

Model	PSNR↑	SSIM↑	L1↓	LPIPS↓
StructFlow [130]	35.199	0.9559	0.475	0.0589
EdgeConnect [38]	36.419	0.9635	0.441	0.0480
RFRNet [64]	36.485	0.9648	0.401	0.0463
JPGNet [131]	37.646	0.9724	0.353	0.0469
MISF [127]	38.383	0.9735	0.341	0.0330
Ours	**38.6705**	**0.9743**	**0.3161**	**0.0286**

for training and testing, respectively. For mask settings, we follow prior work [2, 127] and use irregular masks [15] for CelebA, CelebA-HQ, and Places2. As for Dunhuang Challenge, we use the officially released masks for testing.

4.3.2 Implementation Details

In the 7-level transformer blocks, the number of Sandwich blocks is sequentially set to [4,6,6,8,6,6,4] and the attention head in SCAL are [1,2,4,8,4,2,1]. All experiments are carried out on a single NVidia A100 GPU with a batch size of 4. We adopt the Adam optimiser [128] with $\beta_1 = 0.9$, $\beta_2 = 0.999$. The learning rate is initially set to $1e^{-4}$ and is halved at the 75% milestone of the training progress. Compared to the state of the art in the existing literature [2, 72, 129], our approach is more robust against small changes in the training procedure, making it more generalizable and easier to deploy. Our training pipeline does not rely on warm-up step [72], pre-training requirements [129] or fine-tuning [2].

4.3.3 Comparison with the State of the Art

In assessing our HINT, designed to generate high-quality, fine-grained images, we follow [127] to employ a suite of evaluation metrics: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity (SSIM), L1 and Perceptual Similarity (LPIPS). These chosen metrics align with our intent to create a nuanced and comprehensive understanding of the performance of models. PSNR and L1 are used to measure pixel-wise reconstruction accuracy, which reflects the fidelity of the inpainted output. SSIM [86] evaluates structural similarity, ensuring the inpainted segments remain coherent within the image contextually.

Table 4.2: Number of parameter and inference time

Model	Param $\times 10^6$	Infer. Time/per img
DeepFill v1 [18]	3	7 ms
DeepFill v2 [16]	4	10 ms
Wavefill [132]	49	70 ms
CTSDG [2]	52	20 ms
WNet [133]	46	35 ms
MISF [127]	26	10 ms
MAT [8]	62	70 ms
LAMA [7]	51	25 ms
Stable Diffusion	860	880 ms
LDM [62]	387	6000 ms
Repaint [134]	552	250000 ms
Ours	139	125 ms

Table 4.3: Comparison with diffusion models.

Plces2	PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓
LDM [62]	19.6476	0.7052	4.6895	27.3619	0.2675
Stable Diffusion*	19.4812	0.7185	4.5729	27.8830	0.2416
Ours	**20.8579**	**0.7227**	**4.3814**	**26.7895**	**0.2102**
CelebA-HQ	PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓
RePaint [134]	21.8321	0.7791	3.9427	8.9637	0.1943
Ours	**24.1287**	**0.8241**	**2.778**	**7.5793**	**0.1449**

*: The officially released Stable Diffusion inpainting model pretrained on high-quality LAION-Aesthetics V2 5+ dataset.

We also include LPIPS [135], a learned perceptual metric, capable of detecting complex distortions that mirror human perceptual differences, a crucial attribute when the aim is to produce high-quality imagery.

We categorise the masks into three groups based on the mask ratio, i.e., small (0.01%-20%), medium (20%-40%) and large (40%-60%), referring to the extent of missing regions.

Quantitative Results As shown in Tab. 4.11 and Tab. 4.1, HINT achieves a better overall performance across all datasets and mask ratios than the state of the arts [2, 8, 72, 127, 131, 132]. Compare to the latest transformer-based MAT [8] on CelebA-HQ, HINT improves PSNR by 5.7%, 3.3% and 3.4% at the increasing mask ratios respectively, demonstrating that it preserves more high-fidelity details in reconstructed images. In Places2, compared with the latest high-quality inpainting method MISF [127], HINT achieves a 12.6%, 13.8% and 7.2% decrease for LPIPS, showcasing its effectiveness in perceptual recovery. Since the Dunhuang Challenge provides standard masks, we crawled the benchmark from [127] for comparison. HINT outperforms existing models across all metrics.

For a comprehensive and robust evaluation, we also compare our model with the state-of-the-art diffusion model-based methods with large masks, which are well-known for their prowess in generating high-quality images [136]. Three prominent diffusion models, LDM [62], Stable Diffusion (SD) and RePaint [134], are chosen for comparison. To allow for a fair comparison, all experiments are conducted on officially released pretrained models on the corresponding datasets. It is important to note that SD does not provide models pretrained on either CelebA-HQ or Places2, so, we chose the LAION v2 5+ pretrained model, as its data distribution is similar to that of the Places2 dataset, but it is much larger and of higher quality. Tab. 4.3 and Tab. 4.2 underscore the superior performance of our model across all metrics and signify the efficiency in image inpainting tasks. Ideally, we wish to assess all diffusion model on Places2. However, due to the significant inference time required by RePaint (Tab. 4.2), a single evaluation on the Places2 dataset for three mask ratios demands around one GPU-year, making it computationally intractable. As a result, we chose to evaluate LDM on Places2, given its relatively more manageable inference time, and focused our analysis of RePaint on the CelebA-HQ dataset.

Qualitative Results We provide the exemplar visual results to further demonstrate

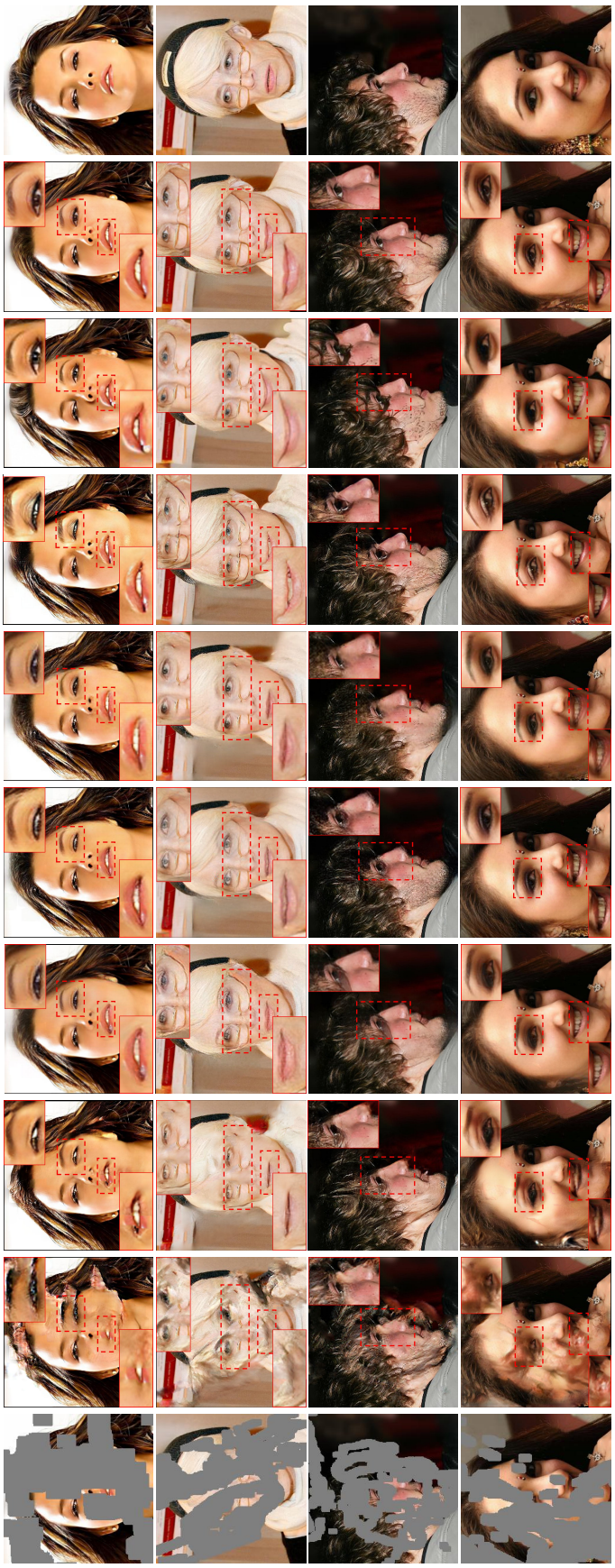

Figure 4.4: Comparisons on CelebA-HQ [4] with visualisations (256×256).

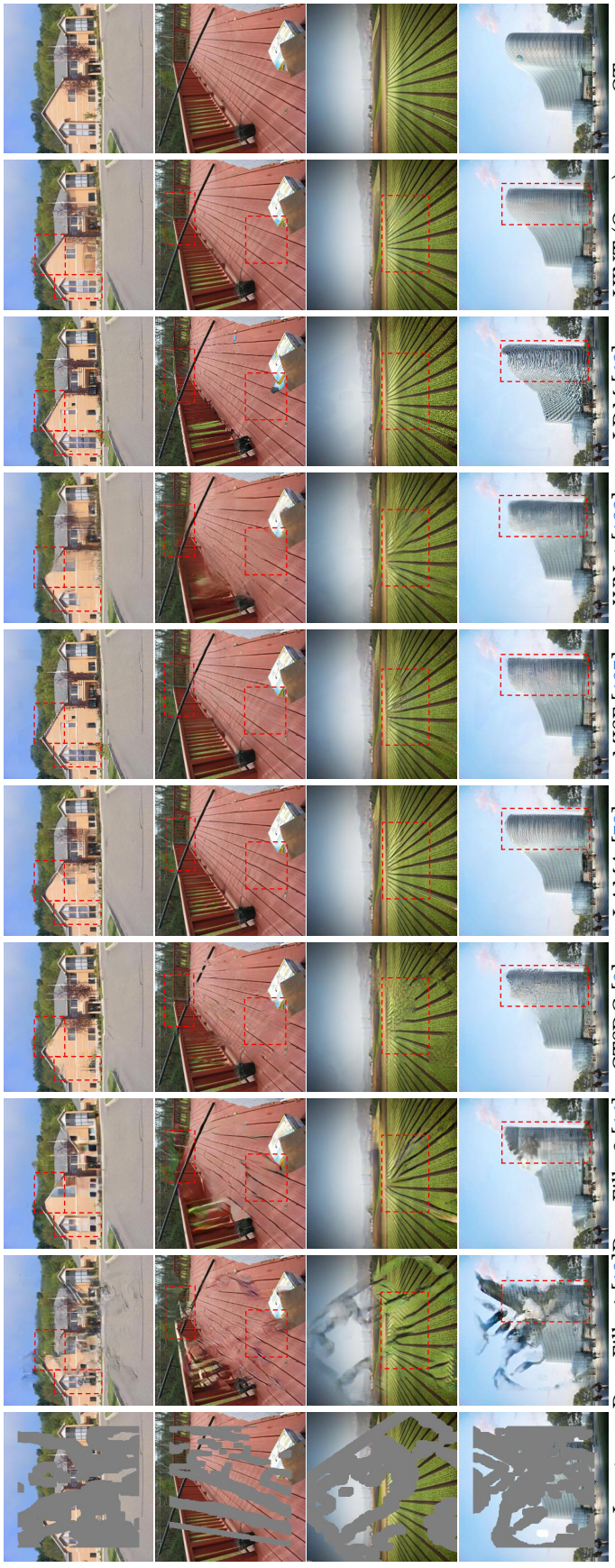

Figure 4.5: Comparisons on Places2 [5] with visualisations (256×256).

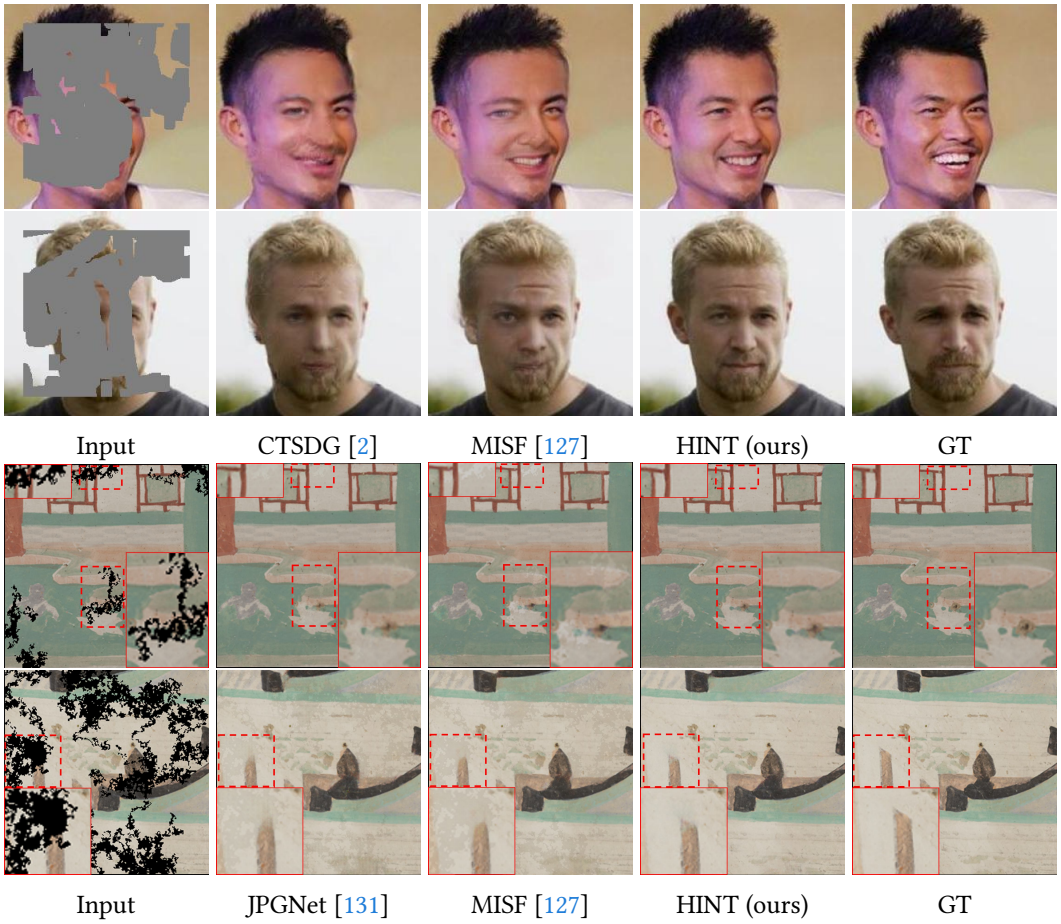

Figure 4.6: Comparisons with visualisations (256×256) showing that our results are more coherent in structure and sharper in texture and semantic details. The top two rows are from CelebA-HQ [4] and the bottom two rows are from Places2 [5].

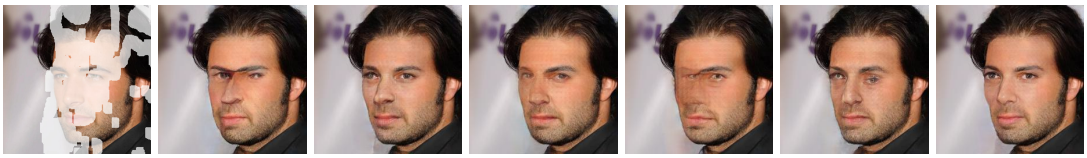

Figure 4.7: Visual results of our ablation studies. A refers to replacing MPD with conventional PD, B removes the first FFN in "sandwich", C replaces SCAL with a single channel-wise self-attention design, D ablates HINT to only include channel self-attention, a single FFN, and convolutional down-sampling. E replaces our spatial branch with the basic gated mechanism from [6].

the advantages of HINT over comparators. As shown in Fig. 4.4 and Fig. 4.5, our model generates high-quality images with more coherent structures and fewer artifacts, such as roofs and planks. For face restoration, our model better recovers finer-grained details, such as eye features, compared to the current state of the art [8, 72, 132]. We also provide qualitative results for CelebA [35] and Dunhuang datasets [36] in Fig. 4.6, Fig. 4.12, and Fig. 4.13, to indicate our superior performance in global context modelling. The proposed HINT recovers high-quality faces with clear textures and plausible semantics, even with a large mask covering almost all facial attributes. The results on Dunhuang show that our model suppresses the generation of light mottle, and demonstrates the effectiveness of our model in handling small scratch masks.

Efficiency Comparison Our model uniquely incorporates spatial awareness into the channel-wise self-attention, a design innovation that maintains linear complexity, $\mathcal{O}(C^2)$, with C being the channel number. It manages to strike an impressive balance between complexity and efficiency. As shown in Tab. 4.2, our model, carrying 139 million parameters, still situates itself within the parameter counts seen among state-of-the-art methods. More significantly, our model upholds an inference time of 125ms per image, ensuring practicality with millisecond-level response time. This efficiency does not come at the expense of performance since our model outshines competing methods in both qualitative and quantitative evaluations.

4.3.4 Ablation Study and Parameter Analysis

We conducted a series of ablation experiments on the CelebA-HQ dataset to evaluate the impact of each proposed component by downgrading them. All models are trained for

Table 4.4: Ablation studies. Setup A replaces MPD with conventional PD, B removes the first FFN in "Sandwich", C replaces SCAL with single channel-wise self-attention design, D is a HINT variant with the spatial branch replaced by [6]'s gated mechanism.

Setup	Model	0.01%-20%					20%-40%					40%-60%				
		PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓	PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓	PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓
A	w/o MPD	34.5955	0.9649	0.4780	1.7458	0.0381	26.9292	0.8863	1.6320	4.9815	0.1084	22.5618	0.7813	3.4982	8.2196	0.1951
B	w/o Sandwich	34.7272	0.9658	0.4661	1.4687	0.0361	27.0914	0.8893	1.5796	4.7625	0.1050	22.7027	0.7853	3.4185	7.9138	0.1912
C	w/o SCAL	34.7951	0.9659	0.4624	1.7568	0.0364	27.1193	0.8895	1.5732	4.8769	0.1057	22.7206	0.7856	3.4021	8.1627	0.1925
D	U-Net w self-attention	34.0204	0.9538	0.5129	2.0152	0.0497	26.0814	0.8754	1.8547	5.1029	0.1277	21.6149	0.7679	3.6912	8.9314	0.2104
E	Full†	34.3155	0.9636	0.4891	1.3968	0.0393	26.7534	0.8837	1.6521	4.7358	0.1122	22.4632	0.7772	3.5221	7.9637	0.1999
Ours	Full	**35.0436**	**0.9671**	**0.4489**	**1.3542**	**0.0345**	**27.2954**	**0.8924**	**1.5363**	**4.6891**	**0.1016**	**22.8473**	**0.7895**	**3.3403**	**7.8697**	**0.1867**

Table 4.5: "Attention-FFN" structure vs. "FFN-Attention-FFN" structure (Sandwich) with the same number of parameters.

Model	0.01%-20%					20%-40%					40%-60%				
	PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓	PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓	PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓
SCAL-FFN	34.7272	0.9658	0.4661	1.3716	0.0361	27.0914	0.8893	1.5796	4.7174	0.1050	22.7027	0.7853	3.4185	7.8970	0.1912
Conformer	34.5125	0.9576	0.4729	1.4028	0.3914	26.9672	0.8804	1.6760	4.7597	0.1083	21.2186	0.7218	3.6829	8.9506	0.2147
Thin-Sandwich (Ours)	**34.7843**	**0.9661**	**0.4614**	**1.3697**	**0.0357**	**27.1070**	**0.8911**	**1.5763**	**4.6993**	**0.1047**	**22.7075**	**0.7872**	**3.4077**	**7.8863**	**0.1908**

Table 4.6: Ablation study of using 1×1 convolution after the last skip connection.

Model	0.01%-20%					20%-40%					40%-60%				
	PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓	PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓	PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓
w 1×1 conv	34.5246	0.9646	0.4780	1.3693	0.0386	26.8984	0.8863	1.6267	4.7131	0.1101	22.4694	0.7792	3.5434	7.9647	0.1997
w/o 1×1 conv (Ours)	**35.0436**	**0.9671**	**0.4489**	**1.3542**	**0.0345**	**27.2954**	**0.8924**	**1.5363**	**4.6891**	**0.1016**	**22.8473**	**0.7895**	**3.3403**	**7.8697**	**0.1867**

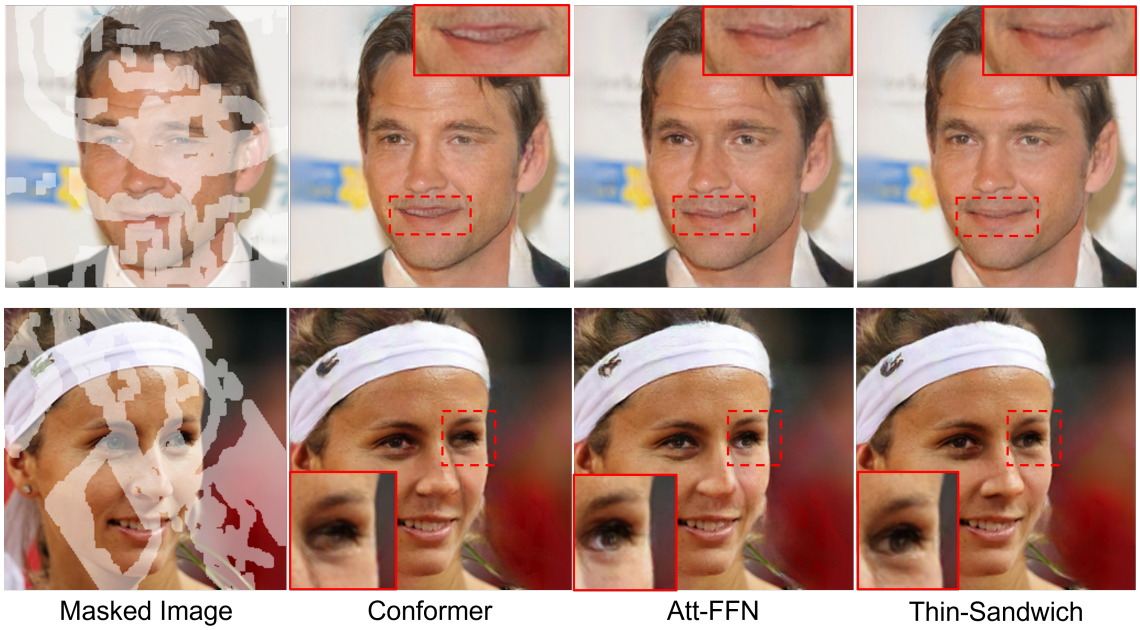

Figure 4.8: Visual results of the variants of sandwich.

Table 4.7: Different kernal size in the embedding layer.

Model	0.01%-20%					20%-40%					40%-60%				
	PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓	PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓	PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓
7×7 emb	34.6389	0.9657	0.4681	1.4034	0.0366	27.0422	0.8905	1.5803	4.9783	0.1043	22.6667	0.7865	3.3950	7.9168	0.1898
3×3 emb (Ours)	**35.0436**	**0.9671**	**0.4489**	**1.3542**	**0.0345**	**27.2954**	**0.8924**	**1.5363**	**4.6491**	**0.1016**	**22.8473**	**0.7895**	**3.3403**	**7.8697**	**0.1867**

Table 4.8: Comparison of alterantive design of mask-aware pixel-shuffle down-sampling

Model	0.01%-20%					20%-40%					40%-60%				
	PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓	PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓	PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓
CD	34.3159	0.9641	0.4842	1.4875	0.0380	26.6809	0.8846	1.6499	4.9362	0.1088	22.2408	0.7761	3.5828	8.2493	0.1979
PD	34.5229	0.9647	0.4787	1.3729	0.0393	26.8446	0.8865	1.6328	4.8756	0.1116	22.4448	0.7795	3.5466	8.0196	0.2023
MPD (Ours)	**34.5820**	**0.9649**	**0.4733**	**1.3542**	**0.0375**	**26.9327**	**0.8867**	**1.6089**	**4.6891**	**0.1085**	**22.4769**	**0.7812**	**3.5001**	**7.8697**	**0.1972**

30,000 iterations. Our quantitative comparison results, which are presented in Tab. 4.12, demonstrate the effectiveness of our key contributions. "U-Net w self-attention" (model D) is the variant in which we ablate HINT to only include channel self-attention [10], a single FFN, and convolutional down-sampling. We also present visual results for a more intuitive demonstration in Fig. 4.7.

Spatially-activated Channel Attention Layer Our proposed SCAL captures channel-wise long-range dependencies while complementing the spatial attention in an efficient manner. We suggest that introducing the spatial attention identifies "where" the important regions are. As illustrated in Fig. 4.7 (model C), after removing the spatial attention, the model is not confident enough to determine if an eye is missing on the left, thus generating a very blurry left eye. We substituted our spatial branch with the basic gated mechanism from [6] (model E) to evaluate our superiority. In Tab. 4.10, we replace the spatial branch with traditional spatial self-attention (SSA), denoted as 'w SSA', to evaluate our efficiency. However, due to the significant computational cost of SSA, we have to resize the image to 64×64 to train on a single A100. For a fair comparison, all experiments in Tab. 4.10 are conducted on 64×64 images. We notice that the significant computational cost of SSA does not bring better performance, which is reflected in the ambiguous features with the blur texture (shown in Fig. 4.9).

Mask-aware Pixel-shuffle Down-sampling Our novel downsampling method based on pixel shuffling maintains a consistent flow of valid information within the transformer. First, to demonstrate the feasibility of pixel-shuffle down-sampling, we compare the performance of convolutional downsampling, conventional PD and the proposed MPD on the baseline. We ablate all proposed designs to build the baseline,

Table 4.9: Hyper-parameter tuning on the weights associated with different losses.

Model	0.01%-20%					20%-40%					40%-60%				
	PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓	PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓	PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓
Sample A	34.6959	0.9581	0.4417	1.3143	0.0355	26.8916	0.8358	1.6470	4.8173	0.1073	22.7519	0.7850	3.4011	7.9715	0.1902
Sample B	33.5762	0.9233	0.4256	1.0158	0.0384	26.3527	0.8657	1.5419	4.9836	0.1216	22.4893	0.7754	3.4581	8.0381	0.1972
Sample C (ours)	**35.0436**	**0.9671**	**0.4489**	**1.3542**	**0.0345**	**27.2954**	**0.8924**	**1.5363**	**4.6891**	**0.1016**	**22.8473**	**0.7895**	**3.3403**	**7.8697**	**0.1867**

Table 4.10: Ablation study of using traditional spatial self-attention in the SCAL on the 64×64 resolution.

Model	0.01%-20%					20%-40%					40%-60%				
	PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓	PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓	PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓
w SSA	34.4915	0.9515	0.6812	1.5641	0.0295	26.2375	0.8726	2.3536	5.0318	0.0720	21.3534	0.7726	4.3687	8.4753	**0.1312**
SCAL (Ours)	**34.5849**	**0.9538**	**0.6715**	**1.5119**	**0.0271**	**26.3195**	**0.8781**	**2.2328**	**4.9513**	**0.0707**	**21.5242**	**0.7774**	**4.2918**	**8.4518**	**0.1312**

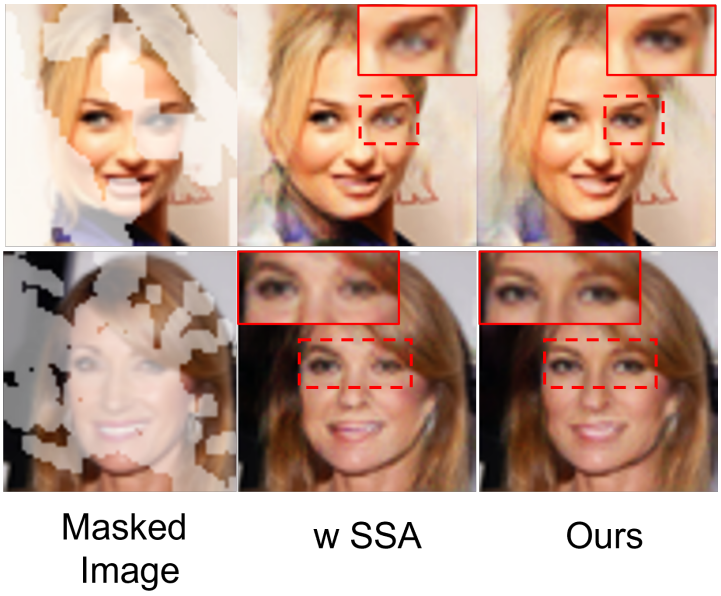

Figure 4.9: Visual results of the variants of SCAL.

including the "Attention-FFN" structure, single channel-wise self-attention branch, and conventional PD. As shown in Tab. 4.8, directly using conventional PD provides an overall improvement compared to convolutional down-sampling, but leads to a decline in LPIPS. We attribute this degradation to the incoherence of invalid information, which causes inaccurate transfer of high-level feature representations. MPD solves this problem and improves LPIPS significantly. Correspondingly, in Tab. 4.12, the performance of HINT suffers the largest drop when we replace the MPD with conventional PD. As shown in Fig. 4.7 (model A), the facial attributes are severely drifting when MPD is removed.

Sandwich-shaped Transformer Block We introduce an FFN-SCAL-FFN block to effectively manage the limited flow of information. As evidenced by the results in Tab. 4.12, removing the first FFN in the sandwich leads to a notable decrease across all four metrics. In Fig. 4.7 (model B), the model fails to learn a good enough feature representation of the eyeball and nose, resulting in unclear textures for the generated left eye and nose. Furthermore, to confirm that the effectiveness of the proposed Sandwich Network is not merely attributed to an increase in the number of parameters, we implemented a lightweight variant that diminishes the parameter count in both Feedforward Neural Networks (FFNs) by 50%. This thin "Sandwich" configuration possesses an equivalent number of parameters as the "Attention-FFN" architecture. Furthermore, we substituted our "FFN-SCAL-FFN" with the Conformer structure (FFN-Attention-CONV-FFN) [126] to evaluate our superiority. As shown in Fig. 4.8, the proposed Thin-Sandwich helps the model to learn a better feature representation of the eyeball and mouth to provide clearer texture details. Although Conformer also has a "sandwich" structure, it moves the convolutional layer that can extract local spatial feature behind the attention. Therefore, it does not embed good enough features for the following attention layer, making it difficult to generate clear texture and structure in the generated area. As shown in Table 4.5, the experimental results substantiate that, given an equal parameter quantity, the Sandwich module enhances the overall performance of the model.

Decision for the Last Skip Connection To harness the low-level texture and structural features derived from the encoder, we refrain from utilising 1×1 convolution for modulating the number of channels post the last skip connection. The contrast between the two strategies is enumerated in Tab. 4.6.

Embedding Layer In the embedding layer, we adopt a gated convolutional layer with padding to embed the input without downsampling. In contrast to prior works using 7×7 convolutional layers to project the input [2, 38, 127], a smaller kernel size (3×3) is employed in our embedding layer to obtain more fine-grained features. As illustrated in Table 4.7, the smaller kernel gains better performance.

Parameter Tuning To tune HINT, we employ Optuna [107] to identify the best set of hyper-parameters in terms of different values of weights of our loss components. The top three sets of combinations are λ_i: sample A [1,1,0.5,2], sample B [1,60,1,2], sample C [1,250,0.1,0.001], as shown in Tab. 4.9. We implement the sample C for all of the experiments.

4.4 Conclusion and Discussions

We propose HINT, an end-to-end Transformer for image inpainting with the proposed MPD module to ensure information remains intact and consistent throughout the encoding process. The MPD is a plug-and-play module, which is easy to adopt to the other multimedia tasks that require masking process, such as video edit, and animation edit. Our SCAL, enhanced by the proposed "sandwich" module, captures long-range dependencies while remaining spatial awareness, to boosting the capacity of representation learning in a cheap approach, which could potentially benefit multimedia tasks that are based on channel self-attention.

The proposed components contribute to each other and drive HINT to recover high-quality completed images. Experimental results demonstrate that HINT overall surpasses the current state of the art on four datasets [4, 5, 35, 36], with particularly notable improvements observed on facial datasets [4, 35]. Extensive qualitative evaluations demonstrate the superior image quality achieved by our framework.

As a direction of future research, HINT can be improved by employing geometric information [38, 137] by simply adding an indicator or incorporating a multi-task architecture, to get better structural consistency. Furthermore, considering the success of existing work [138], HINT can be potentially upgraded to a text-guided image inpainting system by introducing the pre-trained multi-model features to interpret the text feature

into the latent space.

Furthermore, unlike existing multi-step approaches [8, 72, 73], as HINT is already able to recover high-quality completed images without requiring additional refinement process, a second stage of reconstruction could further enhance the quality of the results. Constrained by the current limited computing resources, we will implement another refinement network in the second step, utilising the results from HINT as inputs and fine-tuning them in the same scale. Two networks are trained separately, thereby avoiding the large number of parameters introduced by joint training.

Table 4.11: Comparison results on (a, top) CelebA-HQ, (b, middle) CelebA and (c, bottom) Places2. The **bold** and underline indicate the best and the second best respectively.

CelebA-HQ	0.01%-20%					20%-40%					40%-60%				
Method	PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓	PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓	PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓
DeepFill v1 [18]	34.2507	0.9047	1.7433	2.2141	0.1184	26.8796	0.8271	2.3117	9.4047	0.1329	21.4721	0.7492	4.6285	15.4731	0.2521
DeepFill v2 [16]	34.4735	0.9533	0.5211	1.4374	0.0429	27.3298	0.8657	1.7687	5.5498	0.1064	22.6937	0.7962	3.2721	8.8673	0.1739
LaMa [7]	35.5656	0.9685	0.4029	1.4309	0.0319	28.0348	0.8983	1.3722	4.4295	0.0903	23.9419	0.8003	2.8646	8.4538	0.1620
WNet [133]	35.3591	0.9647	0.4957	1.2759	0.0287	28.1736	0.8872	1.4495	4.7299	0.0833	23.8357	0.7872	2.9316	9.4926	0.1649
MAT [8]	35.5466	0.9689	0.3961	1.2428	0.0268	27.6684	0.8957	1.3852	3.4677	0.0832	23.3371	0.7964	2.9816	5.7284	0.1575
WaveFill [132]	31.4695	0.9290	1.3228	6.0638	0.0802	27.1073	0.8668	2.1159	8.3804	0.1231	23.3569	0.7817	3.5617	13.0849	0.1917
Ours	**36.5725**	**0.9777**	**0.3942**	**1.1128**	**0.0228**	**28.6247**	**0.9195**	**1.2885**	**3.3915**	**0.0745**	**24.1287**	**0.8241**	**2.7778**	**5.6179**	**0.1449**

Places2	0.01%-20%					20%-40%					40%-60%				
Method	PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓	PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓	PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓
DeepFill v1 [18]	30.2958	0.9532	0.6953	26.3275	0.0497	24.2983	0.8426	2.4927	31.4296	0.1472	19.3751	0.6473	5.2092	46.4936	0.3145
DeepFill v2 [16]	31.4725	0.9558	0.6632	23.6854	0.0446	24.7247	0.8572	2.2453	27.3259	0.1362	19.7563	0.6742	4.9284	36.5458	0.2891
CTSDG [2]	32.111	0.9565	0.6216	24.9852	0.0458	24.6502	0.8536	2.1210	29.2158	0.1429	20.2962	0.7012	4.6870	37.4251	0.2712
WNet [133]	32.3276	0.9372	0.5913	20.4925	0.0387	25.2198	0.8617	2.0765	24.7436	0.1136	20.4375	0.6727	4.6371	32.6729	0.2416
MISF [127]	32.9873	0.9615	0.5931	21.7526	0.0357	25.3843	0.8681	1.9460	30.5499	0.1183	20.7260	0.7187	4.4383	44.4778	0.2278
LaMa [7]	32.4660	0.9584	0.5969	14.7288	0.0354	25.0921	0.8635	2.0048	22.9381	0.1079	20.6796	0.7245	4.4060	25.9436	0.2124
WaveFill [132]	29.8598	0.9468	0.9008	30.4259	0.0519	23.9875	0.8395	2.5329	39.8519	0.1365	18.4017	0.6130	7.1015	56.7527	0.3395
Ours	**33.0276**	**0.9689**	**0.5612**	**13.9128**	**0.0307**	**25.4216**	**0.8807**	**1.9270**	**20.0241**	**0.1003**	**20.9243**	**0.7470**	**4.3296**	**25.7150**	**0.2041**

CelebA	0.01%-20%					20%-40%					40%-60%				
Method	PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓	PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓	PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓
CTSDG [2]	36.465	0.9732	0.5871	2.5876	0.0334	29.1393	0.9159	1.38	7.4925	0.0935	23.8371	0.8157	3.04	9.8473	0.1815
MISF [127]	36.8981	0.9747	0.3441	3.3598	0.0333	28.9270	0.9103	1.227	8.0249	0.1031	23.5355	0.8033	3.182	13.2475	0.2012
Ours	**37.5696**	**0.9754**	**0.3402**	**1.0270**	**0.0232**	**29.8525**	**0.9208**	**1.220**	**4.1359**	**0.0689**	**24.4538**	**0.8270**	**2.7802**	**5.3612**	**0.1408**

Table 4.12: Ablation studies. Setup A replaces MPD with conventional PD, B removes the first FFN in "Sandwich", C replaces SCAL with single channel-wise self-attention design, D is a HINT variant with the spatial branch replaced by [6]'s gated mechanism.

Setup	Model	0.01%-20%					20%-40%					40%-60%				
		PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓	PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓	PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓
A	w/o MPD	34.5955	0.9649	0.4780	1.7458	0.0381	26.9292	0.8863	1.6320	4.9815	0.1084	22.5618	0.7813	3.4982	8.2196	0.1951
B	w/o Sandwich	34.7272	0.9658	0.4661	1.4687	0.0361	27.0914	0.8893	1.5796	4.7625	0.1050	22.7027	0.7853	3.4185	7.9138	0.1912
C	w/o SCAL	34.7951	0.9659	0.4624	1.7568	0.0364	27.1193	0.8895	1.5732	4.8769	0.1057	22.7206	0.7856	3.4021	8.1627	0.1925
D	U-Net w self-attention	34.0204	0.9538	0.5129	2.0152	0.0497	26.0814	0.8754	1.8547	5.1029	0.1277	21.6149	0.7679	3.6912	8.9314	0.2104
E	Full†	34.3155	0.9636	0.4891	1.3968	0.0393	26.7534	0.8837	1.6521	4.7358	0.1122	22.4632	0.7772	3.5221	7.9637	0.1999
Ours	Full	**35.0436**	**0.9671**	**0.4489**	**1.3542**	**0.0345**	**27.2954**	**0.8924**	**1.5363**	**4.6891**	**0.1016**	**22.8473**	**0.7895**	**3.3403**	**7.8697**	**0.1867**

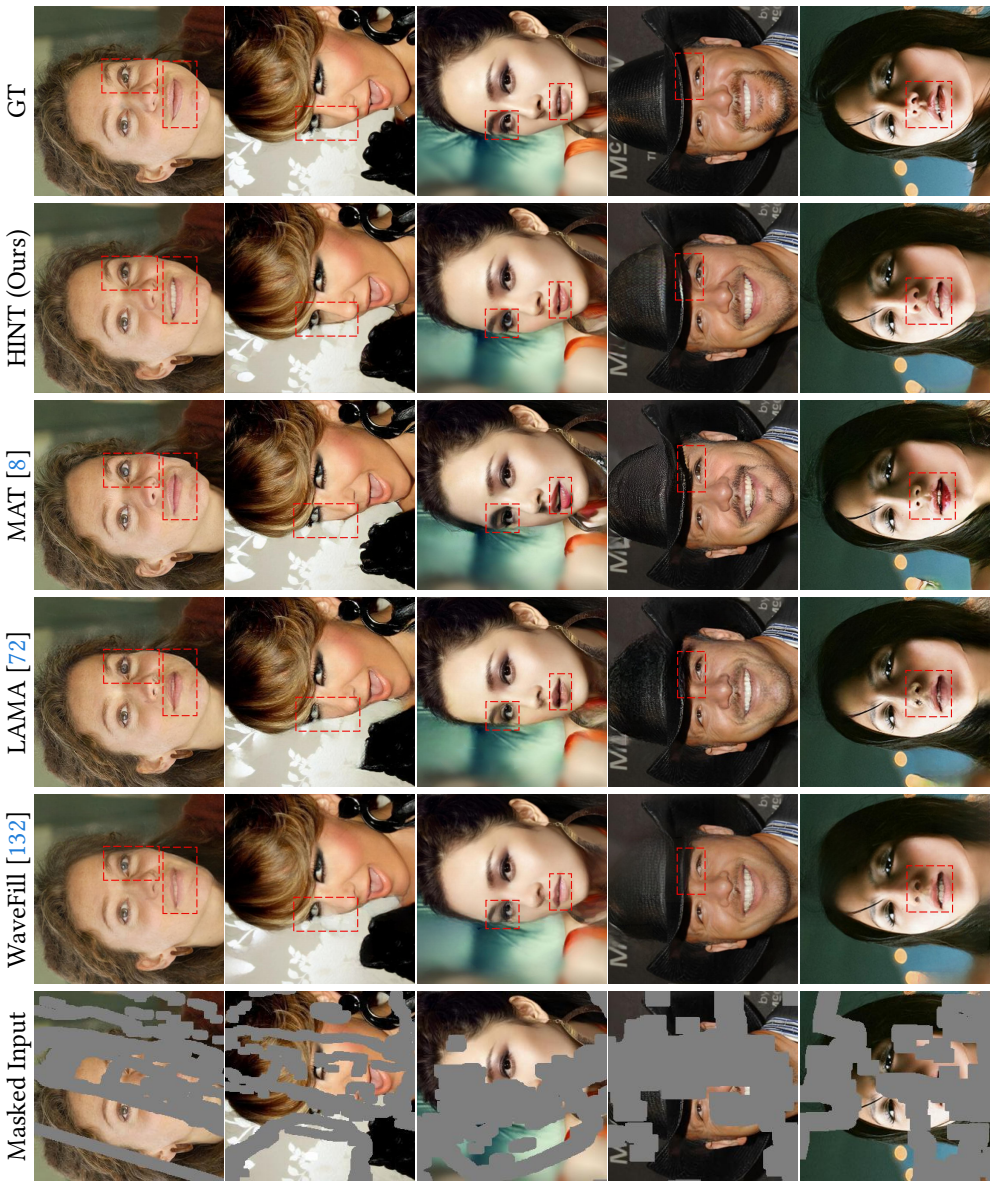

Figure 4.10: More visualisations (256×256) on the CelebA-HQ dataset. Please zoom in to see details.

Figure 4.11: More visualisations (256×256) on the Places2 dataset. Please zoom in to see details.

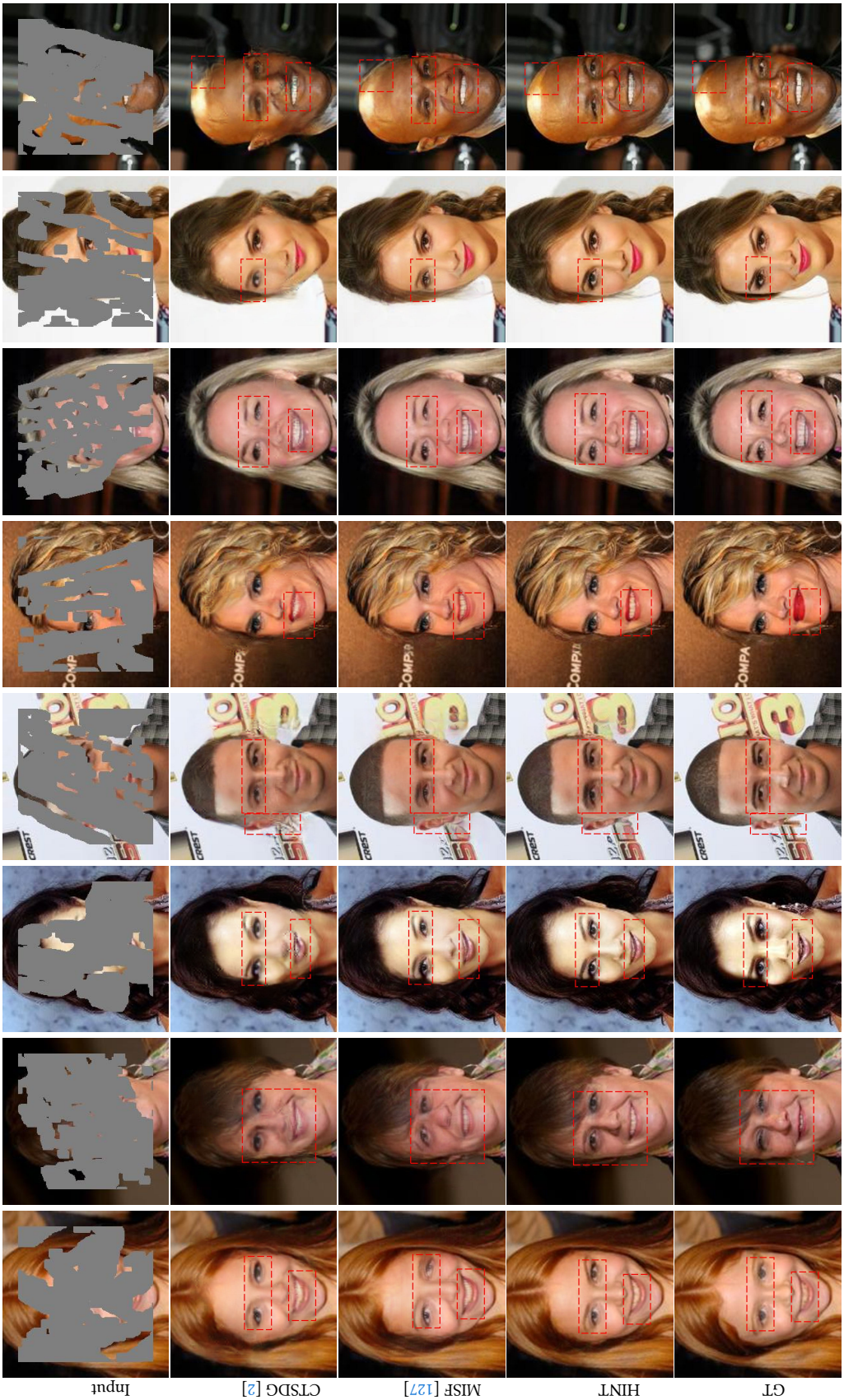

Figure 4.12: More visualisations (256×256) on the CelebA dataset. Please zoom in to see details.

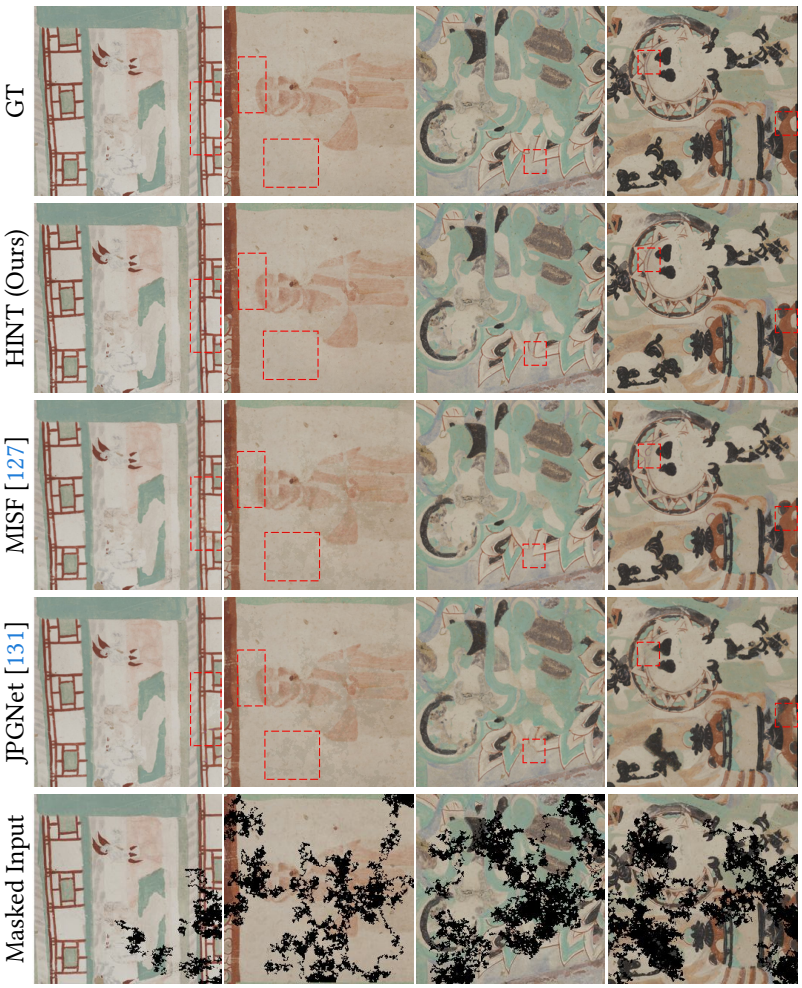

Figure 4.13: More visualisations (256×256) on the Dunhuang dataset. Please zoom in to see details.

CHAPTER 5

Long-Range Dependency Capture and Pixel-Level Sequential Modelling

Portions of this chapter have previously been published in the following peer-reviewed publications:

- **Shuang Chen**, Amir Atapour-Abarghouei, Haozheng Zhang, Hubert P. H. Shum, "$M \times T$Mamba $\times$ Transformer for Image Inpainting", In *the 2024 British Machine Vision Conference* (BMVC). 2024.
- **Shuang Chen**, Haozheng Zhang, Amir Atapour-Abarghouei, Hubert P. H. Shum, "SEM-Net: Efficient Pixel Modelling for Image Inpainting with Spatially Enhanced SSM", In *the 2025 IEEE/CVF Winter Conference on Applications of Computer Vision* (WACV). IEEE, 2025.

Image inpainting, or image completion, is a crucial task in computer vision that aims to restore missing or damaged regions of images with semantically coherent content. This technique requires a precise balance of local texture replication and global contextual understanding to ensure the restored image integrates seamlessly with its surroundings. Traditional methods using Convolutional Neural Networks (CNNs) are effective at capturing local patterns but often struggle with broader contextual relationships due

to the limited receptive fields. Recent advancements have incorporated transformers, leveraging their ability to understand global interactions. However, these methods struggle to maintain fine-grained details and face computational inefficiencies, which limit the ability to capture long-range dependencies. To overcome these challenges, in this chapter, we introduce $M \times T$ and SEM. $M \times T$ is composed of the proposed Hybrid Module (HM), which combines Mamba with the transformer in a synergistic manner, given the observation of that, Mamba is adept at efficiently processing long sequences with linear computational costs, making it an ideal complement to the transformer for handling long-scale data interactions. SEM is a novel visual State Space model (SSM) vision network, modelling corrupted images at the pixel level while capturing long-range dependencies (LRDs) in state space, achieving a linear computational complexity. Such SEM is able to address the inherent lack of spatial awareness in SSM. We evaluate both $M \times T$ and SEM on the widely-used CelebA-HQ and Places2-standard datasets, where they consistently outperformed existing state-of-the-art methods.

5.1 Introduction

Convolutional Neural Networks (CNNs) have been employed for image inpainting, capitalizing on their ability to capture local patterns and textures. However, CNNs-based methods are inherently limited by the slow-grown receptive field, which limits the ability to grasp broader image context [8, 10]. To solve this issue, recent advancements [8, 58] have seen the integration of transformer or self-attention into image inpainting, leveraging their capability to capture global correlations across entire images. However, transformer-based methods are often constrained by quadratic computational complexity, prompting most methods to process images in smaller patches to reduce the spatial dimension [72, 73], to learn the interaction in patch-level. This patch-based approach hinders the learning of fine-grained details, often resulting in artifacts in the generated images.

Mamba [34], merging from the domain of long-sequence modelling, offers promising advantages for handling long sequential data and capturing long-range dependency efficiently, all at a linear computational cost. This capability makes Mamba particu-

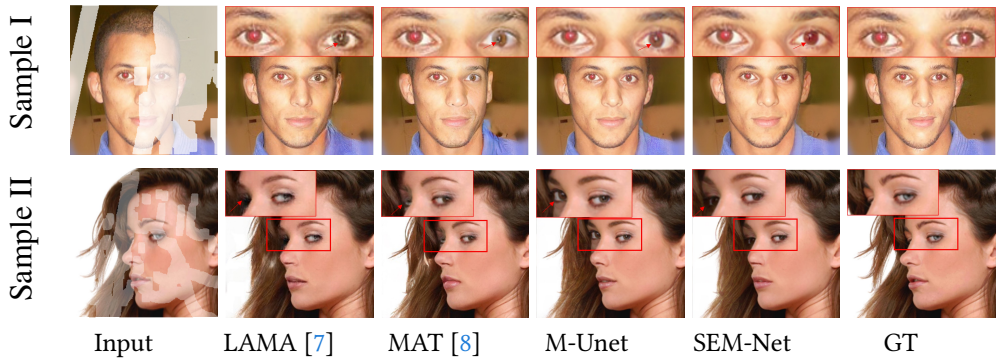

Figure 5.1: Comparisons with the state-of-the-art CNN-based method [7] and transformer-based method [8]. M-Unet is a variant of directly applying the Mamba model [9] followed by a feedforward network [10] in a U-Net. Red boxes and arrows highlight major differences. Our SEM-Net demonstrates the strong capability to capture LRDs visualised by the consistent eye colors and patterns, and addresses the challenge of lack of spatial awareness in M-Unet. Please refer to the supplementary material for more quantitative results.

larly suitable for globally learning interactions at the pixel level, thus complementing transformers by adding detailed context.

We observe that, Mamba and transformer exhibit complementary strengths: Mamba is good at learning long-range pixel-wise dependency, which is computationally expensive for the transformer. Conversely, transformer is good at capturing global interactions between localized patches, such spatial awareness is an area that Mamba lacks due to it being designed for sequence modelling. Based on this observation, we propose $M \times T$, consisting of proposed Hybrid Modules that synergistically combine the strengths of both transformer and Mamba. This novel approach allows for dual-level interaction learning from the patch level and pixel level.

Apart from introduce extra spatial information to SSM by using transformer, is there another way to enhance SSM in spatial awareness? Our hypothesis is that, as SSM is good at capturing long-range dependencies, it should be able to learn more plausible spatial associations with a proper strategy for modeling the visual data. Such hypothesis is proposed from the visualisation in Fig 5.1: The prominent CNN-based method [7] and transformer-based method [8] (Sample I of Fig. 5.1) struggle in learning consistent the eye colours and patterns, where the visible red eye fails to guide the accurate reconstruction of the other eye. Although, directly adopting SSM [9] (M-Unet) captures LRDs effectively and achieves more consistent eye colour, it lacks 2D spatial awareness, making the way to model pixels in SSM crucial. This is because the vanilla SSM scans the data as a

sequence with a single fixed direction. As illustrated in Sample II of Fig. 5.1, a vanilla SSM model [9] shows positional drifting of the inpainted left eye (upper than the right eye). This insight introduces two key challenges: (i) how to maintain the continuity and consistency of pixel adjacency for pixel-level dependencies learning while processing the SSM recurrence; and (ii) how to effectively integrate 2D spatial awareness to the predominant linear recurrent-based SSMs.

To solve these challenges, we propose **SEM-Net**: **S**patially-**E**nhanced SS**M** **Net**work for image inpainting, which is a simple yet effective encoder-decoder architecture incorporating four-stage Snake Mamba Blocks (SMB). The proposed SMB is assembled by two novel modules, which holistically integrate local and global spatial awareness into the model. Specifically, we introduce the Snake Bi-Directional Modelling module (SBDM) in place of vanilla SSM. It brings the crucial spatial context into a linear recurrent system, modelling images in two directions by consistently scanning each pixel with a snake shape. Moreover, we explicitly incorporate positional embedding into the sequences via a Position Enhancement Layer (PE layer) to strengthen the long-range positional awareness and improve the sensitivity to specific parts of the sequence(e.g., masked regions). We further propose Spatially- Enhanced Feedforward Network (SEFN) to complement the local spatial dependencies. aiming to leverage spatial information stored in the feature before SBDM, to refine the feature after SBDM with a gating mechanism.

Comparative experiments show that SEM-Net outperforms state-of-the-art approaches across two distinct datasets, i.e, CelebA-HQ [4] and Places2 [5]. Detailed qualitative comparison demonstrates that our method achieves a significant improvement in capturing spatial LRDs while preserving better spatial structure. In addition, SEM-Net achieves state-of-the-art performance on two motion-deblurring datasets, further demonstrating our method's generalisability in image representation learning.

Our main contributions are summarised as follows:

- We propose $M \times T$, to introduce Mamba combined with transformer focused for image inpainting.

- In $M \times T$, we design a novel Hybrid Module to capture the feature interaction at both the pixel level an patch level.

- We propose a novel U-shaped Spatially-Enhanced SSM architecture focused on capturing short- and long-range spatial dependencies in image inpainting.
- We propose a Snake Mamba Block (SMB), involving a Snake Bi-Directional Modelling (SBDM) module and a Position Enhancement Layer (PE layer), to implicitly integrate crucial spatial context awareness into a linear recurrent SSM, and explicitly enhance the long-range positional awareness.
- We propose a Spatially-Enhanced Feedforward Network (SEFN) to complement local spatial dependencies learning among pixels, enhancing the spatial awareness throughout image representation learning.
- Both $M \times T$ and SEM suppress the state-of-the-art methods on both CelebA-HQ and Places2 dataset.
- Both $M \times T$ and SEM are able to adapt to high-resolution images with only training on low-resolution data.

5.2 Mamba $\times$ Transformer

The overall pipeline of the proposed $M \times T$ is illustrated in Fig. 5.2, which is a U-Net shape architecture formed with 7 Hybrid Blocks. Formally, the masked image $I_{masked} \in \mathbb{R}^{H \times W \times 3}$ concatenated with a mask $M \in \mathbb{R}^{H \times W \times 1}$ as the input I_{in}. We first use an overlapped convolution to embed I_{in}, then feed into the following 7 Hybrid Blocks with 3 times downsampling and 3 times upsampling. At the end, one convolution layer projects the final output I_{out}. Each Hybrid Block consists of n Hybrid Modules, as shown in Fig. 5.2 (b), each Hybrid Module has a Transformer block, a Mamba block and a Context Broadcasting Feedforward Network (CBFN), which will be detailed in section 5.2.1.

5.2.1 Hybrid Module

Each of the seven Hybrid Modules involves a pair of SRSA (Spatial Reduced Self-Attention) and Mamba modules for capturing long-range dependency, followed by a Context Broadcasting Feedforward Network (CBFN) to enhance the local context and control data flow

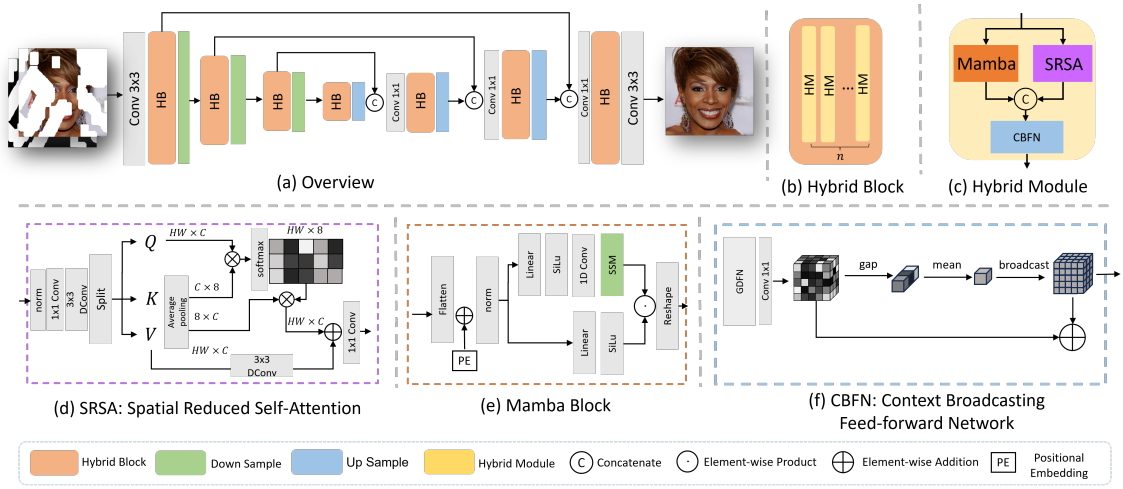

Figure 5.2: (a) The architecture overview of the proposed $M \times T$. (b) The Hybrid Block is composed of n proposed Hybrid Modules. (c) The proposed Hybrid Module, consisted of a Mamba Block, a Spatial Reduced Self-Attention and a Context Broadcasting Feed-forward Network. (d) The Spatial Reduced Self-Attention provides spatial awareness. (e) The Mamba Block captures pixel-level interaction. (f) The Context Broadcasting Feed-forward Network transfers the features.

consistency.

Spatial Reduced Self-Attention.

We introduce the Spatial Reduced Self-Attention (SRSA) module, designed to leverage the capability of the transformer for capturing global correlation while enriching local context detail. Reduced self-attention means that the attention operation is computed on a compressed set of tokens and channels rather than the full $H \times W$ feature map. Concretely, we form queries at the native resolution but build keys/values from a downsampled feature map (strided convolution or pooling with factor s), and we bottleneck the $Q/K/V$ channels by a ratio r (i.e., $C \to C/r$). Optionally, keys/values are further windowed to the masked neighborhood (local attention) instead of the whole image. This changes the complexity from $\mathcal{O}\big((HW)^2\big)$ to $\mathcal{O}\big(HW \cdot HW/s^2\big)$ (with an additional $1/r$ factor from channel reduction), and proportionally lowers memory.

Specifically, given a input feature F, we first apply layer normalisation followed by a 1×1 convolution and a 3×3 depth-wise convolution to extract the local features:

$$F' = DConv_{3\times3}(Conv_{1\times1}(LayerNorm(F))). \tag{5.1}$$

The feature F' is then split along the channel dimensions to form the Query Q, Key K and Value V. To address the traditional quadratic computational complexity of self-attention, we share the idea with PVTv2 [139] to adopt average pooling for K and V to a fixed dimension.

$$\begin{aligned} K', V' &= AvgPool(K), AvgPool(V), \\ Att &= softmax(K' \cdot Q), \end{aligned} \tag{5.2}$$

where Att is the attention map. In this work, the spatial dimension is reduced to 8. After multiplying Att and V', we get the initial output F''. To further enhance local context, we incorporate a Local Enhancement operation $LE(V)$ as proposed in [140], which is implemented using a 3×3 depth-wise convolution, to effectively balances capturing extensive global interactions with detailed local features. After an element-wise addition, the final output of SRSA is:

$$\begin{aligned} LE(V) &= DConv_{3\times3}(V), \\ Output_{srsa} &= LE(V) + Att \cdot V'. \end{aligned} \tag{5.3}$$

Mamba with Positional Embedding

Mamba showcases a strong capacity to handle long sequence data with linear computational complexity, making it highly effective for modelling interactions between adjacent pixels. In this work, we propose leveraging the Mamba module to modelling the flattened feature, thereby capturing long-range dependency at the pixel level, which is expensive to capture by self-attention. To adapt Mamba more aptly for vision tasks and enhance its ability to maintain positional awareness, we incorporate positional embedding into the module. Within the Mamba module, given an input feature F with the shape of (B, C, H, W), the process begins by flatting and transposing it to (B, C, L), where $L = H \times W$:

$$F' = transpose(reshape(F, (B, C, L))). \tag{5.4}$$

Subsequently, we introduce cosine positional embedding [68] to the transformed feature, enhancing the capacity to maintain positional awareness:

$$F'' = F' + PE(L). \tag{5.5}$$

After applying layer normalisation, mamba implements a gated mechanism to further refine the feature representation. The body branch involves a linear layer, a SiLU activation function [118], $1D$ convolutional layer and the SSM (State Space Sequence Models) layer.

$$F_{body} = SSM(Conv1D(SiLU(Linear(F'')))) \tag{5.6}$$

The gate branch involves a linear layer and a SiLU activation function [118]. After the gate branch re-weight the body branch, the output will be reshaped to the shape of (B, C, H, W):

$$\begin{aligned} G &= SiLU(Linear(F'')), \\ F_{gated} &= G \cdot F_{body}, \\ Output_{mamba} &= reshape(F_{gated}, (B, C, H, W)), \end{aligned} \tag{5.7}$$

where G is the gate matrix, F_{gated} is the output from gate mechanism, $Output_{mamba}$ is the final output from Mamba module.

Context Broadcasting Feed-forward Network.

We propose Context Broadcasting Feedforward Network (CBFN) by improving the Gated-Dconv Feed-Forward Network (GDFN) [10]. The GDFN is recognised for its efficacy in enhancing local context through a gated mechanism with depth-wise convolution. To build upon this, our CBFN integrates a global processing stage post-GDFN. Specifically, we implement global average pooling followed by channel-wise averaging to obtain the overall mean value of the input feature F, denoted as $\mu = GlobalAvgPool(F)$, where F is the output from GDFN. This μ is then broadcast to the dimensions of F and added to

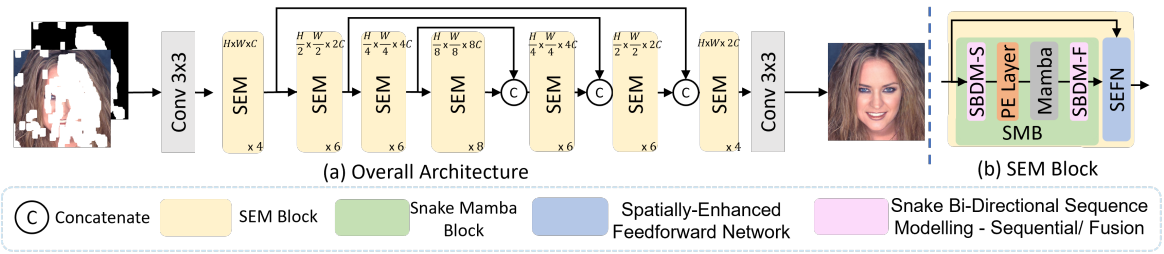

Figure 5.3: (a) Architecture overview of the proposed SEM-Net with multi-scale SEM blocks. (b) The details in each SEM block with core designs in SMB and SEFN, which holistically enhance the spatial awareness and improve the capability to capture LRDs.

it. The output of CBFN is represented as F':

$$F' = F + broadcast(\mu). \tag{5.8}$$

This global processing is designed to facilitate the learning of dense interactions within the self-attention layers [141], thereby enhancing the effectiveness of the Hybrid Module.

5.2.2 Loss Functions

To achieve superior inpainting outcomes, we adopt a multi-component loss strategy as delineated in the previous research [38, 58, 127]. This strategy includes an $\mathcal{L}_1$ loss, a style loss $\mathcal{L}_{\text{style}}$, a perceptual loss $\mathcal{L}_{\text{perc}}$ and an adversarial loss $\mathcal{L}_{\text{adv}}$. The composite loss function is formulated as:

$$\mathcal{L}_{all}(I_{out}, I_{gt}) = \alpha_1\mathcal{L}_1 + \alpha_2\mathcal{L}_{\text{style}} + \alpha_3\mathcal{L}_{\text{perc}} + \alpha_4\mathcal{L}_{\text{adv}}, \tag{5.9}$$

where I_{out} and I_{gt} are the reconstructed image and ground truth, respectively. α_1=1, α_2=250, α_3=0.1, and α_4=0.001 are the weighting factors for each component.

5.3 Spatially-Enhanced Mamba Network

Given an image I with pixels $\hat{x}_1, \hat{x}_2, ..., \hat{x}_{N\times N}$ in $N \times N$ resolution. Image inpainting is a task for learning the mapping from the input masked image $I_{in} = \text{concat}[I \odot M, M]$ to the semantically accurate output image I_{out}, where M is the mask. The overall pipeline

of the proposed SEM-Net is illustrated in Fig. 5.3. Our framework comprises two key components to address the two identified challenges in a synergistic manner. The first component, a Snake Mamba Block (Sec. 5.3.1), aims at effectively preserving the continuity and consistency of pixel adjacency for pixel-level dependency learning during the linear recurrence in SSMs. The second component, a Spatially-Enhanced Feedforward Network (Sec. 5.3.2), is proposed to further complement the 2D spatial awareness of the 1D linear recurrent based SSMs.

Our SEM-Net adopts the encoder-decoder based U-Net architecture formed with four-stage SEM blocks to learn hierarchical multi-scale representation. Given a masked image $I_{in} \in \mathbb{R}^{H\times W\times 3}$, where $H \times W$ is spatial dimension and 3 denotes the RGB channels. SEM-Net first employs a 3×3 convolution to extract low-level feature embedding $\mathbf{h_0} \in \mathbb{R}^{H\times W\times C}$. Then, these features $\mathbf{h_0}$ pass through the four-scale encoder SEM blocks, which gradually decrease in spatial size while increasing in channel capacity, to generate latent features $\mathbf{h_l} \in \mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times 8C}$. Next, the decoder takes $\mathbf{h_l}$ to progressively reconstruct high-resolution representations. Every stage contains multiple SEM blocks, each SEM block has a pair of proposed Snake Mamba Block (SMB) and Spatially-Enhanced Feedforward Network (SEFN) for refining image representation learning while effectively capturing spatial LRDs. During this process, we use skip connections to link mirrored SEM blocks at the end of each stage and use a 1×1 convolution to half the channels after each connection, preserving the shared features learned by the encoder and then supporting the decoder. The cost-efficient pixel-unshuffle and pixel-shuffle operations [120] are employed to achieve feature downsampling and upsampling, respectively. In the final step, a convolutional layer projects the decoded features to the output.

5.3.1 Snake Mamba Block

In each snake mamba block (SMB), we propose a holistic framework to preserve continuity and ensure the comprehensiveness of pixel adjacency for pixel-level dependency learning during 1D linear recurrence in SSMs. This is achieved through two novel designs: the implicit Snake Bi-Directional Modelling (SBDM) and the explicit Position Enhancement Layer (PE layer).

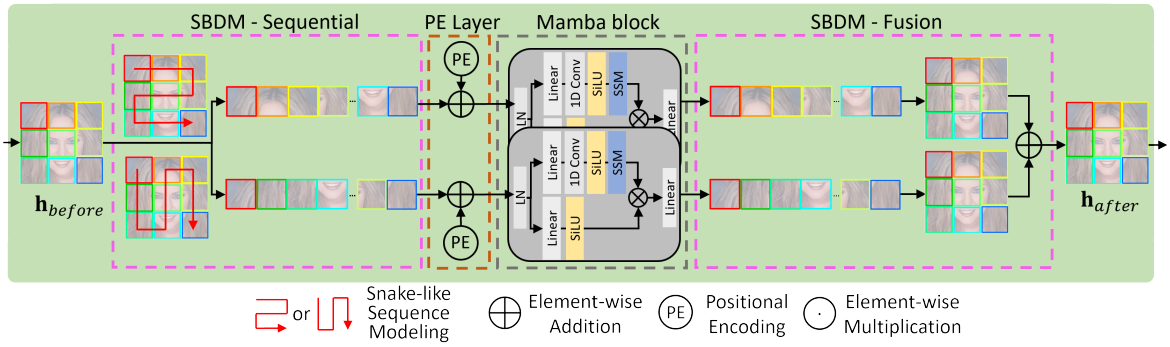

Figure 5.4: The architecture of proposed SMB. The input feature is modelled to sequences in two directions with snake-like traverses in SBDM-Sequential, enhancing the spatial awareness implicitly. Then, the PE layer explicitly enhances the long-range positional awareness through positional embeddings. The features after Mamba are restructured and aggregated by SBDM-Fusion to generate the output.

Snake Bi-Directional Modelling

Directly leveraging the predominant 1D linear SSMs by feeding the flattened spatial features is prone to an inevitable loss of pixel adjacency continuity and spatial information, resulting in a degradation in image representation learning. To alleviate this challenge, SBDM mainly contains two sequence modelling techniques: snake-like sequence modelling and bi-directional sequence modelling.

Snake-like Sequence Modelling

Snake-like sequence modelling aims to maintain the continuity in pixel adjacency when flattening spatial features across each channel from a shape of $H \times W$ to $1 \times HW$. This is crucial as we observe that the conventional flattening operation directly connects the end of one row to the start of the next, forcing SSMs to recognise recurrent connections between spatially distant pixels rather than adjacent ones, leading to a loss of pixel adjacency continuity and constrains the dependency-reasoning capacity. To address this issue, our snake-like sequence modelling ensures consistent connections among neighbouring pixels both within and across rows by reordering pixels and concatenating rows, illustrated by the red arrows in Fig. 5.4.

Specifically, given an input feature $\mathbf{h}_{in} \in \mathbb{R}^{H \times W \times C}$, where H is the number of rows (lines), W is the number of columns (pixels in a line), and C is the dimension for each pixel. $p_{i,j} \in \mathbb{R}^{1 \times 1 \times C}$ denotes the pixel value at the position of i-th row and j-th column.

Then, the horizontal snake-like sequence modelling process is represented as:

$$S_i = \begin{cases} [p_{i,0}, p_{i,1}, \ldots, p_{i,W-1}], & i = 0, 2, 4, \ldots, \\ [p_{i,W-1}, p_{i,W-2}, \ldots, p_{i,0}], & i = 1, 3, 5, \ldots, \end{cases} \tag{5.10}$$

$$S = \text{concat}[S_0, S_1, S_2, S_3, \ldots, S_{H-1}], \tag{5.11}$$

where the 1D sequence S maintains the pixel adjacency continuity by concatenating the sequences S_i for $i \in [0, H-1]$, each S_i represents the reordered pixel position in that row.

Bi-directional Sequence Modelling

To further complement the comprehensiveness of pixel adjacency and implicitly enhance spatial awareness, we propose a bi-directional sequence modelling involving two processes: SBDM-Sequential (SBDM-S) and SBDM-Fusion (SBDM-F). As shown in 5.4, SBDM-S simultaneously traverse pixels in a snake-like manner in two directions: horizontally and vertically across all pixels, enabling the SMB to generate sequences that capture discriminative dependencies. Specifically, in a snake-like manner, SBDM-S vertically traverses pixels to 1-D sequences $S = \text{concat}[S_0, S_1, \ldots, S_{H-1}]$, and horizontally traverses pixels to 1-D sequences $T = \text{concat}[S_0^\top, S_1^\top, \ldots, S_{W-1}^\top]$, where each $S_j^\top$ for $j \in [0, W-1]$ contains reordered pixels in that column. These two directions are designed since they are spatially complementary to each other and are computationally efficient in multi-directional traversals. After processing through Mamba, SBDM-F restructures the 1D sequences back to 2D via the inverse function of Eq. 5.11 and fuse them by element-wise aggregation to retain their spatial information, enriching the spatial awareness in image representation learning.

Position Enhancement Layer

To further explicitly complement the implicit approach of SBDM in enhancing spatial dependency reasoning, we propose a simple yet effective strategy of integrating 1D positional embeddings to enhance position awareness. Specifically, we incorporate the

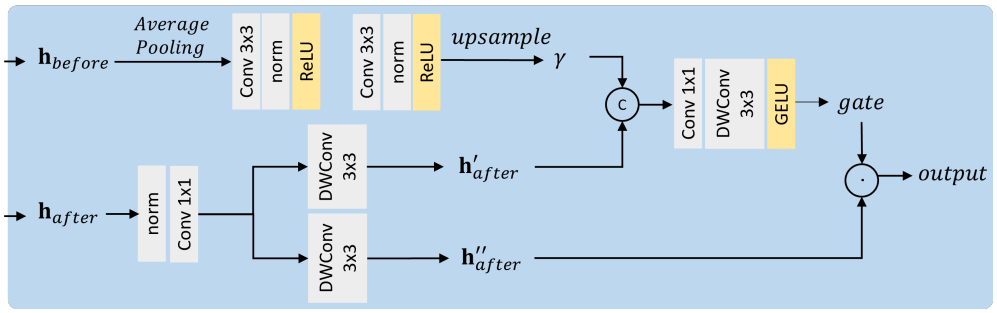

Figure 5.5: The architecture of proposed Spatially-Enhanced Feedforward Network (SEFN)

1D positional embeddings directly into 1D sequences in the position enhancement layer (PE layer) before processing with Mamba, assigning absolute positional information to each element within the sequences for providing the positional context and maintaining the pixel adjacency relationships. Formally, assume $S(n)$ for $n \in [0, N^2 - 1] \cap \mathbb{Z}$ is the element at positional coordinate n, $PE(n)$ is the corresponding cosine positional embedding [68]. Then, the elements in 1D sequence S with 1D positional embeddings $\bar{S}(n)$ are integrated by aggregation:

$$\bar{S}(n) = S(n) + PE(n), n = 0, 1, 2, 3, \ldots, N^2 - 1. \tag{5.12}$$

5.3.2 Spatially-Enhanced Feedforward Network

To complement local spatial information in regions spanning multiple rows and columns that are subject to inherent design limitations of SSMs, we propose a Spatially-Enhanced Feedforward Network (SEFN) for refining spatial awareness in image representation learning. The key idea of SEFN lies in leveraging spatial information extracted from the feature representations prior to the SEM block, subsequently applying it in a gating mechanism to inform the features post-SMB, thereby facilitating the integration of spatial awareness and LRDs learning to the entire SEM block.

Specifically, SEFN first snatches $\mathbf{h}_{before}$ and $\mathbf{h}_{after}$ at the entrance and exit of the Mamba block. Then, SEFN uses the average pooling to expand the receptive field, followed by two $\{Conv\text{-}LN\text{-}ReLU\}$ blocks to capture a broader spatial perception. The subsequent upsampling yields a spatial awareness indicator γ preserving spatial relationships from

$\mathbf{h}_{before}$. The gating mechanism starts from $\mathbf{h}_{after}$, which is divided into $\mathbf{h}'_{after}$ and $\mathbf{h}''_{after}$. The $\mathbf{h}'_{after}$ is informed by γ to form a '$gate$' via a linear transformation and a GELU non-linear activation. '$gate$' then modulates $\mathbf{h}''_{after}$ through a point-wise product, significantly enhancing the spatial awareness of $\mathbf{h}''_{after}$. The whole process is formulated as:

$$\mathbf{h}'_{after} = W_{d3}W_1 LN(\mathbf{h}_{after}), \tag{5.13}$$

$$\mathbf{h}''_{after} = W'_{d3}W'_1 LN(\mathbf{h}_{after}), \tag{5.14}$$

$$\gamma = Up(f(AveragePooling(\mathbf{h}_{before}))), \tag{5.15}$$

$$gate = GELU(W_{d3}W_1\gamma||\mathbf{h}'_{after}), \tag{5.16}$$

$$output = gate \odot \mathbf{h}''_{after}, \tag{5.17}$$

where W_1, W'_1 are 1×1 convolutions, W_{d3}, W'_{d3} are 3×3 depth-wise convolutions to reduce computational cost while refining features, LN is a layer normalization, f denotes two $\{Conv\text{-}LN\text{-}ReLU\}$ blocks, Up is upsampling.

5.3.3 Loss Function

To achieve superior inpainting outcomes, we optimize our SEM-Net with the loss combination of $\mathcal{L}_{total} = \lambda_1\mathcal{L}_1 + \lambda_2\mathcal{L}_{style} + \lambda_3\mathcal{L}_{perc} + \lambda_4\mathcal{L}_{adv}$, where λ_1 = 1, λ_2 = 250, λ_3 = 0.1, λ_4 = 0.001. $\mathcal{L}_1$ is the pixel-wise reconstruction loss, $\mathcal{L}_{style}$ is style loss, $\mathcal{L}_{perc}$ is the perceptual loss, and $\mathcal{L}_{adv}$ is the adversarial loss. We define the I_{gt} as the ground truth, I_{out} is the completed image, G is the SEM-Net and D is the discriminator. The formulation for each loss is shown below:

$$\mathcal{L}_1 = \mathbb{E}\left[\|\boldsymbol{I}_{out} - \boldsymbol{I}_{gt}\|_1\right], \tag{5.18}$$

$$\mathcal{L}_{perc} = \mathbb{E}\left[\sum_i \|\phi_i(\boldsymbol{I}_{out}) - \phi_i(\boldsymbol{I}_{gt})\|_1\right], \tag{5.19}$$

$$\mathcal{L}_{style} = \mathbb{E}\left[\sum_i \|(\psi_i(\boldsymbol{I}_{out}) - \psi_i(\boldsymbol{I}_{gt}))\|_1\right], \tag{5.20}$$

$$\mathcal{L}_{adv} = \min_G \max_D \mathbb{E}_{\boldsymbol{I}_{gt}}[\log D(\boldsymbol{I}_{gt})] + \mathbb{E}_{\boldsymbol{I}_{out}} \log[1 - D(\boldsymbol{I}_{out})], \tag{5.21}$$

where $\phi i(\cdot)$ indicates the activation map from the i-th pooling layer of VGG-16. $\psi_i(\cdot) = \phi_i(\cdot)^T \phi_i(\cdot)$ denotes the Gram matrix. The loss combination of $\mathcal{L}_{total} = \lambda_1 \mathcal{L}_1 + \lambda_2 \mathcal{L}_{style} + \lambda_3 \mathcal{L}_{perc} + \lambda_4 \mathcal{L}_{adv}$, where λ_1 = 1, λ_2 = 250, λ_3 = 0.1, λ_4 = 0.001.

5.4 Experiment Results

5.4.1 Datasets

We evaluate $M \times T$ and SEM on two diverse datasets, CelebA-HQ [4] and Places2-standard [5], to ensure a comprehensive comparison. CelebA-HQ is a dataset consisting of high-quality human face images. For CelebA-HQ, we train our model on the first 28000 images and reserve the remaining 2000 for testing. Places2 comprises a wide range of natural and indoor scene images. For Places2, we employ the standard training set, which includes 1.8 million images, and test on its validation set of 30000 images. We follow [2, 58, 127] to conduct all experiments with the widely used irregular mask [15] in three mask ratios.

5.4.2 Implementation Details

Except where specified differently, all experiments are conducted on a single Nvidia A100 GPU. We adopt the following set of parameters for our experiments: a batch size of 6 and a patch size of 256×256. We use Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$) optimizer with learning rate $= 1e^{-4}$.

5.4.3 Baselines and Metrics

We choose the following baselines for inpainting comparison: CNN-based methods with DeepFill v1 [18], DeepFill v2 [16], CTSDG [2] and *MISF* [127], WaveFill [132] and *LAMA* [7]; Transformer-based methods with MAT [8] and *CMT* [142]; Expensive diffusion models [62, 134]. *Italic* denotes the SOTA methods.

5.4.4 Comparison with State of the Art

Quantitative Comparison

For a fair comparison, we employ the officially released models and test them with the same test sets and masks. As shown in table 5.6, $M \times T$ and SEM outperforms in all metrics across different mask ratios. Especially on CelebA-HQ, at the increasing mask ratios, $M \times T$ improves PSNR by 2.0%, 2.3% and 1.8% respectively, and decreases LPIPS by 12.3%, 11.6% and 10.5% respectively. As a more advanced model, SEM-Net achieves (i) a substantial gain of 0.7743 (2.15%↑), 0.7187 (2.55%↑), and 0.5386 (2.25%↑) PSNR; (ii) and a significant reduction of 0.0192 (5.14%↓), 0.074 (5.72%↓), and 0.1636 (5.84%↓) L1, over the second methods [7, 142] on three mask ratios, respectively. The improvements in these two specific metrics indicate a significant boost in the pixel-wise reconstruction accuracy. In addition, the LPIPS of SEM-Net appreciably drops than the second-best method [142] in CelebA-HQ dataset by 0.0035 (13.41%↓), 0.0101 (12.36%↓), and 0.0199 (12.70%↓) on three mask ratios, respectively. It demonstrates a significant improvement in high-quality image inpainting with lower perceptual differences.

Qualitative Comparison

The qualitative results of $M \times T$ and SEM-Net are shown in Fig. 5.9 and Fig. 5.9. Each sample is the inpainted result where the mask ratio exceeds 40%, to more intuitively demonstrate the advantages of our methods in handling challenging cases.

In Fig. 5.9, for human face samples, $M \times T$ maintains consistency from the visible regions to the missing regions, such as effectively reconstructing elements like a missing hat. Additionally, $M \times T$ renders features like eyes with improved fine-grained details, showcasing its strong capability in learning complex representations. In the Places2 dataset, $M \times T$ effectively captures the spatial layouts in indoor environments and excels at maintaining the architectural integrity of the road surfaces. Such examples highlight our $M \times T$ has superior spatial perceptions.

In Fig. 5.9, for facial inpainting, generating one eye in masked regions (masked eye) based on another eye in visible regions (visible eye) is more challenging than directly generating two eyes, because it requires the model to have a solid ability to capture long-

Table 5.1: Ablation studies for $M \times T$ of each component in 40% - 60 %. MB is the Mamba Block with positonal embedding. SRSA is the Spatial Reduced Self-Attention. GDFN is the feed-forward network in [10]. CBFN is the Context Broadcasting Feed-forward Network. Our $M \times T$ corresponds to configuration (e).

	Components				40%-60%				
	MB	SRSA	GDFN [10]	CBFN	PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓
(a)					21.6134	0.7308	4.1254	8.1732	0.2464
(b)	✓		✓		21.8573	0.7614	3.8649	8.0315	0.2197
(c)		✓	✓		21.8914	0.7682	3.8587	7.9974	0.2157
(d)	✓	✓	✓		22.1377	0.7687	3.7679	7.9910	0.2067
(e)	✓	✓		✓	**22.1704**	**0.7699**	**3.6337**	**7.9905**	**0.2053**

Table 5.2: Ablation studies of each component in $40\% - 60\%$ mask ratio. Refer to supplementary material for all mask ratios. Our SEM-Net corresponds to configuration (g)

Net	Components					40%-60%				
	MB	FN [10]	SEFN	SBDM	PE	PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓
(a)						21.6134	0.7308	4.1254	8.1732	0.2464
(b)	✓	✓				21.7828	0.7587	3.9117	8.0742	0.2227
(c)	✓		✓			22.0510	0.7682	3.7649	7.9871	0.2132
(d)	✓	✓		✓		21.9064	0.7653	3.7679	8.0214	0.2102
(e)	✓		✓	✓		22.0926	0.7692	3.7634	7.9174	0.2091
(f)	✓	✓		✓	✓	22.1776	0.7708	3.6747	7.9125	0.2095
(g)	✓		✓	✓	✓	**22.1780**	**0.7725**	**3.6274**	**7.8915**	**0.2038**

range dependency to learn from another eye. Compared with current state-of-the-art techniques, SEM-Net successfully transfers features in the visible eye to the masked eye, including eyeball colour and shape, while preserving finer-grained features. In Places2, SEM-Net generates fewer artefacts and more coherent structures, such as the white lines in the road and the edges of coloured cardboard, ensuring the contextual consistency of the texture and structure of the image.

5.4.5 Ablation Study

In our comprehensive ablation study conducted on CelebA-HQ, we incrementally enhance the baseline U-Net shape model, observing significant performance improvements with the integration of each component. Results are shown in the Tab. 5.1 (for $M \times T$) and Tab. 5.2 (for SEM-Net). We followed [143] to build $M \times T$ and SEM-Net with a halved parameter for an efficient evaluation. All ablation experiments are trained for 30,000

Input a b c d e f g GT

Figure 5.6: The qualitative visualisation of ablation studies. Zoom in for the details.

iterations.

M × T

For $M \times T$, in the Tab. 5.1, the addition of the Mamba Block in configuration (b) and the addition of self-attention in configuration (c) both demonstrated improvement across all evaluated metrics compared to the baseline (a). Notably, self-attention improves in SSIM, suggesting its superior capability in capturing spatial interactions. Mamba showcases the superior in capturing pixel-level interactions, demonstrated by the better PSNR and L1 values. The simultaneous use of Mamba and self-attention in configuration (d) lead to further improvements, indicating that these components effectively complement each other and contribute to a robust model. Configuration (e) is our final model, where we optimise GDFN to CBFN. The overall metrics are further improved.

SEM-Net

For SEM-Net, Tab. 5.2 and Fig. 5.6 present the improvement of each component quantitatively and qualitatively. Based on the U-Net shape baseline (Tab. 5.2a), integrating the Mamba Block (MB) and Feedforward Network (FN) [10] (Tab. 5.2b) results in noticeable improvements across all metrics. Fig. 5.6d→b and Fig. 5.6e→c shows that degrading SBDM, model struggle in capture the relations of vertically adjacent pixels, resulting in artefacts between the left eyebrow and left eye. Fig. 5.6b→c, Fig. 5.6d→e and Fig. 5.6f→g revealed the effect of SEFN by resulting sharper jaw and less artefacts, demonstrated by the improvement in SSIM score. Tab. 5.2d→f and Tab. 5.2e→g showcase that introducing positional embedding significantly improves L1 and PSNR in larger masks, which is evidenced by the clearer texture at the mouth and eye.

Comparing SMB with Transformer Blocks. We evaluate the effectiveness of our proposed SMB in image representation learning by comparing it with two typical and widely used transformer blocks that claimed to have strong capability in capturing

Table 5.3: Comparison between our proposed SMB with transformer-based methods in $40\% - 60\%$ mask ratio. Refer to supplementary material for all mask ratios.

Input Resolution	Model	40%-60%				
		PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓
256*256	CSA [10]	21.5362	0.7543	4.0471	8.1652	0.2326
	SSA [69]	Out of memory				
	SMB	**22.1776**	**0.7708**	**3.6747**	**7.9125**	**0.2095**
64*64	SSA [69]	20.1655	0.7265	5.2256	5.5547	0.1702
	SMB	**20.1716**	**0.7352**	**5.1332**	**5.3158**	**0.1617**

LRDs: channel-wise self-attention [10] and Spatial-wise self-attention (SSA) [69]. For fair comparisons, all models use vanilla feedforward networks [10] instead of our novel SEFN, with only differences between SMB, CSA and SSA. From Table. 5.3, we observe that our SMB consistently outperforms two distinct transformer blocks across all metrics in all mask ratios. In addition, our SMB is shown to be efficient enough to process original resolution (256×256) images while SSA can only be trained on the degraded 64×64 images with a single A100 due to its significant computational cost. Furthermore, compared with the diffusion-based models [62, 134] with a very long inference time [136], our model has better performance while the inference time is still in milliseconds, which is suitable for real-time scenarios (shown in Tab. 5.6).

Comparison of Sequential Modelling. We provide the illustration in Fig. 5.7 to showcase the difference between the proposed Snake Bi-Directional Modelling and simple sequential modelling. Tab. 5.4 showcases the quantitative results on CelebA-HQ in $40\% - 60\%$ mask ratio to compare with other optimisations of the SSM-based sequential modelling [9, 79, 144], demonstrating our superiority across all metrics.

Table 5.4: Comparison of different SSM-based modelling.

Mask	PSNR ↑	SSIM↑	L_1↓	FID↓	LPIPS↓
2-D SSM [144]	24.1153	0.7877	3.0950	5.8556	0.1672
VMamba [79]	24.1409	0.8031	2.9168	5.9508	0.1739
U-Mamba [9]	24.2077	0.8119	2.7440	5.6034	0.1466
Ours	**24.4805**	**0.8240**	**2.6389**	**5.5972**	**0.1368**

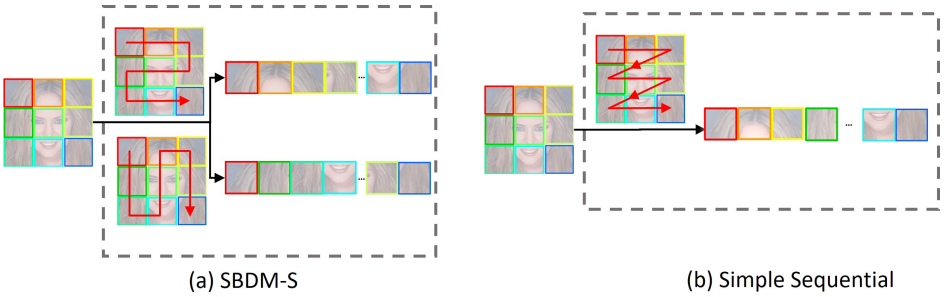

Figure 5.7: Comparison between (a) the proposed Snake Bi-Directional Modelling - Sequential (SBDM-S) and (b) the simple sequential approach. Our SBDM implicitly models bi-directional positional context by horizontally and vertically scanning the tokens, while the snake-shape design preserves the relations within adjacent tokens.

Ablation study for Snake Bi-Directional Modelling (SBDM) To further evaluate each decision in designing the proposed Snake Bi-Directional Modelling (SBDM) module, we conduct the experiment by ablating each component. As shown in Tab. 5.7. $Bi\text{-}D$ means horizontal and vertical direction modelling. The model without $Bi\text{-}D$ only contains single horizontal direction modelling. $Snake$ denotes the Snake-like Sequence Modelling. The model without $Snake$ contains simple sequential modelling. We notice that the proposed snake-like design and bidirectional design overall improve the performance. An interesting observation is that at the largest mask ratio, individually integrating each of the two designs degrades the FID. But the FID at the largest mask ratio gets better when both snake-like design and bidirectional design are used together. This may indicate that when the damaged region is large and challenging, both complementary methods need to be used simultaneously to achieve better inpainting results without fully convergent training.

5.4.6 Generalization Ability

Unseen High Resolution Images.

We examine the scalability and generalisability of $M \times T$ and SEM-Net trained on 256×256 Places2 images in processing unseen images of higher resolution, since these abilities are crucial for practical applications where image resolutions can significantly vary. Fig. 5.11 showcases examples of unseen real-world high-resolution applications. While [31] performs similarly with larger masks, its upsampling strategy causes narrow

Table 5.5: Performance in generalising to image motion deblurring task. Our SEM-Net is trained only on the GoPro dataset [11] and directly applied to the HIDE [12].

Method	GoPro [11]		HIDE [12]	
	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
DeblurGAN-v2 [145]	29.55	0.934	26.61	0.875
Shen *et al.* [12]	-	-	28.89	0.930
Gao *et al.* [146]	30.90	0.935	29.11	0.913
DBGAN [147]	31.10	0.942	28.94	0.915
MT-RNN [148]	31.15	0.945	29.15	0.918
DMPHN [149]	31.20	0.940	29.09	0.924
Suin *et al.* [150]	31.85	0.948	29.98	0.930
SPAIR [151]	32.06	0.953	30.29	0.931
MIMO-UNet+ [152]	32.45	0.957	29.99	0.930
IPT [153]	32.52	-	-	-
MPRNet [154]	32.66	0.959	30.96	0.939
HINet [155]	32.71	0.959	30.32	0.932
Restormer [10]	32.92	<u>0.961</u>	**31.22**	**0.942**
Stripformer [156]	<u>33.08</u>	**0.962**	31.03	0.940
Ours	**33.11**	**0.962**	<u>31.12</u>	<u>0.941</u>

mask drifting, leading to artefacts. Our $M \times T$ and SEM-Net captures finer details without artefacts by modelling at the pixel level to offer the community a better, more resource-efficient solution for processing large-resolution images. More examples with different resolutions are included in the supplementary.

Low-level Vision Tasks

To further evaluate the capability of representation learning and generalisation ability, we directly apply SEN-Net to another low-level vision task, image motion deblurring, through the necessary learning of the residual between clear images and blurred images without any other task-specific modifications.

The image deblur task is formulated as $I_{out} = I_{in} + SEM - Net(I_{in})$, where I_{in} is the blurred image, I_{out} is the clear image. To train our deblurring model, we follow [143] to use a joint loss consisting of a reconstruction loss and a frequency loss. The formulation for each loss is shown below:

$$\mathcal{L}_{rec} = \mathbb{E}\left[\left\|\boldsymbol{I}_{out} - \boldsymbol{I}_{gt}\right\|_1\right], \tag{5.22}$$

$$\mathcal{L}_{frequency} = \mathbb{E}\left[\|F(\boldsymbol{I}_{out}) - F(\boldsymbol{I}_{gt})\|_1\right], \tag{5.23}$$

where $F(\cdot)$ is the Fast Fourier transform. The total loss for image deblurring is $L_{total} = L_{rec} + 0.1 \times L_{frequency}$.

Tab. 5.5 shows that SEM-Net overall outperforms the restoration models on two synthetic benchmark datasets GoPro [11] and HIDE [12]. Especially on GoPro, SEM-Net improves PSNR by 0.19 compared to the strong restoration baseline model Restormer [10]. Notably, our SEM is trained on GoPro and directly applied to HIDE, without progressive learning [10] or Test-time Local Converter [157] such external optimization, showcasing strong generalization ability. Refer to supplementary materials for qualitative results.

5.5 Summary

In this paper, we introduce $M \times T$ and SEM-Net for image inpainting designed to reconstruct high-quality images with fine-grained details and spatial coherence. For the $M \times T$, the proposed Hybrid Module effectively combines transformer and Mamba, leveraging the capacity of Mamba for capturing pixel-wise long-range interaction along with the spatial perception provided by the transformer. This integration enables $M \times T$ to maintain linear computational complexity, which is particularly advantageous for handling high-resolution images. For the SEM-Net, it demonstrates strong capabilities in capturing LRDs and addresses the challenge of lack of spatial awareness in SSMs. In the SEM-Net, we propose two key designs, SMB and SEFN, for improved image representation learning.

The two proposed models outperform state-of-the-art approaches on two image inpainting datasets, especially on CelebA-HQ. This could be due to dataset characteristics, CelebA-HQ's structured, human-centric images benefit more from our model's ability to capture long-range dependencies and spatial awareness, shown in quantitative results. Also, we showcases strong generalisability to higher-resolution images and another low-level visual task, image deblurring.

In this chapter, we contrast three complementary families. Vanilla SSMs provide linear-time sequence operators that propagate long-range context with minimal memory,

making them effective as lightweight backbones when efficiency is paramount. SBDM (Snake Bi-Directional Modelling), introduced in this work, augments sequence modelling with a snake-like raster and bi-directional passes, enabling richer horizontal–vertical context aggregation than plain SSMs while retaining favourable computational complexity. This is advantageous when coherent structure must be propagated across irregular masks without resorting to full attention. SEM-Net is our geometry-aware, multi-task inpainting framework that couples reduced attention/SCAL with a landmark branch. It injects explicit anatomical priors and is trained on open data with clinical images reserved for validation, emphasising privacy and anatomical plausibility. Practically, SSMs are preferred for fast global context propagation under tight budgets, SBDM for stronger bidirectional context without attention overhead, and SEM-Net for privacy-aware, structure-constrained facial repair.

Table 5.6: Quantitative comparison with the state-of-the-arts on CelebA-HQ (top), and Places2 (bottom). **Bold** and underline are the best and the second-best respectively. Number of parameters (Param.) and inference time (Inf.) are based on the inpainting evaluation conducted on 256 × 256 images. ^{C}, ^{T} and ^{D} indicate CNN-based, Transformer-based and Diffusion-based methods, respectively.

CelebA-HQ	Param.$\times 10^6$	0.01%-20%					20%-40%					40%-60%				
Method	/ Inf. Time	PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓	PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓	PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓
DeepFill v1 [18]^{C}	3 / 7ms	34.2507	0.9047	1.7433	2.2141	0.1184	26.8796	0.8271	2.3117	9.4047	0.1329	21.4721	0.7492	4.6285	15.4731	0.2521
DeepFill v2 [16]^{C}	4 / 10ms	34.4735	0.9533	0.5211	1.4374	0.0429	27.3298	0.8657	1.7687	5.5498	0.1064	22.6937	0.7962	3.2721	8.8673	0.1739
WaveFill [132]^{C}	49 / 70ms	31.4695	0.9290	1.3228	6.0638	0.0802	27.1073	0.8668	2.1159	8.3804	0.1231	23.3569	0.7817	3.5617	13.0849	0.1917
RePaint [134]^{D}	552 / 250000ms	-	-	-	-	-	-	-	-	-	-	21.8321	0.7791	3.9427	8.9637	0.1943
LaMa [7]^{C}	51 / 25ms	35.5656	0.9685	0.4029	1.4309	0.0319	28.0348	0.8983	1.3722	4.4295	0.0903	23.9419	0.8003	2.8646	8.4538	0.1620
MISF [127]^{C}	26 / 10 ms	35.3591	0.9647	0.4957	1.2759	0.0287	27.4529	0.8899	2.0118	4.7299	0.1176	23.4476	0.7970	3.4167	8.1877	0.1868
MAT [8]^{T}	62 / 70ms	35.5466	0.9689	0.3961	1.2428	0.0268	27.6684	0.8957	1.3852	3.4677	0.0832	23.3371	0.7964	2.9816	5.7284	0.1575
CMT [142]^{C}	143 / 60ms	36.0336	0.9749	0.3739	1.1171	0.0261	28.1589	0.9109	1.2938	3.3915	0.0817	23.8183	0.8141	2.8025	5.6382	0.1567
$M \times T$ (Ours)	180 / 110ms	36.7394	0.9737	0.3614	1.1142	0.0229	28.8098	0.9112	1.2413	3.3890	0.0722	24.3784	0.8220	2.6739	5.6041	0.1402
Ours	163 / 240ms	**36.8079**	**0.9774**	**0.3547**	**1.1070**	**0.0226**	**28.8776**	**0.9192**	**1.2198**	**3.3878**	**0.0716**	**24.4805**	**0.8240**	**2.6389**	**5.5972**	**0.1368**

Places2	Param.$\times 10^6$	0.01%-20%					20%-40%					40%-60%				
Method	/ Inf. Time	PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓	PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓	PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓
DeepFill v1 [18]^{C}	3 / 7ms	30.2958	0.9532	0.6953	26.3275	0.0497	24.2983	0.8426	2.4927	31.4296	0.1472	19.3751	0.6473	5.2092	46.4936	0.3145
DeepFill v2 [16]^{C}	4 / 10ms	31.4725	0.9558	0.6632	23.6854	0.0446	24.7247	0.8572	2.2453	27.3259	0.1362	19.7563	0.6742	4.9284	36.5458	0.2891
CTSDG [2]^{C}	52 / 20ms	32.1110	0.9565	0.6216	24.9852	0.0458	24.6502	0.8536	2.1210	29.2158	0.1429	20.2962	0.7012	4.6870	37.4251	0.2712
WaveFill [132]^{C}	49 / 70ms	29.8598	0.9468	0.9008	30.4259	0.0519	23.9875	0.8395	2.5329	39.8519	0.1365	18.4017	0.6130	7.1015	56.7527	0.3395
LDM [62]^{D}	387 / 6000 ms	-	-	-	-	-	-	-	-	-	-	19.6476	0.7052	4.6895	27.3619	0.2675
Stable Diffusion^{D*}	860 / 880 ms	-	-	-	-	-	-	-	-	-	-	19.4812	0.7185	4.5729	27.8830	0.2416
MISF [127]^{C}	26 / 10ms	32.9873	0.9615	0.5931	21.7526	0.0357	25.3843	0.8681	1.9460	30.5499	0.1183	20.7260	0.7187	4.4383	44.4778	0.2278
LaMa [7]^{C}	51 / 25ms	32.4660	0.9584	0.5969	14.7288	0.0354	25.0921	0.8635	2.0048	22.9381	0.1079	20.6796	0.7245	4.4060	25.9436	0.2124
CMT [142]^{T}	143 / 60ms	32.5765	0.9624	0.5915	22.1841	0.0364	24.9765	0.8666	2.0277	32.0184	0.1184	20.4888	0.7111	4.5484	35.1688	0.2378
$M \times T$ (Ours)	180 / 110ms	32.9940	0.9672	0.5950	15.3980	0.0334	25.3278	**0.8756**	1.9404	23.7109	0.1106	20.7022	**0.7319**	4.3379	26.9155	0.2372
SEM (Ours)	163 / 240ms	**33.0106**	**0.9631**	**0.5902**	**14.5163**	**0.0328**	**25.4159**	0.8736	**1.9275**	**22.7814**	**0.1054**	**20.8265**	0.7279	4.3614	**25.7049**	**0.2120**

*: The officially released Stable Diffusion inpainting model pretrained on high-quality LAION-Aesthetics V2 5+ dataset.

Table 5.7: Ablation study of each component in SBDM trained on CelebA-HQ [4].

Net	Components				0.01%-20%				20%-40%				40%-60%			
	MB	Bi-D	Snake	PE	SEFN	PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓	PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓	PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓
(a)	✓			✓	✓	34.1114	0.9624	0.5046	1.4134	0.0418	26.3305	0.8769	1.7231	4.3533	0.1186	22.0760	0.7688	3.7643	7.9868	0.2125
(b)	✓	✓		✓	✓	34.1428	0.9625	0.5016	1.3560	0.0416	26.4725	0.8802	1.7078	4.2751	0.1178	22.1351	0.7715	3.6742	8.0395	0.2078
(c)	✓		✓	✓	✓	34.1172	0.9624	0.5040	1.3564	0.0419	26.4619	0.8793	1.7105	4.2830	0.1185	22.1720	0.7692	3.6890	8.0372	0.2108
(e)	✓	✓	✓	✓	✓	**34.1437**	**0.9627**	**0.4986**	**1.3548**	**0.0403**	**26.4728**	**0.8808**	**1.6947**	**4.2718**	**0.1145**	**22.1780**	**0.7725**	**3.6274**	**7.8915**	**0.2038**

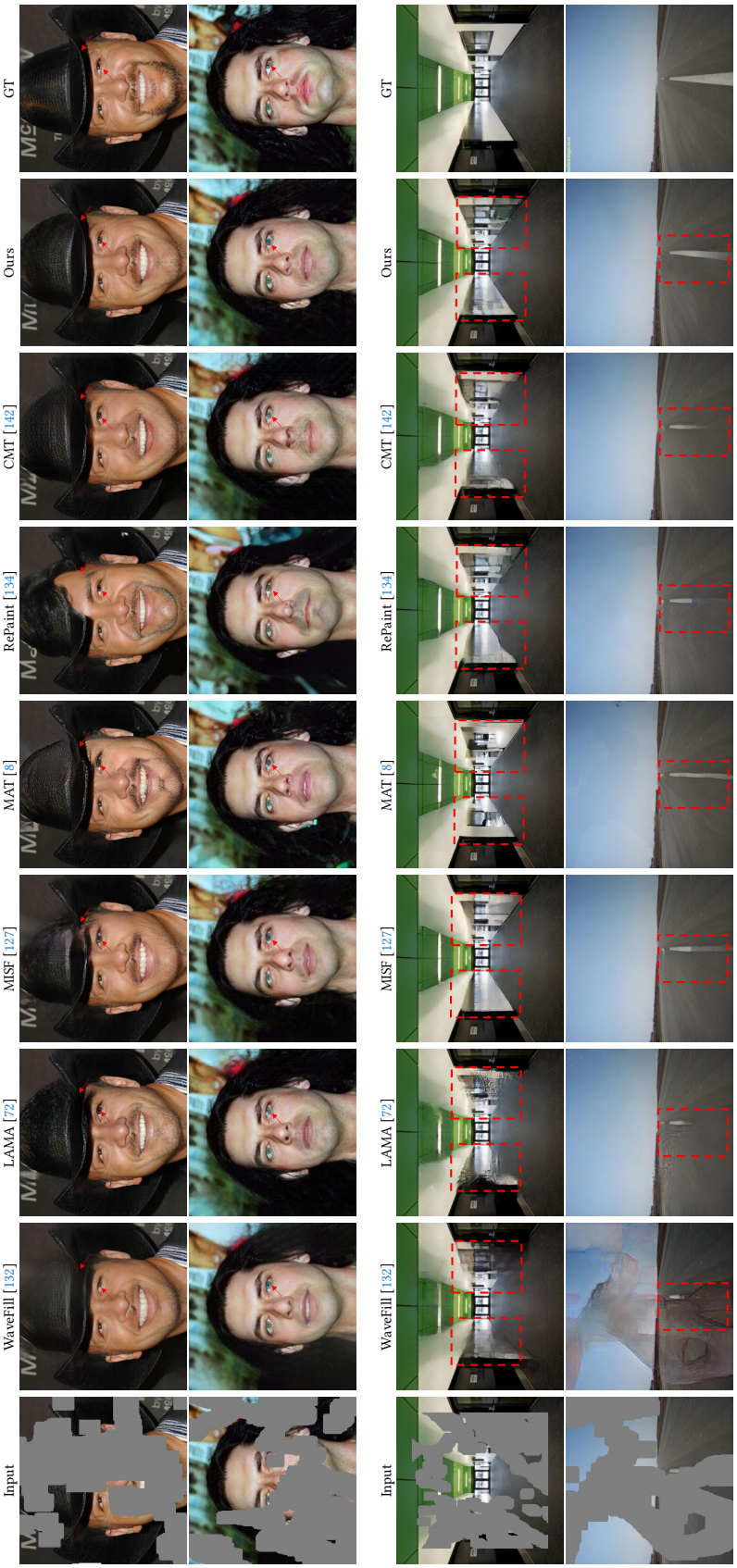

Figure 5.8: Visual comparisons at (256×256) resolution against the state-of-the-art methods on CelebA-HQ [4] (first two rows) and Places2 [5] (last two rows).

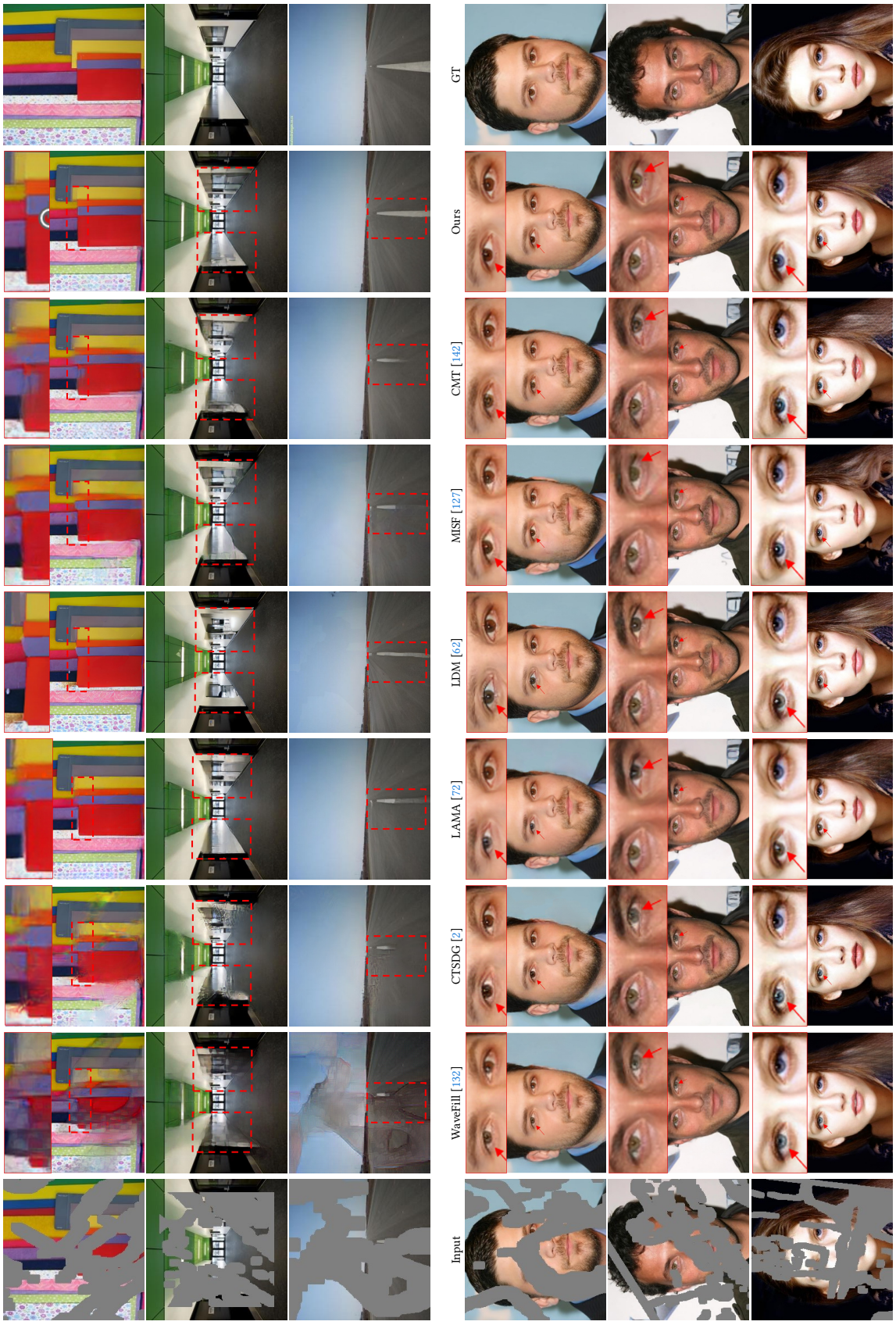

Figure 5.9: Comparisons with visualisations (256×256) showing that our results are more coherent in structure and sharper in texture and semantic details. The top three rows are from Places2 [5] and the bottom three rows are from CelebA-HQ [4].

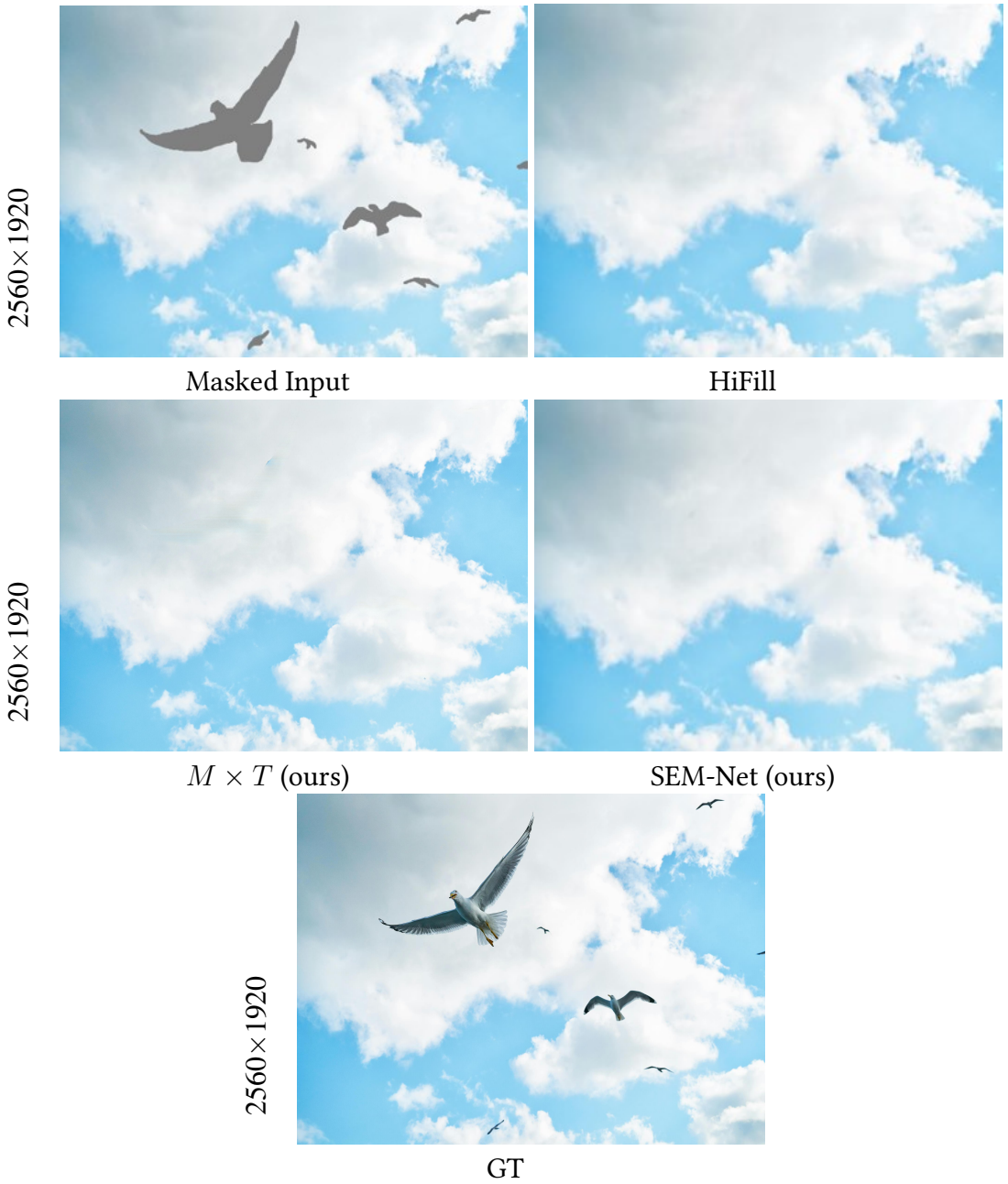

Figure 5.10: Examples of generalisation to real-world high-resolution images of 2560×1920.

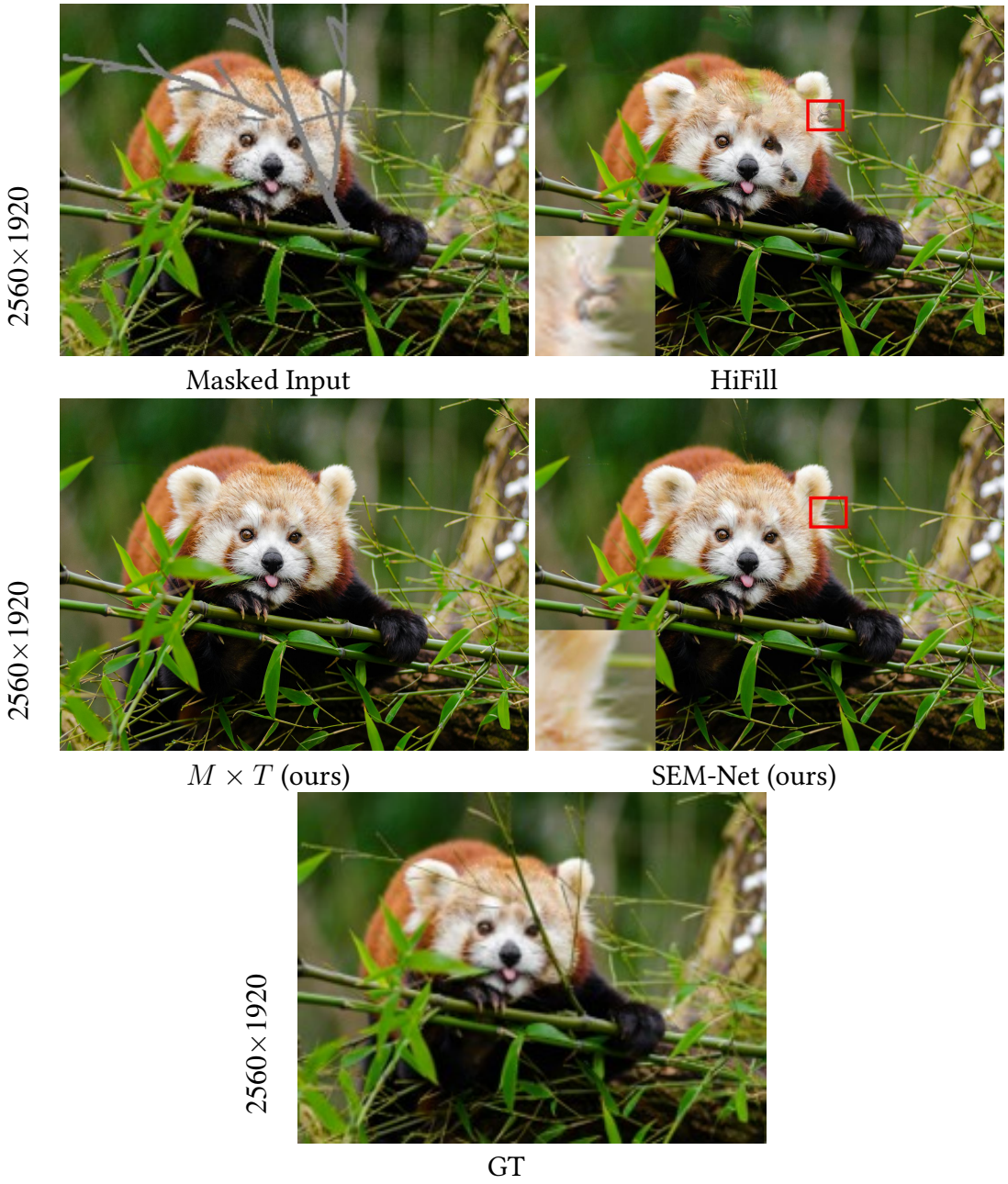

Figure 5.11: Examples of generalisation to real-world high-resolution images of 2560×1920.

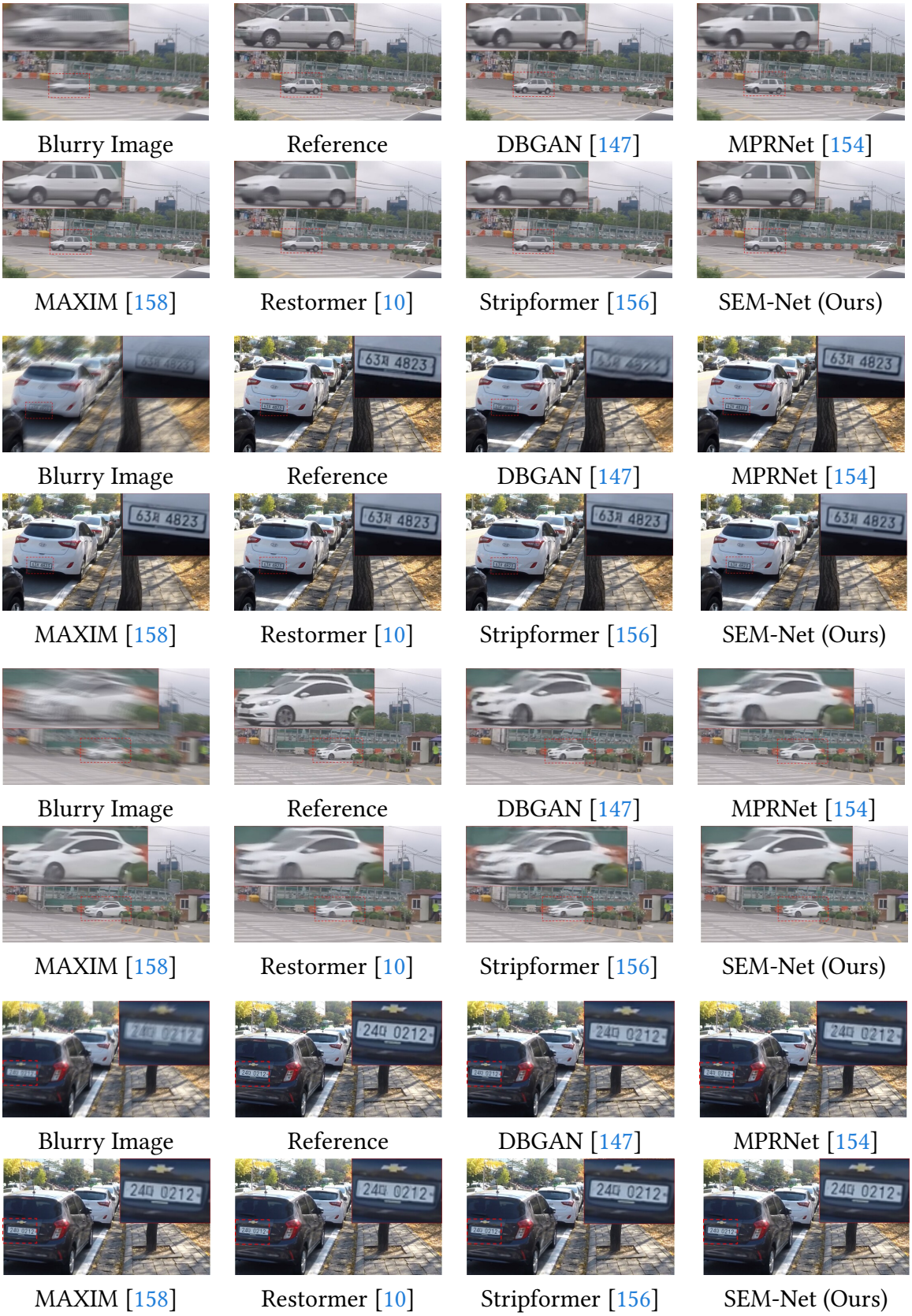

Figure 5.12: Image motion deblurring comparisons on GoPro [11]. Our method generates sharper results with higher visual fidelity.

CHAPTER 6

Conclusion

This thesis presents a comprehensive exploration of advanced techniques for image inpainting, focusing on addressing the critical challenges of "effectively using insufficient information" (Chapter 3) and "capturing long-range dependencies" (Chapter 4). By introducing novel models and methodologies, this work enchances the reconstruction quality of corrupted images for various scenarios with irregular or large missing regions. Additionally, we adopt the image inpainting technique in real-world applications for non-cleft lip facial image generation (Chapter 5), demonstrating its potential to address medical challenges and improve patient outcomes.

Chapter 3 demonstrates a single-stage multi-task framework that shares an encoder between an image-generation branch and a landmark-prediction branch. Fusing predicted landmarks back into generation yields anatomically plausible non-cleft lips from cleft-lip photographs while training exclusively on open data and reserving clinical images for validation, thereby reducing leakage risk. Expert surgeons' assessments and CelebA experiments support its feasibility and superiority over baselines. Chapter 4 introduces HINT, which preserves visible information with a mask-aware pixel-shuffle downsampling (MPD) and models long-range dependencies with a Spatially-activated Channel Attention Layer (SCAL) embedded in a "Sandwich" (FFN–SCAL–FFN) trans-

former block. Ablations and benchmarks on CelebA/CelebA-HQ/Places2/Dunhuang show consistent state-of-the-art performance. Chapter 5 proposes M×T—a hybrid of Mamba and Transformer for dual-level (pixel- and patch-wise) interaction at near-linear cost—and SEM-Net, which adds Snake Bi-Directional Modelling and a Spatially-Enhanced FFN to restore spatial awareness in state-space models. Both outperform prior art, scale to high resolutions, and generalise to image deblurring.

6.1 Review of Contributions

In Chapter 3, we develop and adopt an image inpainting technique to the real-world application of generating a non-cleft lip image from a baby with a cleft lip, focusing on protecting patient privacy. Facial inpainting tasks are evaluated based on the semantic plausibility of facial structure and image quality. Existing methods, such as EdgeConnect [1] and Lafin [3], rely on multi-stage processes that introduce redundancies and dependency on the first stage's accuracy. To address these limitations, we propose a single-stage, end-to-end, multi-task framework that integrates adaptive feature fusion and landmark prediction. This approach enhances parameter sharing, leverages masked images and partial inpainted features, and produces precise geometric indicators (landmark points) for reconstructing facial attributes. Furthermore, patient privacy is preserved as no sensitive data is used for training.

In Chapter 4, we observe that, effective image inpainting requires addressing the challenges of modelling valid information in visible regions while minimising information loss. Existing methods often struggle with information degradation due to convolutional down-sampling or the computational inefficiencies of spatial self-attention. To overcome these issues, we propose High-quality INpainting Transformer (HINT), involving a tailor-made pixel-shuffle down-sampling module to preserve data consistency and reduce information loss. Additionally, HINT employs the Spatially-activated Channel Attention Layer (SCAL) to balance spatial and channel-level information, maintaining spatial awareness while reducing computational complexity. A novel "Sandwich" structure integrates SCAL with feed-forward networks for enhanced efficiency and effectiveness.

In Chapter 5, based on the Mamba's capability of handling long-sequence modelling,

we propose two novel approaches, $M \times T$ and SEM-Net. $M \times T$ combines the strengths of transformers and Mamba for dual-level interaction learning. SEM-Net introduces the Snake Bi-Directional Modelling module (SBDM) for spatial consistency and the Spatially-Enhanced Feedforward Network (SEFN) for local refinement. These innovations address spatial awareness and adjacency continuity while maintaining computational efficiency, making it feasible to handle the large-resolution image inpainting task.

Together, these contributions reflect a holistic advancement in image inpainting, from addressing sensitive real-world needs in medical imagery to innovating core architectural components for scalable, high-fidelity reconstruction. These works bridge the gap between practical utility and theoretical innovation, laying a solid foundation for future research in both applied and general-purpose inpainting.

6.2 Limitation

While this thesis contributes new insights into medical image inpainting, transformer-based frameworks, and Mamba-augmented architectures, several limitations remain that define the boundaries of the presented work.

Data availability and diversity. The proposed cleft lip inpainting framework is trained and validated on relatively small, domain-specific datasets. Although expert evaluations confirm its feasibility, the limited diversity of facial appearances, age groups, and clinical variations restricts generalisability. Similarly, the benchmark datasets used in subsequent chapters (CelebA-HQ, Places2, Dunhuang) do not fully represent the complexities of unconstrained real-world imagery.

Scalability and computational efficiency. Our claims are supported by evidence in earlier chapters: (i) HINT (Chapter 4) attains state-of-the-art accuracy across CelebA-HQ, CelebA, Places2, and Dunhuang, yet its inference remains costlier than some CNN baselines and still significant versus diffusion variants at high resolution (see the diffusion/runtime comparison in Tab. 4.2), so scaling to very large images is non-trivial. (ii) M×T (Chapter 5) combines Mamba with a transformer and preserves (near-)linear complexity, which improves scalability to high-resolution settings, though the hybrid design adds architectural complexity. (iii) SEM-Net (Chapter 5) captures long-range dependen-

cies with linear-time state-space modelling and demonstrates cross-modality transfer by directly applying the trained model to image deblurring on GoPro/HIDE—outperforming strong restoration baselines without task-specific redesign—yet the full U-Net-style stack can still challenge edge deployment.

Evaluation metrics. As shown in Tab. 3.3, Tab. 4.1, Tab. 4.11, and Tab. 5.6, standard quantitative metrics (PSNR, SSIM, LPIPS, FID) capture fidelity and perceptual realism but remain imperfect proxies for human judgement, particularly in sensitive clinical scenarios. The absence of user-centric evaluation frameworks constrains the ability to measure true downstream impact.

Generalisability across domains. While the methods demonstrate strong results on facial and natural scene imagery, their performance in specialised areas such as medical radiology, satellite imaging, or scientific visualisation remains unexplored. Adapting inpainting to such domains requires rethinking assumptions about texture, geometry, and semantics.

Ethical and interpretability concerns. Applying inpainting to human faces, especially in clinical settings, raises important ethical questions regarding bias, fairness, and the interpretability of generated outputs. These issues are not comprehensively addressed within the scope of this thesis but remain critical to responsible deployment.

6.3 Future Research Directions

Looking ahead, several promising research directions can extend the contributions of this thesis and shape the broader field of image inpainting.

6.3.1 Towards richer and more representative datasets.

Expanding datasets for both clinical and general inpainting tasks is a pressing need. Collecting large-scale, demographically balanced medical image datasets (e.g., including varied cleft lip cases across age, ethnicity, and severity) will improve robustness. For general-purpose inpainting, curating multi-domain datasets (combining natural images, scientific data, and artwork) will encourage more versatile models.

6.3.2 Scalable and efficient architectures.

The increasing resolution and complexity of modern imagery demands architectures that scale linearly (or sub-linearly) with image size. Future work could focus on lightweight variants of HINT, SEM-Net, and M×T, exploring pruning, quantisation, and knowledge distillation to reduce training and inference costs without sacrificing fidelity.

6.3.3 Unified modelling of geometry, semantics, and context.

Current approaches often trade off between preserving fine local geometry and capturing global semantics. Future inpainting frameworks should aim for unified representations that balance structural integrity with contextual plausibility, potentially through hybrid models that integrate geometric priors, 3D reasoning, or multimodal signals (e.g., text, depth, or temporal cues).

6.3.4 Controllability and user-guided inpainting.

A critical future direction is enabling controllable inpainting where users (or downstream systems) can specify constraints such as structure, style, or semantics. This may involve conditioning on textual prompts, sketches, or high-level attributes, bridging inpainting with the rapidly developing field of generative AI controllability.

6.3.5 Evaluation beyond metrics.

New evaluation frameworks are required that capture perceptual realism, functional correctness, and ethical considerations. For example, in medical applications, assessments should be co-designed with clinicians to ensure that inpainted outputs are safe, interpretable, and clinically meaningful. In creative domains, human perceptual studies and task-oriented evaluations will better reflect the value of generated imagery.

6.3.6 Generalising across modalities and tasks.

The principles explored here—geometric guidance, mask-aware downsampling, hybrid transformer-SSM modelling—are not confined to image inpainting. They may be extended

to related restoration and generation tasks, such as video inpainting, motion deblurring, super-resolution, and cross-modal translation. Future work should explore unifying these under a single generative framework.

6.3.7 Responsible and ethical deployment.

Finally, inpainting research must actively engage with ethical questions. Future work should develop mechanisms to detect and prevent misuse (e.g., deepfakes), address fairness across demographics, and provide interpretable uncertainty estimates. Embedding these considerations at the model-design level will be essential for trustworthy applications.

Bibliography

[1] K. Nazeri, E. Ng, T. Joseph, F. Qureshi, and M. Ebrahimi, "Edgeconnect: Structure guided image inpainting using edge prediction," in *The IEEE International Conference on Computer Vision (ICCV) Workshops*, Oct 2019. xi, 32, 34, 36, 39, 43, 44, 45, 114

[2] 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*, pp. 14134–14143, 2021. xi, 1, 18, 32, 43, 44, 45, 49, 63, 64, 65, 67, 68, 75, 77, 80, 81, 97, 106, 109

[3] Y. Yang, X. Guo, J. Ma, L. Ma, and H. Ling, "Lafin: Generative landmark guided face inpainting," *arXiv preprint arXiv:1911.11394*, 2019. xi, 11, 32, 34, 36, 39, 43, 44, 45, 49, 114

[4] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017. xi, xii, xiii, xv, 7, 49, 51, 53, 62, 66, 68, 75, 86, 97, 107, 108, 109

[5] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 6, pp. 1452–1464, 2017. xi, xii, xiii, 7, 51, 53, 62, 67, 68, 75, 86, 97, 108, 109

[6] Y. Deng, S. Hui, S. Zhou, D. Meng, and J. Wang, "T-former: An efficient transformer for image inpainting," in *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 6559–6568, 2022. xii, xiv, xv, 21, 69, 70, 72, 78

[7] R. Suvorov, E. Logacheva, A. Mashikhin, A. Remizova, A. Ashukha, A. Silvestrov, N. Kong, H. Goka, K. Park, and V. Lempitsky, "Resolution-robust large mask inpainting with fourier convolutions," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 2149–2159, 2022. xii, 64, 77, 85, 97, 98, 106

[8] W. Li, Z. Lin, K. Zhou, L. Qi, Y. Wang, and J. Jia, "Mat: Mask-aware transformer for large hole image inpainting," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10758–10768, 2022. xii, 3, 18, 20, 21, 52, 58, 62, 64, 65, 66, 69, 76, 77, 79, 84, 85, 97, 106, 108, 109

[9] J. Ma, F. Li, and B. Wang, "U-mamba: Enhancing long-range dependency for biomedical image segmentation," *arXiv preprint arXiv:2401.04722*, 2024. xii, 4, 23, 85, 86, 101

[10] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Restormer: Efficient transformer for high-resolution image restoration," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5728–5739, 2022. xii, xv, 4, 21, 24, 52, 58, 61, 72, 84, 85, 90, 99, 100, 101, 103, 104, 112

[11] S. Nah, T. Hyun Kim, and K. Mu Lee, "Deep multi-scale convolutional neural network for dynamic scene deblurring," in *CVPR*, 2017. xiii, xv, 103, 104, 112

[12] Z. Shen, W. Wang, X. Lu, J. Shen, H. Ling, T. Xu, and L. Shao, "Human-aware motion deblurring," in *ICCV*, 2019. xv, 103, 104

[13] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2536–2544, 2016. 1, 18

[14] S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Globally and locally consistent image completion," *ACM Transactions on Graphics (ToG)*, vol. 36, no. 4, pp. 1–14, 2017. 1, 18

[15] G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro, "Image inpainting for irregular holes using partial convolutions," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 85–100, 2018. 1, 18, 37, 43, 46, 63, 97

[16] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Free-form image inpainting with gated convolution," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4471–4480, 2019. 1, 18, 36, 38, 54, 64, 66, 67, 77, 97, 106

[17] C. Cao and Y. Fu, "Learning a sketch tensor space for image inpainting of man-made scenes," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14509–14518, 2021. 1, 18, 38

[18] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5505–5514, 2018. 1, 11, 18, 64, 66, 67, 77, 97, 106

[19] N. Wang, J. Li, L. Zhang, and B. Du, "Musical: Multi-scale image contextual attention learning for inpainting.," in *IJCAI*, pp. 3748–3754, 2019. 1, 18

[20] H. Sun, W. Li, Y. Duan, J. Zhou, and J. Lu, "Learning adaptive patch generators for mask-robust image inpainting," *IEEE Transactions on Multimedia*, 2022. 1

[21] G. Sridevi and S. Srinivas Kumar, "Image inpainting based on fractional-order nonlinear diffusion for image reconstruction," *Circuits, Systems, and Signal Processing*, vol. 38, pp. 3802–3817, 2019. 1, 17

[22] A. Atapour-Abarghouei, G. P. de La Garanderie, and T. P. Breckon, "Back to butterworth-a fourier basis for 3d surface relief hole filling within rgb-d imagery," in *2016 23rd International Conference on Pattern Recognition (ICPR)*, pp. 2813–2818, 2016. 1, 17

[23] C. Barnes, E. Shechtman, A. Finkelstein, and D. B. Goldman, "Patchmatch: A randomized correspondence algorithm for structural image editing," *ACM Trans. Graph.*, vol. 28, no. 3, p. 24, 2009. 1, 17

[24] G. Zhao, J. Wang, Z. Zhang, *et al.*, "Random shifting for cnn: a solution to reduce information loss in down-sampling layers.," in *IJCAI*, pp. 3476–3482, 2017. 2, 51

[25] S. A Sharif, R. A. Naqvi, and M. Biswas, "Beyond joint demosaicking and denoising: An image processing pipeline for a pixel-bin image sensor," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 233–242, 2021. 2, 51

[26] Z. Yue, J. Xie, Q. Zhao, and D. Meng, "Semi-supervised video deraining with dynamical rain generator," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 642–652, 2021. 2, 51

[27] L. Wang, Y. Wang, Z. Lin, J. Yang, W. An, and Y. Guo, "Learning a single network for scale-arbitrary super-resolution," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4801–4810, 2021. 2, 51

[28] Y. Zhou, J. Jiao, H. Huang, Y. Wang, J. Wang, H. Shi, and T. Huang, "When awgn-based denoiser meets real noises," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 13074–13081, 2020. 2, 51

[29] T. Yu, Z. Guo, X. Jin, S. Wu, Z. Chen, W. Li, Z. Zhang, and S. Liu, "Region normalization for image inpainting," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, pp. 12733–12740, 2020. 3, 18

[30] M. Suin, K. Purohit, and A. Rajagopalan, "Distillation-guided image inpainting," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2481–2490, 2021. 3

[31] Z. Yi, Q. Tang, S. Azizi, D. Jang, and Z. Xu, "Contextual residual aggregation for ultra high-resolution image inpainting," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7508–7517, 2020. 3, 102

[32] J. Peng, D. Liu, S. Xu, and H. Li, "Generating diverse structure for image inpainting with hierarchical vq-vae," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10775–10784, 2021. 3, 18

[33] H. Chen, Y. Wang, T. Guo, C. Xu, Y. Deng, Z. Liu, S. Ma, C. Xu, C. Xu, and W. Gao, "Pre-trained image processing transformer," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12299–12310, 2021. 3

[34] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023. 4, 22, 23, 84

[35] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of the IEEE international conference on computer vision*, pp. 3730–3738, 2015. 7, 32, 37, 43, 46, 51, 53, 62, 69, 75

[36] T. Yu, S. Zhang, C. Lin, S. You, J. Wu, J. Zhang, X. Ding, and H. An, "Dunhuang grottoes painting dataset and benchmark," *arXiv preprint arXiv:1907.04589*, 2019. 7, 51, 53, 62, 69, 75

[37] Y. Zhao, W. Chen, J. Xing, X. Li, Z. Bessinger, F. Liu, W. Zuo, and R. Yang, "Identity preserving face completion for large ocular region occlusion," *arXiv preprint arXiv:1807.08772*, 2018. 11

[38] K. Nazeri, E. Ng, T. Joseph, F. Z. Qureshi, and M. Ebrahimi, "Edgeconnect: Generative image inpainting with adversarial edge learning," *arXiv preprint arXiv:1901.00212*, 2019. 11, 18, 49, 61, 63, 75, 91

[39] L. Liao, J. Xiao, Z. Wang, C.-W. Lin, and S. Satoh, "Image inpainting guided by coherence priors of semantics and textures," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6539–6548, 2021. 11, 32, 34

[40] R. Mehmood, R. Bashir, and K. Giri, "Deep generative models: a review," *Indian Journal of Science and Technology*, vol. 16, no. 7, pp. 460–467, 2023. 13

[41] J. Xu, H. Li, and S. Zhou, "An overview of deep generative models," *IETE Technical Review*, vol. 32, no. 2, pp. 131–139, 2015. 13

[42] D. P. Kingma, M. Welling, *et al.*, "Auto-encoding variational bayes," 2013. 13

[43] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *arXiv e-prints*, pp. arXiv–1406, 2014. 13, 14, 18

[44] A. Van Den Oord, N. Kalchbrenner, and K. Kavukcuoglu, "Pixel recurrent neural networks," in *International conference on machine learning*, pp. 1747–1756, PMLR, 2016. 13

[45] A. Van den Oord, N. Kalchbrenner, L. Espeholt, O. Vinyals, A. Graves, *et al.*, "Conditional image generation with pixelcnn decoders," *Advances in neural information processing systems*, vol. 29, 2016. 13

[46] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," *Advances in neural information processing systems*, vol. 32, 2019. 13

[47] C. G. Turhan and H. S. Bilge, "Recent trends in deep generative models: a review," in *2018 3rd International Conference on Computer Science and Engineering (UBMK)*, pp. 574–579, IEEE, 2018. 14

[48] T. Sakirin and S. Kusuma, "A survey of generative artificial intelligence techniques," *Babylonian Journal of Artificial Intelligence*, vol. 2023, pp. 10–14, 2023. 14

[49] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015. 14

[50] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein gans," *Advances in neural information processing systems*, vol. 30, 2017. 14

[51] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4401–4410, 2019. 14

[52] A. Brock, J. Donahue, and K. Simonyan, "Large scale gan training for high fidelity natural image synthesis," *arXiv preprint arXiv:1809.11096*, 2018. 14

[53] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *Proceedings of the 32nd International Conference on Machine Learning* (F. Bach and D. Blei, eds.), vol. 37 of *Proceedings of Machine Learning Research*, (Lille, France), pp. 2256–2265, PMLR, 07–09 Jul 2015. 16

[54] F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah, "Diffusion models in vision: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 10850–10869, 2023. 16

[55] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020. 16

[56] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020. 16

[57] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in neural information processing systems*, vol. 34, pp. 8780–8794, 2021. 16

[58] S. Chen, A. Atapour-Abarghouei, and H. P. Shum, "Hint: High-quality inpainting transformer with mask-aware encoding and enhanced attention," *IEEE Transactions on Multimedia*, 2024. 16, 17, 50, 84, 91, 97

[59] X. Yang, D. Zhou, J. Feng, and X. Wang, "Diffusion probabilistic model made slim," in *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pp. 22552–22562, 2023. 16, 17

[60] W. Wang, D. Yang, Q. Ye, B. Cao, and Y. Zou, "Nadiffuse: Noise-aware diffusion-based model for speech enhancement," in *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 2416–2423, IEEE, 2023. 17

[61] T. Salimans and J. Ho, "Progressive distillation for fast sampling of diffusion models," *arXiv preprint arXiv:2202.00512*, 2022. 17

[62] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, June 2022. 17, 64, 65, 67, 97, 101, 106, 109

[63] S. Chen, A. Atapour-Abarghouei, H. Zhang, and H. P. Shum, "Mxt: Mamba x transformer for image inpainting," *arXiv preprint arXiv:2407.16126*, 2024. 17

[64] J. Li, N. Wang, L. Zhang, B. Du, and D. Tao, "Recurrent feature reasoning for image inpainting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7760–7768, 2020. 18, 63

[65] H. Wu, J. Zhou, and Y. Li, "Deep generative model for image inpainting with local binary pattern learning and spatial attention," *IEEE Transactions on Multimedia*, vol. 24, pp. 4016–4027, 2021. 18

[66] S. Chen, A. Atapour-Abarghouei, E. S. Ho, and H. P. Shum, "Inclg: Inpainting for non-cleft lip generation with a multi-task image processing network," *Software Impacts*, vol. 17, p. 100517, 2023. 18, 30

[67] H. Zheng, Z. Lin, J. Lu, S. Cohen, E. Shechtman, C. Barnes, J. Zhang, N. Xu, S. Amirghodsi, and J. Luo, "Image inpainting with cascaded modulation gan and object-aware training," in *European Conference on Computer Vision*, pp. 277–296, Springer, 2022. 18

[68] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017. 18, 20, 90, 95

[69] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020. 18, 20, 58, 101

[70] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021. 18, 20, 58

[71] Y. Yu, F. Zhan, R. Wu, J. Pan, K. Cui, S. Lu, F. Ma, X. Xie, and C. Miao, "Diverse image inpainting with bidirectional and autoregressive transformers," in *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 69–78, 2021. 18, 20

[72] Z. Wan, J. Zhang, D. Chen, and J. Liao, "High-fidelity pluralistic image completion with transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4692–4701, 2021. 18, 20, 21, 52, 58, 62, 63, 65, 66, 67, 69, 76, 79, 80, 84, 108, 109

[73] C. Zheng, T.-J. Cham, J. Cai, and D. Phung, "Bridging global context interactions for high-fidelity image completion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11512–11522, 2022. 18, 20, 21, 51, 52, 58, 62, 76, 84

[74] Y. Zhang, Y. Liu, R. Hu, Q. Wu, and J. Zhang, "Mutual dual-task generator with adaptive attention fusion for image inpainting," *IEEE Transactions on Multimedia*, 2023. 18, 20

[75] Y. Deng, S. Hui, S. Zhou, D. Meng, and J. Wang, "Learning contextual transformer network for image inpainting," in *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 2529–2538, 2021. 18, 20

[76] A. Gu, K. Goel, and C. Re, "Efficiently modeling long sequences with structured state spaces," in *International Conference on Learning Representations*, 2021. 23

[77] A. Gu, T. Dao, S. Ermon, A. Rudra, and C. Ré, "Hippo: Recurrent memory with optimal polynomial projections," *Advances in neural information processing systems*, vol. 33, pp. 1474–1487, 2020. 23

[78] L. Zhu, B. Liao, Q. Zhang, X. Wang, W. Liu, and X. Wang, "Vision mamba: Efficient visual representation learning with bidirectional state space model," *arXiv preprint arXiv:2401.09417*, 2024. 23

[79] Y. Liu, Y. Tian, Y. Zhao, H. Yu, L. Xie, Y. Wang, Q. Ye, and Y. Liu, "Vmamba: Visual state space model," *arXiv preprint arXiv:2401.10166*, 2024. 23, 101

[80] B. N. Patro, V. P. Namboodiri, and V. S. Agneeswaran, "Spectformer: Frequency and attention is what you need in a vision transformer," *arXiv preprint arXiv:2304.06446*, 2023. 23

[81] B. Patro and V. Agneeswaran, "Scattering vision transformer: Spectral mixing matters," *Advances in Neural Information Processing Systems*, vol. 36, 2024. 23

[82] T. Yao, Y. Pan, Y. Li, C.-W. Ngo, and T. Mei, "Wave-vit: Unifying wavelet and transformers for visual representation learning," in *European Conference on Computer Vision*, pp. 328–345, 2022. 23

[83] Z. Xiang and C. Huang, "Image inpainting based on mamba-gan network," in *International Conference on Remote Sensing, Mapping, and Image Processing (RSMIP 2025)*, vol. 13650, pp. 651–656, SPIE, 2025. 24

[84] J. Wen, W. Hou, L. Van Gool, and R. Timofte, "Matir: A hybrid mamba-transformer image restoration model. arxiv 2025," *arXiv preprint arXiv:2501.18401*. 24

[85] A. Sandooghdar and F. Yaghmaee, "Enhancing image restoration: parameter-assisted and edge-based inpainting with u-net mamba," *Signal, Image and Video Processing*, vol. 19, no. 6, p. 466, 2025. 24

[86] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004. 27, 45, 63

[87] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017. 29, 45

[88] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016. 29

[89] S. Chen, A. Atapour-Abarghouei, J. Kerby, E. S. L. Ho, D. C. G. Sainsbury, S. Butterworth, and H. P. H. Shum, "A feasibility study on image inpainting for non-cleft lip generation from patients with cleft lip," in *Proceedings of the 2022 IEEE-EMBS International Conference on Biomedical and Health Informatics*, IEEE, 2022. 30

[90] T. Karras, M. Aittala, S. Laine, E. Härkönen, J. Hellsten, J. Lehtinen, and T. Aila, "Alias-free generative adversarial networks," *Advances in Neural Information Processing Systems*, vol. 34, 2021. 31

[91] L. Zhu, Z. Liu, and S. Han, "Deep leakage from gradients," *Advances in Neural Information Processing Systems*, vol. 32, 2019. 31, 37, 47

[92] J. Shi, A. Samal, and D. Marx, "How effective are landmarks and their geometry for face recognition?," *Computer vision and image understanding*, vol. 102, no. 2, pp. 117–133, 2006. 32, 34

[93] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks)," in *International Conference on Computer Vision*, 2017. 32, 38, 43, 46

[94] "Cleft lip and palate," 2019. 33

[95] W. Wellens and V. Poorten, "Keys to a successful cleft lip and palate team," *B ENT*, p. 3, 2006. 33

[96] D. G. Mosmuller, L. M. Mennes, C. Prahl, G. J. Kramer, M. A. Disse, G. M. Van Couwelaar, B. N. Frank, and J. Don Griot, "The development of the cleft aesthetic rating scale: a new rating scale for the assessment of nasolabial appearance in complete unilateral cleft lip and palate patients," *The Cleft Palate-Craniofacial Journal*, vol. 54, no. 5, pp. 555–561, 2017. 33

[97] T. A. Patel and K. G. Patel, "Comparison of the fisher anatomical subunit and modified millard rotation-advancement cleft lip repairs," *Plastic and reconstructive surgery*, vol. 144, no. 2, pp. 238e–245e, 2019. 33

[98] C. Asher-Mcdade, C. Roberts, W. C. Shaw, and C. Gallager, "Development of a method for rating nasolabial appearance in patients with clefts of the lip and palate," *The Cleft Palate-Craniofacial Journal*, vol. 28, no. 4, pp. 385–391, 1991. 33

[99] Y. Li, J. Cheng, H. Mei, H. Ma, Z. Chen, and Y. Li, "Clpnet: cleft lip and palate surgery support with deep learning," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 3666–3672, IEEE, 2019. 34, 36

[100] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 681–685, 2001. 35

[101] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1867–1874, 2014. 35

[102] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *European Conference on Computer Vision*, pp. 483–499, Springer, 2016. 35

[103] Y. Sun, X. Wang, and X. Tang, "Deep convolutional network cascade for facial point detection," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3476–3483, 2013. 35

[104] C. Zhao, Y. Zhang, Y. Zhang, and B. Zhang, "Automatic detection of anatomical landmarks in ultrasound images: A review," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 5, pp. 1390–1401, 2020. 35

[105] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2242–2251, 2017. 36

[106] N. Nozawa, H. P. H. Shum, Q. Feng, E. S. L. Ho, and S. Morishima, "3d car shape reconstruction from a contour sketch using gan and lazy learning," *Visual Computer*, vol. 38, no. 4, pp. 1317–1330, 2022. 36

[107] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 2623–2631, 2019. 41, 75

[108] L. Zhang, H. P. Shum, L. Liu, G. Guo, and L. Shao, "Multiview discriminative marginal metric learning for makeup face verification," *Neurocomputing*, vol. 333, pp. 339–350, 2019. 47

[109] C. Rathgeb, D. Dogan, F. Stockhardt, M. De Marsico, and C. Busch, "Plastic surgery: An obstacle for deep face recognition?," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 3510–3517, 2020. 47

[110] D. Organisciak, E. S. L. Ho, and H. P. H. Shum, "Makeup style transfer on low-quality images with weighted multi-scale attention," in *Proceedings of the 2020 International Conference on Pattern Recognition*, ICPR '20, pp. 6011–6018, Jan 2020. 47

[111] F. Tramèr, F. Zhang, A. Juels, M. K. Reiter, and T. Ristenpart, "Stealing machine learning models via prediction {APIs}," in *25th USENIX security symposium (USENIX Security 16)*, pp. 601–618, 2016. 47

[112] N. Hesse, S. Pujades, J. Romero, M. J. Black, C. Bodensteiner, M. Arens, U. G. Hofmann, U. Tacke, M. Hadders-Algra, R. Weinberger, W. Muller-Felber, and A. S. Schroeder, "Learning an infant body model from RGB-D data for accurate full body motion analysis," in *Int. Conf. on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, Sept. 2018. 47

[113] H. Zhang, H. P. H. Shum, and E. S. L. Ho, "Cerebral palsy prediction with frequency attention informed graph convolutional networks," in *Proceedings of the 2022 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, EMBC '22, IEEE, 2022. 47

[114] S. Uddin and Y. J. Jung, "Global and local attention-based free-form image inpainting," *Sensors*, vol. 20, no. 11, p. 3204, 2020. 51

[115] M. Li, Y. Fu, and Y. Zhang, "Spatial-spectral transformer for hyperspectral image denoising," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, pp. 1368–1376, 2023. 52

[116] S.-I. Jang, T. Pan, Y. Li, P. Heidari, J. Chen, Q. Li, and K. Gong, "Spach transformer: spatial and channel-wise transformer based on local and global self-attentions for pet image denoising," *IEEE Transactions on Medical Imaging*, 2023. 52

[117] L. Wang, M. Cao, Y. Zhong, and X. Yuan, "Spatial-temporal transformer for video snapshot compressive imaging," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 52

[118] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," *arXiv preprint arXiv:1606.08415*, 2016. 54, 90

[119] T. Xiao, M. Singh, E. Mintun, T. Darrell, P. Dollár, and R. Girshick, "Early convolutions help transformers see better," *Advances in Neural Information Processing Systems*, vol. 34, pp. 30392–30400, 2021. 54

[120] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "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*, pp. 1874–1883, 2016. 56, 92

[121] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1251–1258, 2017. 57, 61

[122] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3146–3154, 2019. 58

[123] K. Fukushima, "Cognitron: A self-organizing multilayered neural network," *Biological cybernetics*, vol. 20, no. 3-4, pp. 121–136, 1975. 59

[124] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *Proc. Interspeech 2019*, pp. 2613–2617, 2019. 60

[125] D. S. Park, Y. Zhang, C.-C. Chiu, Y. Chen, B. Li, W. Chan, Q. V. Le, and Y. Wu, "Specaugment on large scale datasets," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6879–6883, IEEE, 2020. 60

[126] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020. 60, 61, 74

[127] X. Li, Q. Guo, D. Lin, P. Li, W. Feng, and S. Wang, "Misf: Multi-level interactive siamese filtering for high-fidelity image inpainting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1869–1878, 2022. 61, 63, 64, 65, 67, 68, 75, 77, 80, 81, 82, 91, 97, 106, 108, 109

[128] D. P. Kingma, J. A. Ba, and J. Adam, "A method for stochastic optimization. arxiv 2014," *arXiv preprint arXiv:1412.6980*, vol. 106, 2020. 63

[129] Y. Zeng, J. Fu, H. Chao, and B. Guo, "Learning pyramid-context encoder network for high-quality image inpainting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1486–1494, 2019. 63

[130] Y. Ren, X. Yu, R. Zhang, T. H. Li, S. Liu, and G. Li, "Structureflow: Image inpainting via structure-aware appearance flow," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 181–190, 2019. 63

[131] Q. Guo, X. Li, F. Juefei-Xu, H. Yu, Y. Liu, and S. Wang, "Jpgnet: Joint predictive filtering and generative network for image inpainting," in *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 386–394, 2021. 63, 65, 68, 82

[132] Y. Yu, F. Zhan, S. Lu, J. Pan, F. Ma, X. Xie, and C. Miao, "Wavefill: A wavelet-based generation network for image inpainting," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 14114–14123, 2021. 64, 65, 66, 69, 77, 79, 80, 97, 106, 108, 109

[133] R. Zhang, W. Quan, Y. Zhang, J. Wang, and D.-M. Yan, "W-net: Structure and texture interaction for image inpainting," *IEEE Transactions on Multimedia*, 2022. 64, 66, 67, 77

[134] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, and L. Van Gool, "Repaint: Inpainting using denoising diffusion probabilistic models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11461–11471, June 2022. 64, 65, 66, 97, 101, 106, 108, 109

[135] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018. 65

[136] Z. Chang, G. A. Koulieris, and H. P. H. Shum, "On the design fundamentals of diffusion models: A survey," *arXiv*, 2023. 65, 101

[137] S. Chen, A. Atapour-Abarghouei, J. Kerby, E. S. L. Ho, D. C. G. Sainsbury, S. Butterworth, and H. P. H. Shum, "A feasibility study on image inpainting for non-cleft lip generation from patients with cleft lip," in *Proceedings of the 2022 IEEE-EMBS International Conference on Biomedical and Health Informatics*, BHI '22, pp. 1–4, IEEE, 9 2022. 75

[138] M. Ni, X. Li, and W. Zuo, "Nuwa-lip: Language-guided image inpainting with defect-free vqgan," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14183–14192, June 2023. 75

[139] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pvt v2: Improved baselines with pyramid vision transformer," *Computational Visual Media*, vol. 8, no. 3, pp. 415–424, 2022. 89

[140] S. Ren, D. Zhou, S. He, J. Feng, and X. Wang, "Shunted self-attention via multi-scale token aggregation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10853–10862, 2022. 89

[141] N. Hyeon-Woo, K. Yu-Ji, B. Heo, D. Han, S. J. Oh, and T.-H. Oh, "Scratching visual transformer's back with uniform attention," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5807–5818, 2023. 91

[142] K. Ko and C.-S. Kim, "Continuously masked transformer for image inpainting," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13169–13178, 2023. 97, 98, 106, 108, 109

[143] Y. Cui, Y. Tao, Z. Bing, W. Ren, X. Gao, X. Cao, K. Huang, and A. Knoll, "Selective frequency network for image restoration," in *The Eleventh International Conference on Learning Representations*, 2022. 99, 103

[144] E. Baron, I. Zimerman, and L. Wolf, "2-d ssm: A general spatial layer for visual transformers," *arXiv preprint arXiv:2306.06635*, 2023. 101

[145] O. Kupyn, T. Martyniuk, J. Wu, and Z. Wang, "DeblurGAN-v2: Deblurring (orders-of-magnitude) faster and better," in *ICCV*, 2019. 103

[146] H. Gao, X. Tao, X. Shen, and J. Jia, "Dynamic scene deblurring with parameter selective sharing and nested skip connections," in *CVPR*, 2019. 103

[147] K. Zhang, W. Luo, Y. Zhong, L. Ma, B. Stenger, W. Liu, and H. Li, "Deblurring by realistic blurring," in *CVPR*, 2020. 103, 112

[148] D. Park, D. U. Kang, J. Kim, and S. Y. Chun, "Multi-temporal recurrent neural networks for progressive non-uniform single image deblurring with incremental temporal training," in *ECCV*, 2020. 103

[149] H. Zhang, Y. Dai, H. Li, and P. Koniusz, "Deep stacked hierarchical multi-patch network for image deblurring," in *CVPR*, 2019. 103

[150] M. Suin, K. Purohit, and A. N. Rajagopalan, "Spatially-attentive patch-hierarchical network for adaptive motion deblurring," in *CVPR*, 2020. 103

[151] K. Purohit, M. Suin, A. Rajagopalan, and V. N. Boddeti, "Spatially-adaptive image restoration using distortion-guided networks," in *ICCV*, 2021. 103

[152] S.-J. Cho, S.-W. Ji, J.-P. Hong, S.-W. Jung, and S.-J. Ko, "Rethinking coarse-to-fine approach in single image deblurring," in *ICCV*, 2021. 103

[153] H. Chen, Y. Wang, T. Guo, C. Xu, Y. Deng, Z. Liu, S. Ma, C. Xu, C. Xu, and W. Gao, "Pre-trained image processing transformer," in *CVPR*, 2021. 103

[154] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, M.-H. Yang, and L. Shao, "Multi-stage progressive image restoration," in *CVPR*, 2021. 103, 112

[155] L. Chen, X. Lu, J. Zhang, X. Chu, and C. Chen, "Hinet: Half instance normalization network for image restoration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 182–192, 2021. 103

[156] F.-J. Tsai, Y.-T. Peng, Y.-Y. Lin, C.-C. Tsai, and C.-W. Lin, "Stripformer: Strip transformer for fast image deblurring," in *European Conference on Computer Vision*, pp. 146–162, Springer, 2022. 103, 112

[157] X. Chu, L. Chen, C. Chen, and X. Lu, "Improving image restoration by revisiting global information aggregation," in *European Conference on Computer Vision*, pp. 53–71, 2022. 104

[158] Z. Tu, H. Talebi, H. Zhang, F. Yang, P. Milanfar, A. Bovik, and Y. Li, "Maxim: Multi-axis mlp for image processing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5769–5780, 2022. 112

APPENDIX A

Content Acknowledgements

In addition to those individuals outlined in the Acknowledgments of this thesis, specific assistance in terms of hardware, software and suggestions have been received as acknowledged below.

We acknowledge Durham NVIDIA CUDA Center (NCC) GPU system [1]. The research and experiments in this thesis has used Durham University's NCC cluster. NCC has been purchased through Durham University's strategic investment funds, and is installed and maintained by the Department of Computer Science.

A.1 Image Inpainting for Non-Cleft Lip Generation

We acknowledge Dr. Edmond S. L. Ho in the University of Glasgow for his suggestions and discussions on various methods and management of cleft lip datasets. We also acknowledge the surgeons Dr. Jane Kerby, Dr. David C. G. Sainsbury and Sophie Butterworth in the Newcastle Upon Tyne Hospitals NHS Foundation Trust for their effort on data collection and method evaluation.

[1] https://nccadmin.webspace.durham.ac.uk

Moreover, We acknowledge the use of the following public resources, during the course of Chapter 3:

- Lafin[2] . . . . . CC BY-NC-SA 4.0.
- BerHu CelebA[3] . . . . . CC BY-NC-SA 4.0
- face-alignment[4] . . . . . BSD-3-Clause license

A.2 High Quality Image Inpainting with Enhanced Transformer

We acknowledge the use of the following public resources, during the course of Chapter 4:

- CelebA[5] . . . . . CC BY-NC-SA 4.0
- Places2[6] . . . . . CC BY-NC-SA 4.0
- Dunhuang[7] . . . . . CC BY-NC-SA 4.0
- apex[8] . . . . . BSD-3-Clause license
- restormer[9] . . . . . CC BY-NC-SA 4.0

A.3 Long-Range Dependency Capture and Pixel-Level Sequential Modelling

we acknowledge the use of the following public resources, during the course of Chapter 5:

- CelebA[10] . . . . . CC BY-NC-SA 4.0

[2] https://github.com/YaN9-Y/lafin.
[3] http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html.
[4] https://github.com/1adrianb/face-alignment.
[5] http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html.
[6] http://places2.csail.mit.edu/download.html.
[7] https://www.cvl.iis.u-tokyo.ac.jp/e-Heritage2019/index.php?id=challenge.
[8] https://github.com/NVIDIA/apex.
[9] https://github.com/swz30/Restormer.
[10] http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html.

- Places2[11] . . . . . CC BY-NC-SA 4.0
- apex[12] . . . . . BSD-3-Clause license
- restormer[13] . . . . . CC BY-NC-SA 4.0
- mamba[14] . . . . . Apache License 2.0
- U-Mamba[15] . . . . . Apache-2.0 license
- Vim[16] . . . . . Apache License 2.0

[11] http://places2.csail.mit.edu/download.html.
[12] https://github.com/NVIDIA/apex.
[13] https://github.com/swz30/Restormer.
[14] https://github.com/state-spaces/mamba.
[15] https://github.com/bowang-lab/U-Mamba.
[16] https://github.com/hustvl/Vim.