# SEM-Net: Efficient Pixel Modelling for image inpainting with Spatially Enhanced SSM

Shuang Chen[1] Haozheng Zhang[1] Amir Atapour-Abarghouei[1] Hubert P. H. Shum[1†]
[1]{shuang.chen, haozheng.zhang, amir.atapour-abarghouei, hubert.shum}@durham.ac.uk
[†]Corresponding Author

## Abstract

*Image inpainting aims to repair a partially damaged image based on the information from known regions of the images. Achieving semantically plausible inpainting results is particularly challenging because it requires the reconstructed regions to exhibit similar patterns to the semanticly consistent regions. This requires a model with a strong capacity to capture long-range dependencies. Existing models struggle in this regard due to the slow growth of receptive field for Convolutional Neural Networks (CNNs) based methods and patch-level interactions in Transformer-based methods, which are ineffective for capturing long-range dependencies. Motivated by this, we propose SEM-Net, 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. To address the inherent lack of spatial awareness in SSM, we introduce the Snake Mamba Block (SMB) and Spatially-Enhanced Feedforward Network. These innovations enable SEM-Net to outperform state-of-the-art inpainting methods on two distinct datasets, showing significant improvements in capturing LRDs and enhancement in spatial consistency. Additionally, SEM-Net achieves state-of-the-art performance on motion deblurring, demonstrating its generalizability. Our source code is available:* https://github.com/ChrisChen1023/SEM-Net.

## 1. Introduction

Image inpainting is a highly challenging low-level vision task in computer vision due to its ill-posed nature. It aims to repair partially damaged or missing regions by leveraging information from the known areas [58]. Successful inpainting relies heavily on advanced image representation learning, particularly in capturing both short-range and long-range dependencies [11, 58], to ensure consistent reconstruction between the filled and visible contents.

Convolutional Neural Networks (CNNs) are widely used as backbone networks in image inpainting due to their strong performance in learning generalizable representations from images and their effective mining of short-range dependencies through convolution operations [34, 40, 49, 52]. 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 [2, 25] 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. [54] 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.

LRDs are critical in image inpainting, as a lack of LRDs often results in low-quality outcomes due to insufficient context capturing. This is evidenced by the inconsistent eye colours and patterns, as shown in the visualization of the prominent CNN-based method [42] and transformer-based method [25] (Sample I of Fig. 1), where the visible red eye fails to guide the accurate reconstruction of the other eye.

One feasible solution to this challenge is from [14], which proposes an emerging Selective State Space Model (SSM), known as Mamba. SSM has demonstrated its efficient and effective capacity in learning LRDs, and good adaptability in computer vision [29]. As shown in Sample I of Fig. 1, directly adopting SSM [31] (M-Unet) captures LRDs effectively and achieves more consistent eye colour.

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. 1, a vanilla SSM model [31] shows positional drifting of the inpainted left eye (upper than the right eye). This insight introduces two key chal-

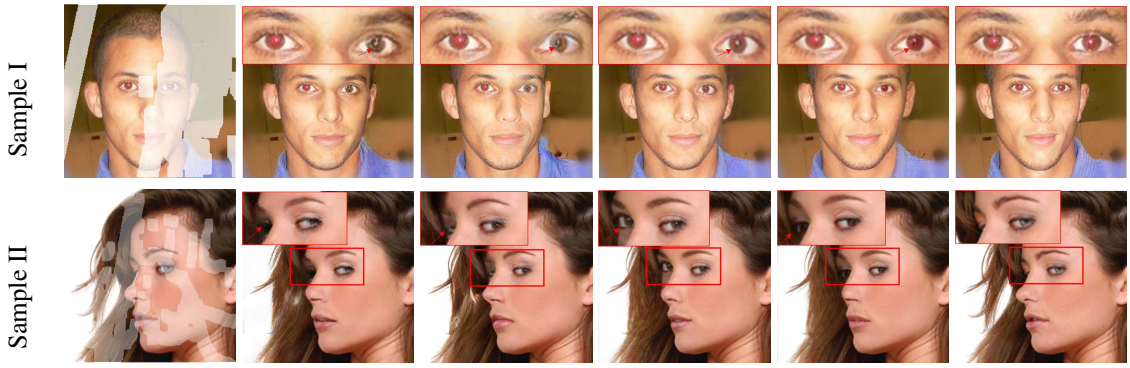


Figure 1. Comparisons with the state-of-the-art CNN-based method [42] and transformer-based method [25]. M-Unet is a variant of directly applying the Mamba model [31] followed by a feedforward network [54] in a U-Net. Red boxes and arrows highlight major differences. Our SEM-Net demonstrates the strong 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.

lenges: (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 [22] and Places2 [60]. 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 generalizability in image representation learning.

Our main contributions are summarized as follows:

- We propose a novel U-shaped Spatially-Enhanced SSM architecture focused on capturing short- and long-range spatial dependencies in image inpainting. To the best of our knowledge, SEM-Net is the first SSM-based model in this research field.
- 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.

## 2. Related Work

### 2.1. Image Inpainting

Image inpainting is a classic ill-posed low-level vision task that requires inferring the missing or damaged areas of the image based on the known pixels. The rapid advancements in CNNs have led to the introduction of diverse techniques that have significantly improved image inpainting by using CNN-based encoder-decoder architectures [40, 48, 52] or CNN-based Generative Adversarial Networks (GANs) [21, 34, 37, 47, 49, 50]. Nonetheless, limited receptive fields of convolution hinder the capturing of long-range dependencies [46, 54], which motivates the researchers to enlarge the receptive fields by employing convolution in frequency domain [7, 42] or developing transformer-based models [2, 5, 25, 27]. However, practical applications are limited by the computational complexity of self-attention, specifically, as it is still challenging to achieve pixel-level self-attention at relatively high resolutions [9]. In this paper, we primarily focus on the field of selective SSM for achieving image pixel-level LRDs learning with strong spatial awareness.

### 2.2. SSM in Computer Vision

SSMs serve as fundamental models in various fields, including control theory and computational neuroscience. However, the application of SSMs in deep learning is limited due to the risk of vanishing gradients when using the linear first-order Ordinary Differential Equations to solve an exponential function [17]. To solve this problem, Gu et al. [16] proposed a structured SSM model with the HiPPO framework [15] to address the significant computational challenges and model LRDs with rigorous theoretical proofs. Recently, [14] further proposed a selective structured SSM, namely Mamba, to allow for context-based reasoning using long sequence inputs. Inspired by them, several researchers have begun to explore the selective SSMs in computer vision tasks, including image segmentation [31], image classification and object detection [29,61]. However, these works have not effectively leveraged the potential of Mamba in capturing long-range pixel-level spatial dependencies for image representation learning. [20] uses a zigzag scanning approach to improve spatial continuity in patched images. However, the fixed scanning directions can result in a loss of joint information between pixels, particularly as the patch size increases.

### 2.3. Preliminary

While SSMs are generally regarded as linear time-invariant systems that map a 1-dimensional sequence $x(t) \in \mathbb{R}$ to response $y(t) \in \mathbb{R}$ via an implicit latent state $h(t) \in \mathbb{R}^N$ (Eq.1), the structured SSMs use zero-order hold discretisation rule (Eq.2) to transform the continues parameter $(\Delta, \boldsymbol{A}, \boldsymbol{B})$ to discrete parameters $(\overline{\boldsymbol{A}}, \overline{\boldsymbol{B}})$ for allowing efficient linear recurrence in (Eq.3).

$$h'(t) = \boldsymbol{A}h(t) + \boldsymbol{B}x(t), \quad y(t) = \boldsymbol{C}h(t), \tag{1}$$

$$\overline{\boldsymbol{A}} = \exp(\Delta\boldsymbol{A}), \quad \overline{\boldsymbol{B}} = (\Delta\boldsymbol{A})^{-1}(\exp(\Delta\boldsymbol{A}) - \boldsymbol{I}) \cdot \Delta\boldsymbol{B}, \tag{2}$$

$$h_t = \overline{\boldsymbol{A}}h_{t-1} + \overline{\boldsymbol{B}}x_t, \quad y_t = \boldsymbol{C}h_t, \tag{3}$$

where $\boldsymbol{A} \in \mathbb{R}^{N\times N}$, $\boldsymbol{B} \in \mathbb{R}^{N\times 1}$, $\boldsymbol{C} \in \mathbb{R}^{1\times N}$, and $\Delta$ is the step size.

To enhance the hidden state dimension while avoiding the trade-offs in speed and memory, Gu et al. [14] further proposes an architecture (i.e., Mamba) with a selection mechanism on SSMs to transform parameters $\Delta$,$\boldsymbol{B}$,$\boldsymbol{C}$, into functions that depend on the input, such that introducing a length dimension to these parameters, making the model changed from time-invariant to time-varying.

## 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. 2. Our framework comprises two key components to address the two identified challenges in a synergistic manner. The first component, a Snake Mamba Block (Sec. 3.1.1), 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. 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 $3 \times 3$ 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 $1 \times 1$ 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 [39] are employed to achieve feature downsampling and upsampling, respectively. In the final step, a convolutional layer projects the decoded features to the output.

### 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).

#### 3.1.1 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

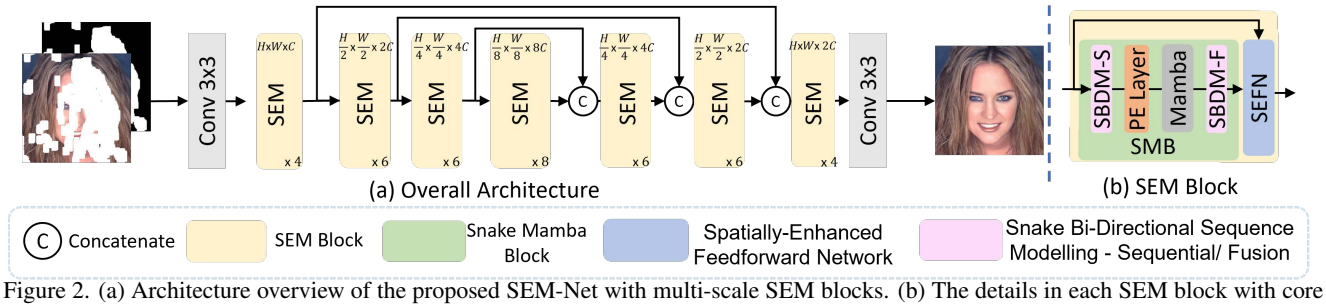


Figure 2. (a) Architecture overview of the proposed SEM-Net with multi-scale SEM blocks. (b) The details in each SEM block with core designs in SMB and SEFN, which holistically enhance the spatial awareness and improve the capability to capture LRDs.

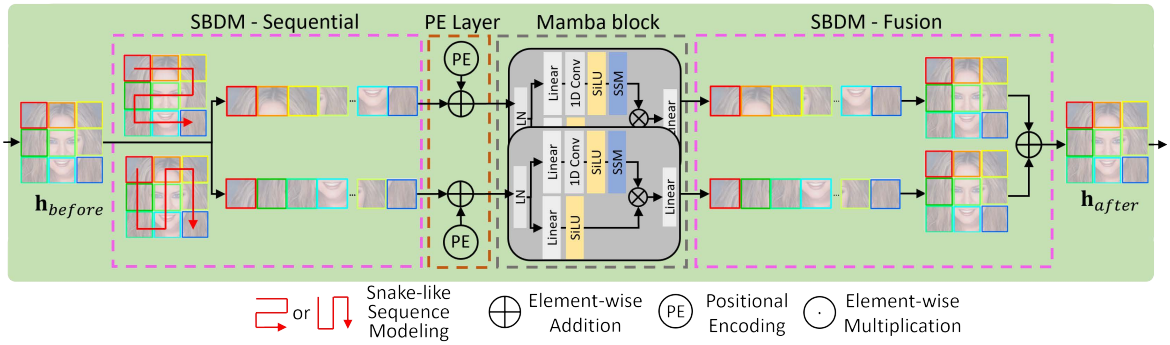


Figure 3. The architecture of proposed SMB. The input feature is modelled to sequences in two directions with snake-like traverses in SBDM-Sequential, enhancing the spatial awareness implicitly. Then, the PE layer explicitly enhances the long-range positional awareness through positional embeddings. The features after Mamba are restructured and aggregated by SBDM-Fusion to generate the output.

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 recognize 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 neighboring pixels both within and across rows by reordering pixels and concatenating rows, illustrated by the red arrows in Fig. 3.

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:

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

$$S = \text{concat}[S_0, S_1, S_2, S_3, \ldots, S_{H-1}], \quad (5)$$

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 3, 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 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 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 [44]. Then, the elements in 1D sequence $S$ with 1D positional embeddings $\bar{S}(n)$ are integrated by aggregation:

$$\bar{S}(n) = S(n) + PE(n), n = 0, 1, 2, 3, \ldots, N^2 - 1. \quad (6)$$

### 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 $\gamma$ 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 $\gamma$ 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:

$$\mathbf{h}'_{after} = W_{d3} W_1 LN(\mathbf{h}_{after}), \quad (7)$$

$$\mathbf{h}''_{after} = W'_{d3} W'_1 LN(\mathbf{h}_{after}), \quad (8)$$

$$\gamma = Up(f(AveragePooling(\mathbf{h}_{before}))), \quad (9)$$

$$gate = GELU(W_{d3} W_1 \gamma || \mathbf{h}'_{after}), \quad (10)$$

$$output = gate \odot \mathbf{h}''_{after}, \quad (11)$$

where $W_1$, $W'_1$ are $1 \times 1$ convolutions, $W_{d3}$, $W'_{d3}$ are $3 \times 3$ 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.

## 4. Experiments

We quantitatively and qualitatively prove the superiority of our proposed inpainting method by comparing it with the state-of-the-art methods on two widely used datasets [25, 45, 59]: CelebA-HQ [22] and Places2-Standard [60] in Sec. 4.1 and 4.2. We carefully evaluate each of the proposed novelties by a comprehensive ablation and component analysis in Sec. 4.3. In Sec. 4.5, we demonstrate the capability of our model in generalising to both higher resolution and unseen images. In addition, we further evaluate the image representation learning capability and generalisation ability of SEM-Net by directly applying it to another low-level vision task - image motion debluring. The implementation setting is detailed in the supplementary material.

**Baselines and Metrics** We choose the following baselines for inpainting comparison: CNN-based methods with DeepFill v1 [50], DeepFill v2 [51], CTSDG [18] and *MISF* [26], WaveFill [53] and *LAMA* [42]; Transformer-based methods with MAT [25] and *CMT* [23]; Expensive diffusion models [30, 36]. *Italic* denotes the SOTA methods. We followed [25, 26] to evaluate our SEM-Net on Peak Signal-to-Noise Ratio (PSNR), Structural Similarity (SSIM), L1, Fréchet Inception Distance (FID) [19] and Perceptual Similarity (LPIPS).

### 4.1. Quantitative Comparison

We employ officially released models and test with the same masks [28] for fair comparisons. The quantitative results are shown in Tab 1. Our method consistently outperforms the state-of-the-art approaches [23, 26, 42] across all mask ratios in both CelebA-HQ and Places2 datasets. Particularly on the CelebA-HQ dataset, 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 [23, 42] 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 [23] 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.

### 4.2. Qualitative Comparison

We showcase the qualitative image inpainting results on both datasets in Fig. 5. Each sample is the inpainted result where the mask ratio exceeds 40%, to more intuitively demonstrate the advantages of SEM-Net in handling challenging cases. In 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-range dependency

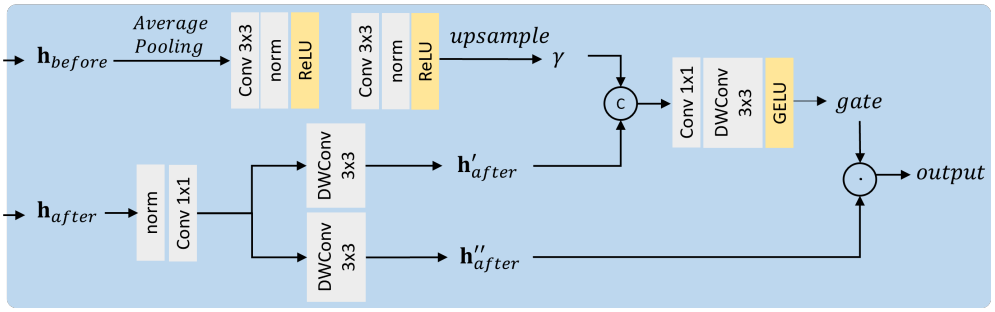


Figure 4. The architecture of proposed Spatially-Enhanced Feedforward Network (SEFN)

Table 1. 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 [50]$^{C}$ | 3 / 7$ms$ | 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 [51]$^{C}$ | 4 / 10$ms$ | 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 [53]$^{C}$ | 49 / 70$ms$ | 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 [30]$^{D}$ | 552 / 250000$ms$ | - | - | - | - | - | - | - | - | - | - | 21.8321 | 0.7791 | 3.9427 | 8.9637 | 0.1943 |
| LaMa [42]$^{C}$ | 51 / 25$ms$ | 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 [26]$^{C}$ | 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 [25]$^{T}$ | 62 / 70$ms$ | 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 [23]$^{C}$ | 143 / 60$ms$ | 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 |
| Ours | 163 / 240$ms$ | **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 [50]$^{C}$ | 3 / 7$ms$ | 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 [51]$^{C}$ | 4 / 10$ms$ | 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 [18]$^{C}$ | 52 / 20$ms$ | 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 [53]$^{C}$ | 49 / 70$ms$ | 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 [36]$^{D}$ | 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 [26]$^{C}$ | 26 / 10$ms$ | 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 [42]$^{C}$ | 51 / 25$ms$ | 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 [23]$^{T}$ | 143 / 60$ms$ | 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 |
| Ours | 163 / 240$ms$ | **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.

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.

Table 2. Ablation studies of each component in $40\% - 60\%$ mask ratio. Refer to supplementary material for all mask ratios.

| Net | Components | | | | | 40%-60% | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | MB | FN [54] | 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** |

## 4.3. Ablation Study and Component Analysis

To efficiently verify the proposed modules, we followed [10] to conduct the ablation and component analysis experiments on a lighter version of SEM-Net with the halved number of SMB in each U-Net stage. Each experiment is trained on CelebA-HQ for 30K iterations.

Table 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 [54] | 21.5362 | 0.7543 | 4.0471 | 8.1652 | 0.2326 |
| | SSA [12] | Out of memory | | | | |
| | SMB | **22.1776** | **0.7708** | **3.6747** | **7.9125** | **0.2095** |
| 64*64 | SSA [12] | 20.1655 | 0.7265 | 5.2256 | 5.5547 | 0.1702 |
| | SMB | **20.1716** | **0.7352** | **5.1332** | **5.3158** | **0.1617** |

**Improvement in Each Component.** Tab. 2 and Fig. 6 present the improvement of each component quantitatively and qualitatively. Based on the U-Net shape baseline (Tab. 2a), integrating the Mamba Block (MB) and Feedforward Network (FN) [54] (Tab. 2b) results in noticeable improvements across all metrics. Fig. 6d→b and Fig. 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. 6b→c,

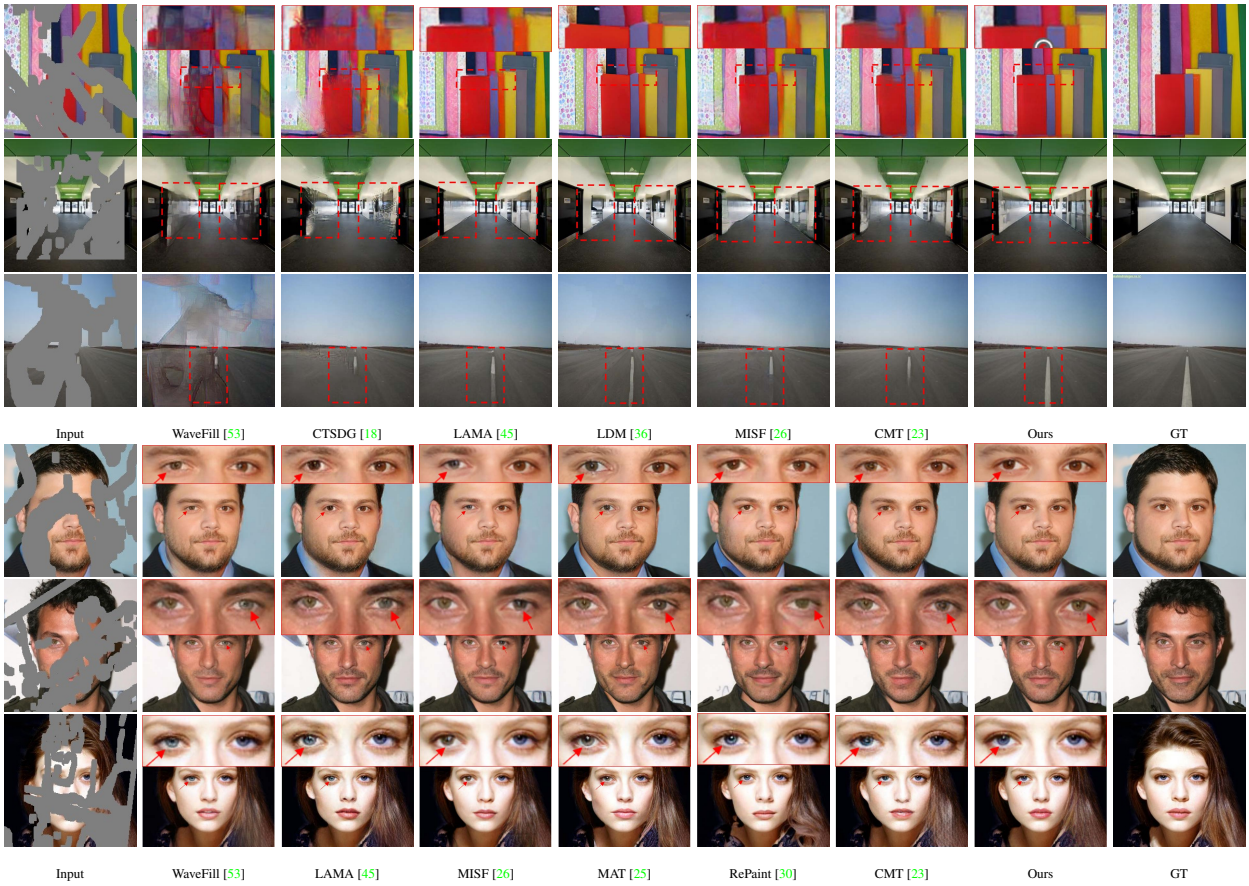


Figure 5. Comparisons with visualisations $(256 \times 256)$ showing that our results are more coherent in structure and sharper in texture and semantic details. The top three rows are from Places2 [60] and the bottom three rows are from CelebA-HQ [22].


Figure 6. The qualitative visualisation of ablation studies. Zoom in for the details.

Fig. 6d→e and Fig. 6f→g revealed the effect of SEFN by resulting sharper jaw and less artefacts, demonstrated by the improvement in SSIM score. Tab. 2d→f and Tab. 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.

### 4.4. 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 LRDs: channel-wise self-attention [54] and Spatial-wise self-attention (SSA) [12]. For fair comparisons, all models use vanilla feedforward networks [54] instead of our novel SEFN, with only differences between SMB, CSA and SSA. From Table. 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 $(256 \times 256)$ images while SSA can only be trained on the degraded $64 \times 64$ images with a single A100 due to its significant computational cost. Furthermore, compared with the diffusion-based models [30, 36] with a very long inference time [1], our model has better performance while the inference time is still in milliseconds, which is suitable for real-time scenarios (shown in Tab. 1).

### 4.5. Generalisation Ability

**Unseen High Resolution Images.** We examine the scalability and generalizability of SEM-Net trained on $256 \times 256$

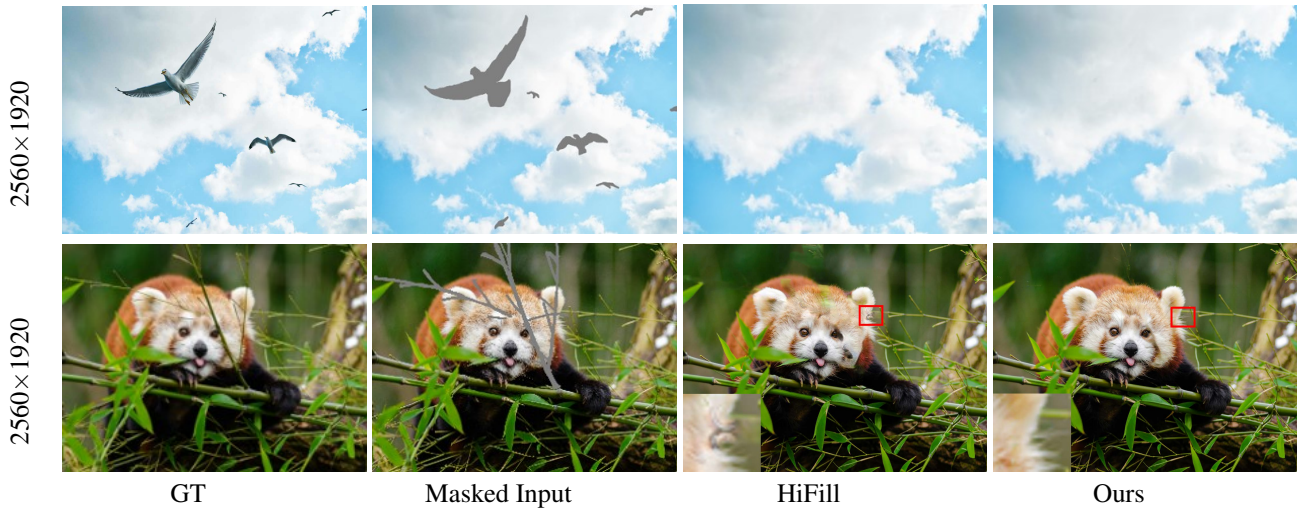


Figure 7. Examples of generalisation to real-world high-resolution images of 2560 × 1920.

Table 4. Performance in generalising to image motion deblurring task. Our SEM-Net is trained only on the GoPro dataset [32] and directly applied to the HIDE [38].

| Method | GoPro [32] | | HIDE [38] | |
|---|---|---|---|---|
| | PSNR ↑ | SSIM ↑ | PSNR ↑ | SSIM ↑ |
| DeblurGAN-v2 [24] | 29.55 | 0.934 | 26.61 | 0.875 |
| Shen *et al*. [38] | - | - | 28.89 | 0.930 |
| Gao *et al*. [13] | 30.90 | 0.935 | 29.11 | 0.913 |
| DBGAN [57] | 31.10 | 0.942 | 28.94 | 0.915 |
| MT-RNN [33] | 31.15 | 0.945 | 29.15 | 0.918 |
| DMPHN [56] | 31.20 | 0.940 | 29.09 | 0.924 |
| Suin *et al*. [41] | 31.85 | 0.948 | 29.98 | 0.930 |
| SPAIR [35] | 32.06 | 0.953 | 30.29 | 0.931 |
| MIMO-UNet+ [6] | 32.45 | 0.957 | 29.99 | 0.930 |
| IPT [3] | 32.52 | - | - | - |
| MPRNet [55] | 32.66 | 0.959 | 30.96 | 0.939 |
| HINet [4] | 32.71 | 0.959 | 30.32 | 0.932 |
| Restormer [54] | 32.92 | 0.961 | **31.22** | **0.942** |
| Stripformer [43] | 33.08 | **0.962** | 31.03 | 0.940 |
| Ours | **33.11** | **0.962** | 31.12 | 0.941 |

Places2 images in processing unseen images of higher resolution, since these abilities are crucial for practical applications where image resolutions can significantly vary. Fig. 7 showcases examples of unseen real-world high-resolution applications. While [49] performs similarly with larger masks, its upsampling strategy causes narrow mask drifting, leading to artefacts. SEM-Net, by modelling at the pixel level, captures finer details without artefacts, offering 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 of SEN-Net, 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. Tab. 4 shows that SEM-Net overall outperforms the restoration models on two synthetic benchmark datasets GoPro [32] and HIDE [38]. Especially on GoPro, SEM-Net improves PSNR by 0.19 compared to the strong restoration baseline model Restormer [54]. Notably, our SEM is trained on GoPro and directly applied to HIDE, without progressive learning [54] or Test-time Local Converter [8] such external optimization, showcasing strong generalization ability. Refer to supplementary materials for qualitative results.

## 5. Conclusion and Discussion

We present an SSM-based image inpainting model, SEM-Net, which demonstrates strong capabilities in capturing LRDs and addresses the challenge of lack of spatial awareness in SSMs. We propose two key designs, SMB and SEFN, for improved image representation learning. With these designs, our model outperforms 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 generalizability to higher-resolution images and another low-level visual task, image deblurring. Our future work aims to build a controllable image inpainting model based on the proposed SMB to handle large-resolution images.

## References

[1] Ziyi Chang, George Alex Koulieris, and Hubert P. H. Shum. On the design fundamentals of diffusion models: A survey. *arXiv*, 2023. 7

[2] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF conference on computer vi-*

*sion and pattern recognition*, pages 12299–12310, 2021. 1, 2

[3] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *CVPR*, 2021. 8

[4] Liangyu Chen, Xin Lu, Jie Zhang, Xiaojie Chu, and Chengpeng Chen. Hinet: Half instance normalization network for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 182–192, 2021. 8

[5] Shuang Chen, Amir Atapour-Abarghouei, and Hubert PH Shum. Hint: High-quality inpainting transformer with mask-aware encoding and enhanced attention. *IEEE Transactions on Multimedia*, 2024. 2

[6] Sung-Jin Cho, Seo-Won Ji, Jun-Pyo Hong, Seung-Won Jung, and Sung-Jea Ko. Rethinking coarse-to-fine approach in single image deblurring. In *ICCV*, 2021. 8

[7] Tianyi Chu, Jiafu Chen, Jiakai Sun, Shuobin Lian, Zhizhong Wang, Zhiwen Zuo, Lei Zhao, Wei Xing, and Dongming Lu. Rethinking fast fourier convolution in image inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23195–23205, 2023. 2

[8] Xiaojie Chu, Liangyu Chen, Chengpeng Chen, and Xin Lu. Improving image restoration by revisiting global information aggregation. In *European Conference on Computer Vision*, pages 53–71, 2022. 8

[9] Yuning Cui, Wenqi Ren, Xiaochun Cao, and Alois Knoll. Image restoration via frequency selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2

[10] Yuning Cui, Yi Tao, Zhenshan Bing, Wenqi Ren, Xinwei Gao, Xiaochun Cao, Kai Huang, and Alois Knoll. Selective frequency network for image restoration. In *The Eleventh International Conference on Learning Representations*, 2022. 6

[11] Ye Deng, Siqi Hui, Sanping Zhou, Wenli Huang, and Jinjun Wang. Context adaptive network for image inpainting. *IEEE Transactions on Image Processing*, 2023. 1

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 6, 7

[13] Hongyun Gao, Xin Tao, Xiaoyong Shen, and Jiaya Jia. Dynamic scene deblurring with parameter selective sharing and nested skip connections. In *CVPR*, 2019. 8

[14] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023. 1, 3

[15] Albert Gu, Tri Dao, Stefano Ermon, Atri Rudra, and Christopher Ré. Hippo: Recurrent memory with optimal polynomial projections. *Advances in neural information processing systems*, 33:1474–1487, 2020. 3

[16] Albert Gu, Karan Goel, and Christopher Re. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations*, 2021. 3

[17] Albert Gu, Isys Johnson, Karan Goel, Khaled Saab, Tri Dao, Atri Rudra, and Christopher Ré. Combining recurrent, convolutional, and continuous-time models with the structured learnable linear state space layer. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 3

[18] Xiefan Guo, Hongyu Yang, and Di Huang. Image inpainting via conditional texture and structure dual generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14134–14143, 2021. 5, 6, 7

[19] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 5

[20] Vincent Tao Hu, Stefan Andreas Baumann, Ming Gui, Olga Grebenkova, Pingchuan Ma, Johannes Fischer, and Bjorn Ommer. Zigma: Zigzag mamba diffusion model. *arXiv preprint arXiv:2403.13802*, 2024. 3

[21] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. *ACM Transactions on Graphics (ToG)*, 36(4):1–14, 2017. 2

[22] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 2, 5, 7

[23] Keunsoo Ko and Chang-Su Kim. Continuously masked transformer for image inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13169–13178, 2023. 5, 6, 7

[24] Orest Kupyn, Tetiana Martyniuk, Junru Wu, and Zhangyang Wang. DeblurGAN-v2: Deblurring (orders-of-magnitude) faster and better. In *ICCV*, 2019. 8

[25] Wenbo Li, Zhe Lin, Kun Zhou, Lu Qi, Yi Wang, and Jiaya Jia. Mat: Mask-aware transformer for large hole image inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10758–10768, 2022. 1, 2, 5, 6, 7

[26] Xiaoguang Li, Qing Guo, Di Lin, Ping Li, Wei Feng, and Song 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*, pages 1869–1878, 2022. 5, 6, 7

[27] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1833–1844, 2021. 2

[28] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *Proceedings of the European conference on computer vision (ECCV)*, pages 85–100, 2018. 5

[29] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, and Yunfan Liu. Vmamba: Visual state space model. *arXiv preprint arXiv:2401.10166*, 2024. 1, 3

[30] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting

using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11461–11471, June 2022. 5, 6, 7

[31] Jun Ma, Feifei Li, and Bo Wang. U-mamba: Enhancing long-range dependency for biomedical image segmentation. *arXiv preprint arXiv:2401.04722*, 2024. 1, 2, 3

[32] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *CVPR*, 2017. 8

[33] Dongwon Park, Dong Un Kang, Jisoo Kim, and Se Young Chun. Multi-temporal recurrent neural networks for progressive non-uniform single image deblurring with incremental temporal training. In *ECCV*, 2020. 8

[34] Jialun Peng, Dong Liu, Songcen Xu, and Houqiang Li. Generating diverse structure for image inpainting with hierarchical vq-vae. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10775–10784, 2021. 1, 2

[35] Kuldeep Purohit, Maitreya Suin, AN Rajagopalan, and Vishnu Naresh Boddeti. Spatially-adaptive image restoration using distortion-guided networks. In *ICCV*, 2021. 8

[36] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, June 2022. 5, 6, 7

[37] Andranik Sargsyan, Shant Navasardyan, Xingqian Xu, and Humphrey Shi. Mi-gan: A simple baseline for image inpainting on mobile devices. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7335–7345, 2023. 2

[38] Ziyi Shen, Wenguan Wang, Xiankai Lu, Jianbing Shen, Haibin Ling, Tingfa Xu, and Ling Shao. Human-aware motion deblurring. In *ICCV*, 2019. 8

[39] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan 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*, pages 1874–1883, 2016. 3

[40] Maitreya Suin, Kuldeep Purohit, and AN Rajagopalan. Distillation-guided image inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2481–2490, 2021. 1, 2

[41] Maitreya Suin, Kuldeep Purohit, and A. N. Rajagopalan. Spatially-attentive patch-hierarchical network for adaptive motion deblurring. In *CVPR*, 2020. 8

[42] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2149–2159, 2022. 1, 2, 5, 6

[43] Fu-Jen Tsai, Yan-Tsung Peng, Yen-Yu Lin, Chung-Chi Tsai, and Chia-Wen Lin. Stripformer: Strip transformer for fast image deblurring. In *European Conference on Computer Vision*, pages 146–162. Springer, 2022. 8

[44] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 5

[45] Ziyu Wan, Jingbo Zhang, Dongdong Chen, and Jing Liao. High-fidelity pluralistic image completion with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4692–4701, 2021. 5, 7

[46] Jingyuan Xu, Hongtao Xie, Chuanbin Liu, Fang Yang, Sicheng Zhang, Xun Chen, and Yongdong Zhang. Hip landmark detection with dependency mining in ultrasound image. *IEEE Transactions on Medical Imaging*, 40(12):3762–3774, 2021. 2

[47] Li Xu, Jimmy S Ren, Ce Liu, and Jiaya Jia. Deep convolutional neural network for image deconvolution. *Advances in neural information processing systems*, 27, 2014. 2

[48] Zhaoyi Yan, Xiaoming Li, Mu Li, Wangmeng Zuo, and Shiguang Shan. Shift-net: Image inpainting via deep feature rearrangement. In *Proceedings of the European conference on computer vision (ECCV)*, pages 1–17, 2018. 2

[49] Zili Yi, Qiang Tang, Shekoofeh Azizi, Daesik Jang, and Zhan Xu. Contextual residual aggregation for ultra high-resolution image inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7508–7517, 2020. 1, 2, 8

[50] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5505–5514, 2018. 2, 5, 6

[51] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4471–4480, 2019. 5, 6

[52] Tao Yu, Zongyu Guo, Xin Jin, Shilin Wu, Zhibo Chen, Weiping Li, Zhizheng Zhang, and Sen Liu. Region normalization for image inpainting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 12733–12740, 2020. 1, 2

[53] Yingchen Yu, Fangneng Zhan, Shijian Lu, Jianxiong Pan, Feiying Ma, Xuansong Xie, and Chunyan Miao. Wavefill: A wavelet-based generation network for image inpainting. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 14114–14123, 2021. 5, 6, 7

[54] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5728–5739, 2022. 1, 2, 6, 7, 8

[55] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. In *CVPR*, 2021. 8

[56] Hongguang Zhang, Yuchao Dai, Hongdong Li, and Piotr Koniusz. Deep stacked hierarchical multi-patch network for image deblurring. In *CVPR*, 2019. 8

[57] Kaihao Zhang, Wenhan Luo, Yiran Zhong, Lin Ma, Bjorn Stenger, Wei Liu, and Hongdong Li. Deblurring by realistic blurring. In *CVPR*, 2020. 8

[58] Lei Zhao, Qihang Mo, Sihuan Lin, Zhizhong Wang, Zhiwen Zuo, Haibo Chen, Wei Xing, and Dongming Lu. Uctgan: Diverse image inpainting based on unsupervised cross-space translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5741–5750, 2020. 1

[59] Chuanxia Zheng, Tat-Jen Cham, Jianfei Cai, and Dinh Phung. Bridging global context interactions for high-fidelity image completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11512–11522, 2022. 5

[60] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017. 2, 5, 7

[61] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*, 2024. 3