MxT: Mamba x Transformer for Image Inpainting

Shuang Chen
shuang.chen@durham.ac.uk
Amir Atapour-Abarghouei
amir.atapour-abarghouei@durham.ac.uk
Haozheng Zhang
haozheng.zhang@durham.ac.uk
Hubert P. H. Shum†
hubert.shum@durham.ac.uk

The department of Computer Science
Durham University
Durham, UK

†Corresponding Author: Hubert P. H. Shum

Abstract

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 face computational inefficiencies and struggle to maintain fine-grained details. To overcome these challenges, we introduce M×T composed of the proposed Hybrid Module (HM), which combines Mamba with the transformer in a synergistic manner. 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. Our HM facilitates dual-level interaction learning at both pixel and patch levels, greatly enhancing the model to reconstruct images with high quality and contextual accuracy. We evaluate M×T on the widely-used CelebA-HQ and Places2-standard datasets, where it consistently outperformed existing state-of-the-art methods.

1 Introduction

Image inpainting, as known as image completion, aims at restoring missing or damaged parts of images with semantically plausible context. This demands accurate modeling of both global and local information within the corrupted image, which is crucial as the global and local interaction maintains the coherence of both the content and style of the missing areas, ensuring seamless integration with the surrounding image regions [42].

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 [17, 46]. To solve this issue, recent advancements [9, 17] 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 [36, 48], 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 [7], merging from the domain of long-sequence modeling, 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.

In this paper, we introduce M×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. Our comparative experiments demonstrate that our M×T outperforms existing state-of-the-art methods on two wildly used datasets, CelebA-HQ and Places2. We summarize our contributions as: 1) We propose M×T to introduce Mamba combined with transformer for image inpainting. 2) We design a novel Hybrid Module to capture the feature interaction at both the pixel level and patch level. 3) Our M×T overall suppress the state-of-the-art methods on both CelebA-HQ and Places2 datasets. 4) M×T is able to adapt to high-resolution images with only training on low-resolution data.

2 Related Work

2.1 Image Inpainting

Image inpainting is an ill-posed low-level vision task that aims to infer the missing regions of the image via the undamaged pixels. Conventional works employ diffusion-based approaches to inpaint the missing regions by deriving by neighbouring visible pixels [32], or filling the missing areas by using the good match patches from the background or external sources such as the depth information [1, 2]. Although these methods can effectively complete small missing regions, they face challenges in precisely reconstructing more complex scenes due to limited global understanding of the image. Recently, deep learning based image inpainting studies propose to use CNN-based encoder-decoder architecture [33, 39, 44] or CNN-based Generative Adversarial Networks (GANs) [13, 28, 31, 38, 41, 42], which significantly improves the visual plausibility and diversity of inpainted images. However, the limited convolutional receptive field hinders the model's learning of long-range dependencies, which motivates studies on expanding the receptive field by applying the frequency convolution techniques [5, 34] or developing transformer-based models [3, 9, 17, 19]. Nonetheless, practical training and application of the transformer-based models are still constrained by the exponential complexity of self-attention calculations. In particular, it is still challeng-

ing to use pixel-level self-attention to achieve relatively high-resolution image inpainting. To this end, we turn our eyes to the field of selective SSM (i.e., Mamba [7]) for achieving image long-range pixel-wise dependency learning with robust spatial awareness.

2.2 SSMs in Computer Vision

Recently, State Space Models (SSMs) have demonstrated promising advantages of long sequence modeling and linear-time complexity in Natural Language Processing (NLP) [9]. This work specifically tackles the problem of vanishing gradients in SSMs when solving the exponential function by the linear first-order Ordinary Differential Equations [8]. Building on the rigorous theoretical proofs of the HiPPO framework that enables SSMs to capture long-range dependencies, Gu et al. [7] 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 [50] and V-Mamba [22], propose visual SSMs designs for computer vision tasks including image classification and object detection [22, 50]. However, their performance is still behind of the state-of-the-art transformer-based models like SpectFormer [27], SVT [26], and WaveViT [40]. U-Mamba [24] 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.

3 Preliminary

3.1 State Space Models and Mamba

The State Space Model (SSM) is generally known as a linear time-invariant system that maps a 1-dimensional input sequence $x(t) \in \mathbb{R}$ to a response $y(t) \in \mathbb{R}$ via a hidden latent state $h(t) \in \mathbb{R}^N$ (Eq.1). For efficient linear-complexity deep learning model training, the structured SSMs employ a zero-order hold discretization rule (Eq.2) to transform the continues parameter $(\Delta, \boldsymbol{A}, \boldsymbol{B})$ to discrete parameters $(\overline{\boldsymbol{A}}, \overline{\boldsymbol{B}})$, as shown in Eq.3.

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

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

$$h_t = \overline{\boldsymbol{A}}h_{t-1} + \overline{\boldsymbol{B}}x_t \quad (3a), \qquad y_t = \boldsymbol{C}h_t \quad (3b). \tag{3}$$

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 Δ is a time-scale parameter. Mamba [7], one of the most recent selective SSMs, introduces a gated selective mechanism to propagate or eliminate selected information based on the current state, significantly improving the content-reasoning performance. Specifically, Mamba changes the model from time-invariant to time-varying via converting parameters Δ, $\boldsymbol{B}$, $\boldsymbol{C}$ into input-dependent functions.

4 Method

The overall pipeline of the proposed M×T is illustrated in Fig. 1, which is a U-Net shape architecture formed with 7 Hybrid Blocks. Formally, the masked image $I_{masked} \in \mathbb{R}^{H\times W\times 3}$

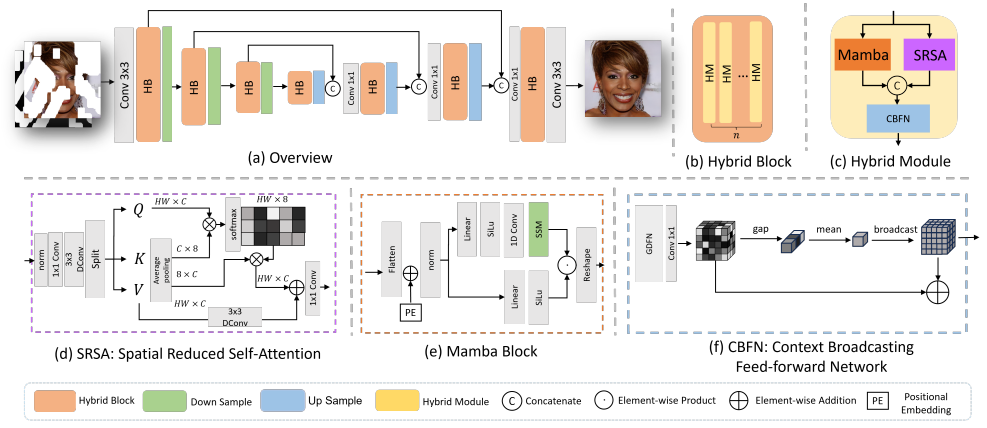

Figure 1: (a) The architecture overview of the proposed M×T. (b) The Hybrid Block is composed of n proposed Hybrid Modules. (c) The proposed Hybrid Module, consisted of a Mamba Block, a Spatial Reduced Self-Attention and a Context Broadcasting Feed-forward Network. (d) The Spatial Reduced Self-Attention provides spatial awareness. (e) The Mamba Block captures pixel-level interaction. (f) The Context Broadcasting Feed-forward Network transfers the features.

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 HBs 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 (HM), as shown in Fig. 1 (b), where n is the number of HMs. Each HM has a Transformer block, a Mamba block and a Context Broadcasting Feedforward Network (CBFN), which will be detailed in section 4.1.

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

$$F' = DConv_{3\times 3}(Conv_{1\times 1}(LayerNorm(F))). \tag{4}$$

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 [57] to adopt average pooling for K and V to a fixed dimension.

$$\begin{aligned} K', V' &= AvgPool(K), AvgPool(V), \\ Att &= softmax(K' \cdot Q), \end{aligned} \tag{5}$$

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 [29], 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:

$$\begin{aligned} LE(V) &= DConv_{3\times3}(V), \\ Output_{srsa} &= LE(V) + Att \cdot V', \end{aligned} \tag{6}$$

Mamba with Positional Embedding. Mamba showcases a strong capacity to handle long sequence data with linear computational complexity, making it highly effective for modeling interactions between adjacent pixels. In this work, we propose leveraging the Mamba module to modeling 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 $L = H \times W$:

$$F' = transpose(reshape(F,(B,C,L))). \tag{7}$$

Subsequently, we introduce cosine positional embedding [55] to the transformed feature, enhancing the capacity to maintain positional awareness:

$$F'' = F' + PE(L). \tag{8}$$

After applying layer normalization, mamba implements a gated mechanism to further refine the feature representation. The body branch involves a linear layer, a SiLU activation function [11], $1D$ convolutional layer and the SSM (State Space Sequence Models) layer.

$$F_{body} = SSM(Conv1D(SiLU(Linear(F'')))) \tag{9}$$

The gate branch involves a linear layer and a SiLU activation function [11]. After the gate branch re-weight the body branch, the output will be reshaped to the shape of (B,C,H,W):

$$\begin{aligned} G &= SiLU(Linear(F'')), \\ F_{gated} &= G \cdot F_{body}, \\ Output_{mamba} &= reshape(F_{gated},(B,C,H,W)), \end{aligned} \tag{10}$$

where G is the gate matric, 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 Feed-forward Network (CBFN) by improving the Gated-Dconv Feed-Forward Network (GDFN) [46].

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

$$F' = F + broadcast(\mu). \tag{11}$$

This global processing is designed to facilitate the learning of dense interactions within the self-attention layers [12], thereby enhancing the effectiveness of the Hybrid Module.

4.2 Loss Functions

To achieve superior inpainting outcomes, we adopt a multi-component loss strategy as delineated in the previous research [4, 18, 25]. 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:

$$\mathcal{L}_{all}(I_{out}, I_{gt}) = \alpha_1 \mathcal{L}_1 + \alpha_2 \mathcal{L}_{\text{style}} + \alpha_3 \mathcal{L}_{\text{perc}} + \alpha_4 \mathcal{L}_{\text{adv}}, \tag{12}$$

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 Experiment Results

Datasets. We evaluate our M×T on two diverse datasets, CelebA-HQ [14] and Places2-standard [49], 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 [4, 10, 18] to conduct all experiments with the widely used irregular mask [20] in three mask ratios.

Implementation Details. All experiments are carried out on one Nvidia A100 GPU. During training, we adopt the Adam optimiser [15] with $\beta_1 = 0.9$, $\beta_2 = 0.999$. The learning rate is set to 1×10^{-4} and the batch size is 4. In the Hybrid Blocks, the numbers of Hybrid Modules are [4,6,6,8,6,6,4].

Evaluation Metrics. We followed [4, 18] to evaluate the image generation quality with Peak Signal-to-Noise ratio (PSNR), Structural similarity (SSIM), L1, Frechet inception distance (FID) and Perceptual Similarity (LPIPS). We use these metrics as they offer insights into the quality of the images generated by the model. PSNR, SSIM, and L1 metrics evaluate reconstruction quality at the pixel level, assessing fine-grained details and structural context. FID quantifies the distributional differences between generated images and the original dataset. LPIPS is employed to reflect differences in human perception.

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 [■]	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 [■]	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 [■]	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 [■]	-	-	-	-	-	-	-	-	-	-	21.8321	0.7791	3.9427	8.9637	0.1943
LaMa [■]	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 [■]	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 [■]	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 [■]	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	**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**

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[■]	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[■]	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 [■]	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 [■]	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 [■]	-	-	-	-	-	-	-	-	-	-	19.6476	0.7052	4.6895	27.3619	0.2675
Stable Diffusion*	-	-	-	-	-	-	-	-	-	-	19.4812	0.7185	4.5729	27.8830	0.2416
WNet [■]	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 [■]	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 [■]	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 [■]	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	**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

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

Table 1: Quantitative comparison against state-of-the-art methods on CelebA-HQ (top), and Places2 (bottom). The best and second-best are indicated by **Bold** and underline, respectively.

5.1 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 1, our M×T outperforms in all metrics across different mask ratios. Especially on CelebA-HQ, at the increasing mask ratios, M×T improves PSNR by 2.0%, 2.3% and 1.8% respectively, and decreases LPIPS by 12.3%, 11.6% and 10.5% respectively. Moreover, for Places2, our model demonstrated comparable performance to SOTAs such as MISF and LAMA. While our training utilized the Places2-Standard dataset with 1.8 million images, MISF and LAMA were trained on the Places2-Challenge dataset, which contains 8 million images. Despite using only 22.5% of the images employed by these benchmarks, our model achieved comparable results, showcasing its robustness and efficiency.

Qualitative Comparison. The qualitative comparisons are presented in Fig. 2. For human face samples, M×T maintains consistency from the visible regions to the missing regions, such as effectively reconstructing elements like a missing hat. Additionally, M×T renders features like eyes with improved fine-grained details, showcasing its strong capability in learning complex representations. In the Places2 dataset, M×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×T has superior spatial perceptions.

Efficiency Comparison. Our M×T efficiently reduces spatial dimensions within the transformer and leverages Mamba's inherent capabilities, both achieving linear complexity. This synergy optimizes operational simplicity while enhancing effectiveness. As shown in Tab. 3, our model comprises 180 million parameters and achieves 110 ms to infer one image, making it suitable for real-time applications.

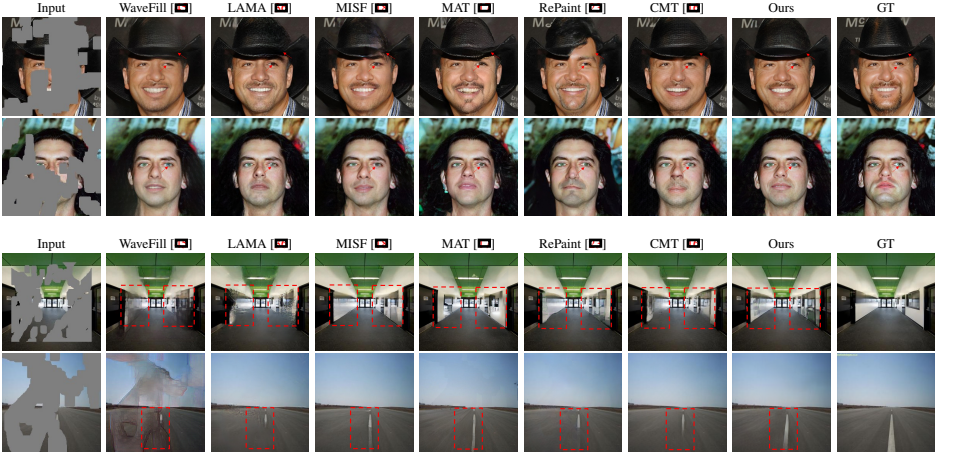

Figure 2: Visual comparisons at (256×256) resolution against the state-of-the-art methods on CelebA-HQ [14] (first two rows) and Places2 [49] (last two rows).

	Components				0.01%-20%					20%-40%					40%-60%				
	MB	SRSA	GDFN [46]	CBFN	PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓	PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓	PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓
(a)					33.5812	0.9537	0.5385	1.4877	0.0513	25.8971	0.8527	1.9786	4.4025	0.1480	21.6134	0.7308	4.1254	8.1732	0.2464
(b)	✓		✓		33.7640	0.9567	0.5244	1.4604	0.0425	26.1093	0.8736	1.8007	4.3721	0.1236	21.8573	0.7614	3.8649	8.0315	0.2197
(c)		✓	✓		33.7408	0.9598	0.5295	1.4181	0.0422	26.1274	0.8760	1.8092	4.3577	0.1196	21.8914	0.7682	3.8587	7.9974	0.2157
(d)	✓	✓	✓		33.9042	0.9610	0.5129	1.4362	0.0419	26.2847	0.8768	1.7498	4.3115	0.1178	22.1377	0.7687	3.7679	7.9910	0.2067
(e)	✓	✓		✓	**34.1393**	**0.9618**	**0.5010**	**1.3896**	**0.0411**	**26.3231**	**0.8780**	**1.7043**	**4.2927**	**0.1170**	**22.1704**	**0.7699**	**3.6337**	**7.9905**	**0.2053**

Table 2: Ablation studies of each component. MB is the Mamba Block with positonal embedding. SRSA is the Spatial Reduced Self-Attention. GDFN is the feed-forward network in [46]. CBFN is the Context Broadcasting Feed-forward Network. Our M×T corresponds to configuration (e).

5.2 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. 2. 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 optimize GDFN to CBFN. The overall metrics are further improved. All

Model	Wavefill [58]	WNet [62]	MISF [36]	MAT [39]	LAMA [64]	CMT [41]	SD [51]	LDM [51]	Repaint [46]	Ours
Param $\times 10^6$	49	46	26	62	51	143	860	387	552	180
Infer. Time (ms)	70	35	10	70	25	60	880	6000	250000	110

Table 3: Comparison of parameter count and inference time from the evaluations conducted on 256×256 images.

Figure 3: Illustration of the application on real-world high-resolution images with resolution of 2560 × 1920.

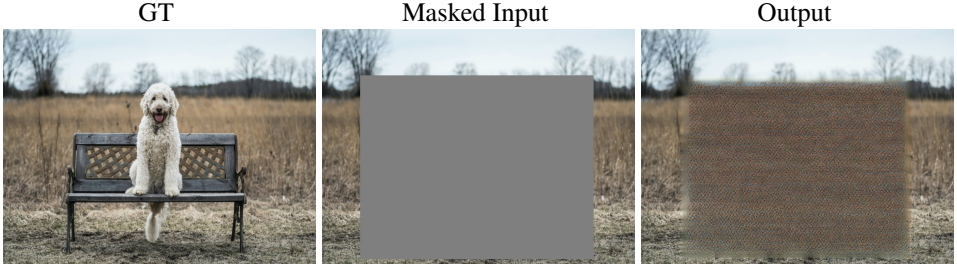

Figure 4: Illustration of a fail case with a much larger mask ratio.

ablation experiments are trained for 30K iterations. In addition, we followed [6] to build a light M×T with a halved parameter for an efficient evaluation.

5.3 Application: High-Resolution Image Inpainting

Our M×T is designed with linear computational complexity, enabling it to effectively handle high-resolution image inpainting tasks. We directly apply our model, pre-trained on the Places2-standard dataset, to real-world high-resolution images, demonstrating its capability. The example is illustrated in Fig. 3.

5.4 Limitation and Discussion

Fail Cases. Our model is trained using a wildly used irregular mask dataset [21], where the largest mask ratio is 60%, which restricts the ability to effectively handle images with much larger missing regions., particularly when the missing regions are concentrated in large shapes, such as very large rectangular masks (as shown in Fig. 4).

Future Work. In this work, our goal is to develop an inpainting model capable of reconstructing high-quality images, emphasizing fine-grained details and contextually plausible structures. Moving forward, our next objective is to enhance its capability by integrating multimodal foundational models, such as CLIP, to make the inpainting process controllable through text guidance, allowing users to influence the reconstruction results with descriptive input, while maintaining high quality.

6 Conclusion

In this paper, we introduce M×T, a hybrid model for image inpainting designed to reconstruct high-quality images with fine-grained details and spatial coherence. 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×T to maintain linear computational complexity, which is particularly advantageous for handling high-resolution images. We validate M×T on the widely-used CelebA-HQ and Places2 datasets, where it demonstrated superior or comparable performance to existing state-of-the-art methods.

7 Acknowledgement

This research is supported in part by the EPSRC NortHFutures project (ref: EP/X031012/1).

References

[1] Amir Atapour-Abarghouei, Gregoire Payen de La Garanderie, and Toby 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)*, pages 2813–2818, 2016.

[2] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. Patchmatch: A randomized correspondence algorithm for structural image editing. *ACM Trans. Graph.*, 28(3):24, 2009.

[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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12299–12310, 2021.

[4] 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.

[5] 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.

[6] 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.

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

[8] 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.

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

[10] 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.

[11] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

[12] Nam Hyeon-Woo, Kim Yu-Ji, Byeongho Heo, Dongyoon Han, Seong Joon Oh, and Tae-Hyun Oh. Scratching visual transformer's back with uniform attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5807–5818, 2023.

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

[14] 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.

[15] Diederik P Kingma, J Adam Ba, and J Adam. A method for stochastic optimization. arxiv 2014. *arXiv preprint arXiv:1412.6980*, 106, 2020.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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 *The European Conference on Computer Vision (ECCV)*, 2018.

[22] 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.

[23] 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.

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

[25] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Z Qureshi, and Mehran Ebrahimi. Edgeconnect: Generative image inpainting with adversarial edge learning. *arXiv preprint arXiv:1901.00212*, 2019.

[26] Badri Patro and Vijay Agneeswaran. Scattering vision transformer: Spectral mixing matters. *Advances in Neural Information Processing Systems*, 36, 2024.

[27] Badri N Patro, Vinay P Namboodiri, and Vijay Srinivas Agneeswaran. Spectformer: Frequency and attention is what you need in a vision transformer. *arXiv preprint arXiv:2304.06446*, 2023.

[28] 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.

[29] Sucheng Ren, Daquan Zhou, Shengfeng He, Jiashi Feng, and Xinchao Wang. Shunted self-attention via multi-scale token aggregation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10853–10862, 2022.

[30] 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.

[31] 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.

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

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8(3):415–424, 2022.

[38] 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.

[39] 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.

[40] Ting Yao, Yingwei Pan, Yehao Li, Chong-Wah Ngo, and Tao Mei. Wave-vit: Unifying wavelet and transformers for visual representation learning. In *European Conference on Computer Vision*, pages 328–345, 2022.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] 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.

[47] Ruisong Zhang, Weize Quan, Yong Zhang, Jue Wang, and Dong-Ming Yan. W-net: Structure and texture interaction for image inpainting. *IEEE Transactions on Multimedia*, 2022.

[48] 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.

[49] 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.

[50] 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.