BrushNet: A Plug-and-Play Image Inpainting Model with Decomposed Dual-Branch Diffusion

Xuan Ju[1,2], Xian Liu[1,2], Xintao Wang[1⋆], Yuxuan Bian[2], Ying Shan[1], and Qiang Xu[2⋆]

[1]ARC Lab, Tencent PCG [2]The Chinese University of Hong Kong

https://github.com/TencentARC/BrushNet

Fig. 1: Performance comparisons of *BrushNet* and previous image inpainting methods across various inpainting tasks: (I) Random Mask ($<50\%$ masked), (II) Random Mask ($>50\%$ masked), (III) Segmentation Mask Inside-Inpainting, (IV) Segmentation Mask Outside-Inpainting. Each group of results contains an artificial image (left) and a natural image (right) with 6 inpainting methods: (b) Blended Latent Diffusion (BLD) [1], (c) Stable Diffusion Inpainting (SDI) [33], (d) HD-Painter (HDP) [25], (e) PowerPaint (PP) [56], (f) ControlNet-Inpainting (CNI) [51], and (g) Ours.

Abstract. Image inpainting, the process of restoring corrupted images, has seen significant advancements with the advent of diffusion models (DMs). Despite these advancements, current DM adaptations for inpainting, which involve modifications to the sampling strategy or the development of inpainting-specific DMs, frequently suffer from semantic inconsistencies and reduced image quality. Addressing these challenges, our work introduces a novel paradigm: the division of masked image features and noisy latent into separate branches. This division dramatically diminishes the model's learning load, facilitating a nuanced incorporation of essential masked image information in a hierarchical fashion. Herein, we present *BrushNet*, a novel plug-and-play dual-branch model engineered to embed pixel-level masked image features into any pre-trained DM, guaranteeing coherent and enhanced image inpainting outcomes. Additionally, we introduce *BrushData* and *BrushBench* to facilitate segmentation-based inpainting training and performance assessment. Our extensive experimental analysis demonstrates BrushNet's superior performance over existing models across seven key metrics, including image quality, mask region preservation, and textual coherence.

Keywords: Image Inpainting · Diffusion Models · Image Generation

⋆ Corresponding author.

1 Introduction

Image inpainting [45] aims at restoring the missing regions of an image while maintaining the overall coherence. As a long-standing computer vision problem, it facilitates numerous applications such as virtual try-on [18] and image editing [15]. Recently, diffusion models [12, 36] have demonstrated impressive performance in image generation, enabling flexible user control with semantic and structural conditions [33, 51]. To this end, researchers resort to diffusion-based pipelines for high-quality image inpainting that aligns with given text prompts.

Commonly used diffusion-based text-guided inpainting methods can be roughly divided into two categories: (1) *Sampling strategy modification* [1, 2, 6, 20, 23, 47, 50], which modifies the standard denoising process by sampling the masked regions from a pre-trained diffusion model, and the unmasked areas are simply copy-pasted from the given image in each denoising step. Although they can be used in arbitrary diffusion backbones, the limited perceptual knowledge of mask boundaries and the unmasked image region context leads to incoherent inpainting results. (2) *Dedicated inpainting models* [4, 33, 37, 42, 43, 46, 49, 56], which fine-tune a specially designed image inpainting model by expanding the input channel dimension of base diffusion models to incorporate provided corrupted image and mask. While they enable the diffusion model to generate more satisfying results with specialized content-aware and shape-aware models, we argue, **is this architecture the best fit for diffusion-based inpainting?**

As shown in Fig. 2, dedicated inpainting models fuse noisy latent, masked image latent, mask, and text at an early stage. This architectural design makes the masked image feature easily influenced by the text embedding, preventing subsequent layers in the UNet from obtaining pure masked image features due to the influence of text. Additionally, handling the condition and generation in a single branch imposes extra burdens on the UNet framework. These approaches also necessitate fine-tuning in different variations of diffusion backbones, which can be time-consuming with limited transferability.

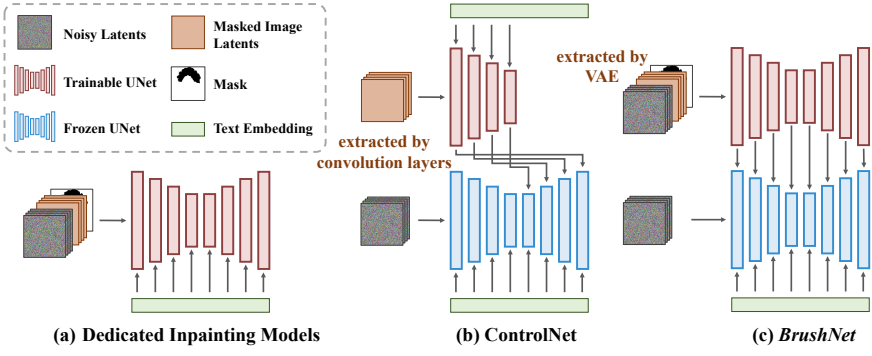

Fig. 2: Comparison of previous inpainting architectures and *BrushNet*.

Adding an additional branch dedicated to masked image feature extraction could be a promising approach to tackle the issue above. However, the existing solutions such as ControlNet [51] lead to inadequate information extraction

and insertion when directly applied to inpainting, which originates from differences between controllable image generation and inpainting: inpainting requires pixel-to-pixel constraints with strong restrictive information instead of sparse structural control relying on text for content completion. As a result, ControlNet yields unsatisfactory results compared to dedicated inpainting models.

To address this issue, we propose *BrushNet*, which introduces an additional branch to the diffusion framework, creating a more suitable architecture for image inpainting. Specifically, our designs are three-fold: (1) To improve the extraction of image features for adaptation to the UNet distribution, we use a VAE encoder instead of randomly initialized convolution layers to process the masked image. (2) To enable dense per-pixel control, we adopt a hierarchical approach by gradually incorporating the full UNet feature layer-by-layer into the pre-trained UNet. (3) To ensure pure image information is considered in the additional branch, we remove text cross-attention from UNet. This design further offers plug-and-play capabilities and flexible unmasked region controllability to the inpainting process. For better consistency and a larger range of unmasked region controllability, we additionally propose a blurred blending strategy.

To ensure a comprehensive evaluation for real-life applications, we categorize inpainting tasks into two distinct types based on mask shape: random brush masks and segmentation-based masks. We utilize EditBench [37] as the comparing benchmark for random brush mask inpainting. Additionally, we introduce a new training dataset *BrushData* and a new benchmark *BrushBench* for training and evaluating segmentation-based mask inpainting. Results show *BrushNet* achieve state-of-the-art performance across 7 metrics encompassing image quality, masked region preservation, and text alignment.

2 Related Work

Image inpainting is a classic problem in computer vision, aiming to restore masked regions of an image with plausible and natural content [29, 45]. Previous methods based on traditional techniques [3, 7], Variational Auto-Encoders (VAEs) [27, 54] and Generative Adversarial Networks (GANs) [21, 53, 55] often require auxiliary hand-engineered features but yield poor results. Recently, diffusion-based methods [1,2,20,23,31,42] gain popularity due to their impressive high-quality generation, fine-grained control, and output diversity [12, 13, 33].

Initial attempts at utilizing diffusion models for text-guided inpainting [1, 2, 6, 20, 23, 47, 50], such as Blended Latent Diffusion, modifies the standard denoising strategy by sampling the masked regions from a pre-trained diffusion model and the unmasked areas from the given image, which is commonly used as the default inpainting choice in widely-used image generation libraries like Diffusers [28]. Although these methods demonstrate satisfactory results in simple image inpainting tasks and can be plug-and-play to any diffusion model, they struggle with complex mask shapes, image contents, and text prompts, leading to results that lack coherence. This is primarily attributed to their limited perceptual knowledge of mask boundaries and the unmasked image region context.

Previous works [4,33,37,42,43,46,49,56] address this issue by fine-tuning the base models into content-aware and shape-aware models specifically designed

for image inpainting. Specifically, SmartBrush [42] augments the diffusion U-Net with object-mask prediction to guide the sampling process with mask boundaries information. Stable Diffusion Inpainting [33] fine-tunes a diffusion model specifically designed for inpainting tasks, taking the mask, masked image, and noisy latent as inputs to UNet architecture. HD-Painter [25] and PowerPaint [56] build upon the foundation of Stable Diffusion Inpainting, separately enhancing the generation quality and enabling the model to perform multiple tasks.

However, these approaches make it difficult to effectively transfer their inpainting ability to arbitrary pre-trained models, restricting their applicability. To enable any diffusion model with inpainting capabilities, the community fine-tunes ControlNet [51] on inpainting image pairs. However, the model design of ControlNet exhibits limitations in its perceptual understanding of masks and masked images, which consequently leads to unsatisfactory outcomes. Compared with previous methods (shown in Tab. 1), *BrushNet* is plug-and-play, content-aware, and shape-aware, with a flexible preserving degree for unmasked regions.

Table 1: Comparison of *BrushNet* with Previous Image Inpainting Methods. *BrushNet* offers the advantage of being plug-and-play with any pretrained diffusion model. Moreover, it allows for flexible control over the scale of inpainting and is designed to be aware of both the mask shape and the unmasked content. Note that we only list commonly used text-guided diffusion methods in this table.

Model	Plug-and-Play	Flexible-Scale	Content-Aware	Shape-Aware
Blended Diffusion [1, 2]	✓			
SmartBrush [42]				✓
SD Inpainting [33]			✓	✓
PowerPaint [56]			✓	✓
HD-Painter [25]			✓	✓
ReplaceAnything [4]			✓	✓
Imagen [37]			✓	✓
ControlNet-Inpainting [51]	✓	✓	✓	
BrushNet	✓	✓	✓	✓

3 Preliminaries and Motivation

In this section, we will first introduce diffusion models in Sec. 3.1. Then, Sec. 3.2 would review previous inpainting techniques based on sampling strategy modification and special training. Finally, the motivation is outlined in Section 3.3.

3.1 Diffusion Models

Diffusion models include a forward process that adds Gaussian noise ϵ to convert clean sample z_0 to noise sample z_T, and a backward process that iteratively performs denoising from z_T to z_0, where $\epsilon \sim \mathcal{N}(0, 1)$, and T represents the total number of timesteps. The forward process can be formulated as:

$$z_t = \sqrt{\alpha_t} z_0 + \sqrt{1 - \alpha_t}\epsilon \tag{1}$$

z_t is the noised feature at step t with $t \sim [1, T]$, and α is a hyper-parameter.

In the backward process, given input noise z_T sampled from a random Gaussian distribution, learnable network ϵ_θ estimates noise at each step t conditioned on C. After T progressively refining iterations, z_0 is derived as the output sample:

$$z_{t-1} = \frac{\sqrt{\alpha_{t-1}}}{\sqrt{\alpha_t}} z_t + \sqrt{\alpha_{t-1}} \left(\sqrt{\frac{1}{\alpha_{t-1}} - 1} - \sqrt{\frac{1}{\alpha_t} - 1} \right) \epsilon_\theta \left(z_t, t, C \right) \tag{2}$$

The training of diffusion models revolves around optimizing the denoiser network ϵ_θ to conduct denoising with condition C, guided by the objective:

$$\min_\theta E_{z_0, \epsilon \sim \mathcal{N}(0,I), t \sim U(1,T)} \left\| \epsilon - \epsilon_\theta \left(z_t, t, C \right) \right\| \tag{3}$$

3.2 Previous Inpainting Models

Sampling Strategy Modification. This line of research achieves inpainting by gradually blending masked images with the generated results. The most used method among them is Blended Latent Diffusion (BLD) [1], serving as the default choice for inpainting in widely-used diffusion-based image generation libraries (*e.g.*, Diffusers [28]). Given a binary mask m and a masked image x_0^{masked}, BLD first extracts the latent representation z_0^{masked} of the masked image using VAE. Subsequently, the mask m is resized to $m^{resized}$ to match the size of the latent representation. To formulate the inpainting process, BLD adds Gaussian noise to z_0^{masked} for T steps and gets z_t^{masked}, where $t \sim [1, T]$. Then, denoising steps start from z_T^{masked}, where each sampling step in eq. 2 is followed by:

$$z_{t-1} \leftarrow z_{t-1} \cdot \left(1 - m^{resized}\right) + z_{t-1}^{masked} \cdot m^{resized} \tag{4}$$

Despite its simplicity in implementation, BLD exhibits suboptimal performance in terms of both unmasked region preservation and generation content alignment. This is due to (1) the resize of the mask preventing it from correctly blending the noisy latent, (2) the diffusion model lacking perceptual knowledge of mask boundaries and the unmasked image region context.

Dedicated Inpainting Models. To enhance the performance of inpainting, previous works fine-tune the base models by expanding the input UNet channel to include the mask and masked image inputs, turning it into an architecture specifically designed for image inpainting. Though having better generation results compared with BLD, they still have several drawbacks: (1) These models merge the noisy latent, masked image latent, and mask at the initial convolution layer of the UNet architecture, where they are collectively influenced by the text embedding. Consequently, subsequent layers in the UNet model struggle to obtain pure masked image features due to the text's influence. (2) Incorporating both the condition processing and generation within a single branch places additional burdens on the UNet framework. (3) These approaches require extensive fine-tuning across various variations of diffusion backbones, which is computationally intensive and lacks transferability to custom diffusion models.

3.3 Motivation

Based on the analysis presented in Section 3.2, a more effective architecture design of inpainting would be introducing an additional branch specifically dedicated to masked image processing. ControlNet [51] is one of the widely adopted

strategies that exemplifies this idea. However, it should be noted that directly fine-tuning ControlNet, which is originally designed for controllable image generation, on the inpainting task yields unsatisfactory results. ControlNet designs a lightweight encoder to incorporate out-of-domain structural conditions (*e.g.*, skeleton) and relies on text guidance for content generation, which is unsuitable for pixel-level inpainting image feature injection. Furthermore, ControlNet typically relies on sparse control, meaning that merely adding control to residuals in the UNet framework would be sufficient, while inpainting requires pixel-to-pixel constraints with strong restrictive information. Thus a new architecture specifically designed for inpainting is urgently needed.

4 Method

An overview of *BrushNet* is shown in Fig. 3. We employ a dual-branch strategy for masked image guidance insertion (Sec. 4.1). Blending operations with a blurred mask is used to ensure better-unmasked region preservation (Sec. 4.2). Notably, *BrushNet* can achieve flexible control by adjusting the added scale.

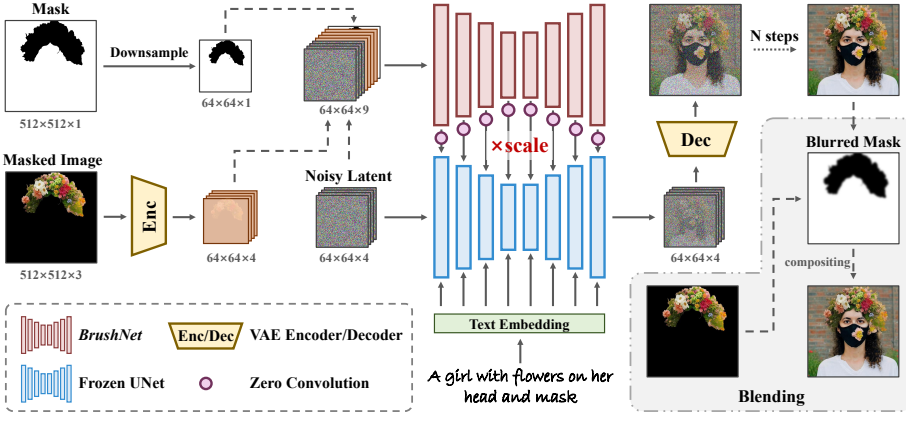

Fig. 3: Model overview. Our model outputs an inpainted image given the mask and masked image input. Firstly, we downsample the mask to accommodate the size of the latent, and input the masked image to the VAE encoder to align the distribution of latent space. Then, noisy latent, masked image latent, and downsampled mask are concatenated as the input of *BrushNet*. The feature extracted from *BrushNet* is added to pretrained UNet layer by layer after a zero convolution block [51]. After denoising, the generated image and masked image are blended with a blurred mask.

4.1 Masked Image Guidance

The insertion of the masked image feature into the pre-trained diffusion network is accomplished through an additional branch, which explicitly separates the feature extraction of masked images from the image-generating processes. The input to the additional branch includes noisy latent, masked image latent, and downsampled mask, which are concatenated together to form the input.

Specifically, the noisy latent provides generation information during the present generation process, helping *BrushNet* enhance the semantic coherence of the masked image feature. The masked image latent is extracted from the masked image using VAE, which is aligned with the data distribution of the pre-trained UNet. To ensure the alignment of the mask size with the noisy latent and masked image latent, we employ cubic interpolation to downsample the mask.

To process the masked image features, *BrushNet* utilizes a clone of the pre-trained diffusion model while excluding its cross-attention layers. The pre-trained weights of the diffusion model serve as a strong prior for extracting the masked image features, while the removal of the cross-attention layers ensures that only pure image information is considered within this additional branch. *BrushNet* feature is inserted into the frozen diffusion model layer-by-layer, enabling dense per-pixel control hierarchically. Similar to ControlNet [51], we employ zero convolution layers to establish a connection between the locked model and the trainable *BrushNet*. This ensures that harmful noise does not influence the hidden states in the trainable copy during the initial stages of training.

The feature insertion operation is shown in Eq. 5. Specifically, $\epsilon_\theta\left(z_t, t, C\right)_i$ indicates the feature of the i-th layer in network ϵ_θ with $i \sim [1, n]$, where n is the number of layers. The same notation applies to $\epsilon_\theta^{BrushNet}$. $\epsilon_\theta^{BrushNet}$ takes the concatenated noisy latent z_t, masked image latent z_0^{masked}, and downsampled mask $m^{resized}$ as input, with the concatenation operation denoted as $[\cdot]$. $\mathcal{Z}$ is the zero convolution operation. w is the preservation scale used to adjust the influence of *BrushNet* on pretrained diffusion model.

$$\epsilon_\theta\left(z_t, t, C\right)_i = \epsilon_\theta\left(z_t, t, C\right)_i + w \cdot \mathcal{Z}\left(\epsilon_\theta^{BrushNet}\left(\left[z_t, z_0^{masked}, m^{resized}\right], t\right)_i\right) \quad (5)$$

4.2 Blending Operation

As mentioned in Section 4.2, the blending operation conducted in latent space can result in inaccuracies due to the resizing of the mask. Similarly, in our approach, a similar issue arises as we resize the mask to match the size of the latent space, which can introduce potential inaccuracies. Additionally, it is important to acknowledge that VAE encoding and decoding operations have inherent limitations and may not ensure complete image reconstruction.

To ensure a fully consistent image reconstruction of the unmasked region, previous works have explored different techniques. Some approaches [4, 56], utilize past-and-copy methods, where the unmasked region is directly copied from the original image. However, this can result in a lack of semantic coherence in the final generation results. On the other hand, adopting latent blending operations inspired by BLD [1, 33] has been observed to face challenges in effectively preserving the desired information in the unmasked regions.

In this work, we present a simple pixel space solution to address this issue by first blurring the mask and then performing copy-and-paste using the blurred mask. Although this approach may result in a slight loss of accuracy in preserving the details of the mask boundary, the error is nearly imperceptible to the naked eye and results in significantly improved coherence in the mask boundary.

4.3 Flexible Control

The architecture design of *BrushNet* inherently makes it suitable for seamless plug-and-play integration to various pretrained diffusion models and enables flexible preservation scale. Specifically, the flexible control of our proposed *BrushNet* includes: (1) Since *BrushNet* does not modify the weights of the pretrained diffusion model, it can be readily integrated as a plug-and-play component with any community fine-tuned diffusion models. This allows for easy adoption and experimentation with different pretrained models. (2) Preservation Scale Adjustment: The preservation scale of the unmasked region can be controlled by incorporating *BrushNet* features into the frozen diffusion model with the weight w. This weight determines the influence of *BrushNet* on the preservation scale, offering the ability to adjust the desired level of preservation. (3) Blurring Scale and Blending Operation: By adjusting the scale of blurring and deciding whether to apply the blending operation, the preservation scale of the unmasked region can be further customized. These features allow for flexible fine-grained control over the inpainting process. More explanation can be found in Sec. 5.5.

5 Experiments

5.1 Evaluation Benchmark and Metrics

Benchmark. Previous commonly used datasets in the image inpainting field include CelebA [22], CelebA-HQ [14], ImageNet [8], MSCOCO [19], Open Images [17], and LSUN-Bedroom [48]. However, these datasets either primarily focus on a small area, such as human faces, or predominantly consist of low-quality, cluttered real-life scene data. As a result, these datasets are not well-suited for training and evaluating diffusion-based inpainting models, which can generate high-quality diverse images that align with text prompts.

Recently proposed EditBench [37] serves as a benchmark specifically designed for text-guided image inpainting for diffusion models. This benchmark consists of a collection of 240 images comprising an equal ratio of natural images and generated images, with mask and caption annotation for each image. However, the annotated masks in EditBench are mostly random shapes without specific object information, neglecting the practical application of inpainting in real scenarios such as replacing an object with an external mask, as commonly seen in E-commerce product displays and image editing.

To fill the gap, we propose *BrushBench* for segmentation-based inpainting, as shown in Fig. 4. *BrushBench* comprises a total of 600 images, with each image accompanied by the human-annotated mask and caption annotation. The images in *BrushBench* are evenly distributed between natural images and artificial images, such as paintings. Furthermore, the dataset ensures an equal distribution among different categories, including humans, animals, indoor scenarios, and outdoor scenarios. This balanced distribution enables a fair evaluation across various categories, promoting better evaluation equity.

To further enhance the analysis of the inpainting task, we categorize it into two distinct types based on the masks used: random brush masks and segmentation-based masks. We use EditBench as the comparison benchmark

Fig. 4: Benchmark overview. I and II separately show natural and artificial images, masks, and caption of *BrushBench*. (a) to (d) show images of humans, animals, indoor scenarios, and outdoor scenarios. Each group of images shows the original image, inside-inpainting mask, and outside-inpainting mask, with an image caption on the top. III show image, mask, and caption from EditBench [37], with (e) for generated images and (f) for natural images. The images are randomly selected from both benchmarks.

for random brush masks and use *BrushBench* for segmentation-based masks. Regarding inpainting with segmentation-based masks, we refine the task by considering two specific scenarios: segmentation mask inside-inpainting and segmentation mask outside-inpainting. By separating these two subtasks, we can better understand the inpainting performance in different image regions.

Dataset. To train segmentation-based mask inpainting, we annotate segmentation mask on Laion-Aesthetic [34] dataset, called *BrushData*. We employ the Grounded-SAM [32] to annotate open-world masks and then filter the masks based on their confidence score, ensuring that only masks with relatively higher confidence scores are retained. Additionally, we consider factors such as reasonable mask size and good mask continuity during the filtering process. [1]

Metrics. We consider 7 metrics from three aspects: image generation quality, masked region preservation, and text alignment.

- *Image Generation Quality.* Metrics most used by previous inpainting methods (*e.g.*, FID [11] and KID [5]) show a poor representation of the rich and varied content generated by modern text-to-image models [16]. Thus, we use Image Reward (**IR**) [44], HPS v2 (**HPS**) [41], and Aesthetic Score (**AS**) [34] that align with human perception. Specifically, ImageReward and HPS v2 are text-to-image human preference evaluation models trained on large-scale datasets of human preference choices on generated images. Aesthetic Score is a linear model trained on image quality rating pairs of real images.
- *Masked Region Preservation.* We follow previous works using standard Peak Signal-to-Noise Ratio (**PSNR**) [39], Learned Perceptual Image Patch Similarity (**LPIPS**) [52], and Mean Squared Error (**MSE**) [38] in the unmasked region among the generated image and the original image.

[1] The proposed *BrushData* and *BrushBench* will be released along with the codes.

- *Text Alignment.* We use CLIP Similarity (**CLIP Sim**) [40] to evaluate text-image consistency between the generated images and corresponding text prompts. CLIP Similarity projects text and images to the same shared space with the CLIP model [30] and evaluates the similarity of their embeddings.

5.2 Implementation Details

We perform the inference of different inpainting methods in the same setting unless specifically clarified, *i.e.*, on NVIDIA Tesla V100 following their open-source code with a base model of Stabe Diffusion v1.5 in 50 steps, with a guidance scale of 7.5. We keep the recommended hyper-parameter for each inpainting method in all images for fair comparison. *BrushNet* and all ablation models are trained for 430 thousands steps on 8 NVIDIA Tesla V100 GPUs, which takes around 3 days. For comparison on *BrushBench*, we use *BrushNet* trained on *BrushData*. For comparison on EditBench, we use the model trained on LAION-5B with random masks. Details can be found in the provided codes.

5.3 Quantitative Comparison

Table 2: Quantitative comparisons among *BrushNet* and other diffusion-based inpainting models in *BrushBench*: Blended Latent Diffusion (BLD) [1], Stable Diffusion Inpainting (SDI) [33], HD-Painter (HDP) [25], PowerPaint (PP) [56], and ControlNet-Inpainting (CNI) [51]. Metrics encompassing image quality, masked region preservation, and text alignment (Text Align) for inside-inpainting and outside-inpainting are shown in the table. All models use Stable Diffusion V1.5 as base model. **Red** stands for the best result, **Blue** stands for the second best result.

	Metrics	Image Quality			Masked Region Preservation			Text Align
	Models	IR$_{\times 10}$↑	HPS$_{\times 10^2}$↑	AS↑	PSNR↑	LPIPS$_{\times 10^3}$↓	MSE$_{\times 10^3}$↓	CLIP Sim↑
Inside	**BLD** [1]	9.78	25.87	6.17	21.33	9.76	49.26	26.15
	SDI [33]	11.72	27.06	6.50	21.52	13.87	48.39	26.17
	HDP [25]	11.68	26.90	6.42	22.61	9.95	43.50	26.37
	PP [56]	11.46	27.35	6.24	21.43	32.73	48.43	**26.48**
	CNI [51]	9.9	26.02	**6.53**	12.39	78.78	243.62	**26.47**
	CNI* [51]	11.21	26.92	6.39	**22.73**	24.58	**43.49**	26.22
	Ours	**12.36**	**27.40**	**6.53**	21.65	**9.31**	48.28	**26.48**
	Ours*	**12.64**	**27.78**	**6.51**	**31.94**	**0.80**	**18.67**	26.39
Outside	**BLD** [1]	7.81	26.77	6.23	15.85	35.86	21.40	26.73
	SDI [33]	10.27	27.99	**6.55**	18.04	19.87	15.13	27.21
	HDP [25]	9.66	27.79	6.46	18.03	22.99	15.22	26.96
	PP [56]	7.45	28.01	6.26	18.04	31.78	15.13	26.72
	CNI [51]	9.26	27.68	6.42	11.91	83.03	58.16	**27.29**
	CNI* [51]	9.57	27.76	6.28	17.50	37.72	19.95	26.92
	Ours	**10.82**	**28.02**	**6.64**	**18.06**	**22.86**	**15.08**	**27.33**
	Ours*	**10.88**	**28.09**	**6.64**	**27.82**	**2.25**	**4.63**	27.22

* with blending operation

Tab. 2 and Tab. 3 show the quantitative comparison on *BrushBench* and EditBench [37]. We compare the inpainting results of sampling strategy modification method Blended Latent Diffusion [1], dedicated inpainting models Stable Diffusion Inpainting [33], HD-Painter [25], and PowerPaint [56], as well as plug-and-play method ControlNet [51] trained on inpainting data.

Results demonstrate *BrushNet*'s effectiveness across image quality, masked region preservation, and image-text alignment. Blended Latent Diffusion [1] shows the poorest results in image quality and text alignment, derived from the incoherence between the generated masked and the unmasked given image. At the same time, its performance in masked region preservation is also not satisfactory because of the loss incurred by resized mask blending operation in the latent space. Modified from Stable Diffusion Inpainting [33], HD-Painter [25] and PowerPaint [56]demonstrate comparable performance to Stable Diffusion Inpainting in the task of inside-inpainting. However, when it comes to outside-inpainting, their results in terms of image quality and text alignment are significantly poorer compared to Stable Diffusion Inpainting, which can be attributed to their exclusive emphasis on the inside-inpainting task.

ControlNet [51] trained on inpainting has the most similar experimental configuration to ours. Due to the mismatch between its model design and the inpainting task, ControlNet shows poor results in masked region preservation and image quality, necessitating its combination with Blended Latent Diffusion [1] to generate satisfying inpainted images. However, even with this combination, it still falls short compared to dedicated inpainting models and *BrushNet*.

Table 3: Quantitative comparisons among *BrushNet* and other diffusion-based inpainting models in EditBench. A detailed explanation of compared methods and metrics can be found in the caption of Tab. 2. **Red** stands for the best result, **Blue** stands for the second best result.

Metrics	Image Quality			Masked Region Preservation			Text Align
Models	IR$_{\times 10}$↑	HPS$_{\times 10^2}$↑	AS↑	PSNR↑	LPIPS$_{\times 10^3}$↓	MSE$_{\times 10^3}$↓	CLIP Sim↑
BLD [1]	0.90	23.81	5.44	20.89	10.93	31.90	28.62
SDI [33]	1.86	24.24	5.69	23.25	6.94	24.30	28.00
HDP [25]	1.74	24.20	5.64	23.07	**6.70**	24.32	28.34
PP [56]	1.24	24.50	5.44	23.34	20.12	24.12	27.80
CNI [51]	1.49	24.46	**5.82**	12.71	69.42	159.71	28.16
CNI* [51]	0.90	23.79	5.46	22.61	35.93	26.14	27.74
Ours	**4.40**	**25.10**	**5.84**	**23.35**	6.81	**24.11**	**28.67**
Ours*	**4.46**	**25.24**	**5.82**	**33.66**	**0.63**	**10.12**	**28.87**

* with blending operation

The performance on the EditBench is roughly consistent with the overall performance on *BrushBench*, which similarly shows *BrushNet*'s superior performance. This indicates that our method exhibits strong performance across a range of inpainting tasks with various mask types, including random masks, inside-inpainting masks, and outside-inpainting masks.

5.4 Qualitative Comparison

The qualitative comparison with previous image inpainting methods is shown in Fig. 1. We provide results on artificial images and natural images across various inpainting tasks, including random mask inpainting, segmentation mask inside-inpainting, and segmentation mask outside-inpainting. *BrushNet* consistently

show exceptional results in the coherent of generated region and unmasked region, considering content (I, II right, III right, IV), color (II left), and text (III left). Interestingly, Fig. 1 III left requires the model to generate a cat and a goldfish. All previous methods fail to recognize that a goldfish is already present in the masked image, resulting in the generation of an additional fish within the masked region. *BrushNet* successfully realized the awareness of background information due to the design of the dual-branch decoupling.

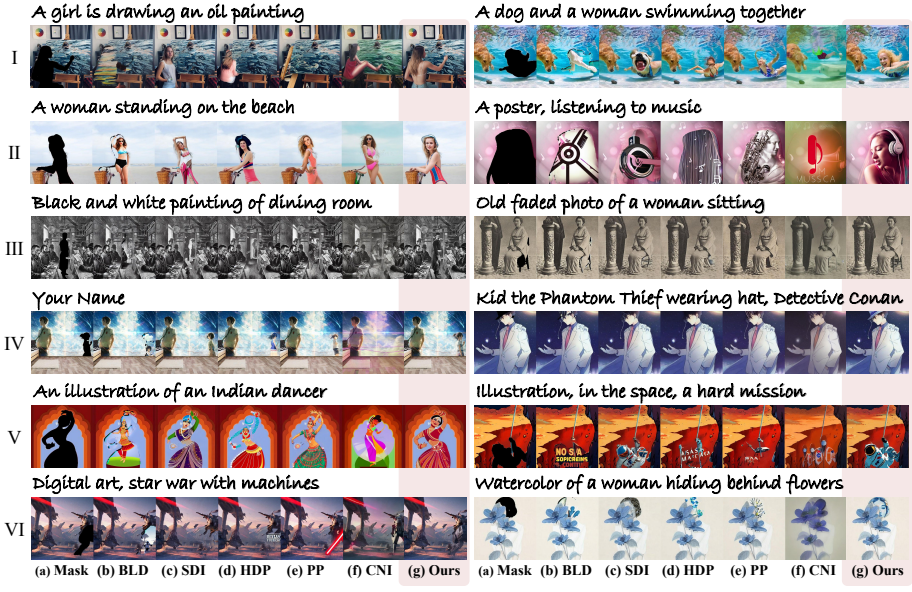

Fig. 5: Comparison of previous inpainting methods and *BrushNet* on various image domain. A detailed explanation of compared methods is in Fig. 1.

The untouched pre-trained diffusion branch also provides the advantage of better coverage across different data domains, such as painting and anime. As shown in Fig. 5, *BrushNet* demonstrates superior performance across various image categories, including natural image (I, II), pencil painting (III), anime (IV), illustration (V), digital art (VI left), and watercolor (VI right). Due to the page limit, more qualitative comparison results are in supplementary files.

5.5 Flexible Control Ability

Fig. 6 and Fig. 7 illustrate the flexible control provided by *BrushNet* in two aspects: base diffusion model selection and controlling scale. In Fig. 6, we showcase the ability to combine *BrushNet* with different diffusion models fine-tuned by the community. This allows users to select a specific model that best suits their inpainting requirements, enabling users to achieve the desired inpainting results based on their specific needs. Fig. 7 demonstrates the adjustment of the control scale of *BrushNet*. This control scale parameter allows users to effectively control the extent of unmasked region protection during the inpainting process. By manipulating the scale parameter, users can achieve fine-grained control over the inpainting process, enabling precise and customizable inpainting.

Fig. 6: Integrating *BrushNet* to community fine-tuned diffusion models. We use five popular community diffusion models fine-tuned from stable diffusion v1.5: DreamShaper (DS) [24], epiCRealism (ER) [9], Henmix_Real (HR) [10], MeinaMix (MM) [26], and Realistic Vision (RV) [35]. MM is specifically designed for anime images.

Fig. 7: Flexible control scale of *BrushNet*. (a) shows the given masked image, (b)-(h) show adding *BrushNet* with control scale w from 1.0 to 0.2. Results show a gradually diminishing controllable ability from precise to rough control.

5.6 Ablation Study

Table 4: Ablation on dual-branch design. Stable Diffusion Inpainting (SDI) use single-branch design, where the entire UNet is fine-tuned. We conducted an ablation analysis by training a dual-branch model with two variations: one with the base UNet fine-tuned, and another with the base UNet forzened. Results demonstrate the superior performance achieved by adopting the dual-branch design. **Red** is the best result.

Metrics	Image Quality			Masked Region Preservation			Text Align
Model	IR$_{\times 10}$↑	HPS$_{\times 10^2}$↑	AS↑	PSNR↑	LPIPS$_{\times 10^2}$↓	MSE$_{\times 10^3}$↓	CLIP Sim↑
SDI	11.00	27.53	6.53	19.78	16.87	31.76	26.69
w/o fine-tune	11.59	27.71	6.59	19.86	16.09	31.68	26.91
w/ fine-tune	**11.63**	**27.73**	**6.60**	**20.13**	**15.84**	**31.57**	**26.93**

We conducted ablation studies to investigate the impact of different model designs. Tab. 4 compares the dual-branch and single-branch designs. Tab. 5 shows the ablation study focusing on the additional branch architecture. The ablation studies are conducted on *BrushBench*, averaging the performance of inside-inpainting and outside-inpainting. The results presented in Tab. 4 demonstrate that the dual-branch design significantly outperforms the single-branch design. Additionally, fine-tuning the base diffusion model in dual-branch design yields better results compared to freezing it. However, fine-tuning the base diffusion model may restrict flexibility and control over the model. Considering this trade-off between performance and flexibility, we decide to adopt the frozen

dual-branch design as our model design. Tab. 5 presents the rationale behind the design choices for (1) using a VAE encoder instead of randomly initialized convolution layers to process the masked image, (2) incorporating the full UNet feature layer-by-layer into the pre-trained UNet, (3) removing text cross-attention in *BrushNet*, which avoids masked image features influenced by text.

Table 5: Ablation on model architecture. We ablate on the following components: the image encoder (Enc), selected from a random initialized convolution (Conv) and a VAE; the inclusion of mask in input (Mask), chosen from adding (w/) and not adding (w/o); the presence of cross-attention layers (Attn), chosen from adding (w/) and not adding (w/o); the type of UNet feature addition (UNet), selected from adding the full UNet feature (full), adding half of the UNet feature (half), and adding the feature like ControlNet (CN); and finally, the blending operation (Blend), chosen from not adding (w/o), direct pasting (paste), and blurred blending (blur). **Red** is the best result.

Metrics					Image Quality			Masked Region Preservation			Text Align
Enc	Mask	Attn	UNet	Blend	IR$_{\times 10}$↑	HPS$_{\times 10^2}$↑	AS↑	PSNR↑	LPIPS$_{\times 10^2}$↓	MSE$_{\times 10^3}$↓	CLIP Sim↑
Conv	w/	w/o	full	w/o	11.05	26.23	6.55	14.89	37.23	64.54	26.76
VAE	w/o	w/o	full	w/o	11.55	27.70	6.57	17.96	26.38	49.33	26.87
VAE	w/	w/	full	w/o	11.25	27.62	6.56	18.69	19.44	34.28	26.63
Conv	w/	w/	CN	w/o	9.58	26.85	6.47	12.15	80.91	150.89	26.88
VAE	w/	w/	CN	w/o	10.53	27.42	6.59	18.28	24.36	41.63	26.89
VAE	w/	w/o	CN	w/o	11.42	27.69	6.58	18.49	24.09	36.33	26.86
VAE	w/	w/o	half	w/o	11.47	27.70	6.57	19.01	23.77	33.57	26.87
VAE	w/	w/o	full	w/o	11.59	27.71	**6.59**	19.86	16.09	31.68	**26.91**
VAE	w/	w/o	full	paste	11.72	27.93	6.58	-	-	-	26.80
VAE	w/	w/o	full	blur	**11.76**	**27.94**	6.58	**29.88**	**1.53**	**11.65**	26.81

6 Discussion

Conclusion. This paper proposes a plug-and-play image inpainting method *BrushNet* with a pixel-level masked image feature insertion architectural design. Quantitative and qualitative results on our proposed benchmark, *BrushBench*, and EditBench show the superior performance of *BrushNet* considering image generation quality, masked region preservation, and image-text alignment.

Limitations and Future Work. However, *BrushNet* still has some limitations: (1) The quality and content generated by our model are heavily dependent on the chosen base model. As shown in Figure 6, the results of Model MeinaMix [26] exhibit incoherence because the given image is a natural image while the generation model primarily focuses on anime. (2) Even with *BrushNet*, we still observe poor generation results in cases where the given mask has an unusually shaped or irregular form, or when the given text does not align well with the masked image. In our future work, we will continue to address these challenges and further improve upon the identified problems.

Negative Social Impact. Image inpainting models present exciting opportunities for content creation, but they also carry potential risks to individuals and society. Their reliance on internet-collected training data can amplify social biases, and there is a specific risk of generating persuasive misinformation by manipulating human images with offensive elements. To address these concerns, it is crucial to emphasize responsible use and establish ethical guidelines when utilizing these models. This is also a key focus for our future model releases.

References

1. Avrahami, O., Fried, O., Lischinski, D.: Blended latent diffusion. ACM transactions on graphics (TOG) **42**(4), 1–11 (2023)
2. Avrahami, O., Lischinski, D., Fried, O.: Blended diffusion for text-driven editing of natural images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 18208–18218 (2022)
3. Bertalmio, M., Sapiro, G., Caselles, V., Ballester, C.: Image inpainting. In: International Conference and Exhibition on Computer Graphics and Interactive Techniques (SIGGRAPH). pp. 417–424 (2000)
4. Binghui, C., Chao, L., Chongyang, Z., Wangmeng, X., Yifeng, G., Xuansong, X.: Replaceanything as you want: Ultra-high quality content replacement (2023), `https://aigcdesigngroup.github.io/replace-anything/`
5. Bińkowski, M., Sutherland, D.J., Arbel, M., Gretton, A.: Demystifying MMD GANs. arXiv preprint arXiv:1801.01401 (2018)
6. Corneanu, C., Gadde, R., Martinez, A.M.: Latentpaint: Image inpainting in latent space with diffusion models. In: IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 4334–4343 (2024)
7. Criminisi, A., Pérez, P., Toyama, K.: Region filling and object removal by exemplar-based image inpainting. IEEE Transactions on Image Processing **13**(9), 1200–1212 (2004)
8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 248–255. Ieee (2009)
9. epinikion: epicrealism (2023), `https://civitai.com/models/25694?modelVersionId=143906`
10. heni29833: Henmixreal (2024), `https://civitai.com/models/20282?modelVersionId=305687`
11. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a local Nash equilibrium. Advances in Neural Information Processing Systems (NIPS) **30** (2017)
12. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems (NIPS) **33**, 6840–6851 (2020)
13. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
14. Huang, H., He, R., Sun, Z., Tan, T., et al.: Introvae: Introspective variational autoencoders for photographic image synthesis. Advances in Neural Information Processing Systems (NIPS) **31** (2018)
15. Huang, Y., Huang, J., Liu, Y., Yan, M., Lv, J., Liu, J., Xiong, W., Zhang, H., Chen, S., Cao, L.: Diffusion model-based image editing: A survey. arXiv preprint arXiv:2402.17525 (2024)
16. Jayasumana, S., Ramalingam, S., Veit, A., Glasner, D., Chakrabarti, A., Kumar, S.: Rethinking fid: Towards a better evaluation metric for image generation. arXiv preprint arXiv:2401.09603 (2023)
17. Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., et al.: The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. International Journal of Computer Vision (IJCV) **128**(7), 1956–1981 (2020)
18. Li, Z., Wei, P., Yin, X., Ma, Z., Kot, A.C.: Virtual try-on with pose-garment keypoints guided inpainting. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 22788–22797 (2023)

19. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European Conference on Computer Vision (ECCV). pp. 740–755. Springer (2014)
20. Liu, A., Niepert, M., Broeck, G.V.d.: Image inpainting via tractable steering of diffusion models. arXiv preprint arXiv:2401.03349 (2023)
21. Liu, H., Wan, Z., Huang, W., Song, Y., Han, X., Liao, J.: Pd-GAN: Probabilistic diverse GAN for image inpainting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9371–9381 (2021)
22. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: IEEE/CVF International Conference on Computer Vision (ICCV) (December 2015)
23. Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., Van Gool, L.: RePaint: Inpainting using denoising diffusion probabilistic models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11461–11471 (2022)
24. Lykon: Dreamshaper (2022), `https://civitai.com/models/4384?modelVersionId=128713`
25. Manukyan, H., Sargsyan, A., Atanyan, B., Wang, Z., Navasardyan, S., Shi, H.: Hd-painter: High-resolution and prompt-faithful text-guided image inpainting with diffusion models. arXiv preprint arXiv:2312.14091 (2023)
26. Meina: Meinamix (2023), `https://civitai.com/models/7240?modelVersionId=119057`
27. Peng, J., Liu, D., Xu, S., Li, H.: Generating diverse structure for image inpainting with hierarchical vq-vae. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10775–10784 (2021)
28. von Platen, P., Patil, S., Lozhkov, A., Cuenca, P., Lambert, N., Rasul, K., Davaadorj, M., Wolf, T.: Diffusers: State-of-the-art diffusion models. `https://github.com/huggingface/diffusers` (2022)
29. Quan, W., Chen, J., Liu, Y., Yan, D.M., Wonka, P.: Deep learning-based image and video inpainting: A survey. International Journal of Computer Vision (IJCV) pp. 1–34 (2024)
30. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning (ICML). pp. 8748–8763. PMLR (2021)
31. Razzhigaev, A., Shakhmatov, A., Maltseva, A., Arkhipkin, V., Pavlov, I., Ryabov, I., Kuts, A., Panchenko, A., Kuznetsov, A., Dimitrov, D.: Kandinsky: an improved text-to-image synthesis with image prior and latent diffusion. arXiv preprint arXiv:2310.03502 (2023)
32. Ren, T., Liu, S., Zeng, A., Lin, J., Li, K., Cao, H., Chen, J., Huang, X., Chen, Y., Yan, F., et al.: Grounded sam: Assembling open-world models for diverse visual tasks. arXiv preprint arXiv:2401.14159 (2024)
33. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10684–10695 (June 2022)
34. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. Advances in Neural Information Processing Systems (NIPS) **35**, 25278–25294 (2022)

35. SG161222: Realisticvision (2023), https://civitai.com/models/4201?modelVersionId=130072
36. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
37. Wang, S., Saharia, C., Montgomery, C., Pont-Tuset, J., Noy, S., Pellegrini, S., Onoe, Y., Laszlo, S., Fleet, D.J., Soricut, R., et al.: Imagen editor and editbench: Advancing and evaluating text-guided image inpainting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 18359–18369 (2023)
38. Wikipedia contributors: Mean squared error — Wikipedia, the free encyclopedia (2024), https://en.wikipedia.org/w/index.php?title=Mean_squared_error&oldid=1207422018, [Online; accessed 4-March-2024]
39. Wikipedia contributors: Peak signal-to-noise ratio — Wikipedia, the free encyclopedia (2024), https://en.wikipedia.org/w/index.php?title=Peak_signal-to-noise_ratio&oldid=1210897995, [Online; accessed 4-March-2024]
40. Wu, C., Huang, L., Zhang, Q., Li, B., Ji, L., Yang, F., Sapiro, G., Duan, N.: GODIVA: Generating open-domain videos from natural descriptions. arXiv preprint arXiv:2104.14806 (2021)
41. Wu, X., Hao, Y., Sun, K., Chen, Y., Zhu, F., Zhao, R., Li, H.: Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. arXiv preprint arXiv:2306.09341 (2023)
42. Xie, S., Zhang, Z., Lin, Z., Hinz, T., Zhang, K.: Smartbrush: Text and shape guided object inpainting with diffusion model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 22428–22437 (2023)
43. Xie, S., Zhao, Y., Xiao, Z., Chan, K.C., Li, Y., Xu, Y., Zhang, K., Hou, T.: Dreaminpainter: Text-guided subject-driven image inpainting with diffusion models. arXiv preprint arXiv:2312.03771 (2023)
44. Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., Tang, J., Dong, Y.: Imagereward: Learning and evaluating human preferences for text-to-image generation (2023)
45. Xu, Z., Zhang, X., Chen, W., Yao, M., Liu, J., Xu, T., Wang, Z.: A review of image inpainting methods based on deep learning. Applied Sciences **13**(20), 11189 (2023)
46. Yang, S., Chen, X., Liao, J.: Uni-paint: A unified framework for multimodal image inpainting with pretrained diffusion model. In: ACM International Conference on Multimedia (MM). pp. 3190–3199 (2023)
47. Yang, S., Zhang, L., Ma, L., Liu, Y., Fu, J., He, Y.: Magicremover: Tuning-free text-guided image inpainting with diffusion models. arXiv preprint arXiv:2310.02848 (2023)
48. Yu, F., Seff, A., Zhang, Y., Song, S., Funkhouser, T., Xiao, J.: Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. arXiv preprint arXiv:1506.03365 (2015)
49. Yu, T., Feng, R., Feng, R., Liu, J., Jin, X., Zeng, W., Chen, Z.: Inpaint anything: Segment anything meets image inpainting. arXiv preprint arXiv:2304.06790 (2023)
50. Zhang, G., Ji, J., Zhang, Y., Yu, M., Jaakkola, T., Chang, S.: Towards coherent image inpainting using denoising diffusion implicit models (2023)
51. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: IEEE/CVF International Conference on Computer Vision (ICCV) (2023)

52. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 586–595 (2018)
53. Zhao, S., Cui, J., Sheng, Y., Dong, Y., Liang, X., Chang, E.I., Xu, Y.: Large scale image completion via co-modulated generative adversarial networks. arXiv preprint arXiv:2103.10428 (2021)
54. Zheng, C., Cham, T.J., Cai, J.: Pluralistic image completion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1438–1447 (2019)
55. Zheng, H., Lin, Z., Lu, J., Cohen, S., Shechtman, E., Barnes, C., Zhang, J., Xu, N., Amirghodsi, S., Luo, J.: Image inpainting with cascaded modulation GAN and object-aware training. In: European Conference on Computer Vision (ECCV). pp. 277–296. Springer (2022)
56. Zhuang, J., Zeng, Y., Liu, W., Yuan, C., Chen, K.: A task is worth one word: Learning with task prompts for high-quality versatile image inpainting. arXiv preprint arXiv:2312.03594 (2023)