MAD-paint: Mask-Aware Diffusion Sampling for Image Inpainting

Shipeng Jiang
Southwest University
College of Computer and Information Science
Chongqing, China
baozzi@email.swu.edu.cn

Jingwei Qu
Southwest University
College of Computer and Information Science
Chongqing, China
qujingwei@swu.edu.cn

Bingyao Huang*
Southwest University
College of Computer and Information Science
Chongqing, China
bhuang@swu.edu.cn

Abstract

Image inpainting aims to repair digital images with defects such as holes and scratches at both semantic and textural levels. Diffusion models have shown great success in image inpainting, delivering high-quality results. However, existing diffusion-based methods often overlook the shape of defective regions/masks, applying a uniform sampling strategy across varying shapes. This oversight may lead to low-quality or semantically inappropriate restored images. In this paper, we propose MAD-paint (Mask-Aware Diffusion sampling for inpainting), and show that applying different noise types tailored to specific defect regions/mask shapes during the reverse diffusion process can significantly improve the inpainting quality. We begin by introducing a metric for mask uncertainty to assess the impact of different masks on inpainting quality. Using this metric, we propose a mask-aware sampling approach that automatically adjusts its sampling strategy according to different mask shapes, as indicated by the mask uncertainty. In addition, leveraging the known image texture consistency, we propose a known region-guided iterative refinement mechanism to condition texture restoration. The experimental results demonstrate the advantages of our method over other diffusion-based inpainting methods.

CCS Concepts

• **Computing methodologies** → **Computer vision**.

Keywords

Image inpainting, Diffusion model, Mask uncertainty

ACM Reference Format:
Shipeng Jiang, Jingwei Qu, and Bingyao Huang. 2025. MAD-paint: Mask-Aware Diffusion Sampling for Image Inpainting. In *Proceedings of the 2025 International Conference on Multimedia Retrieval (ICMR '25), June 30-July 3, 2025, Chicago, IL, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3731715.3733381

*Corresponding author.

ICMR '25, Chicago, IL, USA

ACM ISBN 979-8-4007-1877-9/2025/06
https://doi.org/10.1145/3731715.3733381

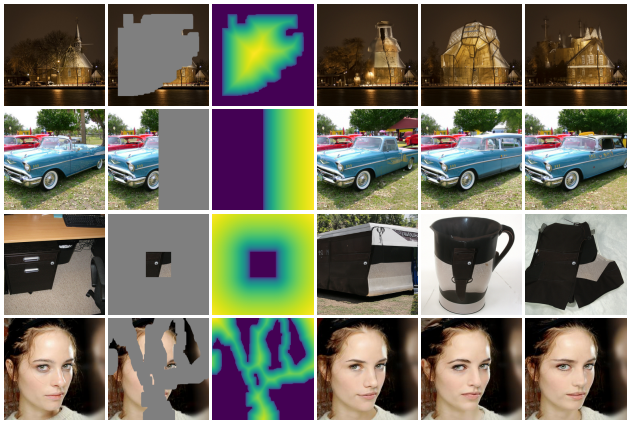


Figure 1: Our inpainting results under varying masks. From left to right: original image, masked image, mask uncertainty, and multiple stochastic inpainted results.

1 Introduction

In modern multimedia applications, images and videos play a central role in content creation, communication, and entertainment. However, multimedia data can suffer from various types of corruption, such as scratches and artifacts due to storage issues, transmission errors, or intentional editing needs. To address these issues, image inpainting techniques are usually applied to restore corrupted images by filling missing regions in a way that ensures both semantic coherence and texture consistency, as we show in Figure 1. Over the years, numerous approaches have been proposed, ranging from traditional methods to deep learning-based frameworks. However, existing techniques still face challenges when dealing with complex structures, fine details, or diverse masks in multimedia data. In recent years, diffusion models [15, 32] have demonstrated superior performance in image generation tasks [29, 37], surpassing traditional methods in producing high-fidelity and diverse results. By progressively denoising random noise through a learned reverse diffusion process, diffusion models have been successfully extended to various vision applications, including image editing [16], super-resolution [12], and image inpainting [7, 18, 24, 25, 36, 38]. However, diffusion-based image inpainting approaches still face challenges in handling complex structures and fine details.

The main difficulty of diffusion-based image inpainting lies in handling the noise introduced during the sampling process. Specifically, the noise must be aligned with the shape and structure of the missing regions defined by the mask. Unlike unconditional image

generation, inpainting needs to fill in missing pixels by leveraging the surrounding known ones, without breaking the semantic coherence and texture continuity. However, the sampling process in standard diffusion models often treats the entire image uniformly, without considering the structure of the missing regions. As a result, the added noise may overwhelm the available contextual information, leading to blurry textures or semantic/structural inconsistencies. These challenges highlight the need for more fine-grained control over the sampling strategy in missing regions, as well as better strategies to preserve both global semantics and local details throughout the reverse diffusion process.

In this paper, we propose MAD-paint, a novel mask-aware sampling and known region-guided iterative refinement framework for diffusion-based image inpainting. First, we analyze how mask shapes affect inpainting results and propose a metric to evaluate the uncertainty of masks in reconstructing missing areas. Then, we propose a mask-aware sampling strategy that adjusts the per-pixel ratio of predicted and random noise during the reverse diffusion process to improve semantic consistency and restoration quality. Furthermore, we design a known region-guided iterative refinement mechanism that enforces texture consistency between the restored and known regions to optimize the texture details during the reverse diffusion process. As shown in Figure 1, our approach effectively adapts to varying masks, ensuring high-quality and diverse restorations with well-preserved structure and texture.

We summarize our contributions as follows.

- We introduce a metric for mask uncertainty to assess the impact of mask shapes on inpainting quality.
- We propose a mask-aware sampling approach that automatically adjusts the sampling noises according to different mask shapes, as indicated by the mask uncertainty.
- We propose a known region-guided iterative refinement mechanism focused on improving texture restoration specifically.

2 Related Work

2.1 Image Inpainting

Image inpainting has been extensively explored with traditional and deep learning techniques. Early methods primarily rely on handcrafted priors and local information propagation. PDE-based approaches [2, 4, 34] first formulate inpainting as a diffusion process to smoothly interpolate missing regions. Exemplar-based methods [3, 8] search for similar patches in the known regions and copy them into missing areas to achieve texture continuity. Hybrid methods [1, 5] combine low-level cues and structural information to guide the filling process. Although these traditional approaches can yield plausible results in relatively simple scenarios, they often struggle with large missing regions, complex textures, or semantically meaningful structures.

With the rise of deep learning, Generative Adversarial Networks (GANs) [13] have become dominant in the field. Context Encoders [27] pioneered the use of encoder-decoder structures with adversarial loss for inpainting, inspiring a range of subsequent improvements [9, 14, 19, 21, 22, 39–41] aimed at enhancing realism and flexibility. While GAN-based methods generate semantically plausible completions, they often suffer from training instability and mode collapse, affecting the diversity and quality of results.

Another branch of deep learning approaches leverages Variational Autoencoders (VAEs). Methods such as [28, 43] employ VAE frameworks to improve diversity through probabilistic modeling. However, due to the inherent limitations of latent space sampling, VAE-based models tend to produce blurrier outputs compared to GAN-based methods, especially when reconstructing fine textures [35]. While these approaches represent important progress, achieving high-quality, semantically coherent, and texture-consistent inpainting under diverse mask conditions remains an open challenge.

2.2 Diffusion Models

Diffusion models are a powerful class of generative models that generate high-quality and diverse images by gradually adding noise to images and then learning to reverse this process. Representative methods include DDPM [15], DDIM [33], and their variants. Due to their strong generative capabilities and stable training, diffusion models have been widely adapted to various downstream tasks, including image inpainting.

Diffusion-based inpainting methods can be classified as supervised and unsupervised. Briefly, supervised methods [23, 26, 29, 31, 38] are trained specially for inpainting, where the mask is explicitly provided as input during training to guide the generation process. In contrast, unsupervised ones [18, 24, 25, 36, 38] repurpose pre-trained unconditional diffusion models without incorporating mask information during training. Our approach falls into the latter category, so we primarily discuss the unsupervised methods. It is worth noting that the "mask-aware" in our title refers to guiding the sampling process using the mask, rather than incorporating mask information during training as in supervised approaches, which often further leverage the mask throughout the model design [6, 20].

Image inpainting can be seen as the restoration of an image from certain degradation operators. Kawar et al. [18] use SVD to decompose these operators, which can be incorporated into the range-null space decomposition [36]. However, these methods often suffer from an inharmony issue as explained in [24, 30], primarily because unconditional DDPM is designed for generative tasks. Several approaches have modified the reverse diffusion process to address these limitations. SDEdit [25], for instance, introduces noise to the corrupted image only up to an intermediate level, rather than fully degrading it to Gaussian noise, and then repeatedly performs the reverse diffusion process to generate improved results. Repaint [24] extends this idea by incorporating a time-travel resampling strategy that repeatedly revisits intermediate states during the reverse diffusion process. This helps align the distributions between the known and unknown regions, improving consistency. However, this iterative resampling significantly increases the computational cost.

Despite the promising results of these methods, recent studies [30] indicate that diffusion-based inpainting still tends to produce overly smooth or contextually inconsistent completions, and often lacks fine-grained detail or local coherence in restored regions.

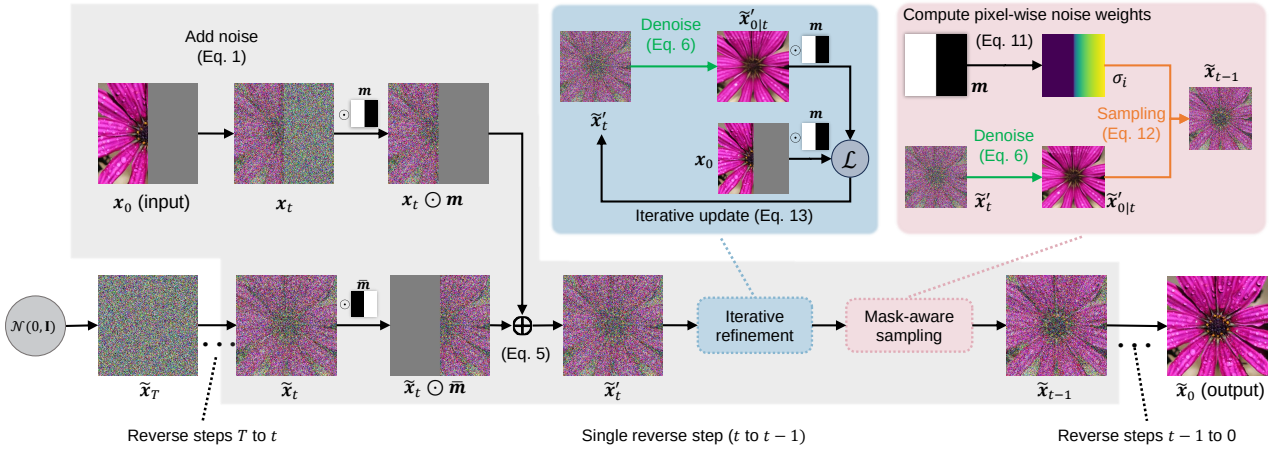


Figure 2: Overview of our method. We sample $\tilde{x}_T, \tilde{x}_{T-1}, \ldots, \tilde{x}_t, \ldots, \tilde{x}_1, \tilde{x}_0$ step by step through the reverse diffusion process. The sampling process at step t is as follows: First, we obtain a composite image $\tilde{x}'_t$ by concatenating the unknown region of $\tilde{x}_t$ with the known region of x_t, which is obtained from x_0 by applying time step t-level noise. Then, we use a U-Net to obtain a rough estimation $\tilde{x}'_{0|t}$. The texture loss between the known region of $\tilde{x}'_{0|t}$ and x_0 is computed for iterative refining $\tilde{x}'_t$. Next, we calculate the pixel-wise noise weights σ_i from the mask. Finally, we apply mask-aware sampling with Eq. (12) to obtain x_{t-1}. This reverse diffusion step is repeated until $t \to 0$, and the final inpainted image is $\tilde{x}_0$.

3 Method

In this section, we first provide the preliminaries of diffusion models and introduce the problem formulation with the notations used throughout our method. We then analyze the respective influences of predicted and random noise during the reverse diffusion process and discuss how mask shapes affect the restoration results. Based on these observations, we propose a metric to quantify the uncertainty of inpainting based on the certainty of unknown pixels in different mask regions. Following this, we present a mask-aware inpainting strategy that dynamically adjusts the sampling process according to the spatial characteristics of the mask. Finally, to further enhance texture restoration, we introduce a known region-guided iterative refinement mechanism that updates the intermediate results within the reverse diffusion process.

3.1 Preliminaries

To make the paper more accessible to a broader audience, we present a brief introduction to diffusion models. In DDPM [15], two processes, which are called the forward process and the reverse process, are built with the Markov chain. The forward process adds noise to a normal image $\boldsymbol{x}_0$ step by step, making it eventually converge to standard Gaussian noise $\boldsymbol{x}_T$:

$$\boldsymbol{x}_t = \sqrt{\overline{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\overline{\alpha}_t}\boldsymbol{\epsilon}_t, \quad \boldsymbol{\epsilon}_t \sim \mathcal{N}(0,\mathbf{I}), \tag{1}$$

where integer $t \in [1,T]$ represents the time step, and $\boldsymbol{\epsilon}_t \sim \mathcal{N}(0,\mathbf{I})$ is standard Gaussian noise independent of $\boldsymbol{x}_t$, $\alpha_t \in [0,1]$ is a parameter associated with time step t, and $\overline{\alpha}_t = \prod_{i=0}^{t} \alpha_i$. Here, $\mathcal{N}(0,\mathbf{I})$ denotes the standard Gaussian noise. In the reverse process, the denoised image at time step $t-1$ is given by:

$$\tilde{\boldsymbol{x}}_{t-1|t} = \frac{1}{\sqrt{\alpha_t}}\left(\tilde{\boldsymbol{x}}_t - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha}_t}}\boldsymbol{\epsilon}_\theta\left(\tilde{\boldsymbol{x}}_t, t\right)\right) + \sigma_t\boldsymbol{\epsilon}_t, \tag{2}$$

where $\boldsymbol{\epsilon}_\theta$ is a noise prediction model, e.g., U-Net, parameterized by θ, and $\boldsymbol{\epsilon}_\theta(\tilde{\boldsymbol{x}}_t, t)$ is the predicted noise added to the image $\tilde{\boldsymbol{x}}_t$ from the 0-th to the t-th time steps. Notably, this predicted noise is unweighted standard Gaussian noise. The term σ_t is set to $\sqrt{1-\alpha_t}$ in DDPM. Additionally, the symbol $\tilde{}$ denotes that results are produced during the reverse diffusion process, differentiating them from those of the forward process.

By constructing non-Markov processes, DDIM [33] proposes a more generalized and faster sampling approach. In our implementation, we adopt a reformulated version of the DDIM sampling process, which allows σ_t to take values in the range $[0,1]$ for more flexible control over the added noise during sampling. We can estimate the clean image $\tilde{\boldsymbol{x}}_{0|t}$ from the noisy image $\tilde{\boldsymbol{x}}_t$ based on the forward diffusion process formulation in Eq. (1) by:

$$\tilde{\boldsymbol{x}}_{0|t} = \frac{1}{\sqrt{\overline{\alpha}_t}}\left(\tilde{\boldsymbol{x}}_t - \sqrt{1-\overline{\alpha}_t}\boldsymbol{\epsilon}_\theta(\tilde{\boldsymbol{x}}_t, t)\right), \tag{3}$$

where subscript $_{0|t}$ indicates that the clean image $\tilde{\boldsymbol{x}}_{0|t}$ is estimated based on the noisy input $\tilde{\boldsymbol{x}}_t$. Then, by adding two types of noise, we can obtain $\tilde{\boldsymbol{x}}_{t-1|t}$:

$$\tilde{\boldsymbol{x}}_{t-1|t} = \sqrt{\overline{\alpha}_{t-1}}\tilde{\boldsymbol{x}}_{0|t} + \sqrt{1-\overline{\alpha}_{t-1}}\left(\sigma_t\boldsymbol{\epsilon}_t + \sqrt{1-\sigma_t^2}\boldsymbol{\epsilon}_\theta\left(\tilde{\boldsymbol{x}}_t, t\right)\right), \tag{4}$$

where coefficient σ_t is a tunable parameter rather than a fixed value in DDPM, allowing for control over the level of randomness in the generation process.

3.2 Problem Formulation

Based on whether an explicit mask is provided, image inpainting can be divided into two settings: without or with a mask. In the first setting, only a corrupted image $\boldsymbol{x}_0$ is provided, and the model must implicitly infer the whole image, including the missing regions. In the second setting, a binary mask $\boldsymbol{m}$ is given together with the

(a) Original **(b) Masked** **(c) $\sigma_t = 1$** **(d) $\sigma_t = 0$**

Figure 3: An example to illustrate the extreme scenarios of σ_t in § 3.3.1: the last two images are the inpainting results when we set $\sigma_t = 1$ and $\sigma_t = 0$ during the whole reverse diffusion process.

corrupted image, and only pixels within the complement mask $\overline{\boldsymbol{m}}$ are to be restored. In this work, we focus on the with mask setting [24]. Given a corrupted image $\boldsymbol{x}_0$ and a binary mask $\boldsymbol{m}$, our goal is to restore a semantically coherent and visually plausible image $\tilde{\boldsymbol{x}}_0$, based on the known regions $\boldsymbol{x}_0 \odot \boldsymbol{m}$.

From the perspective of diffusion models, image inpainting can be seen as a generation task, where the known region $\boldsymbol{x}_0 \odot \boldsymbol{m}$ serves as conditions. Similar to Repaint [24], rather than retraining a diffusion model from scratch, we directly integrate condition into the reverse diffusion process of DDIM [33] by

$$\tilde{\boldsymbol{x}}'_t = \underbrace{\boldsymbol{x}_t \odot \boldsymbol{m}}_{\text{known}} + \underbrace{\tilde{\boldsymbol{x}}_t \odot \overline{\boldsymbol{m}}}_{\text{unknown}}, \tag{5}$$

$$\tilde{\boldsymbol{x}}'_{0|t} = \frac{1}{\sqrt{\overline{\alpha}_t}}\left(\tilde{\boldsymbol{x}}'_t - \sqrt{1-\overline{\alpha}_t}\boldsymbol{\epsilon}_\theta(\tilde{\boldsymbol{x}}'_t, t)\right), \tag{6}$$

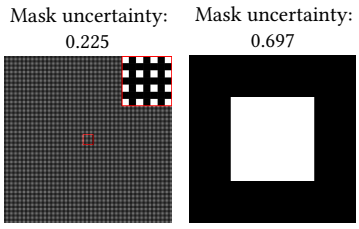
$$\tilde{\boldsymbol{x}}_{t-1|t} = \sqrt{\overline{\alpha}_{t-1}}\tilde{\boldsymbol{x}}'_{0|t} + \sqrt{1-\overline{\alpha}_{t-1}}\left(\sigma_t\boldsymbol{\epsilon}_t + \sqrt{1-\sigma_t^2}\boldsymbol{\epsilon}_\theta\left(\tilde{\boldsymbol{x}}'_t, t\right)\right). \tag{7}$$

The reverse diffusion process above is also illustrated in Figure 2. Note that two types of noise, i.e., $\boldsymbol{\epsilon}_t$ and $\boldsymbol{\epsilon}_\theta\left(\tilde{\boldsymbol{x}}'_t, t\right)$ are added to the predicted image $\tilde{\boldsymbol{x}}'_{0|t}$ to restore $\tilde{\boldsymbol{x}}_{t-1|t}$, and they play different roles. The random noise $\boldsymbol{\epsilon}_t$, independent of the prior known region or the previous step's result, may increase ambiguity and uncertainty in the reverse process. When $\sigma_t \to 1$, this can lead to blurred inpainted results, as illustrated in Figure 3(c). In contrast, the U-Net predicted noise $\boldsymbol{\epsilon}_\theta\left(\tilde{\boldsymbol{x}}'_t, t\right)$ may push the inpainted results toward more deterministic directions, resulting in a completely deterministic output when $\sigma_t \to 0$. While potentially sharp, such results can exhibit biases from the training data, causing semantic inconsistency with known regions, as shown in Figure 3(d).

In image inpainting, we aim for the missing pixels that are more certain about the known regions to be smooth, whereas those less certain about the known regions should be more random. Hence, determining σ_t according to the certainty of known regions is vital to our approach. Next, we will show how to quantify the missing pixel uncertainty using the mask shape.

3.3 Mask-aware Inpainting

To balance the ratio of random noise $\boldsymbol{\epsilon}_t$ and U-Net predicted noise $\boldsymbol{\epsilon}_\theta\left(\tilde{\boldsymbol{x}}'_t, t\right)$, we propose to compute per-pixel weights σ_t according to the pixel uncertainty, and use these weights to guide the reverse diffusion process. Since our σ_t is independent of the time step, we omit the subscript t from it in the following formulation.



(a) Grid Mask (b) Expand Mask

Figure 4: Mask uncertainties of two masks with identical unknown areas but different shapes. Both masks have 75% unknown pixels. Left: a grid mask where known pixels are evenly distributed with a stride of 2 (zoom in for more details and to avoid moiré artifacts). Right: an expand mask with all unknown pixels off the center. Despite having the same unknown area, the expand mask on the right is significantly harder to inpaint as shown by the mask uncertainty calculated by Eq. (9).

3.3.1 Mask Uncertainty. Intuitively, the uncertainty of a mask is closely related to its shape. Larger masks often lead to greater uncertainty in inpainting because they offer less known information. While one might simply consider the size of the unknown region as a measure of uncertainty, our findings suggest that the distance to known regions is more critical. Specifically, unknown pixels further away from known regions show higher uncertainty, whereas those near known regions are more reliable. Masks with identical unknown regions but different shapes can yield distinct results. As shown in Figure 4, mask (b) is clearly more challenging to inpaint than mask (a). Therefore, we propose a mask-aware pixel certainty metric $\mathcal{I}$ and an uncertainty metric for the entire mask $D_{\boldsymbol{m}}$. Given a set of unknown pixels $\mathcal{U} = \{i|\boldsymbol{m}(i) = 0\}$ to be inpainted, we have:

$$\mathcal{I}(i) = \min\left(\sum_{j \in \mathcal{N}_k(i)} \frac{1}{d(i,j)}, c\right), \tag{8}$$

$$D_{\boldsymbol{m}} = 1 - \frac{\sum_{i \in \mathcal{U}} \mathcal{I}(i)}{|\boldsymbol{m}| \cdot c}, \tag{9}$$

where $\mathcal{N}_k(i)$ denotes the set of the k nearest known pixels to pixel i. $d(i,j)$ represents the Euclidean distance between pixel i and j, with c acting as the upper bound of known information. where $|\boldsymbol{m}|$ denotes the total number of pixels of mask $\boldsymbol{m}$. A visual comparison of two masks using our mask uncertainty is shown in Figure 4.

3.3.2 Mask-aware Sampling. Given the unknown pixel and mask uncertainty, we can compute a global noise weight $\sigma_{\boldsymbol{m}}$ below to control the ratio of two types of noise applied during the reverse diffusion process in Eq. (7).

$$\sigma_{\boldsymbol{m}} = \sqrt{1 - D_{\boldsymbol{m}}^{\gamma}}, \tag{10}$$

where γ serves as an exponential scaling factor, ensuring that the mask with higher uncertainty receives more deterministic noise $\boldsymbol{\epsilon}_\theta\left(\tilde{\boldsymbol{x}}'_t, t\right)$. Finally, our pixel-wise noise weights are given by:

$$\sigma_i = \sqrt{\sigma_{\boldsymbol{m}}^2 + \frac{(1-\sigma_{\boldsymbol{m}}^2)\,\mathcal{I}(i)}{\max_{i \in \mathcal{U}} \mathcal{I}(i)}}. \tag{11}$$

where $\mathcal{I}(i)$ represents the unknown pixel's certainty about the known regions, and is defined in Eq. (8). $\max_{i\in\mathcal{U}}\mathcal{I}(i)$ denotes the maximum value of $\mathcal{I}(i)$ over all unknown pixels.

Once we have determined σ_i, we can perform mask-aware sampling during the reverse diffusion process:

$$\tilde{\boldsymbol{x}}_{t-1|t} = \sqrt{\overline{\alpha}_{t-1}}\tilde{\boldsymbol{x}}'_{0|t} + \sqrt{1-\overline{\alpha}_{t-1}}\left(\sigma_i\boldsymbol{\epsilon}_t + \sqrt{1-\sigma_i^2}\boldsymbol{\epsilon}_\theta\left(\tilde{\boldsymbol{x}}'_t, t\right)\right). \quad (12)$$

This strategy ensures that pixels with lower uncertainty receive more random noise $\boldsymbol{\epsilon}_t$, thereby preserving semantic consistency and smoothness near the boundary of the known and unknown regions. In contrast, pixels with higher uncertainty receive more deterministic noise $\boldsymbol{\epsilon}_\theta$. This allows the pixels to rely more on the training data prior, resulting in more complex and sharper structures. The pipeline of our mask-aware sampling is detailed in the pink block of Figure 2.

3.4 Known Region-guided Iterative Refinement

During the early stages of the reverse diffusion process, where t nears T, the input $\tilde{\boldsymbol{x}}'_t$ is highly noisy. At this point, the model heavily depends on the U-Net $\boldsymbol{\epsilon}_\theta$ learned priors and the noisy input. However, learned priors alone are often insufficient to guide the restoration effectively, resulting in poorly constrained predictions and suboptimal inpainting results [23, 42].

To address this issue, we introduce a known region-guided iterative refinement mechanism that takes advantage of the known regions of the image $\boldsymbol{x}_0 \odot \boldsymbol{m}$ to condition the unknown regions during the reverse diffusion process. Instead of relying solely on the U-Net to estimate the restored image (Eq. (6)), we also leverage texture loss against the known region and gradient descent to iteratively refine the inpainted regions, thereby improving texture consistency and visual quality. The key idea is that as the U-Net generates an initial prediction for the fully restored image $\tilde{\boldsymbol{x}}'_{0|t}$, the known regions can be directly validated against the same areas of $\boldsymbol{x}_0$. Enforcing texture consistency in these regions, we can guide the network towards more faithful reconstructions, rather than solely depending on learned training data priors. Our known region-guided iterative refinement approach is shown in the blue block of Figure 2. In particular, given a noisy image $\tilde{\boldsymbol{x}}_t$, we obtain $\tilde{\boldsymbol{x}}'_t$ with Eq. (5). Then, we predict an original image $\tilde{\boldsymbol{x}}'_{0|t}$ with Eq. (6).

Since the known regions of $\tilde{\boldsymbol{x}}'_{0|t}$ can be directly compared with the corresponding pixels from the original image $\mathbf{x_0}$, we apply a texture loss $\mathcal{L}(\cdot)$ to capture discrepancies within these regions. In particular, it measures the difference between the reconstructed and original pixels in the known areas. It could be an L_1 or L_2 loss, or perceptual loss. The performances of different loss functions are compared in Table 3. Once this texture error is computed, we apply gradient descent to optimize $\tilde{\boldsymbol{x}}'_t$:

$$\tilde{\boldsymbol{x}}'_t \leftarrow \tilde{\boldsymbol{x}}'_t - \eta\frac{\partial\mathcal{L}(\tilde{\boldsymbol{x}}'_{0|t}\odot\boldsymbol{m},\ \boldsymbol{x}_0\odot\boldsymbol{m})}{\partial\tilde{\boldsymbol{x}}'_t}, \quad (13)$$

where η is a learning rate that controls the step size for refinement. This optimization process effectively guides $\tilde{\boldsymbol{x}}'_t$ toward a solution that better aligns with the known regions of the original image.

Through this process, our method enhances the reconstructed image quality in two ways: (1) it reinforces fidelity in the known regions, ensuring that they match the original content as closely as possible, and (2) by improving the local consistency of the known parts, it indirectly influences the restoration of the unknown regions, leading to better global coherence in the final output.

Ultimately, by incorporating our known region-guided iterative refinement into the reverse diffusion process, we achieve sharper details and more structurally consistent inpainted results. Our complete algorithm is as shown in Alg. 1.

Algorithm 1 Mask-aware sampling and known region-guided iterative refinement

Input: Image $\boldsymbol{x}_0$, mask $\boldsymbol{m}$, time step T
Initialize $\boldsymbol{x}_T \sim \mathcal{N}(0, \mathbf{I})$
for $t = T \rightarrow 1$ **do**
  $\tilde{\boldsymbol{x}}'_t = \boldsymbol{x}_0 \odot \boldsymbol{m} + \tilde{\boldsymbol{x}}_t \odot \overline{\boldsymbol{m}}$ // Eq. (5)
  // Known region-guided iterative refinement (§ 3.4)
  for $n = 1 \rightarrow N$ **do**
    $\tilde{\boldsymbol{x}}'_t \leftarrow \tilde{\boldsymbol{x}}'_t - \eta\frac{\partial f}{\partial\tilde{\boldsymbol{x}}'_t}$
  end for
  // Mask-aware sampling § 3.3.2
  $\sigma_{\boldsymbol{m}} = \sqrt{1 - D^\gamma(\boldsymbol{m})}$
  $\sigma_i = \sqrt{\sigma_{\boldsymbol{m}}^2 + \frac{(1-\sigma_{\boldsymbol{m}}^2)\,\mathcal{I}(i)}{\max_{i\in\mathcal{U}}\mathcal{I}(i)}}$
  Obtain $\tilde{\mathbf{x}}_{t-1}$ with Eq. (12)
end for
Return: $\boldsymbol{x}_0$

4 Experiments

4.1 Implementation Details

Datasets. We evaluated our method on CelebA-HQ [17] and ImageNet [10]. All images were resized or cropped to 256×256, and experiments were conducted using one NVIDIA GeForce RTX 4060 Ti GPU. To evaluate our method on masks of different proportions, we adopted the mask datasets provided by Liu et al. [21] in which the masks are categorized into subsets based on their coverage proportion from 0% to 60%. In addition, we also tested two large-area masks: a right-half mask, where only the left half of the image is known, and an expanded center mask, where only the central 64×64 region of the 256×256 image is known. These settings were designed to simulate the restoration of extensive and contiguous missing regions.

Methods. For CelebA-HQ, we utilized the U-Net model trained by Lugmayr et al. [24], and for ImageNet, we utilized the U-Net model provided by Dhariwal et al. [11]. We compare our method against four existing approaches: DDNM [36], Repaint [24], DDRM [18], and Copaint [42]. For Copaint, we conducted experiments using its default settings. We also followed DDNM and DDRM's default settings except for the respacing step, which controls the number of sampling steps for faster inference. Instead of setting the respacing step to 250, we set it to 1,000 for a fair comparison. For Repaint, we kept the same total steps, jump lengths, and number of resampling as they did. Empirically, we set $\gamma = 0.3$ in Eq. (10) and $k = 10$, $c = 7$ in Eq. (8), respectively. The learning rate for known region-guided iterative refinement is set to $\eta = 0.02$. We set our time step to 250,

Figure 5: Qualitative comparison of our method against other diffusion-based methods on CelebA-HQ. The first row denotes different mask types, including masks with 0–20%, 20–40%, and 40–60% missing areas from [21], as well as large and coherent masks such as *right-half* and *expand*. The numbers in parentheses indicate the mask uncertainty.

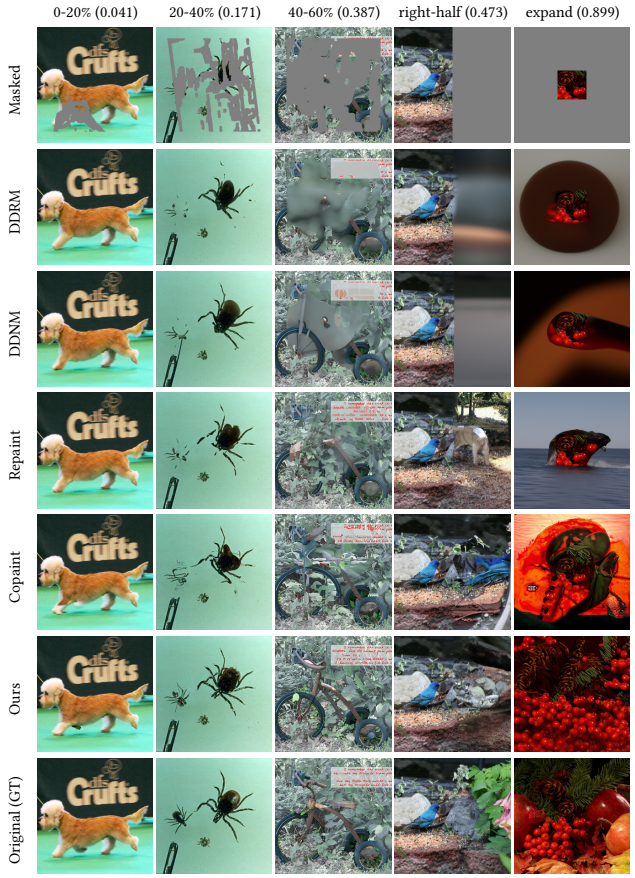


Figure 6: Qualitative comparison of our method against other diffusion-based methods on ImageNet. The first row denotes different mask types, including masks with 0–20%, 20–40%, and 40–60% missing areas from [21], as well as large and coherent masks such as *right-half* and *expand*. The numbers in parentheses indicate the mask uncertainty.

and at each step of the reverse diffusion process, we perform this iterative refinement twice, i.e., $N = 2$ in Alg. 1.

Metrics. In experiments, we used two perceptual metrics to evaluate inpainting performance: LPIPS and FID. LPIPS (Learned Perceptual Image Patch Similarity) measures perceptual similarity between two images by comparing high-level features extracted from a neural network, focusing on human visual perception rather than pixel-level accuracy. Lower LPIPS values indicate better perceptual quality. FID (Fréchet Inception Distance) evaluates the quality of generated images by comparing their feature distributions to those of real images using a pre-trained Inception network. Lower FID indicates better image quality and diversity. The two metrics allow us to assess both perceptual quality and consistency with the original data distribution.

4.2 Results

We evaluated our method by both qualitative and quantitative comparisons with several state-of-the-art (SOTA) diffusion-based inpainting approaches. Our experiments are conducted on 100 test images from CelebA-HQ and 100 test images from ImageNet datasets.

As shown in Figure 5 and Figure 6, previous diffusion-based inpainting baselines perform well for small mask areas, i.e., 0% to 40%, but as the mask size increases, they struggle to effectively integrate information from the unknown regions into the known regions. Specifically, these models often fail to propagate contextual information from the known regions into the missing ones, leading to blurry or semantically implausible completions. For instance, the generated textures of hair in DDNM and DDRM results appear over-smoothed, as seen in the second and third rows of Figure 5. As for other models, they may even hallucinate incorrect semantic content under large or irregular masks, as shown in the last two columns in Figure 6. In contrast, our proposed method continues to produce sharp, semantically coherent results even in challenging

Table 1: Quantitative comparison against other diffusion-based methods on CelebA-HQ. All experiments were conducted on 100 test images. For each metric, the best score is highlighted in bold, and the second best score is underlined.

Metric	Method	Mask Type					Mean
		0-20%	20-40%	40-60%	right-half	expand	
LPIPS↓	DDRM	0.014	0.057	0.127	0.197	0.515	0.182
	DDNM	0.012	0.051	0.119	0.192	0.504	0.176
	Copaint	0.014	0.052	0.117	0.189	0.478	0.170
	Repaint	**0.010**	0.046	0.114	0.197	0.489	0.171
	Ours	0.012	**0.044**	**0.105**	**0.171**	**0.459**	**0.158**
FID↓	DDRM	6.72	21.26	43.17	38.86	104.42	42.89
	DDNM	6.60	19.54	37.68	38.44	103.02	41.06
	Copaint	9.43	20.54	39.06	40.64	90.54	40.04
	Repaint	**5.64**	**17.95**	38.37	40.83	86.54	37.87
	Ours	5.96	18.23	**34.85**	**37.42**	**82.95**	**35.88**
Mean D_m		0.057	0.197	0.383	0.473	0.899	0.402

Table 2: Quantitative comparison against other diffusion-based methods on ImageNet. All experiments were conducted on 100 test images. For each metric, the best score is highlighted in bold, and the second best score is underlined.

Metric	Method	Mask Type					Mean
		0-20%	20-40%	40-60%	right-half	expand	
LPIPS↓	DDRM	0.027	0.113	0.264	0.363	0.780	0.309
	DDNM	0.026	0.103	0.250	0.359	0.770	0.302
	Copaint	0.027	0.101	0.226	0.296	0.638	0.258
	Repaint	**0.020**	**0.086**	0.214	0.329	0.697	0.269
	Ours	0.026	0.087	**0.194**	**0.266**	**0.620**	**0.239**
FID↓	DDRM	10.95	49.52	122.08	100.42	253.91	107.38
	DDNM	9.83	46.26	107.42	99.42	249.30	102.45
	Copaint	13.13	45.95	114.62	114.58	271.46	111.95
	Repaint	**9.25**	41.40	117.34	133.60	251.09	110.54
	Ours	10.56	**32.38**	**84.37**	**86.49**	**235.96**	**89.95**
Mean D_m		0.057	0.197	0.383	0.473	0.899	0.402

cases. By introducing a mask- and pixel-level noise weights, our method dynamically adjusts the noise ratio during sampling, allowing for better guidance in challenging regions, such as the precise reconstruction of the hat brim in the third column of Figure 5. Our mask-aware sampling strategy can better adapt to different mask shapes, enhancing semantic consistency in diverse contexts, particularly evident in the fourth and fifth columns of Figure 6. Our known region-guided iterative refinement approach further enhances details, as shown in the insect limbs (the second column of Figure 6) and the clear text (the third column of Figure 6). These observations are also reflected in the quantitative results Table 1 and Table 2: for LPIPS, our method achieves performance comparable to that of other approaches for small masks, i.e., 0% to 40%. However, as the mask size increases, our method consistently outperforms the other baselines, demonstrating its robustness in handling more challenging inpainting scenarios. Our method achieves a lower FID score compared to the baseline methods in most cases, demonstrating better alignment with the statistical properties of real images. This improvement is due to our method's ability to effectively restore both global structure and local details, ensuring not only perceptual accuracy but also distributional consistency.

In summary, our approach provides both quantitative and qualitative improvements across a wide range of inpainting scenarios. Its strength lies in spatially adapting to the complexity of the masked region and enhancing details with awareness of the mask shape. These advantages highlight the potential of our framework in practical applications involving large occlusions or complex textures.

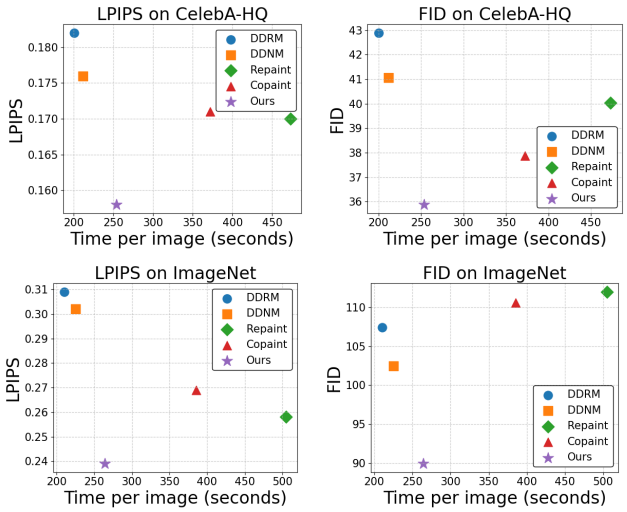


Figure 7: Time-effect trade-off chart of our and other methods. The horizontal axis represents the time required for inpainting one image, and the vertical axis represents the average LPIPS and FID of 100 images on 5 mask setups.

4.3 Time-effect Trade-off

Figure 7 shows the trade-off between runtime and performance for our method and several baselines. The X-axis represents the average running time required to process a single image, while the Y-axis shows the average LPIPS or FID score across 100 images from CelebA-HQ or ImageNet on 5 mask types. A lower position on the y-axis indicates better perceptual quality or consistent distribution, and a position closer to the left on the x-axis represents greater time efficiency. Our method, marked "Ours" in Figure 7, consistently demonstrates a strong time-effect trade-off. It achieves lower LPIPS and FID values than all the baselines, suggesting superior perceptual similarity to the original images and better alignment with the real image distribution. Meanwhile, the computational cost remains moderate, significantly lower than some recent high-performing methods such as Repaint or Copaint.

4.4 Ablation Study

To evaluate the effectiveness of our method, we conduct an ablation study on two components: texture loss used in the know region-guided iterative refinement (§ 3.4) and mask-aware sampling strategy (§ 3.3). Both components are designed to enhance different aspects of the generation process, with texture loss for fine-grained detail restoration and semantic consistency, and mask-aware sampling for adaptive noise control based on the spatial

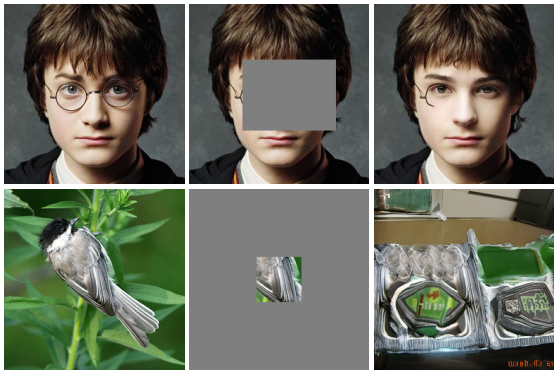
Table 3: Ablation results on CelebA-HQ and ImageNet, comparing different mask-aware sampling strategies (columns) and texture losses (rows). Metrics are averaged over 100 images under 5 random masks. "No ref." stands for no known region-guided iterative refinement. "✗" indicates no mask-aware sampling at all; "Mask-wise" uses a global noise weight σ_M in § 3.3.2; "Pixel-wise" uses pixel-wise noise weights σ_i in § 3.3.2. For each dataset and each metric, the best score is highlighted in bold, and the second-best score is underlined.

Dataset	Texture loss	Mask-aware Sampling					
		✗		Mask-wise		Pixel-wise	
		LPIPS↓	FID↓	LPIPS↓	FID↓	LPIPS↓	FID↓
CelebA-HQ	No ref.	0.269	75.00	0.173	39.59	0.170	39.43
	L1	0.176	40.67	0.162	36.76	<u>0.160</u>	<u>36.15</u>
	L2	0.173	40.06	0.163	36.94	**0.158**	**35.88**
	SSIM	0.205	53.83	0.170	38.18	0.165	37.26
	LPIPS	0.241	67.71	0.167	37.57	0.168	37.55
ImageNet	No ref.	0.356	135.60	0.270	109.73	0.253	94.01
	L1	0.273	113.55	0.251	93.42	<u>0.242</u>	**87.83**
	L2	0.277	116.38	0.245	91.79	**0.239**	<u>89.95</u>
	SSIM	0.308	120.89	0.262	99.50	0.248	93.35
	LPIPS	0.331	128.16	0.266	101.67	0.249	95.48

context of masks. Our mask-aware sampling is further analyzed with two variations: mask-wise noise weights, where the noise scaling is applied uniformly across the entire masked region, and pixel-wise noise weights, where the noise ratio is adjusted on a per-pixel basis, enabling finer control and local adaptivity. As shown in Table 3, incorporating known region-guided iterative refinement improves the performance of the model compared with no iterative refinement ("No ref."), demonstrating its effectiveness in improving inpainting quality, especially when the texture loss is set to L1 or L2 loss. Moreover, the introduction of mask-aware sampling significantly boosts performance by making the noise adjustment aware of spatial uncertainty. The variant using mask-wise noise weights already outperforms the method without mask-aware sampling ("✗"), indicating that adjusting noise weights based on global mask uncertainty improves model performance. Additionally, pixel-wise noise weights further enhance inpainting quality, suggesting that spatial noise adjustment using per-pixel uncertainty helps effectively reconstruct unknown regions. In summary, the ablation study confirms the effectiveness of both proposed modules and the benefit of their integration. They enhance robustness across various mask types and sizes while ensuring visually and semantically coherent results.

4.5 Limitations and Future Work

First, although our method balances inpainting quality and efficiency well, it still encounters challenges in cases with limited semantic context and highly structured content, as illustrated in Figure 8. In the first row, the model fails to recover the partially occluded glasses worn by the man, likely due to their small size, structural complexity, and insufficient surrounding information. Similarly, in the second row, although the central feathers of the bird are preserved, the inpainted result is semantically incorrect, resembling a dish rather than a bird. These failures are primarily attributed to two factors: first, the insufficient known information



(a) Original image (b) Masked (c) Our failed result

Figure 8: Failure cases of our method.

makes it difficult for the model to infer accurate content; second, the inherent characteristics of diffusion models, which generate images through progressive denoising, lose high-frequency details in early steps, leading to degraded or misleading outcomes.

In addition, although our adaptive sampling strategy improves efficiency and robustness across various masks, the hyperparameters γ, k, c used to control this process may be sensitive to image resolution. Inappropriate settings can degrade performance on very large or small images. Future work may explore resolution-invariant or self-adaptive strategies to enhance generalization.

Finally, while our current method focuses on static images, extending the framework to video inpainting remains a challenging direction. Future work can explore temporal consistency constraints and efficient sampling strategies for sequential data, further expanding the applicability of our approach.

5 Conclusion

In this paper, we address key challenges in applying diffusion models to image inpainting, particularly focusing on the impact of mask shapes on sampling noise types. We introduce an uncertainty metric to quantify the influence of different masks on inpainting quality, which is then used to guide a mask-aware sampling strategy that adjusts the sampling process based on mask shapes. This approach improves both semantic consistency and texture restoration, ensuring more accurate and realistic inpainting results. Additionally, we implement a known region-guided iterative refinement mechanism to further enhance texture details in each reverse diffusion step. Experimental results demonstrate that our method outperforms existing diffusion-based inpainting methods both qualitatively and quantitatively, achieving better restoration quality with acceptable computational overhead.

Acknowledgments

This work was supported by National Natural Science Foundation of China (NSFC) under Grant No. 62302401. The authors would like to thank Jijiang Li and Qingyue Deng for their help in proofreading.

References

[1] C. Allene and N. Paragios. 2006. Image Renaissance Using Discrete Optimization. In *18th International Conference on Pattern Recognition (ICPR'06)*, Vol. 3. IEEE, Hong Kong, China, 631–634.

[2] C. Ballester, M. Bertalmio, V. Caselles, G. Sapiro, and J. Verdera. 2001. Filling-in by Joint Interpolation of Vector Fields and Gray Levels. *IEEE Transactions on Image Processing* 10, 8 (2001), 1200–1211.

[3] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B. Goldman. 2009. PatchMatch: A randomized correspondence algorithm for structural image editing. *ACM Transactions on Graphics* 28, 3 (2009), 1–11.

[4] Marcelo Bertalmio, Guillermo Sapiro, Vincent Caselles, and Coloma Ballester. 2000. Image Inpainting. In *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques*. ACM, USA, 417–424.

[5] M. Bertalmio, L. Vese, G. Sapiro, and S. Osher. 2003. Simultaneous Structure and Texture Image Inpainting. *IEEE Transactions on Image Processing* (2003).

[6] Shuang Chen, Amir Atapour-Abarghouei, and Hubert P. H. Shum. 2024. HINT: High-Quality INpainting Transformer With Mask-Aware Encoding and Enhanced Attention. *IEEE Transactions on Multimedia* 26 (2024), 7649–7660.

[7] Ciprian Corneanu, Raghudeep Gadde, and Aleix M Martinez. 2024. Latentpaint: Image inpainting in latent space with diffusion models. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. IEEE, Waikoloa, Hawaii, USA, 4334–4343.

[8] A. Criminisi, P. Perez, and K. Toyama. 2004. Region Filling and Object Removal by Exemplar-Based Image Inpainting. *IEEE Transactions on Image Processing* 13, 9 (2004), 1200–1212.

[9] Ugur Demir and Gozde Unal. 2018. Patch-Based Image Inpainting with Generative Adversarial Networks. arXiv:1803.07422 [cs]

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Miami, FL, USA, 248–255.

[11] Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems* 34 (2021), 8780–8794.

[12] Sicheng Gao, Xuhui Liu, Bohan Zeng, Sheng Xu, Yanjing Li, Xiaoyan Luo, Jianzhuang Liu, Xiantong Zhen, and Baochang Zhang. 2023. Implicit diffusion models for continuous super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. IEEE, Vancouver, BC, Canada, 10021–10030.

[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative Adversarial Networks. *Commun. ACM* 63, 11 (2020), 139–144.

[14] Xiefan Guo, Hongyu Yang, and Di Huang. 2021. Image Inpainting via Conditional Texture and Structure Dual Generation. In *2021 IEEE/CVF International Conference on Computer Vision*. IEEE, Montreal, QC, Canada, 14114–14123.

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising Diffusion Probabilistic Models. *Advances in Neural Information Processing Systems* 33 (2020), 6840–6851.

[16] Jiancheng Huang, Mingfu Yan, Yifan Liu, and Shifeng Chen. 2024. SBCR: Stochasticity Beats Content Restriction Problem in Training and Tuning Free Image Editing. In *Proceedings of the 2024 International Conference on Multimedia Retrieval* (Phuket, Thailand) *(ICMR '24)*. ACM, New York, NY, USA, 878–887.

[17] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. 2018. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In *Proceedings of International Conference on Learning Representations*. OpenReview.net, Vancouver, BC, Canada.

[18] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. 2022. Denoising Diffusion Restoration Models. In *Advances in Neural Information Processing Systems*, Vol. 35. New Orleans, Louisiana, USA, 23593–23606.

[19] Jingyuan Li, Ning Wang, Lefei Zhang, Bo Du, and Dacheng Tao. 2020. Recurrent Feature Reasoning for Image Inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, Seattle, WA, USA, 7760–7768.

[20] Wenbo Li, Zhe Lin, Kun Zhou, Lu Qi, Yi Wang, and Jiaya Jia. 2022. MAT: Mask-Aware Transformer for Large Hole Image Inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, New Orleans, LA, USA, 10758–10768.

[21] Guilin Liu, Fitsum A. Reda, Kevin J. Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. 2018. Image Inpainting for Irregular Holes Using Partial Convolutions. In *Proceedings of the European Conference on Computer Vision*. Springer, Munich, Germany, 85–100.

[22] Hongyu Liu, Ziyu Wan, Wei Huang, Yibing Song, Xintong Han, and Jing Liao. 2021. PD-GAN: Probabilistic Diverse GAN for Image Inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, Virtual Event, 9371–9381.

[23] Haipeng Liu, Yang Wang, Biao Qian, Meng Wang, and Yong Rui. 2024. Structure Matters: Tackling the Semantic Discrepancy in Diffusion Models for Image Inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, Seattle, WA, USA, 8038–8047.

[24] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. 2022. RePaint: Inpainting Using Denoising Diffusion Probabilistic Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, New Orleans, LA, USA, 11461–11471.

[25] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. 2022. SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations. In *International Conference on Learning Representations*. OpenReview.net, Virtual Event.

[26] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. 2022. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In *Proceedings of the 39th International Conference on Machine Learning*, Vol. 162. PMLR, Baltimore, Maryland, USA, 16784–16804.

[27] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. 2016. Context Encoders: Feature Learning by Inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Las Vegas, NV, USA, 2536–2544.

[28] Jialun Peng, Dong Liu, Songcen Xu, and Houqiang Li. 2021. Generating Diverse Structure for Image Inpainting With Hierarchical VQ-VAE. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, Virtual Event, 10775–10784.

[29] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, New Orleans, LA, USA, 10684–10695.

[30] Litu Rout, Advait Parulekar, Constantine Caramanis, and Sanjay Shakkottai. 2023. A Theoretical Justification for Image Inpainting Using Denoising Diffusion Probabilistic Models. arXiv:2302.01217 [cs, math, stat]

[31] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. 2022. Palette: Image-to-Image Diffusion Models. In *ACM SIGGRAPH 2022 Conference Proceedings*. ACM, New York, NY, USA, 1–10.

[32] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. 2015. Learning Structured Output Representation Using Deep Conditional Generative Models. In *Advances in Neural Information Processing Systems*. 3483–3491.

[33] Jiaming Song, Chenlin Meng, and Stefano Ermon. 2021. Denoising Diffusion Implicit Models. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, Virtual Event.

[34] David TschumperlÉ. 2006. Fast Anisotropic Smoothing of Multi-Valued Images Using Curvature-Preserving PDE's. *International Journal of Computer Vision* 68, 1 (2006), 65–82.

[35] Sanchayan Vivekananthan. 2024. Comparative Analysis of Generative Models: Enhancing Image Synthesis with VAEs, GANs, and Stable Diffusion. arXiv:2408.08751 [cs.CV]

[36] Yinhuai Wang, Jiwen Yu, and Jian Zhang. 2023. Zero-Shot Image Restoration Using Denoising Diffusion Null-Space Model. In *International Conference on Learning Representations*. OpenReview.net, Kigali, Rwanda.

[37] Yankun Wu, Yuta Nakashima, and Noa Garcia. 2023. Not Only Generative Art: Stable Diffusion for Content-Style Disentanglement in Art Analysis. In *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval* (Thessaloniki, Greece) *(ICMR '23)*. ACM, New York, NY, USA, 199–208.

[38] Shaoan Xie, Zhifei Zhang, Zhe Lin, Tobias Hinz, and Kun Zhang. 2023. Smartbrush: Text and shape guided object inpainting with diffusion model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. IEEE, Vancouver, BC, Canada, 22428–22437.

[39] Chao Yang, Xin Lu, Zhe Lin, Eli Shechtman, Oliver Wang, and Hao Li. 2017. High-Resolution Image Inpainting Using Multi-Scale Neural Patch Synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Honolulu, HI, USA, 6721–6729.

[40] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S. Huang. 2018. Generative Image Inpainting With Contextual Attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Salt Lake City, UT, USA, 5505–5514.

[41] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S. Huang. 2019. Free-Form Image Inpainting With Gated Convolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. IEEE, Seoul, Korea (South), 4471–4480.

[42] Guanhua Zhang, Jiabao Ji, Yang Zhang, Mo Yu, Tommi Jaakkola, and Shiyu Chang. 2023. Towards coherent image inpainting using denoising diffusion implicit models. In *Proceedings of the 40th International Conference on Machine Learning* (Honolulu, Hawaii, USA) *(ICML'23)*. JMLR.org.

[43] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. 2019. Pluralistic Image Completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, Long Beach, CA, USA, 1438–1447.