Structure Matters: Tackling the Semantic Discrepancy in Diffusion Models for Image Inpainting

Haipeng Liu[1]   Yang Wang[1]*   Biao Qian[1]   Meng Wang[1]   Yong Rui[2]

[1]Hefei University of Technology, China   [2]Lenovo Research, China

hpliu_hfut@hotmail.com, yangwang@hfut.edu.cn,

{hfutqian,eric.mengwang}@gmail.com, yongrui@lenovo.com

Abstract

*Denoising diffusion probabilistic models (DDPMs) for image inpainting aim to add the noise to the texture of the image during the forward process and recover the masked regions with the unmasked ones of the texture via the reverse denoising process. Despite the meaningful semantics generation, the existing arts suffer from the semantic discrepancy between the masked and unmasked regions, since the semantically dense unmasked texture fails to be completely degraded while the masked regions turn to the pure noise in diffusion process, leading to the large discrepancy between them. In this paper, we aim to answer how the unmasked semantics guide the texture denoising process; together with how to tackle the semantic discrepancy, to facilitate the consistent and meaningful semantics generation. To this end, we propose a novel structure-guided diffusion model for image inpainting named **StrDiffusion**, to reformulate the conventional texture denoising process under the structure guidance to derive a simplified denoising objective for image inpainting, while revealing: 1) the semantically sparse structure is beneficial to tackle the semantic discrepancy in the early stage, while the dense texture generates the reasonable semantics in the late stage; 2) the semantics from the unmasked regions essentially offer the time-dependent structure guidance for the texture denoising process, benefiting from the time-dependent sparsity of the structure semantics. For the denoising process, a structure-guided neural network is trained to estimate the simplified denoising objective by exploiting the consistency of the denoised structure between masked and unmasked regions. Besides, we devise an adaptive resampling strategy as a formal criterion as whether the structure is competent to guide the texture denoising process, while regulate their semantic correlations. Extensive experiments validate the merits of StrDiffusion over the state-of-the-arts. Our code is available at* https://github.com/htyjers/StrDiffusion.

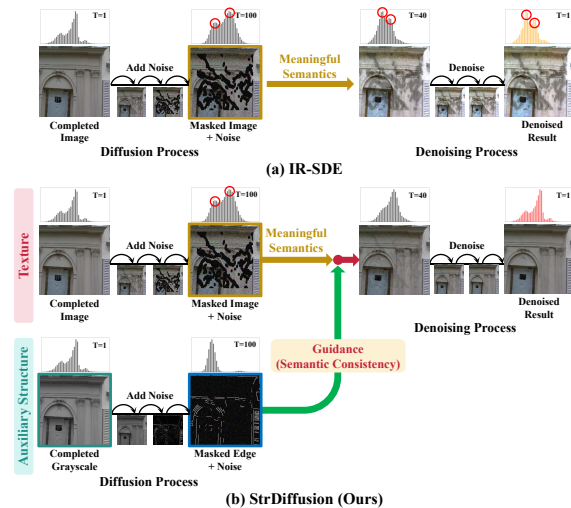*Yang Wang is the corresponding author.

Figure 1. Existing arts, *e.g.*, IR-SDE [23] (a), suffer from the semantic *discrepancy* (◯) between the masked and unmasked regions despite of the meaningful semantics for the masked regions during the denoising process. Our StrDiffusion (b) tackles the semantic discrepancy issue via the *guidance* of the auxiliary *sparse* structure, yielding the consistent and meaningful denoised results. The experiments are conducted on PSV [6].

1. Introduction

Recently, image inpainting has supported a wide range of applications, *e.g.*, photo restoration and image editing, which aims to recover the masked regions of an image with the semantic information of unmasked regions, where the *principle* mainly covers two aspects: the *reasonable semantics* for the masked regions and their *semantic consistency* with the unmasked regions. The prior work mainly involves the diffusion-based [2, 8] and patch-based [1, 4, 10] scheme, which tends to inpaint the small mask or repeated patterns of the image via simple color information, while fails to handle irregular or complicated masks. To address such issue, substantial attention [15, 17, 31, 36] has shifted to the convolutional neural networks (CNNs), which devote themselves to encoding the local semantics

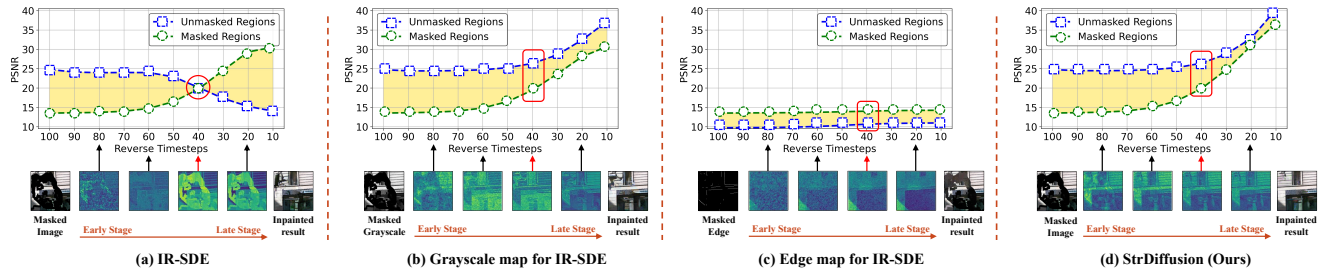| (a) IR-SDE | (b) Grayscale map for IR-SDE | (c) Edge map for IR-SDE | (d) StrDiffusion (Ours) |

Figure 2. Illustration of the motivating experiments about whether the sparse structure is beneficial to alleviating the discrepancy issue during the denoising process for image inpainting. Apart from the dense texture for IR-SDE [23] (a), the unmasked semantics combined with the Gaussian noise is further set as the sparse structure, *e.g.*, the grayscale map (b) and edge map (c). Our StrDiffusion (d) can tackle the semantic discrepancy via the progressively sparse structure. The shadow area indicates the discrepancy between the masked and unmasked regions during the denoising process. The PSNR (higher is better) reflects the recovered semantics for the masked (unmasked) regions compared to the completed image (*i.e.*, ground truth) by calculating the semantic similarity between them. The inpainted results are obtained by combining the masked regions of denoised results with the original masked images.

around the masked regions, while ignore the global information from the unmasked regions, incurring that the regions far away from the mask boundary remain burry. Recently, self-attention mechanism [5, 14, 16, 32, 35] is proposed to globally associate the masked regions with unmasked ones in the form of the segmented image patches, enhancing the semantic consistency between them. However, such strategies suffer from poor semantic correlation between the varied patches inside the masked regions. To this end, the semantically sparse structure is exploited [3, 7, 9, 13, 18, 20, 27, 28, 33, 34] to strength their correlations, which, however, implies the heavy reliance on the semantic consistency between the structure and texture, hence inevitably bears the artifact in the inpainted output.

Fortunately, the denoising diffusion probabilistic models (DDPMs) [12, 25] have emerged as the powerful generative models to deliver remarkable gains in the semantics generation and mode converge, hence well remedy the poor semantics generation for image inpainting [22, 23, 29, 38]. Specifically, instead of focusing on the training process, [22] proposed to adopt the pre-trained DDPMs as the prior and develop a resampling strategy to condition the reverse denoising process during the inference process. Further, [23] attempted to model the diffusion process for image inpainting by exploiting the semantics from the unmasked regions, yielding an optimal reverse solution to the denoising process. These methods mostly exhibit the semantically meaningful inpainted results with the advantages of DDPMs, whereas *overlook* the semantic consistency between the masked and unmasked regions. The intuition is that the semantically dense *unmasked* texture is degraded into the combination of the unmasked texture and the Gaussian noise, while the *masked* regions turn to the pure noise during the diffusion process, leading to the large discrepancy between *them* (see Fig.1(a) and Fig.2(a)), and *motivates* the followings:

(1) *How does the unmasked semantics guide the texture denoising process for image inpainting?* Our motivating ex-

periments (see Fig.2) suggest that, when the unmasked semantics combined with the noise turns to be sparser, *e.g.*, utilizing the grayscale or edge map of the masked image as an alternative, the discrepancy issue is largely alleviated, meanwhile, bears the large semantic information loss in the inpainted results; see Fig.2(b)(c). Therefore, the invariable semantics from the unmasked regions over time fail to guide the texture denoising process well.

(2) Following (1), one naturally wonders *how to yield the denoised results with consistent and meaningful semantics.* It is apparent that the sparse structure benefits the semantic consistency in the early stage while the dense texture tends to generate the meaningful semantics in the late stage during denoising process, implying a balance between the consistent and reasonable semantics of the denoised results. To further yield the desirable results, we take into account the guidance of the structure as an auxiliary over the texture denoising process; see Fig.1(b).

To answer the above questions, we propose a novel structure-guided texture diffusion model, named StrDiffusion, for image inpainting, to tackle the semantic discrepancy between masked and unmasked regions, together with the reasonable semantics for masked regions; technically, we reformulate the conventional texture denoising process under the guidance of the structure to derive a *simplified* denoising objective for image inpainting. For the denoising process, a structure-guided neural network is trained to estimate the simplified denoising objective during the denoising process, which exploits the time-dependent consistency of the denoised structure between the masked and unmasked regions to mitigate the discrepancy issue. Our StrDiffusion *reveals*: 1) the semantically *sparse* structure encourages the consistent semantics for the denoised results in the early stage, while the *dense* texture carries out the semantic generation in the late stage; 2) the semantics from the unmasked regions essentially offer the time-dependent structure guidance to the texture denoising process, benefiting from the time-dependent sparsity of the structure se-
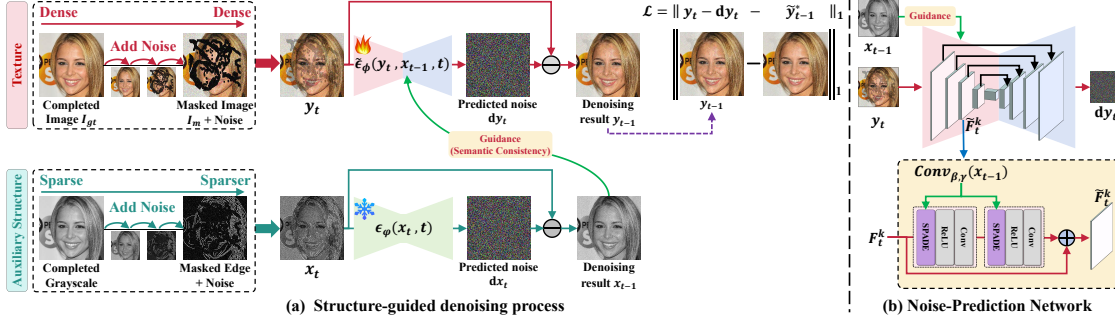
Figure 3. Illustration of the proposed StrDiffusion pipeline. Our basic idea is to tackle the semantic discrepancy between masked and unmasked regions via the guidance of the progressively sparse structure (a), which guides the texture denoising network (b) to generate the consistent and meaningful denoised results.

mantics. Concurrently, we remark that whether the structure guides the texture well greatly depends on the semantic correlation between them. As inspired, an adaptive resampling strategy comes up to monitor the semantic correlation and regulate it via the resampling iteration. Extensive experiments on typical datasets validate the merits of StrDiffusion over the state-of-the-arts.

2. Structure-Guided Texture Diffusion Models

Central to our method lies in three aspects: 1) how to tackle the semantic discrepancy between the masked and unmasked regions (Sec.2.2); 2) structure-guided denoising network for image inpainting (Sec.2.3); and 3) more insights on whether the structure guides the texture well (Sec.2.4). Before shedding light on our technique, we elaborate the DDPMs for image inpainting.

2.1. Preliminaries: Denoising Diffusion Probabilistic Models for Image Inpainting

Given a completed image $I_{gt} \in \mathbb{R}^{3 \times H \times W}$ and the binary mask $M \in \{0,1\}^{1 \times H \times W}$ (0 for masked regions, 1 for unmasked regions), image inpainting aims to recover the masked image $I_m = I_{gt} \odot M$ towards a inpainted image. For diffusion models, the inpainted results are obtained by merging the masked regions of denoised results with the original masked images. The typical DDPMs for inpainting [23] basically involve the forward texture diffusion process and reverse texture denoising process:

Forward Texture Diffusion Process. During the forward diffusion process, the semantic information from unmasked regions is utilized to guide the texture denoising process for image inpainting. Specifically, for the diffusion process with T timesteps, we set the initial texture state $y_0 = I_{gt}$ and the terminal texture state y_T as the combination of the masked image $\mu_y = I_m$ and the Gaussian noise ε. For any state $t \in [0, T]$, the diffusion process $\{y_t\}_{t=0}^{T}$ via the mean-reverting stochastic differential equation (SDE) [26] is defined as

$$\mathrm{d}y = \theta_t(\mu_y - y)dt + \eta_t \mathrm{d}w, \qquad (1)$$

where θ_t and η_t are time-dependent positive parameters that characterize the speed of the mean-reversion and the stochastic volatility during the diffusion process, which are constrained by $\eta_t^2/\theta_t = 2\lambda^2$ for the stationary variance since λ is fixed as the noise level applied to y_T. In particular, θ_t can be adjusted to construct different noise schedules for the texture; w is a standard Wiener process [25] (known as Brownian motion), which brings the randomness to the differential equation.

Reverse Texture Denoising Process. To recover the completed image y_0 from the terminal texture state y_T, we reverse the diffusion process by simulating the SDE backward in time, formulated as

$$\mathrm{d}y = [\theta_t(\mu_y - y) - \eta_t^2 \nabla_y \log q_t(y)]\mathrm{d}t + \eta_t \mathrm{d}\hat{w}, \qquad (2)$$

where $\hat{w}$ is a reverse-time Wiener process and $q_t(y)$ stands for the marginal probability density function of texture y_t at time t. The score function $\nabla_y \log p_t(y)$ is unknown, which can be approximated by training a conditional time-dependent neural network $\epsilon_\phi(y_t, t)$ [12]. Alternatively, we find the optimal reverse texture state y_{t-1}^* from y_t in $(t-1)$-th timestep via the maximum likelihood learning, which is achieved by minimizing the negative log-likelihood:

$$y_{t-1}^* = \arg\min_{y_{t-1}} [-\log q(y_{t-1}|y_t, y_0)]. \qquad (3)$$

By solving the above objective, we have:

$$\begin{aligned} y_{t-1}^* =& \frac{1 - e^{-2\overline{\theta}_{t-1}}}{1 - e^{-2\overline{\theta}_t}} e^{-\theta_t'}(y_t - \mu_y) \\ &+ \frac{1 - e^{-2\theta_t'}}{1 - e^{-2\overline{\theta}_t}} e^{-\overline{\theta}_{t-1}}(y_0 - \mu_y) + \mu_y, \end{aligned} \qquad (4)$$

where $\theta_t' = \int_{t-1}^{t} \theta_i \mathrm{d}i$ and $\overline{\theta}_t = \int_{0}^{t} \theta_z \mathrm{d}z$. Thus, we can optimize ϵ_ϕ via the following training objective:

$$\mathcal{L}_\beta^t(\phi) = \sum_{t=1}^{T} \beta_t \mathbb{E}\left[\|y_t - (\mathrm{d}y_t)_{\epsilon_\phi} - y_{t-1}^*\|_p\right], \qquad (5)$$

where β_t is positive weight; $(\mathrm{d}y_t)_{\epsilon_\phi}$ denotes the reverse-time SDE in Eq.(2) and its score is predicted via the noise network $\epsilon_\phi(y_t, t)$. $\|\cdot\|_p$ stands for the ℓ_p norm.
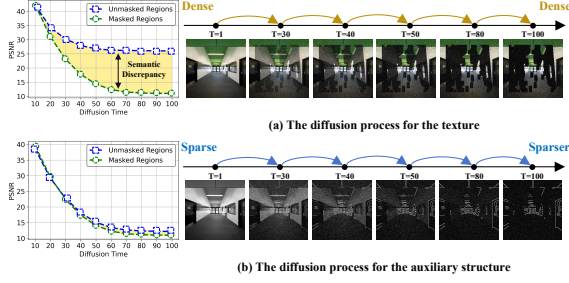
Figure 4. Illustration of the diffusion process for the dense texture (a) and sparse structure (b). In particular, the semantic sparsity of the structure is strengthened over time.

2.2. Tackling the Semantic Discrepancy in Diffusion Models for Image Inpainting

Based on the above, the existing diffusion models for image inpainting [23] suffer from the semantic discrepancy between the masked and unmasked regions, owing to the *dense* semantics of the texture. As opposed to the texture, the previous work [20, 34] suggests that the structure are generally considered to be *sparse*, which motivates us to tackle the discrepancy issue via the structure.

2.2.1 Structure Matters: Sparse Semantics benefits the Semantic Consistency

To validate such intuition, we design a series of motivation experiments over [23], where we replace the masked image μ_y in the terminal texture state y_T with the masked image μ_x from the *sparse* structure, *e.g.*, the grayscale map and edge map. Our motivation experiments (see Fig.2) reveal the followings:

(1) With the dense semantics, the texture performs well in producing meaningful semantics, while suffers from the large semantic discrepancy between the masked and unmasked regions; see Fig.1(a) and Fig.2(a).

(2) *The semantically sparse structure benefits the consistent semantics*, while incurs the semantics loss in the inpainted results; see the grayscale and edge map version of [23] in Fig.2(b)(c).

The above observations indicate that, unlike [23] that fixes the texture guidance from the unmasked regions, the structure guidance from the unmasked regions *varies* to the texture over time; meanwhile, *merely* involving the semantics of either dense texture or sparse structure fails to obtain the desirable inpainted results, since it actually attempts to perform the trade-off between the consistent and meaningful semantics. Motivated by that, we propose to exploit structure as an *auxiliary*, which exhibits the sparser semantics than the texture, so as to tackle the discrepancy issue, as remarked below.

Remark 1. Strengthening the semantic sparsity of the structure over time. Throughout the texture diffusion process, the semantic discrepancy between the masked and unmasked regions *progressively* increases as the timestep goes

towards to a totally masked status (see Fig.4(a)). To decrease the semantic discrepancy, we aim to strengthen the semantic sparsity of the structure over time, as opposed to the texture, during diffusion process. Concretely, the *sparse* grayscale map is degraded into the combination of the *sparser* masked edge map and the Gaussian noise via the noise schedule; see Fig.4(b).

Our basic idea is to tackle the semantic discrepancy issue via the guidance of the structure during the denoising process, thus simplify the texture denoising objective, i.e., the optimal solution (the ground truth) to the denoising process, which is discussed in the next.

2.2.2 Optimal Solution to the Denoising Process Under the Guidance of the Structure

As per the above observations, during the texture diffusion process, the distribution of the added noise in each timestep is disturbed by the semantic discrepancy between the masked and unmasked regions, incurring the difficulty to estimate such noise distribution via a noise-prediction network during the denoising process. Instead of focusing on texture, we exploit both structure and texture for diffusion process, where the sparsity of structure alleviates the semantic discrepancy, and the texture attempts to preserve the meaningful semantic. Unlike the diffusion process to simultaneously exploit structure and texture, for denoising process, the structure *firstly* generates the consistent semantics (x_{t-1}) between masked and unmasked regions, which *subsequently* guides the texture to generate the meaningful semantics (y_{t-1}); see Fig.3. By the assistance of structure for StrDiffusion, the texture diffusion process generates both consistent and meaningful inpainting output.

Remark 2. Different from the typical DDPMs for inpainting [22, 23], our StrDiffusion exploits the semantic consistency of the denoised structure to guide the texture denoising process, yielding both consistent and meaningful denoised results. *In particular*, we view the structure as an auxiliary to guide the texture rather than the inverse; the reason is two-fold: 1) during the inference denoising process, the masked regions mainly focus on the texture to be inpainted rather than the structure; 2) the structure focuses on the consistency between the masked and unmasked regions, lacking the meaningful semantics, hence still relies on the texture for inpainting.

Thanks to the guidance of the structure, the reverse transition for the texture in Eq.(3) can be reformulated via Bayes' rule as follows:

$$q(y_{t-1}|y_t, x_{t-1}) = \frac{q(y_{t-1}|y_t)q(x_{t-1}|y_{t-1}, y_t)}{q(x_{t-1}|y_t)}$$
$$= \frac{q(y_t|y_{t-1})q(y_{t-1})q(x_{t-1}|y_{t-1})}{q(x_{t-1}, y_t)}, \quad (6)$$

where the texture y_t and y_{t-1}, together with the structure x_{t-1}, are actually conditioned on their initial states, *i.e.*,

y_0 and x_0, during the diffusion process, hence $q(y_{t-1}) = q(y_{t-1}|y_0)$ and $q(x_{t-1}, y_t) = q(x_{t-1})q(y_t|x_{t-1}) = q(x_{t-1}|x_0)q(y_t|x_{t-1}, y_0)$. Eq.(6) can thus be simplified as

$$
\begin{aligned}
& q(y_{t-1}|y_t, y_0, x_{t-1}, x_0) \\
&= \frac{q(y_t|y_{t-1})q(x_{t-1}|y_{t-1})}{q(y_t|x_{t-1}, y_0)} \times \frac{q(y_{t-1}|y_0)}{q(x_{t-1}|x_0)} \\
&\approx \frac{q(y_{t-1}|y_0)}{q(x_{t-1}|x_0)},
\end{aligned}
\tag{7}
$$

where $q(y_t|y_{t-1}) \approx q(y_t|x_{t-1}, y_0)$ and $q(x_{t-1}|y_{t-1}) \approx 1$, since the texture y_t can be obtained via the conditioned auxiliary structure x_{t-1} during the diffusion process and initial texture y_0 as Remarked above. Based on Eq.(7), the optimal reverse state is naturally acquired by minimizing the negative log-likelihood (*see the supplementary material for more inference details*):

$$
\begin{aligned}
\tilde{y}_{t-1}^* &= \arg\min_{y_{t-1}}[-\log q(y_{t-1}|y_t, y_0, x_{t-1}, x_0)] \\
&= \arg\min_{y_{t-1}}[-\log \frac{q(y_{t-1}|y_0)}{q(x_{t-1}|x_0)}],
\end{aligned}
\tag{8}
$$

where $\tilde{y}_{t-1}^*$ denotes the ideal state reversed from $\tilde{y}_t$ under the structure guidance. To solve the above objective, we compute its gradient and setting it to be zero:

$$
\begin{aligned}
& \nabla_{\tilde{y}_{t-1}^*}\{-\log q(\tilde{y}_{t-1}^*|y_t, y_0, x_{t-1}^*, x_0)\} = \\
& -\nabla_{\tilde{y}_{t-1}^*}\log q(\tilde{y}_{t-1}^*|y_0) + \nabla_{x_{t-1}^*}\log q(x_{t-1}^*|x_0) = 0,
\end{aligned}
\tag{9}
$$

where x_{t-1}^* is the ideal state reversed from x_t for the structure, similar in spirit to Eq.(4), given as

$$
\begin{aligned}
x_{t-1}^* =& \frac{1 - e^{-2\bar{\delta}_{t-1}}}{1 - e^{-2\bar{\delta}_t}} e^{-\delta_t'}(x_t - \mu_x) \\
& + \frac{1 - e^{-2\delta_t'}}{1 - e^{-2\bar{\delta}_t}} e^{-\bar{\delta}_{t-1}}(x_0 - \mu_x) + \mu_x,
\end{aligned}
\tag{10}
$$

where structure μ_x is the masked version of its initial state x_0 and δ_t is time-dependent parameter that characterizes the speed of the mean-reversion. To simplify the notation, we have $\delta_t' = \int_{t-1}^t \delta_i di$, $\bar{\delta}_{t-1} = \int_0^{t-1} \delta_z dz$ and $\bar{\delta}_t = \bar{\delta}_{t-1} + \delta_t'$. Since Eq.(9) is linear, we can derive $\tilde{y}_{t-1}^*$ as

$$
\begin{aligned}
\tilde{y}_{t-1}^* =& \underbrace{\left(\frac{1 - e^{-2\bar{\theta}_{t-1}}}{1 - e^{-2\bar{\delta}_t}} e^{-\delta_t'}\right)(x_t - \mu_x)}_{\text{Consistency term for masked regions}} \\
& - \underbrace{\left(\frac{1 - e^{-2\bar{\theta}_{t-1}}}{1 - e^{-2\bar{\delta}_t}} e^{-\delta_t'}\right) e^{-\bar{\delta}_t}(x_0 - \mu_x)}_{\text{Balance term for masked regions}} \\
& + \underbrace{e^{-\bar{\theta}_{t-1}}(y_0 - \mu_y)}_{\text{Semantics term for masked regions}} + \underbrace{\mu_y}_{\text{Unmasked regions}}.
\end{aligned}
\tag{11}
$$

Remark 3. Eq.(11) implies that the ideal reverse state $\tilde{y}_{t-1}^*$ focuses more on the semantic consistency based on the semantic generation for image inpainting, to alleviate

the semantic discrepancy issues. Under the guidance of the sparse structure, $\tilde{y}_{t-1}^*$ tends to maintain the consistency between masked and unmasked regions in the early stage of the denoising process, while recovering the reasonable semantics in the late stage. Specifically, $\tilde{y}_{t-1}^*$ is mainly composed of four parts: 1) the consistency term for the structure $x_t - \mu_x$ is utilized to form the masked regions inside the denoised texture and keep the semantic *consistency* with the unmasked regions; 2) the semantics term for the texture $y_0 - \mu_y$ progressively provides *meaningful* semantics for masked regions as the denoising process progresses; 3) the negative balance term $x_0 - \mu_x$ adaptively surpasses the semantic information from $y_0 - \mu_y$ in the early stage and the consistency information from $x_t - \mu_x$ in the late stage; 4) the unmasked texture μ_y retains the semantics of the *unmasked* regions for the denoised texture. All four parts together constitute the *optimal solution* to the structure-guided denoising process in the $(t-1)$-th timestep, *i.e.*, the ground truth to optimize the noise-prediction network, thus yielding the denoised results with the consistent and desirable semantics (see Fig.6), as discussed in the next section.

2.3. Structure-Guided Denoising Process: How does the Structure Guide the Texture Denoising Process?

With the above optimal solution (*i.e.*, the ground truth) in Eq.(11), the goal of the structure-guided denoising process is to estimate it via a noise-prediction neural network $\tilde{\epsilon}_\phi(y_t, x_{t-1}, t)$ under the guidance of the structure, where $\tilde{\epsilon}_\phi(y_t, x_{t-1}, t)$ is trained to predict the mean of the noise with the fixed variance of noise [25]. Unlike the conventional texture denoising process in [23], the structure denoising network is pre-trained as an *auxiliary* to obtain the denoised structure, *e.g.*, x_{t-1} in the t-th timestep. To fulfill the structure guidance, a *multi-scale spatially adaptive normalization strategy* is devised to condition a time-dependent noise network $\tilde{\epsilon}_\phi(y_t, x_{t-1}, t)$ for the texture, which deploys the spatially adaptive normalization and denormalization strategy [24] to incorporate the statistical information over the feature map across varied layers; see Fig.3. Formally, given x_{t-1} from a pre-trained structure denoising network, we achieve the structure guidance by reconstructing the feature map on the k-th layer of CNNs in time-dependent noise network $\tilde{\epsilon}_\phi$ for the texture (*i.e.*, $F_k^t \in \mathbb{R}^{C_k \times H_k \times W_k}$) as

$$
\begin{aligned}
& \tilde{F}_k^t = conv_\gamma(x_{t-1}) \frac{F_k^t - \mu_k^t}{\sigma_k^t} + conv_\beta(x_{t-1}), \\
& \mu_k^t(h_k, w_k) = \frac{1}{C_k} \sum_{c_k=1}^{C_k} F_k^t(h_k, w_k, c_k), \\
& \sigma_k^t(h_k, w_k) = \sqrt{\frac{1}{C_k} \sum_{c_k=1}^{C_k} (F_k^t(h_k, w_k, c_k) - \mu_k^t(h_k, w_k))^2}, \\
& h_k = 1, 2, ..., H_k, w_k = 1, 2, ..., W_k,
\end{aligned}
\tag{12}
$$

where $conv_\gamma(\cdot)$ and $conv_\beta(\cdot)$ denote the mappings that convert the input x_{t-1} to the scaling and bias values.

$\mu_k^t(h_k, w_k)$ and $\sigma_k^t(h_k, w_k)$ are the statistical mean and variance of the pixels across different channels at the position (h_k, w_k). Based on that, we turn to optimize the noise network $\tilde{\epsilon}_\phi(y_t, x_{t-1}, t)$ to estimate the optimal solution $\tilde{y}_{t-1}^*$ (*i.e.*, the ground truth) to the denoising process. To this end, the overall training objective is formulated as

$$\mathcal{L}_{\tilde{\beta}}^t(\phi) = \sum_{t=1}^{T} \tilde{\beta}_t \mathbb{E}[\| \underbrace{y_t - (\mathrm{d}y_t)_{\tilde{\epsilon}_\phi}}_{\text{Denoised } y_{t-1}} - \tilde{y}_{t-1}^* \|_p], \quad (13)$$

where $\tilde{\beta}_t$ is positive weight; $(\mathrm{d}y_t)_{\tilde{\epsilon}_\phi}$ denotes the estimated reverse noise by the noise-prediction network. By minimizing Eq.(13), the inpainted results with consistent and meaningful semantics can be generated during the structure-guided denoising process.

2.4. More Insights on Eq.(13)

To validate the effectiveness of the guidance from the structure to the texture denoising process, we provide more insights on Eq.(13) by studying the semantic correlation between the denoised texture and structure during inference.

2.4.1 Correlation Measure via a Discriminator

Apart from the semantic structure sparsity, the semantic correlation between the structure and texture is crucial to achieve the semantic consistency between the masked and unmasked regions. To measure such semantic correlation, we adopt a discriminator network D to yield the correlation score $D(y_{t-1}, x_{t-1}, t-1)$ (abbreviated as $D(y_{t-1}, x_{t-1})$), where the denoised texture y_{t-1} and structure x_{t-1}, together with the timestep $t-1$, serve as the inputs. To this end, we present a discriminator loss and a triplet loss to optimize D, as described below:

Discriminator Loss. During the denoising process, to provide the desirable guidance for the texture, the current structure x_{t-1} is expected to be closely correlated with the texture y_{t-1} in the current timestep, instead of the previous one y_t. Hence, we aim to discriminate y_{t-1} and y_t via a discriminator loss below:

$$\mathcal{L}_{dis} = -\mathbb{E}_{y_{t-1}}\left[\log\left(D(y_{t-1}, x_{t-1})\right)\right] \\ -\mathbb{E}_{y_t}\left[\log\left(1 - D(y_t, x_{t-1})\right)\right]. \quad (14)$$

Triplet Loss. Since the unmasked regions almost possess much denser semantics information compared to the masked ones during the denoising process, to well evaluate the correlation, Eq.(14) is encouraged to focus more on the unmasked regions, which is achieved via a triplet loss:

$$\mathcal{L}_{tri} = \max(\|D(\bar{y}_{t-1}, x_{t-1}) - D(y_{t-1}, x_{t-1}))\|_2 \\ - \|D(\bar{y}_{t-1}, x_{t-1}) - D(y_t, x_{t-1})\|_2 + \alpha, 0), \quad (15)$$

where $\bar{y}_{t-1} = y_{t-1} \odot M + y_t \odot (1-M)$ is constructed to associate the unmasked regions of y_{t-1} and the masked ones of y_t. α is a margin parameter and $\|\cdot\|_2$ denotes the ℓ_2 norm.
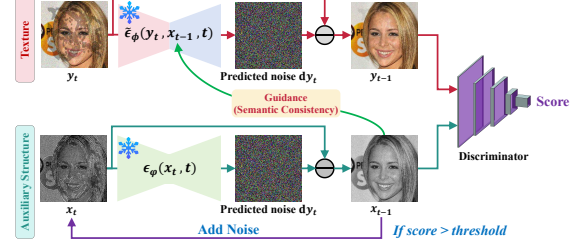

Figure 5. Illustration of the adaptive resampling strategy, which adaptively regulates the semantic correlation between the denoised texture and structure according to the score from the discriminator.

Total Loss. In summary, we finalize the total loss as

$$\mathcal{L} = \mathcal{L}_{dis} + \lambda_{tri}\mathcal{L}_{tri}, \quad (16)$$

where λ_{tri} denotes the balance paramneter, which is empirically set as 1 in our experiments. With Eq.(16) to be minimized, D can be trained to evaluate the semantic correlation, where the small score value from D indicates that the correlation between the texture and structure is not strong, weakening the guidance of the structure. To enhance the semantic correlation, an adaptive resampling strategy comes up, which is elaborated in the next section.

2.4.2 Adaptive Resampling Strategy

Different from the previous work [22] that directly performs the resampling strategy without the clear measurement, we devise an *adaptive* resampling strategy, which regulates the semantic correlation between the texture (*e.g.*, y_{t-1}) and structure (*e.g.*, x_{t-1}) according to the score value from a discriminator in Sec.2.4.1. Specifically, when the score value of the semantic correlation between y_{t-1} and x_{t-1} is smaller than a specific threshold Δ (*see the supplementary material for more intuitions*), we first obtain $\tilde{x}_t$ by adding the noise to x_{t-1}, then receive a updated $\tilde{x}_{t-1}$ by denoising $\tilde{x}_t$, which are described as follows:

$$\tilde{x}_t = x_{t-1} + \delta_{t-1}(\mu_x - x_{t-1})dt + \sigma_{t-1}\mathrm{d}w, \\ \tilde{x}_{t-1} = \tilde{x}_t - (\mathrm{d}x_t)_{\tilde{\epsilon}_\varphi}. \quad (17)$$

Based on Eq.(17), $\tilde{x}_{t-1}$ is utilized to generate an updated $\tilde{y}_{t-1}$ via the trained noise network $\tilde{\epsilon}_\phi(y_t, \tilde{x}_{t-1}, t)$. Following that, we evaluate the correlation between $\tilde{y}_{t-1}$ and $\tilde{x}_{t-1}$ again, and repeat the above procedure; see Fig.5.

3. Experiments

3.1. Implementation Details

We experimentally evaluate StrDiffusion over three typical datasets, including: **Paris StreetView (PSV)** [6] is a collection of street view images in Paris, consisting of 14.9k images for training and 100 images for validation; **CelebA** [21] is the human face dataset with 30k aligned face images,
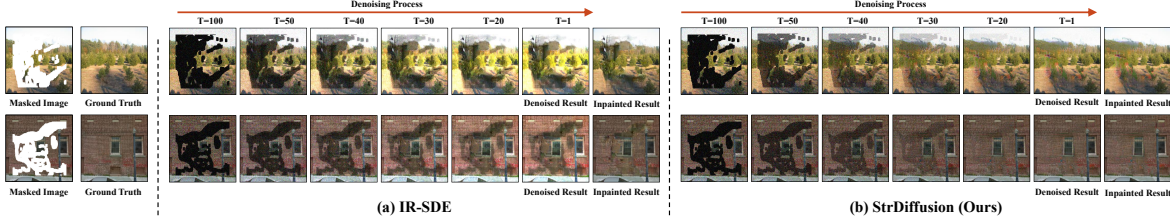
Figure 6. Illustration of the denoised results for IR-SDE [23] (a) and StrDiffusion (b) in the varied timesteps during the denoising process.

Table 1. Comparison of quantitative results (*i.e.*, PSNR, SSIM [30] and FID [11]) under varied mask ratios on Places2 with irregular mask dataset. ↑: Higher is better; ↓: Lower is better; - : no results are reported. The percentage (%) represents the ratios of varied irregular masked regions to the completed image. The best results are reported with **boldface**.

Metrics		PSNR↑				SSIM↑				FID↓			
Method	Venue	10-20%	20-30%	30-40%	40-50%	10-20%	20-30%	30-40%	40-50%	10-20%	20-30%	30-40%	40-50%
RFR [15]	CVPR' 20	27.26	24.83	22.75	21.11	0.929	0.891	0.830	0.756	17.88	22.94	30.68	38.69
PENNet [35]	CVPR' 19	26.78	23.53	21.80	20.48	0.946	0.894	0.841	0.781	20.40	37.22	53.74	67.08
HiFill [32]	CVPR' 20	27.50	24.64	22.81	21.23	0.954	0.914	0.873	0.816	16.95	29.30	42.88	57.12
FAR [3]	ECCV' 22	28.36	25.48	23.60	22.18	0.942	0.903	0.861	0.818	-	-	-	-
HAN [5]	ECCV' 22	28.93	25.44	23.06	21.38	0.957	0.903	0.839	0.762	12.01	20.15	28.85	37.63
CMT [14]	ICCV' 23	29.28	25.88	23.56	21.70	0.960	0.924	0.885	0.841	-	-	-	-
MEDFE [13]	ECCV' 20	28.13	25.02	22.98	21.53	0.958	0.919	0.874	0.825	14.41	27.52	38.45	53.05
CTSDG [9]	ICCV' 21	28.91	25.36	22.94	21.21	0.952	0.901	0.834	0.755	15.72	27.88	42.44	57.78
ZITS [7]	CVPR' 22	28.31	25.40	23.51	22.11	0.942	0.902	0.860	0.817	-	-	-	-
*RePaint** [22]	CVPR' 22	29.16	26.02	24.01	22.11	0.967	0.933	0.915	0.862	11.16	20.62	28.49	41.16
IR-SDE [23]	ICML' 23	28.98	25.37	23.93	21.68	0.961	0.928	0.908	0.864	12.03	21.34	29.37	39.77
StrDiffusion (Ours)	-	**29.45**	**26.28**	**24.43**	**22.37**	**0.973**	**0.941**	**0.917**	**0.874**	**8.750**	**16.16**	**23.72**	**36.93**

which is split into 26k training images and 4k validation images; **Places2** [37] contains more than 1.8M natural images from varied scenes. The network is trained using 256×256 images with irregular masks [19]. The noise-prediction network is constructed by removing group normalization layers and self-attention layers in the U-Net in DDPM [12] for inference efficiency. We adopt the Adam optimizer with β_1 = 0.9 and β_2 = 0.99. We set the timestep $T = 100$ for the diffusion model. All the experiments are implemented with the pytorch framework and run on 4 NVIDIA 2080TI GPU.

3.2. Why can StrDiffusion Work?

In this section, the core goal is to confirm that the intuition of StrDiffusion — the progressively sparse structure over time is beneficial to alleviate the semantic discrepancy issue between the masked and unmasked regions during the denoising process. We conduct the experiments on PSV and Places2 dataset. Fig.6 illustrates that, as the denoising process progresses, the denoised results from IR-SDE [23] always address the clear semantic discrepancy between the masked and unmasked regions (see Fig.6(a)); while for StrDiffusion, such discrepancy *progressively* degraded until *vanished*, yielding the consistent semantics (see Fig.6(b)), confirming our *proposals* in Sec.2.2.2 — the guidance of the structure benefits the *consistent semantics* between the masked and unmasked regions while the texture retains the *reasonable semantics* in the inpainted results; *see the supplementary material for more results.*

3.3. Comparison with State-of-the-arts

To validate the superiority of StrDiffusion, we compare it with the typical inpainting models, including: RFR [15] adopts the CNNs to encoding the local information;

Figure 7. Comparison of the inpainted results with the state-of-the-arts under varied irregular masks on PSV, CelebA and Places2. Our StrDiffusion delivers the desirable inpainted results (marked as the red box) against others.

PENNet[35], HiFill[32], FAR[3], HAN[5] and CMT[14] globally associate the masked regions with unmasked ones via self-attention mechanism; MEDFE[13], CTSDG[9], ZITS[7] and DGTS[20] focus on the guidance of the structure, while ignore the semantic consistency between the structure and texture; RePaint[22] and IR-SDE[23] benefit from DDPMs whereas overlook the semantic consistence. For fair comparison, we reproduce RePaint by utilizing the pre-trained models from IR-SDE, denoted as RePaint*. Notably, the inpainted results are obtained by merging the masked regions from the denoised results with the original masked image.

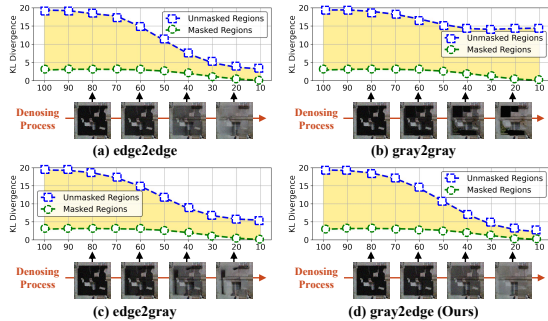**Quantitative and Qualitative analysis.** The quantitative

Figure 8. Ablation study about the the progressive sparsity for the structure over time on the PSV dataset. The KL divergence (lower is better) measures the distribution difference between the masked (unmasked) regions and the completed image (*i.e.*, ground truth).

results in Table 1 apart from the qualitative results in Fig.7 summarize our findings below: StrDiffusion enjoys a much smaller FID score, together with larger PSNR and SSIM than the competitors, confirming that StrDiffusion effectively addresses semantic discrepancy between the masked and unmasked regions, while yielding the reasonable semantics. Notably, RePaint* and IR-SDE still remain the large performance margins (at most 1.2% for PSNR, 1.4% for SSIM and 10% for FID) compared to StrDiffusion, owing to the semantic discrepancy (e.g., the white wall in the row 1 of Fig.7) in the denoised results incurred by the dense texture. Albeit MEDFE, CTSDG and ZITS focus on the guidance of the structure similar to StrDiffusion, they suffer from a performance loss due to the semantic discrepancy between the structure and texture, *e.g.*, the woman's face for CTSDG in the row 4 of Fig.7, which further verifies our *proposal* in Sec.2.2 — *the progressively sparse structure provides the time-dependent guidance for texture denoising process*; *see the supplementary material for more results.*

3.4. Ablation Studies

3.4.1 Why Should the Semantic Sparsity of the Structure be Strengthened over Time?

Recalling Sec.2.2, the progressive sparsity for the structure is utilized to mitigate the semantic discrepancy, by setting the initial structure state as the *grayscale* and the terminal structure state as the combination of the *edge* and the noise during the structure diffusion process, denoted as **gray2edge**. To shed more light on the semantic sparsity of the structure, we replace the initial and terminal structure state, yielding several alternatives as **gray2gray**, **edge2edge** and **edge2gray**; and perform the experiments on the PSV dataset. The structure sparsity of **gray2gray** and **edge2edge** is fixed over time. Fig.8 illustrates that **gray2edge** (Fig.8(d)) exhibits better consistency with meaningful semantics in the inpainted results against others, especially for **edge2gray**, implying the *benefits* of strengthening structure sparsity over time. Notably, for **gray2gray**, the discrepancy issue in the denoised re-

Table 2. Ablation study about the adaptive resampling strategy on the PSV dataset. The best results are reported with **boldface**.

Cases	Hard mask			Easy mask		
	PSNR↑	SSIM↑	FID↓	PSNR↑	SSIM↑	FID↓
A	24.65	0.851	55.73	27.69	0.913	33.73
B	25.47	0.863	46.44	28.31	0.921	28.43
C	25.59	0.867	47.31	28.64	0.925	28.20
D	**25.94**	**0.875**	**42.28**	**28.85**	**0.929**	**26.74**

sults still suffers (Fig.8(b)), while **edge2edge** receives the poor semantics (Fig.8(a)), which attributes to their *invariant* semantic sparsity over time, indicating *time-dependent* impacts of the unmasked semantics on the texture denoising process (Sec.2.2.1); *see the supplementary material for more results.*

3.4.2 How does Adaptive Resampling Strategy Benefit Inpainting?

To verify the effectiveness of the adaptive resampling strategy (Sec.2.4), we perform the ablation study on the PSV dataset from the following cases: **A**: refining the inpainted results of IR-SDE[23] via the resampling strategy in RePaint; **B**: removes the discriminator, which performs the adaptive resampling strategy without considering the semantic correlation between the texture and structure; **C**: StrDiffusion without the adaptive resampling strategy; **D**: the proposed StrDiffusion. Table 2 suggests that case **D** delivers great gains (at most 13.45%) to other cases, which is *evidenced* that the adaptive resampling strategy succeeds in regulating the semantic correlation for the desirable results (Sec.2.4); case **D** upgrades beyond case **A** despite of the combination with the resampling strategy, which, in turn, verifies the *necessity* of the time-dependent guidance from the structure. In particular, case **B** bears a large performance degradation (at most 4.16%), confirming the importance of *adaptively* resampling strategy as per the semantic correlation between the structure and texture in Sec.2.4.1.

4. Conclusion

In this paper, we propose a structure-guided diffusion model for inpainting (StrDiffusion), to reformulate the conventional texture denoising process under the guidance of the structure to derive a simplified denoising objective; while revealing: 1) the semantics from the unmasked regions essentially offer the time-dependent guidance for the texture denoising process; 2) the progressively sparse structure well tackle the semantic discrepancy in the inpainted results. Extensive experiments verify the superiority of StrDiffusion to the state-of-the-arts.

References

[1] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. Patchmatch: A randomized correspondence algorithm for structural image editing. *ACM Trans. Graph.*, 28(3):24, 2009. 1

[2] Marcelo Bertalmio, Guillermo Sapiro, Vincent Caselles, and Coloma Ballester. Image inpainting. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 417–424, 2000. 1

[3] Chenjie Cao, Qiaole Dong, and Yanwei Fu. Learning prior feature and attention enhanced image inpainting. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XV*, pages 306–322. Springer, 2022. 2, 7

[4] Soheil Darabi, Eli Shechtman, Connelly Barnes, Dan B Goldman, and Pradeep Sen. Image melding: Combining inconsistent images using patch-based synthesis. *ACM Transactions on graphics (TOG)*, 31(4):1–10, 2012. 1

[5] Ye Deng, Siqi Hui, Rongye Meng, Sanping Zhou, and Jinjun Wang. Hourglass attention network for image inpainting. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVIII*, pages 483–501. Springer, 2022. 2, 7

[6] Carl Doersch, Saurabh Singh, Abhinav Gupta, Josef Sivic, and Alexei Efros. What makes paris look like paris? *ACM Transactions on Graphics*, 31(4), 2012. 1, 6

[7] Qiaole Dong, Chenjie Cao, and Yanwei Fu. Incremental transformer structure enhanced image inpainting with masking positional encoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11358–11368, 2022. 2, 7

[8] Alexei A Efros and William T Freeman. Image quilting for texture synthesis and transfer. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 341–346, 2001. 1

[9] Xiefan Guo, Hongyu Yang, and Di Huang. Image inpainting via conditional texture and structure dual generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14134–14143, 2021. 2, 7

[10] James Hays and Alexei A Efros. Scene completion using millions of photographs. *ACM Transactions on Graphics (ToG)*, 26(3):4–es, 2007. 1

[11] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 7

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2, 3, 7

[13] Yibing Song Wei Huang Hongyu Liu, Bin Jiang and Chao Yang. Rethinking image inpainting via a mutual encoder-decoder with feature equalizations. In *Proceedings of the European Conference on Computer Vision*, 2020. 2, 7

[14] Keunsoo Ko and Chang-Su Kim. Continuously masked transformer for image inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13169–13178, 2023. 2, 7

[15] Jingyuan Li, Ning Wang, Lefei Zhang, Bo Du, and Dacheng Tao. Recurrent feature reasoning for image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7760–7768, 2020. 1, 7

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

[17] Xiaoguang Li, Qing Guo, Di Lin, Ping Li, Wei Feng, and Song Wang. Misf: Multi-level interactive siamese filtering for high-fidelity image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1869–1878, 2022. 1

[18] Liang Liao, Jing Xiao, Zheng Wang, Chia-Wen Lin, and Shin'ichi Satoh. Image inpainting guided by coherence priors of semantics and textures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6539–6548, 2021. 2

[19] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *Proceedings of the European conference on computer vision (ECCV)*, pages 85–100, 2018. 7

[20] Haipeng Liu, Yang Wang, Meng Wang, and Yong Rui. Delving globally into texture and structure for image inpainting. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1270–1278, 2022. 2, 4, 7

[21] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015. 6

[22] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11461–11471, 2022. 2, 4, 6, 7

[23] Ziwei Luo, Fredrik K Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas B Schön. Image restoration with mean-reverting stochastic differential equations. *arXiv preprint arXiv:2301.11699*, 2023. 1, 2, 3, 4, 5, 7, 8

[24] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2337–2346, 2019. 5

[25] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2020. 2, 3, 5

[26] Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models. *Advances in Neural Information Processing Systems*, 34: 1415–1428, 2021. 3

[27] Ziyu Wan, Jingbo Zhang, Dongdong Chen, and Jing Liao. High-fidelity pluralistic image completion with transform-

ers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4692–4701, 2021. 2

[28] Wentao Wang, Jianfu Zhang, Li Niu, Haoyu Ling, Xue Yang, and Liqing Zhang. Parallel multi-resolution fusion network for image inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14559–14568, 2021. 2

[29] Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-shot image restoration using denoising diffusion null-space model. *The Eleventh International Conference on Learning Representations*, 2023. 2

[30] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 7

[31] Chaohao Xie, Shaohui Liu, Chao Li, Ming-Ming Cheng, Wangmeng Zuo, Xiao Liu, Shilei Wen, and Errui Ding. Image inpainting with learnable bidirectional attention maps. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8858–8867, 2019. 1

[32] Zili Yi, Qiang Tang, Shekoofeh Azizi, Daesik Jang, and Zhan Xu. Contextual residual aggregation for ultra high-resolution image inpainting, 2020. 2, 7

[33] Yingchen Yu, Fangneng Zhan, Rongliang Wu, Jianxiong Pan, Kaiwen Cui, Shijian Lu, Feiying Ma, Xuansong Xie, and Chunyan Miao. Diverse image inpainting with bidirectional and autoregressive transformers. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 69–78, 2021. 2

[34] Yongsheng Yu, Dawei Du, Libo Zhang, and Tiejian Luo. Unbiased multi-modality guidance for image inpainting. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVI*, pages 668–684. Springer, 2022. 2, 4

[35] Yanhong Zeng, Jianlong Fu, Hongyang Chao, and Baining Guo. Learning pyramid-context encoder network for high-quality image inpainting. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1486–1494, 2019. 2, 7

[36] Yu Zeng, Zhe Lin, Huchuan Lu, and Vishal M. Patel. Cr-fill: Generative image inpainting with auxiliary contextual reconstruction. In *Proceedings of the IEEE International Conference on Computer Vision*, 2021. 1

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

[38] Zixin Zhu, Xuelu Feng, Dongdong Chen, Jianmin Bao, Le Wang, Yinpeng Chen, Lu Yuan, and Gang Hua. Designing a better asymmetric vqgan for stablediffusion. *arXiv preprint arXiv:2306.04632*, 2023. 2