A Structure-Guided Diffusion Model for Large-Hole Diverse Image Completion

Daichi Horita\textsuperscript{1} Jiaolong Yang\textsuperscript{2} Dong Chen\textsuperscript{2} Yuki Koyama\textsuperscript{3} Kiyoharu Aizawa\textsuperscript{1}

\textsuperscript{1}The University of Tokyo \textsuperscript{2}Microsoft Research Asia \textsuperscript{3}National Institute of Advanced Industrial Science and Technology (AIST)

\{horita,aizawa\}@hal.t.u-tokyo.ac.jp \{jiaoyan,doch\}@microsoft.com koyama.y@aist.go.jp

Figure 1. We propose a structure-guided diffusion model (SGDM), which uses structural guidance for large-hole diverse image completion. The SGDM first generates an edge image within missing regions, indicated by blue regions. Then, the SGDM generates images using the edge image as the structural guidance. The SGDM can generate diverse variations thanks to its stochastic capabilities.

Abstract

Diverse image completion, a problem of generating various ways of filling incomplete regions (\textit{i.e.} holes) of an image, has made remarkable success. However, managing input images with large holes is still a challenging problem due to the corruption of semantically important structures. In this paper, we tackle this problem by incorporating explicit structural guidance. We propose a structure-guided diffusion model (SGDM) for the large-hole diverse completion problem. Our proposed SGDM consists of a structure generator and a texture generator, which are both diffusion probabilistic models (DMs). The structure generator generates an edge image representing a plausible structure within the holes, which is later used to guide the texture generation process. To jointly train these two generators, we design a strategy that combines optimal Bayesian denoising and a momentum framework. In addition to the quality improvement, auxiliary edge images generated by the structure generator can be manually edited to allow user-guided image editing. Our experiments using datasets of faces (CelebA-HQ) and natural scenes (Places) show that our method achieves a comparable or superior trade-off between visual quality and diversity compared to other state-of-the-art methods.

1. Introduction

Image completion (also known as image inpainting) is a task to fill missing regions (\textit{i.e.} holes) of the target image with visually plausible contents. Image completion methods have been crucial in many applications, such as user-guided image editing [24, 80] and object removal [3, 9]. Humans possess the creative ability to guess the content of the missing regions, especially when the regions are large, in various rational ways. Therefore, image completion methods should ideally produce plausible yet diverse results while maintaining integrity and consistency with the visible regions.

How can we fill in holes in images with rational structure and visually realistic textures? Bertalmio \textit{et al}. [4]
describe how expert conservators restore damaged artworks as 1) figure out what content to put in missing regions, 2) draw contour edges for structural continuity, and 3) paint the regions with reasonable textures guided by the contours. Related to the second step, previous work [6,13,43,77] has introduced explicit structure guidance in image completion, resulting in better-quality image completion. An additional benefit of using structural guidance is to provide the opportunity for user-guided image editing [24,80], such as using sketching tools (see Fig. 1). However, it is still challenging to estimate reasonable structures when holes are large. Other previous work [46] has formulated the completion process as conditional image synthesis by the input condition of holes.

In this paper, we focus on diffusion probabilistic models (DMs) and explore incorporating structural guidance in the image completion process. We propose a structure-guided diffusion model (SGDM), which is the first DM that explicitly considers structural guidance using edge information; that is, we condition the textured image generation process on an edge image. Our framework consists of two networks: a structure generator that generates plausible edges and a texture generator that generates textures with the guidance of the edges, which aim to fill missing regions.

We present a novel joint-training strategy for these DM-based networks. The output of the network with time information prevents us from direct end-to-end joint training. To overcome it, we propose using optimal Bayesian denoising, in particular, Tweedie’s formula [14, 25, 59], which can denoise noisy edge images by a single step. However, this technique generates overly blurred edges depending on time. Therefore, we propose adopting a momentum framework [18,61] to perform the joint training effectively with blurred edges. That is, we prepare two networks during the joint training, the texture generator and the momentum texture generator, and update the weights of the momentum one as an exponential moving average (EMA) of the weights of the texture one. This framework allows us to use generated denoised edges and ground-truth edges simultaneously, which is the key to achieving stable training. In our experiments with datasets of faces (CelebA-HQ [26]) and natural scenes (Places [90]), we show that our method achieves a comparable or superior trade-off performance between visual quality and diversity using CelebA-HQ [26] and Places [90] datasets.

2. Related Work

2.1. Deterministic Image Completion

Early image completion approaches such as diffusion- [2,4,5,30] and patch-based [3,8–10,19] methods focused on non-learning prior knowledge. In particular, Bertalmio et al. [5] first considered texture and structure contexts simultaneously in missing regions. However, the traditional method suffered from complex scenes, including many different kinds of texture types or large missing regions.

Deep-learning-based methods using GANs [1,17,48] have demonstrated tremendous success in image completion [22,23,33–35,43,46,49,67–70,73–77,79,81,82,89,91]. To achieve fine-grained textures, numerous works proposed task-specific operations such as global and local discriminators [23], attention blocks [35,69,70,75,76,79,89], partial [34], and gated [77] convolutions. Concurrently, several works utilized explicit clues such as object edges [6,13,43,77], foreground contours [24,72,80], smoothed images [49], reference images [91], and semantic segmentation maps [32,33,58]. Nazeri et al. [43] first proposed a two-stage framework for edges and textures to introduce structure guidance. In this paper, we also employ edge images to explicitly provide structural clues for a texture generator.

2.2. Diverse Image Completion

Recent image completion studies have addressed more challenging issues, which fill up multiple visually plausible and diverse contents in a large hole [31,36,38,40,47,50,52,66,78,84,85,87,88]. Zheng et al. [87] first demonstrated the diverse image completion task. They and Zhao et al. [84] proposed a variational auto-encoder (VAE)-based framework, although their synthesized quality was limited because of variational training [86]. CoModGAN [85] and MAT [31] achieved a high-fidelity quality by introducing stochastic style representation [27,28], although their diversity was restricted due to the conditional training procedure.

Recent studies [38,66,78] have focused on an AR transformer [11,63] and achieved high-fidelity quality and diversity. However, ICT [66] and BAT-Fill [78] have a problem of information loss due to “low-resolution” and “quantization,” which are down-sampled into a much lower resolution (e.g. $32 \times 32$) and RGB values $256^3$ are quantized into a much lower dimension 512. To solve it, PUT [38] used a patch-based autoencoder with VQVAE [62] and applied the AR transformer to the vector quantized tokens. Nonetheless, the AR transformer-based approach has limitations on sampling orders [15], computational costs [16], and the lack of introducing explicit structural information. Because ARs
use only the sequence of preceding pixels, they cannot use a structure guidance that exists in the focused pixel.

2.3. Image Completion with Diffusion Models

Diffusion and score-based models have emerged as the family of likelihood-based models and shown remarkable success in quality, diversity, mode coverage, and generality in the training objective [12, 21, 44, 54, 56, 57]. Most previous studies [40, 54, 57] have shown the image completion results using unconditional image generation models. The completion can be performed by replacing the known region with the given image after each sampling step.

A major limitation of these methods is to produce images in which the content type matches the known region but is not semantically consistent and harmonious. To solve it, several methods have proposed [40, 52] to explicitly learn or deal with the completion problem. Palette [52] has outperformed task-specific baselines without any specific architecture and loss function by learning the completion problem. RePaint [40] proposed a conditional sampling method, which alternately performs the forward and reverse diffusion processes, for pre-trained unconditional generation models and showed the DMs could generate a variety of samples for large holes. However, these methods often fail to synthesize structural contents that satisfy the given context. Our method overcomes this limitation by explicitly estimating the structure of missing regions and using it as guidance.

3. Preliminaries

Diffusion probabilistic models (DMs) formulate a mapping from an empirical data distribution to a standard Gaussian distribution via an iterative denoising process of varying noise scales. Here we describe denoising diffusion probabilistic models (DDPM) [21, 44], which we adopt in this paper. We also describe optimal Bayesian denoising, which we use to enable our joint training.

3.1. Diffusion Models

DDPM defines a discrete Markov chain between two processes: forward and reverse processes. Starting from (noiseless) data $x_0$, the forward process at timestep $t$ adds Gaussian noise to the previous data $x_{t-1}$ to obtain the current data $x_t$. This process is modeled as $q(x_t|x_0) \sim \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t \mathbf{1})$, where $\beta_t$ is a pre-defined noise scale depending on timestep $t$. By accumulating the timesteps, we can write $q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\alpha_t}x_0, (1-\alpha_t) \mathbf{1})$, where $\alpha_t := \prod_{s=1}^{t}(1-\beta_s)$. The reverse process iteratively samples $x_{t-1}$ from $x_t$, starting from $x_T$. This process is modeled as $p_{\theta}(x_{t-1}|x_t) \sim \mathcal{N}(x_{t-1}; \mu_{\theta}(x_t, t), \Sigma_{\theta}(x_t, t))$. DDPM uses a UNet [51] to approximate this posterior. For conditional cases, the UNet $e_\theta$ can be trained using a denoising score matching loss $\|e_\theta(x_t, t, y) - \epsilon\|^2_2$, where $\epsilon$ is a noise added to $x_0$ to create $x_t, y$ is a condition such as an image and an edge image with a hole in our case, and $e_\theta(x_t, t, y)$ represents a score function of the perturbed data distribution, $\nabla_{x_t} \log p_{\theta}(x_t)$ [56]. After the training, DDPM can produce samples with annealed Langevin dynamics [56, 65].

3.2. Optimal Bayesian Denoising

Optimal Bayesian denoising is a technique for performing a minimum mean square error (MMSE) denoising in a single step. To perform denoising for a Gaussian variable $z \sim \mathcal{N}(z; \mu, \Sigma)$, an MMSE estimator is given by Tweedie’s formula [14, 25, 59]; that is, $\mathbb{E}[\mu|z] = z + \Sigma \nabla_z \log p(z)$.

In DDPM, the forward step is modeled as $q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\alpha_t}x_0, (1-\alpha_t) \mathbf{1})$ as described in Sec. 3.1. Thus, we can apply Tweedie’s formula here by substituting $\sqrt{\alpha_t}x_0$ and $(1-\alpha_t) \mathbf{1}$ for $\mu$ and $\Sigma$, respectively. This allows us to determine a single-step denoising operation as

$$F(x_t) \coloneqq \hat{x}_t^0 = x_t + (1-\alpha_t) \nabla_{x_t} \log p(x_t) \sqrt{\alpha_t},$$

where $x_0^t$ represents a denoised sample. With this operation, we can convert the noisy sample (at time $t$) into the denoised one (at time 0) by a single step, as long as the optimal score function $\nabla_{x_t} \log p(x_t)$ is known. Refer to [7, 29, 41] for more detailed discussions.

4. Structure-Guided Diffusion Model

Given an input image with missing regions (i.e., holes), we aim to generate a semantically reasonable image that respects the context of the visible regions. We denote the target image by $I \in \mathbb{R}^{3 \times H \times W}$, the binary mask representing the missing regions by $M \in \{0, 1\}^{1 \times H \times W}$, and the generated image by $\hat{I} \in \mathbb{R}^{3 \times H \times W}$, where $H$ and $W$ represent a spatial resolution. With this notation, the goal is to generate $\hat{I}$ from $I_M = I \odot M$. The SGDM uses structural guidance in its generation process. Specifically, it generates a hole-filled edge image $\hat{E}$ and then uses it as structural guidance to generate $\hat{I}$. This edge image $\hat{E}$ is generated using an edge image with missing regions, denoted by $E_M$, which is generated from $I_M$ using an existing edge detection algorithm, Holistically-Nested Edge Detection (HED) [71].

4.1. Framework Architecture

Our framework consists of two DM-based networks: a structure generator $f_\theta$ and a texture generator $g_\theta$. The structure generator aims to generate an edge image that guides the texture generator. We employ UNet [12] for both networks. We attach five additional channels in the first layer of both networks following [45, 53] to take the conditions of the image and edge image with missing regions.

First, the structure generator fills in the holes of the edge image $E_M$ to produce the hole-filled edge image $\hat{E}$. Then,
the texture generator produces the plausible texture with the guidance of $E$ maintaining the context of the visible regions of $I$. These generations use the iterative sampling of DMs as described in Sec. 3.1.

### 4.2. Individual Training

We describe the data preparation and the training procedure. Suppose we have a ground-truth image $I$ from a training dataset. Then, we extract an edge image $E$ using HED. We degrade the image $I$ and the edge image $E$ using a binary mask $M$ that is randomly drawn for each sample, denoted as the masked image $I_M = I \cap M$ and the masked edge image $E_M = E \cap M$, respectively. We fill the masked region with zeros. To train DMs, we create a noisy image $I_t$ and a noisy edge image $E_t$ at timestep $t$ (out of $T$ timesteps) using Gaussian noises $\epsilon_I$ and $\epsilon_E$ with $I$ and $E$, respectively.

Figure 2 shows our individual training process. The structure generator $f_\theta$ generates, from the noisy edge image at the timestep $t$, a less noisy edge image at the previous timestep $t-1$. More specifically, given the noisy edge image $E_t \in \mathbb{R}^{3 \times H \times W}$, masked image $I_M$, mask $M$, masked edge image $E_M$, and timestep $t$, it outputs a less noisy edge image $\hat{E}_{t-1} \in \mathbb{R}^{3 \times H \times W}$. Similarly, given the noisy image $I_t \in \mathbb{R}^{3 \times H \times W}$, masked image $I_M$, mask $M$, edge image (without noise or masked regions) $E$, and timestep $t$, the texture generator $g_\phi$ outputs $\hat{I}_{t-1} \in \mathbb{R}^{3 \times H \times W}$. These processes can be written as

$$f_\theta(E_t, I_M, M, E_M, t) = \hat{E}_{t-1},$$

$$g_\phi(I_t, I_M, M, E, t) = \hat{I}_{t-1}.$$  

Both networks can be trained via the denoising score matching loss [21] in a closed form as described in Sec. 3.1,

$$L_f = \mathbb{E}_{I, M, E, t, \epsilon_E} \|f_\theta(E_t, I_M, M, E_M, t) - \epsilon_E \|^2_2,$$

$$L_g = \mathbb{E}_{I, M, E, t, \epsilon_I} \|g_\phi(I_t, I_M, M, E, t) - \epsilon_I \|^2_2,$$

where the noises $\epsilon_E$ and $\epsilon_I$ are sampled from Gaussian distribution to create $E_t$ and $I_t$, respectively.

### 4.3. Joint Training

The individually trained structure generator sometimes generates semantically unreasonable edges. This is because edge images are sparse compared to textured images, so the modeling is more difficult. To mitigate this issue, after the individual training, we jointly fine-tune both networks in an end-to-end manner. For this, we propose a novel joint-training strategy using optimal Bayesian denoising and a momentum framework, as shown in Fig. 3. The joint training of DMs is not performed in a straightforward fashion, such as in the training of GANs [13, 43]. This is because the texture generator needs to take a noisecess edge image as input, but the structure generator cannot generate a noiseless edge image without iterations. Even if we produce a noiseless edge image via iterative sampling, backpropagation is intractable due to the cause of gradient accumulation and computational costs. We solve this issue by applying the single-step denoising operation in Eq. (1); that is, we obtain a noiseless estimate by $\hat{E}_0 = F(E_t)$. We further average it in the channel dimension and binarize it using a threshold of 0.5. This allows us to perform the backpropagation in an end-to-end manner.

However, the denoised edge image $\hat{E}_0$ tends to be overly blurred (especially when $t$ is close to $T$), as shown in Fig. 4. We observe that the gap between the original edge image
where \( weight \) updating. To avoid corruption, we introduce without a sufficient difference between the prediction and ground-truth representations. Specifically, we replicate the texture generator for these networks and train them with the structure generator's output as an EMA of the weights of the network being trained. Other is the momentum network where the weight is updated between visual quality and diversity. We also analyzed the effectiveness of the joint training via an ablation study. We considered different sizes of missing regions conditions to see the performance change.

Datasets. The experiments were conducted with CelebA-HQ [26] and Places [90], which cover different degrees of context (face only vs. diverse natural scenes). The image resolution of both datasets was \( 256 \times 256 \). For CelebA-HQ, we prepared a train set and a test set with 24,183 and 5,000 images, respectively. The test set was newly created by combining the original split and validation set and then selecting 5,000 images randomly. For Places, we prepared a train set and a test set with 8 million (M) and 5,000 images. The test set of Places was created by randomly sampling 5,000 images.
images from the original validation set. When evaluating the diversity, we randomly selected 50 images from each test set. For a better understanding of the performances for holes with various sizes, we prepared three different masks (i.e., large, small, and center masks) and applied them for each image; see the supplementary material for details.

**Implementation details.** Before the training, we first initialized our generators’ weights with the pre-trained weights from [40], which were trained using an unconditional image generation task. For the individual training, each network was trained for 10M images on CelebA-HQ and 20M images on Places, respectively. Additionally, we carried out the joint training with 10M images. The batch size was fixed to 1. Both trainings were performed with AdamW optimizer [39] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ and a learning rate of $10^{-4}$. We conducted all experiments with four NVIDIA A100 GPUs. For the evaluation, we generated edge images and images using 1,000 sampling steps and 4,570 sampling steps using RePaint [40], respectively.

**Compared methods.** We compared our method with the following methods: MAT [31], CoModGAN (CMGAN) [85], BAT-Fill [85], PUT [38], RePaint [40], ADM [12], and Palette [52] using their pre-trained weights. MAT and CMGAN used a resolution of $512 \times 512$ while other methods used a resolution of $256 \times 256$. See Tab. 1 for what modeling each method used. RePaint, ADM, and ours used the same pre-trained weights. The difference between RePaint and ADM was the sampling procedure. We did not use CMGAN for CelebA-HQ since pre-trained weights were not publicly available.

**Evaluation metrics.** We considered two different aspects: visual quality and diversity. For the quality evaluation, we used two perceptual metrics: Fréchet inception distance (FID) [20] and Learned Perceptual Image Patch Similarity (LPIPS) [83]. FID measures the discrepancy between two sets of ground truth and generated image. LPIPS measures the perceptual similarity between each sample of them. For each mask setting, we calculated the mean value of these perceptual metrics for the 5,000 test images. For the diversity evaluation, we defined a diversity score for a given set of generated images, $X = \{x_i\}_{i=1}^N$, as

$$\text{Div}(X) = \frac{2}{N(N-1)} \sum_{i<j} (1 - \text{CosSim}(\Phi(x_i), \Phi(x_j))),$$

where $\text{CosSim}$ represents a cosine similarity between two image feature vectors and we set $N = 100$. We extracted image features by InceptionV3 [60] and applied the global average pooling to obtain the feature vector, as represented by $\Phi(\cdot)$. A higher score indicates that the generated set $X$ is more diverse. We then calculated the mean and standard deviation of the diversity scores for the 50 test images.

**5.1. Quantitative Comparisons**

Tables 1 and 2 show the completion performance with different masked regions on CelebA-HQ and Places, respectively. MAT and CMGAN far outperformed other methods in all visual qualities. These GANs-based methods, however, showed the lowest diversity. Our method achieved superior or comparable performance to BAT-Fill and PUT, which used ARs. Compared with RePaint, our method showed the advantage for all metrics on CelebA-HQ. However, on Places, RePaint achieved superior or comparable results to our method. ADM consistently showed the highest diversity score over all datasets. This implies that there was a trade-off between visual quality and diversity, as shown in Fig. 5. We conjecture that this was because, for visual quality, our method often generated some artifacts when edge generation was inaccurate. For diversity, the explicit introduction of structure would suppress the generation of samples that were not contextually relevant. As a result, our method tended to be less diverse than the methods using DMs, especially for Places with complex natural structures.

**5.2. Qualitative Comparisons**

Figure 6 shows a qualitative comparison of the competing methods. Each image was generated using a different seed. Our method generated various rational edge images and realistic images. From our variations from a single edge image, we observed that our method was able to synthesize texture and color variations while maintaining the structural context. In particular, our variations 3 and 4 successfully completed the reasonable structure and texture, even at the edges where no details were drawn. This implies that the texture generator could learn the semantic context from the partially visible region as well as from the edges. RePaint, ADM, and Palette, which use diffusion models, generated plausible images, but the semantic consistency was insufficient.
Figure 6. **Qualitative comparisons of the proposed SGDM with the state-of-the-art methods.** The SGDM generates rational structure and visually realistic textures. In addition, the SGDM produces multiple plausible images for each structural guidance by the edge images.

We also show comparisons between our method and existing methods with structural guidance, DeepFillv2 [77] and ZITS [13], in Fig. 8. Both comparison methods used GANs for generative modeling. For DeepFillv2, we had the generator to generate edge images for obtaining the guidance and the image generator. For ZITS, we used the officially provided pre-trained weights. As shown in Fig. 8, they failed to generate valid edge images for large holes. Although ZITS used the transformer for the global receptive field, it generated incompleted edges and blurred textures. In contrast, our diffusion-based method could generate plausible edges and images even for large masked regions, achieving higher quality.

5.3. Ablation Study

We conducted an ablation study as shown in Tab. 3. We generated edge images and images using DDIM [55] and PNDM [37] sampler with 100 sampling steps, respectively. We compared SGDMs with several settings: (a) no structural guidance, (b) only the individual training, (c) joint training after the individual training (d) using $L_c$ in Eq. (6) (e) with the EMA-based weight update strategy.
The visual quality of condition (a), which used no edge images, was worse than others (b-d), which used structural guidance provided by an edge. In contrast, the diversity of (a) achieved better because there were unstructured low-quality results. We see that the structural guidance could improve visual quality at the expense of diversity. Compared (b) to (c), we see that the joint training brought quality improvements. Furthermore, introducing \( \mathcal{L}_c \) and the EMA clearly boosted the visual quality. We note that condition (d) was numerically unstable, and we failed to produce results. This suggests that the use of the EMA was the key to achieving stable training. Figure 7 shows the visual comparison among conditions (a), (b), and (e). We can see that our method could generate more visually plausible images thanks to the structural guidance as edges and the joint training. After the joint training, we observed that our method tended to produce flatter edge images thanks to the improvement of robustness.

5.4. User-Editing Capability

Figure 1 shows user-edited results, where we edited the structure and shape of the eyes, expression, rock, and buildings of the target images. Our method was sufficiently robust so that we could easily edit the target image and perform quick trial-and-error procedures.

6. Limitation and Discussion

Failure cases. Our method often failed to generate semantically correct and adequately closed edges. For example, our method failed to generate a plausible structure for the human hand in Fig. 7 (g). In the right side of “Ours 2” in Fig. 8, while our method followed the edge guidance, the image did not have a reasonable structure. To avoid such artifacts, it is necessary to manually correct inadequate edges, as shown in the user-edited examples in Fig. 1.

Computational costs. Our method requires the iterative denoising process. In contrast, GANs can generate images in a single step, meaning that our method inherently takes more time than GAN-based methods. ARs-based methods such as BAT-Fill and PUT also require iterative inference, but their computational cost is lower than ours, thanks to their low-resolution modeling. To better understand the computational cost issue, we measured the time needed to complete a center-masked image for each method. As a result, MAT (GANs-based) needed 0.098 seconds, and PUT and BAT-Fill (ARs-based) required 4.06 and 4.49 seconds, respectively. Our method with 1,000 steps and with the sampling method of RePaint, on the other hand, took 71.15 and 157.53 seconds, respectively. To mitigate the computational cost problem, we can use acceleration methods such as DDIM and PLMS; our method with DDIM and PLMS took 6.80 and 12.47 seconds, respectively. This was more than 10× faster, and it could provide previews at a user-friendly and practical speed.

Potential societal impact. Our method inherits the potential societal impact of previous image completion methods (e.g. [40]). On the negative side, generated images may directly reflect the biases in the datasets, including gender, age, ethnicity, and so on. Additionally, our method can easily edit images by providing desired edge structures; thus, it could facilitate DeepFake creation [42, 64]. On the positive side, image completion may enhance privacy protection by replacing any identity-related information from photographs taken in public spaces.

7. Conclusion

We have presented the first diffusion-based model that considers structural guidance in the image generation process, called the structure-guided diffusion model (SGDM). The SGDM can generate rational structures and visually realistic textures. We have proposed a novel training strategy to enable effective end-to-end training. Extensive experiments show that the SGDM achieves a comparable or superior visual quality and diversity trade-off on CelebA-HQ and Places as compared with the state-of-the-art. Explicitly incorporating structural guidance using edge information has not only improved the visual quality but also enabled user-guided image editing.

Table 3. Result of the ablation study on CelebA-HQ with large holes. Indiv. and Joint indicate individual and joint training settings, respectively. The condition (e) represents the proposed method. In the condition (d), the training was unstable and failed.

| Condition                        | CelebA-HQ | Large hole |
|----------------------------------|-----------|------------|
| FID ↓                            | LPIPS ↓   | Div ↑      |
| w/o Edge                         | 24.70     | 0.153      | 0.083 ± 0.083 |
| Indiv. only                      | 22.20     | 0.136      | 0.068 ± 0.060 |
| Indiv. + Joint + EMA             | 22.05     | 0.134      | 0.068 ± 0.058 |
| Indiv. + Joint + \( \mathcal{L}_c \) | N/A       | N/A        | N/A         |
| Indiv. + Joint + \( \mathcal{L}_c \) + EMA | 21.42     | 0.131      | 0.087 ± 0.059 |

Figure 8. Visual comparison among image completion methods with structural guidance [13, 77]. Our method generates rational structures and rational and realistic textures, even with large holes.
References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein Generative Adversarial Networks. In ICML, 2017. 2

[2] C. Ballester, M. Bertalmio, V. Caselles, G. Sapiro, and J. Verdera. Filling-In by Joint Interpolation of Vector Fields and Gray Levels. IEEE TIP, 10(8):1200–1211, 2001. 2

[3] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B. Goldman. PatchMatch: A Randomized Correspondence Algorithm for Structural Image Editing. ACM TOG, 28(3), 2009. 1, 2

[4] Marcelo Bertalmio, Guillermo Sapiro, Vincent Caselles, and Coloma Ballester. Image Inpainting. In SIGGRAPH, 2000. 1, 2

[5] M. Bertalmio, L. Vese, G. Sapiro, and S. Osher. Simultaneous Structure and Texture Image Inpainting. IEEE TIP, 12(8):882–889, 2003. 2

[6] Chenjie Cao and Yanwei Fu. Learning a Sketch Tensor Space for Image Inpainting of Man-made Scenes. In ICCV, 2021. 2

[7] Hyungjin Chung, Byeongsu Sim, Dohoon Ryu, and Jong Chul Ye. Improving Diffusion Models for Inverse Problems using Manifold Constraints. In NeurIPS, 2022. 3

[8] A. Criminisi, P. Perez, and K. Toyama. Object Removal by Exemplar-Based Inpainting. In CVPR, 2003. 2

[9] A. Criminisi, P. Perez, and K. Toyama. Region filling and object removal by exemplar-based image inpainting. IEEE TIP, 13(9):1200–1212, 2004. 1, 2

[10] Soheil Darabi, Eli Shechtman, Connelly Barnes, Dan B. Goldman, and Pradeep Sen. Image Melding: Combining Inconsistent Images Using Patch-Based Synthesis. ACM TOG, 31(4), 2012. 2

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In NAACL, 2019. 2

[12] Prafulla Dhariwal and Alex Nichol. Diffusion Models Beat GANs on Image Synthesis. In NeurIPS, 2021. 3, 5, 6

[13] Qiaole Dong, Chenjie Cao, and Yanwei Fu. Incremental Transformer Structure Enhanced Image Inpainting with Masking Positional Encoding. In CVPR, 2022. 2, 4, 7, 8

[14] Bradley Efron. Tweedie’s Formula and Selection Bias. Journal of the American Statistical Association, 106(496):1602–1614, 2011. 2, 3

[15] Patrick Esser, Robin Rombach, Andreas Blattmann, and Björn Ommer. ImageBART: Bidirectional Context with Multimodal Diffusion for Autoregressive Image Synthesis. In NeurIPS, 2021. 2

[16] Patrick Esser, Robin Rombach, and Björn Ommer. Taming Transformers for High-Resolution Image Synthesis. In CVPR, 2021. 2

[17] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. In NeurIPS, 2014. 2

[18] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning. In NeurIPS, 2020. 2, 5

[19] James Hays and Alexei A. Efros. Scene Completion Using Millions of Photographs. ACM TOG, 26(3):4–12, 2007. 2

[20] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, Günter Klambauer, and Sepp Hochreiter. GANs Trained by a Two Time-Scale Update Rule Converge to a Nash Equilibrium. In NeurIPS, 2017. 6

[21] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. In NeurIPS, 2020. 3, 4

[22] Yibing Song Wei Huang Hongyu Liu, Bin Jiang and Chao Yang. Rethinking Image Inpainting via a Mutual Encoder-Decoder with Feature Equalizations. In ECCV, 2020. 2

[23] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and Locally Consistent Image Completion. ACM TOG, 36(4), 2017. 2

[24] Younggjo Jo and Jongyoul Park. SF-FEGAN: Face Editing Generative Adversarial Network With User’s Sketch and Color. In ICCV, 2019. 1, 2

[25] Miyasawa K. An empirical Bayes estimator of the mean of a normal population. Bull. Internat. Statist., 38:181–188, 1961. 2, 3

[26] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In ICLR, 2018. 2, 3, 12

[27] Tero Karras, Samuli Laine, and Timo Aila. A Style-Based Generator Architecture for Generative Adversarial Networks. In CVPR, 2019. 2

[28] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and Improving the Image Quality of StyleGAN. In CVPR, 2020. 2

[29] Bahjat Kawar, Gregory Vaksman, and Michael Elad. SNIPS: Solving Noisy Inverse Problems Stochastically. In NeurIPS, 2021. 3

[30] Levin, Zomet, and Weiss. Learning How to Inpaint from Global Image Statistics. In ICCV, 2003. 2

[31] Wenbo Li, Zhe Lin, Kun Zhou, Lu Qi, Yi Wang, and Jiaya Jia. MAT: Mask-Aware Transformer for Large Hole Image Inpainting. In CVPR, 2022. 2, 5, 6, 7, 12, 13, 14, 15, 16, 17

[32] Liang Liao, Jing Xiao, Zheng Wang, Chia-Wen Lin, and Shin’ichi Satoh. Guidance and evaluation: Semantic-aware image inpainting for mixed scenes. In ECCV, 2020. 2

[33] Liang Liao, Jing Xiao, Zheng Wang, Chia-Wen Lin, and Shin’ichi Satoh. Image Inpainting Guided by Coherence Priors of Semantics and Textures. In CVPR, 2021. 2

[34] 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 Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, ECCV, 2018. 2

[35] Hongyu Liu, Bin Jiang, Yi Xiao, and Chao Yang. Coherent Semantic Attention for Image Inpainting. In ICCV, 2019. 2

[36] Hongyu Liu, Ziyu Wan, Wei Huang, Yibing Song, Xintong Han, and Jing Liao. PD-GAN: Probabilistic Diverse GAN for Image Inpainting. In CVPR, 2021. 2
Zhaoyi Yan, Xiaoming Li, Mu Li, Wangmeng Zuo, and Shiguang Shan. Shift-Net: Image Inpainting via Deep Feature Rearrangement. In ECCV, 2018. 2

Raymond A. Yeh∗, Chen Chen∗, Teck Yian Lim, Schwing Alexander G., Mark Hasegawa-Johnson, and Minh N. Do. Semantic Image Inpainting with Deep Generative Models. In CVPR, 2017. 2

Zili Yi, Qiang Tang, Shekoofeh Azizi, Daesik Jang, and Zhan Xu. Contextual Residual Aggregation for Ultra High-Resolution Image Inpainting. In CVPR, 2020. 2

Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S. Huang. Generative Image Inpainting with Contextual Attention. In CVPR, 2018. 2

Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S. Huang. Free-Form Image Inpainting With Gated Convolution. In ICCV, 2019. 2, 7, 8, 12

Yingchen Yu, Fangneng Zhan, Rongliang WU, Jianxiong Pan, Kaivwen Cui, Shijian Lu, Feiyiing Ma, Xuansong Xie, and Chunyan Miao. Diverse Image Inpainting with Bidirectional and Autoregressive Transformers. In ACM MM, 2021. 2, 5, 6, 7, 13, 14, 15, 16, 17

Yanhong Zeng, Jianlong Fu, Hongyang Chao, and Baining Guo. Learning Pyramid-Context Encoder Network for High-Quality Image Inpainting. In CVPR, 2019. 2

Yu Zeng, Zhe Lin, and Vishal M. Patel. SketchEdit: Mask-Free Local Image Manipulation with Partial Sketches. In CVPR, 2022. 1, 2

Yu Zeng, Zhe Lin, Jimei Yang, Jianming Zhang, Eli Shechtman, and Huchuan Lu. High-Resolution Image Inpainting with Iterative Confidence Feedback and Guided Upsampling. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, ECCV, 2020. 2

Haoran Zhang, Zhenzhen Hu, Changzhi Luo, Wangmeng Zuo, and Meng Wang. Semantic Image Inpainting with Progressive Generative Networks. In ACM MM, 2018. 2

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In CVPR, 2018. 6

Lei Zhao, Qihang Mo, Sihuan Lin, Zhizhong Wang, Zhiwen Zuo, Haibo Chen, Wei Xing, and Dongming Lu. UCTGAN: Diverse Image Inpainting Based on Unsupervised Cross-Space Translation. In CVPR, 2020. 2

Shengyu Zhao, Jonathan Cui, Yilun Sheng, Yue Dong, Xiao Liang, Eric I Chang, and Yan Xu. Large Scale Image Completion via Co-Modulated Generative Adversarial Networks. In ICLR, 2021. 2, 5, 6

Shengjia Zhao, Jiaming Song, and Stefano Ermon. Towards Deeper Understanding of Variational Autoencoding Models. arXiv preprint arXiv:1702.08658, 2017. 2

Chuanxiao Zheng, Tat-Jen Cham, and Jianfei Cai. Pluralistic image completion. In CVPR, 2019. 2

Chuanxiao Zheng, Tat-Jen Cham, and Jianfei Cai. Pluralistic Free-Form Image Completion. IJCV, 129(10):2786–2805, 2021. 2

Chuanxiao Zheng, Tat-Jen Cham, Jianfei Cai, and Dinh Phung. Bridging Global Context Interactions for High-Fidelity Image Completion. In CVPR, 2022. 2

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 Million Image Database for Scene Recognition. IEEE TPAMI, 40(6):1452–1464, 2018. 2, 5, 12

Yuqian Zhou, Connelly Barnes, Eli Shechtman, and Sohrab Amirghodsi. TransFill: Reference-Guided Image Inpainting by Merging Multiple Color and Spatial Transformations. In CVPR, 2021. 2
Appendix

A. Statistics for Masks

We followed MAT [31] and DeepFill v2 [77] to create masks. Figures A and B show statistics of large and small masks for visual quality and diversity evaluations, respectively. In the training, we created zero to five full-size or half-size rectangles and zero to nine random strokes with 12 to 48 width and 4 to 18 vertex. In the evaluation, we additionally created small masks. For the small masks, we created zero to three full-size or half-size rectangles and zero to four random strokes with 12 to 48 width and 4 to 18 vertex. For the center masks, we removed the top and bottom center area of the 256 × 256 image from 68 to 192 pixels.

B. Detailed Settings of Ablation Study

The goal of this ablation study was to investigate how each technical element of our model affects the performance. We evaluated the performance of the visual quality and diversity metrics in several different settings. To conduct this ablation study, we newly created a subset of the test of CelebA-HQ [26] (see Sec. 5 in the main text) for the ablation study as follows. For evaluating the visual quality metrics, we randomly selected 1,000 images from the test set that consisted of 5,000 images. For evaluating the diversity metric, we randomly selected 20 images from the test set that consisted of 50 images. We used the large mask setting for this study. For each test image, we generated a completed image (visual quality) and 50 completed images (diversity), as well as their corresponding edge images. Then, we calculated the scores of visual quality and diversity.

C. Quality Improvement by User Correction

The SGDM sometimes generated semantically unreasonable edges, which is our limitation. However, one of the advantages of the SGDM, which explicitly incorporates structural guidance, is that the user can easily correct such failures by editing the generated edge images. As shown in Fig. C, the user could resolve the artifacts as indicated by the orange arrows by user correcting.

D. Additional Results

We present more qualitative comparisons of the proposed SGDM with the state-of-the-art methods. Figures D and E show the results using large and center masks on CelebA-HQ [26], respectively. Figures F and G and Fig. H show the results using large and center masks on Places [90], respectively.
Figure D. Qualitative comparisons of the proposed SGDM with the state-of-the-art methods using CelebA-HQ and large masks.
Figure E. Qualitative comparisons of the proposed SGDM with the state-of-the-art methods using CelebA-HQ and center masks.
Figure F. Qualitative comparisons of the proposed SGDM with the state-of-the-art methods using Places and large masks.
Figure G. Qualitative comparisons of the proposed SGDM with the state-of-the-art methods using Places and large masks.
Figure H. Qualitative comparisons of the proposed SGDM with the state-of-the-art methods using Places and center masks.