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

Daichi Horita[1]
horita@hal.t.u-tokyo.ac.jp
Jiaolong Yang[2]
jiaoyan@microsoft.com
Dong Chen[2]
doch@microsoft.com
Yuki Koyama[3]
koyama.y@aist.go.jp
Kiyoharu Aizawa[1]
aizawa@hal.t.u-tokyo.ac.jp
Nicu Sebe[4]
niculae.sebe@unitn.it

[1] The University of Tokyo
[2] Microsoft Research Asia
[3] National Institute of Advanced Industrial Science and Technology (AIST)
[4] University of Trento

Abstract

Image completion techniques have made significant progress in filling missing regions (*i.e.*, holes) in images. However, large-hole completion remains challenging due to limited structural information. In this paper, we address this problem by integrating explicit structural guidance into diffusion-based image completion, forming our structure-guided diffusion model (SGDM). It consists of two cascaded diffusion probabilistic models: structure and texture generators. The structure generator generates an edge image representing plausible structures within the holes, which is then used for guiding the texture generation process. To train both generators jointly, we devise a novel strategy that leverages optimal Bayesian denoising, which denoises the output of the structure generator in a single step and thus allows backpropagation. Our diffusion-based approach enables a diversity of plausible completions, while the editable edges allow for editing parts of an image. Our experiments on natural scene (Places) and face (CelebA-HQ) datasets demonstrate that our method achieves a superior or comparable visual quality compared to state-of-the-art approaches. The code is available for research purposes at `https://github.com/UdonDa/Structure_Guided_Diffusion_Model`.

1 Introduction

Image completion aims to fill missing regions (*i.e.* holes) in images with visually coherent content. Prior work has used guidance clues such as edges [34, 60] and semantic maps [22, 23] to divide the problem into structure and texture generation. These attempts have enabled various image-editing applications like object removal [35], insertion [38], and manipulation [12, 64]. However, *large-hole* completion remains a challenge because of the difficulty in generating useful guidance clues.

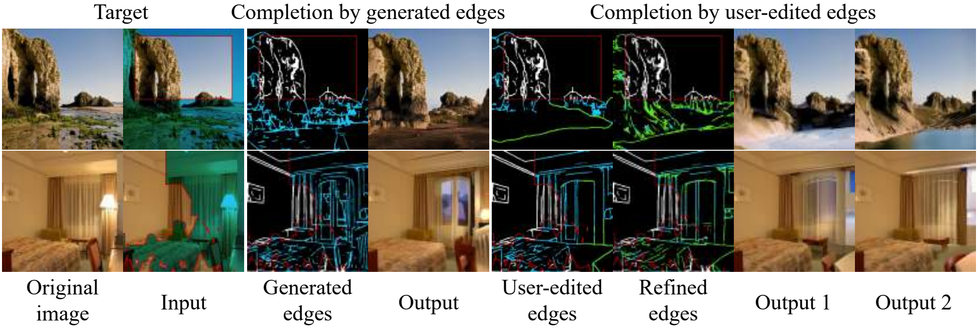

Figure 1: The SGDM first generates edges within missing regions, indicated by blue. Then, it produces textured images using the edges as structural guidance. Optionally, the edges can be manually edited, which are then refined by SDEdit [31] using the SGDM's prior, represented by green. The SGDM's stochastic process allows for generating diverse outputs.

Image completion has been achieved using various techniques, including convolutional neural networks [18] and generative adversarial networks (GANs) [8, 15]. While GAN-based methods [19, 69] are good at filling large holes in flat texture patterns, they often produce images with distorted structures. Researchers have tried to integrate structural guidances [5, 34, 60] to fill holes with rational structure; however, they still struggle to generate reasonable clues for large holes. Recently, autoregressive (AR) transformers [51, 55] and diffusion models (DMs) [10, 48] have gained attention as promising techniques in image completion. Their ability to generate diverse results is an additional strength from an application viewpoint. Still, these techniques struggle to fill holes with coherent structures. Thus, providing diverse ways to complete large missing regions with coherent structures remains challenging.

To address this challenge, we focus on DMs and explore the incorporation of structural guidance into the image completion process. We propose a *structure-guided diffusion model* (SGDM), which explicitly considers structural guidance using *edge* information. Our framework consists of a structure generator that generates plausible edges and a texture generator that completes textures guided by the edges. Leveraging structural guidance and DMs, our SGDM can complete large holes with diverse and structurally coherent results. Additionally, structural guidance provides opportunities for user-guided image editing, such as using sketching tools (see Fig. 1).

To train the structure and texture generators simultaneously, we present a novel joint-training strategy using *optimal Bayesian denoising*, specifically Tweedie's formula [6, 13, 49], which can denoise noisy images in a single step and thus allows backpropagation. End-to-end training of two cascaded DM-based generators is a non-trivial problem; the SGDM's training is conditioned by noise, preventing the texture generator from directly using the noisy edge map from the structure generator. We solve this problem by using optimal Bayesian denoising, improving the generalizability as in multi-task learning [33, 42]. Our experiments with natural scene (Places [73]) and face (CelebA-HQ [14]) datasets demonstrate that our method achieves a superior or comparable visual quality compared to state-of-the-art methods.

Our contributions are summarized as follows. 1) We propose the structure-guided diffusion model (SGDM) for large-hole image completion. It consists of structure and texture generators producing coherent, realistic contexts in large holes. As far as we know, this is the first work combining structural generation and guidance with diffusion models for image completion. 2) We design a novel joint-training strategy using optimal Bayesian denoising to enable end-to-end training of two cascaded DM-based generators. 3) We show that SGDM achieves state-of-the-art or comparable visual quality on both Places [73] and CelebA-HQ [14] datasets.

2 Related Work

2.1 Deterministic Image Completion

The advent of deep learning brought significant success to image completion, especially GAN-based methods [11, 72]. To achieve fine-grained textures, many works proposed task-specific operations such as global and local discriminators [11], attention mechanisms [25, 57, 59, 62], partial [24], and gated [60] convolutions. Concurrently, several works utilized explicit clues such as object edges [3, 5, 34], foreground contours [12, 58, 64], smoothed images [39], reference images [74], confidence maps [63], and semantic segmentation maps [22, 23]. Nazeri *et al*. [34] first proposed a two-stage framework for edges and textures, introducing structure guidance. ZITS [5] used an attention-based transformer [53] to predict structural guidance. However, these methods still have difficulties in predicting guidance within large missing regions. In contrast, the use of DMs with strong model capacity and novel joint training generates reasonable edges.

2.2 Diverse Image Completion

Recent image completion studies have addressed more challenging issues, *i.e*., filling large holes in images with multiple visually plausible and diverse contents [20, 26, 36, 56, 71]. Variational-auto-encoder-based methods [67, 70] demonstrated diverse image completion, although their synthesized quality was limited due to variational training [68]. Subsequently, CoModGAN [69] and MAT [19] successfully filled large holes, particularly in flat texture patterns. However, they often produce images with distorted or unrealistic structures and limited diversity.

Recent studies [27, 55, 61] have focused on an AR transformer [53], achieving high-fidelity quality and diversity. However, AR-based methods [55, 61] faced information loss issues due to “low resolution” and “quantization,” which down-sample images to a much lower resolution (*e.g*. 32×32) and quantize RGB values 256^3 into a much lower dimension 512. To address this, PUT [27] used a patch-based autoencoder with VQVAE [52] and applied the AR transformer to vector quantized tokens. Nonetheless, AR transformer-based approaches struggle with handling high-resolution images and sampling orders [7] due to their pixel-by-pixel autoregressive sampling. In contrast, our proposed method overcomes these limitations by leveraging diffusion models and explicit structural guidance, which allows for generating coherent structures and diverse images.

2.3 Image Completion with Diffusion Models

Diffusion and score-based models have emerged as a family of likelihood-based models, showing remarkable success in quality, diversity, mode coverage, and generality in their training objective [10, 47]. Most previous studies [30, 45, 48] have demonstrated image completion using unconditional image generation models by replacing the known region with a designated hole at each sampling step. However, a major limitation of these methods is their inability to produce harmonious images that match the known regions. To address this, Palette [43] learned conditional completion, and RePaint [30] introduced a conditional sampling method, which alternately performs the forward and reverse diffusion processes for pre-trained unconditional models. Nonetheless, these methods often generate irrelevant content for large holes. Our method overcomes this limitation by explicitly estimating the structure of missing regions and using it as guidance. ControlNet [65] proposed an encoder to support additional input conditions for pretrained diffusion models. It also uses an edge map as structural guidance to fill holes. However, edges within holes must be prepared beforehand. In contrast, our method learns to generate edges within missing regions and then synthesize textures.

3 Preliminaries

This section provides an overview of diffusion models, focusing on denoising diffusion probabilistic models (DDPM) and latent diffusion models (LDM), which underpin our proposed method. We also introduce optimal Bayesian denoising, which is crucial for our joint training.

3.1 Diffusion Models

DDPM is built upon a discrete Markov chain between two processes: forward and reverse processes. The forward process, initiated from a noiseless data x_0, adds Gaussian noise to a previous data x_{t-1} at each timestep t to generate a current data x_t. The reverse process, on the other hand, iteratively samples x_{t-1} from x_t, starting from a Gaussian noise x_T. The forward process is modeled as $q(x_t|x_{t-1}) := \mathcal{N}(x_t; \sqrt{1-\beta_t} x_{t-1}, \beta_t \mathbf{I})$, where β_t is a pre-defined noise scale depending on timestep t. By accumulating the timesteps, we can express the forward to any time t from a data as a reparameterization trick [17]: $q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\alpha_t} x_0, (1-\alpha_t)\mathbf{I})$, where $\alpha_t := \prod_{i=1}^{t}(1-\beta_i)$. The reverse process is modeled as $p_\theta(x_{t-1}|x_t) := \mathcal{N}(x_{t-1}; \mu_\theta(x_t,t), \Sigma_\theta(x_t,t))$. DDPM employs UNet [41] to approximate this posterior. For conditional cases, the UNet ε_θ can be trained using a denoising score matching $\|\varepsilon_\theta(x_t,t,y) - \varepsilon\|_2^2$, where ε is a Gaussian 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 $\varepsilon_\theta(x_t,t,y)$ represents a score function of the perturbed data distribution, $\nabla_{x_t} \log p_\theta(x_t)$. After the training, DDPM can generate images using annealed Langevin dynamics [47].

3.2 Latent Diffusion Model

DMs typically learn denoising in an RGB pixel space, which requires high-computational costs for high-resolution images. In contrast, LDM [40] learns denoising in the latent space of pretrained autoencoders, significantly improving both training and sampling efficiency without compromising quality compared to pixel-based DMs.

LDM consists of an autoencoder with an encoder $\mathcal{E}$ and decoder $\mathcal{D}$ in an RGB space as well as denoising autoencoders in the latent space. Given a data $x \in \mathbb{R}^{H \times W \times 3}$ in an RGB space, the encoder $\mathcal{E}$ encodes x into a latent representation $z = \mathcal{E}(x)$, and the decoder $\mathcal{D}$ reconstructs the data from the latent $z \in \mathbb{R}^{h \times w \times c}$, denoted as $\tilde{x} = \mathcal{D}(z) = \mathcal{D}(\mathcal{E}(x))$, where h and w are downsampled latent size. LDM learns the denoising autoencoder ε_θ in the latent space using DDPM.

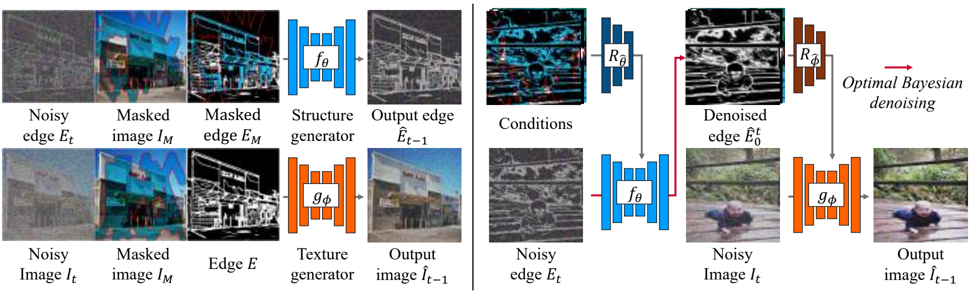

Figure 2: Overview of our individual training (left) and joint training (right).

3.3 Optimal Bayesian Denoising

Optimal Bayesian denoising is a technique that performs minimum mean square error (MMSE) denoising in a single step. Given a Gaussian noise $\varepsilon \sim \mathcal{N}(\varepsilon; \mu, \Sigma)$, its MMSE estimator is given by Tweedie's formula [6, 13, 49]; that is, $\mathbb{E}[\mu|\varepsilon] = \varepsilon + \Sigma\nabla_\varepsilon \log p(\varepsilon)$. In DDPM, the forward step is modeled as $q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\alpha_t}x_0, (1-\alpha_t)\mathbf{I})$ as described in Section 3.1. Therefore, we can apply Tweedie's formula by substituting $\sqrt{\alpha_t}x_0$ and $(1-\alpha_t)\mathbf{I}$ for μ and Σ, respectively. This allows us to determine a single-step denoising operation as

$$F(x_t) := \hat{x}_0^t = \frac{x_t + (1-\alpha_t)\nabla_{x_t} \log p(x_t)}{\sqrt{\alpha_t}}, \tag{1}$$

where $\hat{x}_0^t$ represents a denoised sample. This operation enables us to convert the noisy sample (at time t) into a denoised one (at time 0) in a single step, provided that the optimal score function $\nabla_{x_t} \log p(x_t)$ is known. To reconstruct a denoised sample in an RGB space, the operation is written as $\mathcal{D}(F(x_t))$; however, we will omit the decoder $\mathcal{D}$ for simplicity. We note that previous studies have used the single-step denoising technique for sampling [46] or a formulation of DMs [16] while we aim to use it for a component of training.

4 Structure-Guided Diffusion Model

Given an input image with missing regions (*i.e.* holes), our goal is to generate a structurally reasonable image that respects the context of the visible regions. We denote the target image by $I \in \mathbb{R}^{H\times W\times 3}$, the binary mask representing the missing regions by $M \in \{0,1\}^{H\times W\times 1}$, and the generated image by $\hat{I} \in \mathbb{R}^{H\times W\times 3}$, where H and W represent a spatial resolution. With this notation, the objective is to generate $\hat{I}$ from $I_M = I \odot M$. Our proposed SGDM utilizes structural guidance during the generation process. Specifically, it generates a hole-filled edge image $\hat{E}$ and then uses it as structural guidance to generate the output image $\hat{I}$. The edge image $\hat{E}$ is generated using an edge image with missing regions, denoted by E_M, which is produced from I_M using an existing edge detection algorithm [37].

4.1 Framework Architecture

Our framework consists of two DM-based networks: a structure generator f_θ and a texture generator g_ϕ, where θ and ϕ are learnable parameters. In particular, we employ LDM [40] to reduce computational cost in high-resolution images as described in Section 3.2. We train a condition encoder R to encode input conditions with five channels for each network, including a mask, masked image, masked edge image, and hole-filled edge image, using [65]. The encoded condition is incorporated into the generator.

The framework procedure for hole-filling follows these steps: 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 creates plausible textures with the guidance of $\hat{E}$ while maintaining the context of the visible regions of the input image. These generations use the iterative sampling of DMs as described in Section 3.1. Compared to GAN and AR methods, the SGDM using DMs has two advantages: 1) input data with noise have no holes and 2) iterative generation. In other words, the SGDM generates an output by recursively denoising a noisy input without holes, whereas GAN and AR methods predict the output from a masked input that lacks most contextual information.

4.2 Individual Training

We describe the data preparation and training procedure. Suppose we have a ground-truth image I from a training dataset. Then, we extract an edge image E using the edge detection algorithm [37]. We degrade the image I and the edge image E using a binary mask M, which is randomly drawn for each sample, denoted as the masked image $I_M = I \odot M$ and the masked edge image $E_M = E \odot M$, respectively. We fill the masked region with zeros. To train DMs, we create a noisy image $I_t \in \mathbb{R}^{h \times w \times 3}$ and a noisy edge image $E_t \in \mathbb{R}^{h \times w \times 3}$ in a latent space at timestep t (out of T timesteps) using Gaussian noises ε_I and encoded image ε_E with edge map $\mathcal{E}(I)$ and $\mathcal{E}(E)$, respectively, following the reparameterization trick in Section 3.1.

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

$$f_\theta(E_t, R_{\hat{\theta}}(I_M, M, E_M), t) = \hat{E}_{t-1}, \;\; g_\phi(I_t, R_{\hat{\phi}}(I_M, M, E), t) = \hat{I}_{t-1}, \tag{2}$$

where $\hat{\theta}$ and $\hat{\phi}$ are parameters of condition encoders R for the structure f_θ and texture generator g_ϕ, respectively. These networks can be trained via the denoising score matching [10] in a closed form as described in Section 3.1,

$$\mathcal{L}_f = \mathbb{E}_{I,M,E,t,\varepsilon_E}\left[\|f_\theta(E_t, R_{\hat{\theta}}(I_M, M, E_M), t) - \varepsilon_E\|^2\right], \tag{3}$$

$$\mathcal{L}_g = \mathbb{E}_{I,M,E,t,\varepsilon_I}\left[\|g_\phi(I_t, R_{\hat{\phi}}(I_M, M, E), t) - \varepsilon_I\|^2\right], \tag{4}$$

where the noises ε_E and ε_I are sampled from Gaussian distribution to create E_t and I_t.

4.3 Joint Training

The individually trained structure generator sometimes generates unreasonable edges. This is because edge images are sparse compared to textured images, making the modeling more difficult. To mitigate this issue and improve generalization, we propose joint fine-tuning of both networks in an end-to-end manner after the individual training. For this purpose, we propose a novel joint-training strategy using optimal Bayesian denoising, as shown in Fig. 2.

Method	Modeling	Places (512 × 512)						CelebA-HQ (512 × 512)					
		Small mask			Large mask			Small mask			Large mask		
		FID↓	P-IDS↑	U-IDS↑	FID↓	P-IDS↑	U-IDS↑	FID↓	P-IDS↑	U-IDS↑	FID↓	P-IDS↑	U-IDS↑
SGDM (ours)	DM	3.85	25.54	38.53	6.96	18.12	31.78	2.58	22.01	33.56	4.72	13.99	24.97
Stable Diffusion [1]	DM	5.36	16.32	32.05	7.21	15.34	30.80	-	-	-	-	-	-
LDM [40]	DM	5.64	13.42	30.66	8.74	11.7	27.00	-	-	-	-	-	-
MAT [19]	GAN	4.10	25.56	37.73	7.11	18.40	32.46	2.81	19.24	31.33	5.04	11.42	24.13
MISF [21]	GAN	14.52	3.58	16.23	18.05	2.8	13.19	9.92	4.27	14.53	21.04	0.43	1.88
CoordFill [28]	GAN	6.32	10.4	27.74	14.52	3.58	16.23	4.27	7.35	20.13	10.54	1.57	5.52
ZITS [5]	GAN	4.25	19.56	34.56	8.24	10.84	25.74	-	-	-	-	-	-
MAE-FAR [4]	GAN	4.06	22.18	36.82	7.71	15.06	28.74	-	-	-	-	-	-
LaMa [50]	GAN	4.09	22.18	36.58	8.00	13.54	27.47	4.06	8.55	21.34	8.59	2.17	7.41
CoModGAN [69]	GAN	4.87	22.44	35.99	8.73	15.60	30.10	-	-	-	-	-	-
PUT [27]	AR	7.73	2.68	18.35	15.17	2.54	12.89	-	-	-	-	-	-

Table 1: Quantitative comparisons on Places [73] and CelebA-HQ [14]. The best and second best results are in red and blue. Stable Diffusion inpainting model is trained on LAION-Aesthetics V2 5+ [1].

The joint training of DMs cannot be performed in a straightforward manner, such as in the training of GANs [5, 34]. This is because the texture generator requires a *noiseless* edge image as input, but the structure generator cannot generate a noiseless edge image without iterative sampling. Even if we produce a noiseless edge image via iterative sampling, backpropagation becomes intractable due to gradient accumulation and computational costs. We tackle this issue by applying the single-step denoising operation in Eq. (1); that is, we obtain a noiseless estimate by $\hat{E}_0^t = F(E_t)$. This approach allows us to perform backpropagation in an end-to-end manner. Finally, we formulate our total loss for the joint training as

$$\mathcal{L}_{jt} = \mathcal{L}_f + \mathcal{L}_{g\_o} + \mathcal{L}_{g\_d}, \tag{5}$$

where $\mathcal{L}_{g\_o}$ and $\mathcal{L}_{g\_d}$ are both calculated by Eq. (4), but with different edge images. $\mathcal{L}_{g\_d}$ uses edge images generated by the structure generator (Eq. (2)) and Tweedie's formula (Eq. (1)), encouraging the structure generator to learn texture-aware edge prediction (see Fig. 4). $\mathcal{L}_{g\_o}$ uses original (ground truth) edge images, regularizing the texture generator and preventing overfitting to edge images generated by the structure generator and Tweedie's formula.

5 Experiments

Datasets. The experiments were conducted with Places [73] and CelebA-HQ [14], which cover different degrees of context (natural scenes only vs. face). The image resolution was 512×512 for all experiments. For Places, we prepared a train set and a test set with 8 million (M) and 5,000 images. The test set was created from the official test set for our evaluation. For CelebA-HQ, we prepared a train set and a test set with 24,183 and 2,993 images, respectively. For a better understanding of the performances for holes with various sizes, we prepared two different masks (*i.e.* large and small masks) following MAT [19].

Evaluation metrics. Following [19, 69], we used FID [9], P-IDS, and U-IDS [69] to measure a perceptual fidelity between ground truth and hole-filled images for evaluation. P-IDS and U-IDS robustly assess perceptual fidelity and correlate well with human preferences [69]. Similarity-based metrics such as PSNR and SSIM fail to measure completion, thus, we did not use these metrics.

Places		Large mask	
Model	Samples	FID ↓	LPIPS ↓
(a) Indiv. only	25M	32.28	0.188
(b) + Joint w/o OBd	0.1M	28.68	0.175
(c) + Joint w OBd	0.1M	27.47	0.170
	1M	27.81	0.168

Table 2: Result of the ablation study on Places with large masks. Indiv. and Joint indicate individual and joint training settings, respectively. OBd represents optimal Bayesian denoising.

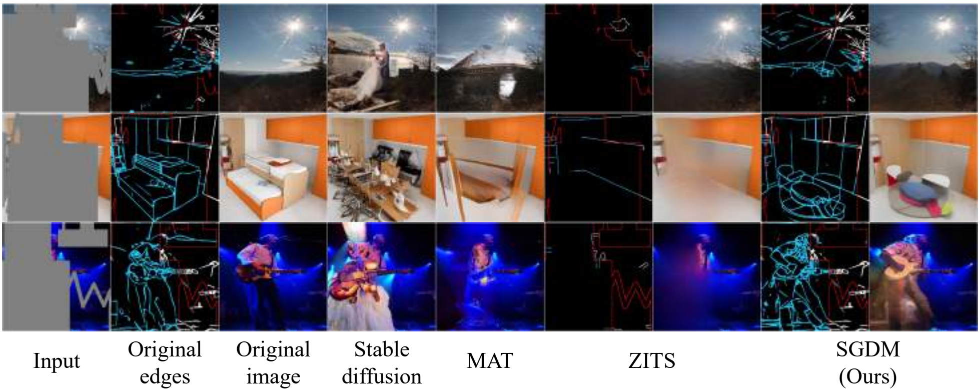

Figure 3: Qualitative comparison between existing methods and the SGDM on Places.

Implementation details. Before the training, we initialized our generators' weights with stable-diffusion-2-1-base [2], which was trained using LAION-5B dataset [44]. We did not use any prompt inputs. For the individual training, each network was trained for 25M images on Places and CelebA-HQ. Additionally, we carried out the joint training with 1M images. The batch size was fixed to 1. Both trainings were performed with AdamW optimizer [29] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ and a learning rate of 10^{-5}. We conducted all experiments with four NVIDIA A100 GPUs. To generate images, we used RePaint [30] sampler.

5.1 Comparison with State-of-the-Art Methods

Quantitative comparisons. We provide the quantitative performance with different masked regions on Places and CelebA-HQ, respectively. Only ZITS [5] used an edge map as structure guidance. Table 1 shows the SGDM achieved the best performance in all metrics under both small and large masks on CelebA-HQ. However, on Places, the SGDM yielded the best FID, but demonstrated P-IDS and U-IDS comparable to MAT.

Qualitative comparisons. Figure 3 shows the qualitative comparison of the competing methods. We see that the proposed SGDM was able to produce rational edges and coherent textures in large holes. Stable diffusion (SD) and MAT produced content without blurring but generated messy results, especially for the region of the human and tables. We observed that SD often generated persons unrelated to the context of the input. ZITS failed to generate reasonable edges for large holes and the generated textures contain a lot of blur. All these results demonstrate that the SGDM was superior to the current state-of-the-art methods.

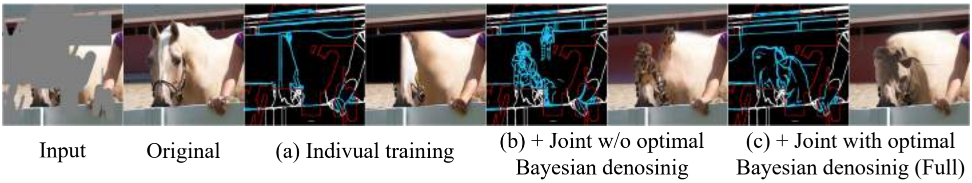
Input
Original
(a) Indivual training
(b) + Joint w/o optimal Bayesian denosinig
(c) + Joint with optimal Bayesian denosinig (Full)

Figure 4: Visual comparison among the models of the ablation study using the same seed.

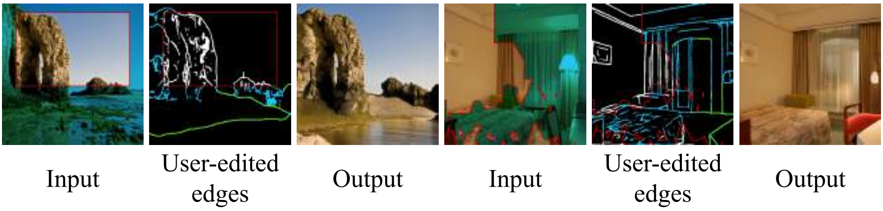

Figure 5: Image completion results using user-edited edges directly (without refinement). The input images and user-edited edge images are the same as in Fig. 1.

5.2 Ablation Study

We conducted an ablation study with large masks using FID and LPIPS [66] to evaluate the image quality. We created a subset with 1,000 images from the test set on Places. We compared SGDMs with different settings: (a) only individual training with 25M training images, (b) joint training after the individual training with 0.1M images without optimal Bayesian denoising, and (c) joint training after individual training with optimal Bayesian denoising using 0.1M and 1M images as our full model. Table 2 shows the result. Comparing models (b) and (c), we see that the joint training with optimal Bayesian denoising improved the metrics. Figure 4 visually compares the settings. Model (a) could not generate reasonable edges and textures. Model (b) could generate edges of the animal's head but failed to synthesize visually coherent textures. In contrast, model (c) could generate realistic content. We conjecture that the texture generator in model (b) did not learn contextual correspondences between edges and textures well because it was conditioned by noisy edges in training. Joint training with optimal Bayesian denoising effectively improved the image quality.

5.3 Applications

Sketch-guided image completion. Figure 1 shows image completion results using user-edited edge images, highlighting SGDM's potential of use as a user-guided image editing tool. First, the edge images generated by our structure generator (the third column) were manually edited by the user with sketching tools (the fifth column). Then, they were further refined using the prior of the structure generator by SDEdit [31] (the sixth column); specifically, we first perturbed the user-edited edge images with Gaussian noise at timestep 500 and 200 out of 1,000, respectively, and then progressively removed the noise via the reverse process. This process, called SDEdit, can refine (potentially) unrealistic user-edited edges, making them more compatible with the texture generator. We found this approach to be sufficiently robust and well-suited for interactive editing. Note that our SGDM can also generate plausible images directly from raw edges manually drawn by users, *i.e.* without the refinement by SDEdit, as demonstrated in Fig. 5.

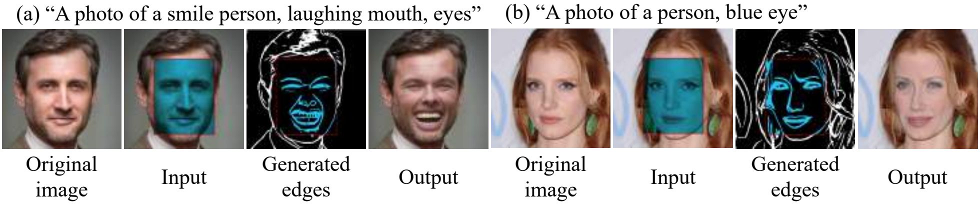
(a) “A photo of a smile person, laughing mouth, eyes”
(b) “A photo of a person, blue eye”
Original image
Input
Generated edges
Output
Original image
Input
Generated edges
Output

Figure 6: Language-guided image completion for (a) structure and (b) texture modifications.

Language-guided image completion. Figure 6 demonstrates a language-guided image completion application where user-provided text prompts successfully control structure and texture generation. Note, however, that SGDM may not work well with more complex prompts because SGDM has not been explicitly trained using prompt input.

6 Limitation and Discussion

Failure cases. Our method sometimes struggled to generate structurally rational edges. For example, in the second row of Fig. 1, the auto-generated edges were not very rational, especially in the bottom right region. Nonetheless, users can manually correct such irrational edges if necessary; this flexibility is one of the SGDM's strengths.

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 PUT also require iterative inference, but their computational cost is lower than ours. To investigate the computational cost issue, we measured the time needed to complete a center-masked image for each method. As a result, MAT (GAN-based) needed 0.098 seconds, and PUT (AR-based) required 4.06 seconds, respectively. Our method with the RePaint sampler, on the other hand, took 133 seconds.

Potential societal impacts. Our method inherits the potential societal impact of previous image completion methods (*e.g.* [30]). Generated images may reflect the biases in the datasets, such as gender, age, ethnicity, *etc*. Moreover, the image-editing capability of our method could aid in DeepFake creation [32, 54]. On the other hand, image completion may enhance privacy protection by removing identifiable information from public-space photographs.

Future work. Even though our SGDM was not explicitly trained using prompts, our SGDM could reasonably perform language-guided image completion for simple prompts, as shown in Fig. 6. The original ControlNet [65], on the other hand, presents a training method that explicitly uses prompt input to achieve more sophisticated language guidance. Exploring such training methods that utilize prompt input is a promising direction for future research.

7 Conclusion

We have presented a structure-guided diffusion model (SGDM), which uses structural guidance in image completion. We have proposed a novel training strategy to enable effective end-to-end training. Extensive experiments show that the SGDM achieves a superior or comparable visual quality on both Places and CelebA-HQ as compared to state-of-the-art methods. Incorporating structural guidance has not only improved the visual quality but also enabled user-guided image editing.

Acknowledgement

This research was partly supported by JSPS KAKENHI 22KJ1014, JSPS KAKENHI 21H03460, and D-CORE Grant from Microsoft Research Asia. We would like to thank Seiji Kurokoshi and Atsushi Honda for the helpful discussions.

References

[1] runwayml/stable-diffusion-inpainting. https://huggingface.co/runwayml/stable-diffusion-inpainting, accessed 29 August, 2023.

[2] stabilityai/stable-diffusion-2-1-base. https://huggingface.co/stabilityai/stable-diffusion-2-1-base, Last accessed 29 August, 2023.

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

[4] Chenjie Cao, Qiaole Dong, and Yanwei Fu. Learning Prior Feature and Attention Enhanced Image Inpainting. In *ECCV*, 2022.

[5] Qiaole Dong, Chenjie Cao, and Yanwei Fu. Incremental Transformer Structure Enhanced Image Inpainting with Masking Positional Encoding. In *CVPR*, 2022.

[6] Bradley Efron. Tweedie's Formula and Selection Bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011.

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

[8] 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.

[9] 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.

[10] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. In *NeurIPS*, 2020.

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

[12] Youngjoo Jo and Jongyoul Park. SC-FEGAN: Face Editing Generative Adversarial Network With User's Sketch and Color. In *ICCV*, 2019.

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

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

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

[16] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the Design Space of Diffusion-Based Generative Models. In *NeurIPS*, 2022.

[17] Diederik P Kingma and Max Welling. Auto-Encoding Variational Bayes. In *ICLR*, 2014.

[18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *NeurIPS*, 2012.

[19] 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.

[20] Wenbo Li, Xin Yu, Kun Zhou, Yibing Song, Zhe Lin, and Jiaya Jia. Image Inpainting via Iteratively Decoupled Probabilistic Modeling. *arXiv preprint arXiv:2212.02963*, 2023.

[21] 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 *CVPR*, 2022.

[22] 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.

[23] 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.

[24] 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.

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

[26] Hongyu Liu, Ziyu Wan, Wei Huang, Yibing Song, Xintong Han, and Jing Liao. PD-GAN: Probabilistic Diverse GAN for Image Inpainting. In *CVPR*, 2021.

[27] Qiankun Liu, Zhentao Tan, Dongdong Chen, Qi Chu, Xiyang Dai, Yinpeng Chen, Mengchen Liu, Lu Yuan, and Nenghai Yu. Reduce Information Loss in Transformers for Pluralistic Image Inpainting. In *CVPR*, 2022.

[28] Weihuang Liu, Xiaodong Cun, Chi-Man Pun, Menghan Xia, Yong Zhang, and Jue Wang. CoordFill: Efficient High-Resolution Image Inpainting via Parameterized Coordinate Querying. In *AAAI*, 2023.

[29] Ilya Loshchilov and Frank Hutter. Fixing Weight Decay Regularization in Adam. In *ICLR*, 2019.

[30] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. RePaint: Inpainting Using Denoising Diffusion Probabilistic Models. In *CVPR*, 2022.

[31] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations. In *ICLR*, 2022.

[32] Yisroel Mirsky and Wenke Lee. The Creation and Detection of Deepfakes: A Survey. *ACM Comput. Surv.*, 54(1), 2021.

[33] Ishan Misra, Abhinav Shrivastava, Abhinav Gupta, and Martial Hebert. Cross-stitch Networks for Multi-task Learning. In *CVPR*, 2016.

[34] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Qureshi, and Mehran Ebrahimi. EdgeConnect: Structure Guided Image Inpainting using Edge Prediction. In *ICCVW*, 2019.

[35] Evangelos Ntavelis, Andrés Romero, Iason Kastanis, Luc Van Gool, and Radu Timofte. SESAME: Semantic Editing of Scenes by Adding, Manipulating or Erasing Objects. In *ECCV*, 2020.

[36] Jialun Peng, Dong Liu, Songcen Xu, and Houqiang Li. Generating Diverse Structure for Image Inpainting With Hierarchical VQ-VAE. In *CVPR*, 2021.

[37] Xavier Soria Poma, Ángel D. Sappa, Patricio Humanante, and Arash Akbarinia. Dense extreme inception network for edge detection. *Pattern Recognition*, 139:109461, 2023.

[38] Jiayan Qiu, Yiding Yang, Xinchao Wang, and Dacheng Tao. Hallucinating Visual Instances in Total Absentia. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *ECCV*, 2020.

[39] Yurui Ren, Xiaoming Yu, Ruonan Zhang, Thomas H. Li, Shan Liu, and Ge Li. StructureFlow: Image Inpainting via Structure-aware Appearance Flow. In *ICCV*, 2019.

[40] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. In *CVPR*, 2022.

[41] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *MICCAI*, 2015.

[42] Sebastian Ruder. An Overview of Multi-Task Learning in Deep Neural Networks. *arXiv preprint arXiv:1706.05098*, 2017.

[43] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-Image Diffusion Models. In *SIGGRAPH*, 2022.

[44] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: An open large-scale dataset for training next generation image-text models. In *NeurIPS*, 2022.

[45] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. In *ICML*, 2015.

[46] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising Diffusion Implicit Models. In *ICLR*, 2021.

[47] Yang Song and Stefano Ermon. Generative Modeling by Estimating Gradients of the Data Distribution. In *NeurIPS*, 2019.

[48] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-Based Generative Modeling through Stochastic Differential Equations. In *ICLR*, 2021.

[49] Charles Stein. Estimation of the Mean of a Multivariate Normal Distribution. *Annals of Statistics*, 9(6):1135–1151, 1981.

[50] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust Large Mask Inpainting with Fourier Convolutions. In *WACV*, 2022.

[51] Aäron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel Recurrent Neural Networks. In *ICML*, 2016.

[52] Aaron van den Oord, Oriol Vinyals, and koray kavukcuoglu. Neural Discrete Representation Learning. In *NeurIPS*, 2017.

[53] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In *NeurIPS*, 2017.

[54] Luisa Verdoliva. Media Forensics and DeepFakes: An Overview. *IEEE Journal of Selected Topics in Signal Processing*, 14(5):910–932, 2020.

[55] Ziyu Wan, Jingbo Zhang, Dongdong Chen, and Jing Liao. High-Fidelity Pluralistic Image Completion with Transformers. In *ICCV*, 2021.

[56] Cairong Wang, Yiming Zhu, and Chun Yuan. Diverse Image Inpainting with Normalizing Flow. In *ECCV*, 2022.

[57] 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 *ICCV*, 2019.

[58] Wei Xiong, Jiahui Yu, Zhe Lin, Jimei Yang, Xin Lu, Connelly Barnes, and Jiebo Luo. Foreground-Aware Image Inpainting. In *CVPR*, 2019.

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

[60] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S. Huang. Free-Form Image Inpainting With Gated Convolution. In *ICCV*, 2019.

[61] 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 *ACM MM*, 2021.

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

[63] 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 *ECCV*, 2020.

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

[65] Lvmin Zhang and Maneesh Agrawala. Adding Conditional Control to Text-to-Image Diffusion Models. *arXiv preprint arXiv:2302.05543*, 2023.

[66] 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.

[67] 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.

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

[69] 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.

[70] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. Pluralistic image completion. In *CVPR*, 2019.

[71] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. Pluralistic Free-Form Image Completion. *IJCV*, 129(10):2786–2805, 2021.

[72] Haitian Zheng, Zhe Lin, Jingwan Lu, Scott Cohen, Eli Shechtman, Connelly Barnes, Jianming Zhang, Ning Xu, Sohrab Amirghodsi, and Jiebo Luo. CM-GAN: Image Inpainting with Cascaded Modulation GAN and Object-Aware Training. In *ECCV*, 2022.

[73] 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.

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

Appendix

A Mask Statistics

We followed MAT [19] and DeepFill v2 [60] to create masks. Figure A show statistics of large and small masks of Places [73] and CelebA-HQ [14] dataset for the quantitative evaluation. 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 512 × 512 images from 136 to 394 pixels.

B Additional Qualitative Comparisons

We present more qualitative comparisons of the proposed SGDM with comparison methods. Figure B show other variations of Fig. 1. Figure C and Fig. D show the qualitative comparisons using Places. Figure E provides the qualitative comparisons using CelebA-HQ.

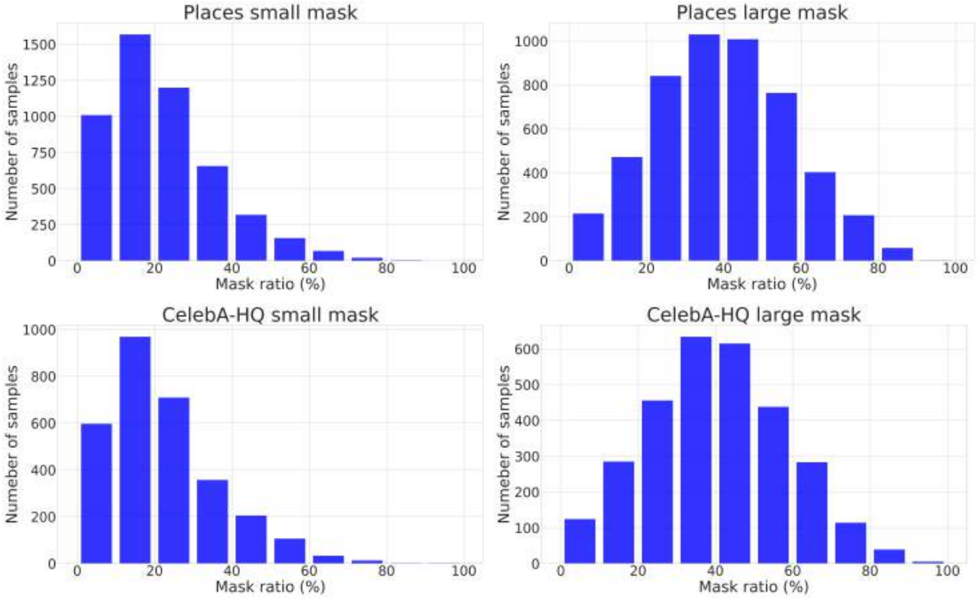

Figure A: Statistics of large and small masks of Places and CelebA-HQ dataset for the quantitative evaluation.

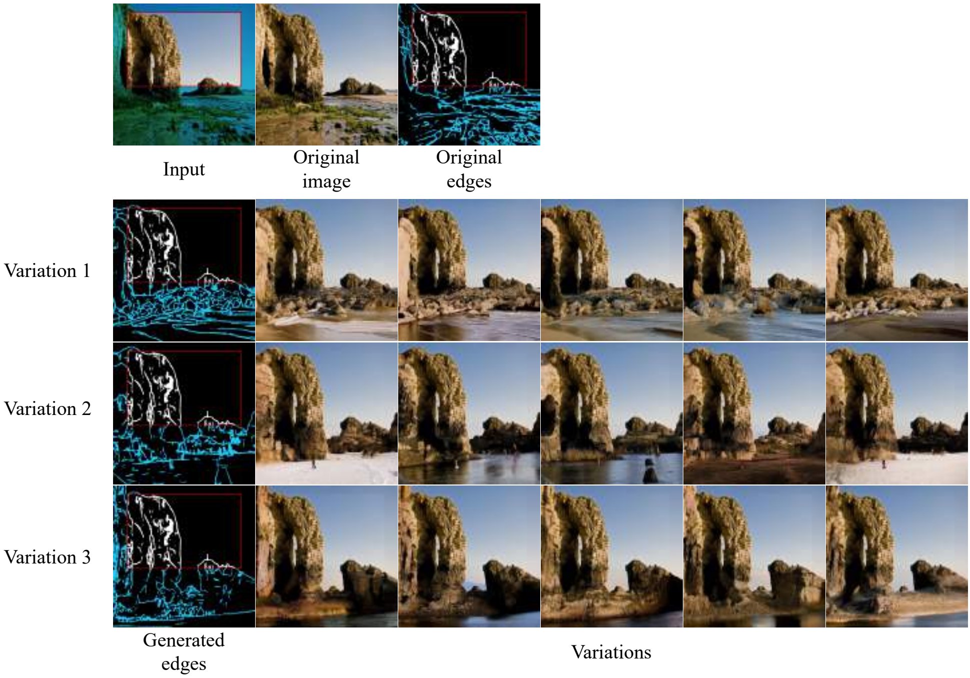

Figure B: Qualitative comparisons of the generated different outputs by the SGDM.

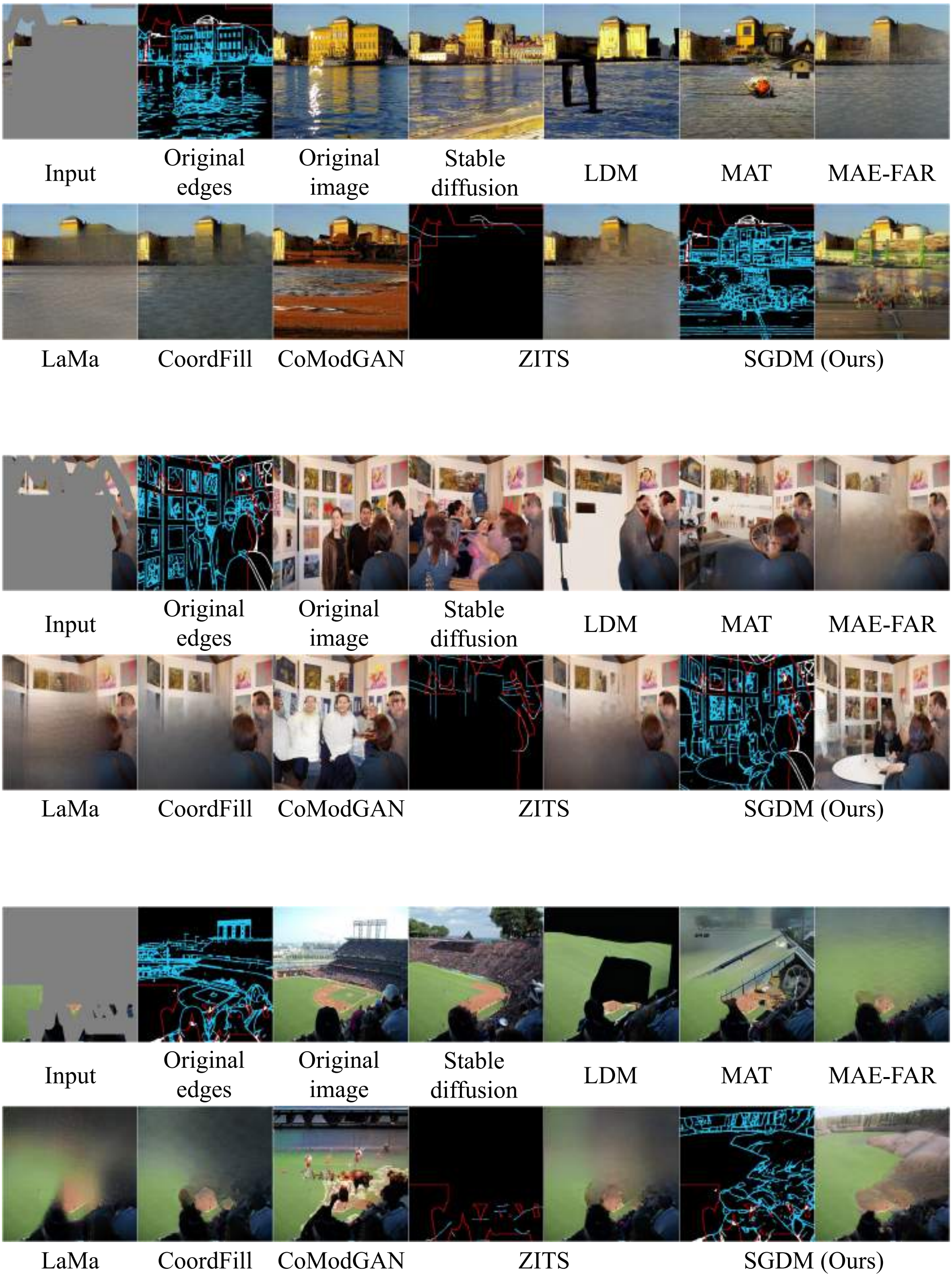

Figure C: Qualitative comparisons of the proposed SGDM with the comparison methods using Places.

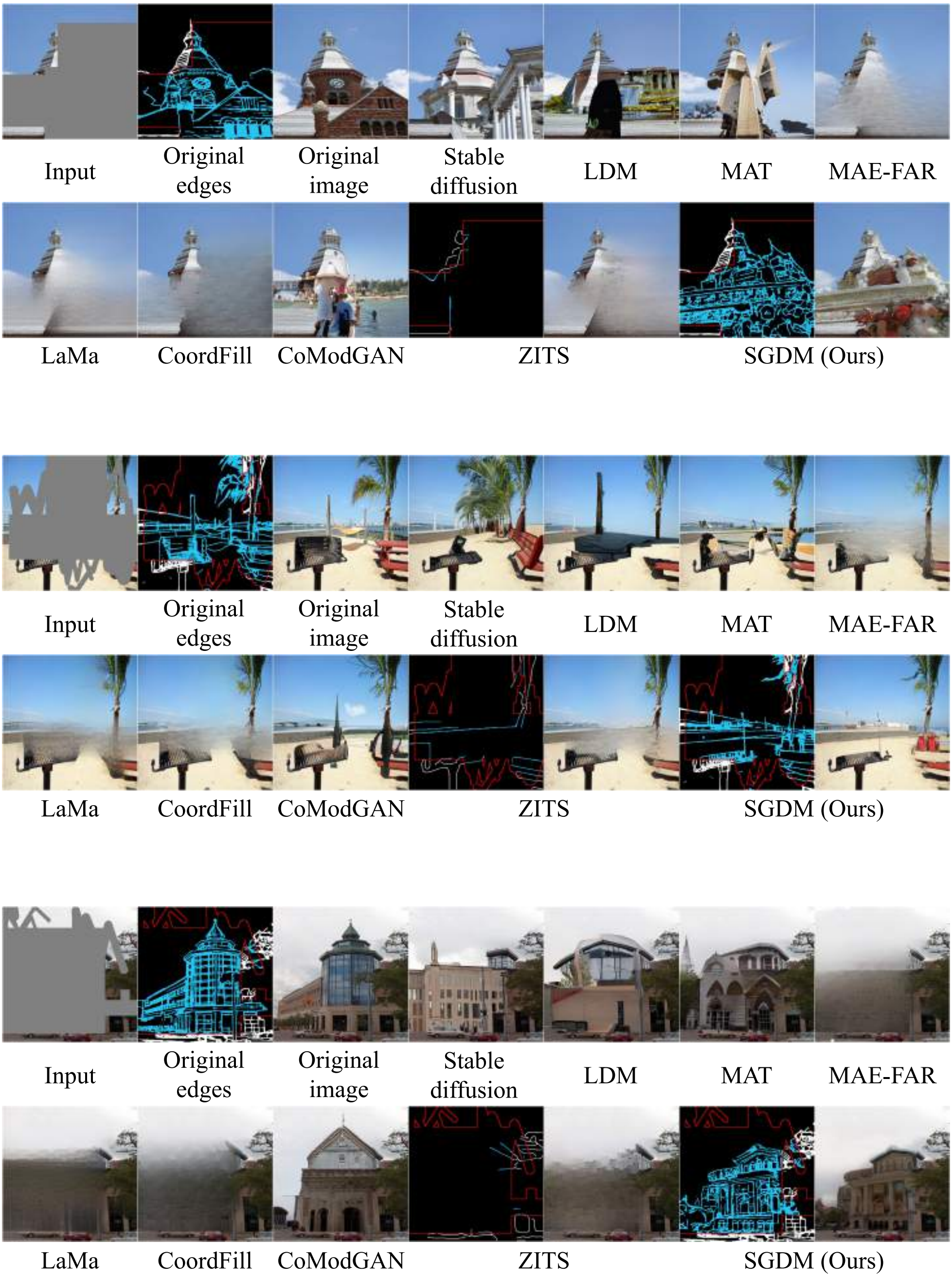

Figure D: Qualitative comparisons of the proposed SGDM with the comparison methods using Places.

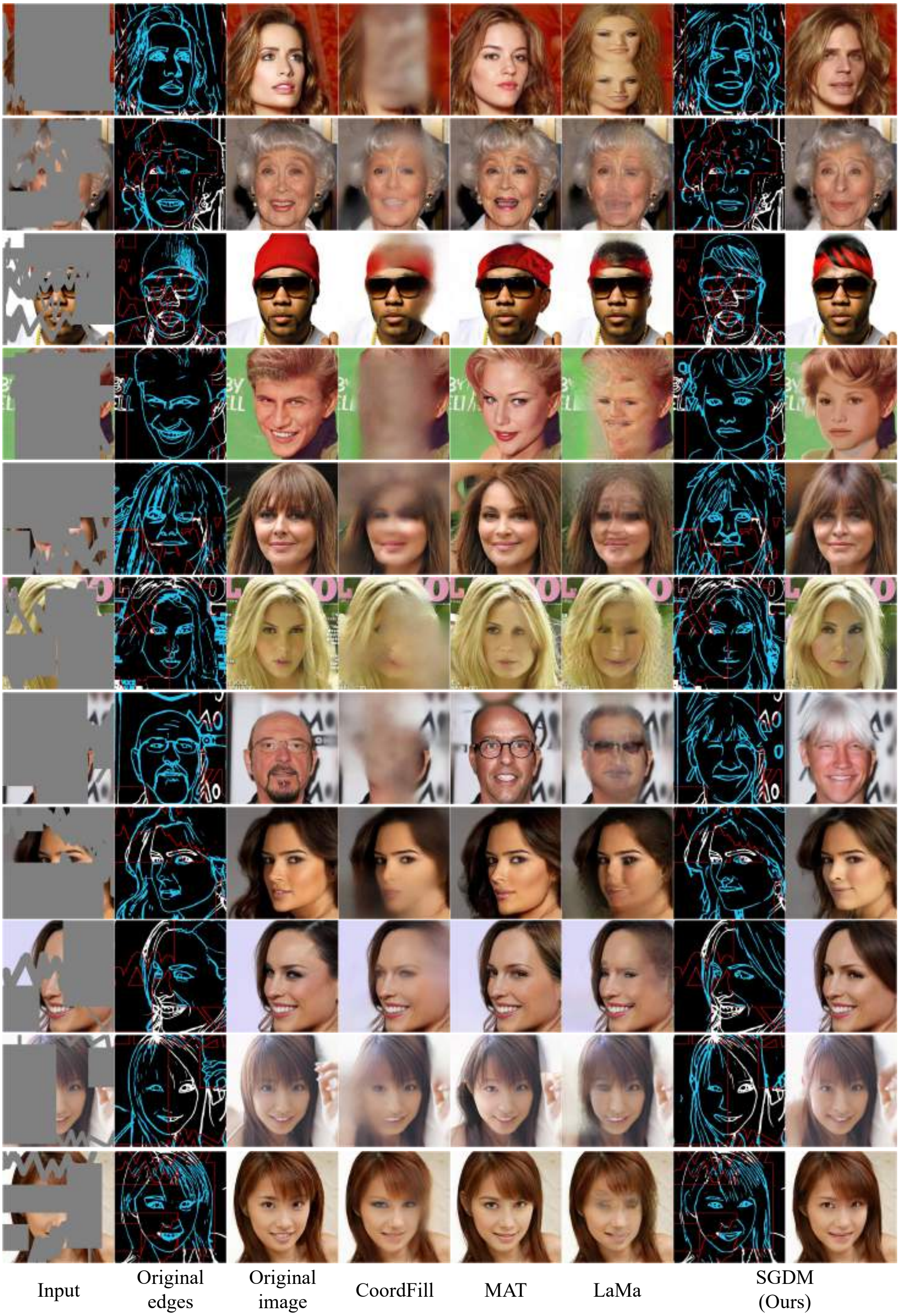

Figure E: Qualitative comparisons of the proposed SGDM with the comparison methods using CelebA-HQ.