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

Daichi Horita[1] Jiaolong Yang[2] Dong Chen[2] Yuki Koyama[3] Kiyoharu Aizawa[1]
[1]The University of Tokyo [2]Microsoft Research Asia
[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

Abstract

*Diverse image completion, a problem of generating various ways of filling incomplete regions (*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 is a task to fill missing regions (*i.e.* holes) of the target image. Humans possess the creative ability to guess the content of the missing regions in various rational ways. Therefore, image completion methods should ideally produce plausible yet diverse results while maintaining consistency with the visible regions.

How can we fill in holes in images? Bertalmio *et al*. [1] describe how expert conservators restore damaged artworks as 1) figure out what content to put in missing regions, 2) draw contour edges, and 3) paint the regions guided by the contours. Related to the second step, previous work [4, 22, 35] has introduced explicit structure guidance. An additional benefit of using the guidance is to provide a user-guided image editing [13, 37] (see Fig. 1). However, it is still challenging to estimate reasonable structures for large holes.

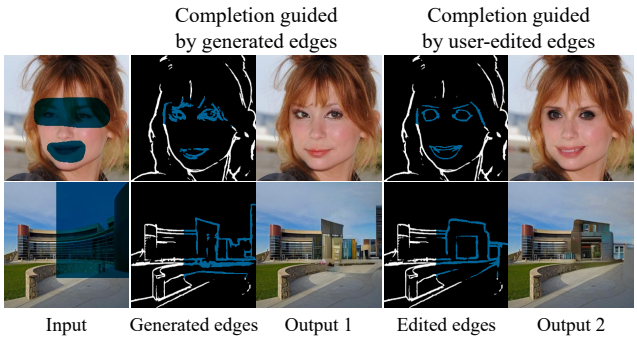

Figure 1. **We propose a structure-guided diffusion model (SGDM).** The SGDM first generates an edge image within missing regions, indicated by blue regions. Then, the SGDM generates images using the edge image as the structural guidance.

In this paper, we focus on diffusion probabilistic models (DMs) and explore incorporating structural guidance. We propose a *structure-guided diffusion model* (SGDM), which 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. To achieve it, we propose using *optimal Bayesian denoising*, in particular, Tweedie's formula [5, 14, 29], 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 [9, 31]. 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). This framework allows us to use generated denoised edges and ground-truth edges simultaneously. In our experiments with datasets of faces (CelebA-HQ [15]) and natural

scenes (Places [40]), we show that our method achieves a comparable or superior trade-off between visual quality and diversity compared to state-of-the-art methods.

Our contributions are summarized as follows. (1) We propose the structure-guided diffusion probabilistic model (SGDM) for large-hole diverse image completion. (2) We design a novel joint-training strategy using optimal Bayesian denoising and a momentum framework to enable end-to-end training of the two generators. (3) We show that the SGDM achieves a comparable or superior trade-off performance between visual quality and diversity using CelebA-HQ [15] and Places [40] datasets.

2. Related Work

2.1. Deterministic Image Completion

Deep-learning-based methods using GANs [8] have demonstrated tremendous success in image completion [12, 24]. Several works utilized explicit clues such as object edges [4, 22], foreground contours [34, 37], reference images [41], and semantic segmentation maps [19]. Nazeri *et al*. [22] first proposed a two-stage framework for edges and textures to introduce structure guidance. However, both GAN- [35] and transformer-based [4] methods often failed to produce valid edge maps, while the SGDM can generate natural results as shown in Fig. 4.

2.2. Large-Hole 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 [18, 38, 39]. Zheng *et al*. [39] first demonstrated the diverse image completion task. CoModGAN [38] and MAT [18] achieved a high-fidelity quality by introducing stochastic style representation [16,17], although their diversity was restricted due to the conditional training procedure. BAT-Fill [36] and PUT [20] have focused on an AR transformer [2, 32]. Their approaches have limitations on sampling orders [6], computational costs [7], and the lack of introducing explicit structural information.

2.3. Image Completion with Diffusion Models

Previous studies [26, 28] 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 them is to produce non-semantically consistent results. To solve it, Palette [25] learned DMs as a conditional image synthesis. RePaint [21] proposed a conditional sampling method, which alternately performs the forward and reverse diffusion processes. 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. Preliminary: Optimal Bayesian Denoising

Here we describe optimal Bayesian denoising, which we use to enable our joint training. For more details of DMs, please refer to [11, 23]. 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 $z \sim \mathcal{N}(z; \boldsymbol{\mu}, \boldsymbol{\Sigma})$, an MMSE estimator is given by Tweedie's formula [5, 14, 29]; that is, $\mathbb{E}[\boldsymbol{\mu}|z] = z + \boldsymbol{\Sigma}\nabla_z \log p(z)$. In DDPM, the forward step is modeled as $q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\alpha_t}x_0, (1-\alpha_t)\mathbf{I})$. Thus, we can apply Tweedie's formula here by substituting $\sqrt{\alpha_t}x_0$ and $(1-\alpha_t)\mathbf{I}$ for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, respectively. This allows us to determine a single-step denoising operation as

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

where x_0^t represents a denoised sample. We can convert the noisy sample 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. DDIM [27] also relies on the formulation.

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) [33].

4.1. Framework Architecture

Our framework consists of two DM-based networks: a structure generator f_θ and a texture generator g_ϕ. The structure generator aims to generate an edge image that guides the texture generator. We attach five additional channels in the first layer of both networks 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 $\hat{E}$ maintaining the context of the visible regions of I. These generations use the iterative sampling of DMs.

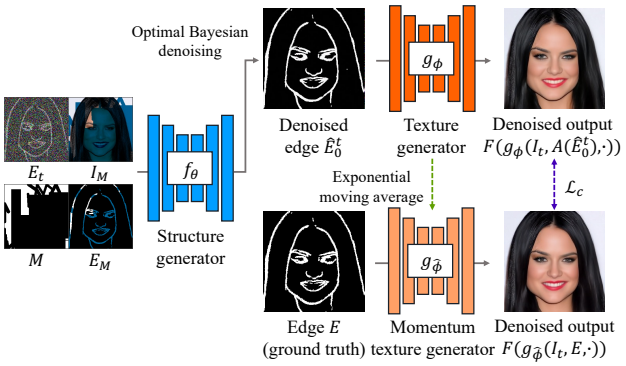

Figure 2. **Overview of our joint training** for training a structure f_θ, a texture g_ϕ, and a momentum texture generator $g_{\hat{\phi}}$. The weights of the momentum texture generator $g_{\hat{\phi}}$ is updated as an exponential moving average of the weights of texture generator g_ϕ.

4.2. Individual Training

We describe the data preparation and the training procedure. Suppose we have a ground-truth image I. Then, we extract an edge image E using HED. We degrade the image I and the edge image E using a random binary mask M, denoted as the masked image $I_M = I \odot M$ and the masked edge image $E_M = E \odot M$, respectively. We create a noisy image I_t and a noisy edge image E_t at timestep t using Gaussian noises ϵ_I and ϵ_E with I and E, respectively. Given the above inputs, the generators predict outputs as follows:

$$f_\theta(E_t, I_M, M, E_M, t) = \hat{E}_{t-1}, \tag{2}$$

$$g_\phi(I_t, I_M, M, E, t) = \hat{I}_{t-1}. \tag{3}$$

Both networks can be trained via the denoising score matching loss [11] individually,

$$\mathcal{L}_f = \mathbb{E}_{I,M,E,t,\epsilon_E} \|f_\theta(E_t, I_M, M, E_M, t) - \epsilon_E\|_2^2, \tag{4}$$

$$\mathcal{L}_g = \mathbb{E}_{I,M,E,t,\epsilon_I} \|g_\phi(I_t, I_M, M, E, t) - \epsilon_I\|_2^2. \tag{5}$$

4.3. Joint Training

The individually trained structure generator sometimes generates semantically unreasonable edges. To mitigate this issue, we propose a joint-training strategy using optimal Bayesian denoising and a consistency loss, as shown in Fig. 2. The joint training is cannot be performed in a straightforward fashion due to time information. We apply the single-step denoising operation in Eq. (1); that is, we obtain a noiseless estimate by $\hat{E}_0^t = F(E_t)$.

However, the denoised edge image $\hat{E}_0^t$ tends to be overly blurred especially when t is close to T. We observe that the gap between the original edge image E and the denoised edge image $\hat{E}_0^t$ leads to poor visual quality. To overcome these challenges, we use the momentum framework for the texture generator. We have observed that the training also becomes unstable without the two differences and the EMA-based weight updating. To avoid corruption, we introduce $\hat{E}_0^t$ to an augmentation A that erases the region randomly. The consistency loss can be formulated as

$$\mathcal{L}_c = \|F(g_\phi(I_t, A(\hat{E}_0^t),\cdot)) - F(g_{\hat{\phi}}(I_{\hat{t}}, E,\cdot))\|_2^2, \tag{6}$$

where $\cdot$ denotes the same condition as in Eq. (3). Finally, we formulate our total loss for the joint training as

$$\mathcal{L}_{jt} = \mathcal{L}_f + \mathcal{L}_g + \mathcal{L}_c. \tag{7}$$

CelebA-HQ		Large hole		Small hole	
Method	Modeling	FID ↓	Div ↑	FID ↓	Div ↑
MAT [18]	GAN	**3.63**	0.024	**1.97**	0.014
BAT-Fill [36]	AR	7.52	0.043	5.24	0.025
PUT [20]	AR	10.37	0.039	3.96	**0.039**
RePaint [21]	DM	8.06	**0.056**	4.71	**0.034**
Palette [25]	DM	**6.59**	**0.056**	**2.88**	0.032
Ours	DM	**5.71**	**0.057**	**2.66**	**0.038**

Table 1. **Quantitative comparison on CelebA-HQ.** **RBG** shows good performance in this order.

Places		Large hole		Small hole	
Method	Modeling	FID ↓	Div ↑	FID ↓	Div ↑
MAT [18]	GAN	**8.15**	0.044	**4.39**	0.032
BAT-Fill [36]	AR	21.16	0.079	8.33	0.051
PUT [20]	AR	17.17	0.092	7.37	0.063
RePaint [21]	DM	**10.71**	**0.113**	**5.36**	**0.079**
Palette [25]	DM	26.61	**0.102**	10.95	**0.074**
Ours	DM	**12.30**	**0.095**	**5.95**	**0.064**

Table 2. **Quantitative comparison on Places.**

5. Experiments

We conducted experiments to compare our method with other state-of-the-art methods in terms of the trade-off between visual quality and diversity.

Datasets. The experiments were conducted with CelebA-HQ [15] and Places [40] The image resolution of both datasets was 256×256. For CelebA-HQ, we prepared a train set and a test set with 24,183 and 5,000 images, respectively. For Places, we prepared a train set and a test set with 8 million (M) and 5,000 images. To evaluate the diversity, we randomly selected 50 images from each test set. 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. For the evaluation, we generated edge images and images using 1,000 sampling steps and 4,570 sampling steps using RePaint [21], respectively.

Compared methods. We compared our method with the following methods: MAT [18], BAT-Fill [36], PUT [20], RePaint [21], ADM [3], and Palette [25] using their pre-trained weights. RePaint, ADM, and ours used the same pre-trained weights. The difference between RePaint and ADM was the sampling procedure.

Evaluation metrics. We considered two different aspects: visual quality and diversity. For the quality evaluation, we used FID [10] using 5,000 test images. For the diversity evaluation, we defined a diversity score for a given set of generated images, $\mathcal{X} = \{x_i\}_{i=1}^{N}$, as

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

where CosSim represents a cosine similarity between two image feature vectors and we set $N = 100$. We extracted image features by InceptionV3 [30] and applied the global average pooling to obtain the feature vector, as represented by $\Phi(\cdot)$. A higher score indicates that the generated set $\mathcal{X}$ is more diverse. We then calculated the mean of the diversity scores. The diversity score is not a common metric, but it can measure how well a large hole can be filled.

5.1. Quantitative Comparisons

Tables 1 and 2 show the completion performance on CelebA-HQ and Places, respectively. MAT achieved the best visual qualities although the lowest diversity. Compared with RePaint, our method showed the advantage on CelebA-HQ. However, on Places, RePaint achieved superior or comparable results to our method. This implies that there was a trade-off between visual quality and diversity. 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. 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 3 shows a qualitative comparison of the competing methods. We observed that our method was able to synthesize texture variations while maintaining the structural context. This implies that the texture generator could learn the semantic context from the partially visible region as well as from the edges. RePaint and Palette, which use diffusion models, generated plausible images, but the semantic consistency was insufficient.

We also show comparisons between our method and existing GAN-based methods with structural guidance, DeepFillv2 [35] and ZITS [4], in Fig. 4. 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.

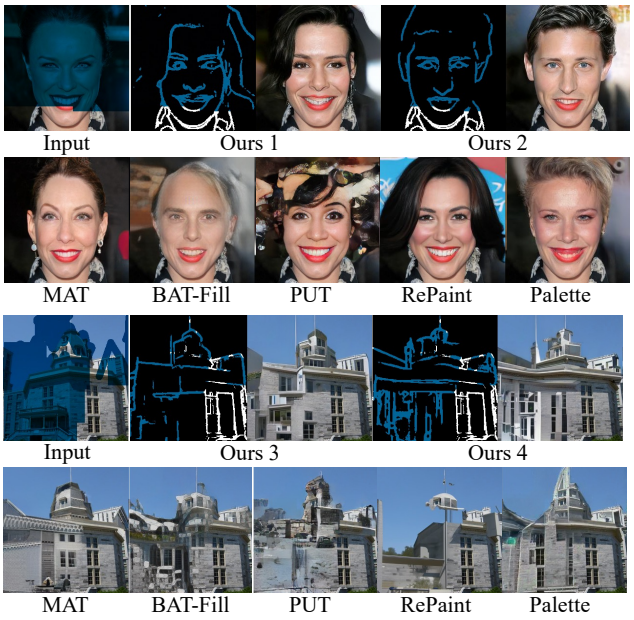

Figure 3. **Qualitative comparisons of the proposed SGDM with the state-of-the-art methods.**

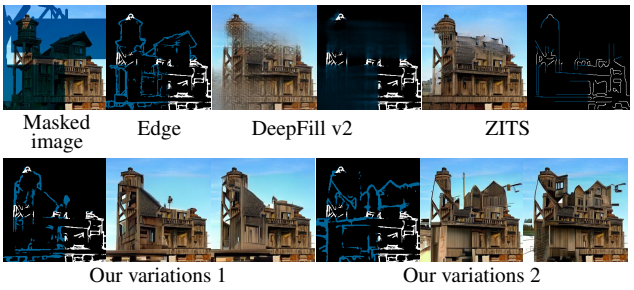

Figure 4. **Visual comparison among image completion methods with structural guidance [4, 35].**

6. 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.

References

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

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

[3] Prafulla Dhariwal and Alex Nichol. Diffusion Models Beat GANs on Image Synthesis. In *NeurIPS*, 2021. 4

[4] Qiaole Dong, Chenjie Cao, and Yanwei Fu. Incremental Transformer Structure Enhanced Image Inpainting with Masking Positional Encoding. In *CVPR*, 2022. 1, 2, 4

[5] Bradley Efron. Tweedie's Formula and Selection Bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011. 1, 2

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

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

[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. 2

[9] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila 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. 1

[10] 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. 4

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. In *NeurIPS*, 2020. 2, 3

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

[13] Youngjoo Jo and Jongyoul Park. SC-FEGAN: Face Editing Generative Adversarial Network With User's Sketch and Color. In *ICCV*, 2019. 1

[14] Miyasawa K. An empirical Bayes estimator of the mean of a normal population. *Bull. Inst. Internat. Statist.*, 38:181–188, 1961. 1, 2

[15] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In *ICLR*, 2018. 1, 2, 3

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

[17] 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

[18] 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, 3, 4

[19] 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

[20] 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. 2, 3, 4

[21] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. RePaint: Inpainting Using Denoising Diffusion Probabilistic Models. In *CVPR*, 2022. 2, 3, 4

[22] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Qureshi, and Mehran Ebrahimi. EdgeConnect: Structure Guided Image Inpainting using Edge Prediction. In *ICCVW*, 2019. 1, 2

[23] Alex Nichol and Prafulla Dhariwal. Improved Denoising Diffusion Probabilistic Models. In *ICML*, 2021. 2

[24] Deepak Pathak, Philipp Krähenbühl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. Context Encoders: Feature Learning by Inpainting. In *CVPR*, 2016. 2

[25] 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. 2, 3, 4

[26] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. In *ICML*, 2015. 2

[27] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising Diffusion Implicit Models. In *ICLR*, 2021. 2

[28] 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. 2

[29] Charles Stein. Estimation of the Mean of a Multivariate Normal Distribution. *Annals of Statistics*, 9(6):1135–1151, 1981. 1, 2

[30] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the Inception Architecture for Computer Vision. In *CVPR*, 2016. 4

[31] Antti Tarvainen and Harri Valpola. Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*, 2017. 1

[32] 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. 2

[33] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *ICCV*, 2015. 2

[34] Wei Xiong, Jiahui Yu, Zhe Lin, Jimei Yang, Xin Lu, Connelly Barnes, and Jiebo Luo. Foreground-Aware Image Inpainting. In *CVPR*, 2019. 2

[35] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S. Huang. Free-Form Image Inpainting With Gated Convolution. In *ICCV*, 2019. 1, 2, 4

[36] 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. 2, 3, 4

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

[38] 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

[39] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. Pluralistic image completion. In *CVPR*, 2019. 2

[40] 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, 3

[41] 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