Diffusion-based Semantic Image Synthesis from Sparse Layouts*

Yuantian Huang[1], Satoshi Iizuka[1], and Kazuhiro Fukui[1]

University of Tsukuba, Ibaraki 305-8577, Japan
{huang_yuantian@cvlab., iizuka@, kfukui@}cs.tsukuba.ac.jp

Abstract. We present an efficient framework that utilizes diffusion models to generate landscape images from sparse semantic layouts. Previous approaches use dense semantic label maps to generate images, where the quality of the results is highly dependent on the accuracy of the input semantic layouts. However, it is not trivial to create detailed and accurate semantic layouts in practice. To address this challenge, we carefully design a random masking process that effectively simulates real user input during the model training phase, making it more practical for real-world applications. Our framework leverages the Semantic Diffusion Model (SDM) as a generator to create full landscape images from sparse label maps, which are created randomly during the random masking process. Missing semantic information is complemented based on the learned image structure. Furthermore, we achieve comparable inference speed to GAN-based models through a model distillation process while preserving the generation quality. After training with the well-designed random masking process, our proposed framework is able to generate high-quality landscape images with sparse and intuitive inputs, which is useful for practical applications. Experiments show that our proposed method outperforms existing approaches both quantitatively and qualitatively.

Keywords: Semantic Image Synthesis · Sparse Input · Diffusion Models.

1 Introduction

Semantic image synthesis refers to a subfield of image synthesis, which aims to generate new images conditioned on semantic layouts that contains required features and structural information of the image. It can be utilized for a variety of applications, including image editing, content creation, and artificial intelligence art. The earliest attempt at this task goes back at least to image analogies [8], where they automatically learned filters from training data based on a simplistic multi-scale autoregression.

* This study was supported by the Japan Science and Technology Agency Support for Pioneering Research Initiated by the Next Generation (JST SPRING); Grant Number JPMJSP2124.

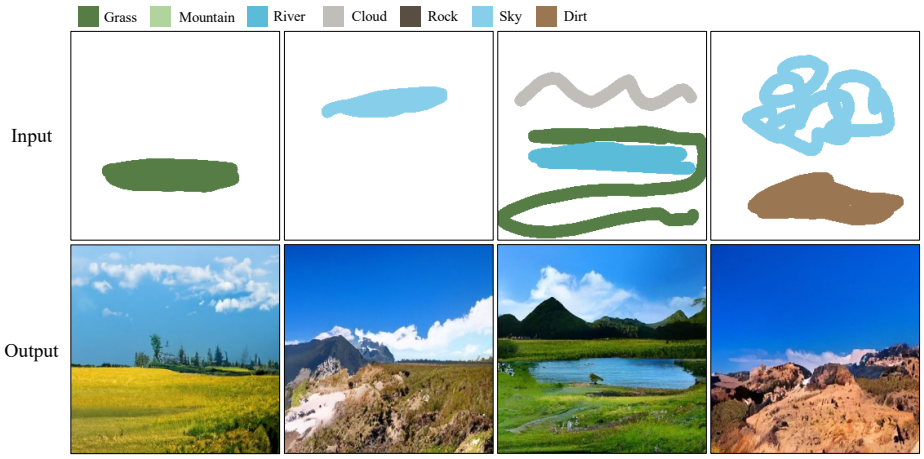

Fig. 1. Our approach can synthesize landscape images from more sparse and intuitive semantic layouts rather than detailed and precise layouts while maintaining the same level of generation quality.

With the recent success of generative adversarial networks (GANs [7]), data-driven generation approaches have become a dominant trend. Several GAN-based methods have achieved good results by directly learning image-to-image mapping, such as pix2pixGAN, which first established a common framework for learning the mapping of paired data and successfully generated realistic images. Furthermore, Park *et al.* [16] proposed SPatially-ADaptive DEnormalization (SPADE), which outperforms previous methods on the task of semantic image synthesis. Subsequent models, including SEAN [32], SMIS [33], and OASIS [28], extend the SPADE model in different aspects. Recently, Wang *et al.* [30] developed a semantic image synthesis approach based on Denoising Diffusion Probabilistic Models (DDPMs) [11]. This method integrates SPADE into the residual blocks of the decoder networks to better leverage the information in the input semantic mask, leading to improvements in the overall quality of the generated images.

Existing approaches for semantic image synthesis have focused on processing a comprehensive semantic layout that captures the entire scene, which is equivalent to the reverse task of semantic segmentation. However, in real-world applications that rely on human-authored layouts as input, creating detailed and precise layouts that accurately represent real-world scenes can be quite challenging. To address this issue, we carefully design a random masking process that enables the simulation to mimic actual user input, thereby making the inputs better suited for use in real-world settings.

Our generator is constructed based on the Semantic Diffusion Model (SDM) [30]. It feeds noisy images to the encoder of the U-Net structure, similar to other diffusion-based image generation methods. Additionally, the semantic layouts are injected into the SPADE [16] layers of the decoder. By progressively refining

the generated results, it achieves state-of-the-art performance on the task of semantic image synthesis. Moreover, we find that the methodology of diffusion is most appropriate for our masking design and achieves the best results for more sparse and intuitive inputs. We also accomplished comparable inference speed to GAN-based models through a model distillation process while preserving generation quality.

After training with the well-designed random masking process, our proposed framework can generate high-quality landscape images with sparse and intuitive inputs, which enables the practical application of semantic image synthesis in scenarios that require human-authored layouts. We evaluate our approach quantitatively with both automatic metrics and a perceptual user study, in addition to qualitative results. Results demonstrate that our approach outperforms existing approaches in all metrics.

This study presents the following contributions:

- A well-designed masking strategy that simulates human-authored sparse layouts, improving the practical application of semantic image synthesis.
- A diffusion-based generator with a model distillation process that enables fast sampling while preserving better generation quality compared to GAN-based models.
- In-depth evaluation of our method with both qualitative and quantitative comparisons with existing approaches.

2 Related Work

2.1 Semantic Image Synthesis

Semantic image synthesis focuses on generating new images conditioned on semantic image maps and is mainly dominated by GAN-based [7] approaches in recent years. For example, the pix2pixGAN [12] established a common framework for learning the mapping of paired data and successfully generated realistic images. This approach has been extended in the following research [3, 17, 29]. Furthermore, Park *et al.* [16] proposed the SPatially-ADaptive DEnormalization (SPADE) approach, which outperforms previous methods in generating photo-realistic images conditioned on semantic layouts. Various improvements have been made to the SPADE architecture in different aspects, such as SMIS [33], which produces semantically multimodal images by replacing all regular convolution layers in the generator with group convolutions, and SEAN [32], which uses style input images to create spatially varying normalization parameters per semantic region. Additionally, OASIS [28] surpasses SPADE in diversity while maintaining similar quality by redesigning the discriminator. Besides GAN-based models, Chen *et al.* [2] proposed the Cascaded Refinement Network (CRN) for high-resolution semantic image synthesis. It has been extended by subsequent methods [14, 18]. However, these approaches still underperform compared to state-of-the-art GAN-based models. Recently, with the advent of diffusion models, SDM [30] has incorporated SPADE with Denoising Diffusion Probabilistic

Models (DDPMs), enabling better generation fidelity and diversity simultaneously. We construct our generator based on SDM, which outperforms GAN-based methods in terms of generation quality and is most appropriate for our masking design.

2.2 Sparse and Intuitive User Input on Image Synthesis

In real-world computer graphics applications, users often prefer to provide intuitive inputs rather than comprehensive ones, as producing the latter can be highly challenging. Intuitive inputs may include text, sketches, scene graphs, and semantic layouts for image synthesis tasks. For instance, text-to-image synthesis models [19–21, 31] generate photorealistic images from text descriptions, while other works focus on generating images from edges and sketches [5, 6, 12] and scene graphs [1, 13]. These inputs are easy to create but lack precise control and struggle to produce high-quality results due to the difficulty of training. Among the many types of inputs, we believe that semantic layouts offer the most control and interactivity while providing plausible results, as the semantic labels provide precise shape and content. However, creating a semantic layout that perfectly matches the inputs to the training set is always challenging. In contrast to previous methods that require users to reproduce detailed semantic label maps, in this paper, we propose simulating human-authored semantic layouts during training using a well-designed random masking strategy. Notably, our proposed method still provides precise control if the input semantic layout is dense enough, unlike sketches-based models and scene graphs-based models that only accept sparse inputs, as discussed in Section 3.1.

2.3 Diffusion Probabilistic Models

Diffusion Models (DMs) can be defined as Markov chains trained using variational inference that gradually transition from random noise to the target distribution through a series of diffusion steps [26]. Ho et al. [11] proposed Denoising Diffusion Probabilistic Models (DDPMs), establishing an explicit connection between diffusion models and denoising score matching, which leads to improved image generation quality. Subsequently, many studies [23, 27] have explored the potential of DMs in various aspects, such as unconditional image synthesis [4], image-to-image translation [22, 25], and text-to-image generation [19, 21]. These studies demonstrate the superiority of DMs over GAN-based models in terms of both quality and stability. Recently, Wang *et al.* [30] developed a semantic image synthesis approach that integrates SPADE [16] into the residual blocks of the decoder networks to better leverage the information in the input semantic mask, leading to improvements in overall quality on the task of semantic image synthesis. In this paper, we utilize a network structure similar to the SDM model and modify it to accommodate the requirements of progressive distillation [24]. This modification results in a significant boost in inference speed, making the approach more practical for real-world applications.

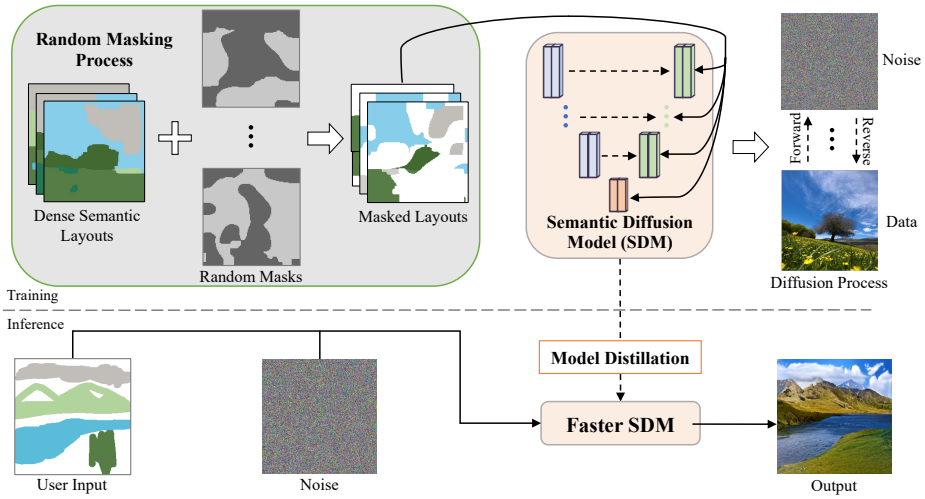

Fig. 2. Our proposed framework. The semantic layout is processed by our random masking process to simulate the actual user input and then passed to our generator, which learns to predict a landscape image from random noise through the diffusion process. During the inference phase, after the model distillation process, a distilled generator can quickly sample an output landscape corresponding to the sparser user input.

3 Proposed Framework

Our proposed framework consists of three components: a) a random masking process that simulates actual user input, improving generation quality in practical applications; b) a diffusion-based generator that we found to be most suitable for our masking design while also surpassing previous GAN-based models in both fidelity and diversity; and c) a progressive model distillation process that significantly reduces diffusion steps during the inference stage, making our framework interactive and broadly applicable.

During training, a detailed semantic layout is processed by our random masking process for each iteration to simulate user input. The masked layout is then passed to our generator, which learns to predict a landscape image from random noise through the diffusion process while conditioning on the masked layout. In the inference phase, a distilled generator can quickly sample an output landscape image from random noise, corresponding to the sparser input semantic layouts provided by the users. An overview of the model is depicted in Figure 2.

3.1 Masking Strategy

We try to simulate user input in different strategies as follows:

- **Random Blocks** generate multiple coordinates, with random width and height.

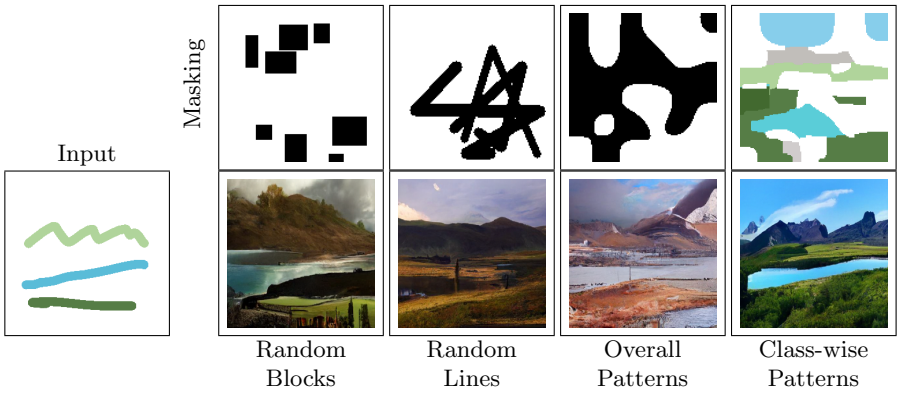

Fig. 3. Comparisons of different strategies of random masking.

- **Random Lines** generate multiple paired coordinates as the start and end points of a stroke line, then generate random middle points. Finally, draw lines with random stroke width.
- **Random Patterns** generate a low-resolution random pattern to meet an average percentage of binary masks. For example, we set the average masking percentage of our generated patterns from 15% $\sim$ 75%, which is also randomized in every iteration.
- **Class-wise Random Patterns** are similar to the previous one but generate a different mask for each class at the same percentage to avoid disregarding classes that are smaller than the average size, such as trees and rocks. These smaller-sized classes are more likely to be completed masked if we use an overall mask, which may lead to biased learning. Examples of different random masking strategies are shown in Fig. 3.

Based on our experiments, we find that class-wise random patterns achieve the most appropriate simulation for actual user inputs. Thus, we adopt this masking strategy in the following experiments for our proposed model and baseline models. Additionally, to improve performance when the input semantic is dense, we set a random 15% of the input semantic to be left intact, enhancing the stability of the training and providing more precise control over the generated images. For a fair comparison, we apply the same settings during the training of all baseline models.

3.2 Diffusion-based Generator

We use the same network architecture as Semantic Diffusion Model [30], which builds upon DDPMs [11]. This architecture features a U-Net structure comprising an encoder and a decoder. It is specifically designed for the diffusion process, incorporating attention blocks, skip-connections, and a timestep embedding module. Most importantly, in contrast to previous models that feed both input

Table 1. Evaluation of Model Distillation (MD). We evaluate the models before and after model distillation in different scales. Notably, we substantially increase inference speed to 64 times while maintaining a consistent FID metric.

	Diffusion Steps	Inference Time (s)	FID↓
Original	1024	183.84	38.37
MD×16	64	10.27	39.60
MD×64	**16**	**2.49**	**41.94**

semantic layouts and random noise into the encoder, the SDM injects semantic label maps into the decoder to condition the semantic information. This is achieved by introducing SPADE (SPatially-ADaptive DEnormalization) [16], a highly effective normalization technique for the task of semantic image synthesis, into the decoder part of the networks.

Our generator has been modified in two aspects that have been proven to be more effective for the progressive distillation [24] process, which we conduct after the training. Firstly, we adopt a cosine noise schedule where $\alpha_t = \cos(0.5\pi t)$ and $\sigma_t^2 = 1 - \alpha_t^2$ similar to that introduced by improved DDPMs [15]. Furthermore, our generator is modified to predict $v = \alpha_t \epsilon - \sigma_t x$ instead of a random noise ϵ that is sampled from the standard Gaussian distribution, where $x \sim p(x)$ denotes the input image and $t \in [0, T]$ indicates timestep. We set $T = 1024$ in our proposed framework for better alignment during the distillation.

Training Objective. Given a sample image x, a noisy sample $\widetilde{x}$ is produced as follows:

$$\widetilde{x} = \sqrt{\alpha_t} x + \sqrt{1 - \alpha_t} v \tag{1}$$

Our objective function during the training includes a simple mean-squared error loss L_{simple} to predict v described above, which can be defined as:

$$L_{simple} = \mathbb{E}_{x,v,t} \left[\| v - v_\theta(\sqrt{\alpha_t} x + \sqrt{1 - \alpha_t} v, l, t) \|_2 \right] \tag{2}$$

, where l indicates the input semantic layout.

3.3 Model Distillation

One of the main limitations of diffusion-based models is the extremely slow sampling process, which requires simulating a Markov chain for numerous steps to generate a sample. To address this concern, we apply progressive distillation [24] to our proposed model. We briefly review the concept of progressive distillation, which is quite simple and straightforward. Assume we have a teacher model, which is the model obtained after the main training. We establish a student model initially copied from the teacher model and attempt to learn one student diffusion step to match two teacher diffusion steps. After the first distillation training, we obtain a student model that takes $T/2$ sampling steps to replace

the original teacher model that required T sampling steps. This process can be conducted repeatedly until the final model only requires a few steps to generate a sample of similar quality to the original one.

As mentioned in Sec. 3.2, our generator is modified to meet the necessary requirements discovered by the authors of the progressive distillation, enabling it to accelerate the sampling process significantly while still preserving plausible generation quality. A brief evaluation of our distillation process in terms of generation quality and inference time cost are shown in Table 1. In addition to the distillation loss used in the original progressive distillation paper, we incorporate an additional simple loss, as defined in Eq.2, to evaluate how well the student model matches the initial training. This strategy is widely used in distillation and was first introduced by Hinton *et al.* [10].

4 Results

We evaluate our approach and compare it with other methods both qualitatively and quantitatively.

4.1 Experiments Setup

We adopt this masking strategy in the following experiments for our proposed model and baseline models.

Datasets. We conduct experiments on Flickr Landscapes. We first collect 50,000 landscape photos that include various outdoor scenes from Flickr. We then remove 19,000 samples according to their generated label maps if they have irrelevant labels, such as people or animals. We are left with 31,000 images, of which 1,000 are left as a validation set.

Baselines. We chose three existing models as baselines for semantic image synthesis: the SPADE [16], the SEAN [32] model, and the OASIS [28] model. All of the baseline models are trained with the implementations provided by the authors. In the inference phase, we note that the SEAN model requires a style code as input, which we pre-compute from the mean style codes for all the training data. It is important to note that the same masking process is applied to all baseline models and our proposed model, ensuring a fair comparison.

4.2 Implementation Details

We train all networks from scratch. Specifically, we set the learning rate to 0.0001 in the first 600,000 iterations, which is then reduced to 0.00002 for the subsequent 200,000 iterations. To better facilitate the progressive distillation process, we set the diffusion steps to 1024 instead of the original 1000 steps. We adopt a cosine scheduler instead of a linear one for the same reason. The image size is set to be 256×256 pixels in all training.

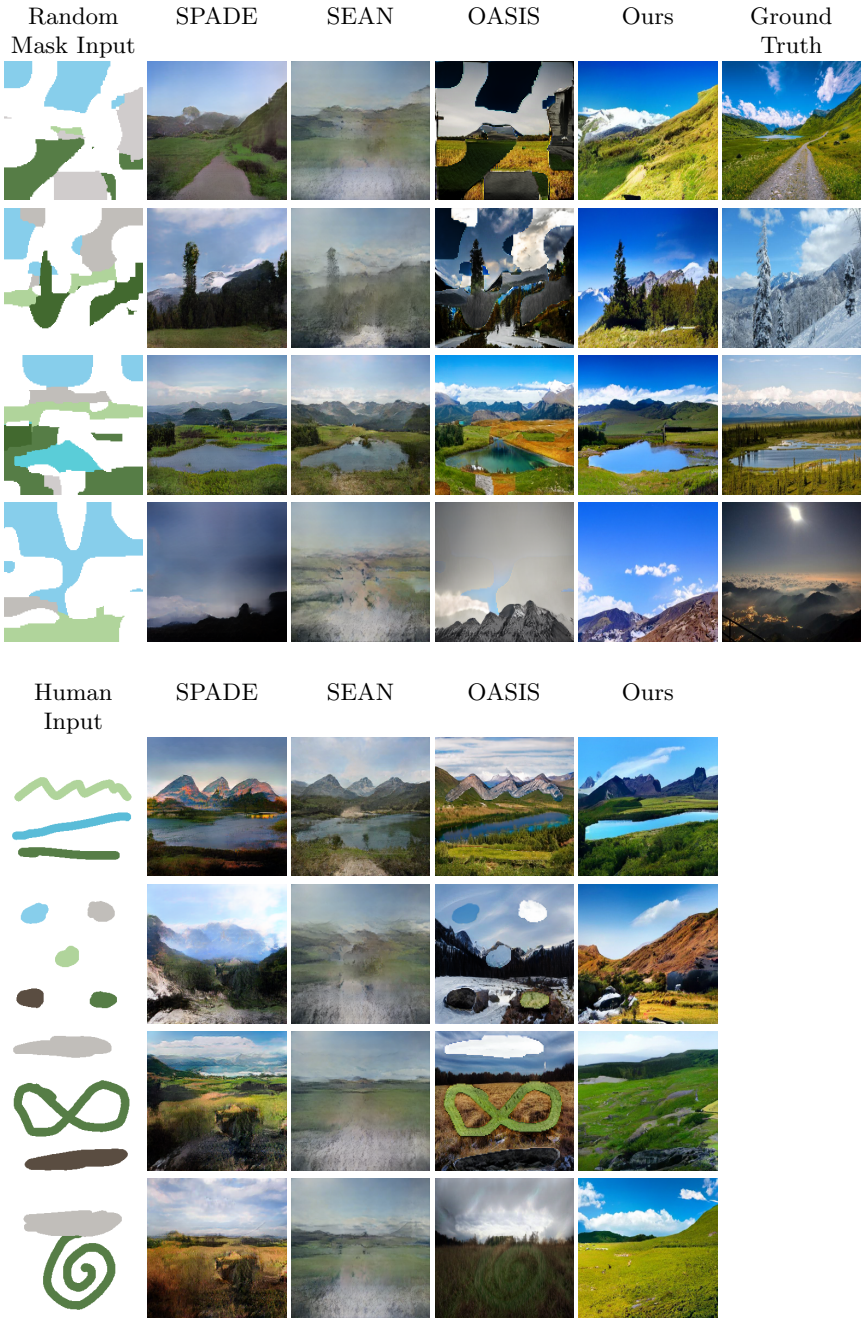

Fig. 4. Qualitative comparison. We compare our approach against existing approaches in both random mask input using our random masking strategy and actual human input.

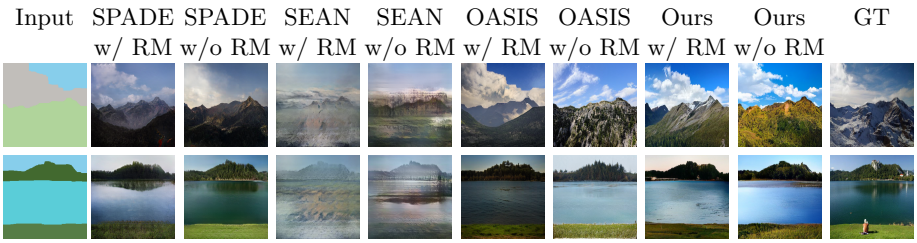

Fig. 5. Examples from ablation study. We compare our approach with existing methods for generating images conditioned on unmasked dense semantic layouts. All models are trained under two different settings: with or without our random masking (RM) process, and ours trained without RM is equivalent to the original SDM model.

Table 2. Quantitative evaluation. We compare with existing approaches for conditional image generation using the FID metric.

	SPADE	SEAN	OASIS	ours
FID↓	57.82	148.32	44.47	**38.37**

4.3 Quantitative Comparison

We use the **Fréchet Inception Distance (FID)** [9] as our primary evaluation metric, which captures the perceptual similarity of generated images with real ones. We calculated FID between the real images and generated images from random masked semantic layouts. We use a total of 4,000 validation semantic layouts as input, which are randomly masked four times based on the validation set described in Sec. 4.1. As shown in Table 2, we can see that our approach outperforms existing approaches in terms of generating quality.

4.4 Visual Quality Comparison

We provide a qualitative comparison of generation results from both random mask input and actual human input, against our baseline models, as shown in Fig. 4. It can be observed that the results of SEAN fail to generate meaningful outcomes, as it generates an average style derived from the training data based on the paired label maps. However, due to the random masking during the training stage, this approach fails to generate satisfactory results. A similar issue is observed for OASIS, which tends to generate obviously different textures in masked areas due to its semantically-aware discriminator. Although SPADE seems unaffected by the random masks, the generation quality is limited. In contrast, our approach automatically completes the missing areas from the input label while preserving a high-quality generation.

Table 3. Perceptual user study results. The numbers indicate the percentage of users that prefer our method with respect to existing approaches.

vs.	SPADE	OASIS	Real
Ours	73.42%	65.22%	24.56%

Table 4. Abalation Study. We compare our approach with existing methods for generating images conditioned on unmasked dense semantic layouts. All models are trained under two different settings: with or without our random masking (RM) process, and ours trained without RM is equivalent to the original SDM model.

Method	trained w/ RM	FID↓	Increase↓
SPADE [16]	✗	38.70	
	✓	55.49	43.39%
SEAN [32]	✗	143.20	
	✓	152.86	6.75%
OASIS [28]	✗	32.31	
	✓	40.46	25.22%
ours	✗	32.67	
	✓	**36.41**	**11.45**%

4.5 Perceptual User Study

We evaluate our method with a perceptual user study conducted with 10 participants. We use all 4,000 generated images from each approach and their paired real images, which are the same settings used in the quantitative comparisons. In each round of the study, two images are shown to the user, both randomly selected from different approaches or real images. Participants are asked to choose which image appears more realistic for a total of 500 rounds per user. As shown in Table 3, our approach is preferred over existing approaches. Furthermore, when compared against real images, our approach is considered better 24.56% of the time, consistent with the quantitative comparison results.

4.6 Ablation Study

We also conduct an ablation study to verify the effectiveness of our masking design, as shown in Table 4 and Fig. 5. All models are trained under two different settings: with or without our random masking process. During the inference phase, only complete semantic layouts are used as input to generate results for evaluation. We are able to evaluate how it affects generation quality by observing the increase in FID when utilizing our random masking strategy. Notably, our proposed model is minimally affected by the masking process, indicating its robustness and effectiveness in handling masked inputs.

(a) Input (b) Output

Fig. 6. Generation conditioned on unnatural inputs. When provided inputs are unnatural, such as clouds beneath mountains on the left or 90 degree rotations, the model might face difficulty in producing results as the input is significantly different from anything seen during training.

4.7 Limitations and Discussion

Although our framework can generate high-quality landscape images from sparse and intuitive semantic label maps, it comes with several constraints due to the data-driven approach. Specifically, the application is limited to known labels, and new data must be acquired to extend the model to new labels, such as animals. Furthermore, the model learns a mapping from realistic semantic maps to artwork images and can fail if the input semantic maps diverge significantly from the training data, as shown in Fig. 6.

5 Conclusion

We have presented a novel framework for generating landscape images from sparse semantic layouts. Our approach consists of a well-designed masking strategy that simulates actual user input, thereby avoiding the challenging task of producing detailed semantic layouts and improving generation quality in real-world applications. We employ a diffusion-based generator tailored to our masking design, which outperforms existing models in terms of both fidelity and diversity Furthermore, an additional model distillation process makes our framework more interactive and applicable for practical use.

References

1. Ashual, O., Wolf, L.: Specifying object attributes and relations in interactive scene generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2019)
2. Chen, Q., Koltun, V.: Photographic image synthesis with cascaded refinement networks. In: Proceedings of the IEEE international conference on computer vision. pp. 1511–1520 (2017)
3. Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J.: Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8789–8797 (2018)
4. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in Neural Information Processing Systems **34**, 8780–8794 (2021)
5. Gao, C., Liu, Q., Xu, Q., Wang, L., Liu, J., Zou, C.: Sketchycoco: Image generation from freehand scene sketches. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5174–5183 (2020)
6. Ghosh, A., Zhang, R., Dokania, P.K., Wang, O., Efros, A.A., Torr, P.H.S., Shechtman, E.: Interactive sketch & fill: Multiclass sketch-to-image translation. In: Proceedings of the IEEE international conference on computer vision (2019)
7. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks (2014)
8. Hertzmann, A.: Can computers create art? Arts **7**(2), 18 (May 2018). https://doi.org/10.3390/arts7020018
9. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Advances in neural information processing systems. pp. 6626–6637 (2017)
10. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
11. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems **33**, 6840–6851 (2020)
12. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1125–1134 (2017)
13. Johnson, J., Gupta, A., Fei-Fei, L.: Image generation from scene graphs. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
14. Li, K., Zhang, T., Malik, J.: Diverse image synthesis from semantic layouts via conditional imle. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4220–4229 (2019)
15. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: International Conference on Machine Learning. pp. 8162–8171. PMLR (2021)
16. Park, T., Liu, M.Y., Wang, T.C., Zhu, J.Y.: Semantic image synthesis with spatially-adaptive normalization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2337–2346 (2019)
17. Qi, X., Chen, Q., Jia, J., Koltun, V.: Semi-parametric image synthesis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8808–8816 (2018)
18. Qi, X., Chen, Q., Jia, J., Koltun, V.: Semi-parametric image synthesis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8808–8816 (2018)

19. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 (2022)
20. Razavi, A., van den Oord, A., Vinyals, O.: Generating diverse high-fidelity images with vq-vae-2. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 32. Curran Associates, Inc. (2019)
21. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10684–10695 (June 2022)
22. Saharia, C., Chan, W., Chang, H., Lee, C., Ho, J., Salimans, T., Fleet, D., Norouzi, M.: Palette: Image-to-image diffusion models. In: ACM SIGGRAPH 2022 Conference Proceedings. pp. 1–10 (2022)
23. Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D.J., Norouzi, M.: Image super-resolution via iterative refinement. IEEE Transactions on Pattern Analysis and Machine Intelligence (2022)
24. Salimans, T., Ho, J.: Progressive distillation for fast sampling of diffusion models. In: International Conference on Learning Representations (2022)
25. Sasaki, H., Willcocks, C.G., Breckon, T.P.: Unit-ddpm: Unpaired image translation with denoising diffusion probabilistic models. arXiv preprint arXiv:2104.05358 (2021)
26. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International Conference on Machine Learning. pp. 2256–2265. PMLR (2015)
27. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456 (2020)
28. Sushko, V., Schönfeld, E., Zhang, D., Gall, J., Schiele, B., Khoreva, A.: You only need adversarial supervision for semantic image synthesis. arXiv preprint arXiv:2012.04781 (2020)
29. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8798–8807 (2018)
30. Wang, W., Bao, J., Zhou, W., Chen, D., Chen, D., Yuan, L., Li, H.: Semantic image synthesis via diffusion models (2022). https://doi.org/10.48550/ARXIV.2207.00050
31. Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., Metaxas, D.N.: Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (Oct 2017)
32. Zhu, P., Abdal, R., Qin, Y., Wonka, P.: Sean: Image synthesis with semantic region-adaptive normalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5104–5113 (2020)
33. Zhu, Z., Xu, Z., You, A., Bai, X.: Semantically multi-modal image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5467–5476 (2020)