Depth and Semantic Aware Image Generation

Pei-Mao Sun*
Department of Electrical Engineering(Program A)
Yuan Ze University
Taoyuan City, Taiwan
jacky_51@kimo.com

Han Fu*
Department of Electrical Engineering(Program A)
Yuan Ze University
Taoyuan City, Taiwan
s1080661@mail.yzu.edu.tw

Yu-Hui Huang
Department of Electrical Engineering(Program A)
Yuan Ze University
Taoyuan City, Taiwan
yhhuang@saturn.yzu.edu.tw

Abstract

Conditional image generation has been widely studied in computer vision. In this paper, we investigate the possibility of integrating depth information into this task. By generating depth maps as an intermediate step, our proposed model successfully produces realistic images on ADE20K [2] in better quality than the baseline model.

I. Introduction

Generative adversarial networks (GANs) [3] play an important role in image generation tasks. It has wide applications such as different view synthesis [13] and image inpainting [14]. Moreover, it can be applied as a form of data augmentation to improve the training of recognition tasks, such as synthesizing images under different weather conditions for driving scenes to help visual understanding tasks such as image semantic segmentation [15].

In this paper, we focus on a specific type of image synthesis task called semantic image synthesis. It is a subclass of image-to-image translation where a realistic image is generated from a semantic segmentation mask. This type of practice has a wide range of applications, including assisting interior designers in designing interior layouts. Users can render a realistic view directly from semantic layouts.

Recent methods [1,4,5] utilize deep neural networks to directly learn the mapping between both input and output space. They rely on large numbers of training pairs to successfully learn the mapping. Zhang et al. [10] proposed CocosNet to improve image quality while reducing the number of training images required. In a different way, Pang et al. [12] presented SDM in order to improve the quality and semantic interpretability in semantic image synthesis, while Shi et al. [11] proposed RESAIL to improve the structure and texture clarity of the generated image details.

Unlike recent studies, we aim to improve the learning of this task by incorporating more geometric information. In addition to the semantic constraint, we add an intermediate step to produce a depth map, and the generated depth map together with the corresponding semantic labels serve as the input for our image generation model. As such, we hope that

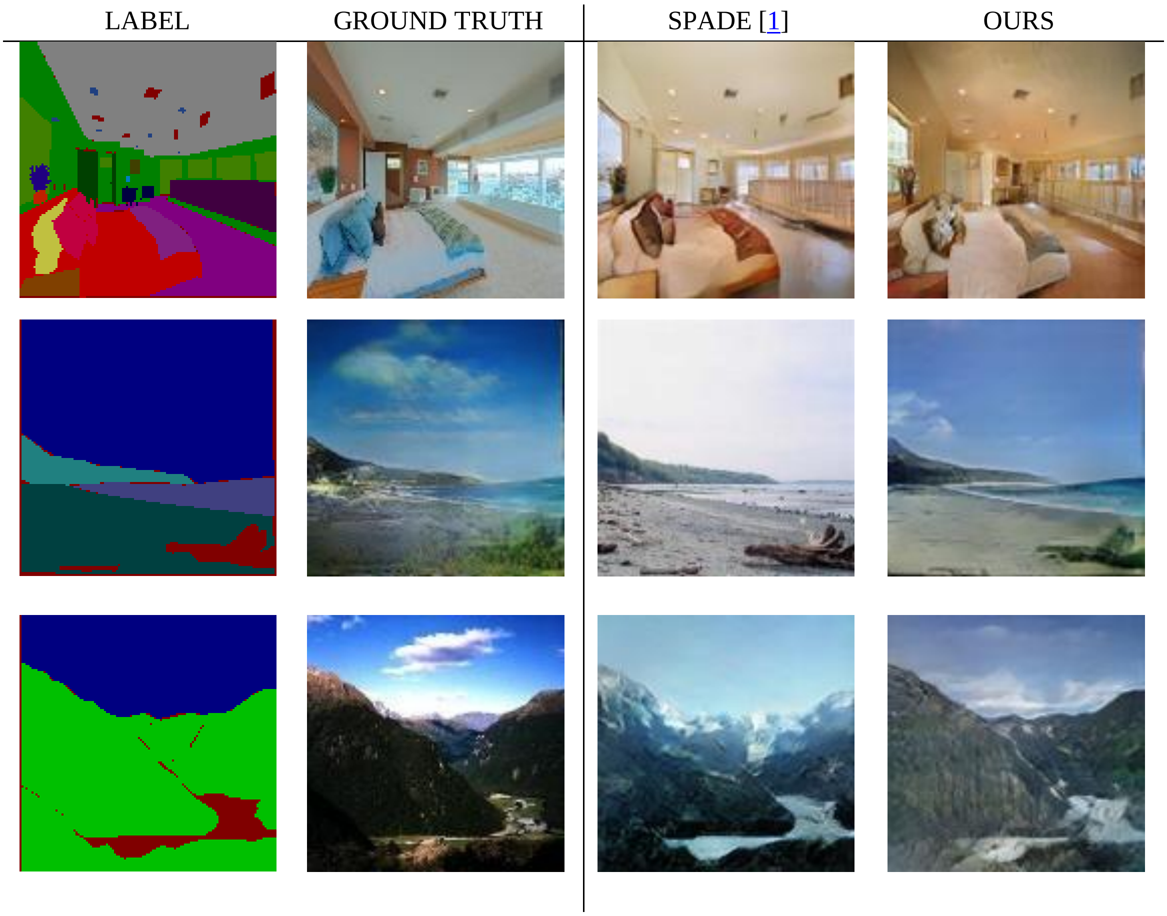

Fig 1: Visual comparison of semantic image synthesis results on the ADE20K. Our method successfully produces more realistic images.

*Authors contributed equally.

the model can explore more structural information for image generation.

Similar to other synthesis works, we use the SPADE [1] as our base model and extend it to learn the mapping between semantic labels and depth information. In addition, we further extend it to incorporate both semantic and depth information for image generation. Using our approach, we improve our baseline model by obtaining sharper boundaries as shown in Fig 1.

II. Methods

Here we explain the pipeline of our approach. Our model is divided into two stages as illustrated in Fig 2. In the Stage 1, a segmentation map is sent to the model to predict a depth map. After that, the generated depth map is concatenated with the input segmentation map as the input to the model of the Stage 2.

As described earlier, for both stages we use the SPADE [1] as our base model and modify it accordingly to deal with our different input formats and targets. The SPADE model is a GAN based model which is composed of a generator and a discriminator. Further details on the architecture of both generator and discriminator and the loss can be found in [1].

For Stage 1, the generator (G_1) takes the segmentation maps as input fed into different stages of SPADE residual blocks and finally predicts a depth map. At the same time, the discriminator (D_1) is trained jointly to tell if the predicted depth map is generated or not.

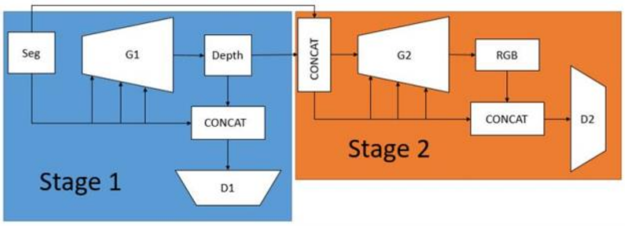

Fig 2: The architecture of our method. During Stage 1, we only used the original SPADE model to generate depth map images. We modified the original SPADE model in Stage 2 to make it take the concatenation of semantic segmentation and depth map as its input.

After the training loss of the Stage 1 is converged, we start training for the Stage 2. In Stage 2, we train another generator (G_2) which takes both segmentation maps and generated depth maps as inputs to generate final RGB images. The two inputs are concatenated, aiming to enforce the geometric information for the image generation task. In a similar manner, we train the discriminator (D_2) which takes the concatenation of all the inputs from Stage 2 and the corresponding generated image, or ground truth image, to differentiate whether the RGB image is generated or not.

III. Experiments

A. Dataset

We chose the ADE20K dataset to test our methods. ADE20K consists of 20,210 training and 2,000 validation images with corresponding fine-grained semantic labels. Similar to the COCO [9], ADE20K contains challenging scenes with 150 semantic classes. In the experiments, we adopted the training set for training, and used the validation set for evaluation.

B. Implementation details

In the training of Stage 1, we adopt two different settings of input size, 64*64 and 128*128, for depth image synthesis. While in the Stage 2, we use an input size of 64*64 to generate the same size of output due to the hardware constraints. For training, we use an NVIDIA GeForce RTX 3060 12GB GPU and set the batch size to 64 or 12 in Stage 1, and 64 in Stage 2. The weights of both the generator and discriminator were initialized randomly without any pretraining.

C. Performance Metrics

The Fréchet inception distance (FID) [8] score was used to evaluate the experimental results. It measures the image quality of generative models by comparing the distribution of generated image with the real one used to train the generator. The lower the FID value, the better the image quality.

As ADE20K does not contain depth information, we used a depth prediction model (monodepth2) from [6] pretrained on the KITTI dataset [7] to generate corresponding pseudo depth ground truth. In Stage 1, we train the first original SPADE model using the depth map obtained above as the ground truth and the semantic segmentation maps in ADE20K as the input. In Stage 2, we modify the original SPADE model and train it with the depth maps obtained in the Stage 1 along with the semantic segmentation masks in ADE20K as input and the original RGB images in ADE20K as ground truth.

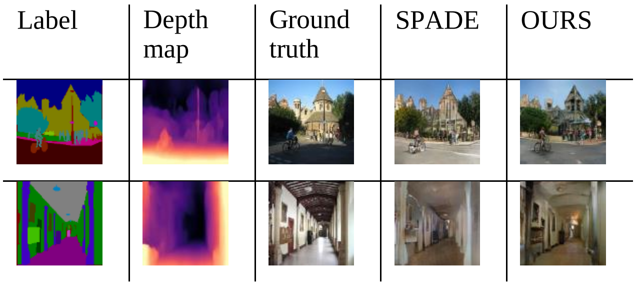

Fig 3: Qualitative results from Stage 2, while depth maps are obtained from Stage 1.

Table 1 shows the depth generation results from the ADE20K validation set. We calculate the FID score between the generated depth maps and the pseudo ground truth depth maps produced by monodepth2. In this experiment, we utilized two different output sizes of the depth map and observed that the higher resolution (128*128) gives a better FID score (16.06) compared with 19.96 for 64*64. The reason may be because the larger depth maps contain more detail, allowing the network to learn more information. This can be further confirmed from the qualitative results in Fig 4.

Table 2 presents the image generation results via the FID score on ADE20K validation set. Following the setting of Stage 1, we conducted our experiments in two different settings, 64*64 and 128*128, for the resolution of depth maps. As shown in Table 2, our proposed method always gives a lower FID score (46.42 and 46.01) than the baseline (46.47). In addition, we observed that the model generates better semantic image synthesis results. when we utilize a higher resolution of depth map (128x128). It is because higher resolution of depth map contains more structural information which is helpful for the image generation.

The qualitative results are shown in Fig 3. From the first row of the figure, we can see that the SPADE model does not properly handle the street light in front of the building while our method successfully generates the image following the semantic and depth constraints. Further results can be found in Fig 5. In most of the rows of Fig 5, our model generates a sharper image than the original SPADE model. Both quantitative and qualitative results prove that our proposed method can effectively consider the structural information from depth map and thus improve the image quality.

TABLE I. FID SCORE OF DEPTH MAP IMAGES

64*64	***128*128***
19.96	16.06

Table 1: The larger size of depth map image gives us a lower score, which indicates that larger maps contain more detail and will be more advantageous in Stage 2.

IV. CONCLUSION

In this paper, we proposed a method for considering depth information as an additional constraint for semantic image synthesis. For the situation where the depth maps are not available, we train a model to predict it directly from semantic segmentation masks utilizing a pseudo ground truth generated by a pretrained model. Experimental results illustrate the effectiveness of our proposed model. For future work, we plan to utilize a multi-task model to incorporate depth information.

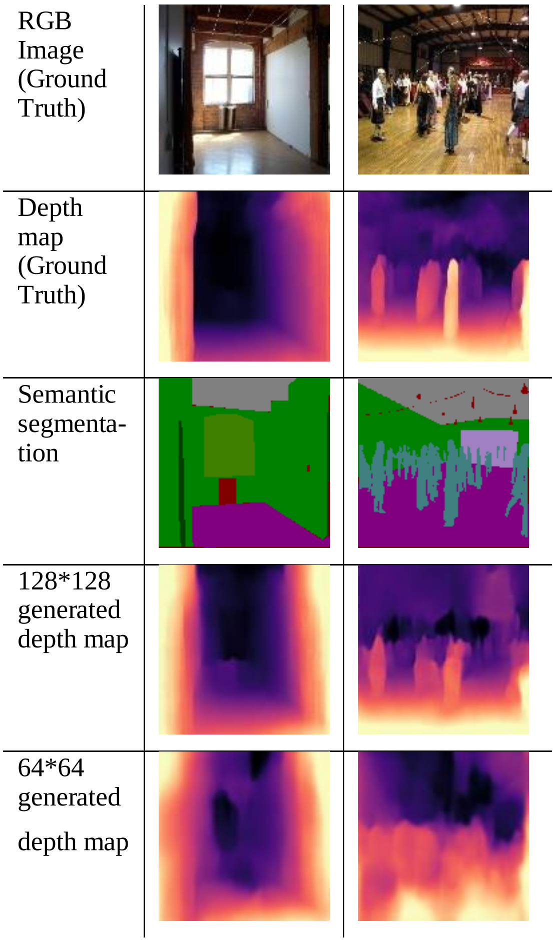

Fig 4: We use pre-made model monodepth2 to generate depth map ground truth with RGB images as a input. And then we train two new SPADE models to generate different sizes of generated depth map images.

TABLE II. FID SCORE OF GENERATED IMAGES

SPADE	***OURS(64, 64)***	***OURS(128, 64)***
46.47	46.42	**46.01**

Table 2: (X, Y) shows that we generated size X*X depth map images during Stage 1, and obtained size Y*Y realistic images in the end. The output size of the original SPADE model is therefore also 64*64.

REFERENCES

[1] T. Park, M. Liu, T. Wang, and J. Zhu. Semantic image synthesis with spatially-adaptive normalization. In *proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2337-2346,2019.

[2] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba. Scene parsing through ade20k dataset. In *IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2017.

[3] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2014.

[4] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision* (ICCV), 2017.

[5] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional GANs," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8798–8807.

[6] Clément Godard, Oisin Mac Aodha, Michael Firman and Gabriel J. Brostow. Digging Into Self-Supervised Monocular Depth Estimation. In *International Conference on Computer Vision* (ICCV),2019.

[7] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vison meets Robotics: The KITTI Dataset. In *International Journal of Robotics Research* (IJRR), 2013.

[8] M. He usel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems*,2017.

[9] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision* (ECCV), 2014.

[10] P. Zhang, B. Zhang, D. Chen, L. Yuan, and F. Wen. Cross-domain correspondence learning for exemplar-based image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5143–5153, 2020.

[11] Y. Shi, X. Liu, Y. Wei, Z. Wu and W. Zuo. Retrieval-based Spatially Adaptive Normalization for Semantic Image Synthesis. In *Computer Vision and Pattern Recognition* (CVPR), 2022.

[12] W. Wang, J. Bao and W. Zhou. Semantic Image Synthesis via Diffusion Models. In *Computer Vision and Pattern Recognition* (CVPR), 2022

[13] B. Park, H. Go and C. Kim. Bridging Implicit and Explicit Geometric Transformations for Single-Image View Synthesis. In *Computer Vision and Pattern Recognition*(CVPR), 2022

[14] Y. Yu, L. Zhang, H. Fan and T. Luo. High-Fidelity Image Inpainting with GAN Inversion. In *Computer Vision and Pattern Recognition*(CVPR), 2022

[15] Y. Du, Y. Shen, H. Wang, J. Fei, W. Li, L. Wu, R. Zhao, Z. Fu and Q. Liu. Learning from Future: A Novel Self-Training Framework for Semantic Segmentation. In *Computer Vision and Pattern Recognition*(CVPR), 2022

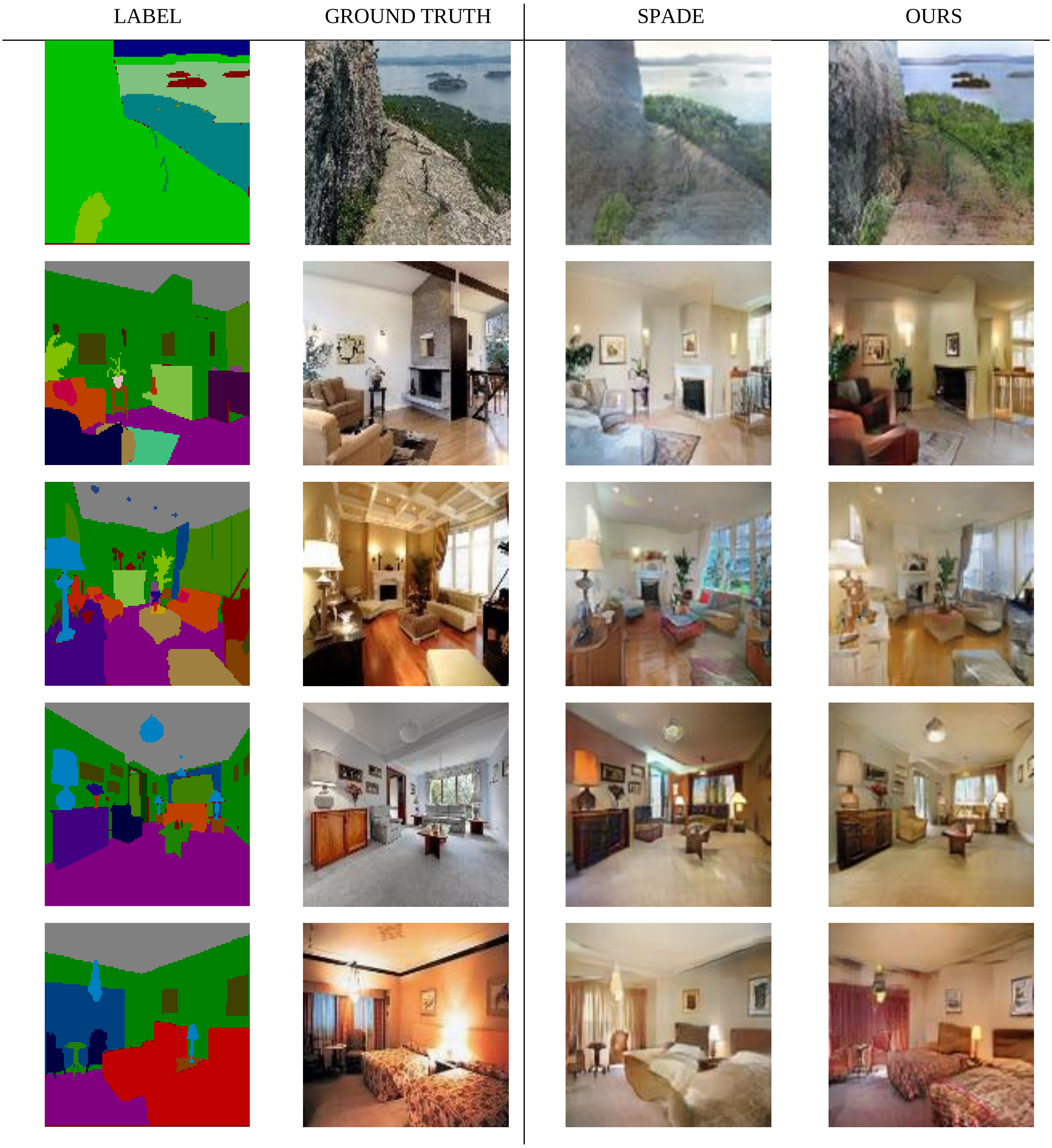

Fig 5: More visual comparison of semantic image synthesis results on the ADE20K. Our method does perform better than the original SPADE model.