UNCONDITIONAL SYNTHESIS OF COMPLEX SCENES USING A SEMANTIC BOTTLENECK

Samaneh Azadi[1], Michael Tschannen[2], Eric Tzeng[1], Sylvain Gelly[2], Trevor Darrell[1], Mario Lucic[2]
[1]University of California, Berkeley [2]Google Research, Brain Team

ABSTRACT

Coupling the high-fidelity generation capabilities of label-conditional image synthesis methods with the flexibility of unconditional generative models, we propose a semantic bottleneck GAN model for unconditional synthesis of complex scenes especially when the number of training examples is small. We assume pixel-wise segmentation labels are available during training and use them to learn the scene structure. During inference, our model first synthesizes a realistic segmentation layout from scratch, then synthesizes a realistic scene conditioned on that layout. For the former, we use an unconditional progressive segmentation generation network that captures the distribution of realistic semantic scene layouts. For the latter, we use a conditional segmentation-to-image synthesis network that captures the distribution of photo-realistic images conditioned on the semantic layout. When trained end-to-end, the resulting model outperforms state-of-the-art generative models in unsupervised image synthesis on two challenging domains in terms of the Fréchet Inception Distance and perceptual evaluations. Moreover, we demonstrate that the end-to-end training significantly improves the segmentation-to-image synthesis sub-network, which results in superior performance over the state-of-the-art when conditioning on real segmentation layouts.

1 INTRODUCTION

Significant strides have been made on generative models for image synthesis, with a variety of methods based on Generative Adversarial Networks (GANs) achieving state-of-the-art performance. At lower resolutions or in specialized domains, GAN-based methods are able to synthesize samples which are near-indistinguishable from real samples (Brock et al., 2019). However, generating complex, high resolution scenes from scratch remains a challenging problem. As image resolution and complexity increase, the coherence of synthesized images decreases — samples contain convincing local textures, but lack a consistent global structure.

Stochastic decoder-based models, such as conditional GANs, were recently proposed to alleviate some of these issues. In particular, both Pix2PixHD (Wang et al., 2018) and SPADE (Park et al., 2019) are able to synthesize high-quality scenes using a strong conditioning mechanism based on semantic segmentation labels during the scene generation process. Global structure encoded in the segmentation layout of the scene is what allows these models to focus primarily on generating convincing local content consistent with that structure. A key drawback of such conditional models is that they require full segmentation layouts as input. Thus, unlike unconditional generative approaches which synthesize images from randomly sampled noise, these models are limited to generating images from a set of scenes that is prescribed in advance, typically either through segmentation labels from an existing dataset, or scenes that are hand-crafted by experts.

To overcome these limitations, we propose Semantic Bottleneck GAN which couples high-fidelity generation capabilities of label-conditional models with the flexibility of unconditional image generation. This in turn enables our model to synthesize an unlimited number of novel complex scenes, while still maintaining high-fidelity output characteristic of image-conditional models.

Our Semantic Bottleneck GAN first unconditionally generates a pixel-wise semantic label map of a scene, and then generates a realistic scene image by conditioning on that semantic map. By factorizing the task into these two steps, we are able to separately tackle the problems of producing convincing segmentation layouts (i.e. a useful global structure) and filling these layouts with

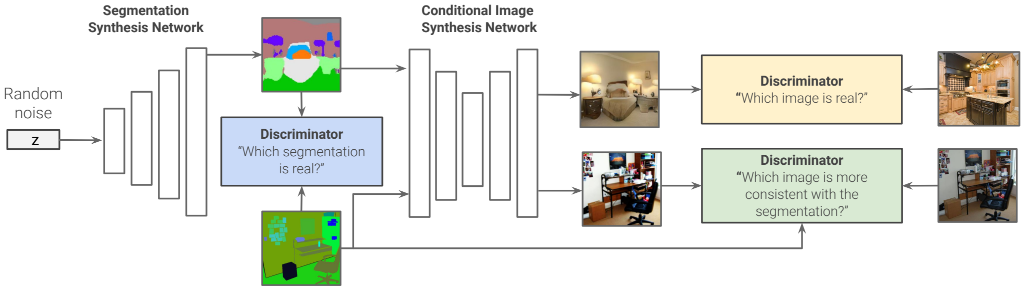

Figure 1: Schematic of Semantic Bottleneck GAN. Starting from random noise, we synthesize a segmentation layout and use a discriminator to bias the segmentation synthesis network towards realistic looking segmentation layouts. The generated layout is then provided as input to a conditional image synthesis network to synthesize the final image. A second discriminator is used to bias the conditional image synthesis network towards realistic images paired with real segmentation layouts. Finally, a third unconditional discriminator is used to bias the conditional image synthesis network towards generating images that match the ground truth.

convincing appearances (i.e. local structure). When trained end-to-end, the model yields samples which have a coherent global structure as well as fine local details. Empirical evaluation shows that our Semantic Bottleneck GAN achieves a new state-of-the-art on two complex datasets with relatively small number of training examples, Cityscapes and ADE-Indoor, as measured both by the Fréchet Inception Distance (FID) and by perceptual evaluations. Additionally, we observe that the conditional segmentation-to-image synthesis component of our SB-GAN jointly trained with segmentation layout synthesis significantly improves the state-of-the-art semantic image synthesis network (Park et al., 2019), resulting in higher-quality outputs when conditioning on ground truth segmentation layouts.

2 Semantic Bottleneck GAN (SB-GAN)

We propose an unconditional Semantic Bottleneck GAN architecture to learn the distribution of complex scenes. We assume the ground truth segmentation masks are available for all or part of the target scene dataset.

Semantic bottleneck synthesis We first learn a coarse estimate of the scene distribution from samples corresponding to real segmentation maps with K semantic categories. Starting from random noise, we generate a tensor $Y \in [\![1, K]\!]^{N \times 1 \times H \times W}$ which represents a per-pixel segmentation class, with H and W indicating the height and width, respectively, of the generated map and N the batch size. In practice, we progressively train this model from a low to a high resolution on the ground truth discrete-valued segmentation maps. We apply a softmax function to the last layer of the generator leading to an output that can be interpreted as a probability score for each pixel belonging to each of the K semantic classes. To have a differentiable sampling scheme to predict per-pixel semantic classes from this estimated probability map, we apply the Gumbel-softmax trick coupled with a straight-through estimator (Jang et al., 2017).

Semantic image synthesis Our second sub-network converts the synthesized semantic layouts into photo-realistic images using spatially-adaptive normalization (Park et al., 2019). The segmentation masks are employed to spread the semantic information throughout the generator by modulating the activations with a spatially adaptive learned transformation. We train this network using pairs of real RGB images and their corresponding segmentations from the target data set.

End-to-end framework After training the semantic bottleneck synthesis and semantic image synthesis models, we adversarially fine-tune the parameters of both networks in an end-to-end approach by introducing an unconditional discriminator network, D_2, on top of the SPADE generator (see Figure 1). This discriminator is designed to distinguish between real RGB images and the fake ones generated from the *synthesized* semantic layouts.In contrast to the conditional discriminator in

SPADE, which enforces consistency between the input semantic map and the output image, D_2 is primarily concerned with the overall quality of the final output.Through a joint fine-tuning of the two networks, the gradients with respect to RGB images synthesized by SPADE are back-propagated to the segmentation synthesis model, thereby encouraging it to synthesize segmentation layouts that lead to higher quality final images. Thus, SPADE plays the role of a loss function for synthesizing segmentations, but in the RGB space. Similarly, fine-tuning SPADE with synthesized segmentations allows it to adapt to a more diverse set of scene layouts improving the quality of generated samples.

3 Experiments and Results

We evaluate the performance of the proposed approach on Cityscapes and ADE-Indoor datasets containing images with complex scenes. Cityscapes-5K is a subset of Cityscapes-25K, where fine ground truth annotations are only provided for this subset. We extract the corresponding fine annotations for the rest of training images in Cityscapes-25K using the segmentation model (Yu et al., 2017; Yu & Koltun, 2016) trained on the training annotated samples from Cityscapes-5K. We compare the performance of SB-GAN against the SOTA BigGAN model (Brock et al., 2019) as well as a ProGAN (Karras et al., 2017) baseline that has been trained on the RGB images directly.

In Figures 2, 3 and 4, we provide qualitative comparisons of the competing methods. We observe that both Cityscapes-5K and ADE-Indoor are very challenging for the state-of-the-art ProGAN and BigGAN models, likely due to the complexity of the data and small number of training instances. BigGAN suffers from mode collapse, as illustrated in the last row of Figure 4. In contrast, SB-GAN significantly improves the structure of the scene distribution, and provides samples of higher quality. On Cityscapes-25K, the performance improvement of SB-GAN is more modest due to the large number of training images available. It is worth emphasizing that in this case only 3K ground truth segmentations for training SB-GAN are available. Compared to BigGAN, images synthesized by SB-GAN are sharper and contain more structural details (e.g., one can zoom-in on the synthesized cars). We also report FID and perceptual evaluation scores in Table 1 revealing the superiority of SB-GAN to the state-of-the-art baselines. In our perceptual study, evaluators were asked to select a quality score from 1 to 4, indicating terrible and high quality images, respectively.

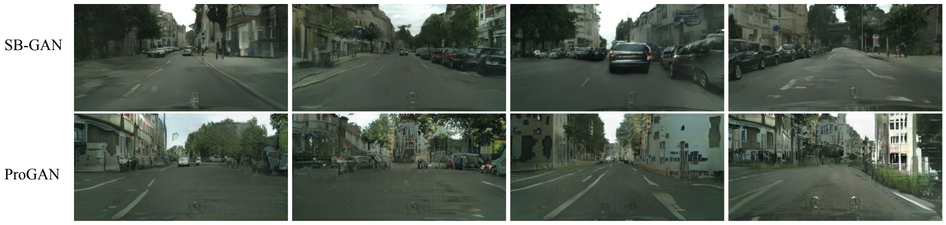

Figure 2: Cityscapes-5K. Zoom in for more detail. SB-GAN (1st row) generates more convincing objects, e.g. buildings and cars.

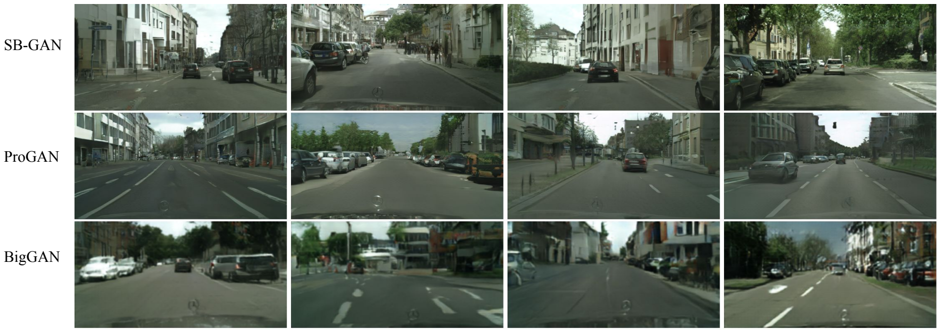

Figure 3: Cityscapes-25K. Zoom in for more detail. Images synthesized by BigGAN (3rd row) are blurry and sometimes defective in local structures.

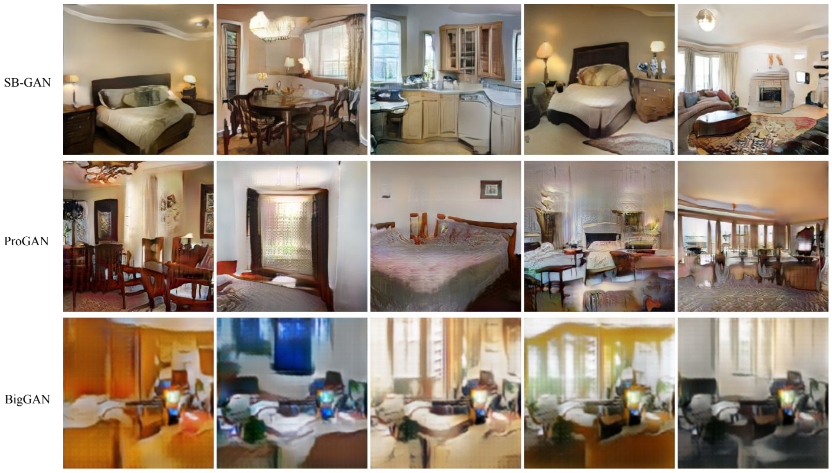

Figure 4: ADE-Indoor: A challenging dataset causing mode collapse for the BigGAN model (3rd row). In contrast, samples generated by SB-GAN (1st row) are generally of higher quality and much more structured than those of ProGAN (2nd row).

Table 1: FID and perceptual evaluation scores of the synthesized samples at two resolutions of 128x256 and 256x512 for Cityscapes and resolutions of 128x128 and 256x256 for ADE-Indoor. BigGAN fails to capture the distribution of Cityscapes-5K. Moreover, BigGAN could not be successfully trained at a 256 resolution due to instability observed during training and mode collapse.

	CITYSCAPES-5K			CITYSCAPES-25K			ADE-INDOOR		
Method	$\text{FID}_{256}\downarrow$	$\text{FID}_{128}\downarrow$	perc ↑	$\text{FID}_{256}\downarrow$	$\text{FID}_{128}\downarrow$	perc ↑	$\text{FID}_{256}\downarrow$	$\text{FID}_{128}\downarrow$	perc ↑
ProGAN	92.57	178.19	2.08	63.87	56.7	2.53	104.83	85.94	2.35
BigGAN	-	-	-	-	64.82	2.27	-	156.65	1.96
SB-GAN	**65.49**	**57.48**	**2.48**	**62.97**	**54.92**	**2.61**	**85.27**	**81.39**	**2.49**

Table 2: FID of the synthesized samples when conditioned on the ground truth labels.

	CITYSCAPES-5K	CITYSCAPES-25K	ADE-INDOOR
SPADE	72.12	60.83	50.30
SB-GAN	**60.39**	**54.13**	**48.15**

Generating by conditioning on real segmentations To independently assess the impact of end-to-end training on the conditional image synthesis sub-network, we evaluate the quality of generated samples when conditioning on ground truth validation segmentations from each dataset. Comparisons to the baseline network SPADE (Park et al., 2019) are provided in Table 2. We observe that the image synthesis component of SB-GAN consistently outperforms SPADE across all three datasets, indicating that fine-tuning on data sampled from the segmentation generator improves the conditional image generator.

4 CONCLUSION

We proposed an end-to-end Semantic Bottleneck GAN model that synthesizes semantic layouts from scratch, and then generates photo-realistic scenes conditioned on the synthesized layouts. Through extensive quantitative and qualitative evaluations, we showed that this novel end-to-end training pipeline significantly outperforms the state-of-the-art models in unconditional synthesis of complex scenes. In addition, Semantic Bottleneck GAN strongly improves the performance of the state-of-the-art semantic image synthesis model in synthesizing photo-realistic images from ground truth segmentations.

REFERENCES

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. 2019. 1, 3

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with Gumbel-softmax. In *ICLR*, 2017. 2

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *ICLR*, 2017. 3

Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *CVPR*, 2019. 1, 2, 4

Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *CVPR*, 2018. 1

Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. In *ICLR*, 2016. 3

Fisher Yu, Vladlen Koltun, and Thomas Funkhouser. Dilated residual networks. In *CVPR*, 2017. 3