YOU ONLY NEED ADVERSARIAL SUPERVISION FOR SEMANTIC IMAGE SYNTHESIS

Edgar Schönfeld[1], **Vadim Sushko**[1], **Dan Zhang**[1], **Jürgen Gall**[2], **Bernt Schiele**[3], **Anna Khoreva**[1]
[1]Bosch Center for Artificial Intelligence, [2]University of Bonn, [3]Max Planck Institute for Informatics

ABSTRACT

Despite their recent successes, GAN models for semantic image synthesis still suffer from poor image quality when trained with only adversarial supervision. Historically, additionally employing the VGG-based perceptual loss has helped to overcome this issue, significantly improving the synthesis quality, but at the same time limited the progress of GAN models for semantic image synthesis. In this work, we propose a novel, simplified GAN model, which needs only adversarial supervision to achieve high quality results. We re-design the discriminator as a semantic segmentation network, directly using the given semantic label maps as the ground truth for training. Moreover, we enable multi-modal image synthesis through global and local sampling of a 3D noise tensor, which allows complete or partial image editing. With the proposed modifications we achieve synthesis of higher quality and diversity compared to previous state-of-the-art models. Furthermore, we show that our trained discriminator allows generating multiple versions of a given image by predicting its label map and using it as input to the generator, potentially reducing the amount of label map annotations needed during inference.

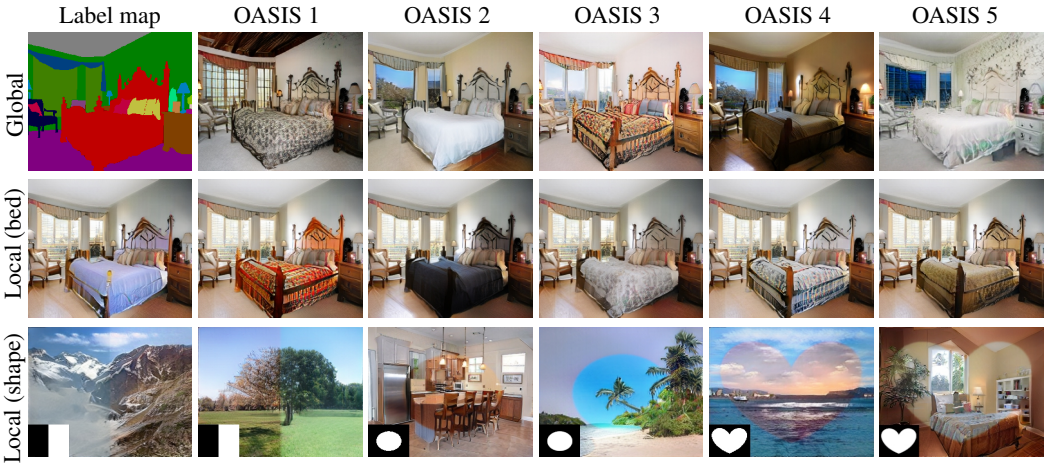

Figure 1: OASIS multi-modal synthesis results. The 3D noise can be sampled globally, or locally. For the latter, we only re-sample noise in the bed segment area (in red) or arbitrary drawn area.

1 INTRODUCTION

Semantic image synthesis is the task of generating realistic images from semantic label maps (Wang et al., 2018; Park et al., 2019). State-of-the-art models (Wang et al., 2018; Park et al., 2019; Liu et al., 2019) still suffer from training instabilities and poor image quality when trained only with adversarial supervision. This is commonly overcome by using an additional perceptual loss (Wang et al., 2018) that matches intermediate generator features of the synthetic and real images, estimated via an external VGG perception network (Simonyan & Zisserman, 2015) pre-trained on ImageNet (Deng et al., 2009). However, the perceptual loss causes an additional computational overhead and a potential bias towards ImageNet, which decreases image diversity and impacts quality, as we show in our experiments. Hence, in this work we propose a novel, simplified GAN model that improves the state of the art without requiring a perceptual loss. We argue that the previously used encoder-shaped discriminators require perceptual supervision as they alone are not capable of providing the rich supervision necessary for the task. Therefore, we re-design the discriminator as an encoder-decoder

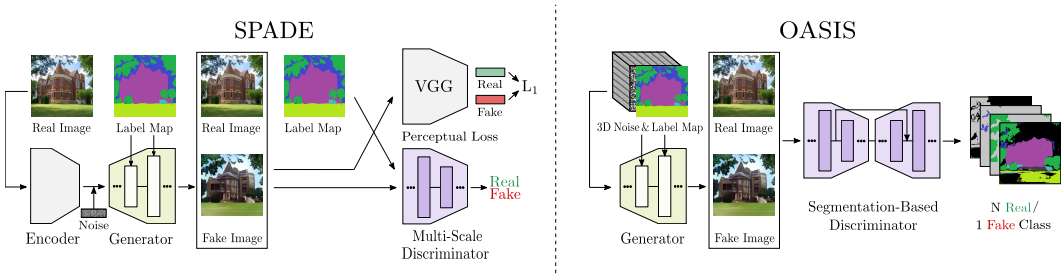

Figure 2: SPADE (left) vs. OASIS (right). OASIS outperforms SPADE, while being simpler and lighter: it uses only adversarial loss supervision and a single segmentation-based discriminator, without relying on heavy external networks. Furthermore, OASIS learns to synthesize multi-modal outputs by directly re-sampling the 3D noise tensor, instead of using an image encoder as in SPADE.

semantic segmentation network (Ronneberger et al., 2015), which exploits the given semantic label maps as ground truth via an N+1-class cross-entropy loss (N real semantic classes and 1 fake class). Since the discriminator has to look at each patch in detail to assign it to the right semantic class, it learns semantic-aware fine-grained representations. In contrast, the previous encoder-shaped discriminators require concatenating the label map to the input image, allowing them to ignore parts of the label map, as they are not penalized for predicting wrong semantic classes. Second, we introduce a LabelMix regularization, which helps the discriminator to focus more on the semantic and structural differences of real and synthetic images. Our changes lead to a much stronger discriminator which makes the perceptual loss supervision superfluous. Third, we enable multi-modal image synthesis of the generator via a new noise sampling scheme. Previously, direct input noise to the generator was either ignored or led to images of poor quality (Wang et al., 2018; Park et al., 2019). In contrast, our model can synthesize diverse multi-modal outputs by simply re-sampling an input 3D noise tensor. As our noise tensor is spatially sensitive, we can re-sample it both globally (channel-wise) and locally (pixel-wise), allowing to change not only the appearance of the whole scene, but also of specific semantic classes or any chosen areas (see Fig. 1). We call our model OASIS, as it needs **o**nly **a**dversarial supervision for **s**emantic **i**mage **s**ynthesis.

2 OASIS MODEL

Segmentation-based discriminator. Our model builds on SPADE (Park et al., 2019). The comparison of OASIS and SPADE is shown in Fig. 2. The discriminator of SPADE follows multi-scale PatchGAN (Isola et al., 2017), adopting two image classification networks. Instead, we use only one network, and propose to cast the discriminator task as multi-class semantic segmentation, redesigning its architecture to follow a U-Net (Ronneberger et al., 2015). In the proposed setting, the discriminator D segments real images using the given semantic label maps with N classes as the ground truth, and aims to assign a separate class label to all pixels of fake images, synthesized by the generator G. Overall, we have $N+1$ classes in the semantic segmentation problem, and thus propose to train the discriminator using a weighted (N+1)-class cross-entropy loss:

$$\mathcal{L}_D = -\mathbb{E}_{(x,t)}\left[\sum_{c=1}^{N}\alpha_c\sum_{i,j}^{H\times W} t_{i,j,c}\log D(x)_{i,j,c}\right] - \mathbb{E}_{(z,t)}\left[\sum_{i,j}^{H\times W}\log D(G(z,t))_{i,j,c=N+1}\right], \quad (1)$$

where x denotes the real image, (z,t) is the noise-label map pair, and $D(x)$ is per-pixel (N+1)-class prediction. The ground truth label map t has three dimensions: the spatial position $(i,j) \in H\times W$, and one-hot vector encoding the class $c \in \{1,..,N+1\}$. To balance contributions of different semantic classes, we add a balancing weight α_c, which is the inverse of per-pixel class frequency. In accordance to the discriminator loss, the OASIS generator loss is

$$\mathcal{L}_G = -\mathbb{E}_{(z,t)}\left[\sum_{c=1}^{N}\alpha_c\sum_{i,j}^{H\times W} t_{i,j,c}\log D(G(z,t))_{i,j,c}\right]. \quad (2)$$

LabelMix regularization. To encourage our discriminator to focus on structural differences between fake and real classes, we propose a LabelMix regularization. Based on the semantic layout, we generate a binary mask M to mix a pair $(x,\hat{x})$ of real and fake images conditioned on the same label map: $\text{LabelMix}(x,\hat{x},M) = M\odot x + (1-M)\odot\hat{x}$. We train the discriminator to be equivariant under the LabelMix operation by adding $\mathcal{L}_{cons}$ to Eq. 1: $\mathcal{L}_{cons} = \left\| D_{\text{logits}}\Big(\text{LabelMix}(x,\hat{x},M)\Big) - \text{LabelMix}\Big(D_{\text{logits}}(x), D_{\text{logits}}(\hat{x}), M\Big)\right\|^2$. LabelMix is different to Cut-

Method	# param	VGG	ADE20K		ADE-outd.		Cityscapes		COCO-stuff	
			FID↓	mIoU↑	FID↓	mIoU↑	FID↓	mIoU↑	FID↓	mIoU↑
Pix2pixHD	183M	✓	81.8	20.3	97.8	17.4	95.0	58.3	111.5	14.6
CC-FPSE	131M	✓	31.7	43.7	n/a	n/a	54.3	65.5	19.2	41.6
SPADE	102M	✓	33.9	38.5	63.3	30.8	71.8	62.3	22.6	37.4
SPADE+	102M	✓	32.9	42.5	51.1	32.1	47.8	64.0	21.7	38.8
		✗	60.7	21.0	65.4	22.7	61.4	47.6	99.1	16.1
OASIS	94M	✗	**28.4**	**50.6**	**48.4**	**41.5**	**47.7**	**69.3**	**16.7**	**45.5**

Table 1: Comparison with other methods across datasets. Bold denotes best performance.

Mix (Yun et al., 2019), which randomly samples the binary mask M. Instead, we generate M according to the label map and therefore encourage the generator to respect semantic class boundaries.

3D noise. To achieve multi-modality, SPADE model resorted to using an image encoder. We propose a simpler solution, enabling multi-modal synthesis through sampling of a 3D noise. We construct a 64-dimensional noise vector and replicate it to match the spatial dimensions of the label map. The channel-wise concatenation of the noise and label map forms a 3D tensor used as input and also for conditioning at the SPADE-norm layers. In doing so, intermediate feature maps are conditioned on both the semantic labels and the noise. As the 3D noise is channel-wise and pixel-wise sensitive, at test time, one can sample the noise tensor globally, per-channel, and locally, per-pixel, allowing control of the whole scene but also of specific semantic classes (Fig. 1). Lastly, we remove the first generator block, decreasing the parameter count by 24M without a noticeable performance loss.

3 Experiments

Experimental setup. We conduct experiments on ADE20K, ADE-Outdoors (Zhou et al., 2017), COCO-stuff (Caesar et al., 2018) and Cityscapes (Cordts et al., 2016). We evaluate our model quantitatively using the Fréchet Inception Distance (FID) (Heusel et al., 2017), mean Intersection-over-Union (mIoU) and MS-SSIM (Wang et al., 2003) with LPIPS (Zhang et al., 2018) for multi-modal synthesis. FID and mIoU estimate the quality of generated images and alignment with input label maps, while MS-SSIM and LPIPS evaluate image diversity. We follow the experimental setting of (Park et al., 2019). We did not apply the GAN feature matching loss and used VGG perceptual loss only for ablations with $\lambda_{VGG} = 10$. All our models use an exponential moving average (EMA) with 0.9999 decay (Yaz et al., 2018). We use SPADE (Park et al., 2019) as our baseline, trained without the feature matching loss and using EMA (Yaz et al., 2018) (further referred to as SPADE+). In addition, to investigate how the perceptual loss influences image quality we measure the color and texture similarity between real and generated images. Colors are compared via the earth mover's distance between LAB space color histograms (Rubner et al., 2000), while texture similarity is measured as the χ^2-distance between Local Binary Patterns histograms (Ojala et al., 1996).

Main results. OASIS outperforms the current state of the art on all datasets with an average improvement of 6 FID and 7 mIoU points (Table 1). Importantly, OASIS achieves the improvement via adversarial supervision alone, without a perceptual loss. On the contrary, as seen from the table, SPADE+ does not produce images of high visual quality without the perceptual loss. A strong discriminator is the key factor for good performance: without a rich training signal from the discriminator, the SPADE+ generator has to learn through minimizing the VGG loss. This is visualized in Fig. 3, which shows that SPADE+ struggles to learn the color and texture distribution of real data without VGG. In contrast, with the stronger OASIS discriminator there is no need for additional supervision from the VGG loss. This allows producing images of notably higher quality with better alignment to input label maps, reaching color and texture distributions closer to real data.

In contrast to previous work, OASIS produces diverse images by simply re-sampling input 3D noise, not requiring an additional image encoder. As such noise has spatial dimensions, an image can be re-sampled both globally and locally, for example editing only a selected semantic region (Fig. 1). We observed that OASIS reaches higher diversity compared to SPADE+ (Table 2). As seen from Table 2, a strong quality-diversity tradeoff exists for SPADE+: 3D noise improves diversity at the cost of quality, and the perceptual loss improves quality at the cost of diversity. Such trade-off is not visible for OASIS, where the VGG loss also reduces diversity but does not noticeably affect quality. Overall, we observe that 3D noise is beneficial for diversity, while the VGG perceptual loss, on the

Table 2: Multi-modal synthesis evaluation on ADE20K. Bold and red denote the best and the worst performance.

Method	Multi-mod.	VGG	MS-SSIM↓	LPIPS↑	FID↓	mIoU↑
SPADE+	Encoder	✓	0.85	0.16	33.4	40.2
SPADE+	3D noise	✗	**0.35**	**0.50**	58.4	18.7
		✓	0.53	0.36	34.4	36.2
OASIS	3D noise	✗	0.65	0.35	**28.3**	48.8
		✓	0.88	0.15	31.6	**50.8**

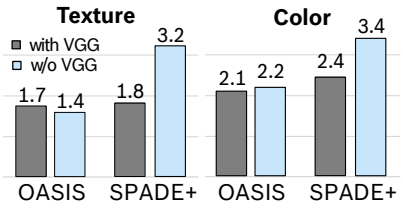

Figure 3: Histogram distances to real data.

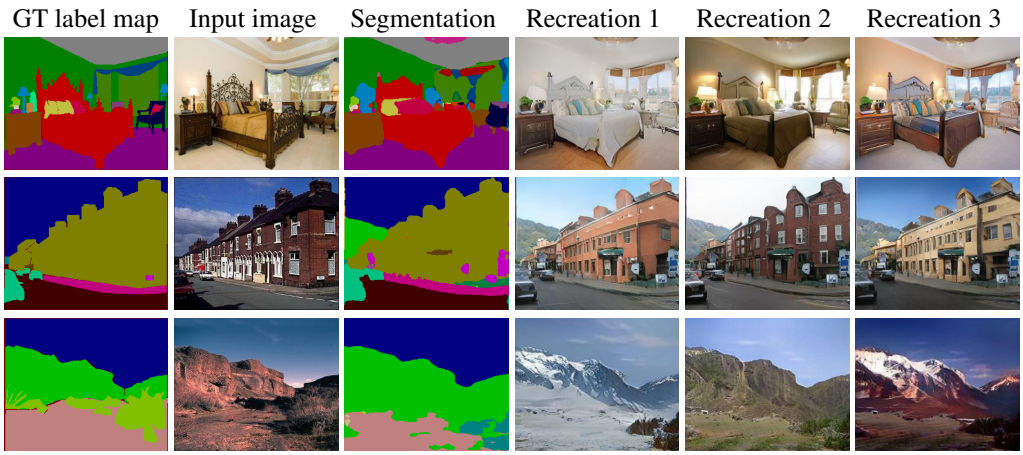

Figure 4: After training, the OASIS discriminator can be used to segment images. Columns 1-3 show the ground truth label map, real image, and segmentation of the discriminator. Using the predicted label map the generator can produce multiple versions of the original image by resampling noise (Recreations 1-3). Note that this alleviates the need of ground truth maps during inference.

other hand, limits the variability among generated samples. Unlike the discriminator features, VGG features are static and do not change in response to the generator. Therefore the generator can easily get stuck at dominant modes imposed by the VGG features space, which limits its diversity.

Application to unlabelled data. OASIS has a unique property that its discriminator is trained to be an image segmenter. We observed that it shows good performance on this task, reaching the mIoU of 40.0 on ADE20K validation set. For comparison, current state of the art on ADE20K is a mIoU of 46.91, achieved by ResNeST (Zhang et al., 2020). Such a good segmentation performance allows OASIS to be applied to unlabelled images: given an unseen image without a ground truth annotation, OASIS can predict a label map via the discriminator. Subsequently feeding this prediction to the generator allows to synthesize a scene with the same layout but different style. This property is shown in Fig. 4. Due to the good segmentation performance, the recreated scenes closely follow the ground truth label map of the original image. The high sensitivity of OASIS to the 3D noise enforces good variability, so the recreations are different from each other. We believe that creating multiple versions of one image while retaining the layout can be useful for data-augmentation.

4 Conclusion

In this work we propose OASIS, a semantic image synthesis model that only relies on adversarial supervision to achieve high fidelity image synthesis. In contrast to previous work, our model eliminates the need for a perceptual loss, which often imposes extra constraints on image quality and diversity. This is achieved via detailed spatial and semantic-aware supervision from our novel segmentation-based discriminator, which uses semantic label maps as ground truth for training. With this powerful discriminator, OASIS can easily generate diverse multi-modal outputs by re-sampling the 3D noise, both globally and locally, allowing to change the appearance of the whole scene and of individual objects. OASIS significantly improves over the state of the art, both in terms of image quality and diversity, while being simpler and more lightweight than previous methods.

REFERENCES

Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Xihui Liu, Guojun Yin, Jing Shao, Xiaogang Wang, et al. Learning to predict layout-to-image conditional convolutions for semantic image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Timo Ojala, Matti Pietikäinen, and David Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern recognition*, 29(1):51–59, 1996.

Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015.

Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision (IJCV)*, 2000.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2015.

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 *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *Asilomar Conference on Signals, Systems & Computers*, 2003.

Yasin Yaz, Chuan-Sheng Foo, Stefan Winkler, Kim-Hui Yap, Georgios Piliouras, Vijay Chandrasekhar, et al. The unusual effectiveness of averaging in gan training. In *International Conference on Learning Representations*, 2018.

Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *International Conference on Computer Vision (ICCV)*, 2019.

Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Zhi Zhang, Haibin Lin, Yue Sun, Tong He, Jonas Mueller, R Manmatha, et al. Resnest: Split-attention networks. *arXiv preprint arXiv:2004.08955*, 2020.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.