Semantic Image Synthesis

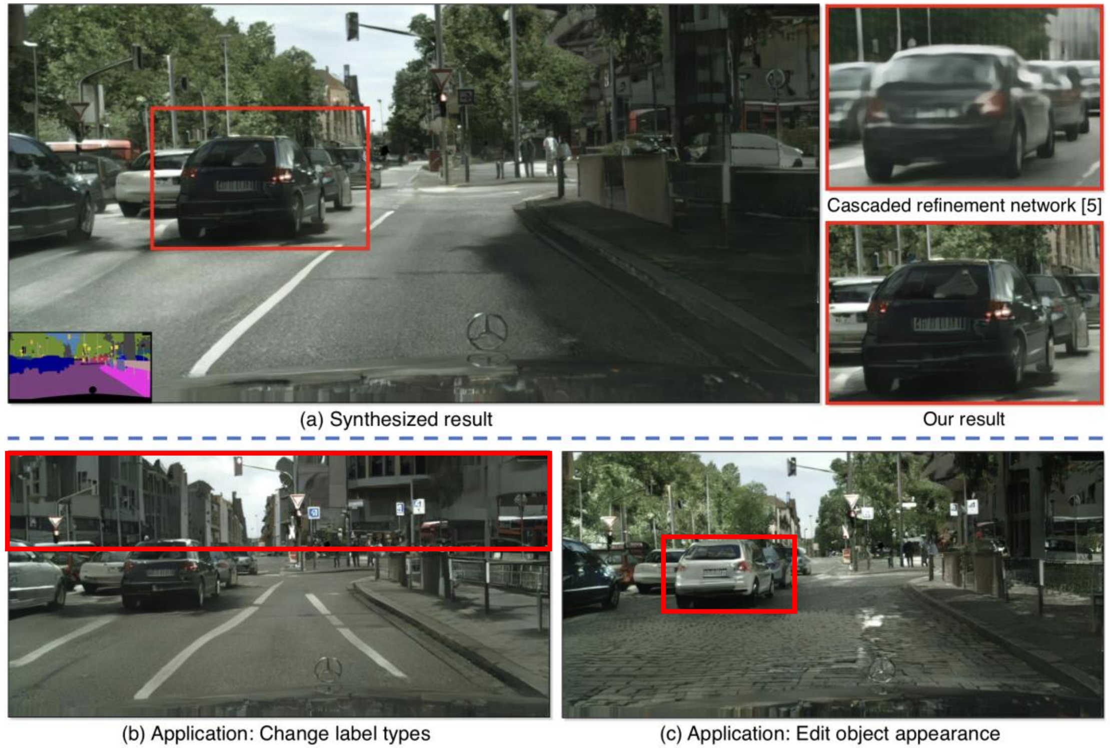
Presented by Minho Park

TOC

- Pix2pixHD (CVPR’18)
- SPADE (CVPR’19)
- CC-FPSE (NeurIPS’19)
- LGGAN (CVPR’20)
- SESAME (ECCV’20)
- OASIS (ICLR’21)
- SAFM (CVPR’22)

Pix2pixHD: Target Task

- Creating synthetic dataset for training visual recognition algorithms.
- New tools for higher-level image editing.
    - E.g., adding objects to images or changing the appearance of existing objects.



Cascaded refinement network [5]

(a) Synthesized result

Our result

(b) Application: Change label types

(c) Application: Edit object appearance

Figure 1: We propose a generative adversarial framework for synthesizing 2048 × 1024 images from semantic label maps (lower left corner in (a)). Compared to previous work [5], our results express more natural textures and details. (b) We can change labels in the original label map to create new scenes, like replacing trees with buildings. (c) Our framework also allows a user to edit the appearance of individual objects in the scene, e.g. changing the color of a car or the texture of a road. Please visit our website for more side-by-side comparisons as well as interactive editing demos.

Wang, Ting-Chun, et al. "High-resolution image synthesis and semantic manipulation with conditional gans." *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2018.

Pix2pixHD: Method

- High-resolution image synthesis and semantic manipulation with conditional GANs.
- Pix2pixHD overcame high-resolution image synthesis **using a residual connection.**

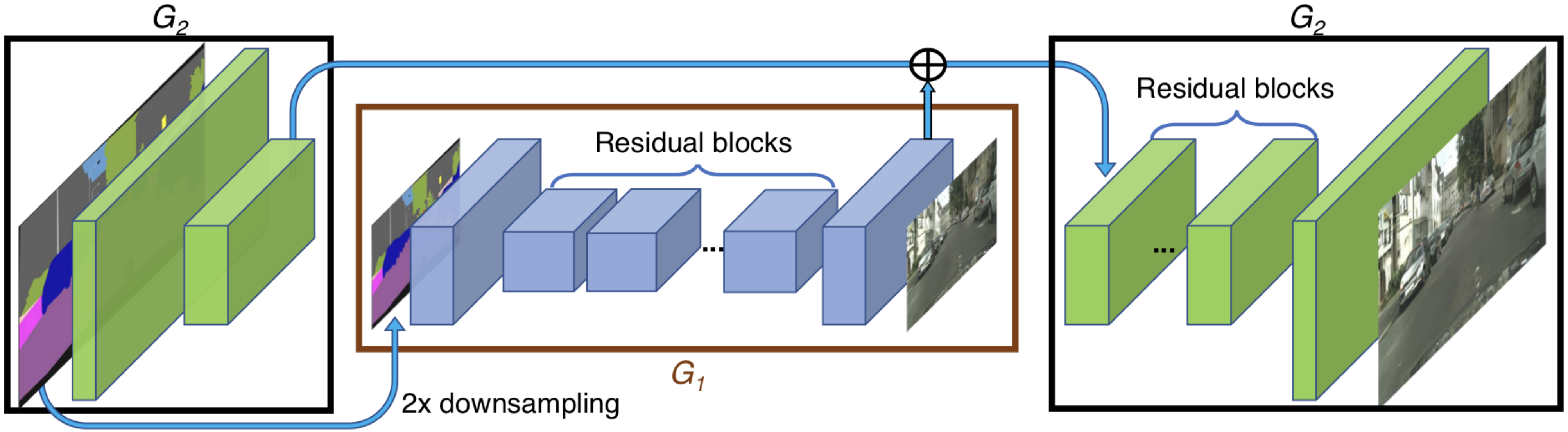
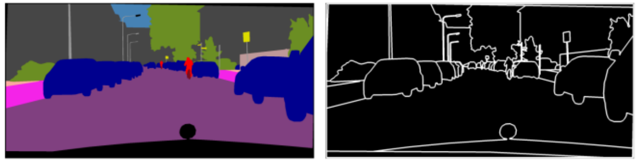


Figure 3: Network architecture of our generator. We first train a residual network G_1 on lower resolution images. Then, another residual network G_2 is appended to G_1 and the two networks are trained jointly on high resolution images. Specifically, the input to the residual blocks in G_2 is the element-wise sum of the feature map from G_2 and the last feature map from G_1.

Wang, Ting-Chun, et al. "High-resolution image synthesis and semantic manipulation with conditional gans." *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2018.

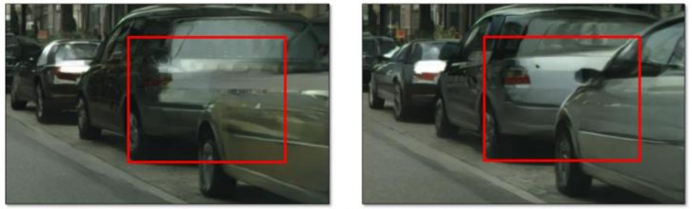
Pix2pixHD: Additional Tricks

- Using instance maps.
- Learning an instance-level feature embedding for diversity.



(a) Semantic labels (b) Boundary map

Figure 4: Using instance maps: (a) a typical semantic label map. Note that all connected cars have the same label, which makes it hard to tell them apart. (b) The extracted instance boundary map. With this information, separating different objects becomes much easier.



(a) Using labels only (b) Using label + instance map

Figure 5: Comparison between results without and with instance maps. It can be seen that when instance boundary information is added, adjacent cars have sharper boundaries.

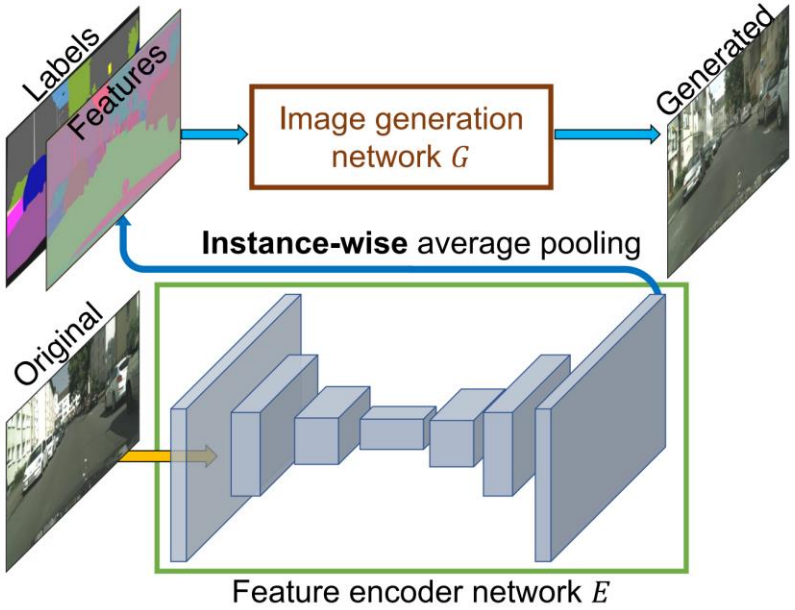

Labels
Features
Image generation network G
Generated
Instance-wise average pooling
Original
Feature encoder network E

Figure 6: Using instance-wise features in addition to labels for generating images.

Wang, Ting-Chun, et al. "High-resolution image synthesis and semantic manipulation with conditional gans." *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2018.

SPADE: Wash Away

- **Instance normalization layers tend to "wash away" semantic information.**
- $y_{bchw} = \frac{x_{bchw} - \mu_{bc}}{\sigma_{bc}}$.

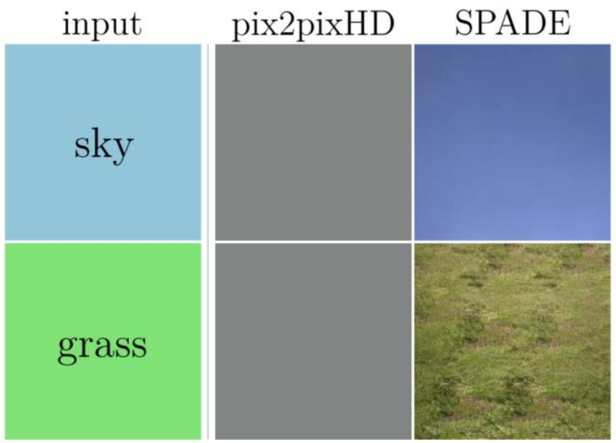
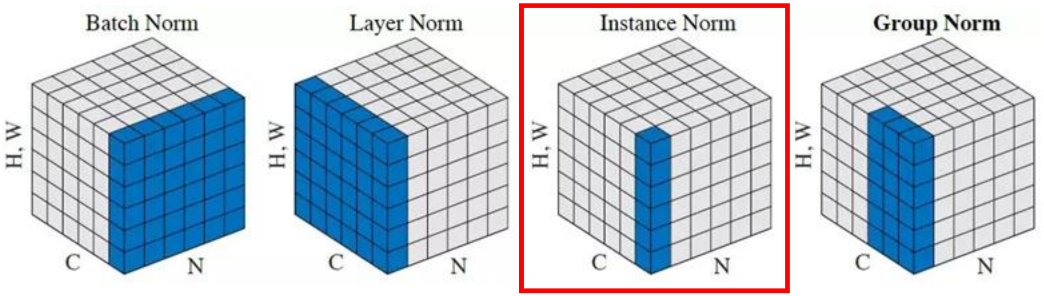


Figure 3: Comparing results given uniform segmentation maps: while the SPADE generator produces plausible textures, the pix2pixHD generator [48] produces two identical outputs due to the loss of the semantic information after the normalization layer.



Various types of normalization.

Park, Taesung, et al. "Semantic image synthesis with spatially-adaptive normalization." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019.
Wu, Yuxin, and Kaiming He. "Group normalization." *Proceedings of the European conference on computer vision (ECCV)*. 2018.

SPADE: SPatially-Adaptive (DE)normalization

- **We have to inject semantic information using normalization.**
- Batch normalization w/ spatial adaptive parameter γ, β
- $\gamma_{chw}(m) \cdot \frac{h_{bchw}-\mu_c}{\sigma_c} + \beta_{chw}(m).$
- m: semantic segmentation mask.

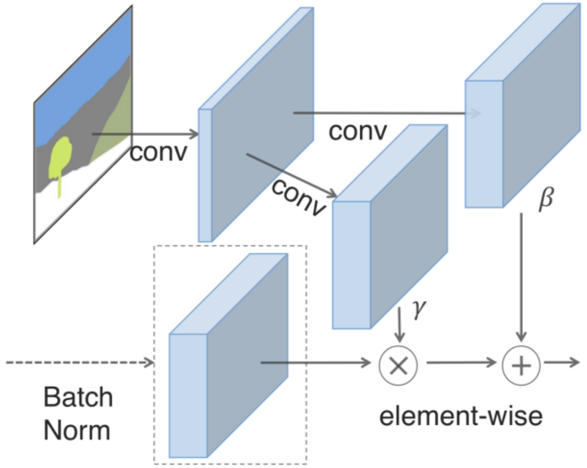
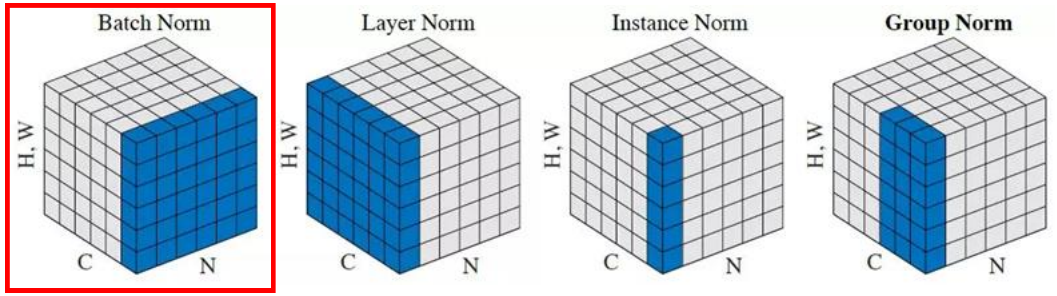


Figure 2: In the SPADE, the mask is first projected onto an embedding space and then convolved to produce the modulation parameters γ and β. Unlike prior conditional normalization methods, γ and β are not vectors, but tensors with spatial dimensions. The produced γ and β are multiplied and added to the normalized activation element-wise.



Various types of normalization.

Park, Taesung, et al. "Semantic image synthesis with spatially-adaptive normalization." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019.
Wu, Yuxin, and Kaiming He. "Group normalization." *Proceedings of the European conference on computer vision (ECCV)*. 2018.

SPADE: SPADE generator (GauGAN)

- SPADE generator does not need encoder.

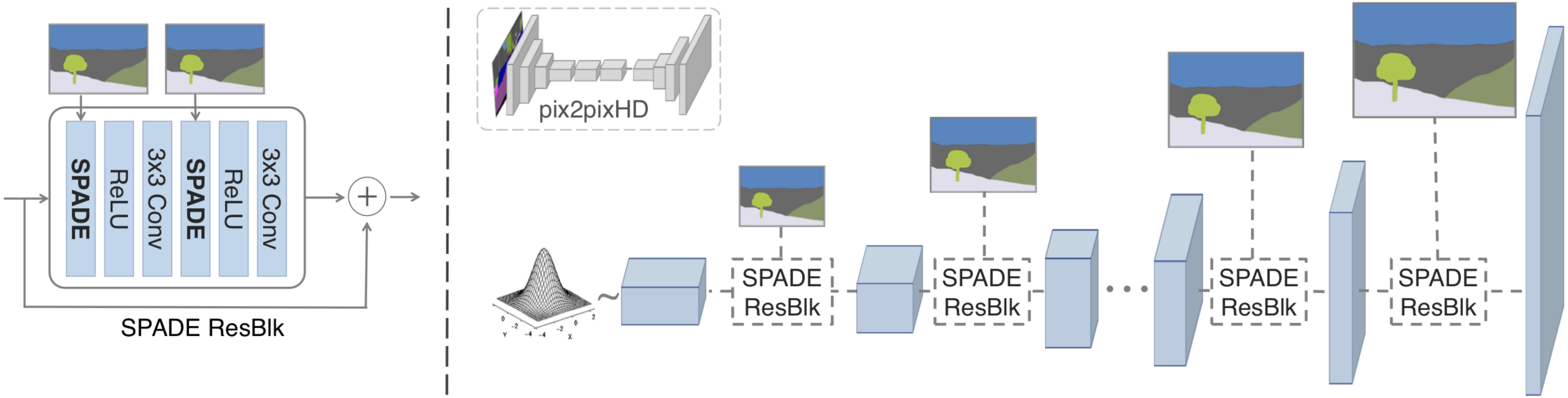


Figure 4: In the SPADE generator, each normalization layer uses the segmentation mask to modulate the layer activations. *(left)* Structure of one residual block with the SPADE. *(right)* The generator contains a series of the SPADE residual blocks with upsampling layers. Our architecture achieves better performance with a smaller number of parameters by removing the downsampling layers of leading image-to-image translation networks such as the pix2pixHD model [48].

Park, Taesung, et al. "Semantic image synthesis with spatially-adaptive normalization." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019.

SPADE: Image Encoder

- Use image encoder with KL Divergence loss for style manipulation.

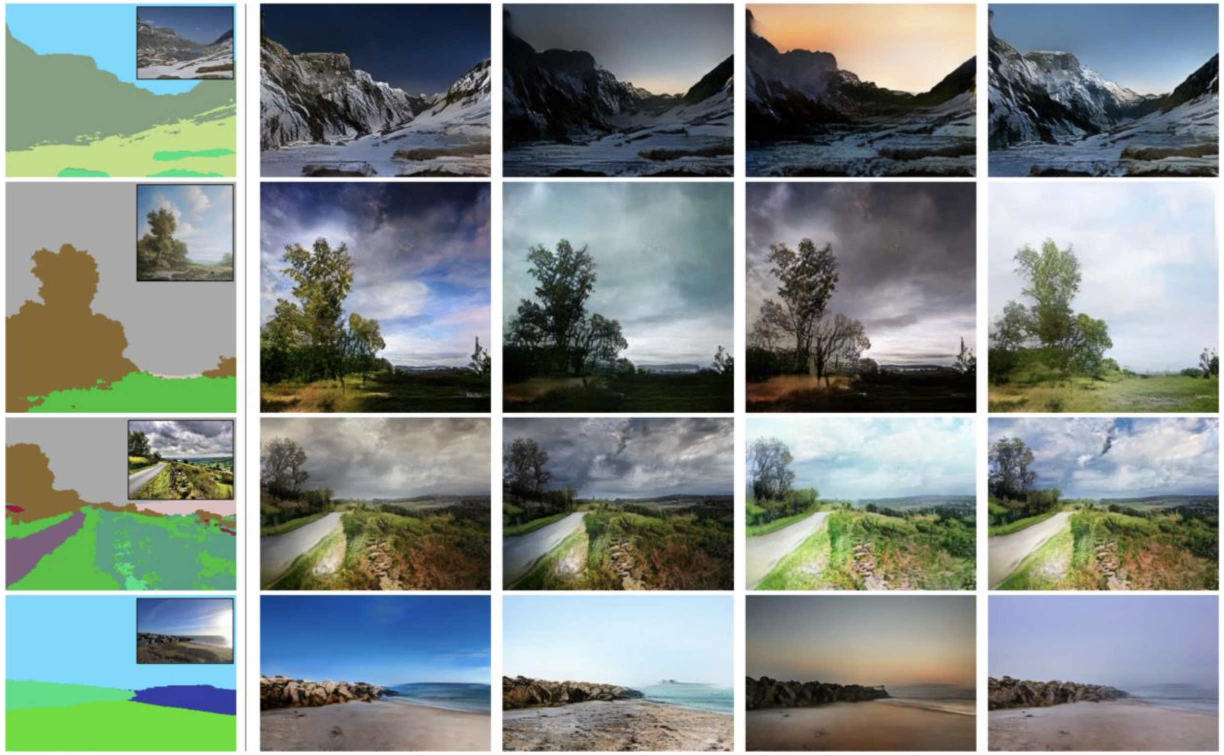


Figure 9: Our model attains multimodal synthesis capability when trained with the image encoder. During deployment, by using different random noise, our model synthesizes outputs with diverse appearances but all having the same semantic layouts depicted in the input mask. For reference, the ground truth image is shown inside the input segmentation mask.

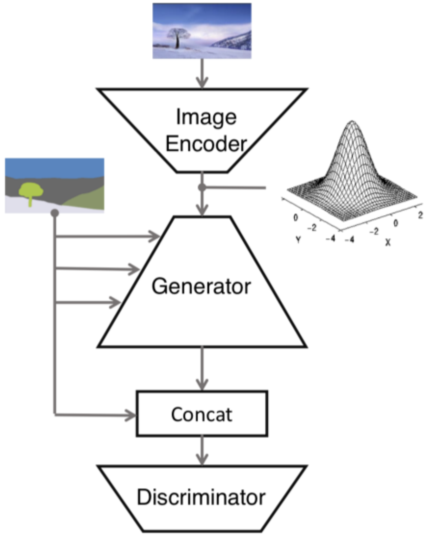


Figure 15: The image encoder encodes a real image to a latent representation for generating a mean vector and a variance vector. They are used to compute the noise input to the generator via the reparameterization trick [28]. The generator also takes the segmentation mask of the input image as input via the proposed SPADE ResBlks. The discriminator takes concatenation of the segmentation mask and the output image from the generator as input and aims to classify that as fake.

Park, Taesung, et al. "Semantic image synthesis with spatially-adaptive normalization." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019.

CC-FPSE: CC Generator

- **Convolution operation is spatial invariant.** ⇒ Hyper-network for convolutional generator.
- Label map encoder needs a larger receptive field (SPADE's is 5×5). ⇒ Use FPN.

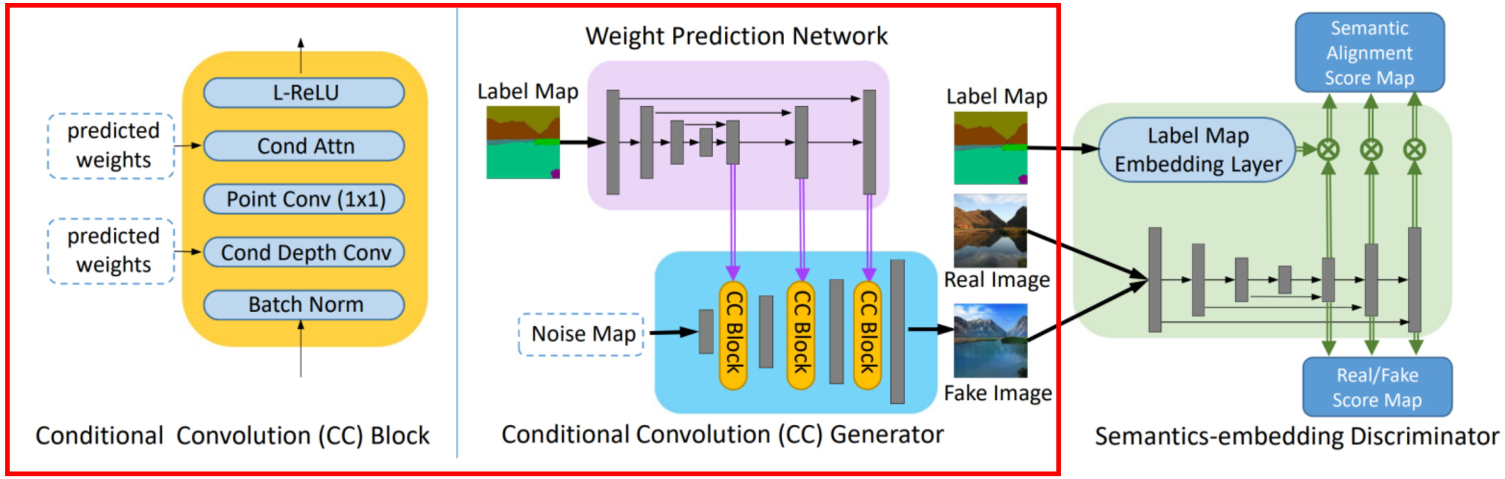

predicted weights
L-ReLU
Cond Attn
Point Conv (1x1)
predicted weights
Cond Depth Conv
Batch Norm
Conditional Convolution (CC) Block
Weight Prediction Network
Label Map
Noise Map
CC Block
CC Block
CC Block
Conditional Convolution (CC) Generator
Label Map
Real Image
Fake Image
Semantic Alignment Score Map
Label Map Embedding Layer
Real/Fake Score Map
Semantics-embedding Discriminator

Figure 2: (Left) The structure of a Conditional Convolution Block (CC Block). (Right) The overall framework of our proposed CC-FPSE. The weight prediction network predicts weights for CC Blocks in the generator. The conditional convolution generator is built up of Conditional Convolution (CC) Blocks shown on the left. The feature pyramid semantics-embedding (FPSE) discriminator predicts real/fake scores as well as semantic alignment scores. L-ReLU in the CC Block denotes Leaky ReLU.

Liu, Xihui, et al. "Learning to predict layout-to-image conditional convolutions for semantic image synthesis." *Advances in Neural Information Processing Systems* 32 (2019).

CC-FPSE: FPSE Discriminator

- Feature Pyramid Semantics-Embedding.

1. High-fidelity details such as texture and edges.
2. Semantic alignment with the input semantic map.

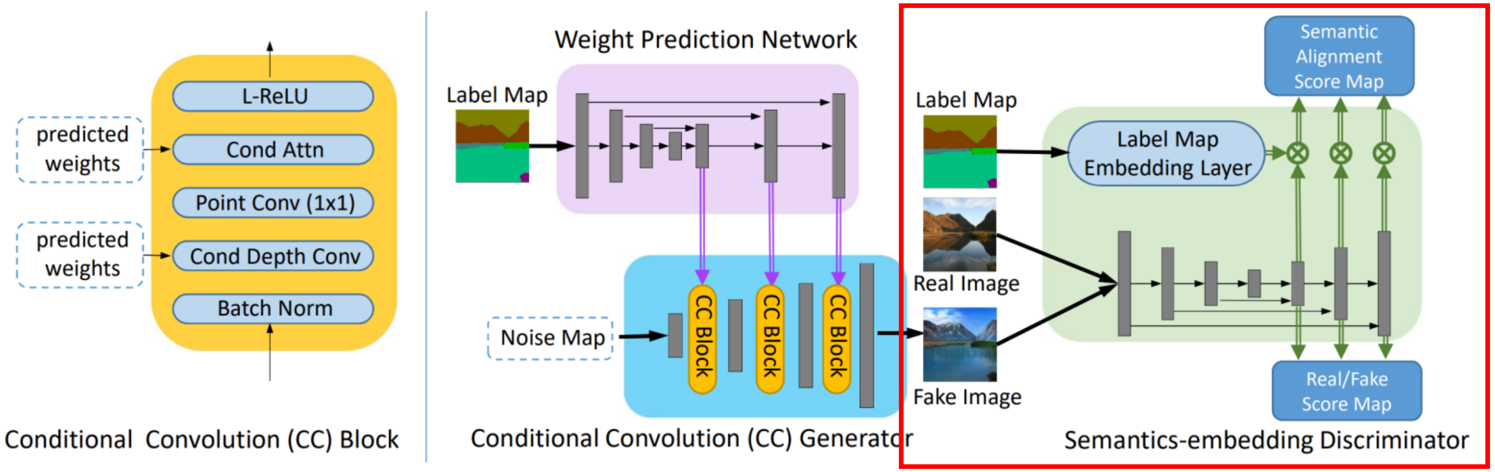

L-ReLU
predicted weights
Cond Attn
Point Conv (1x1)
predicted weights
Cond Depth Conv
Batch Norm
Conditional Convolution (CC) Block
Weight Prediction Network
Label Map
Noise Map
CC Block
CC Block
CC Block
Conditional Convolution (CC) Generator
Label Map
Real Image
Fake Image
Semantic Alignment Score Map
Label Map Embedding Layer
Real/Fake Score Map
Semantics-embedding Discriminator

Figure 2: (Left) The structure of a Conditional Convolution Block (CC Block). (Right) The overall framework of our proposed CC-FPSE. The weight prediction network predicts weights for CC Blocks in the generator. The conditional convolution generator is built up of Conditional Convolution (CC) Blocks shown on the left. The feature pyramid semantics-embedding (FPSE) discriminator predicts real/fake scores as well as semantic alignment scores. L-ReLU in the CC Block denotes Leaky ReLU.

Liu, Xihui, et al. "Learning to predict layout-to-image conditional convolutions for semantic image synthesis." *Advances in Neural Information Processing Systems* 32 (2019).

CC-FPSE: Qualitative Results

- Increase diversity and reality in the intra-class semantic map.

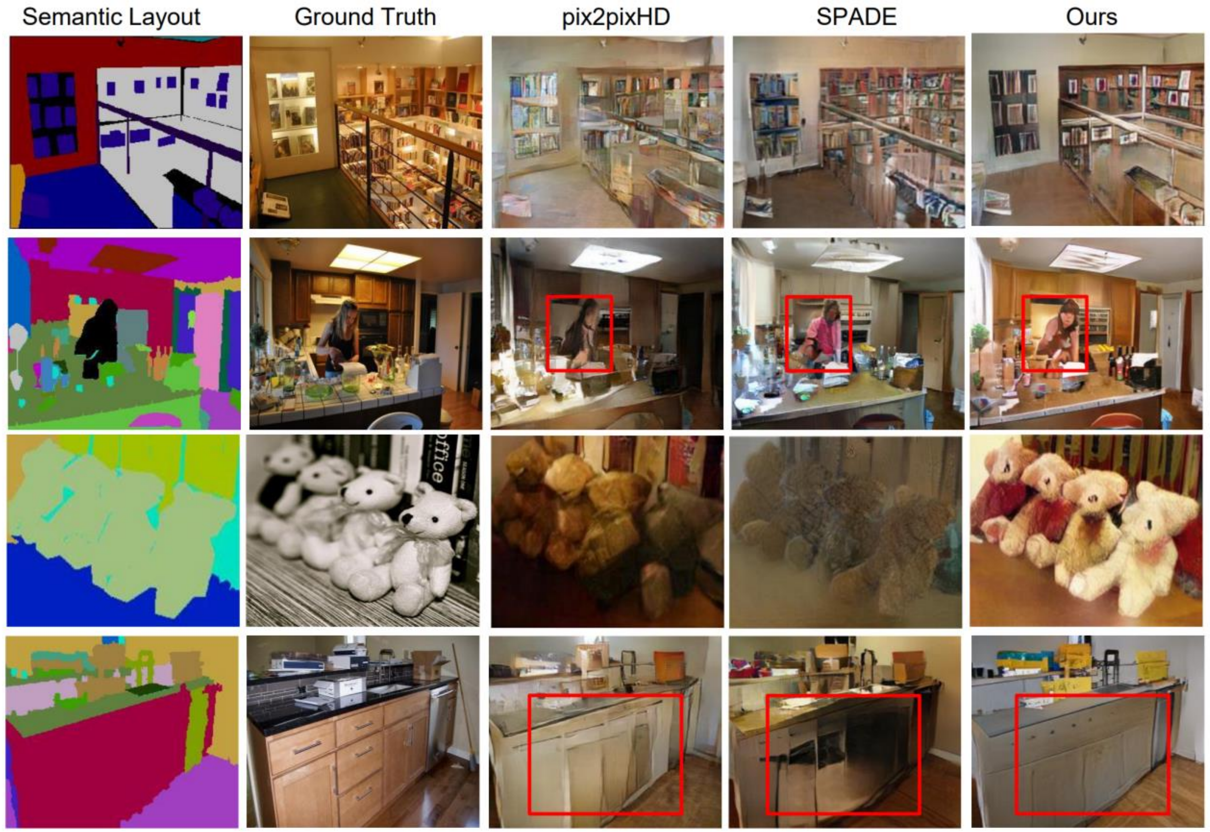


Figure 3: Results comparison with previous approaches. Better viewed in color. Zoom in for details.

Liu, Xihui, et al. "Learning to predict layout-to-image conditional convolutions for semantic image synthesis." *Advances in Neural Information Processing Systems* 32 (2019).

OASIS: Method

- 3D noise as an input.
- Segmentation-based discriminator.

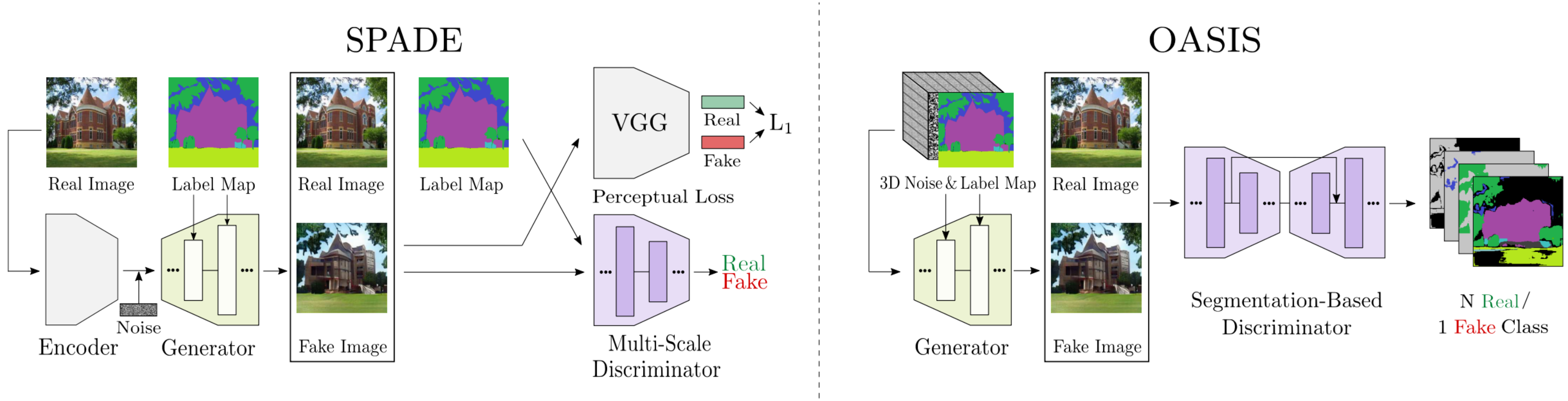


Figure 3: SPADE (left) vs. OASIS (right). OASIS outperforms SPADE, while being simpler and lighter: it uses only adversarial loss supervision and a single segmentation-based discriminator, without relying on heavy external networks. Furthermore, OASIS learns to synthesize multi-modal outputs by directly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

Schönfeld, Edgar, et al. "You only need adversarial supervision for semantic image synthesis." *International Conference on Learning Representations*. 2021.

OASIS: LabelMix

- Regularizer for $N + 1$-th label map.

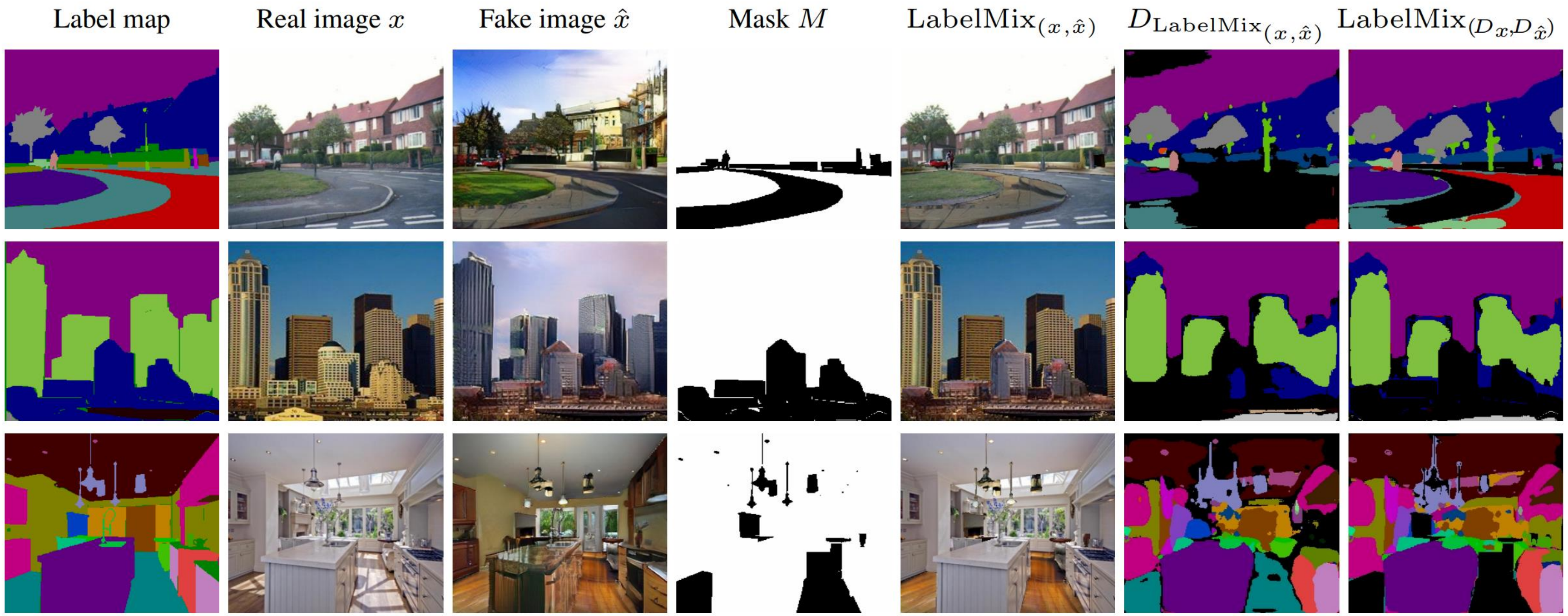

Label map
Real image x
Fake image $\hat{x}$
Mask M
$\mathrm{LabelMix}_{(x,\hat{x})}$
$D_{\mathrm{LabelMix}_{(x,\hat{x})}}$
$\mathrm{LabelMix}_{(D_x,D_{\hat{x}})}$

Figure 4: LabelMix regularization. Real x and fake $\hat{x}$ images are mixed using a binary mask M, sampled based on the label map, resulting in $\mathrm{LabelMix}_{(x,\hat{x})}$. The consistency regularization then minimizes the L2 distance between the logits of $D_{\mathrm{LabelMix}_{(x,\hat{x})}}$ and $\mathrm{LabelMix}_{(D_x,D_{\hat{x}})}$. In this visualization, **black** corresponds to the fake class in the N+1 segmentation output.

Schönfeld, Edgar, et al. "You only need adversarial supervision for semantic image synthesis." *International Conference on Learning Representations*. 2021.

OASIS: Multi-modal Synthesis

- Due to 3D noise, OASIS achieves spatially diverse SIS.

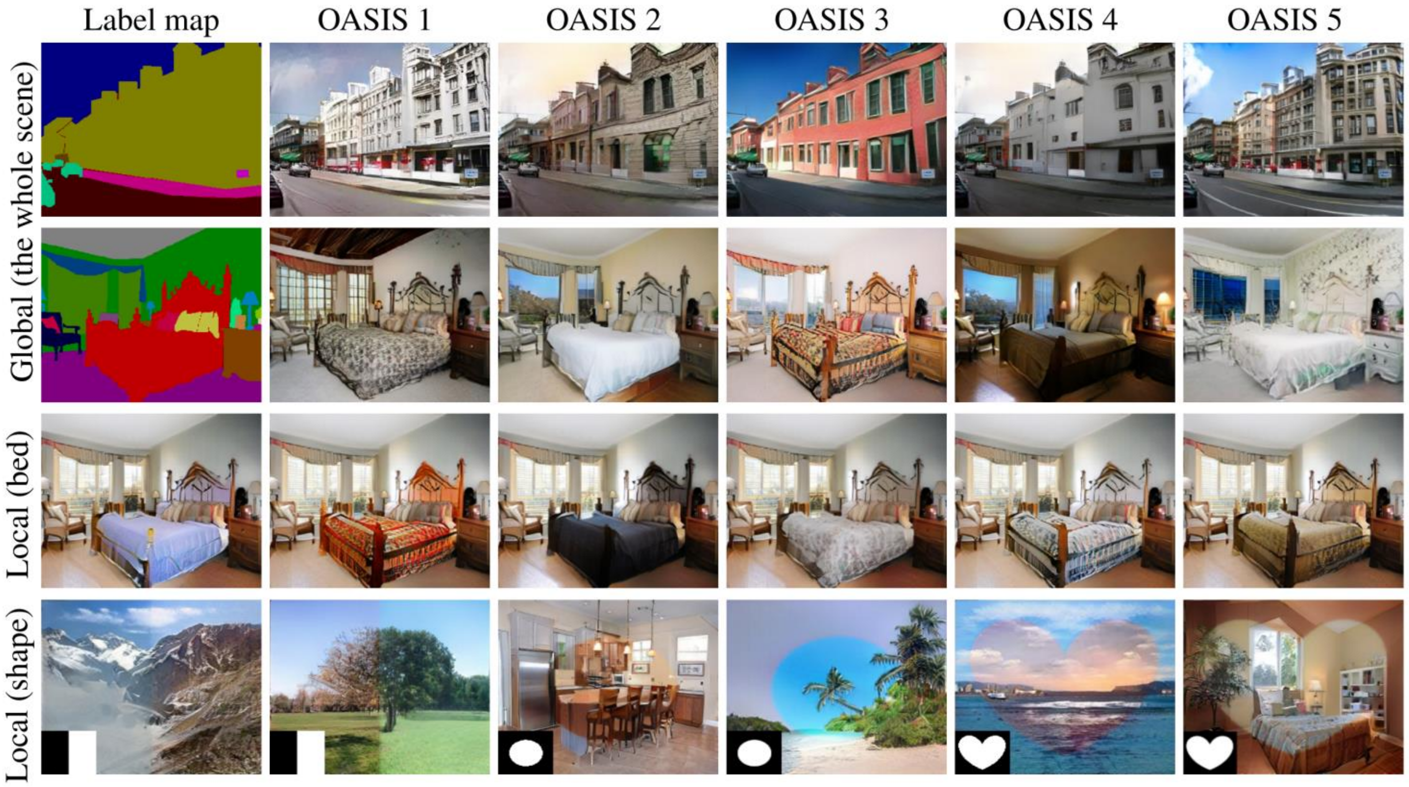


Figure 2: OASIS multi-modal synthesis results. The 3D noise can be sampled globally (first 2 rows), changing the whole scene, or locally (last 2 rows), partially changing the image. For the latter, we sample different noise per region, like the bed segment (in red) or arbitrary areas defined by shapes.

Schönfeld, Edgar, et al. "You only need adversarial supervision for semantic image synthesis." *International Conference on Learning Representations*. 2021.

OASIS: Qualitative Results

- Better perceptual quality and FID score without VGG-perceptual loss.

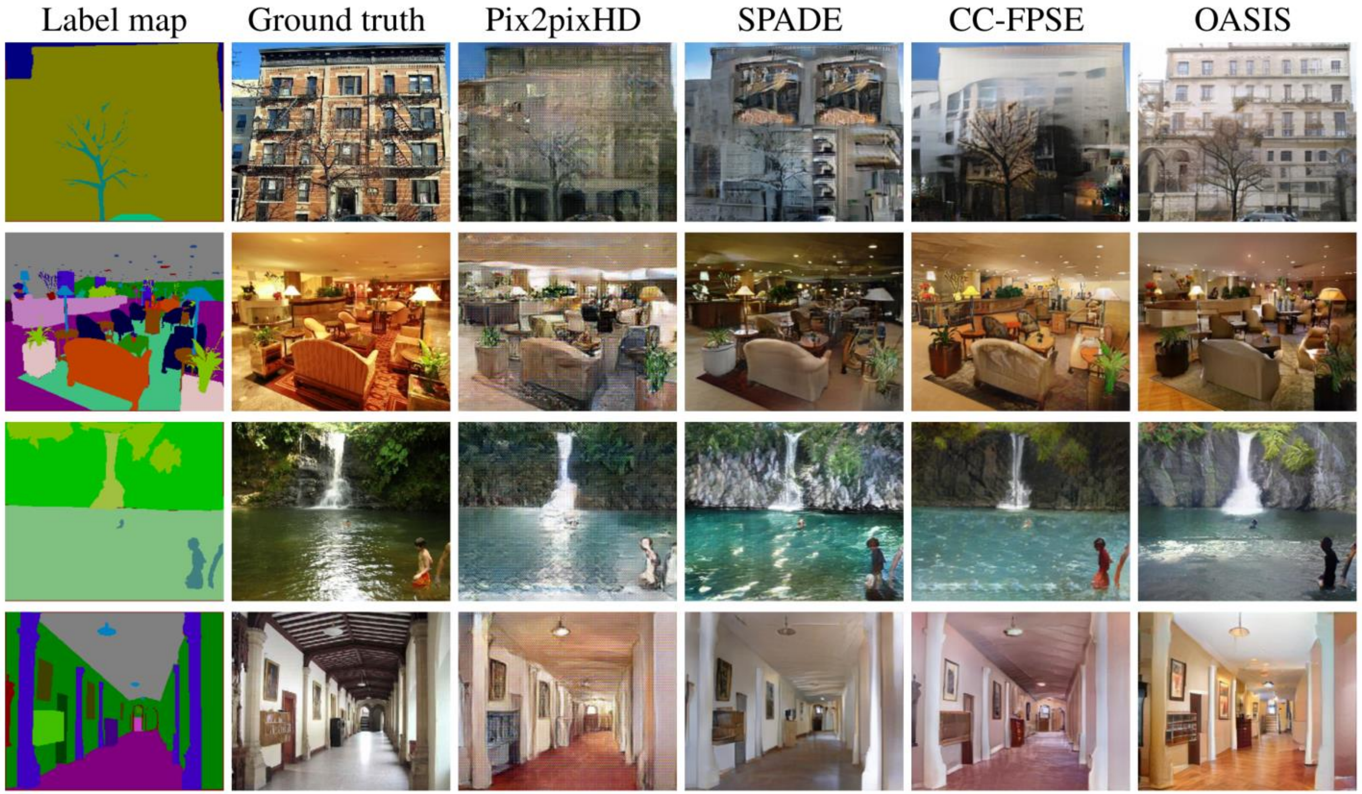


Figure 5: Qualitative comparison of OASIS with other methods on ADE20K. Trained with only adversarial supervision, our model generates images with better perceptual quality and structure.

Schönfeld, Edgar, et al. "You only need adversarial supervision for semantic image synthesis." *International Conference on Learning Representations*. 2021.

OASIS: Qualitative Results

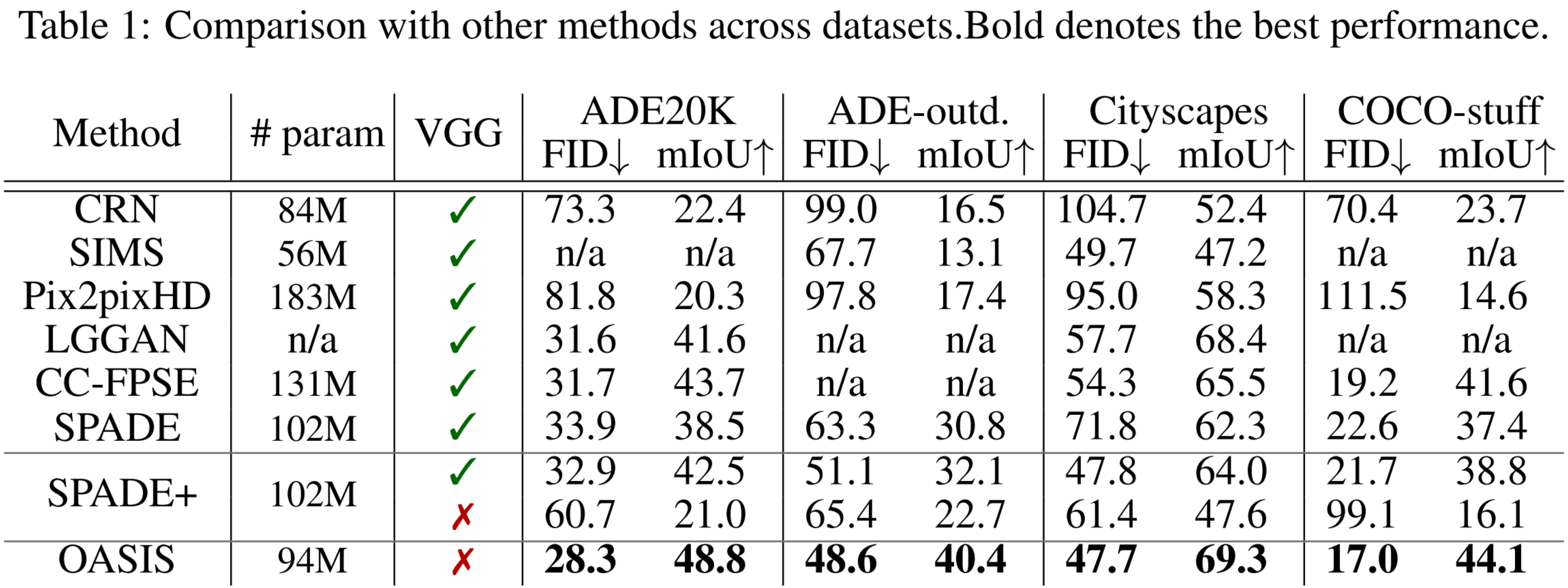


Table 1: Comparison with other methods across datasets.Bold denotes the best performance.

Method	# param	VGG	ADE20K		ADE-outd.		Cityscapes		COCO-stuff	
			FID↓	mIoU↑	FID↓	mIoU↑	FID↓	mIoU↑	FID↓	mIoU↑
CRN	84M	✓	73.3	22.4	99.0	16.5	104.7	52.4	70.4	23.7
SIMS	56M	✓	n/a	n/a	67.7	13.1	49.7	47.2	n/a	n/a
Pix2pixHD	183M	✓	81.8	20.3	97.8	17.4	95.0	58.3	111.5	14.6
LGGAN	n/a	✓	31.6	41.6	n/a	n/a	57.7	68.4	n/a	n/a
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	✗	**28.3**	**48.8**	**48.6**	**40.4**	**47.7**	**69.3**	**17.0**	**44.1**

Schönfeld, Edgar, et al. "You only need adversarial supervision for semantic image synthesis." *International Conference on Learning Representations*. 2021.