

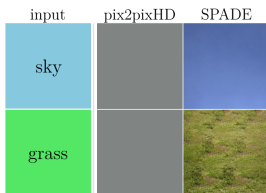
# Semantic Image Synthesis with Spatially-Adaptive Normalization

Taesung Park, Ming-Yu Liu, Ting-Chun Wang, Jun-Yan Zhu  
Presented by: Andrey Yashkin

Big Data AI  
Novosibirsk State University

October 20, 2020

# Introduction



**Figure:** An example of “washing away” semantic information with uniform segmentation maps

The task of generating photorealistic images conditioning on certain input data is an actual problem of machine learning. Recent methods directly learn how to compute the output image using neural networks directly feeding the semantic layout as input to the deep network, which is suboptimal as the normalization layers tend to “wash away” semantic information. Applying **SP**atially-**A**daptive (**DE**)normalization (**SPADE**) is shown to be a simple but effective way for synthesizing photorealistic images given an input semantic layout.

# Generative Adversarial Networks

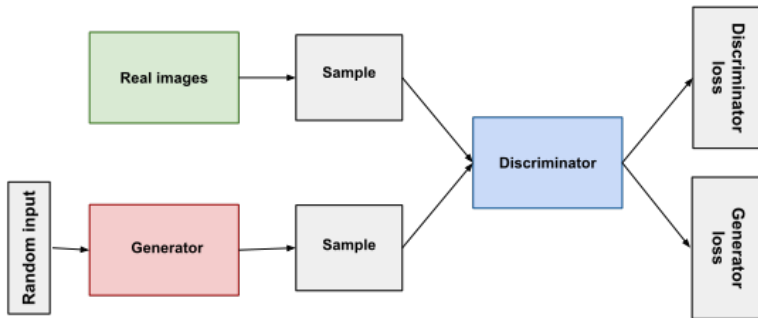


Figure: GAN structure

# Batch Normalization

**Input:** Values of  $x$  over a mini-batch:  $\mathcal{B} = \{x_{1\dots m}\}$ ;

Parameters to be learned:  $\gamma, \beta$

**Output:**  $\{y_i = \text{BN}_{\gamma, \beta}(x_i)\}$

$$\mu_{\mathcal{B}} \leftarrow \frac{1}{m} \sum_{i=1}^m x_i \quad // \text{ mini-batch mean}$$

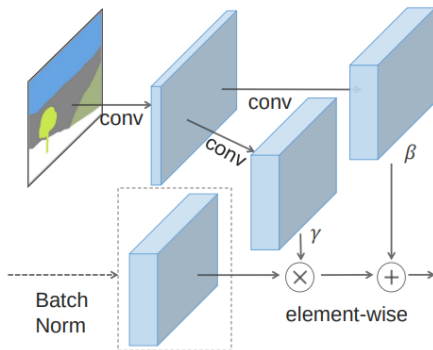
$$\sigma_{\mathcal{B}}^2 \leftarrow \frac{1}{m} \sum_{i=1}^m (x_i - \mu_{\mathcal{B}})^2 \quad // \text{ mini-batch variance}$$

$$\hat{x}_i \leftarrow \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}} \quad // \text{ normalize}$$

$$y_i \leftarrow \gamma \hat{x}_i + \beta \equiv \text{BN}_{\gamma, \beta}(x_i) \quad // \text{ scale and shift}$$

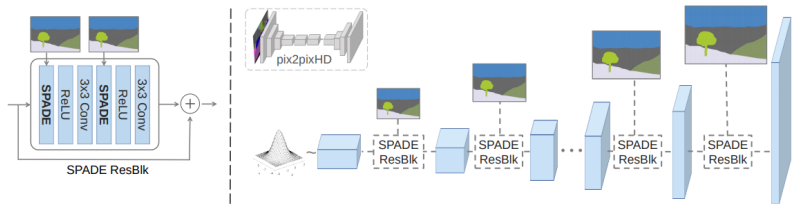
Figure: Equations behind batch normalization

# SPADE



**Figure:** In the SPADE, the mask is first projected onto an embedding space and then convolved to produce the modulation parameters  $\gamma$  and  $\beta$ . Unlike prior conditional normalization methods,  $\gamma$  and  $\beta$  are not vectors, but tensors with spatial dimensions.

# SPADE generator



**Figure:** 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.

# Selecting multivariate normal distribution

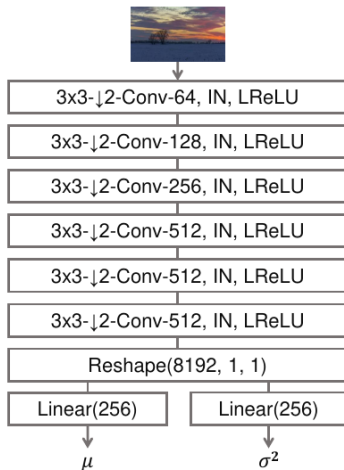


Figure: Image Encoder

# Discriminator

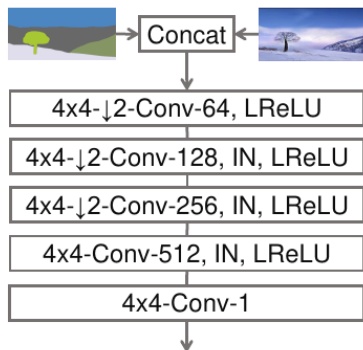


Figure: Patch-GAN based Discriminator



# Datasets

- ▶ COCO-Stuff is derived from the COCO dataset. It has 118,000 training images and 5000 validation images captured from diverse scenes. It has 182 semantic classes.
- ▶ ADE20K consists of 20210 training and 2000 validation images.
- ▶ Similarly to the COCO, the dataset contains challenging scenes with 150 semantic classes. ADE20K-outdoor is a subset of the ADE20K dataset that only contains outdoor scenes.
- ▶ Cityscapes dataset contains street scene images in German cities. The training and validation set sizes are 3000 and 500, respectively. Recent work has achieved photorealistic semantic image synthesis results on the Cityscapes dataset.

## Performance metrics

At first, we run state-of-the-art segmentation networks for each dataset. For measuring the segmentation accuracy, we use both the mean Intersection-over-Union (mIoU) and the pixel accuracy (accu).


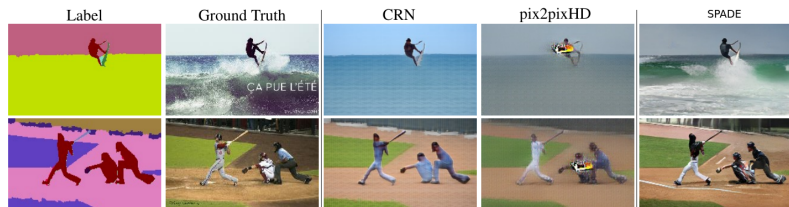
$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}}$$


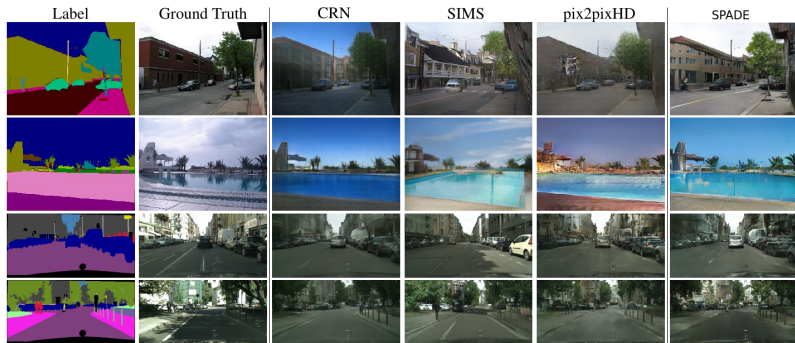
Figure: Intersection-over-Union

# COCO-Stuff



**Figure:** Visual comparison of semantic image synthesis results on the COCO-Stuff dataset. SPADE method successfully synthesizes realistic details from semantic labels.

# CADE20K



**Figure:** Visual comparison of semantic image synthesis results on the ADE20K outdoor and Cityscapes datasets. SPADE method produces realistic images while respecting the spatial semantic layout at the same time.

# Performance

| Method    | COCO-Stuff  |             |             | ADE20K      |             |             | ADE20K-outdoor |             |             | Cityscapes  |             |             |
|-----------|-------------|-------------|-------------|-------------|-------------|-------------|----------------|-------------|-------------|-------------|-------------|-------------|
|           | mIoU        | accu        | FID         | mIoU        | accu        | FID         | mIoU           | accu        | FID         | mIoU        | accu        | FID         |
| CRN       | 23.7        | 40.4        | 70.4        | 22.4        | 68.8        | 73.3        | 16.5           | 68.6        | 99.0        | 52.4        | 77.1        | 104.7       |
| SIMS      | N/A         | N/A         | N/A         | N/A         | N/A         | N/A         | 13.1           | 74.7        | 67.7        | 47.2        | 75.5        | <b>49.7</b> |
| pix2pixHD | 14.6        | 45.8        | 111.5       | 20.3        | 69.2        | 81.8        | 17.4           | 71.6        | 97.8        | 58.3        | 81.4        | 95.0        |
| SPADE     | <b>37.4</b> | <b>67.9</b> | <b>22.6</b> | <b>38.5</b> | <b>79.9</b> | <b>33.9</b> | <b>30.8</b>    | <b>82.9</b> | <b>63.3</b> | <b>62.3</b> | <b>81.9</b> | 71.8        |

**Figure:** SPADE outperforms the current leading methods in semantic segmentation (mIoU and accu) and FID <sup>2</sup> scores on all the benchmark datasets. For the mIoU and accu, higher is better. For the FID, lower is better.

---

<sup>1</sup>arXiv:1706.08500v6

<sup>2</sup>arXiv:1706.08500v6

## Human evaluation

| Dataset        | SPADE vs.<br>CRN | SPADE vs.<br>pix2pixHD | SPADE vs.<br>SIMS |
|----------------|------------------|------------------------|-------------------|
| COCO-Stuff     | 79.76            | 86.64                  | N/A               |
| ADE20K         | 76.66            | 83.74                  | N/A               |
| ADE20K-outdoor | 66.04            | 79.34                  | 85.70             |
| Cityscapes     | 63.60            | 53.64                  | 51.52             |

**Figure:** User preference study. The numbers indicate the percentage of users who favor the results of the proposed method over those of the competing method.



Park, Taesung and Liu, Ming-Yu and Wang, Ting-Chun and Zhu, Jun-Yan, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019

The end