Semantic Image Synthesis with Spatially-Adaptive Normalization

Taesung Park[1,2*] Ming-Yu Liu[2] Ting-Chun Wang[2] Jun-Yan Zhu[2,3]

[1]UC Berkeley [2]NVIDIA [2,3]MIT CSAIL

1. Introduction

- 본 논문은 semantic segmentation mask 를 사실적인 이미지로 변환하는 것에 관심이 있다.
- 본 논문에서 convolutional, normalization, nonlinearity layers 를 쌓아서 만든 전통적인 네트워크 구조[22, 48]가 최적의 방법이 아니라고 주장한다. 그 이유는Normalization layers 가 input semantic masks 에 포함된 정보를 없애버리기 때문이다.
- 이 문제를 해결하기 위해 spatially-adaptive normalization 을 제안한다.

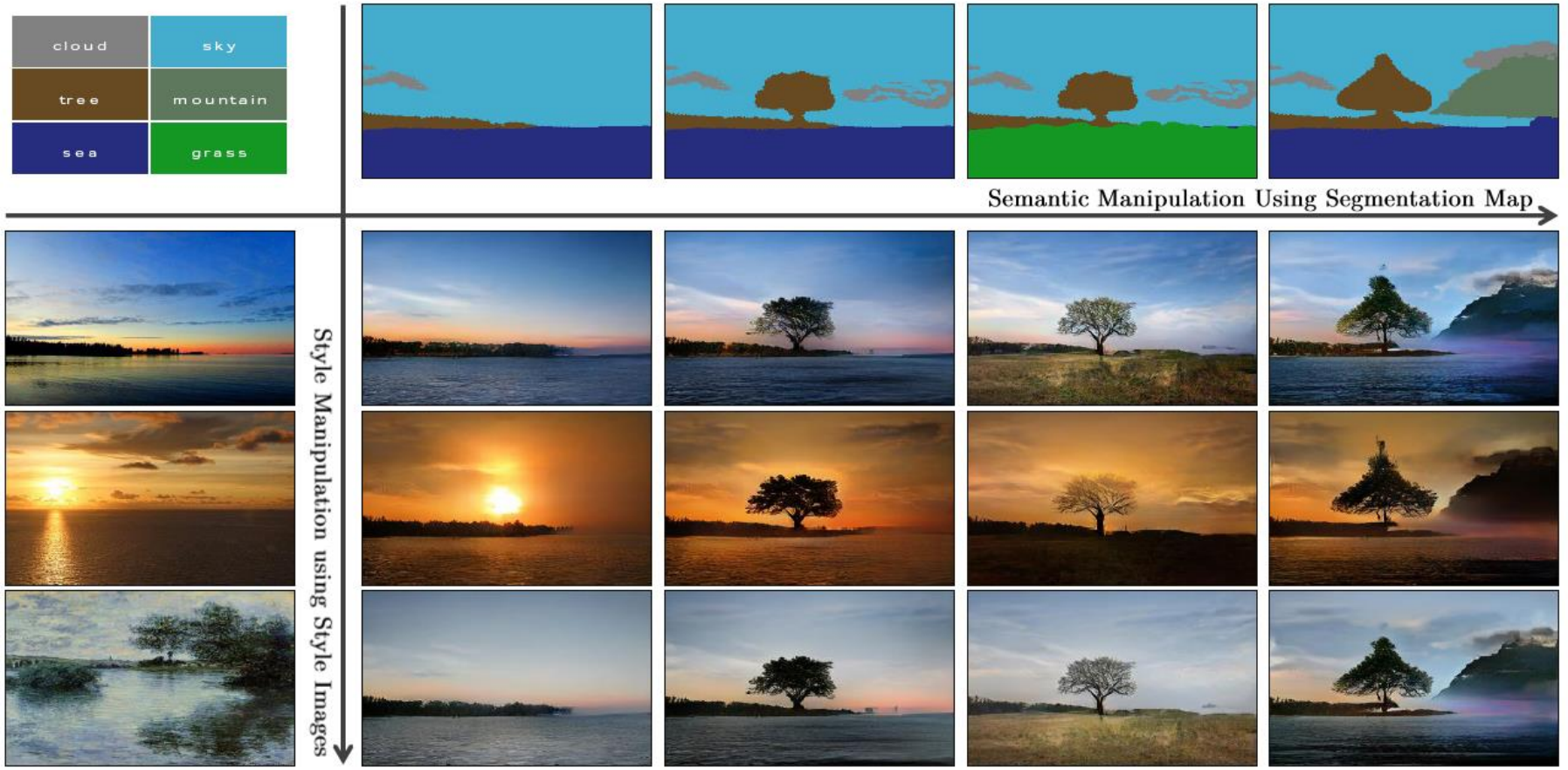

Figure 1: Our model allows user control over both semantic and style as synthesizing an image. The semantic (e.g., the existence of a tree) is controlled via a label map (the top row), while the style is controlled via the reference style image (the leftmost column). Please visit our website for interactive image synthesis demos.

본 논문의 모델은 사용자가 semantic 과 style 모두를 컨트롤 해서 이미지를 생성할 수 있게 한다. Semantic 은 label map (the top row) 을 통해서 컨트롤 되는 반면, style 은 reference style image (the leftmost column) 을 통해서 컨트롤 된다.

2. Related Work

- **Deep generative models**
  - 이미지 합성 모델
  - GAN, VAE
- 본 논문의 연구는 GANs 에 기반하지만 conditional image synthesis task 를 목표로 한다.

- **Conditional image synthesis** 는 입력 데이터의 종류에 따라 나뉜다.
  - Category labels 에 따라 이미지를 합성하는 class-conditional 모델
  - Text 에 기반하여 이미지를 생성하는 모델
  - Conditional GANs 에 기반한 image-to-image translation
- 본 논문은 training dataset 이 segmentation masks 와 images 를 포함하는 것을 가정한다.

- **Unconditional normalization layers**
  - Batch Normalization (BatchNorm), Instance Normalization (InstanceNorm), Layer Normalization, Group Normalization, Weight Normalization
  - 외부 데이터에 의존하지 않기 때문에 무조건 레이블을 지정해야 한다.

- **Conditional normalization layers**
  - Conditional Batch Normalization (Conditional BatchNorm), Adaptive Instance Normalization (AdaIN)
  - conditional normalization layers 는 추가적인 데이터를 요구한다.
  - 일반적인 conditional normalization layers 는 다음과 같이 동작한다.
    - 먼저, layer activations 는 zero mean 과 unit deviation 으로 normalized 된다.
    - 그런 다음 normalized activations 는 외부 데이터로부터 만들어진 파라미터들을 가진 학습된 affine transformation 을 사용하여 activation 을 조절한다.

- 대조적으로, 본 논문의 normalization layer 는 spatially-varing affine transformation 을 적용하고, semantic masks 로부터 이미지를 합성하는 것을 더 안정적으로 한다.
- 본 논문은 normalized activations 를 조절하는 맥락에서 semantic information 을 제공하는 데 중점을 둔다.
- 그리고 다양한 크기의 semantic maps 를 사용하고, 그것은 coarse-to-fine generation 을 가능하게 한다.

3. Semantic Image Synthesis

- $\mathbf{m} \in \mathbb{L}^{H \times W}$
  - $\mathbf{m}$: semantic segmentation mask
  - $\mathbb{L}$: semantic labels 를 표시하는 정수들의 집합
  - H : 이미지의 높이, W : 이미지의 너비
- 본 논문의 목표는 input segmentation mask $\mathbf{m}$ 을 photorealistic image 로 변환하는 mapping function 을 학습하는 것이다.

- **Spatially-adaptive denormalization.**
- h^i : N 개의 samples 이 존재 하는 한 개의 batch 에 대해 deep convolutional network 의 $i-th$ 레이이어의 activations
- C^i : 레이어에서 channels 의 수
- H^i 와 W^i : 레이어에서 activation map 의 높이와 너비
- 본 논문은 새로운 conditional normalization 방법인 SPatially-Adaptive(DE)normalization (SPADE) 를 제안한다.
- Batch Normalization 과 유사하게, activation 은 channel wise 방식으로 normalized 되고 학습된 scale 과 bias 로 조절된다.

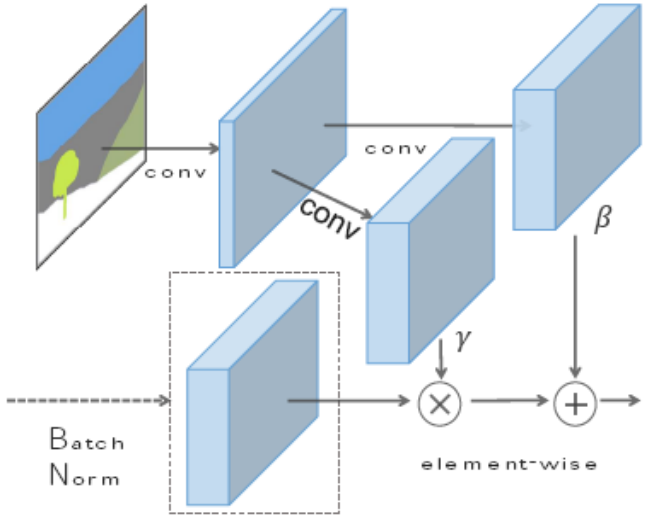

Figure 2: In the SPADE, the mask is first projected onto an embedding space and then convolved to produce the modulation parameters γ and β. Unlike prior conditional normalization methods, γ and β are not vectors, but tensors with spatial dimensions. The produced γ and β are multiplied and added to the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

- Figure 2: SPADE 에서, mask 는 가장 먼저 embedding space 로 project 되고 그런 다음 conv 를 통해 modulation parameters γ 와 β 를 만든다.
- 이전의 conditional normalization 방법들과는 달리, γ 와 β 는 vector 가 아니고 spatial dimension 을 가진 tensor 이다.
- γ 와 β 는 normalized activation 과 element-wise 로 곱해지고 더해진다.

- site $(n \in N, c \in C^i, y \in H^i, x \in W^i)$ 에서 activation value 는 다음과 같다.

$$\gamma^i_{c,y,x}(\mathbf{m}) \frac{h^i_{n,c,y,x} - \mu^i_c}{\sigma^i_c} + \beta^i_{c,y,x}(\mathbf{m}) \qquad (1)$$

- $h^i_{n,c,y,x}$ 는 normalization 전 site 에서 activation 이고 μ^i_c 와 σ^i_c 는 channel c 에서 activations 의 평균과 표준편차이다.

$$\mu^i_c = \frac{1}{NH^iW^i} \sum_{n,y,x} h^i_{n,c,y,x} \qquad (2)$$

$$\sigma^i_c = \sqrt{\frac{1}{NH^iW^i} \sum_{n,y,x} \left((h^i_{n,c,y,x})^2 - (\mu^i_c)^2 \right)}. \qquad (3)$$

- (1) 의 변수 $\gamma^i_{c,y,x}(\mathbf{m})$ 과 $\beta^i_{c,y,x}(\mathbf{m})$ 는 normalization layer 의 학습된 modulation parameters 이다. BatchNorm 과는 대조적으로, γ 와 β 는 input segmentation mask 에 의존하고 위치 (y, x) 에 따라 다양하다.
- $\gamma^i_{c,y,x}(\mathbf{m})$ 과 $\beta^i_{c,y,x}(\mathbf{m})$ 를 $i-th$ activation map 의 site (c, y, x) 에서 $\mathbf{m}$ 을 scaling and bias values 로 변환하는 함수를 나타내는 데에 사용한다.

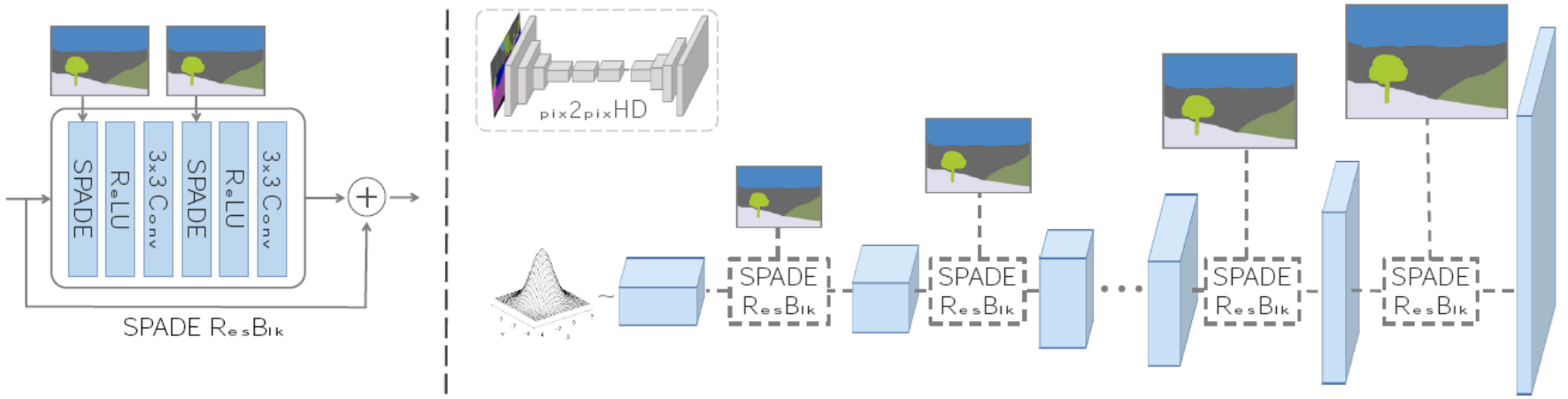

Figure 4: In the SPADE generator, each normalization layer uses the segmentation mask to modulate the layer activations. *(left)* Structure of one residual block with the SPADE. *(right)* The generator contains a series of the SPADE residual blocks with upsampling layers. Our architecture achieves better 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].

- **SPADE generator:**
- SPADE generator 에서, 각 normalization layer 는 layer activations 를 modulated 하는 데에 segmentation mask 를 사용한다.
- (left) SPADE residual block 의 구조
- (right) generator 는 upsampling layers 와 함께 일련의 SPADE residual blocks 를 포함한다.
- 본 논문의 구조는 pix2pixHD 모델과 같은 image-to-image translation 의 downsampling 계층을 제거하여 더 적은 수의 파라미터로 더 나은 성능을 낼 수 있다.

- 본 논문은 least squared loss term 을 hinge loss term 으로 대체한다는 점을 제외하고는 pix2pixHD [48] 에서 사용된 것과 동일한 multi-scale discriminator 와 loss function 을 사용하여 generator 를 훈련한다.

- **왜 SPADE 가 더 잘 동작할까?**
- 그 이유는 common normalization layers 보다 semantic 정보를 더 잘 보존하기 때문이다.

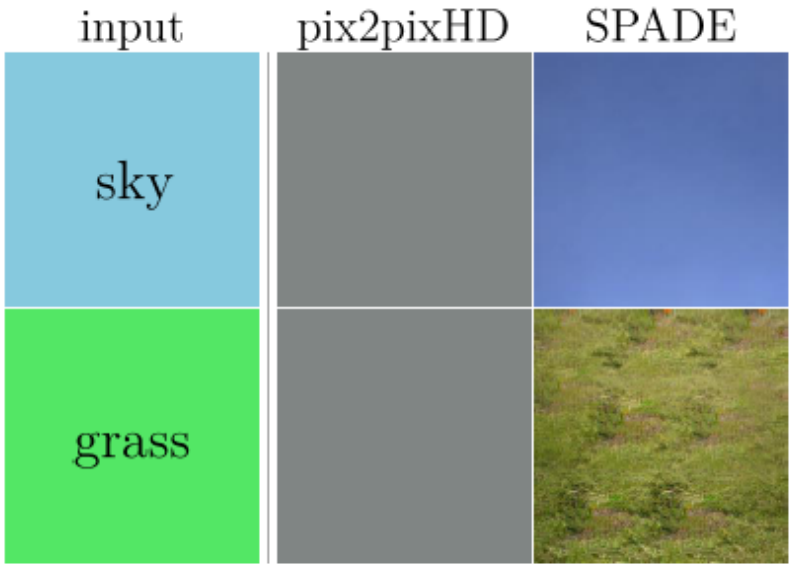

Figure 3: Comparing results given uniform segmentation maps: while the SPADE generator produces plausible textures, the pix2pixHD generator [48] produces two identical outputs due to the loss of the semantic information after the normalization layer.

주어진 uniform segmentation maps 를 비교해보자: SPADE generator 가 plausible textures 를 만들지만, pix2pixHD generator 는 normalization layer 후에 semantic 정보를 잃어버렸기 때문에 두 개의 동일한 결과를 만들었다.

- SPADE Generator 에서 segmentation mask 는 normalization 없이 spatially adaptive modulation 을 통과한다. 오직 이전 레이어로부터의 activations 가 normalized 된다.
- 그러므로, SPADE generator 는 semantic information 을 더 잘 보존할 수 있다.
- 이것은 semantic input information 을 잃어버리지 않고 normalization 의 이득을 취한다.

- **Multi-modal synthesis**

- Generator 의 입력으로 random vector 를 사용해서 multi-modal synthesis 을 할 수 있다.

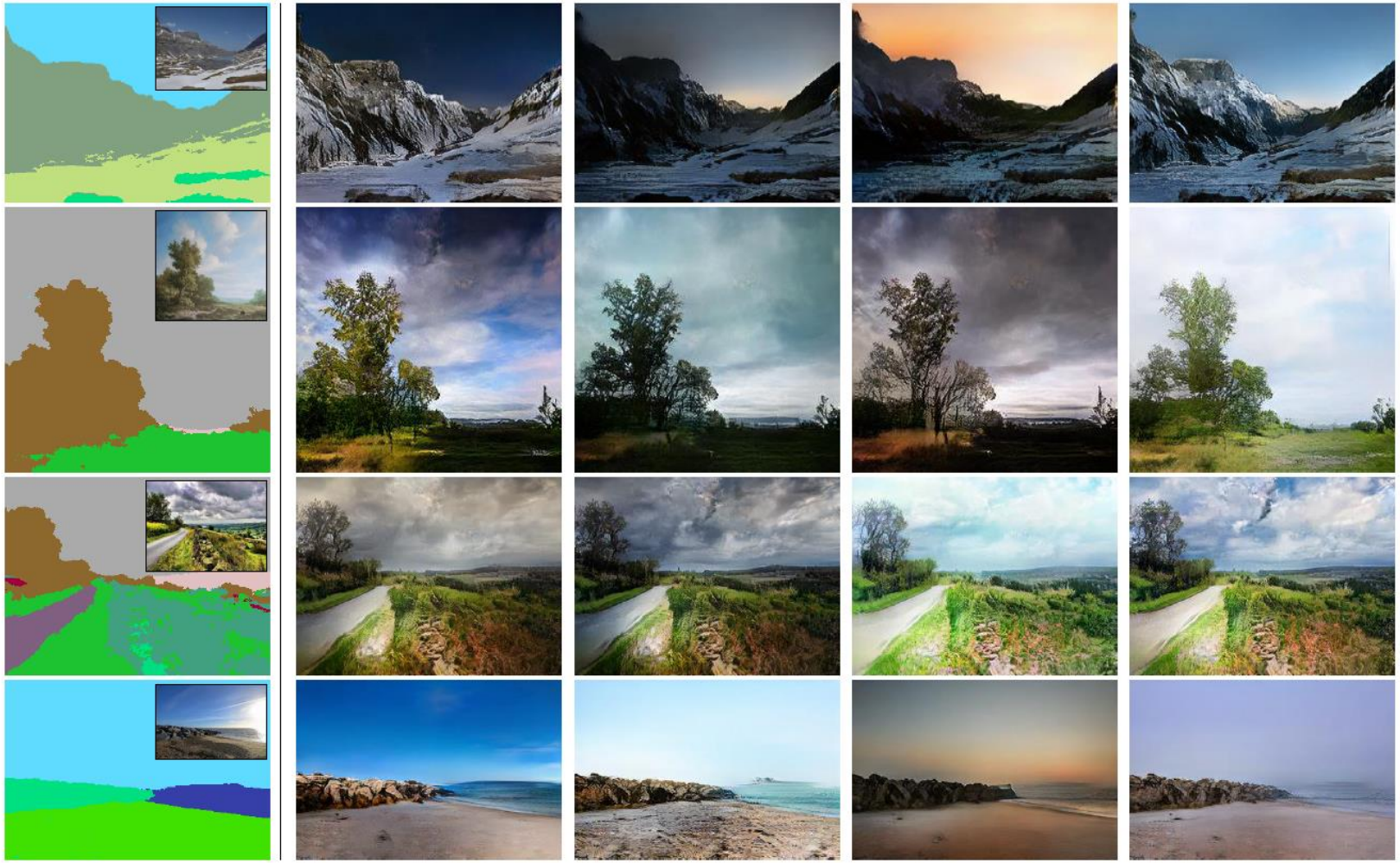

Figure 9: Our model attains multimodal synthesis capability when trained with the image encoder. During deployment, by using different random noise, our model synthesizes outputs with diverse appearances but all having the same semantic layouts depicted in the input mask. For reference, the ground truth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