You Only Need Adversarial Supervision for Semantic Image Synthesis

ICLR 2021

Edgar Schönfeld[1*], Vadim Sushko[1*], Dan Zhang[1], Jürgen Gall[2], Bernt Schiele[3], Anna Khoreva[1]

*Equal contribution

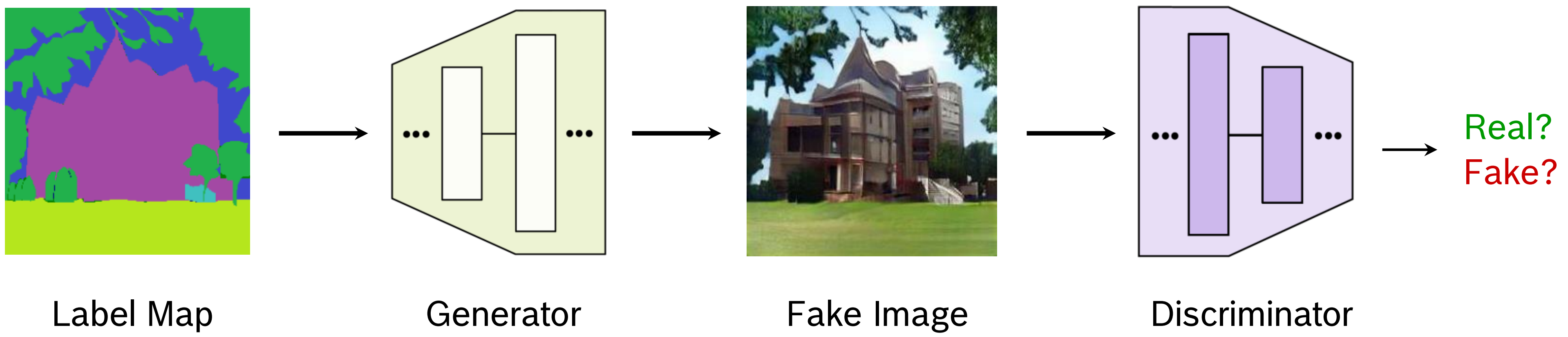
[1]Bosch Center for Artificial Intelligence
[2]University of Bonn
[3]Max Planck Institute for Informatics

BOSCH

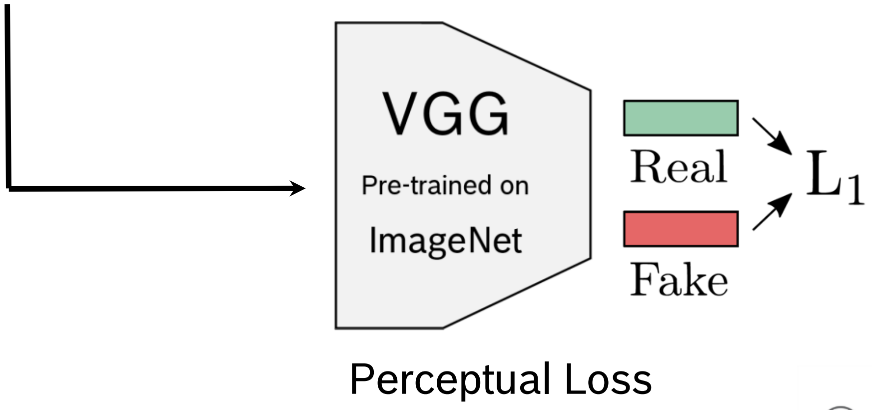
Semantic image synthesis
Problem statement

Our goal: multi-modal photorealistic image generation in alignment with a given semantic label map

Limitations of previous GAN methods
Perceptual loss



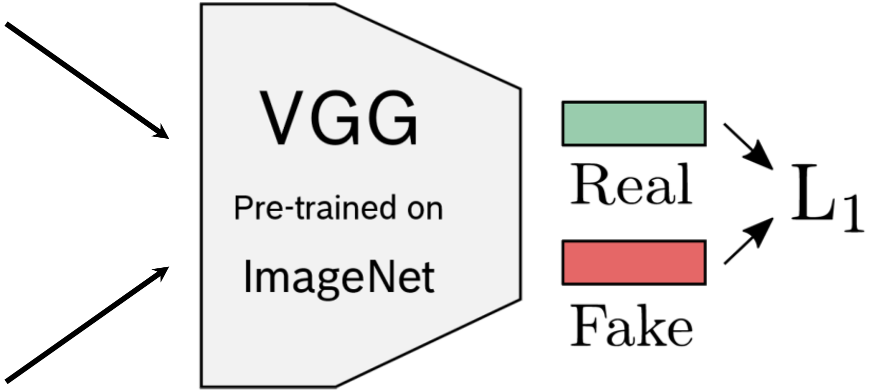
Supervision from discriminator is not enough!



Limitations of previous GAN methods
Perceptual loss

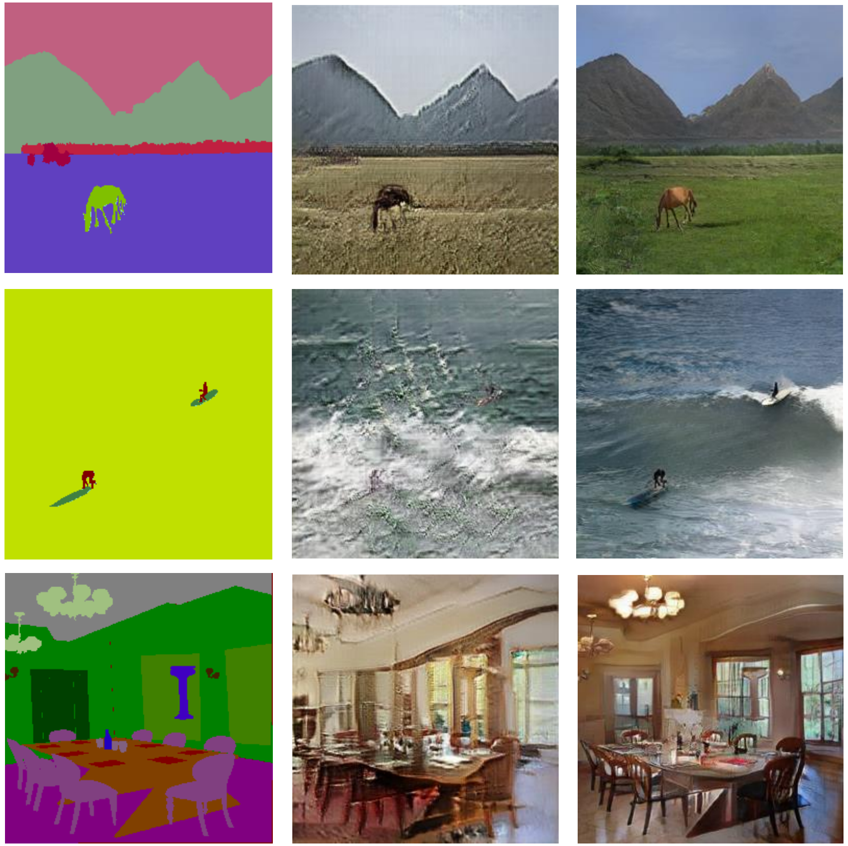
Real Image

Fake Image



Used in:
CRN [Chen et. al, 2017]
Pix2pixHD [Wang et. al, 2018]
SPADE [Park et. al, 2019]
CC-FPSE [Liu et. al, 2020]

Label map | W/o VGG | With VGG

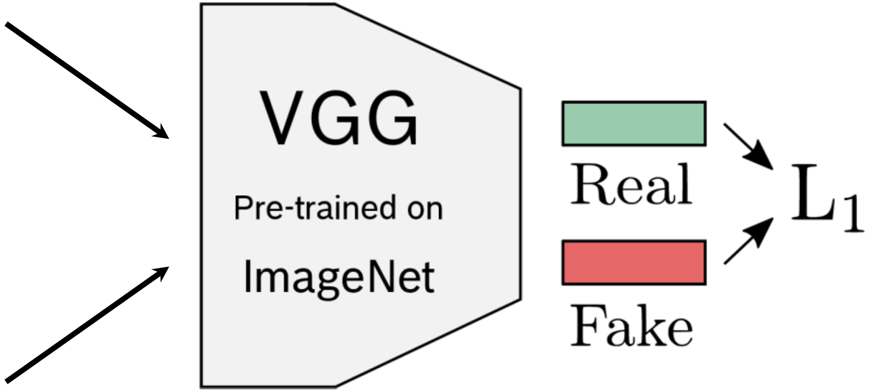


BOSCH

Limitations of previous GAN methods
Perceptual loss

Real Image

Fake Image



Effect of the perceptual loss:

- Stabilized training
- Improved quality of images

Drawbacks:

- Computational overhead
- Texture and color bias
- Constrained diversity

How to achieve high quality without the perceptual loss?

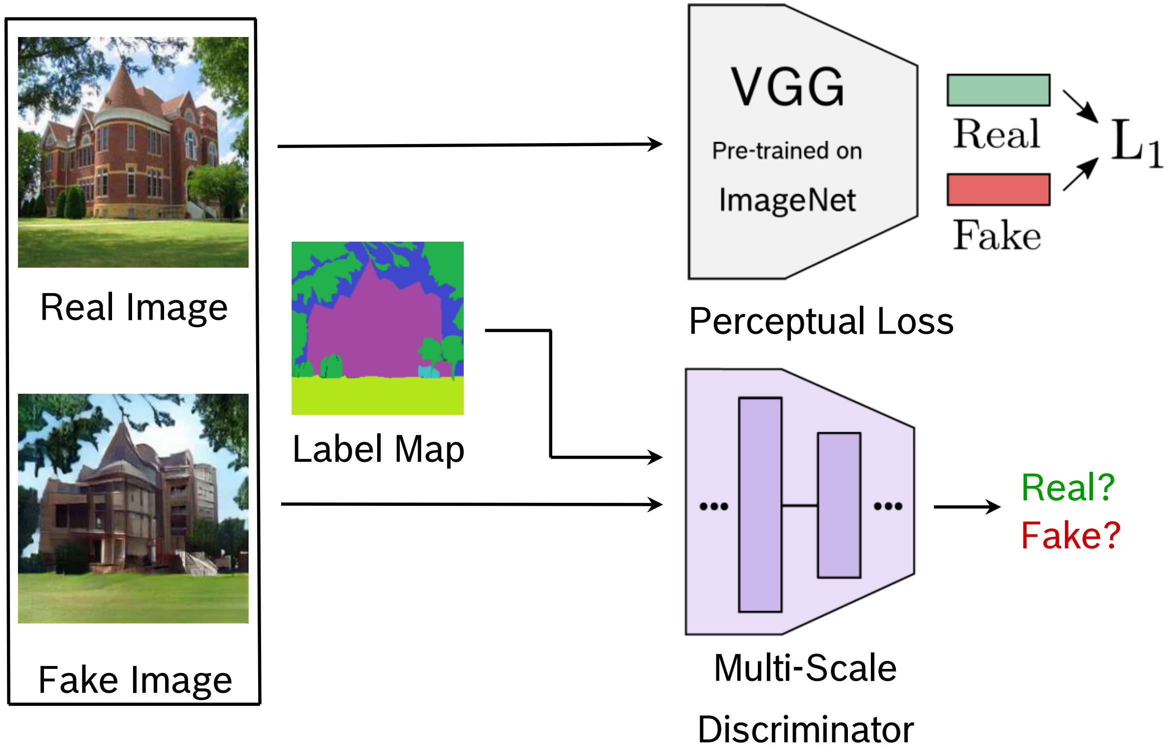
BOSCH

OASIS model
Segmentation-based discriminator

Baseline: SPADE [Park et al., 2019]

Our solution:

- Label map is a target, not input

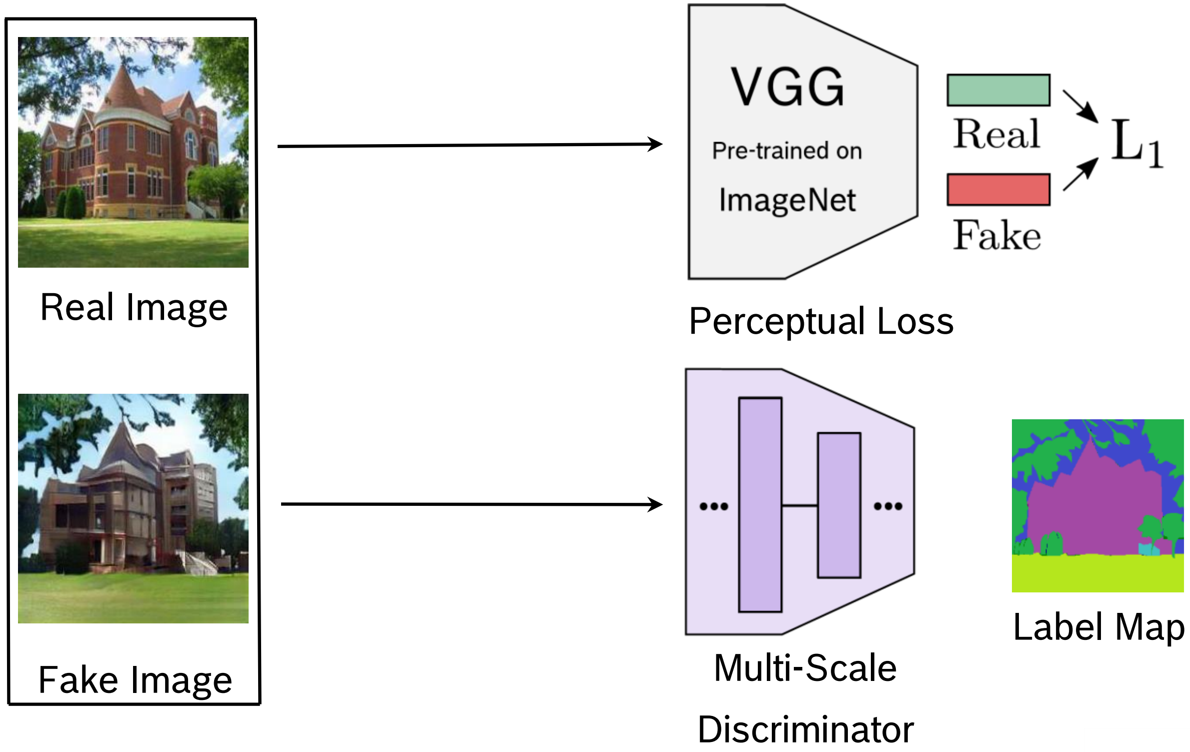


OASIS model
Segmentation-based discriminator

Baseline: SPADE [Park et al., 2019]

Our solution:

- ▶ Label map is a target, not input

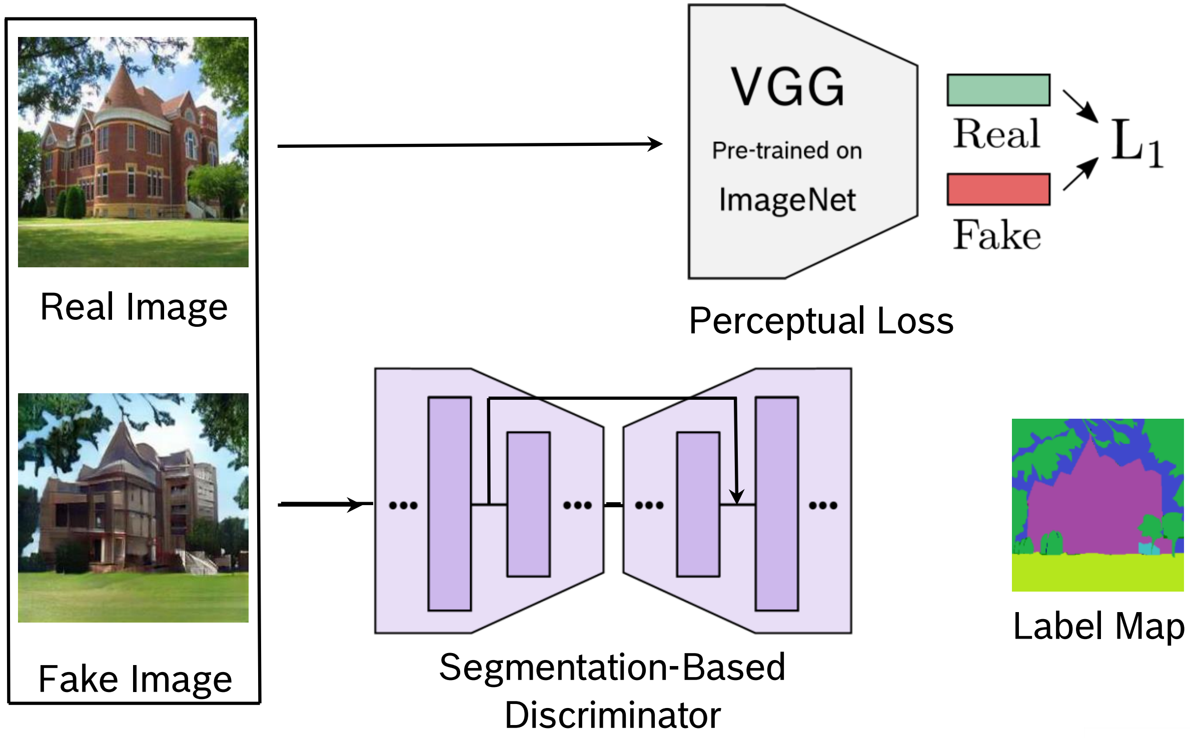


OASIS model
Segmentation-based discriminator

Baseline: SPADE [Park et al., 2019]

Our solution:

- Label map is a target, not input
- Discriminator's task = segmentation



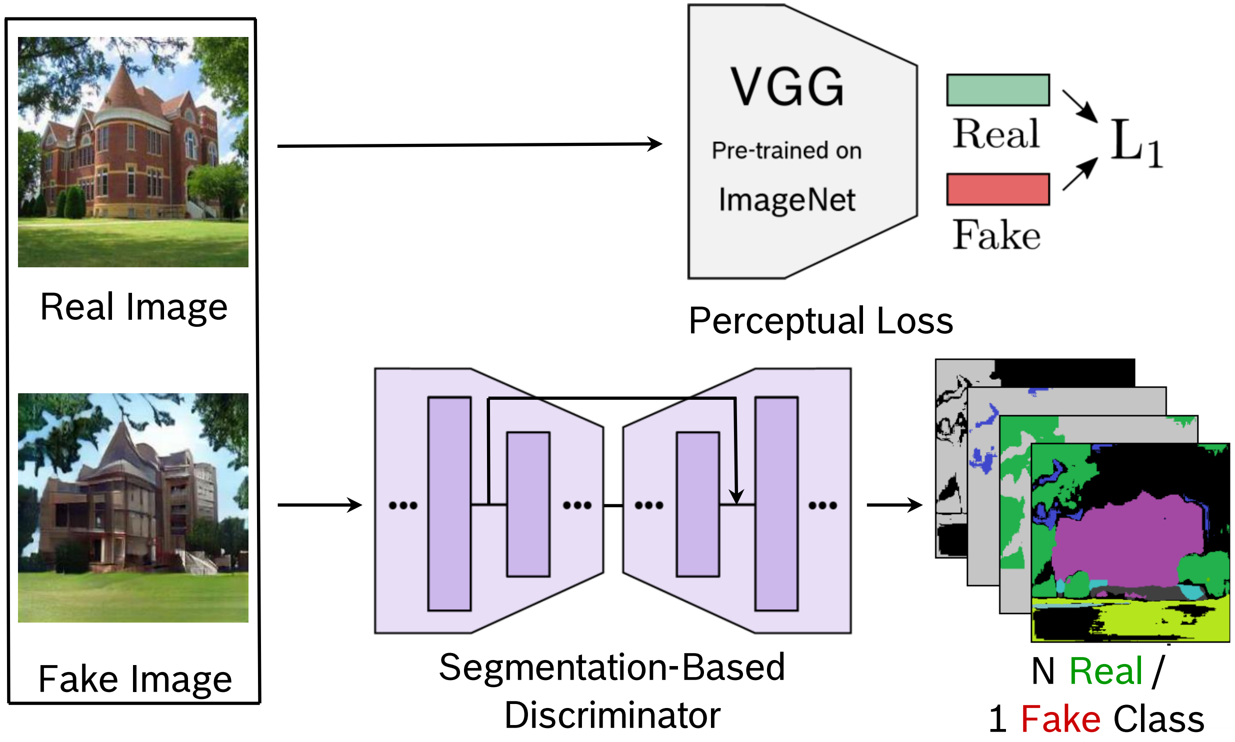
BOSCH

OASIS model
Segmentation-based discriminator

Baseline: SPADE [Park et al., 2019]

Our solution:

- Label map is a target, not input
- Discriminator's task = segmentation
- N+1 loss = adversarial loss



BOSCH

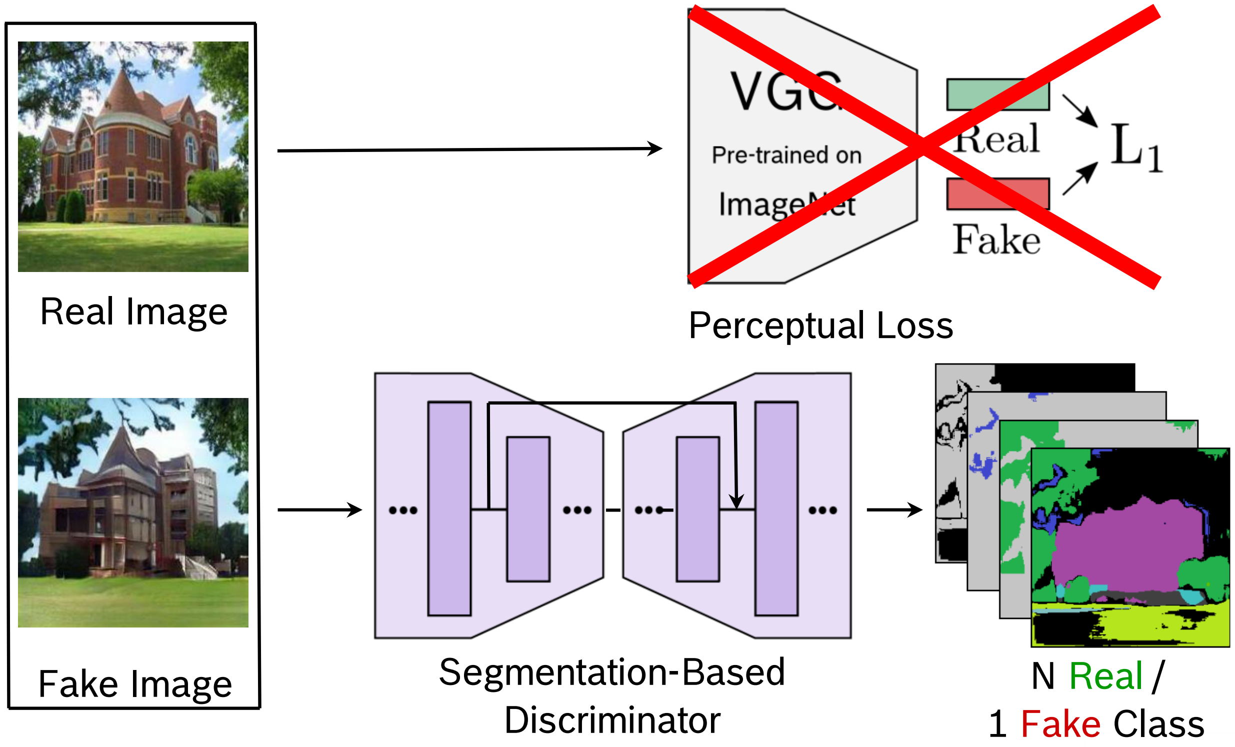
OASIS model
Segmentation-based discriminator

Baseline: SPADE [Park et al., 2019]

Our solution:

- Label map is a target, not input
- Discriminator’s task = segmentation
- N+1 loss = adversarial loss

VGG loss becomes unnecessary!



OASIS model
Segmentation-based discriminator

Baseline: SPADE [Park et al., 2019]

Our solution:

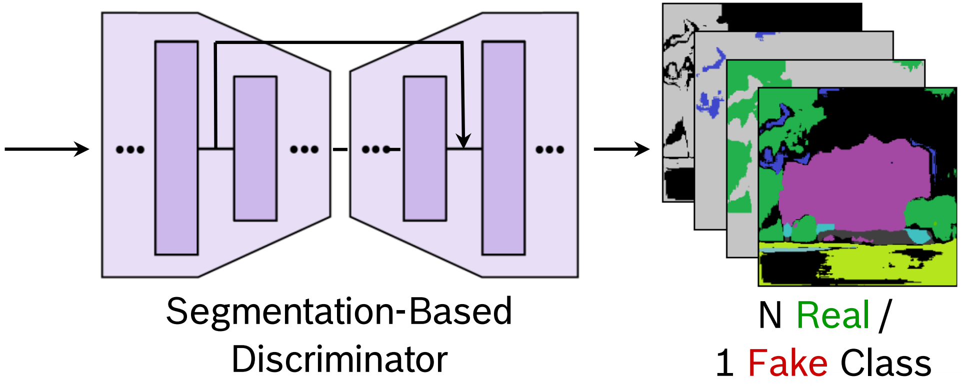
- Label map is a target, not input
- Discriminator's task = segmentation
- N+1 loss = adversarial loss

VGG loss becomes unnecessary!

D architecture	w/o VGG FID↓	w/o VGG mIoU↑	with VGG FID↓	with VGG mIoU↑
SPADE	60.7	21.0	32.9	42.5
OASIS	**29.3**	**51.6**	**29.2**	**51.1**

OASIS

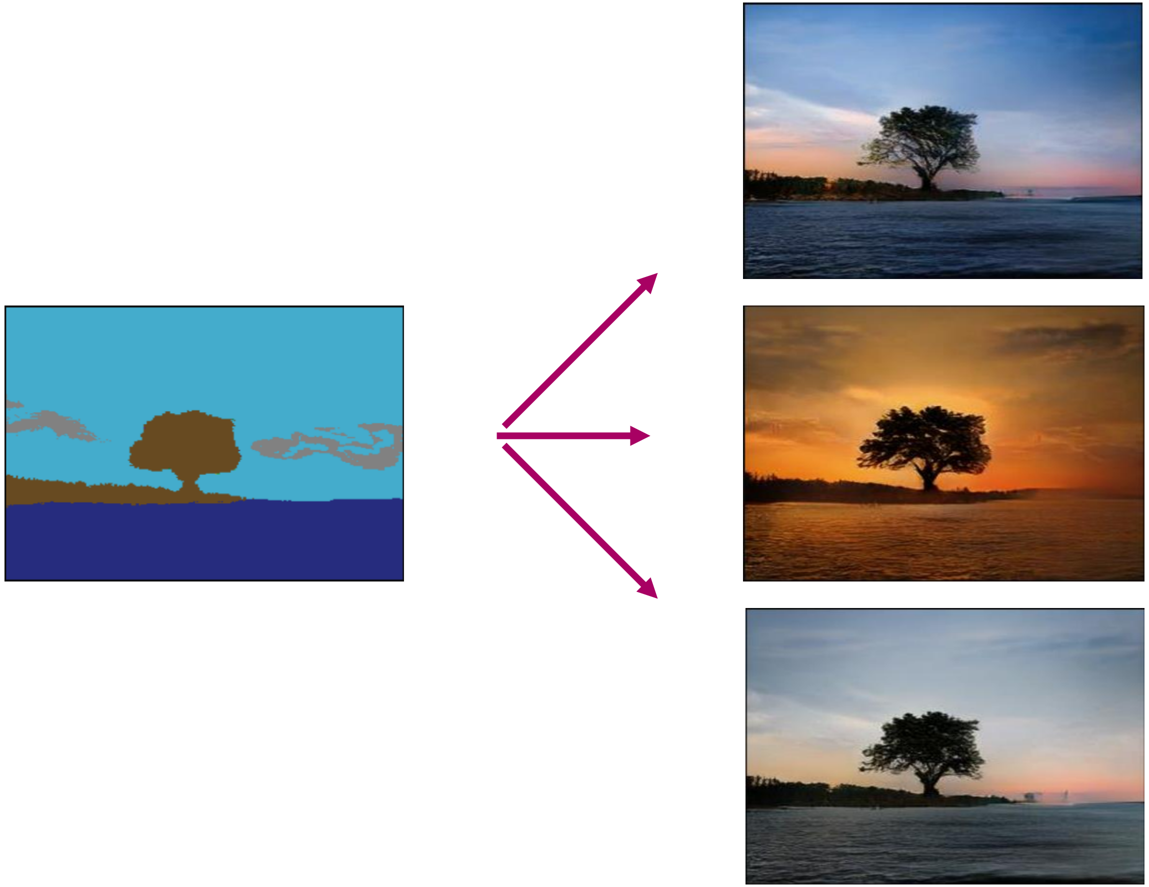
Only **A**dversarial supervision
for **S**emantic **I**mage **S**ynthesis



Results
Comparison to prior art

Label map	Ground truth	SPADE [Park et. al, 2019]	CC-FPSE [Liu et. al, 2020]	OASIS

Semantic image synthesis
Multi-modality

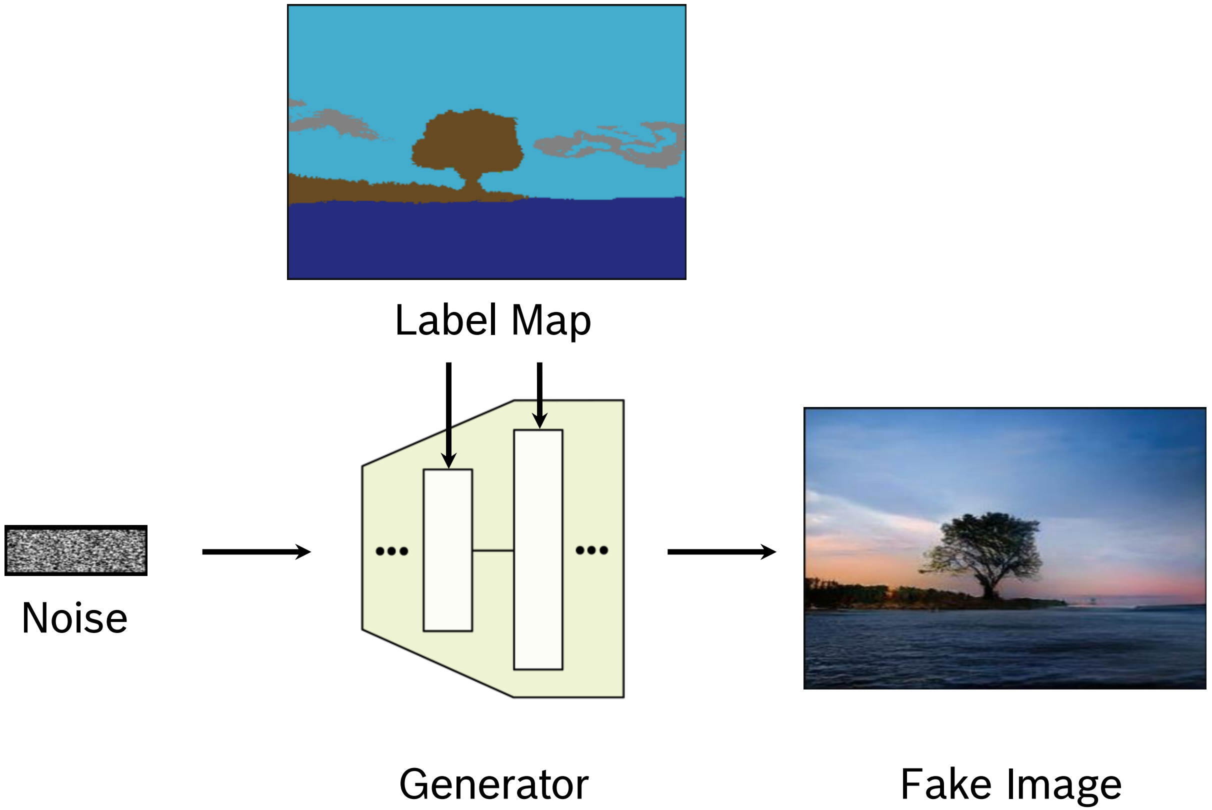


Images from [Park et. al, 2019]

Problems of previous GAN methods
Limited diversity

Noise is ignored!

First reported in:
Pix2pix [Isola et. al, 2017]



How to achieve high diversity through noise sampling?

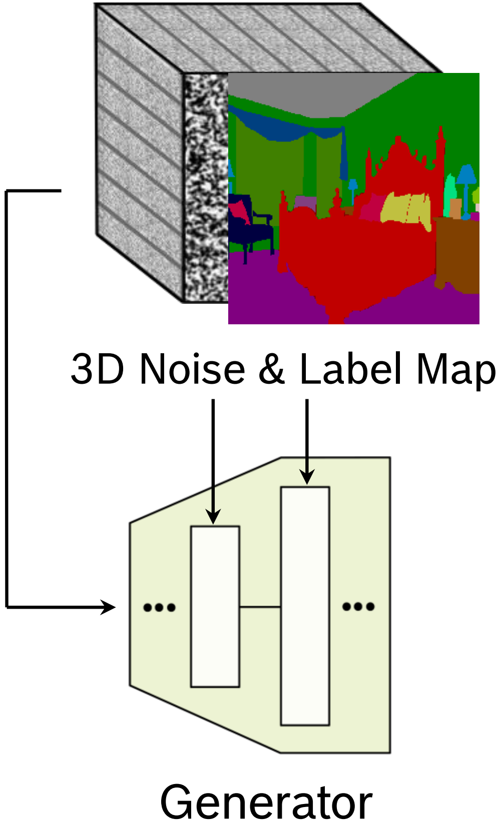
OASIS model
3D noise injection

Step 1: Create a composite tensor

1. Sample a 3D noise tensor
2. Concatenate 3D noise with the (3D) label map

Step 2: Inject the 3D composite tensor

1. Input to 1st generator layer
2. Input at *every* generator layer via the *spatially-adaptive* norm ("SPADE" layer)



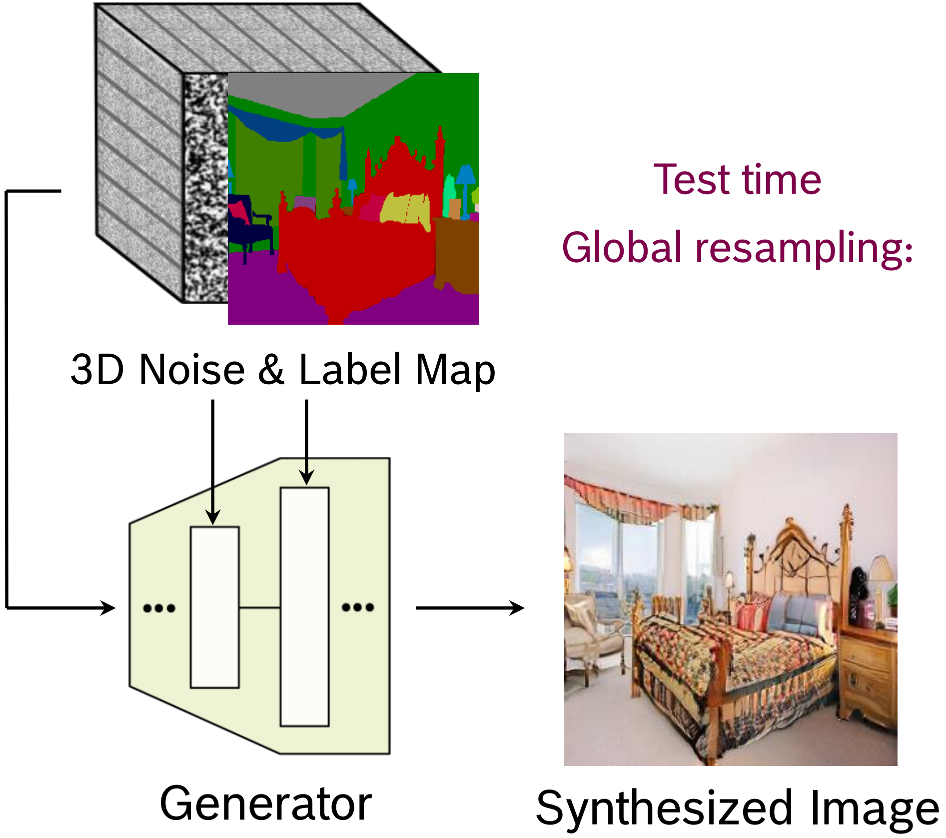
OASIS model
3D noise injection

<u>Step 1: Create a composite tensor</u>

1. Sample a 3D noise tensor
2. Concatenate 3D noise with the (3D) label map

<u>Step 2: Inject the 3D composite tensor</u>

1. Input to 1st generator layer
2. Input at *every* generator layer via the *spatially-adaptive* norm ("SPADE" layer)



Test time

Global resampling:

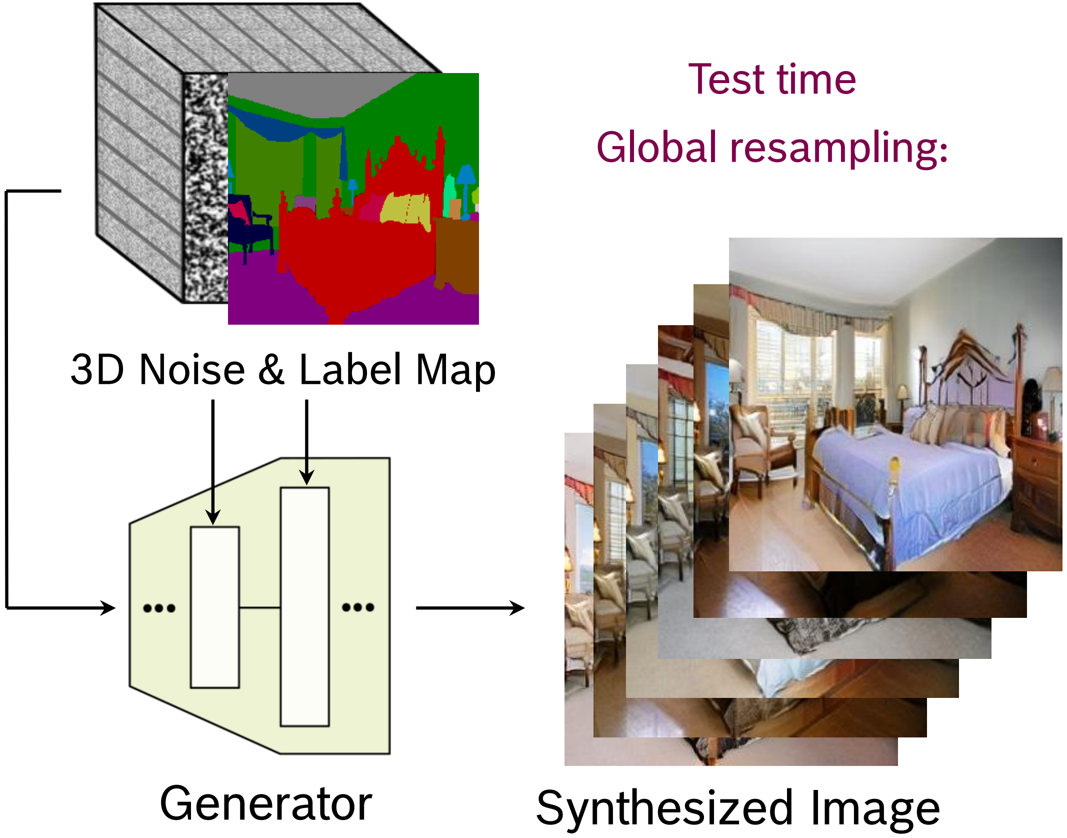
OASIS model
3D noise injection

<u>Step 1: Create a composite tensor</u>

1. Sample a 3D noise tensor
2. Concatenate 3D noise with the (3D) label map

<u>Step 2: Inject the 3D composite tensor</u>

1. Input to 1st generator layer
2. Input at *every* generator layer via the *spatially-adaptive* norm (“SPADE” layer)

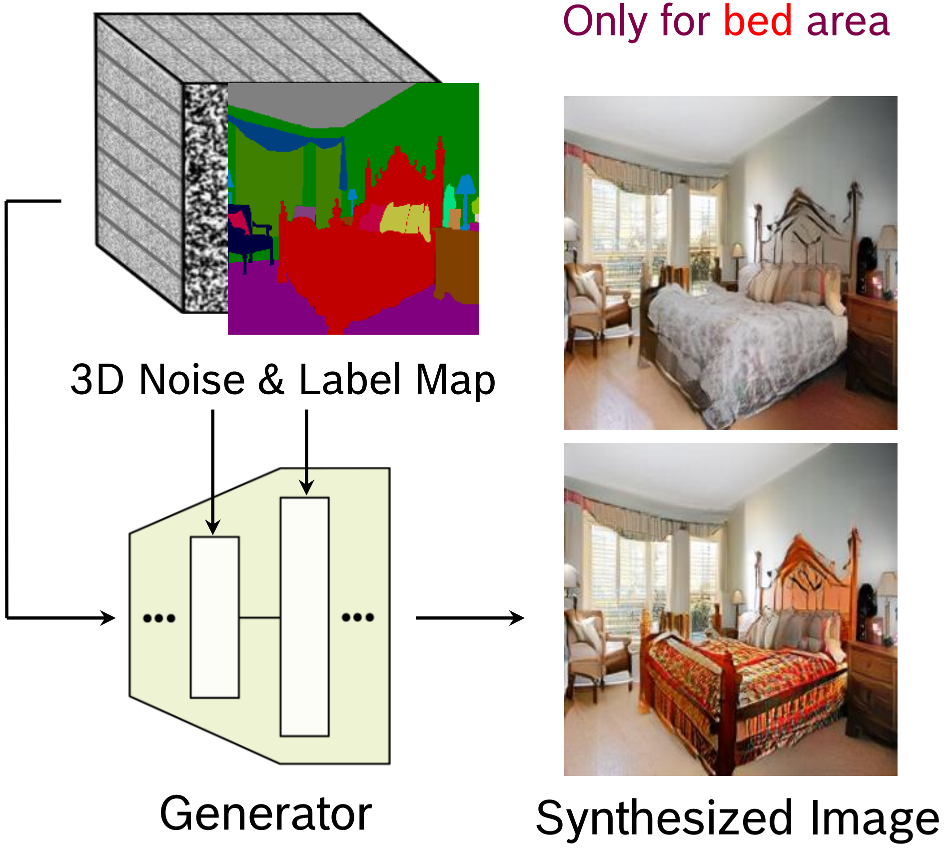


BOSCH

OASIS model
3D noise injection

Test time

Local resampling:
Only for bed area



<u>Step 1: Create a composite tensor</u>

1. Sample a 3D noise tensor
2. Concatenate 3D noise with the (3D) label map

<u>Step 2: Inject the 3D composite tensor</u>

1. Input to 1st generator layer
2. Input at *every* generator layer via the *spatially-adaptive* norm (“SPADE” layer)

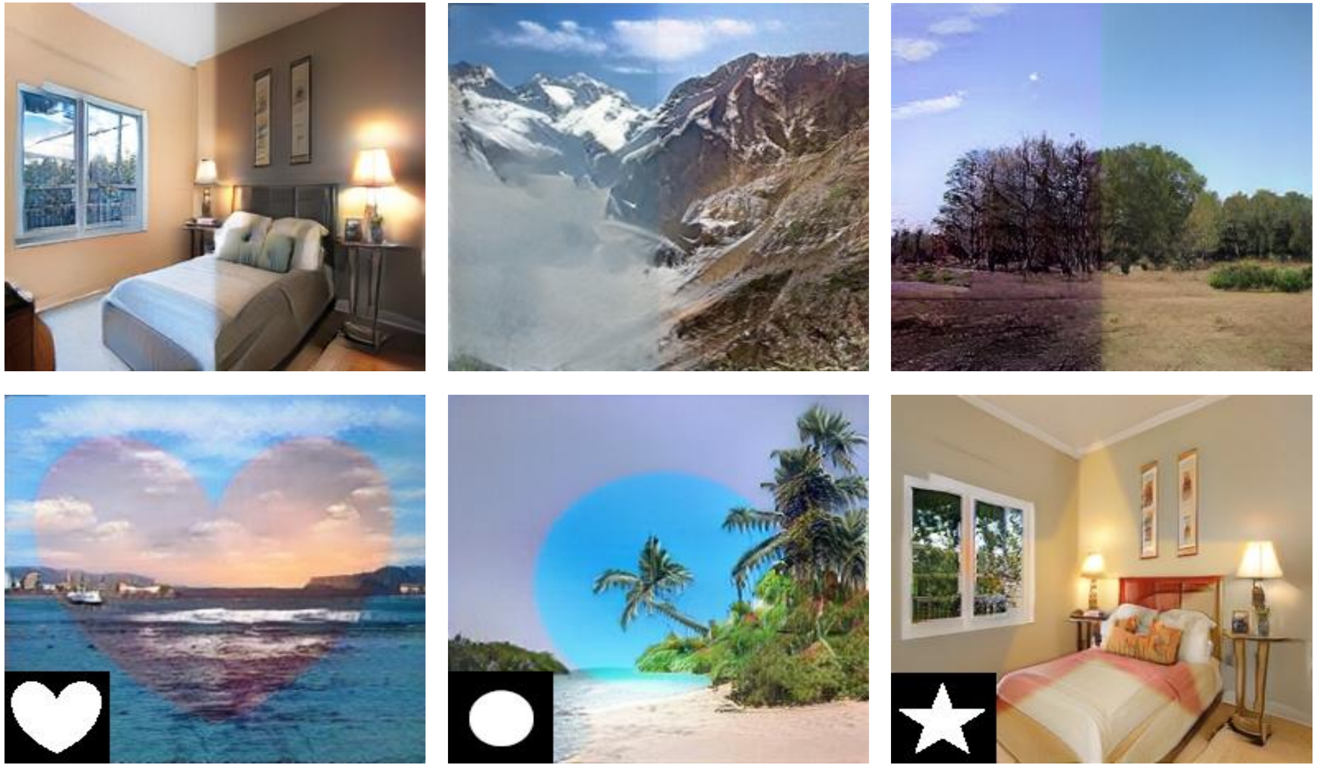
Results
Multi-modal generation

Global

Local

Results
Multi-modal generation

Local
(shape)



Summary
Our contributions

1. New state of the art model
2. Segmentation-based discriminator with an N+1 adversarial loss
3. 3D noise injection scheme

Label map | With VGG loss | W/o VGG loss | W/o VGG loss, sampled with different 3D noise

SPADE [Park et al., 2019] | OASIS (our model)

Thank you!

Paper: https://arxiv.org/abs/2012.04781

Code: https://github.com/boschresearch/OASIS

BOSCH