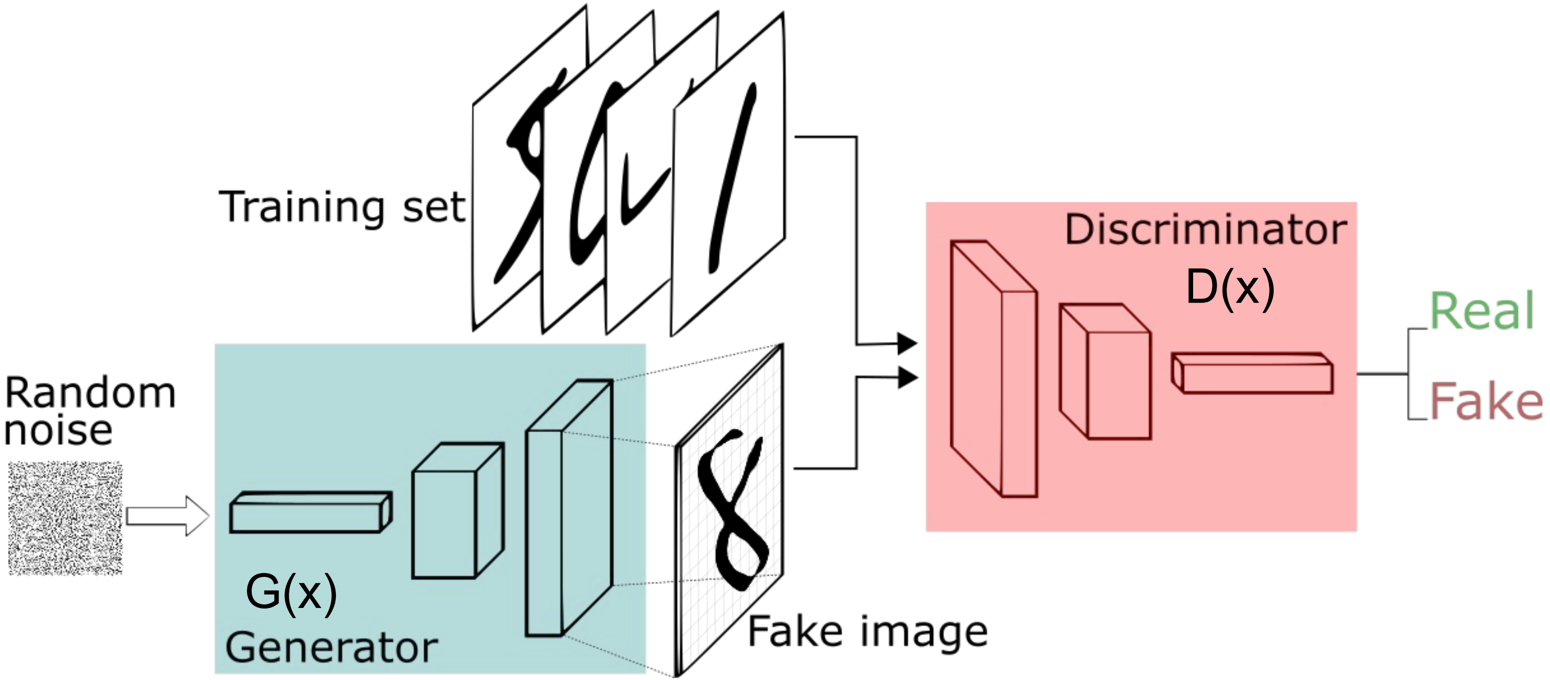
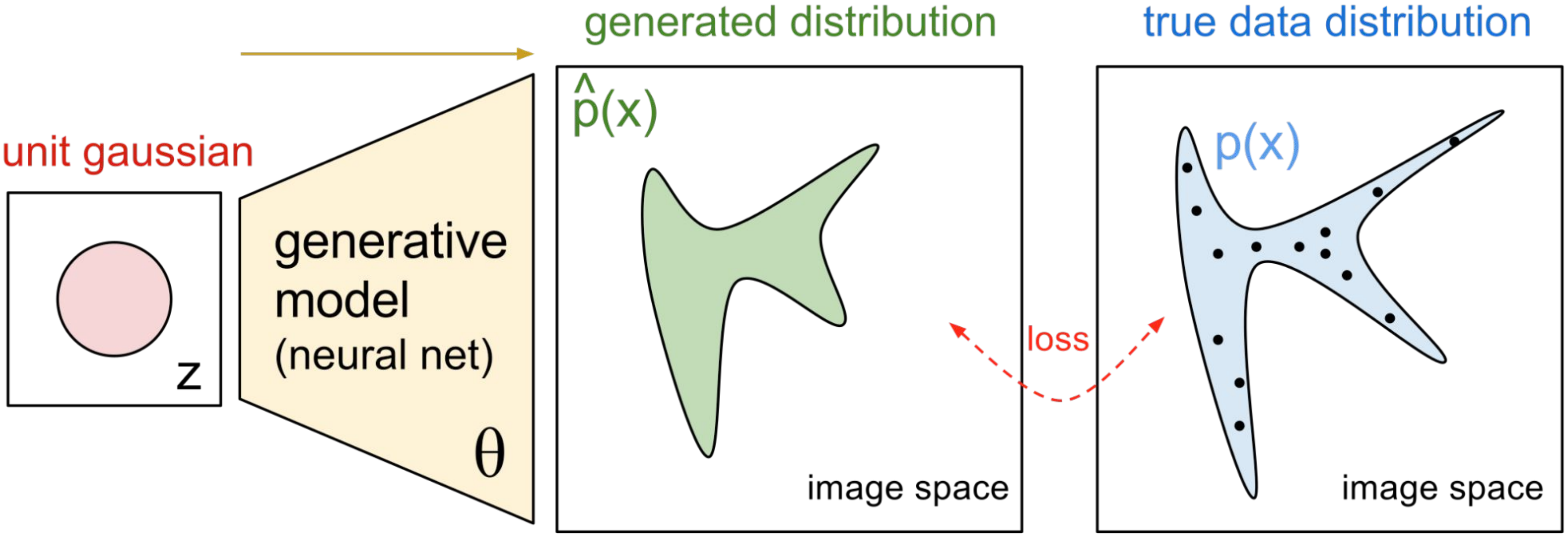
GAN Dissection

Polina Guseva, BAMI181

GANs. Recap


Training set
Random
noise
G(x)
Generator
Fake image
Discriminator
D(x)
Real
Fake

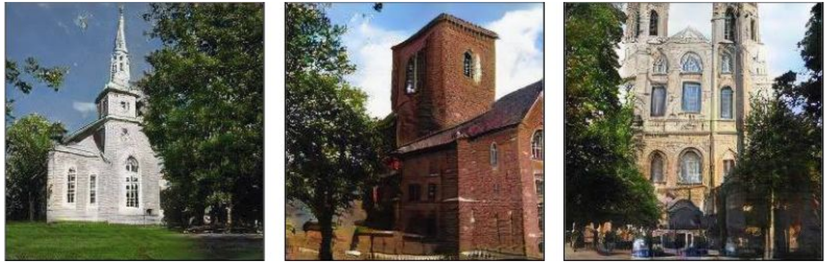
GANs. Recap



generated distribution
true data distribution
unit gaussian
z
generative
model
(neural net)
θ
$\hat{p}(x)$
image space
loss
p(x)
image space

Motivation

- What knowledge does GAN need to learn?



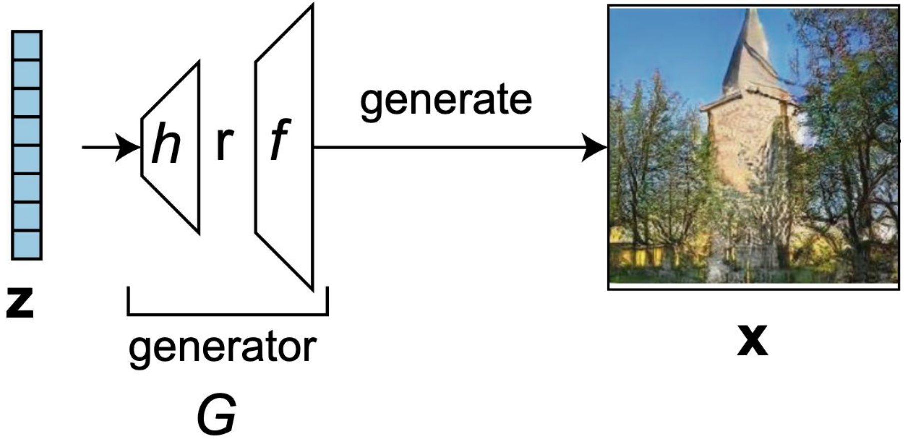
- What causes the mistakes?

- **Does GAN contain internal variables that correspond to the objects that humans perceive?**

Goal

Explain how an image can be generated by a network

Definitions



$\mathbf{z} \in \mathbb{R}^{|z|}$ Latent vector

$\mathbf{x} \in \mathbb{R}^{H \times W \times 3}$ Generated image

$G: \mathbf{z} \rightarrow \mathbf{x}$ Generator

$\mathbf{r} = h(\mathbf{z})$ Representation

$$\mathbf{x} = f(\mathbf{r}) = G(\mathbf{z})$$

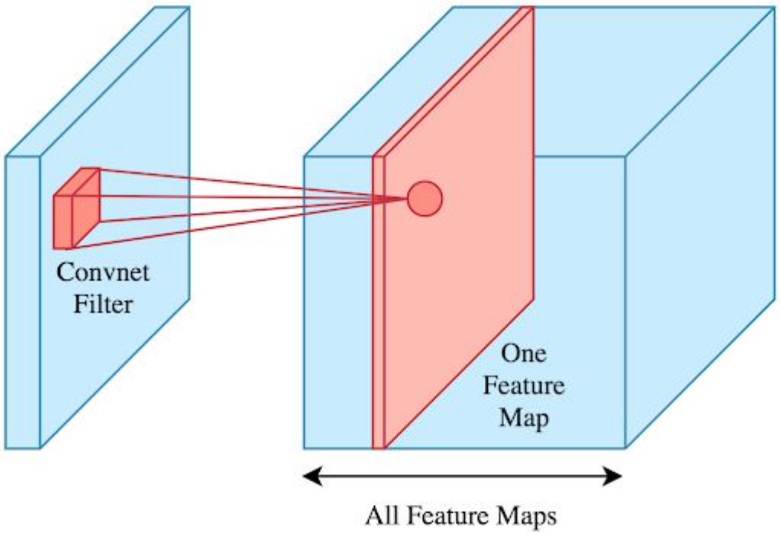
Definitions

U A set of units (channels)

P A set of pixels in featuremap

$\mathbb{U}$ All units

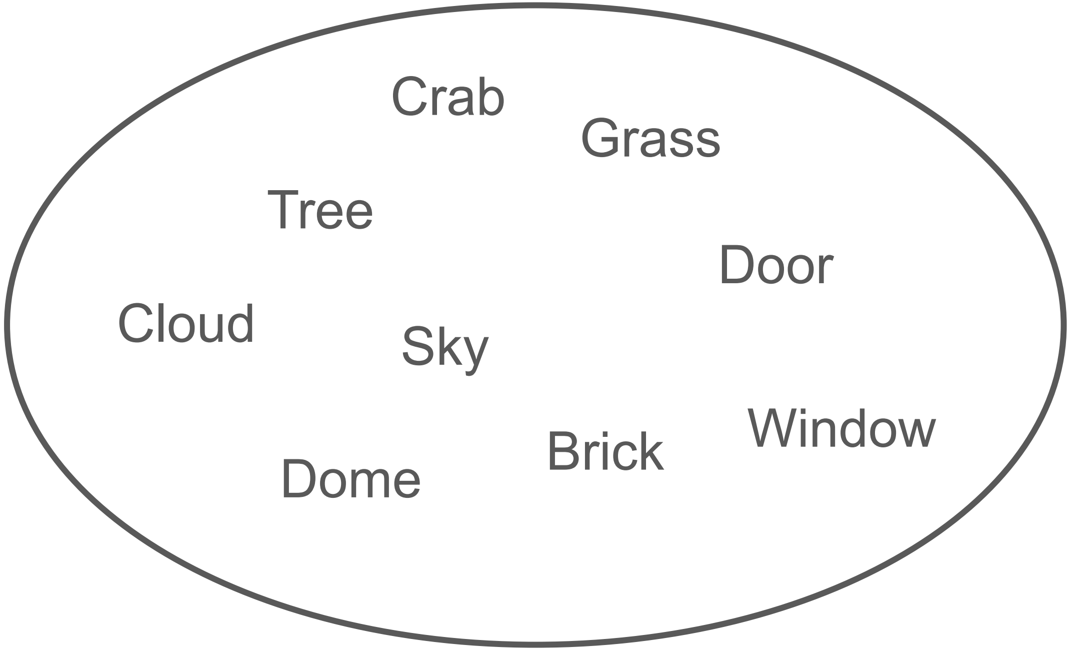
$\mathbb{P}$ All pixels



Definitions

$\mathcal{C}$ Universe of concepts

$c \in \mathcal{C}$ Concept



Task

Factor representation **r** at location P into components

$$\mathbf{r}_{\mathbb{U},\mathrm{P}} = \left(\mathbf{r}_{\mathrm{U},\mathrm{P}}, \mathbf{r}_{\overline{\mathrm{U}},\mathrm{P}}\right)$$

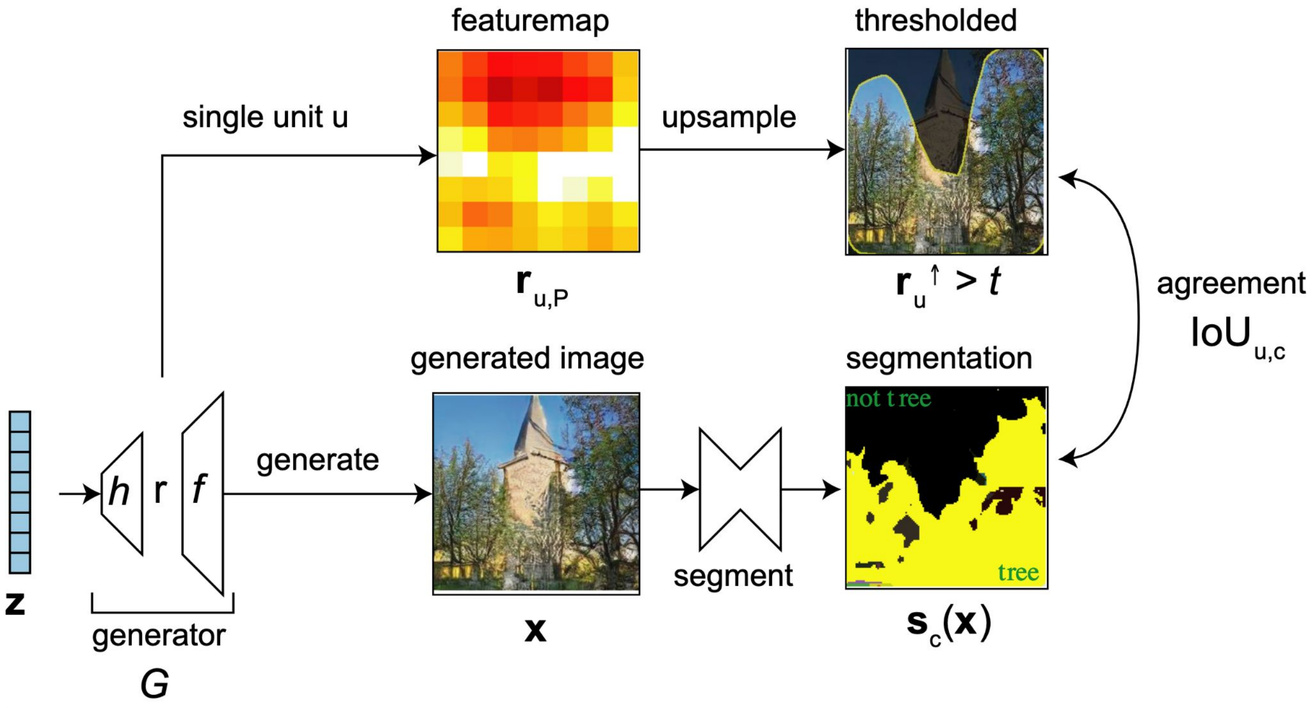
such that generation of object *c* is dependent on the units in first components and is insensitive to units in second component.

Proposed method

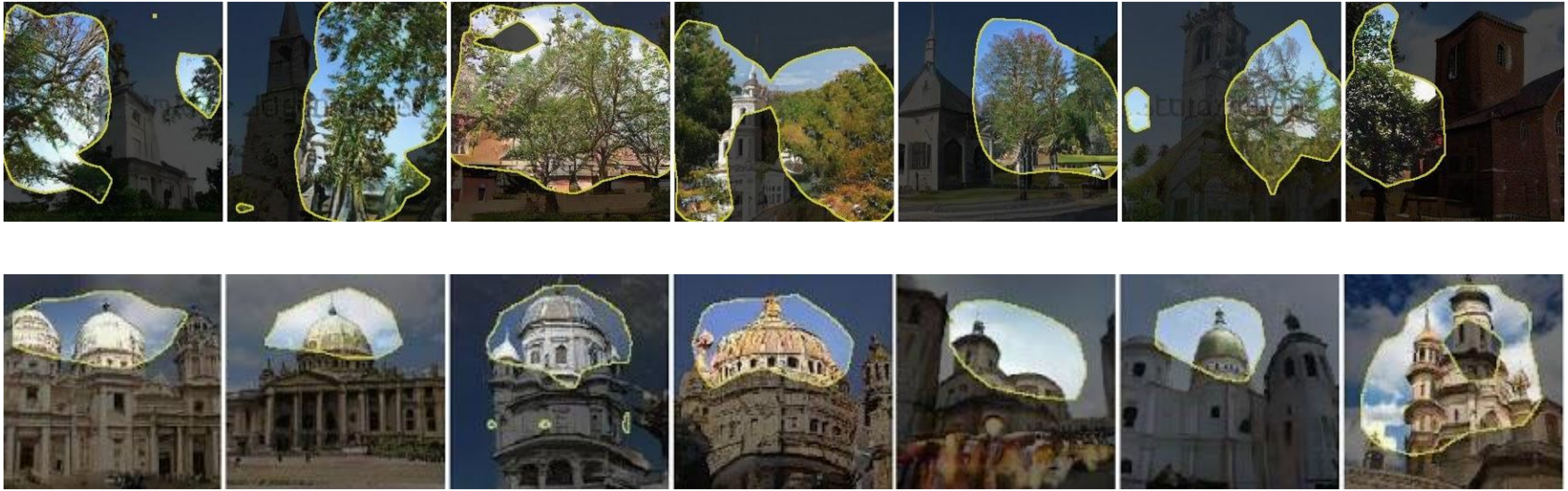
Two phases:

- Dissection. Select classes with explicit representations.
- Intervention. Identify causal sets of units.

Method. Dissection



Method. Dissection



Method. Dissection

$$\mathrm{\mathbf{IoU}}_{u,c} \equiv \frac{\mathbb{E}_{\mathbf{z}} \left| (\mathbf{r}^{\uparrow}_{u,\mathbb{P}} > t_{u,c}) \wedge \mathbf{s}_c(\mathbf{x}) \right|}{\mathbb{E}_{\mathbf{z}} \left| (\mathbf{r}^{\uparrow}_{u,\mathbb{P}} > t_{u,c}) \vee \mathbf{s}_c(\mathbf{x}) \right|}$$

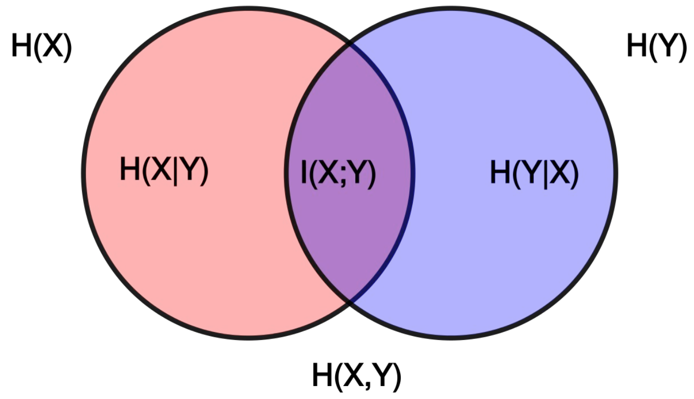
Method. Dissection

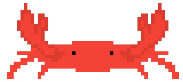
Q. How to select threshold?

A. Maximize **information quality ratio**

$$t_{u,c} = \arg\max_{t} \frac{\mathbf{I}(\mathbf{r}^{\uparrow}_{u,\mathbb{P}} > t; \mathbf{s}_c(\mathbf{x}))}{\mathbf{H}(\mathbf{r}^{\uparrow}_{u,\mathbb{P}} > t, \mathbf{s}_c(\mathbf{x}))}$$

$$I(X;Y) = H(X) + H(Y) - H(X,Y)$$





Method. Dissection

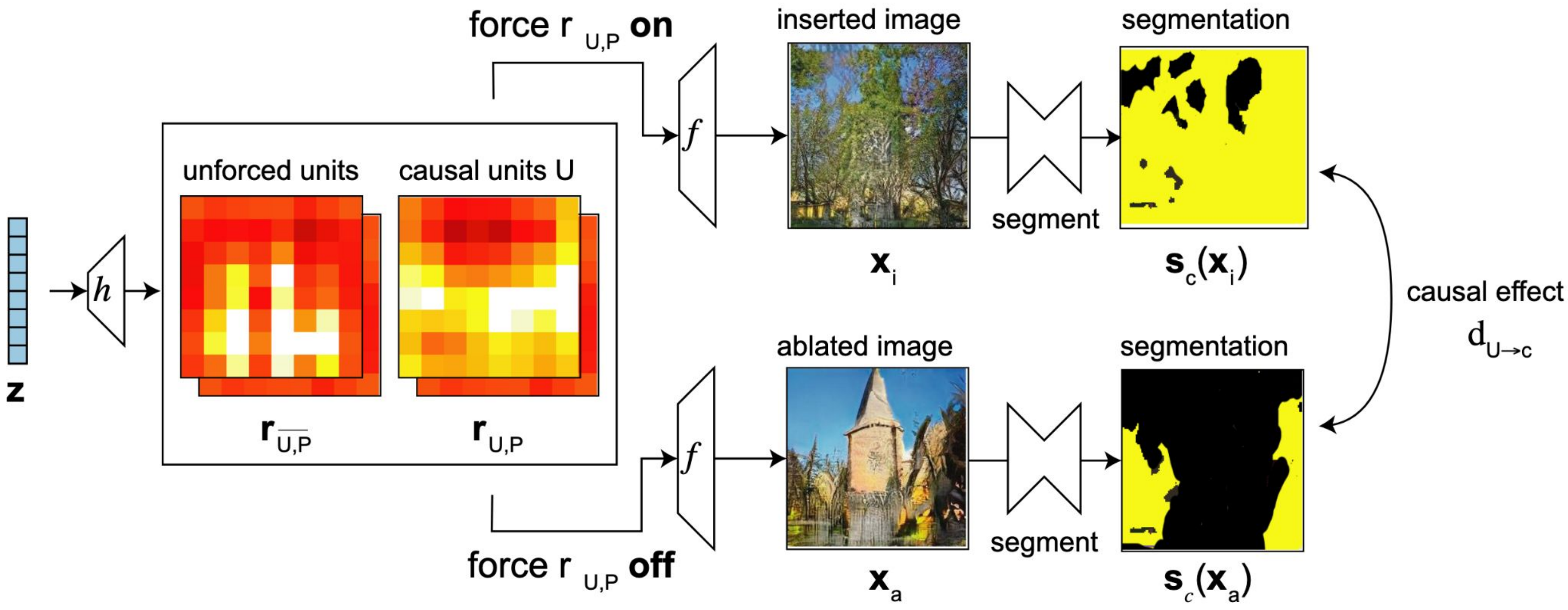
Unit №65

IoU=0.34

Unit №37

IoU=0.29

Method. Intervention



Method. Intervention

Original image : $\mathbf{x} = G(\mathbf{z}) \equiv f(\mathbf{r}) \equiv f(\mathbf{r}_{\mathrm{U,P}}, \mathbf{r}_{\overline{\mathrm{U,P}}})$

Image with U ablated at pixels P : $\mathbf{x}_a = f(\mathbf{0}, \mathbf{r}_{\overline{\mathrm{U,P}}})$

Image with U inserted at pixels P : $\mathbf{x}_i = f(\mathbf{k}, \mathbf{r}_{\overline{\mathrm{U,P}}})$

Method. Intervention

An object is caused by U if the object appears in x_i and disappears from x_a.

The measure is average causal effect (ACE)

$$\delta_{\mathbf{U}\rightarrow c} \equiv \mathbb{E}_{\mathbf{z},\mathbf{P}}[\mathbf{s}_c(\mathbf{x}_i)] - \mathbb{E}_{\mathbf{z},\mathbf{P}}[\mathbf{s}_c(\mathbf{x}_a)]$$

Method. Intervention

r contains d units

$\alpha \in [0, 1]^d$ — continious intervention

α_u — degree of intervention for u

$$\cancel{\mathbf{x}_a = f(\mathbf{0}, \mathbf{r}_{\overline{\mathrm{U,P}}})}$$

$$\cancel{\mathbf{x}_i = f(\mathbf{k}, \mathbf{r}_{\overline{\mathrm{U,P}}})}$$

$$\cancel{\delta_{\mathrm{U}\to c} \equiv \mathbb{E}_{\mathbf{z},\mathrm{P}}[\mathbf{s}_c(\mathbf{x}_i)] - \mathbb{E}_{\mathbf{z},\mathrm{P}}[\mathbf{s}_c(\mathbf{x}_a)]}$$

$$\longrightarrow$$

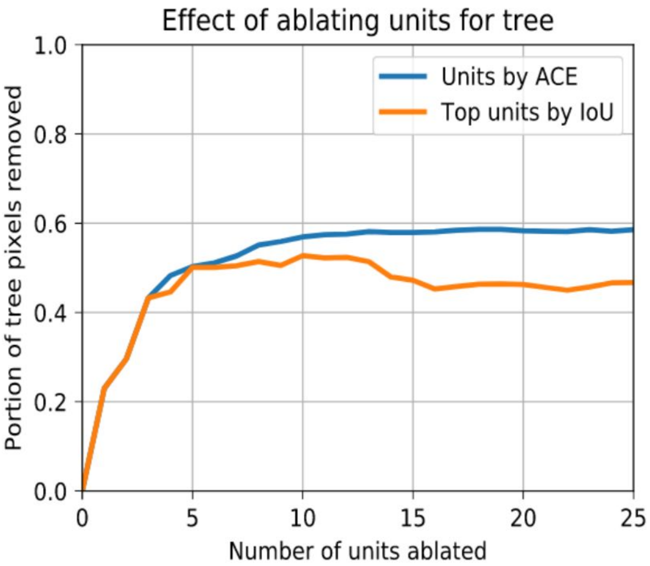
$$\mathbf{x}'_a = f((\mathbf{1} - \boldsymbol{\alpha}) \odot \mathbf{r}_{\mathbb{U},\mathrm{P}},\ \mathbf{r}_{\mathbb{U},\overline{\mathrm{P}}})$$

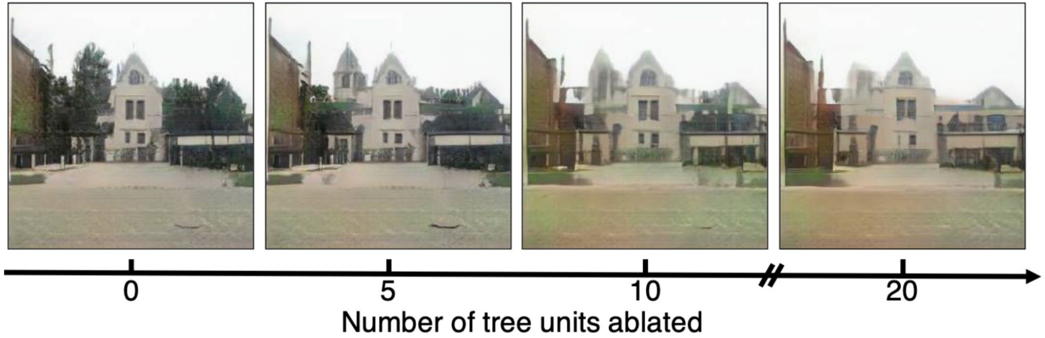
$$\mathbf{x}'_i = f(\boldsymbol{\alpha} \odot \mathbf{k} + (\mathbf{1} - \boldsymbol{\alpha}) \odot \mathbf{r}_{\mathbb{U},\mathrm{P}},\ \mathbf{r}_{\mathbb{U},\overline{\mathrm{P}}})$$

$$\delta_{\boldsymbol{\alpha}\to c} = \mathbb{E}_{\mathbf{z},\mathrm{P}}\left[\mathbf{s}_c(\mathbf{x}'_i)\right] - \mathbb{E}_{\mathbf{z},\mathrm{P}}\left[\mathbf{s}_c(\mathbf{x}'_a)\right],$$

$$\boldsymbol{\alpha}^* = \arg\min_{\boldsymbol{\alpha}}(-\delta_{\boldsymbol{\alpha}\to c} + \lambda||\boldsymbol{\alpha}||_2)$$

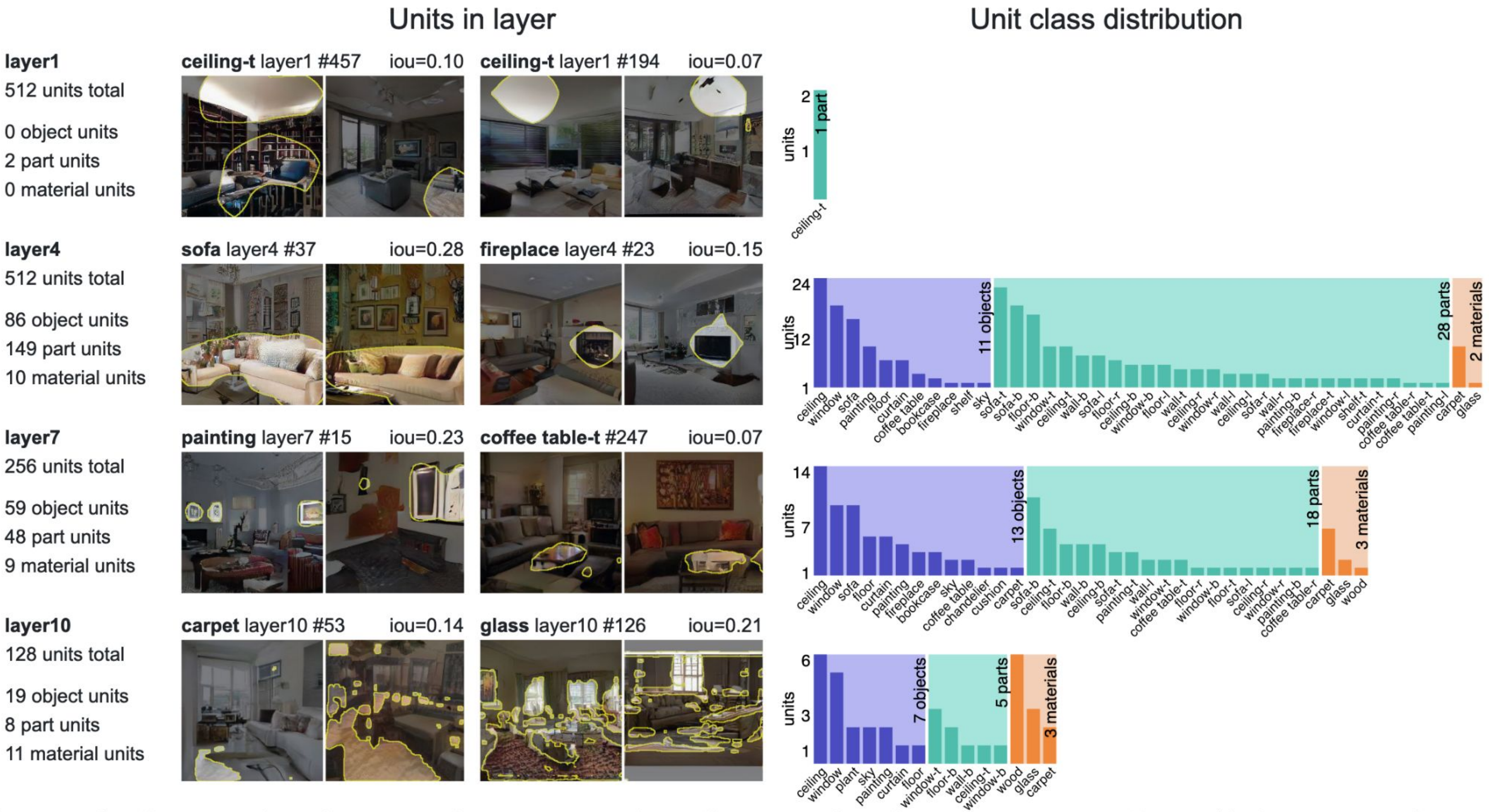
Method. Intervention





Results. Interpretable units for different scene types

Results. Units for different network layers



Results. Units for various networks

interpretable units | SWD | Best "bed" unit | Best "window" unit | Unit class distribution

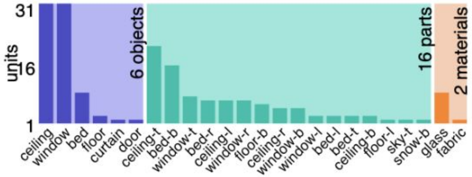
base prog GAN
512 units total
74 object units
84 part units
9 material units

167 units

7.60

bed layer4 #253 iou=0.18

window layer4 #142 iou=0.19



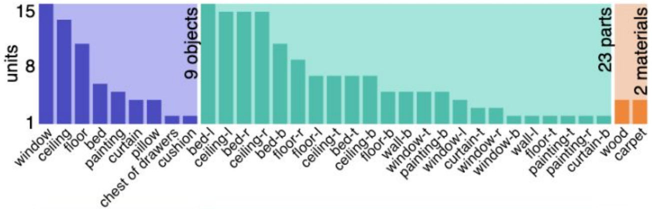
+batch stddev
512 units total
55 object units
128 part units
6 material units

189 units

6.48

bed layer4 #88 iou=0.11

window layer4 #422 iou=0.25



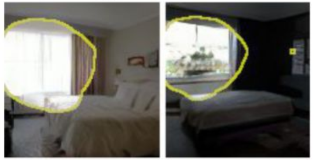
+pixelwise norm
512 units total
82 object units
128 part units
16 material units

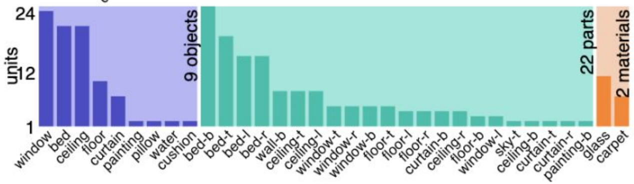
226 units

4.01

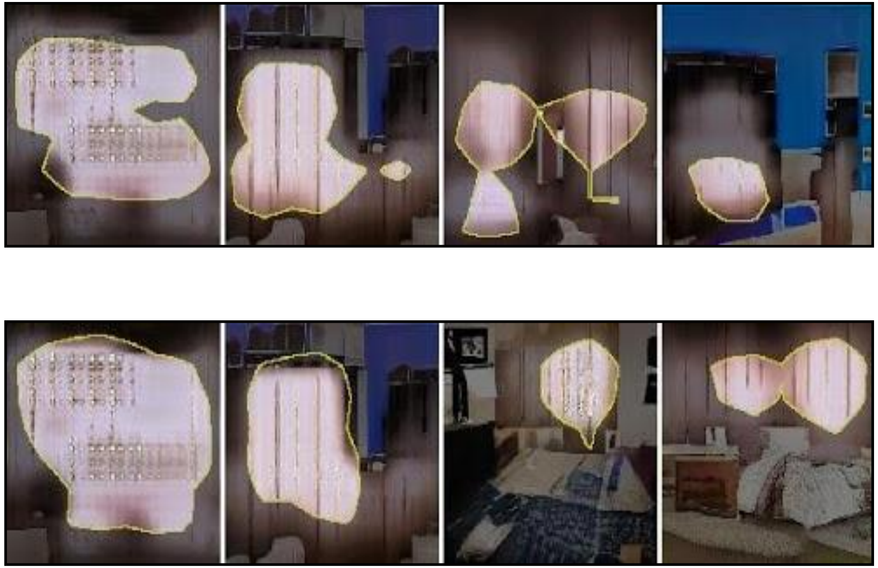
bed layer4 #129 iou=0.29

window layer4 #494 iou=0.26





Results. Debugging GANs

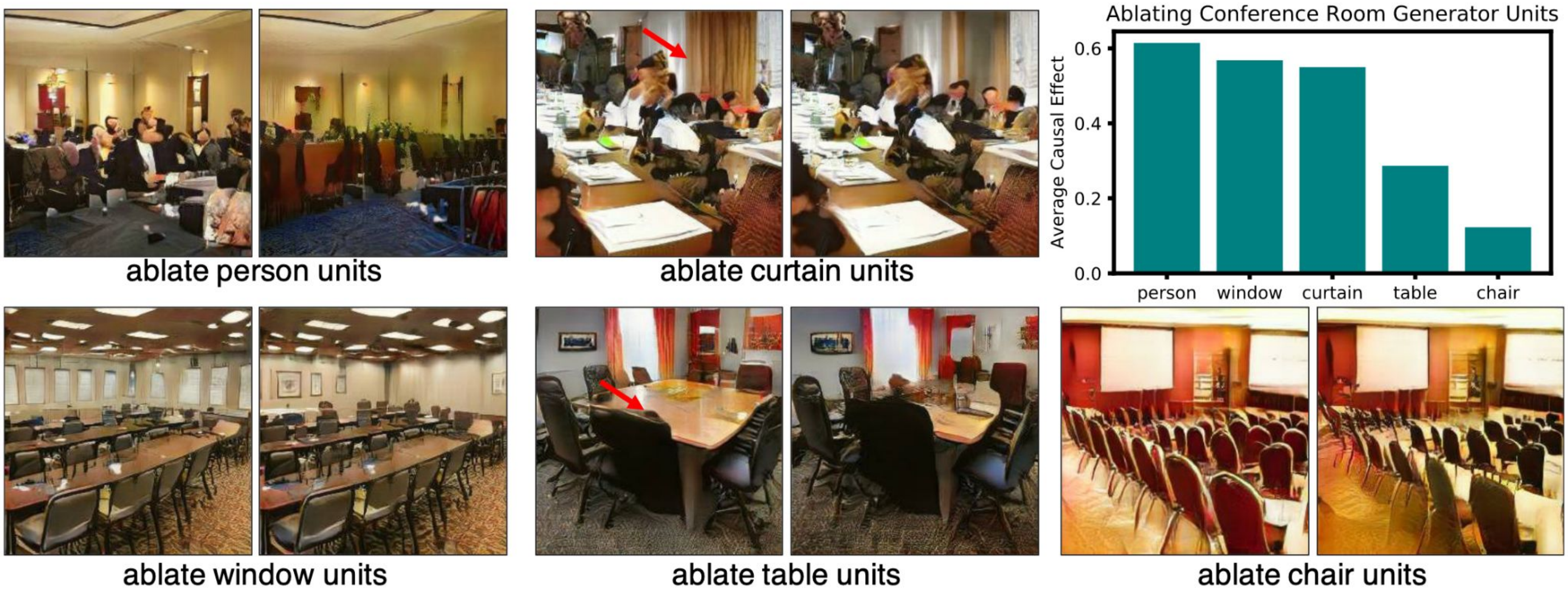


Example artifact-causing units

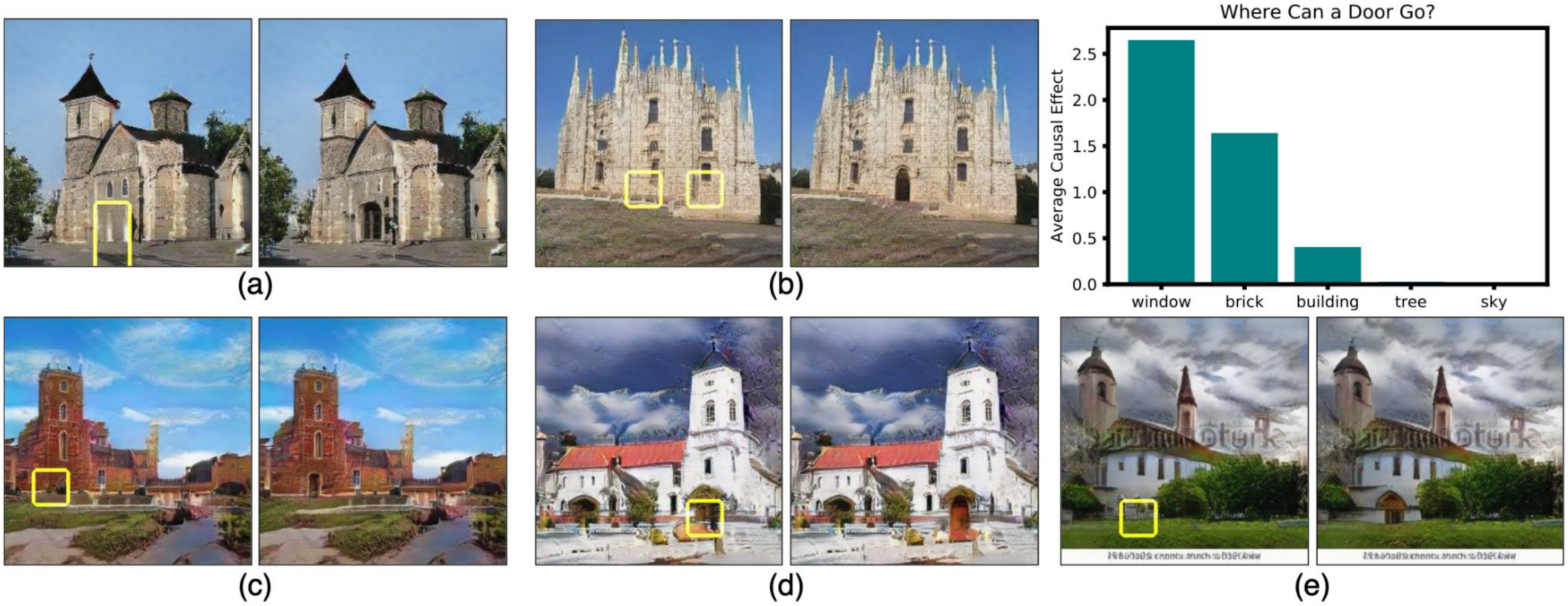
Bedroom images with artifacts

Ablating "artifact" units improves results

Results. Erasing objects



Results. Inserting objects



Summary

- GANs have **sets of neurons** that **explicitly control object generation**
- Suggested that GANs learned some **aspects of composition**
- Some **artifacts** may be triggered by **specific sets of neurons** — easy fix

References

1. David Bau et al. GAN Dissection: Visualizing and Understanding Generative Adversarial Networks. 2019
2. David Bau et al. Understanding the role of individual units in a deep neural network. 2020
3. OpenAI Blog. Generative models. 2016
4. David Bau, Bolei Zhou et al. Network dissection: Quantifying interpretability of deep visual representations. 2017