GAN Dissection

Based on : David Bau, Jun-Yan Zhu, Hendrik Strobelt, Bolei Zhou, Joshua B. Tenenbaum, William T. Freeman, Antonio Torralba. GAN Dissection: Visualizing and Understanding Generative Adversarial Networks. arXiv preprint arxiv 1811.10597, 2018.
A slight modification of slides found at : https://gandissect.csail.mit.edu/

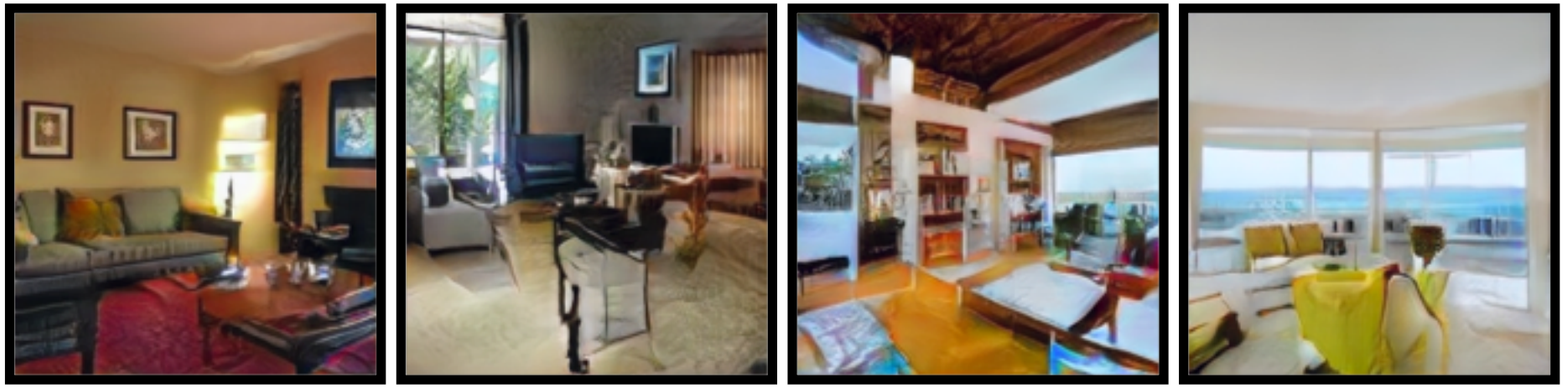
Presented by : Mani Sotoodeh
Date: 02.19.2019

Some motivating questions?

- Does a GAN model have an explicit variable for representing objects, or it just detects a pattern of pixels?
- What happens within the structure of a GAN when it produces unrealistic results or exceptionally good results?
- How does a GAN represent relationship of objects?
- Goals: find neurons, objects, contextual relationships between object that causes certain real objects to appear and disappear. Leverage these to manipulate pictures.

Church

Living room



Restaurant

256x256 images synthesized by a Progressive GAN [Karras, et al 2017]

Church

To render a beautiful scene,

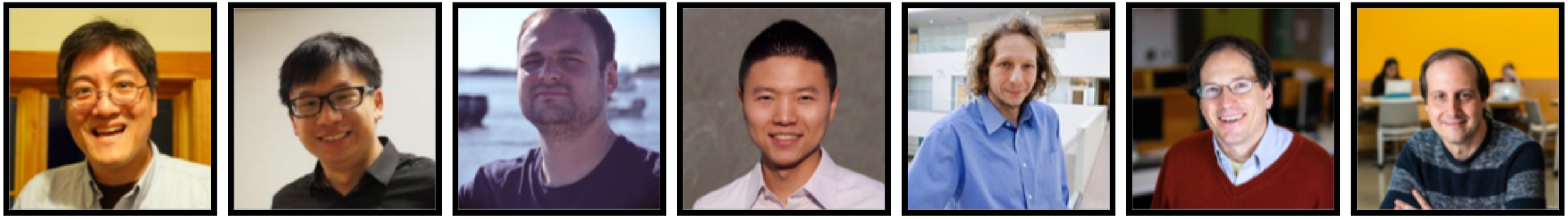
What does a GAN need to know?

Bedroom

What causes the mistakes?

GAN Dissection: Visualizing and Understanding Generative Adversarial Networks

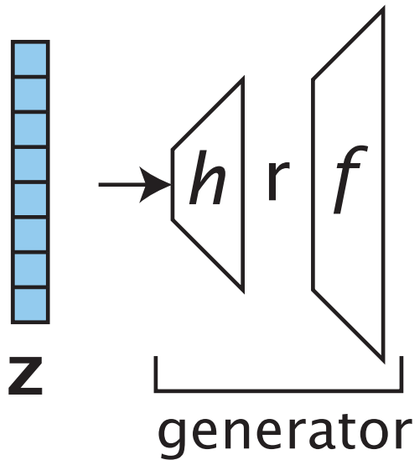
David Bau, Jun-Yan Zhu, Hendrik Strobelt, Bolei Zhou
Joshua B. Tenenbaum, William T. Freeman, Antonio Torralba



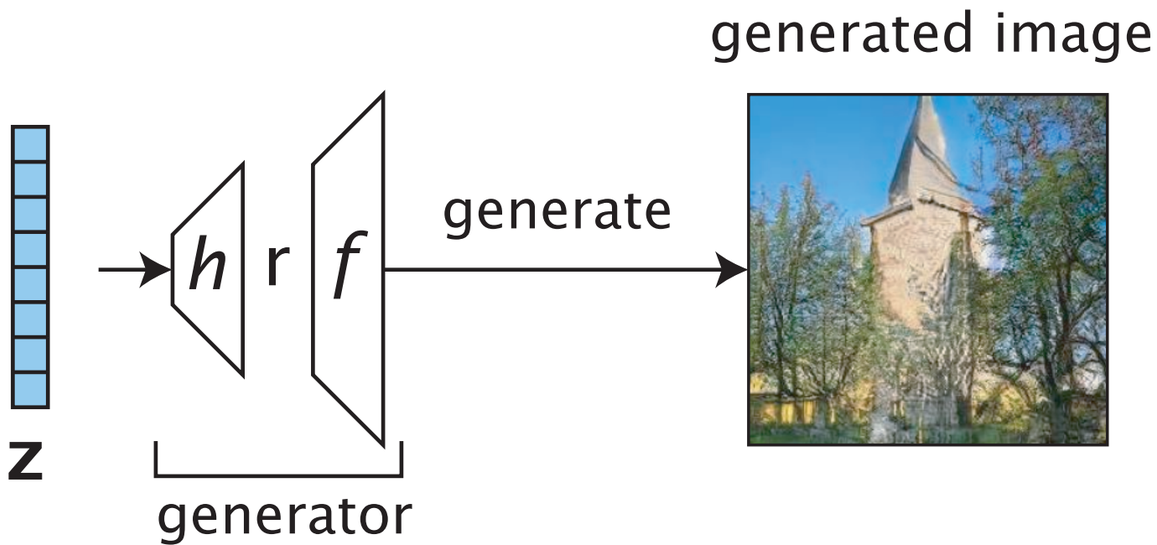
Which units correlate to an object class?

Which units correlate to an object class?

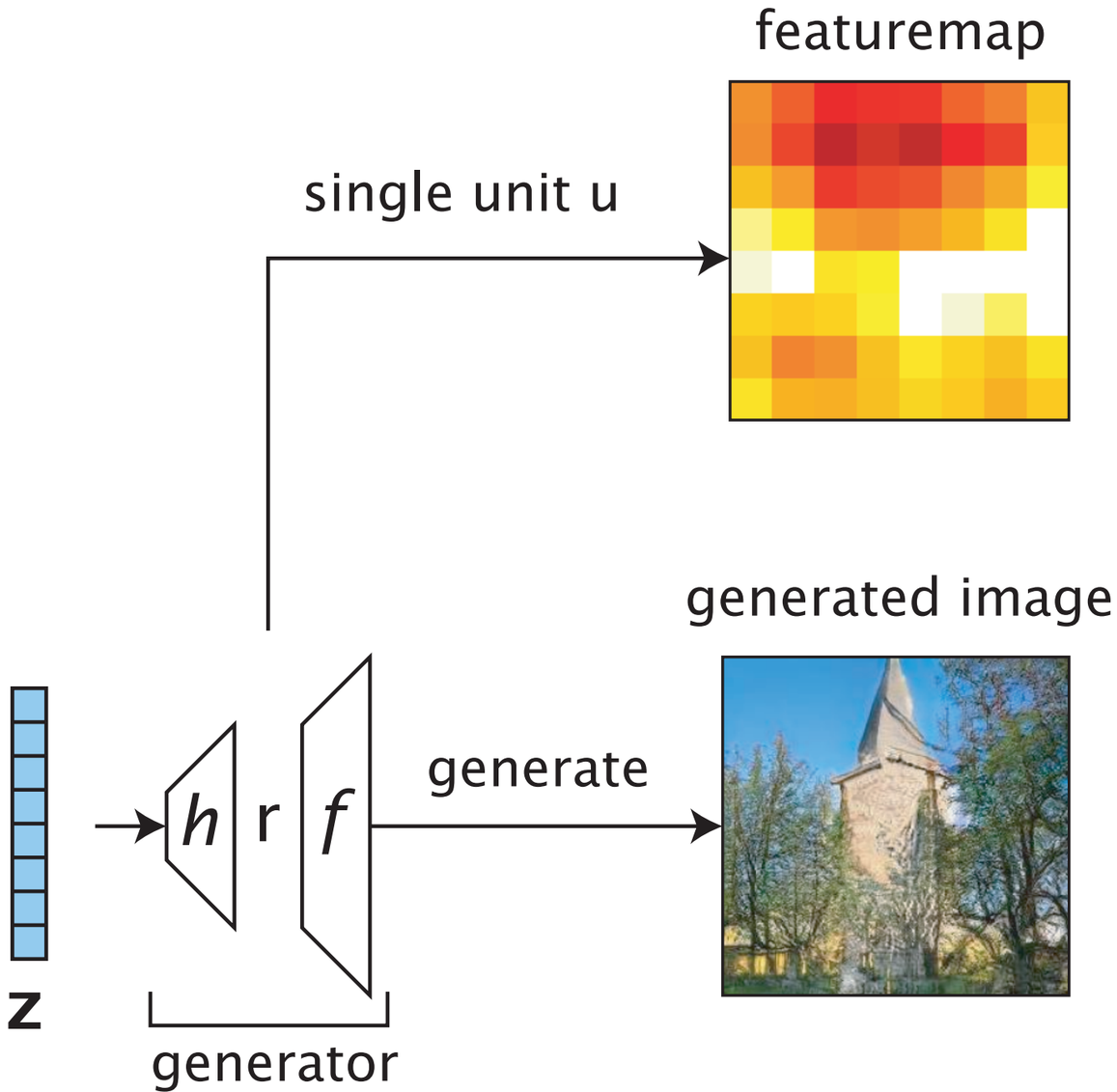
r: the current layer
h: the first half
f: the second half



Which units correlate to an object class?



Which units correlate to an object class?



Notation

$$\text{IoU}_{u,c} \equiv \frac{\mathbb{E}_{\mathbf{z}} \left| (\mathbf{r}^{\uparrow}_{u,\mathbb{P}} > t_{u,c}) \wedge \mathbf{s}_c(\mathbf{x}) \right|}{\mathbb{E}_{\mathbf{z}} \left| (\mathbf{r}^{\uparrow}_{u,\mathbb{P}} > t_{u,c}) \vee \mathbf{s}_c(\mathbf{x}) \right|}, \text{where } t_{u,c} = \arg\max_t \frac{\text{I}(\mathbf{r}^{\uparrow}_{u,\mathbb{P}} > t; \mathbf{s}_c(\mathbf{x}))}{\text{H}(\mathbf{r}^{\uparrow}_{u,\mathbb{P}} > t, \mathbf{s}_c(\mathbf{x}))},$$

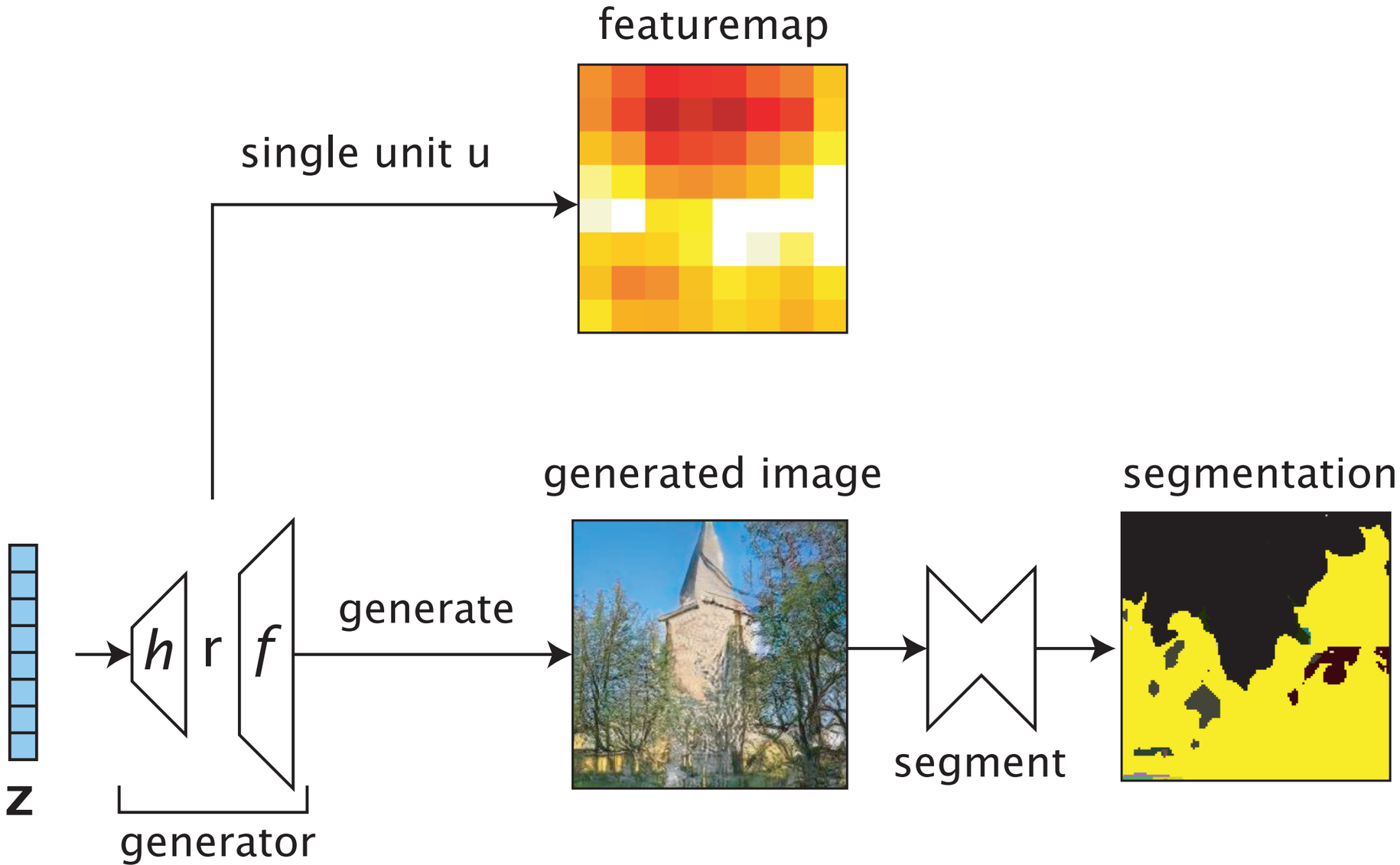
The internal representations of a GAN generator : $G: \mathbf{z} \rightarrow \mathbf{x}$

A latent vector sampled from a low-dimensional distribution $\mathbf{z} \in \mathbb{R}^{|z|}$

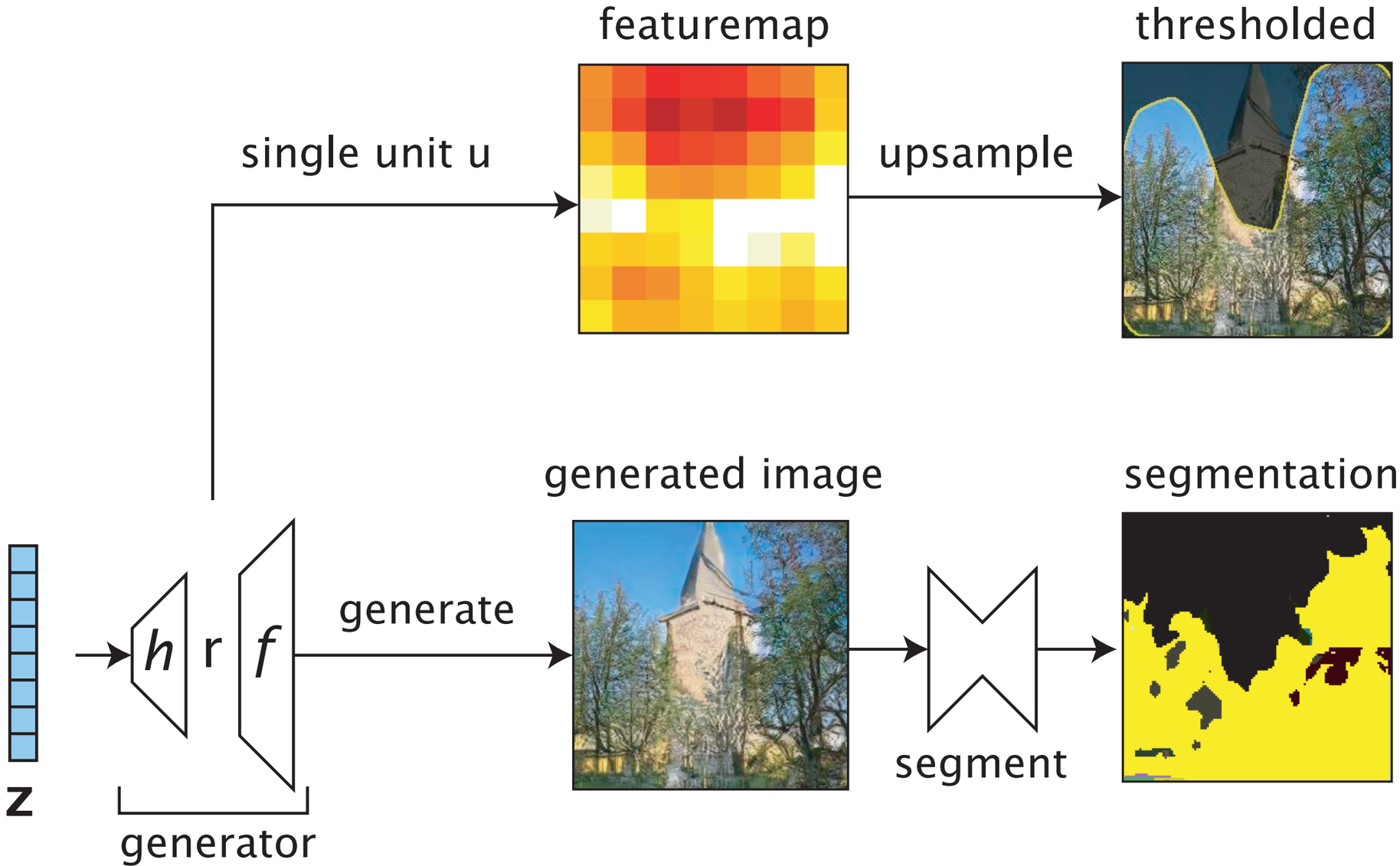
$\mathbf{x} \in \mathbb{R}^{H \times W \times 3}$ denotes an $H \times W$ generated image.

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

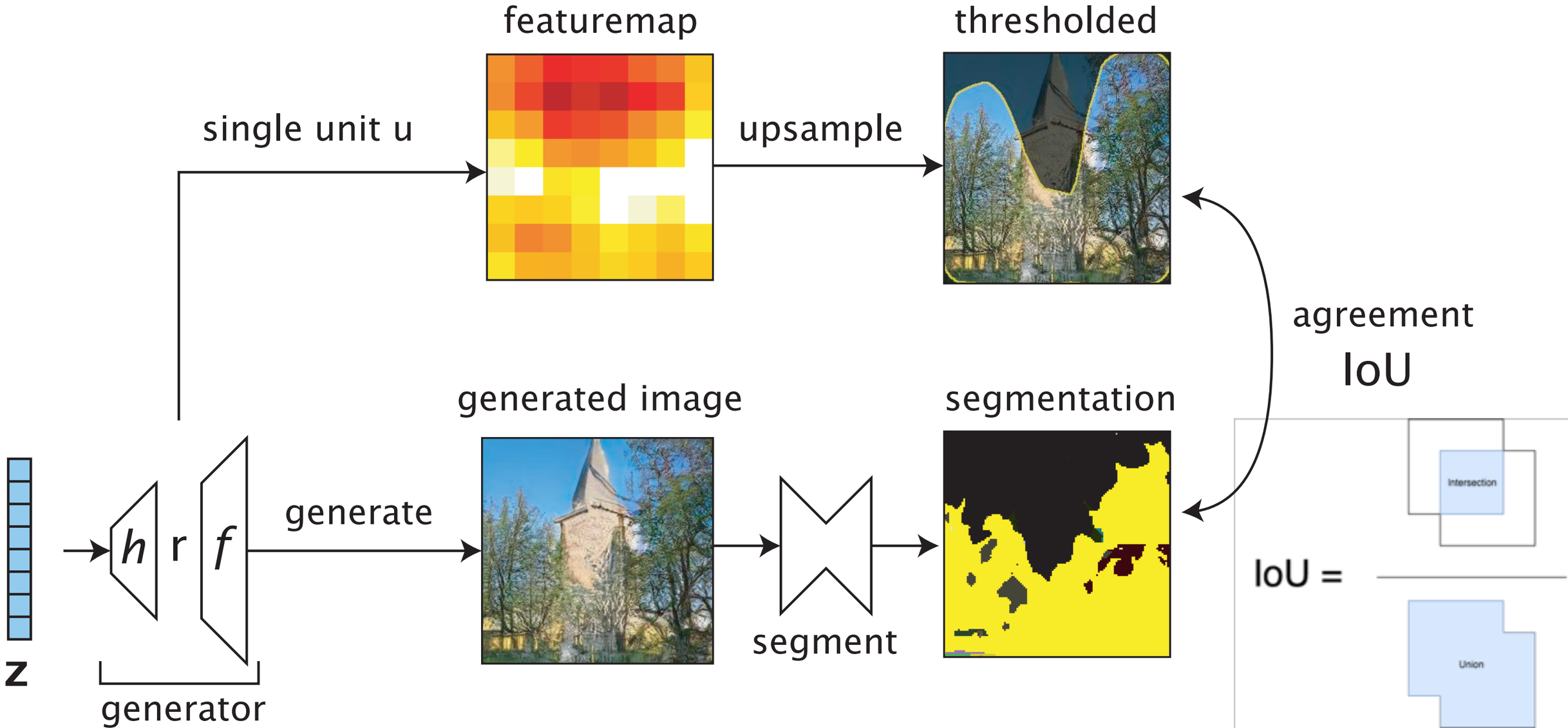
Which units correlate to an object class?



Which units correlate to an object class?



Which units correlate to an object class?



Which units **correlate** to an object class?

Church samples

Unit #119
Tree

Unit #32
Dome

Which units **correlate** to an object class?

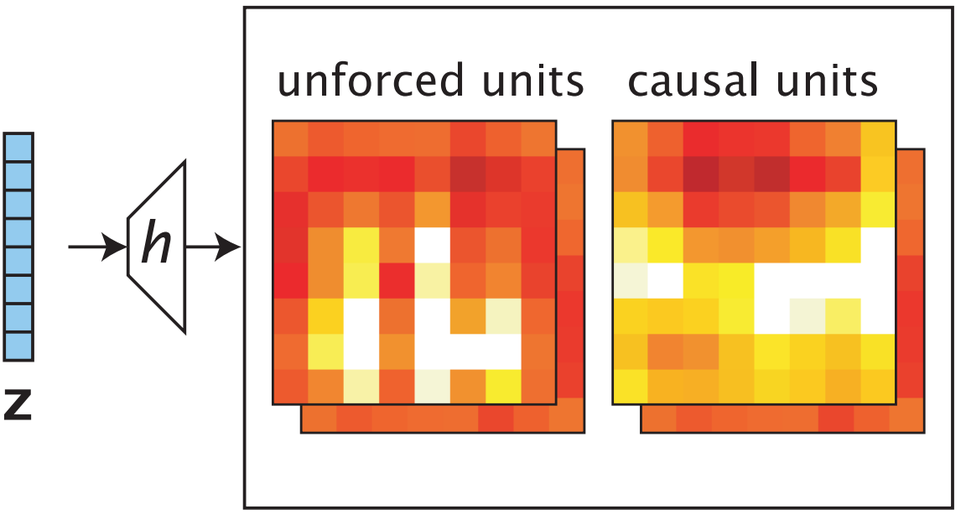
Dining room samples

Unit #139
Window

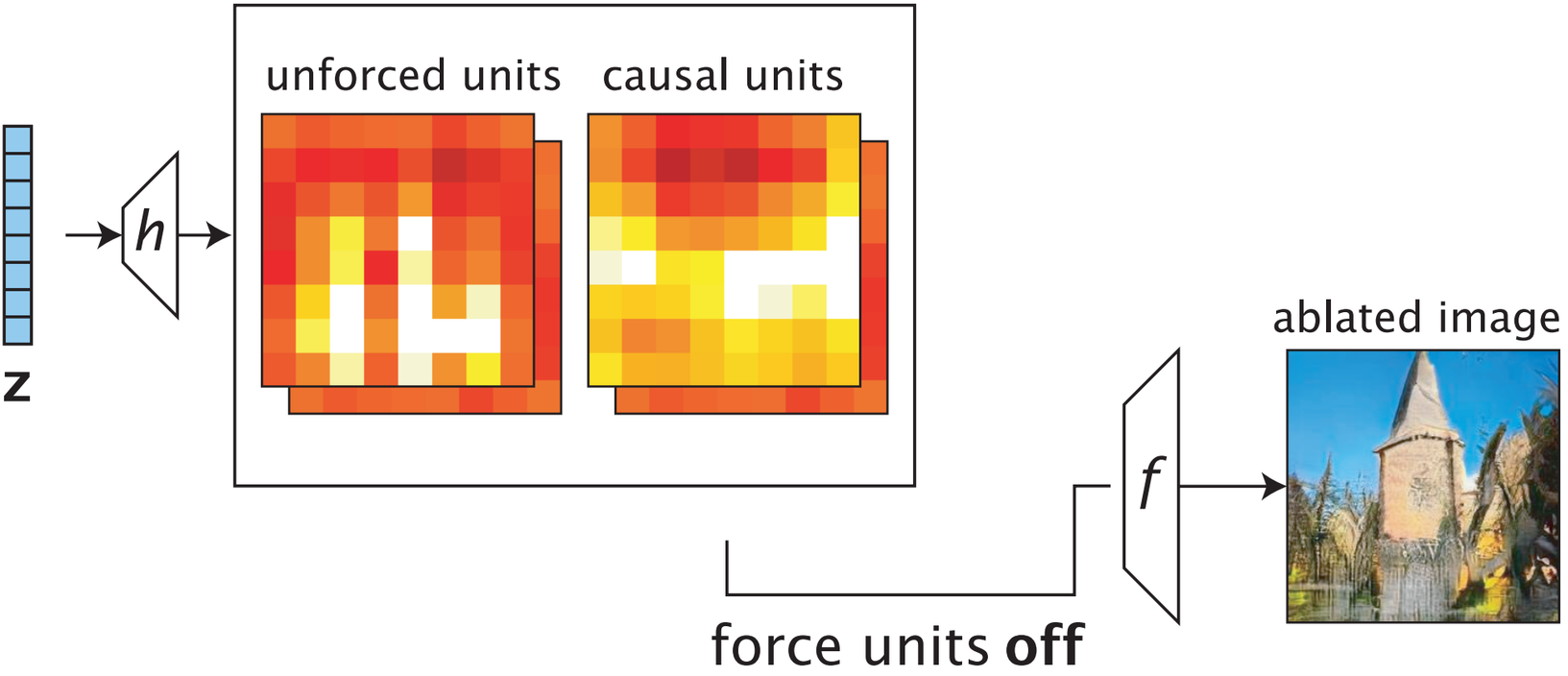
Unit #65
Table

Which units **cause** an object class?

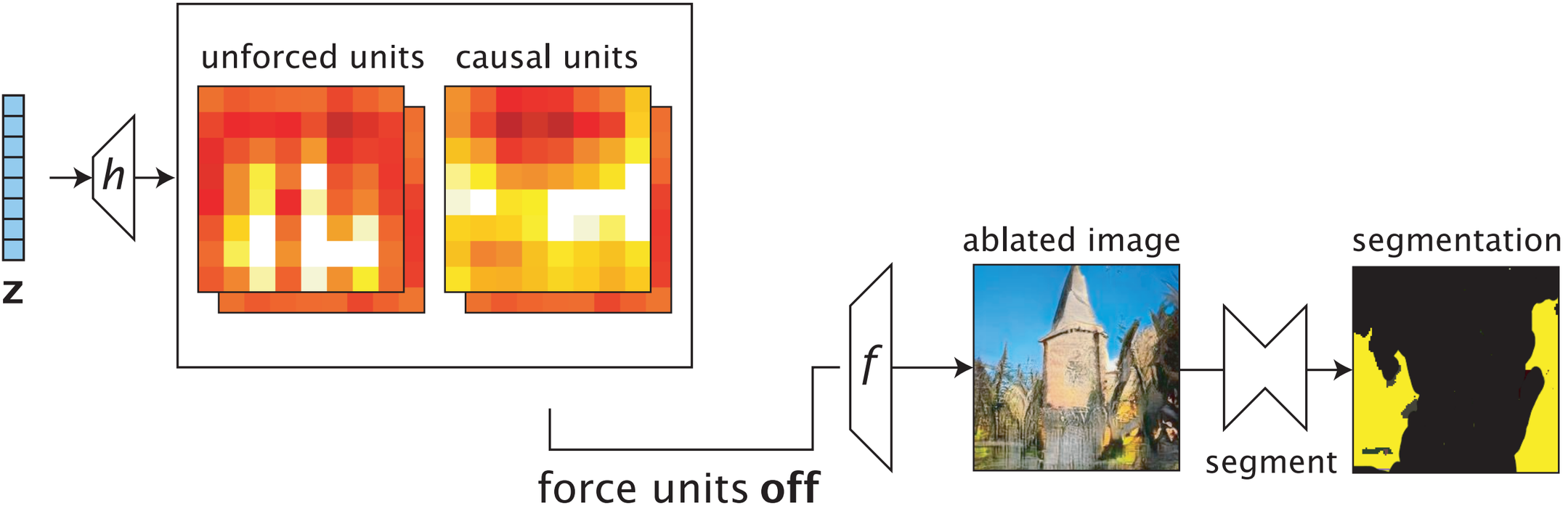
Which units **cause** an object class?



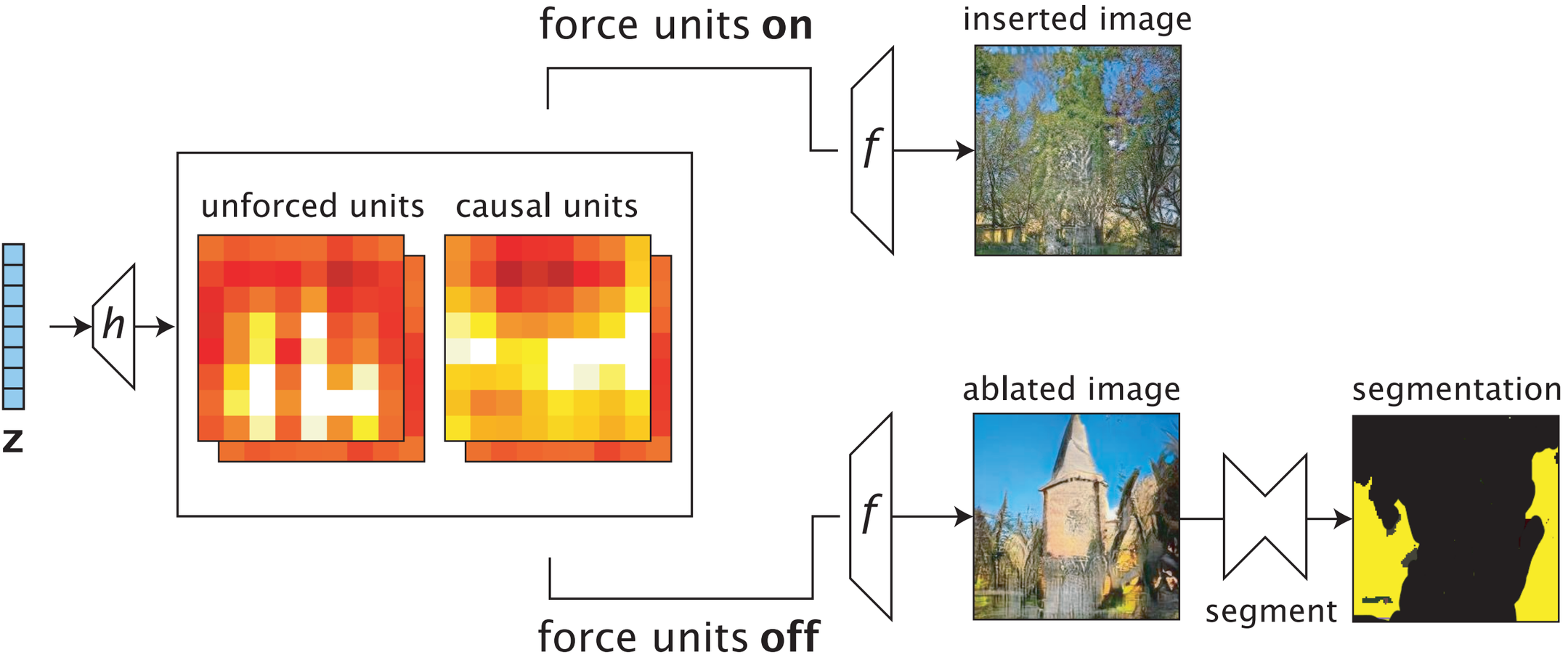
Which units **cause** an object class?



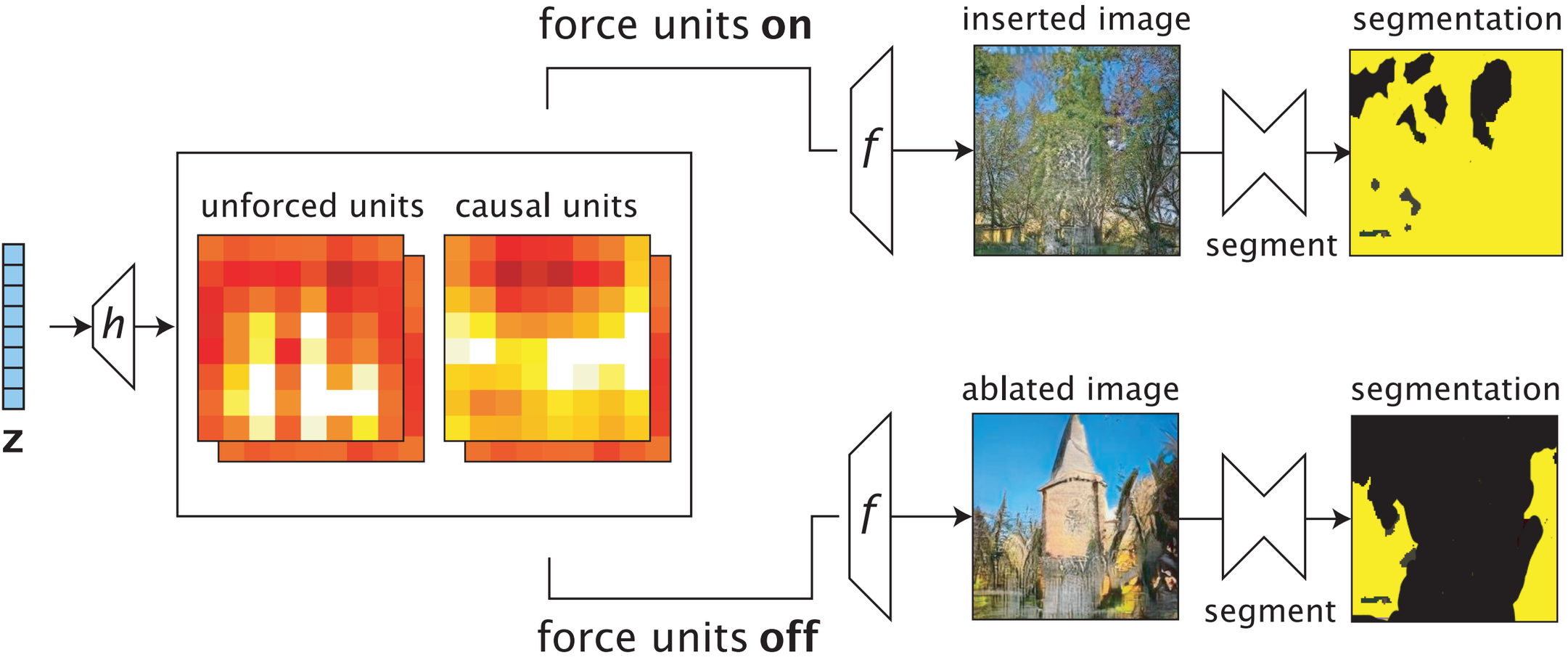
Which units **cause** an object class?



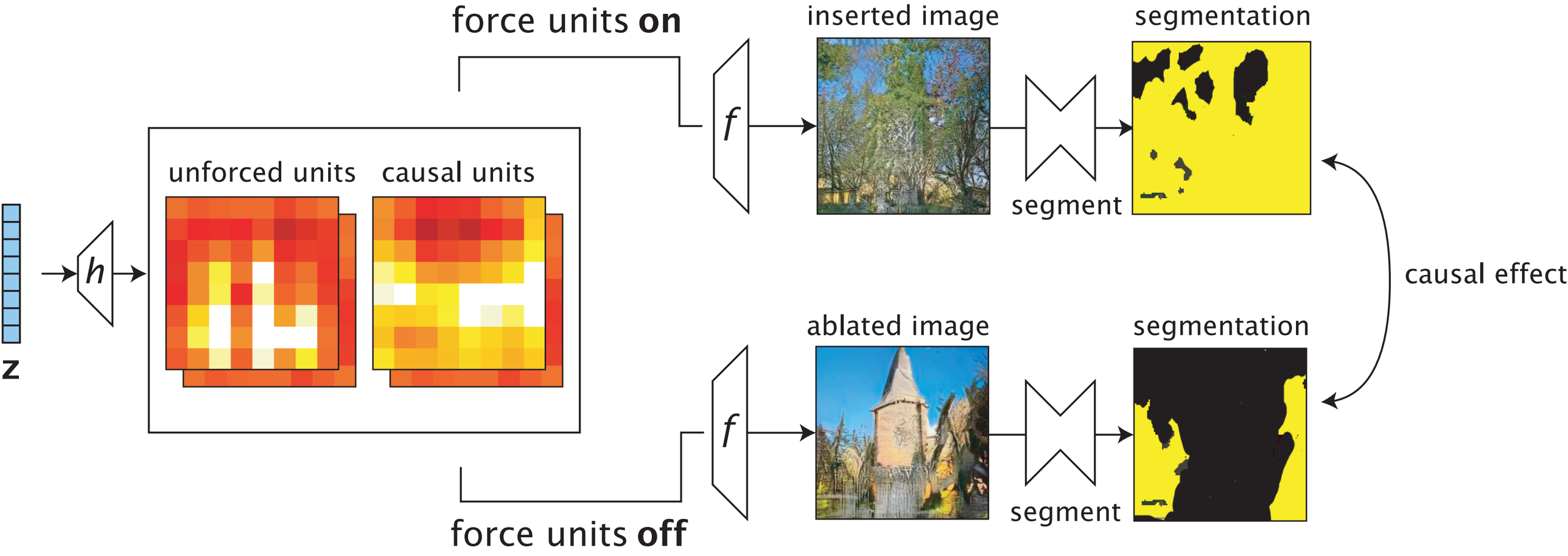
Which units **cause** an object class?



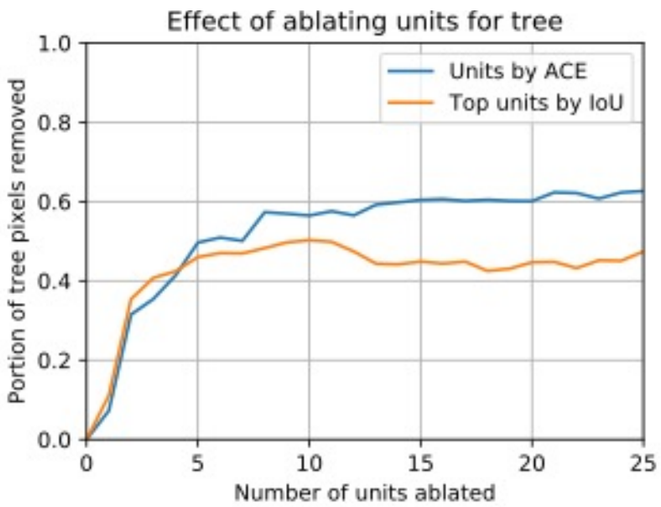
Which units **cause** an object class?



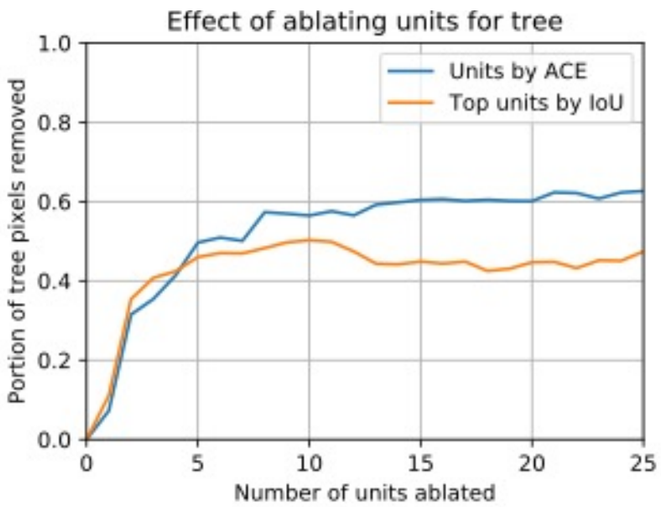
Which units **cause** an object class?



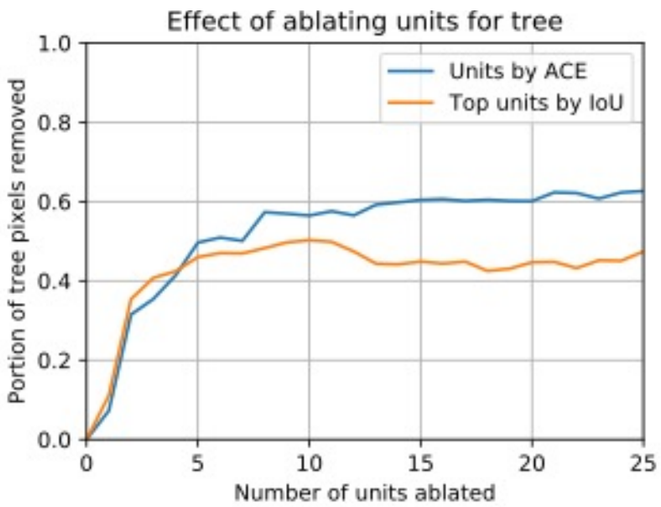
Removing or Adding Units



Removing or Adding Units



Removing or Adding Units



Underlying math!

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

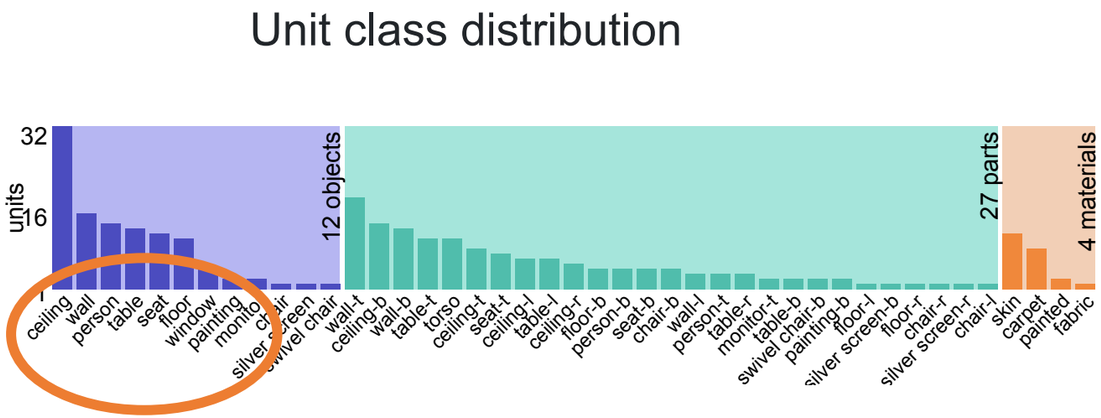
Image with U ablated at pixels P : $\mathbf{x}_a = f(\mathbf{0}, \mathbf{r}_{\overline{\mathrm{U,P}}})$

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

the average causal effect (ACE) of units U on the generation of on class c

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

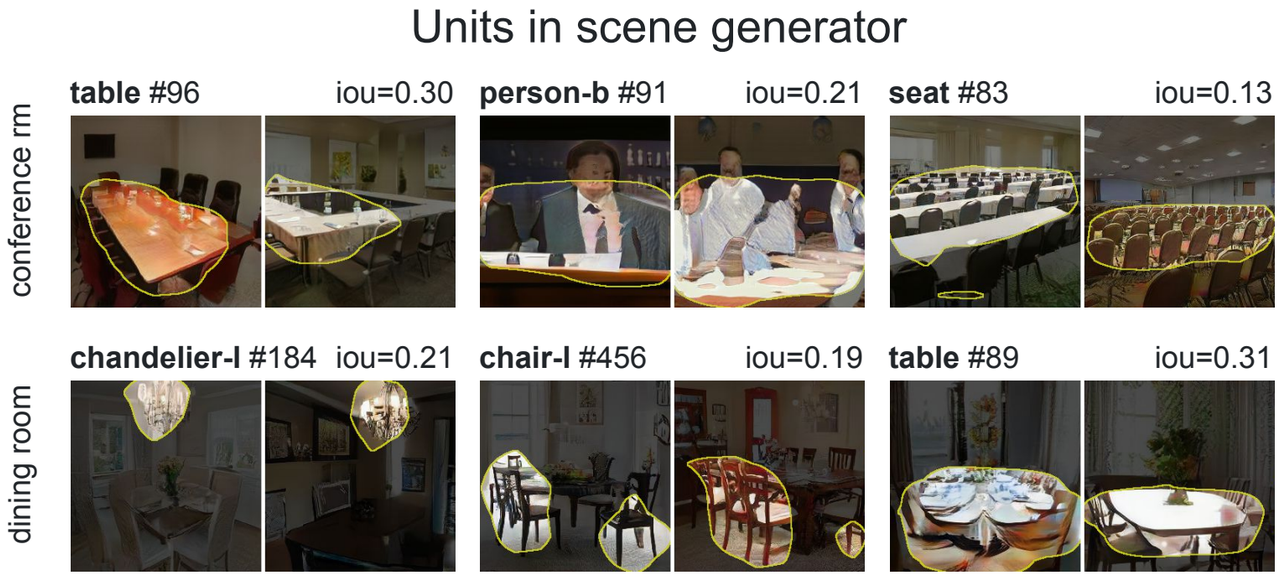
GAN Dissection: Comparing Datasets

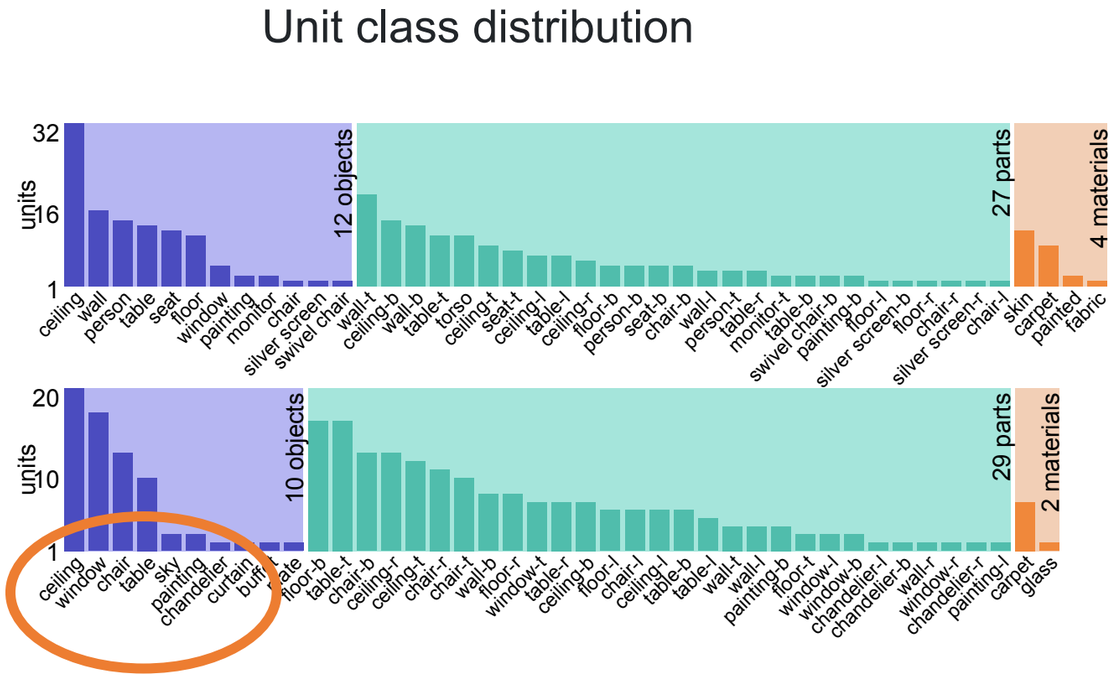


Top objects: ceiling, wall, person, table…

Scene: conference room

GAN Dissection: Comparing Datasets





Top objects: ceiling, window, chair, table…

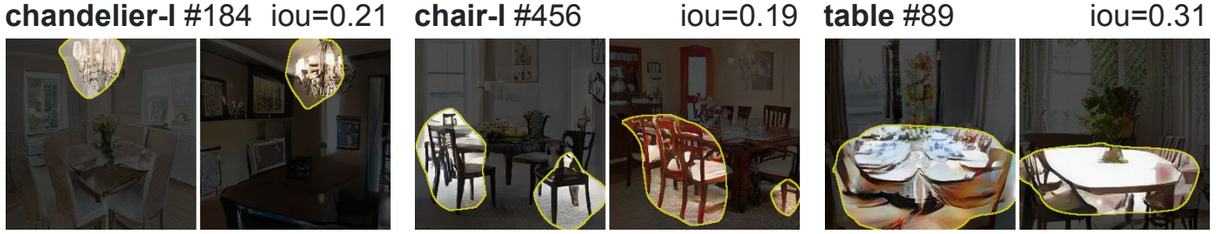
Scene: dining room

GAN Dissection: Comparing Datasets

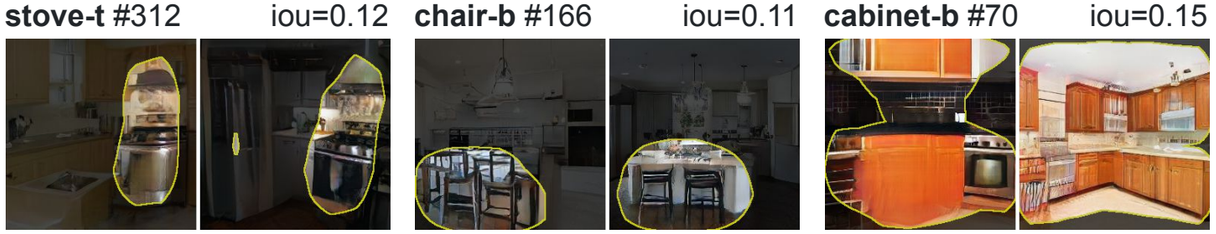
Units in scene generator

conference rm

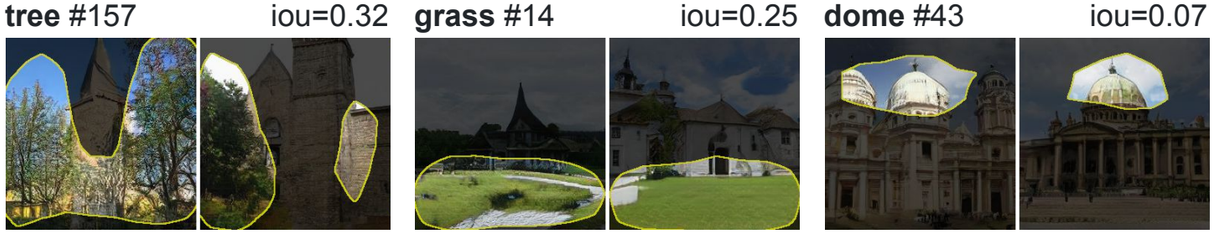
dining room



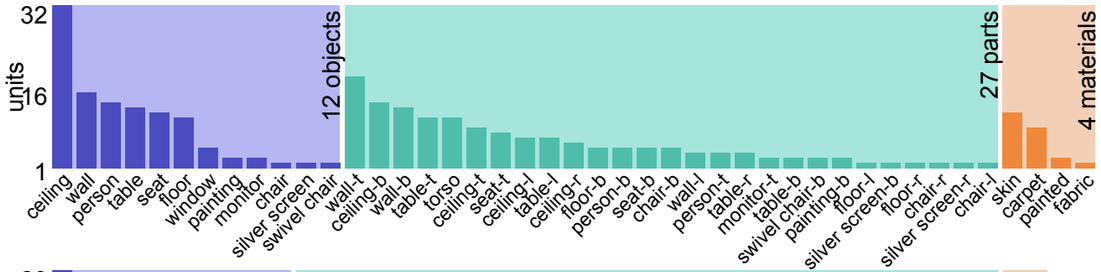
kitchen

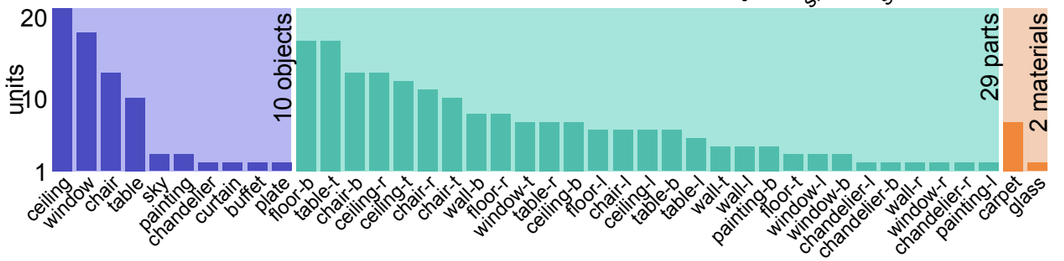


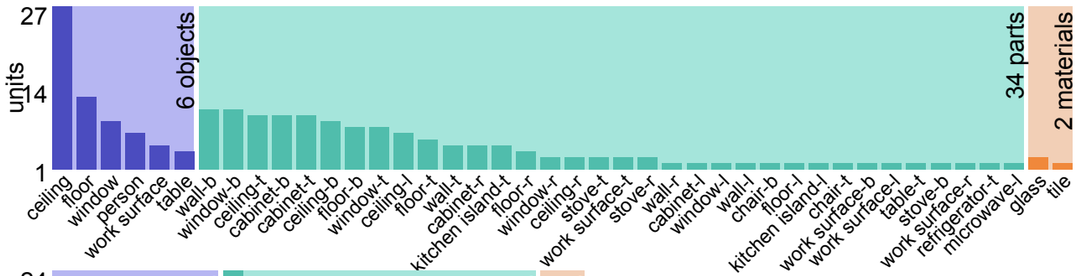
church/outdoor

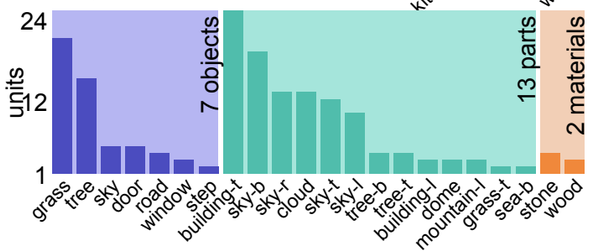


Unit class distribution

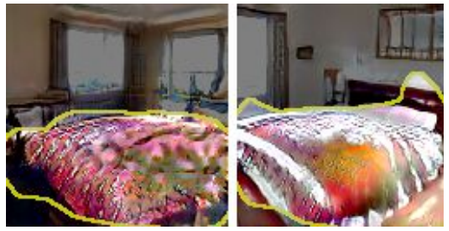
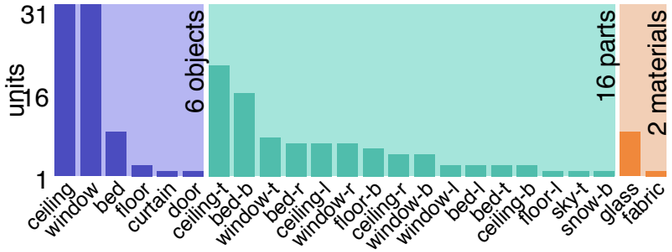








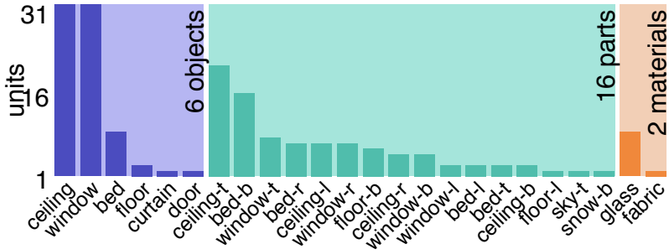
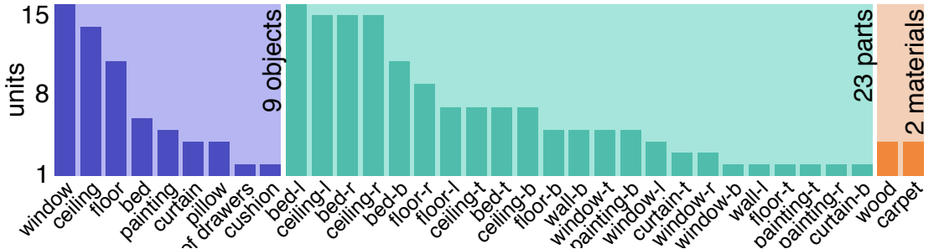
GAN Dissection: Comparing Models

interpretable units		SWD	Best "bed" unit	Best "window" unit	Unit class distribution
base prog GAN 512 units total			**bed** layer4 #253 iou=0.18	**window** layer4 #142 iou=0.19	
74 object units 84 part units 9 material units	167 units	7.60			

SWD is inversely related to how good the approximation is to reality

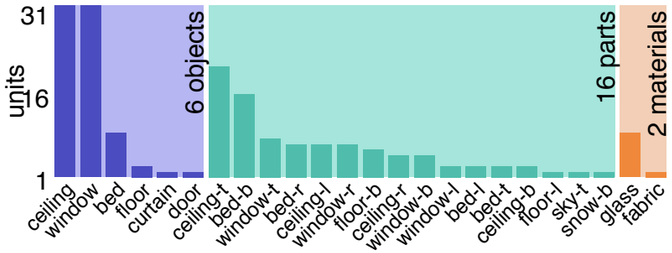
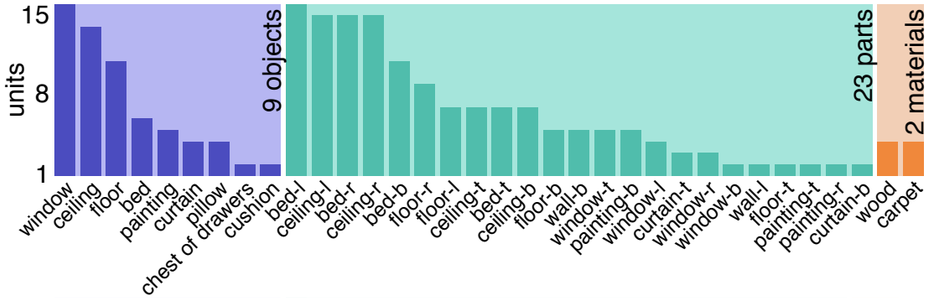
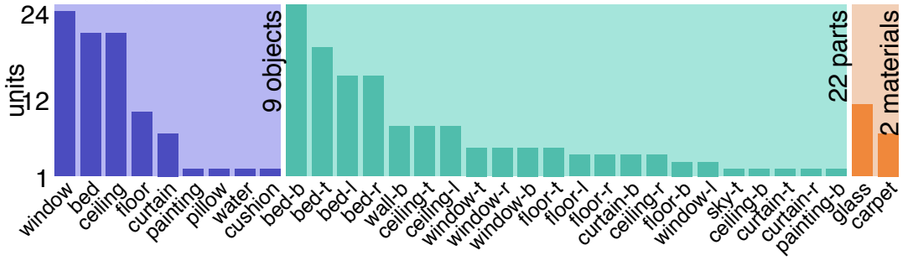
SWD (Sliced Wasserstein Distance) [Karras, et al 2017]: the **lower**, the better

GAN Dissection: Comparing Models

	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	

SWD (Sliced Wasserstein Distance) [Karras, et al 2017]: the **lower**, the better

GAN Dissection: Comparing Models

	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	

SWD (Sliced Wasserstein Distance) [Karras, et al 2017]: the **lower**, the better

GAN Dissection: Comparing Layers

Units in layer

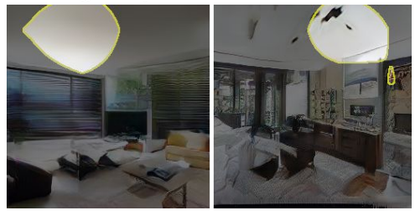
Unit class distribution

layer1
512 units total

0 object units
2 part units
0 material units

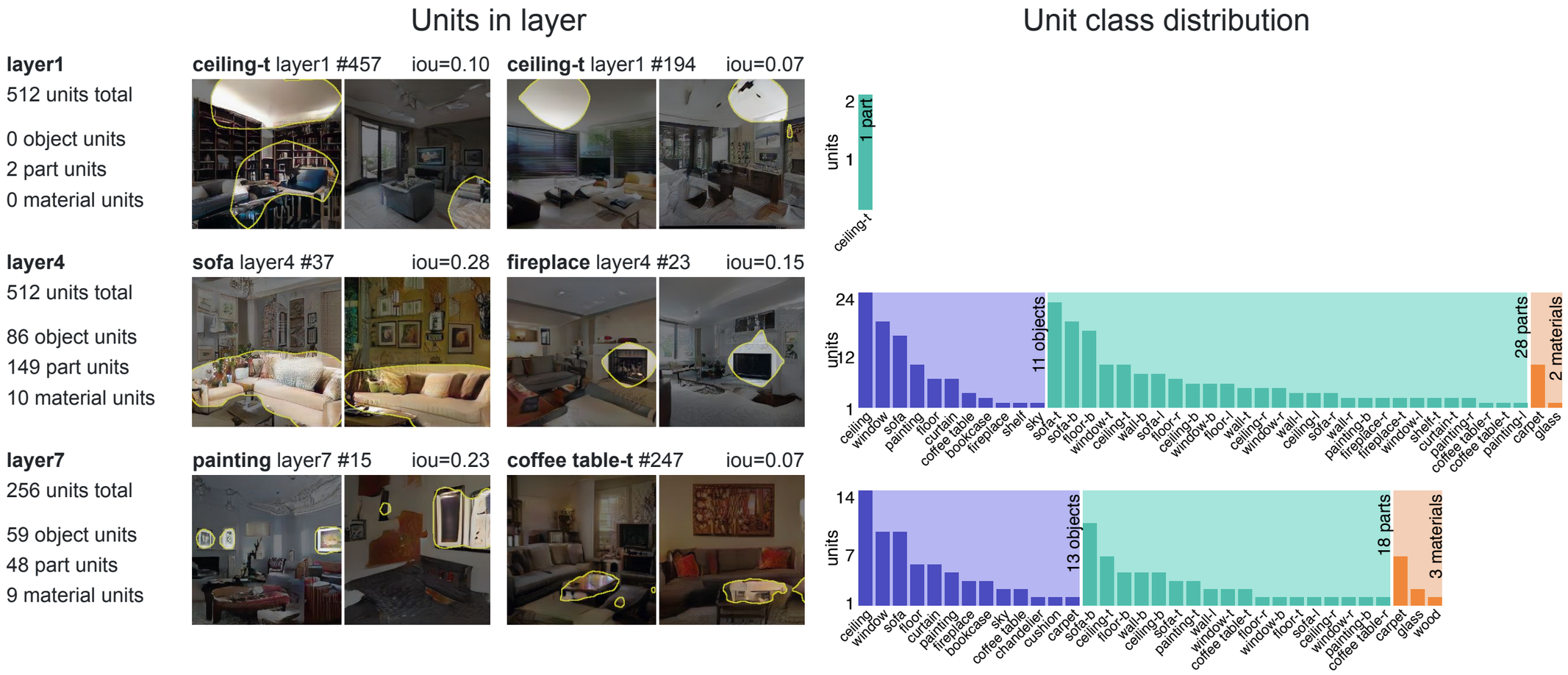
ceiling-t layer1 #457 iou=0.10

ceiling-t layer1 #194 iou=0.07



Hard to find object concepts

GAN Dissection: Comparing Layers



Objects and object parts

GAN Dissection: Comparing Layers

Units in layer

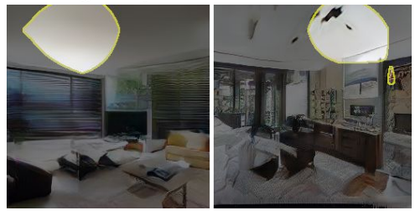
Unit class distribution

layer1
512 units total

0 object units
2 part units
0 material units

ceiling-t layer1 #457 iou=0.10

ceiling-t layer1 #194 iou=0.07

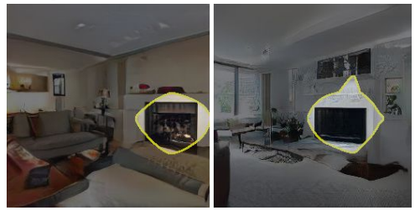


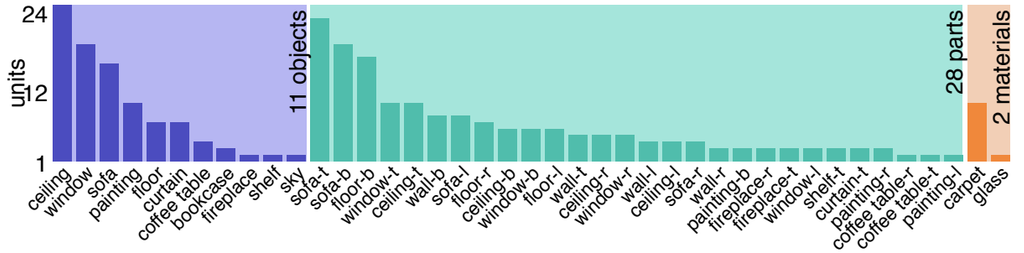
layer4
512 units total

86 object units
149 part units
10 material units

sofa layer4 #37 iou=0.28

fireplace layer4 #23 iou=0.15



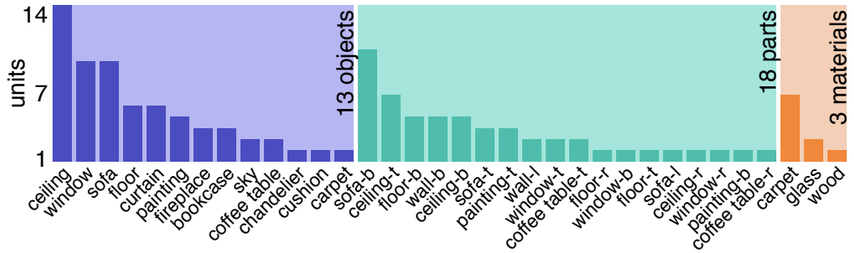


layer7
256 units total

59 object units
48 part units
9 material units

painting layer7 #15 iou=0.23

coffee table-t #247 iou=0.07

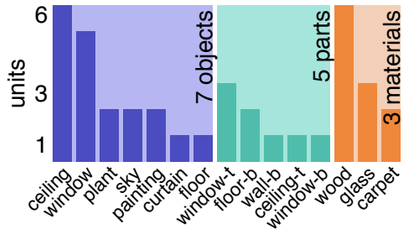


layer10
128 units total

19 object units
8 part units
11 material units

carpet layer10 #53 iou=0.14

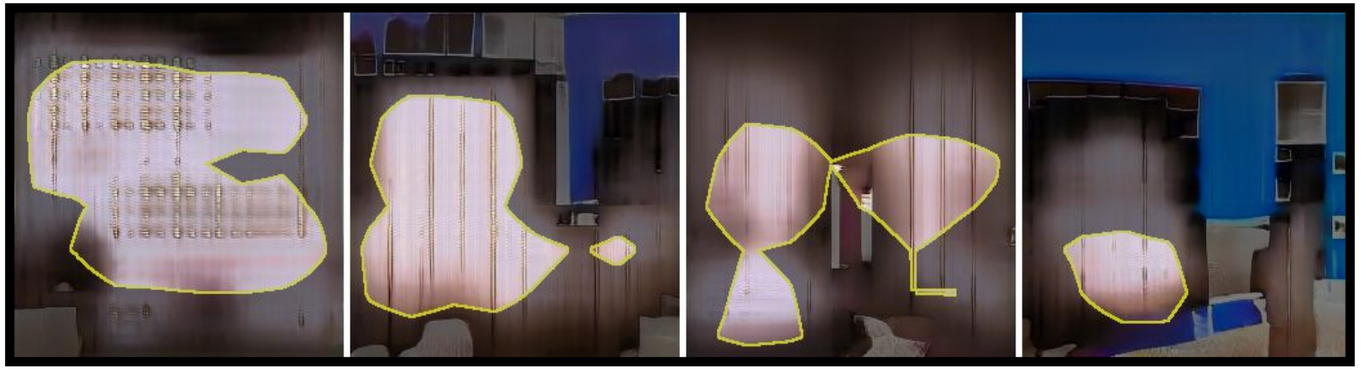
glass layer10 #126 iou=0.21



Texture, color, edges

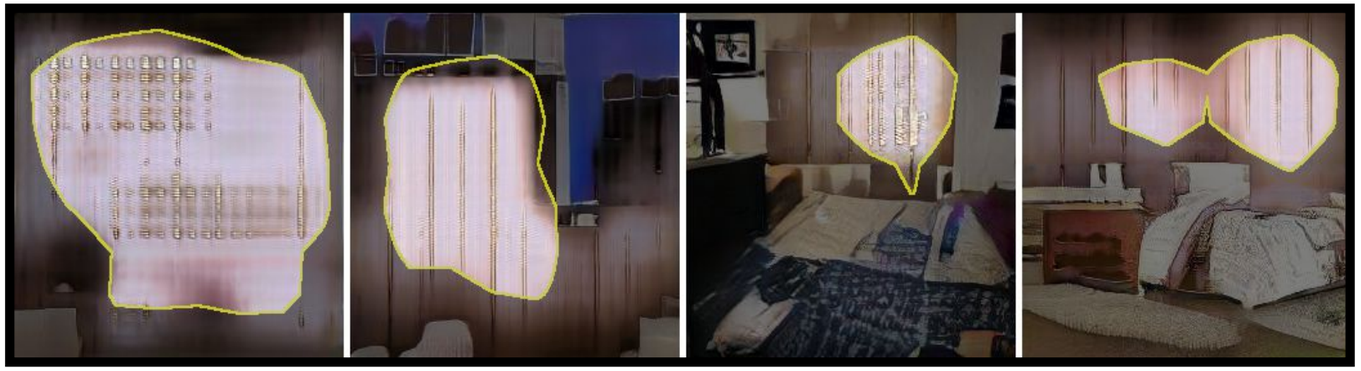
Debugging and Improving GANs

Unit #63



Bedroom images with artifacts

Unit #231



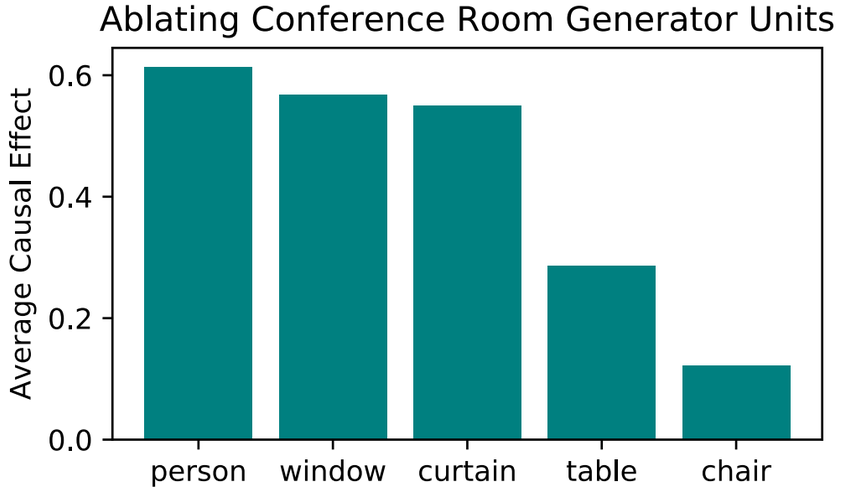
Example artifact-causing units

Ablating “artifact” units improves results

Object-Scene Relationship

ablate person units

ablate curtain units



ablate window units

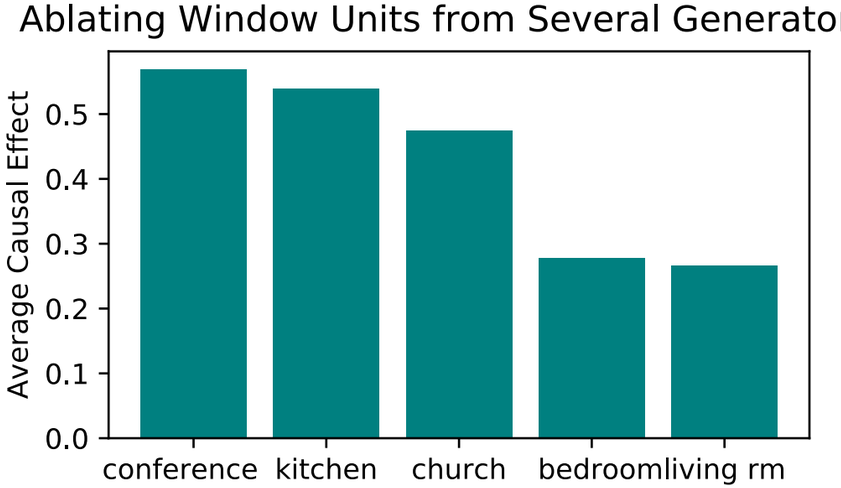
ablate table units

ablate chair units

Object-Scene Relationship

conference room

church

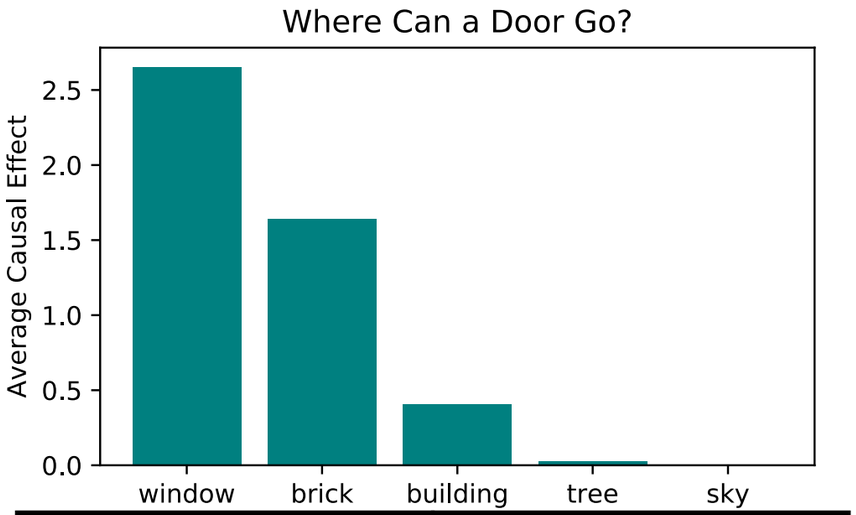


kitchen

living room

bedroom

Object-Scene Relationship



Yellow bounding box: highlight every location where we can insert doors.