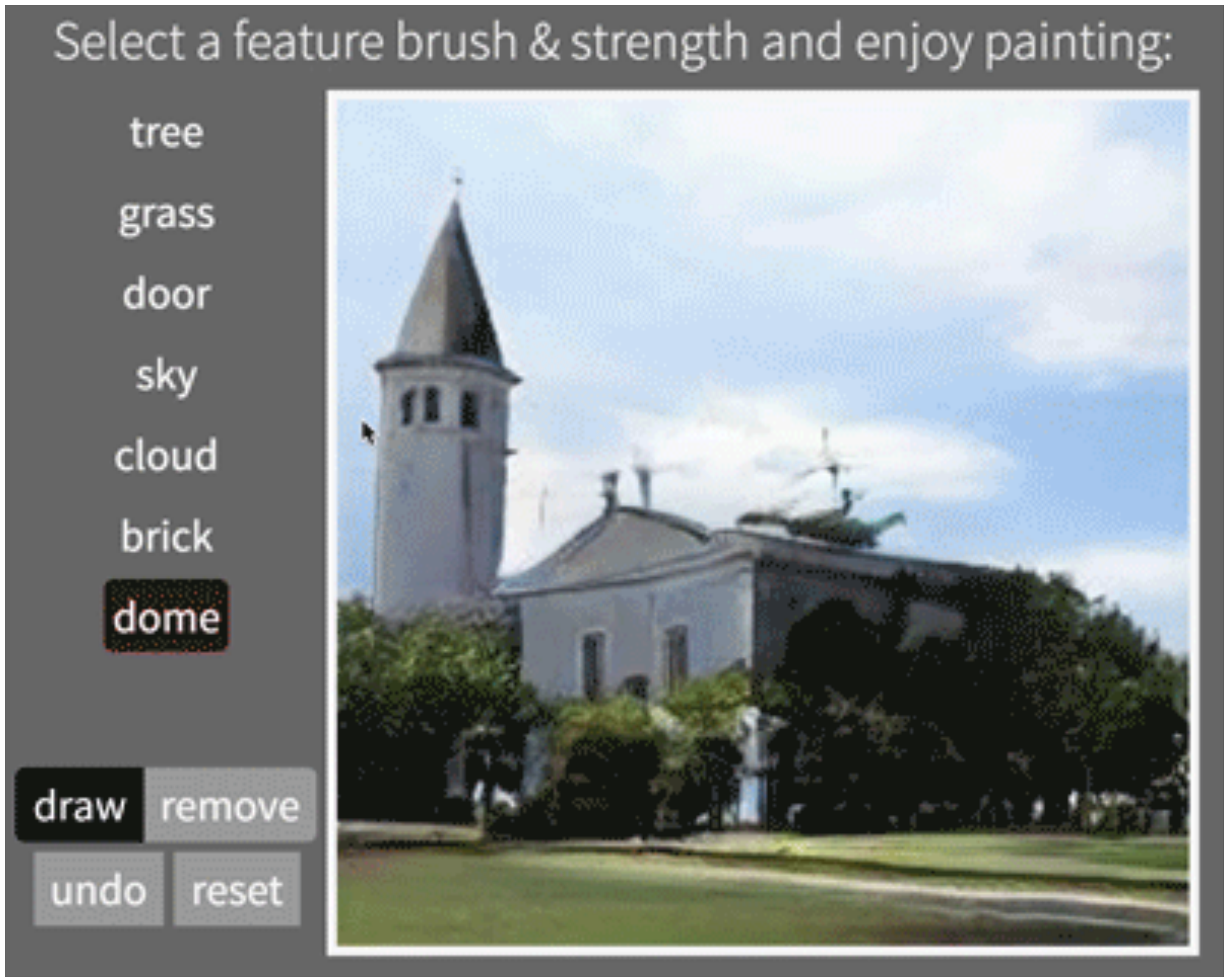
GAN DISSECTION: VISUALIZING AND UNDERSTANDING GENERATIVE ADVERSARIAL NETWORKS

David Bau[1,2], Jun-Yan Zhu[1], Hendrik Strobelt[2,3], Bolei Zhou[4], Joshua B. Tenenbaum[1], William T. Freeman[1], Antonio Torralba[1,2]
[1]Massachusetts Institute of Technology, [2]MIT-IBM Watson AI Lab,
[3]IBM Research, [4]The Chinese University of Hong Kong

Presented by Nolan Dey

Slides are heavily borrowed from https://gandissect.csail.mit.edu/slides/tutorial.pptx

Demo



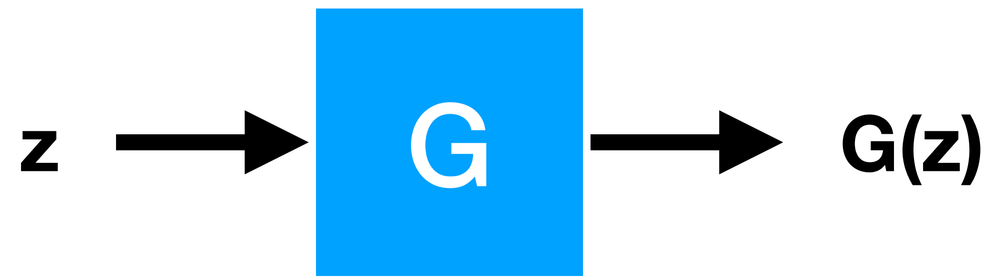
https://ganpaint.io/demo/?project=church

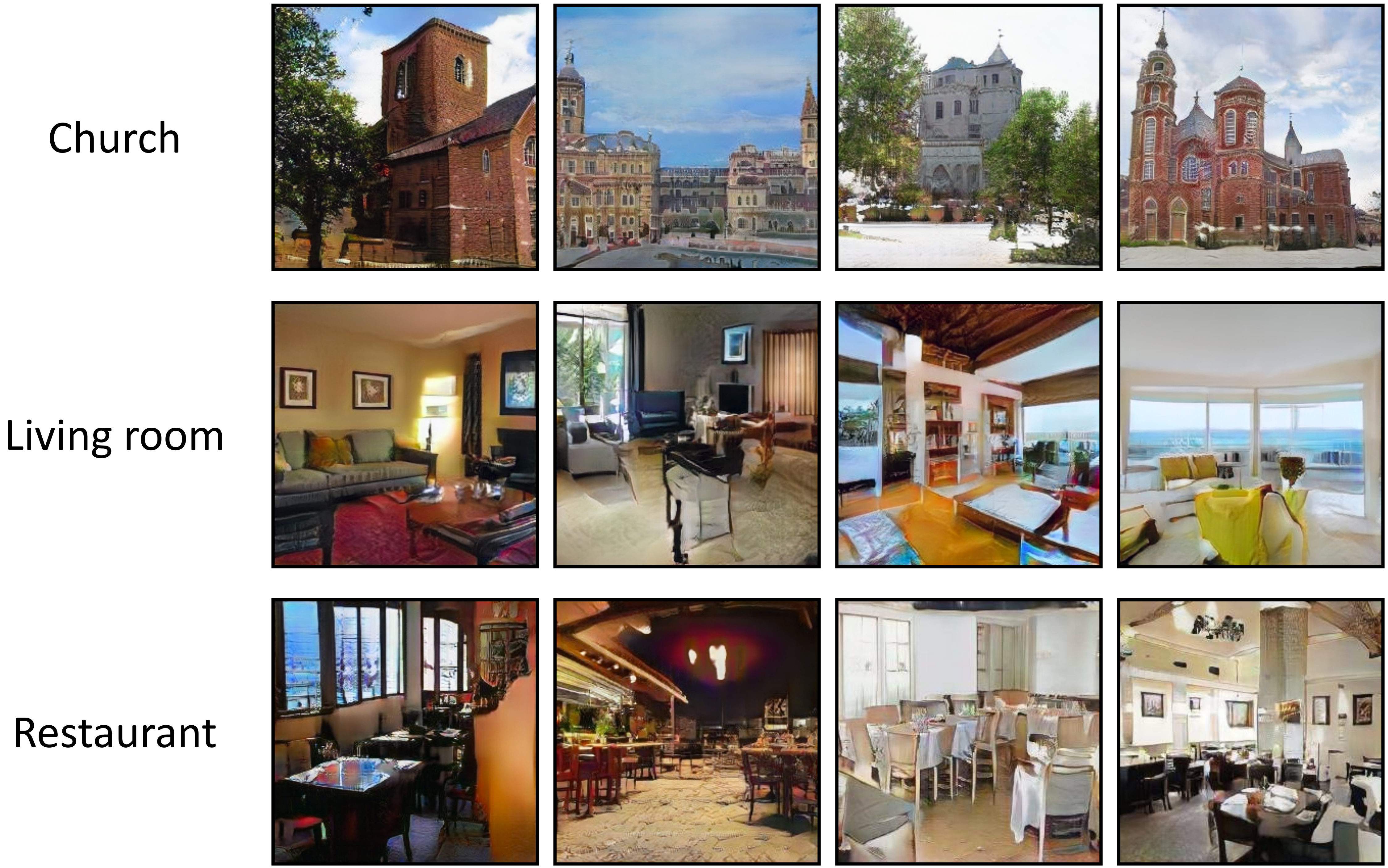
Motivation

- What do GANs learn in order to generate realistic-looking images?
    - Do they memorize pure pixel patterns?
    - Do they learn to compose a scene out of concepts it has learned to detect?

What are GANs?

- Generator G
  - Input: Latent vector z
  - Output: Generated image G(z)
- Discriminator D
  - Input: Real image x or a generated image G(x)
  - Output: Guess if input was real or generated
- Train G and D together until generated images look realistic





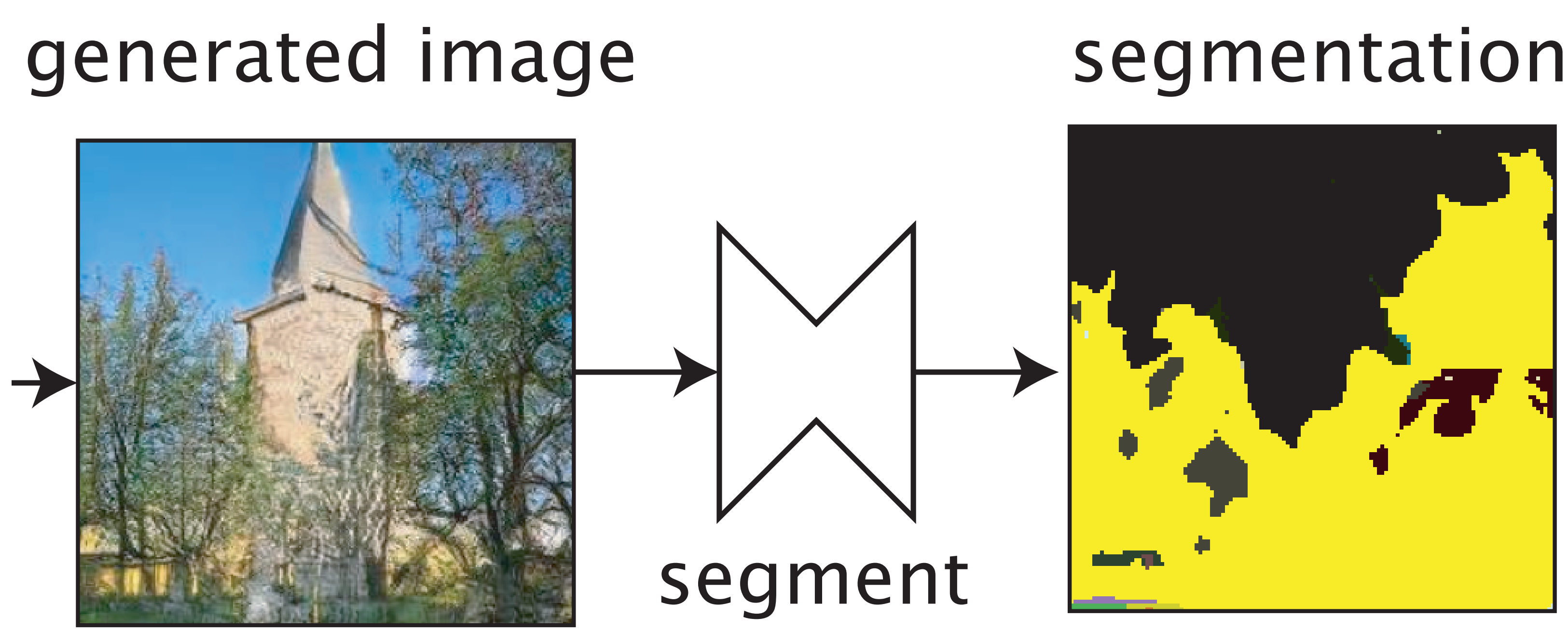
256x256 images synthesized by a Progressive GAN [Karras, et al 2017]

Method Overview

1. Dissection: What units **correlate** with a concept?
2. Intervention: What units **cause** a concept?
3. GANPaint: Add/remove visual concepts from images!

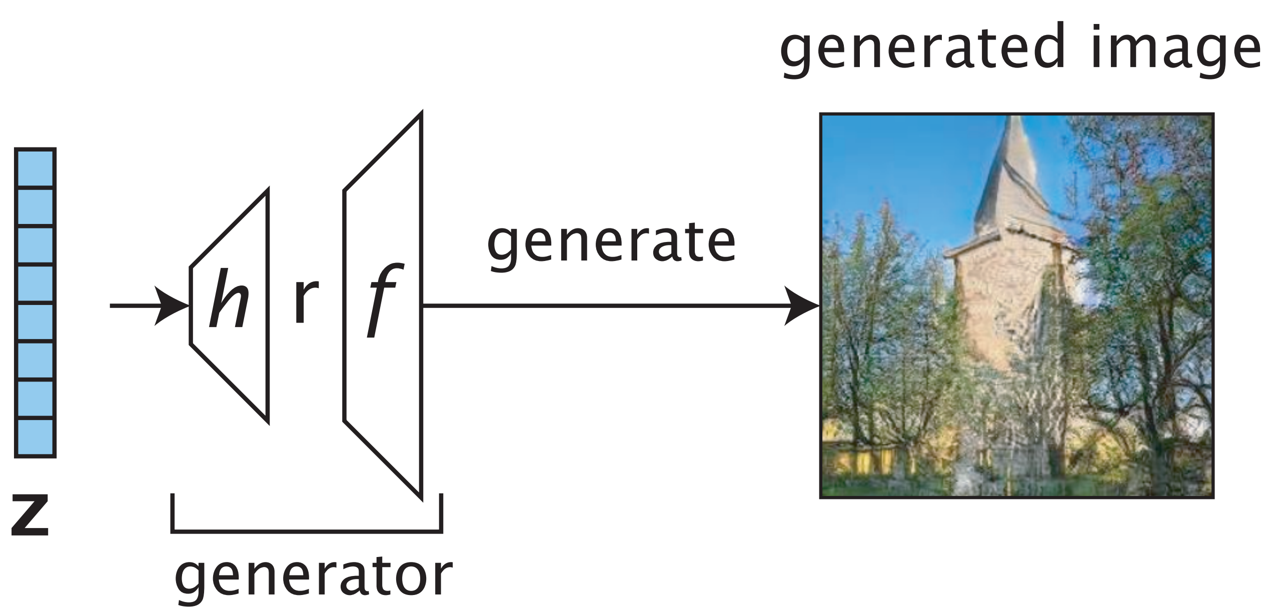
Segmentation Network

- Segmentation network was trained on the ADE20K dataset
- Outputs a pixel-wise segmentation map $S_c(x)$ for a concept c and image x
- Segments 336 objects, 29 large object parts, 25 materials

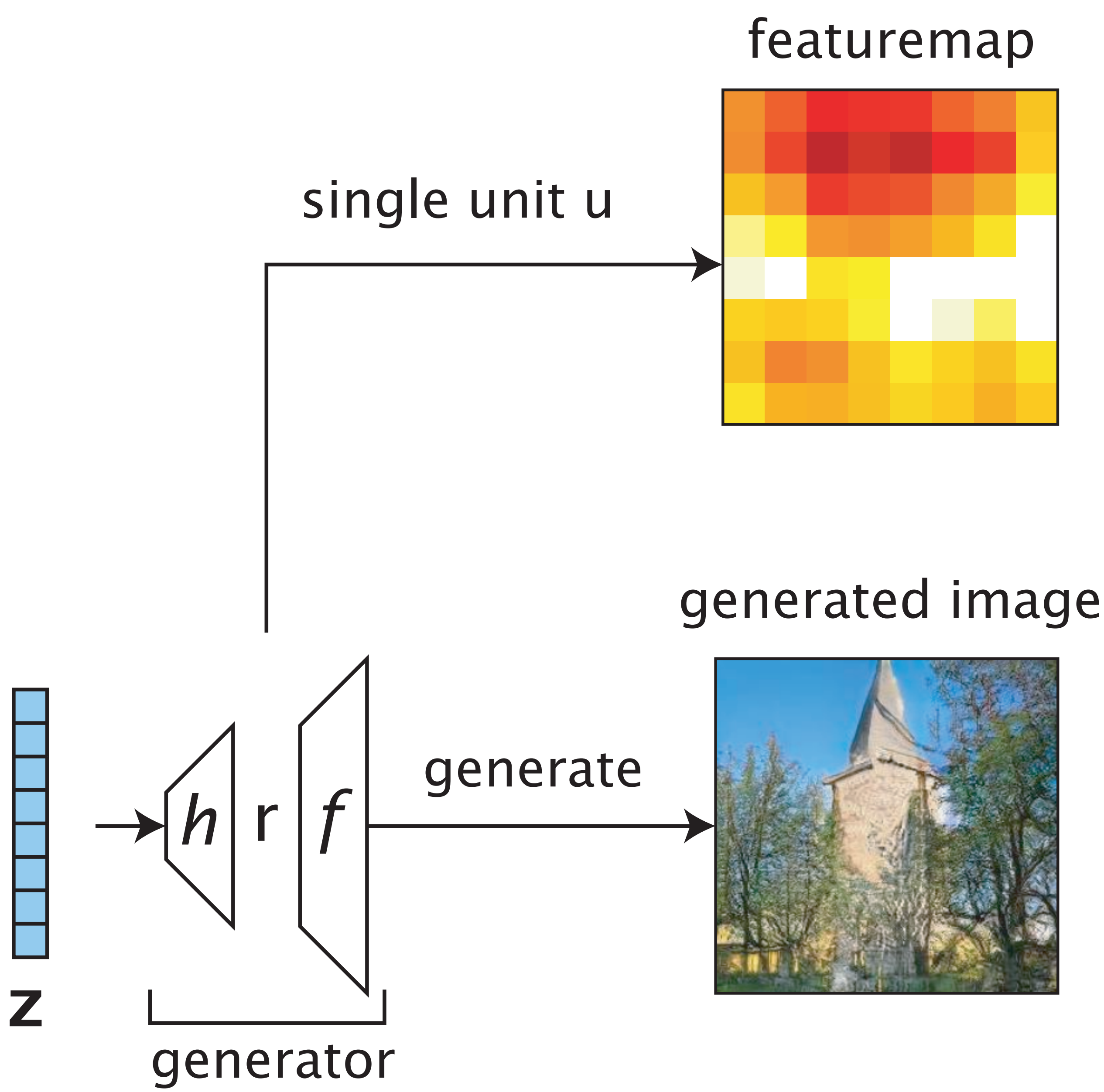


(1/3) Dissection: What units correlate with a concept?

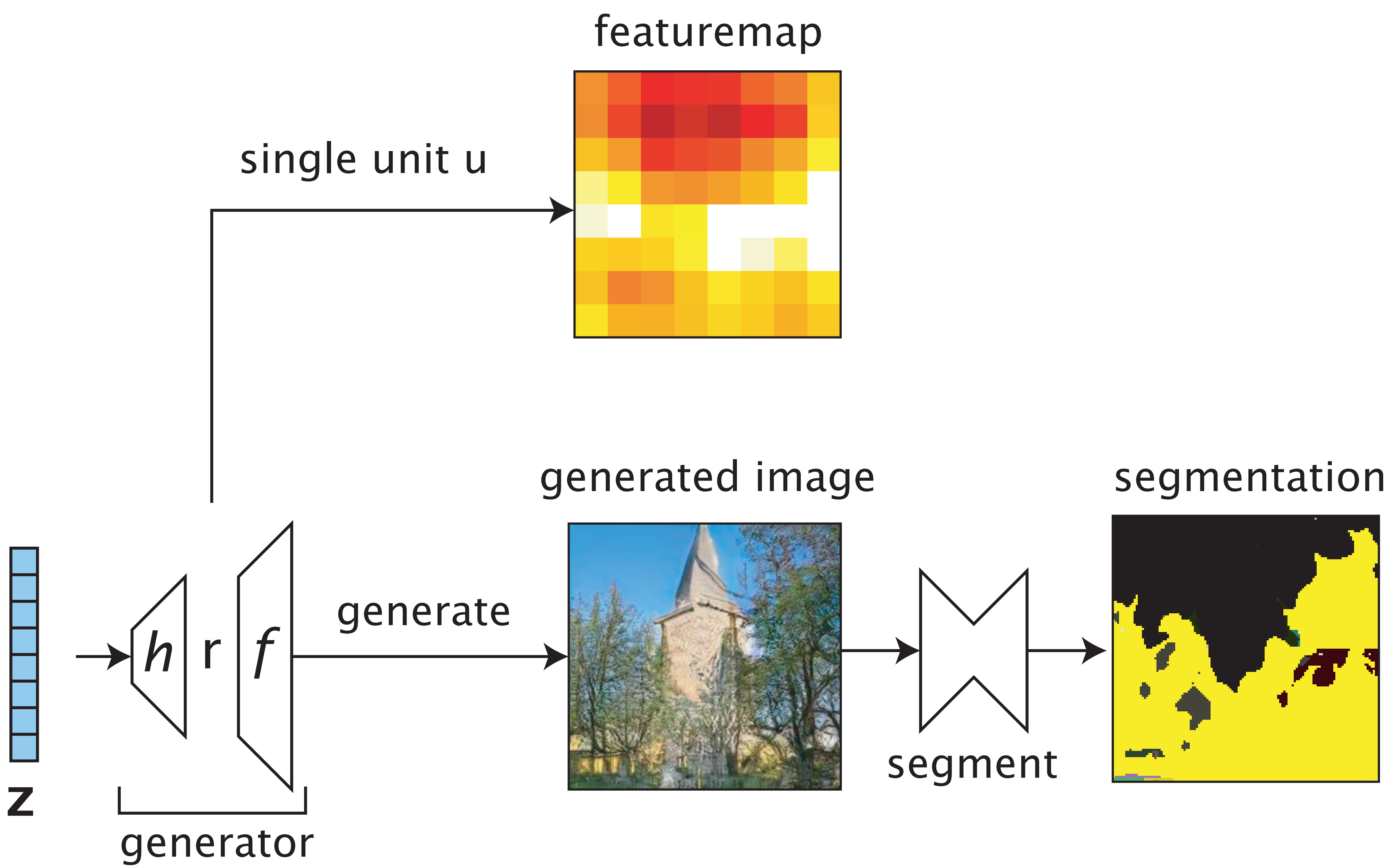
Dissection: What units correlate with a concept?



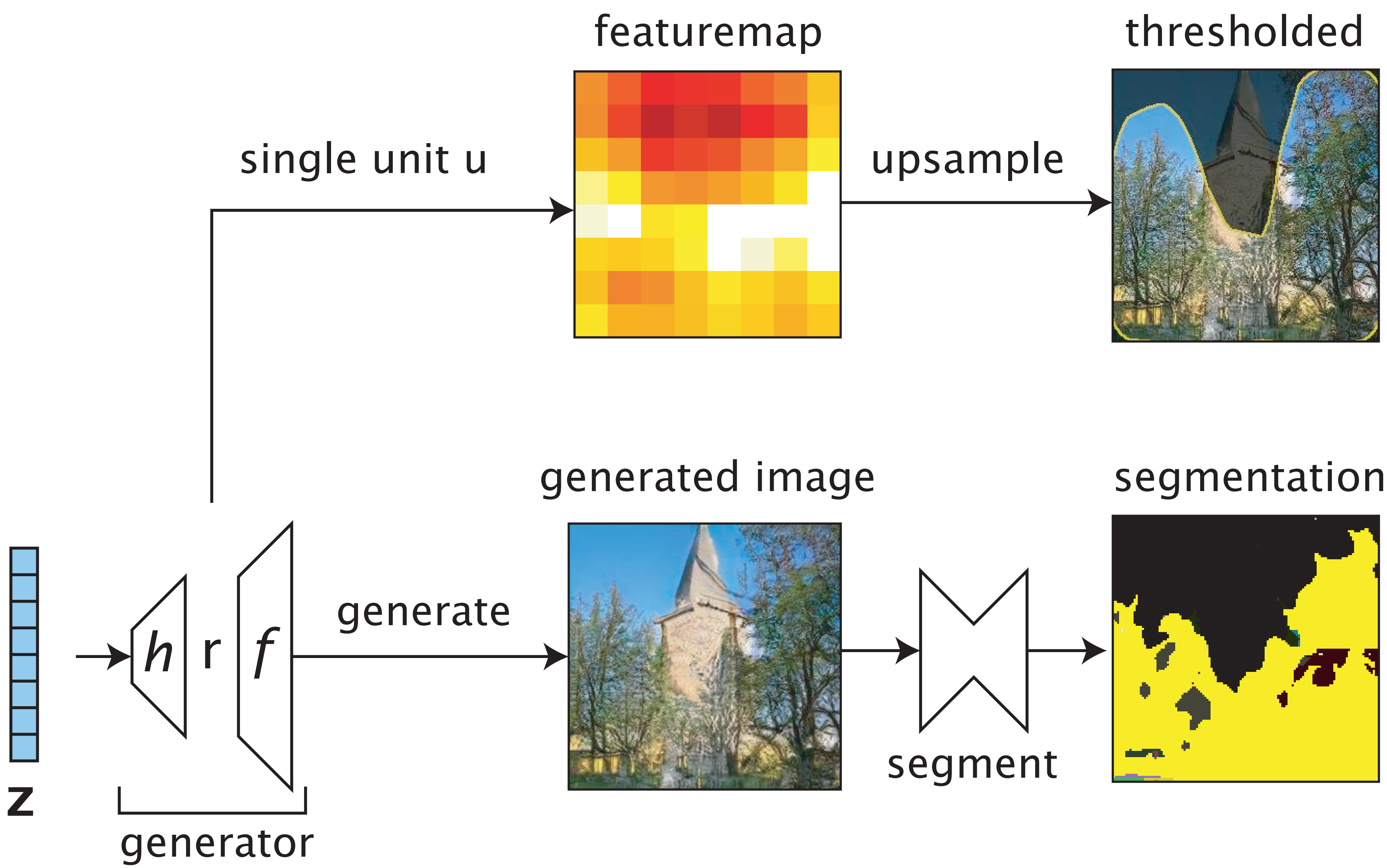
Dissection: What units correlate with a concept?



Dissection: What units correlate with a concept?

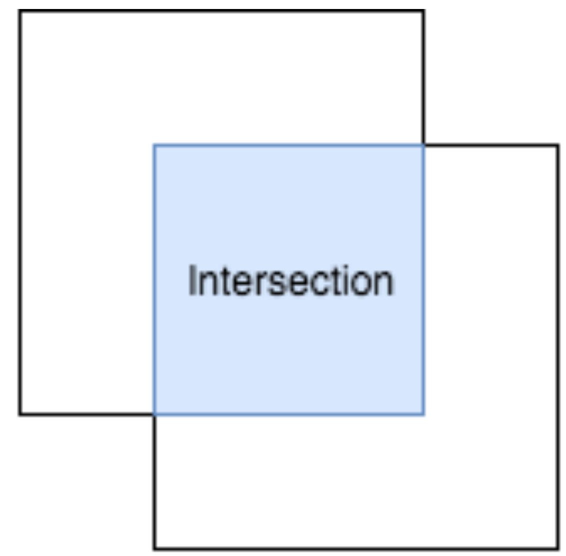


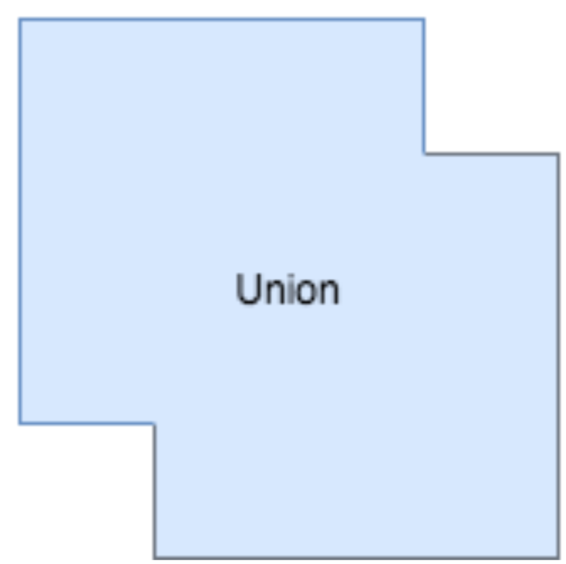
Dissection: What units correlate with a concept?



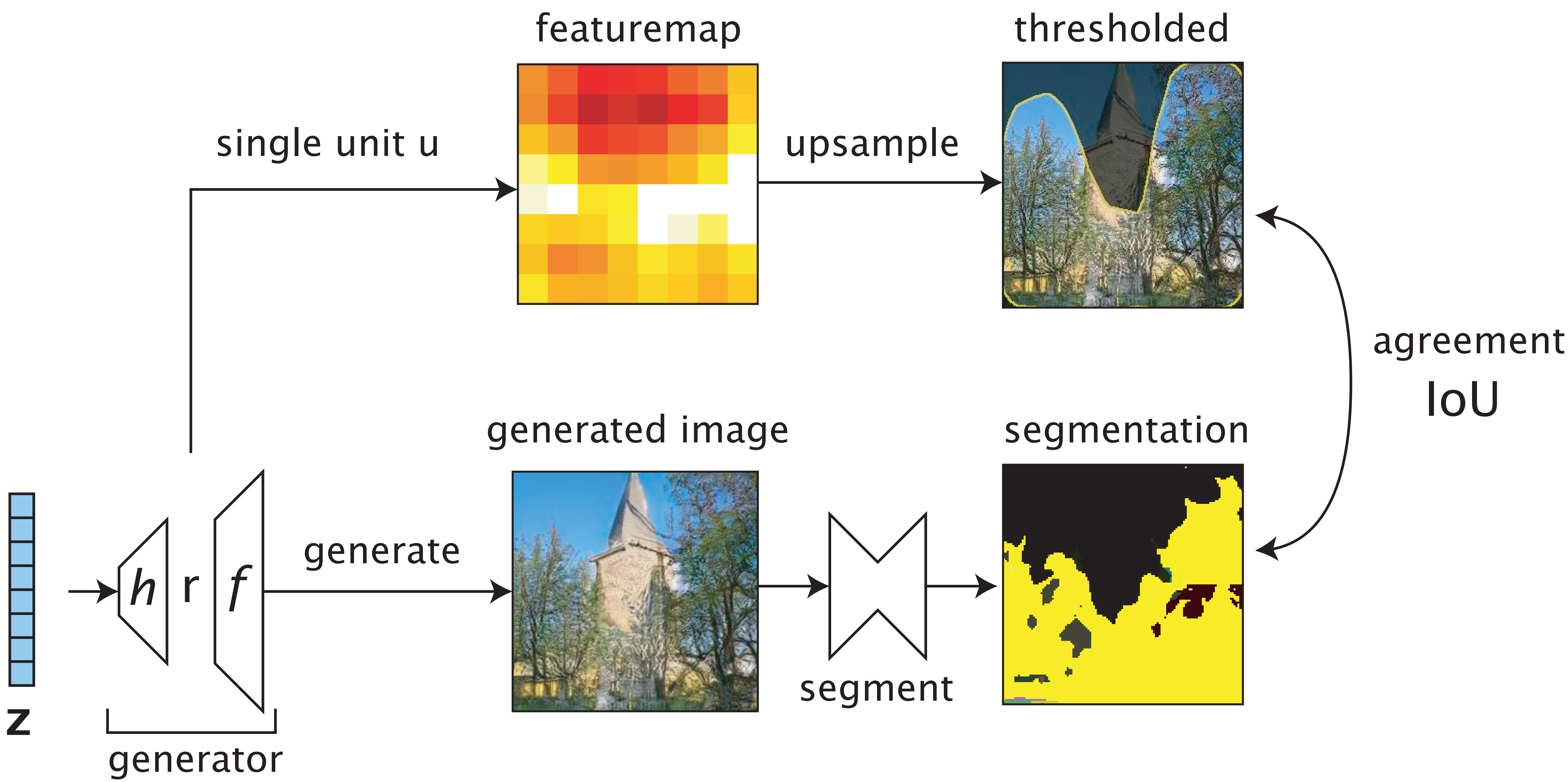
Intersection over union

IoU =

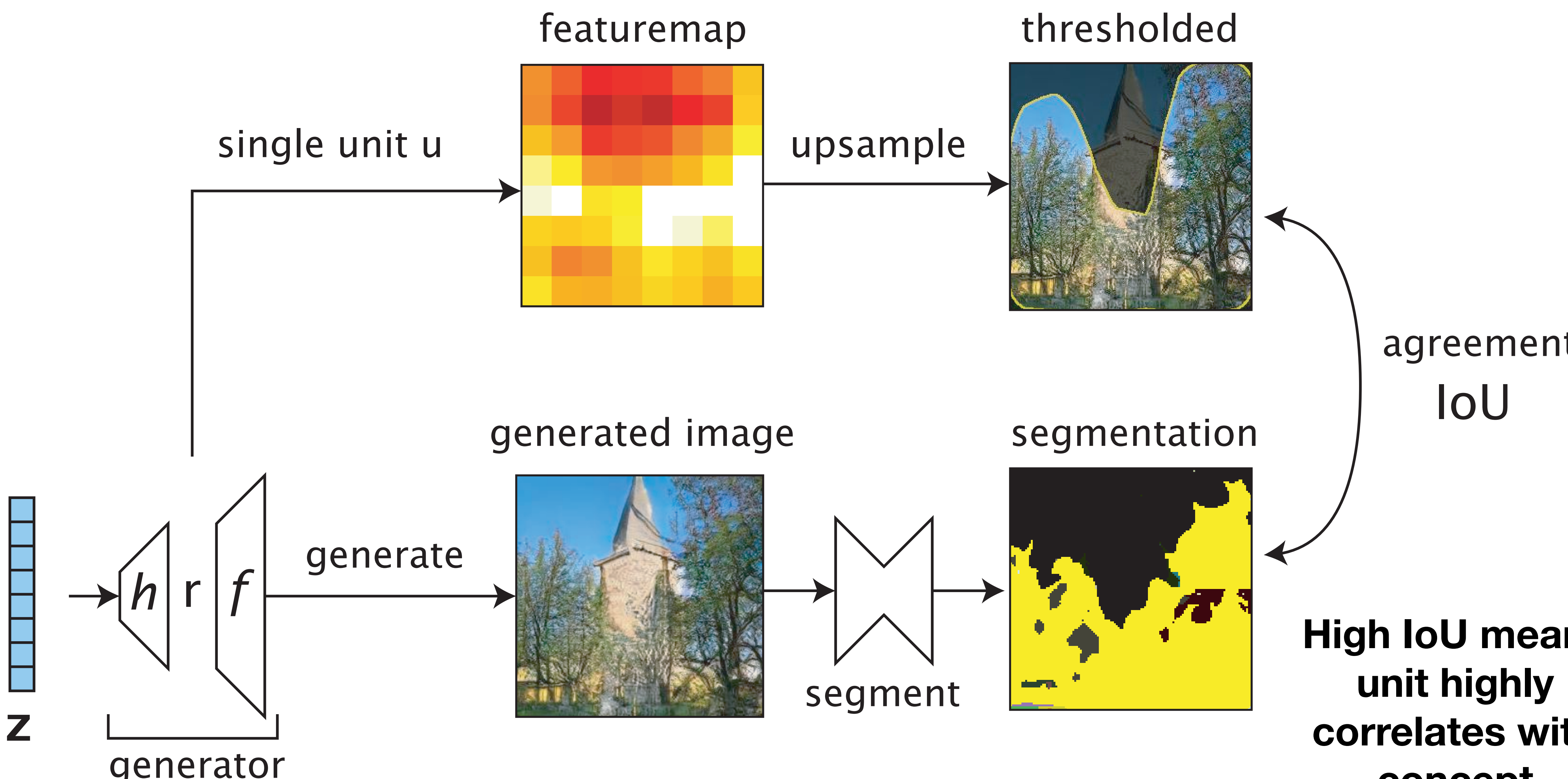




Dissection: What units correlate with a concept?



Dissection: What units correlate with a concept?



High IoU means unit highly correlates with concept

Dissection examples

Church samples

Unit #119
Tree

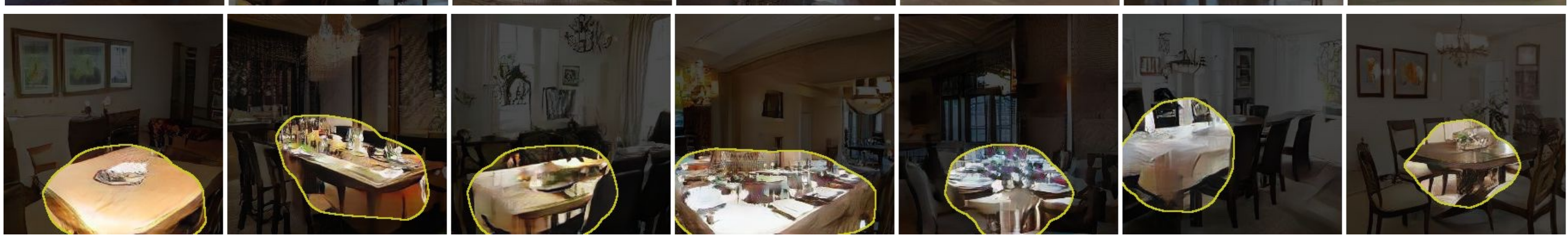
Unit #32
Dome

Dissection examples

Dining room samples

Unit #139
Window

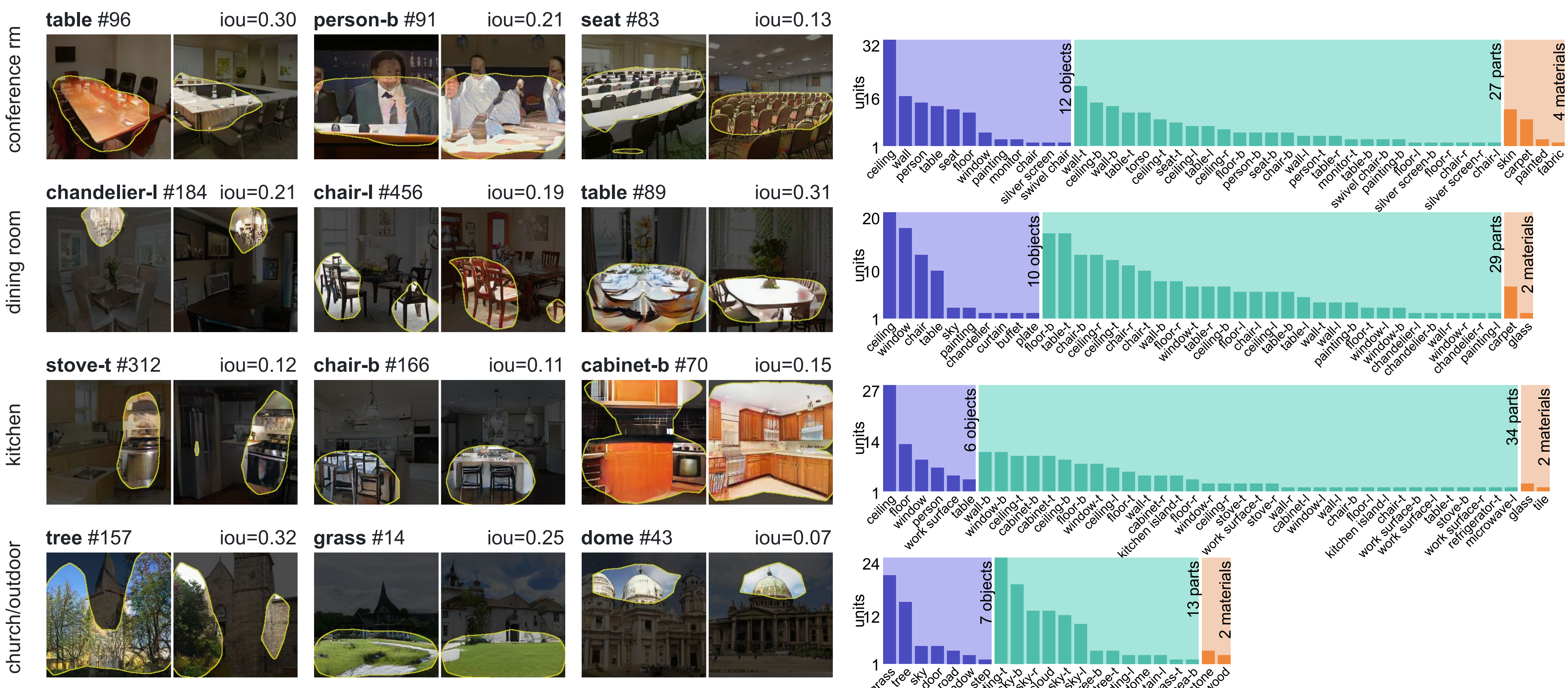
Unit #65
Table



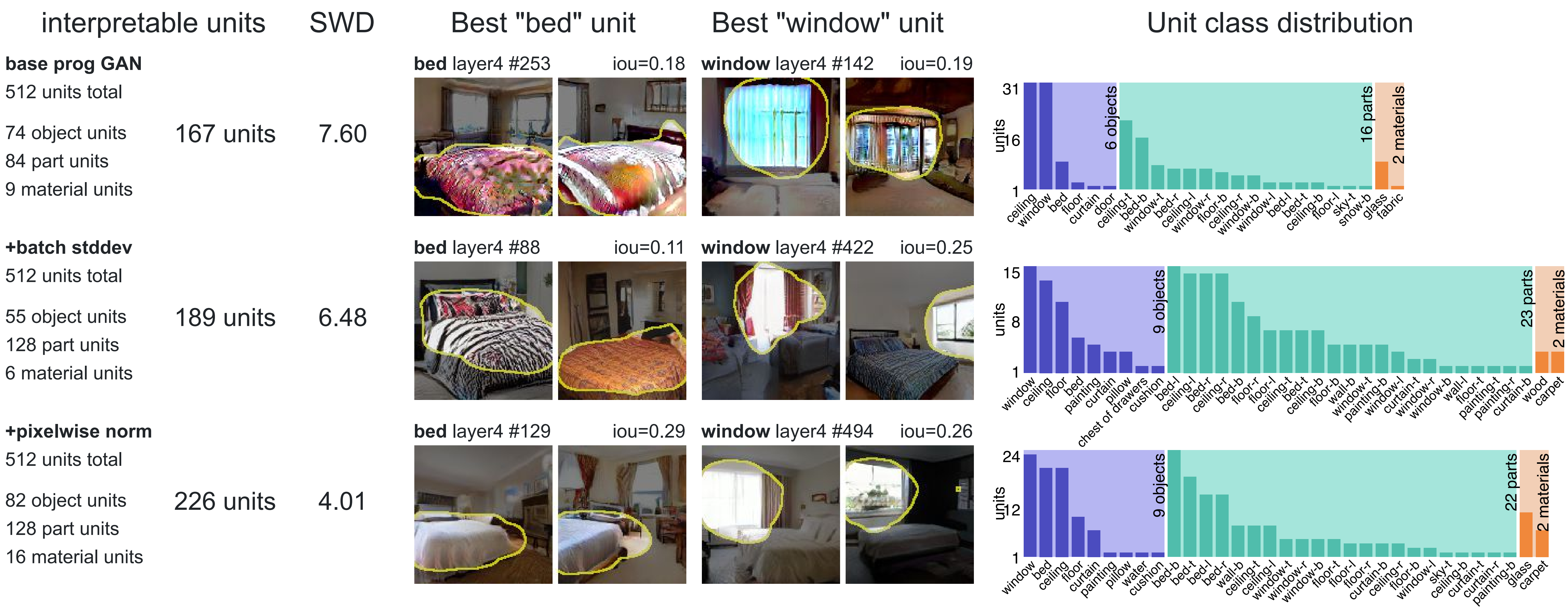
Dissection: Comparing datasets

Units in scene generator

Unit class distribution



Dissection: Comparing models



Dissection: Comparing layers

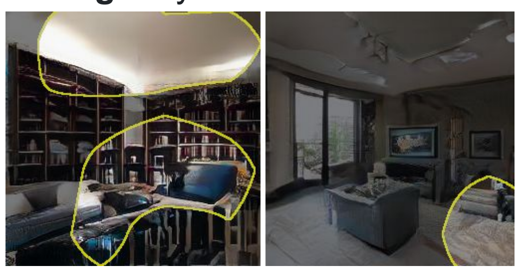
Units in layer

Unit class distribution

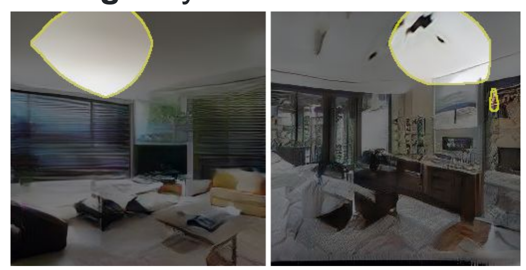
layer1
512 units total

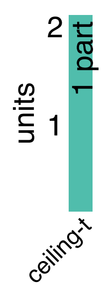
0 object units
2 part units
0 material units

ceiling-t layer1 #457 iou=0.10



ceiling-t layer1 #194 iou=0.07



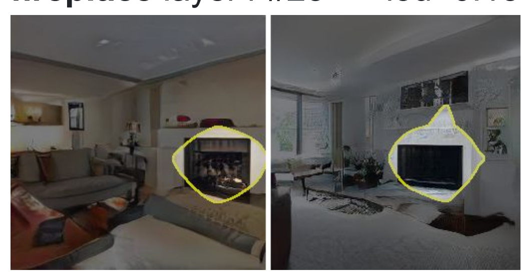


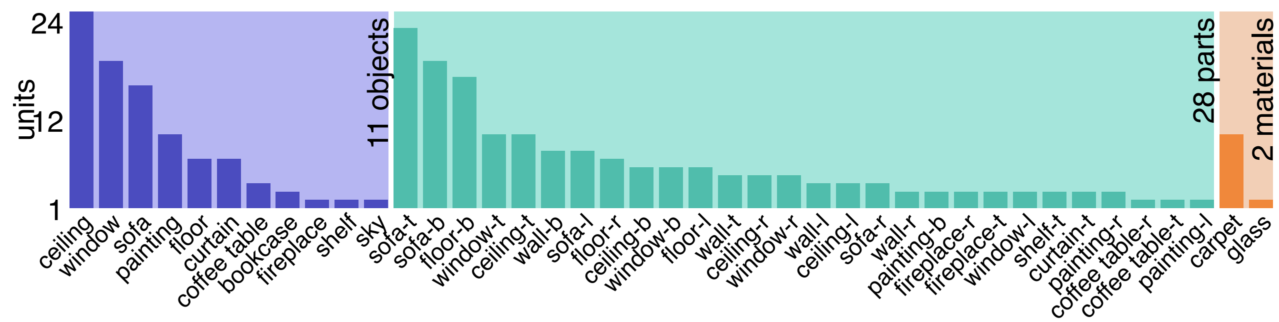
layer4
512 units total

86 object units
149 part units
10 material units

sofa layer4 #37 iou=0.28

fireplace layer4 #23 iou=0.15

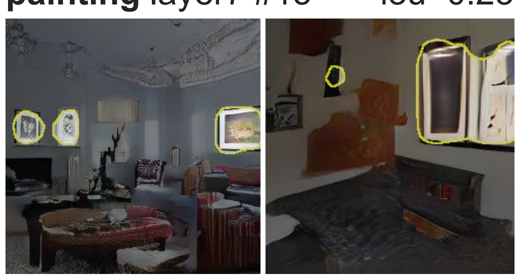




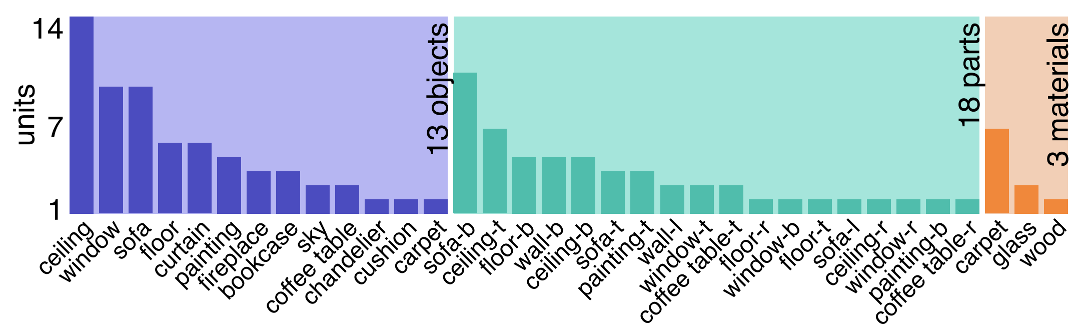
layer7
256 units total

59 object units
48 part units
9 material units

painting layer7 #15 iou=0.23



coffee table-t #247 iou=0.07

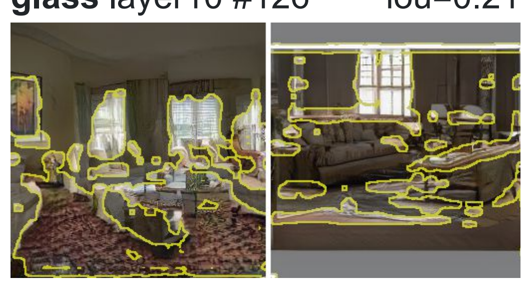


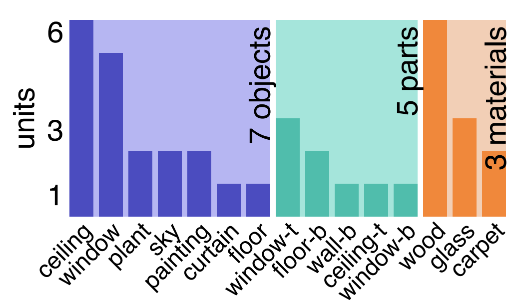
layer10
128 units total

19 object units
8 part units
11 material units

carpet layer10 #53 iou=0.14

glass layer10 #126 iou=0.21

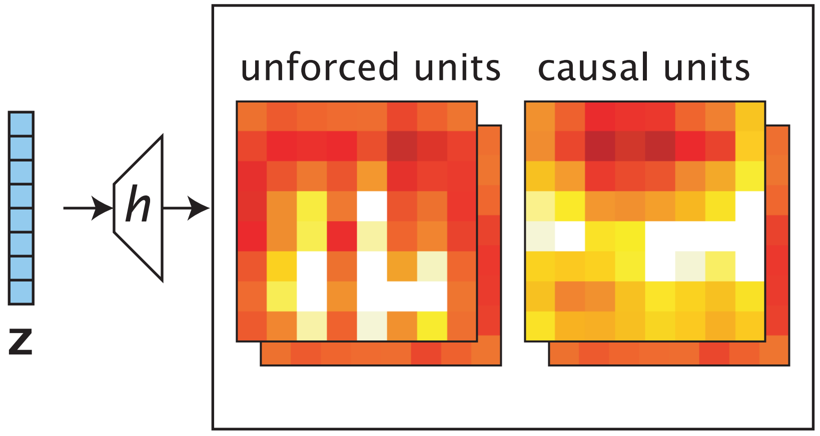




(2/3) Intervention: What units cause a concept?

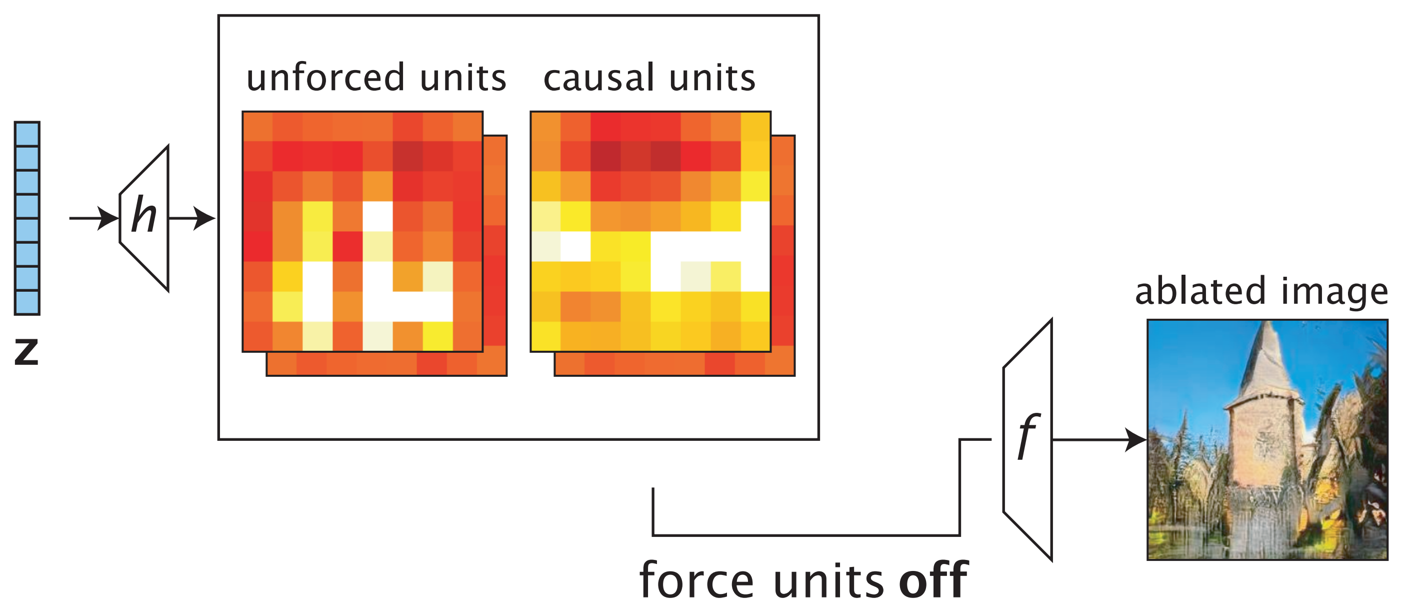
Intervention: What units cause a concept?

- Choose ~20 units at a layer that we suspect are causal (based on dissection results)



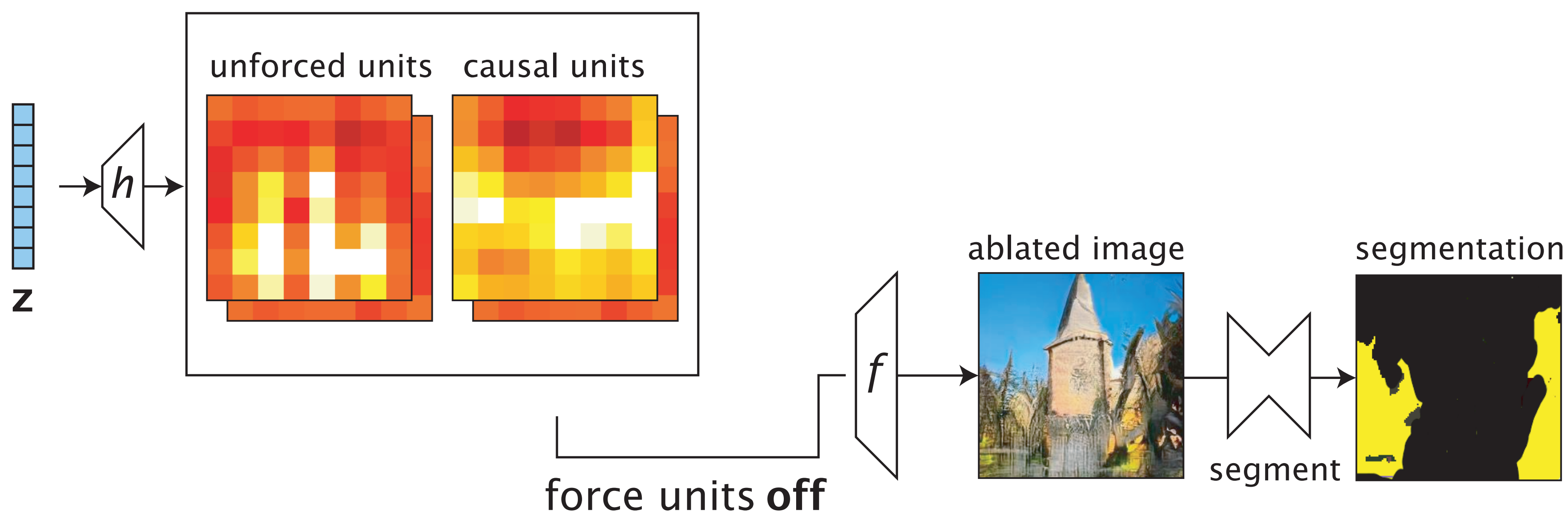
Intervention: What units cause a concept?

- Force feature map of suspected causal units to 0 and forward propagate to obtain ablated image x_a



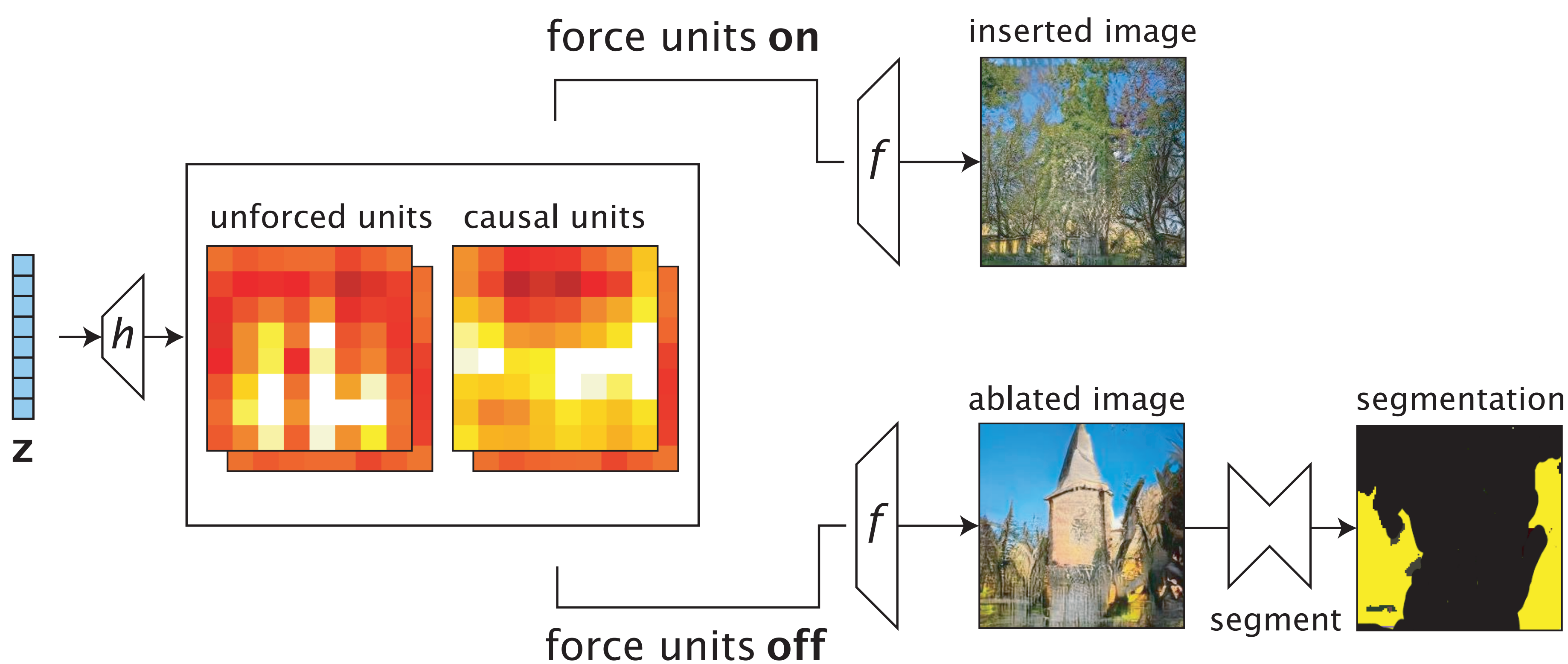
Intervention: What units cause a concept?

- Obtain segmentation $S_c(x_a)$



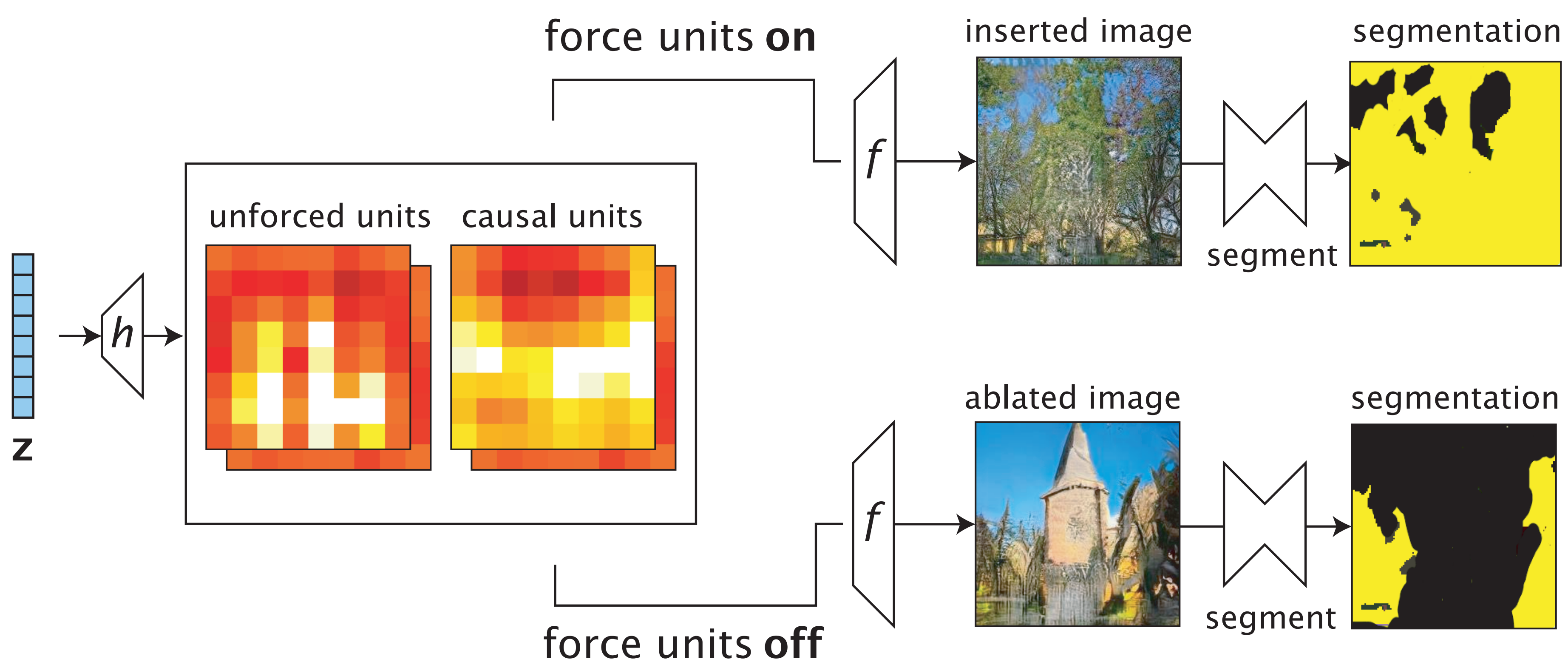
Intervention: What units cause a concept?

- Force feature map of suspected causal units to positive value and forward propagate to obtain inserted image x_i



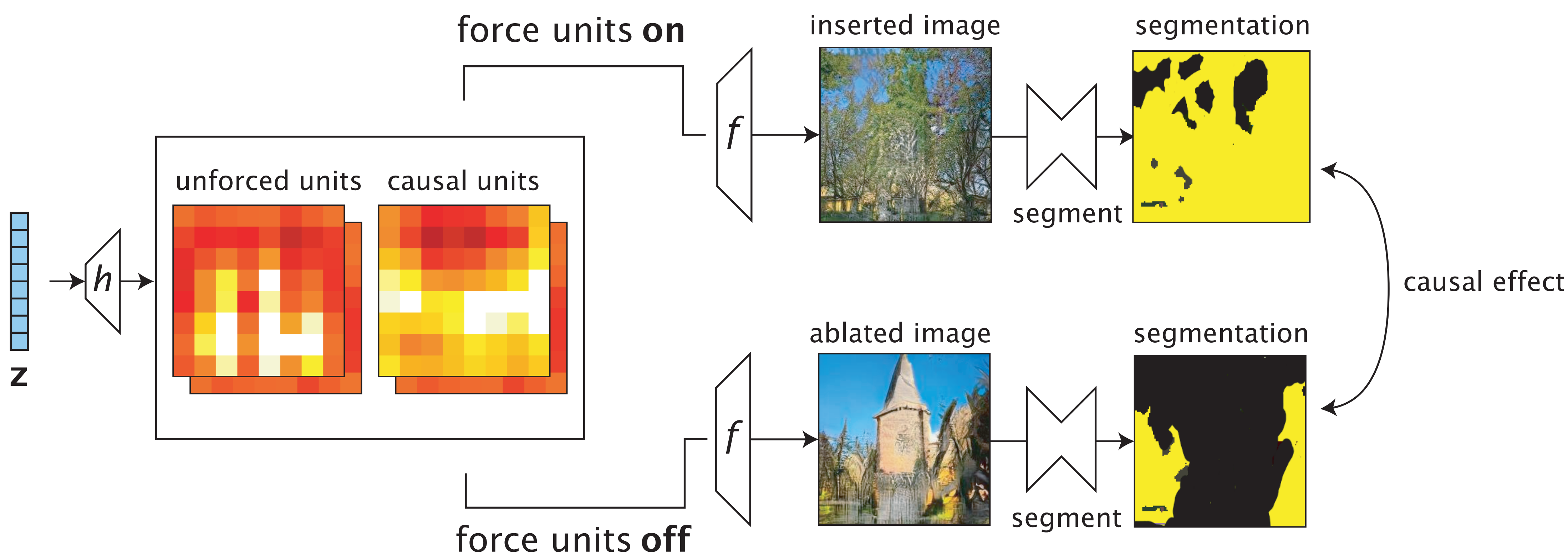
Intervention: What units cause a concept?

- Obtain segmentation $S_c(x_i)$



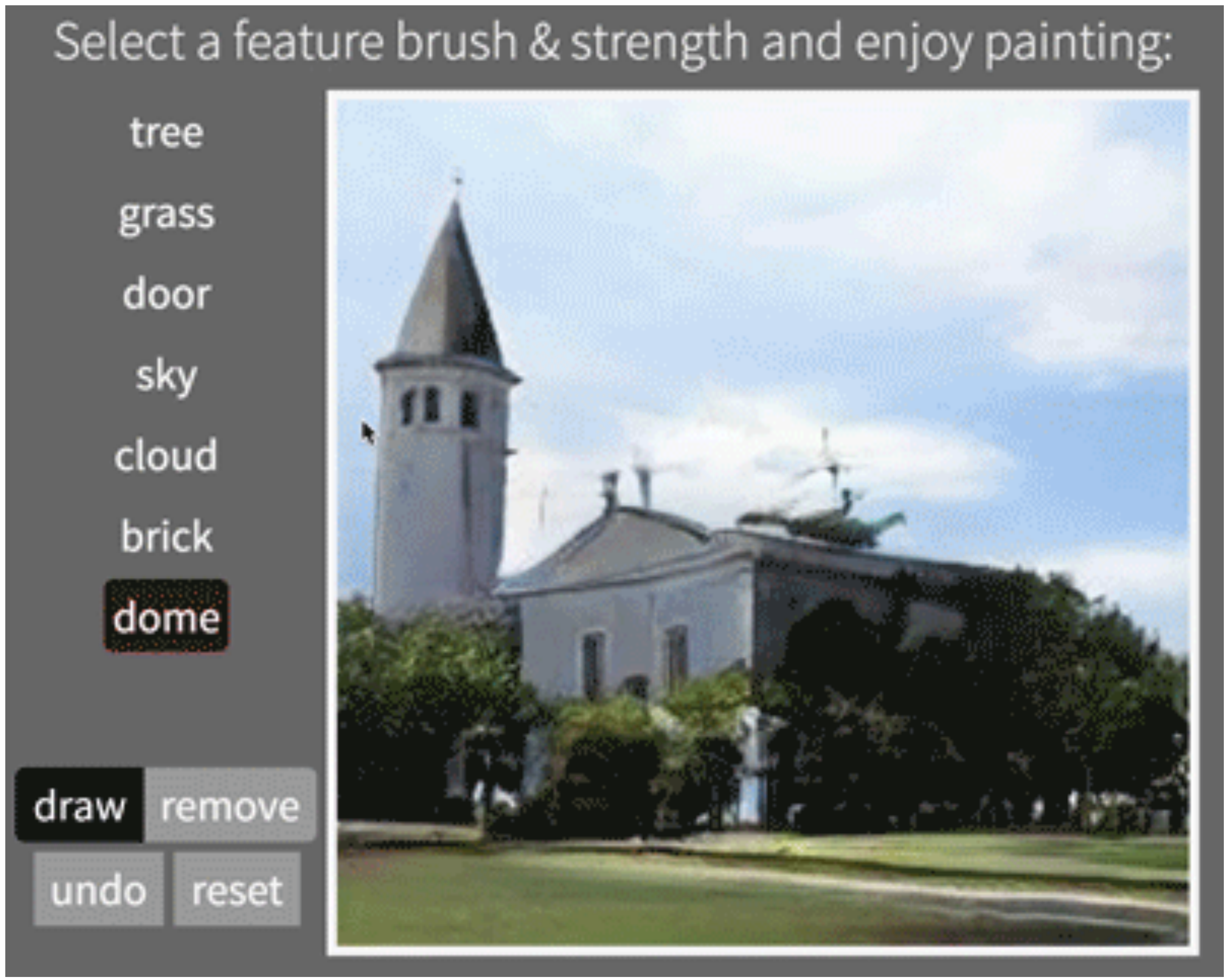
Intervention: What units cause a concept?

- Average causal effect: $\delta_{u\rightarrow c} = \mathbb{E}[S_c(x_i)] - \mathbb{E}[S_c(x_a)]$
- Average $\delta_{u\rightarrow c}$ over many images
- Use optimization to find how strongly each unit should be inserted or ablated to cause a concept



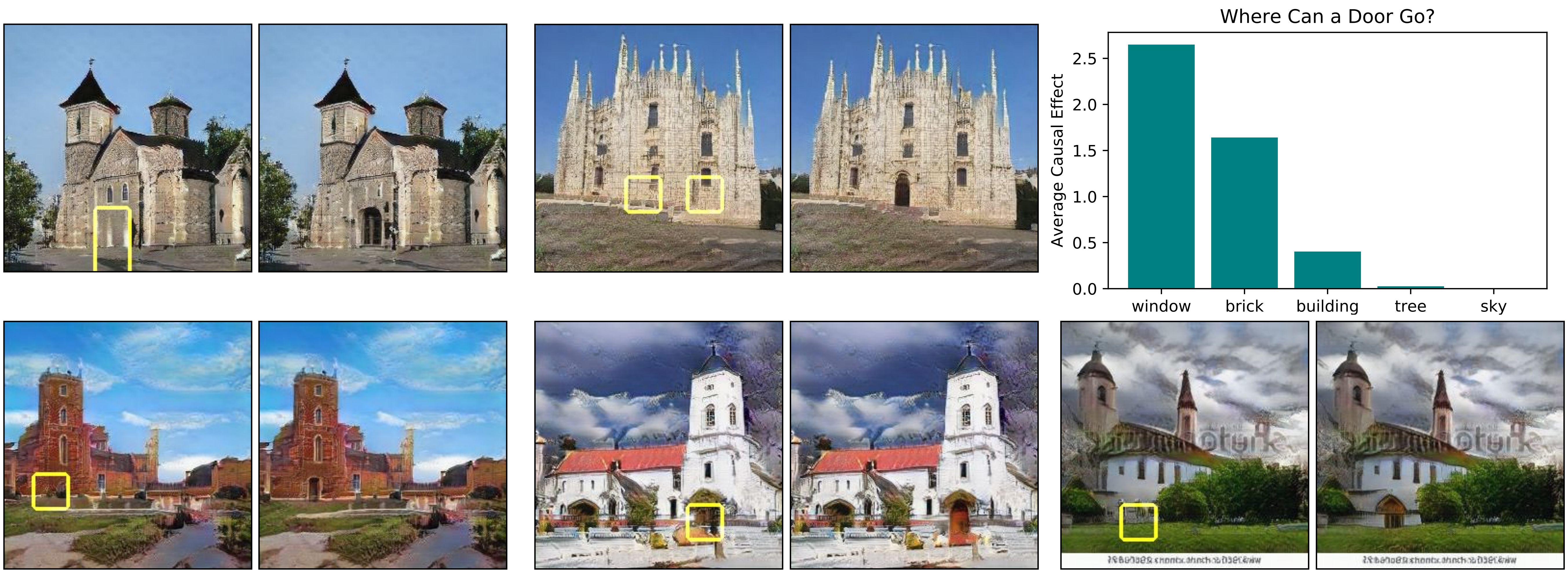
(3/3) GANPaint: Add/ remove visual concepts from images!

Demo



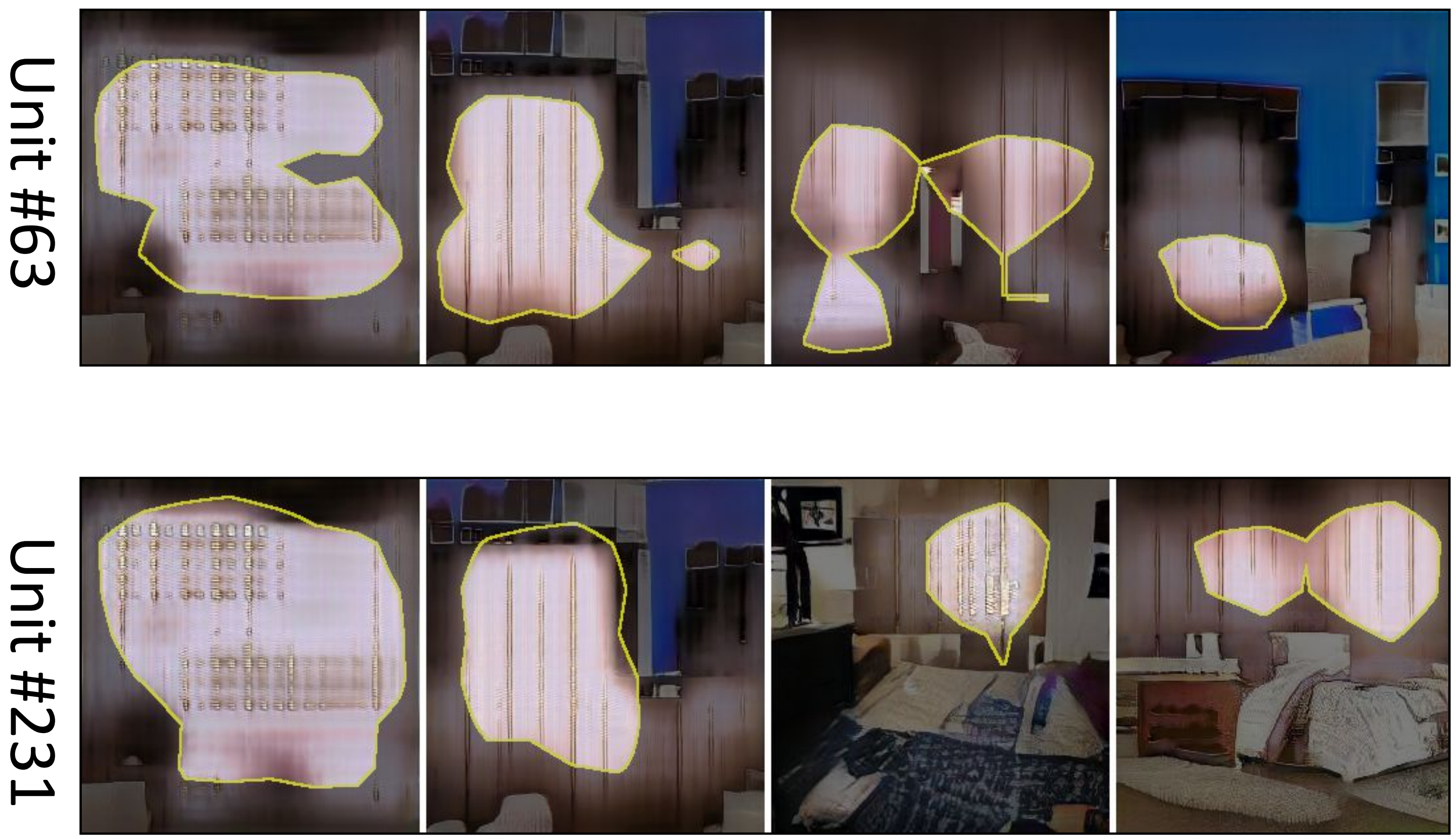
https://ganpaint.io/demo/?project=church

Object-scene relationship



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

Removing artifacts



Example artifact-causing units

Bedroom images with artifacts

Ablating "artifact" units improves results

Thank you

Extra slides

Intervention: What units cause a concept?

Finding sets of units with high ACE. Given a representation $\mathbf{r}$ with d units, exhaustively searching for a fixed-size set U with high $\delta_{\mathrm{U}\to c}$ is prohibitive as it has $\binom{d}{|\mathrm{U}|}$ subsets. Instead, we optimize a continuous intervention $\boldsymbol{\alpha} \in [0,1]^d$, where each dimension $\boldsymbol{\alpha}_u$ indicates the degree of intervention for a unit u. We maximize the following average causal effect formulation $\delta_{\boldsymbol{\alpha}\to c}$:

$$\text{Image with partial ablation at pixels P}: \quad \mathbf{x}'_a = f((\mathbf{1}-\boldsymbol{\alpha}) \odot \mathbf{r}_{\mathbb{U},\mathrm{P}},\ \mathbf{r}_{\mathbb{U},\overline{\mathrm{P}}}) \tag{5}$$

$$\text{Image with partial insertion at pixels P}: \quad \mathbf{x}'_i = f(\boldsymbol{\alpha} \odot \mathbf{k} + (\mathbf{1}-\boldsymbol{\alpha}) \odot \mathbf{r}_{\mathbb{U},\mathrm{P}},\ \mathbf{r}_{\mathbb{U},\overline{\mathrm{P}}})$$

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

where $\mathbf{r}_{\mathbb{U},\mathrm{P}}$ denotes the all-channel featuremap at locations P, $\mathbf{r}_{\mathbb{U},\overline{\mathrm{P}}}$ denotes the all-channel featuremap at other locations $\overline{\mathrm{P}}$, and $\odot$ applies a per-channel scaling vector $\boldsymbol{\alpha}$ to the featuremap $\mathbf{r}_{\mathbb{U},\mathrm{P}}$. We optimize

$\boldsymbol{\alpha}$ over the following loss with an L2 regularization:

$$\boldsymbol{\alpha}^* = \arg\min_{\boldsymbol{\alpha}}(-\delta_{\boldsymbol{\alpha}\to c} + \lambda||\boldsymbol{\alpha}||_2), \tag{6}$$

where λ controls the relative importance of each term. We add the L2 loss as we seek a minimal set of casual units. We optimize using stochastic gradient descent, sampling over both $\mathbf{z}$ and featuremap locations P and clamping the coefficient $\boldsymbol{\alpha}$ within the range $[0,1]^d$ at each step (d is the total number of units). More details of this optimization are discussed in Section S-6.4. Finally, we can rank units by $\boldsymbol{\alpha}^*_u$ and achieve a stronger causal effect (i.e., removing trees) when ablating successively larger sets of tree-causing units as shown in Figure 4.

Network Dissection

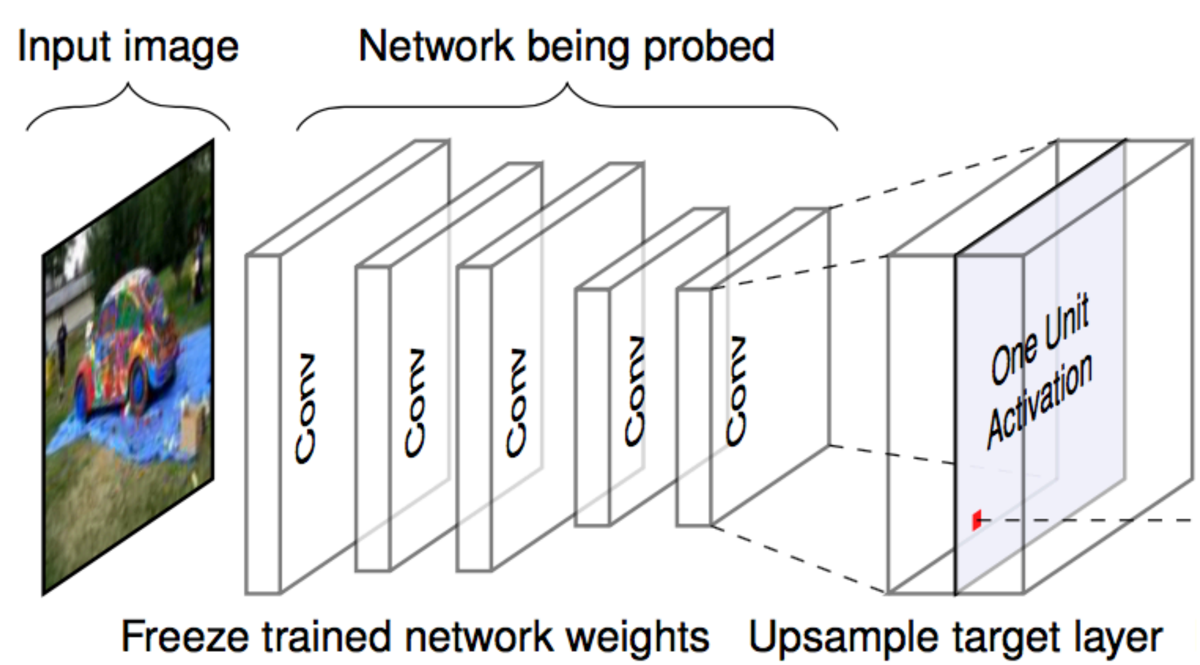
1. Identify a broad set of human-labeled visual concepts
2. Gather hidden variables' response to known concepts
3. Quantify alignment of hidden variable - concept pairs

1. Identify a broad set of human-labeled visual concepts

- Broden dataset: Broadly and densely labelled dataset
- 63,305 images with 1197 visual concepts
- Concept labels are assigned pixel-wise

2. Gather hidden variables' response to known concepts

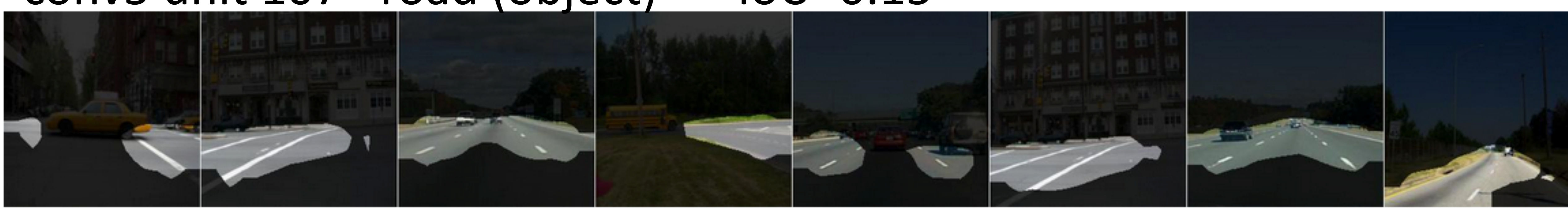
- For convolutional neurons, compute their activation map
- In other words, what is the output of a particular convolutional filter for a given image
- Threshold this activation map to convert it to a binary activation map



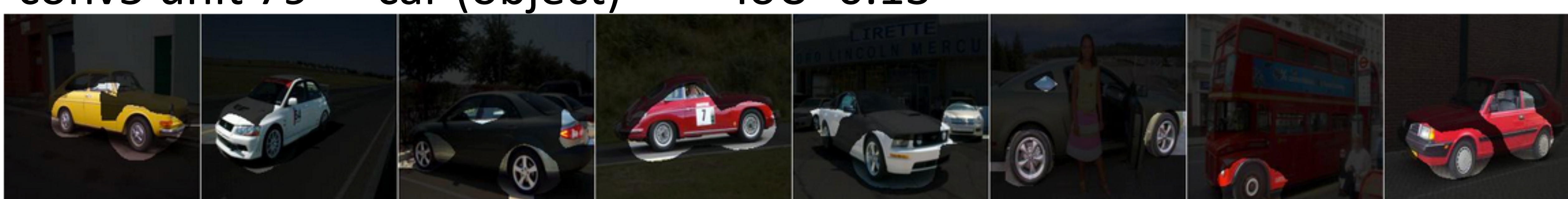
3. Quantify alignment of hidden variable - concept pairs

- Measure the IoU between the binary activation map and the labelled concept images
- If activation map overlaps highly with a concept, the neuron is a detector for that concept

conv5 unit 107 road (object) IoU=0.15

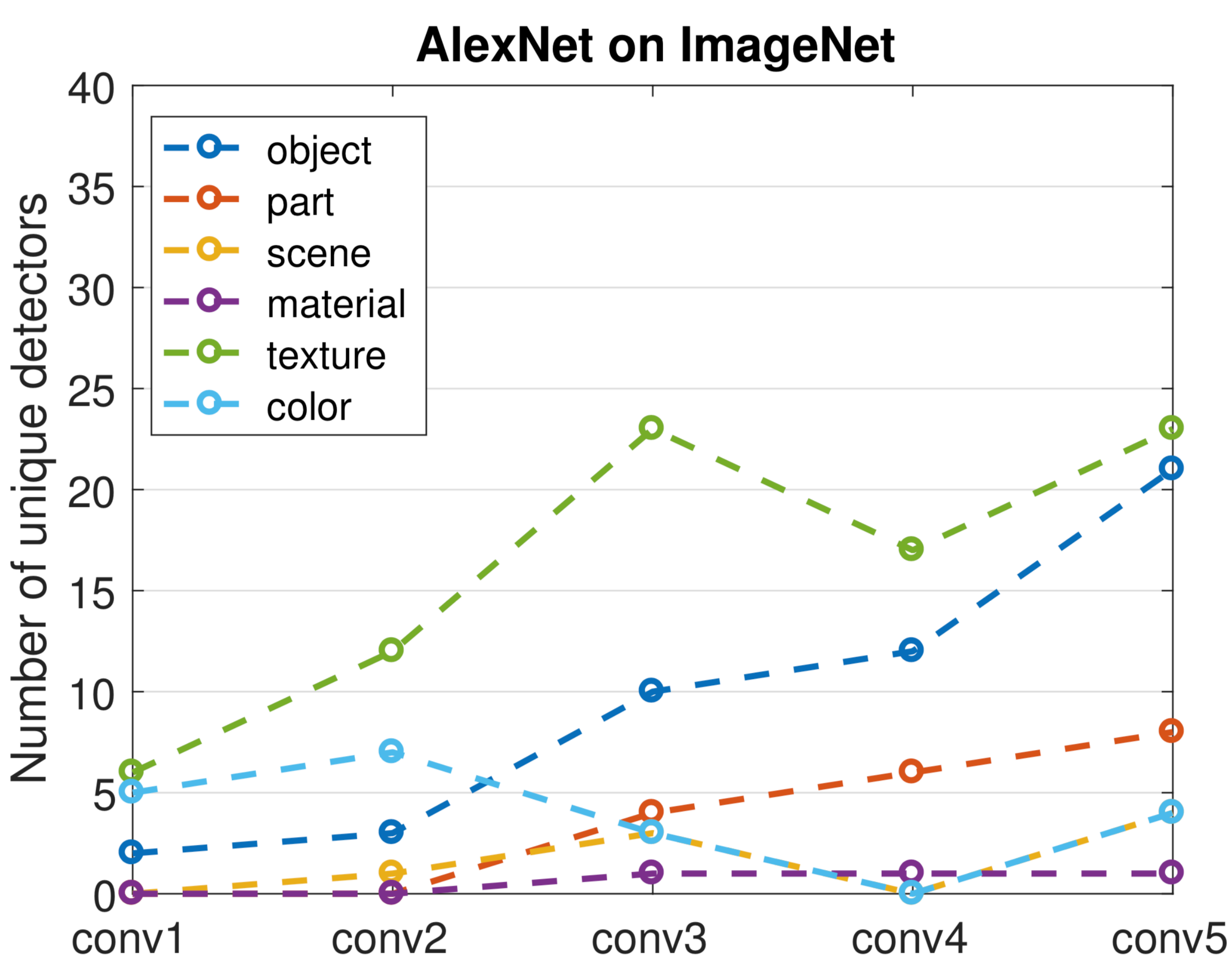


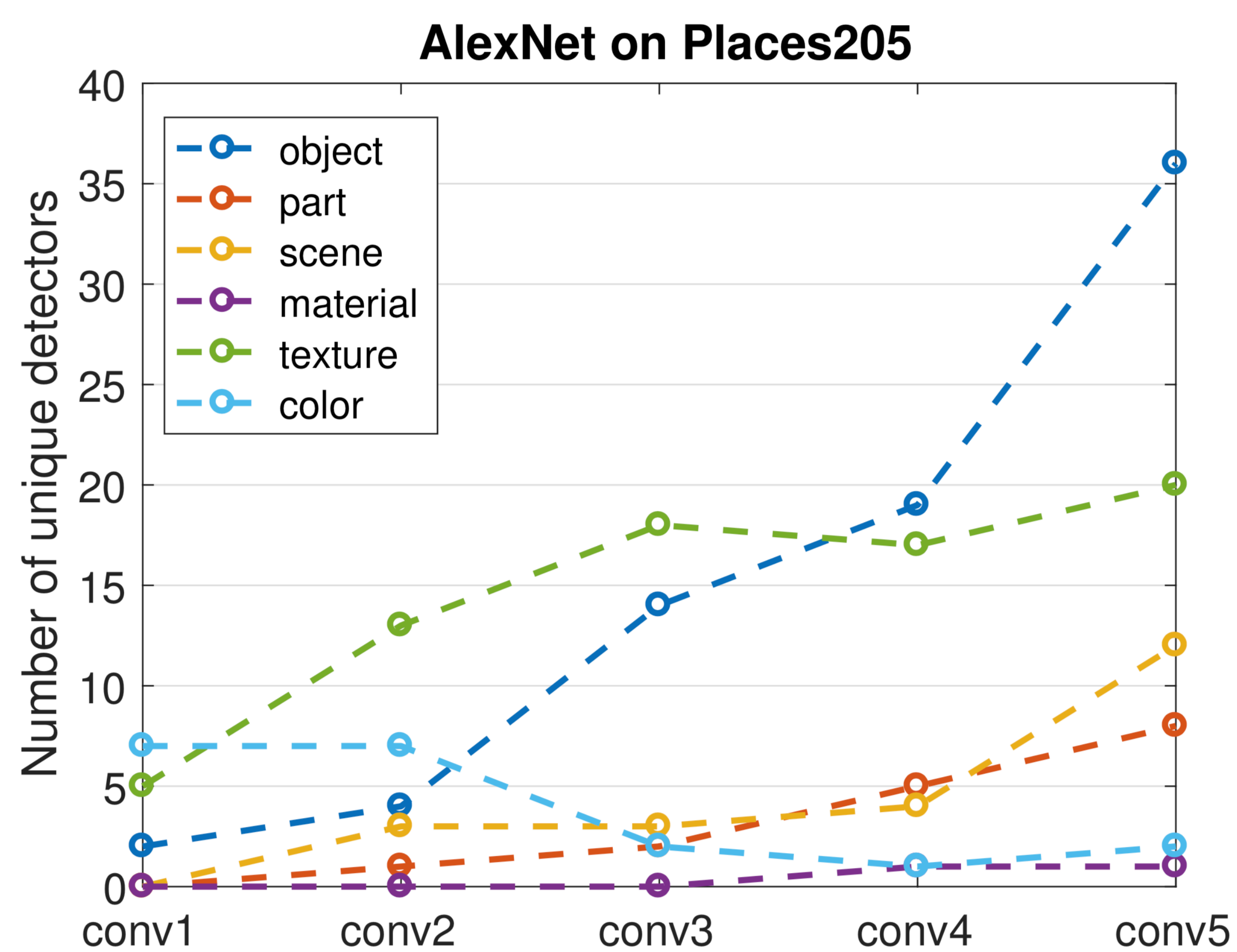
conv5 unit 79 car (object) IoU=0.13



Quantifying interpretability of deep visual representations

- Interpretability is quantified by how well the network aligns with a set of human interpretable concepts





Effect of regularization on interpretability

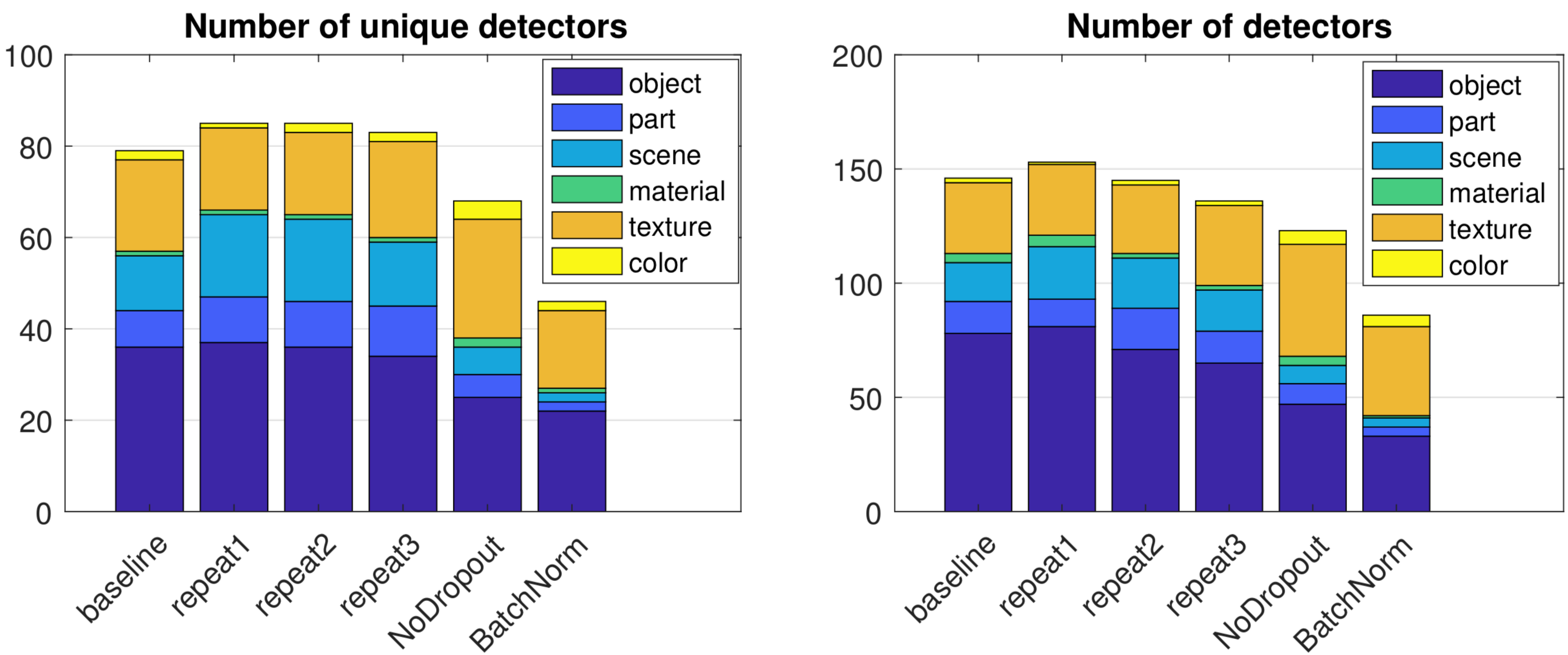
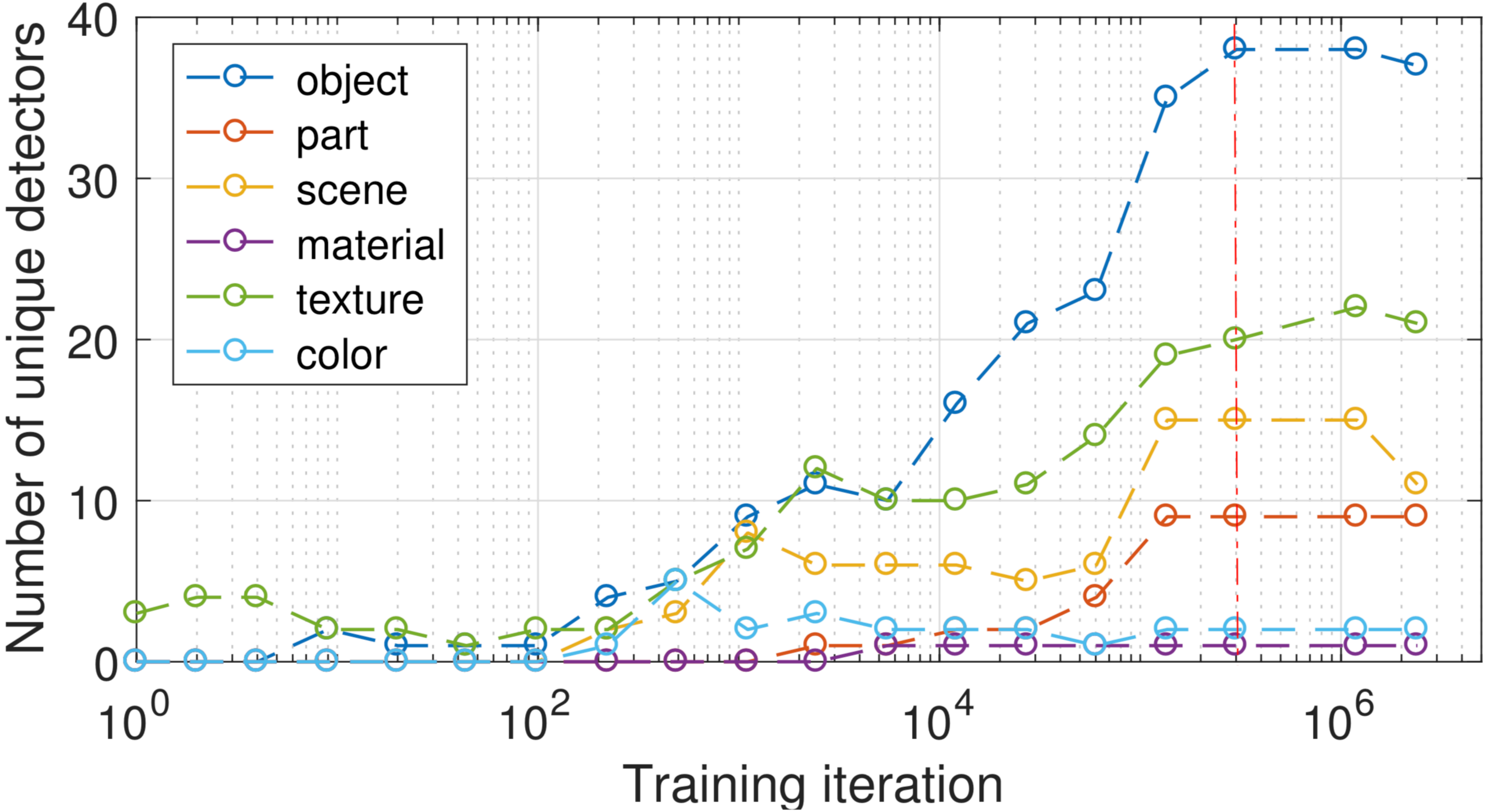


Figure 11. Effect of regularizations on the interpretability of CNNs.

Number of detectors vs epoch



Other experiments

- Random initialization does not seem to affect interpretability
- Widening of AlexNet showed an increase in the number of concept detectors