GAN Dissection

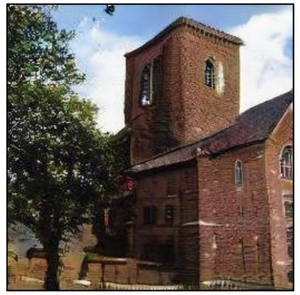
David Bau, Jun-Yan Zhu, Hendrik Strobelt, Bolei Zhou, Joshua B. Tenenbaum, William T. Freeman, Antonio Torralba

Julia Semavina, 171

Motivation

Motivation

- What knowledge does a GAN need to learn to produce a church image?

- What causes the mistakes when GAN produces unrealistic images?


- Why does one GAN variant work better than another?
- **Does the GAN contain internal variables that correspond to the objects that humans perceive?**
- **If so, do those variables cause the generation of those objects, or do they merely correlate?**

Main Goal

Explain how an image can be generated by a network.

Method

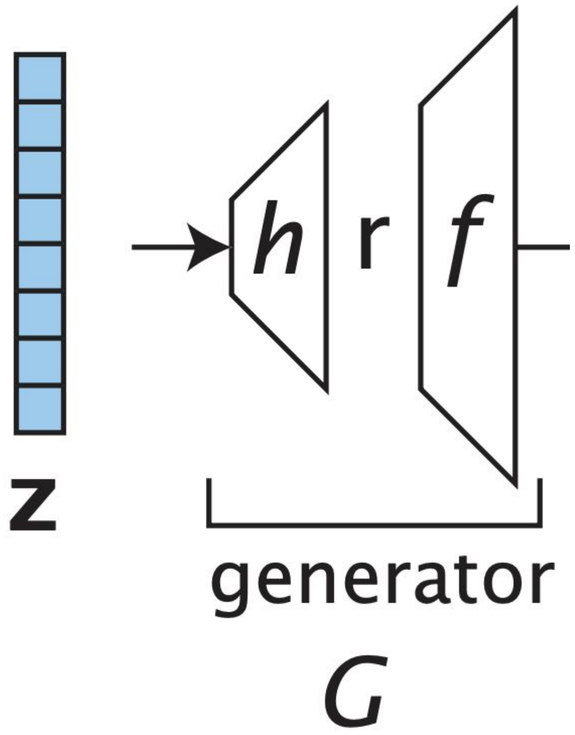
Notation - Representation

- $G : z \to x$ - generator.
- $z \in \mathbb{R}^{|z|}$ - latent vector
- $x \in \mathbb{R}^{H \times W \times 3}$ - generated image

Notation - Representation

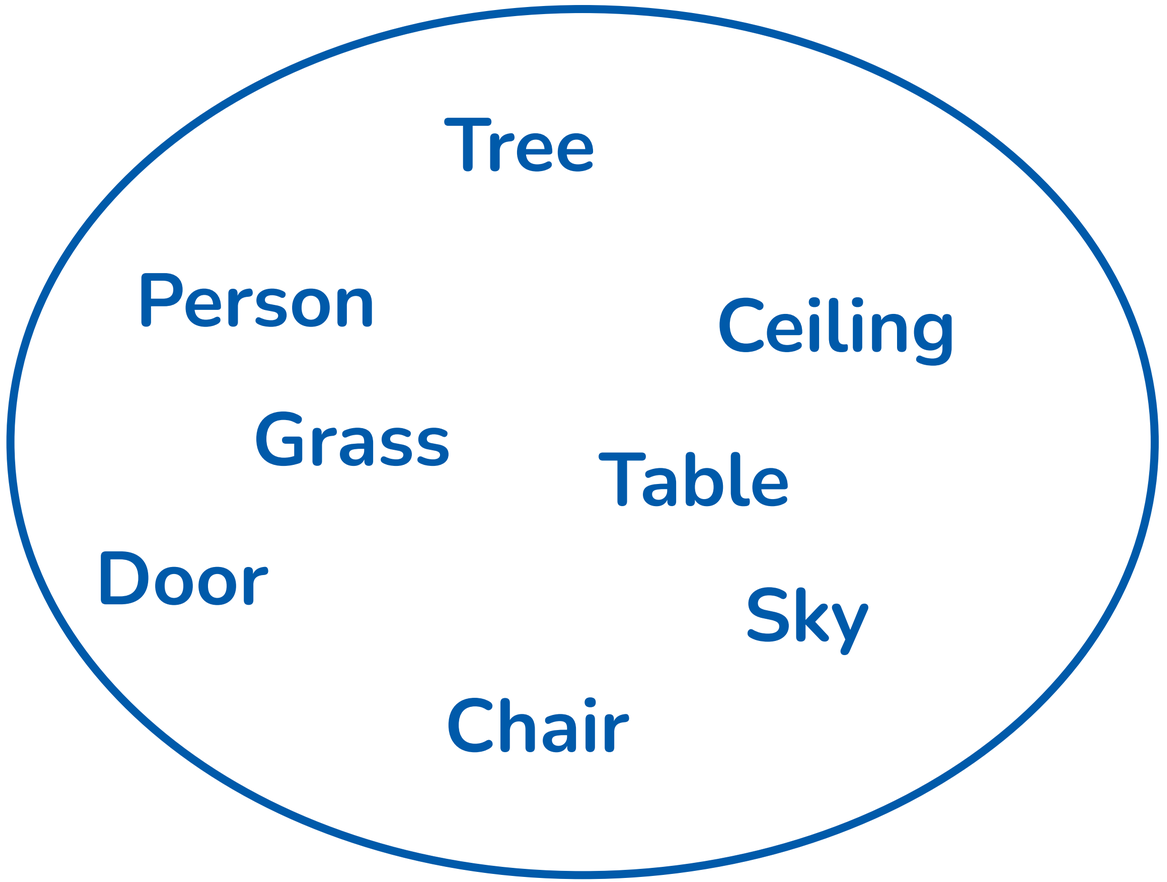
- $G : z \to x$ - generator.
- $z \in \mathbb{R}^{|z|}$ - latent vector
- $x \in \mathbb{R}^{H \times W \times 3}$ - generated image

- $G(z) = f(h(z))$ - decomposition of G
- $r = h(z)$ - representation



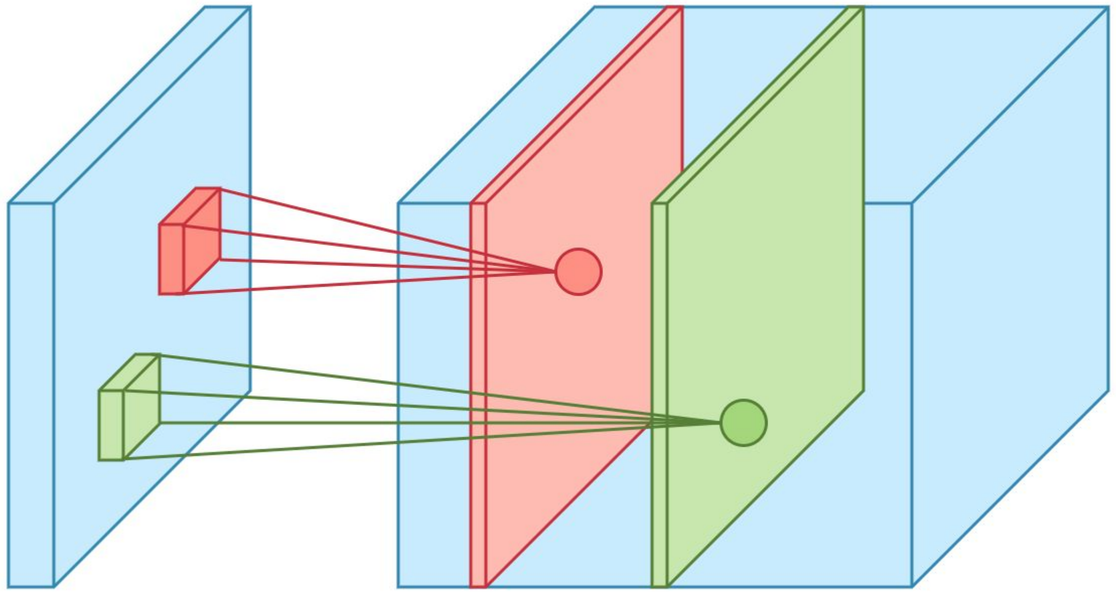
Notation - Concept

- $\mathbf{C}$ - a universe of concepts
- $c \in \mathbf{C}$ - a concept



Notation - Units and Pixels

- U - a set of units (channels of the featuremap)
- P - a set of featuremap pixels
- $\mathbb{U}, \mathbb{P}$ - entire set of units and featuremap pixels of r



Main Question

Given concept $c \in \mathbb{C}$ and representation $r = h(z)$.

Can we find a subset of units responsible for the generation of c at locations P?

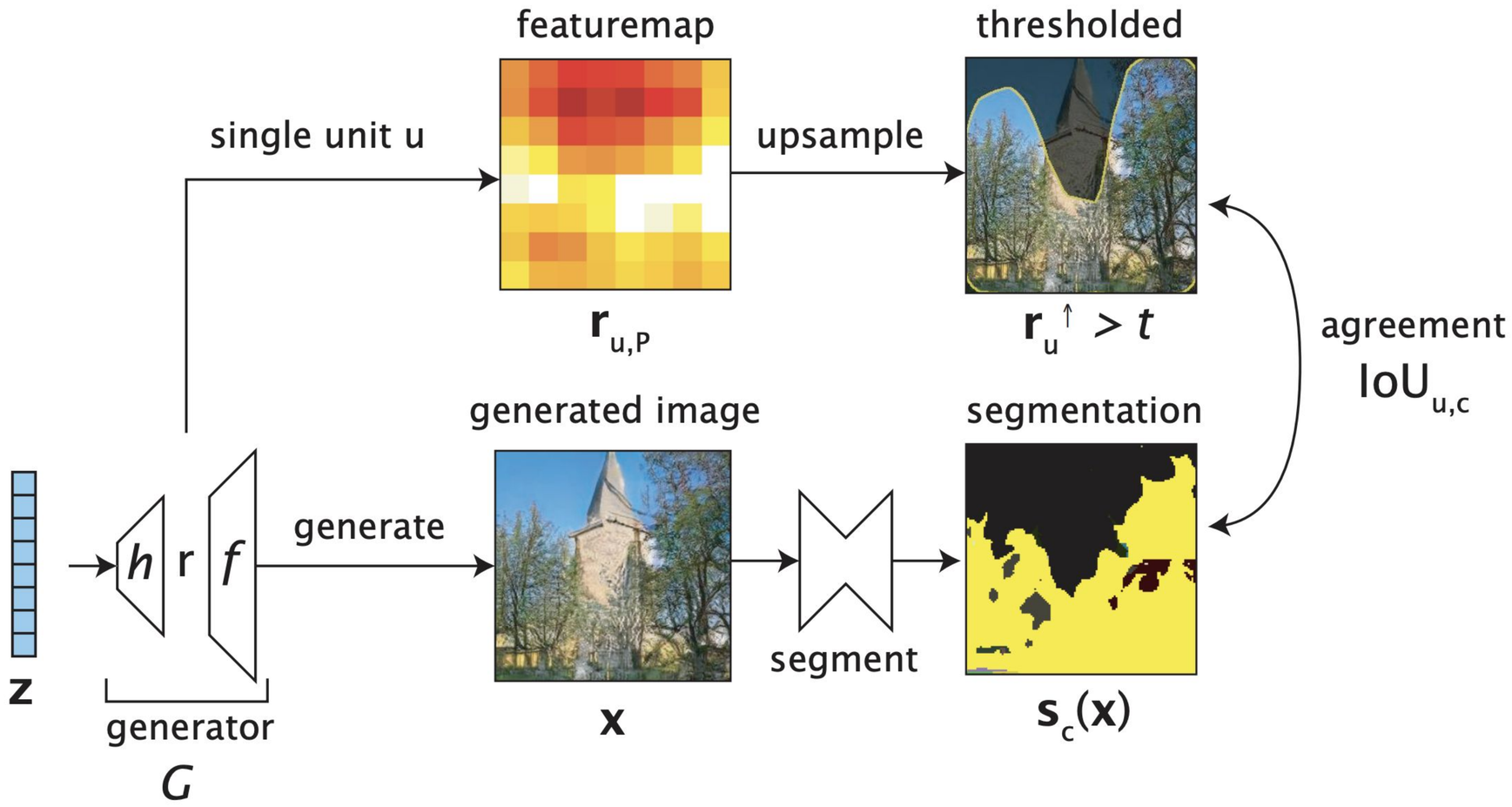
$$r_{\mathbb{U},P} = (r_{U,P}, r_{\overline{U},P})$$

Method

2 Stages:

- **Dissection** - find relevant concepts - IoU metric
- **Intervention** - find subsets of units that cause the objects - ACE metric

Stage 1 - Dissection



Stage 1 - Dissection

Visualizing unit heatmaps

Stage 1 - Dissection

$r_{u,\mathbb{P}}$ - one-channel $h \times w$ featuremap of unit u.
Does $r_{u,\mathbb{P}}$ encode a concept c?

Let $s_c(x)$ - semantic segmentation for the concept c. Spacial agreement between $r_{u,\mathbb{P}}$ and c is quantified as intersection-over-union measure:

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

Stage 1 - Dissection

Question: How to choose thresholding level $t_{u,c}$ to make it as informative as possible?

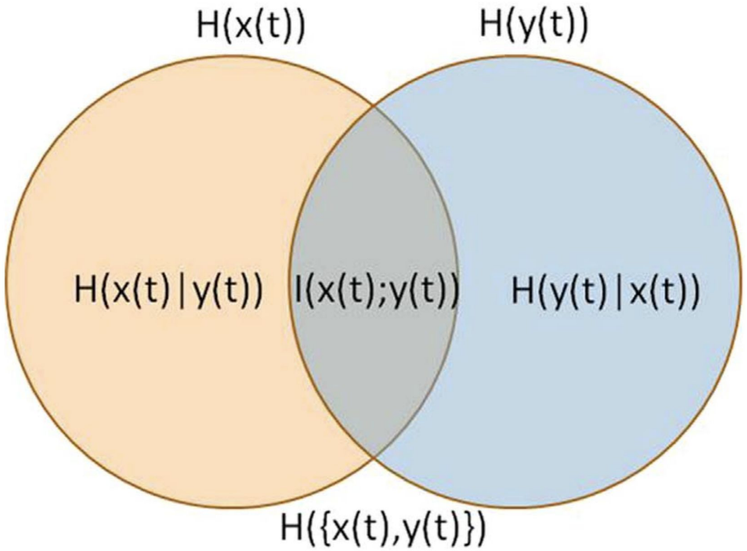
Proposed Approach: Maximize the information quality ratio I/H (the portion of the joint entropy H which is mutual information I).

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

Stage 1 - Dissection

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

- $t_{u,c} \in [0, 1]$
- $t_{u,c} = 1$ - perfect segmentation
- We want segmentation to be as good as possible



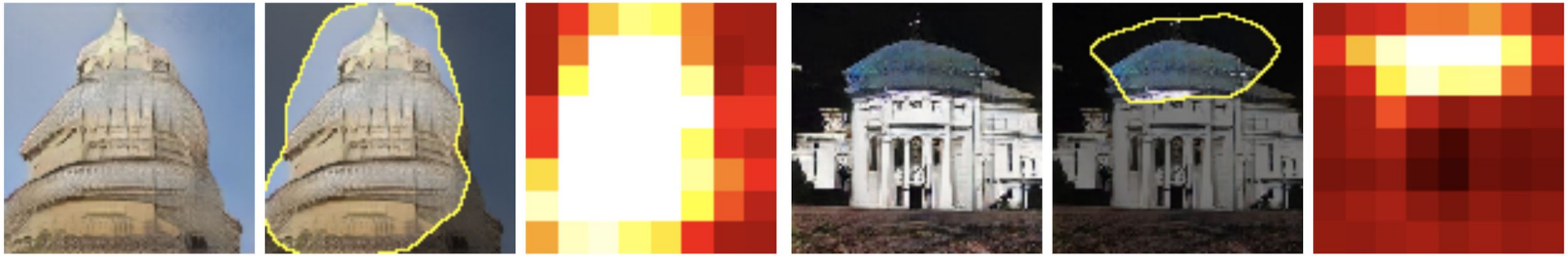
$$H(\{x(t), y(t)\}) = -\sum_{x_i \in x(t)} \sum_{y_j \in y(t)} p(x_i, y_j) \log_2 (p(x_i, y_j))$$

$$I(x(t); y(t)) = \sum_{x_i \in x(t)} \sum_{y_j \in y(t)} p(x_i, y_j) \log_2 \left(\frac{p(x_i, y_j)}{p(x_i) p(y_j)} \right)$$

Stage 1 - Dissection

unit 380: grass (iou 0.27)

unit 233: dome (iou 0.22)

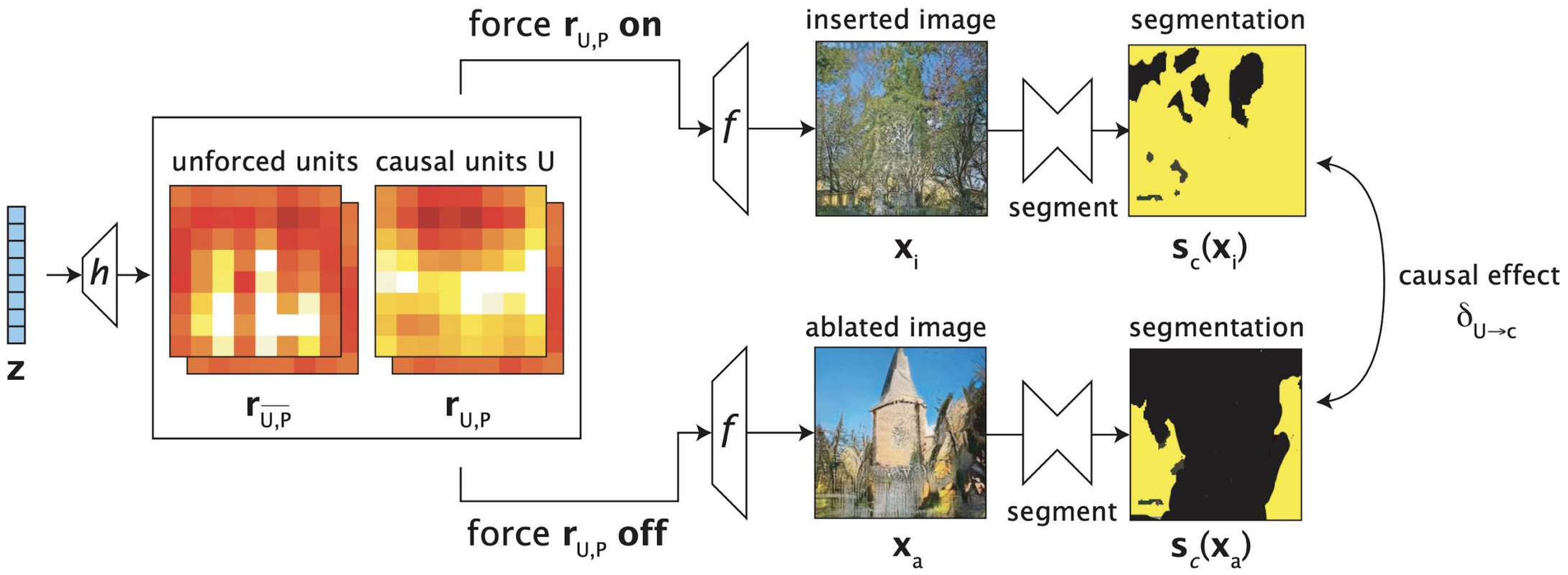


Stage 1 - Dissection

unit 119, trees iou 0.12

unit 408, trees iou 0.08

Stage 2 - Intervention



Stage 2 - Intervention

Which combination of units cause an object?

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

Stage 2 - Intervention

An object is caused by U if the object appears in x_i and disappears from x_a. Average causal effect (ACE) of units U on the generation of concept c:

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

Stage 2 - Intervention

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

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

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

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

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

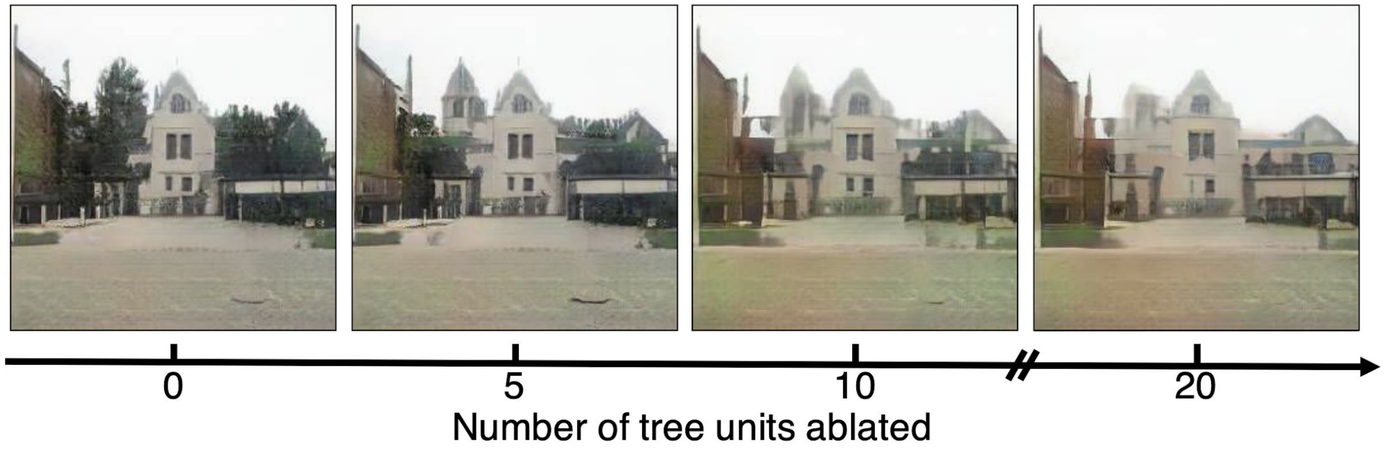
$$\delta_{\boldsymbol{\alpha}\rightarrow c} = \mathbb{E}_{\mathbf{z},\mathrm{P}}\left[\mathbf{s}_c(\mathbf{x}'_i)\right] - \mathbb{E}_{\mathbf{z},\mathrm{P}}\left[\mathbf{s}_c(\mathbf{x}'_a)\right]$$

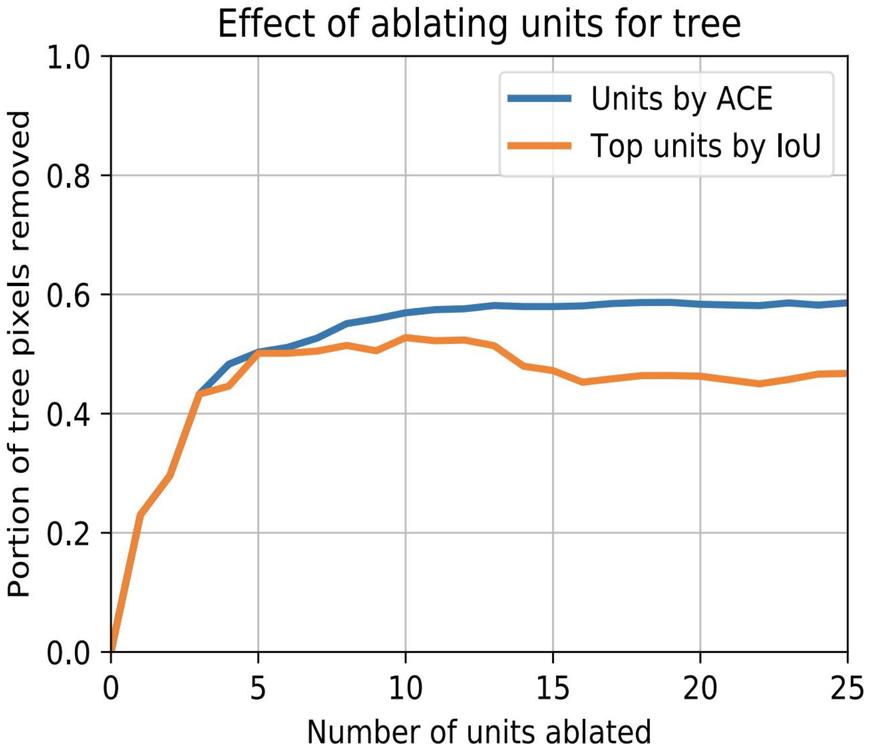
$$\boldsymbol{\alpha}^* = \arg\min_{\boldsymbol{\alpha}}(-\delta_{\boldsymbol{\alpha}\rightarrow c} + \lambda||\boldsymbol{\alpha}||_2)$$

- r consists of d units.
- $\alpha \in [0, 1]^d$ - intervention
- α_u - degree of intervention for unit u

Stage 2 - Intervention

ACE vs IoU





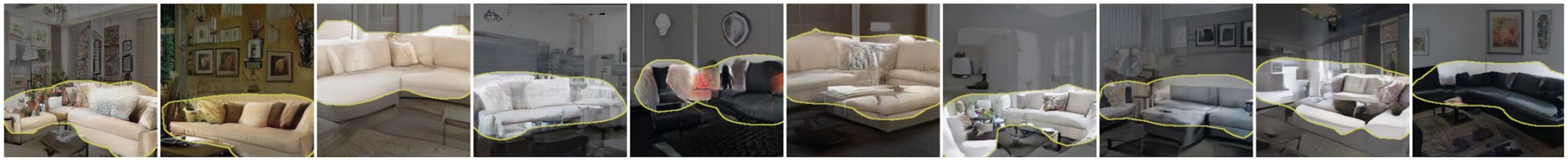
Results

Set Up

- **3 different GANs**:
    - baseline progressive GAN
    - progressive GAN + minibatch stddev statistics
    - progressive GAN + pixelwise normalization
- **336** object classes, **29** parts of large objects, and **25** materials

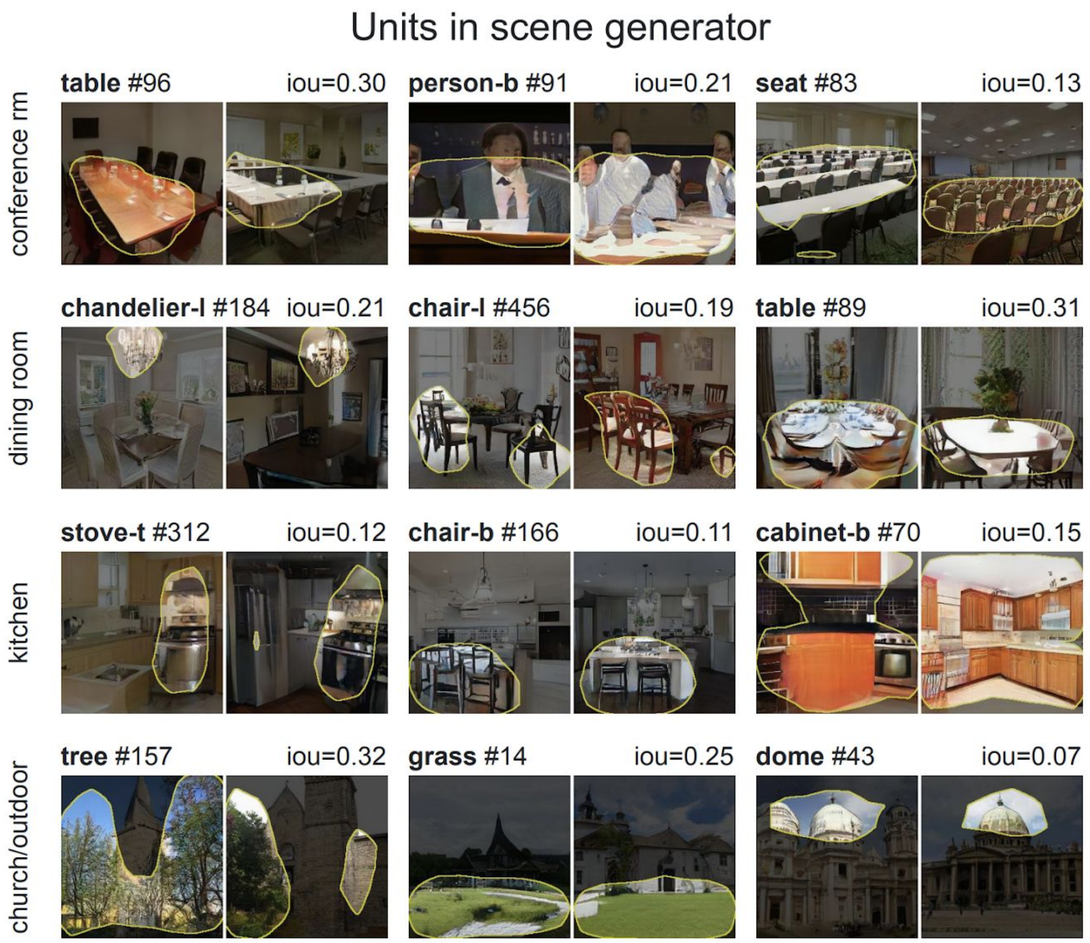
Finding units correlated with instances of a class

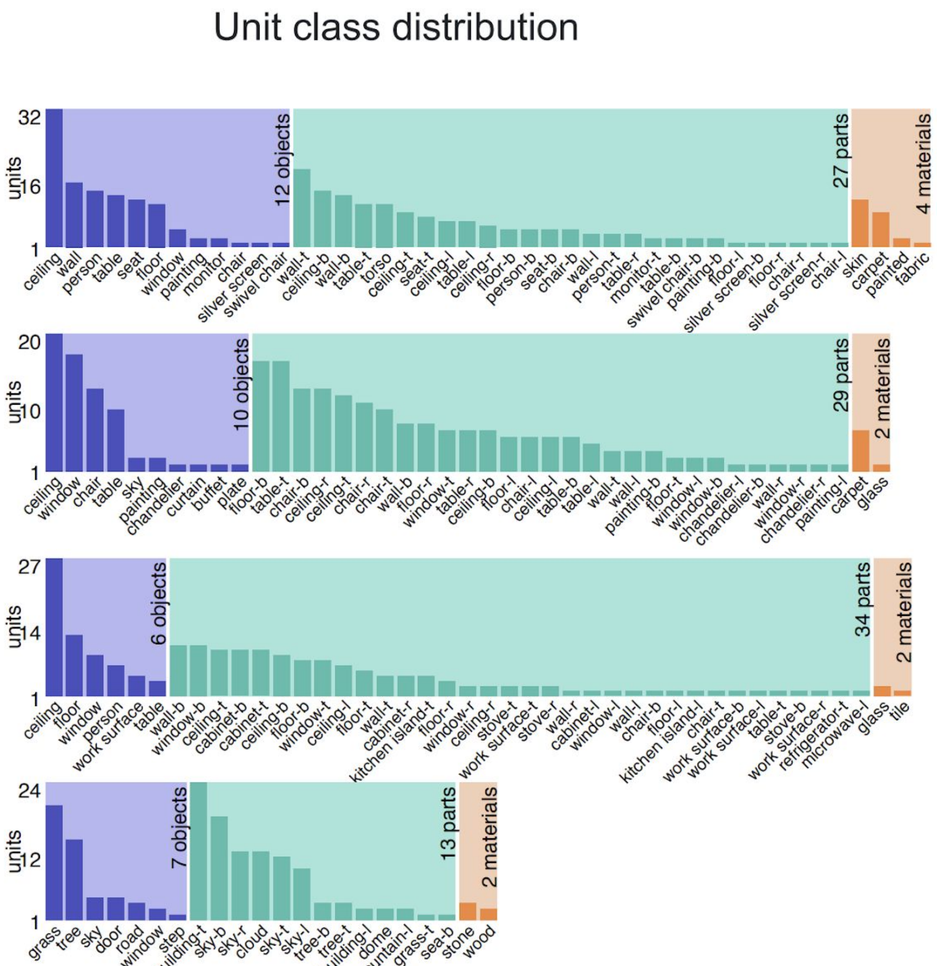
Thresholding unit #65 layer 3 of a dining room generator matches ‘table’ segmentations with IoU=0.34.



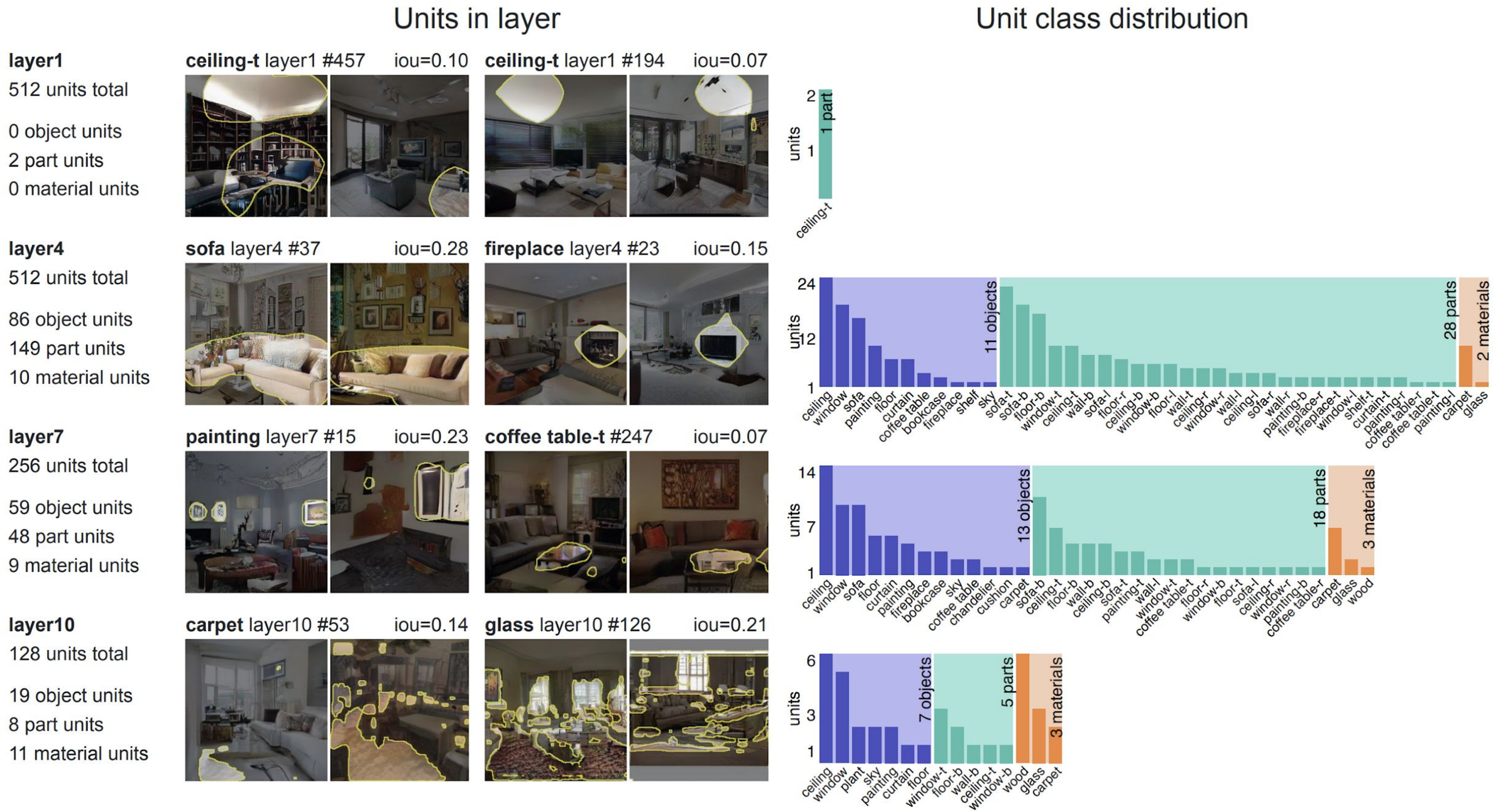
Thresholding unit #37 layer 4 of a living room generator matches ‘sofa’ segmentations with IoU=0.29.

Objects found on different scene types

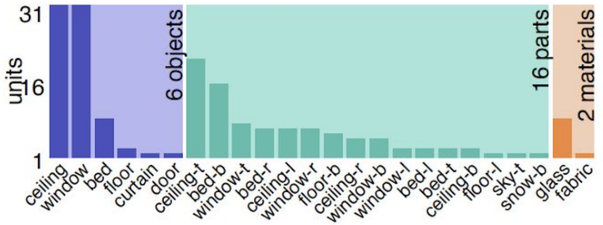
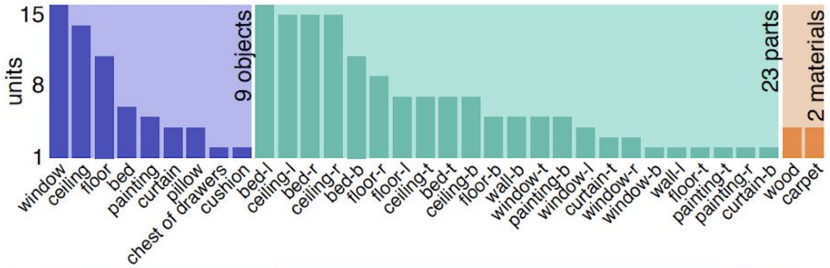
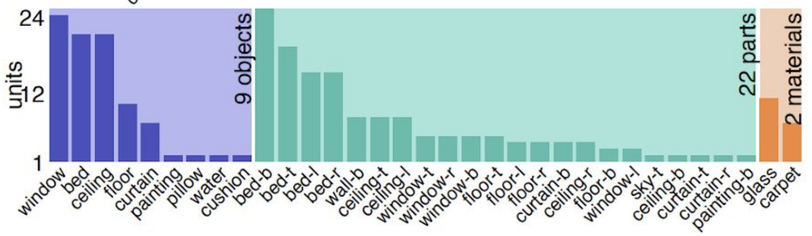




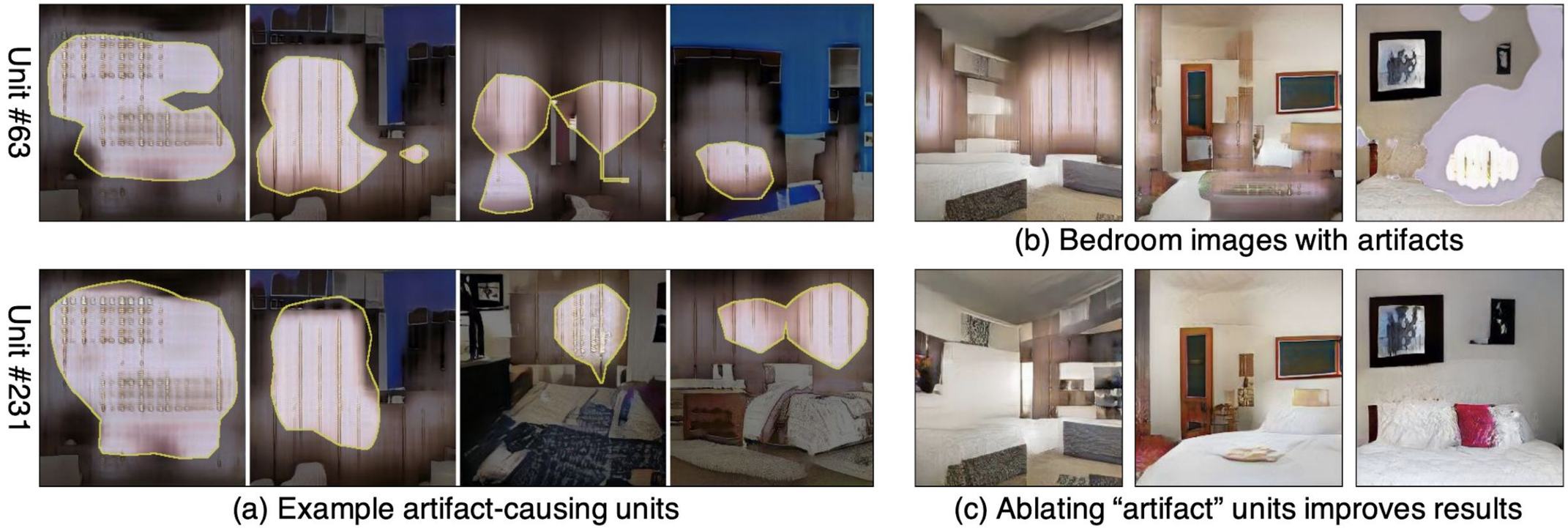
Different network layers



Different GAN 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	

Diagnosing and Improving GANs



(a) Example artifact-causing units

(b) Bedroom images with artifacts

(c) Ablating "artifact" units improves results

Table 1: We compare generated images before and after ablating 20 "artifacts" units. We also report a simple baseline that ablates 20 randomly chosen units.

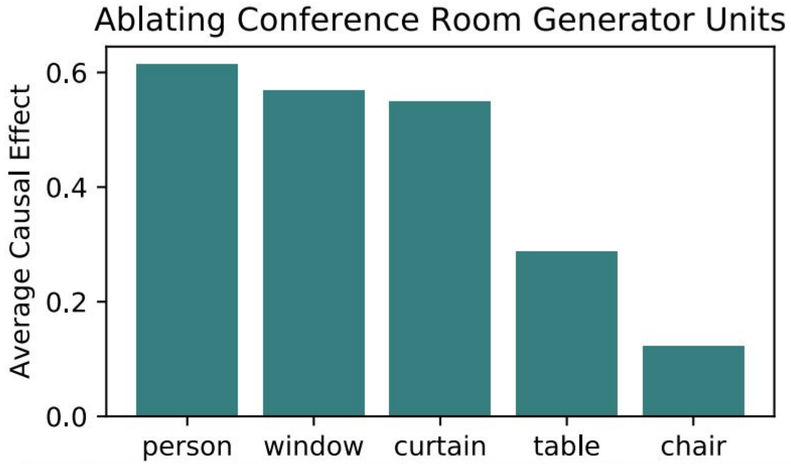
Fréchet Inception Distance (FID)	
original images	43.16
"artifacts" units ablated (ours)	**27.14**
random units ablated	43.17

Human preference score	original images
"artifacts" units ablated (ours)	**72.4%**
random units ablated	49.9%

Use case: Removing objects from an image

ablate person units

ablate curtain units



ablate window units

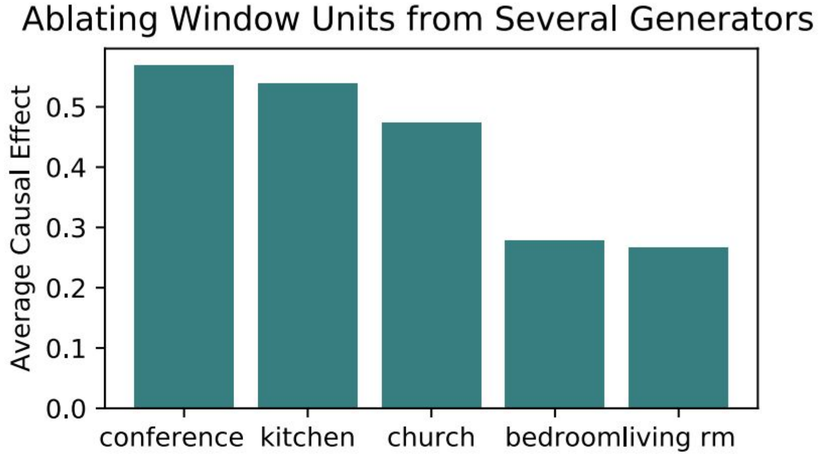
ablate table units

ablate chair units

Effect of ablating objects

conference room

church



kitchen

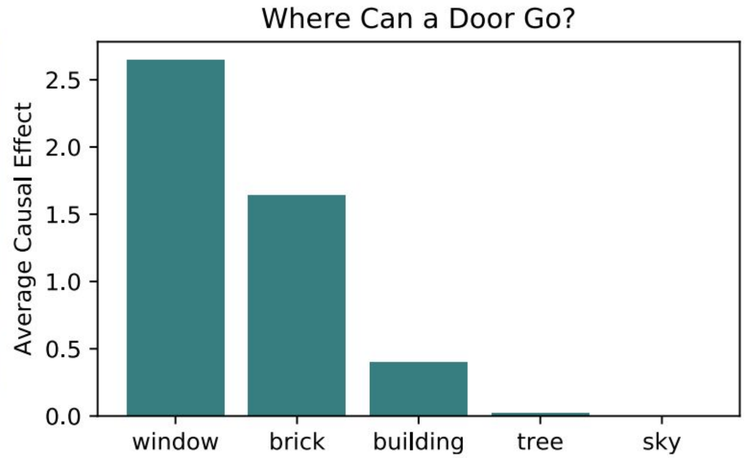
living room

bedroom

Effect of inserting objects

(a)

(b)



(c)

(d)

(e)

References

- https://arxiv.org/pdf/1811.10597.pdf
- https://lbeifits.files.wordpress.com/2018/03/2017-dedy-elsevier-1-s2-0-s0169743916304907-main.pdf
- http://gandissect.res.ibm.com/ganpaint.html?project=churchoutdoor&layer=layer4
- https://gandissect.csail.mit.edu/
- https://colab.research.google.com/github/SIDN-IAP/global-model-repr/blob/master/notebooks/gandissect_solutions.ipynb

Вопросы

1. Чем предлагаемый в статье метод интерпретации нейросетей радикально отличается от ранее существовавших методов?
2. Как авторы количественно измеряют сходство между отдельным юнитом (картой признаков из 1 канала) и некоторым классом объектов (например, деревьев)?
3. Как авторы статьи предлагают решать задачу поиска подмножества юнитов, имеющего наибольшее влияние (ACE) на появление определенного объекта на сгенерированной картинке?

Thank you :)