GAN Dissection: Visualizing and Understanding Generative Adversarial Networks

David Bau[1,2], Jun-Yan Zhu[1], Hendrik Strobelt[2,5], Bolei Zhou[3], Joshua B. Tenenbaum[1], William T. Freeman[1,4], Antonio Torralba[1,2]
[1]Massachusetts Institute of Technology, [2]MIT-IBM Watson AI Lab, [3]The Chinese University of Hong Kong, [4]Google Research, [5]IBM Research

ICLR 2019

Presenter: Jack Lanchantin

Introduction

- G: $z \rightarrow x$, where $z \in R^{|z|}$ and $x \in \mathbb{R}^{H \times W \times 3}$
- Tensor r output from a particular layer of G: $r = h(z)$ and $x = f(r) = f(h(z)) = G(z)$
- r certainly contains the information to deduce the presence of any visible class c in the image
- Question is *how* the information about c is encoded in r

Introduction

- In particular, we seek to understand whether r explicitly represents the concept c in some way where it is possible to factor r at locations P into components

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

  where the generation of the object c at locations P depends mainly on the units $r_{U,P}$ and is insensitive to the other units $r_{\bar{U},P}$

- Refer to each channel of the featuremap as a unit; U denotes the set of unit indices of interest and denotes its complement
- we will write $\mathbb{U}$ and $\mathbb{P}$ to refer to the entire set of units and feature map pixels in r

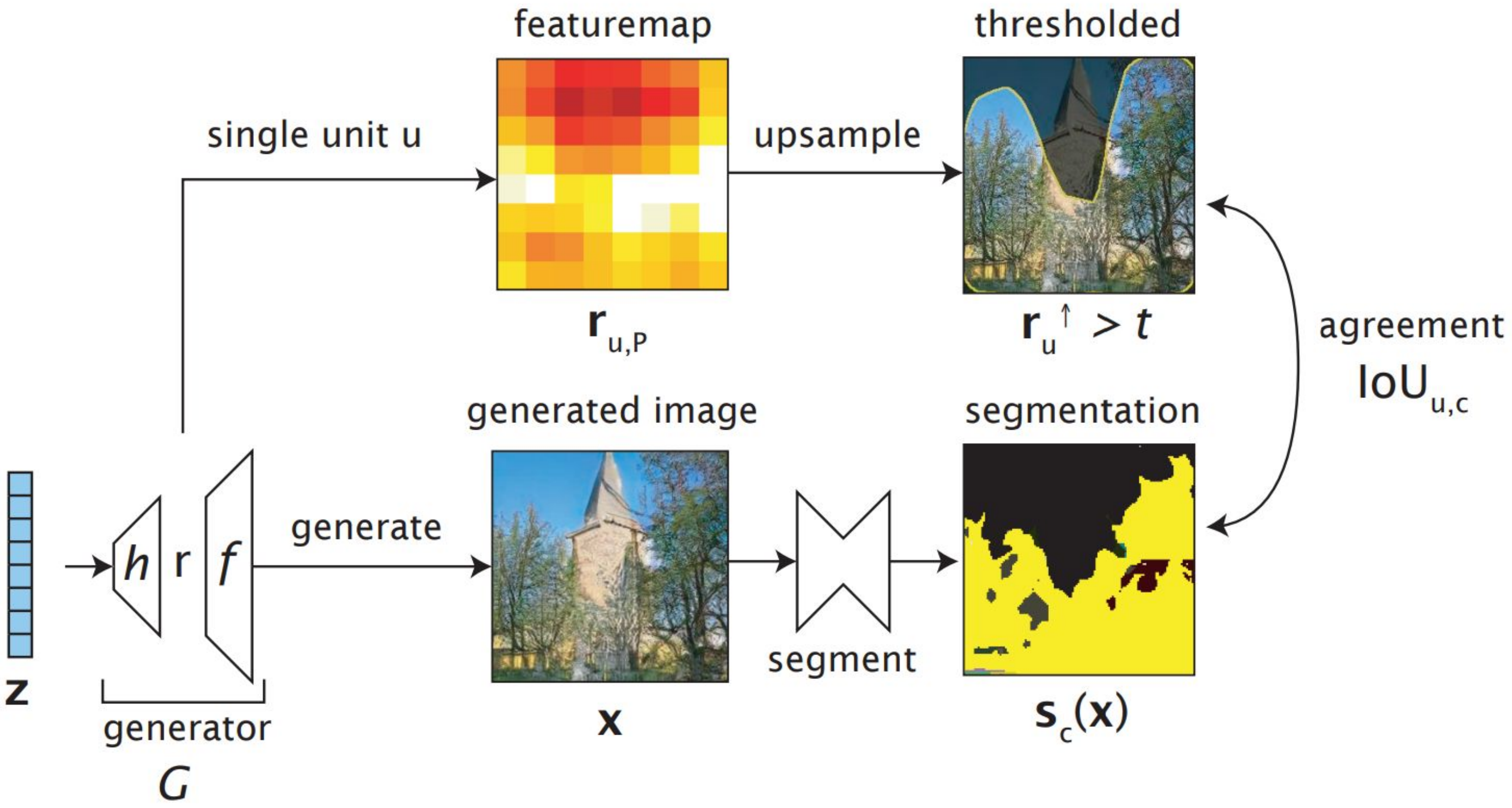
Characterizing Units by Dissection

- Quantify the spatial agreement between the unit U’s thresholded featuremap and a concept c’ segmentation with the following intersection-over-union (IoU) measure:

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

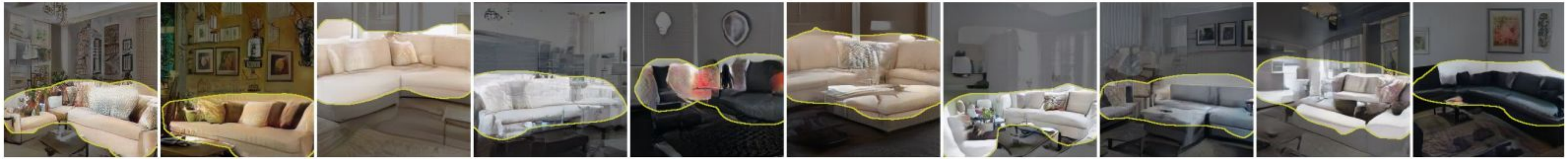
  where $\wedge$ and $\vee$ denote intersection and union, $t_{u,c}$ is a fixed threshold, and $s_c(x)$ is a binary segmentation mask for concept c

Characterizing Units by Dissection



Characterizing Units by Dissection

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.

Figure 3: Visualizing the activations of individual units in two GANs. 10 top activating images are shown, and IoU is measured over a sample of 1000 images. In each image, the unit feature is upsampled and thresholded as described in Eqn. 2.

Measuring Causal Relationships Using Intervention

- Which of those units are actually responsible for triggering the rendering of that object?
  - Correlation != causation
  - Furthermore, any output will jointly depend on several parts of the representation (need to identify combinations)

Measuring Causal Relationships Using Intervention

- Recall that $\mathbf{r}_{U,P}$ denotes the feature map r at unit U and location P
- **Ablate** such unit by forcing $\mathbf{r}_{U,P} = 0$.
- **Insert** such unit by forcing $\mathbf{r}_{U,P} = \mathbf{c}$, where **c** is a big constant.
- Decompose r into two parts ($\mathbf{r}_{U,P}, \mathbf{r}_{\overline{U,P}}$), where $\mathbf{r}_{\overline{U,P}}$ are unforced components of **r**

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

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

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

Measuring Causal Relationships Using Intervention

- Recall that $\mathbf{r}_{U,P}$ denotes the feature map r at unit U and location P
- **Ablate** such unit by forcing $\mathbf{r}_{U,P} = 0$.
- **Insert** such unit by forcing $\mathbf{r}_{U,P} = \mathbf{c}$, where **c** is a big constant.
- Decompose r into two parts ($\mathbf{r}_{U,P}$, $\mathbf{r}_{\overline{U,P}}$), where $\mathbf{r}_{\overline{U,P}}$ are unforced components of **r**

$$
\begin{aligned}
&\text{Original image :} && \mathbf{x} = G(\mathbf{z}) \equiv f(\mathbf{r}) \equiv f(\mathbf{r}_{\mathrm{U,P}}, \mathbf{r}_{\overline{\mathrm{U,P}}}) \\
&\text{Image with U ablated at pixels P :} && \mathbf{x}_a = f(\mathbf{0}, \mathbf{r}_{\overline{\mathrm{U,P}}}) \\
&\text{Image with U inserted at pixels P :} && \mathbf{x}_i = f(\mathbf{c}, \mathbf{r}_{\overline{\mathrm{U,P}}})
\end{aligned}
$$

➢ An object is caused by U if the object appears in $\mathbf{x}_i$ and disappears from $\mathbf{x}_a$

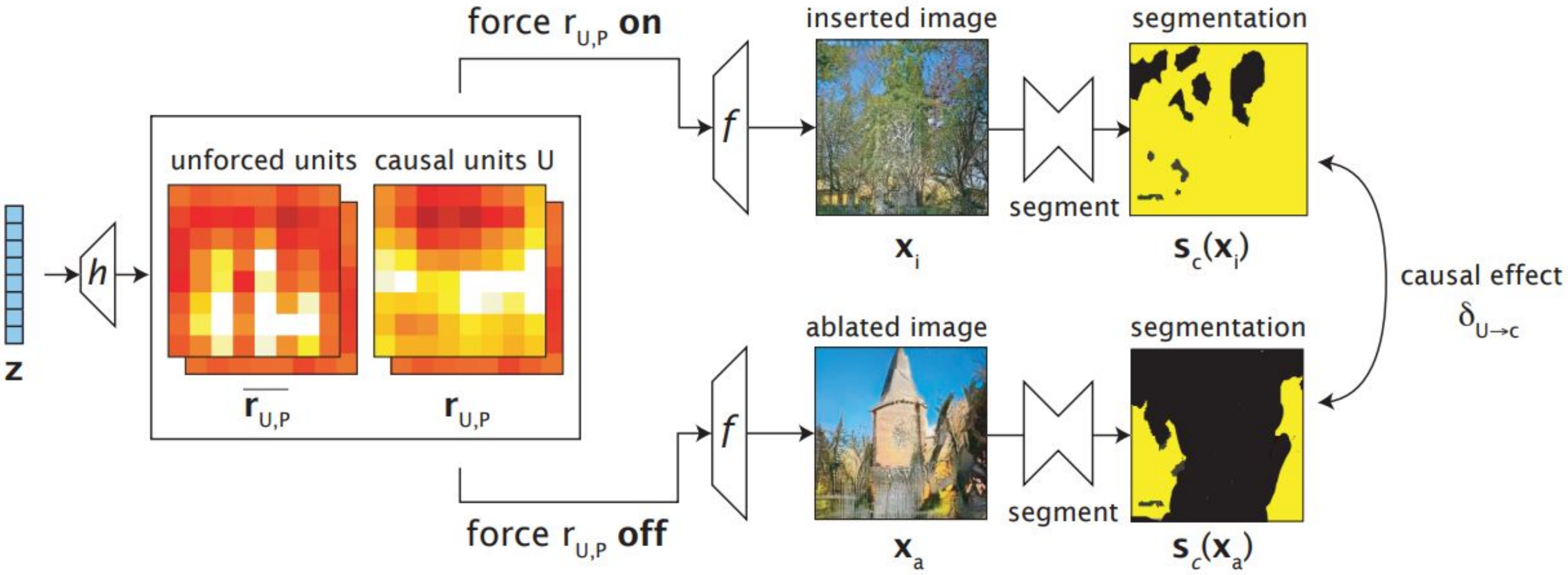
Measuring Causal Relationships Using Intervention

- This causality can be quantified by comparing the presence of an object in x_i and x_a and averaging effects over all locations and images
- Define the average causal effect (ACE) of unit U on the generation of on class c as:

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

where $\mathbf{s}_c(\mathbf{x})$ denotes a segmentation indicating the presence of class c in image $\mathbf{x}$ at P

Measuring Causal Relationships Using Intervention



Sets of Units with High Causal Effect

- Objects tend to depend on more than one unit.
- Thus we need to identify a *set of units* U that maximize the average causal effect $\delta_{U \rightarrow c}$ for a class c

Finding sets of units with high average causal effect

- Given a representation **r** with *d* units, searching for a fixed-size set U with high $\delta_{U \to c}$ requires $\binom{d}{|U|}$ operations
- Instead, we optimize a continuous intervention $\alpha \in [0, 1]^d$, where each dimension α_u indicates the degree of intervention for unit u.

Finding sets of units with high average causal effect

- We maximize the following average causal effect formulation $\delta_{\alpha \to c}$:

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

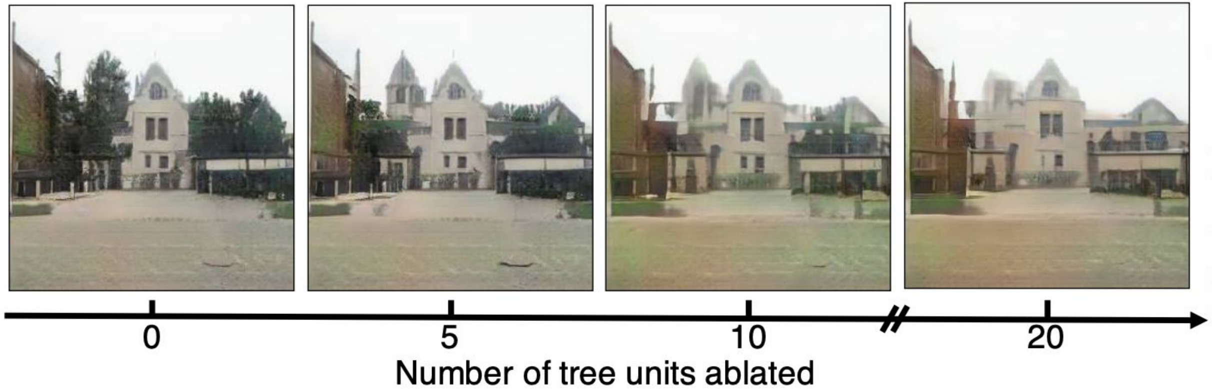
$$\text{Image with partial insertion at pixels P :} \quad \mathbf{x}'_i = f(\boldsymbol{\alpha} \odot \mathbf{c} + (\mathbf{1}-\boldsymbol{\alpha}) \odot \mathbf{r}_{\mathbb{U},\mathrm{P}},\ \mathbf{r}_{\mathbb{U},\overline{\mathrm{P}}})$$

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

  where $r_{U,P}$ denotes the all-channel featuremap at locations P, $r_{U,\overline{P}}$ denotes the all-channel featuremap at other locations $\overline{P}$, and applies a per-channel scaling vector α to the featuremap $r_{U,P}$

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

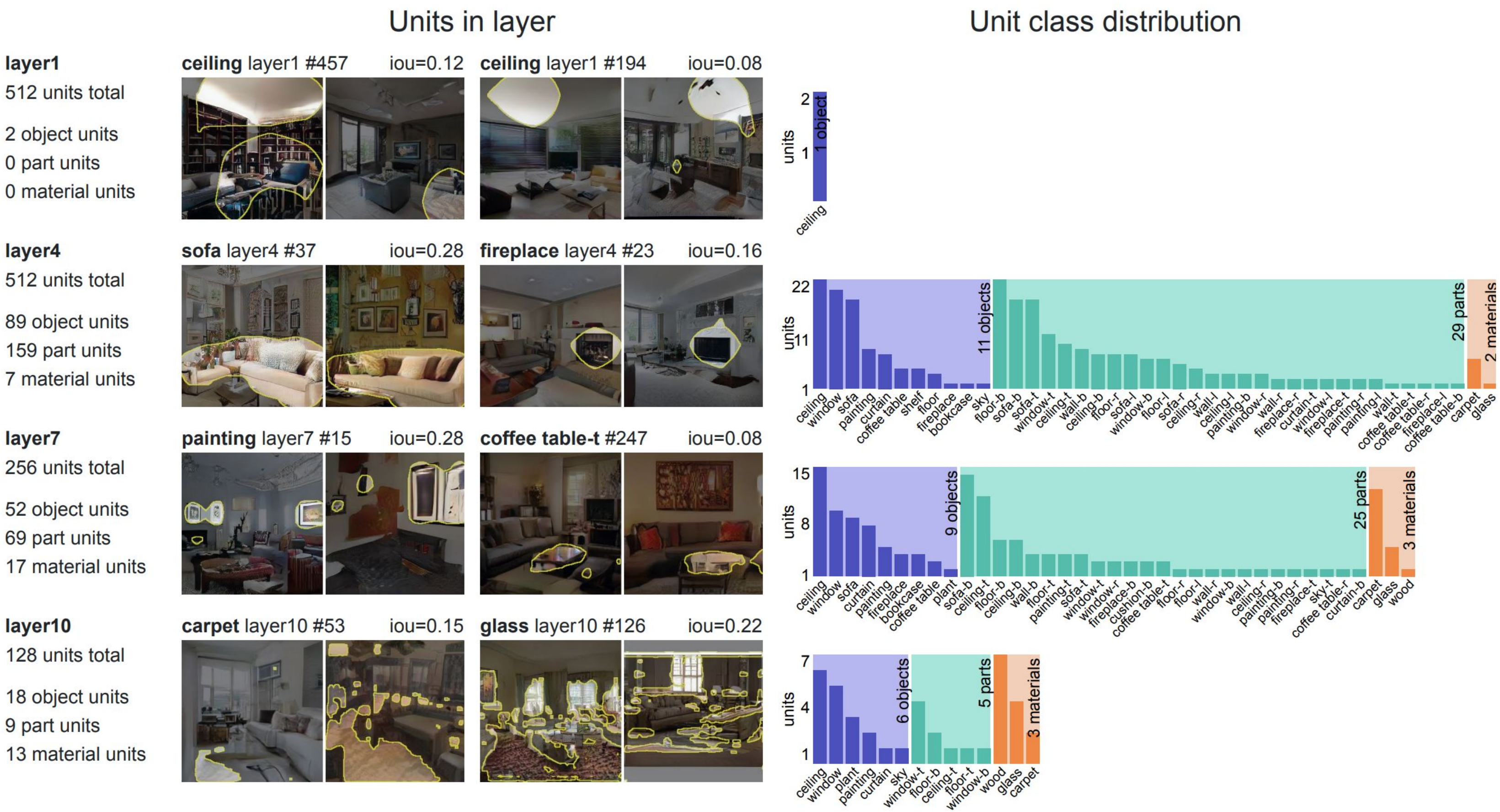
Finding sets of units with high average causal effect



Units that match objects (from layer 4 of trained CNN)

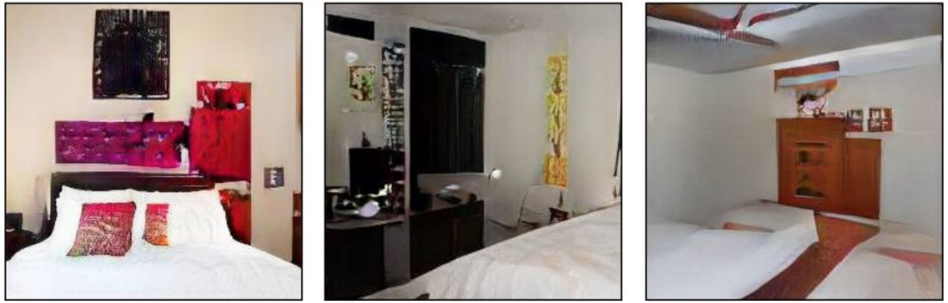
A unit is counted as a class predictor if it matches a supervised segmentation class with pixel accuracy > 0.75 and IoU > 0.05 when upsampled and thresholded.

Comparing Layer Differences



Ablating Artifacts

(f) Bedroom images with artifacts



(g) Ablating “artifacts” units improves results

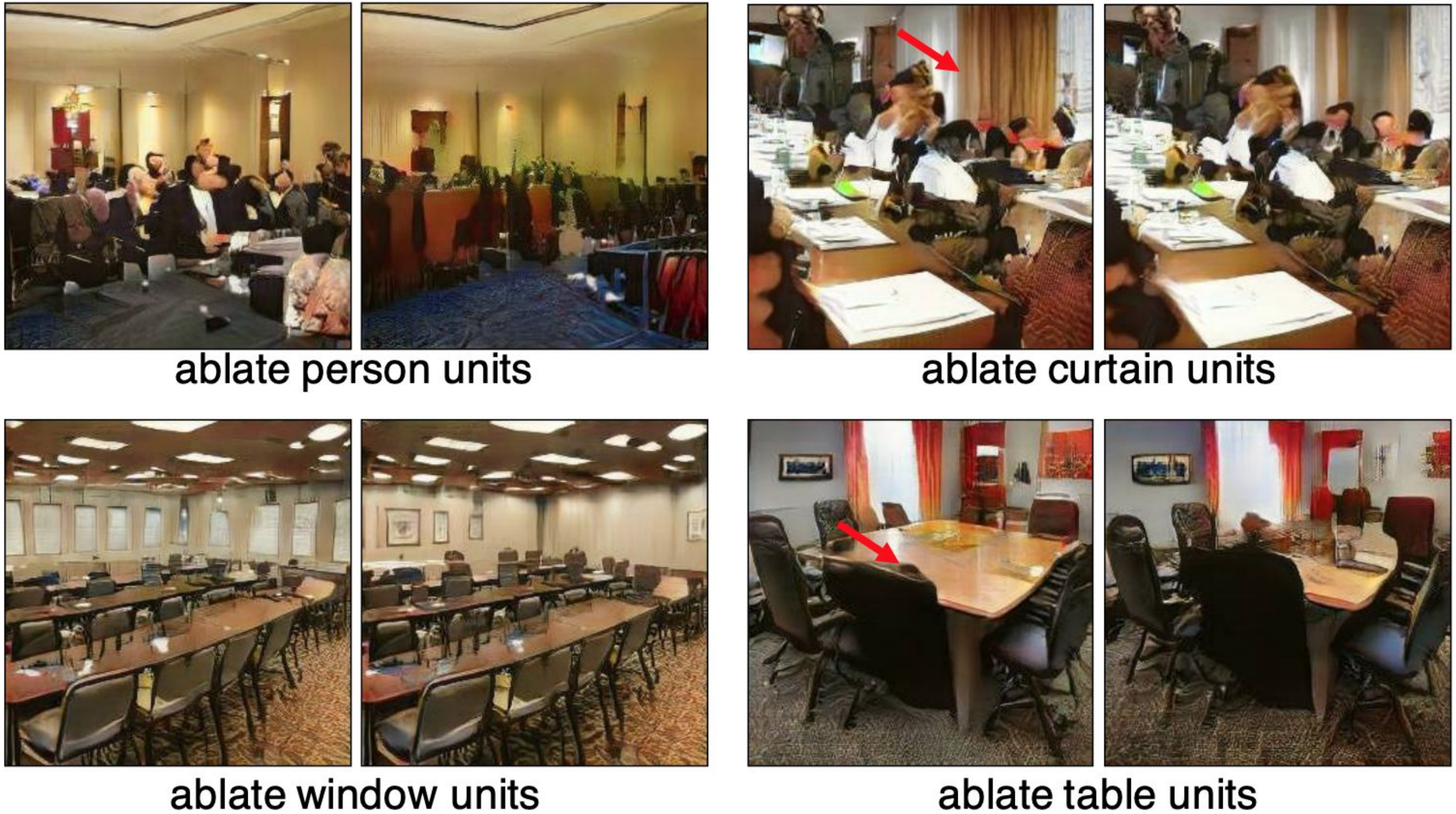
Ablate the 20 artifact-causing units out of 512 units in layer4.

Ablating Artifacts

Fréchet Inception Distance (FID)	
original images	52.87
“artifacts” units ablated (ours)	**32.11**
random units ablated	52.27

Human preference score	original images
“artifacts” units ablated (ours)	**79.0%**
random units ablated	50.8%

Ablating Objects



Inserting Objects

(a)

(b)

(c)

(d)