GAN DISSECTION: VISUALIZING AND UNDERSTANDING GENERATIVE ADVERSARIAL NETWORKS

David Bau, Jun-Yan Zhu,Joshua B. Tenenbaum, William T. Freeman, Antonio Torralba
[davidbau@csail.mit.edu , junyanz@csail.mit.edu , jbt@csail.mit.edu, billf@csail.mit.edu, torralba@csail.mit.edu]

Massachusetts Institute of Technology, Boston

Hendrik Strobelt
hendrik.strobelt@ibm.com

IBM Research, Cambridge MA

Bolei Zhou
bzhou@ie.cuhk.edu.hk

The Chinese University of Hong Kong

Presentation by: Tejas Gaikwad
MT19AI021
Indian Institute of Technology, Jodhpur, India

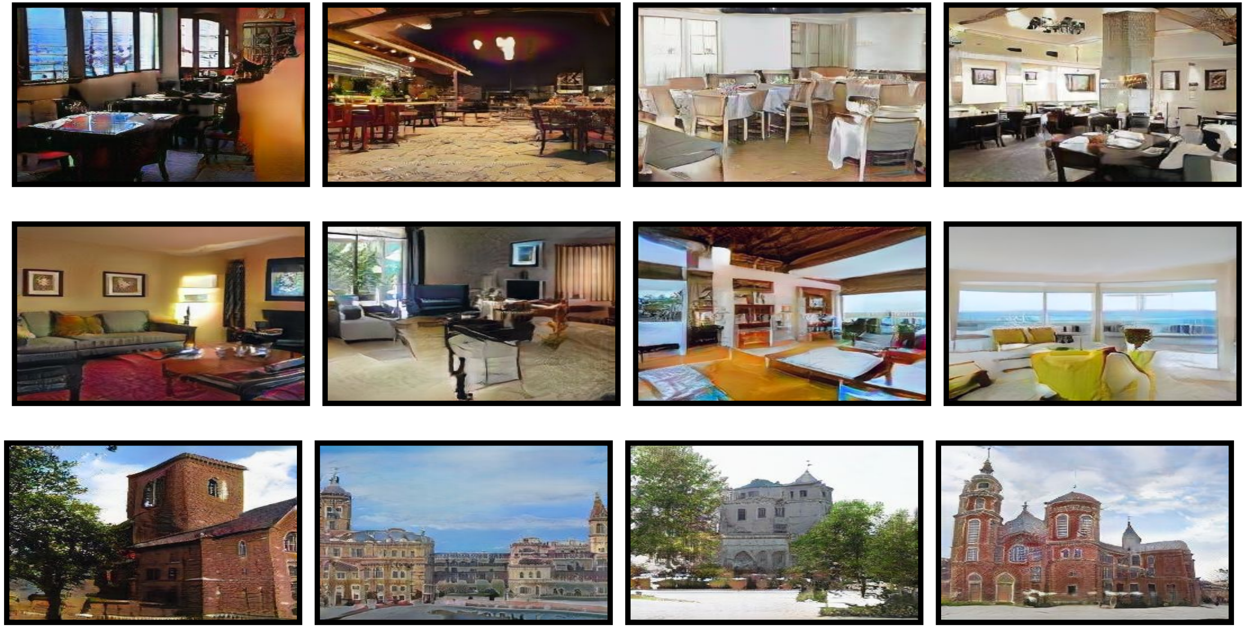
Conference paper at ICLR 2019

Ever Imagined How GAN can give so realistic FAKE Images?.....

Restaurant

Living room

Church



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

Credits : arXiv:1811.10597

To render a beautiful scene, What does a GAN need to know?

Church

Credits : arXiv:1811.10597

And sometimes….
What causes the mistakes?

Bedroom

Credits : arXiv:1811.10597

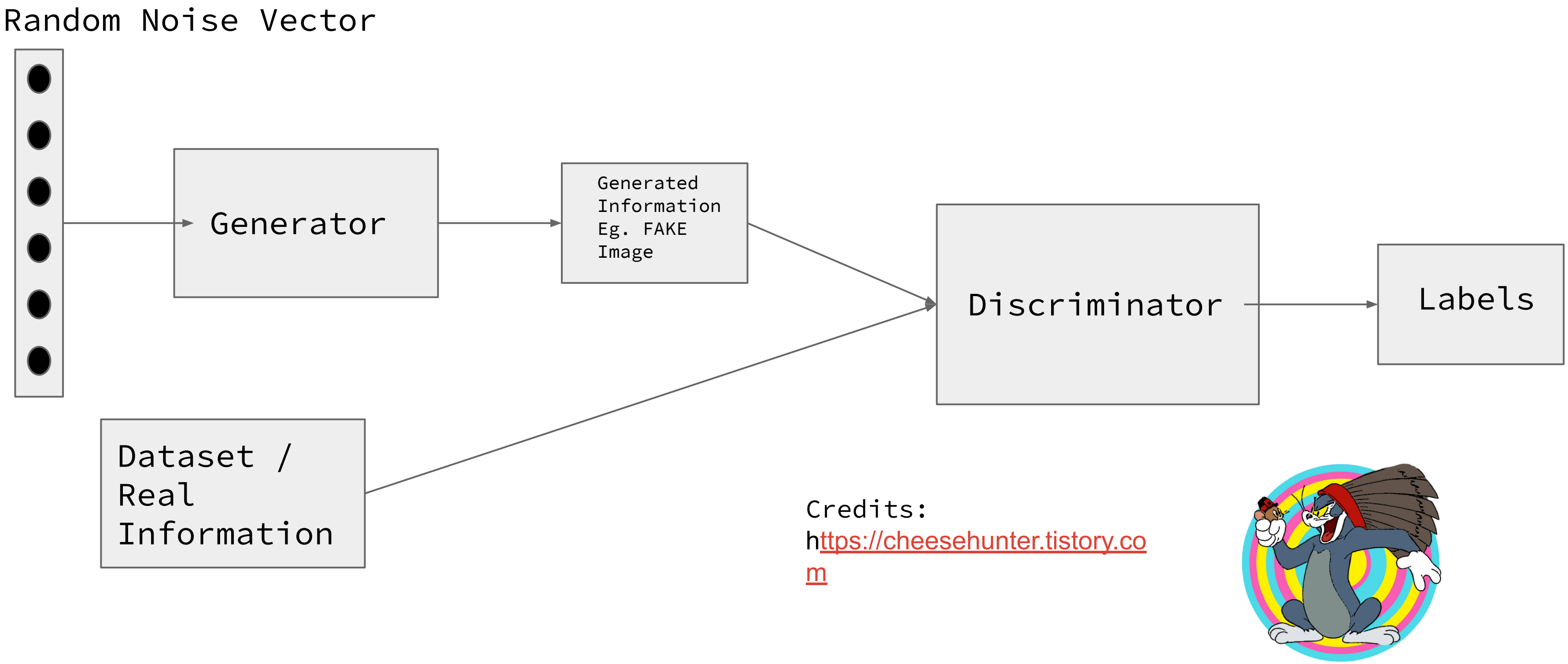
But Before Those Questions...

- What are Generative Adversarial Networks (GAN's) ?
- What makes them so "interesting" ?

GAN's

- These are the networks that belong to the **Generative Model**, **introduced by Ian Goodfellow in the year 2014**.
- GAN's are used to create new images which are not in the dataset, but looks natural.
- A GAME based approach is used to train the model.
- It consists of 2 main blocks named as a **Discriminator** and a **Generator**. Discriminator simply try to discriminate between generated images and images from the dataset. Whereas, generator generates the FAKE image (generated using random noise) and try to match the natural images in the dataset.

Block Diagram for Gan



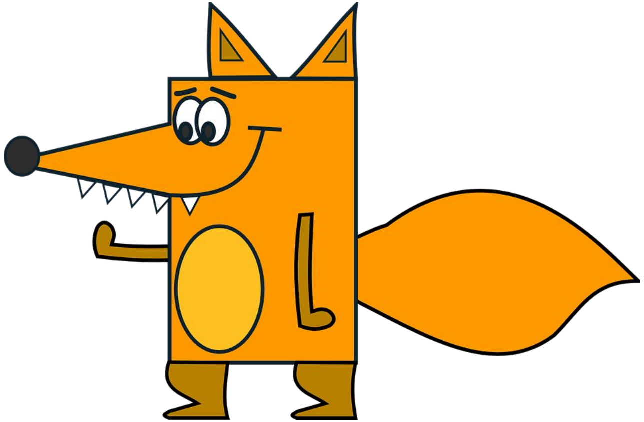
Credits:
https://cheesehunter.tistory.com

Literature Survey

- **GAN(Generative Adversarial Networks) (Goodfellow et al., 2014):** Introduced GAN's
- **Visualizing deep neural networks:** Visualizations for CNNs (Zeiler & Fergus, 2014) and RNNs (Karpathy et al., 2016; Strobelt et al., 2018), by locating and reconstructing salient image features (Simonyan et al., 2014; Mahendran & Vedaldi, 2015) or by mining patches that maximize hidden layers' activations (Zeiler & Fergus, 2014), or we can synthesize input images to invert a feature layer (Dosovitskiy & Brox, 2016)
- **Explaining the decisions of deep neural networks:** Explains individual network decisions using informative heatmaps (Zhou et al., 2018b; 2016; Selvaraju et al., 2017) or modified backpropagation (Simonyan et al., 2014; Bach et al., 2015; Sundararajan et al., 2017)

Credits : arXiv:1811.10597

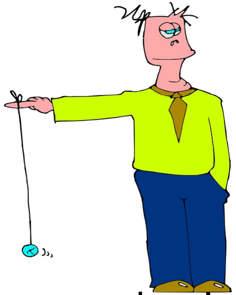
The Proposed model



Goal is **to analyze how objects** such as trees **are encoded by the internal representations** of a GAN generator G: z → x.

Credits : arXiv:1811.10597

Steps...



They’ve presented an analytic framework to visualize and understand GANs at the unit-, object-, and scene-level

- First step is identify a group of interpretable units that are related to Semantic Classes.
- These units’ featuremaps closely match the semantic segmentation of a particular object class (e.g.,trees)
- Then, intervene in units in the network to cause a type of object to disappear or appear
- And Finally, study the contextual relationships by observing where we can insert the object

Credits : arXiv:1811.10597

Now L'il Mathematical explanation

- They have **analysed the internal GAN representations** by decomposing the featuremap r at a layer into positions P ⊂ **P** and unit channels u ∈ **U**
- To identify a unit u with semantic behavior, they have upsampled and thresholded the unit, and measured how well it matches an object class c in the image x as identified by a supervised semantic segmentation network $S_c(x)$
- Each unit is a little segmentation solution. A standard way to measure segmentation accuracy is IoU(Intersection over Union), which is given as

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

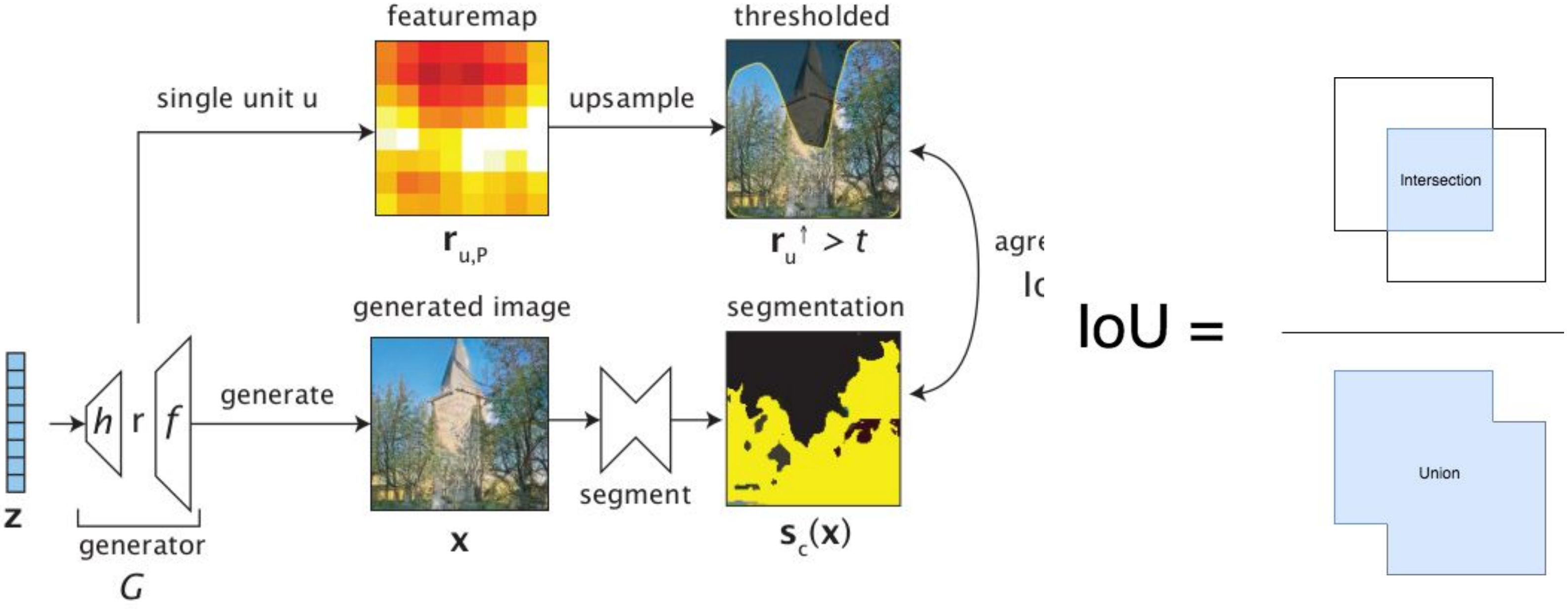
Where the threshold t is given as follows

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

The threshold $t_{u,c}$ is chosen to maximize the information quality ratio, that is, the **portion of the joint entropy H which is mutual information I**

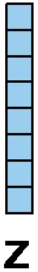
Credits : arXiv:1811.10597

HOW UNITS CORRELATE TO AN OBJECT CLASS?



Credits : arXiv:1811.10597

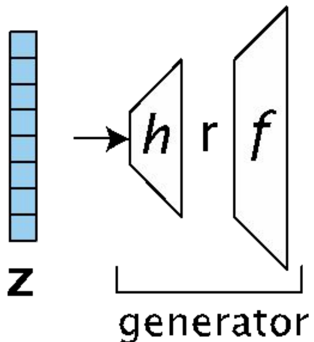
Which units correlate to an object class?



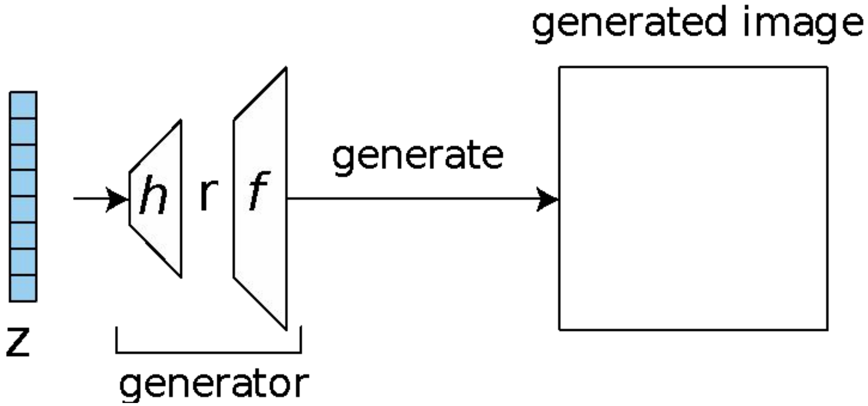
Credits : arXiv:1811.10597

Which units correlate to an object class?

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

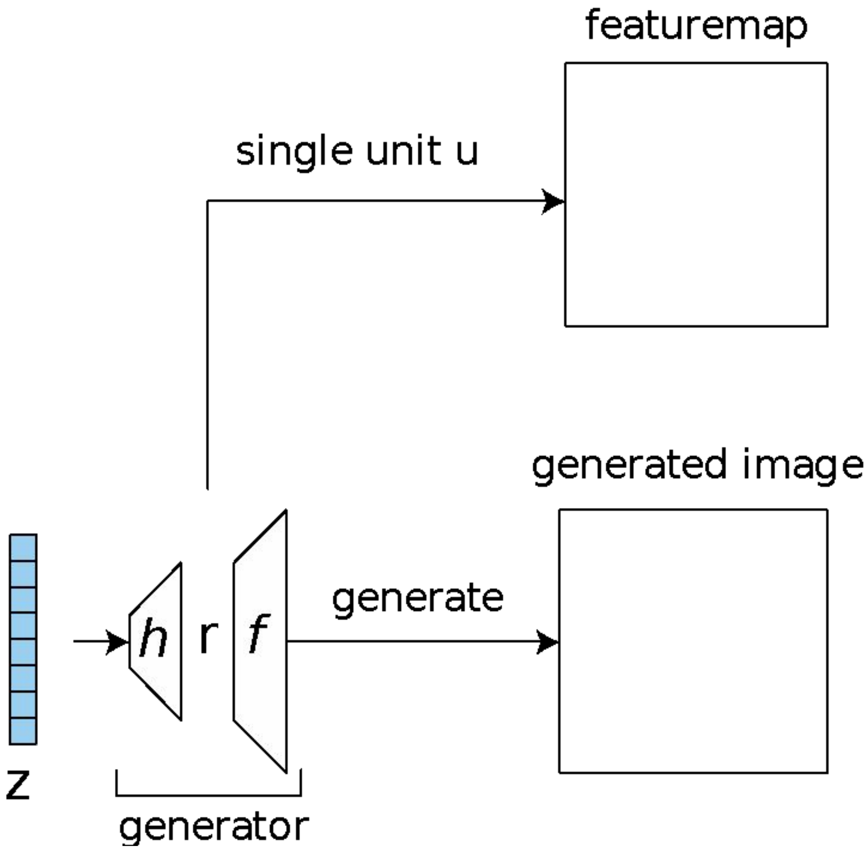


Which units correlate to an object class?



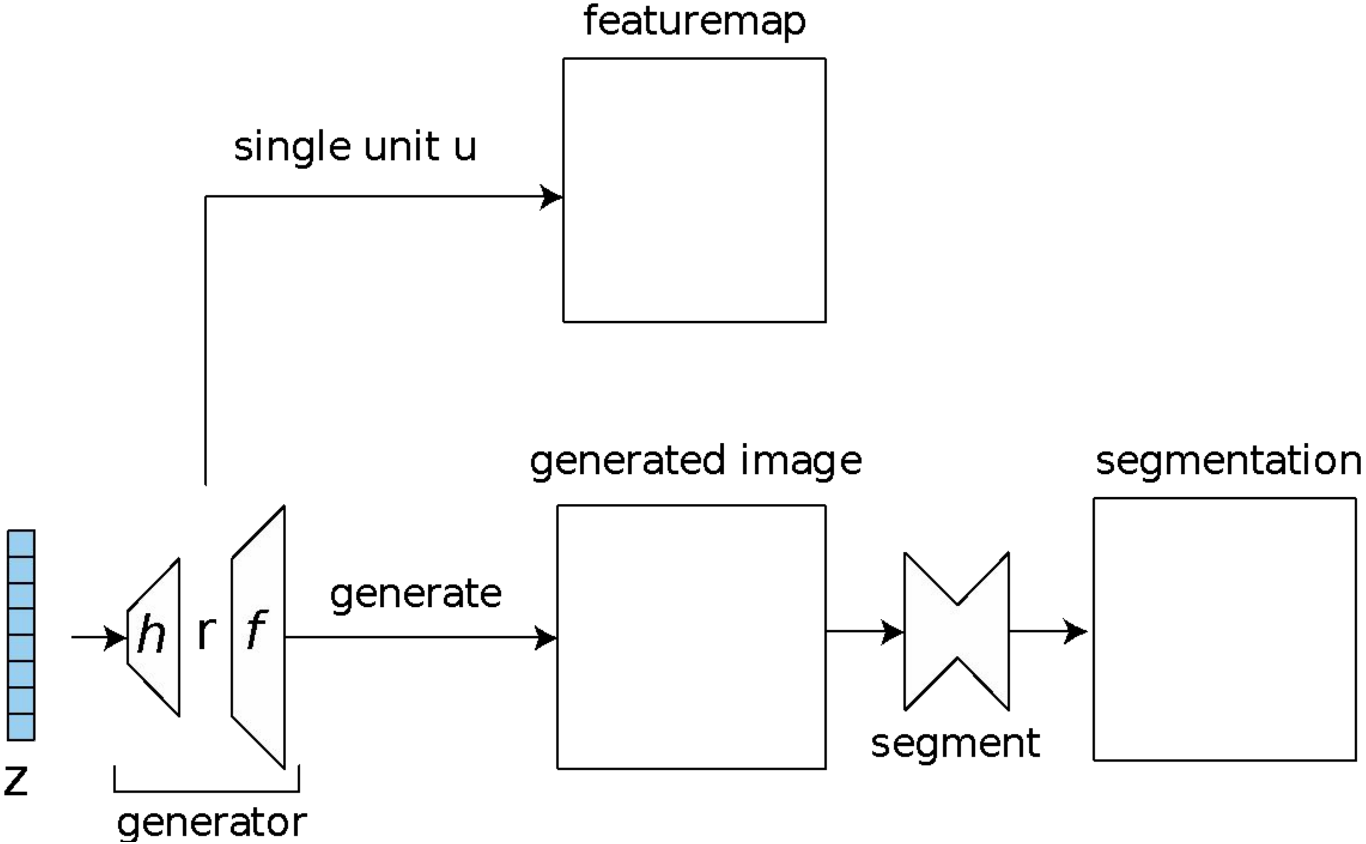
Credits : arXiv:1811.10597

Which units correlate to an object class?



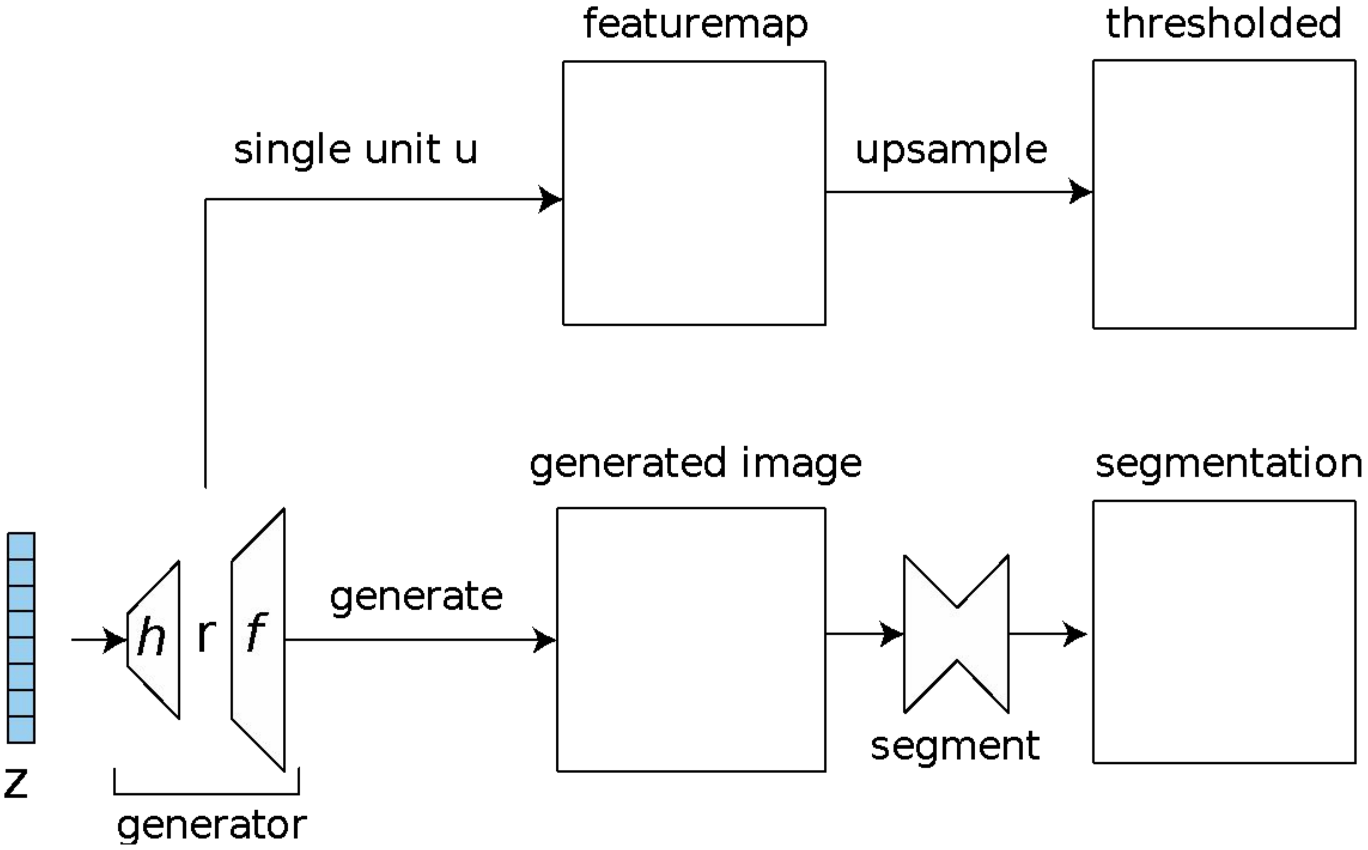
Credits : arXiv:1811.10597

Which units correlate to an object class?



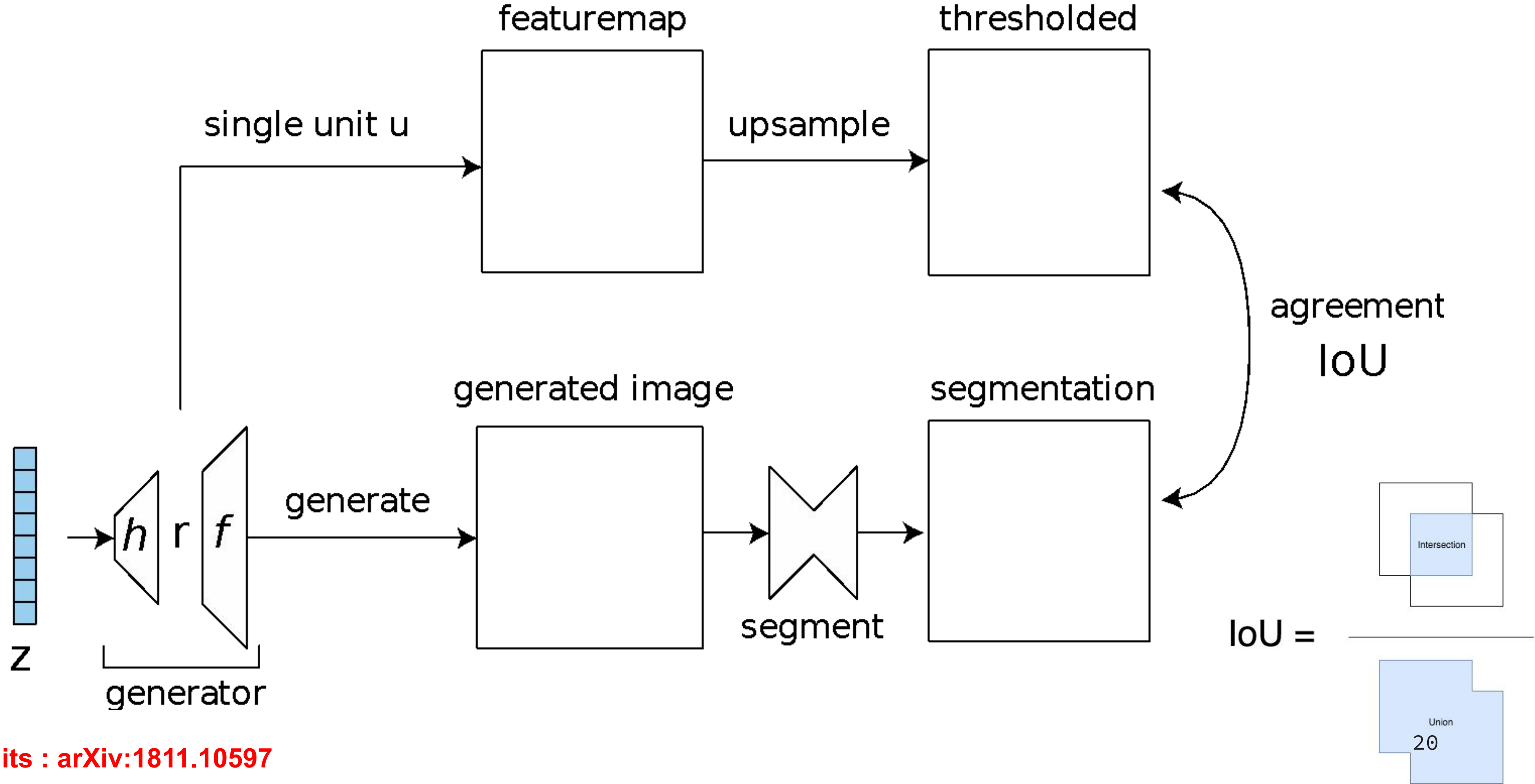
Credits : arXiv:1811.10597

Which units correlate to an object class?



Credits : arXiv:1811.10597

Which units correlate to an object class?



Credits : arXiv:1811.10597

Which units correlate to an object class?

Church samples

Unit: Tree

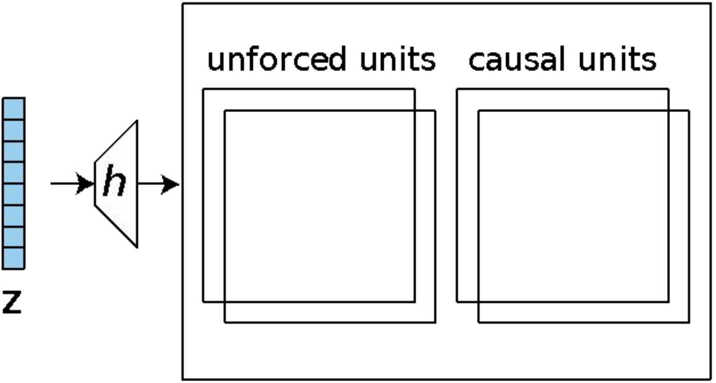
Unit: Dome

Credits : arXiv:1811.10597

Which units cause an object class?

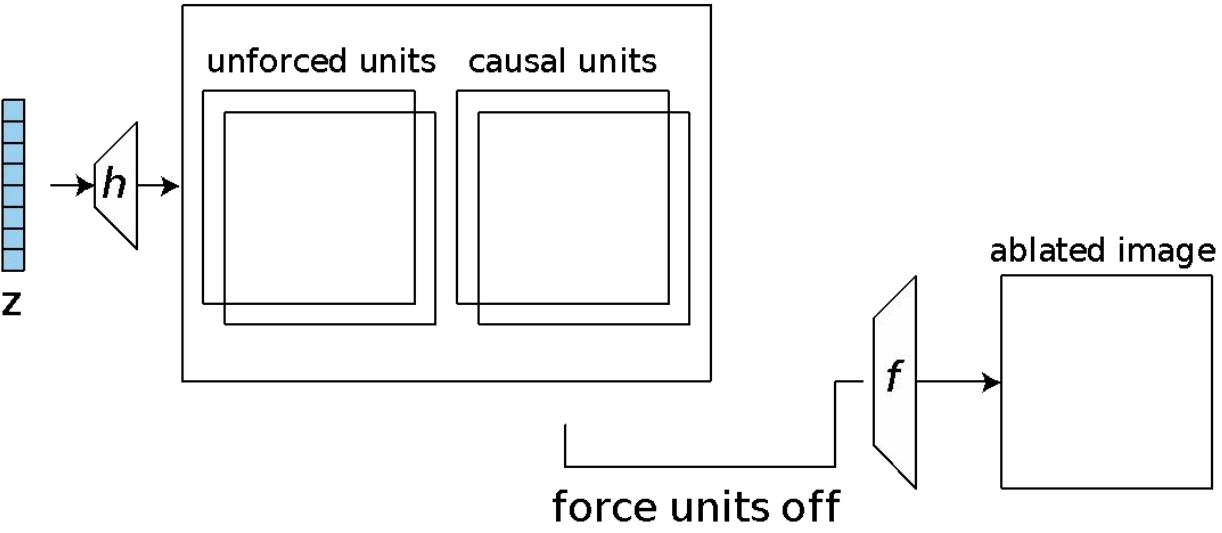
Credits : arXiv:1811.10597

Which units cause an object class?



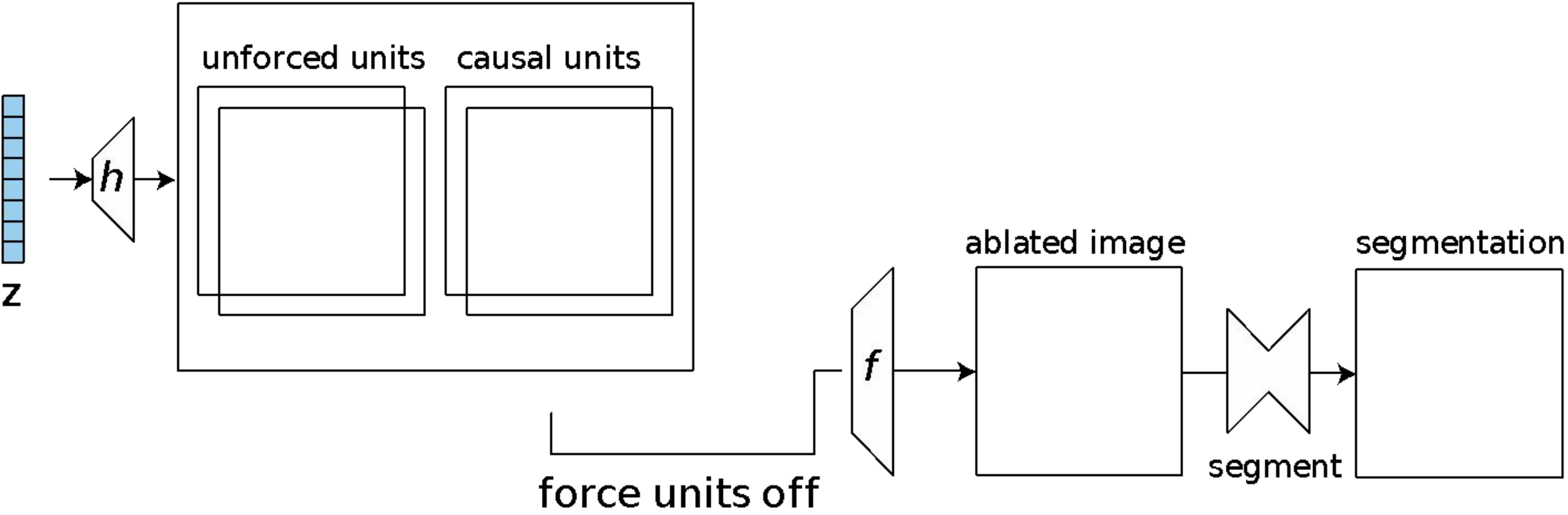
Credits : arXiv:1811.10597

Which units cause an object class?



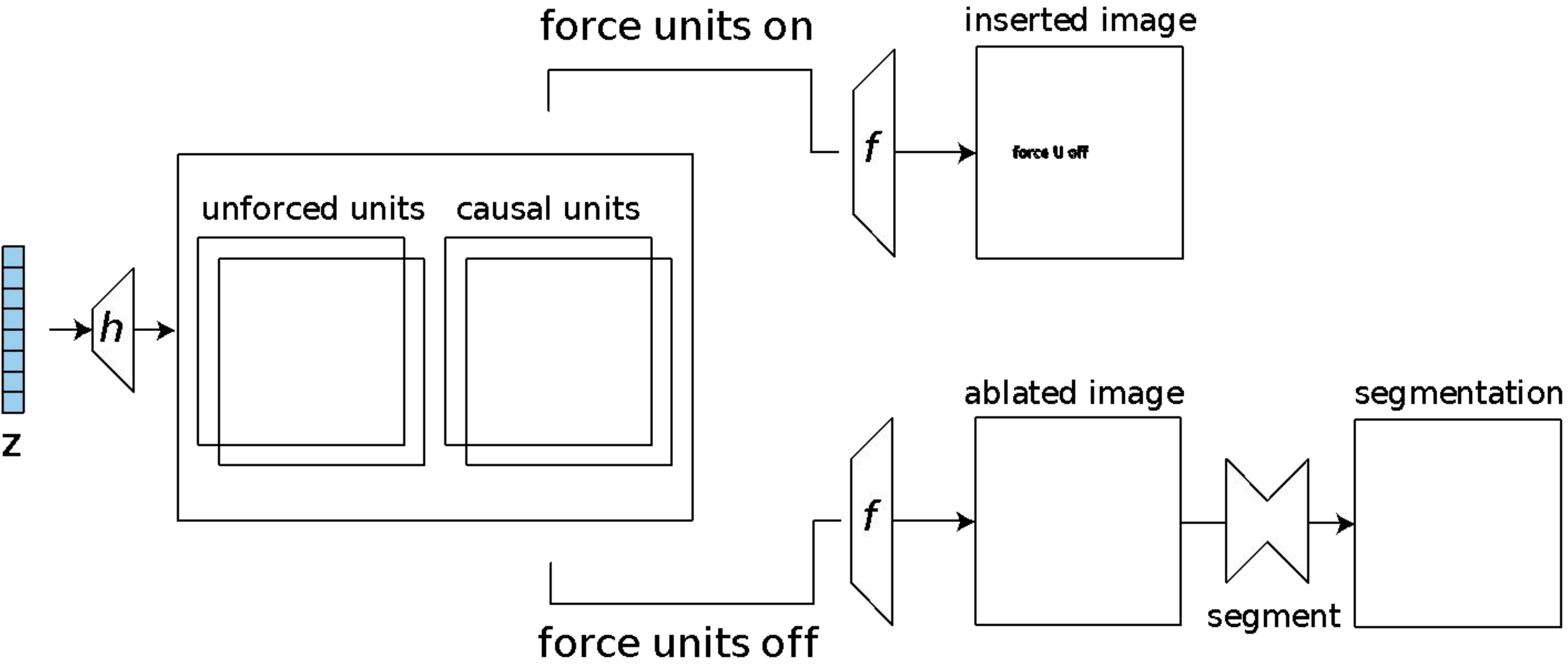
Credits : arXiv:1811.10597

Which units cause an object class?



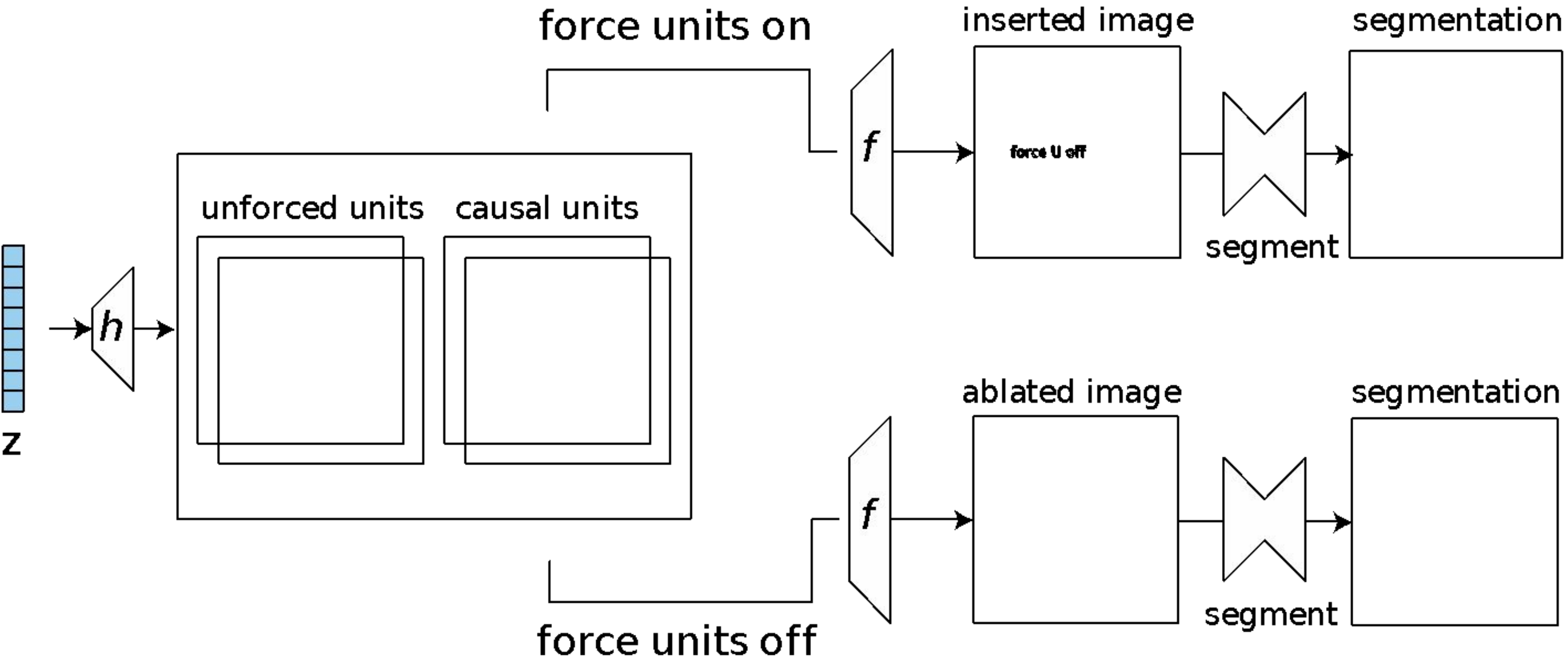
Credits : arXiv:1811.10597

Which units cause an object class?



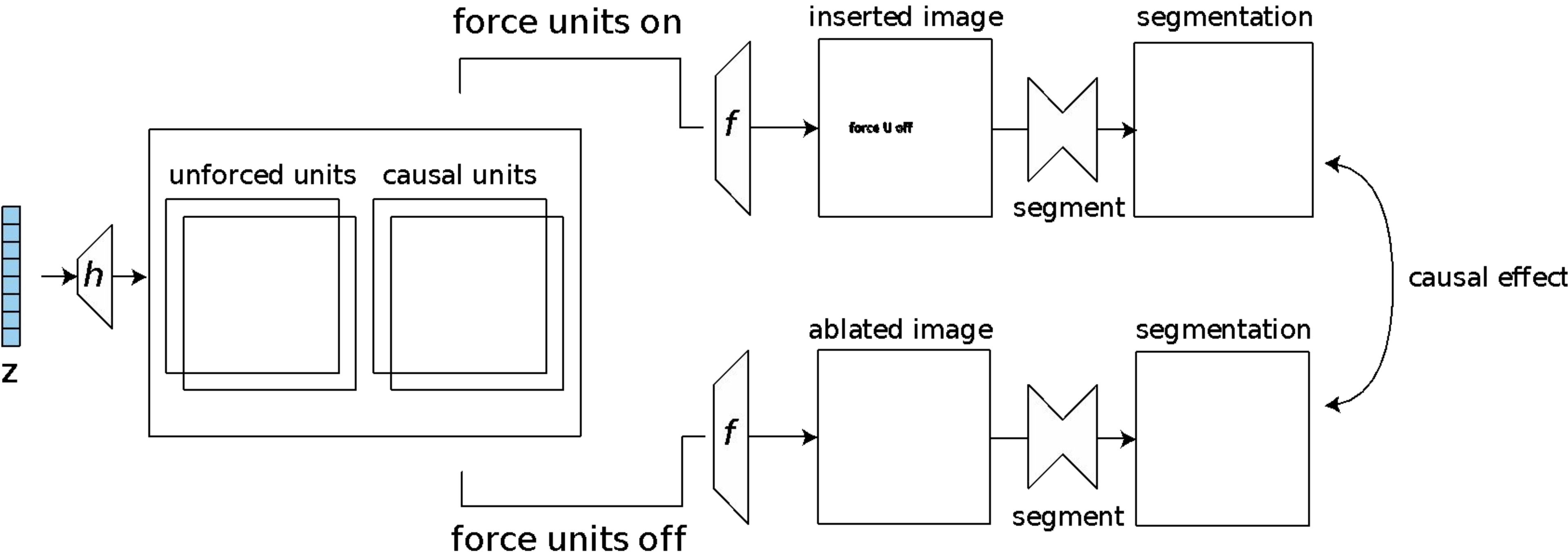
Credits : arXiv:1811.10597

Which units cause an object class?



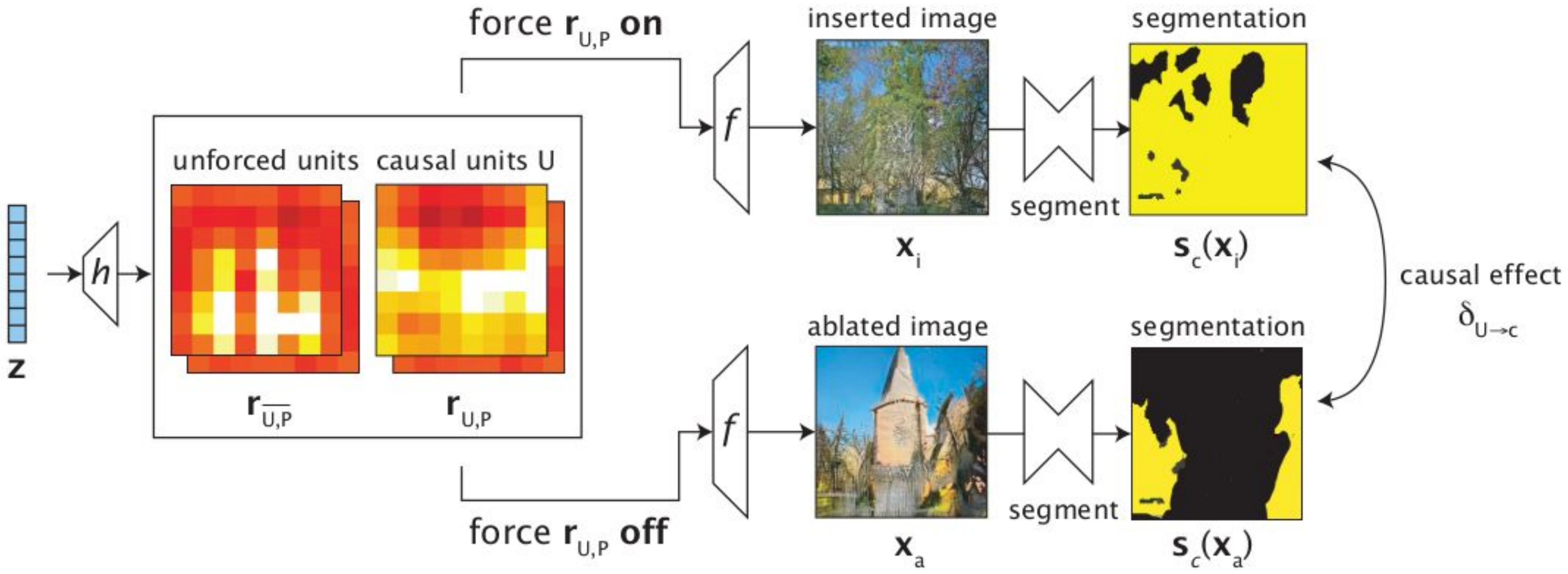
Credits : arXiv:1811.10597

Which units cause an object class?



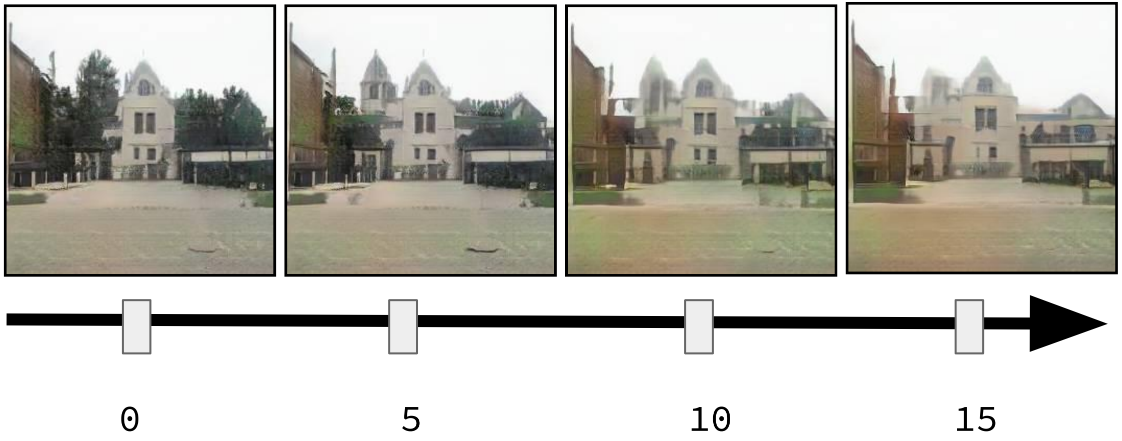
Credits : arXiv:1811.10597

Measuring the relationship between representation units and trees in the output using intervention

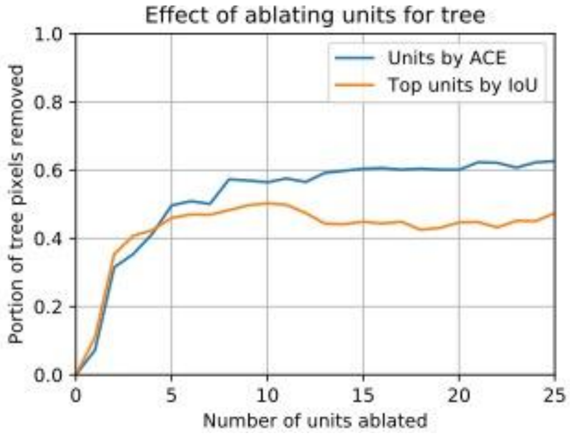


Credits : arXiv:1811.10597

Removing or Adding Units

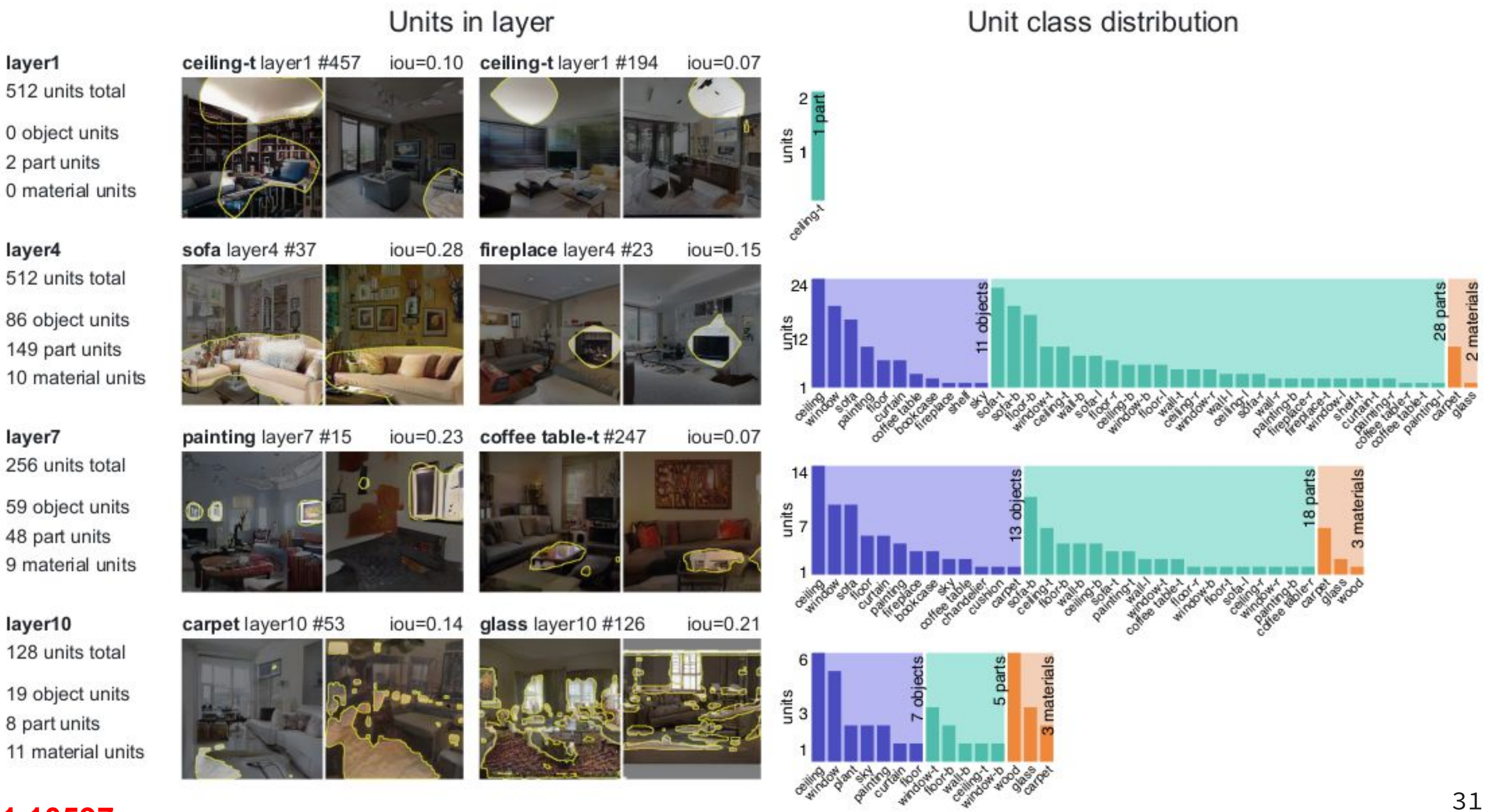


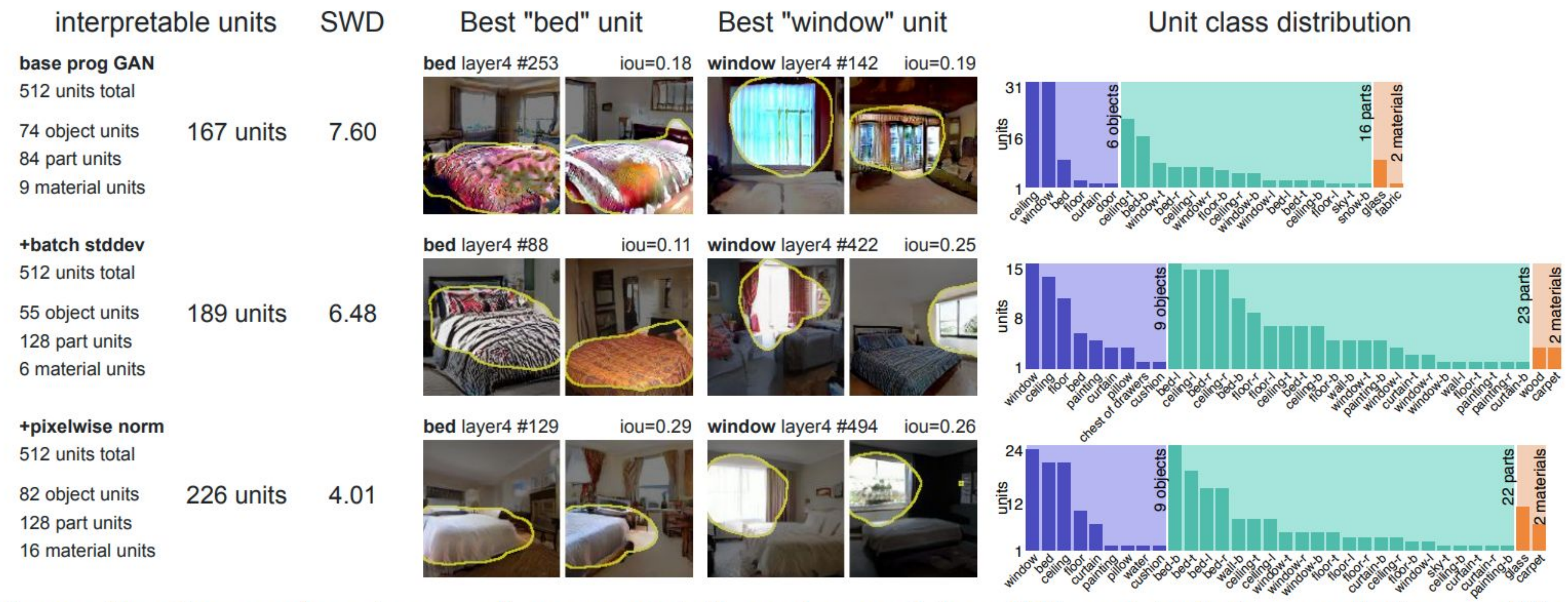
Number of tree units ablated



Credits : arXiv:1811.10597

GAN Dissection: Comparing Datasets , DATASET USED: LSUN





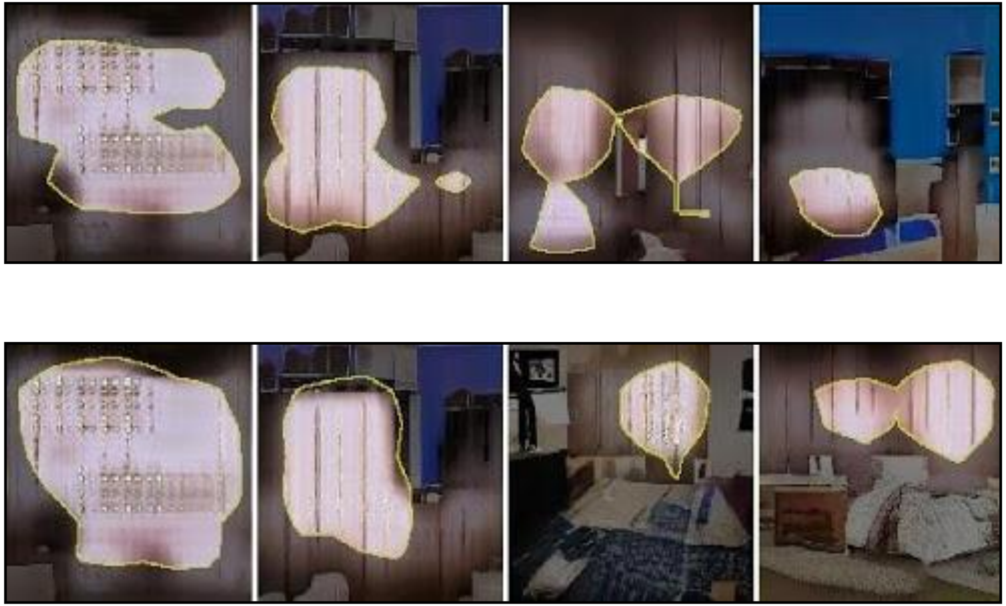
Comparing layer4 representations learned by different training variations. Sliced Wasserstein Distance (SWD) is a GAN quality metric suggested by Karras et al. (2018): lower SWD indicates more realistic image statistics

Credits : arXiv:1811.10597

Debugging and Improving GANs

Unit #63

Unit #231



Example artifact-causing units

Bedroom images with artifacts

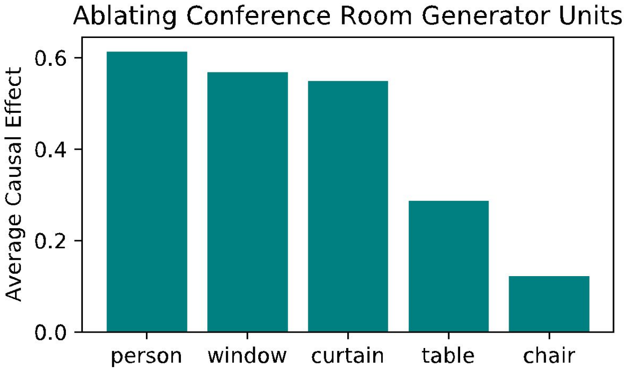
Ablating “artifact” units improves results

Credits : arXiv:1811.10597

Object-Scene Relationship

ablate person

ablate curtain



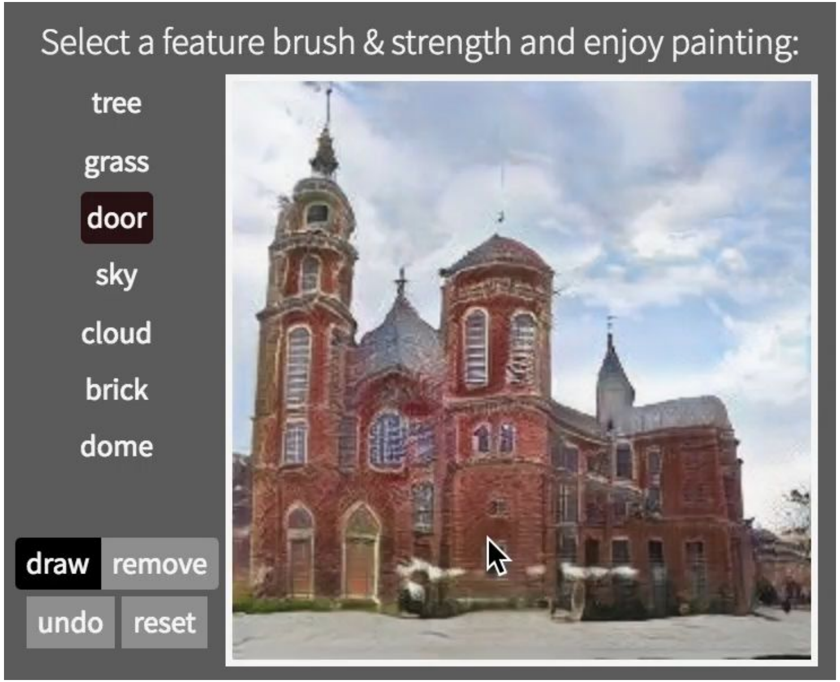
ablate window

ablate table
units

ablate chair
units

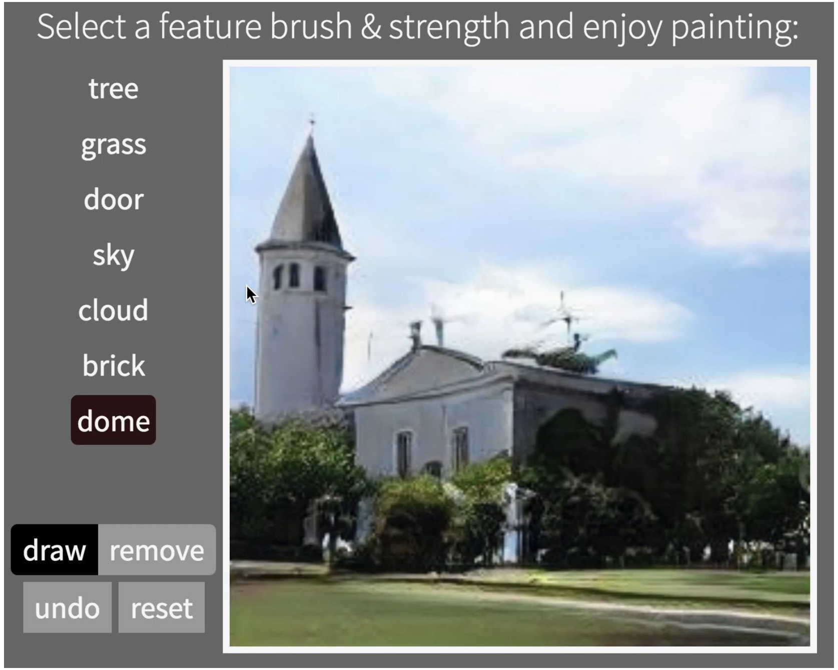
Credits : arXiv:1811.10597

Paint with GANs



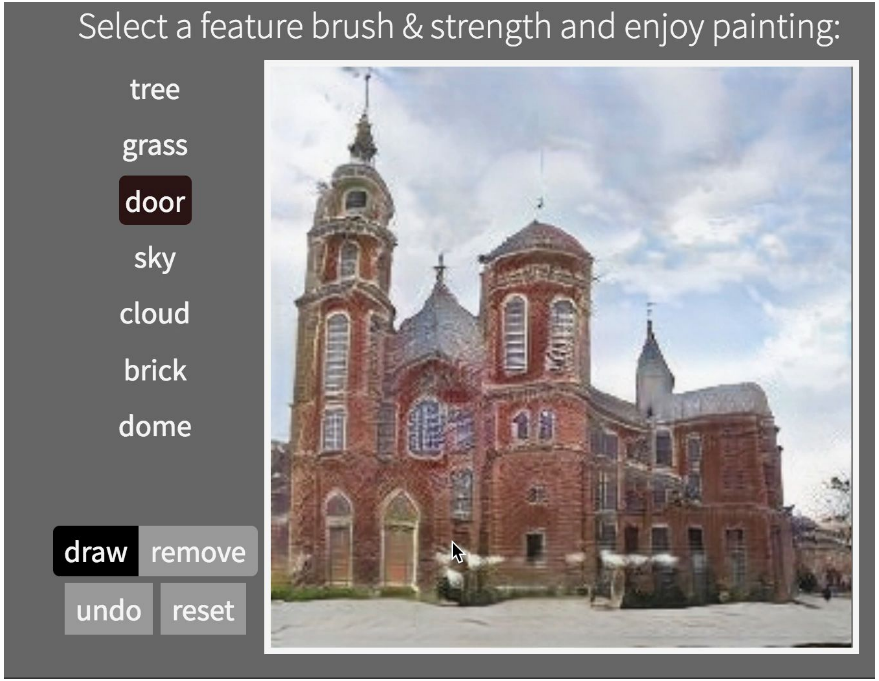
Credits : arXiv:1811.10597

Paint with GANs



Credits : arXiv:1811.10597

Paint with GANs



Credits : arXiv:1811.10597

THANK YOU!

Credits:
https://wallpapersprinted.com/wallpaper/2/minions_14.html

Credits : arXiv:1811.10597