

# GAN Dissection: Visualizing and Understanding Generative Adversarial Networks

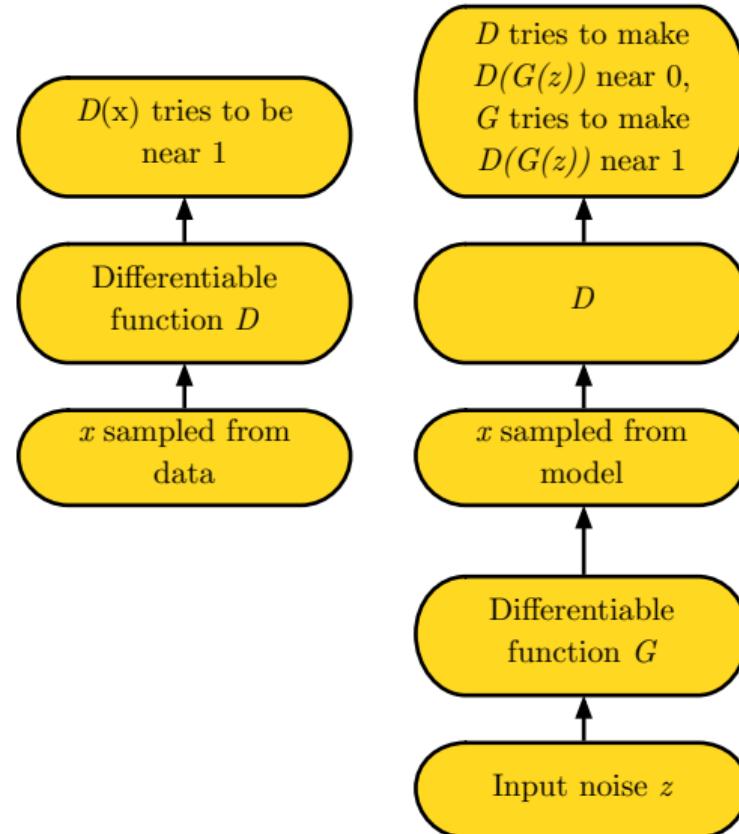
Martin Šafránek

Faculty of Information Technology  
Czech Technical University in Prague



- GAN overview
- GAN dissection (paper)

# Adversarial Nets Framework



# Minimax Game

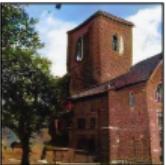
$$J^{(D)} = -\frac{1}{2} \mathbb{E}_{x \sim p_{\text{data}}} \log D(x) - \frac{1}{2} \mathbb{E}_z \log (1 - D(G(z)))$$

$$J^G = -J^D$$

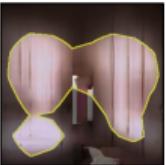
- Generator minimizes the log-probability of the discriminator being correct
- Equilibrium if the discriminator is unable to differentiate between real and generated input

# The paper

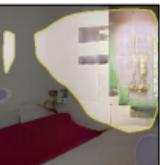
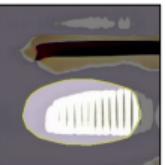
- presents method for visualizing and understanding GAN
- assumes GAN generator is implemented as Convolutional neural network
- presents that learned GAN contains variables for doors, trees, ...
- shows how we can interactively manipulate objects in a scene



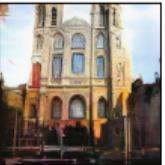
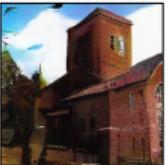
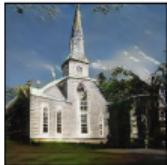
(a) Generate images of churches



(b) Identify GAN units that match trees



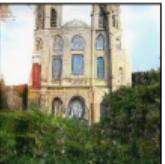
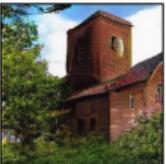
(e) Identify GAN units that cause artifacts



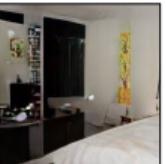
(c) Ablating units removes trees



(f) Bedroom images with artifacts



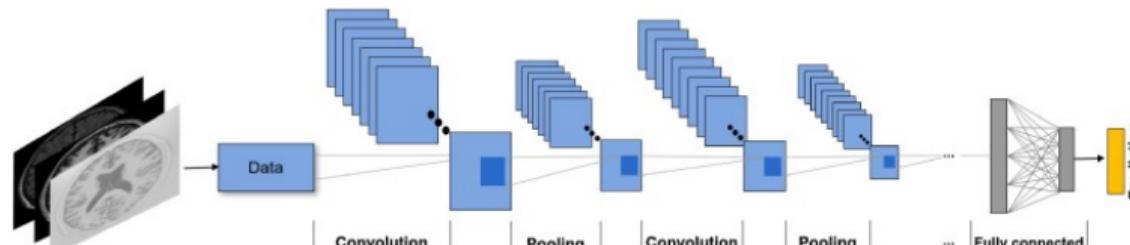
(d) Activating units adds trees



(g) Ablating “artifact” units improves results

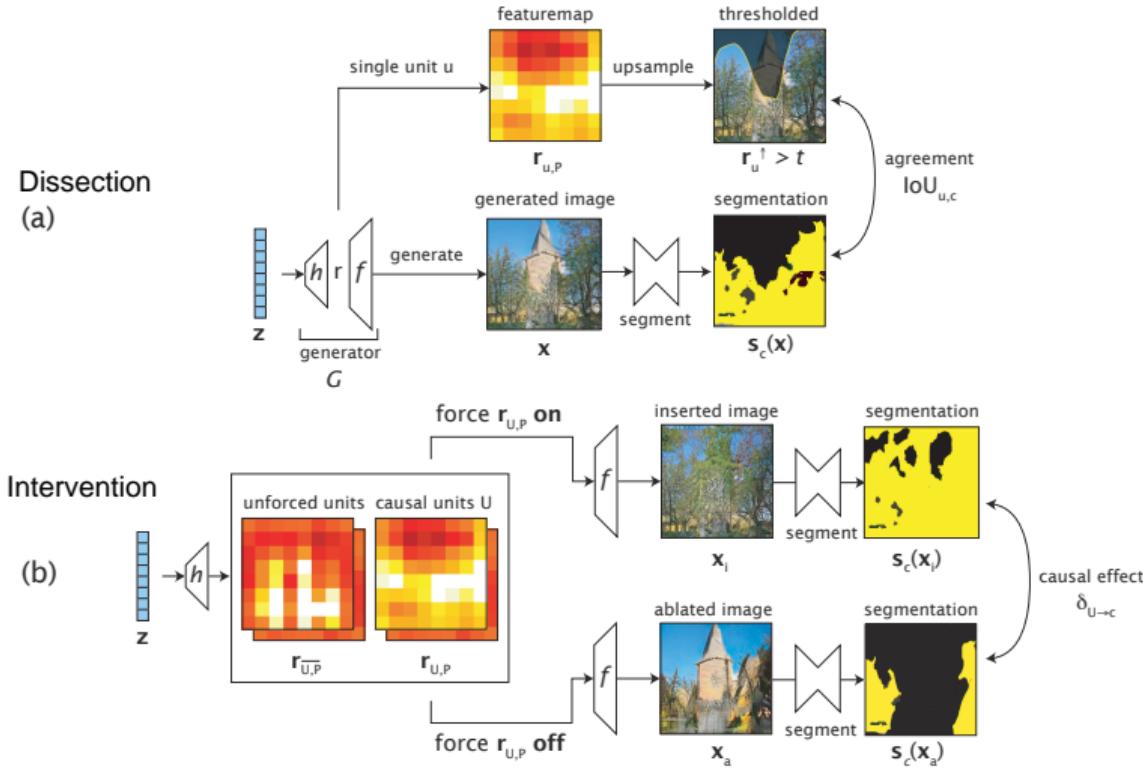
# Terminology

- function  $\mathbf{G} : z \rightarrow x$  (generator  $G$  generates image  $x$  from latent vector  $z$ )
- tensor  $r$ , where  $r = h(z)$ ,  $x = f(r) = f(h(z)) = G(z)$  (featuremaps)
- $r$  has all the data necessary to produce image  $x$
- concept  $c \in \mathbb{C}$  (chair, door, ...)
- $r_{U,P} = (r_{U,P}, r_{\bar{U},P})$
- $U$  selected featuremaps (units)



Covnet example [Source]

# Architecture



# Dissection

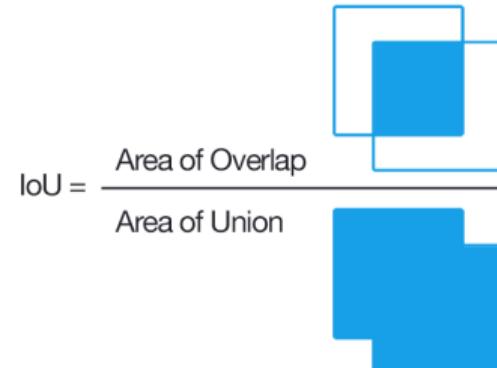
Which units correlate with which context

- 1 Upsample unit to output image resolution
- 2 Threshold unit
- 3 Compute IoU score

threshold is computed by maximizing information quality ratio  $I/H$

label each concept with unit that matches it best (largest IoU score)

$$IoU_{u,c} \equiv \frac{\mathbb{E}_z |(r_{u,\mathbb{P}}^\uparrow > t_{u,c}) \wedge s_c(x)|}{\mathbb{E}_z |(r_{u,\mathbb{P}}^\uparrow > t_{u,c}) \vee s_c(x)|} \subseteq [0, 1]$$



Source

# Intervention 1/2

- unit that correlates highly with a generated output might not cause its generation
  - intervention identifies combination of units that cause object generation
- 1 select some units and one concept
  - 2 force units on/off
  - 3 measure See if concept appears/disappears

Original image :

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

Image with U ablated at pixels P :

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

Image with U inserted at pixels P :

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

## Intervention 2/2 — Finding sets of units with high ACE

- Average causal effect (ACE):  $\delta_{U \rightarrow c} \equiv \mathbb{E}_{z,P}[s_c(x_i)] - \mathbb{E}_{z,P}[s_c(x_a)]$
- ineffective, as there are  $\binom{d}{|U|}$  subsets
- instead optimize  $\alpha \in [0, 1]^d$

Image with partial ablation at pixels P :

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

Image with partial insertion at pixels P :

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

Objective :

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

# Interpretable units for different scene categories

conference rm

**table #96**



iou=0.30

**person-b #91**



iou=0.21

**seat #83**



iou=0.13

dining room

**chandelier-I #184** iou=0.21



**chair-I #456**



iou=0.19

**table #89**



iou=0.31

# Interpretable units for different network layers

**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 layer7 #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**

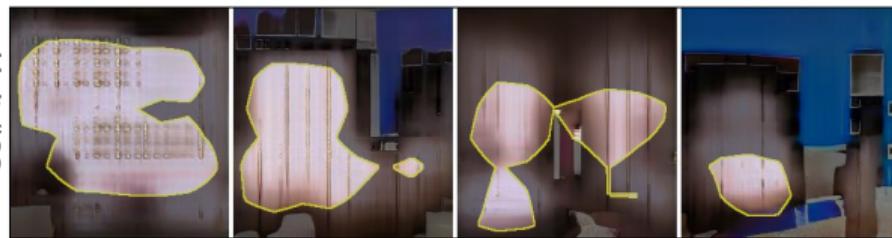
iou=0.21



# Diagnosing and improving GANs

- identify artifacts with human annotation
- 10 minutes to locate 20 artifact-causing units
- ablate found artifacts

Unit #63



Unit #231



(a) Example artifact-causing units

(b) Bedroom images with artifacts



(c) Ablating “artifact” units improves results



# Locating casual units with ablation

- force erasing/size reduction
- Does GAN learns patterns ? eg „*All bedrooms must have windows.*“



Figure 10: Comparing the effect of ablating 20 window-causal units in GANs trained on five scene categories. In each case, the 20 ablated units are specific to the class and the generator and independent of the image. In some scenes, windows are reduced in size or number rather than eliminated, or replaced by visually similar objects such as paintings.

# Characterizing contextual relationships via insertion

- force insertion of features into specific locations in scenes
- GAN forces realtionships eg „Doors can't be added to sky.“, because the choice is vetoed later

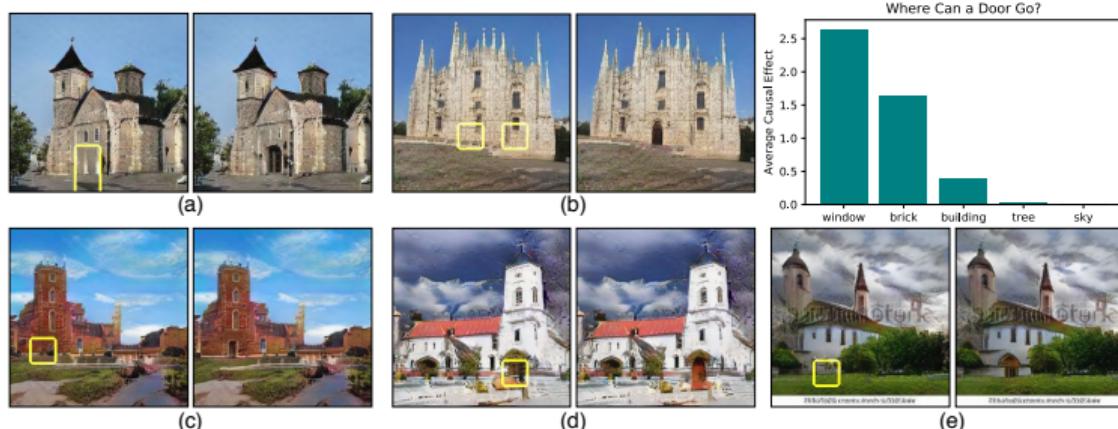


Figure 11: Inserting door units by setting 20 causal units to a fixed high value at one pixel in the representation. Whether the door units can cause the generation of doors is dependent on its local context: we highlight every location that is responsive to insertions of door units on top of the original image, including two separate locations in (b) (we intervene at left). The same units are inserted in every case, but the door that appears has a size, alignment, and color appropriate to the location. Emphasizing a door that is already present results in a larger door (d). The chart summarizes the causal effect of inserting door units at one pixel with different contexts.

# Interactive toolkit

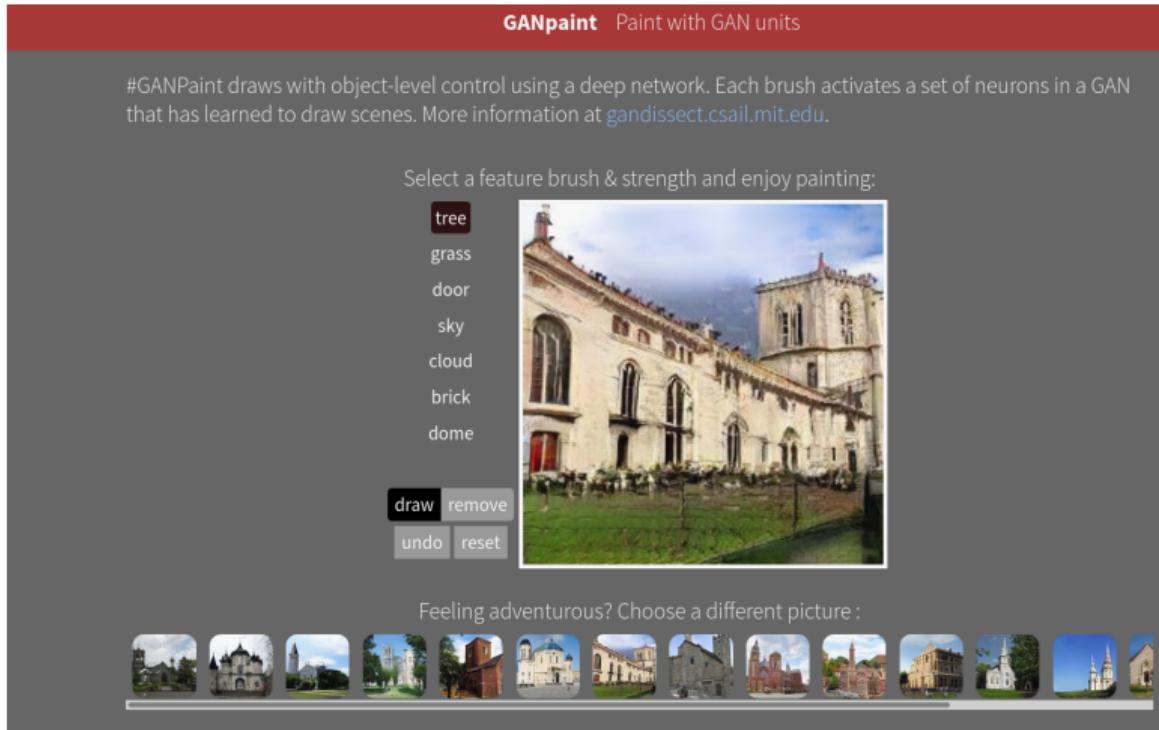
**GANpaint** Paint with GAN units

#GANPaint draws with object-level control using a deep network. Each brush activates a set of neurons in a GAN that has learned to draw scenes. More information at [gandissect.csail.mit.edu](http://gandissect.csail.mit.edu).

Select a feature brush & strength and enjoy painting:

- tree
- grass
- door
- sky
- cloud
- brick
- dome

draw    remove  
 undo    reset



Feeling adventurous? Choose a different picture :



<http://gandissect.res.ibm.com/ganpaint.html>

## Further research suggested by authors

- 1 „*Why can a door not be inserted in the sky ?*“
- 2 „*How does GAN suppress the signal in the later layers ?*“
- 3 „*What are the relationships between layers of a GAN ?*“

# Questions ?