GAN DISSECTION

Аъзам Бехруз

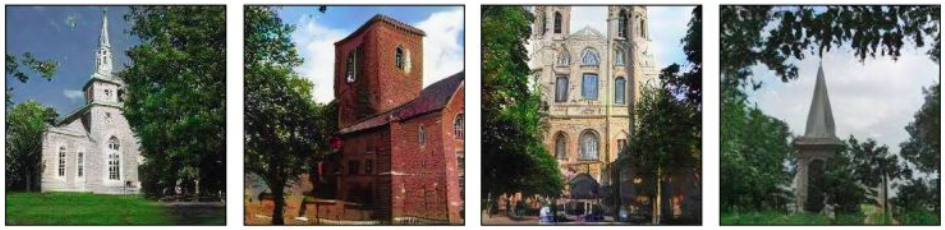
22.02.2022 - палиндром

Интересные вопросы

- Существуют ли во внутренностях GAN понятия об объектах (деревья, двери и т.д.)?
- Есть ли внутри переменные отвечающие за деревья и двери?
- Если да, то каким образом представлено их взаимодействие?

Возможные применения

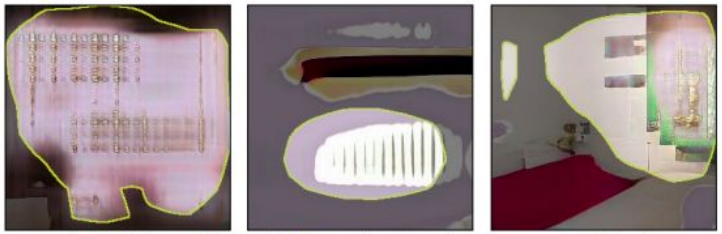
- Манипуляции с объектами
- Удаление артефактов



(a) Generate images of churches

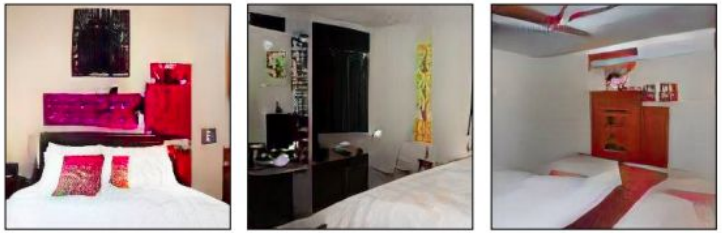
(b) Identify GAN units that match trees

(c) Ablating units removes trees



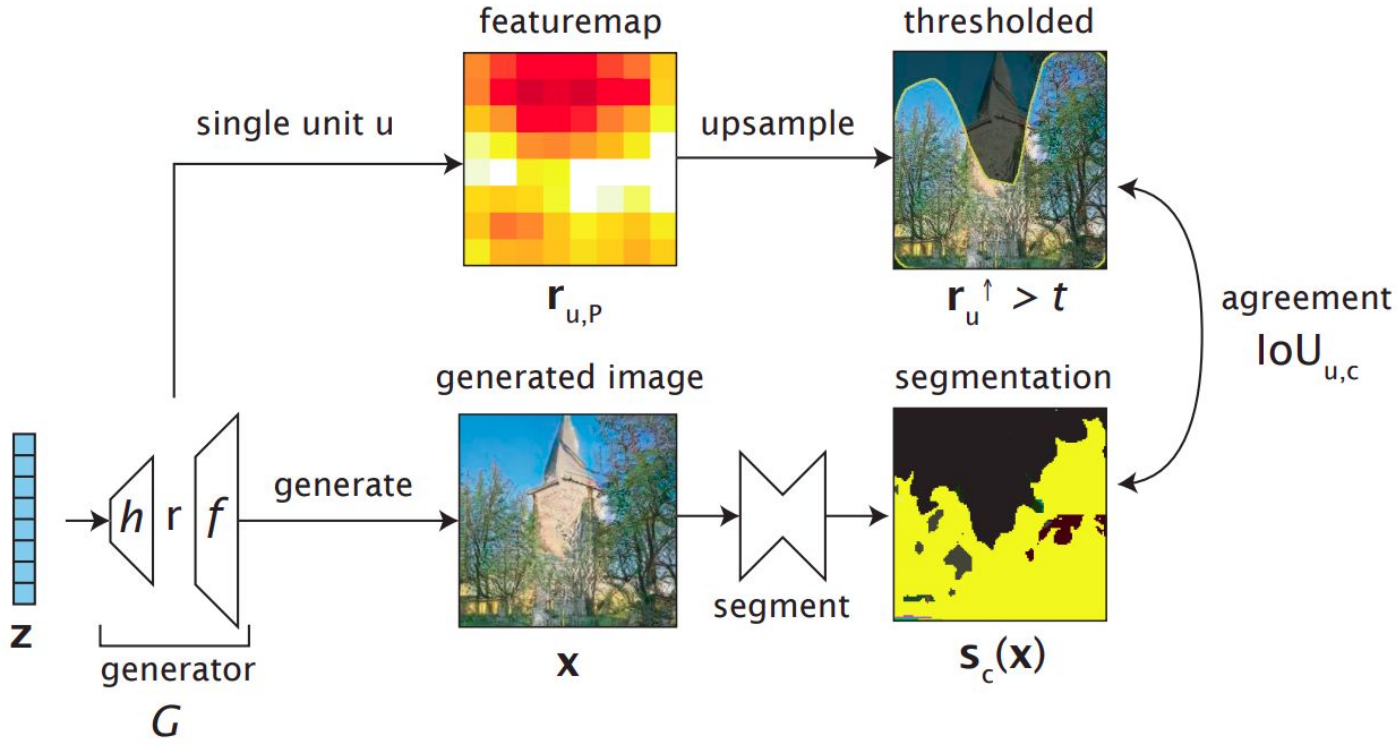
(e) Identify GAN units that cause artifacts

(f) Bedroom images with artifacts



Выявление и вмешательство

Выявление

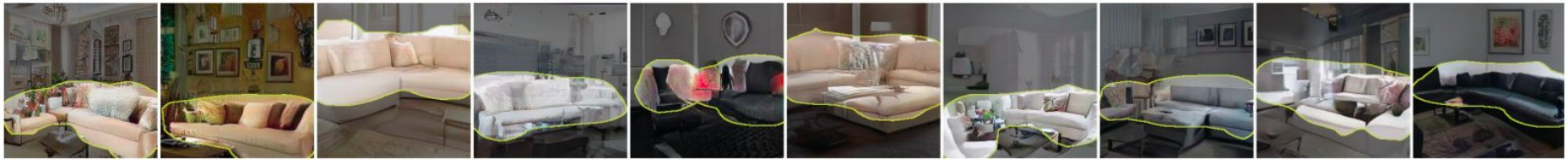

featuremap
thresholded
single unit u
upsample
$\mathbf{r}_{u,P}$
$\mathbf{r}_u^{\uparrow} > t$
agreement
$\mathrm{IoU}_{u,c}$
generated image
segmentation
z
h
r
f
generate
segment
generator
G
x
$\mathbf{s}_c(\mathbf{x})$

Пространственная согласованность

$$\mathrm{IoU}_{u,c} \equiv \frac{\mathbb{E}_{\mathbf{z}} \left| (\mathbf{r}^{\uparrow}_{u,\mathbb{P}} > t_{u,c}) \wedge \mathbf{s}_c(\mathbf{x}) \right|}{\mathbb{E}_{\mathbf{z}} \left| (\mathbf{r}^{\uparrow}_{u,\mathbb{P}} > t_{u,c}) \vee \mathbf{s}_c(\mathbf{x}) \right|}, \text{where } t_{u,c} = \arg\max_t \frac{\mathrm{I}(\mathbf{r}^{\uparrow}_{u,\mathbb{P}} > t; \mathbf{s}_c(\mathbf{x}))}{\mathrm{H}(\mathbf{r}^{\uparrow}_{u,\mathbb{P}} > t, \mathbf{s}_c(\mathbf{x}))},$$

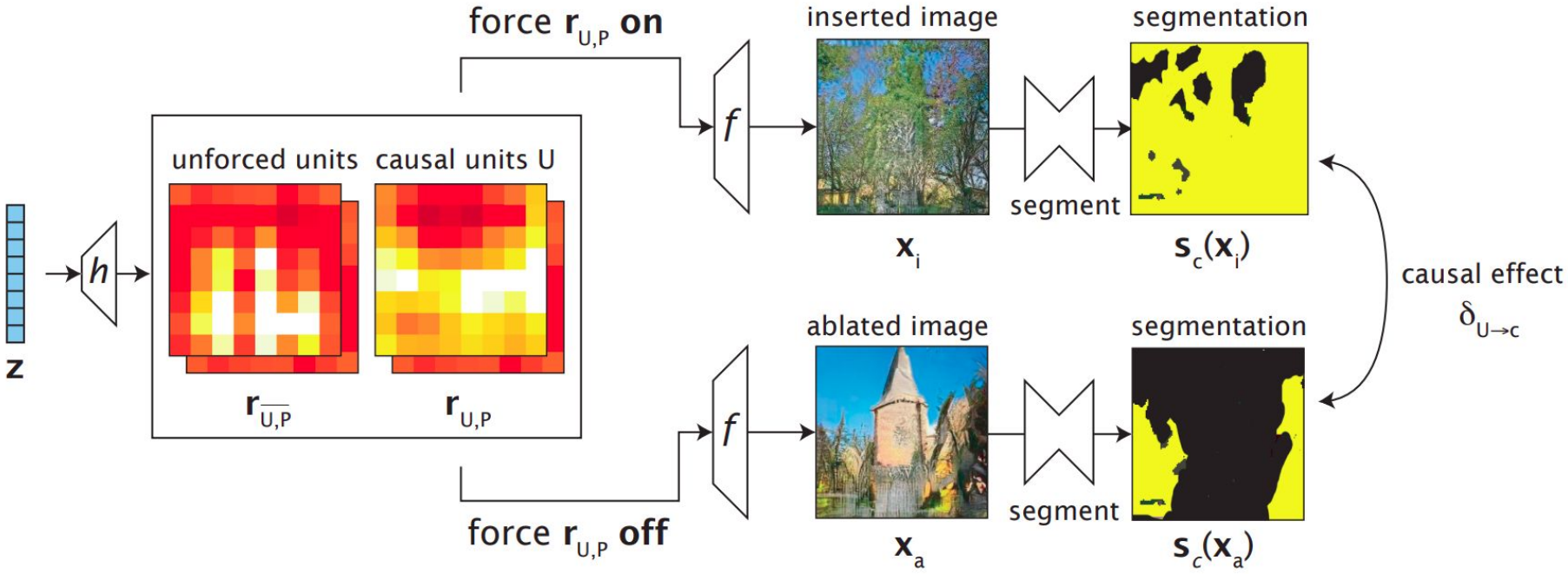
Мы присваиваем каждому слою класс с которым у него наибольшая IoU

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.

Вмешательство



Подсчет каузального эффекта

$$\mathbf{r} = h(\mathbf{z}) \text{ and } \mathbf{x} = f(\mathbf{r}) = f(h(\mathbf{z})) = G(\mathbf{z})$$

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

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

Комбинаторный взрыв!

Вмешательство

$$\boldsymbol{\alpha} \in [0,1]^d$$

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

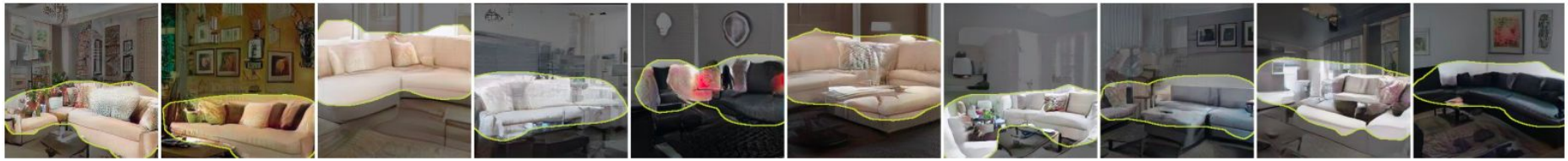
Image with partial insertion at pixels P : $\mathbf{x}'_i = f(\boldsymbol{\alpha} \odot \mathbf{k} + (\mathbf{1} - \boldsymbol{\alpha}) \odot \mathbf{r}_{\mathbb{U},\mathrm{P}},\ \mathbf{r}_{\mathbb{U},\bar{\mathrm{P}}})$

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

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

Выводы

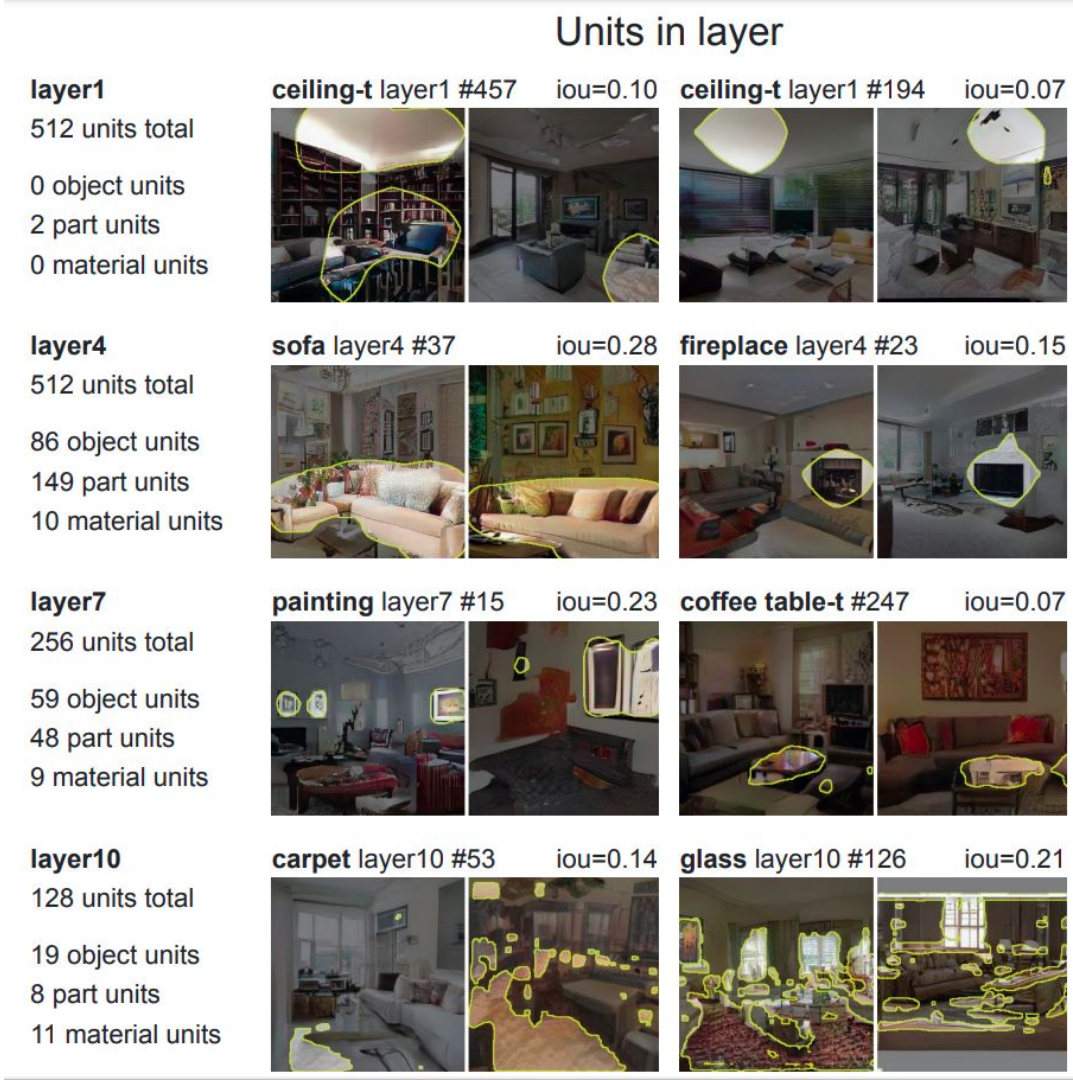
Thresholding unit #65 layer 3 of a dining room generator matches ‘table’ segmentations with IoU=0.34.



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

GAN имеет такое же понятие о столах и диванах что и мы

- Слои с 4-7 коррелируют с семантическими объектами
- 10 слой и далее отвечают за группы пикселей (углы, материал)



(b) Bedroom images with artifacts

(a) Example artifact-causing units

(c) Ablating "artifact" units improves results

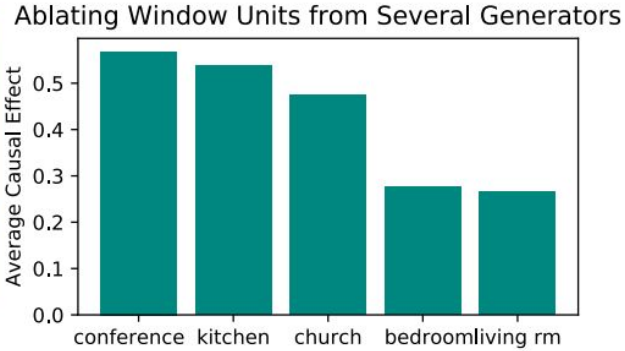
Fréchet Inception Distance (FID)	
original images	43.16
"artifacts" units ablated (ours)	**27.14**
random units ablated	43.17

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

Можем удалять артефакты

conference room

church



kitchen

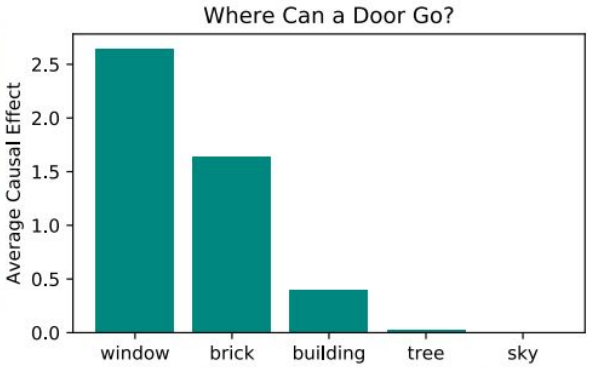
living room

bedroom

Не всегда можно просто взять и удалить окно.
Все зависит от контекста.

(a)

(b)



(c)

(d)

(e)

Можем вставлять двери куда хотим!

Understanding GANs

GANDissection

Interacting with GANs

GANPaint

Спасибо за внимание!