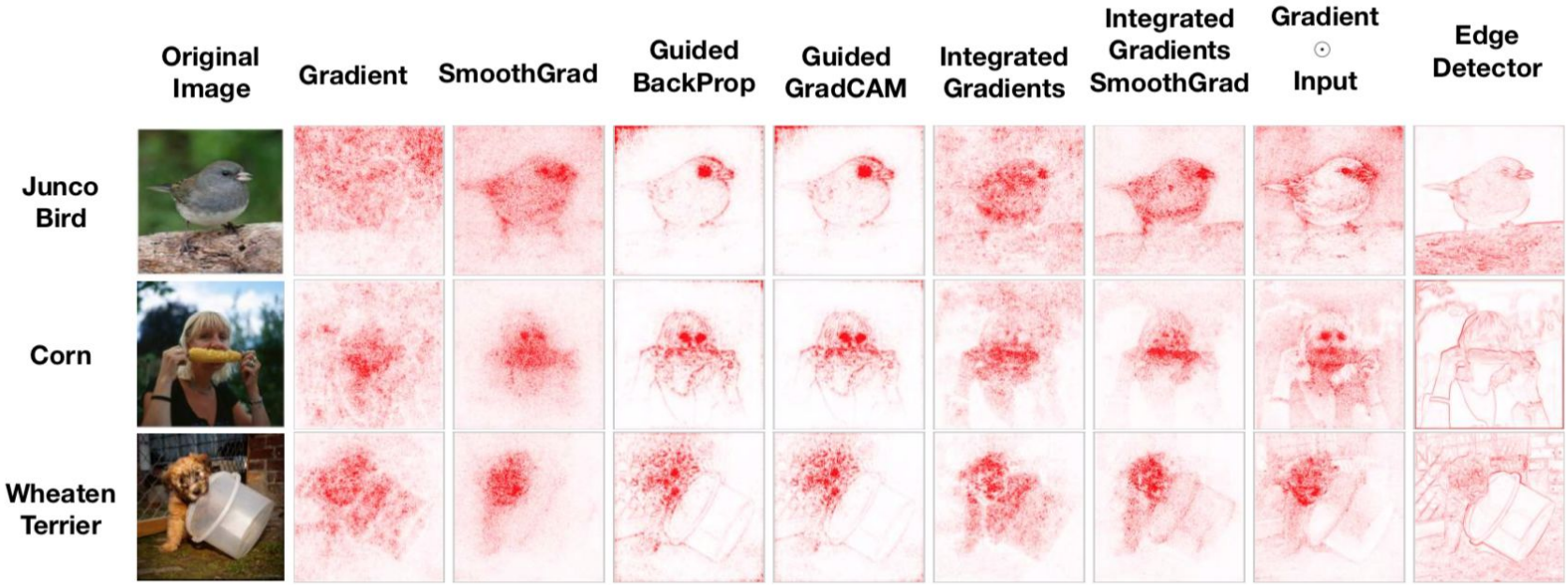
GAN DISSECTION:
VISUALIZING AND UNDERSTANDING GENERATIVE ADVERSARIAL NETWORKS

Романов Игнат

Философия


Original Image
Gradient
SmoothGrad
Guided BackProp
Guided GradCAM
Integrated Gradients
Integrated Gradients SmoothGrad
Gradient ⊙ Input
Edge Detector
Junco Bird
Corn
Wheaten Terrier

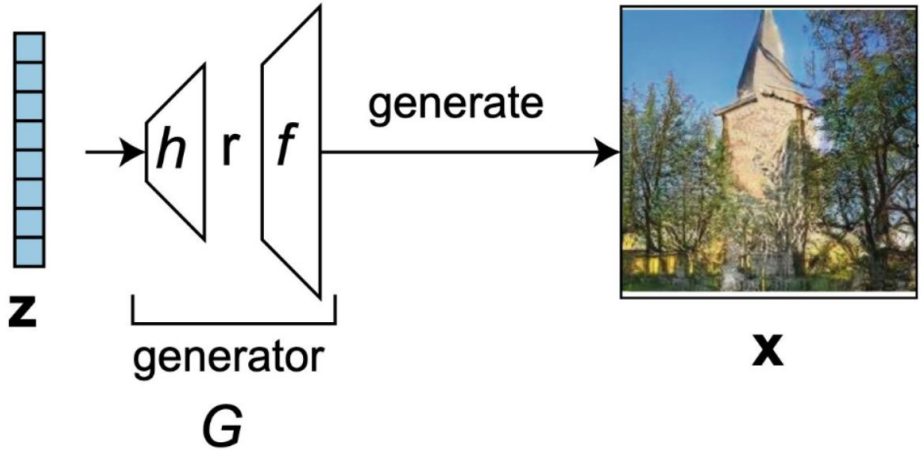
Чуть конкретнее

- Хотим понять что выучивает GAN какие-то пиксельные паттерны или мы можем найти какие-то внутренние переменные, отвечающие за объекты различимые человеком?
- Что порождает артефакты?

(f) Bedroom images with artifacts

(g) Ablating “artifact” units improves results

Определения



r = h(z) and x = f(r) = f(h(z)) = G(z)

Определения

$\mathbb{U}$ - множество всех юнитов (каналов)

$\mathbb{P}$ - множество всех пикселей

U - подмножество юнитов

P - подмножество пикселей

C - множество всех концептов (классов объектов)

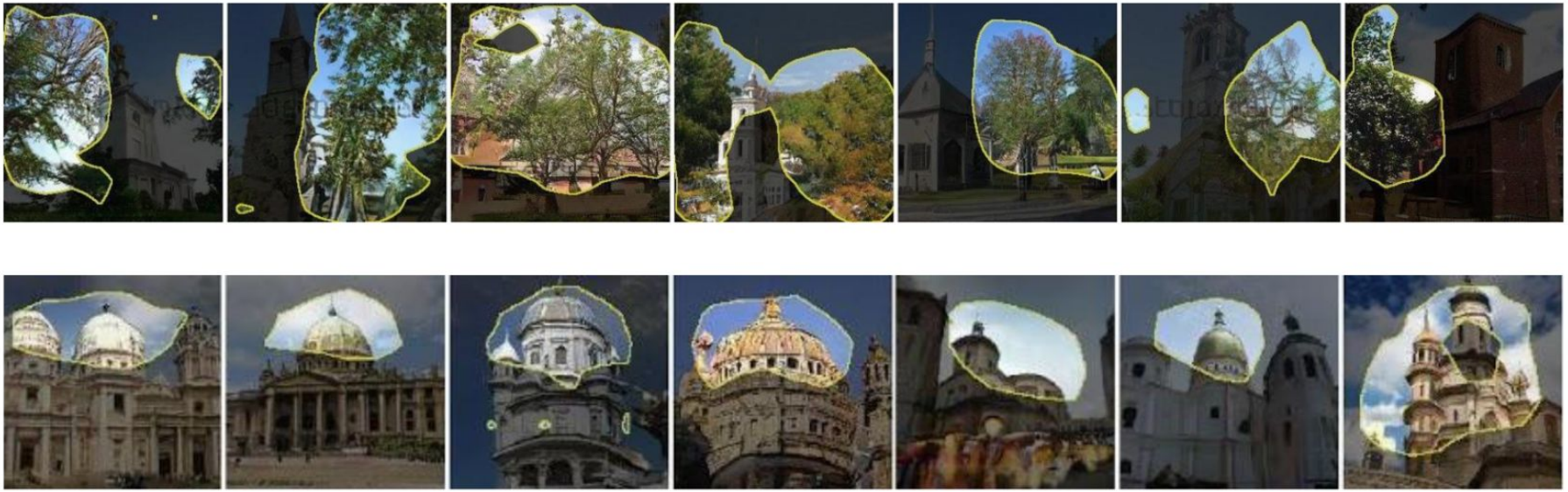
c - конкретный класс

План

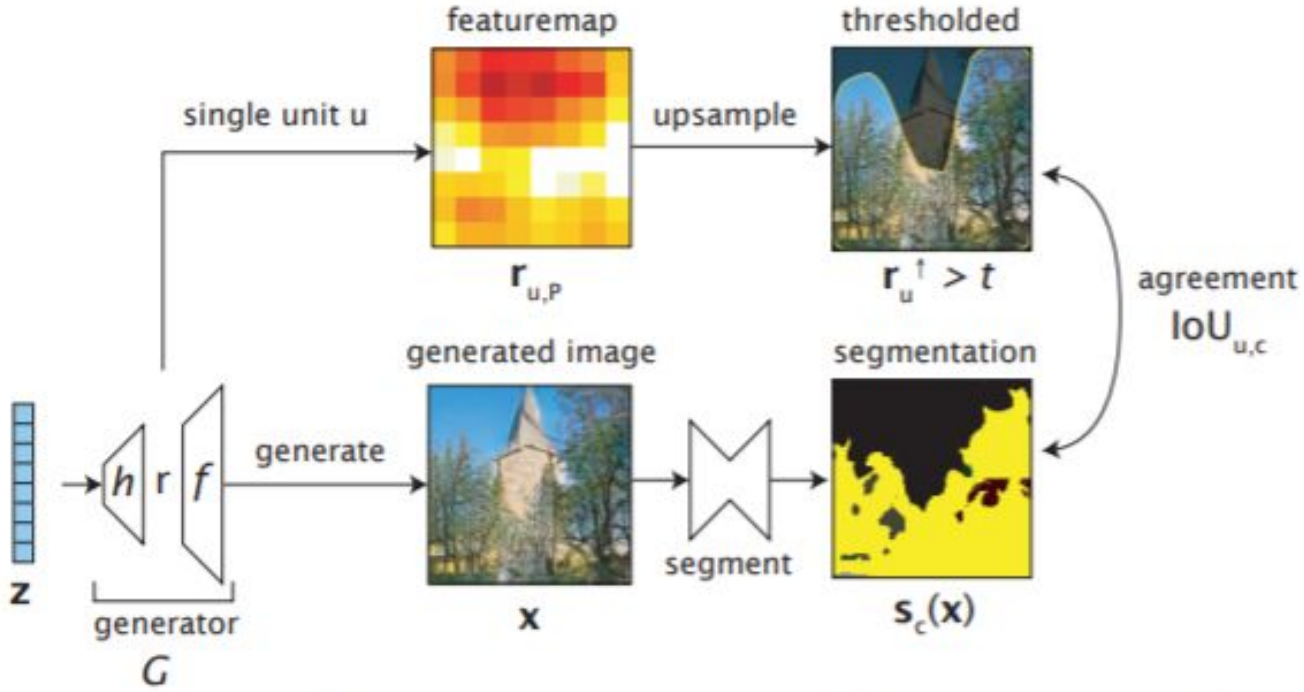
1) Dissection
   Хотим понять представления каких классов вообще присутствуют в гане и найти юниты хорошо скоррелированные с данным классом

2) Intervention
   Выделяем конкретные части изображения влияющие на генерацию объектов

Dissection



Dissection


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
$\mathbf{s}_c(\mathbf{x})$

Back to ПСМО

1) Взаимная информация

$$I(X;Y) = D_{\mathrm{KL}}(P_{(X,Y)} \| P_X \otimes P_Y)$$

2) Совместная энтропия

$$\mathrm{H}(X,Y) = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P(x,y) \log_2 [P(x,y)]$$

Intersection over Union

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

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

Dissection

(e) Identify GAN units that cause artifacts

(f) Bedroom images with artifacts

(g) Ablating “artifact” units improves results

Intervention

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

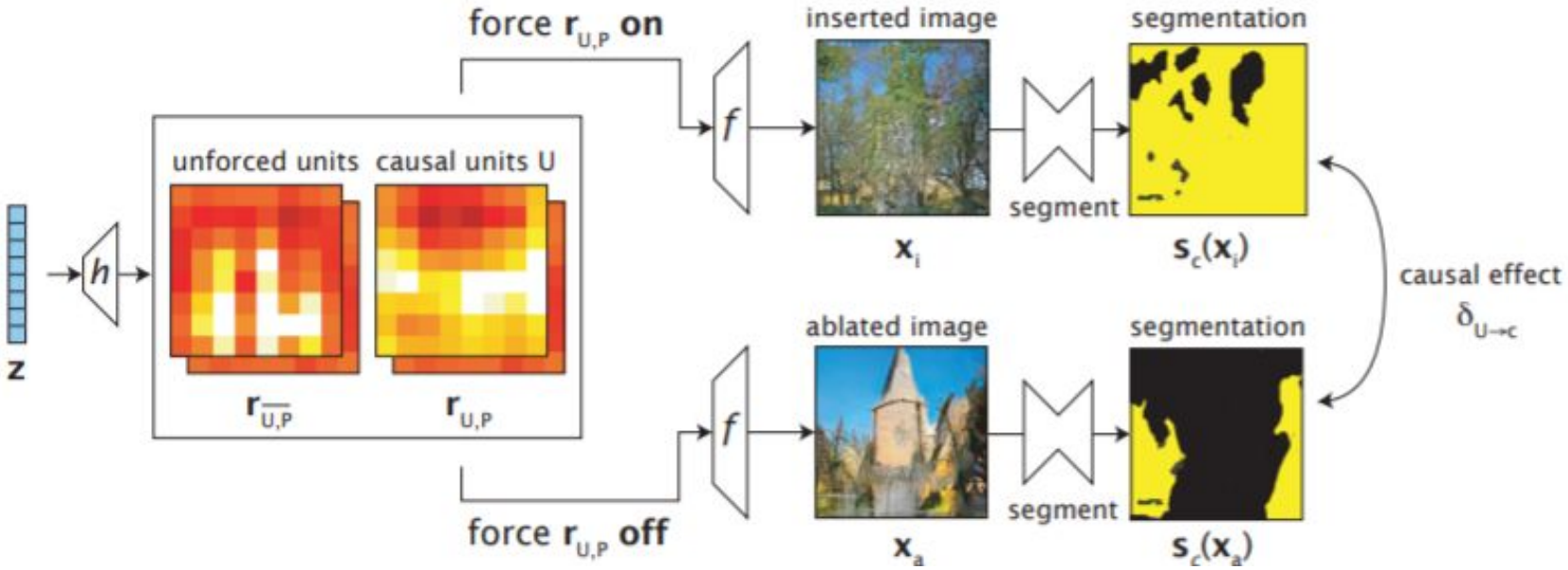
Image with U ablated at pixels P : $\mathbf{x}_a = f(\mathbf{0}, \mathbf{r}_{\overline{\mathrm{U,P}}})$

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

Average Causal Effect

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

Intervention



Intervention

$\alpha \in [0,1]^d$ — continious intervention

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

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

То что осталось за кадром

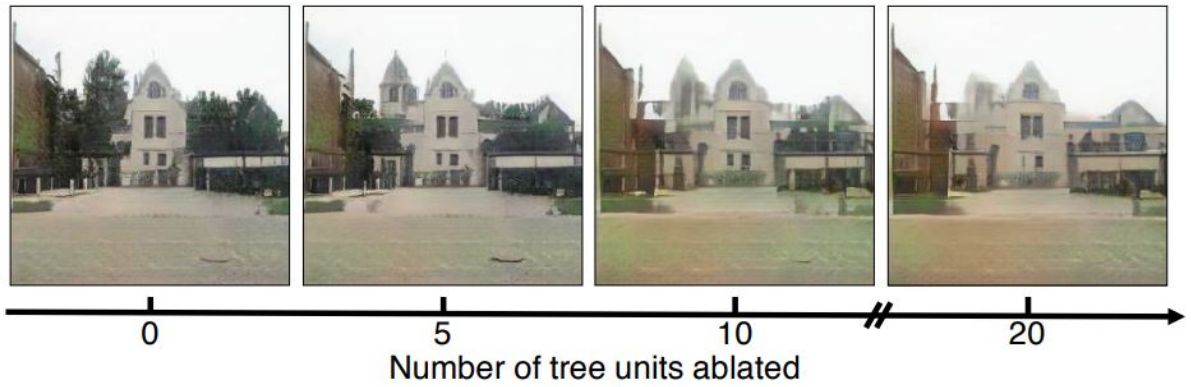
1) Инициализация альфа

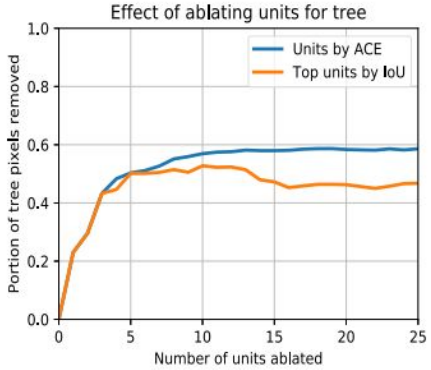
$$\alpha_u = \frac{\mathrm{IoU}_{u,c}}{\max_v \mathrm{IoU}_{v,c}}$$

2) И все-таки P

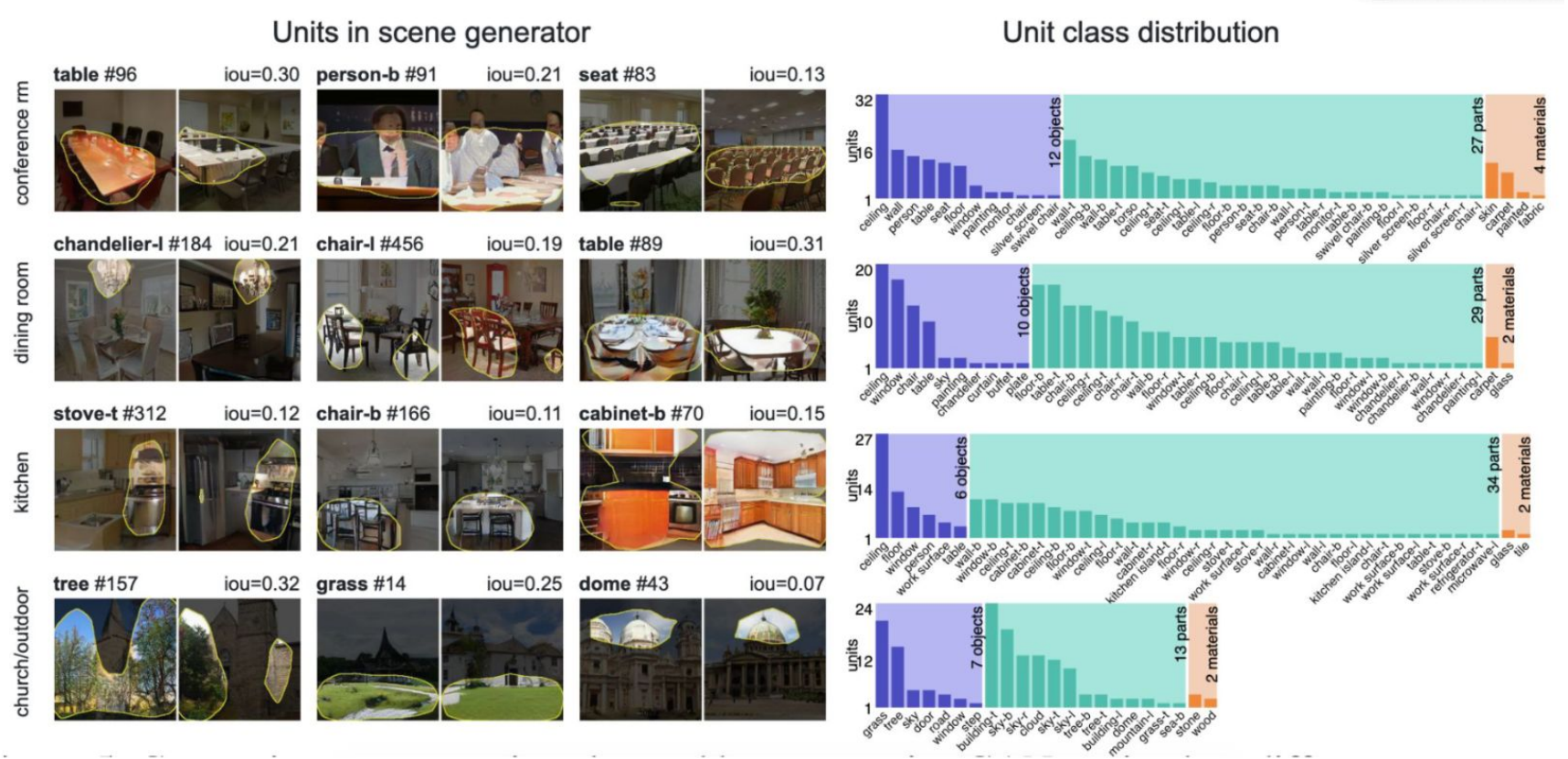
   Будем семплить из масок найденных на первом шаге и сегментаций оригинальных картинок

Intervention

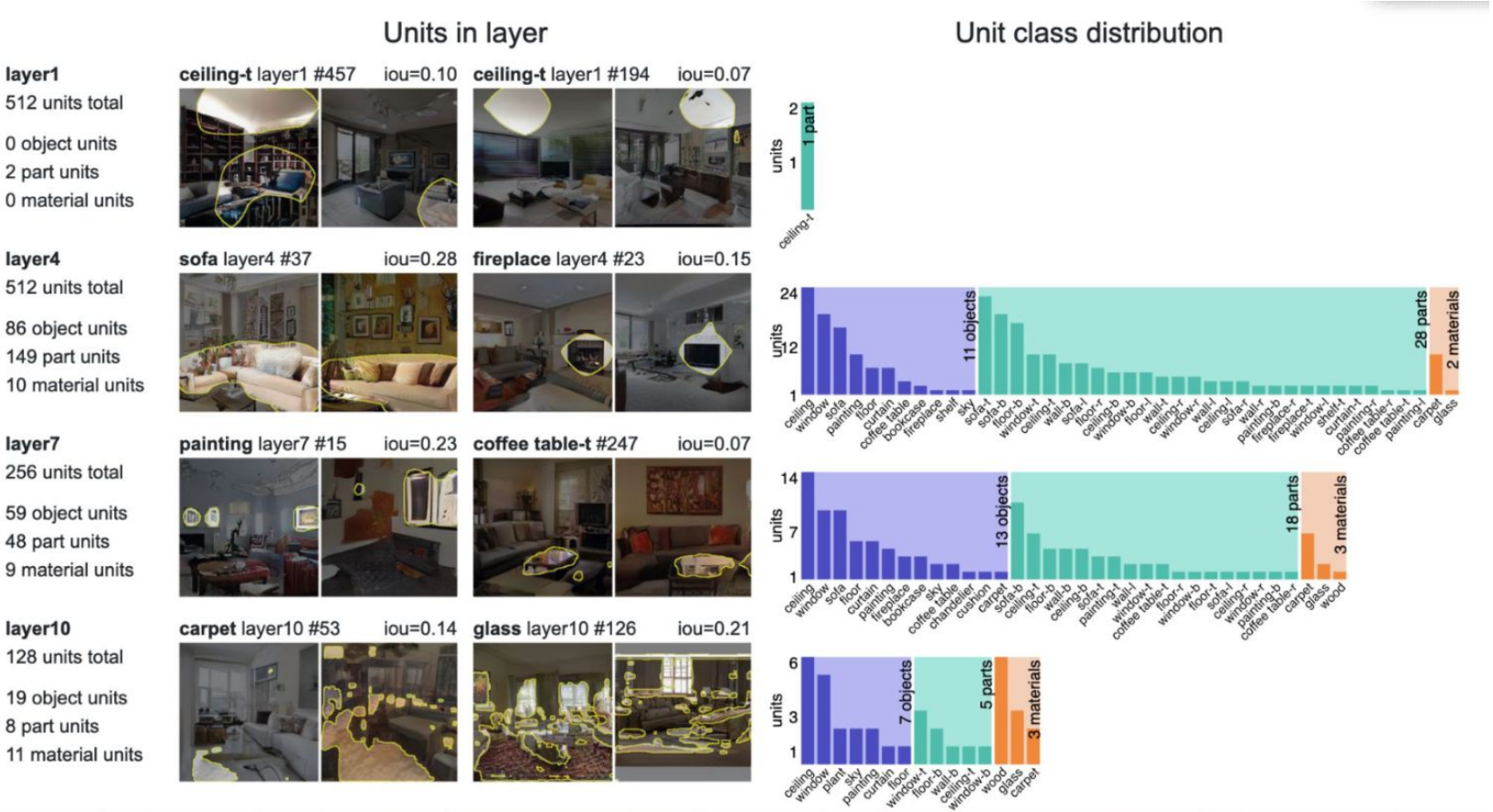




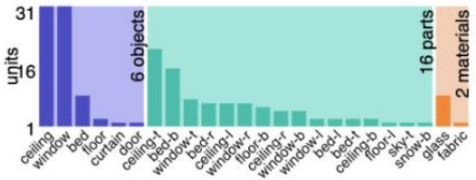
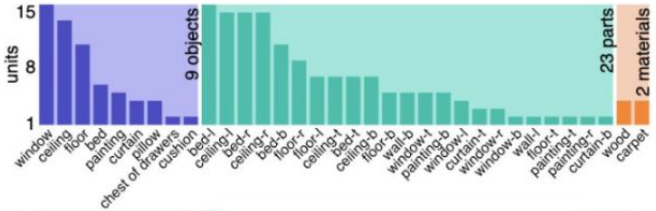
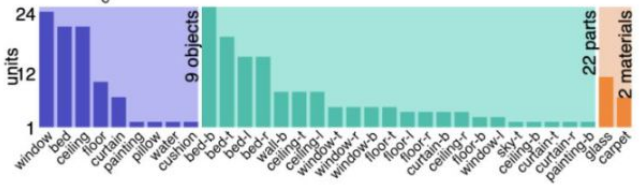
Интерпретируемые юниты для разных данных



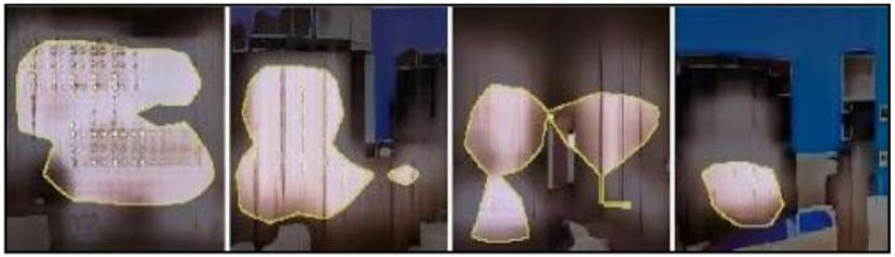
Юниты для разных слоев



Юниты для различных вариантов обучения

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	

Удаление артефактов



Example artifact-causing units

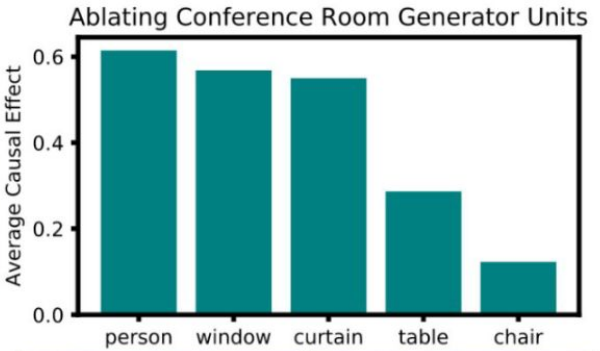
Bedroom images with artifacts

Ablating "artifact" units improves results

Удаление объектов

ablate person units

ablate curtain units

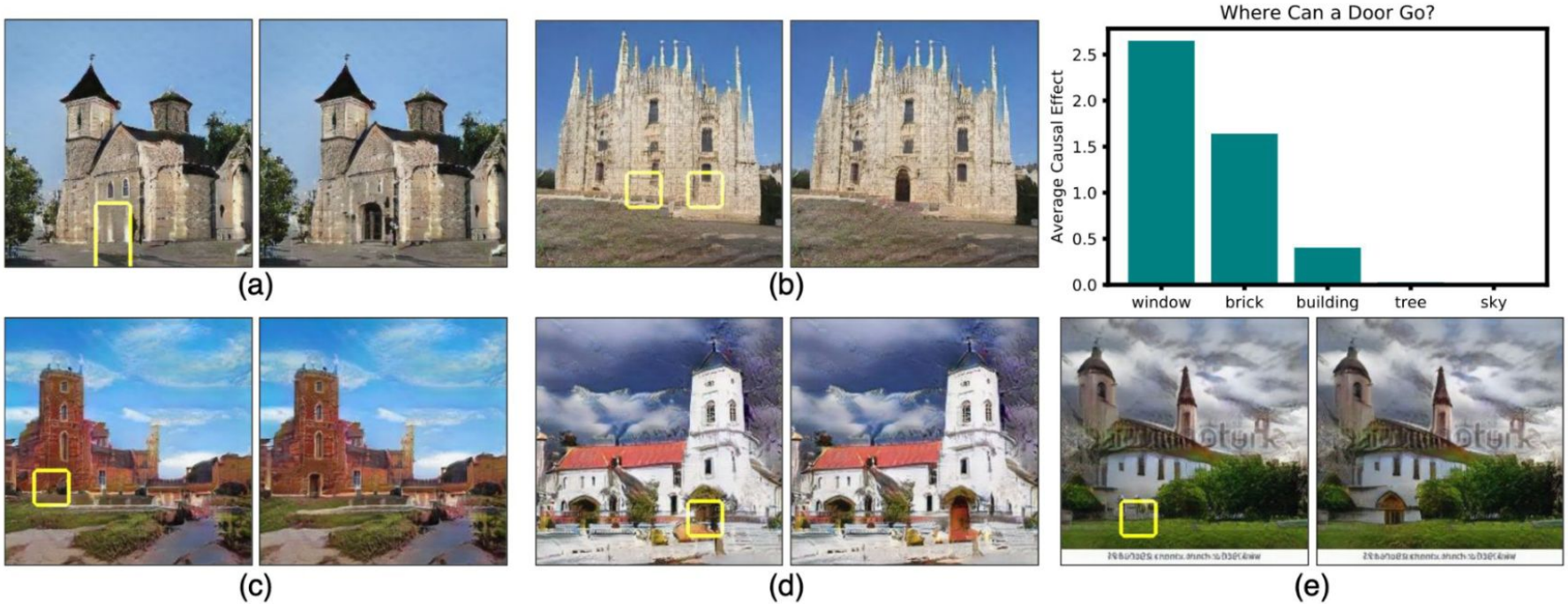


ablate window units

ablate table units

ablate chair units

Вставка объектов



Итоги кратко

- GAN - ы содержат подмножества нейронов, отвечающие за генерацию понятных человеку классов
- GAN - ы могут учить композицию
- Артефакты также могут генерироваться выделенным подмножеством нейронов