Texture vs shape in computer vision networks

Важность

- Устойчивость к разным искажениям изображений
- Важно понимать разницу между человеческим восприятием и восприятием нейросети

(a) Texture image

81.4%	**Indian elephant**
10.3%	indri
8.2%	black swan

(b) Content image

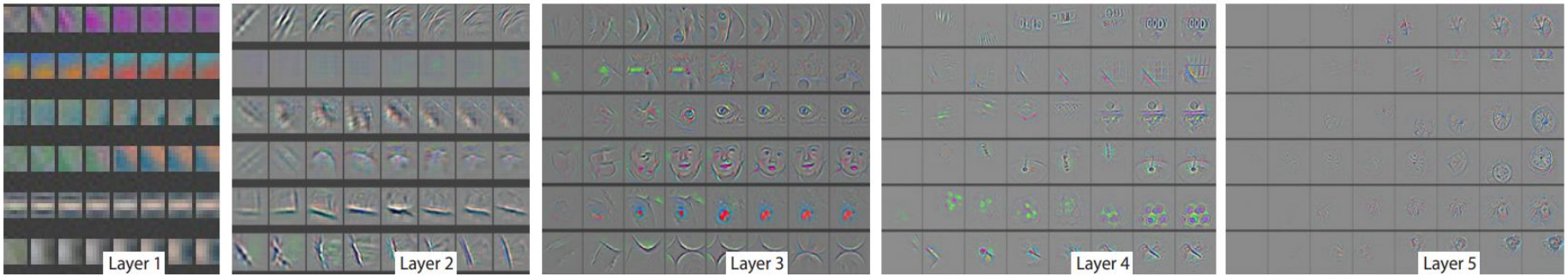
71.1%	**tabby cat**
17.3%	grey fox
3.3%	Siamese cat

(c) Texture-shape cue conflict

63.9%	**Indian elephant**
26.4%	indri
9.6%	black swan

Shape hypothesis

- Это кажется нам более естественным
- Были попытки визуализации:

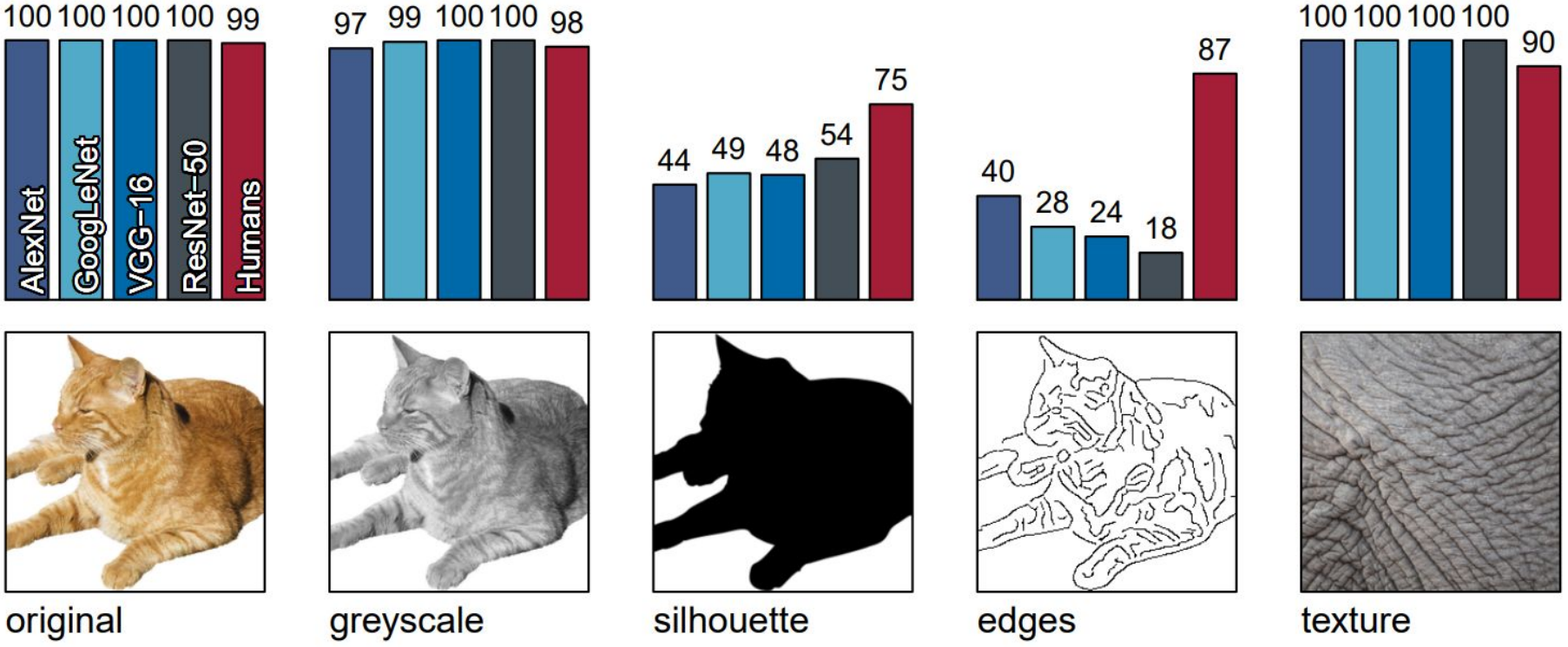


Texture hypothesis

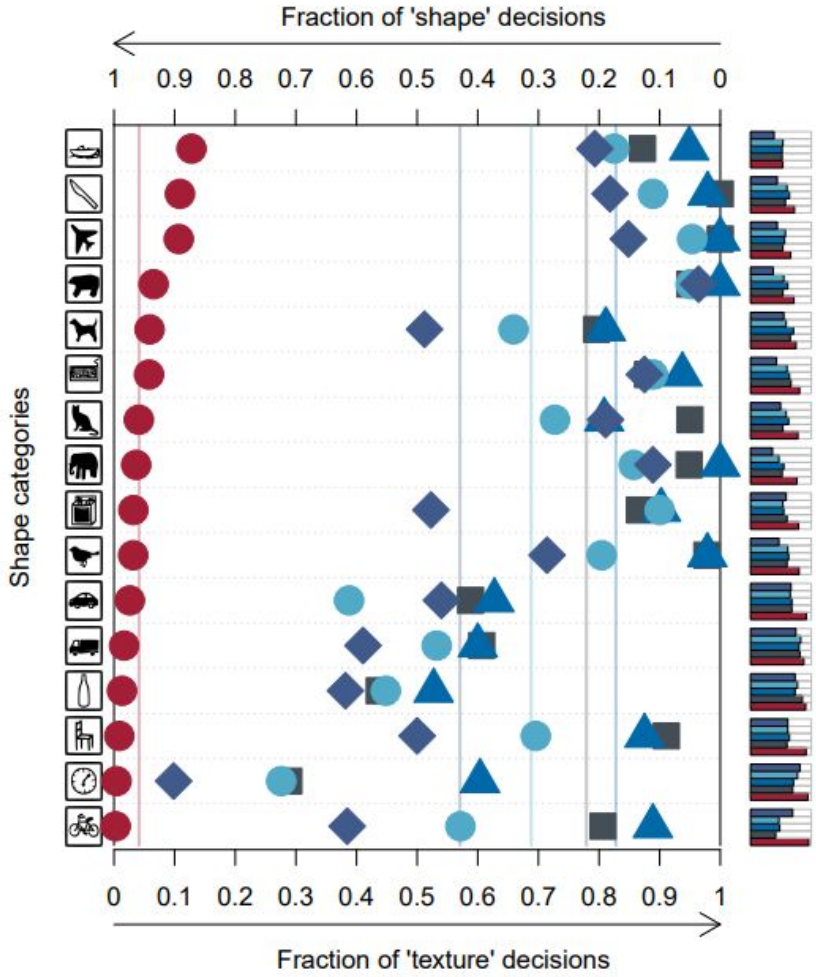
- У CNN получается классифицировать объект даже при разрушении его формы
- CNN плохо классифицируют скетчи (есть форма, нет текстуры)
- Модели с ограниченным receptive field демонстрируют хорошие результаты

Эксперимент

- Original - подмножество из 16 классов ImageNet (10 для каждого класса)
- Texture - датасет из текстур объектов (3 для каждого класса)
- В эксперименте участвуют 4 модели: AlexNet, GoogleLeNet, VGG-16, ResNet-50

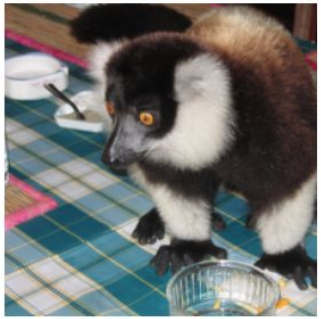

100 100 100 100 99
AlexNet
GoogLeNet
VGG-16
ResNet-50
Humans
97 99 100 100 98
44 49 48 54 75
40 28 24 18 87
100 100 100 100 90
original
greyscale
silhouette
edges
texture

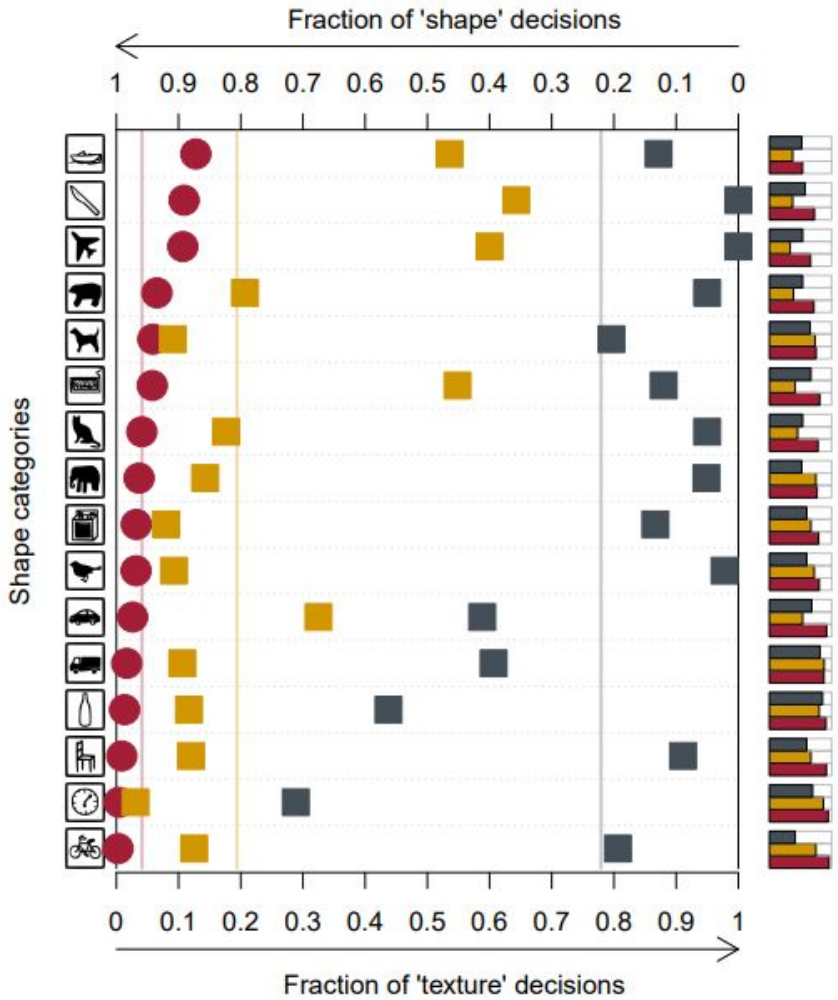

cue conflict (style transfer)
cue conflict (filled silhouettes)
original texture images
original content images



- Участники эксперимента
- GoogLeNet
- AlexNet
- VGG-16
- ResNet-50

Stylized-ImageNet (SIN)





- Участники эксперимента

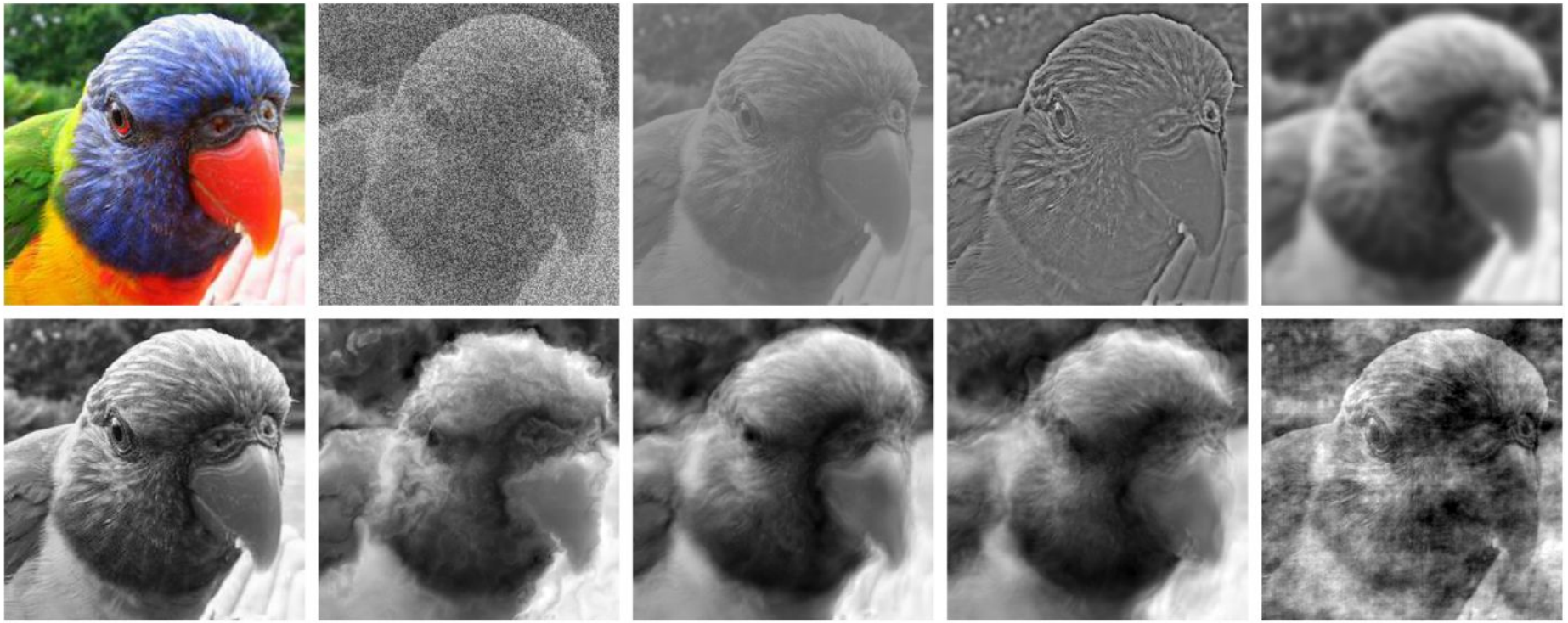
- ResNet-50 Stylized-ImageNet

- ResNet-50 ImageNet

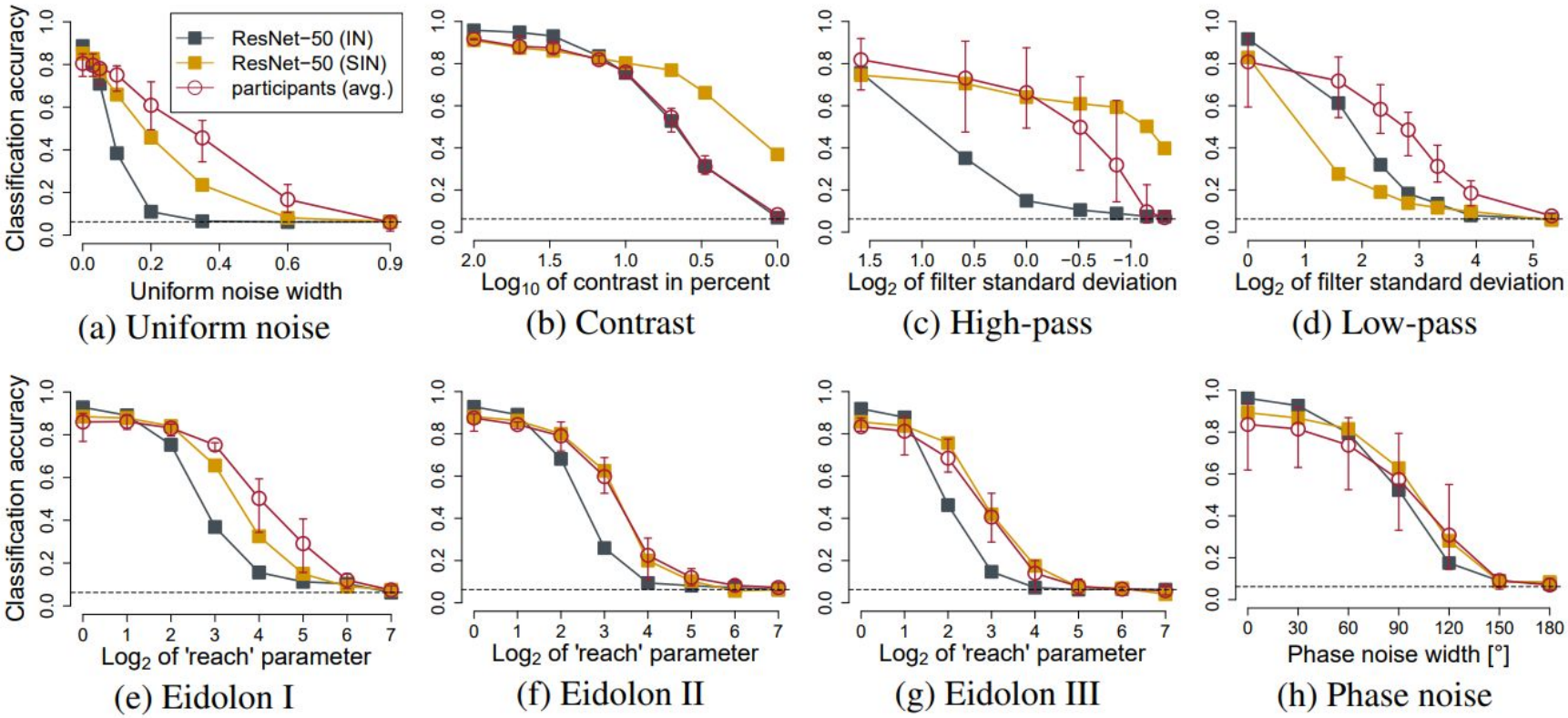
Stylized-ImageNet (SIN)

architecture	IN→IN	IN→SIN	SIN→SIN	SIN→IN
ResNet-50	92.9	16.4	79.0	82.6
BagNet-33 (mod. ResNet-50)	86.4	4.2	48.9	53.0
BagNet-17 (mod. ResNet-50)	80.3	2.5	29.3	32.6
BagNet-9 (mod. ResNet-50)	70.0	1.4	10.0	10.9

Stylized-ImageNet (SIN)



Stylized-ImageNet (SIN)



Причины

Наблюдения

- Модель **может** обучаться со смещением в сторону формы объекта благодаря изменению входных данных
- Характер входных данных определяет на что смотрит модель
- Можно рассмотреть влияние различных аугментаций данных на примере датасета из cue conflict изображений

(a) Original

(b) Crop, resize (and flip)

(c) Color distort. (drop)

(d) Color distort. (jitter)

(e) Rotate $\{90°, 180°, 270°\}$

(f) Cutout

(g) Gaussian noise

(h) Gaussian blur

(i) Sobel filtering

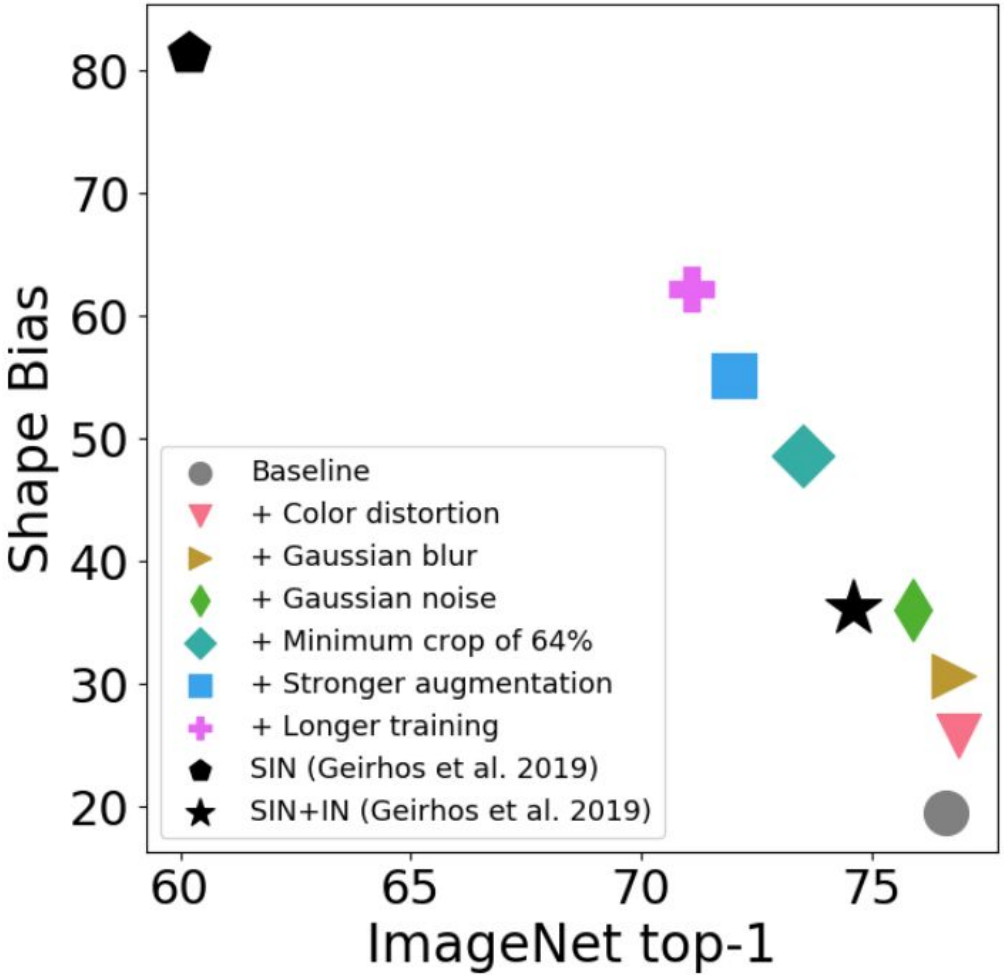
Аугментации данных: random-crop

Model	Shape Bias		Shape Match		Texture Match		ImageNet Top-1 Acc.	
	Random	Center	Random	Center	Random	Center	Random	Center
AlexNet	28.2%	**37.5%**	16.4%	**19.3%**	**41.8%**	32.1%	**56.4%**	50.7%
VGG16	11.2%	**15.8%**	7.6%	**10.7%**	**60.1%**	57.1%	**71.8%**	62.5%
ResNet-50	19.5%	**28.4%**	11.7%	**16.3%**	**48.4%**	41.1%	**76.6%**	70.7%
Inception-ResNet v2	23.1%	**27.9%**	15.1%	**19.8%**	50.2%	**51.2%**	**80.3%**	77.3%

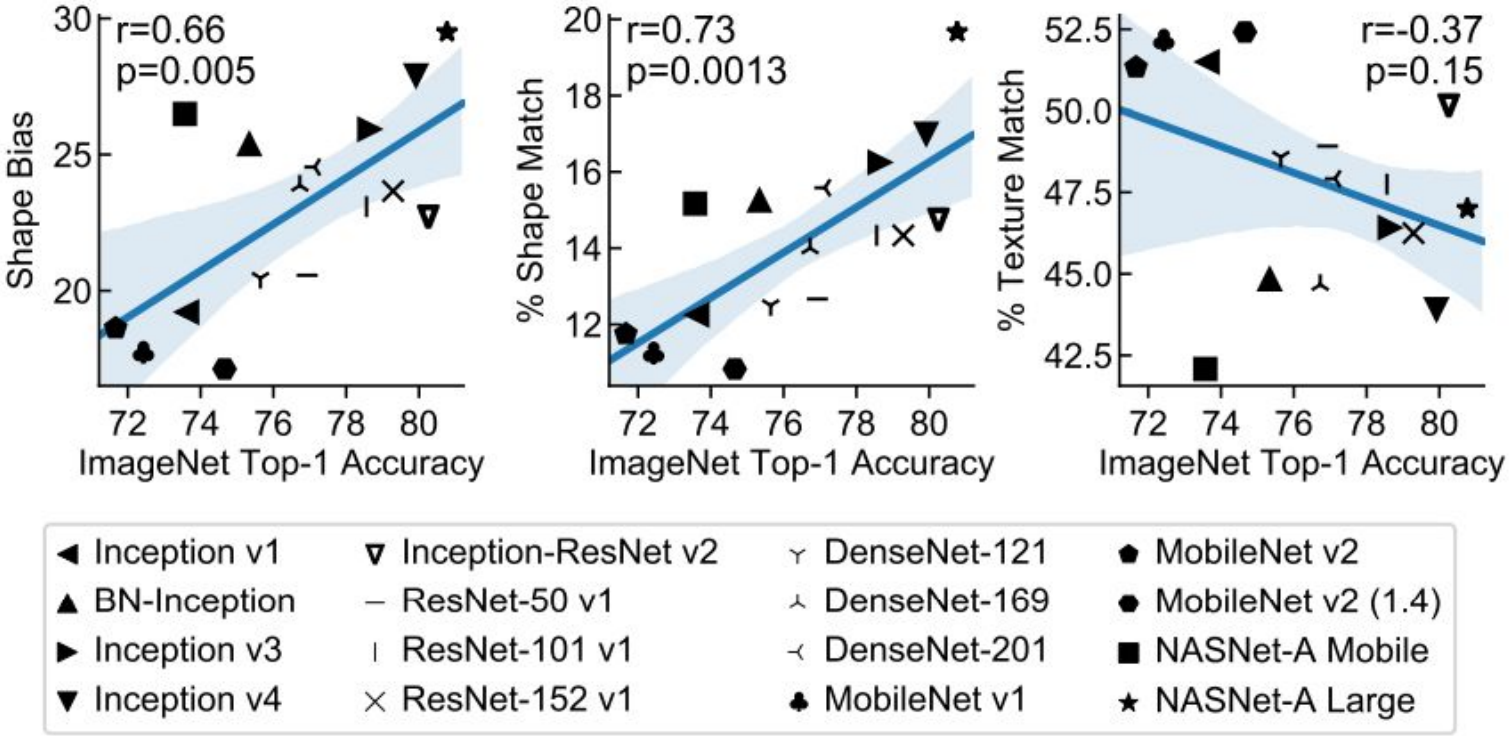
Аугментации данных

Augmentation	Shape Bias	Shape Match	Texture Match	ImageNet Top-1 Acc.
Baseline	19.5%	11.7%	48.4%	76.6%
Rotate 90°, 180°, 270°	19.4%	10.8%	45.1%	75.7%
Cutout	**21.4%**	12.3%	45.2%	76.9%
Sobel filtering	**24.8%**	12.8%	38.9%	71.2%
Gaussian blur	**25.2%**	14.1%	41.7%	75.8%
Color distort.	**25.8%**	15.3%	44.2%	76.9%
Gaussian noise	**30.7%**	17.2%	38.8%	75.6%

Augmentation(s)	Shape Bias	Shape Match	Texture Match	ImageNet		IN-Sketch		SIN	
				top-1	top-5	top-1	top-5	top-1	top-5
Baseline	19.5%	11.7%	**48.4%**	76.6%	**93.3%**	22.4%	39.3%	7.7%	17.0%
+ Color distortion	25.8%	15.3%	44.2%	**76.9%**	**93.3%**	28.1%	46.6%	9.9%	20.5%
+ Gaussian blur	30.7%	17.2%	38.8%	76.8%	**93.3%**	29.0%	47.9%	11.1%	21.9%
+ Gaussian noise	36.1%	20.1%	35.5%	75.9%	92.8%	29.8%	48.9%	12.6%	24.3%
+ Min. crop of 64%	48.7%	29.1%	30.7%	73.5%	91.5%	**30.9%**	**51.4%**	14.5%	28.2%
+ Stronger aug.	55.2%	33.3%	27.1%	72.0%	90.7%	30.4%	50.5%	**15.1%**	**28.8%**
+ Longer training	**62.2%**	**38.3%**	23.3%	71.1%	90.0%	30.5%	50.4%	14.9%	28.4%


Shape Bias
80
70
60
50
40
30
20
60
65
70
75
ImageNet top-1
Baseline
+ Color distortion
+ Gaussian blur
+ Gaussian noise
+ Minimum crop of 64%
+ Stronger augmentation
+ Longer training
SIN (Geirhos et al. 2019)
SIN+IN (Geirhos et al. 2019)

Влияние архитектуры



Влияние гиперпараметров

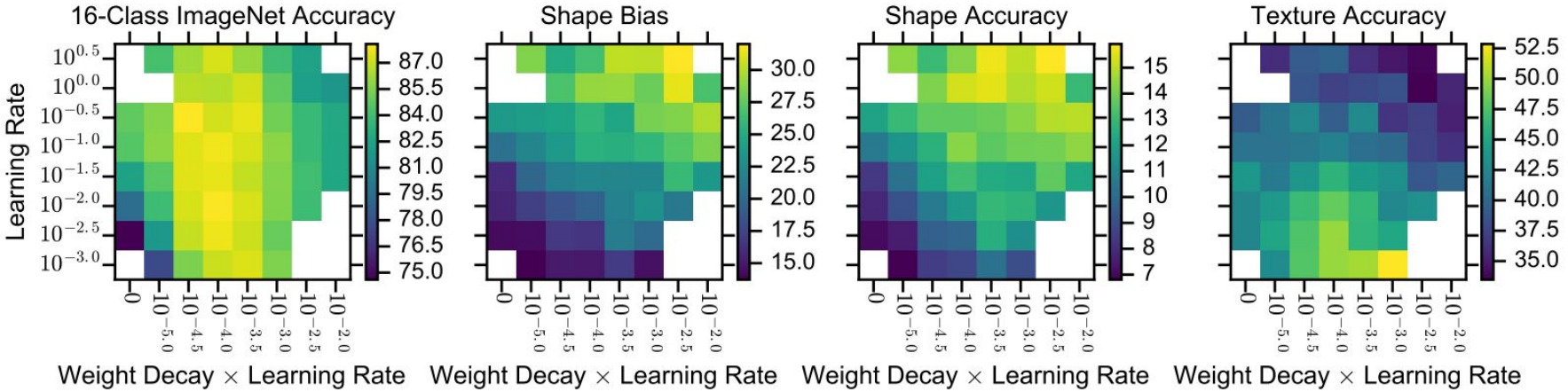


Figure A.1: **Higher learning rates produce greater shape bias.** Plots show mean of 3 runs on 16-class ImageNet. Results for hyperparameter combinations achieving <70% accuracy are masked. We plot weight decay × learning rate because it is more closely related to accuracy than weight decay [56, 14].

Итог

- CNN хотят смотреть на текстуру, а не на форму
- Модель можно “заставить” обучаться по форме объектов
- Смещение в пользу формы/текстуры обосновано входными данными
- Аугментация входных данных позволяет влиять на выбор между формой и текстурой
- Модели со смещением в пользу формы более устойчивы к шумам и другим подобным дефектам

ViT

- human-level perfomance при обучении на Stylized-ImageNet
- Более устойчивы к повреждению входных данных
- В большей степени (по сравнению с CNN) учитывают форму объекта

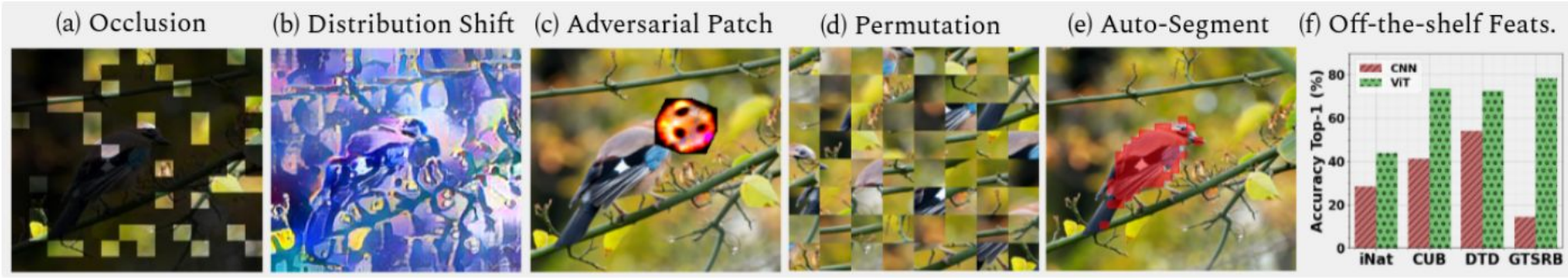
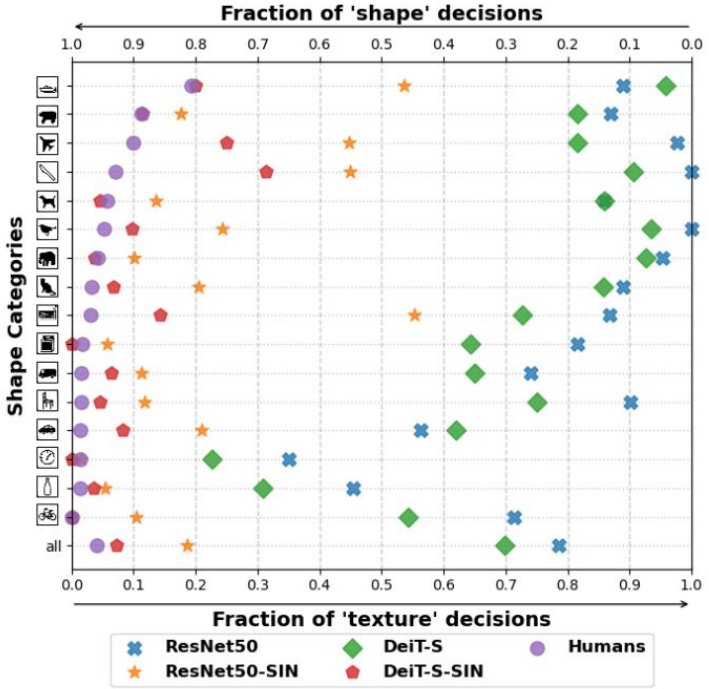


Figure 1: We show intriguing properties of ViT including impressive robustness to (a) severe occlusions, (b) distributional shifts (*e.g.*, stylization to remove texture cues), (c) adversarial perturbations, and (d) patch permutations. Furthermore, our ViT models trained to focus on shape cues can segment foregrounds without any pixel-level supervision (e). Finally, off-the-shelf features from ViT models generalize better than CNNs (f).

ViT



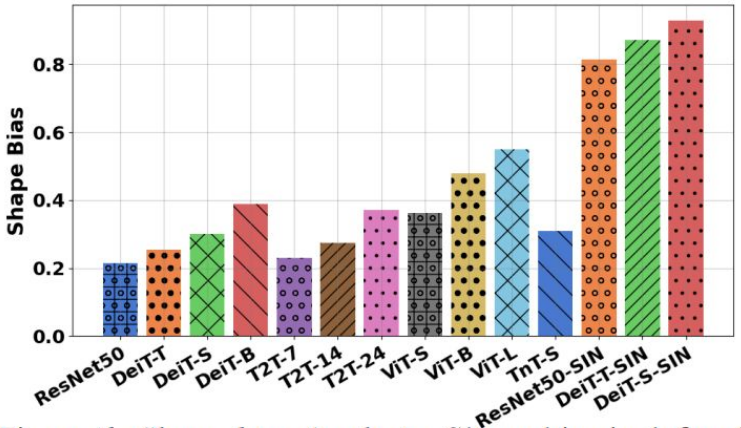


Figure 6: *Shape-bias Analysis:* Shape-bias is defined as the fraction of correct decisions based on object shape. *(Left)* Plot shows shape-texture tradeoff for CNN, ViT and Humans across different object classes. *(Right)* class-mean shape-bias comparison. Overall, ViTs perform better than CNN. The shape bias increases significantly when trained on stylized ImageNet (SIN).

ViT 22-B

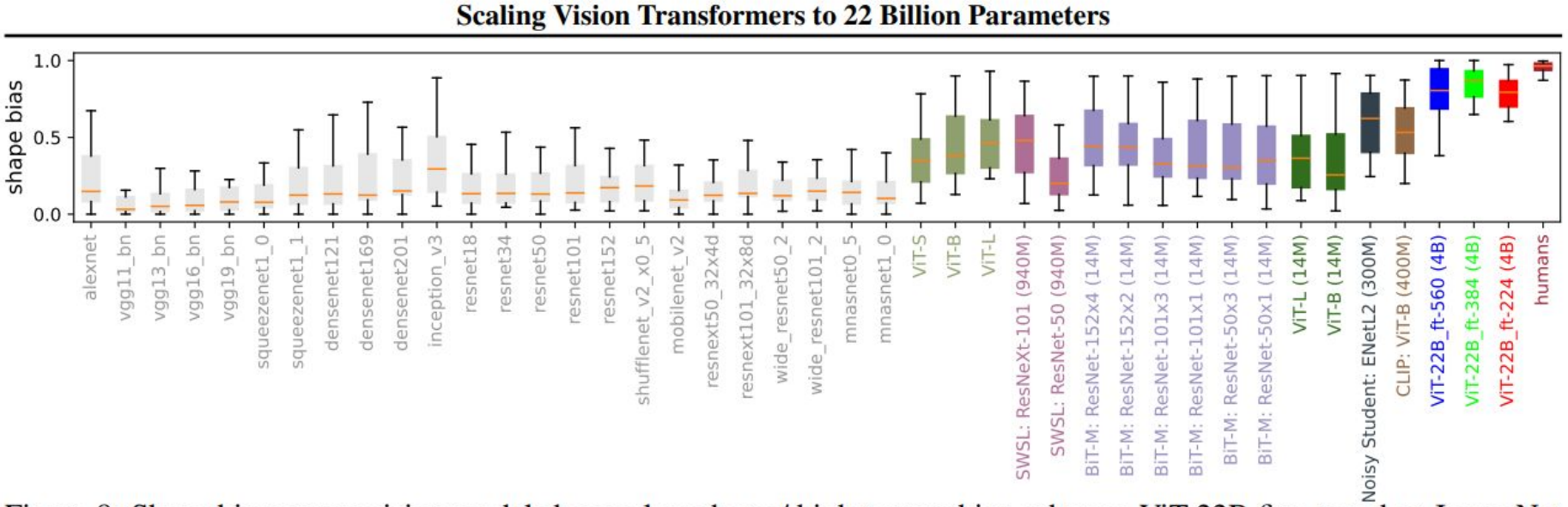


Figure 8: Shape bias: many vision models have a low shape / high texture bias, whereas ViT-22B fine-tuned on ImageNet (red, green, blue trained on 4B images as indicated by brackets after model names, unless trained on ImageNet only) have the highest shape bias recorded in a ML model to date, bringing them closer towards a human-like shape bias.