CNN Shape vs Texture

Присяжнюк Артем

Shape hypothesis: Интуиция

- Слой сопоставляет некоторой категории определенные формы
- Первые слои находят простые формы
- Средние слои находят сложные формы, состоящие из простых
- Последние слои находят формы объектов

Texture hypothesis: Интуиция

Текстура влияет на решение слоя больше, чем форма

(a) Texture image

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

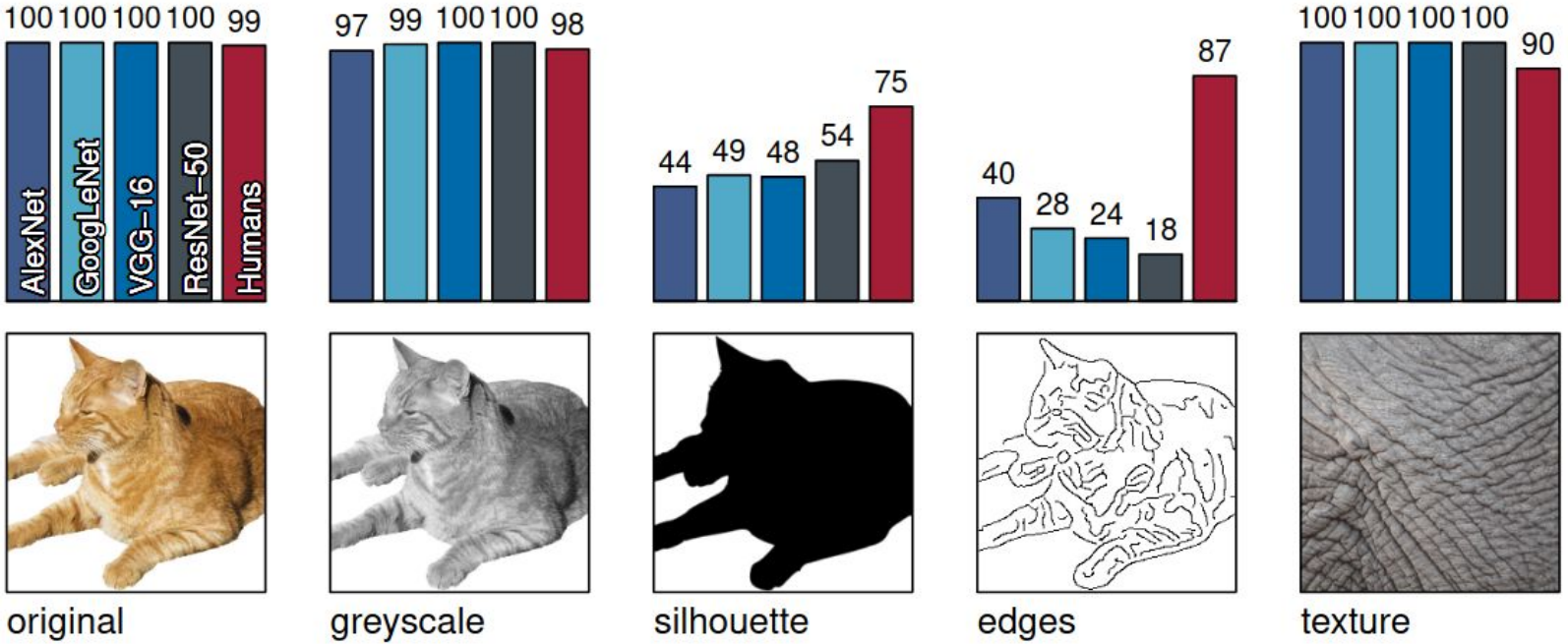
(b) Content image

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

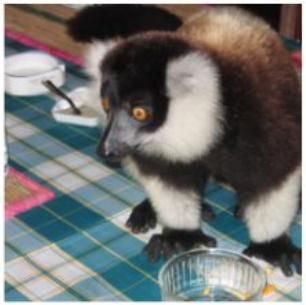
(c) Texture-shape cue conflict

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

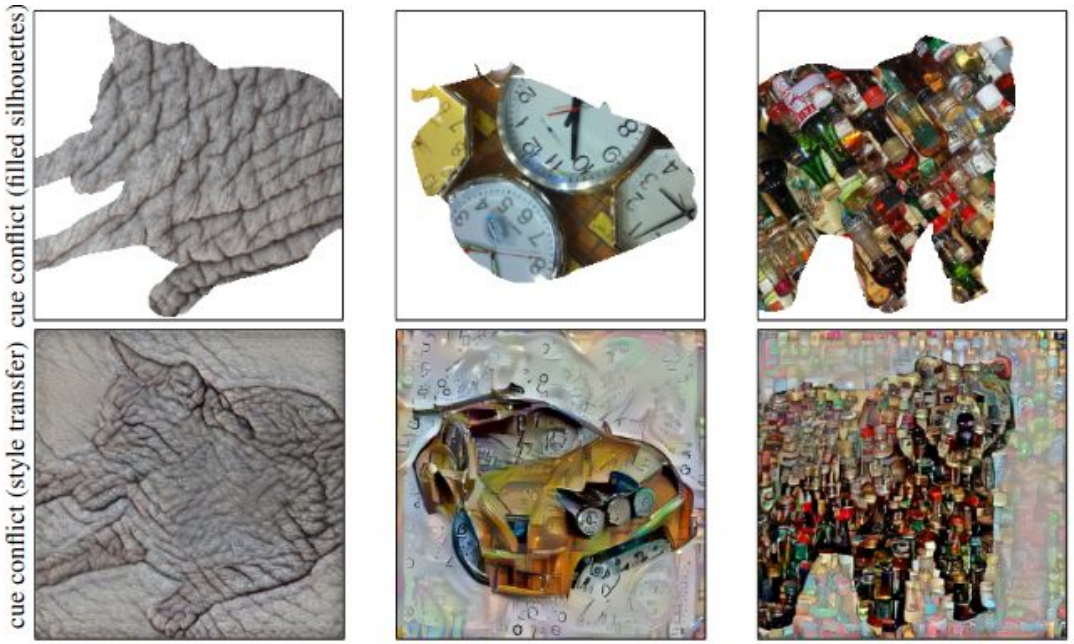
Эксперимент: Датасет



Эксперимент: Style Transfer



Эксперимент: Stylized-ImageNet



Эксперимент: Результаты

Красные - люди

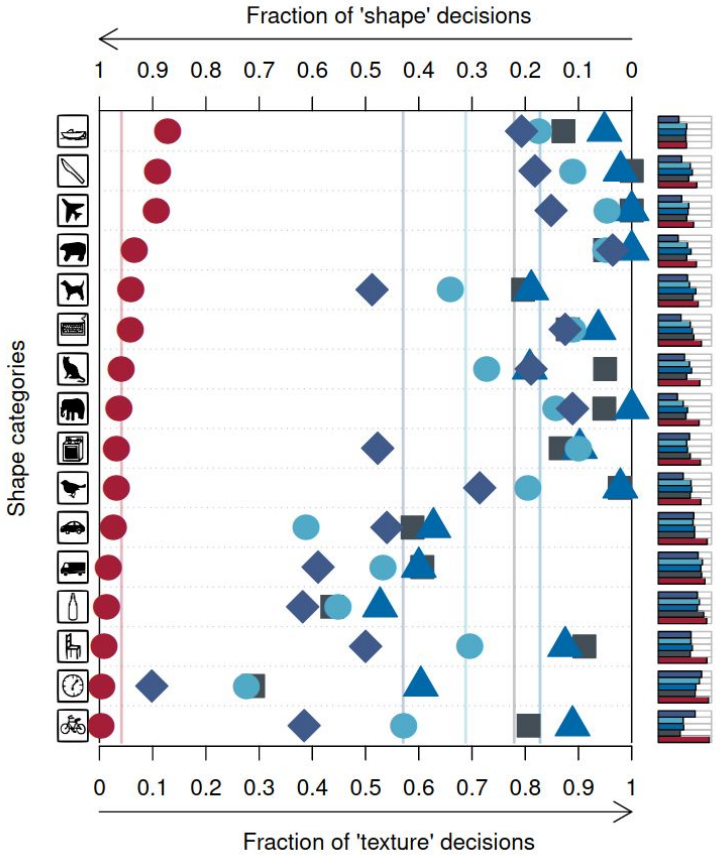
Фиолетовые - AlexNet

Синие - VGG-16

Бирюзовые - GoogLeNet

Серые - ResNet-50

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



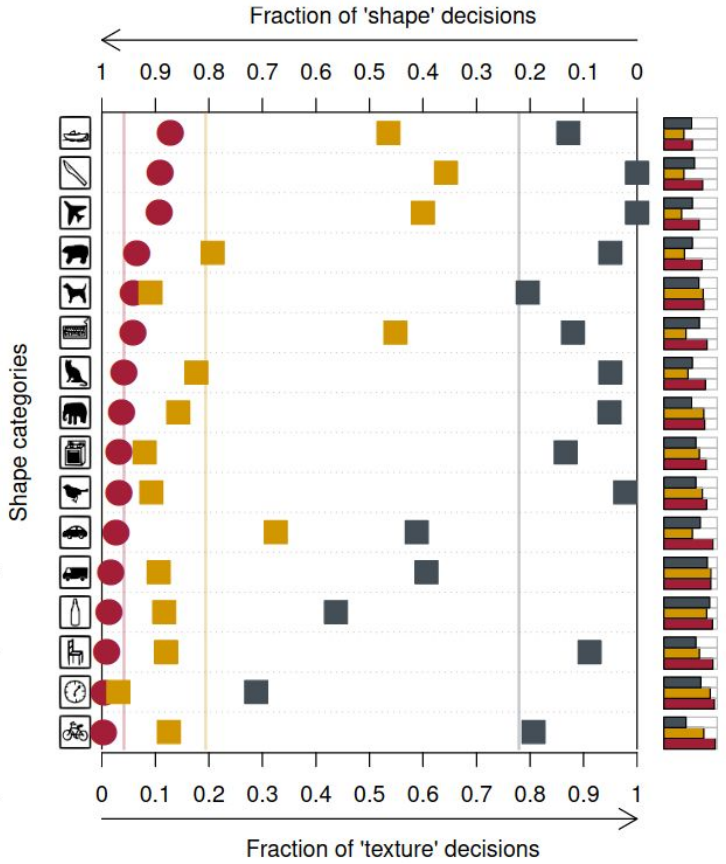
Изменения: Результаты

Красные - люди

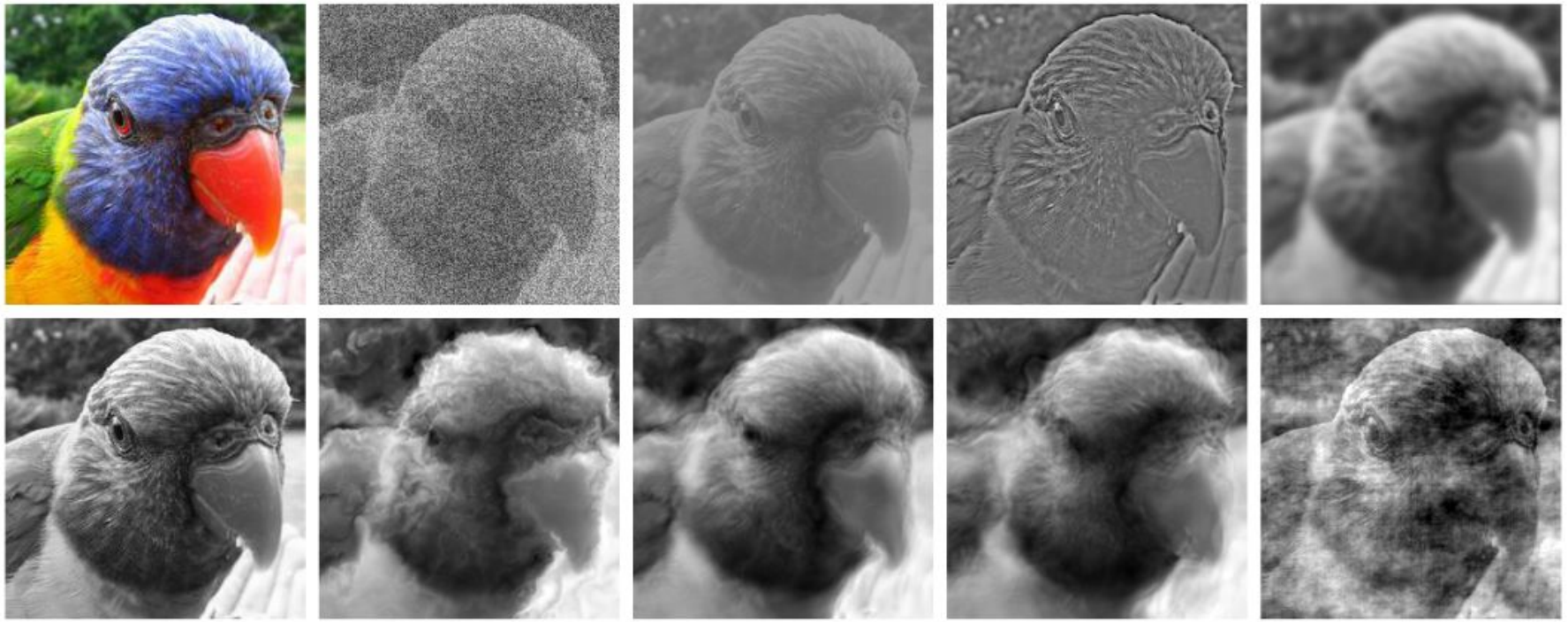
Золотые - Stylized-ImageNet

Серые - ImageNet

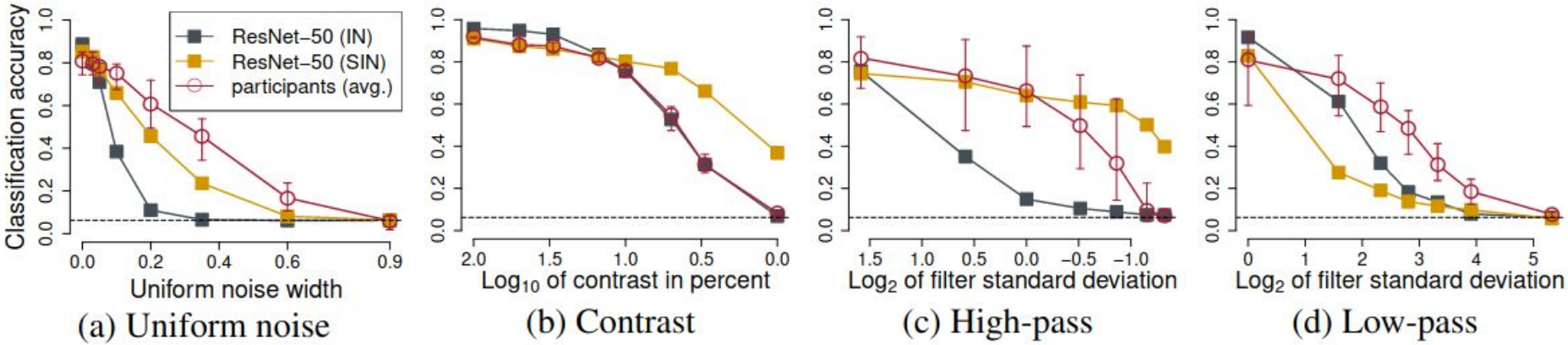
name	training	fine-tuning	top-1 IN accuracy (%)	top-5 IN accuracy (%)	Pascal VOC mAP50 (%)
vanilla ResNet	IN	-	76.13	92.86	70.7
	SIN	-	60.18	82.62	70.6
	SIN+IN	-	74.59	92.14	74.0
Shape-ResNet	SIN+IN	IN	**76.72**	**93.28**	**75.1**



Устойчивость к искажениям



Устойчивость к искажениям



Итоги

- Обучение на ImageNet → texture bias
- Решение - Stylized-ImageNet (SIN)
- Обучение на SIN → устойчивость к искажениям
- Лучший вариант - Shape-ResNet
  - Обучение на SIN + IN
  - Fine-tuning на IN

Сравнение моделей

Convolutional Neural Networks			
Model	top-1(%)	mCE(%)	#params(M)
ResNet-50	76.02	65.54	26
AlexNet	56.44	83.18	61
GoogLeNet	71.70	68.82	7
VGG-16	69.63	75.10	138

Convolutional Neural Networks		
Model	Shape Bias (%)	# params (M)
ResNet-50	26.17	26
AlexNet	29.80	61
GoogLeNet	28.52	7
VGG-16	16.12	138

Vision Transformer Architectures			
Model	top-1(%)	mCE(%)	#params(M)
ViT_base	75.73	58.55	86
ViT_large	79.16	49.02	304
DeiT_base	81.84	42.30	86
DeiT_base-dist.	83.16	41.19	87
DeiT_small	79.68	47.79	22
DeiT_small-dist.	81.05	46.25	22
DeiT_tiny	71.92	**60.08**	**5**
DeiT_tiny-dist.	74.38	57.45	6
CaiT_s24	83.28	40.59	47
CaiT_xxs24	78.38	49.28	11
Swin-T_tiny	80.85	50.70	28
Swin-T_small	82.96	45.51	50
Swin-T_base	84.90	38.52	88
Swin-T_large	85.92	**34.63**	**197**

Vision Transformer Architectures		
Model	Shape Bias (%)	# params (M)
ViT_base	49.10	86
ViT_large	**55.35**	**304**
DeiT_base	42.32	86
DeiT_base-dist.	39.62	87
DeiT_small	38.26	22
DeiT_small-dist.	36.65	22
DeiT_tiny	**29.37**	**5**
DeiT_tiny-dist.	31.06	6
CaiT_s24	38.65	47
CaiT_xxs24	34.24	11
Swin-T_tiny	25.21	28
Swin-T_small	27.43	50
Swin-T_base	36.39	88
Swin-T_large	40.20	197