IMAGENET-TRAINED CNNS ARE BIASED TOWARDS TEXTURE; INCREASING SHAPE BIAS IMPROVES ACCURACY AND ROBUSTNESS

Geirhos et al. (2019)

Introduction

- **ImageNet classification with CNNs**
- **Which image cues are learned**
- **How influential they are**
- **Comparison with humans**

Testing Hypothesis

Shape Hypothesis

"The network acquires **complex knowledge** about the kinds of **shapes** associated with each category. [...] High-level units appear to learn **representations** of **shapes** occurring in **natural images**"

Kriegeskorte (2015)

Intermediate CNN layers recognize "**parts** of familiar objects, and **subsequent** layers [...] detect objects as **combinations** of these parts"

LeCun et al. (2015)

Testing Hypothesis

Shape Hypothesis

“The network acquires **complex knowledge** about the kinds of **shapes** associated with each category. […] High-level units appear to learn **representations** of **shapes** occurring in **natural images**”

Kriegeskorte (2015)

Intermediate CNN layers recognize “**parts** of familiar objects, and **subsequent** layers […] detect objects as **combinations** of these parts”

LeCun et al. (2015)

Texture Hypothesis

CNNs can still classify **texturised** images perfectly well, even if the **global shape structure** is completely **destroyed**

Gatys et al. (2017) and Brendel & Bethge (2019)

Standard CNNs are **bad** at recognizing object **sketches** where object **shapes** are **preserved** yet all **texture** cues are **missing**

Ballester & de Araújo (2016)

Set-up

Psychophysical

- 97 observers
- 48,560 trials
- 300 ms fixation square

  + 200 ms image

  + 200 ms pink noise

  + 1500 ms category selection
- Breaks after every 256 trials
- Practice session of 320 trials

Model experiments

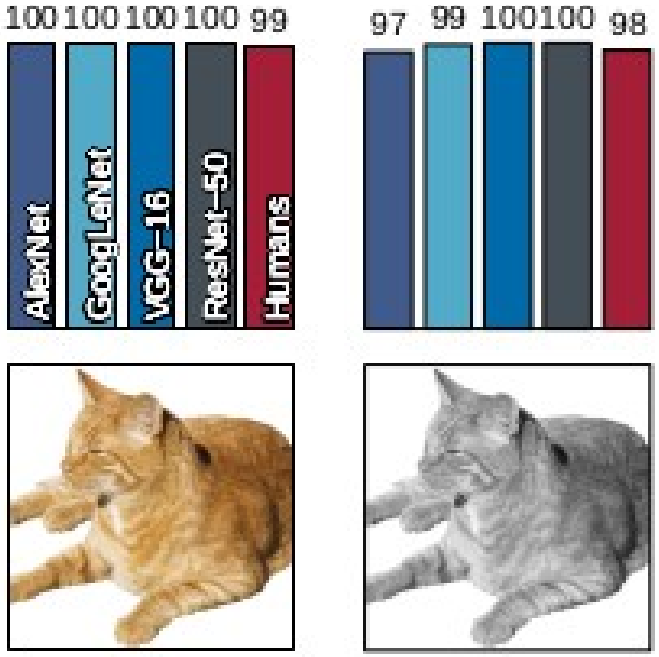
- AlexNet
- GoogLeNet
- VGG-16
- ResNet-50

Experiments

Original

160 color images
10 per category
white background

Experiments



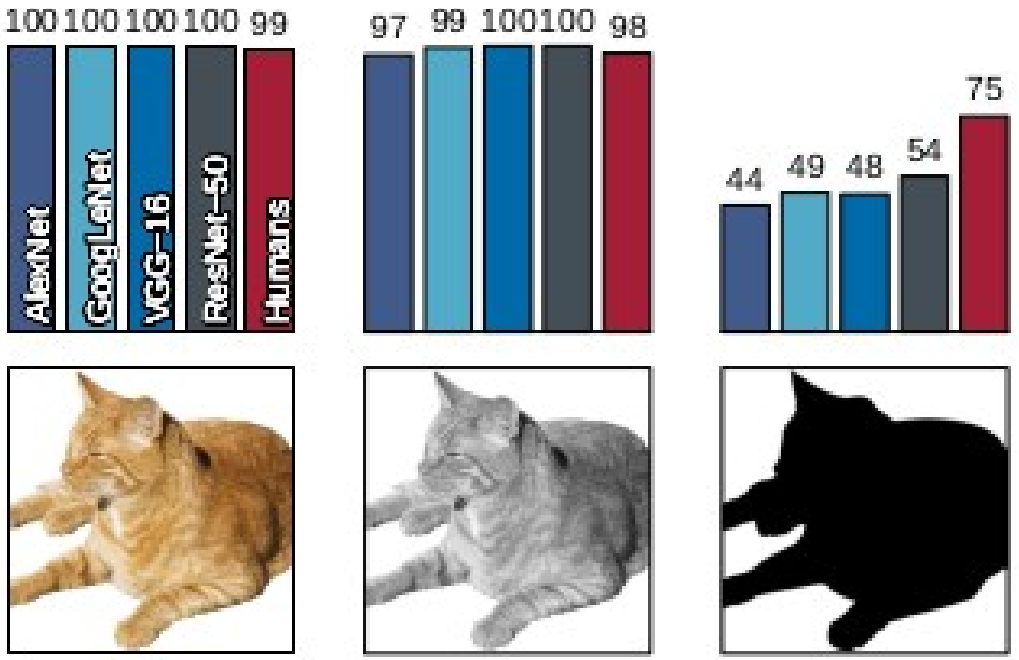
Original

Greyscale

160 color images
10 per category
white background

As original but greyscale

Experiments



Original

160 color images
10 per category
white background

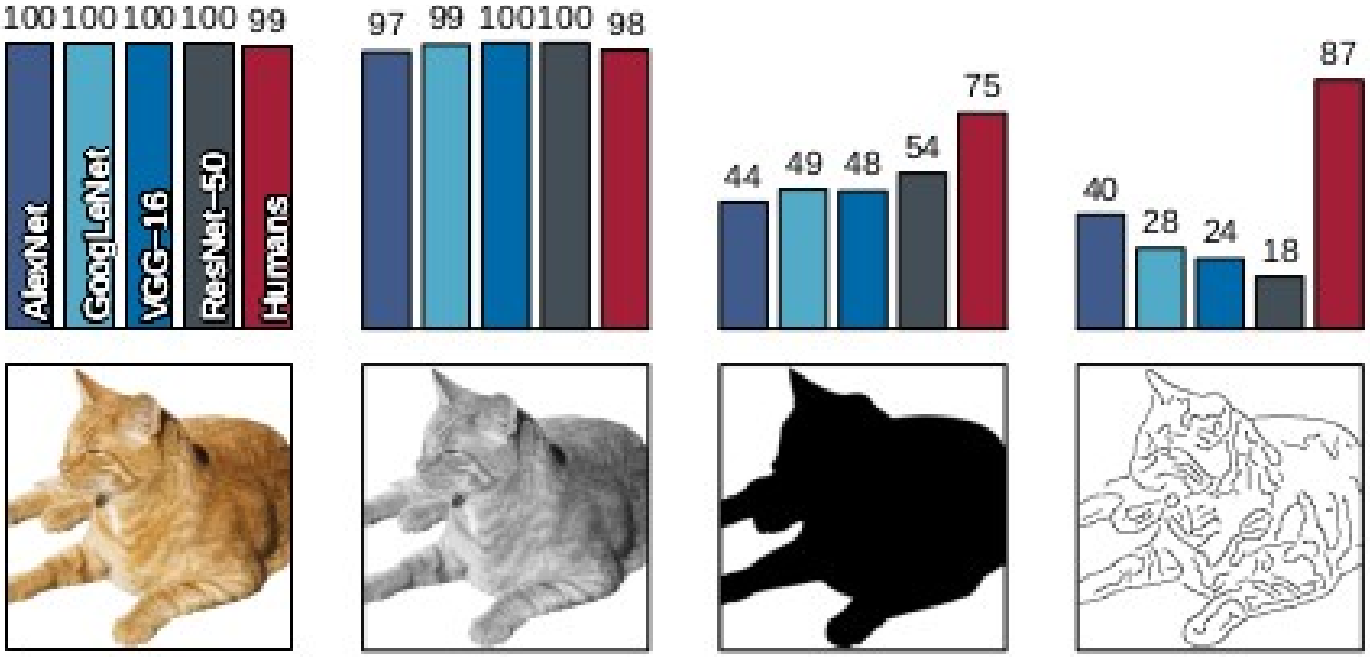
Greyscale

As original but greyscale

Silhouette

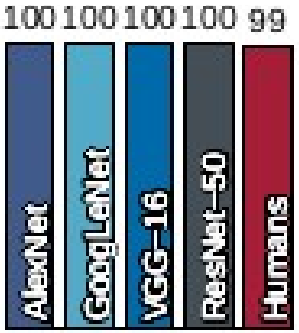
As original but only a manually created black mask

Experiments



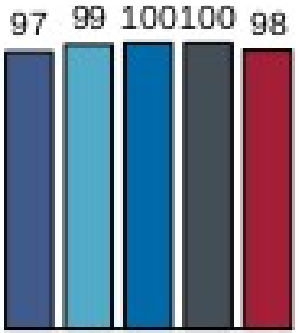
Original	Greyscale	Silhouette	Edge
160 color images 10 per category white background	*As original but greyscale*	*As original but only a manually created black mask*	*Canny edge extractor on original dataset*

Experiments



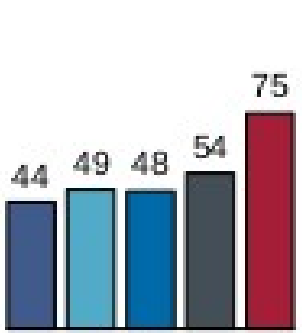
Original

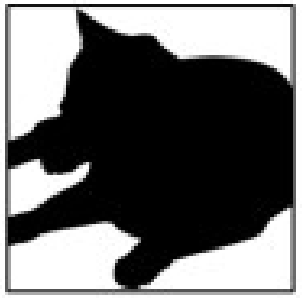
160 color images 10 per category white background



Greyscale

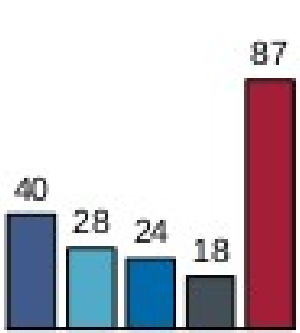
As original but greyscale





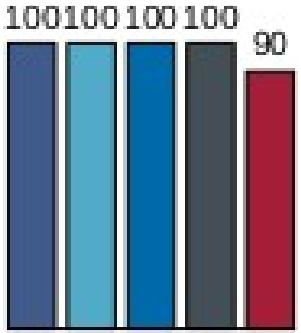
Silhouette

As original but only a manually created black mask



Edge

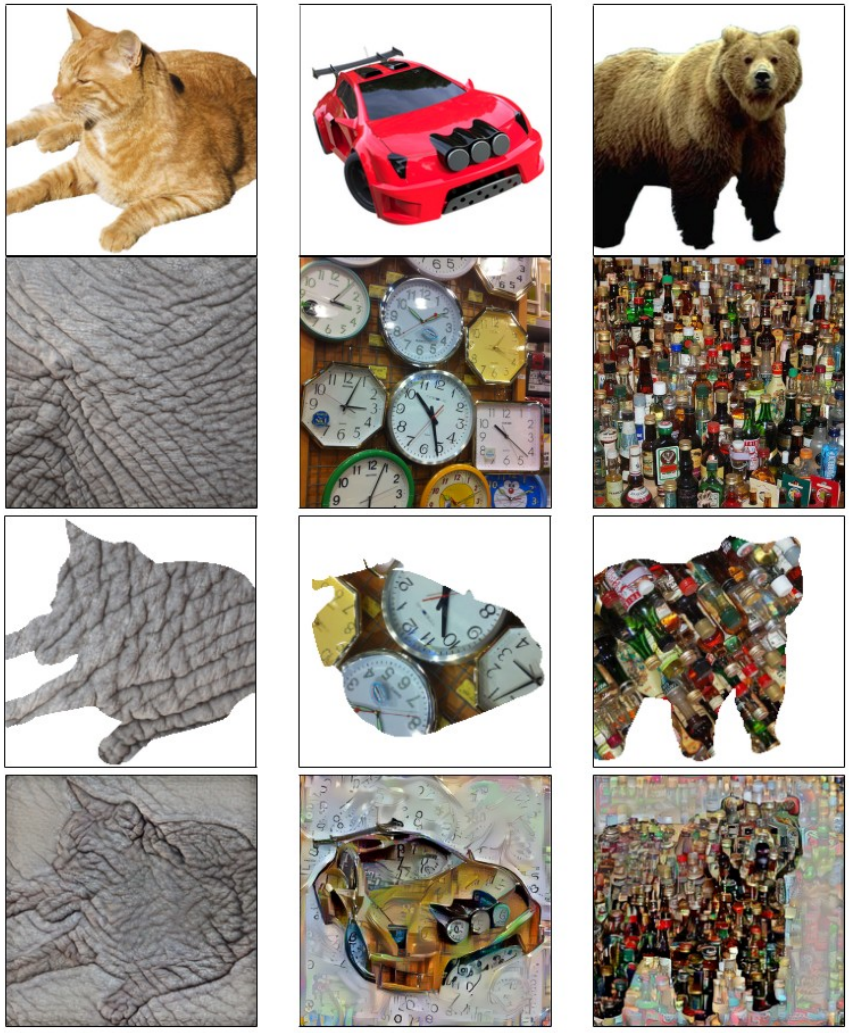
Canny edge extractor on original dataset



Texture

For items with no textured areas, eg "bottles" a cluster of those objects are considered as texture

Experiments



Original content images

Original texture images

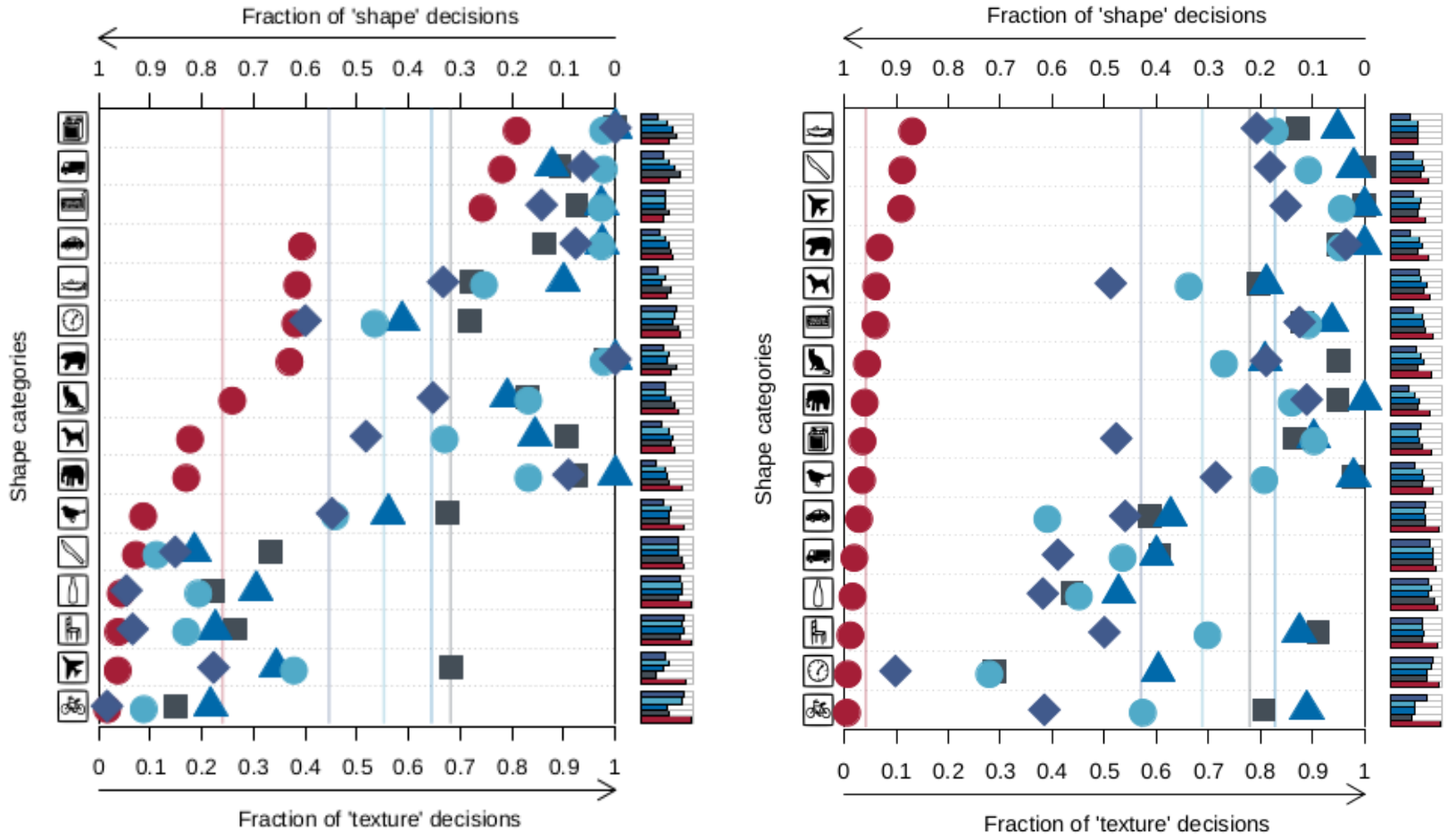
Filled silhouette experiment

Masked texture images inside The silhouettes. The textures had 360 degrees data augmentation

Cue conflict experiment

Using iterative style transfer
Gatys et al. (2016)

Cue Conflict Results



Human observers (**red circles**) AlexNet (**purple diamonds**) VGG-16 (**blue triangles**) GoogLeNet (**turquoise circles**) ResNet-50 (**grey squares**)

Overcoming the texture bias

Stylized-ImageNet (SIN)

Created by applying AdaIN style transfer to ImageNet images
Huang et al. (2017)

Model Metrics

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

Top-5 Accuracy of the stylized-ImageNet trained models compared to the ImageNet trained models

Model Metrics

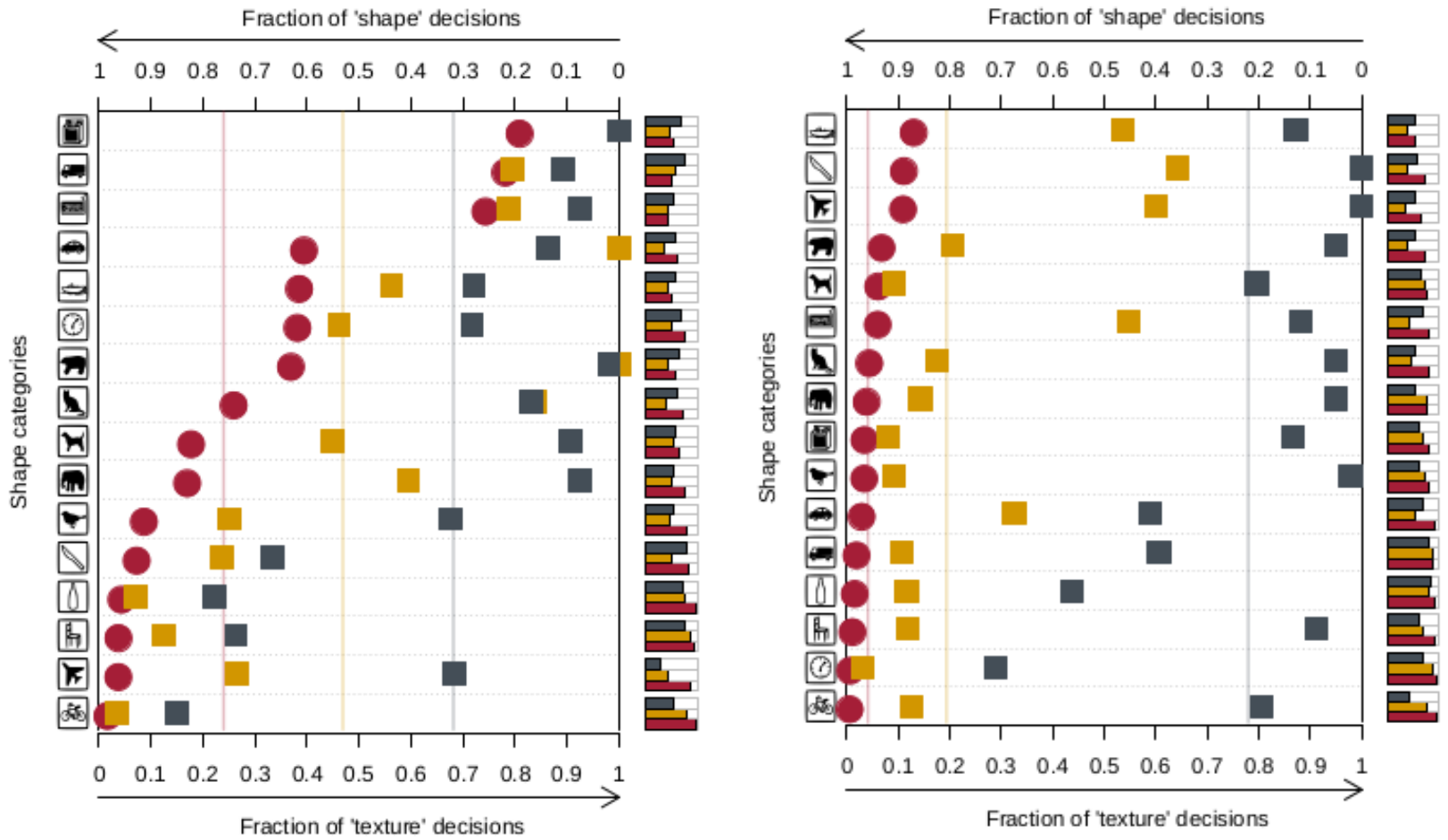
architecture	IN→IN	IN→SIN	SIN→SIN	SIN→IN
ResNet-50	92.9	16.4	79.0	82.6
BagNet-33 (mod. ResNet-50)	86.4	4.2	48.9	53.0
BagNet-17 (mod. ResNet-50)	80.3	2.5	29.3	32.6
BagNet-9 (mod. ResNet-50)	70.0	1.4	10.0	10.9

Top-5 Accuracy of the stylized-ImageNet trained models compared to the ImageNet trained models

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**

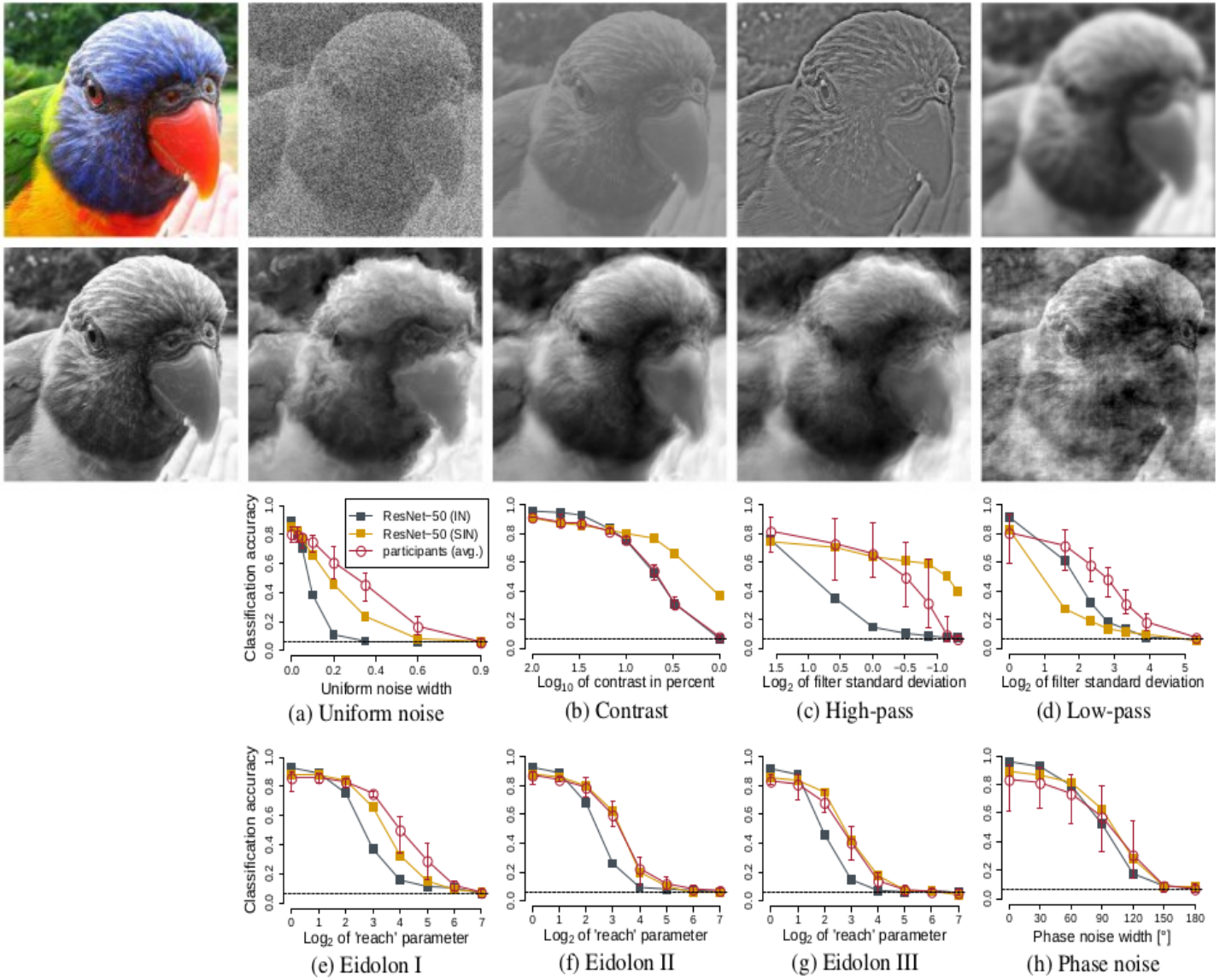
Shape-ResNet is the model trained jointly on SIN and IN and fine-tuned on IN

Bias Results



Human observers (**red circles**) ResNet-50 on Stylized-Imagenet (**orange squares**) ResNet-50 on Imagenet (**grey squares**)

Distortion Robustness Results



Questions?