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

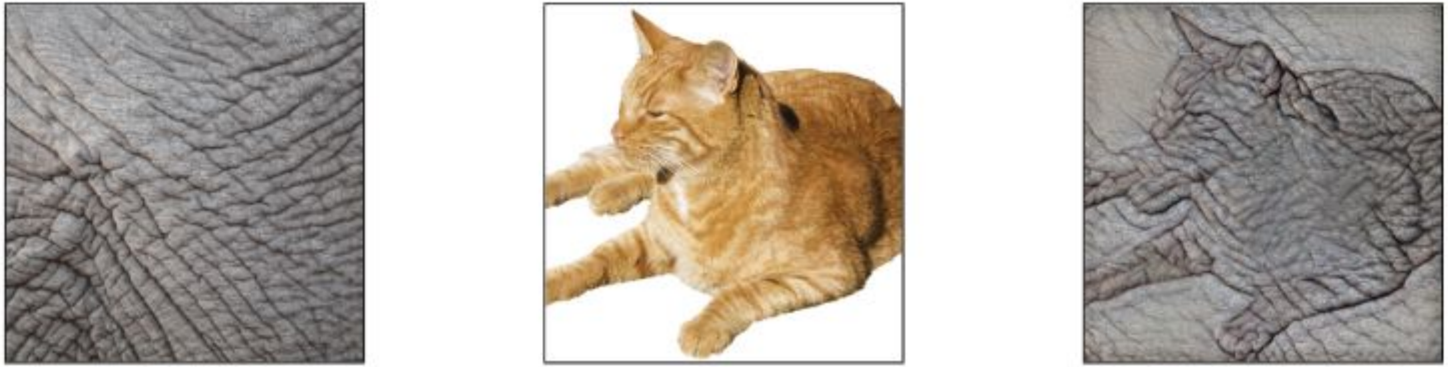
Hypothesis on mechanism of object recognition

- Contemporary view: **shape hypothesis**: progressively combine low level features into more and more complex objects until the object can be classified
  - Neuroscience view
  - Thought that CNNS operate similarly
- Propose an alternative view **texture hypothesis**: Object textures are more important than global object shape for CNN object recognition
  - CNNS look for (combinations of several) textures to identify objects

Contributions

1. Show that Imagenet trained models have a large texture bias.
2. Texture bias can be changed to shape bias by training on stylized imagenet.
3. Shape bias networks are resilient to many image distortions (including unseen distortions).
4. Shape biased networks reach higher performance on classification and object detection

Main Result



(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

Imagenet (IN) trained models - Response to image modifications

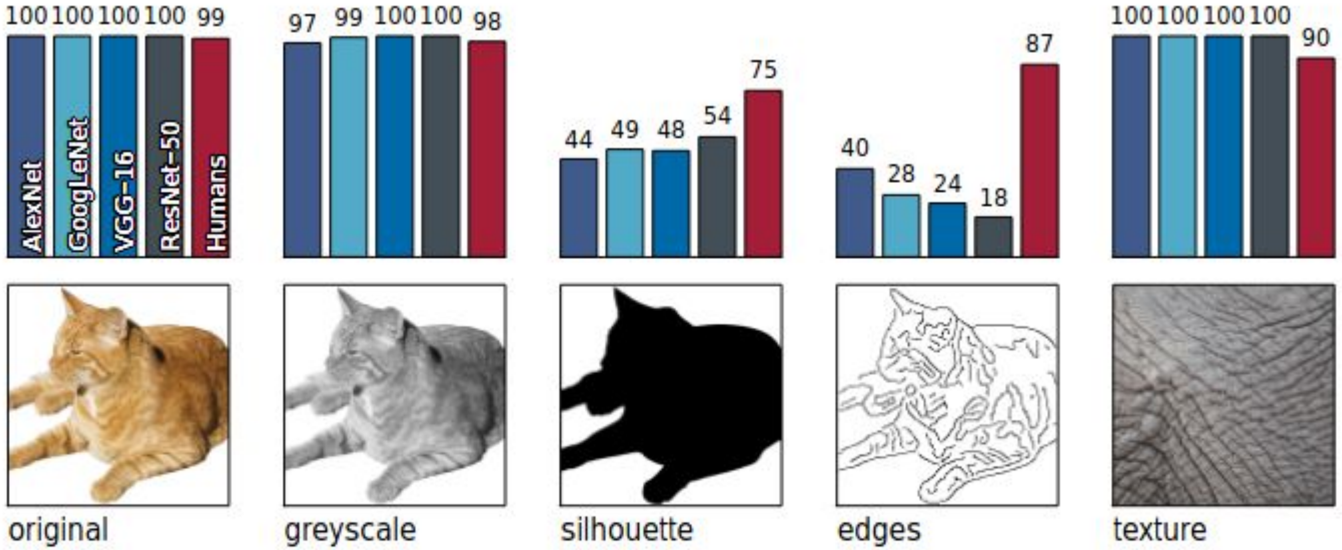
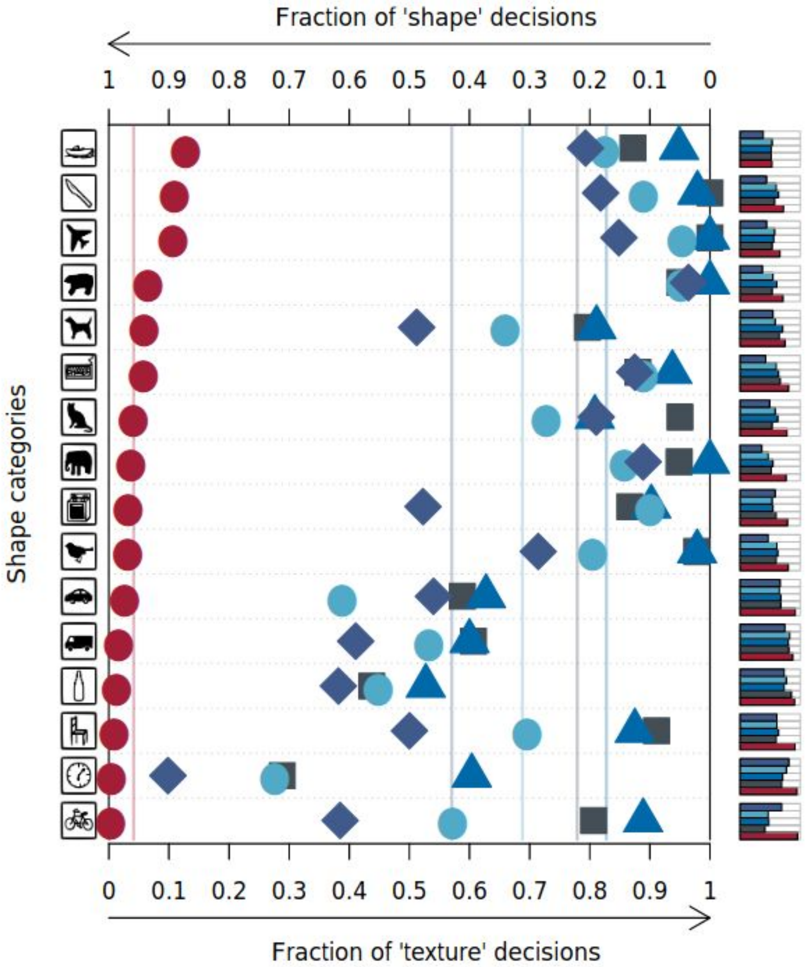


Figure 4: Classification results for human observers (red circles) and ImageNet-trained networks AlexNet (purple diamonds), VGG-16 (blue triangles), GoogLeNet (turquoise circles) and ResNet-50 (grey squares). Shape vs. texture biases for stimuli with cue conflict (sorted by human shape bias). Within the responses that corresponded to either the correct texture or correct shape category, the fractions of texture and shape decisions are depicted in the main plot (averages visualised by vertical lines). On the right side, small barplots display the proportion of correct decisions (either texture or shape correctly recognised) as a fraction of all trials. Similar results for ResNet-152, DenseNet-121 and Squeezenet1_1 are reported in the Appendix, Figure 13.



To Overcome Texture Bias - Train on Stylized Imagenet (SIN)

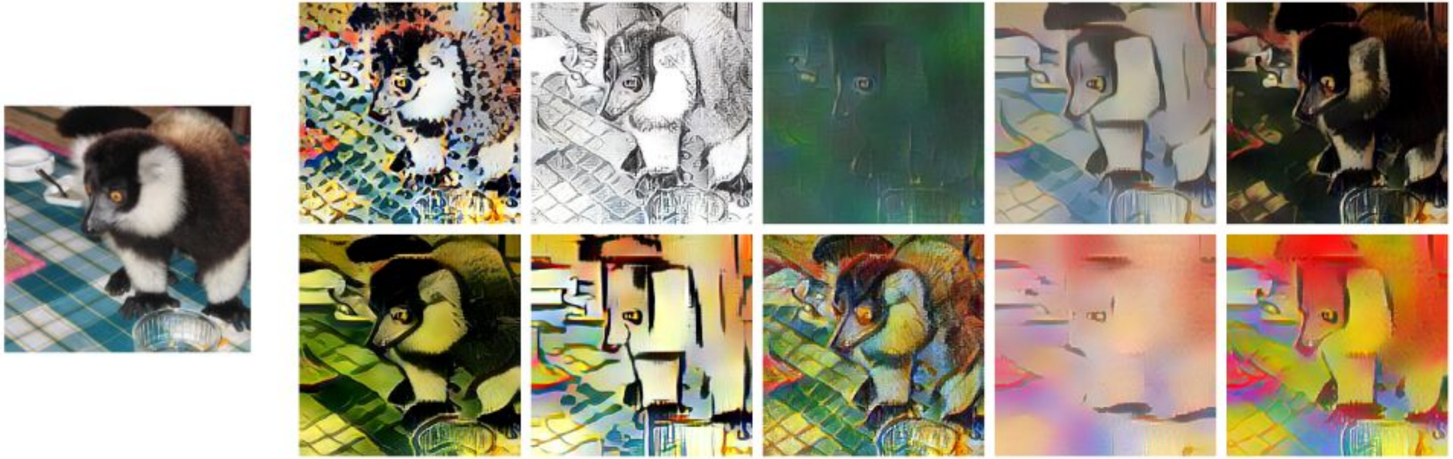
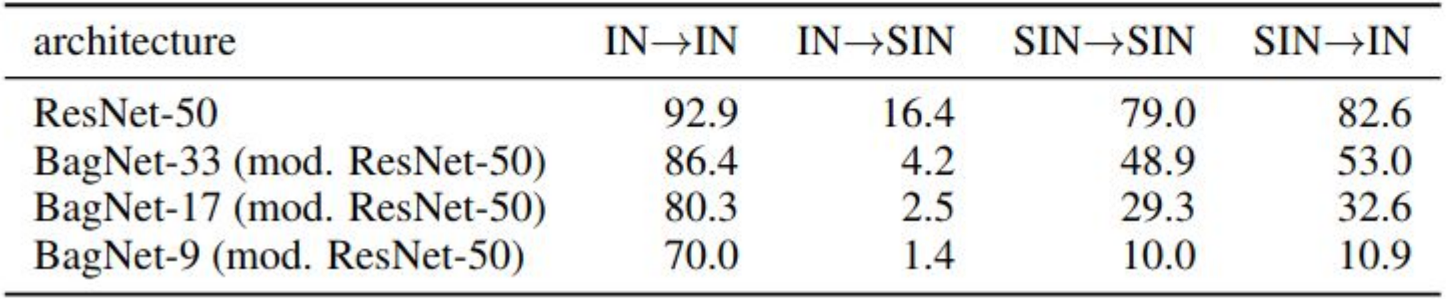


Figure 3: Visualisation of Stylized-ImageNet (SIN), created by applying AdaIN style transfer to ImageNet images. Left: randomly selected ImageNet image of class `ring-tailed lemur`. Right: ten examples of images with content/shape of left image and style/texture from different paintings. After applying AdaIN style transfer, local texture cues are no longer highly predictive of the target class, while the global shape tends to be retained. Note that within SIN, every source image is stylized only once.

Comparison with BagNets (Restricted RF size Nets)



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

Baises after Training with stylized Imagenet

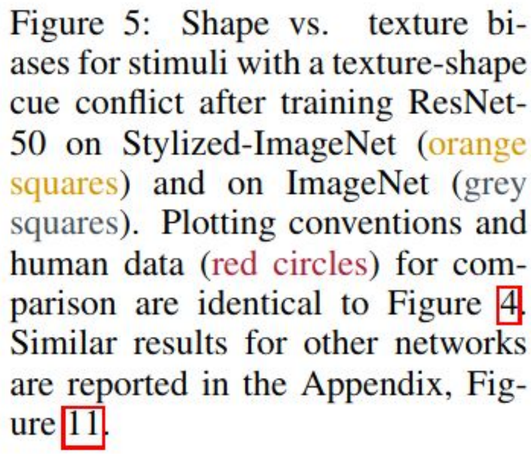
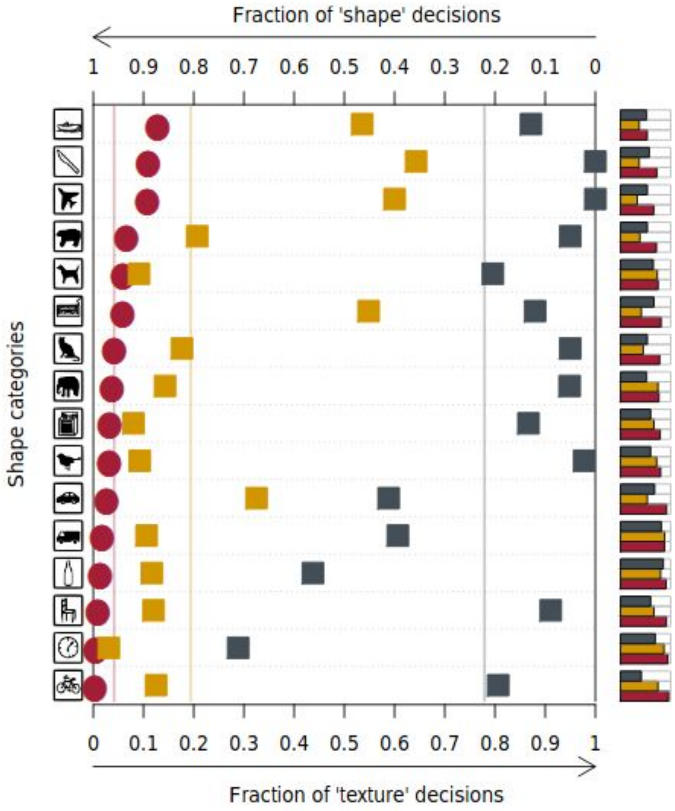


Figure 5: Shape vs. texture biases for stimuli with a texture-shape cue conflict after training ResNet-50 on Stylized-ImageNet (orange squares) and on ImageNet (grey squares). Plotting conventions and human data (red circles) for comparison are identical to Figure 4. Similar results for other networks are reported in the Appendix, Figure 11.



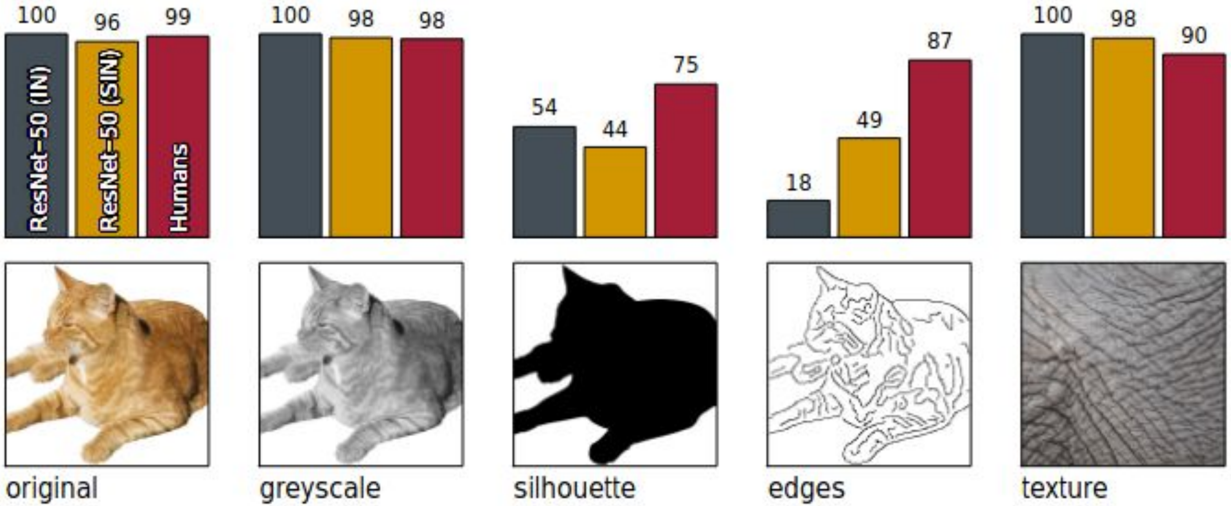
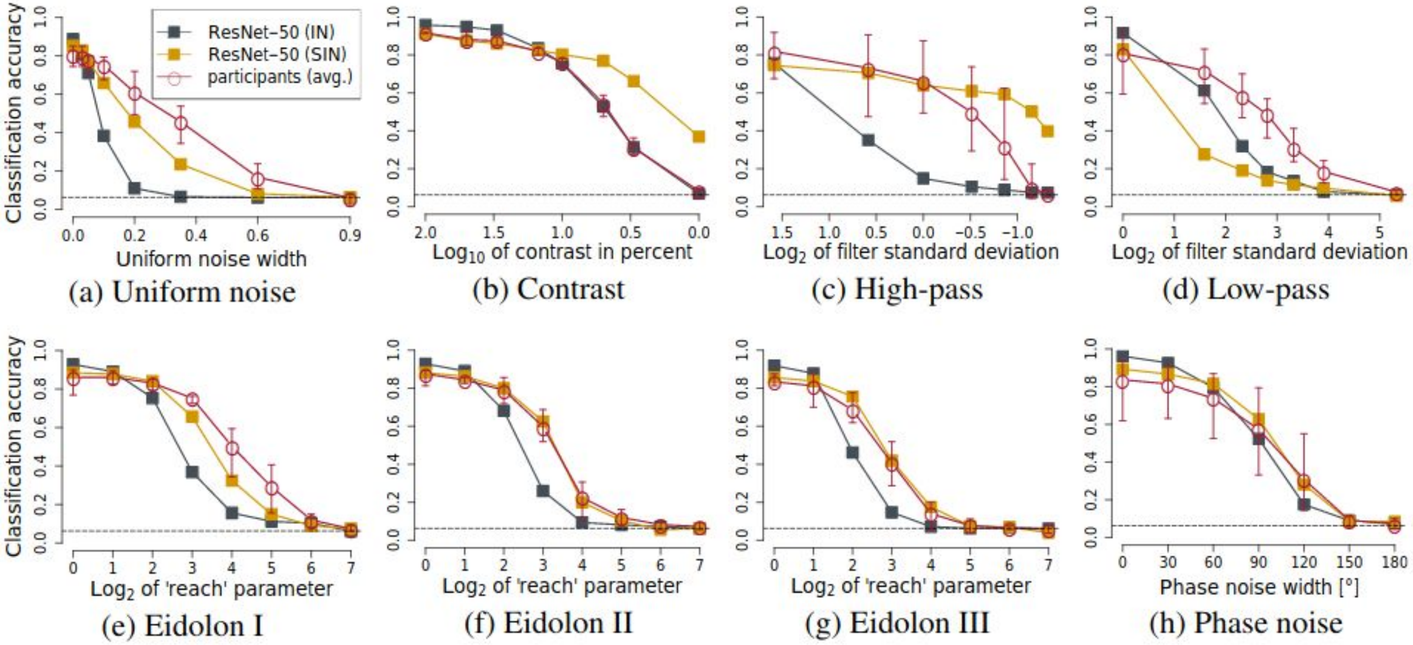


Figure 9: Accuracies and example stimuli for five different experiments without cue conflict, comparing training on ImageNet (IN) to training on Stylized-ImageNet (SIN).

Robustness to unseen image distortions



(a) Uniform noise

(b) Contrast

(c) High-pass

(d) Low-pass

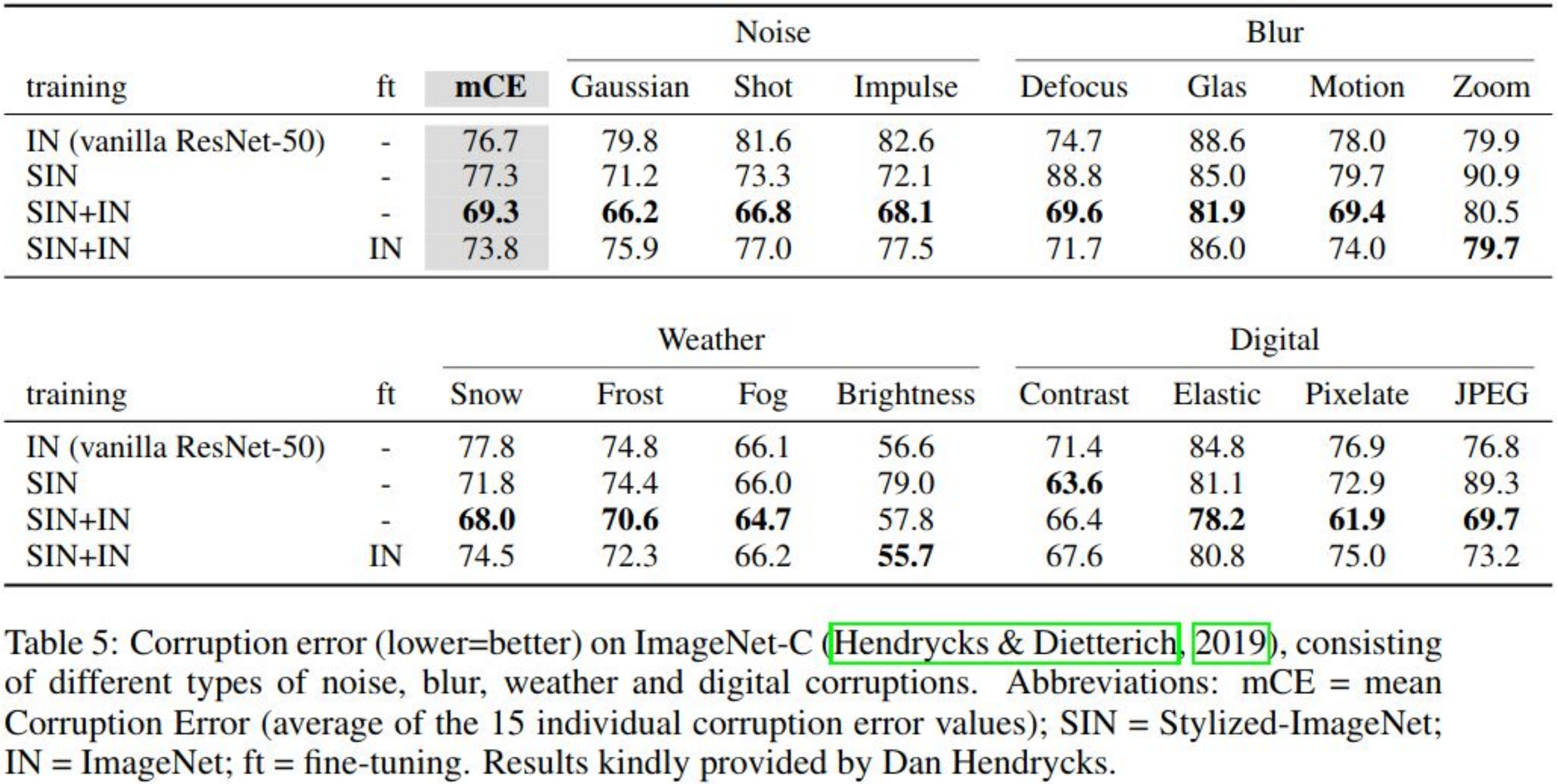
(e) Eidolon I

(f) Eidolon II

(g) Eidolon III

(h) Phase noise

Figure 6: Classification accuracy on parametrically distorted images. ResNet-50 trained on Stylized-ImageNet (SIN) is more robust towards distortions than the same network trained on ImageNet (IN).

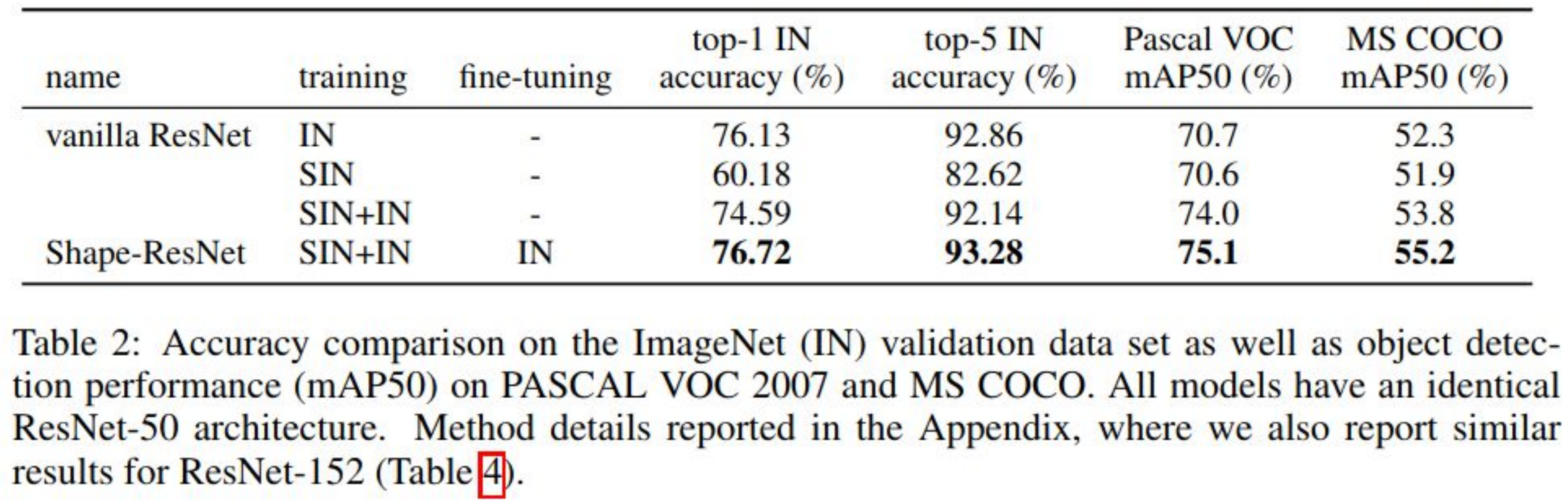


training	ft	mCE	Noise			Blur			
			Gaussian	Shot	Impulse	Defocus	Glas	Motion	Zoom
IN (vanilla ResNet-50)	-	76.7	79.8	81.6	82.6	74.7	88.6	78.0	79.9
SIN	-	77.3	71.2	73.3	72.1	88.8	85.0	79.7	90.9
SIN+IN	-	**69.3**	**66.2**	**66.8**	**68.1**	**69.6**	**81.9**	**69.4**	80.5
SIN+IN	IN	73.8	75.9	77.0	77.5	71.7	86.0	74.0	**79.7**

training	ft	Weather				Digital			
		Snow	Frost	Fog	Brightness	Contrast	Elastic	Pixelate	JPEG
IN (vanilla ResNet-50)	-	77.8	74.8	66.1	56.6	71.4	84.8	76.9	76.8
SIN	-	71.8	74.4	66.0	79.0	**63.6**	81.1	72.9	89.3
SIN+IN	-	**68.0**	**70.6**	**64.7**	57.8	66.4	**78.2**	**61.9**	**69.7**
SIN+IN	IN	74.5	72.3	66.2	**55.7**	67.6	80.8	75.0	73.2

Table 5: Corruption error (lower=better) on ImageNet-C (Hendrycks & Dietterich, 2019), consisting of different types of noise, blur, weather and digital corruptions. Abbreviations: mCE = mean Corruption Error (average of the 15 individual corruption error values); SIN = Stylized-ImageNet; IN = ImageNet; ft = fine-tuning. Results kindly provided by Dan Hendrycks.

Improved Performance



name	training	fine-tuning	top-1 IN accuracy (%)	top-5 IN accuracy (%)	Pascal VOC mAP50 (%)	MS COCO mAP50 (%)
vanilla ResNet	IN	-	76.13	92.86	70.7	52.3
	SIN	-	60.18	82.62	70.6	51.9
	SIN+IN	-	74.59	92.14	74.0	53.8
Shape-ResNet	SIN+IN	IN	**76.72**	**93.28**	**75.1**	**55.2**

Table 2: Accuracy comparison on the ImageNet (IN) validation data set as well as object detection performance (mAP50) on PASCAL VOC 2007 and MS COCO. All models have an identical ResNet-50 architecture. Method details reported in the Appendix, where we also report similar results for ResNet-152 (Table 4).