ImageNet-Trained CNNs are biased towards texture; Increasing shape bias improves accuracy and robustness

Robert Geirhos, Patricia Rubisch, Claudio Michaelis,

Matthias Bethge*, Felix A. Wichmann*, Weiland Brendel*

University of Tubingen

ICLR 2019 (Oral)

Texture vs Shape

(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

Figure 1: Classification of a standard ResNet-50 of **(a)** a texture image (elephant skin: only texture cues); **(b)** a normal image of a cat (with both shape and texture cues), and **(c)** an image with a texture-shape cue conflict, generated by style transfer between the first two images.

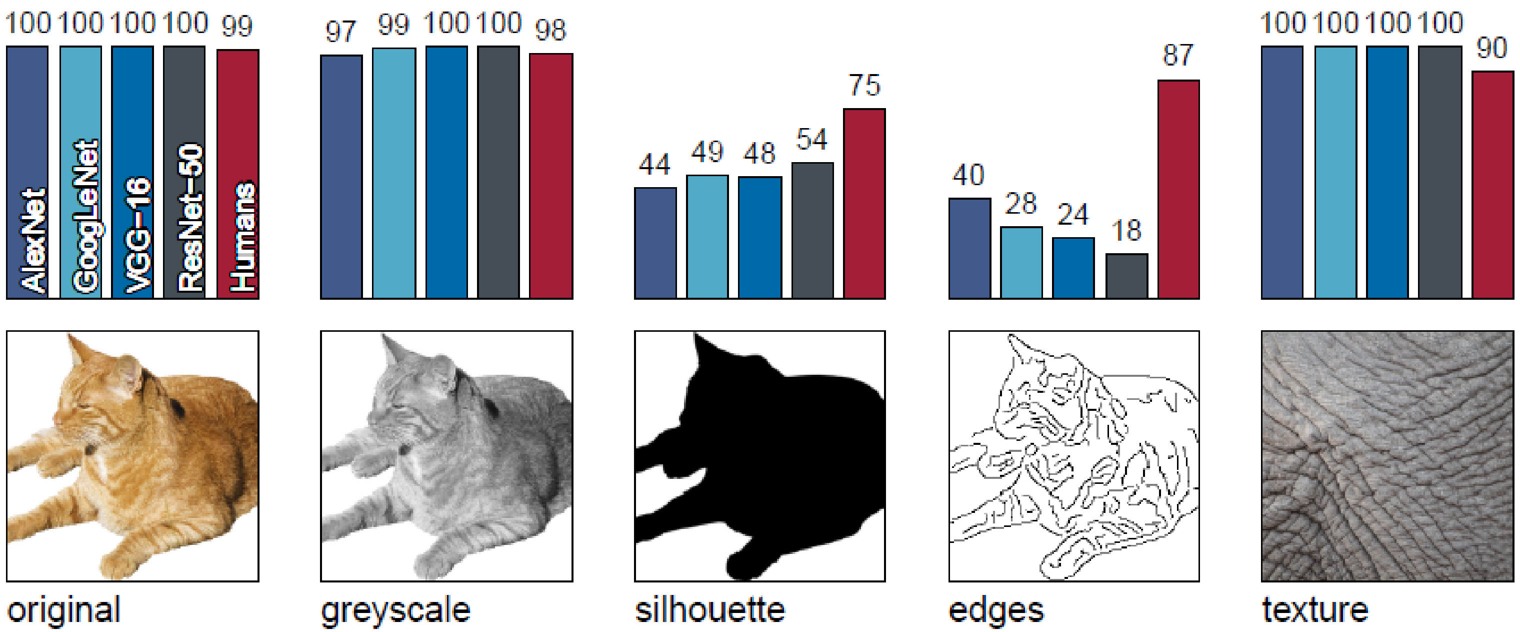
CNN

- **Shape Hypothesis**
  - Combines low-level features(ex: edges) to increasingly complex shapes
- **Texture Hypothesis**
  - Can still classify texturized images perfectly well, even if the global shape structure is completely destroyed
  - Explicitly constrained receptive field sizes throughout all layers are able to reach surprisingly high accuracies on ImageNet

  => ImageNet-Trained CNNs are biased Towards Texture

Human vs CNN

- **Without Cue Conflict**
  - Original : Only selected object that were correctly classified by all four networks



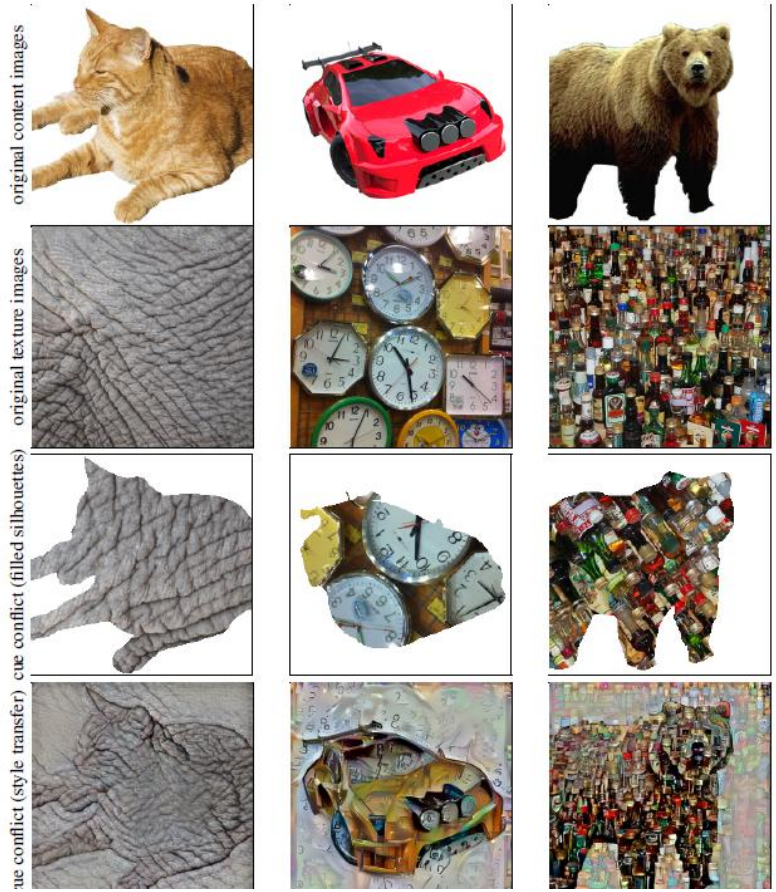
What if Image has a Cue conflict?

- **Generate Cue conflict images using style transfer (80 per category)**
- **Above image, 'Cat' and 'Elephant' are both answer**

Texture-shape cue conflict

What if Image has a Cue conflict?

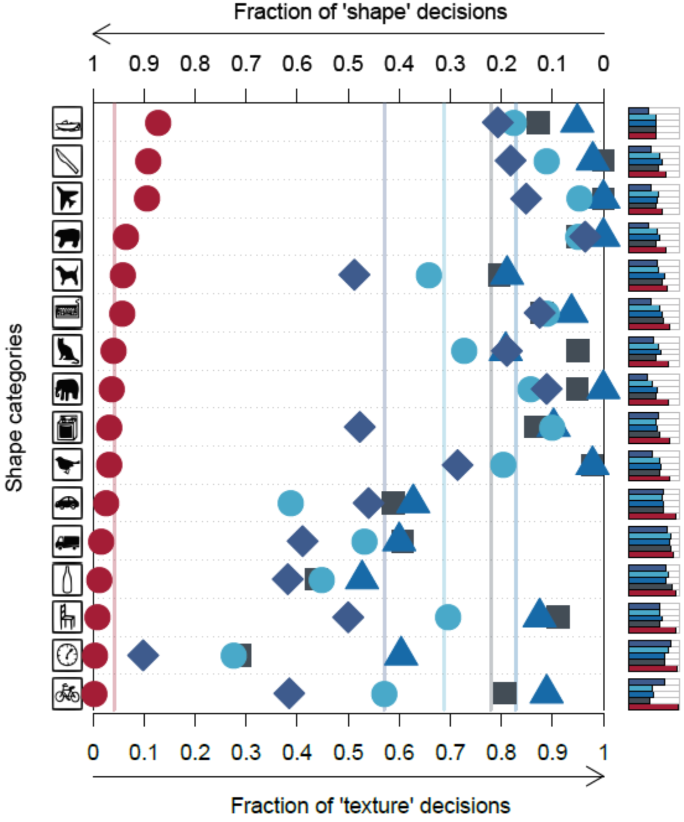
- **Various Image with different texture**



What if Image has a Cue conflict?

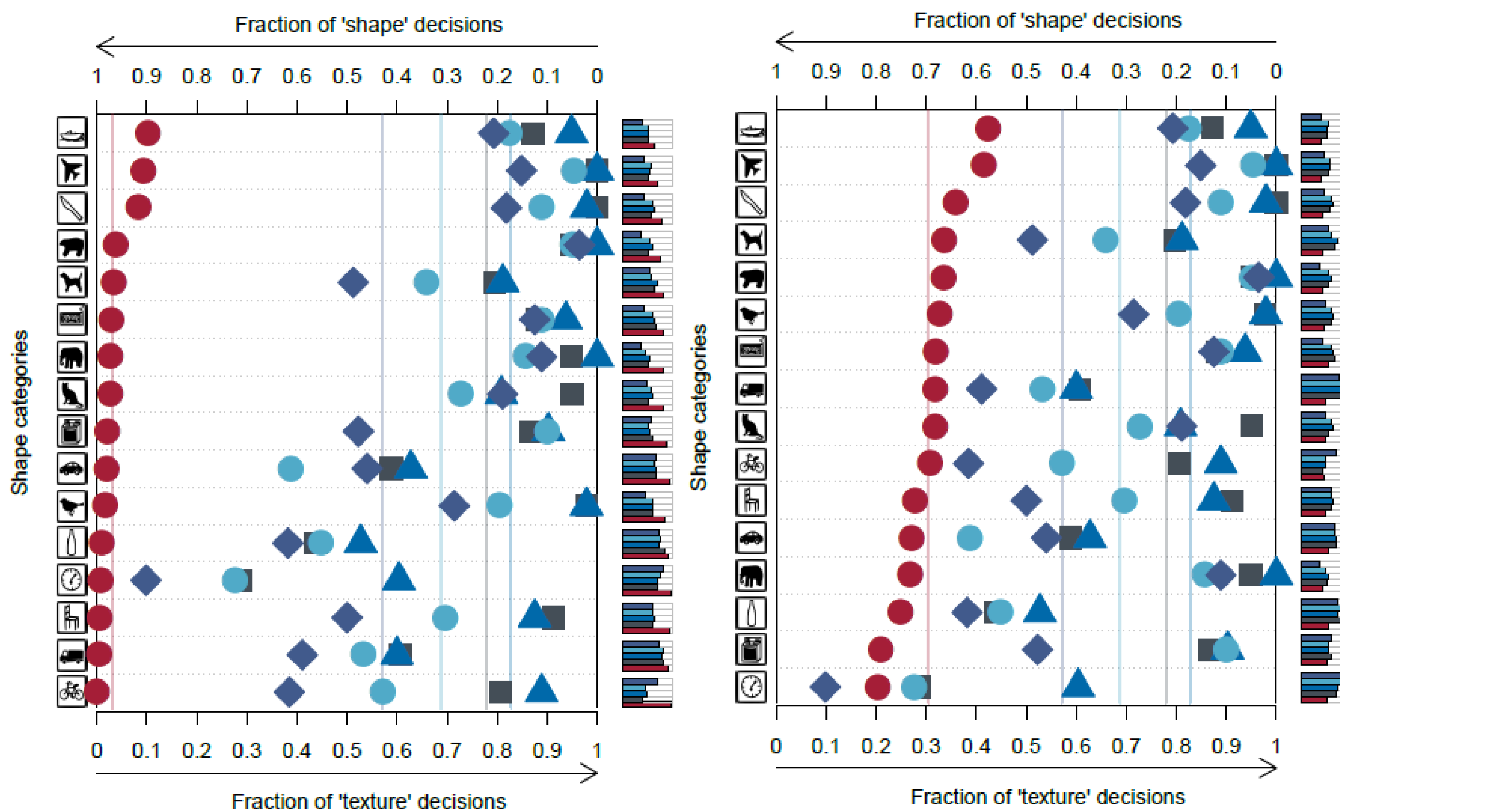
- **Human: Biased towards shape**
- **CNN: Biased towards texture**

Texture-shape cue conflict



What if Image has a Cue conflict?

- **Human with different instructions**



(a) Shape bias instruction

(b) Texture bias instruction

Overcoming the texture bias of CNNs

- **Create images with various textures**

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.

Overcoming the texture bias of CNNs

- **Styled-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

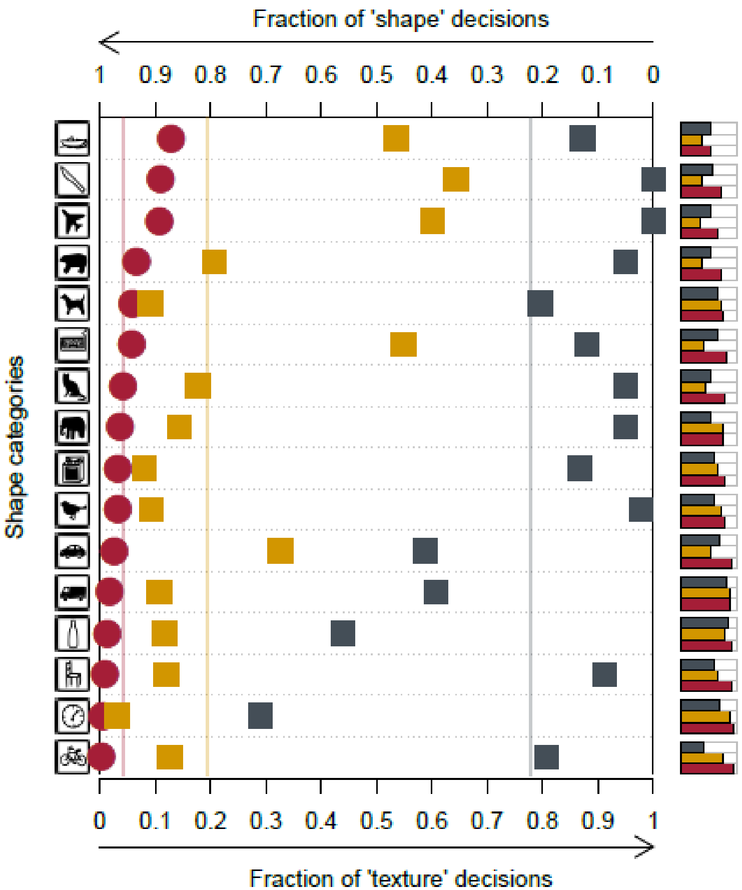
Table 1: Stylized-ImageNet cannot be solved with texture features alone. Accuracy comparison (in percent; top-5 on validation data set) of a standard ResNet-50 with Bag of Feature networks (BagNets) with restricted receptive field sizes of 33×33, 17×17 and 9×9 pixels. Arrows indicate: train data→test data, e.g. IN→SIN means training on ImageNet and testing on Stylized-ImageNet.

Guessing SIN is originally more difficult task than IN
BagNet : In order to test whether local texture features are sufficient
IN → IN : BagNet-9 also shows high accuracy
IN → SIN : Biased towards texture

Overcoming the texture bias of CNNs

▪ Styled-ImageNet (SIN)

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.



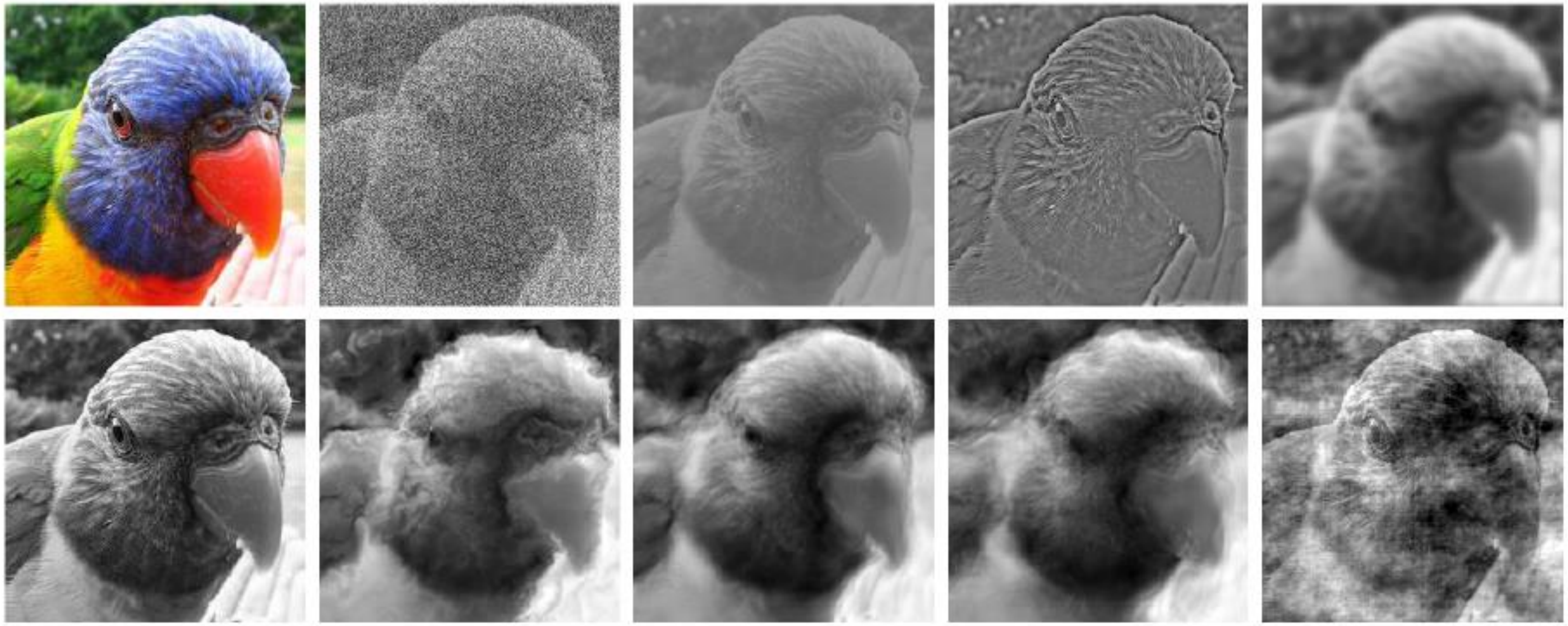
Shape-ResNet

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**

- **3 Advantages**
  - Classification Performance
  - Transfer Learning
  - Robustness against distortions

Robustness

- **Images with various filters**



Robustness

- **Images with various filters**

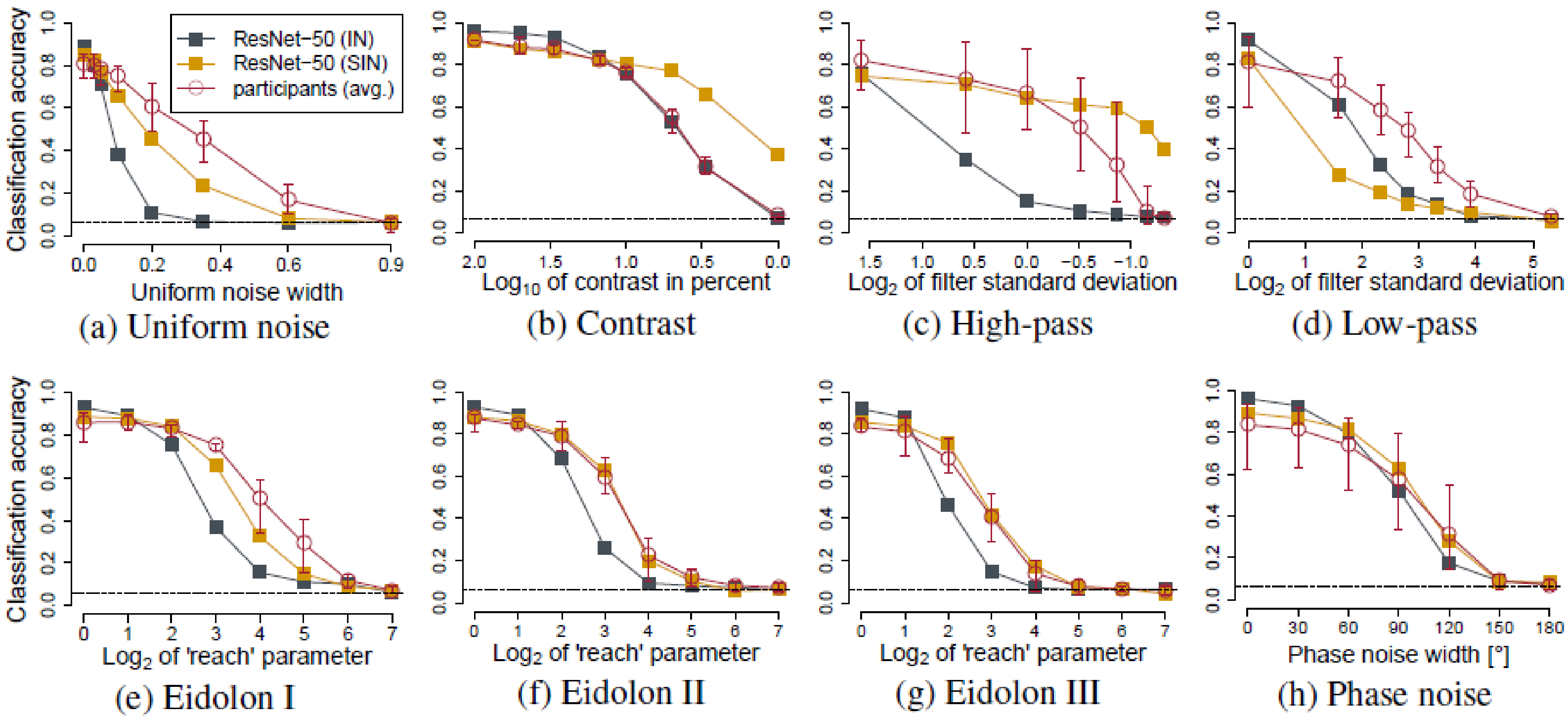


Figure 6: Classification accuracy on parametrically distorted images. ResNet-50 trained on Stylized-ImageNet (SIN) is more robust towards distortions than the same network trained on ImageNet (IN).

Summary

- **ImageNet-trained CNNs are originally biased towards texture**
- **Increasing shape bias improves accuracy and robustness**
- **Stylized-ImageNet(AdaIN style transfer) → Shape ResNet**
  - Classification Performance
  - Transfer Learning
  - Robustness