Texture vs shape in computer vision networks.

(Shape)

Daniil Klochkov

Texture and shape. What is it?

(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.

How texture bias was discovered?

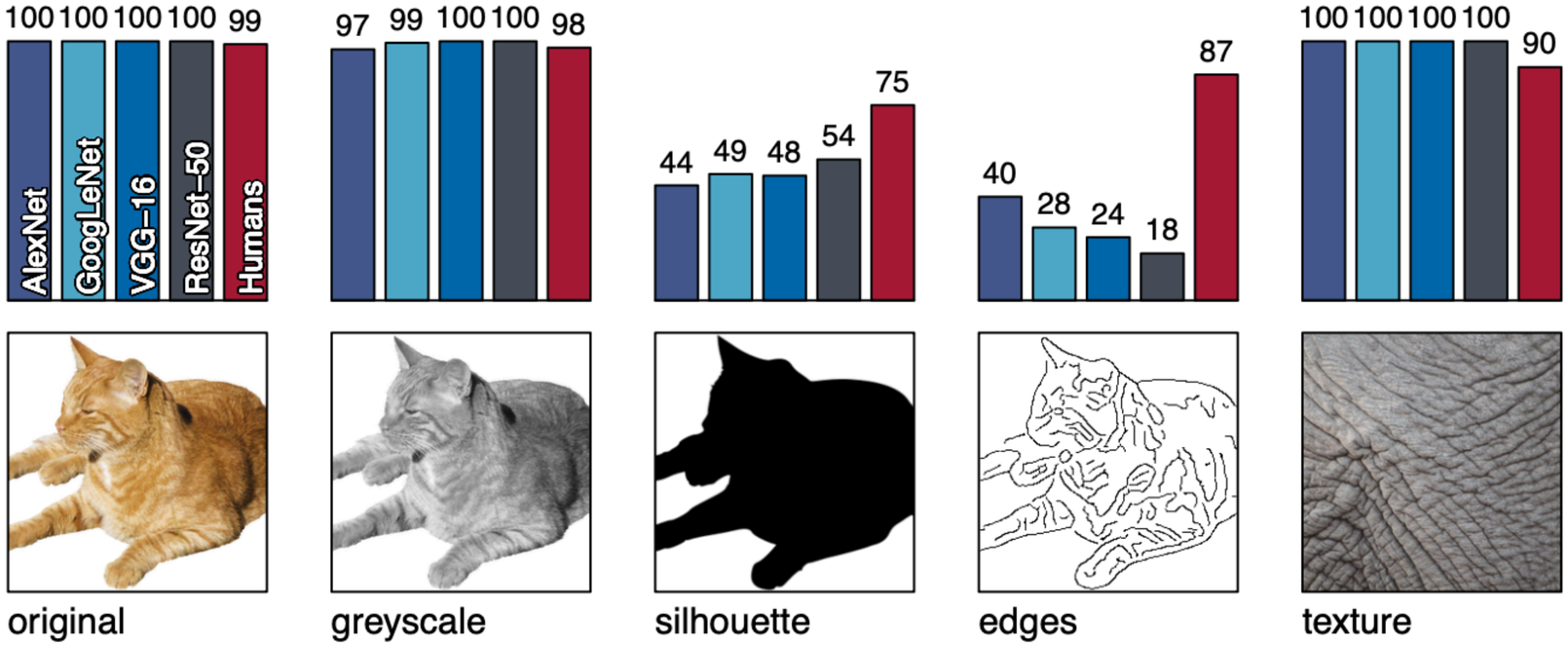
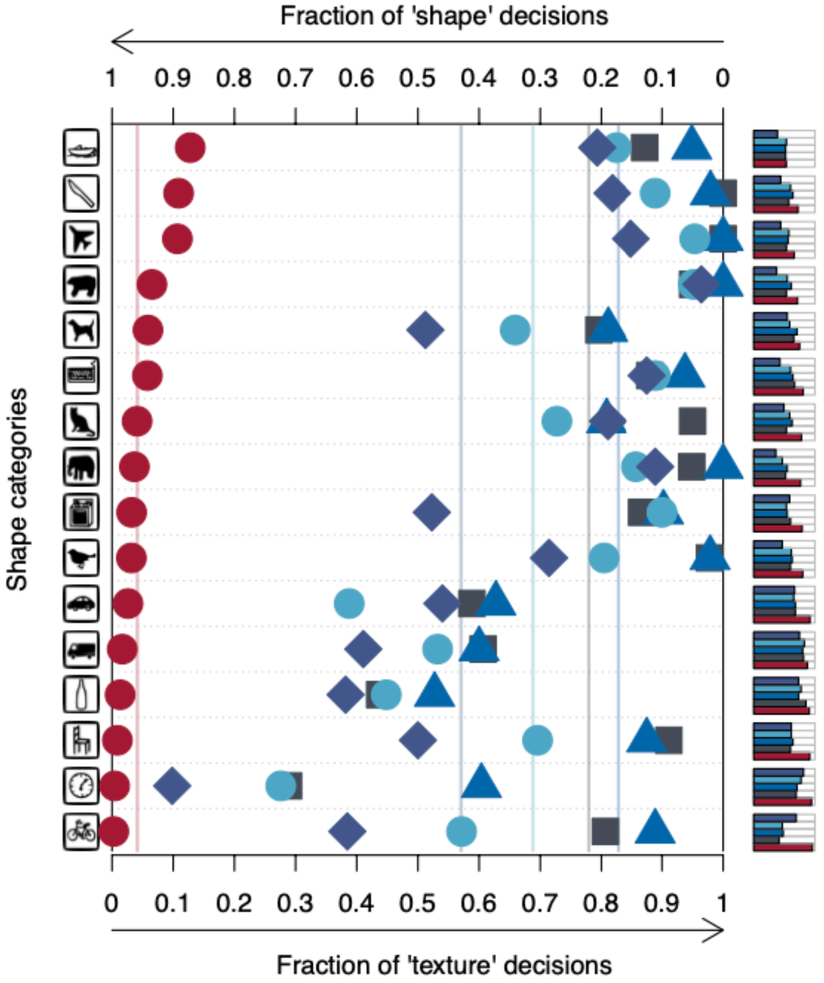


Figure 2: Accuracies and example stimuli for five different experiments without cue conflict.

How to count shape/texture bias

Как численное оценить концентрацию на текстуре или форме

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.



SIN

Stylized-ImageNet

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.

Is the definitely no data on texture in SIN?

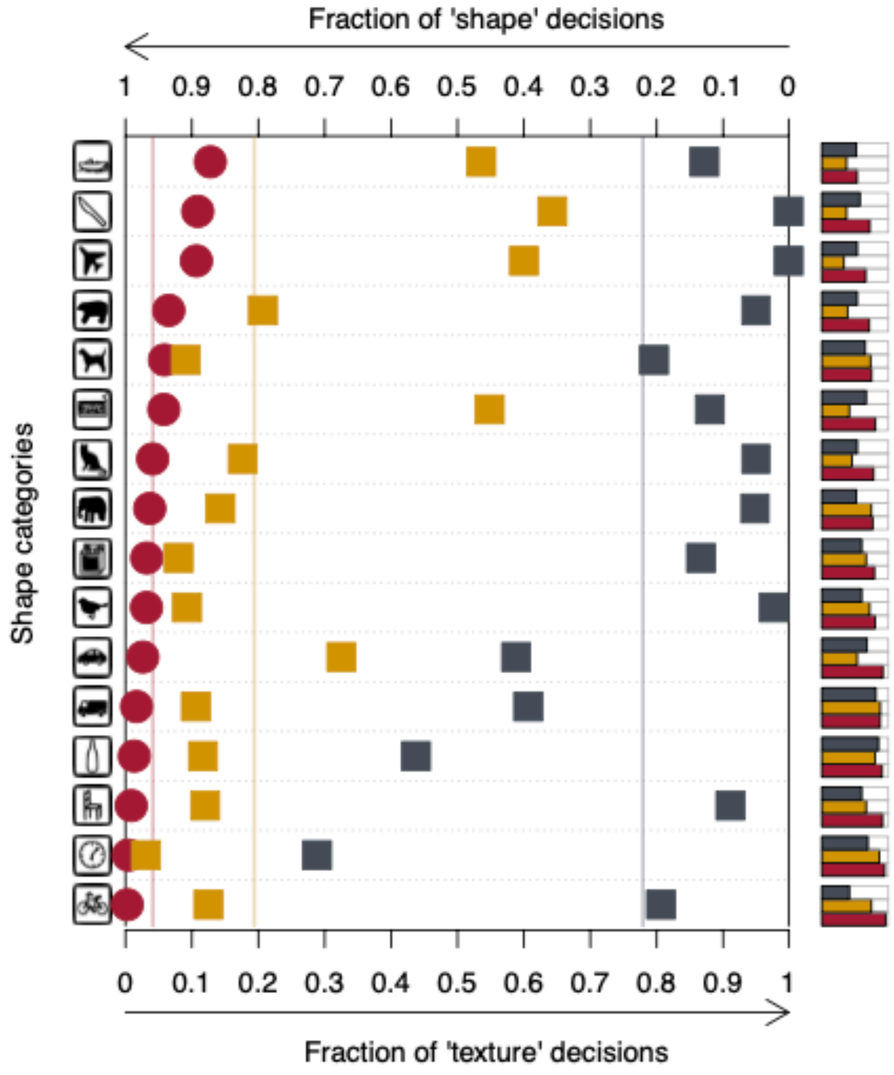
В SIN точно нет данных о текстуре?

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.

What are models trained using SIN focused on and how much

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.



How good are models trained on SIN in general

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).

Stability of a neural network with shape bias

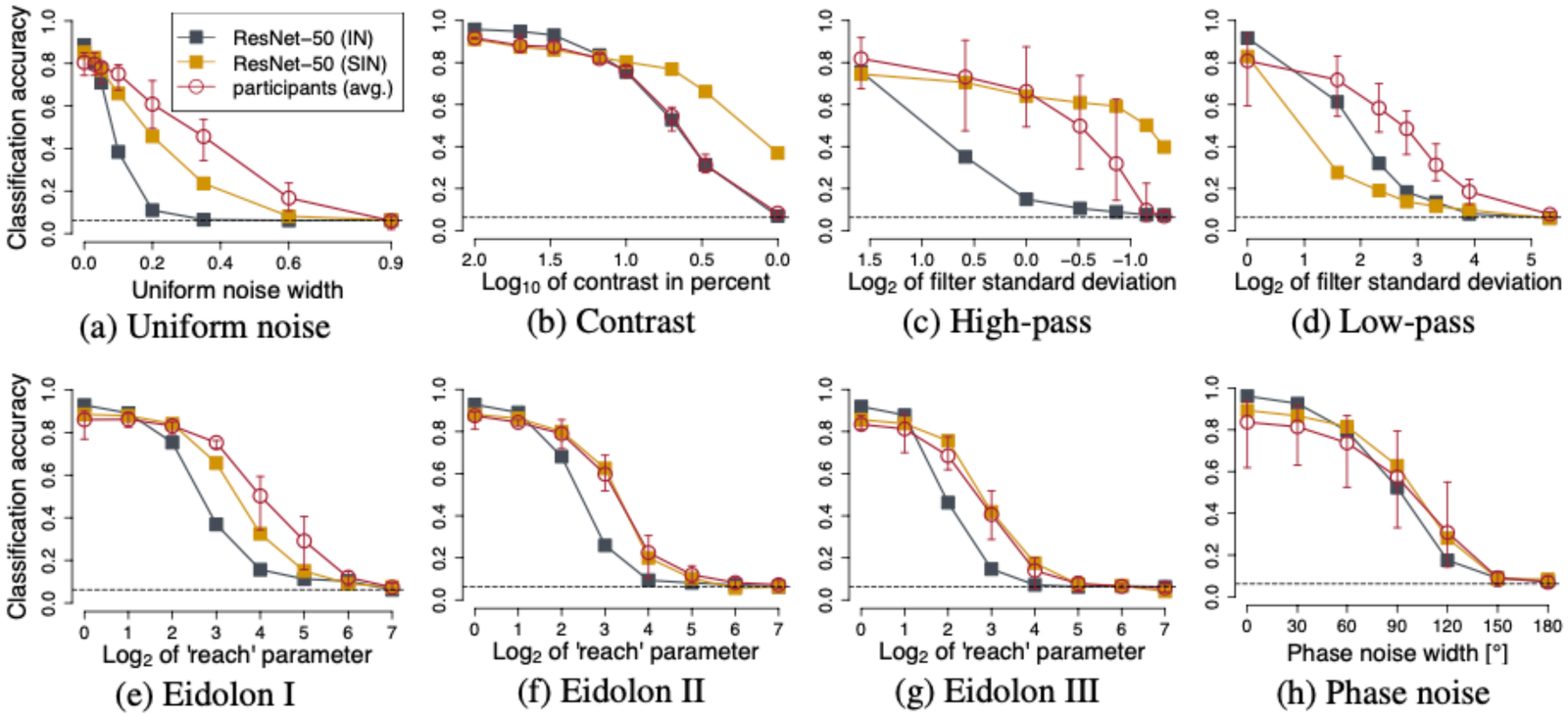


Figure 6: Classification accuracy on parametrically distorted images. ResNet-50 trained on Stylized-ImageNet (SIN) is more robust towards distortions than the same network trained on ImageNet (IN).

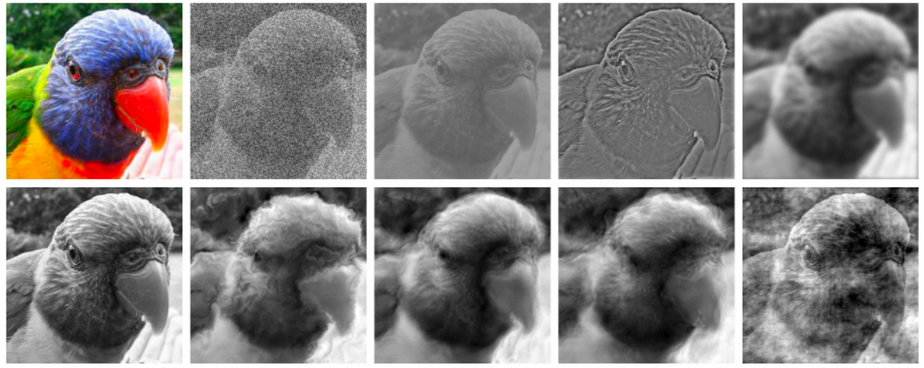


Figure 8: Visualisation of image distortions. One exemplary image (class `bird`, original image in colour at the top left) is manipulated as follows. From left to right: additive uniform noise, low contrast, high-pass filtering, low-pass filtering. In the row below, a greyscale version for comparison; the other manipulations from left to right are: Eidolon manipulations I, II and III as well as phase noise. Figure adapted from Geirhos et al. (2018) with the authors' permission.

More datasets with independent texture

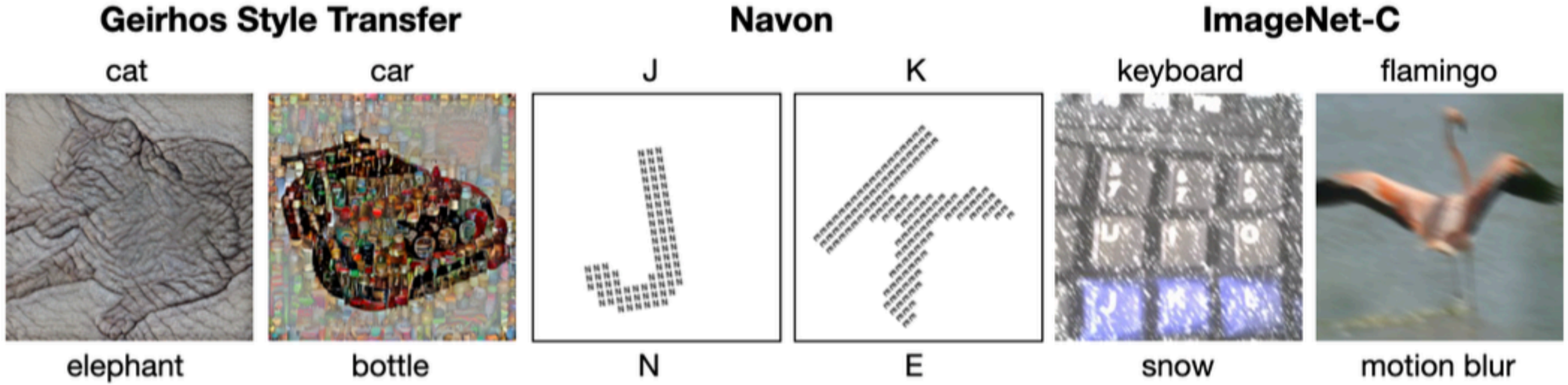


Figure 1: Example items from the three datasets labeled according to shape (top) and texture (bottom). GST items reproduced with permission of [36, 91]. ImageNet-C items from [44].

What does a neural network learn better to recognize?

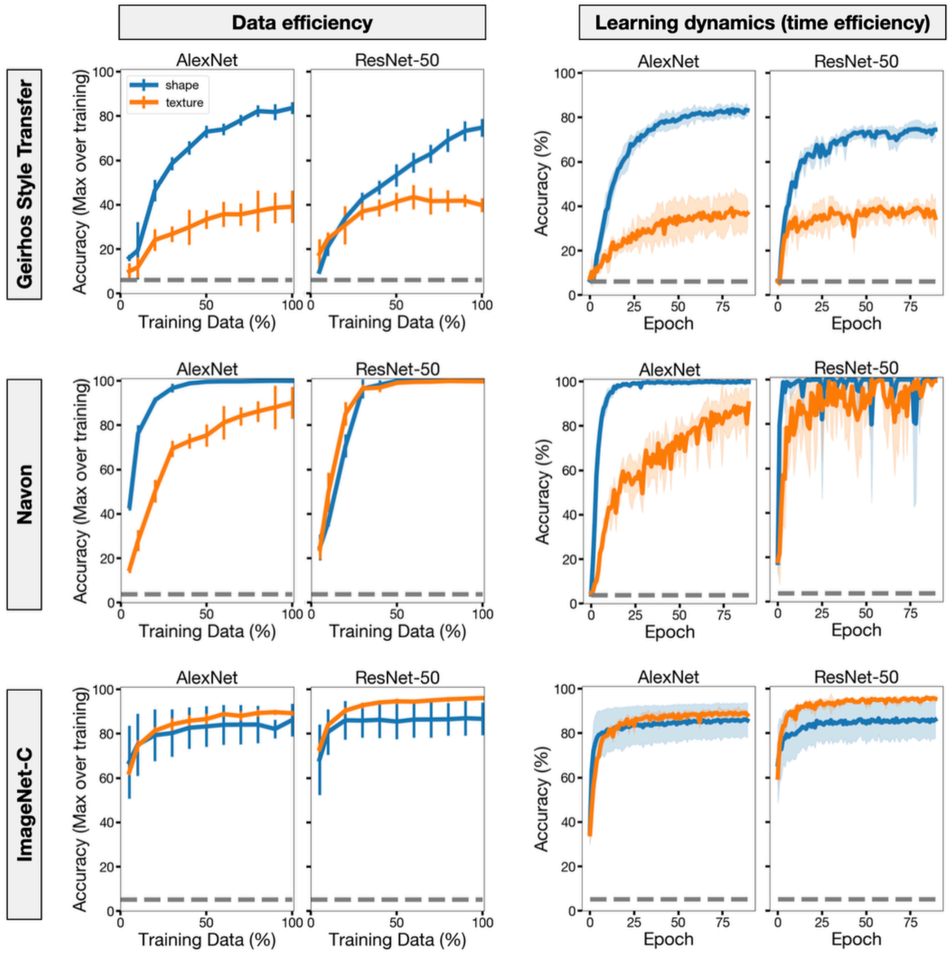


Figure 2: **Task performance of AlexNet and ResNet-50 models as a function of data (first column) and time efficiency (second column), for each dataset (rows). (Data efficiency.)** Performance is the maximum classification accuracy for shape (blue) or texture (orange) over training. All plots show the mean ± SD across 5 splits. Dashed line indicates chance performance. **(Learning dynamics).** Accuracy over training time on 100% of the training data (GST: 716 items, Navon: 2875 items, ImageNet-C: 80750 items).

How does augmentation affect shape bias

Table 2: **Color distortion, Gaussian blur, Gaussian noise, and Sobel filtering reduce texture bias.** ResNet-50 models were trained on ImageNet with random crops for 90 epochs. Augmentations were applied with 50% probability. Bolding indicates significantly greater shape bias than the baseline model ($p < 0.05$, permutation test).

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%

The compromise between IN top-1 and shape bias

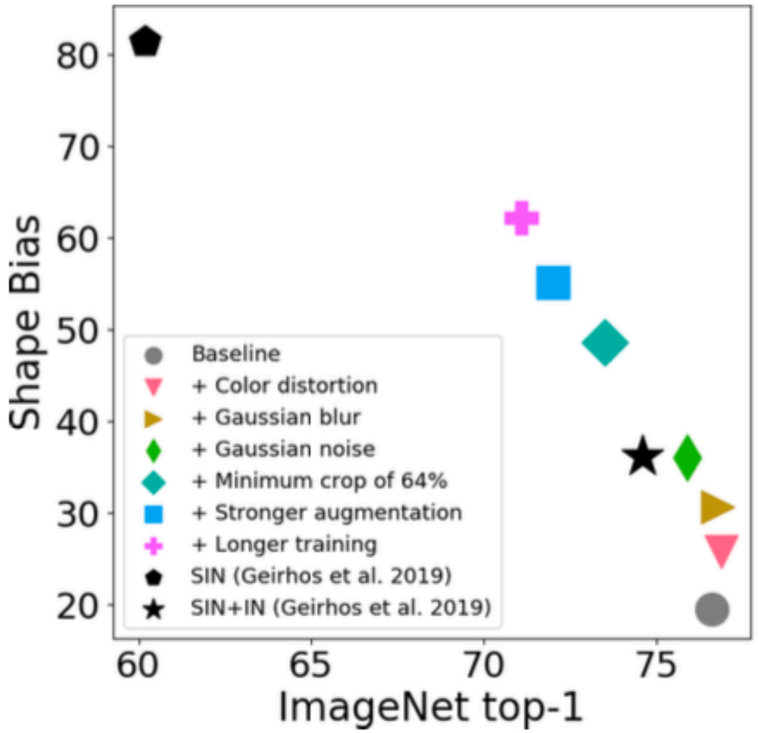


Figure 3: **Tradeoff between ImageNet top-1 and shape bias for models trained with different augmentation.** ResNet-50 models with naturalistic data augmentation achieve comparable tradeoffs to those from Geirhos et al. [36].

What is the correlation between IN top-1 and NN bias

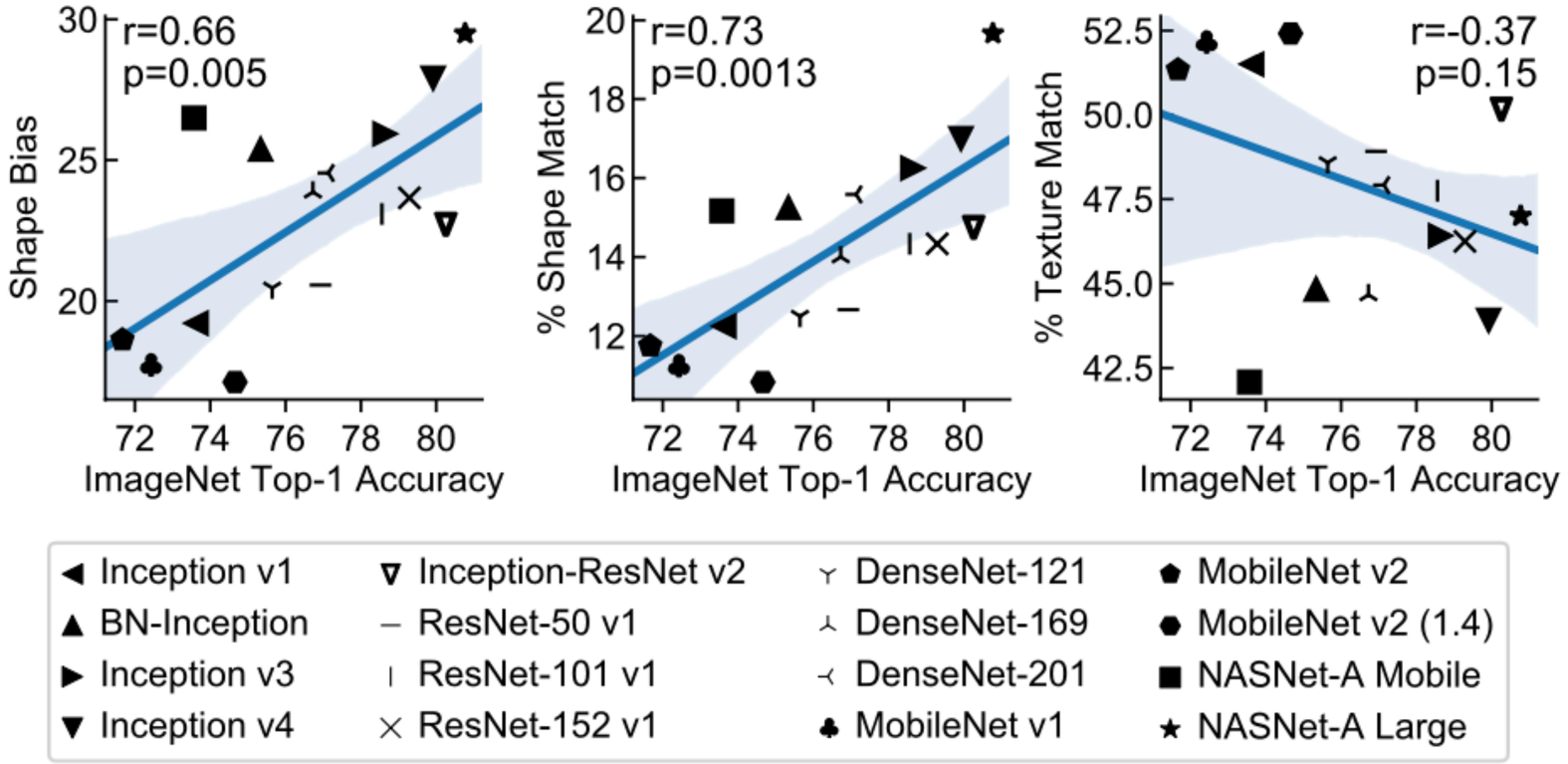


Figure 4: **Among high-performing ImageNet models, shape bias and accuracy correlate with ImageNet accuracy.** p-values indicate significance according to a t distribution. Blue line is a least squares fit to the plotted points; shaded area reflects 95% bootstrap confidence interval.

Does information about the texture and shape remain in the NN

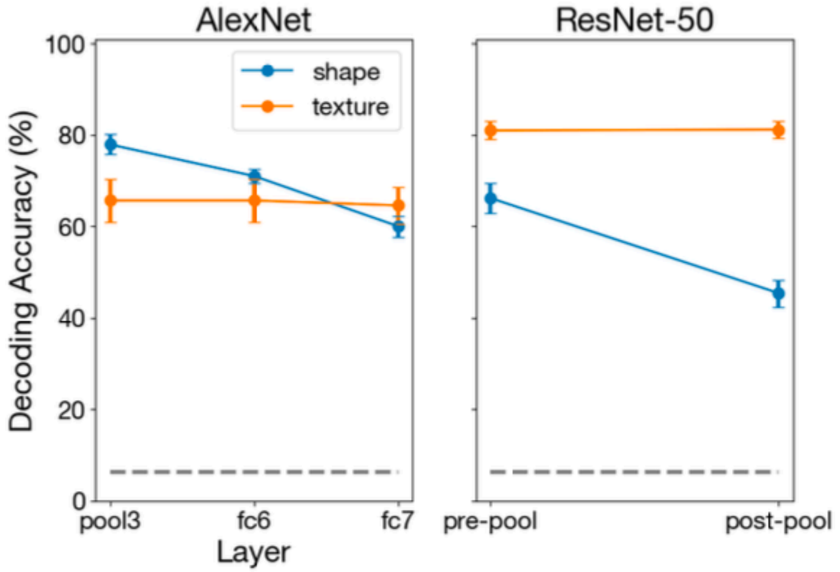


Figure 5: **Both the shape and texture of ambiguous images is decodable from the hidden representations of texture-biased ImageNet-trained models.** Performance of linear classifiers trained to classify shape (blue) or texture (orange) of the GST stimuli given layer activations from frozen ImageNet-trained AlexNet (left) and ResNet-50 (right). Performance is the maximum classification accuracy over the training period (mean $\pm$ SD across 5 splits). Chance is 6.25%. At the final convolutional layer, both models contain substantial shape information, but it decreases subsequently.

What is happening now???

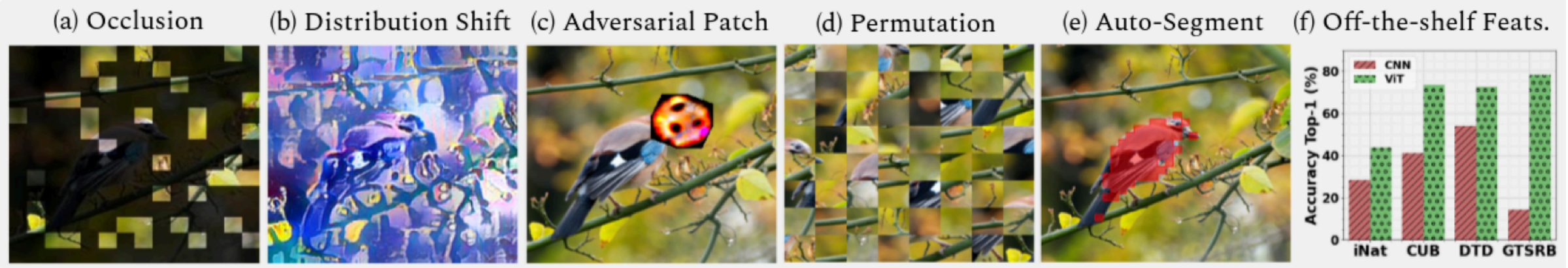


Figure 1: We show intriguing properties of ViT including impressive robustness to (a) severe occlusions, (b) distributional shifts (*e.g.*, stylization to remove texture cues), (c) adversarial perturbations, and (d) patch permutations. Furthermore, our ViT models trained to focus on shape cues can segment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

Conclusion

- IN -> texture bias
- Shape > texture
- Augmentation change bias