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

Abstract

Convolutional Neural Networks (CNNs) are commonly thought to recognise objects by learning increasingly complex representations of object shapes.

We here put these conflicting hypotheses to a quantitative test by evaluating CNNs and human observers on images with a texture-shape cue conflict.

We show that ImageNet-trained CNNs are strongly biased towards recognising textures rather than shapes, which is in stark contrast to human behavioural evidence and reveals fundamentally different classificatior

We then demonstrate that the same standard architecture (ResNet-50) that learns a texture-based representation on ImageNet is able to learn a shape-based representation instead when trained on ‘StylizedImageN

INTRODUCTION

These experiments provide behavioural evidence in favour of the texture hypothesis: A cat with an elephant texture is an elephant to CNNs, and still a cat to humans.

Beyond quantifying existing biases, we subsequently present results for our two other main contributions: changing biases, and discovering emergent benefits of changed biases.

We show that the texture bias in standard CNNs can be overcome and changed towards a shape bias if trained on a suitable dataset.

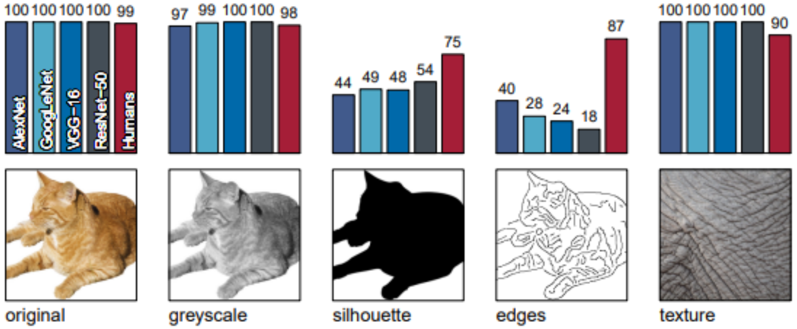
Remarkably, networks with a higher shape bias are inherently more robust to many different image distortions (for some even reaching or surpassing human performance, despite never being trained on any of them

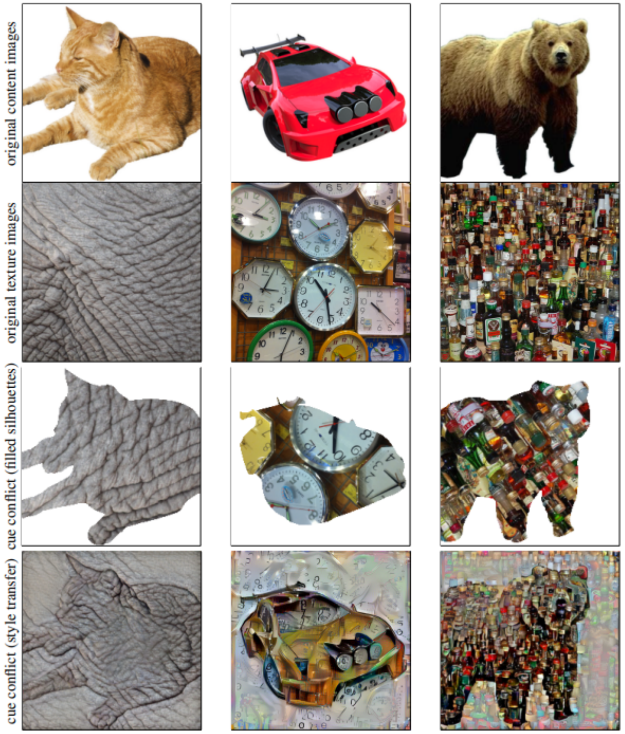
METHODS

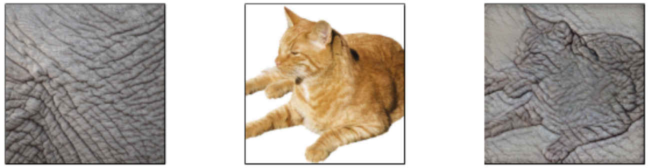
PSYCHOPHYSICAL EXPERIMENTS

Those are the so-called “16-class-ImageNet” categories introduced in Geirhos et al. (2018).

DATA SETS (PSYCHOPHYSICS)




original content images
original texture images
cue conflict (filled silhouettes)
cue conflict (style transfer)



ı) Texture image
81.4% **Indian elephant**
10.3% indri
8.2% black swan

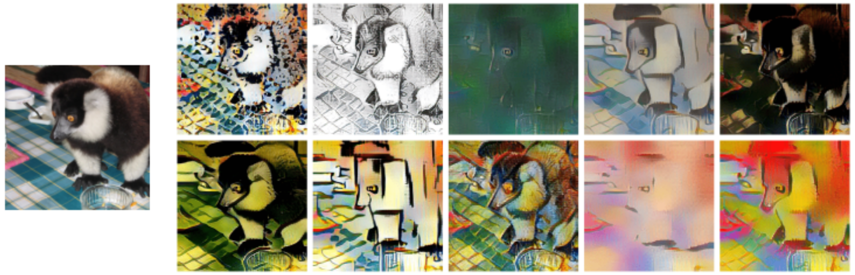
(b) Content image
71.1% **tabby cat**
17.3% grey fox
3.3% Siamese cat

(c) Texture-shape cue conflict
63.9% **Indian elephant**
26.4% indri
9.6% black swan

Briefly, in each trial participants were presented a fixation square for 300 ms, followed by a 200 ms presentation of the stimulus image.
After the stimulus image we presented a full-contrast pink noise mask (1/f spectral shape) for 200 ms to minimise feedback processing in the human visual system and to thereby make the comparison to feedforwar

It is important to note that we only selected object and texture images that were correctly classified by all four networks.
This was made to ensure that our results in the sixth experiment on cue conflicts, which is most decisive in terms of the shape vs texture hypothesis, are fully interpretable.

STYLIZED-IMAGENET



The first five experiments (samples visStarting from ImageNet we constructed a new data set (termed Stylized-ImageNet or SIN) by stripping every single image of its original texture and replacing it with the style of
Secondly, to enable stylizing entire ImageNet, which would take prohibitively long with an iterative approach.

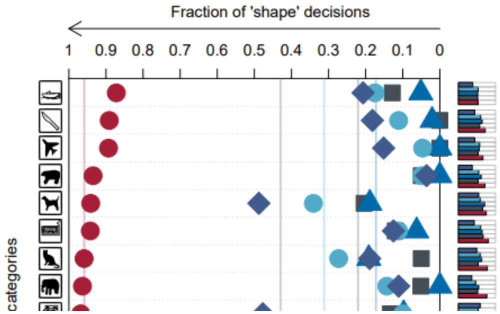
RESULTS

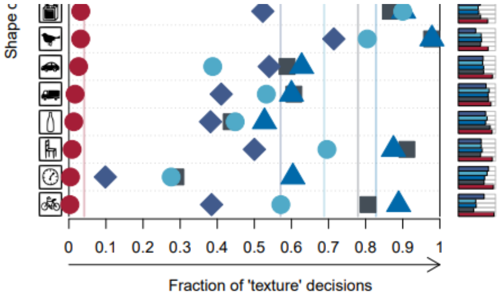
TEXTURE VS SHAPE BIAS IN HUMANS AND IMAGENET-TRAINED CNNS

Almost all object and texture images (Original and Texture data set) were recognised correctly by both CNNs and humans (Figure 2).
Greyscale versions of the objects, which still contain both shape and texture, were recognised equally well.

When object outlines were filled in with black colour to generate a silhouette, CNN recognition accuracies were much lower than human accuracies.
This was even more pronounced for edge stimuli, indicating that human observers cope much better with images that have little to no texture information.

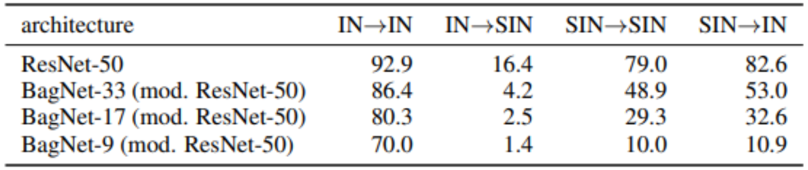
One confound in these experiments is that CNNs tend not to cope well with domain shifts, i.e. the large change in image statistics from natural images (on which the networks have been trained) to sketches (which t
We thus devised a cue conflict experiment that is based on images with a natural statistic but contradicting texture and shape evidence (see Methods).
Participants and CNNs have to classify the images based on the features (shape or texture) that they most rely on.





OVERCOMING THE TEXTURE BIAS OF CNNS

In order to test this hypothesis we train a ResNet-50 on our Stylized-ImageNet (SIN) data set in which we replaced the object-related local texture information with the uninformative style of randomly selected artistic



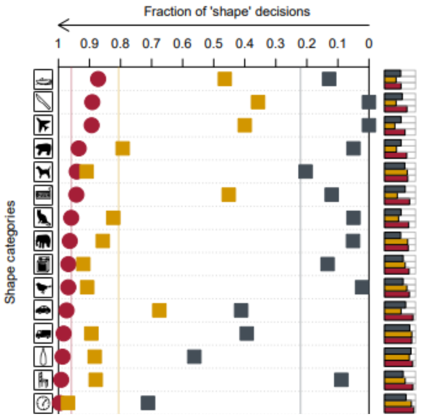
architecture	IN→IN	IN→SIN	SIN→SIN	SIN→IN
ResNet-50	92.9	16.4	79.0	82.6
BagNet-33 (mod. ResNet-50)	86.4	4.2	48.9	53.0
BagNet-17 (mod. ResNet-50)	80.3	2.5	29.3	32.6
BagNet-9 (mod. ResNet-50)	70.0	1.4	10.0	10.9

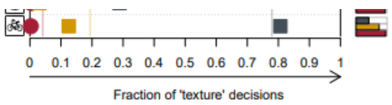
This performance difference indicates that SIN is a much harder task than IN since textures are no longer predictive, but instead a nuisance factor (as desired).
Intriguingly, ImageNet features generalise poorly to SIN (only 16.4% top-5 accuracy); yet features learned on SIN generalise very well to ImageNet (82.6% top-5 accuracy without any fine-tuning).

In order to test whether local texture features are still sufficient to “solve” SIN we evaluate the performance of so-called BagNets.
This precludes BagNets from learning or using any long-range spatial relationships for classification.

This is a clear indication that the SIN data set we propose does actually remove local texture cues, forcing a network to integrate long-range spatial information.

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.





ROBUSTNESS AND ACCURACY OF SHAPE-BASED REPRESENTATIONS

Does the increased shape bias, and thus the shifted representations, also affect the performance or robustness of CNNs?

In addition to the IN- and SIN-trained ResNet-50 architecture we here additionally analyse two joint training schemes:
Training jointly on SIN and IN.
Training jointly on SIN and IN with fine-tuning on IN. We refer to this model as Shape-ResNet.

We then compared these models with a vanilla ResNet-50 on three experiments:

Classification 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**

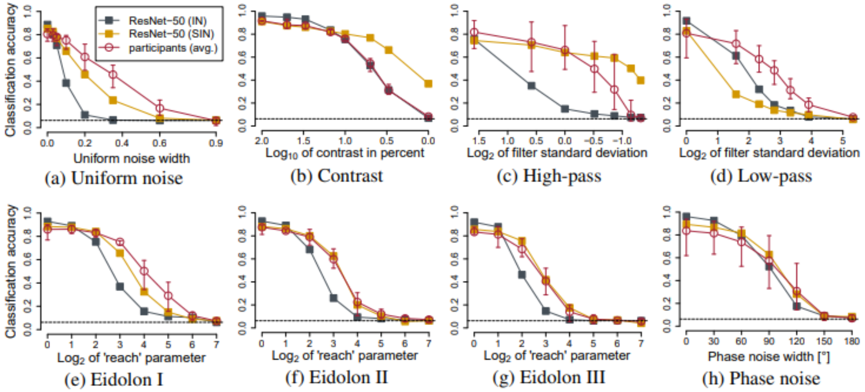
This indicates that SIN may be a useful data augmentation on ImageNet that can improve model performance without any architectural changes.

Transfer learning

This is in line with the intuition that for object detection, a shape-based representation is more beneficial than a texture-based representation, since the ground truth rectangles encompassing an object are by design

Robustness against distortions

We systematically tested how model accuracies degrade if images are distorted by uniform or phase noise, contrast changes, high- and low-pass filtering or eidolon perturbations.

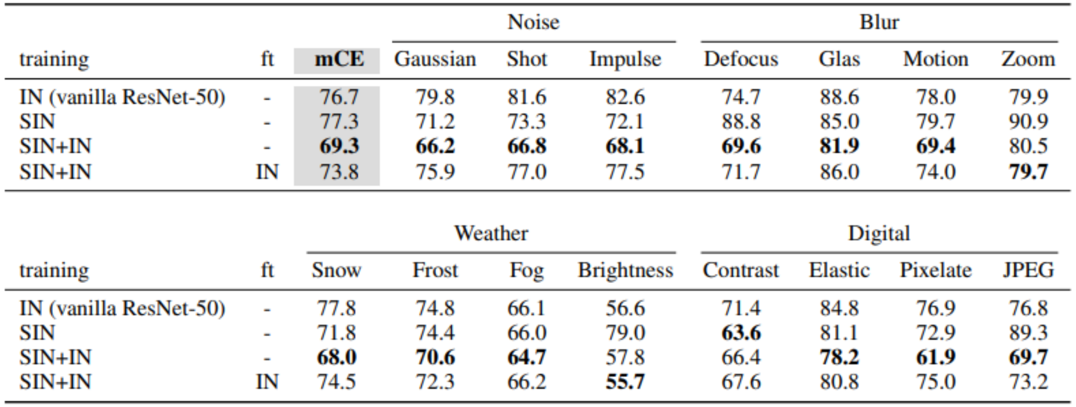


(a) Uniform noise (b) Contrast (c) High-pass (d) Low-pass

(e) Eidolon I (f) Eidolon II (g) Eidolon III (h) Phase noise

While lacking a few percent accuracy on undistorted images, the SIN-trained network outperforms the IN-trained CNN on almost all image manipulations. (Low-pass filtering / blurring is the only distortion type on wh

The SIN-trained ResNet-50 approaches human-level distortion robustness—despite never seeing any of the distortions during training.



training	ft	**mCE**	Noise Gaussian	 Shot	 Impulse	Blur 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 Snow	 Frost	 Fog	 Brightness	Digital 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

Again, none of these corruption types were explicitly part of the training data, reinforcing that incorporating SIN in the training regime improves model robustness in a very general way.

DISCUSSION

As noted in the Introduction, there seems to be a large discrepancy between the common assumption that CNNs use increasingly complex shape features to recognise objects and recent empirical findings which su

In order to explicitly probe this question, we utilised style transfer (Gatys et al., 2016) to generate images with conflicting shape and texture information.

In combination with previous work which showed that changing other major object dimensions such as colour (Geirhos et al., 2018) and object size relative to the context (Eckstein et al., 2017) do not have a strong c

Intriguingly, this offers an explanation for a number of rather disconnected findings: CNNs match texture appearance for humans (Wallis et al., 2017), and their predictive power for neural responses along the human

CNNs can still recognise images with scrambled shapes (Gatys et al., 2017; Brendel & Bethge, 2019), but they have much more difficulties recognising objects with missing texture information (Ballester & de Araujo,

Our hypothesis might also explain why an image segmentation model trained on a database of synthetic texture images transfers to natural images and videos (Ustyuzhaninov et al., 2018).

While both human and machine vision systems achieve similarly high accuracies on standard images (Geirhos et al., 2018), our findings suggest that the underlying classification strategies might actually be very diff

In order to reduce the texture bias of CNNs we introduced Stylized-ImageNet (SIN), a data set that removes local cues through style transfer and thereby forces networks to go beyond texture recognition.

Using this data set, we demonstrated that a ResNet-50 architecture can indeed learn to recognise objects based on object shape, revealing that the texture bias in current CNNs is not by design but induced by Imag

This indicates that standard ImageNet-trained models may be taking a “shortcut” by focusing on local textures, which could be seen as a version of Occam’s razor: If textures are sufficient, why should a CNN learn n

While texture classification may be easier than shape recognition, we found that shape-based features trained on SIN generalise well to natural images.

Our results indicate that a more shape-based representation can be beneficial for recognition tasks that rely on pre-trained ImageNet CNNs.

Furthermore, this finding offers a compellingly simple explanation for the incredible robustness of humans when coping with distortions: a shape-based representation.

CONCLUSION

In summary, we provided evidence that machine recognition today overly relies on object textures rather than global object shapes as commonly assumed.

We demonstrated the advantages of a shape-based representation for robust inference (using our Stylized-ImageNet data set5 to induce such a representation in neural networks).

We envision our findings as well as our openly available model weights, code and behavioral data set (49K trials across 97 observers)6 to achieve three goals:
Firstly, an improved understanding of CNN representations and biases.
Secondly, a step towards more plausible models of human visual object recognition.
Thirdly, a useful starting point for future undertakings where domain knowledge suggests that a shape-based representation may be more beneficial than a texture-based one.