

Vision: Humans vs. Machines

DIFERENCES IN VISION PROCESSING BETWEEN HUMANS AND MACHINES

EMILIO SANSANO

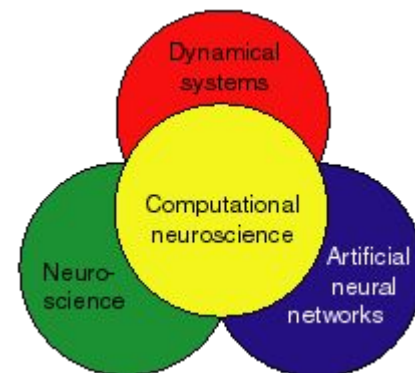
WHAT IS THE POINT OF COMPARING?

ARE DNNs MODELS OF HUMAN NEURAL ARCHITECTURE?

Successful visual perception constitutes a remarkable computational achievement. DNN models of object recognition rival human performance.

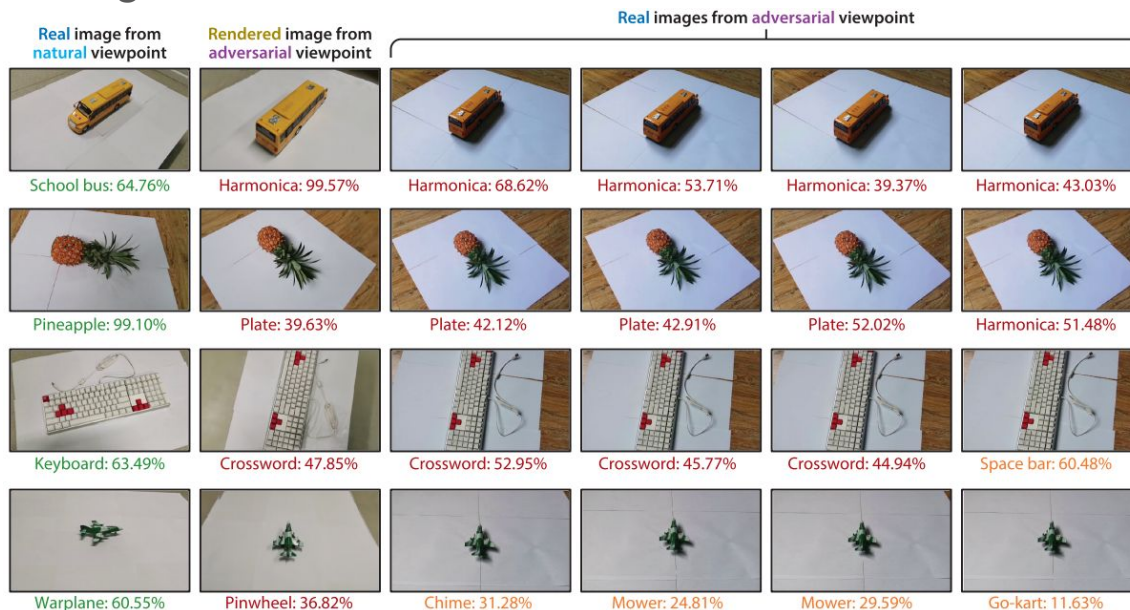
It has been suggested that DNNs might not only be astounding tools for solving computer vision problems, but may also be **good models for the neural architecture of human core object recognition**.

Computational models of vision allow us to specify and test our hypothesized algorithms and computational architectures.



MODELS VS. HUMAN

Find evidence that **models** and **human** observers may be using similar features and processing strategies.









Assess whether the latest computational models show similar input-output behaviour only for tasks for which they are near “ceiling” performance, or whether **their performance degrades similar to human performance** if challenged.

Wichmann F. A. 2023 [Are Deep Neural Networks Adequate Behavioral Models of Human Visual Perception?](#)

OBJECT RECOGNITION

HOW DOES HUMAN OBJECT RECOGNITION WORK?

It is well-known that object shape is the single **most important cue** for human object recognition.

TEST SET			
SIZE CHANGES	1	25"	30"
	2	8.0"	10.0"
	3	24.0"	24.0"
TEXTURE CHANGES	1	blue, cloth	brown, sandpaper
	2	blue, sponge	brown, bubble-pak
	3	blue, wire	brown, beanbag
SHAPE CHANGES	1		
	2		
	3		

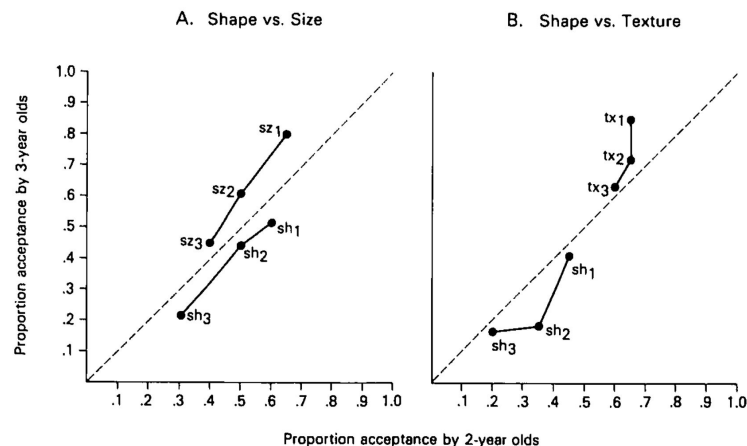


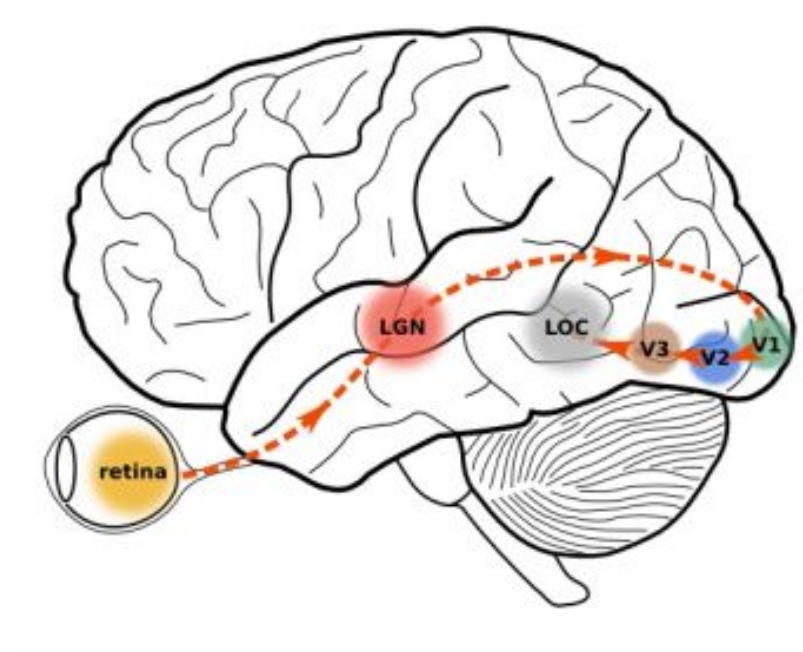
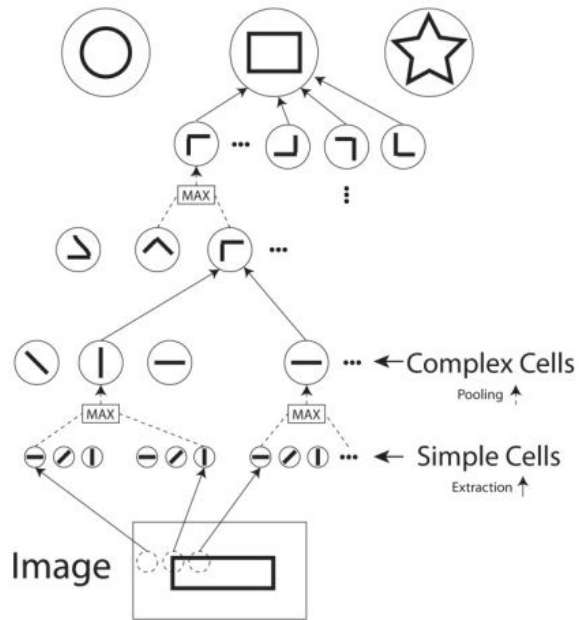
Figure 2. Proportion of same-shape and different-shape choices by 3-year-olds and 2-year-olds for Shape versus Size contrast and Shape versus Texture contrast. Objects chosen denoted by dimension and magnitude of difference from standard.

Differences in shape between a standard and a test object seemed to matter more than differences on either of the other two dimensions.

Landau B. 1988 [The importance of shape in early lexical learning](#)

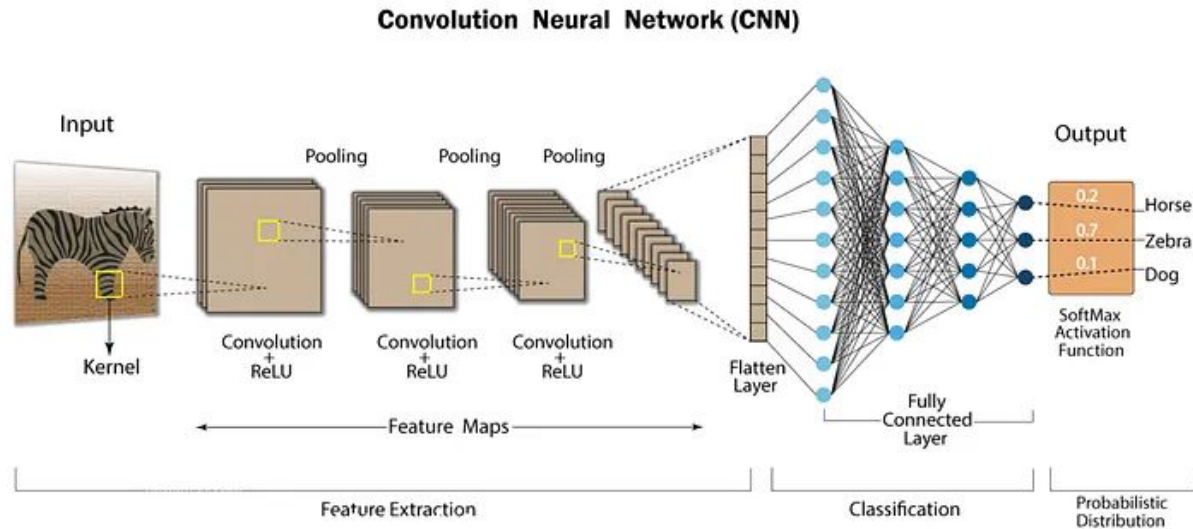
HOW DOES HUMAN OBJECT RECOGNITION WORK?

Hypothesis: Hierarchical processing of vision.



HOW DOES OBJECT RECOGNITION WITH CNNs WORK?

CNNs architecture is structured a series of convolutional layers

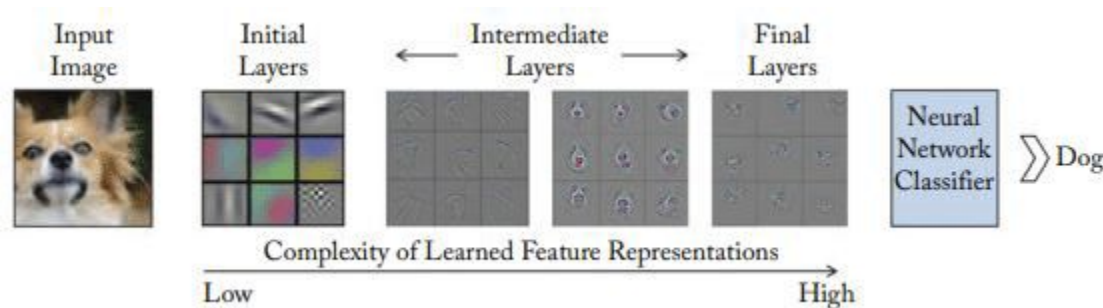


[What is a convolutional neural network?](#)

BUT: there is **no evidence** for backpropagation in the brain It is not biologically plausible.

HOW DOES OBJECT RECOGNITION WITH CNNs WORK?

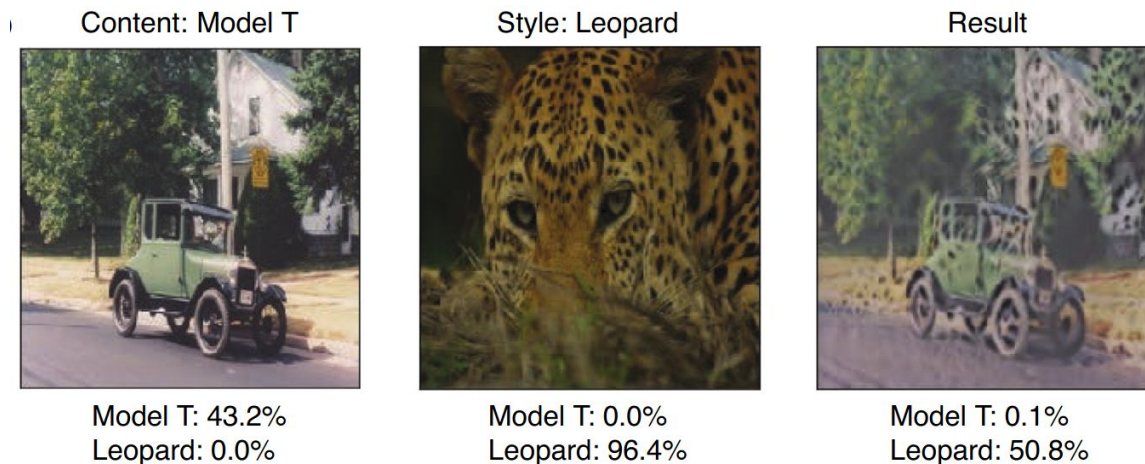
Shape hypothesis: A widely accepted intuition is that CNNs combine low-level features to increasingly complex shapes. High-level units appear to learn representations of shapes occurring in natural images.



Zeiler M. D. 2013 [Visualizing and Understanding Convolutional Networks](#)

HOW DOES OBJECT RECOGNITION WITH CNNs WORK?

Texture hypothesis: CNNs can still classify texturised images perfectly well, even if the global shape structure is completely destroyed. Standard CNNs are bad at recognising object sketches where object shapes are preserved yet all texture cues are missing.



Gatys L. A. 2017 [Texture and art with deep neural networks](#)

EXPERIMENTS WITH CNN-BASED MODELS

A cat with an elephant texture is an elephant to CNNs, and still a cat to humans.



(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

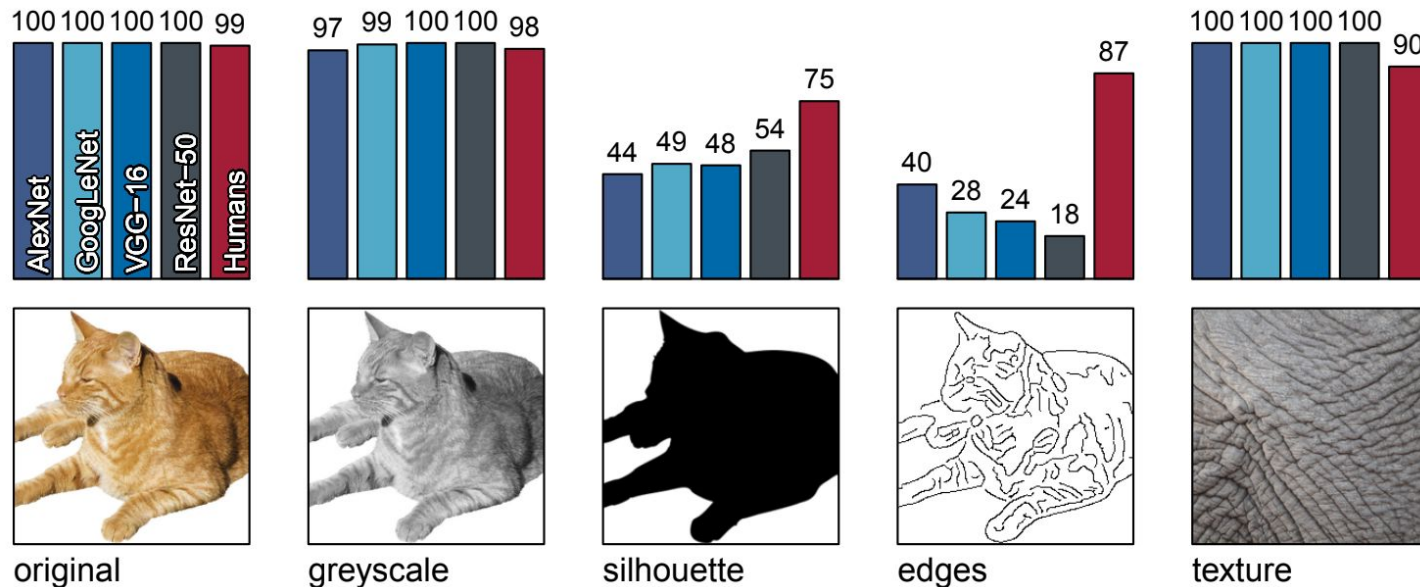
63.9% **Indian elephant**
 26.4% indri
 9.6% black swan

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.

Geirhos R. 2019 [Imagenet-trained CNNs are biased towards texture: increasing shape bias improves accuracy and robustness](#)

EXPERIMENTS WITH CNN-BASED MODELS

Accuracies and example stimuli for five different experiments **without cue conflict**.



Geirhos R. 2019 [Imagenet-trained CNNs are biased towards texture: increasing shape bias improves accuracy and robustness](#)

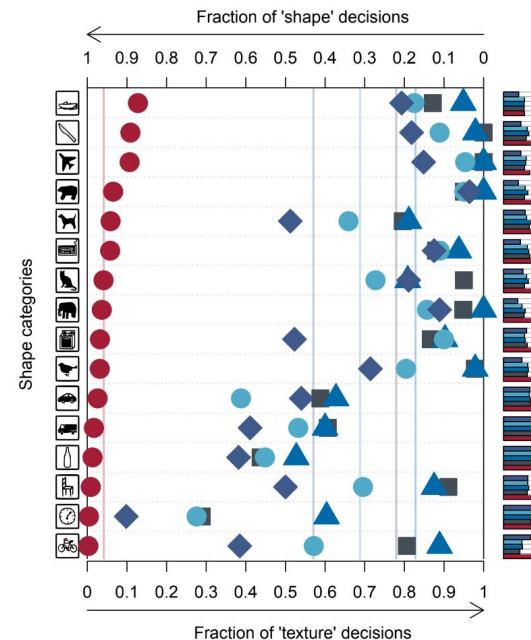
EXPERIMENTS WITH CNN-BASED MODELS

Experiment: strip every image of its original texture and replace it with the style of a randomly selected painting:

- Local texture cues are no longer highly predictive
- The global shape tends to be retained.



Visualisation of Stylized-ImageNet (SIN), created by applying AdaIN style transfer to ImageNet images.



Geirhos R. 2019 [Imagenet-trained CNNs are based towards texture: increasing shape bias improves accuracy and robustness](#)

EXPERIMENTS WITH CNN-BASED MODELS

Making models more “human”

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).

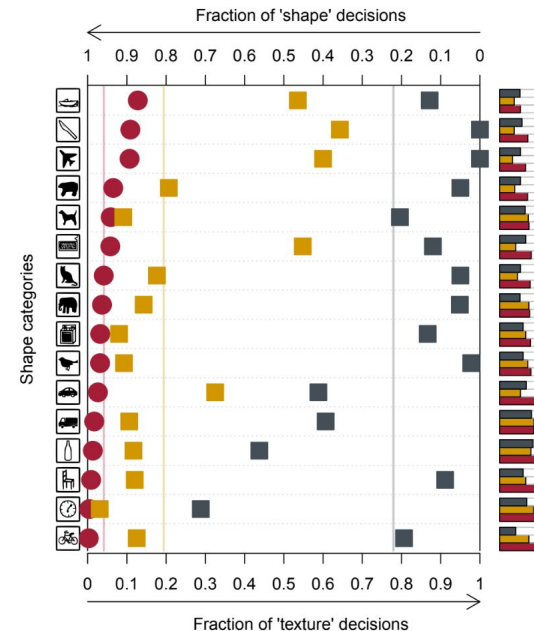


Visualisation of Stylized-ImageNet (SIN), created by applying AdaIN style transfer to ImageNet images.

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
Shape-ResNet	SIN+IN	-	74.59	92.14	74.0	53.8
	SIN+IN	IN	76.72	93.28	75.1	55.2

Accuracy comparison on the ImageNet (IN) validation data set as well as object detection performance (mAP50) on PASCAL VOC 2007 and MS COCO.

Geirhos R. 2019 [Imagenet-trained CNNs are biased towards texture: increasing shape bias improves accuracy and robustness](#)



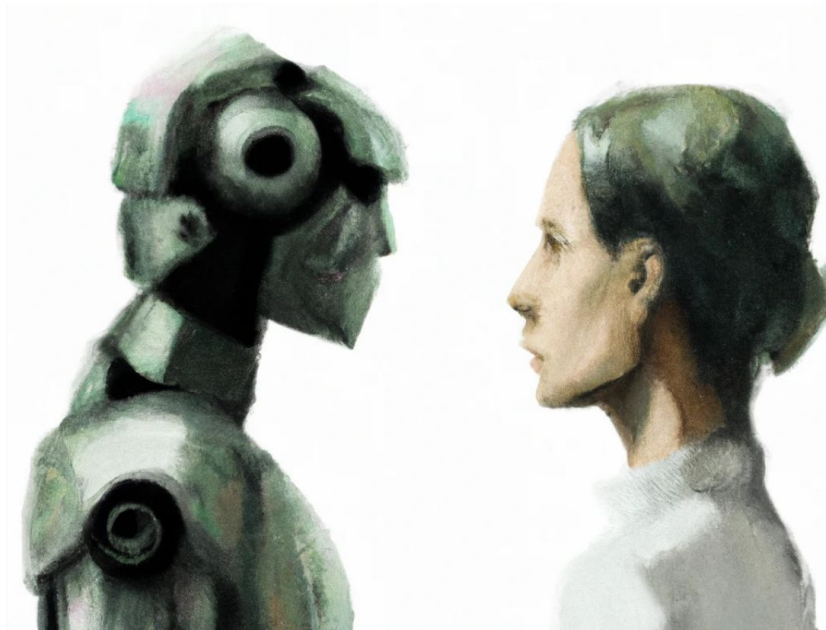
CONCLUSIONS

- Machine recognition today (CNNs) overly **relies on object textures** rather than global object shapes as commonly assumed.
- The texture bias in standard CNNs can be **overcomed** and **changed** towards a shape bias if trained on a suitable data set.
- 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) and reach higher performance on classification and object recognition tasks.

THE GAP BETWEEN MODELS AND HUMANS

ARE TODAY'S MODELS MORE HUMAN?

Are we making progress in closing the gap between human and machine vision?



BENCHMARKING - IID VS. OOD



The gap between human and machine vision has been mainly approximated by comparing benchmark accuracies on IID data.

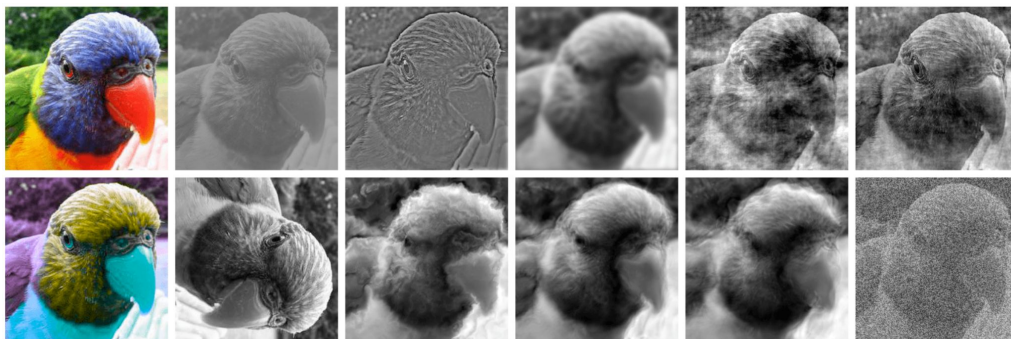
IID: Independent and Identically Distributed

Models are routinely matching and in many cases even outperforming humans on IID data.

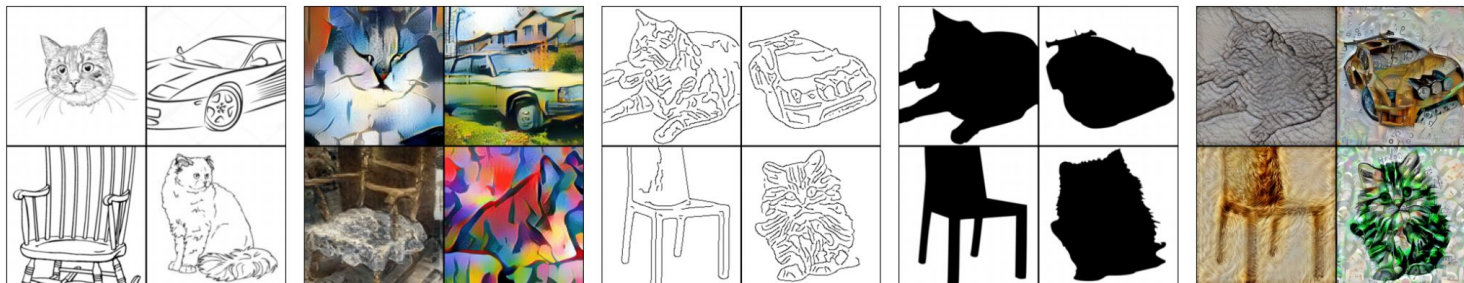
Models systematically **exploit shortcuts** shared between training and test data.

BENCHMARKING - IID VS. OOD

Trend: Shift towards measuring model performance on out-of-distribution (OOD) data rather than IID data alone.



testing models on more challenging test cases where there is still a ground truth category.



MEASURING PERFORMANCE: MODEL VS HUMAN

Many OOD generalisation tests have been proposed: ImageNet-C for corrupted images, ImageNet-Sketch [16] for sketches, Stylized-ImageNet for image style changes, [18] for unfamiliar object poses, and many more [19–29].

Most of these datasets unfortunately lack human comparison data.



Psychophysical experiments:

- Test human observers on a range of OOD datasets.
- Focus: measure distortion robustness.
- Datasets: 17 variations that include changes to image style, texture, and various forms of synthetic noise.
- 90 participants -> 85k trials

Geirhos R. 2021 [Partial success in closing the gap between human and machine vision](#)

MEASURING PERFORMANCE: MODEL VS HUMAN

Comparison along three axis:

- **Objective function:**
 - Supervised
 - Self-supervised
 - Adversarially trained
 - CLIP's joint language-image training ([CLIP](#) = *Contrastive Language-Image Pre-training*)
- **Architecture:**
 - Convolutional
 - Vision transformer
- **Training size:**
 - From 1M to 1000M images

Geirhos R. 2021 [Partial success in closing the gap between human and machine vision](#)

MEASURING PERFORMANCE: METRICS

Metrics:

- **Accuracy difference $A(m)$:** compares the accuracy of a machine (m) to the accuracy of human observers (h) in different OOD tests.

$$A(m) : \mathbb{R} \rightarrow [0, 1], m \mapsto \frac{1}{|D|} \sum_{d \in D} \frac{1}{|H_d|} \sum_{h \in H_d} \frac{1}{|C_d|} \sum_{c \in C_d} (\text{acc}_{d,c}(h) - \text{acc}_{d,c}(m))^2$$

Two models with vastly different image-level decision behaviour might still end up with the same accuracies on each dataset and condition.

Geirhos R. 2021 [Partial success in closing the gap between human and machine vision](#)

MEASURING PERFORMANCE: METRICS

Metrics:

- **Observed consistency** $O(m)$: the fraction of samples for which humans h and a model m get the same sample either both right or both wrong.

$$O(m) : \mathbb{R} \rightarrow [0, 1], m \mapsto \frac{1}{|D|} \sum_{d \in D} \frac{1}{|H_d|} \sum_{h \in H_d} \frac{1}{|C_d|} \sum_{c \in C_d} \frac{1}{|S_{d,c}|} \sum_{s \in S_{d,c}} b_{h,m}(s)$$

where $b_{h,m}(s)$ is one if both a human observer h and m decide either correctly or incorrectly on a given sample s , and zero otherwise.

Geirhos R. 2021 [Partial success in closing the gap between human and machine vision](#)

MEASURING PERFORMANCE: METRICS

Metrics:

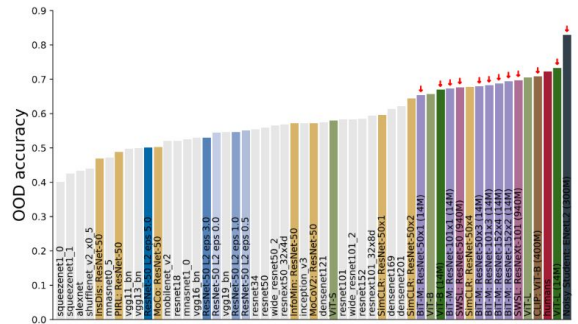
- **Error consistency** $E(m)$: tracks whether there is above-chance consistency. Indicates whether the observed consistency is larger than what could have been expected given two independent binomial decision makers with matched accuracy,

$$E(m) : \mathbb{R} \rightarrow [-1, 1], m \mapsto \frac{1}{|D|} \sum_{d \in D} \frac{1}{|H_d|} \sum_{h \in H_d} \frac{1}{|C_d|} \sum_{c \in C_d} \frac{\left(\frac{1}{|S_{d,c}|} \sum_{s \in S_{d,c}} b_{h,m}(s) \right) - \hat{o}_{h,m}(S_{d,c})}{1 - \hat{o}_{h,m}(S_{d,c})}$$

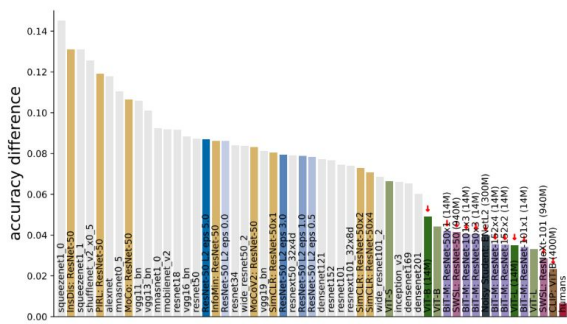
Two decision makers with 95% accuracy each will have at least 90% observed consistency, even if their 5% errors occur on non-overlapping subsets of the test data (intuitively, they both get most images correct and thus observed overlap is high).

Geirhos R. 2021 [Partial success in closing the gap between human and machine vision](#)

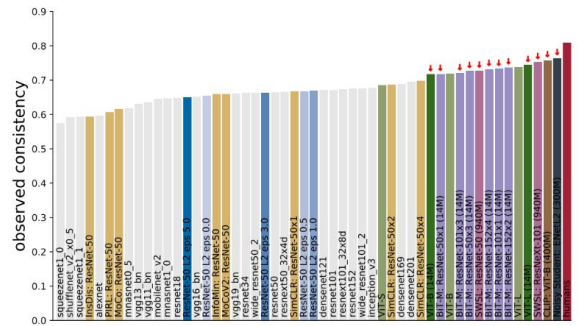
MEASURING PERFORMANCE: RESULTS



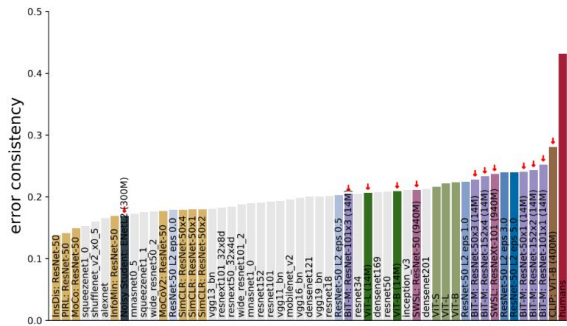
(a) OOD accuracy (higher = better).



(b) Accuracy difference (lower = better).



(c) Observed consistency (higher = better).

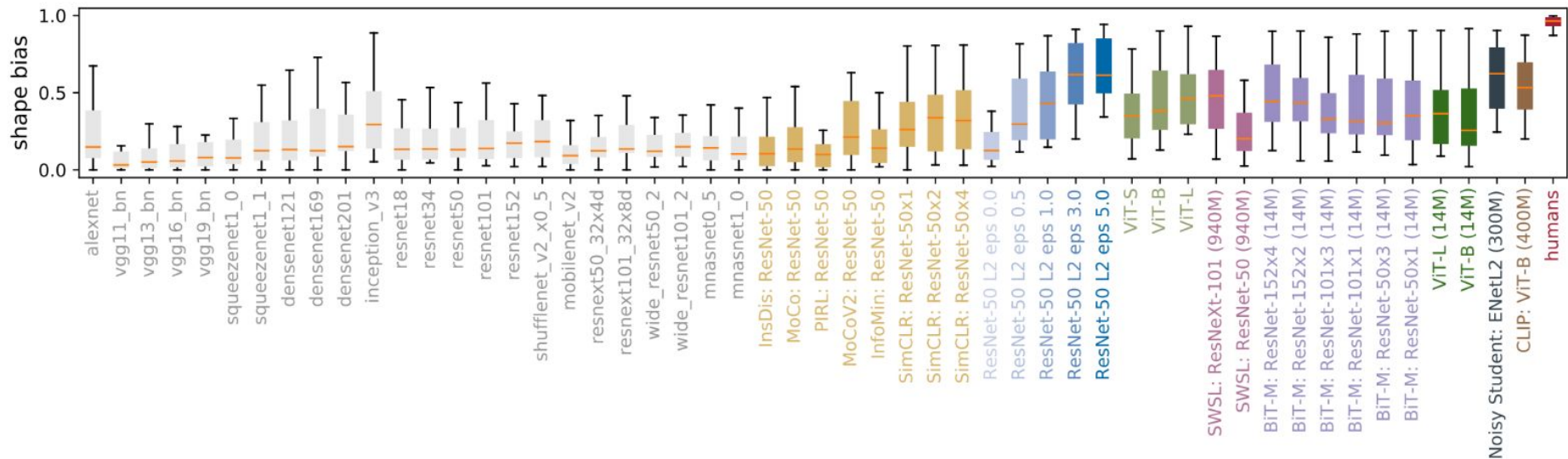


(d) Error consistency (higher = better).

The OOD robustness gap between human and machine vision is closing (top), but an image-level consistency gap remains (bottom). Results compare **humans**, standard supervised CNNs, **self-supervised models**, **adversarially trained models**, **vision transformers**, **noisy student**, **BiT**, **SWSL** and **CLIP**. For convenience, ↓ marks models that are trained on large-scale datasets.

Geirhos R. 2021 [Partial success in closing the gap between human and machine vision](#)

MEASURING PERFORMANCE: RESULTS - SHAPE VS. TEXTURE BIAS

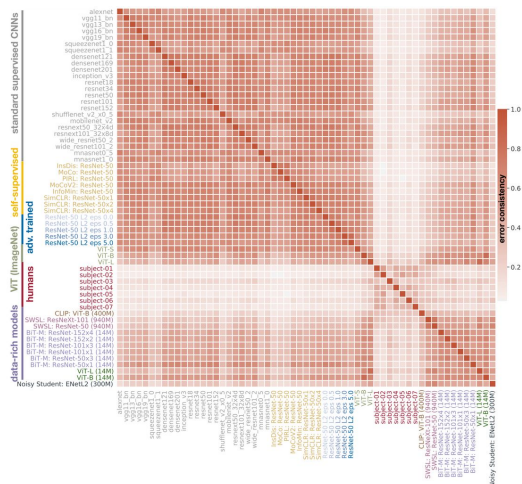


Shape vs. texture biases of different models. While human shape bias is not yet matched, several approaches improve over vanilla CNNs. Box plots show category-dependent distribution of shape / texture biases (shape bias: high values, texture bias: low values).

Geirhos R. 2021 [Partial success in closing the gap between human and machine vision](#)

SUMMARY

- OOD distortion robustness gap between human and machine vision is closing, as the best models now **match or exceed human accuracies**.
- Image-level consistency gap remains, but is narrowing for models trained on large-scale datasets.



To make models more “human” -> simply train with more data -> **disappointing!!!**

Geirhos R. 2021 [Partial success in closing the gap between human and machine vision](#)

OUR CURRENT WORK

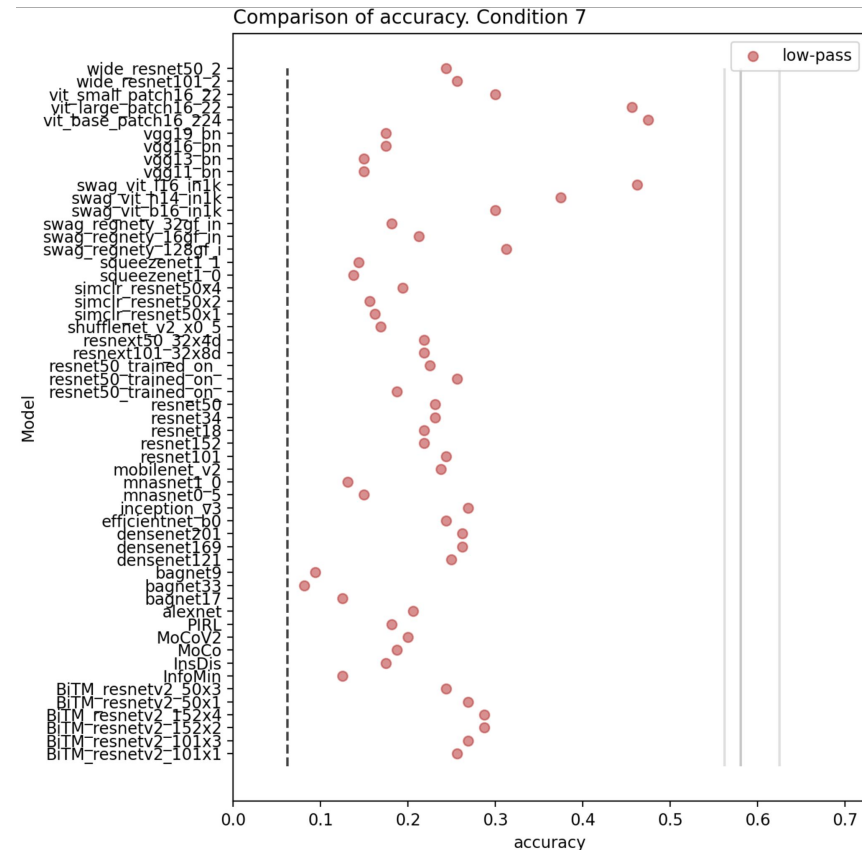
LOW-PASS

Low-pass images are one of the remaining distortion types in which humans are currently still better than all 52 investigated diverse deep neural networks (DNNs) (Geirhos et al. 2021).



“dog” image at blur std = 7 pixels

- Human accuracy: 60%
- DNN accuracy: 10–50%

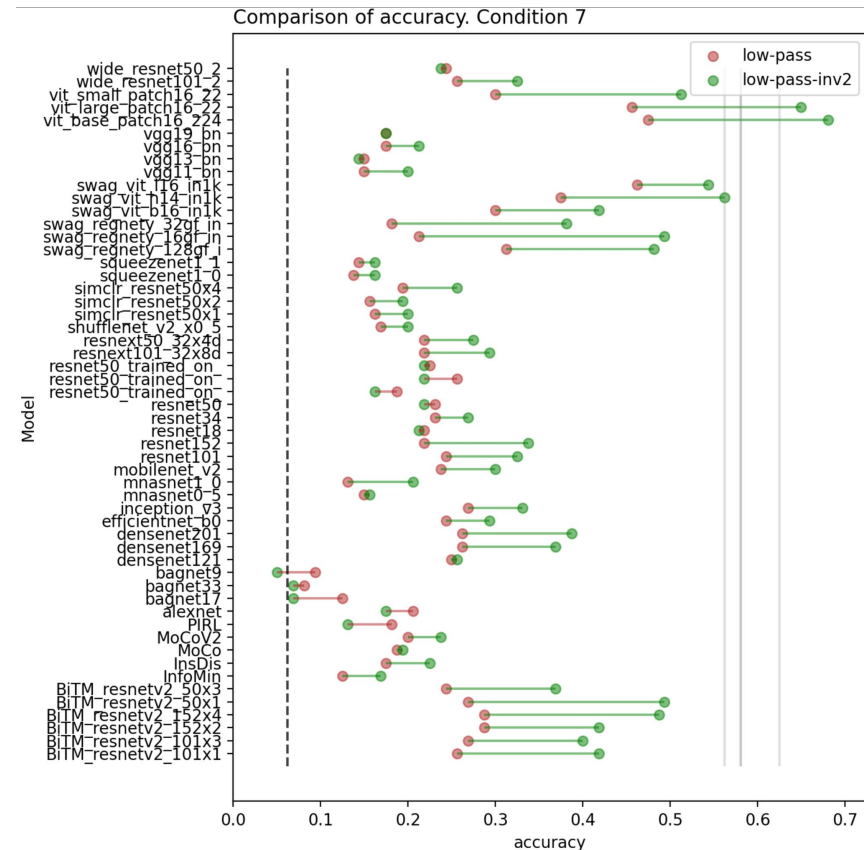


LOW-PASS

Preprocessing with new entropy-based blind deblurring method.

Results are much better for the most “human” models.

Also better results than the s.o.t.a. Blind deblurring methods.



THANKS