Image and AI

ConvNets and humans are not biased towards the same information in images

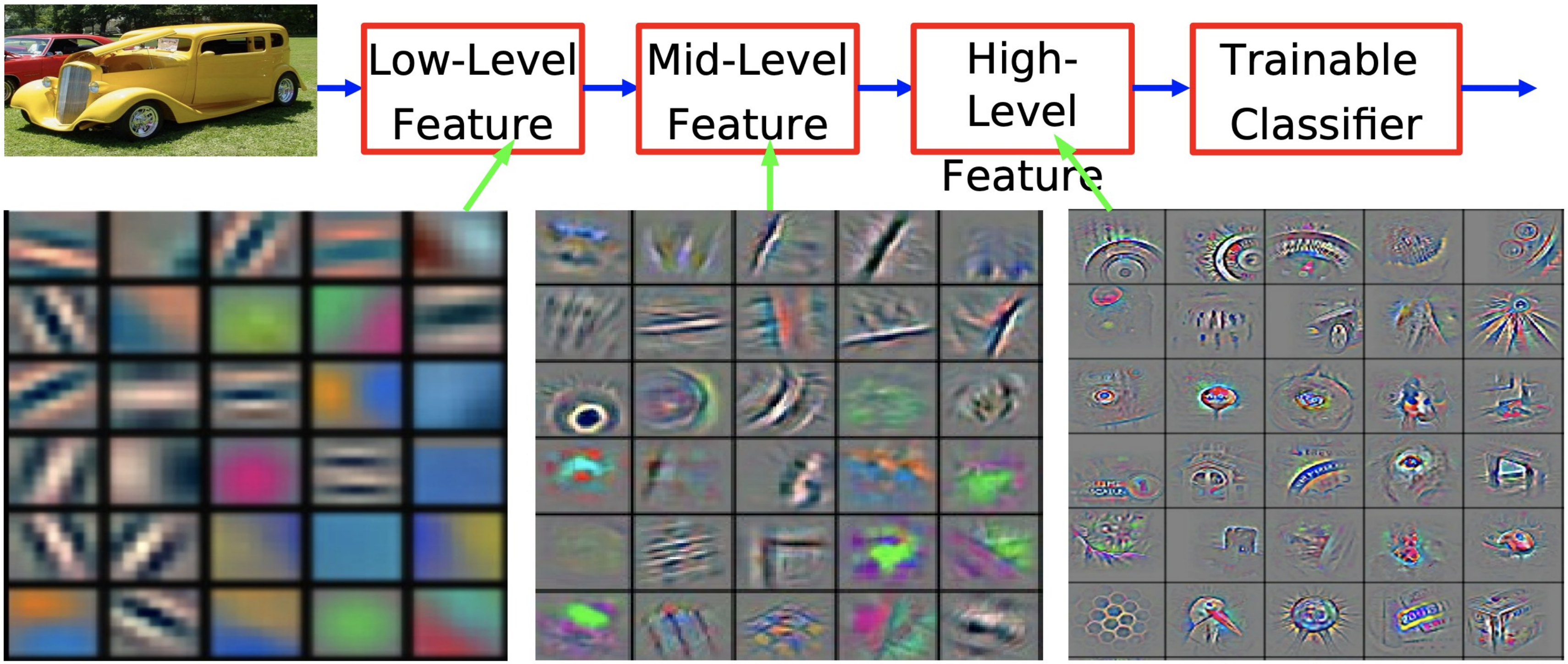
March 4, 2021

MARIE Alban

What is learned by ConvNets
to perform so well on images?

What is learned by ConvNets to perform so well on images?

A common thought…



…from low to high level features, use of global objects shape

Are humans looking at the same kind of features in images?

Not obvious at all...

Are humans looking at the same kind of features in images?

Let’s play a game! **Cat or elephant?**

Are humans looking at the same kind of features in images?

Let's play a game! **Cat or elephant?**

Huh… that's too easy
what's the point of asking this?

Are humans looking at the same kind of features in images?

Let’s play a game! **Cat or elephant?**

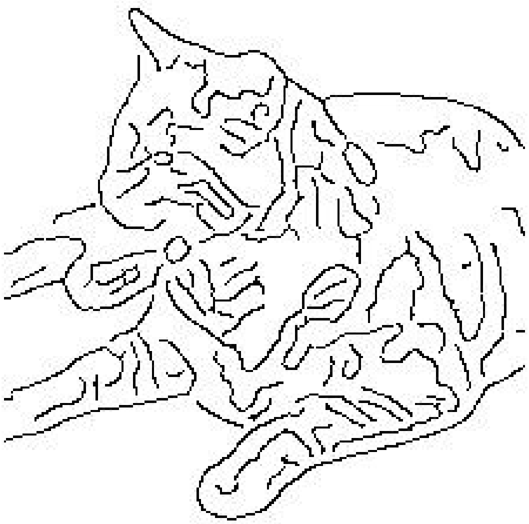
Are humans looking at the same kind of features in images?

Let's play a game! **Cat or elephant?**

Will it ever become hard?

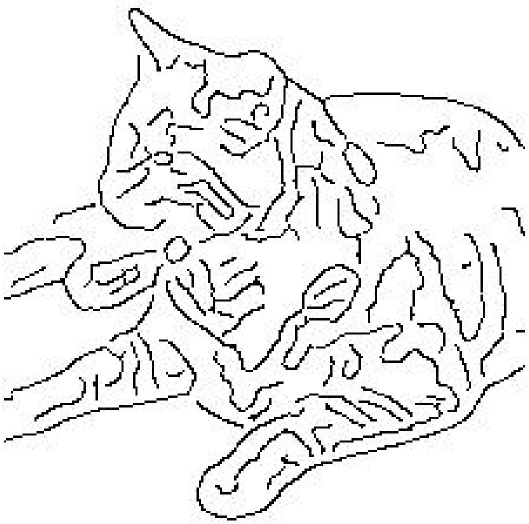
Are humans looking at the same kind of features in images?

Let’s play a game! **Cat or elephant?**



Are humans looking at the same kind of features in images?

Let's play a game! **Cat or elephant?**



A cat I guess ?

Are humans looking at the same kind of features in images?

Let's play a game! **Cat or elephant?**

Are humans looking at the same kind of features in images?

Let's play a game! **Cat or elephant?**

Elephant skin!

Are humans looking at the same kind of features in images?

Let's play a game! **Cat or elephant?**

Are humans looking at the same kind of features in images?

Let's play a game! **Cat or elephant?**

Well.
It depends...

Ok cool. And what opinion ConvNets have on this?

(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

It looks like **ConvNets care about texture to answer**.

Accuracies given by a ResNet-50 classifier trained on ImageNet.

And what about humans?

Does we mostly look at texture too?

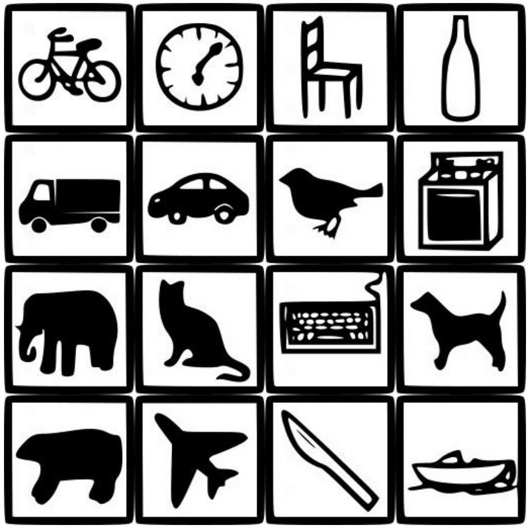
ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness

2018, Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann and Wieland Brendel

→ Comparison between:
- → **ConvNets**: AlexNet / GoogLeNet / VGG-16 / ResNet-50
- → **Humans**

→ 16 classes classification [1] (using WordNet hierarchy)

→ **Five experiments**: original, greyscale, silhouette, edges and texture images



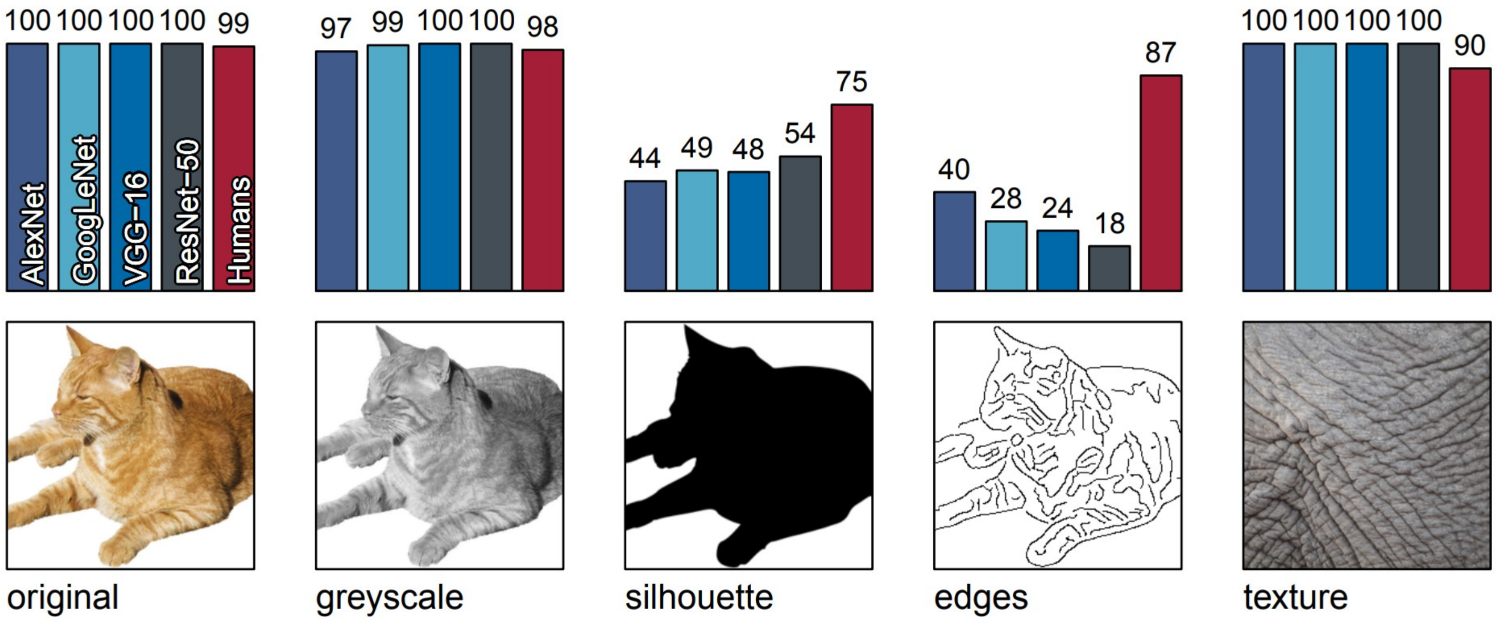
All 16 classes used for this experiment (*)

(*) ConvNets are regular models trained on ImageNet (1000 classes). Each class is mapped to one of the 16 above. See [1] for more details

[1] : Geirhos, R., Temme, C. R. M., Rauber, J., Schütt, H. H., Bethge, M., & Wichmann, F. A. (2018). Generalisation in humans and deep neural networks. arXiv preprint arXiv:1808.08750.

ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness

2018, Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann and Wieland Brendel



Accuracies obtained for humans and ConvNets under 5 different experiments

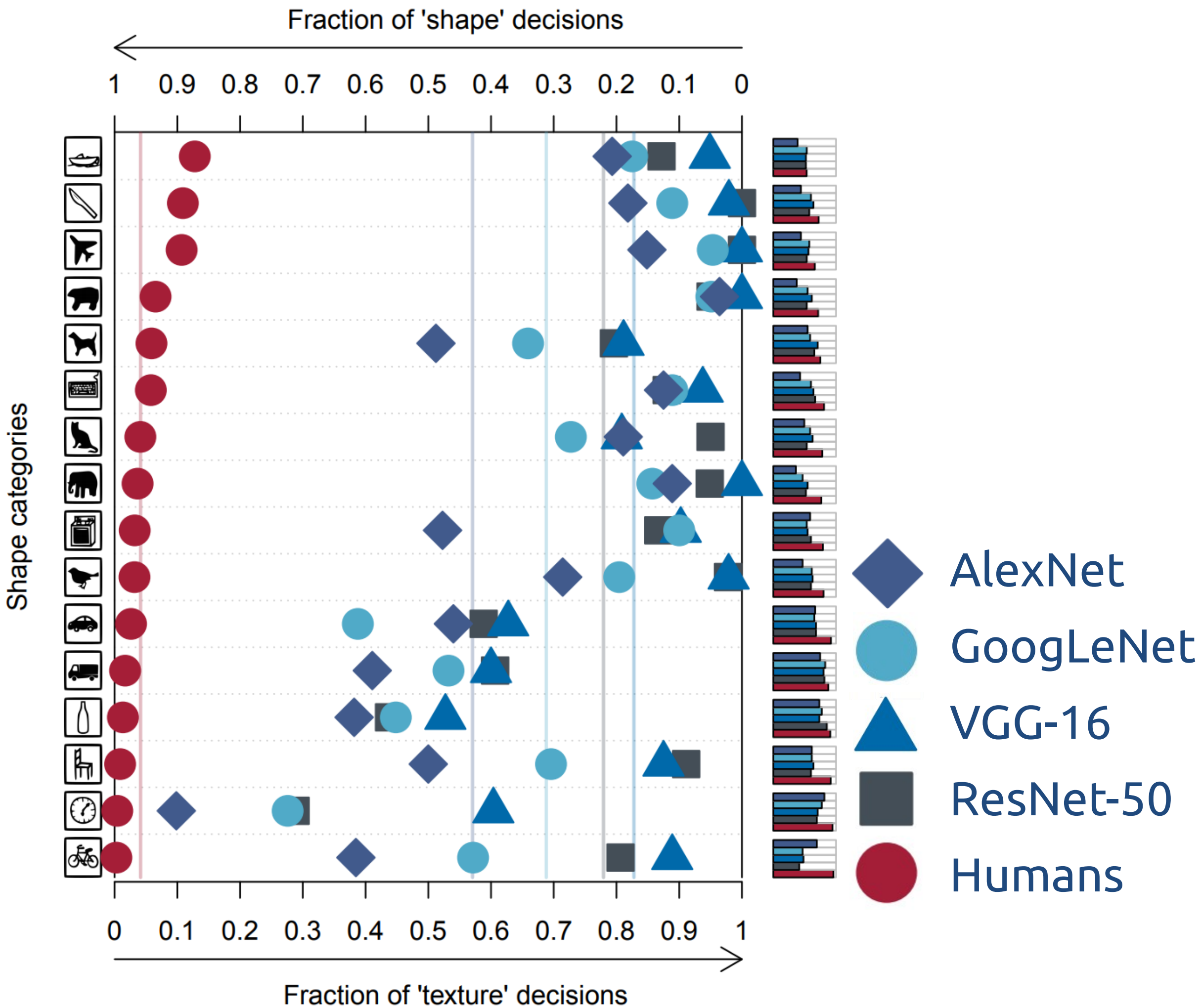
It looks like humans **can answer right without texture**.

For an image obtained with style transfer, there is **no right answer** for classification

However, we can use these images to know **which information** (texture or shape) **is used** by ConvNets and humans

ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness

2018, Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann and Wieland Brendel



Shape vs texture decisions
for humans and ConvNets

Shape decision : cat
Texture decision : elephant

Humans and ConvNets do **NOT** pay attention to the same information

Do we really care? As long at it works…

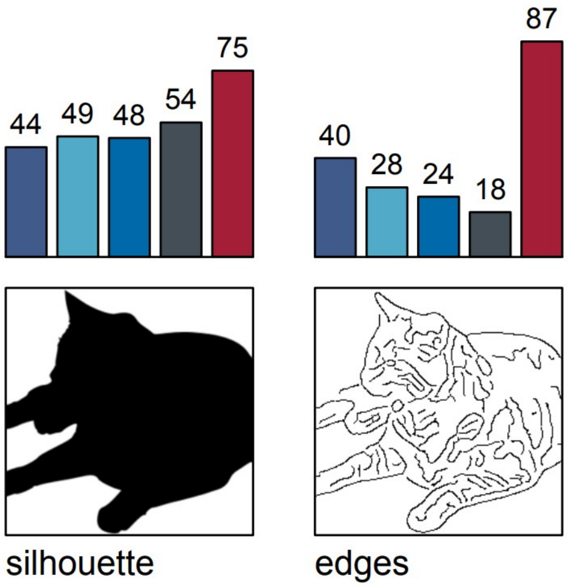
Do we really care? As long at it works...

Well… yes!

→ An **image** is a signal representation **specifically made for humans**
→ Image are almost always adapted to the HVS (*) (*i.e.* though compression)

→ **Humans** still have **more generalisation** ability compared to ConvNets

→ Let's first have similar performances with AI before considering a different approach compared to the HVS



(*) HVS: **H**uman **V**isual **S**ystem

How to change ConvNets texture bais into a shape bais like humans?

→ Style transfer!

→ Texture is now irrelevant, you are forced to use something else to answer right

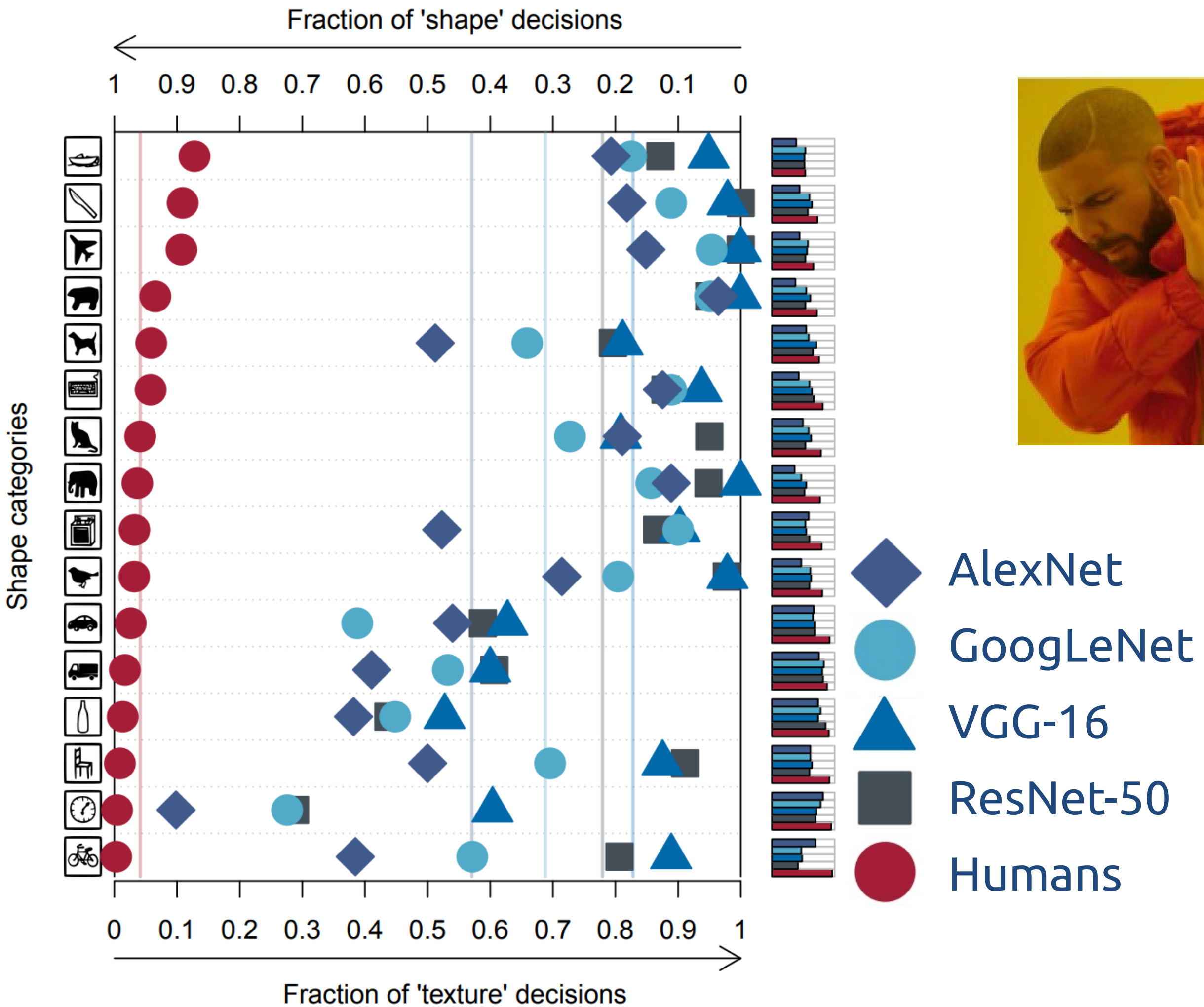
→ We hope to develop a shape bias

Visualisation of Stylized-ImageNet (SIN), created by applying AdaIN [2] style transfer to ImageNet images

[2] : Huang, X., & Belongie, S. (2017). Arbitrary style transfer in real-time with adaptive instance normalization. In Proceedings of the IEEE International Conference on Computer Vision (pp. 1501-1510).

ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness

2018, Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann and Wieland Brendel

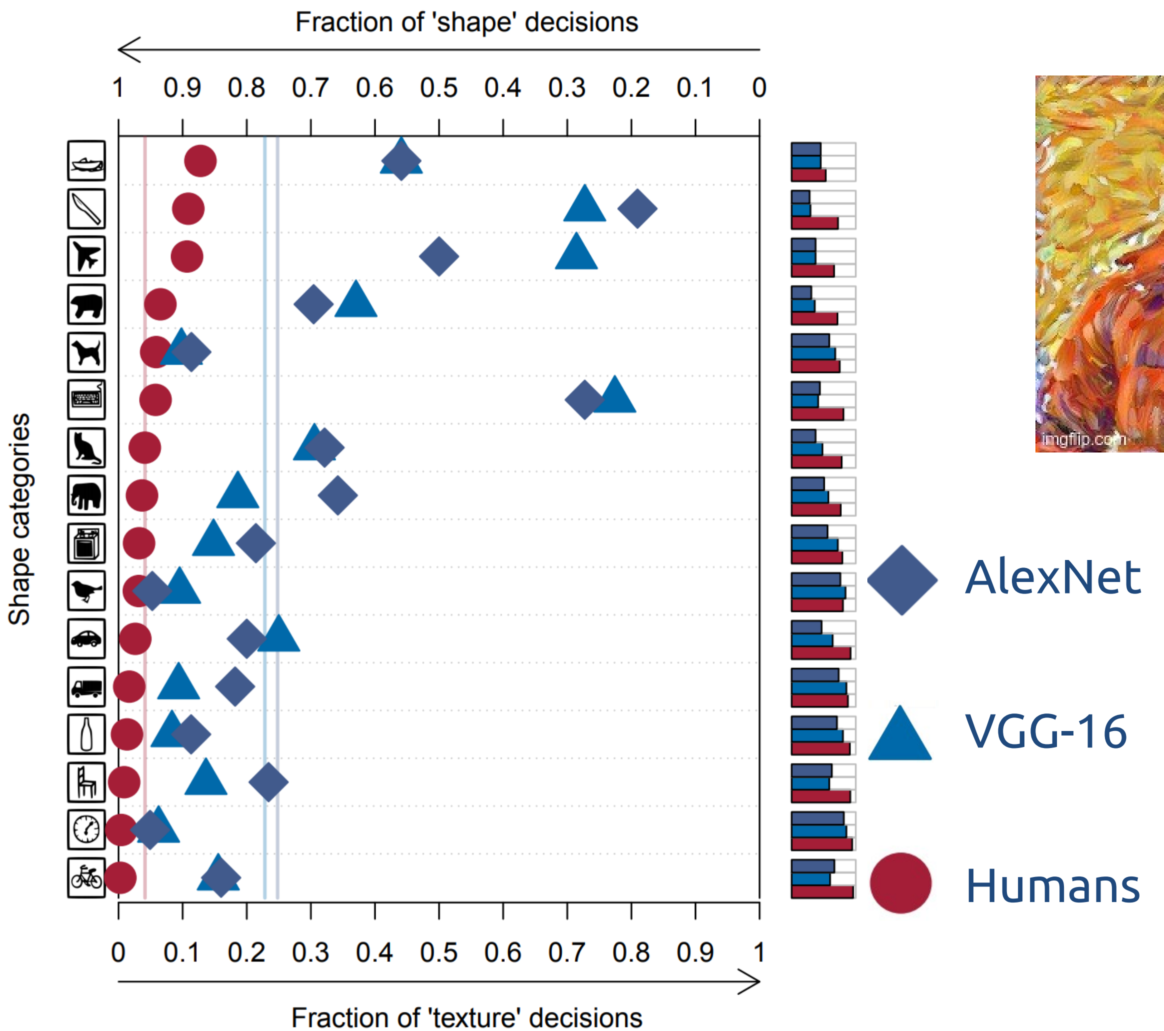


Shape vs texture decisions for humans and ConvNets

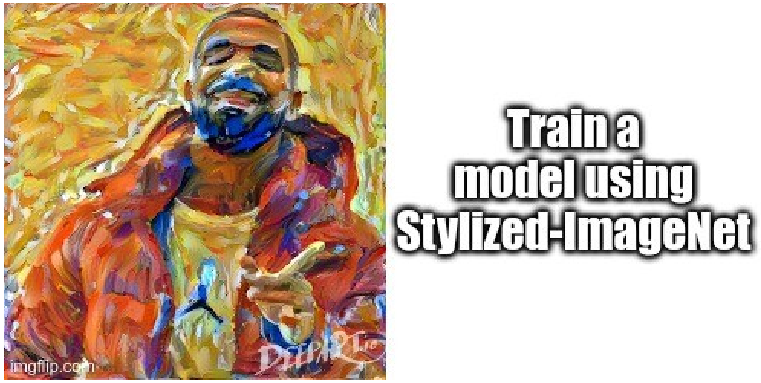
Train a model using ImageNet

ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness

2018, Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann and Wieland Brendel

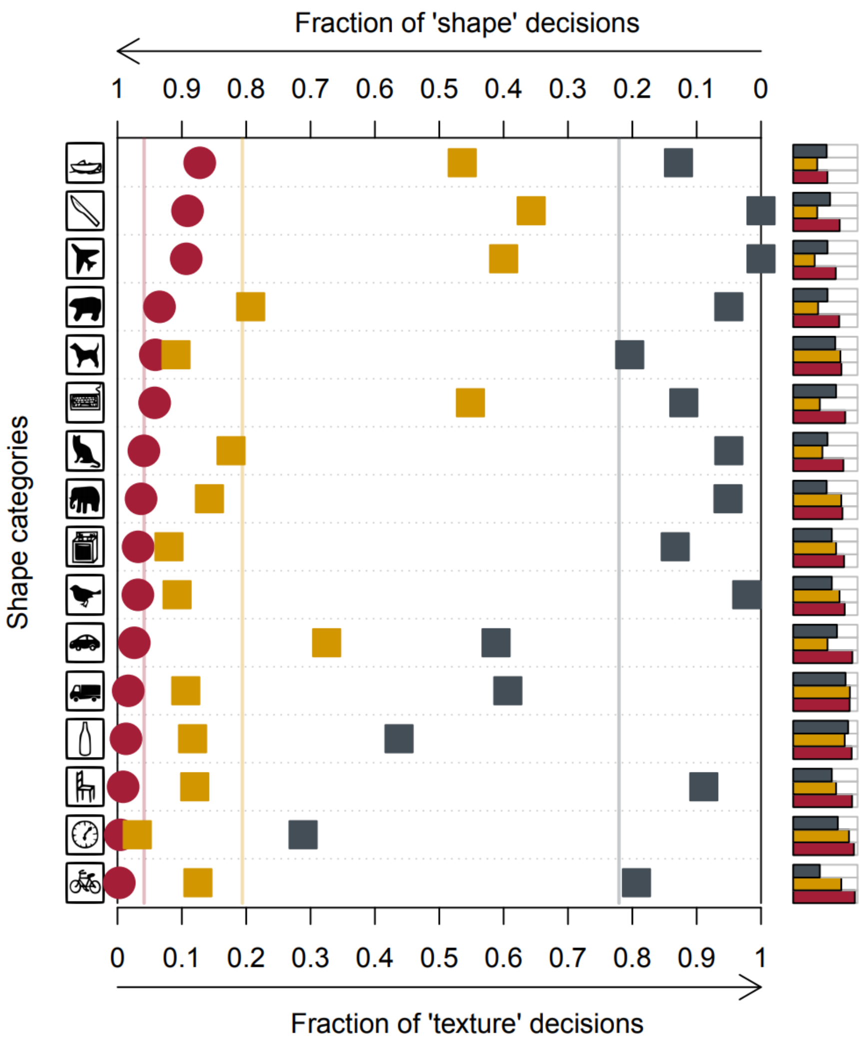


Shape vs texture decisions
for humans and ConvNets



ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness

2018, Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann and Wieland Brendel



ResNet (IN)

ResNet (SIN)

Humans

Shape vs texture decisions
for humans and ConvNets

Train a model using Stylized-ImageNet

Another proof that ConvNets mostly relies on texture

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

Comparison between models trained/evaluated
on ImageNet (IN) and Stylized ImageNet (SIN)

scores are top-5 accuracy on the validation set
(*) train data → test data
(**) BagNet-**33** stands for a model with a maximum receptive field of **33** by **33** pixels

ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness

2018, Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann and Wieland Brendel

Another proof that ConvNets mostly relies on texture

SIN can't be solved with texture features only

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

Comparison between models trained/evaluated
on ImageNet (IN) and Stylized ImageNet (SIN)

scores are top-5 accuracy on the validation set
(*) train data → test data
(**) BagNet-**33** stands for a model with a maximum receptive field of **33** by **33** pixels

ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness

2018, Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann and Wieland Brendel

Another proof that ConvNets mostly relies on texture

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

Comparison between models trained/evaluated on ImageNet (IN) and Stylized ImageNet (SIN)

Smaller receptive field → cannot extract global shapes → lower accuracies

scores are top-5 accuracy on the validation set
(*) train data → test data
(**) BagNet-**33** stands for a model with a maximum receptive field of **33** by **33** pixels

Stylized-ImageNet can be perceived as data augmentation to neglect texture bias

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

Benefits of Stylized-ImageNet (SIN) for classification and object detection

(*) Classifiers (*e.g.* Shape-ResNet) are used as backbone features for Faster R-CNN [3]

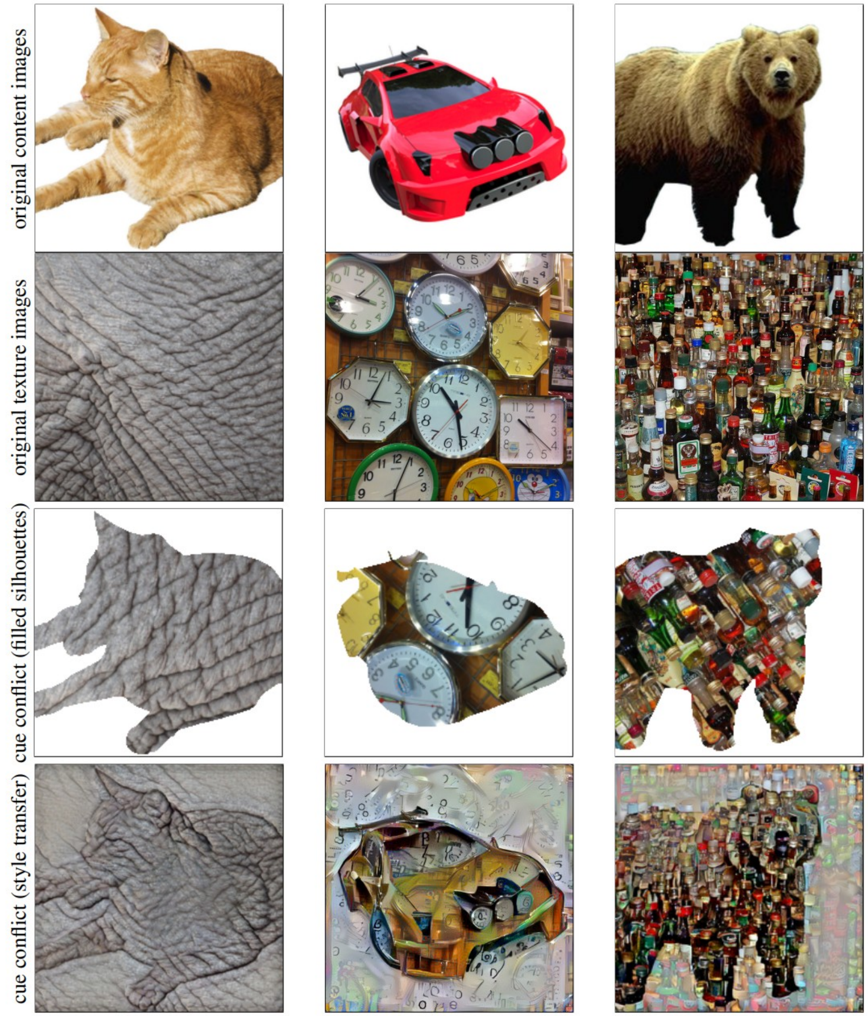
[3] : Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. arXiv preprint arXiv:1506.01497.

That's basically it.

Thank you for your attention!

ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness

2018, Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann and Wieland Brendel



Hack used to obtain texture of an object without texture (*e.g.* glass bottle)