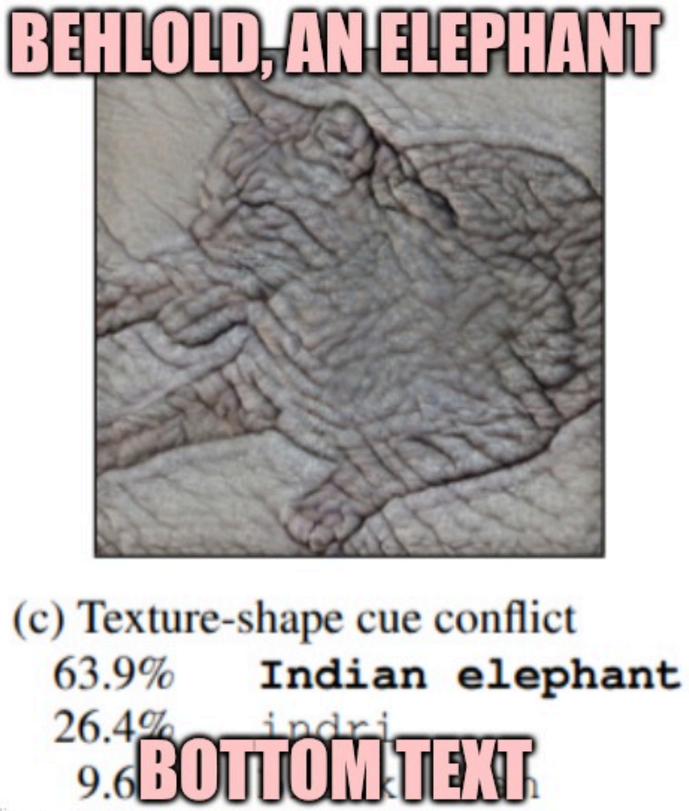
ImageNet-trained CNNs are not biased towards texture: Revisiting feature reliance through controlled suppression

Tom Burgert[1,2], Oliver Stoll[1,2], Paolo Rota[3], Begüm Demir[1,2]
BIFOLD[1], TU Berlin[2], University of Trento[3]

Presented by
Dawid Płudowski

December 1st, 2025

Small teaser



R1: "what is preventing this work from being criticized in the future and replaced by a different methodology that arrives at a different conclusion?" (nothing lol)

Cognitive science context

(so we mention human-CV aligment time by time in this presentation)

A few words about human-CV alignment:

- ***AI does not perceive vision as humans do***

Cognitive science context

(so we mention human-CV aligment time by time in this presentation)

A few words about human-CV alignment:

- ***AI does not perceive vision as humans do***
- ***AI is prone to being manipulated by corruption, biases, adversarial etc., but humans are (generally) not***

Cognitive science context

(so we mention human-CV aligment time by time in this presentation)

A few words about human-CV alignment:

- ***AI does not perceive vision as humans do***
- ***AI is prone to being manipulated by corruption, biases, adversarial etc., but humans are (generally) not***
- ***If we can mimic human perception in AI, we hope the AI will be more robust***

Agenda

- *why?*
- *what?*
- *perspectives*

PART I

Previous works, i.e. what is criticized

Why?

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, Wieland Brendel

ICLR 2019 Conference Blind Submission

Recommendation: Accept (Oral)

Confidence: 5: The area chair is absolutely certain

Rating: 7: Good paper, accept

Confidence: 4: The reviewer is confident but not absolutely certain that the evaluation is correct

Rating: 8: Top 50% of accepted papers, clear accept

Confidence: 4: The reviewer is confident but not absolutely certain that the evaluation is correct

Rating: 8: Top 50% of accepted papers, clear accept

Confidence: 4: The reviewer is confident but not absolutely certain that the evaluation is correct

Cytowane przez 3781

Why?

Robert Geirhos

DeepMind

CV vs human

Why?

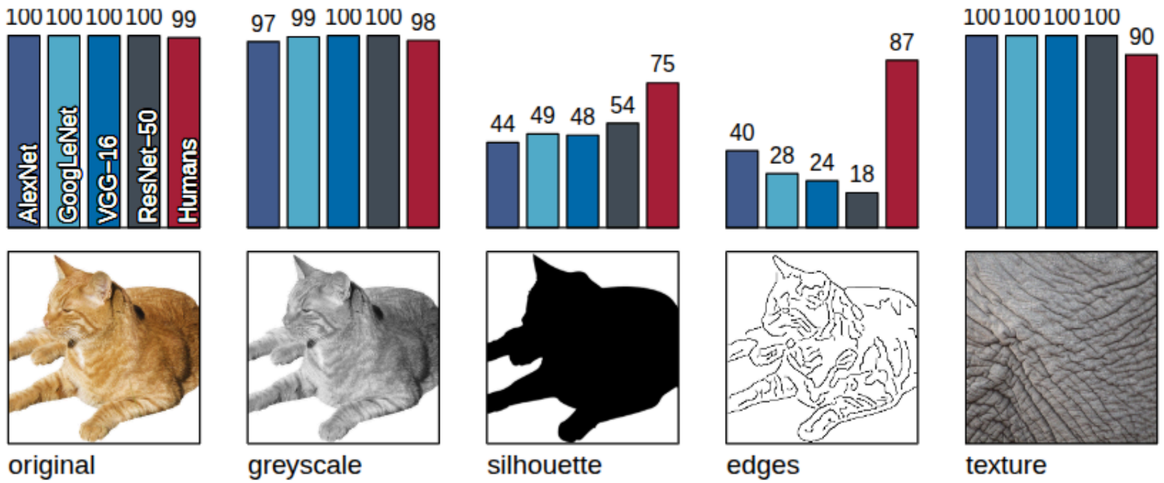
Robert Geirhos
DeepMind

CV vs human

Et.al

Why?

What was done?

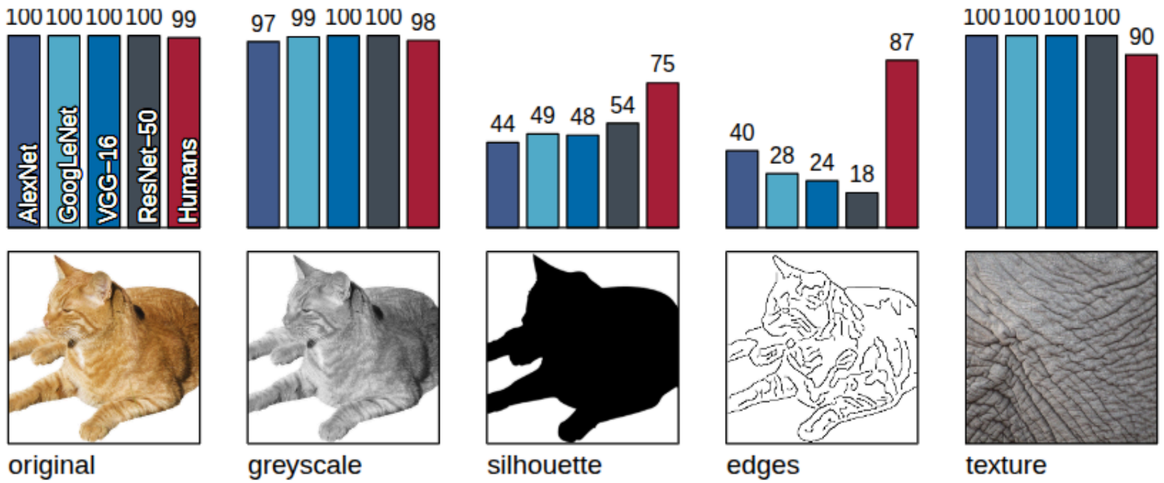


Use models trained on ImageNet and:

- *test them in different scenarios (in figure)*

Why?

What was done?

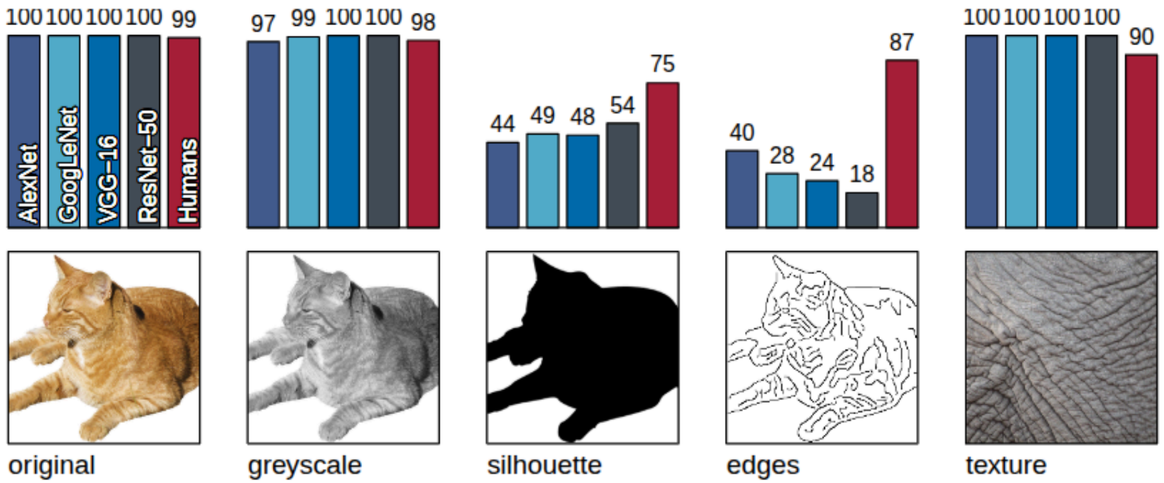


Use models trained on ImageNet and:

- *test them in different scenarios (in figure)*
- *compare them to humans*

Why?

What was done?



Use models trained on ImageNet and:

- *test them in different scenarios (in figure)*
- *compare them to humans*
- *...but this is not realistic case!*

Why?

What was done?

(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

Use models trained on ImageNet and:

- ***style transfer the texture of object X to object Y***

Why?

What was done?

(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

Use models trained on ImageNet and:

- ***style transfer the texture of object X to object Y***
- ***test if models will classify it as object X or object Y***

Why?

What was done?

ImageNet after sylization is conined as StylizedImageNet (SIN)

(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

Use models trained on ImageNet and:

- ***style transfer the texture of object X to object Y***
- ***test if models will classify it as object X or object Y***
- ***compre to humans***

AdaIN style transfer (Huang & Belongie, 2017)

CNN models: AlexNet (Krizhevsky et al., 2012), GoogLeNet (Szegedy et al., 2015), VGG-16 (Simonyan & Zisserman, 2015) and ResNet-50 (He et al., 2015)

Why?

What was done?

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**

Using style transfer as an augmentation to the training routine results in new SoTA solutions - models are forced to ignore texture and focus on the shape only.

Why?

What are limitations?

- *Images are more complex than only shape + texture*

Why?

What are limitations?

- *Images are more complex than only shape + texture*
- *Doesn't style transfer remove the shape cue?*

Why?

What are limitations?

- *Images are more complex than only shape + texture*
- *Doesn’t style transfer remove the shape cue?*
- *How can we distinguish between texture suppression and local shape suppression?*

Why?

What are limitations?

- *Images are more complex than only shape + texture*
- *Doesn’t style transfer remove the shape cue?*
- *How can we distinguish between texture suppression and local shape suppression?*
- *The background is also affected*

Why?

What are limitations?

- *Images are more complex than only shape + texture*
- *Doesn’t style transfer remove the shape cue?*
- *How can we distinguish between texture suppression and local shape suppression?*
- *The background is also affected*
- *Establishing SoTA using style transfer as augmentation doesn’t prove anything*

Why?

What are limitations?

- *Images are more complex than only shape + texture*
- *Doesn't style transfer remove the shape cue?*
- *How can we distinguish between texture suppression and local shape suppression?*
- *The background is also affected*
- *Establishing SoTA using style transfer as augmentation doesn't prove anything*
- *Humans were asked to decide based on shapes - so how can we tell they based their decisions on shape without this stimulus?*

But generally, it is a really good paper!

Why?

A subarea of research was established

(plenty of A rated conference materials)*

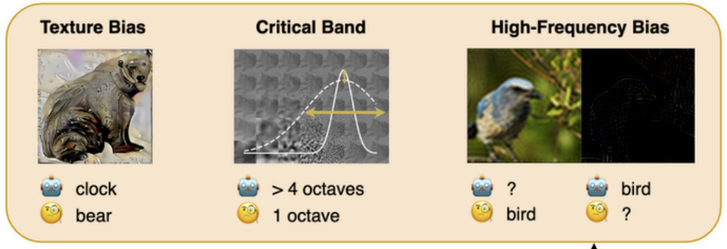
Can Biases in ImageNet Models Explain Generalization?

Improving Robustness to Texture Bias via Shape-focused Augmentation

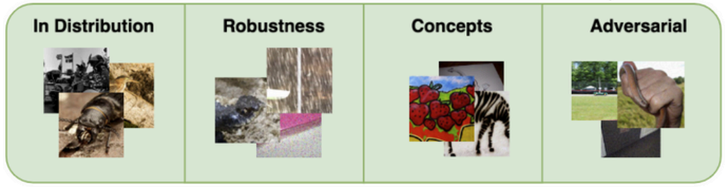
DOES ENHANCED SHAPE BIAS IMPROVE NEURAL NETWORK ROBUSTNESS TO COMMON CORRUPTIONS?

Shape Bias and Robustness Evaluation via Cue Decomposition for Image Classification and Segmentation

The Origins and Prevalence of Texture Bias in Convolutional Neural Networks



Can these model **BIASES** explain **GENERALIZATION?**



(a) Original Image (b) Shape-focused augmentation

PART II

What was done in presented article?

Remember that CV vs human context is still vital

What?

ImageNet-trained CNNs are not biased towards texture: Revisiting feature reliance through controlled suppression

Tom Burgert, Oliver Stoll, Paolo Rota, Begüm Demir

Paper Decision

Decision by Program Chairs

Decision: Accept (oral)

NeurIPS 2025 oral

Rating: 5: Accep
reproducibility, a
Confidence: 3: Y

Rating: 6: Strong A
unaddressed ethica
Confidence: 4: You

Rating: 5: Accept: T
reproducibility, and
Confidence: 3: You

What?

Tom Burgert
BIFOLD

PhD student,
2nd/3rt-ish year

Oliver Stoll
BIFOLD

digital twins

Paolo Rota
Uni@Trento

curriculum leanring

Begüm Demir
BIFOLD

remote sensing,
PhD supervisor

What?

What is the core contribution?

Revisiting types of features:

Shape (local/global)

Path shuffle

Colour

Channel shuffle

Texture

Bilateral filter

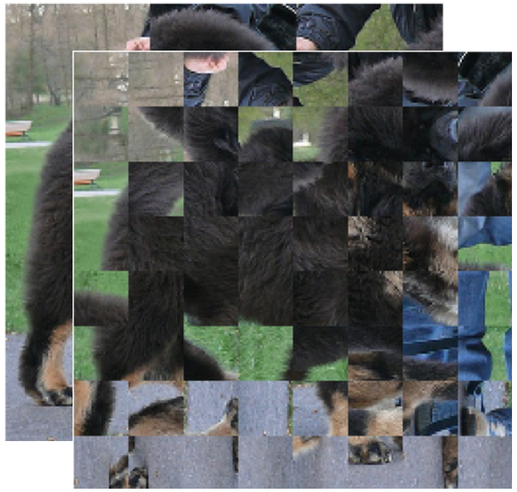
What?

What is the core contribution?

Revisiting types of features:

(alternate transformations)

Shape (local/global)



Path rotation

Colour

Greyscale

Texture

Gausian blur

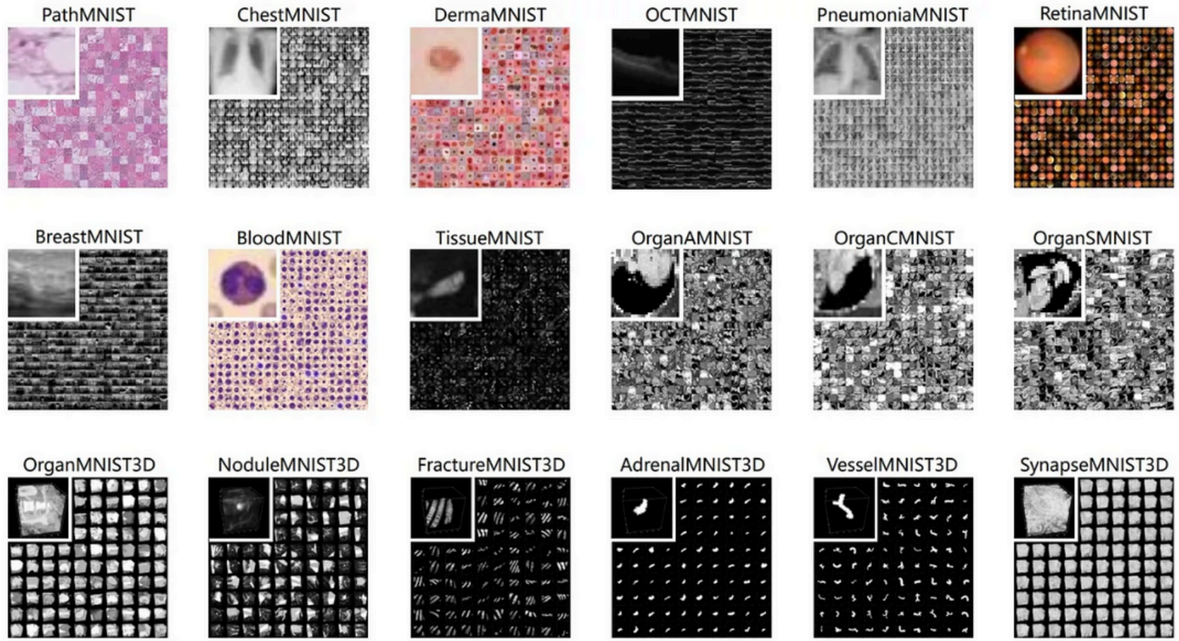
What?

What is the core contribution?

New domains

Remote sensing

MedMNIST (medicine)



What?

Testing pipeline

Use models trained on ImageNet/Remote Sensing/Medical Images and:

- *test them on augmented data*

What?

Testing pipeline

Use models trained on ImageNet/Remote Sensing/Medical Images and:

- *test them on augmented data*
- *compare them to humans*

What?

Testing pipeline

Use models trained on ImageNet/Remote Sensing/Medical Images and:

- *test them on augmented data*
- *compare them to humans*
- *...and show that the conculsions contradict previous work!*

Tested architectures:

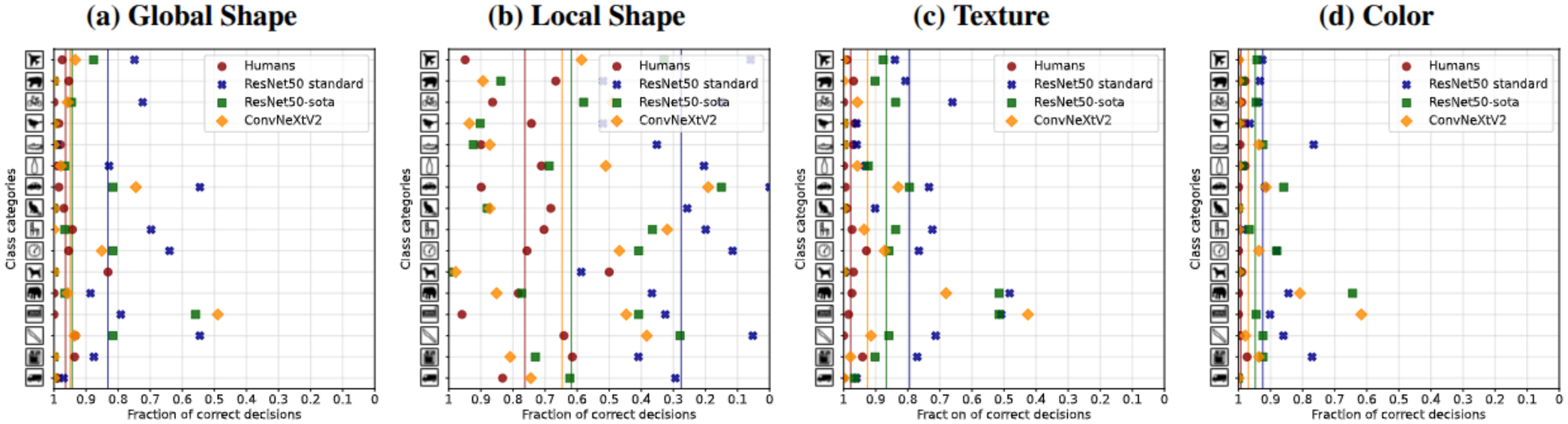
Architecture
Humans
ResNet50-standard [2]
ResNet50-sota [46]
ConvNeXt [4]
ConvNeXtV2 [50]
EfficientNet [3]
EfficientNetV2 [48]
MobileNetV3 [47]
ConvMixer [49]
ViT [51]
DeiT [52]
Swin [53]
CLIP ViT [54]

What?

Outcomes?

Tricky part - correct classification → p(x) > 0.5

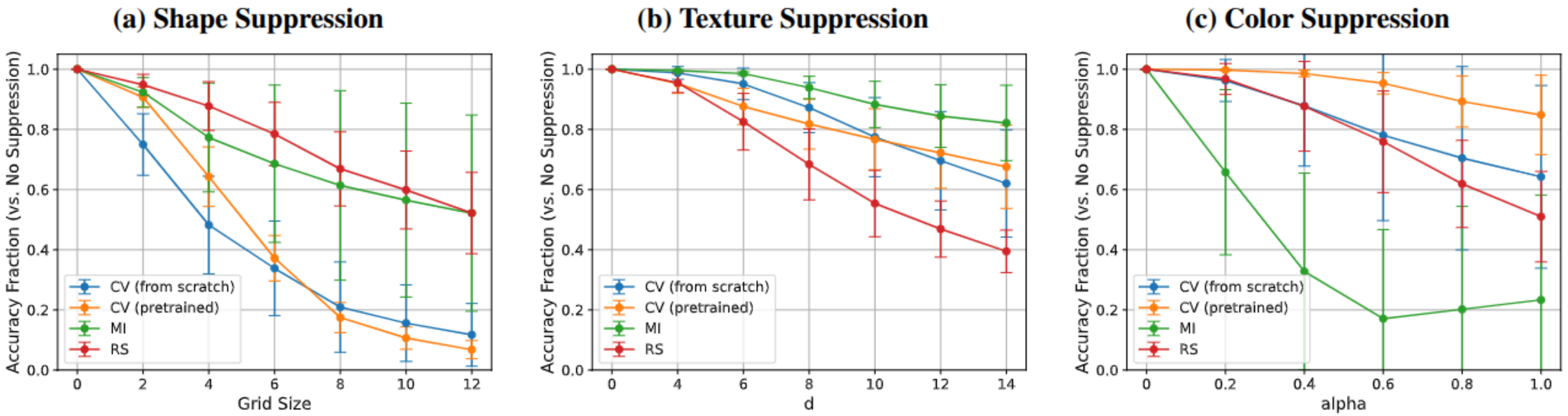
ImageNet only



What?

Outcomes?

All domains
(tested on ResNet50)



What?

Outcomes?

"Using this framework, we find no evidence for an inherent texture bias in CNNs, but instead observed a pronounced reliance on local shape features. [...] Across domains, we find that feature reliance varies substantially: CV models prioritize shape, MI models rely more evenly on color, and RS models exhibit strong texture sensitivity."

PART III

Can we do better?

Can we do better?

Some ideas from my last discussions with you:

(or random stuff that I had dreamed about)

- ***With the current state of genAI, it would be easy to generate realistic samples that suppress some image features (kudos to Hubert)***

Can we do better?

Some ideas from my last discussions with you:

(or random stuff that I had dreamed about)

- *With the current state of genAI, it would be easy to generate realistic samples that suppress some image features (kudos to Hubert)*
- *There is much more than shape, texture, and colour - contours, rotation, relative position, etc (kudos to ChatGPT)*

Can we do better?

Some ideas from my last discussions with you:

(or random stuff that I had dreamed about)

- With the current state of genAI, it would be easy to generate realistic samples that suppress some image features *(kudos to Hubert)*
- There is much more than shape, texture, and colour - contours, rotation, relative position, etc *(kudos to ChatGPT)*
- texture problem (either it exists or not) is a SHORTCUT problem - so we can handle it (maybe?) with Steering Vectors methodology *(kudos to my dream last night)*

Can we do better?

Some ideas from my last discussions with you:

(or random stuff that I had dreamed about)

- With the current state of genAI, it would be easy to generate realistic samples that suppress some image features *(kudos to Hubert)*
- There is much more than shape, texture, and colour - contours, rotation, relative position, etc *(kudos to ChatGPT)*
- texture problem (either it exists or not) is a SHORTCUT problem - so we can handle it (maybe?) with Steering Vectors methodology *(kudos to my dream last night)*
- everything is tested on ImageNet-16 but trained on ImageNet-1k → that definetly creates some bias (adding/removing texture may suggest models sth outside 16 classes)

Can we do better?

Some food for thought:

(maybe)

- *If the model is entirely shape-biased, should the saliency map show edges instead of object area? Are there any potential connections?*
- *Can any image feature be suppressed without going out of the domain - i.e., can the testing-based methodology presented in this article be robust?*
- *Can we distinguish between features of the image using XAI? How to distinguish between saliency of local shape and local texture?*

Questions/discussion time