Deep Learning: (still) Not Robust

Nicholas Carlini
Google

Better Language Models and Thei

Implications

We've trained a large-scale unsupervise
model which generates coherent paragr
text, achieves state-of-the-art performa
many language modeling benchmarks, a
performs rudimentary reading comprehe
machine translation, question answering
summarization—all without task-specifi

February 14, 2019
24 minute read

Deep Speech 2: End-to-E
English an

Baidu Research –
Dario Amodei, Rishita Anubhai, Eric Batten
Jingdong Chen, Mike Chrzanowski, Adam
Linxi Fan, Christopher Fougner, Tony Har
Libby Lin, Sharan Narang, Andrew Ng, S
Sanjeev Satheesh, David Seetapun, Shubho S
Bo Xiao, Dani Yogatan

Ab

We show that an end-to-end deep lea
either English or Mandarin Chinese sp
cause it replaces entire pipelines of han
works, end-to-end learning allows us to
ing noisy environments, accents and different languages. Key to our approach is our application of HPC techniques, resulting in a 7x speedup over our previous system [26]. Because of this efficiency, experiments that previously took weeks now run in days. This enables us to iterate more quickly to identify superior architectures and algorithms. As a result, in several cases, our system is competitive with the transcription of human workers when benchmarked on standard datasets. Finally, using a technique called Batch Dispatch with GPUs in the data center, we show that our system can be inexpensively deployed in an online setting, delivering low latency when serving users at scale.

Facebook

Introducing the First AI Model That Translates 100 Languages Without Relying on English

October 19, 2020
By Angela Fan, Research Assistant

This Talk:

.... however

adversarial
perturbation

88% **tabby cat**

99% **guacamole**

football

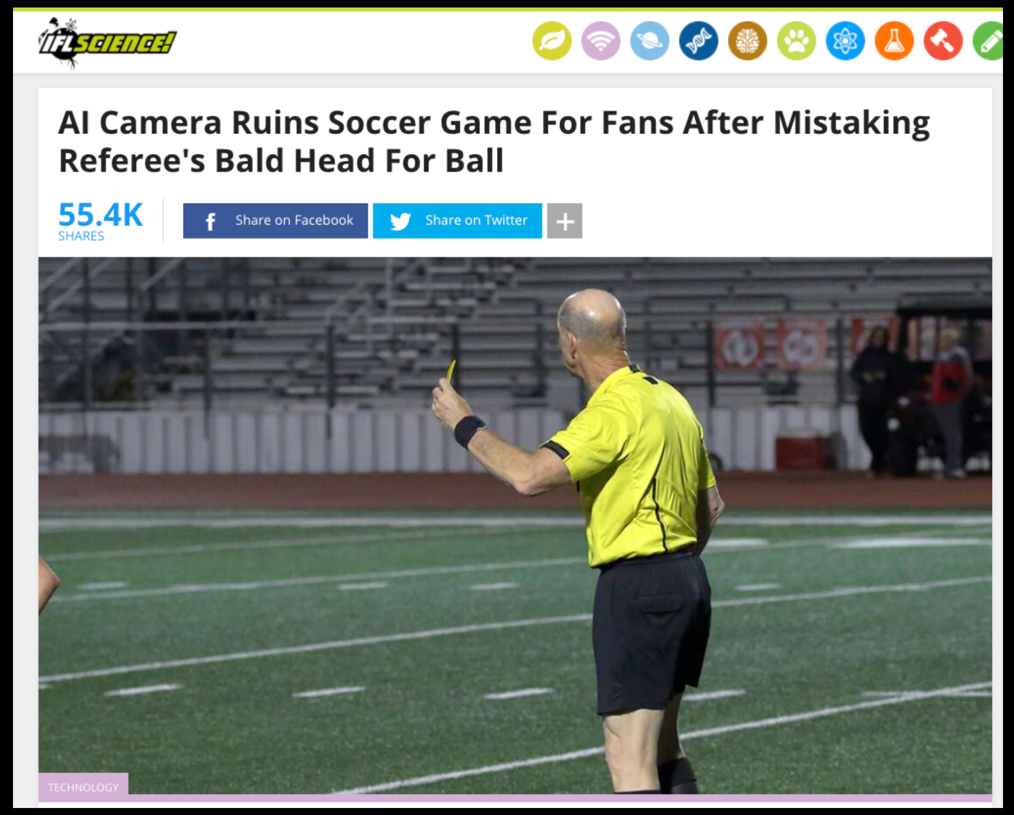
AI Camera Ruins ~~Soccer~~ Game For Fans After Mistaking Referee's Bald Head For Ball

TECHNOLOGY

Adversarial

Distribution Shifts

99% guacamole


AI Camera Ruins Soccer Game For Fans After Mistaking Referee's Bald Head For Ball

55.4K SHARES

Natural

Distribution Shifts

Adversarial

Distribution Shifts

Natural

Distribution Shifts

Adversarial

Distribution Shifts

Adversarial

Distribution Shifts

NeurIPS'20, with Florian Tramer, Wieland Brendel, Aleksander Madry

Adversarial (n.)

Defn: "involving or characterized by conflict or opposition."

GIVEN	a neural network f an input to the network x
FIND	a new input x′
SUCH THAT	f(x′) is classified incorrectly x and x′ are *close*

Adversarial Accuracy

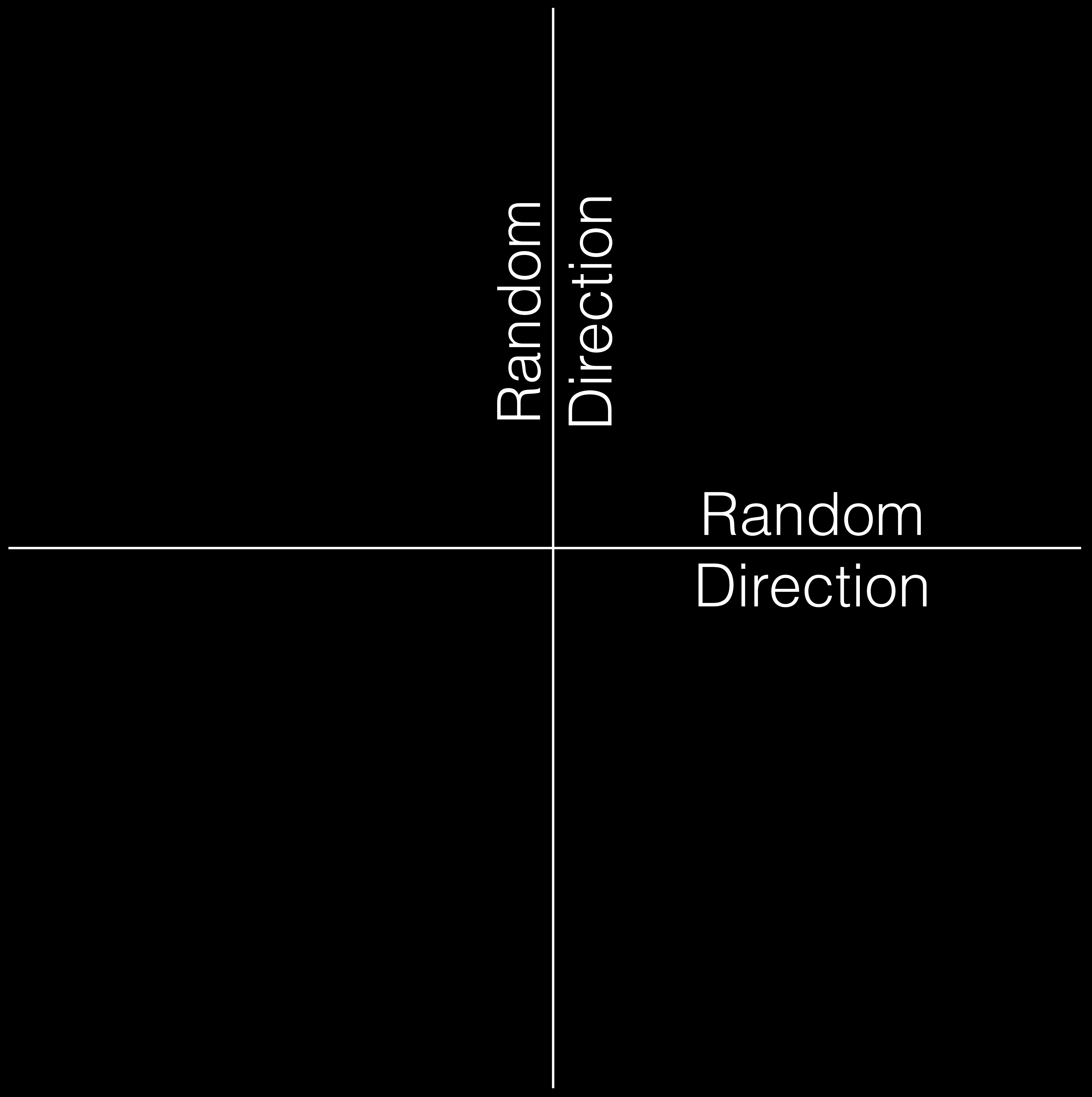
Probability an adversary can succeed at this game

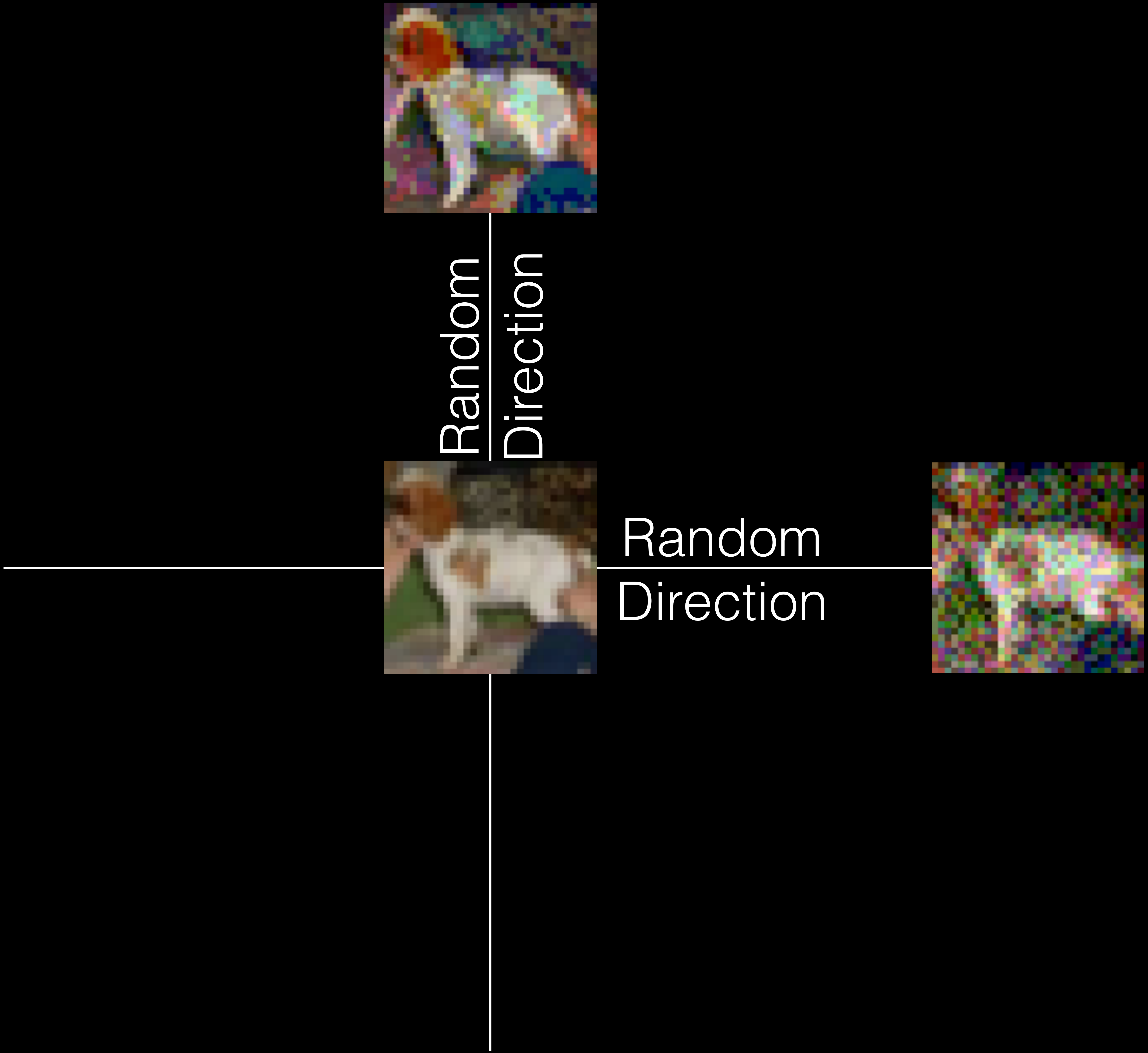
On Adaptive Attacks to Adversarial Example Defenses

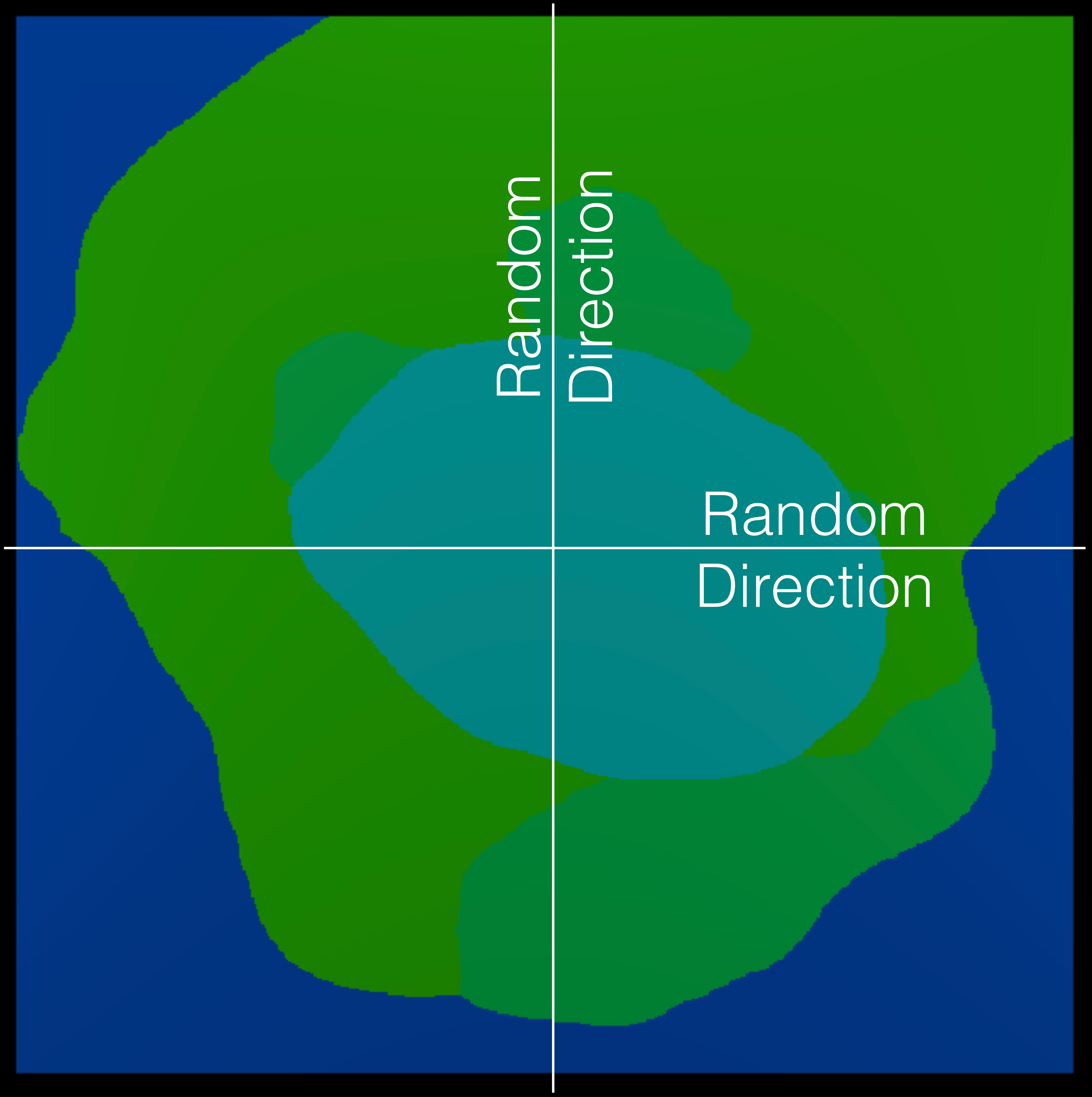
Florian Tramèr* Stanford University
Nicholas Carlini* Google Brain
Wieland Brendel* University of Tübingen
Aleksander Mądry MIT

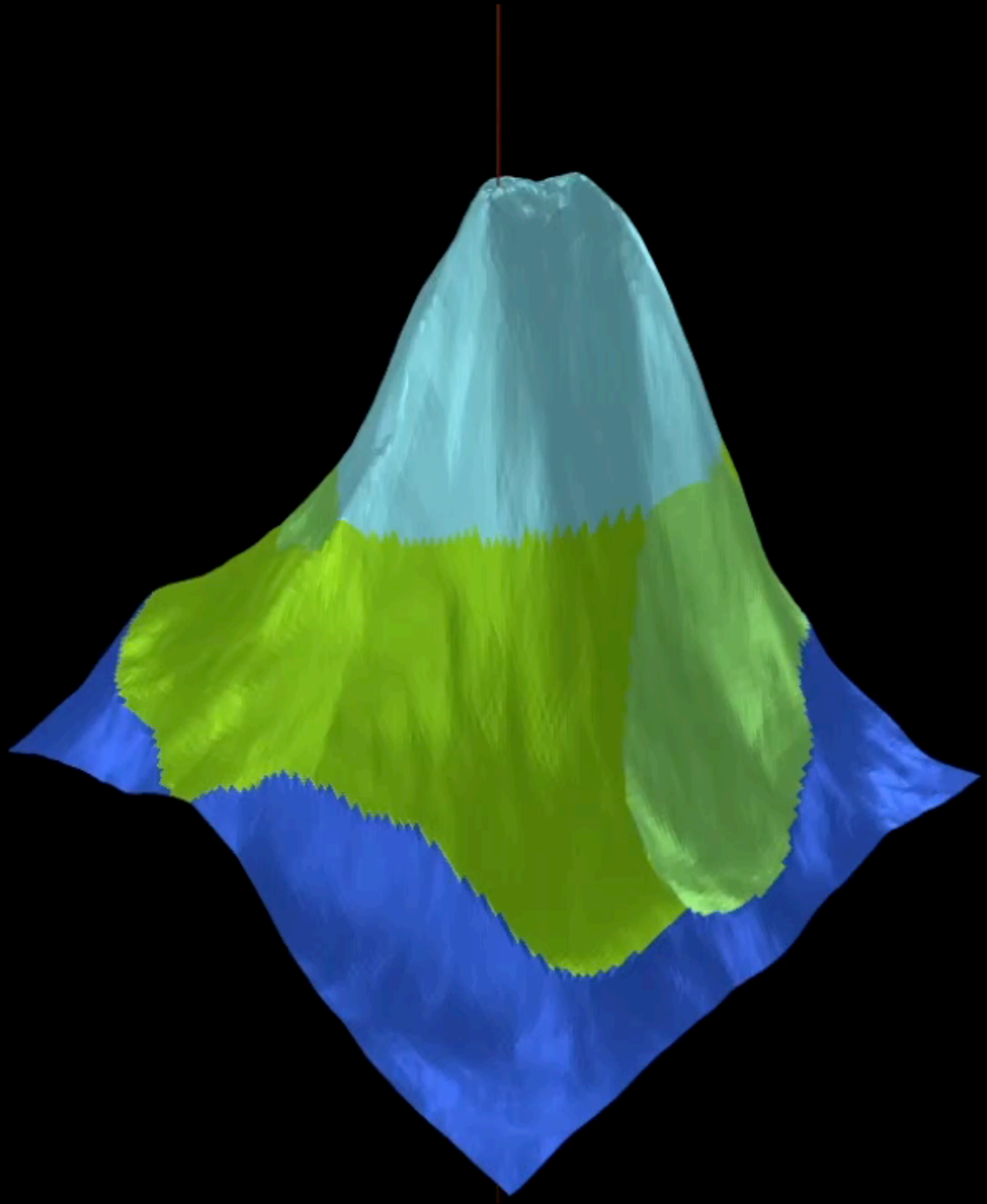
We evaluated 13 defenses proposed at (ICLR|ICML|NeurIPS) 20(18|19|20)

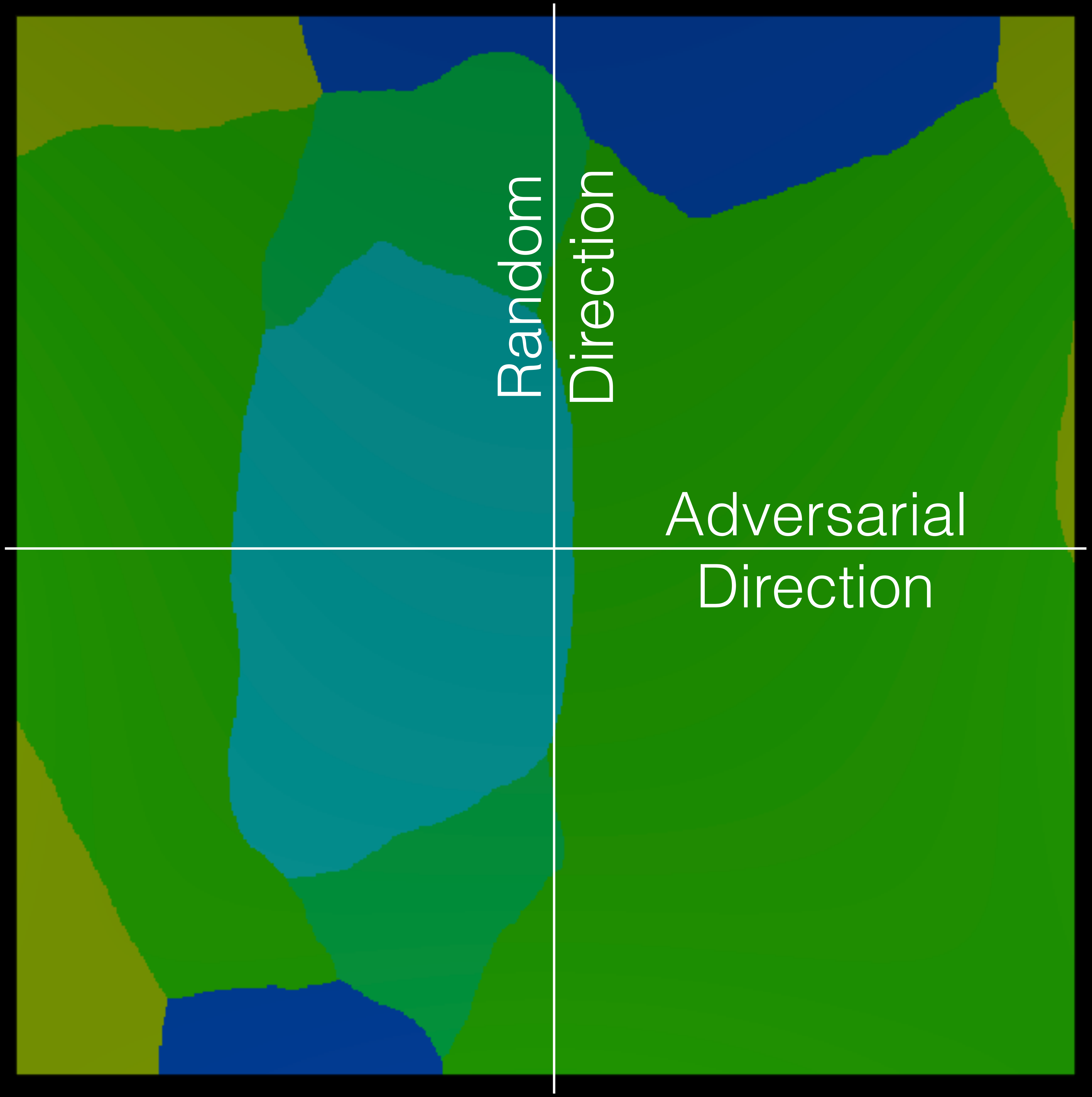
All were broken.
Adversarial accuracy of roughly 0%.

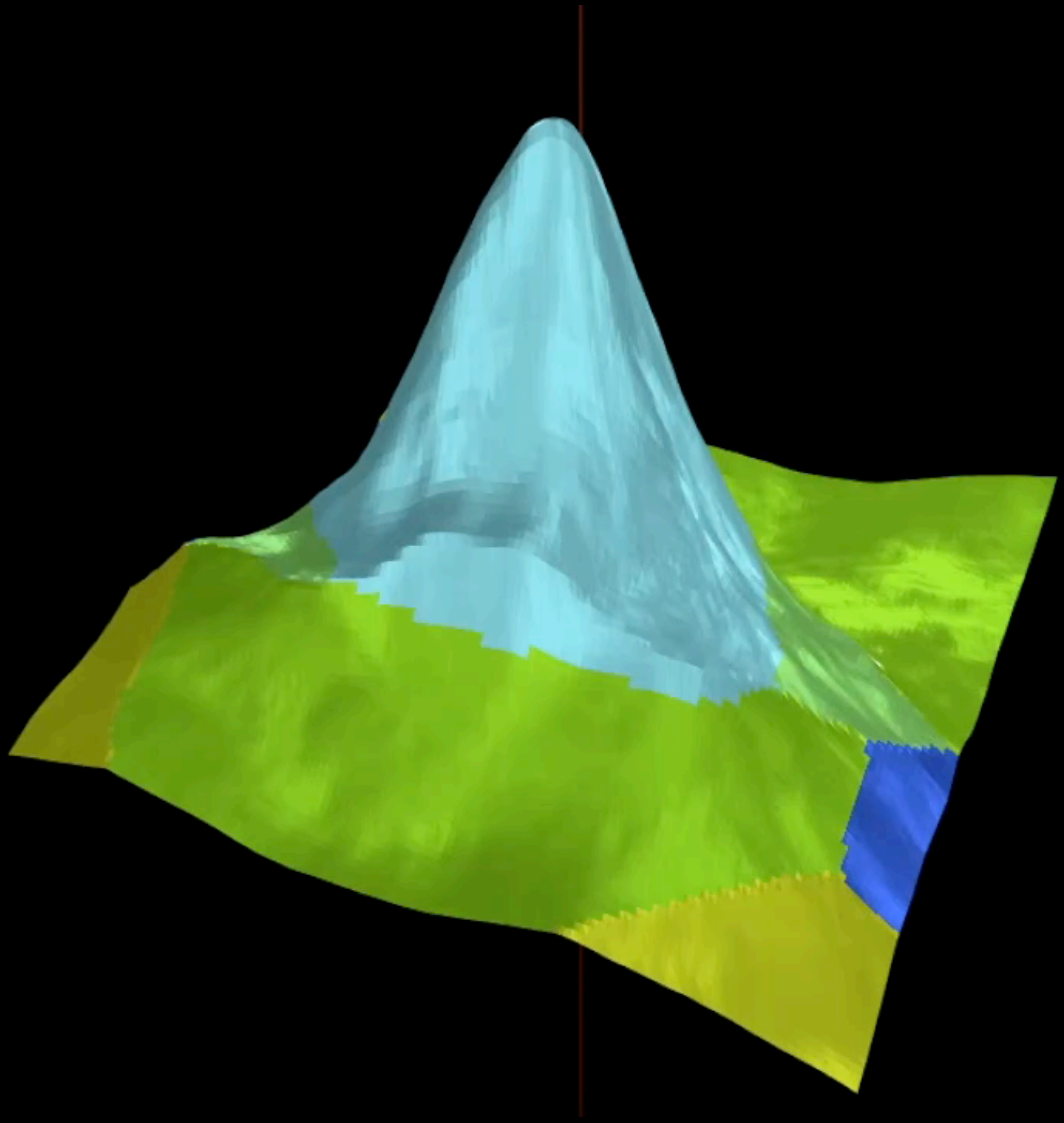

Random Direction
Random Direction


Random Direction
Random Direction

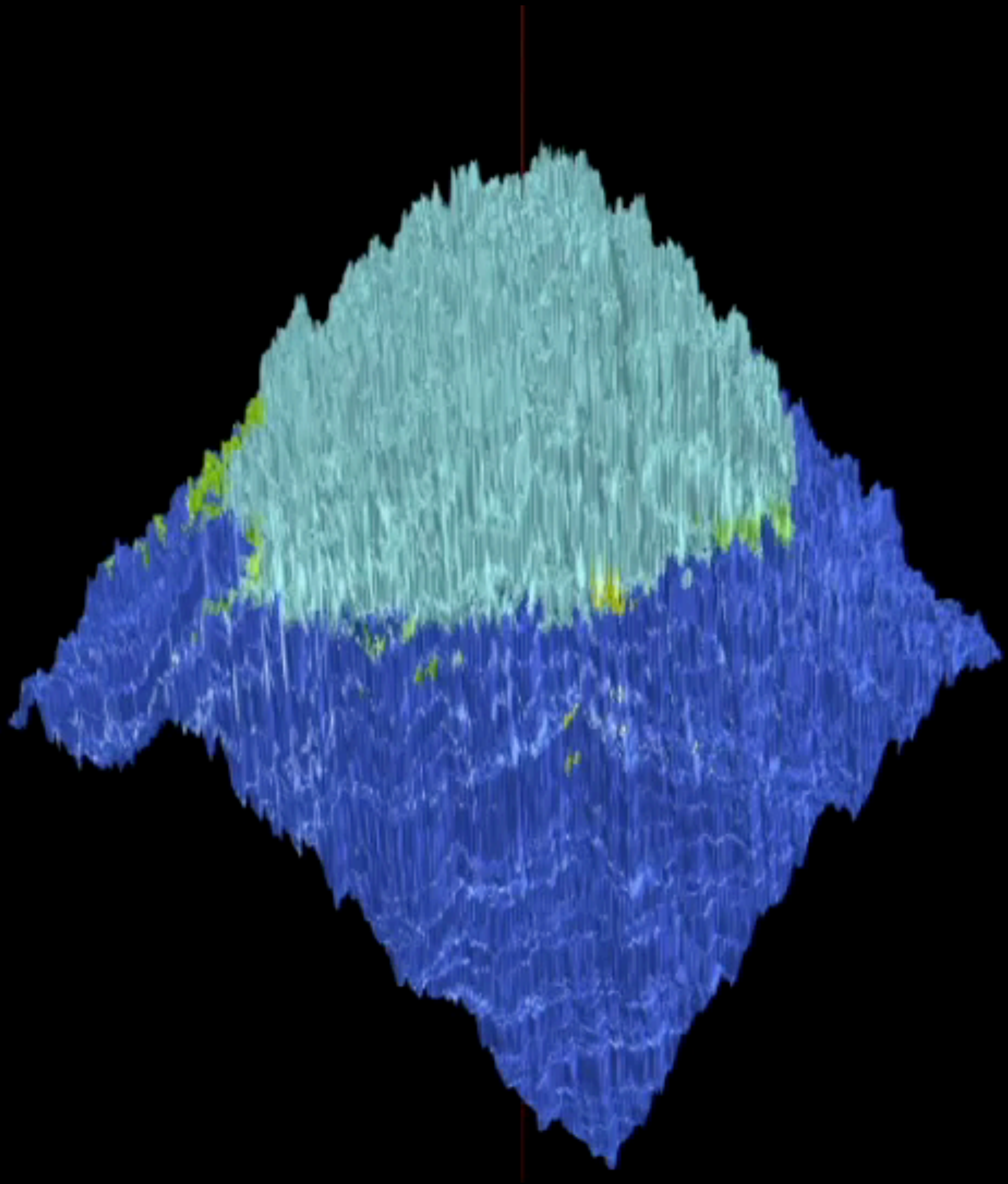

Random Direction
Random Direction

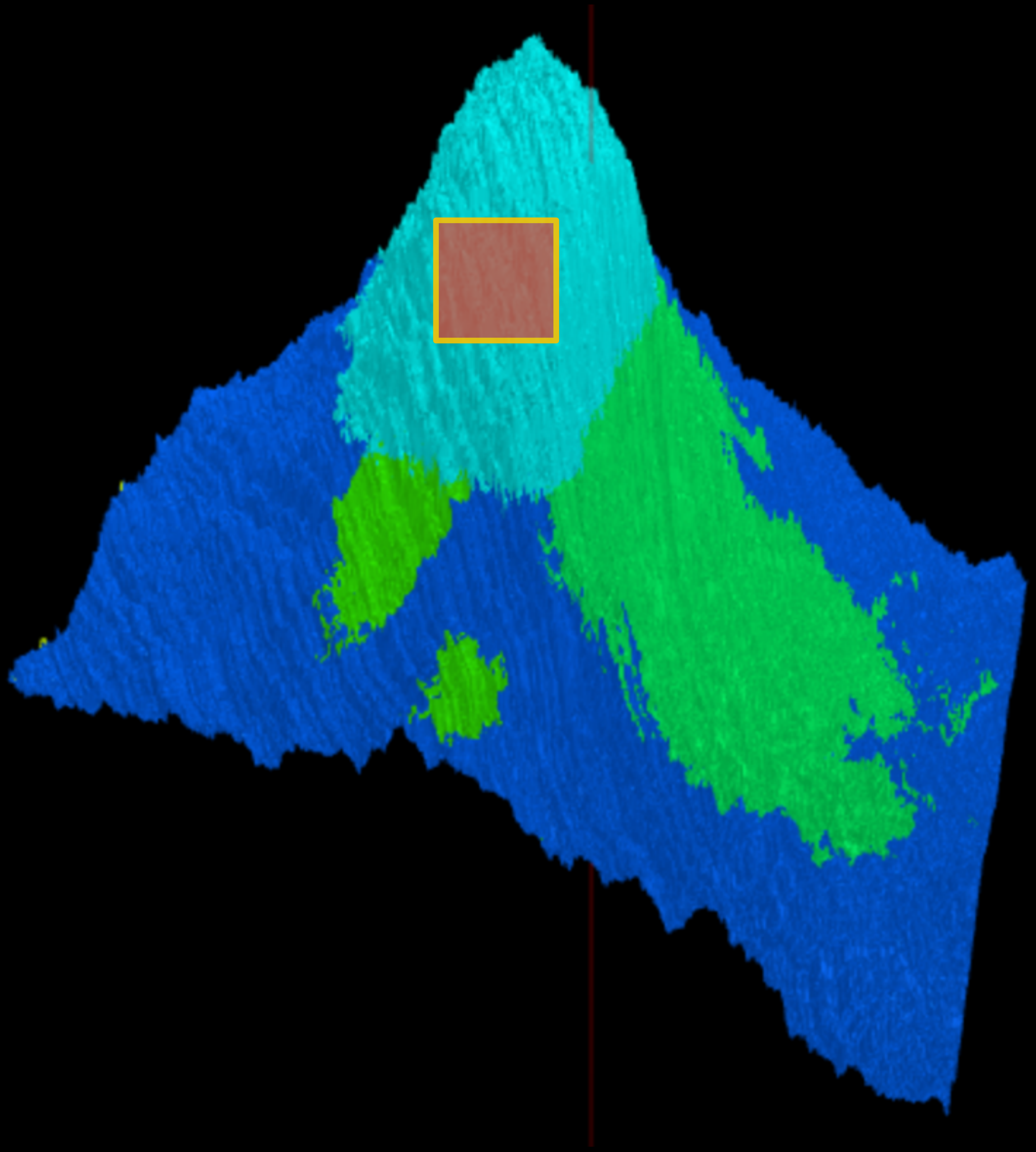


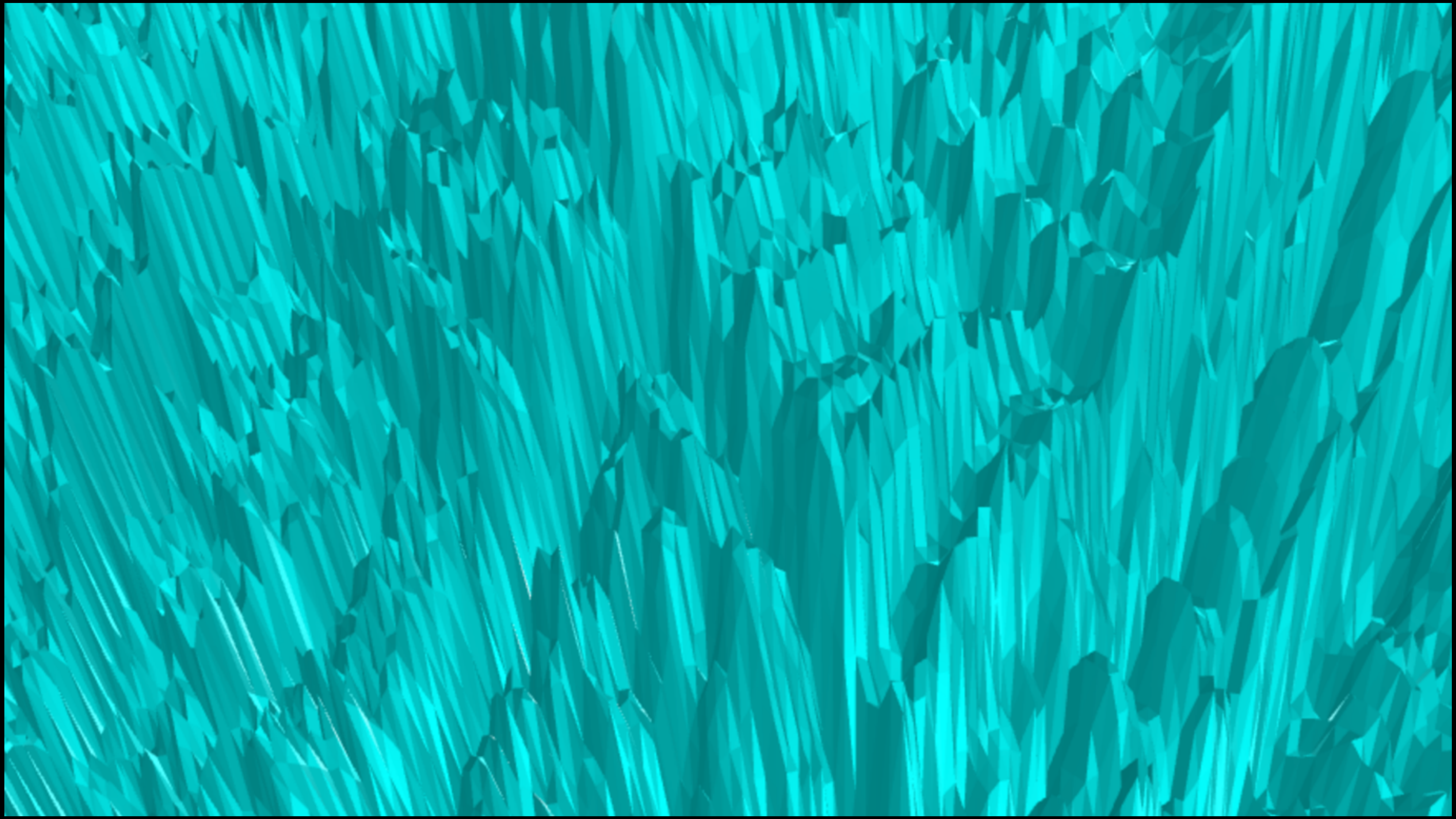

Random Direction
Adversarial Direction



What do defenses do?



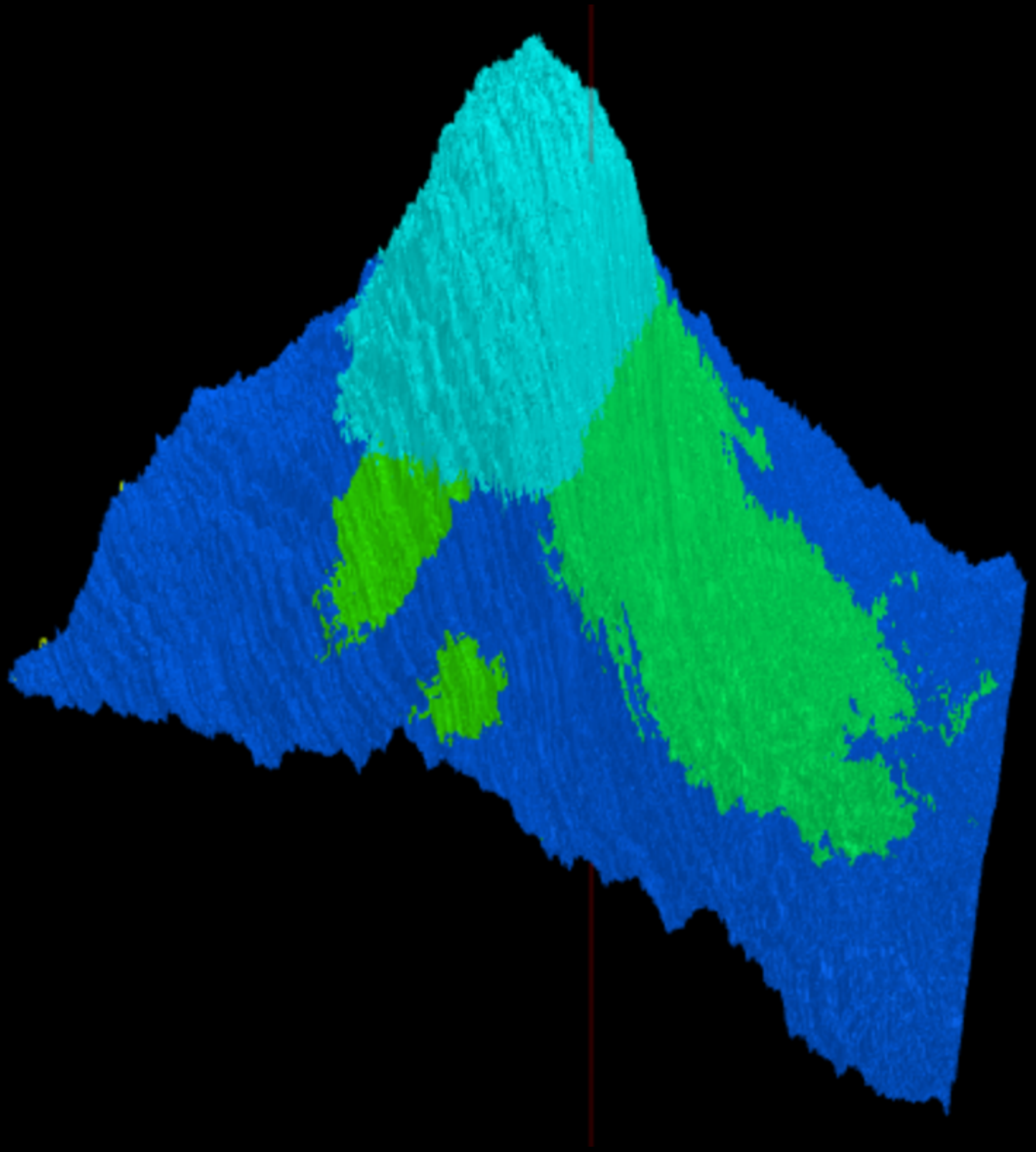


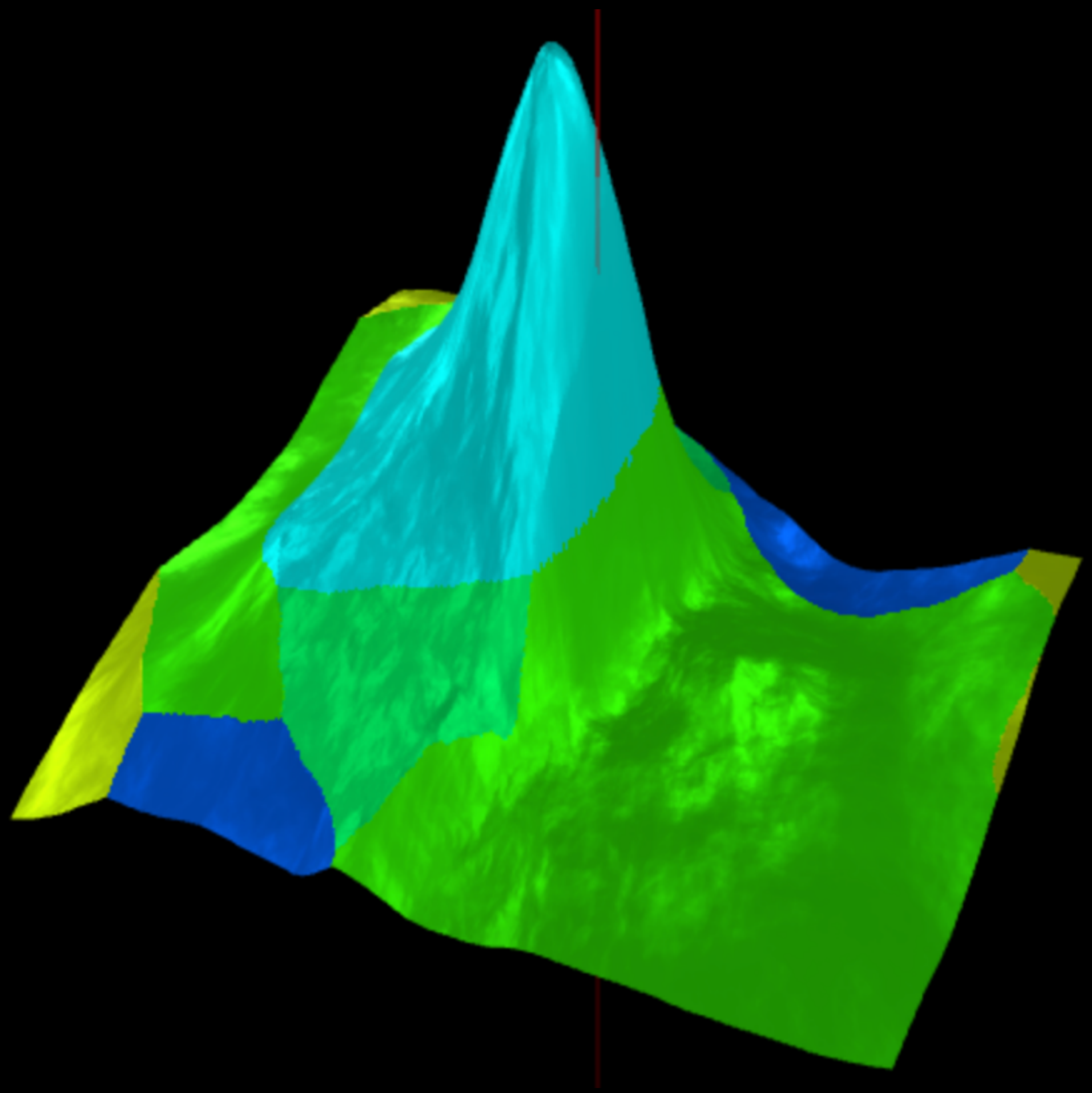


Our paper:
Adaptive Attacks

I'm not going to tell you **how** we broke them.

... it's quite boring.





Instead let's talk about the context of this paper

Previously

Defenses | **Attacks**

New Idea 1 → New Idea A

New Idea 2 → New Idea B

New Idea 3 → New Idea C


Defensive Distillation is Not Robust to Adversarial Examples

Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods

MagNet and "Efficient Defenses Against Adversari… are Not Robust to Adversarial Examples

Obfuscated Gradients Give a False Se… Circumventing Defenses to Adversar…

On the Robustness of the CVPR 2018 …

Is AmI (A… Robust

Comment on *Biologically inspired protection of deep networks from adversarial attacks*

ON THE LIMITATION OF LOCAL INTRINSIC DIMENSIONALITY FOR CHARACTERIZING THE SUBSPACES OF …

Adversarial Risk and the Dangers of Evaluating Against Weak Attacks

The Efficacy of SHIELD under Different Threat Models

Paper Type: Appraisal Paper of Existing Method

Cory Cornelius cory.cornelius@intel.com
Nilaksh Das nilakshdas@gatech.edu
Shang-Tse Chen schen351@gatech.edu

Evaluating and Understanding the Robustness of Adversarial Logit Pairing

Logan Engstrom* Andrew Ilyas* Anish Athalye*
Massachusetts Institute of Technology
{engstrom,ailyas,aathalye}@mit.edu

Abstract

We evaluate the robustness of Adversarial Logit Pairing, a recently proposed defense against adversarial examples. We find that a network trained with Adversarial Logit Pairing achieves 0.6% correct classification rate under targeted adversarial attack, the threat model in which the defense is considered. We provide a brief overview of the defense and the threat models/claims considered, as well as a discussion of the methodology and results of our attack. Our results offer insights into the reasons underlying the vulnerability of ALP to adversarial attack, and are of general interest in evaluating and understanding adversarial defenses.

1 Contributions

For summary, the contributions of this note are as follows:

1. **Robustness**: Under the white-box targeted attack threat model specified in Kannan et al., we upper bound the correct classification rate of the defense to **0.6%** (Table 1). We also perform targeted and untargeted attacks and show that the attacker can reach success rates of 98.6% and 99.9% respectively (Figures 1, 2).

Today …

Defenses		Attacks
New Idea 1	→	New Idea A
New Idea 2	→	New Idea B
New Idea 3	→	New Idea C
New Idea 95	→	just reuse one

On Adaptive Attacks to Adversarial Example Defenses

Florian Tramèr* Stanford University
Nicholas Carlini* Google Brain
Wieland Brendel* University of Tübingen
Aleksander Mądry MIT

Another **weakness** of the paper is that **defenses are broken by existing techniques**. Indeed, at the end of the analysis, most of the defenses are broken either by using EOT, BPDA, or by tuning the parameters of existing attacks such as PGD. **All this techniques already exist in the literature** [1,2,3,4]; hence the technical part is not novel.

Two areas have improved

1. **Code**
   is now always available

2. **Adaptive attacks**
   are at least attempted

The problem
is methodological

Simplicity

for example ... one paper's attack

$$\mathcal{L}_1 = \underbrace{\mathcal{L}(h(\mathbf{x}'), \mathbf{p}^{\text{adv}})}_{\text{misclassify } \mathbf{x}' \text{ as } y_t},$$

$$\mathcal{L}_2 = \underbrace{\mathbb{E}_{\epsilon \sim N(0,\sigma^2 I)} \left[\|h(\mathbf{x}') - h(\mathbf{x}' + \epsilon)\|_1\right]}_{\text{bypass C1}},$$

$$\mathcal{L}_3 = \mathbb{E}_{y' \sim \text{Uniform}, y' \neq y_t} [\mathcal{L}(h(\mathbf{x}' + \alpha\delta_{y'}), y')],$$

$$\mathcal{L}_4 = -\mathcal{L}(h(\mathbf{x}' + \alpha\delta_{y_t}), y_t).$$

$$\mathcal{L}^\star = \lambda\mathcal{L}_1 + \mathcal{L}_2 + \mathcal{L}_3 + \mathcal{L}_4.$$

for example ... one paper's attack

$$\mathcal{L}_1 = \underbrace{\mathcal{L}(h(\mathbf{x}'), \mathbf{p}^{\text{adv}})}_{\text{misclassify } \mathbf{x}' \text{ as } y_t},$$

$$\mathcal{L}_2 = \underbrace{\mathbb{E}_{\epsilon \sim N(0,\sigma^2 I)} \left[\| h(\mathbf{x}') - h(\mathbf{x}' + \epsilon) \|_1 \right]}_{\text{bypass C1}},$$

$$\mathcal{L}_3 = \mathbb{E}_{y' \sim \text{Uniform}, y' \neq y_t} [\mathcal{L}(h(\mathbf{x}' + \alpha \delta_{y'}), y')],$$

$$\mathcal{L}_4 = -\mathcal{L}(h(\mathbf{x}' + \alpha \delta_{y_t}), y_t).$$

$$\mathcal{L}^\star = \lambda \mathcal{L}_1 + \mathcal{L}_2 + \mathcal{L}_3 + \mathcal{L}_4.$$

for example ... our attack

$$\mathcal{L}_1 = \underbrace{\mathcal{L}(h(\mathbf{x}'), \mathbf{p}^{\mathrm{adv}})}_{\text{misclassify } \mathbf{x}' \text{ as } y_t},$$

Adversarial

Distribution Shifts

Adversarial

Distribution Shifts

Natural

Distribution Shifts

Natural

Distribution Shifts

Rohan Taori, Achal Dave, Vaishaal Shankar, Benjamin Recht, Ludwig Schmidt

Natural (adj.)

Defn: "existing in or caused by nature"

What we *want*

1. Someone wants to know what breed of dog they just saw on the street
2. They take out their phone
3. Open up the camera app
4. Take a picture, and run a ResNet on the image

What we *have*

1. Someone wants to know what breed of dog they just saw on the street
2. They take out their phone
3. Open up the camera app
4. Close the camera app. Open up the browser. Visit http://image-net.org/. Download the ILSVRC2012 test set. Select an image of a dog uniformly at random. Ask the resnet model to classify that random image. Ignore the real dog.

Constructing "natural" datasets

Do ImageNet Classifiers Generalize t

Benjamin
UC Berl

We buil
the focus o
re-used tes
extent curr
and find ac
accuracy ga
suggest tha
generalize t

ObjectNet: A large-scale bias-co
pushing the limits of object re

Andrei Barbu* MIT, CSAIL & CBMM
David Mayo* MIT, CSAIL & CBMM
Christopher Wang MIT, CSAIL
Dan Gutfreund MIT-IBM Watson AI
Joshua Ten MIT, BCS &

Abstract

We collect a large real-world test set, ObjectNet, for ob
where object backgrounds, rotations, and imaging vi
scientific experiments have controls, confounds whic
to ensure that subjects cannot perform a task by exp
the data. Historically, large machine learning and co
lacked such controls. This has resulted in models tha
datasets and perform better on datasets than in real
tested on ObjectNet, object detectors show a 40-45%
respect to their performance on other benchmarks, d
Controls make ObjectNet robust to fine-tuning show
increases. We develop a highly automated platform that enables gathering datasets with controls by crowdsourcing image capturing and annotation. ObjectNet is the same size as the ImageNet test set (50,000 images), and by design does not come paired with a training set in order to encourage generalization. The dataset is both easier than ImageNet – objects are largely centered and unoccluded – and harder, due to the controls. Although we focus on object recognition here, data with controls can be gathered at scale using automated tools throughout machine learning to generate datasets that exercise models in new ways thus providing valuable feedback to researchers. This work opens up new avenues for research in generalizable, robust, and more human-like computer vision and in creating datasets where results are predictive of real-world performance.

Do Image Classifiers Generalize Across Time?

Vaishaal Shankar* UC Berkeley
Achal Dave*
Rebecca Roelofs
Deva Ramanan CMU

We study the robustness of image
part of this study, we construct two
containing a total of 57,897 images
Our datasets were derived from Ima
re-annotated by human experts for
pre-trained on ImageNet and show a
datasets. Additionally, we evaluate t
induce both classification as well as lo
of 14 points. Our analysis demonstra
substantial and realistic challenge to
that require both reliable and low-la

Natural Adversarial Examples

Dan Hendrycks
UC Berkeley
hendrycks@berkeley.edu

Kevin Zhao*
University of Washington
kwzhao@cs.washington.edu

Steven Basart*
University of Chicago
steven@ttic.edu

Jacob Steinhardt
UC Berkeley
jsteinhardt@berkeley.edu

Dawn Song
UC Berkeley
dawnsong@berkeley.edu

Abstract

We introduce natural adversarial examples–real-world, unmodified, and naturally occurring examples that cause machine learning model performance to substantially degrade. We introduce two new datasets of natural adversarial examples. The first dataset contains 7,500 natural adversarial examples for ImageNet classifiers and serves as a hard ImageNet classier test set called IMAGENET-A. *We also curate an adversarial out-of-distribution detection dataset called* IMAGENET-O, *which to our knowledge is the first out-of-distribution detection dataset created for ImageNet models. These two datasets provide new ways to measure model robustness and uncertainty. Like ℓ_p adversarial examples, our natural adversarial examples transfer to unseen black-box models. For example, on* IMAGENET-A *a DenseNet-121 obtains around 2% accuracy, an accuracy drop of approximately 90%, and its out-of-distribution detection performance on* IMAGENET-O *is near random chance levels. Popular training techniques for improving robustness have little effect, but some architectural changes provide mild improvements. Future research is required to enable generalization to natural adversarial examples.*

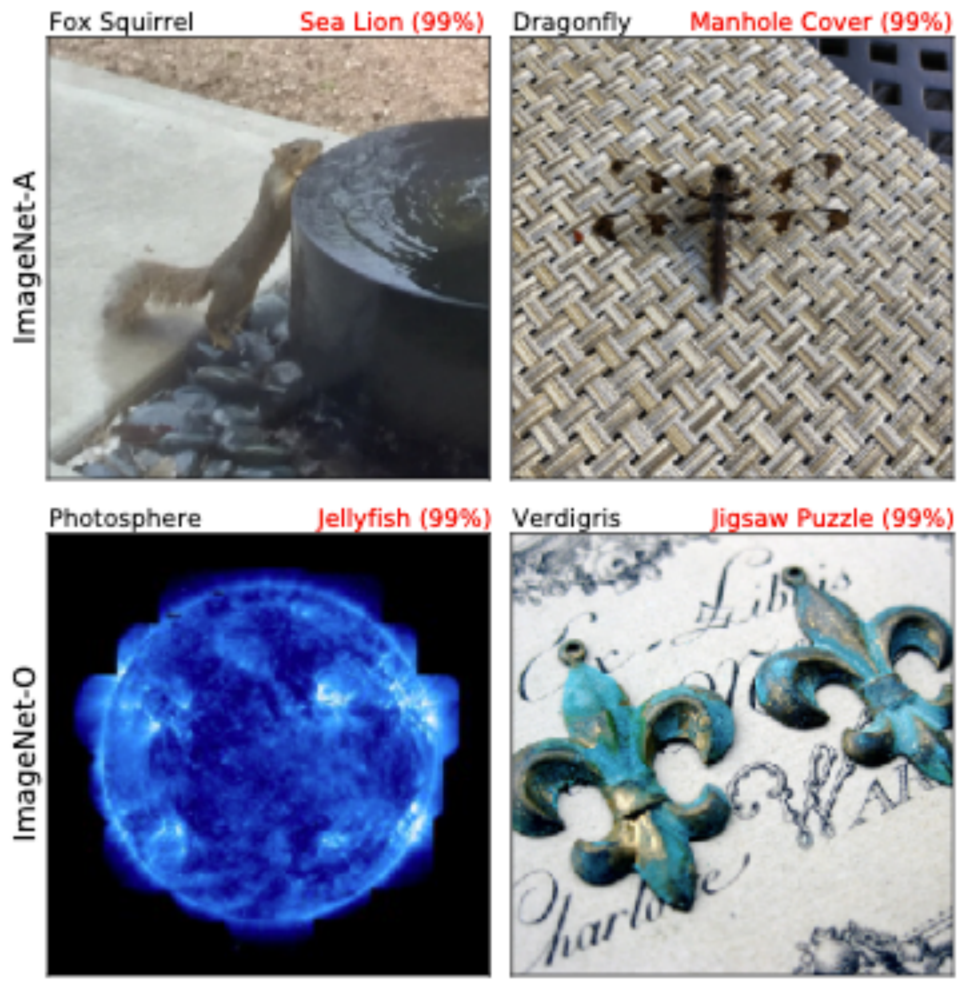


Figure 1: Natural adversarial examples from IMAGENET-A and IMAGENET-O. The black text is the actual class, and the red text is a ResNet-50 prediction and its confidence.

Do ImageNet Classifiers Generalize to ImageNet?

Benjamin Recht* UC Berkeley
Rebecca Roelofs UC Berkeley
Ludwig Schmidt UC Berkeley
Vaishaal Shankar UC Berkeley

Abstract

We build new test sets for the CIFAR-10 and ImageNet datasets. Both benchmarks have been the focus of intense research for almost a decade, raising the danger of overfitting to excessively re-used test sets. By closely following the original dataset creation processes, we test to what extent current classification models generalize to new data. We evaluate a broad range of models and find accuracy drops of 3% – 15% on CIFAR-10 and 11% – 14% on ImageNet. However, accuracy gains on the original test sets translate to larger gains on the new test sets. Our results suggest that the accuracy drops are not caused by adaptivity, but by the models' inability to generalize to slightly "harder" images than those found in the original test sets.

Do ImageNet Classifiers Generalize to ImageNet?

Benjamin Recht*
UC Berkeley

Rebecca Roelofs
UC Berkeley

Ludwig Schmidt
UC Berkeley

Vaishaal Shankar
UC Berkeley

Abstract

We build new test sets for the CIFAR-10 and ImageNet datasets. Both benchmarks have been the focus of intense research for almost a decade, raising the danger of overfitting to excessively re-used test sets. By closely following the original dataset creation processes, we test to what extent current classification models generalize to new data. We evaluate a broad range of models and find accuracy drops of 3% – 15% on CIFAR-10 and 11% – 14% on ImageNet. However, accuracy gains on the original test sets translate to larger gains on the new test sets. Our results suggest that the accuracy drops are not caused by adaptivity, but by the models' inability to generalize to slightly "harder" images than those found in the original test sets.

This research study is being conducted by Ben Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar from UC Berkeley. For question about this study, please contact ludwig@berkeley.edu and roelofs@cs.berkeley.edu. In this study, we will ask you to indicate whether given image belong to a certain object category. Occasionally, the images may contain disturbing or adult content. We would like to remind you that participation in our study is voluntary and that you can withdraw from the study at any time.

Which of these images contain at least one object of type

English foxhound

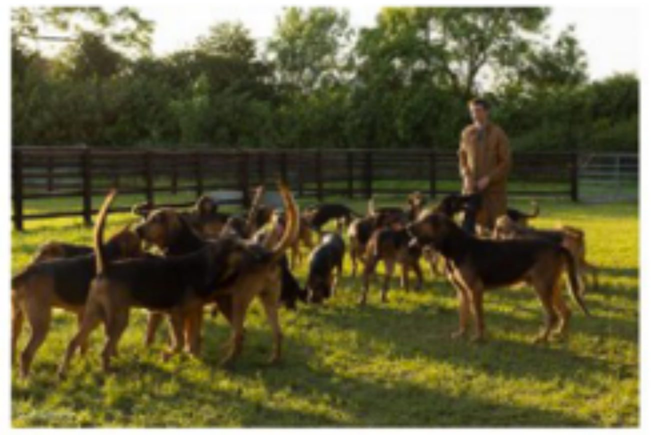
Definition: an English breed slightly larger than the American foxhounds originally used to hunt in packs

Task:
For each of the following images, check the box next to an image if it contains at least one object of type *English foxhound*. Select an image if it contains the object regardless of occlusions, other objects, and clutter or text in the scene. Only select images that are photographs (**no drawings or paintings**).

Please make accurate selections!

If you are unsure about the object meaning, please also consult the following Wikipedia page(s): https://en.wikipedia.org/wiki/English_Foxhound

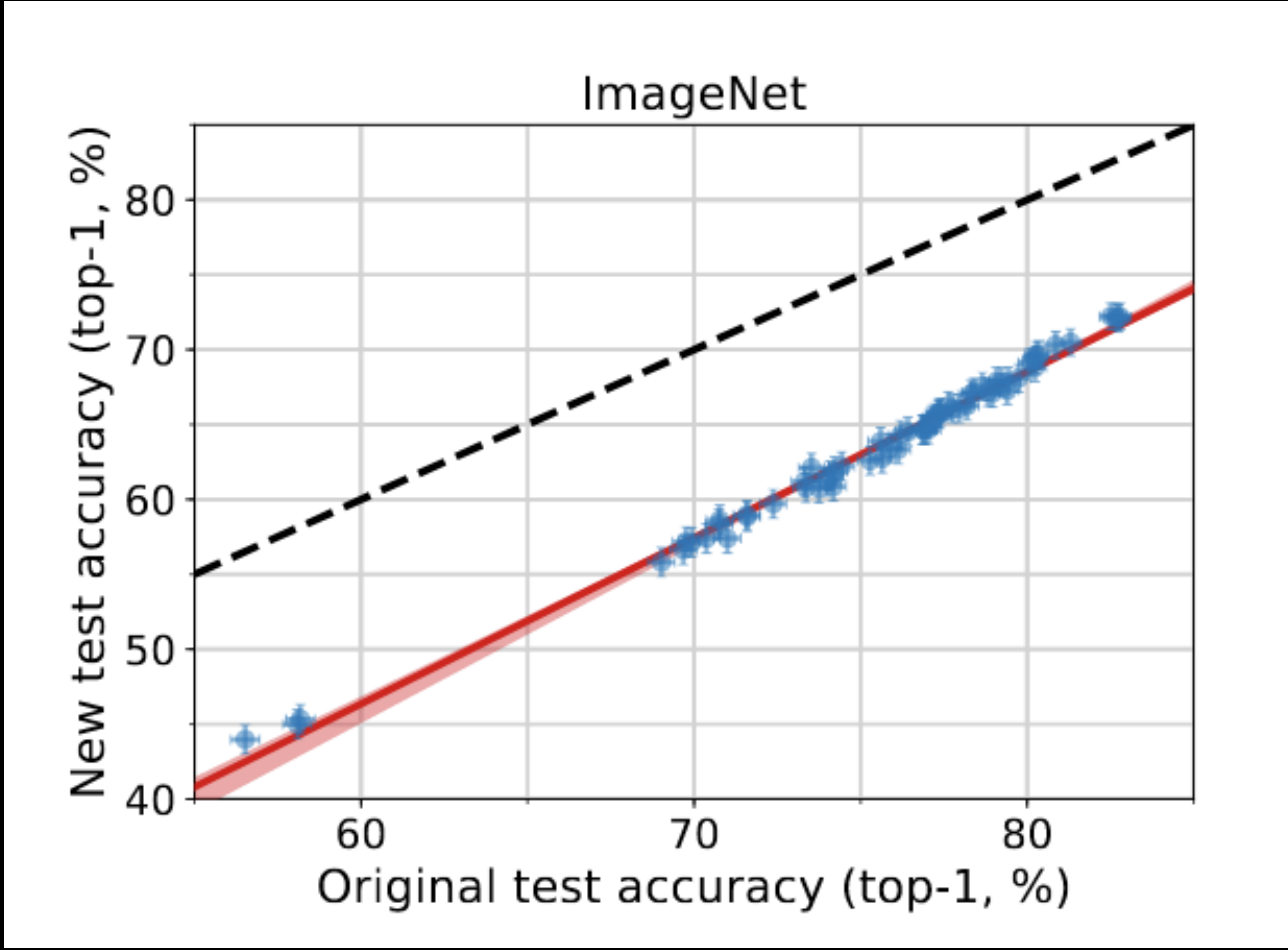
If it is impossible to complete a HIT due to missing data or other problems, please return the HIT.



Now we have a new dataset.

Identical in every way to the original.

How do models do on this new dataset?

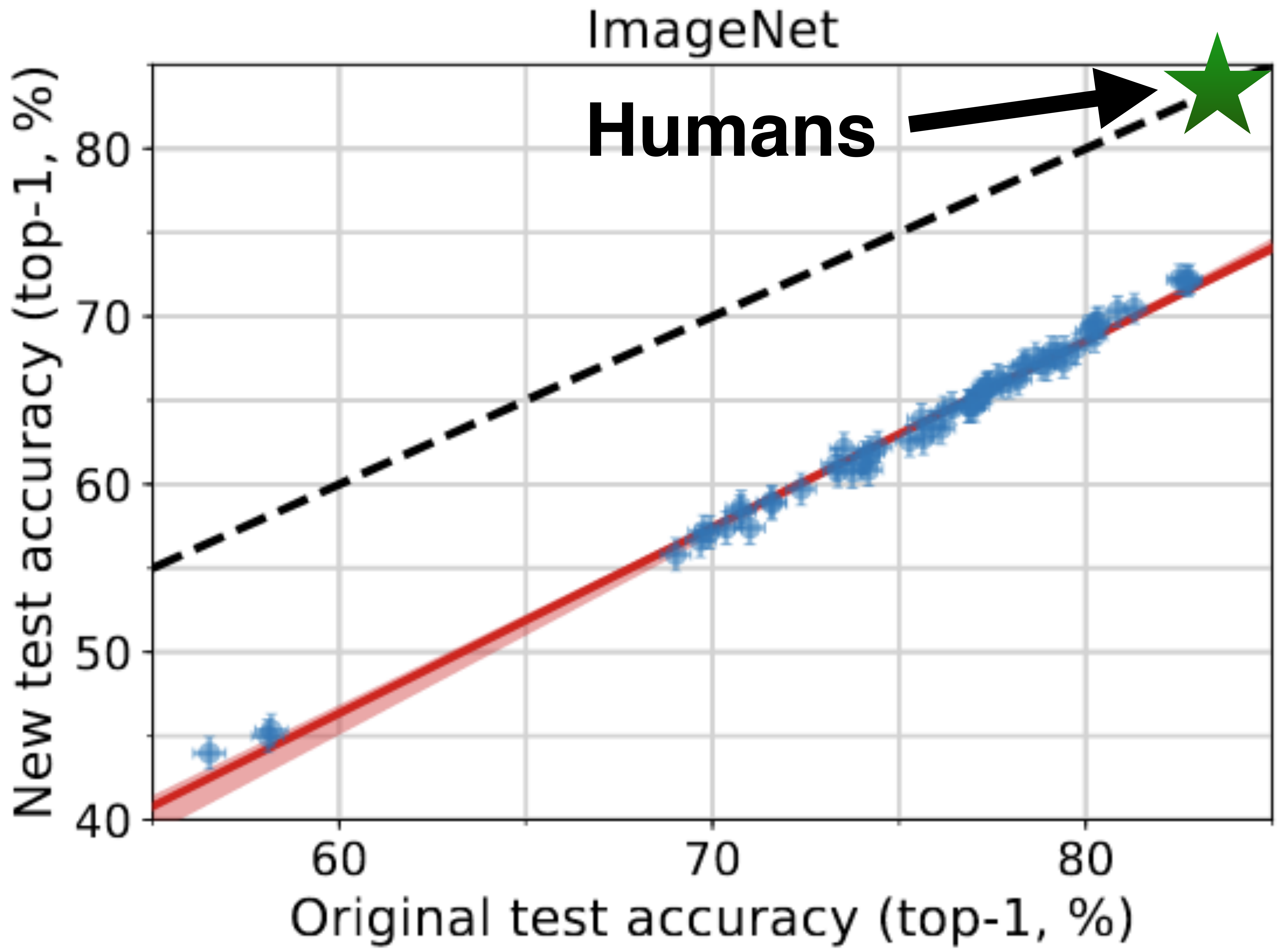

ImageNet
New test accuracy (top-1, %)
Original test accuracy (top-1, %)
40
50
60
70
80
60
70
80

Possible explanations

1. It's just a harder dataset
2. Adaptive overfitting
3. Distribution shift

Possible explanations

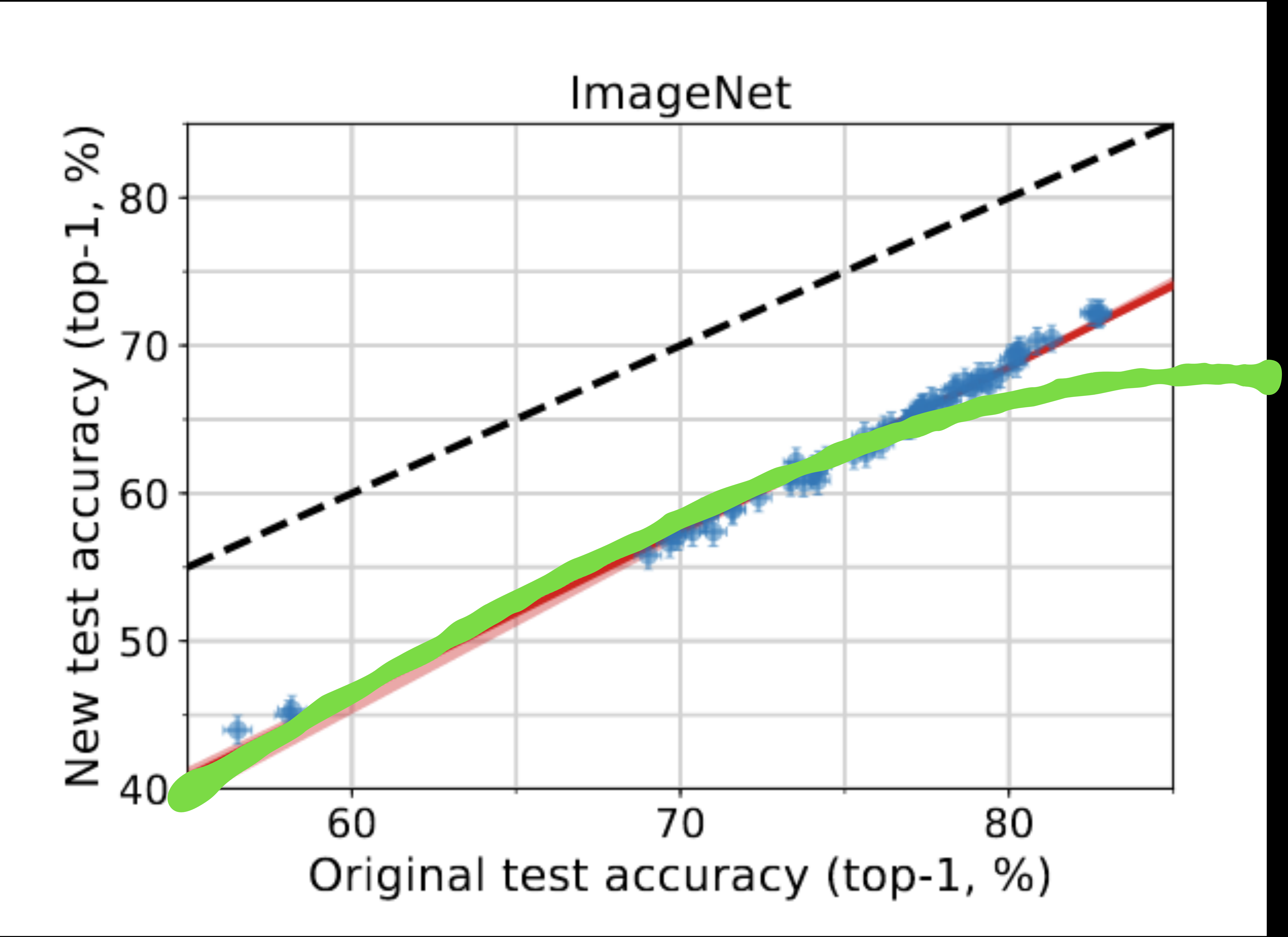
1. **It's just a harder dataset**
2. Adaptive overfitting
3. Distribution shift



ImageNet
Humans
New test accuracy (top-1, %)
Original test accuracy (top-1, %)
40
50
60
70
80
60
70
80

Possible explanations

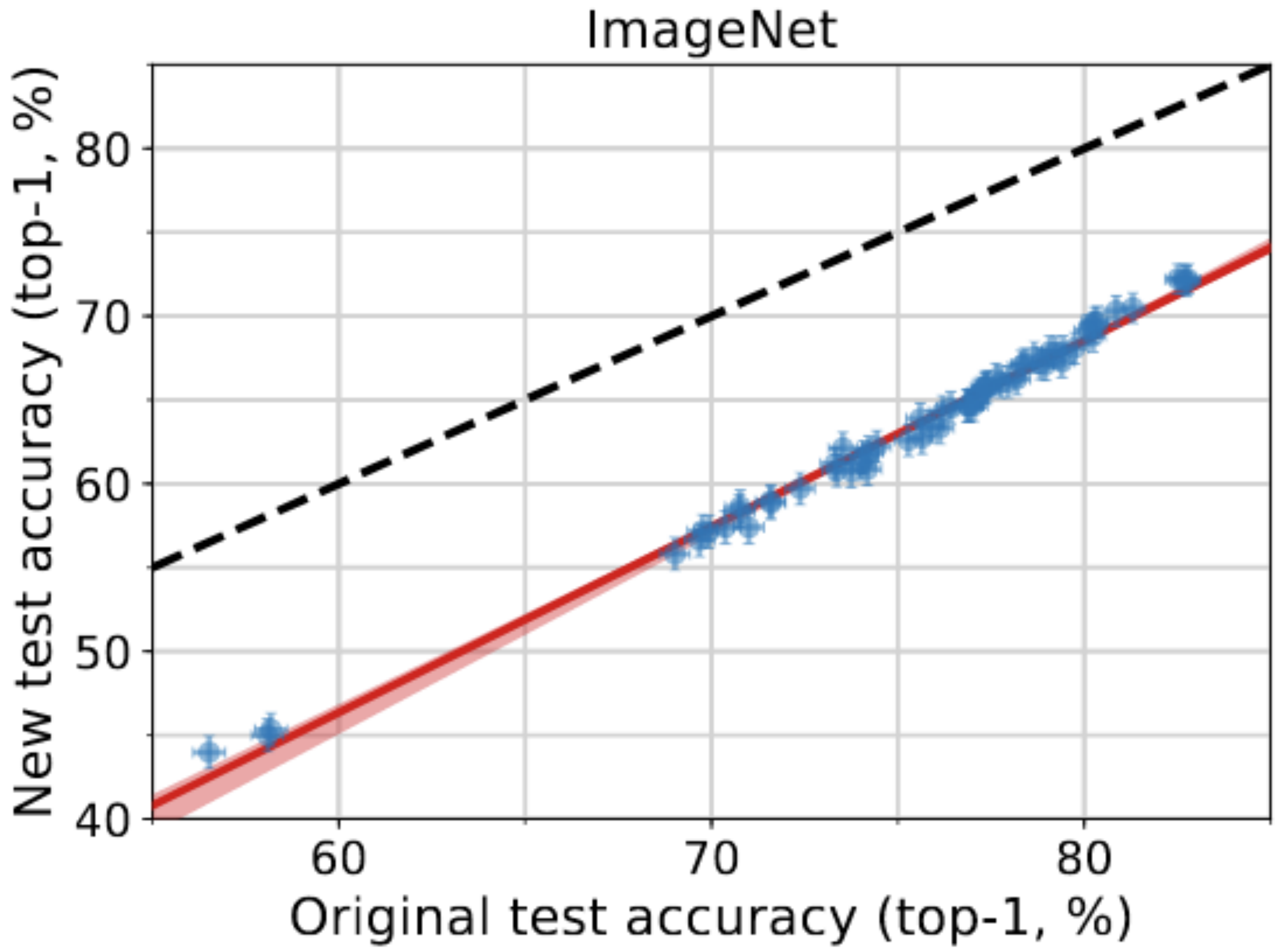
1. It's just a harder dataset
2. **Adaptive overfitting**
3. Distribution shift



Adaptive Overfitting

Possible explanations

1. It's just a harder dataset
2. Adaptive overfitting
3. **Distribution shift**


ImageNet
New test accuracy (top-1, %)
Original test accuracy (top-1, %)
40
50
60
70
80
60
70
80

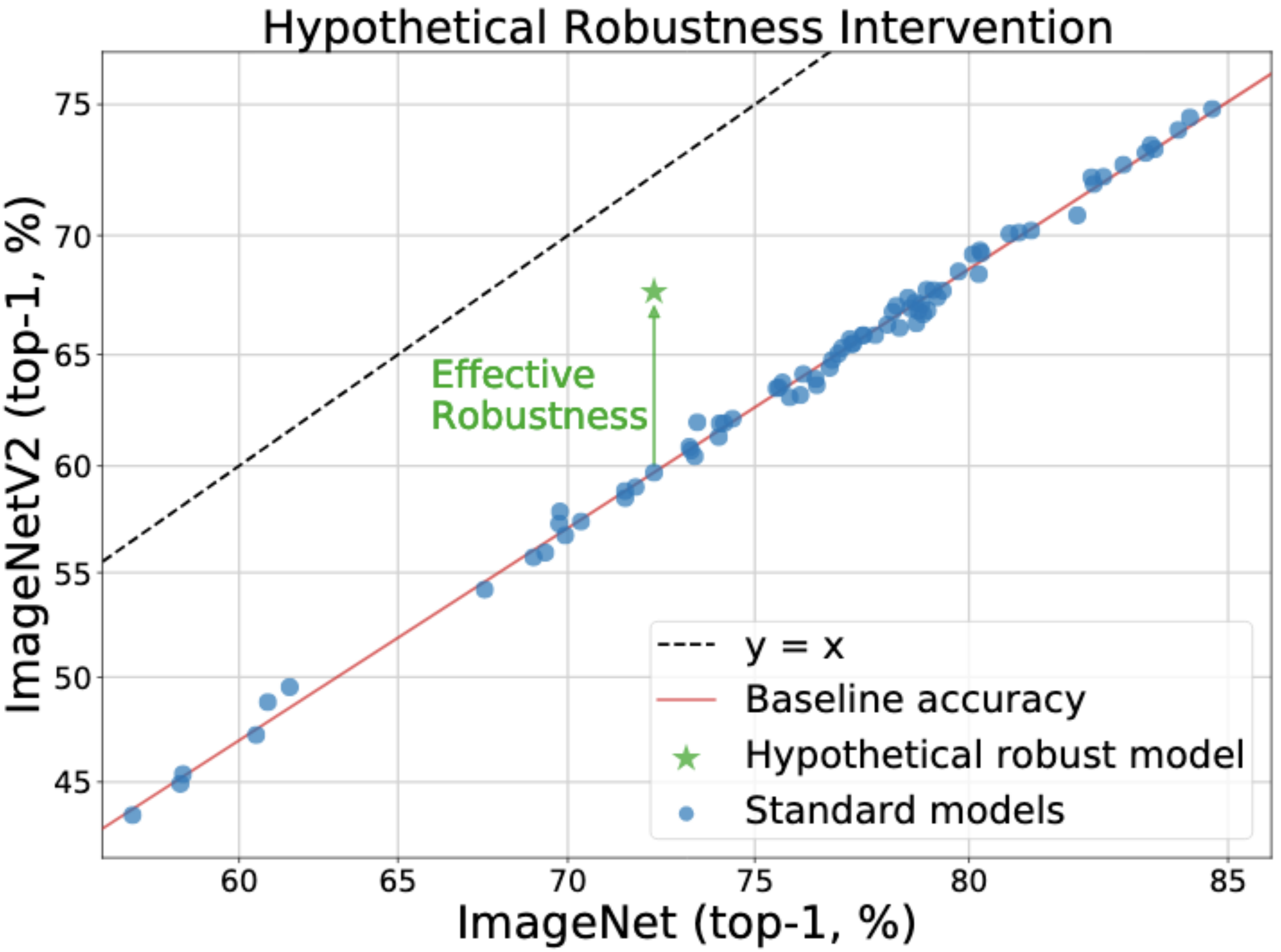
Are there any ways to increase robustness to this distribution shift?

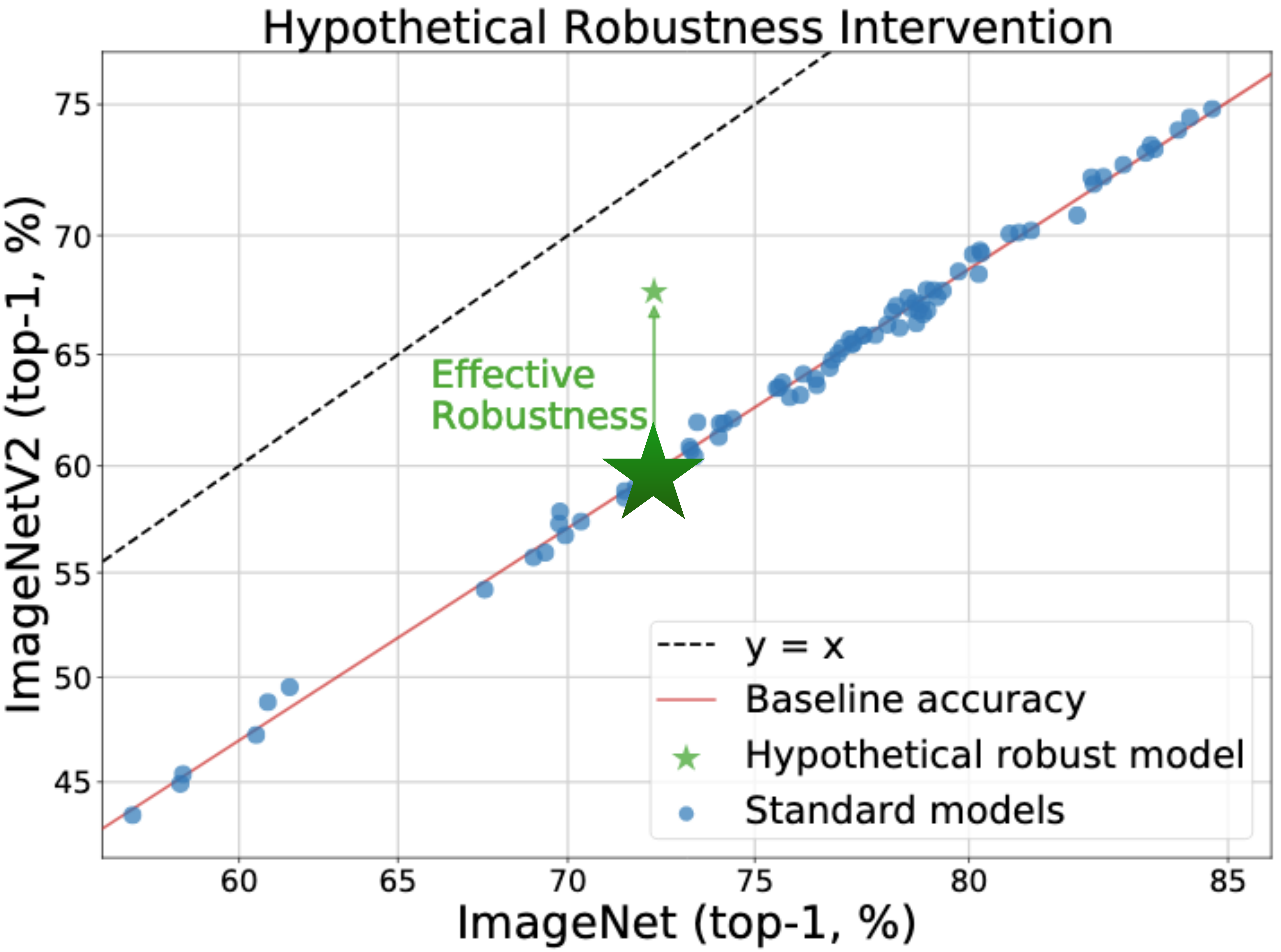
Our Approach:

BIG DATA

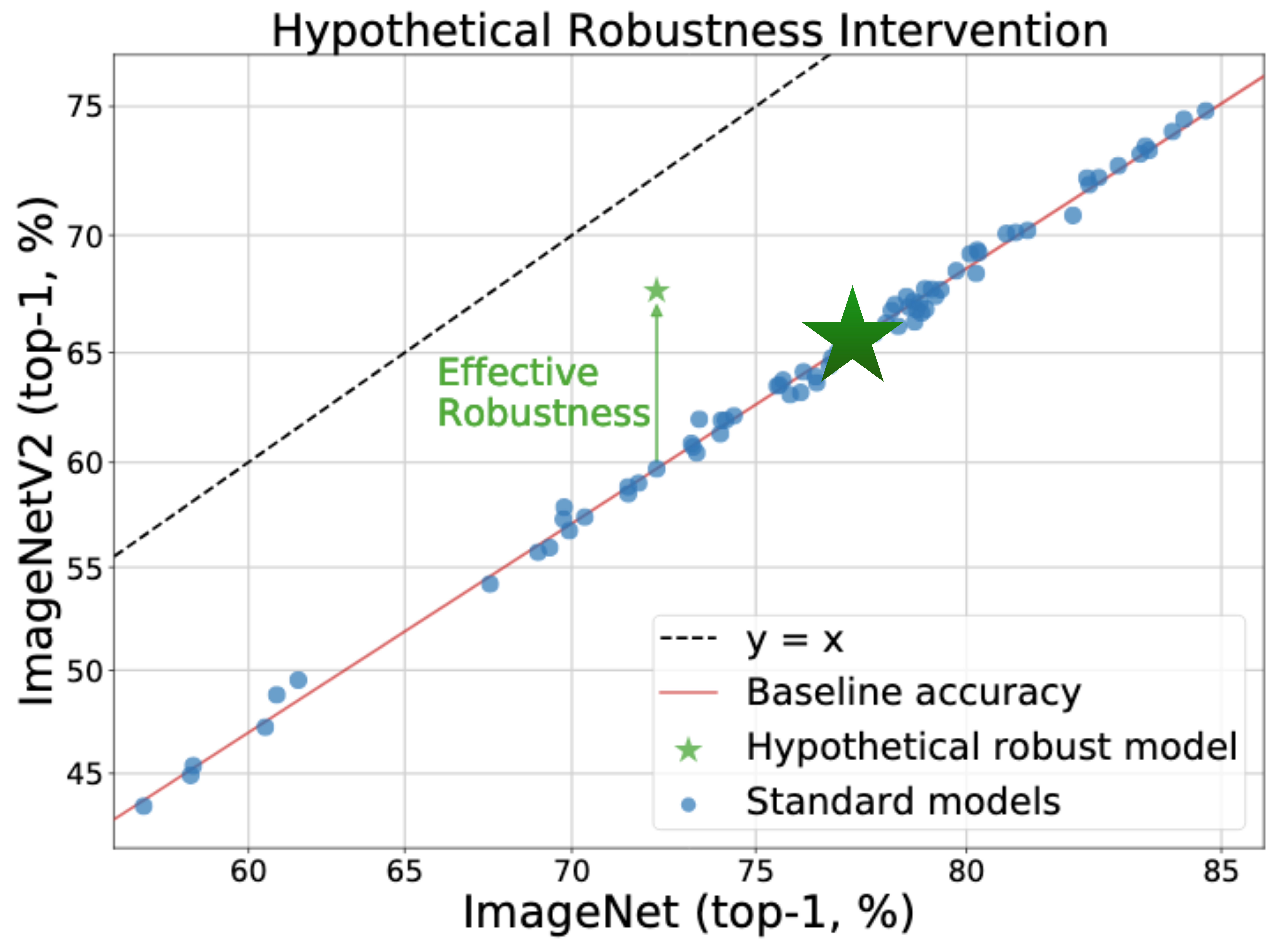
Formalization:

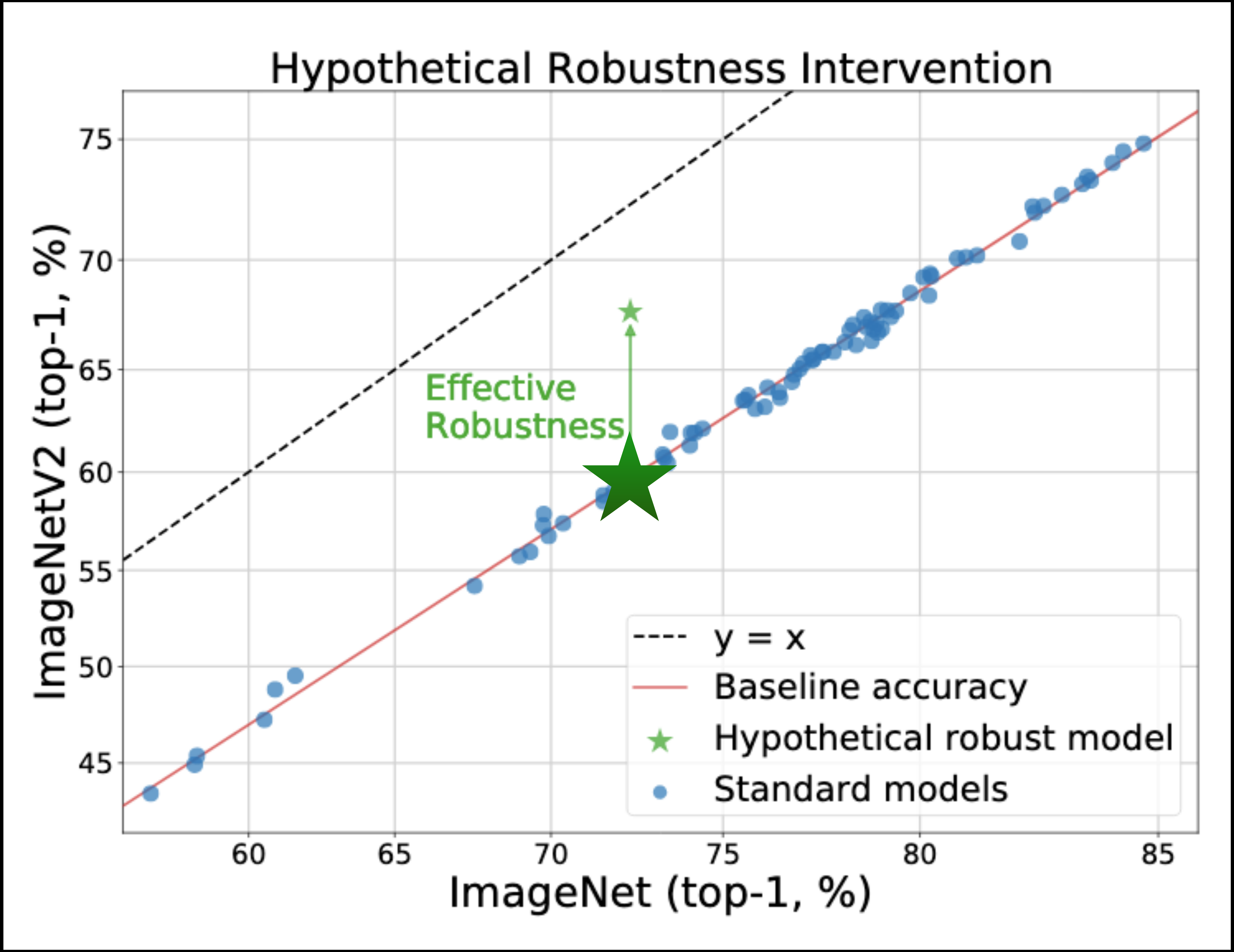
Effective Robustness

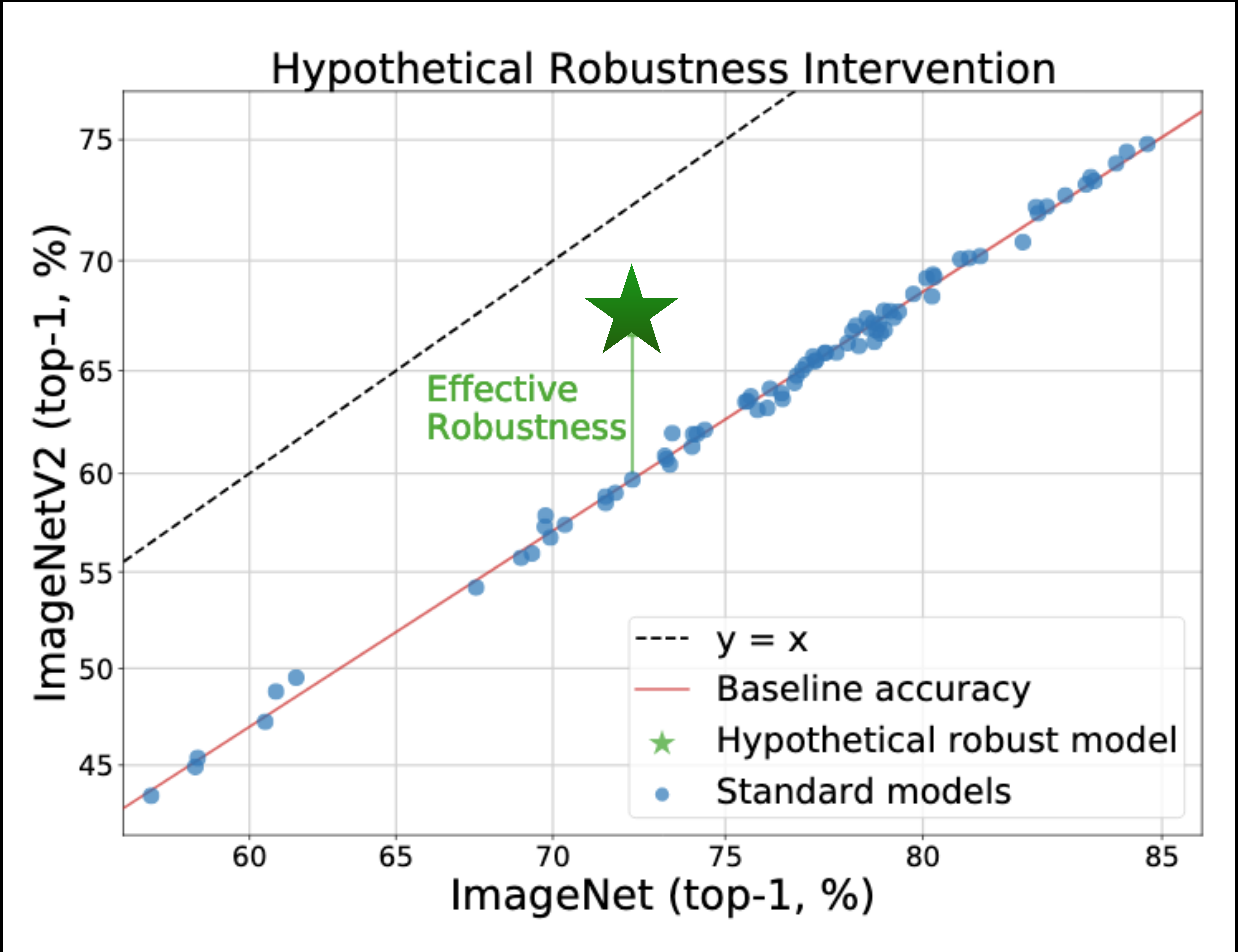

Hypothetical Robustness Intervention
ImageNetV2 (top-1, %)
ImageNet (top-1, %)
75
70
65
60
55
50
45
60
65
70
75
80
85
Effective Robustness
y = x
Baseline accuracy
Hypothetical robust model
Standard models



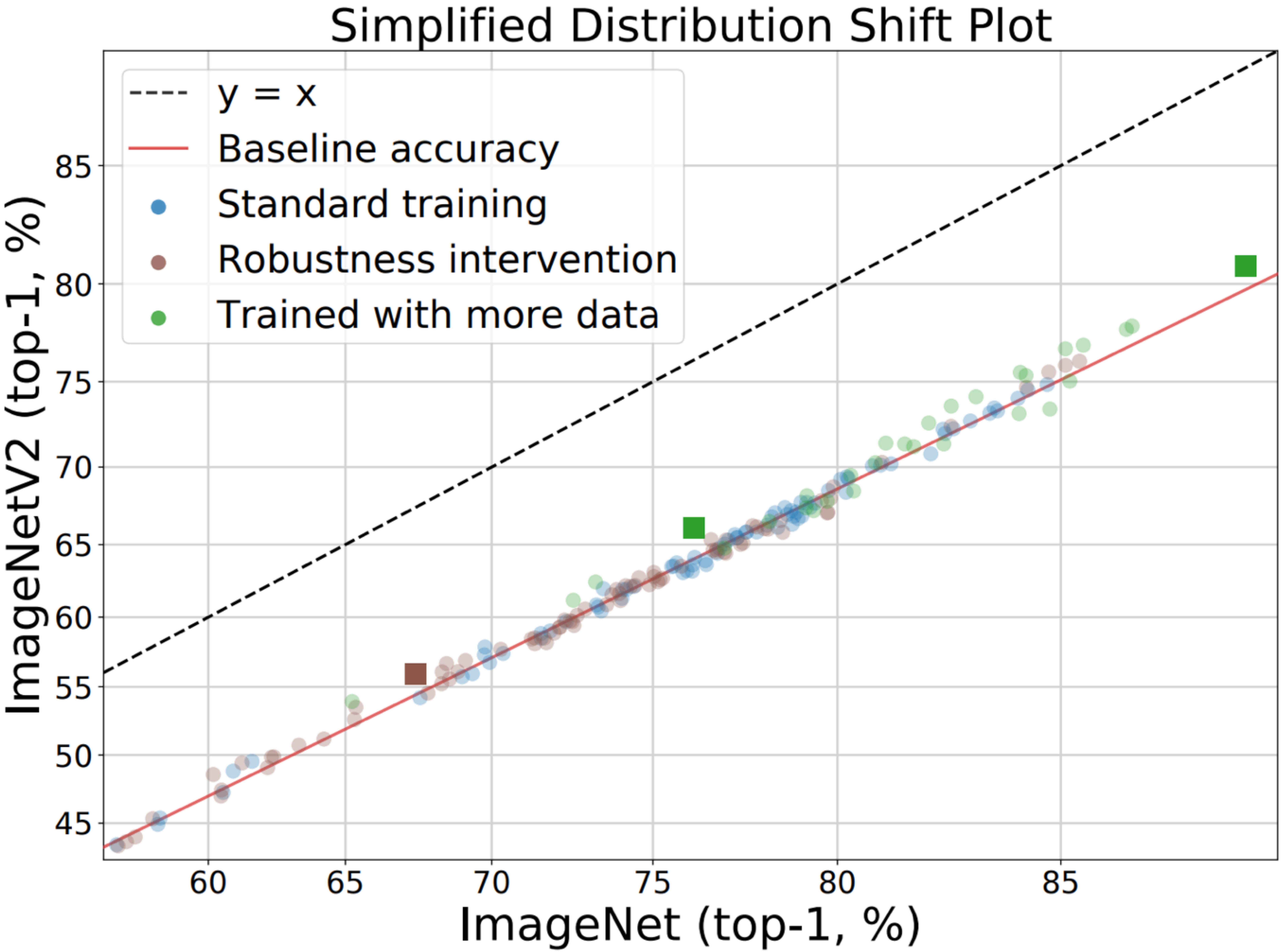
Hypothetical Robustness Intervention
ImageNetV2 (top-1, %)
ImageNet (top-1, %)
75
70
65
60
55
50
45
60
65
70
75
80
85
Effective
Robustness
y = x
Baseline accuracy
Hypothetical robust model
Standard models


Hypothetical Robustness Intervention
ImageNetV2 (top-1, %)
ImageNet (top-1, %)
Effective
Robustness
y = x
Baseline accuracy
Hypothetical robust model
Standard models
45
50
55
60
65
70
75
60
65
70
75
80
85

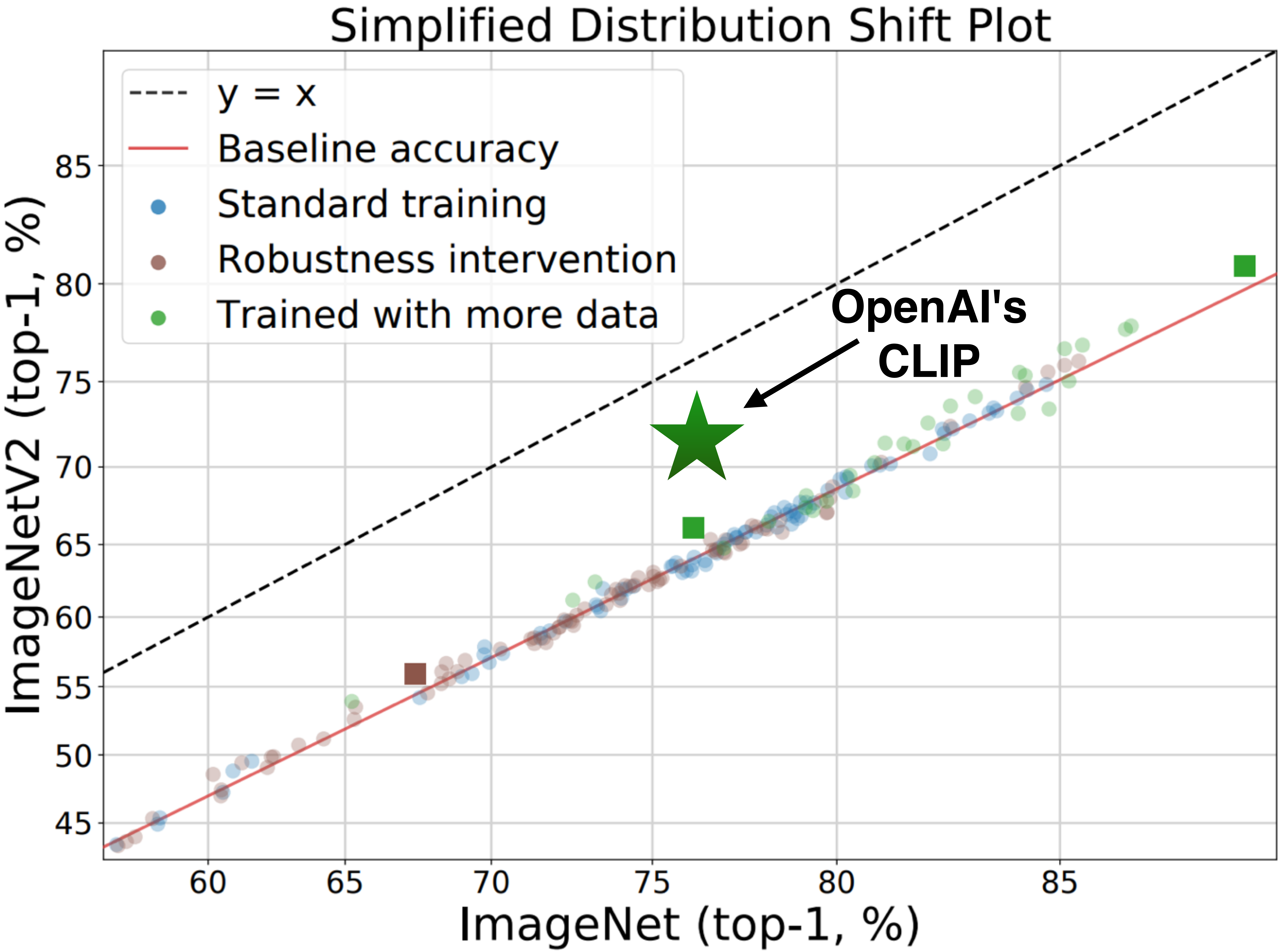

Hypothetical Robustness Intervention
ImageNetV2 (top-1, %)
ImageNet (top-1, %)
75
70
65
60
55
50
45
60
65
70
75
80
85
Effective
Robustness
y = x
Baseline accuracy
Hypothetical robust model
Standard models


Hypothetical Robustness Intervention
ImageNetV2 (top-1, %)
ImageNet (top-1, %)
Effective Robustness
y = x
Baseline accuracy
Hypothetical robust model
Standard models
45
50
55
60
65
70
75
60
65
70
75
80
85

So what helps?



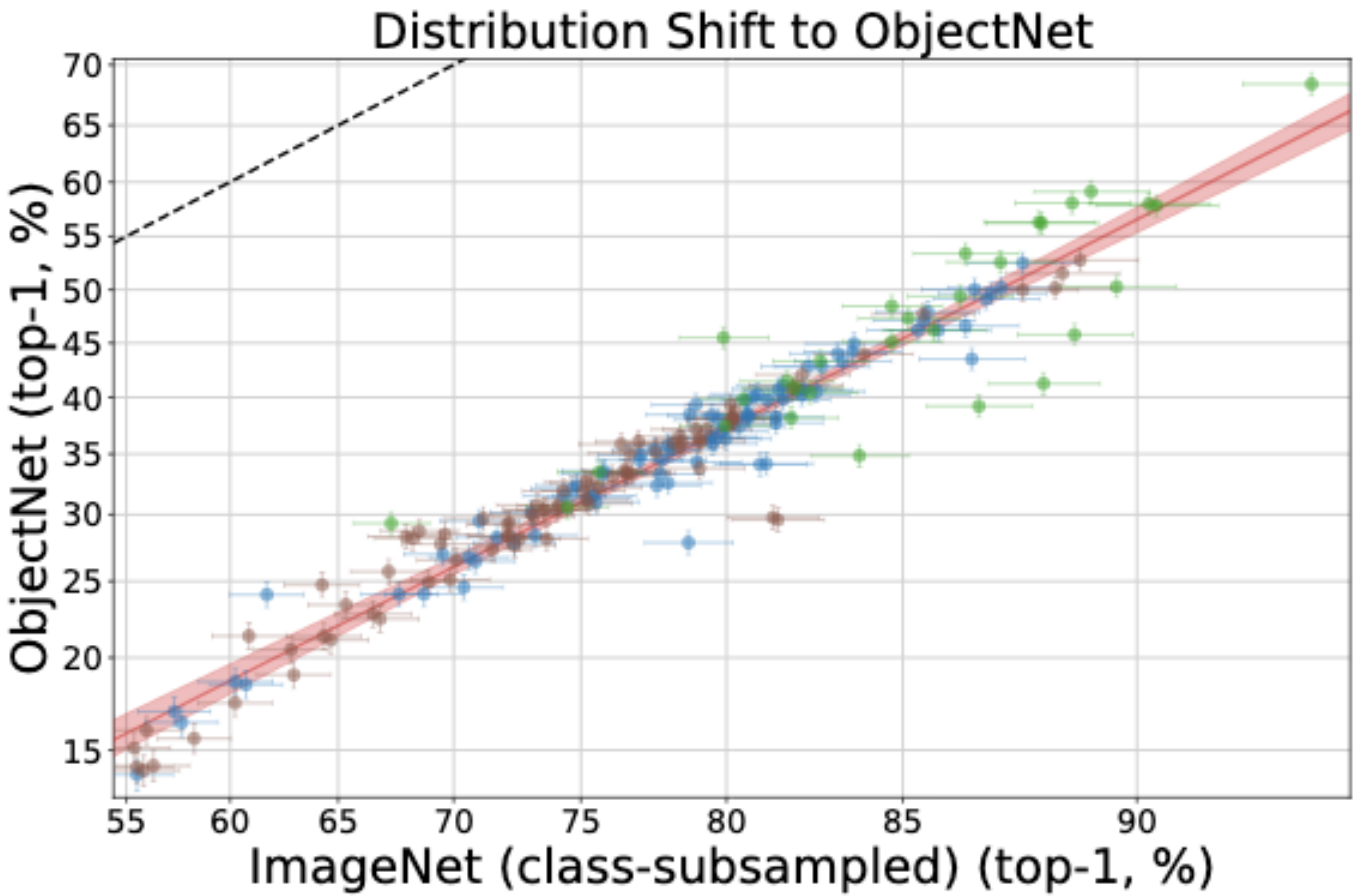
Simplified Distribution Shift Plot
y = x
Baseline accuracy
Standard training
Robustness intervention
Trained with more data
ImageNetV2 (top-1, %)
ImageNet (top-1, %)
45
50
55
60
65
70
75
80
85
60
65
70
75
80
85

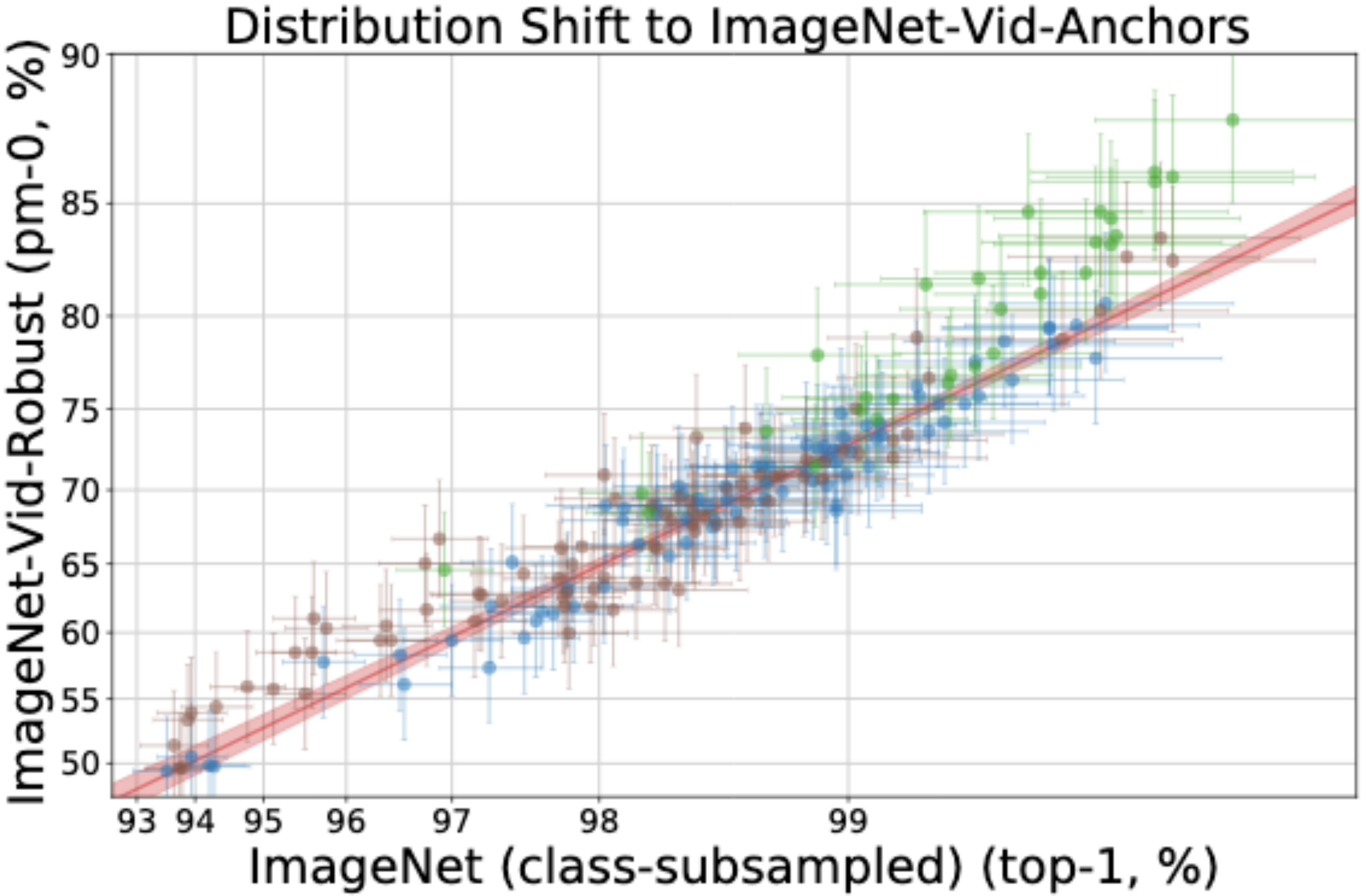


Simplified Distribution Shift Plot
y = x
Baseline accuracy
Standard training
Robustness intervention
Trained with more data
OpenAI's CLIP
ImageNetV2 (top-1, %)
ImageNet (top-1, %)
45
50
55
60
65
70
75
80
85
60
65
70
75
80
85

Possible explanations

1. It's just a harder dataset
2. Adaptive overfitting
3. Distribution shift
4. **It's just a weird dataset**


Distribution Shift to ObjectNet
ObjectNet (top-1, %)
ImageNet (class-subsampled) (top-1, %)
15
20
25
30
35
40
45
50
55
60
65
70
55
60
65
70
75
80
85
90



Distribution Shift to ImageNet-Vid-Anchors
ImageNet-Vid-Robust (pm-0, %)
90
85
80
75
70
65
60
55
50
93
94
95
96
97
98
99
ImageNet (class-subsampled) (top-1, %)

Possible explanations

1. It's just a harder dataset
2. Adaptive overfitting
3. Distribution shift
4. **It's just a weird dataset**

Does synthetic

=>

natural robustness

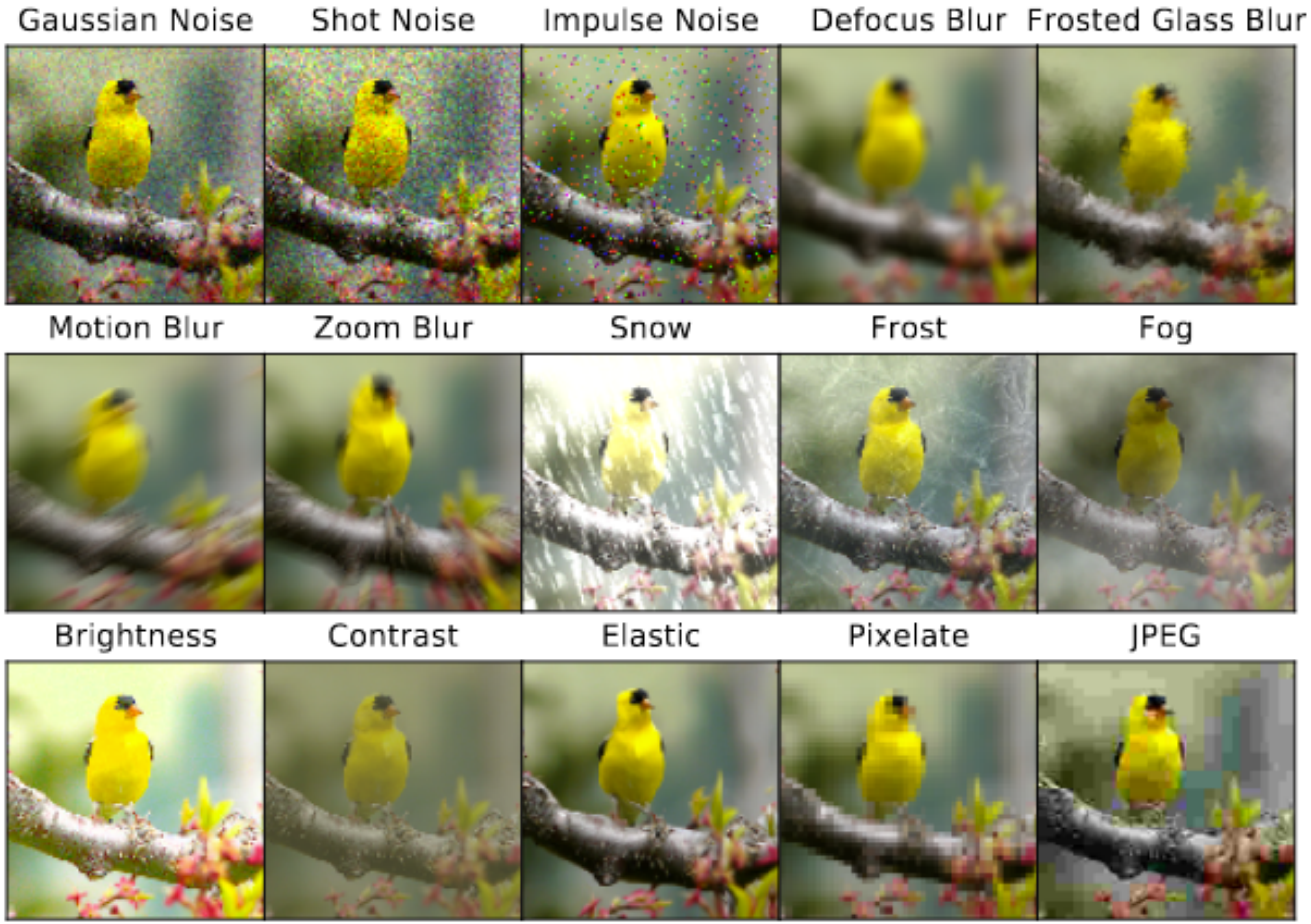
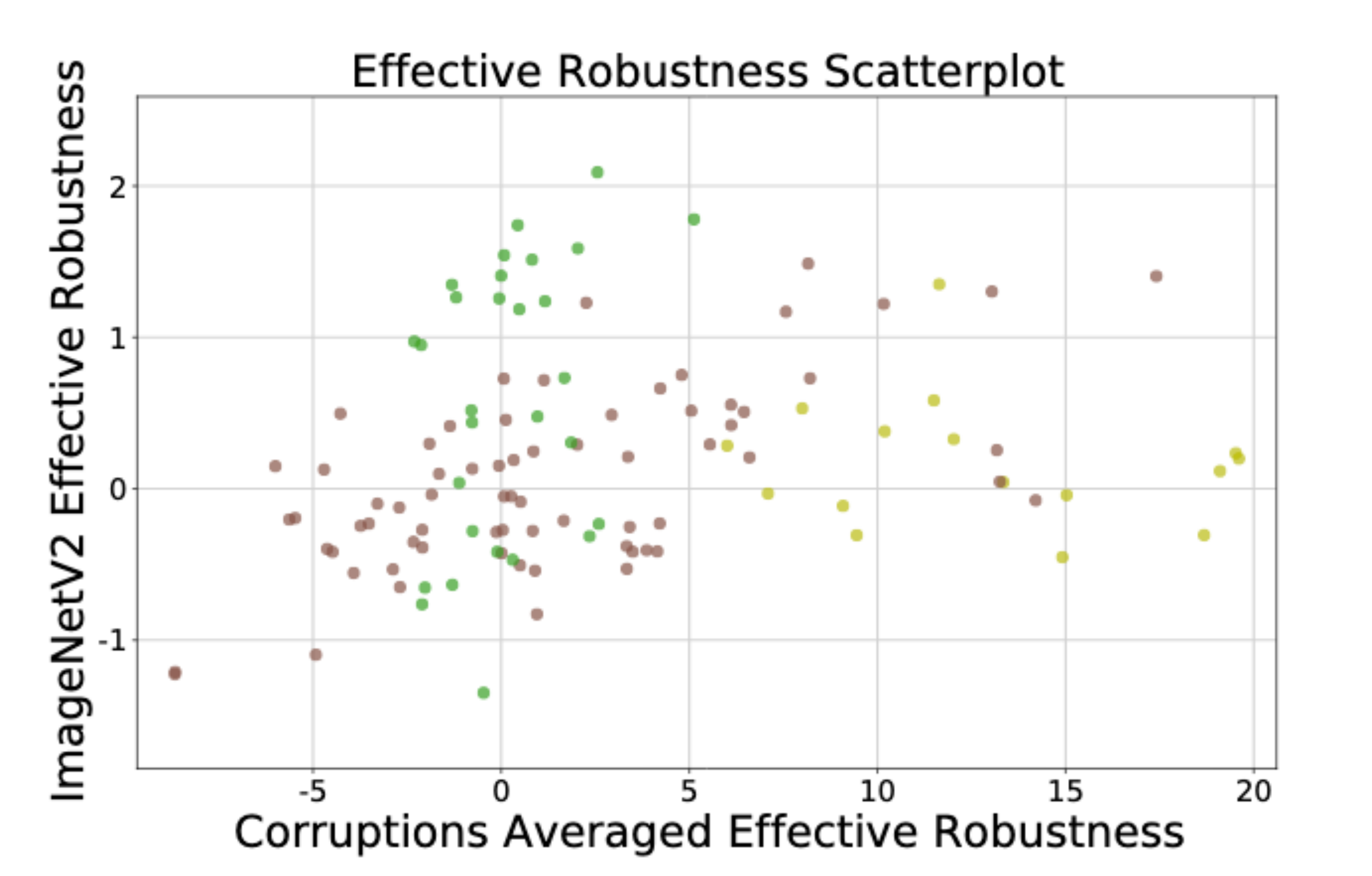


Figure 1: Our IMAGENET-C dataset consists of 15 types of algorithmically generated corruptions from noise, blur, weather, and digital categories. Each type of corruption has five levels of severity, resulting in 75 distinct corruptions. See different severity levels in Appendix B.


Effective Robustness Scatterplot
ImageNetV2 Effective Robustness
Corruptions Averaged Effective Robustness
2
1
0
-1
-5
0
5
10
15
20

If you wan to increase robustness, you can ...

1. Train on more data
2. Train on the distribution shift you care about

If you wan to increase robustness, you can ...

1. Train on more data
2. Train on the distribution shift you care about
3. Train on the distribution shift you care about

If you wan to increase robustness, you can ...

1. Train on more data
2. Train on the distribution shift you care about
3. Train on the distribution shift you care about
4. Train on the distribution shift you care about

And this is what makes adversarial/natural shifts hard to solve

Neural networks are (still) not robust

nicholas@carlini.com https://nicholas.carlini.com