Robust Fine-tuning of Zero-shot Models

Ludwig Schmidt

(Samir Gadre filling in)

UNIVERSITY of WASHINGTON

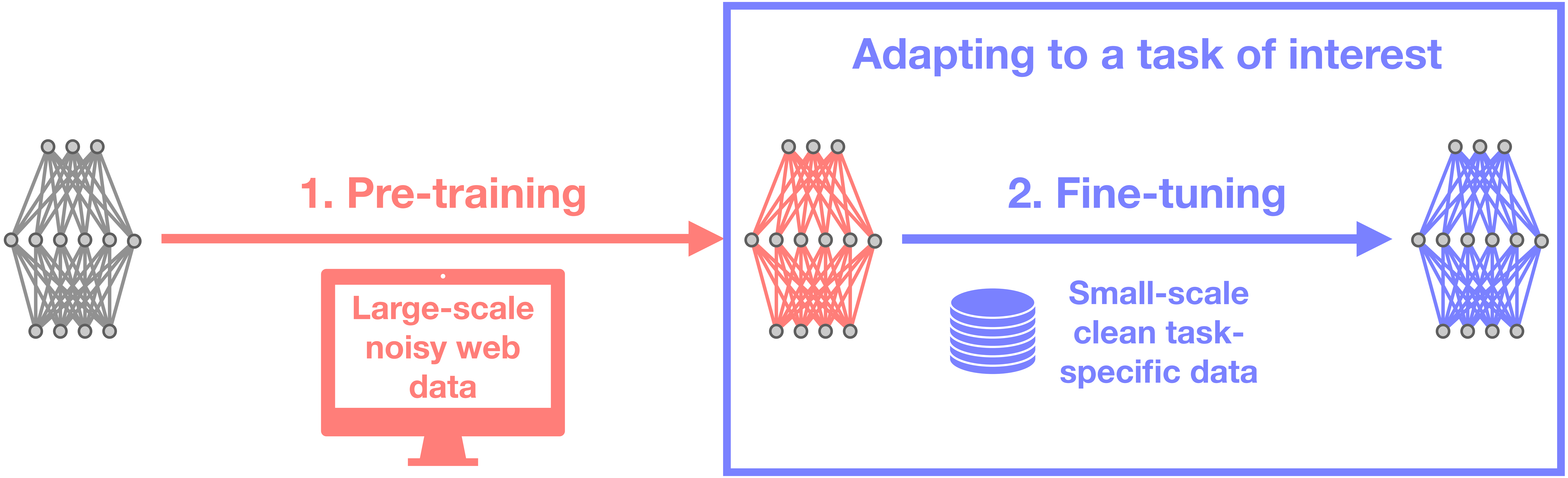
Ai2

LAION

Stanford University

Fine-tuning vs. zero-shot inference

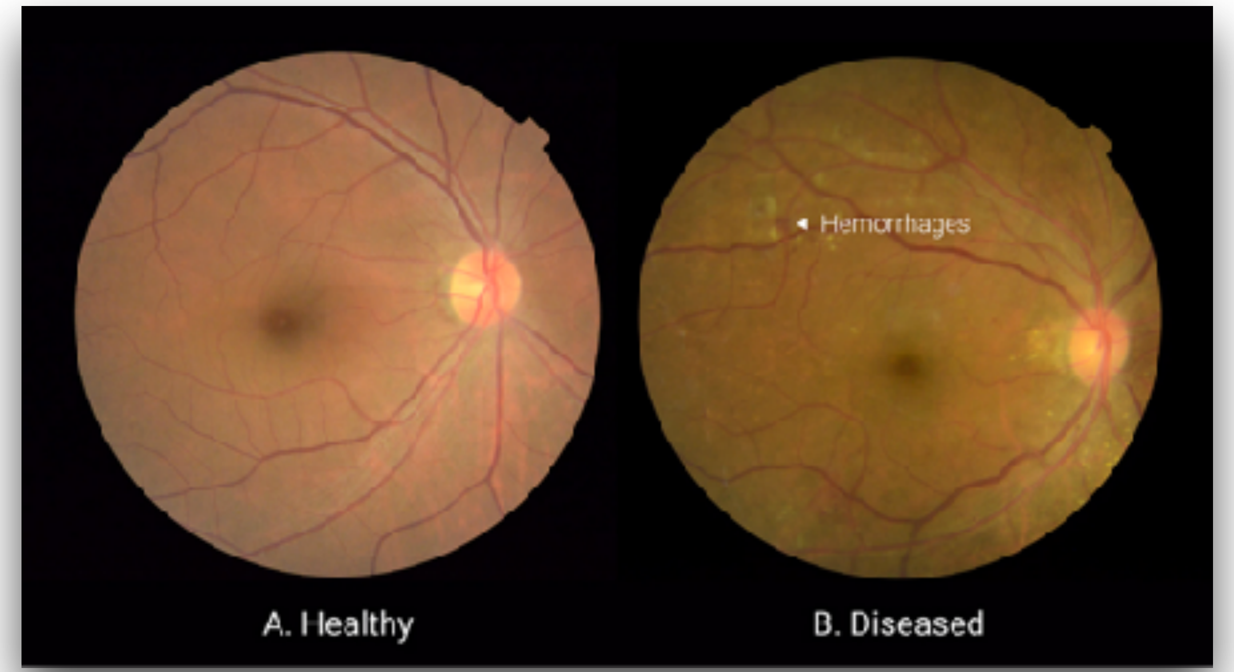
State-of-the-art ML models often come from a **two-step process**.



What is the best way to fine-tune a large pre-trained model?

Focus today: out-of-distribution robustness

Transportation



Health care

Robotics

Chat assistants

➡ Need **reliable** machine learning

Robustness on ImageNet

Lots of progress on ImageNet over the past 10 years, but models are still not robust.

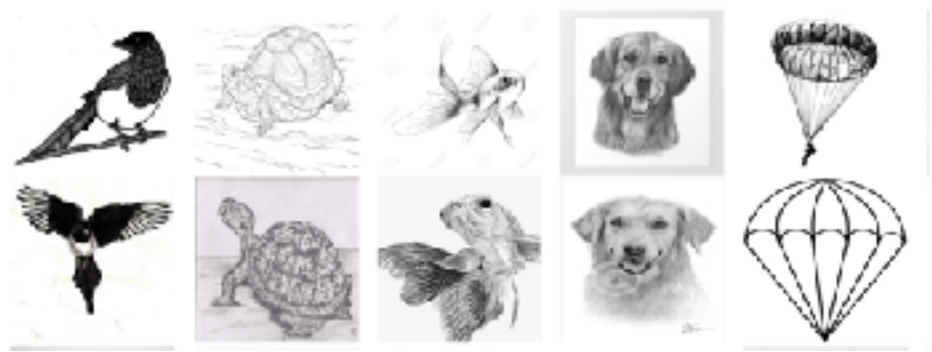
Evaluation: **new test sets**

ImageNetV2

[Recht, Roelofs, Schmidt, Shankar ’19]

ObjectNet

[Barbu, Mayo, Alverio, Luo, Wang, Gutfreund, Tenenbaum, Katz ’19]



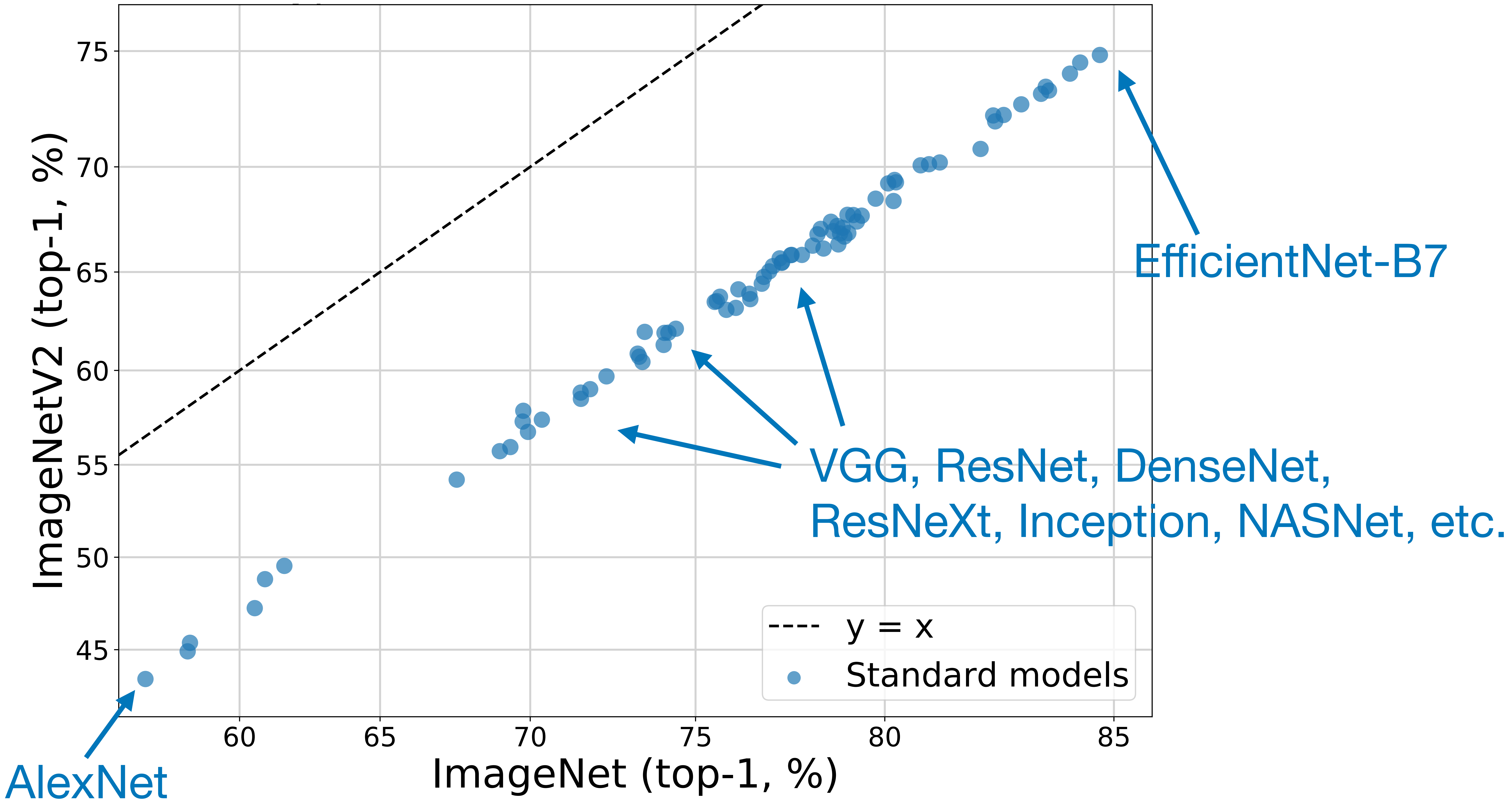
ImageNet-Sketch

[Wang, Ge, Lipton, Xing ’19]

ImageNet-R

[Hendrycks, Basart, Mu, Kadavath, Wang, Dorundo, Desai, Zhu, Parajuli, Guo, Song, Steinhardt, Gilmer ’20]

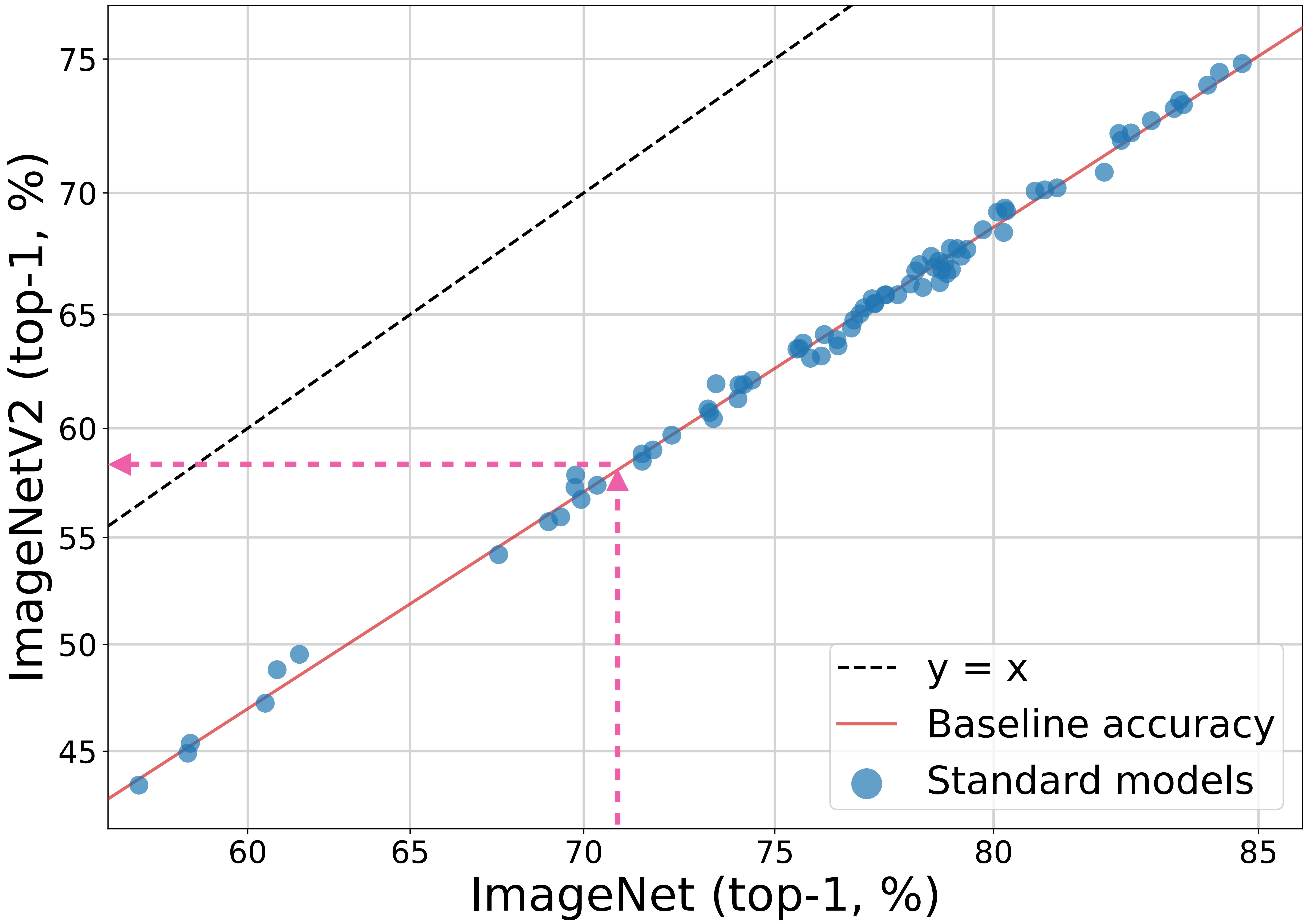
What robustness interventions help?



[Taori, Dave, Shankar, Carlini, Recht, Schmidt ’20]

What robustness interventions help?

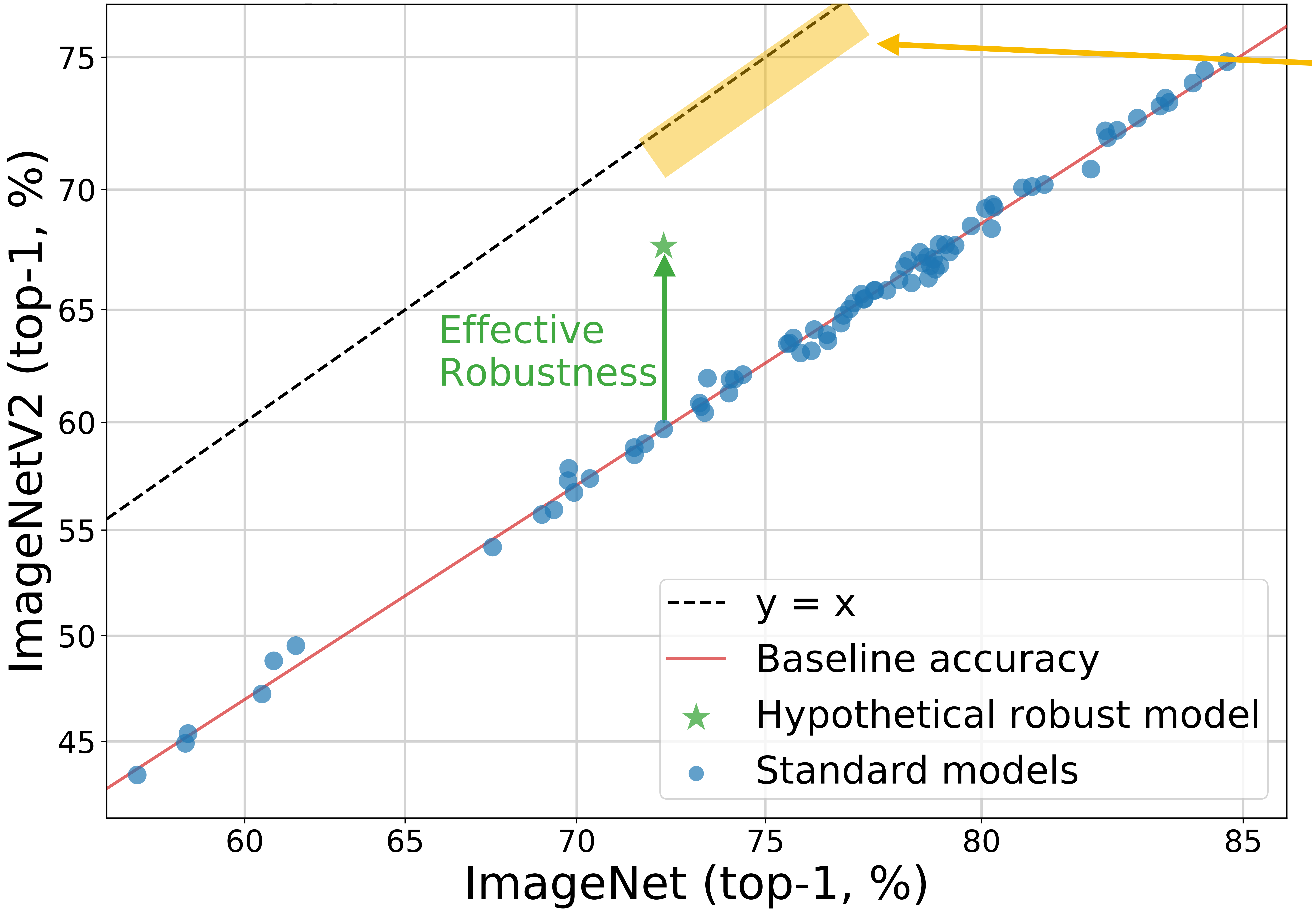
Expected out-of-distribution accuracy



In-distribution accuracy

Baseline **out-of-distribution accuracy** from **in-distribution accuracy.**

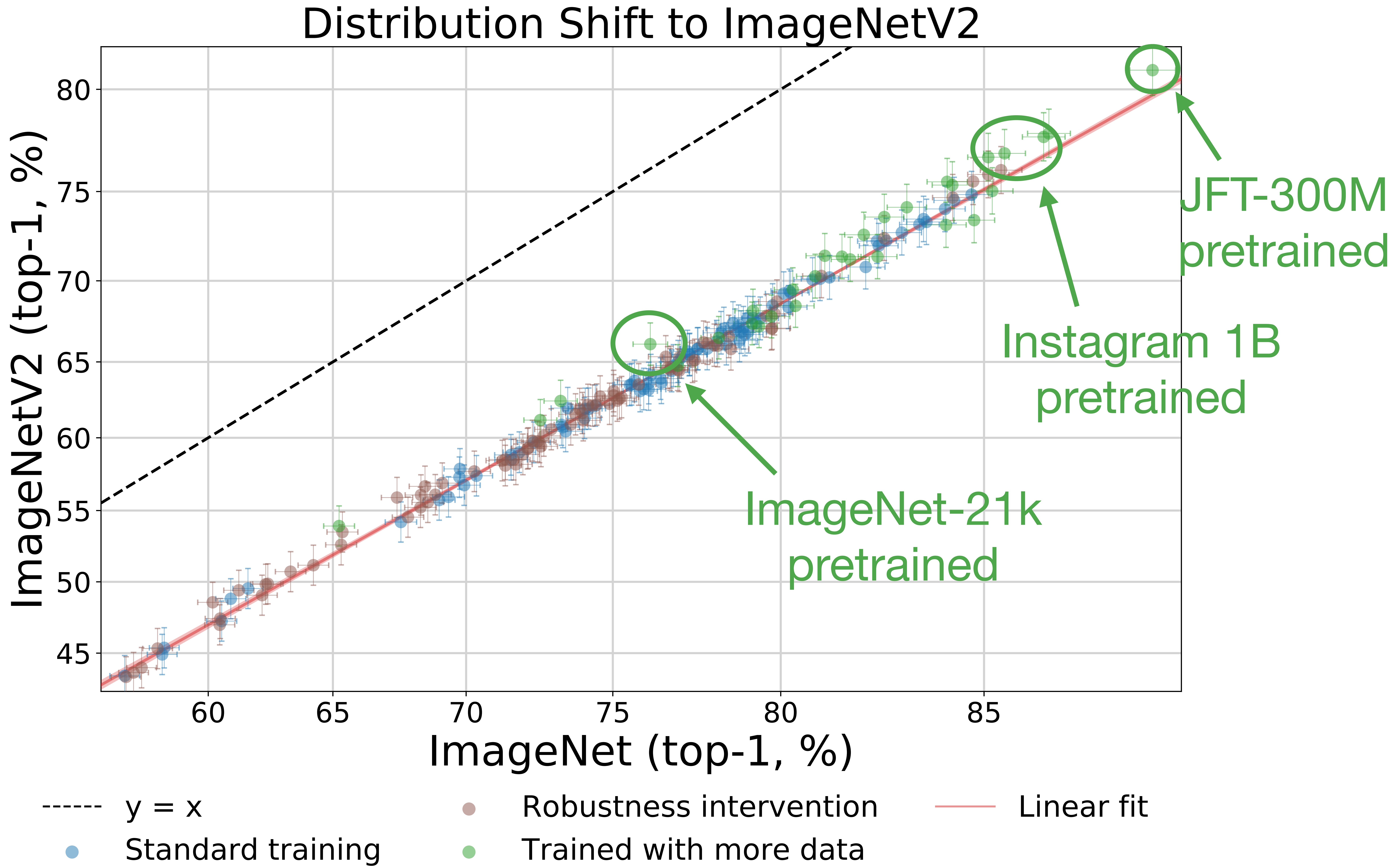
What robustness interventions help?



Humans

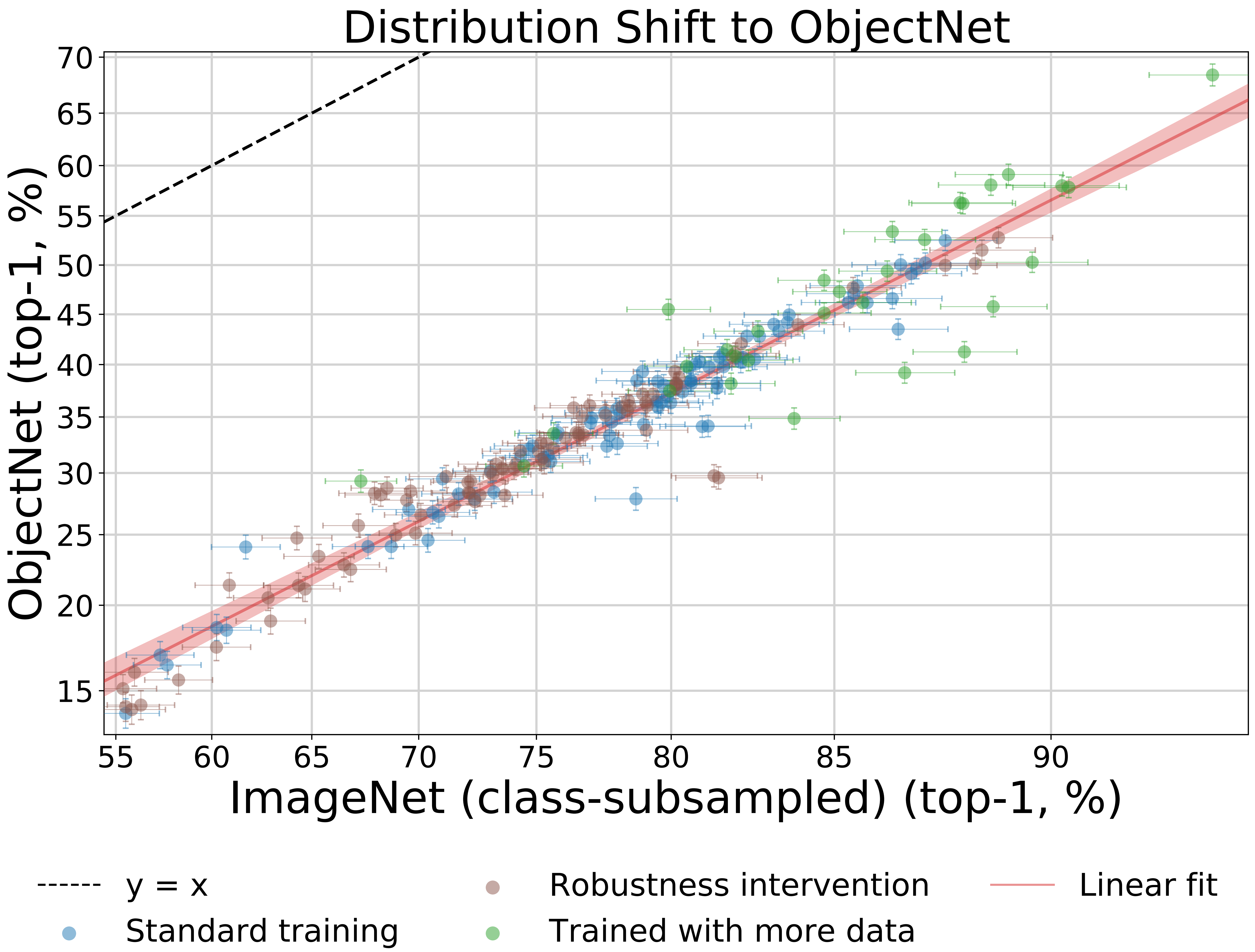
[Shankar, Roelofs, Mania, Fang, Recht, Schmidt ’20]

Do current robustness interventions achieve **effective robustness**?



➡ No current **robustness technique** achieves non-trivial effective robustness.

➡ Only training on (a lot) **more data** gives a small amount of effective robustness.



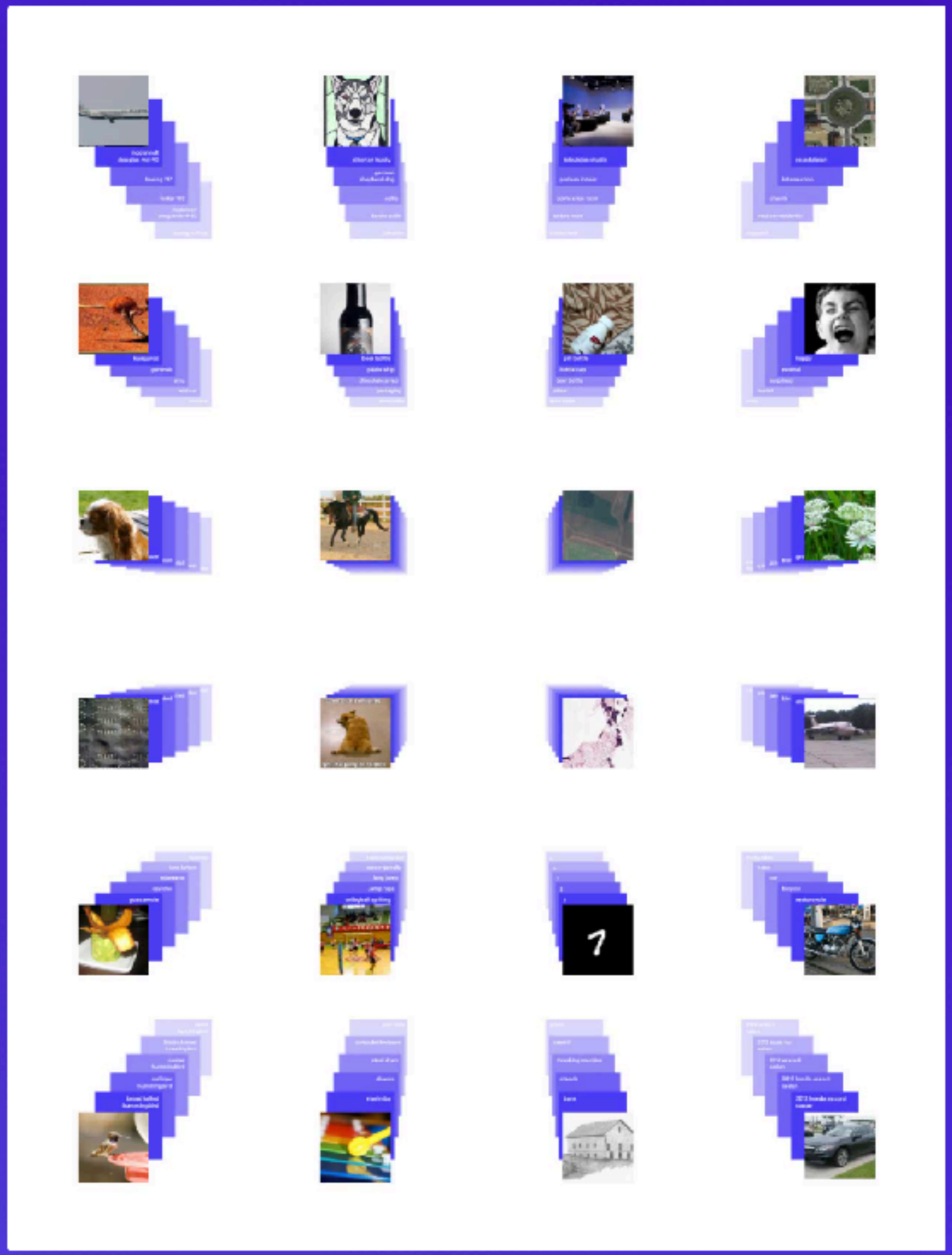
[Barbu, Mayo, Alverio, Luo, Wang, Gutfreund, Tenenbaum, Katz '19]

Same trend: only **more data** gives effective robustness.

OpenAI API PROJECTS BLOG ABOUT

CLIP: Connecting Text and Images

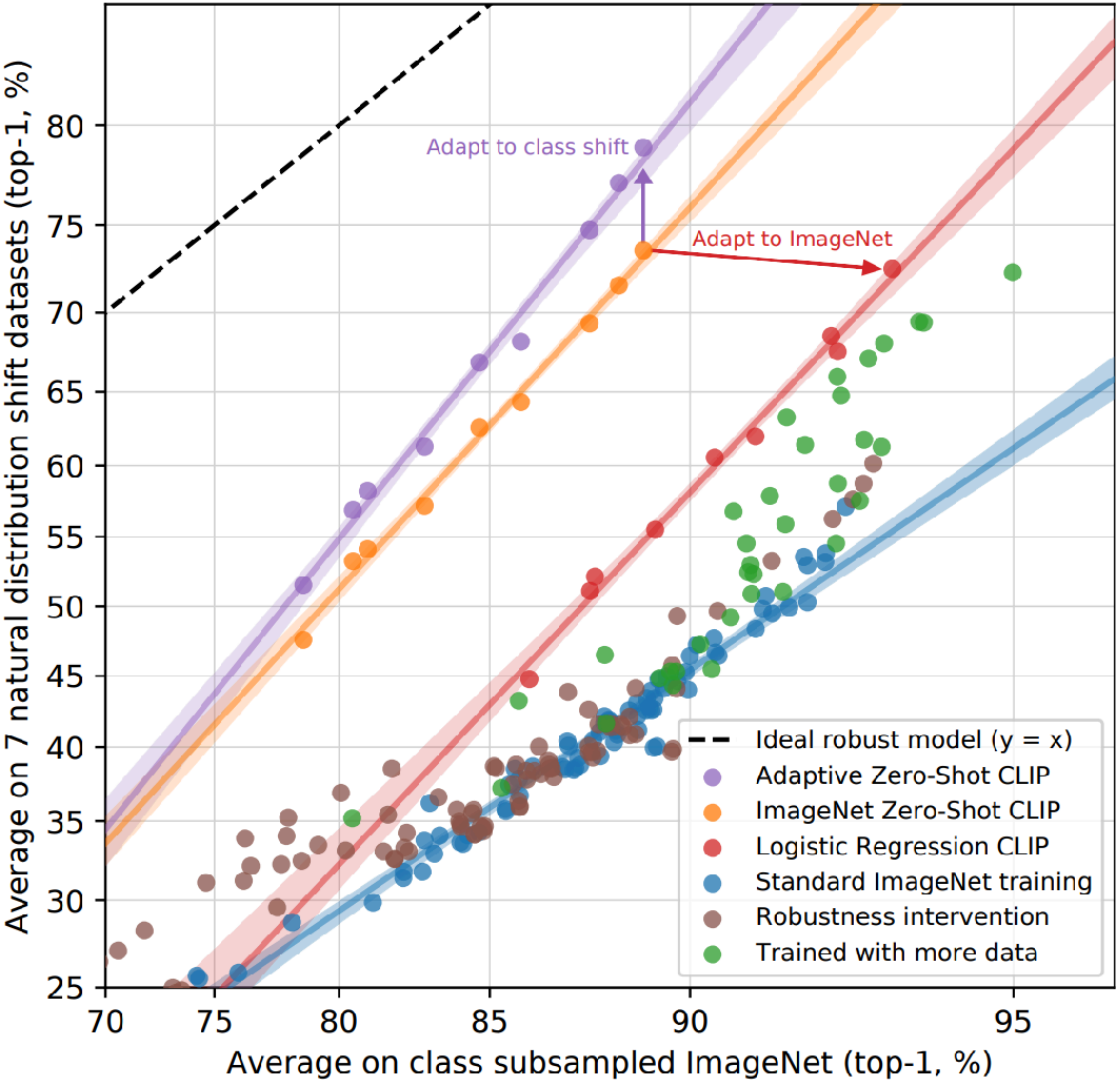
We're introducing a neural network called CLIP which efficiently learns visual concepts from natural language supervision. CLIP can be applied to any visual classification benchmark by simply providing the names of the visual categories to be recognized, similar to the "zero-shot" capabilities of GPT-2 and GPT-3.



January 5, 2021
15 minute read

DATASET	IMAGENET RESNET101	CLIP VIT-L	Effective robustness
 ImageNet	76.2%	76.2%	
 ImageNet V2	64.3%	70.1%	+6%
 ImageNet Rendition	37.7%	88.9%	+51%
 ObjectNet	32.6%	72.3%	+40%
 ImageNet Sketch	25.2%	60.2%	+35%
 ImageNet A	2.7%	77.1%	+74%

Very large improvements in out-of-distribution robustness.



[Radford, Kim, Hallacy, Ramesh, Goh, Agarwal, Sastry, Askell, Mishkin, Clark, Krueger, Sutskever '21]

Large robustness gains

What makes CLIP robust?

But: fine-tuning reduces robustness

Can we get **both** high in-distribution **and** out-of-distribution accuracy?

What makes CLIP robust?

Data Determines Distributional Robustness in Contrastive Language Image Pre-training (CLIP)

Alex Fang† Gabriel Ilharco† Mitchell Wortsman† Yuhao Wan†

Vaishaal Shankar◇ Achal Dave◇ Ludwig Schmidt†○

Abstract

Contrastively trained image-text models such as CLIP, ALIGN, and BASIC have demonstrated unprecedented robustness to multiple challenging natural distribution shifts. Since these image-text models differ from previous training approaches in several ways, an important question is what causes the large robustness gains. We answer this question via a systematic experimental investigation. Concretely, we study five different possible causes for the robustness gains: (i) the training set size, (ii) the training distribution, (iii) language supervision at training time, (iv) language supervision at test time, and (v) the contrastive loss function. Our experiments show that the more diverse training distribution is the main cause for the robustness gains, with the other factors contributing little to no robustness. Beyond our experimental results, we also introduce ImageNet-Captions, a version of ImageNet with original text annotations from Flickr, to enable further controlled experiments of language-image training.

Hypotheses for CLIP's robustness

	CLIP	Standard ImageNet supervised learning
Language supervision	Yes	No
Training distribution	???	ImageNet
Training set size	400M	1.2M
Loss function	Contrastive	Supervised
Test-time prompting	Yes	No
Model architecture	ViTs	CNNs

Hypotheses for CLIP's robustness

	CLIP	Standard ImageNet supervised learning
~~**Language supervision**~~	~~Yes~~	~~No~~
Training distribution	???	ImageNet
~~**Training set size**~~	~~400M~~	~~1.2M~~
~~**Loss function**~~	~~Contrastive~~	~~Supervised~~
~~**Test-time prompting**~~	~~Yes~~	~~No~~
~~**Model architecture**~~	~~ViTs~~	~~CNNs~~

One takeaway: datasets are a key for improving models

DataComp:
In search of the next generation of multimodal datasets

Samir Yitzhak Gadre*[2] Gabriel Ilharco*[1] Alex Fang*[1] Jonathan Hayase[1] Georgios Smyrnis[5]
Thao Nguyen[1] Ryan Marten[7,9] Mitchell Wortsman[1] Dhruba Ghosh[1] Jieyu Zhang[1]
Eyal Orgad[3] Rahim Entezari[10] Giannis Daras[5] Sarah Pratt[1] Vivek Ramanujan[1]
Yonatan Bitton[11] Kalyani Marathe[1] Stephen Mussmann[1] Richard Vencu[6]
Mehdi Cherti[6,8] Ranjay Krishna[1] Pang Wei Koh[1,12] Olga Saukh[10] Alexander Ratner[1,13]
Shuran Song[2] Hannaneh Hajishirzi[1,7] Ali Farhadi[1] Romain Beaumont[6]
Sewoong Oh[1] Alexandros G. Dimakis[5] Jenia Jitsev[6,8]
Yair Carmon[3] Vaishaal Shankar[4] Ludwig Schmidt[1,6,7]

Abstract

Multimodal datasets are a critical component in recent breakthroughs such as Stable Diffusion and GPT-4, yet their design does not receive the same research attention as model architectures or training algorithms. To address this shortcoming in the ML ecosystem, we introduce DataComp, a testbed for dataset experiments centered around a new candidate pool of 12.8 billion image-text pairs from Common Crawl. Participants in our benchmark design new filtering techniques or

Workshop tomorrow at ICCV!

Can we fine-tune CLIP without losing robustness?

Robust fine-tuning of zero-shot models

Mitchell Wortsman[*†] Gabriel Ilharco[*†] Jong Wook Kim[§] Mike Li[‡]

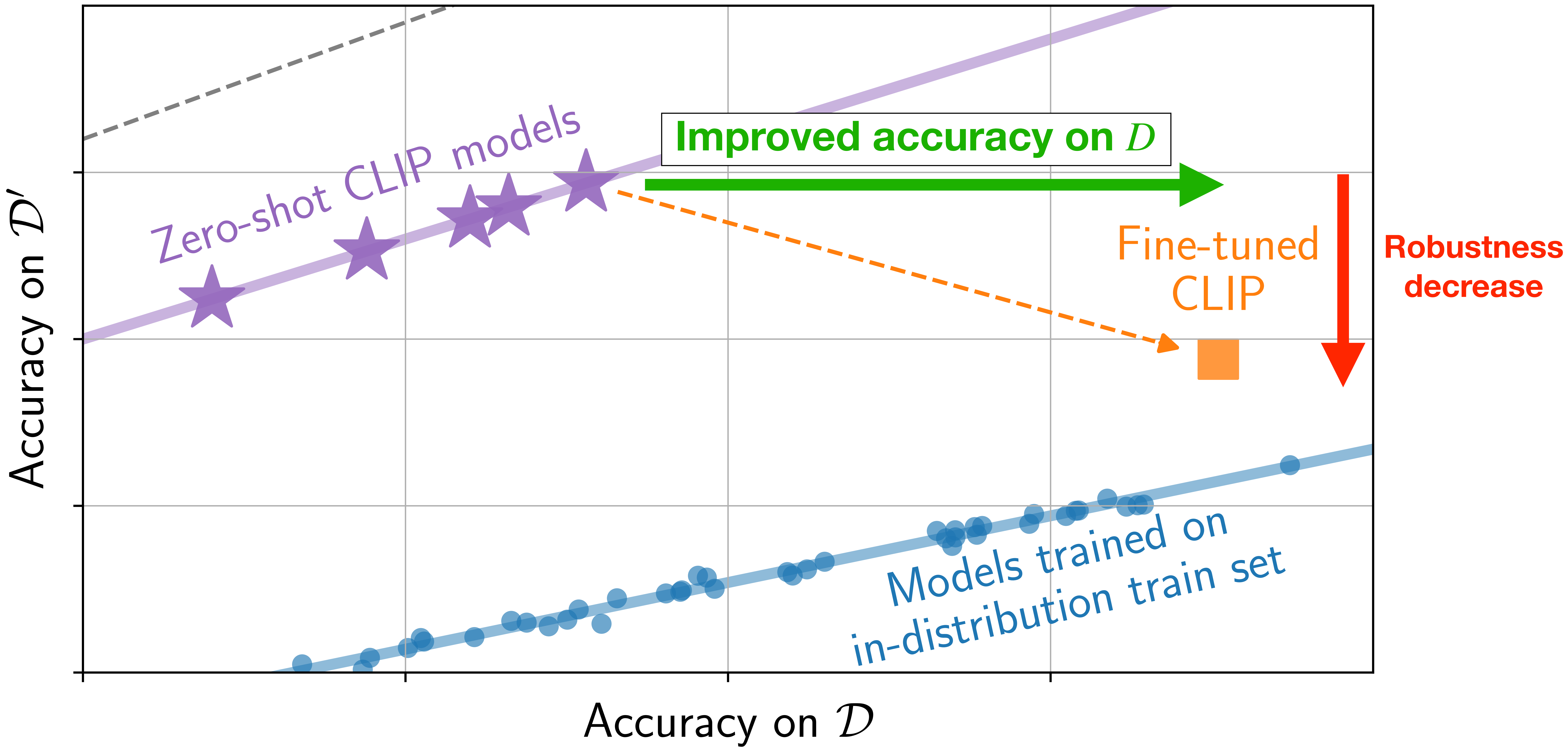
Simon Kornblith[◇] Rebecca Roelofs[◇] Raphael Gontijo-Lopes[◇]

Hannaneh Hajishirzi[†○] Ali Farhadi[⋆†] Hongseok Namkoong[⋆‡] Ludwig Schmidt[†△]

Abstract

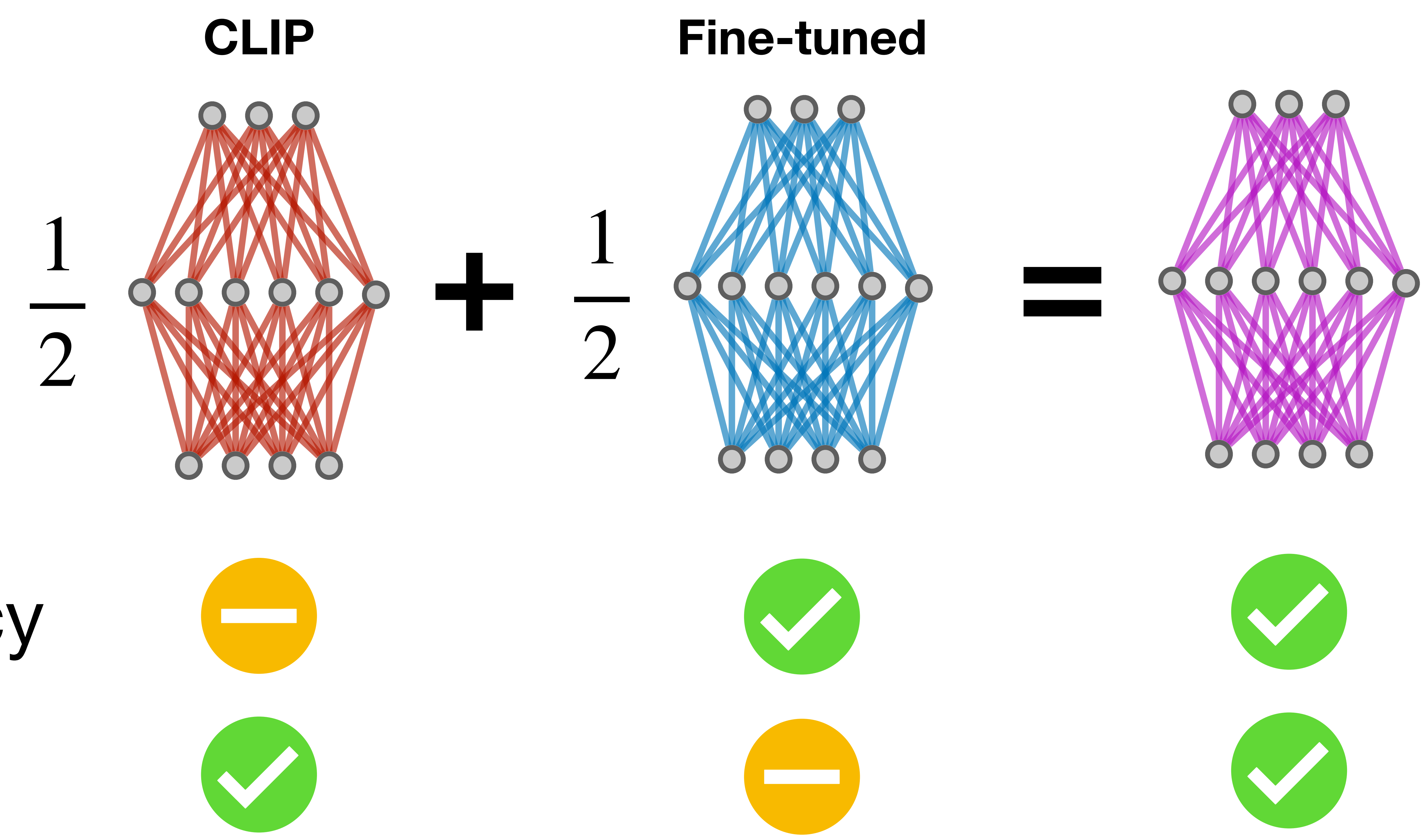
Large pre-trained models such as CLIP or ALIGN offer consistent accuracy across a range of data distributions when performing zero-shot inference (i.e., without fine-tuning on a specific dataset). Although existing fine-tuning methods substantially improve accuracy on a given target distribution, they often reduce robustness to distribution shifts. We address this tension by introducing a simple and effective method for improving robustness while fine-tuning: ensembling the weights of the zero-shot and fine-tuned models (WiSE-FT). Compared to standard fine-tuning, WiSE-FT provides large accuracy improvements under distribution shift, while preserving high accuracy on the target distribution. On ImageNet and five derived distribution shifts, WiSE-FT improves accuracy under distribution shift by 4 to 6 percentage points (pp) over prior work while increasing ImageNet accuracy by 1.6 pp. WiSE-FT achieves similarly large robustness gains (2 to 23 pp) on a diverse set of six further distribution shifts, and accuracy gains of 0.8 to 3.3 pp compared to standard fine-tuning on seven commonly used transfer learning datasets. These improvements come at no additional computational cost during fine-tuning or inference.

The problem with fine-tuning



Raised as an **open problem** by researchers from OpenAI, Stanford, Google, etc.

A simple but effective solution



Task accuracy

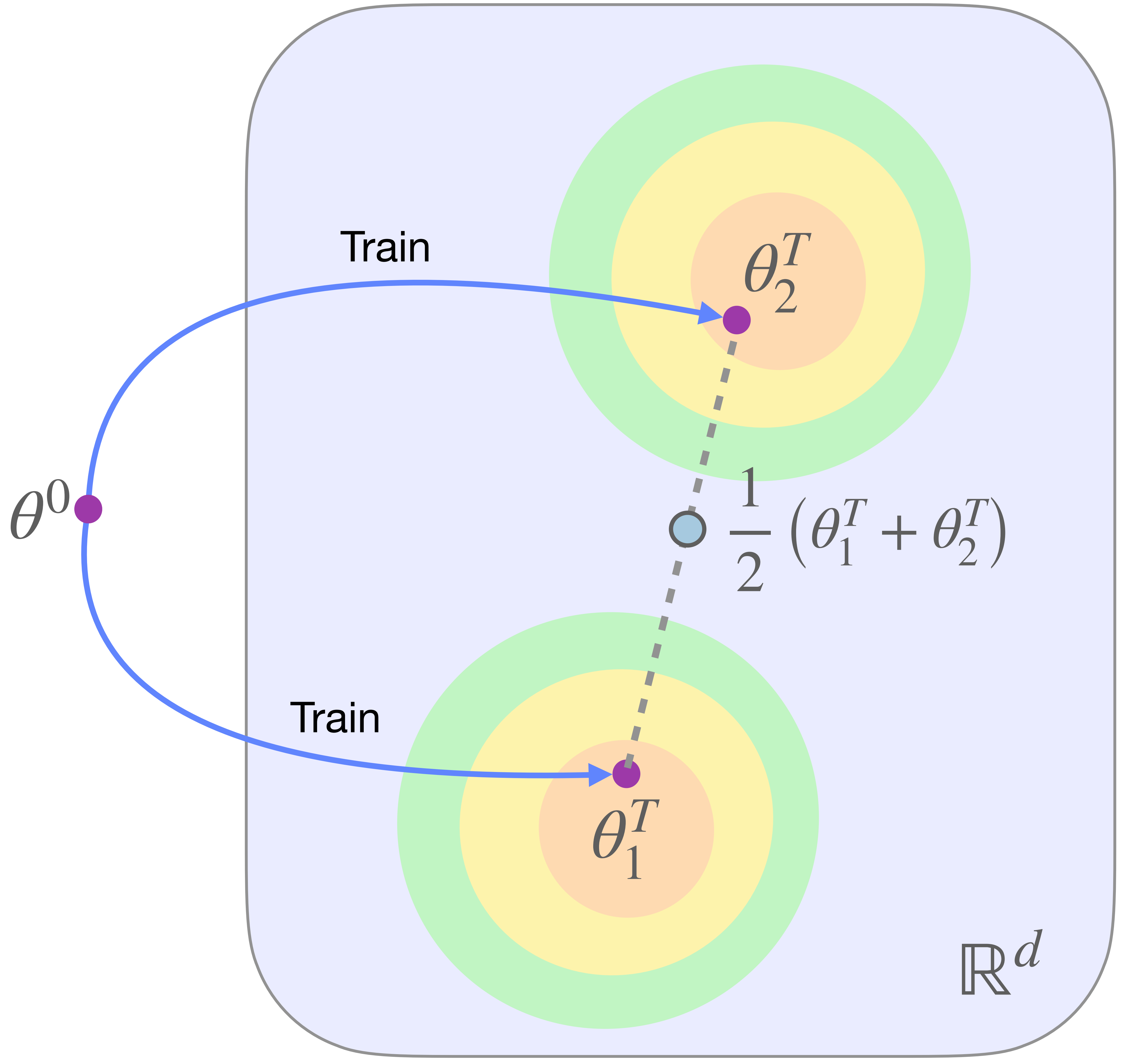
Robustness

Weight-space ensembles for fine-tuning (WiSE-FT)

Building on [Nagarajan, Kolter ’19], [Frankle, Dziugaite, Roy, Carbin ’20], [Neyshabur, Sedghi, Zhang ’20].

Training from scratch

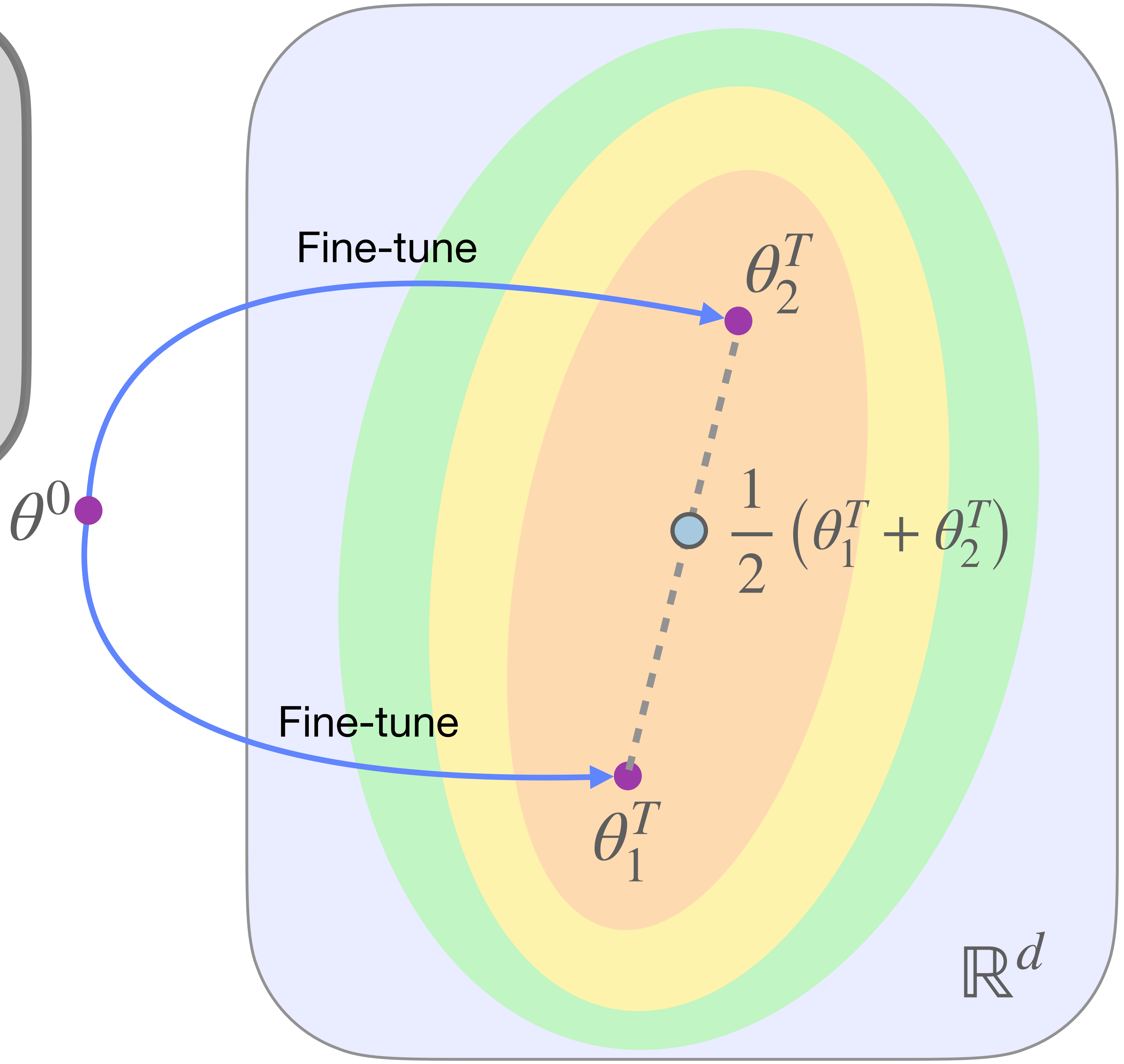
Linearly interpolating the weights of two models trained from scratch encounters a high error barrier (Frankle et al., 2020).



Schematic.

Fine-tuning

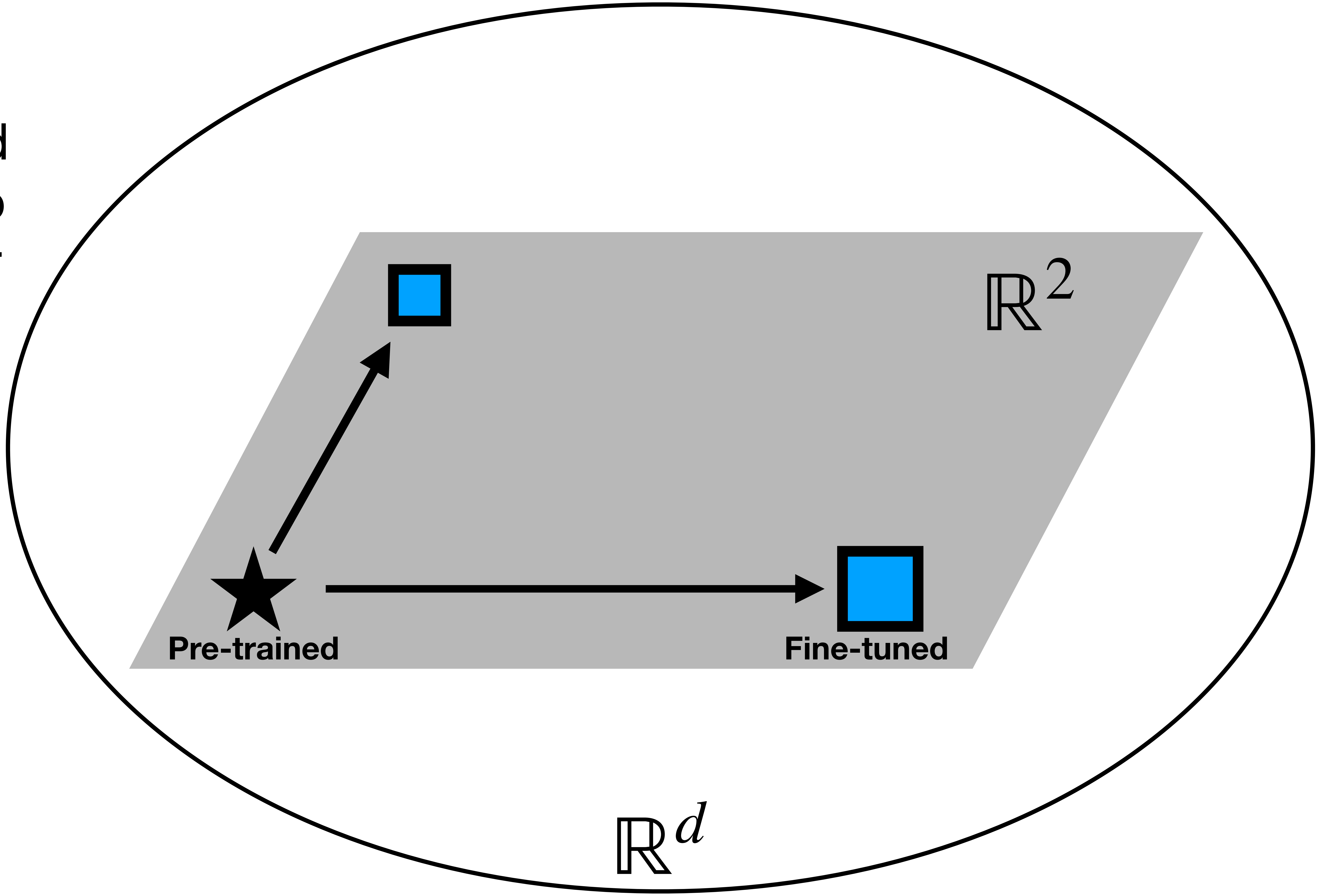
Accuracy remains high when linearly interpolating the weights of two networks fine-tuned from a shared initialization (Neyshabur et al., 2020).

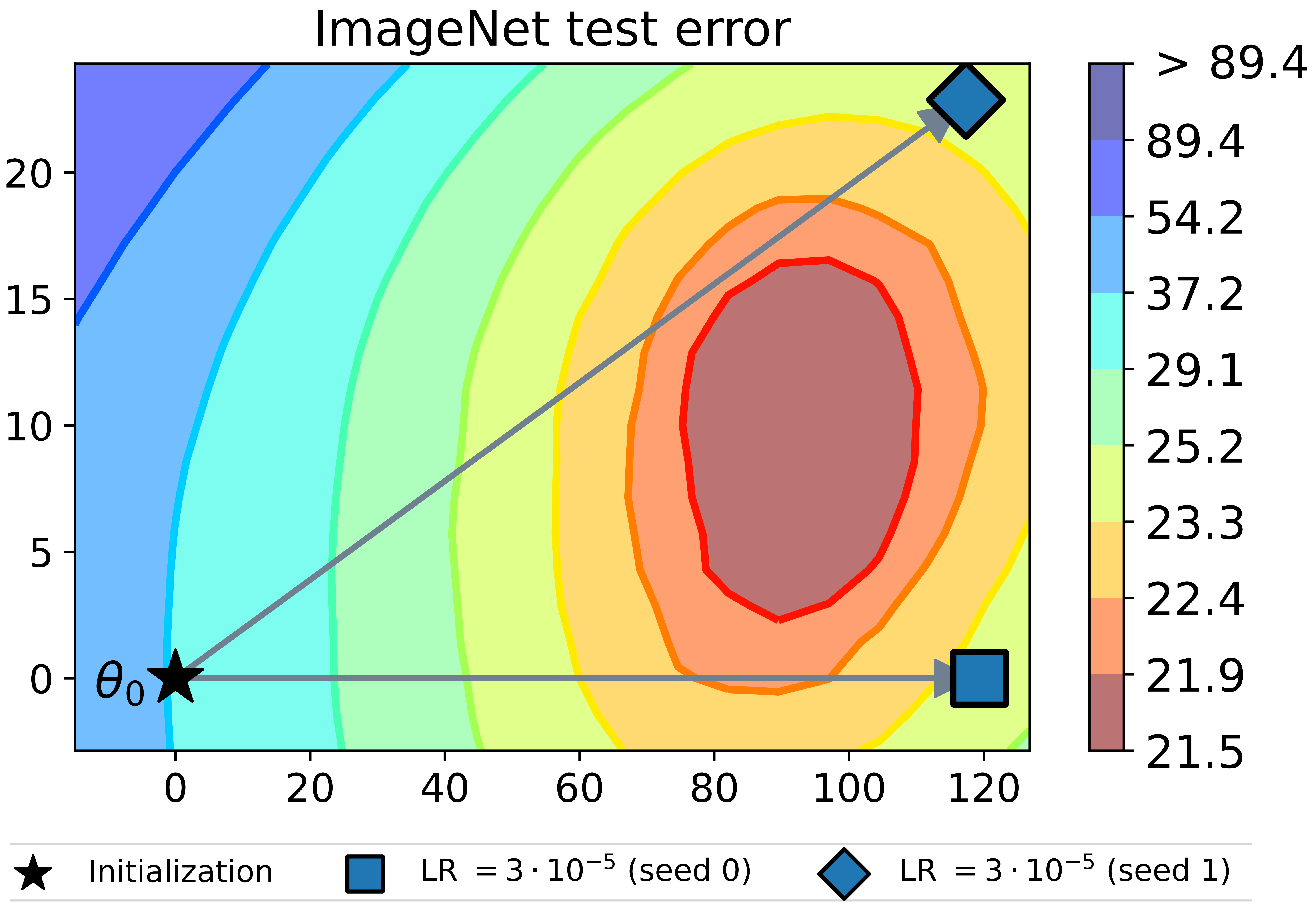


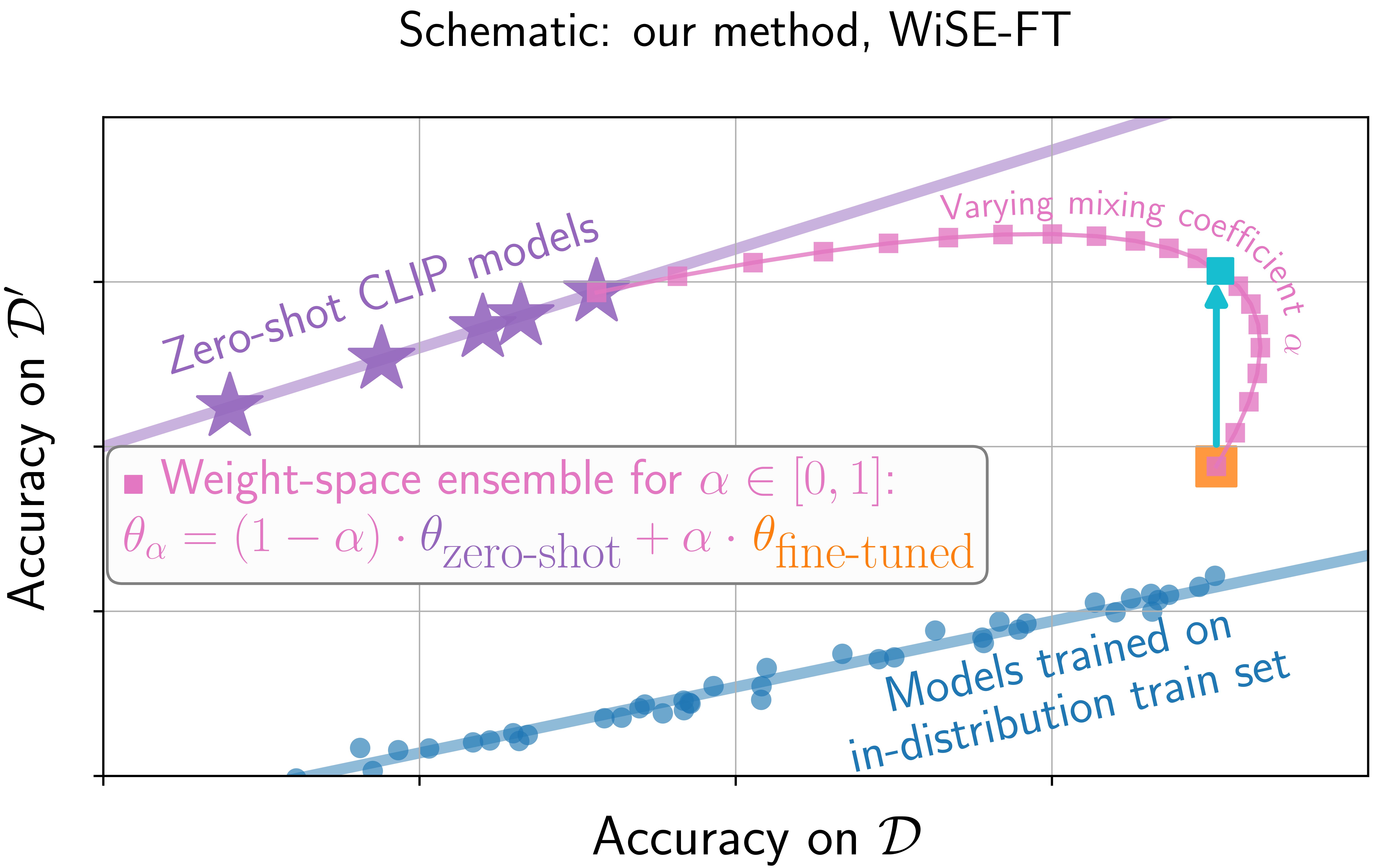
Schematic.

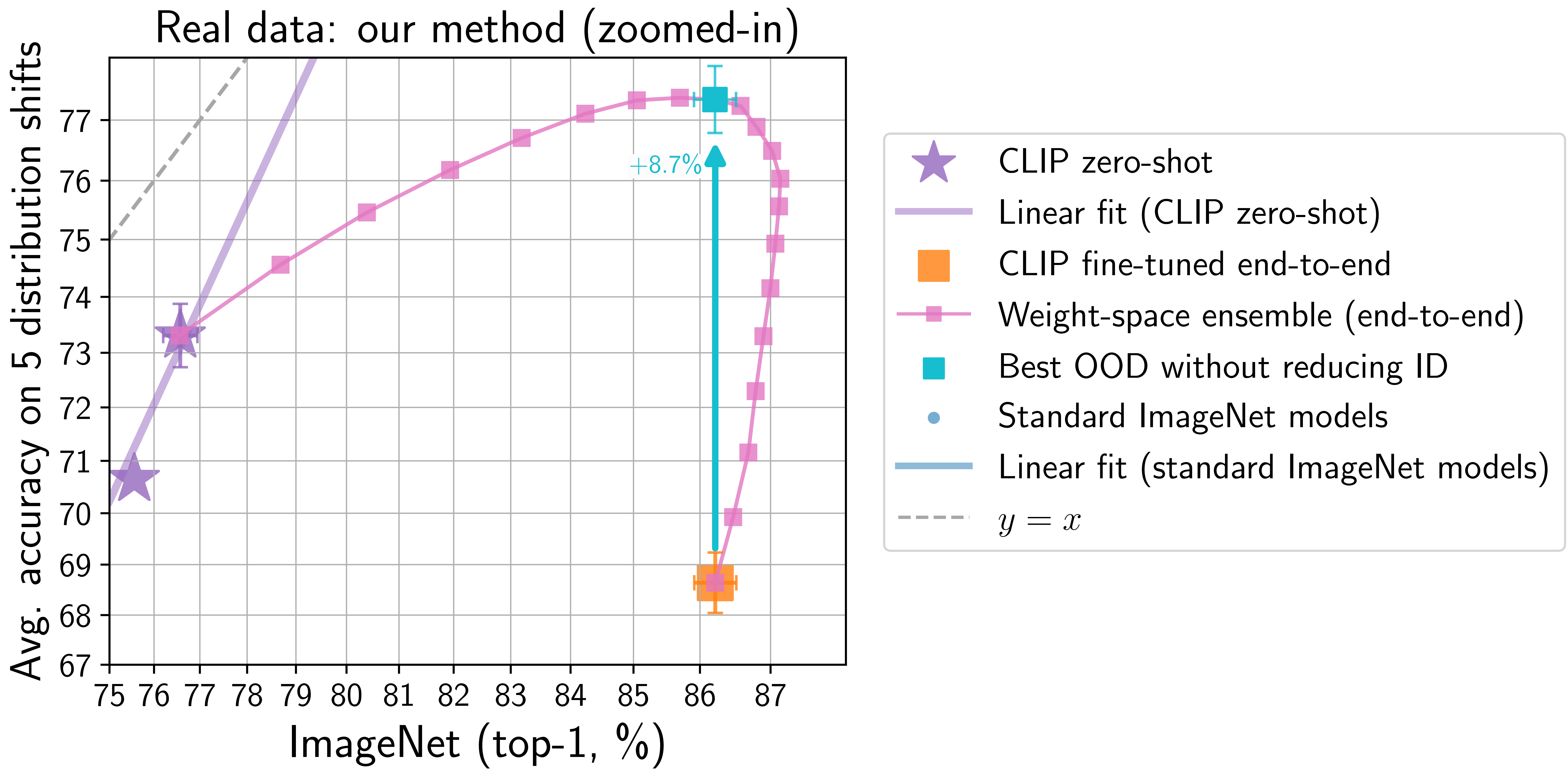
Key difference between fine-tuning and training from scratch

- **From schematic to experiment**: fine-tuned models often appear to lie in a single, low-error region.




ImageNet test error
20
15
10
5
0
0
20
40
60
80
100
120
θ_0
> 89.4
89.4
54.2
37.2
29.1
25.2
23.3
22.4
21.9
21.5
Initialization
LR $= 3 \cdot 10^{-5}$ (seed 0)
LR $= 3 \cdot 10^{-5}$ (seed 1)


Schematic: our method, WiSE-FT
Zero-shot CLIP models
Varying mixing coefficient α
Weight-space ensemble for $\alpha \in [0, 1]$:
$\theta_\alpha = (1 - \alpha) \cdot \theta_{\text{zero-shot}} + \alpha \cdot \theta_{\text{fine-tuned}}$
Models trained on in-distribution train set
Accuracy on $\mathcal{D}'$
Accuracy on $\mathcal{D}$

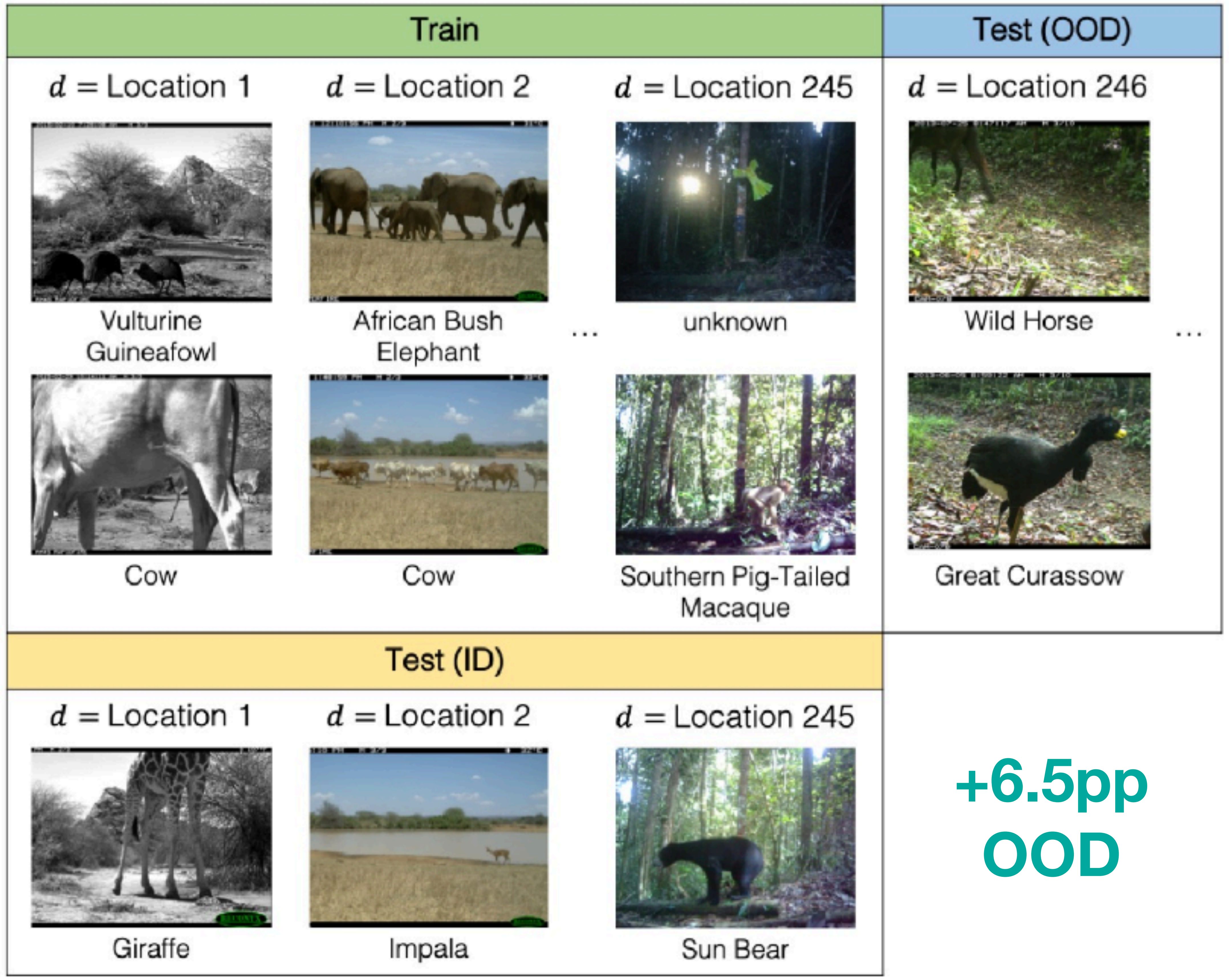


Real data: our method (zoomed-in)
Avg. accuracy on 5 distribution shifts
ImageNet (top-1, %)
67
68
69
70
71
72
73
74
75
76
77
75
76
77
78
79
80
81
82
83
84
85
86
87
+8.7%
CLIP zero-shot
Linear fit (CLIP zero-shot)
CLIP fine-tuned end-to-end
Weight-space ensemble (end-to-end)
Best OOD without reducing ID
Standard ImageNet models
Linear fit (standard ImageNet models)
$y = x$

WILDS

Koh et al., 2021

iWildCam



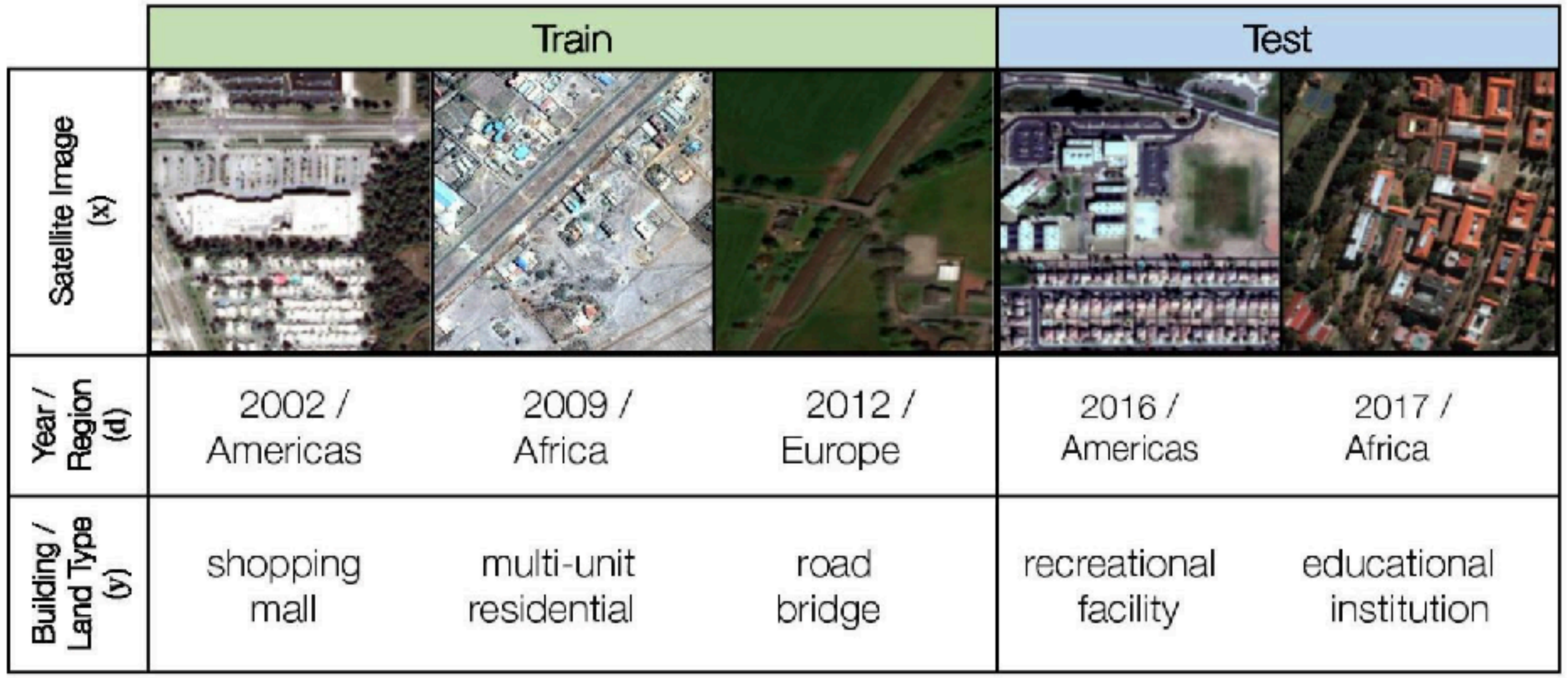
+6.5pp OOD

Beery et al., 2018

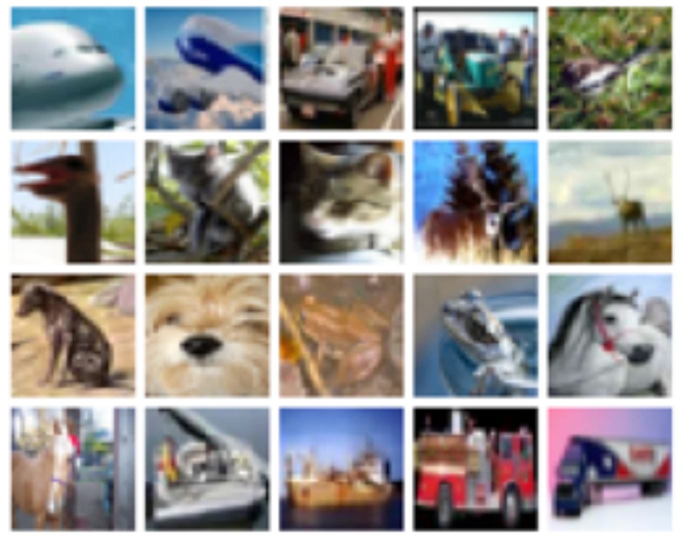
FMoW

+3.7pp OOD

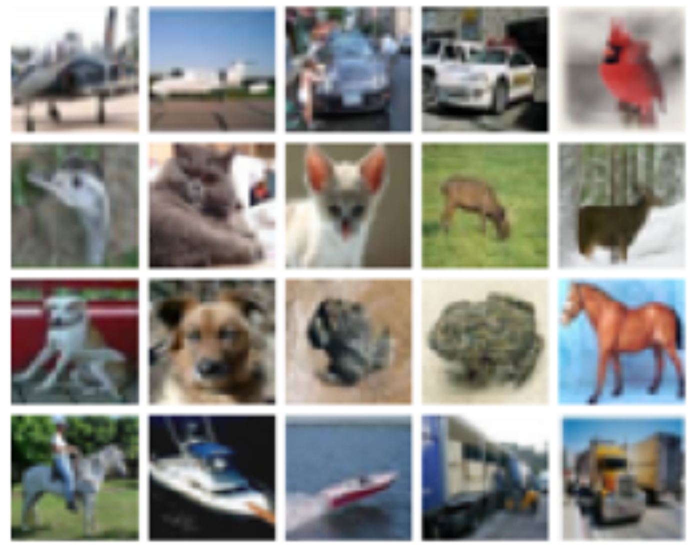
Christie et al., 2018



+2.2pp OOD



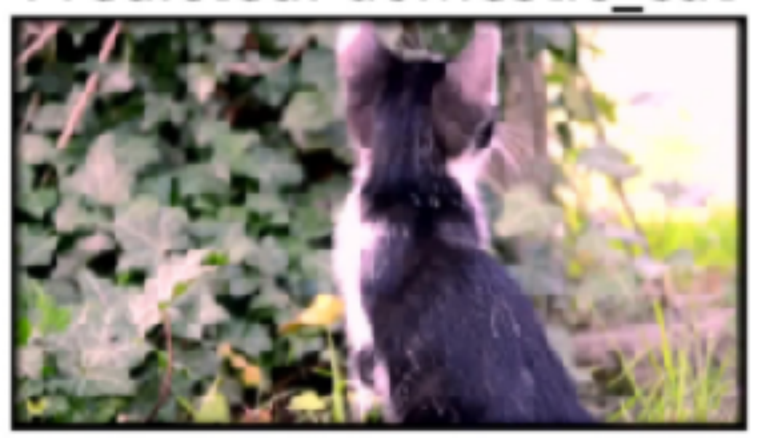
+3.0pp OOD



CIFAR-10.1.
Recht et al., 2019

CIFAR-10.2.
Lu et al., 2020

Predicted: domestic_cat



Predicted: monkey

+8.3pp OOD

ImageNet-Vid-Robust

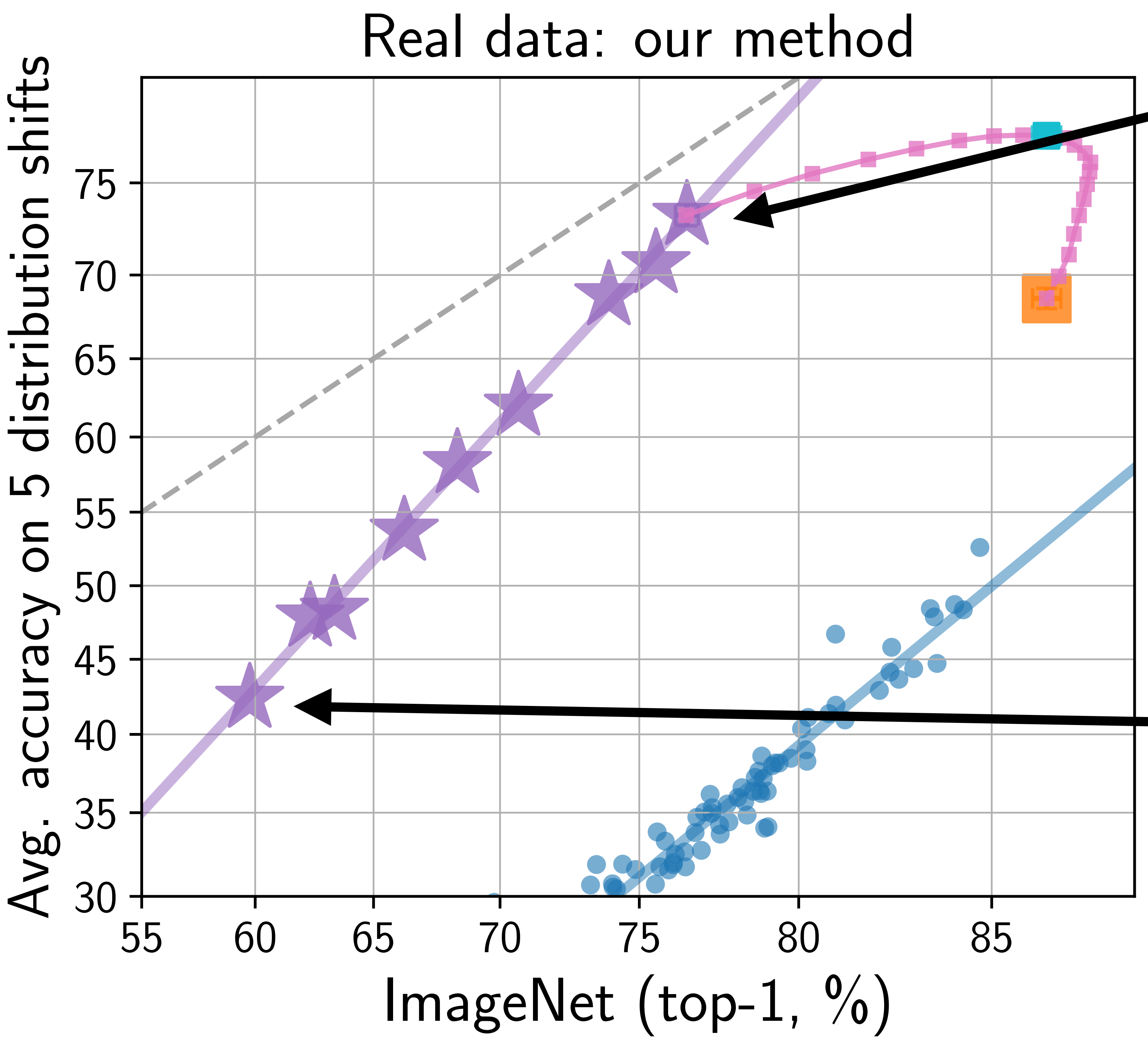
Shankar et al., 2019

YTBBRobust

+14.7pp OOD

Best paper finalist, CVPR 2022

Robustness gains invariant as compute scale increases

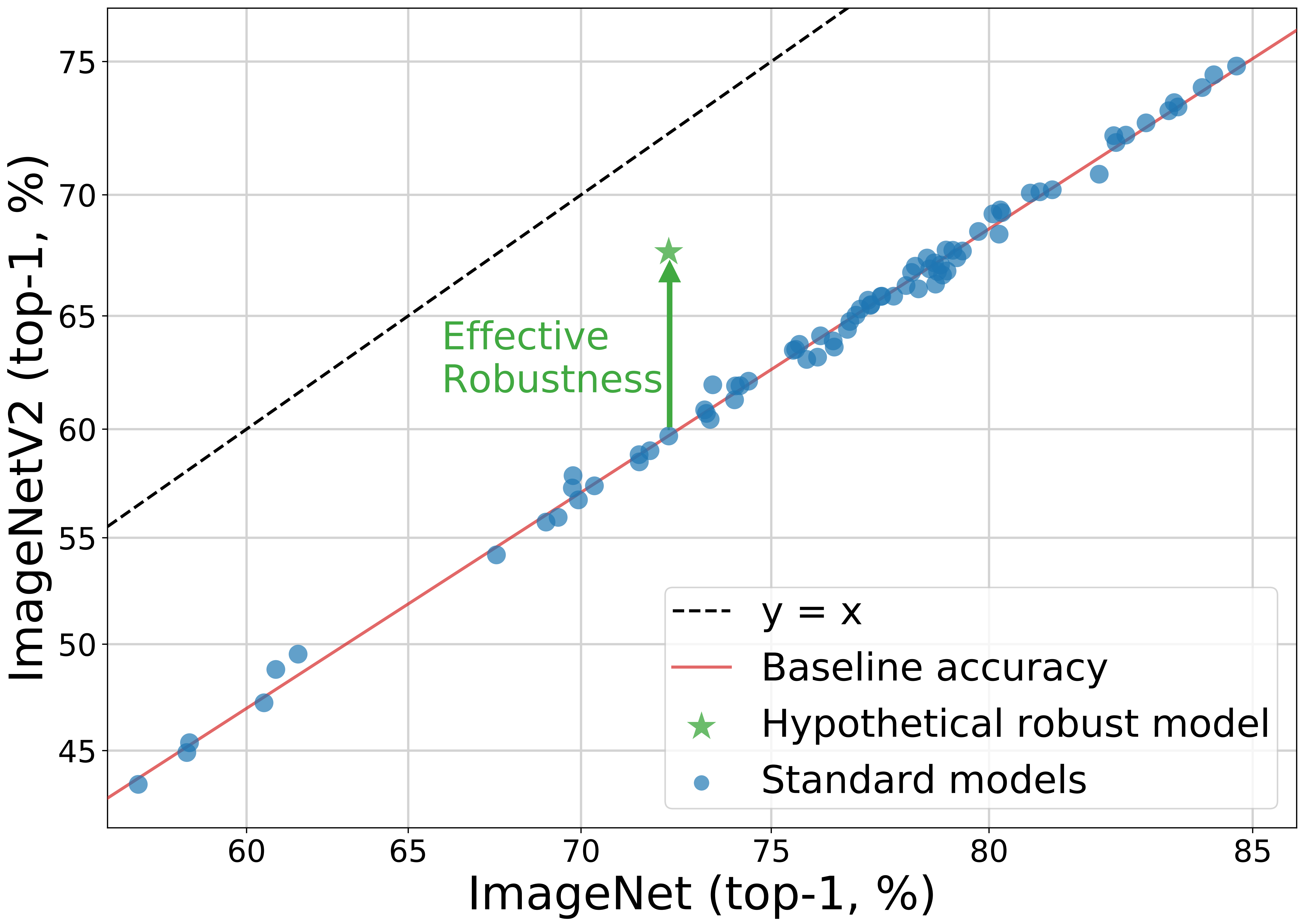


Final result (high accuracy models)

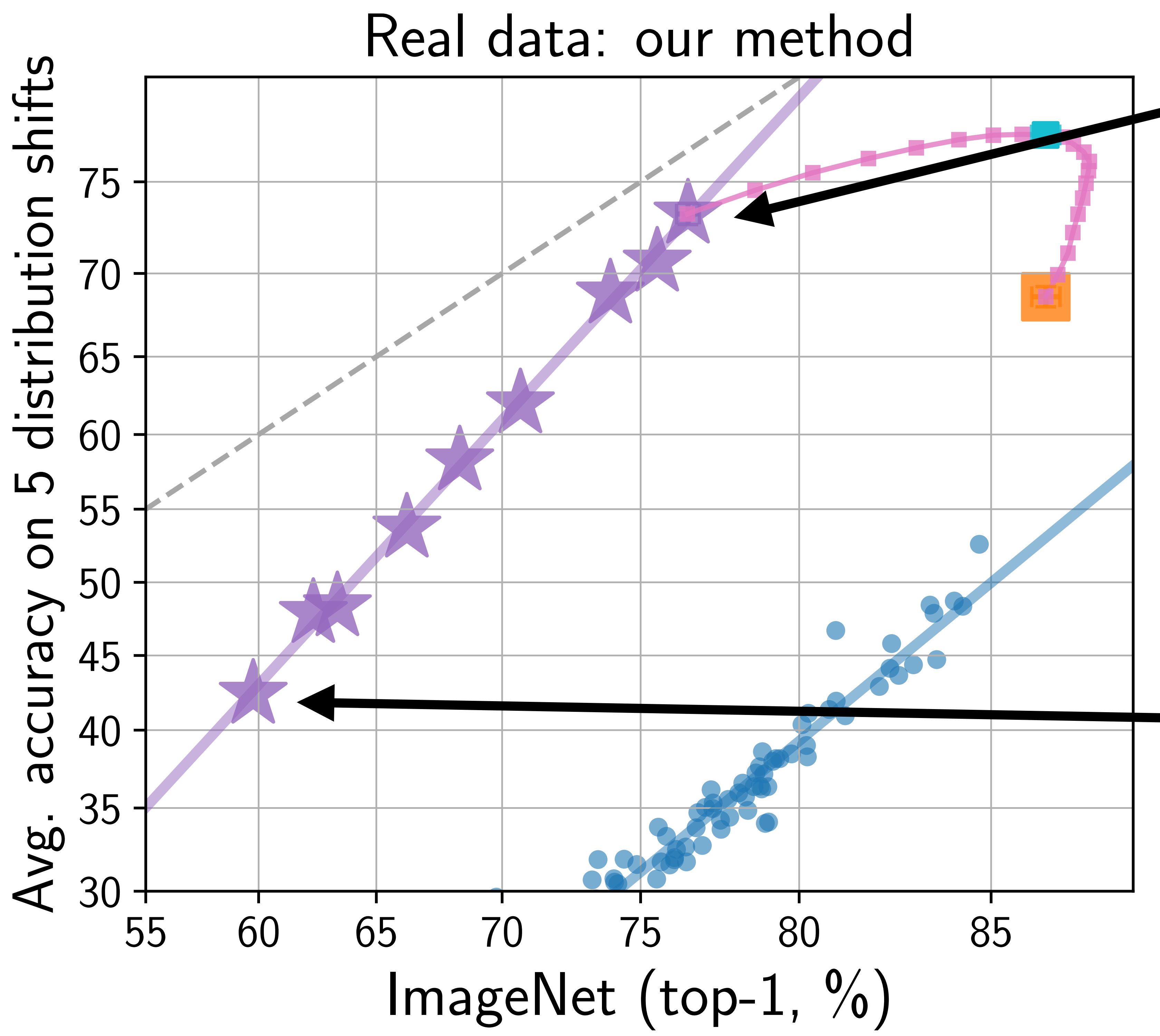
Reliable extrapolation via
“Accuracy on the line”

Where most experiments happened
(low accuracy models)
→ cheaper → faster iteration

All experiments measured effective robustness



Robustness gains invariant as compute scale increases



Final result (high accuracy models)

Reliable extrapolation via "Accuracy on the line"

Where most experiments happened (low accuracy models)

→ cheaper → faster iteration

- Experiment with the full-scale model worked on **first try**
- **ID-OOD trends are a reliable scaling law for model design**

Finetune like you pretrain: Improved finetuning of zero-shot vision models

Sachin Goyal[1], Ananya Kumar[2], Sankalp Garg[1], Zico Kolter[1,3], and Aditi Raghunathan[1]

[1]Carnegie Mellon University
[2]Stanford University
[3]Bosch Center for AI

December 2, 2022

Abstract

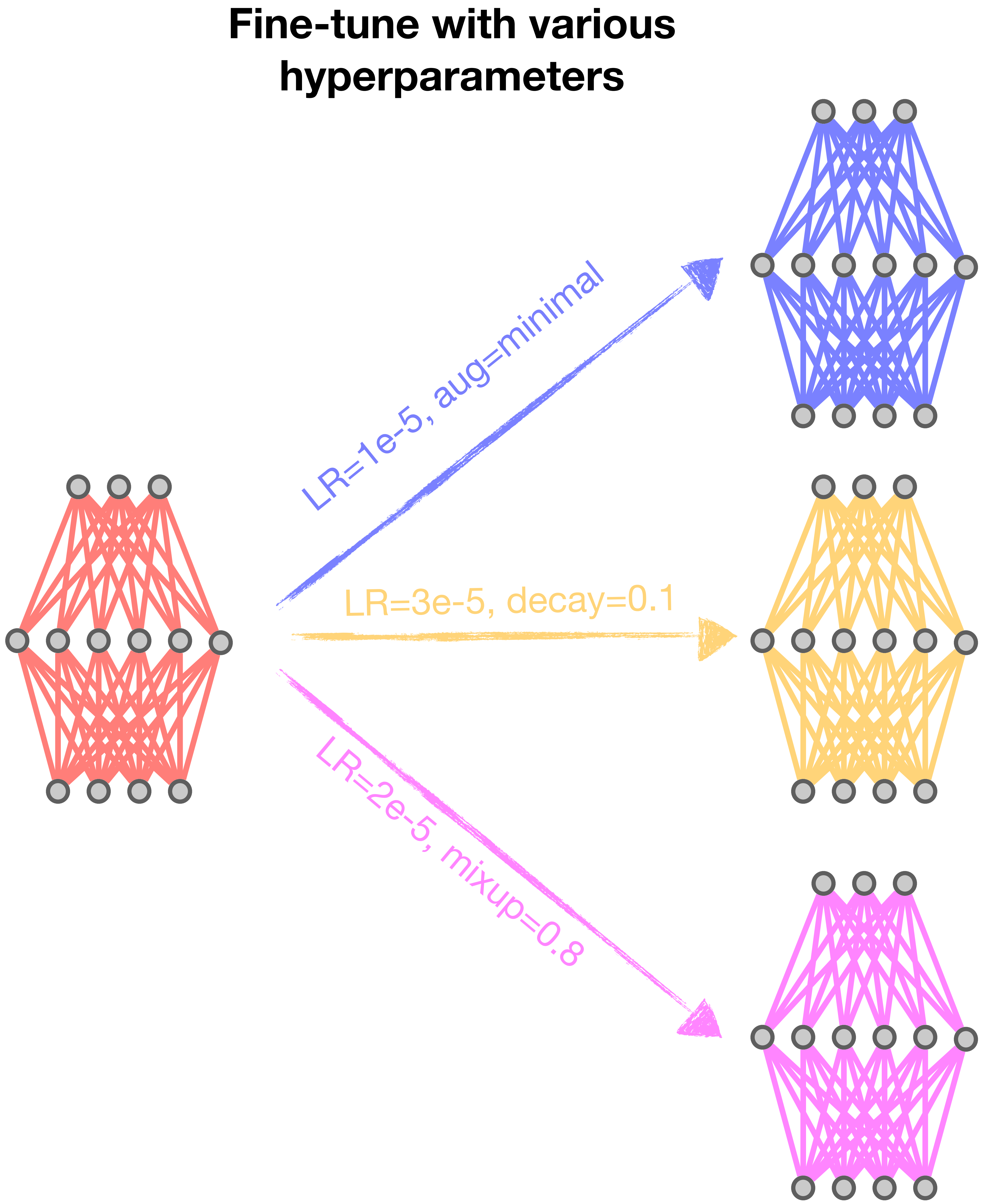
Finetuning image-text models such as CLIP achieves state-of-the-art accuracies on a variety of benchmarks. However, recent works (Wortsman et al., 2021a; Kumar et al., 2022c) have shown that even subtle differences in the finetuning process can lead to surprisingly large differences in the final performance, both for in-distribution (ID) and out-of-distribution (OOD) data. In this work, we show that a natural and simple approach of mimicking contrastive pretraining consistently outperforms alternative finetuning approaches. Specifically, we cast downstream class labels as text prompts and continue optimizing the contrastive loss between image embeddings and class-descriptive prompt embeddings (contrastive finetuning).

Why stop at averaging two models?

Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time

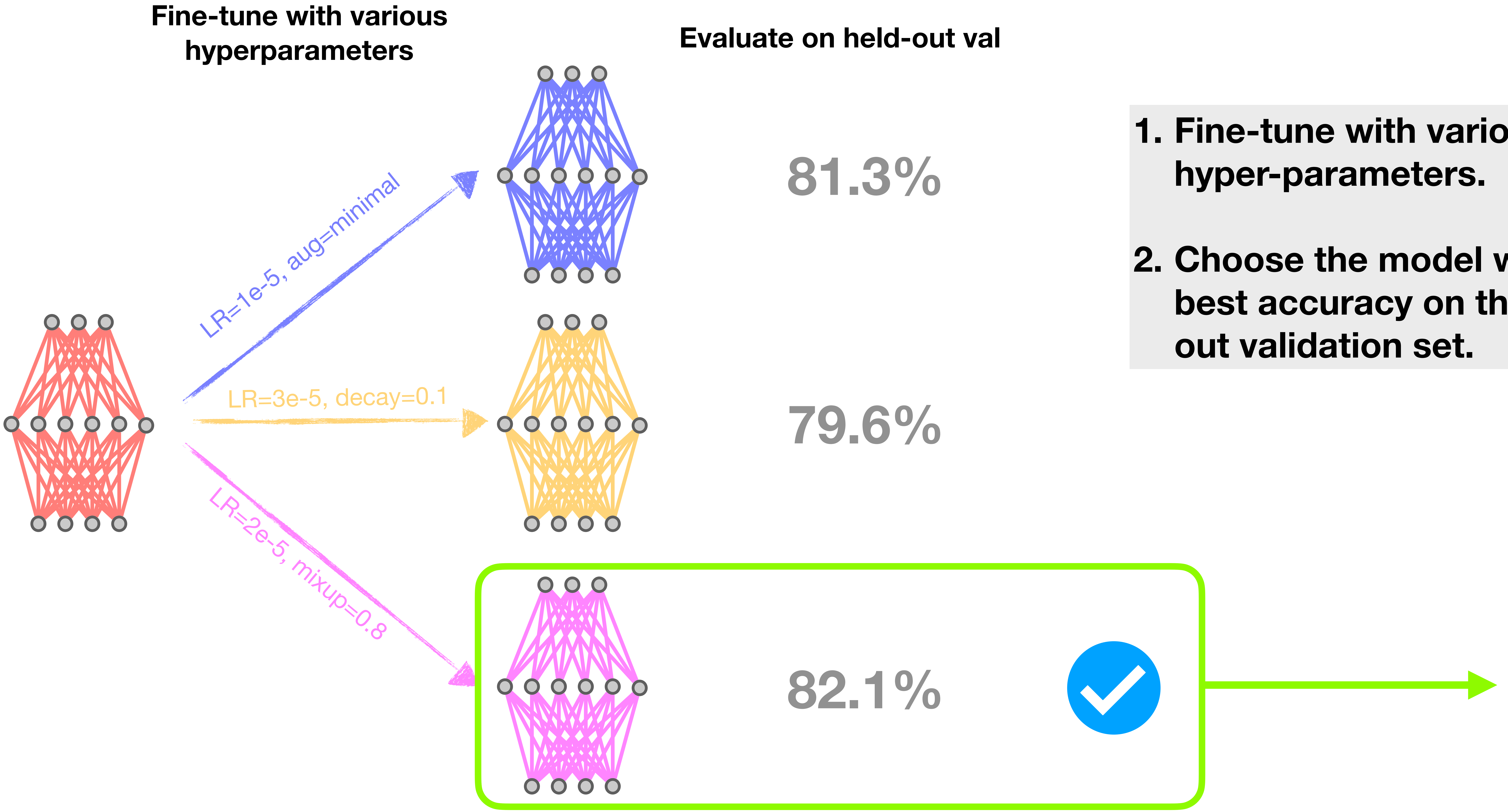
Mitchell Wortsman[1] Gabriel Ilharco[1] Samir Yitzhak Gadre[2] Rebecca Roelofs[3] Raphael Gontijo-Lopes[3]
Ari S. Morcos[4] Hongseok Namkoong[2] Ali Farhadi[1] Yair Carmon[*5] Simon Kornblith[*3] Ludwig Schmidt[*1]

Conventional procedure for maximizing accuracy while fine-tuning



1. **Fine-tune with various hyper-parameters.**

Conventional procedure for maximizing accuracy while fine-tuning



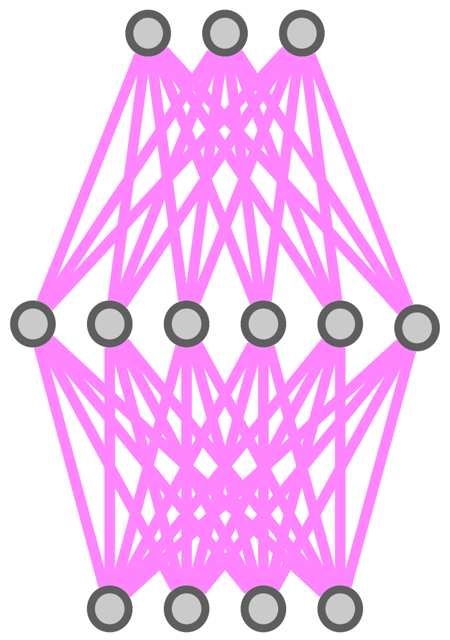
1. Fine-tune with various hyper-parameters.

2. Choose the model with the best accuracy on the held-out validation set.

Downsides of the conventional fine-tuning recipe

Choosing the best individual model on the held-out validation set

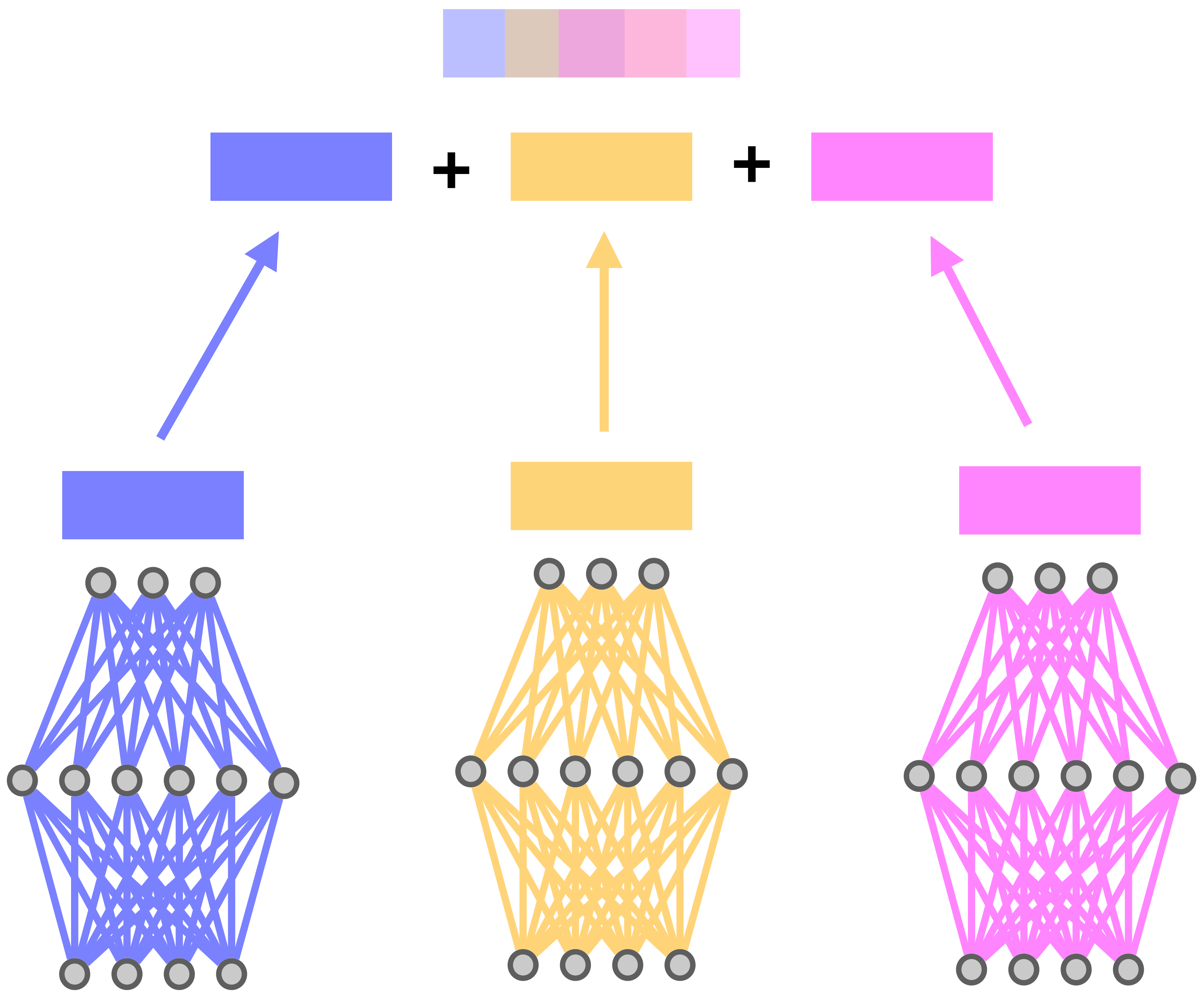
82.1%



Lower accuracy

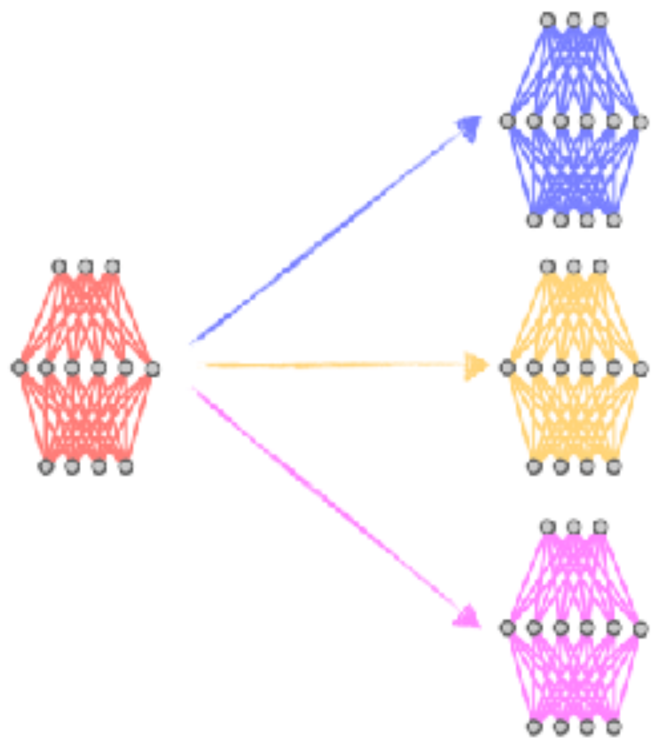
Ensemble

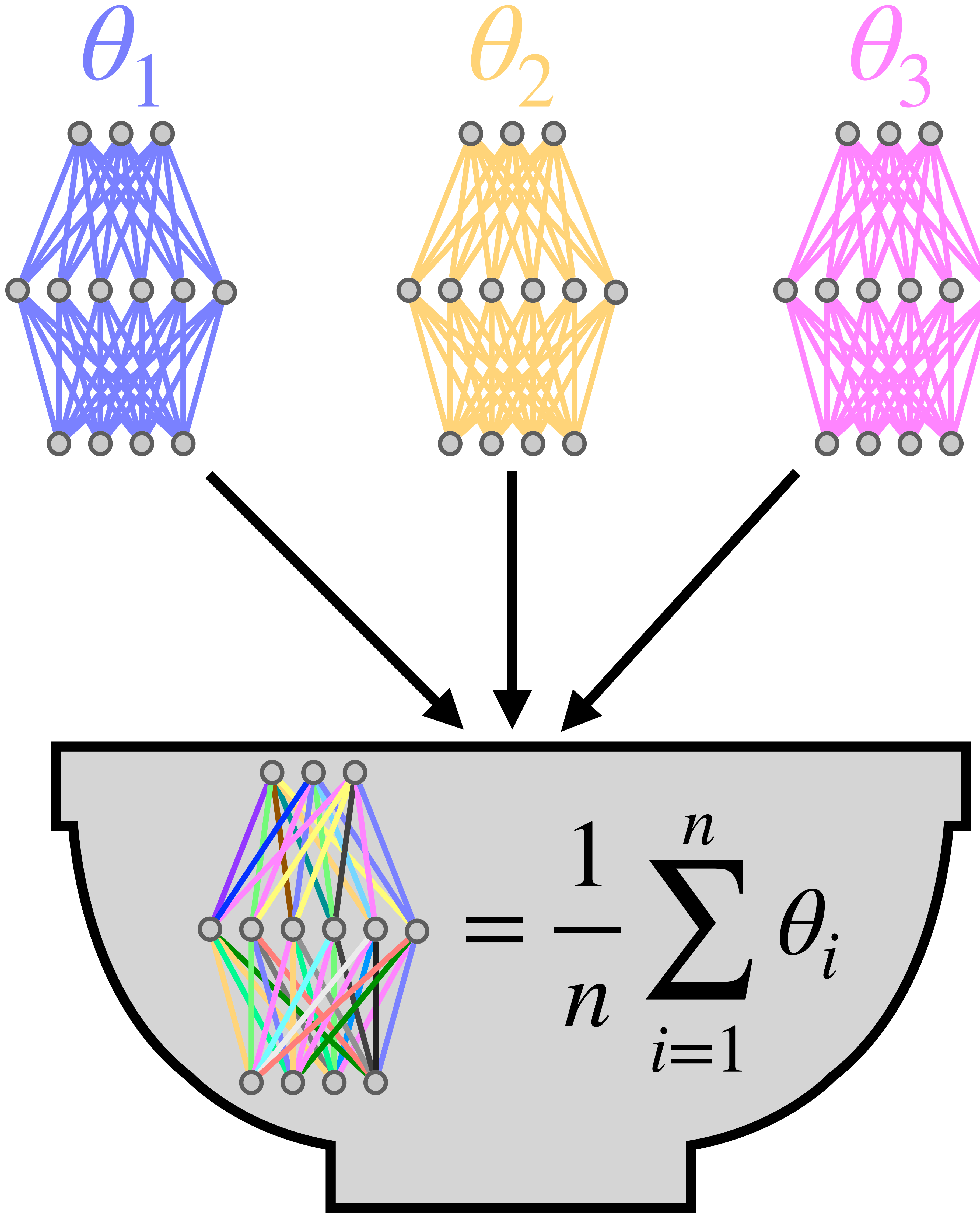
84.0%



Higher inference cost

Model soups





Best of both worlds:

- Same **high accuracy** as the ensemble
- Same **fast inference** time as an individual model

Results

- **ImageNet SotA** 🏆
- Gains on many more dataset
- Widely used for multimodal models

Can we fine-tune a model while preserving its zero-shot abilities?

Patching open-vocabulary models by interpolating weights

Gabriel Ilharco* University of Washington gamaga@cs.washington.edu

Mitchell Wortsman* University of Washington mitchnw@cs.washington.edu

Samir Yitzhak Gadre* Columbia University sy@cs.columbia.edu

Shuran Song Columbia University shurans@cs.columbia.edu

Hannaneh Hajishirzi University of Washington hannaneh@cs.washington.edu

Simon Kornblith Google Research, Brain Team skornblith@google.com

Ali Farhadi University of Washington ali@cs.washington.edu

Ludwig Schmidt University of Washington schmidt@cs.washington.edu

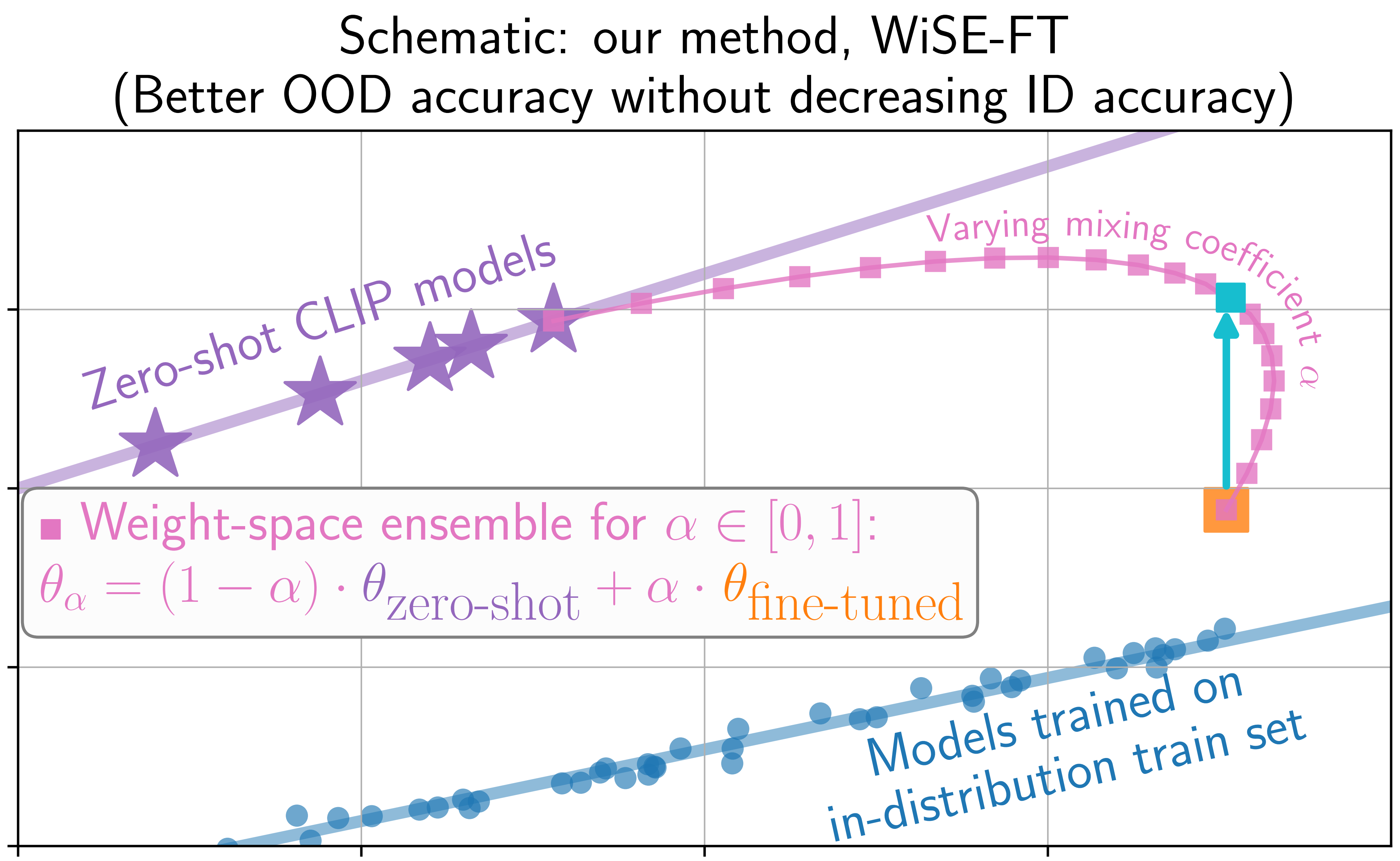
Abstract

Open-vocabulary models like CLIP achieve high accuracy across many image classification tasks. However, there are still settings where their zero-shot performance is far from optimal. We study *model patching*, where the goal is to improve accuracy on specific tasks without degrading accuracy on tasks where performance is already adequate. Towards this goal, we introduce PAINT, a patching method that uses interpolations between the weights of a model before fine-tuning and the weights after fine-tuning on a task to be patched. On nine tasks where zero-

Conclusions

Pre-trained models often can be improved by **fine-tuning on task-specific data**.

➡ Both in **vision** and in **NLP** (instruction tuning, RLHF, etc.)



"Standard" fine-tuning can **negatively affect the capabilities** of the pre-trained model.

➡ Interpolating between the pre-trained and fine-tuned models can preserve robustness while improving task performance.

Open questions

- Simple weight interpolation seems naive → are there better fine-tuning methods?
- Can we remove fine-tuning entirely and improve pre-training instead?