Benchmarking Low-Shot Robustness to Natural Distribution Shifts

Aaditya Singh[*,⋆,†] Kartik Sarangmath[*] Prithvijit Chattopadhyay Judy Hoffman
Georgia Institute of Technology
{singaadi, kartiksarangmath, prithvijit3, judy}@gatech.edu

Abstract

Robustness to natural distribution shifts has seen remarkable progress thanks to recent pre-training strategies combined with better fine-tuning methods. However, such fine-tuning assumes access to large amounts of labelled data, and the extent to which the observations hold when the amount of training data is not as high remains unknown. We address this gap by performing the first in-depth study of robustness to various natural distribution shifts in different low-shot regimes: spanning datasets, architectures, pre-trained initializations, and state-of-the-art robustness interventions. Most importantly, we find that there is no single model of choice that is often more robust than others, and existing interventions can fail to improve robustness on some datasets even if they do so in the full-shot regime. We hope that our work will motivate the community to focus on this problem of practical importance. Our code and low-shot subsets are publicly available at this url.

1. Introduction

In the past decade, Computer Vision has made significant progress due to advanced architectures like Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs), large datasets, and sophisticated training strategies [1, 2, 3, 4]. However, early learning techniques heavily focused their evaluation on ImageNet [5] performance, which raised concerns about their ability to generalize to distribution shifts [6, 7]. To address this, researchers have proposed a wide-range of evaluation datasets [8, 9, 10, 11, 12] that can be used to measure out-of-distribution (OOD) performance of models trained and validated with in-domain (ID) data.

Recent methods [13, 14, 15, 16, 17] use self-supervised or large-scale vision-language pre-trained models (such as CLIP [4]) and fine-tune them on fully labelled ID data to achieve impressive performance on such datasets. Unfortunately, fine-tuning requires large amounts of data and compute that may not be accessible to most practitioners. Moreover, it can be difficult and expensive to collect and consistently annotate such data, especially in settings like camera traps where images can vary significantly in quality, lighting, and pose (e.g iWildCam [18]). Such challenges are also echoed by prior work [19] and compounded by the fact that many images may belong to rare or endangered species, making annotations even more difficult to obtain. Therefore, it is important to study which models and fine-tuning methods provide strong OOD robustness performance when trained with few ID images. We refer to this setting of fine-tuning a pre-trained model on low-shot ID images followed by evaluation on OOD images as the "low-shot robustness" setting (see Fig. 1).

[*]Equal contribution; [⋆]Project lead; [†]Currently affiliated with AWS AI Labs, work done prior to joining.

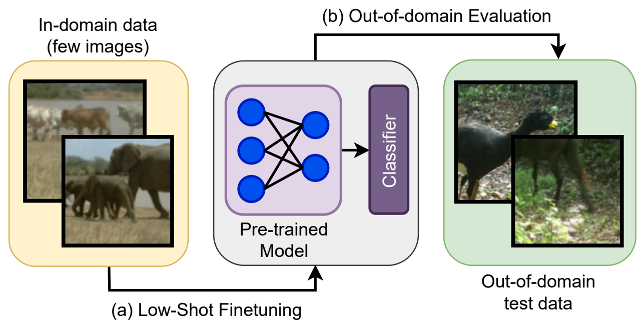

Figure 1: **Low-Shot Robustness Setting.** (a) We assume access to a pre-trained model trained on large-scale datasets such as ImageNet [5] and limited in-domain images (in the order of thousands) for training. We use different kinds of fine-tuning methods that have been shown to improve robustness when there is typically order of magnitudes higher training data. (b) We then evaluate the (low-shot) fine-tuned model on out-of-domain (OOD) data.

From works that demonstrate robustness in the full-shot regime, we seem to arrive at the following conclusions for robustness to natural distribution shifts in the full-shot regime: (1) Amongst ImageNet pre-trained initializations, SSL ViTs are more robust than their supervised and CNN counterparts, with the more recent ones being better [13, 14]. (2) Even without additional robustness interventions (i.e. methods to improve robustness), pre-trained models on large external datasets such as CLIP [4] provide superior robustness [16]. (3) Such models when combined with state-of-the-art robustness interventions lead to significant robustness improvements on several datasets [15, 16, 17]. In this paper, we question to what extent these conclusions hold true when the amount of training data is not as high.

Overall, we perform the first in-depth study of robustness to various natural distribution shifts in different low-shot regimes: spanning datasets, architectures, pre-trained initializations, and state-of-the-art robustness interventions. Through our experiments, we aim to answer the following key questions:

Q1. *For ImageNet pre-trained models, what kind of pre-training strategies and architectures are most effective for robustness in low-shot regimes?*
A: Self-supervised ViTs generally perform better than CNNs and the supervised counterparts (where applicable) on both ID and OOD shifts, but no single initialization or model size works better across datasets.

- For ImageNet and iWildCam [18] datasets, MSN ViT [14] performs better than other models on OOD shifts, however a smaller model size (ViTS-16) works better for iWildCam but not for ImageNet.
- For Camleyon [20] dataset which is non object-centric, DINO ViTS-16 [21] outperforms other models including DINO ViTB-16 and MSN ViTS-16 on both ID and OOD shifts.

Q2. *Do models pre-trained on large external datasets, such as CLIP, provide superior robustness compared to ImageNet pre-trained ones on different datasets?*
A: While we generally conform with the findings of recent works [15, 16, 17] and find that models such as CLIP [4] provide superior robustness on ImageNet and in full-shot regimes, we find that ImageNet pre-trained models can be better on other datasets such as iWildCam and Camelyon in the low-shot regimes.

- Comparing ViTB-16 architecture on these datasets, DINO initialization outperforms CLIP (zero-shot or otherwise) and ImageNet-21k [22] supervised ViT on both ID and OOD shifts.
- ImageNet supervised ViT [23] significantly outperforms ImageNet-21k supervised ViT on OOD shifts.

Q3. *When using robustness interventions, does better robustness in the full-shot regime also imply better robustness in the low-shot regimes?*
A: Not always. We find that depending on the initialization, existing interventions can fail to improve robustness in the full-shot regime or in some of the low-shot regimes for datasets other than ImageNet.

- On iWildCam, interventions often fail to improve robustness with MSN ViTB-16 in the full-shot regime. On the other hand, only WiSE-FT [16] significantly improves robustness with CLIP ViTB-16 in both the full and low-shot regimes.
- On Camelyon, while interventions often improve robustness in the full-shot regime for both MSN and CLIP ViTB-16, they fail to do so either in *extreme* ($\sim$ 3000 images) or in *moderate* ($\sim$ 15000 images) low-shot regimes, except WiSE-FT with CLIP.

As highlighted by our findings, conventional wisdom for robustness to natural distribution shifts in the full-shot regime might not apply in the low-shot regimes, and should be seen as an important challenge for future work.

2. Related Work

Robustness studies. Real-world models may encounter and struggle to generalize on data distributions different than the ones used for training [24, 25]. Previous works have studied such generalization capability of models under synthetic [26, 27, 28, 29, 30, 31] and natural distribution shifts [7, 8, 9, 10, 11, 12]. Researchers have also looked at the effect of architecture, i.e. CNNs and ViTs on robustness to different kinds of shifts [32, 33, 34, 35] and distortion robustness of several models in comparison to humans [36]. In particular, [37] performs a large-scale study of several supervised models and finds that interventions used for synthetic shifts offer little to no robustness gains for natural distribution shifts. On the other hand, accuracy under natural distribution shifts can often be reliably estimated by the in-distribution accuracy [38, 39, 37, 40] except for some shifts [41, 42]. Crucially, these works perform evaluations after training on fully labelled datasets with hundreds of thousands or millions of images which can be out of reach for most practitioners. While some recent works [43, 19, 44] attempt to study the impact of training data amount on out-of-distribution robustness, they do not adopt the recent pre-training strategies [4, 21, 14] and fine-tuning techniques [15, 16, 17] that have led to unprecedented robustness gains. We therefore adopt such methods and perform experiments in the low-shot regime to observe its impact on robustness to various natural distribution shifts.

Self-supervised learning. Researchers have shown self-supervised learning (SSL) to be better or on-par with supervised learning for pre-training deep networks for various downstream tasks [45, 46, 21, 47, 13, 14] and we refer the reader to [48, 49] for thorough literature reviews. Recent methods that leverage ViTs [13, 14] demonstrate superior robustness to some natural distribution shifts [9, 10, 11] compared to previous state-of-the-art methods without additional interventions. However, such evaluations are performed only after fine-tuning on full ImageNet and whether the trend holds for other datasets and in different low-shot regimes remains an open question. We aim to address this gap in our work by evaluating some of the most recent SSL ViTs on a variety of datasets and distribution shifts, also comparing with CNNs and the supervised counterparts.

Few-shot learning. Few-shot learning aims to generalize to novel classes from a few samples belonging to these classes [50, 51, 52]. While meta-learning based approaches used to be popular on standard benchmarks [53, 54, 55, 56], a growing wave of research showed that simpler transfer learning-based approaches can also achieve competitive performance [57, 58]. Recently, [59] conforms with this finding on the more challenging cross-domain few-shot learning (CD-FSL) scenario where the source and novel classes belong to different domains. Since then, works often perform SSL pre-training on the source data followed by low-shot fine-tuning on the few examples of novel classes [47, 60, 61]. However, unlike the (cross-domain) few-shot scenario, we use the target or out-of-distribution (OOD) data only for evaluation purposes similar to most other robustness studies. Nonetheless, we use the classifiers adopted in [59] and present their detailed comparison on different datasets and associated design choices in section 1.2 of appendix. We discuss other related works that are either not applicable or already described in our experiments in appendix Sec. 6.

3. Preliminaries: Robustness Metrics

Although using out-of-distribution (OOD) shifts to measure absolute performance can suggest robustness, it overlooks the in-domain (ID) performance of a model. As pointed in [37], two models with similar OOD performance can have vastly different ID performances. A better definition of robustness should consider the OOD performance beyond what is expected from achieving some level of ID performance. Therefore, to measure robustness, in addition to absolute performance comparison we also adopt the *effective* and *relative* robustness framework used in previous works [7, 37, 16]. We now describe these metrics in detail.

Key to measuring effective robustness is establishing an expected baseline OOD accuracy given some ID accuracy x. This is established by computing a log-linear fit $\beta(x)$ over ID and OOD accuracies, i.e. acc^s_{id} and acc^s_{ood} respectively, for a set of standard models $f^s_1, f^s_2, \dots f^s_n$ as:

$$\beta(x) = \sigma(w \, \text{logit}(x) + b) \tag{1}$$

where $\text{logit}(x) = \ln \frac{1}{1-x}$ and σ is the inverse of the logit function. In practice, $\beta(x)$ is obtained by mapping each point $(x, y) \rightarrow (\text{logit}(x), \text{logit}(y))$ and solving linear regression. This can be visualized by plotting (acc^s_{id}, acc^s_{ood}) on a scatter plot with the x and y axes denoting ID and OOD accuracies respectively.

Once obtained, effective robustness of an "intervention"[1]

Dataset	Low-Shot Regimes (Imgs / Class)		
	Extreme	Moderate	High
1 ImageNet [5]	1	5	~ 13
2 iWildCam [18]	1-480	1-4802	1-9604
3 Camelyon [20]	1500	7500	15000

Table 1: **Different Low-Shot Regimes.** We consider low-shot regimes with similar number of images for different datasets and describe them in more detail in section 4.1.

r applied on the model f^s, i.e. $f^r = (acc^r_{id}, acc^r_{ood})$ can be expressed as:

$$\rho(f^r) = acc^r_{ood} - \beta(acc^r_{id}) \tag{2}$$

which outlines if the intervention leads to OOD accuracy beyond what is expected from having a higher ID accuracy.

While effective robustness is important, it is not enough to provide a comprehensive evaluation of models, especially in the low-shot regimes. An "intervention" on a model may result in high positive $\rho(f^r)$, indicating effective robustness, but it could also decrease both ID and OOD accuracies which is not desirable. Thus, in addition to effective robustness, we measure relative robustness by assessing the impact of an intervention on OOD accuracy as:

$$\tau(f^r) = acc^r_{ood} - acc^s_{ood} \tag{3}$$

Following [37], an intervention r is said to improve the robustness of a model f^s only when it exhibits both effective and relative robustness, that is, $\rho(f^r) > 0$ and $\tau(f^r) > 0$. However, our experiments indicate that interventions frequently lack simultaneous effective and relative robustness across various low-shot regimes. For simplicity, we refer to $\rho(f^r)$ as ρ and $\tau(f^r)$ as τ.

4. Experimental Setting

Following prior work, we assume full label-space overlap and study image classification under natural distribution shifts [4, 16]. Additionally, we refer to low-shot as 10^3-10^4 images, as also shown in Fig. 1 and table 1. We describe our experimental setting with the associated design choices and justifications in this section.

4.1. Datasets and Low-Shot Regimes

Prior studies [37, 40] have observed a linear trend for certain supervised models on ImageNet [5] and iWildCam [18] datasets after applying the logit function (see Eq. 1), while contrasting evidence has been reported for other datasets, such as Camelyon [20], in [42]. To obtain a comprehensive view of robustness in low-shot regimes, where a strong correlation between in-domain (ID) and out-of-distribution (OOD) performances may or may not exist, we conduct experiments on all three datasets.

ImageNet & Distribution Shifts. ImageNet (IN1k) [5] is an extensive dataset for image recognition that consists of objects and scenes belonging to one of the 1000 classes.

[1]For models pre-trained on large external datasets such as CLIP [4], it's unclear what datasets are considered in or out-of-distribution, so we exclude it from the standard set of models and treat it as an intervention.

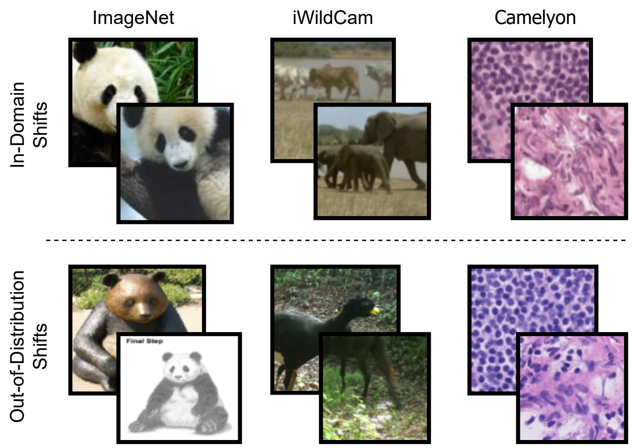

Figure 2: **Datasets & Distribution Shifts.** We show some sample images from ImageNet [5], iWildCam [18], and Camelyon [20] datasets and associated distribution shifts [9, 11].

For training, we use the subsets with 1, 5, and ∼13 images per class (see table 1 row 1) used by [14] for comparison of self-supervised methods based on in-domain (ID) accuracy. We use the IN1k validation split for model validation based on top-1 accuracy.[2] For testing, we report the average top-1 accuracy on the following 5 natural distribution shifts:
ImageNet-R (IN-R) [11] has 200 classes in common with IN1k and rendition images such as sculptures and paintings.
ImageNet-S (IN-S) [9] consists of around 50,000 images of sketches, similar to the size of IN1k's validation set.
ImageNet-A (IN-A) [10] has 200 classes in common with IN1k and images that are classified incorrectly by a supervised ResNet-50 (RN50) [1] trained on IN1k.
ImageNet-v2 (IN-v2) [7] consists of similar images as in IN1k's test set but from a different distribution.
ObjectNet (ON) [8] has 113 common classes with IN1k and images that vary in rotation, background, and viewpoint.
iWildCam. The iWildCam [18] dataset comprises images of 182 animal species captured by various cameras traps, which are treated as different distributions. We use the WILDS benchmark [12] and manually curate low-shot subsets from `train` shift for training which has 129809 images, `val-id` shift for validation which has 7314 images, and `val-ood` shift for testing which has 14961 images. Since these sets have an imbalanced class distribution, We sample images from the different classes in 1%, 10%, and 20% ratios from the `train` shift, ensuring that each class has at least one image. This results in low-shot subsets with 1370, 12973, and 25931 images, respectively (see table 1 row 2). Additionally, we report average per-class accuracy for both validation and testing.

[2]While the optimal model checkpoint for ID performance might not be so for OOD performance, it is a widely adopted practice [37, 15] that allows for a fair comparison across different methods.

Camelyon. The Camelyon [18] dataset consists of 96×96 histopathological images that may or may not contain tumor tissue, resulting in 2 classes. These scans are sourced from different hospitals that are considered different distributions. We again use the WILDS benchmark [12] and manually curate low-shot subsets from `train` shift for training which has 302436 images, `val-id` shift for validation which has 33560 images, and `val-ood` shift for evaluation which has 34904 images.[3] The shifts are well-balanced, so we create subsets containing 1500, 7500, and 15000 images per class (see table 1 row 3). We report the average per-class accuracy for validation and testing and find that it is within 1 percentage point of top-1 accuracy.

4.2. Standard Models

Recall that to establish a baseline out-of-distribution accuracy $\beta(x)$ for a given in-domain accuracy x, fitting Eq. 1 for a set of "standard" models is required. We consider a standard set of ImageNet (IN1k) pre-trained models for this purpose, which are not subjected to additional robustness interventions or pre-training data. The selection of these models is based on their low-shot ID performance comparison on IN1k [62] or average performance on various downstream tasks [47]. We show the architectures used in our experiments ordered by the number of parameters (i.e. size) below. More detailed comparison is also shown in table 6 in appendix.

ViTS-16 ≈ RN50 < ViTB-16 ≈ RN50w2 < ViTL-16

Self-supervised models. We include the following self-supervised (SSL) models for our experiments.
SwAV [63]: SwAV is a SSL method for pre-training CNNs by predicting cluster assignments for different augmented views of an image and enforcing consistency between them. We use the RN50 and RN50w2 checkpoints for all datasets.
DINO [21]: DINO self-trains a student network to match the feature embeddings of augmented local and global views of an image to that of a teacher network which sees only the global view. We use the RN50, ViTS-16, and ViTB-16 checkpoints for all datasets.
MSN [14]: MSN matches the predicted cluster assignments for masked and unmasked augmented views of an image and performs well on low-shot ID evaluation on ImageNet. We use the ViTS-16 and ViTB-16 checkpoints for all datasets and ViTL-16 checkpoint for ImageNet.
Supervised models. For datasets other than ImageNet, we additionally include DEIT [23] (ViTS-16, ViTB-16) and supervised ResNet-50 from PyTorch [64]. Note that for ImageNet, these models violate the "low-shot" condition as they have already been trained with the full labelled dataset.

[3]We emphasize that the `val-ood` shifts are used only for evaluation. While `test-ood` shifts are also available in WILDS benchmark [12], they have similar creation processes but larger number of images than the `val-ood` shifts, so we opt for the latter due to limited compute.

Fitting Standard Models. To obtain the parameters (w and b) of the log-linear curve for $\beta(x)$ in Eq. 1 for a dataset, we first train individual models from the standard set on different low-shot subsets and full-shot subset (details in Sec. 4.2). Fine-tuning and subset details are provided in appendix, Sec. 2. We then evaluate the trained models on both ID validation and OOD test shifts, out of which *only* the former is used for hyperparameter tuning. In case of multiple OOD test shifts, we calculate the average OOD performance following previous work [16]. We assess the quality of the curve fit via mean absolute error (MAE) and coefficient of determination (R^2) of the curve on these data points, as shown in table 4. Finally, the curve $\beta(x)$ is used to calculate effective robustness of an intervention using Eq. 2.

Low-Shot Training. For low-shot training with the standard models, we freeze the pre-trained models and train a classifier on top with the available training data. We compare the following classifiers based on prior work in cross-domain few-shot learning [59] – Logistic Regression [58], Mean-Centroid Classifier [54], and Baseline++ [57] – and select the best-performing one for each dataset. While Logistic Regression performs better or on-par on both ID and OOD shifts for ImageNet and iWildCam, Baseline++ performs better on Camelyon. We provide this comparison and more details in section 1.2 of appendix.

4.3. Robustness Interventions

We consider some of the most recent methods for improving robustness to natural distribution shifts and models pre-trained on large external datasets as robustness interventions (see Sec. 2). We briefly summarize them below:

LP-FT [15]: LP-FT follows a two-stage strategy of first fine-tuning only the randomly initialized linear head followed by fine-tuning the entire model end-to-end on fully labelled datasets.

CLIP [4]: CLIP is a vision-language model that is pre-trained on a large number of ($\sim$ 400M) image-text pairs. It shows strong zero-shot performance on several datasets and is often used as the de-facto initialization by several works [15, 17, 16].

WiSE-FT [16]: WiSE-FT applies a weight-space ensemble between a zero-shot model such as CLIP and this model fine-tuned on fully labelled datasets. For IN1k pre-trained models, we ensemble between the weights of linear-probed (LP) and LP-FT checkpoints due to the absence of a zero-shot head. We use $\alpha = 0.5$ unless mentioned otherwise.

Model Soups [17]: Model Soups uses a weight-space ensemble of several models that are trained with a different epochs, learning rates, weight decay, label smoothing [65], mixup [66], and RandAugment [67]. Due to limited compute and the scale of experiments, we use a greedy soup with 9 models and again use linear-probing for the head initialization. We follow the paper for hyperparameter values.

RobustViT [68]: RobustViT first uses an unsupervised object localization method such as TokenCut [69] to dump offline segmentation maps. It then optimizes a supervised ViT's saliency maps [70] to resemble these offline segmentation maps while maintaining its classification accuracy.

For a uniform comparison across datasets, we apply the relevant interventions on MSN ViTB-16 and use it as the reference model for computing effective and relative robustness (see Sec. 3). Additionally, we include CLIP with LP-FT, WiSE-FT, and Model Soups as interventions, based on their reported performances [15, 16, 17] and strong performance on ID and OOD shifts in our experiments. Despite being amenable to low-shot training, it remains challenging to implement RobustViT on non-object centric datasets such as Camelyon due to its requirement of offline segmentation maps. We provide details on the hyperparameter choices for every intervention in section 3 of appendix.

5. Results

We now present findings for 3 key questions from Sec. 1 – (1) among ImageNet pre-trained models, which ones are more robust in low-shot regimes (see table 1) (2) how do they compare with models pre-trained on larger datasets and (3) do robustness interventions help in the low-shot regimes.

5.1. Comparing ImageNet Pre-trained Models

We compare ImageNet pre-trained models with similar number of parameters (ViTS-16 and RN50) on the basis of absolute ID and OOD performances in Fig. 3. For a uniform comparison, we randomly initialize the classifier head (see Sec. 4.2) and use the same hyperparameters for all models. It can be seen that self-supervised (SSL) ViTs often perform better than SSL CNNs on ImageNet and supervised ViTs and CNNs on iWildCam and Camelyon datasets.

However, the best initialization and model size varies for each dataset as shown in table 2. For a concise comparison, we show the average ID and OOD performances across different low-shot regimes. MSN ViTB-16 outperforms DINO ViTB-16 and MSN ViTS-16 on ImageNet, but not on iWildCam where MSN ViTS-16 performs better on OOD shift. Similarly, DINO ViTS-16 performs better than other models on both ID and OOD shifts on Camelyon.

Thus, while SSL ViTs perform better than SSL CNNs and the supervised counterparts (where applicable) on both ID and OOD shifts in the low-shot regimes, no single initialization or model size performs the best across datasets.

5.2. Pre-training Data Scale and Strategy

We question whether models pre-trained on large external datasets provide superior robustness over ImageNet

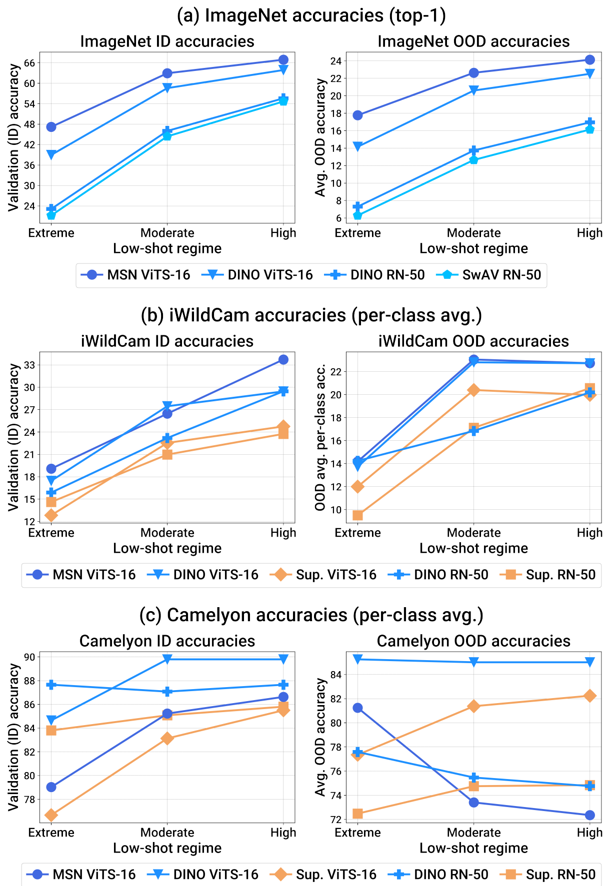

Figure 3: **Comparison of ImageNet pre-trained architectures and initializations.** With similar number of parameters, self-supervised (SSL) ViTs generally perform better on both ID and OOD shifts compared to SSL CNNs and the supervised counterparts where applicable.

	ImageNet		iWildCam		Camelyon	
	ID	OOD	ID	OOD	ID	OOD
1 MSN ViTS-16 [14]	58.99	21.51	26.41	**19.99**	83.62	75.67
2 DINO ViTS-16 [21]	53.78	19.09	24.78	19.75	**88.08**	**85.09**
3 MSN ViTB-16 [14]	**61.40**	**22.81**	24.78	19.65	86.40	78.84
4 DINO ViTB-16 [21]	56.72	21.98	**27.40**	19.82	86.93	84.33

Table 2: **Comparison of ImageNet pre-trained self-supervised ViTs.** On average across low-shot regimes, no single self-supervised initialization or model size outperforms others on ID and OOD shifts across datasets.

(IN1k) pre-trained ones in the low-shot regimes without additional interventions. We compare CLIP ViT and a supervised ViT pre-trained on ImageNet-21k (IN21k) [3, 22] with IN1k pre-trained ViT's. We use the ViTB-16 architecture with the same classifiers described in Sec. 4.2.

We again compare the absolute performance on ID and

	ImageNet		iWildCam		Camelyon	
	ID	OOD	ID	OOD	ID	OOD
1 CLIP zero shot [4, 16]	**67.93**	**57.37**	9.67	16.82	50.48	51.55
2 CLIP [4]	50.8	27.50	23.75	19.10	84.9	77.3
3 Supervised (IN21k) [3]	N/A	N/A	16.84	16.90	85.18	81.07
4 Supervised (IN1k) [23]	N/A	N/A	22.27	18.57	83.35	83.24
5 MSN (IN1k) [14]	61.40	22.81	24.78	19.65	86.40	78.84
6 DINO (IN1k) [21]	56.72	21.98	**27.40**	**19.82**	**86.93**	**84.33**

Table 3: **Comparison between ViTs pre-trained on different datasets.** On average across low-shot regimes, ImageNet (IN) pre-trained SSL ViT's such as DINO are worse than CLIP on ImageNet. However, it performs much better than CLIP and IN-21k supervised ViT on iWildCam and Camelyon datasets.

OOD shifts on average across low-shot regimes. As with the IN1k supervised models, IN21k supervised ViT violates the "low-shot" premise so we don't use it on ImageNet. For CLIP zero-shot results, we match the implementation of [16] and provide additional details in appendix, Sec 1.2.

As shown in table 3, CLIP's zero-shot performance on ID and OOD shifts on ImageNet is significantly better than both CLIP and IN1k pre-trained models. However, CLIP (zero-shot or otherwise) performs worse than IN1k pre-trained models on iWildCam and Camelyon, on which DINO performs better than other models. IN21k supervised ViT often performs significantly worse than IN1k supervised ViT on these datasets, especially on OOD shifts.

Thus, IN1k pre-trained models can perform better on both ID and OOD shifts than the models pre-trained on large external datasets in low-shot regimes, on datasets such as iWildCam and Camelyon.

5.3. Effect of Robustness Interventions

We question the extent to which existing robustness interventions improve robustness in the low-shot regimes, and we examine how the trend compares with the full-shot regime. We present the dataset-wise observations below.

ImageNet. We show the results of this experiment in Fig. 4. With MSN, interventions are largely effectively and relatively robust in the different low-shot regimes, except LP-FT in the *high* low-shot regime. While the interventions are also effectively robust in the full-shot regime, they are often not relatively robust, except RobustViT which improves robustness in all regimes.

When coupled with CLIP, Model Soups and WiSE-FT also become effectively and relatively robust in all data regimes with the latter providing largest robustness improvements. Zero-shot CLIP also improves robustness significantly in low-shot regimes (see table 5), suggesting that not using the limited training data is a better approach. However, we find that it is not the case on other datasets.

iWildCam. We show the results of this experiment in

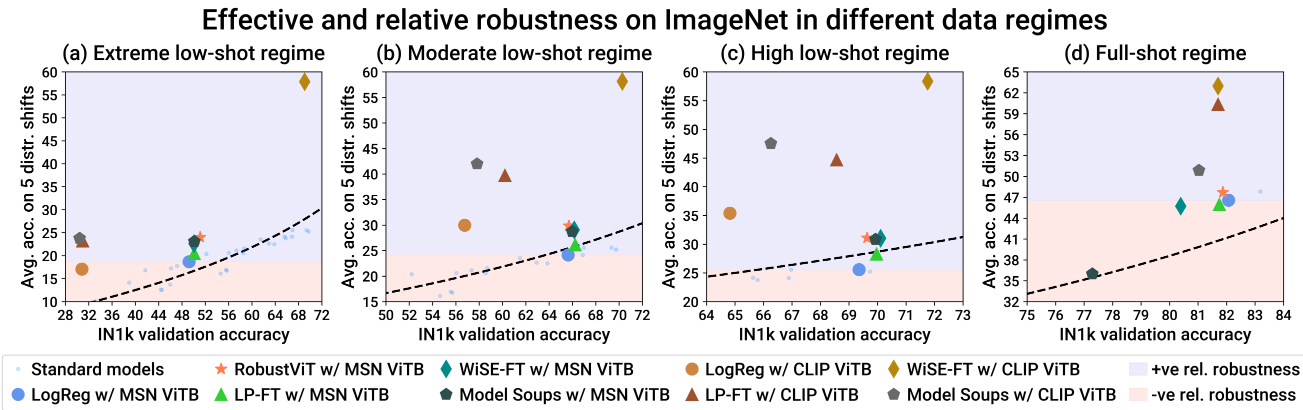

Figure 4: **Effect of robustness interventions on ImageNet.** Plots (a), (b), and (c) show performance of interventions in low-shot regimes (see table 1). Plot (d) shows performance of interventions in the full-shot regime. Interventions located above the line ($\rho > 0$) and in the blue region ($\tau > 0$) are said to improve robustness (see Sec. 3). Interventions largely improve robustness in the low-shot regimes with MSN ViTB-16, and in all data regimes when coupled with CLIP ViTB-16.

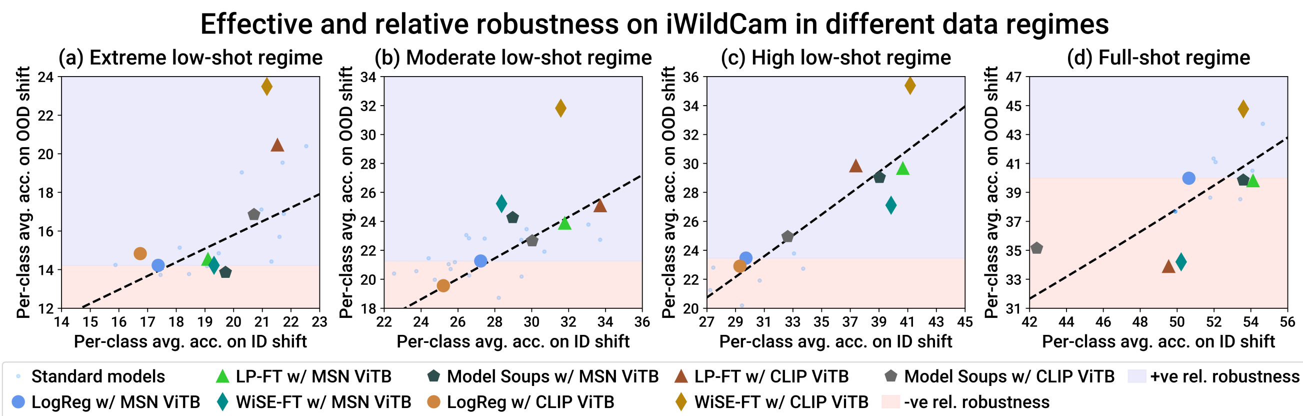

Figure 5: **Effect of robustness interventions on iWildCam.** Interventions often fail to improve robustness in both the full and low-shot regimes with MSN ViTB-16. Only WiSE-FT with CLIP ViTB-16 improves robustness in all data regimes.

Dataset	MAE ↓	R^2 ↑
1 ImageNet [5]	1.75	0.94
2 iWildCam [18]	1.50	0.96
3 Camelyon [20]	3.44	0.50

Table 4: **Quality of curve fit.** Curve $\beta(x)$ fit on the accuracies of standard models (see Sec. 4.2) leads to a relatively higher MAE and lower R^2 on Camelyon, indicating the poor quality of fit.

Fig. 5. With MSN, interventions are often relatively but not effectively robust in the low-shot regimes and neither effectively nor relatively robust in the full-shot regime. Unlike ImageNet, CLIP's zero-shot performance is quite poor (see table 5) and WiSE-FT with CLIP is the only intervention which improves robustness in all data regimes.

Camelyon. We show the results of this experiment in Fig. 6. Note that for Camelyon, the quality of curve $\beta(x)$ fit with the ID and OOD accuracies of *standard* models is relatively low compared to other datasets as shown in table 4, in which case relative robustness should be prioritized since it doesn't rely on the quality of the fit. While interventions improve relative robustness in the full-shot regime with MSN, they fail to do so in the *moderate* low-shot regime. Similarly, interventions improve robustness in the full-shot regime with CLIP, but LP-FT fails to be relatively robust in the *moderate* low-shot regime whereas Model Soups fails to be relatively robust in both the *extreme* and *moderate* low-shot regimes. Only WiSE-FT ($\alpha = 1$) with CLIP improves robustness in all data regimes. CLIP's zero-shot performance is near random as shown in table 3.

We show the effective and relative robustness of the interventions in the full-shot regimes in table 5. To complement our findings, we also highlight the interventions

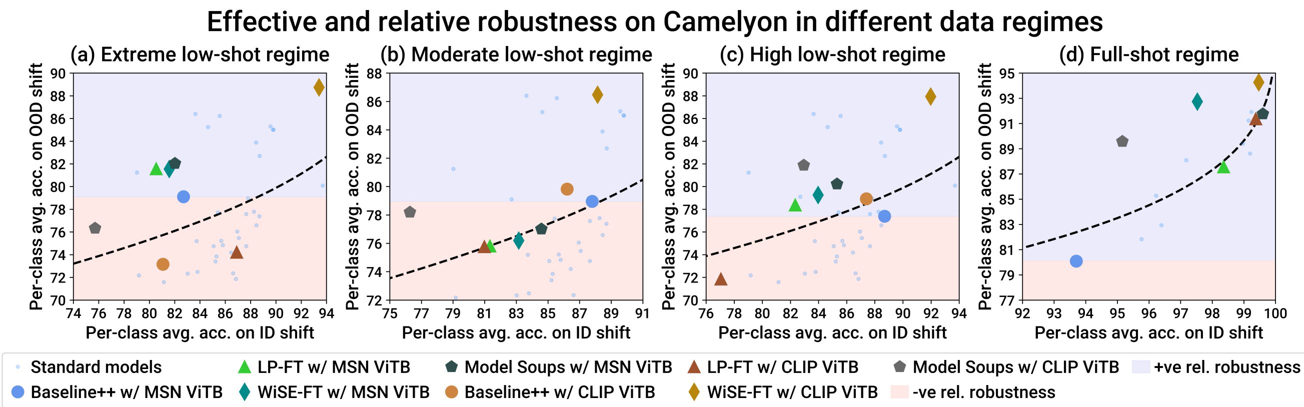

Figure 6: **Effect of robustness interventions on Camelyon.** Interventions often improve robustness in the full-shot regime with both MSN and CLIP ViTB-16 but fail to do so in *extreme* or *moderate* low-shot regimes, except WiSE-FT with CLIP ViTB-16.

	ImageNet		iWildCam		Camelyon	
	$\rho \uparrow$	$\tau \uparrow$	$\rho \uparrow$	$\tau \uparrow$	$\rho \uparrow$	$\tau \uparrow$
Full-Shot Regime						
1 LP-FT [15]	5.16	-0.61	-1.41	-0.17	-0.45	7.48
2 + CLIP	19.60*	13.77*	-3.60	-6.09	0.37	11.28
3 WiSE-FT [16]	6.66	-0.86	-3.84	-5.87	6.22	12.66
4 + CLIP	22.24*	16.41*	3.98	4.78	2.85	14.18
5 Model Soups [16]	0.53	-10.58	-0.93	-0.14	-0.35	11.68
6 + CLIP	11.00†	4.29†	3.20	-4.84	5.93	9.50
7 RobustViT [68]	6.73	1.13	N/A	N/A	N/A	N/A
8 CLIP zero-shot [4, 16]	30.28	10.79	8.46	-23.167	-14.63	-28.54

Table 5: **Robustness intervention comparison.** The table shows effective (ρ) and relative (τ) robustness of different interventions in the full-shot regime. $*$ and † denote numbers obtained from papers for ViTB-16 and ViTB-32 architecture respectively. Interventions that do not improve robustness in the full-shot regime are shown in gray, while interventions that do so are shown in black. Interventions that significantly improve robustness in *both* the full-shot regime and majority of low-shot regimes are highlighted in blue for each dataset. Robustness results for the low-shot regimes (as shown in Fig. 4, 5, and 6) are also provided in the appendix. Most interventions significantly improve robustness on ImageNet but not on other datasets, except WiSE-FT with CLIP.

which significantly[4] improve robustness across both the full-shot regime and majority of low-shot regimes for each dataset. We see that (1) most interventions significantly improve robustness on ImageNet but not on other datasets and (2) no intervention significantly improves robustness across datasets and data regimes, except WiSE-FT with CLIP.

We also measure the statistical significance of our results by obtaining the mean and standard deviation across 2 different runs and show them in table 13 of appendix. We observe that the best performing interventions such as WiSE-FT with CLIP also exhibit small (within 2 pp) variance.

Limitations. We note that there are limitations to our study. First, we were unable to theoretically analyze our results due to the vast and empirical nature of our study. Recent works [71, 72] demonstrate the data specificity of ViTs and the global semantic invariance of SSL approaches such as DINO, which can be helpful for this purpose. Second, we were unable to observe the effects of in-domain SSL pretraining on datasets other than ImageNet. Recent work [62] has also shown that the current objectives of self-supervised methods such as MSN and DINO might not be suitable for class-imbalanced datasets (e.g. iWildCam). Third, while we incorporate different kinds of augmentations and loss functions as a part of interventions such as Model Soups, singly analyzing their effect on robustness in low-shot regimes remains an avenue for future work.

6. Conclusion

We conclude our study of low-shot robustness to several natural distribution shifts, which addresses the gap in the literature and marks the first in-depth study of its kind. Taken together, our results demonstrate that: (1) Modern architectures (i.e. ViT) and pre-training strategies (i.e. self-supervised learning) lead to better robustness in low-shot regimes, but the best initialization and model size is dataset dependent. (2) Without additional interventions, large-scale vision-language pre-training can be underwhelming compared to ImageNet pre-trained models on datasets other than ImageNet. (3) Robustness in the full-shot regime may not imply robustness in low-shot regimes on datasets other than ImageNet. While the performance of interventions is largely dependent on datasets and initializations, ensembling in weight-space seems promising to bridge this gap. We hope that our study will motivate researchers to also focus on this problem of practical importance.

[4] We use the standard deviation of residuals obtained after fitting $\beta(x)$ to determine significance, and provide more details in appendix, Sec. 4.

Acknowledgements

This work was supported in part by NASA ULI, NSF #2144194, and ARL. We thank the anonymous reviewers for their thoughtful comments and suggestions. We also thank our internal reviewers Anisha Pal, Bharat Goyal, and Pratik Ramesh for their constructive suggestions on an early draft of this paper.

References

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016. 1, 4

[2] A. Kolesnikov, L. Beyer, X. Zhai, J. Puigcerver, J. Yung, S. Gelly, and N. Houlsby, "Big transfer (bit): General visual representation learning," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pp. 491–507, Springer, 2020. 1

[3] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020. 1, 6

[4] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, pp. 8748–8763, PMLR, 2021. 1, 2, 3, 5, 6, 8

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, Ieee, 2009. 1, 3, 4, 7

[6] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann, "Shortcut learning in deep neural networks," *Nature Machine Intelligence*, vol. 2, no. 11, pp. 665–673, 2020. 1

[7] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar, "Do imagenet classifiers generalize to imagenet?," in *International conference on machine learning*, pp. 5389–5400, PMLR, 2019. 1, 2, 3, 4

[8] A. Barbu, D. Mayo, J. Alverio, W. Luo, C. Wang, D. Gutfreund, J. Tenenbaum, and B. Katz, "Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models," *Advances in neural information processing systems*, vol. 32, 2019. 1, 2, 4

[9] H. Wang, S. Ge, Z. Lipton, and E. P. Xing, "Learning robust global representations by penalizing local predictive power," *Advances in Neural Information Processing Systems*, vol. 32, 2019. 1, 2, 4

[10] D. Hendrycks, K. Zhao, S. Basart, J. Steinhardt, and D. Song, "Natural adversarial examples," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15262–15271, 2021. 1, 2, 4

[11] D. Hendrycks, S. Basart, N. Mu, S. Kadavath, F. Wang, E. Dorundo, R. Desai, T. Zhu, S. Parajuli, M. Guo, *et al.*, "The many faces of robustness: A critical analysis of out-of-distribution generalization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8340–8349, 2021. 1, 2, 4

[12] P. W. Koh, S. Sagawa, H. Marklund, S. M. Xie, M. Zhang, A. Balsubramani, W. Hu, M. Yasunaga, R. L. Phillips, I. Gao, *et al.*, "Wilds: A benchmark of in-the-wild distribution shifts," in *International Conference on Machine Learning*, pp. 5637–5664, PMLR, 2021. 1, 2, 4

[13] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022. 1, 2

[14] M. Assran, M. Caron, I. Misra, P. Bojanowski, F. Bordes, P. Vincent, A. Joulin, M. Rabbat, and N. Ballas, "Masked siamese networks for label-efficient learning," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXI*, pp. 456–473, Springer, 2022. 1, 2, 4, 6

[15] A. Kumar, A. Raghunathan, R. Jones, T. Ma, and P. Liang, "Fine-tuning can distort pretrained features and underperform out-of-distribution," *arXiv preprint arXiv:2202.10054*, 2022. 1, 2, 4, 5, 8

[16] M. Wortsman, G. Ilharco, J. W. Kim, M. Li, S. Kornblith, R. Roelofs, R. G. Lopes, H. Hajishirzi, A. Farhadi, H. Namkoong, *et al.*, "Robust fine-tuning of zero-shot models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7959–7971, 2022. 1, 2, 3, 5, 6, 8

[17] M. Wortsman, G. Ilharco, S. Y. Gadre, R. Roelofs, R. Gontijo-Lopes, A. S. Morcos, H. Namkoong, A. Farhadi, Y. Carmon, S. Kornblith, *et al.*, "Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time," in *International Conference on Machine Learning*, pp. 23965–23998, PMLR, 2022. 1, 2, 5

[18] S. Beery, E. Cole, and A. Gjoka, "The iwildcam 2020 competition dataset," *arXiv preprint arXiv:2004.10340*, 2020. 1, 2, 3, 4, 7

[19] O. Wiles, S. Gowal, F. Stimberg, S.-A. Rebuffi, I. Ktena, K. D. Dvijotham, and A. T. Cemgil, "A fine-grained analysis on distribution shift," in *International Conference on Learning Representations*, 2022. 1, 2

[20] P. Bandi, O. Geessink, Q. Manson, M. Van Dijk, M. Balkenhol, M. Hermsen, B. E. Bejnordi, B. Lee, K. Paeng, A. Zhong, *et al.*, "From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge," *IEEE transactions on medical imaging*, vol. 38, no. 2, pp. 550–560, 2018. 2, 3, 4, 7

[21] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021. 2, 4, 6

[22] T. Ridnik, E. Ben-Baruch, A. Noy, and L. Zelnik-Manor, "Imagenet-21k pretraining for the masses.[arxiv]," *DOI: https://doi. org/10.48550/arXiv*, vol. 2104, 2021. 2, 6

[23] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *International conference on machine learning*, pp. 10347–10357, PMLR, 2021. 2, 4, 6

[24] J. Quionero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, *Dataset Shift in Machine Learning*. The MIT Press, 2009. 2

[25] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *CVPR 2011*, pp. 1521–1528, IEEE, 2011. 2

[26] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017. 2

[27] F. Tramèr, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel, "Ensemble adversarial training: Attacks and defenses," *arXiv preprint arXiv:1705.07204*, 2017. 2

[28] R. Geirhos, C. R. Temme, J. Rauber, H. H. Schütt, M. Bethge, and F. A. Wichmann, "Generalisation in humans and deep neural networks," *Advances in neural information processing systems*, vol. 31, 2018. 2

[29] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song, "Robust physical-world attacks on deep learning visual classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1625–1634, 2018. 2

[30] M. A. Alcorn, Q. Li, Z. Gong, C. Wang, L. Mai, W.-S. Ku, and A. Nguyen, "Strike (with) a pose: Neural networks are easily fooled by strange poses of familiar objects," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4845–4854, 2019. 2

[31] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," *arXiv preprint arXiv:1903.12261*, 2019. 2

[32] M. M. Naseer, K. Ranasinghe, S. H. Khan, M. Hayat, F. Shahbaz Khan, and M.-H. Yang, "Intriguing properties of vision transformers," *Advances in Neural Information Processing Systems*, vol. 34, pp. 23296–23308, 2021. 2

[33] Y. Bai, J. Mei, A. L. Yuille, and C. Xie, "Are transformers more robust than cnns?," *Advances in Neural Information Processing Systems*, vol. 34, pp. 26831–26843, 2021. 2

[34] D. Zhou, Z. Yu, E. Xie, C. Xiao, A. Anandkumar, J. Feng, and J. M. Alvarez, "Understanding the robustness in vision transformers," in *International Conference on Machine Learning*, pp. 27378–27394, PMLR, 2022. 2

[35] Z. Wang, Y. Bai, Y. Zhou, and C. Xie, "Can cnns be more robust than transformers?," *arXiv preprint arXiv:2206.03452*, 2022. 2

[36] R. Geirhos, K. Narayanappa, B. Mitzkus, T. Thieringer, M. Bethge, F. A. Wichmann, and W. Brendel, "Partial success in closing the gap between human and machine vision," *Advances in Neural Information Processing Systems*, vol. 34, pp. 23885–23899, 2021. 2

[37] R. Taori, A. Dave, V. Shankar, N. Carlini, B. Recht, and L. Schmidt, "Measuring robustness to natural distribution shifts in image classification," *Advances in Neural Information Processing Systems*, vol. 33, pp. 18583–18599, 2020. 2, 3, 4

[38] C. Yadav and L. Bottou, "Cold case: The lost mnist digits," *Advances in neural information processing systems*, vol. 32, 2019. 2

[39] J. Miller, K. Krauth, B. Recht, and L. Schmidt, "The effect of natural distribution shift on question answering models," in *International Conference on Machine Learning*, pp. 6905–6916, PMLR, 2020. 2

[40] J. P. Miller, R. Taori, A. Raghunathan, S. Sagawa, P. W. Koh, V. Shankar, P. Liang, Y. Carmon, and L. Schmidt, "Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization," in *International Conference on Machine Learning*, pp. 7721–7735, PMLR, 2021. 2, 3

[41] A. D'Amour, K. Heller, D. Moldovan, B. Adlam, B. Alipanahi, A. Beutel, C. Chen, J. Deaton, J. Eisenstein, M. D. Hoffman, *et al.*, "Underspecification presents challenges for credibility in modern machine learning," *Journal of Machine Learning Research*, 2020. 2

[42] D. Teney, Y. Lin, S. J. Oh, and E. Abbasnejad, "Id and ood performance are sometimes inversely correlated on real-world datasets," *arXiv preprint arXiv:2209.00613*, 2022. 2, 3

[43] S. Azizi, B. Mustafa, F. Ryan, Z. Beaver, J. Freyberg, J. Deaton, A. Loh, A. Karthikesalingam, S. Kornblith, T. Chen, V. Natarajan, and M. Norouzi, "Big self-supervised models advance medical image classification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3478–3488, October 2021. 2

[44] F. Wenzel, A. Dittadi, P. Gehler, C.-J. Simon-Gabriel, M. Horn, D. Zietlow, D. Kernert, C. Russell, T. Brox, B. Schiele, *et al.*, "Assaying out-of-distribution generalization in transfer learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 7181–7198, 2022. 2

[45] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020. 2

[46] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*, pp. 1597–1607, PMLR, 2020. 2

[47] L. Ericsson, H. Gouk, and T. M. Hospedales, "How well do self-supervised models transfer?," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5414–5423, 2021. 2, 3, 4

[48] L. Jing and Y. Tian, "Self-supervised visual feature learning with deep neural networks: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 11, pp. 4037–4058, 2020. 2

[49] X. Liu, F. Zhang, Z. Hou, L. Mian, Z. Wang, J. Zhang, and J. Tang, "Self-supervised learning: Generative or contrastive," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 1, pp. 857–876, 2021. 2

[50] B. Lake, R. Salakhutdinov, J. Gross, and J. Tenenbaum, "One shot learning of simple visual concepts," in *Proceedings of the annual meeting of the cognitive science society*, vol. 33, 2011. 3

[51] B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum, "Human-level concept learning through probabilistic program induction," *Science*, vol. 350, no. 6266, pp. 1332–1338, 2015. 3

[52] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra, *et al.*, "Matching networks for one shot learning," *Advances in neural information processing systems*, vol. 29, 2016. 3

[53] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *International conference on machine learning*, pp. 1126–1135, PMLR, 2017. 3

[54] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," *Advances in neural information processing systems*, vol. 30, 2017. 3, 5

[55] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1199–1208, 2018. 3

[56] K. Lee, S. Maji, A. Ravichandran, and S. Soatto, "Meta-learning with differentiable convex optimization," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10657–10665, 2019. 3

[57] W.-Y. Chen, Y.-C. Liu, Z. Kira, Y.-C. F. Wang, and J.-B. Huang, "A closer look at few-shot classification," *arXiv preprint arXiv:1904.04232*, 2019. 3, 5

[58] Y. Tian, Y. Wang, D. Krishnan, J. B. Tenenbaum, and P. Isola, "Rethinking few-shot image classification: a good embedding is all you need?," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pp. 266–282, Springer, 2020. 3, 5

[59] Y. Guo, N. C. Codella, L. Karlinsky, J. V. Codella, J. R. Smith, K. Saenko, T. Rosing, and R. Feris, "A broader study of cross-domain few-shot learning," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16*, pp. 124–141, Springer, 2020. 3, 5

[60] Y. Zhang, Y. Zheng, X. Xu, and J. Wang, "How well do self-supervised methods perform in cross-domain few-shot learning?," *arXiv preprint arXiv:2202.09014*, 2022. 3

[61] D. Das, S. Yun, and F. Porikli, "Confess: A framework for single source cross-domain few-shot learning," in *International Conference on Learning Representations*, 2022. 3

[62] M. Assran, R. Balestriero, Q. Duval, F. Bordes, I. Misra, P. Bojanowski, P. Vincent, M. Rabbat, and N. Ballas, "The hidden uniform cluster prior in self-supervised learning," *arXiv preprint arXiv:2210.07277*, 2022. 4, 8

[63] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," *Advances in neural information processing systems*, vol. 33, pp. 9912–9924, 2020. 4

[64] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019. 4

[65] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016. 5

[66] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017. 5

[67] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "Randaugment: Practical automated data augmentation with a reduced search space," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 702–703, 2020. 5

[68] H. Chefer, I. Schwartz, and L. Wolf, "Optimizing relevance maps of vision transformers improves robustness," *arXiv preprint arXiv:2206.01161*, 2022. 5, 8

[69] Y. Wang, X. Shen, S. X. Hu, Y. Yuan, J. L. Crowley, and D. Vaufreydaz, "Self-supervised transformers for unsupervised object discovery using normalized cut," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14543–14553, 2022. 5

[70] H. Chefer, S. Gur, and L. Wolf, "Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 397–406, 2021. 5

[71] N. Park and S. Kim, "How do vision transformers work?," 2022. 8

[72] N. Park, W. Kim, B. Heo, T. Kim, and S. Yun, "What do self-supervised vision transformers learn?," 2023. 8