Towards Better Understanding of Domain Shift on Linear-Probed Visual Foundation Models

Eric Heim
Software Engineering Institute
Carnegie Mellon University
Pittsburgh, PA 15213
`etheim@sei.cmu.edu`

Abstract

Visual foundation models have recently emerged to offer similar promise as their language counterparts: The ability to produce representations of visual data that can be successfully used in a variety of tasks and contexts. One common way this is shown in research literature is through "domain generalization" experiments of linear models trained from representations produced by foundation models (i.e. linear probes). These experiments largely limit themselves to a small number of benchmark data sets and report accuracy as the single figure of merit, but give little insight beyond these numbers as to how different foundation models represent shifts. In this work we perform an empirical evaluation that expands the scope of previously reported results in order to give better understanding into how domain shifts are modeled. Namely, we investigate not just how models generalize across domains, but how models may enable domain transfer. Our evaluation spans a number of recent visual foundation models and benchmarks. We find that not only do linear probes fail to generalize on some shift benchmarks, but linear probes trained on some shifted data achieve low *train* accuracy, indicating that accurate transfer of linear probes is not possible with some visual foundation models.

1 Introduction

An emerging trend in computer vision research is the development of general-purpose neural network models that are meant to be adapted to a variety of tasks and application contexts. These visual "foundation" models can be fine-tuned using application-specific data to perform tasks ranging from object detection to semantic segmentation to image classification. In many cases, high-performant models can be learned by training simple, small models from representations produced by a larger, more complex foundation models and with relatively little training data. As such, these foundation models have the potential to enable computational and data efficient means to build state-of-the-art predictive models, effectively lowering the barrier to powerful computer vision capabilities.

One common adaptation strategy is known as "linear-probing" where a simple linear model is trained to map a foundation model's representation to logits used for classification. While their simplicity has benefits, it also makes linear probes highly reliant on the expressivity of the foundation models they are trained with. In order for linear probes to successfully classify images, the foundation models they are built from must be able to produce representations of images that are discriminative with respect to classes in the application domain.

In this work, we aim to better understand the capability of current visual foundation models when used as a basis for linear probes. More specifically, we focus on the problem of learning under *domain*

I Can't Believe It's Not Better! Failure Modes in the Age of Foundation Models Workshop.
37th Conference on Neural Information Processing Systems (NeurIPS 2023)

shift, where train and test distributions differ. Linear probed foundation models seem uniquely suited for this learning setting, as foundation models are meant to produce generally applicable representations that can be applied to a many different domains, and linear probing does not change these representations, but builds simple models on top of them. Thus, much of the generalization benefits in the original foundation model's representation should remain intact.

We expand current understanding of the performance, utility, and empirical characteristics of linear probed foundation models by performing a series of experiments on a number of current, popular models across a variety of domain shift benchmarks. Through these, we provide new empirical evidence to the limits of current foundation models, as well as some insight into how foundation models differ. We find that 1) perhaps unsurprisingly, linear probes do not generalize to all shifts, but also 2) for some benchmarks, linear probes are not expressive enough to achieve high *train* accuracy, implying even supervised domain transfer of a linear probe would be difficult. Finally, we highlight trends in performance across different foundation model pre-training strategies and architectures.

2 Preliminaries

A *visual foundation model* can be defined as a function $f_\theta : \mathbb{R}^{n_1 \times n_2 \times n_3} \mapsto \mathbb{R}^d$ that maps an image to a vector representation that is meant to be adapted to a down-stream visual prediction task. Most foundation models are characterized by 1) the architecture of f, and 2) the pre-training task used to find parameters θ. Prior to the term "foundation model" becoming widely used, convolutional or residual neural networks pretrained for image classification were often used as "backbones" for visual tasks such as for object detection [14, 31, 25]. More recently, focus has shifted from using repurposed models to ones trained with the explicit goal of being adapted to a variety of visual learning tasks. These are typically transformers [10], and pre-training tasks are mostly either weakly [29] or self-supervised [6, 37, 16] learning tasks on large-scale data scraped from the web. Thus, recent visual foundation models distinguish themselves from traditional backbones largely by pre-training objectives, the scale of pre-training data, and the size and complexity of network architectures.

2.1 Adapting Visual Foundation Models

The two most popular ways of adapting visual foundation models to down-stream tasks are *fine-tuning* and *last layer(s) retraining*. In both cases, model parameters are added to f that map from a foundation model's representation of an input to predictions. Let $g_\phi : \mathbb{R}^d \mapsto \mathbb{R}^{d'}$, be the added task-specific portion of model and $g \circ f$ be the full end-to-end model for the down-stream task. In fine-tuning, all parameters $\theta \cup \phi$ are optimized using an objective function and data relevant for the down-stream application and task, while in last-layer retraining just ϕ are optimized. In this work, we focus on the last-layer retraining case where $g_{\mathbf{w},\mathbf{b}}(\mathbf{z}) = \mathbf{w}\mathbf{z} + \mathbf{b}$, commonly called *linear-probing* for image classification. Here, $\mathbf{w} \in \mathbb{R}^{c \times d}$ and $\mathbf{b} \in \mathbb{R}^c$, where c is the number of classes.

While fine-tuning typically produces more accurate classifiers, there are a number of advantages to linear-probing. First, linear-probes are less computationally demanding to train than an entire end-to-end model, so the computational barrier to create image classifiers is lower than with fine-tuning. Second, with a standard cross-entropy loss, optimizing for $\mathbf{w}$ and $\mathbf{b}$ is a convex optimization problem, for which there are a number of efficient, easy to use linear solvers that can find globally optimal solutions, even for high-dimensional d. Third, because linear-probes are much simpler models, fewer labeled instances are typically required for training. Finally, it has been shown that full fine-tuning can distort the features learned during pre-training [24], resulting in classifiers that do not generalize well to domain shifts. For these reasons, not only is linear-probing attractive for it's advantages in practicality, but also because of its potential to generalize across domains.

2.2 Domain Shifts

In the domain shift setting it is assumed that a classifier is trained on a set of n labeled images $\left\{\left(\mathbf{x}_s^1, y_s^1\right), ..., \left(\mathbf{x}_s^n, y_s^n\right)\right\} \sim \mathcal{D}_s$, where $\mathbf{x}_s^i$ is an image, y_s^i is a label, and $\mathcal{D}_s$ is a *source distribution/domain*. Then, during deployment, the classifier will be tasked to predict the correct class for instances $\left\{\left(\mathbf{x}_t^1, y_t^1\right), ...\right\} \sim \mathcal{D}_t$, where $\mathcal{D}_t$ is a *target distribution/domain*. We assume that the data-generating distributions have *shifted* from train time to deployment ($\mathcal{D}_s \neq \mathcal{D}_t$). We also assume that $\forall_{i,j}\ y_s^i,\ y_t^j \in \{1, ...c\}$, i.e. labels from both domains come from the same closed set of classes.

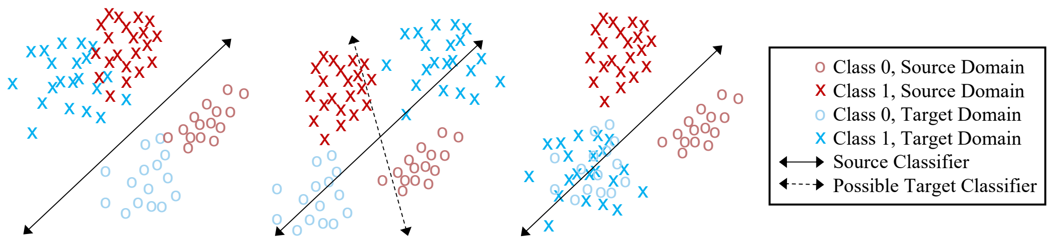

Figure 1: Depictions of the Effects of Domain Shifts on Linear Probes.

Note that for analysis and more focused methodology development it is often useful to assume a formal relationship between $\mathcal{D}_s$ and $\mathcal{D}_t$ (such as covariate or label shift). Because we focus on foundation models that make no formal claims of the kinds of shifts they model, we intentionally make no relationship explicit in our problem definition.

In *domain generalization*, the goal for a classifier is to learn solely from data from $\mathcal{D}_s$ to successfully classify images from $\mathcal{D}_t$. In general without further assumptions on the nature of shifts, a classifier that achieves high accuracy in a source domain does not imply high accuracy in the target domain [13]. While this may make domain generalization seem hopeless, there are a number of techniques that make implicit or explicit assumptions that attempt to solve this problem [33]. On the other hand, *domain transfer* assumes that a limited number of (unlabeled or labeled) images from the target domain $\mathcal{D}_t$ are also available to learn a classifier. Here, the assumption is that data or model parameters from the source domain can lessen the data burden required to learn a model in the target domain. Most recent methods that attempt domain transfer (also sometimes known as domain adaptation) focus on learning end-to-end neural network models [12, 27, 7] instead of building simple, parameter-efficient extensions from foundation models, such as linear probes.

It is important to note that while domain generalization and domain transfer are related, they have different implications for the representations used to learn linear probes. Consider the notional linear classification examples in Figure 1. On the left, the classifier trained on the source data is able to separate both the source and target classes well. This is because both the source and target classes are represented similarly, and as a result, the source classifier generalizes well to the target domain. In the center, the source classifier does not separate the target classes well, as there is a shift that makes many instances of the classes cross the linear decision boundary. However, given data from the target distribution, a linear model could be learned that separates the target classes. This indicates that while domain generalization using a linear model with this representation is challenging, successful domain transfer is possible. Now consider the scenario on the right. Here, the source classsifier is approximately as accurate as in the center example. However, no linear model can separate the target classes well, thus neither accurate domain generalization nor domain transfer is possible with a linear model. We argue that because no foundation model can produce a representation that generalizes to all shifts, it is important to understand whether foundation models produce representations amenable to domain generalization, domain transfer, or neither. In the next section, we perform a series of experiments that 1) expand upon existing published benchmark results in domain generalization for linear probed foundation models and 2) provide some insight into whether popular visual foundation models produce representations amenable to domain transfer of linear probes.

3 Experiments

In our experiments we evaluate the following models, as “base models” for linear probes:

ResNet50 [17] A standard 50 layer residual network pretrained on ImageNet-1K [9].

ConvNextV2 [35] A convolutional network that has been scaled to the size of visual transformers using many of the same advancements including self-supervised pre-training (on ImageNet-22k [9])

CLIP [29] A visual transformer, pre-trained using Contrastive Language-Image Pre-training (CLIP). CLIP learns image representations by co-embedding images and corresponding captions from a data set consisting of 400 million image/caption pairs gathered by querying a web search engine.

Base Model	Size	Val	v2	C	A	R	Cartoon	Drawing
ResNet50		78.204	66.414	5.982	6.320	24.807	56.146	27.116
ConvNeXtV2	Large	86.074	75.874	53.802	38.013	46.623	79.524	56.942
	Tiny	81.316	69.402	41.470	13.480	36.947	69.242	43.794
CLIP	Large	83.246	72.727	22.374	43.600	57.777	71.388	52.312
	Base	79.000	67.970	11.342	25.480	44.690	61.716	37.042
DINOV2	Large	84.920	75.969	49.946	50.307	57.503	80.136	62.766
	Small	80.166	69.467	11.632	18.693	38.743	68.914	34.776

Table 1: Source Model ImageNet Experiment Results

Base Model	Size	Val	Clipart	Quickdraw	Infograph	Painting	Sketch
ResNet50		80.068	34.345	1.960	17.697	34.345	23.726
ConvNeXtV2	Large	88.394	57.527	3.933	27.613	57.527	44.348
	Tiny	84.505	47.481	6.683	21.003	47.481	36.644
CLIP	Large	90.842	74.073	15.337	46.318	74.073	64.453
	Base	88.250	64.972	9.600	41.690	64.972	55.231
DINOV2	Large	88.463	67.708	7.538	34.711	67.708	59.479
	Small	85.504	51.371	6.494	26.397	51.371	43.251

Table 2: Source Model DomainNet Experiment Results

Base Model	Size	iWC(ID)	iWC(OOD)	FMOW(ID)	FMOW(OOD)	C17(ID)	C17(OOD)
ResNet50		67.844	62.684	35.711	31.278	96.332	87.832
ConvNeXtV2	Large	74.614	71.637	44.107	39.262	97.625	92.700
	Tiny	71.315	70.945	38.501	35.191	95.328	90.565
CLIP	Large	72.750	73.279	55.187	49.643	96.812	91.601
	Base	70.530	70.781	49.254	44.179	96.386	90.653
DINOV2	Large	76.049	76.039	48.760	42.962	97.288	91.287
	Small	73.252	74.724	41.414	37.168	95.647	93.563

Table 3: Source Model Wilds Data Sets Experiment Results

DINOV2 [26] A visual transformer pre-trained using a number of separately developed self-supervised learning techniques on a collection of various data sources, together called LVD-142M.

The ResNet50 and ConvNeXtV2 models represent classic "backbone" models orginally trained for classification. CLIP and DINOV2 represent popular, visual transformer-based foundation models learned via weak and self supervised techniques, respectively. We evaluate two versions of each base model (besides ResNet50): "Large" variants that are roughly the same size in terms of number of parameters, and smaller variants (i.e. "Tiny", "Small", or "Base", depending on availability). The large variants allow for direct comparisons across base model types, while their smaller variants allow for comparisons within base model types to see the effect of model size.

Additionally, we utilize three different sets of domain shift benchmark data sets:

ImageNet A popular image classification benchmark with 1,000 classes. For source data we use the train set of ImageNet-1k [9], and for shifted target sets we use the test sets of ImageNetv2 [30], ImageNet-C [19], ImageNet-A [20], ImageNet-R [18], and ImageNet-Cartoon/Drawing [32].

DomainNet [27] A collection of six data sets, each labeled with the same set of 345 coarse-grained classes. For source data we use the train set of the "Real" data set. For shifted target data sets we use the train sets of the "Clipart", "Infograph", "Painting", "Quickdraw", and "Sketch" data sets.

Wilds [23] A collection of ten data sets, of which we use three: iWildCam (iWC) [5], Functional Map of the World (FMOW) [8], and Camelyon17 (C17) [4]. We use the in-distribution (ID) train sets as the source data sets, and the out-of-distribution (OOD) validation sets as the target sets for each.

Though important for fully interpreting results, we omit broad discussion of the shifts induced for each of these benchmarks due to space limitations, but do highlight some of them in subsequent

Base Model	Size	v2	C	A	R	Cartoon	Drawing
ResNet50		99.970	89.694	100	99.157	99.852	99.794
ConvNeXtV2	Large	99.905	99.082	100	99.630	99.498	99.616
	Tiny	99.675	92.856	99.560	94.400	97.706	95.204
CLIP	Large	99.994	84.254	99.987	99.773	99.554	98.530
	Base	99.941	63.864	99.573	98.413	96.986	90.646
DINOV2	Large	99.905	98.480	100	99.667	99.410	99.494
	Small	98.965	57.936	95.467	89.567	91.974	80.850

Table 4: Target Model ImageNet Experiment Results

Base Model	Size	Clipart	Quickdraw	Infograph	Painting	Sketch
ResNet50		98.765	95.019	84.585	98.861	95.904
ConvNeXtV2	Large	98.777	83.316	84.285	98.774	96.797
	Tiny	95.591	74.833	65.975	95.505	86.238
CLIP	Large	98.649	83.421	93.265	98.652	97.256
	Base	96.764	75.115	86.192	96.755	91.809
DINOV2	Large	98.213	84.388	84.973	98.273	95.709
	Small	89.411	69.144	59.876	89.447	80.134

Table 5: Target Model DomainNet Experiment Results

Base Model	Size	iWC(OOD)	FMOW(OOD)	C17(OOD)
ResNet50		97.542	94.839	97.536
ConvNeXtV2	Large	95.576	85.562	98.621
	Tiny	94.609	70.839	96.891
CLIP	Large	95.541	85.467	98.189
	Base	94.127	73.670	97.370
DINOV2	Large	96.069	82.183	98.182
	Small	92.000	60.621	97.089

Table 6: Target Model Wilds Data Sets Experiment Results

sections as part of the discussion of results. Further details pertaining to the base models, the methods used to train linear probes, and data preparation and preprocessing can be found in Appendix A.

3.1 Domain Generalization Experiments

We began our experiments by investigating the question: *Do modern visual foundation models produce representations that can generalize across shifts?* Tables 1, 2, and 3 show results of linear probes trained on the training source sets described above. For each base model and size variant (rows), we report accuracy values of the linear probes on 1) the validation sets of the source data (3rd column), and 2) their corresponding shifted target data sets (columns after the third). Note that for the Wilds experiments in Table 3, the columns designated (ID) are the source validation sets.

We were able to achieve linear probing validation performance comparable (within 1-3 points of accuracy) to those published in the original papers these methods were introduced, as well as results similar to domain generalization results reported in [26]. Consistent with other domain generalization results, almost all models perform worse on the shifted targets than the validation sets. In some cases (e.g. the Wilds data sets), the drop in performance from source validation to target sets is minimal, especially for DINOV2 and CLIP base models. However, on other experiments (e.g. ImageNet-C, DomanNet-Quickdraw) the drop in accuracy is considerably larger. This shows that for some shifts, foundation models tend to generalize well while for others they notably fail.

3.2 Target Class Discriminability Experiments

Since the evaluated foundation models failed to generalize on a number of the benchmarks we tested them on, we then shifted our investigation to whether they had the representational power to enable domain transfer, a potential solution when generalization fails. More specifically we asked: *When foundation models fail to generalize, does there exist* ***any*** *linear probe that can discriminate target classes?* Stated another way: Is accurate linear probe transfer possible in these benchmarks?

Tables 4, 5, and 6 show the train accuracy of linear probes when trained on the target sets themselves, thus showing the upper limit on domain transfer accuracy. For some data sets (e.g. ImageNet-A) the target probes can be trained to perfect or near perfect accuracy despite the generalization accuracy being considerably lower when trained on source data. In these cases, the base models are expressive enough to discriminate target classes, but fail at generalizing from the source set. In other cases (e.g. DomainNet-Quickdraw and FMOW), training linear probes on the target data cannot result in near perfect target accuracy. For example, our results show that no linear probe trained from a CLIP-Large model, learned via transfer from source data or otherwise, can achieve better than $\sim$85% accuracy on the DomainNet-Quickdraw data set. These results indicate a fundamental limitation in these foundation models' ability to represent data in some target domains, which can be seen as a potentially significant shortcoming: for some foundation models and for some shifts, accurate transfer of linear probes is not possible and more sophisticated techniques must be used.

3.3 Discussion of Results

Trends A number of trends emerged from these experiments. First, larger models outperformed their smaller counterparts in terms of generalization accuracy as well as target class discriminability. Second, neither CLIP and DINOV2 uniformly outperformed the other. While deeper investigation into these two models is required to understand why one performs better than another on a given benchmark, their relative generalization accuracy seems to correspond to relative target accuracy. This may indicate that discriminability of classes in the target domains plays a role in domain generalization. Third, ConvNeXtV2, despite being trained for ImageNet classification and not for the specific goal of being used as a foundation model, performs particular well on a number of benchmarks. This may be expected for the ImageNet benchmarks, but it also generalizes better than CLIP on iWC and better than both CLIP and DINOV2 on FMOW. This shows that classical pre-training methods, such as training for large-scale image classification, can still lead to models competitive with those that utilize more recently-developed pre-training techniques.

Value of Evaluating Target Class Discriminability We argue that simply evaluating for domain generalization is insufficient when assessing foundation models. In many of the benchmarks it is unclear whether it is practically reasonable to expect a model to generalize from source to target. For instance, one could argue that it's not only desirable but achievable to build classifiers that robust to the shifts from the Wilds benchmarks (changes in imaging equipment/procedures, geographic location, etc.). It is less clear whether a classifier trained on real images of objects should be expected generalize to hastily drawn, black and white sketches of those objects as in DomainNet. Expecting foundation models to represent classes in a way that universally generalize to all realizable shifts, even in very constrained environments, seems unreasonable. For this reason, we argue that it equally important to evaluate whether a foundation model can be useful for transfer from a source to a target domain as it is to evaluate whether it generalizes across domains. Our target discriminability experiments represents a basic first step in understanding if efficient transfer is possible.

Future Work While we believe this work provides more empirical evidence to the strengths and weaknesses of visual foundation models in representing data across domains, it is limited by not directly measuring the performance of a linear probe learned via transfer from source to target. Indeed, our experiments show the best possible accuracy that a linear probe could achieve in the target domain, but not whether a source domain can be used to learn an accurate target classifier in this setting. To do this, there needs to be further study into the appropriate *mechanism* to transfer a source probe to a target domain, which would likely motivate transfer learning methods specific to linear probing foundation models. We hope this work provides an initial basis for such work.

We argue in this work for the importance not just of evaluating foundation models for their ability to generalize across domains, but whether they are amenable to transfer across them. More generally, we argue that foundation model research could benefit from more well-defined goals. In Appendix B we elaborate on what we feel are possible targets for visual foundation model research.

References

[1] Sanjeev Arora, Nadav Cohen, and Elad Hazan. On the optimization of deep networks: Implicit acceleration by overparameterization. In *International Conference on Machine Learning*, pages 244–253. PMLR, 2018.

[2] Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. *Advances in Neural Information Processing Systems*, 32, 2019.

[3] Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*, pages 322–332. PMLR, 2019.

[4] Peter Bandi, Oscar Geessink, Quirine Manson, Marcory Van Dijk, Maschenka Balkenhol, Meyke Hermsen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao 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*, 38(2):550–560, 2018.

[5] Sara Beery, Arushi Agarwal, Elijah Cole, and Vighnesh Birodkar. The iwildcam 2021 competition dataset. *arXiv preprint arXiv:2105.03494*, 2021.

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[7] Dian Chen, Dequan Wang, Trevor Darrell, and Sayna Ebrahimi. Contrastive test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 295–305, 2022.

[8] Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6172–6180, 2018.

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.

[11] Gintare Karolina Dziugaite and Daniel M Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *arXiv preprint arXiv:1703.11008*, 2017.

[12] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015.

[13] Saurabh Garg, Sivaraman Balakrishnan, Zachary C Lipton, Behnam Neyshabur, and Hanie Sedghi. Leveraging unlabeled data to predict out-of-distribution performance. *arXiv preprint arXiv:2201.04234*, 2022.

[14] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[15] Bogdan Grechuk, Alexander N Gorban, and Ivan Y Tyukin. General stochastic separation theorems with optimal bounds. *Neural Networks*, 138:33–56, 2021.

[16] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[18] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike 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*, pages 8340–8349, 2021.

[19] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2018.

[20] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15262–15271, 2021.

[21] Minyoung Huh, Hossein Mobahi, Richard Zhang, Brian Cheung, Pulkit Agrawal, and Phillip Isola. The low-rank simplicity bias in deep networks. *Transactions on Machine Learning Research*, 2022.

[22] Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. *arXiv preprint arXiv:1912.02178*, 2019.

[23] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR, 2021.

[24] Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations*, 2022.

[25] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[26] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[27] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1406–1415, 2019.

[28] Mohammad Pezeshki, Oumar Kaba, Yoshua Bengio, Aaron C Courville, Doina Precup, and Guillaume Lajoie. Gradient starvation: A learning proclivity in neural networks. *Advances in Neural Information Processing Systems*, 34:1256–1272, 2021.

[29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[30] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pages 5389–5400. PMLR, 2019.

[31] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[32] Xinrui Wang and Jinze Yu. Learning to cartoonize using white-box cartoon representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8090–8099, 2020.

[33] Florian Wenzel, Andrea Dittadi, Peter Gehler, Carl-Johann Simon-Gabriel, Max Horn, Dominik Zietlow, David Kernert, Chris Russell, Thomas Brox, Bernt Schiele, et al. Assaying out-of-distribution generalization in transfer learning. *Advances in Neural Information Processing Systems*, 35:7181–7198, 2022.

[34] Andrew G Wilson and Pavel Izmailov. Bayesian deep learning and a probabilistic perspective of generalization. *Advances in neural information processing systems*, 33:4697–4708, 2020.

[35] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16133–16142, 2023.

[36] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

[37] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. Image bert pre-training with online tokenizer. In *International Conference on Learning Representations*, 2021.

Base Model	Size	Model Link
ResNet50		https://pytorch.org/vision/main/models/generated/torchvision.models.resnet50.html IMAGENET1K_V2 weights
ConvNeXtV2	Large Tiny	https://huggingface.co/facebook/convnextv2-large-22k-224 https://huggingface.co/facebook/convnextv2-tiny-22k-224
CLIP	Large Base	https://huggingface.co/openai/clip-vit-large-patch14 https://huggingface.co/openai/clip-vit-base-patch16
DINOV2	Large Small	https://huggingface.co/facebook/dinov2-large https://huggingface.co/facebook/dinov2-base

Table 7: Links to Base Model Checkpoints used in Experiments

A Experiment Implementation Details

Each image in our experiments was resized to 256x256, center cropped to 224x224, and then normalized using the ImageNet mean and covariance as a means of preprocessing. Though better accuracy could likely be attained using various data augmentation techniques, we chose to not to perform further augmentation in order to more directly measure performance of the base models themselves. Preprocessed images were fed through each base model to get base model specific representations. Table 7 contains the web links to each of the base model checkpoints used in our experiments. The resulting representations of images were then normalized according to the mean and covariance specific to the train data set they belonged to and base model that was used to extract the representation. We trained linear probes using stochastic gradient descent with momentum (momentum parameter set to 0.9) and no weight decay or weight regularization. We found that optimizing for 100 epochs with an initial learning rate of 0.1 and a cosine annealing learning rate scheduler was sufficient for convergence across base models and data sets. All code to run our experiments will be released publicly after publication of this work. Finally, ImageNet-C is not a single data set but a number of corruptions to ImageNet images. For our experiments we used the Gaussian Blur corruption with difficulty of 5, the highest allowed by the data set code.

B Possible Goals for Future Visual Foundation Model Research

In most published work that propose new foundation models, the models are evaluated on downstream task performance, which is the most direct way of measuring their practical utility. However, we feel such empirical, end-task driven pursuits can benefit from both more principled focus into what makes a "good" foundation model, as well as more rigorous investigation into the data and objectives used for pre-train them. In the remainder of this section, we highlight three main empirical findings of this work and use them to highlight possible ways forward in foundation model research.

Finding #1: Some linear probed foundation models achieve high accuracy in domains different than that in which they were trained, but fail in others. Our work highlights that visual foundation models do not represent class structure in such a general way that any conceivable definition of a class as defined in a domain is distinct from potential other classes in that domain. For instance, our results indicate that a linear probe trained using DINOV2 on ImageNet generalizes well to cartoon renderings of the ImageNet test set. On the other hand, probes trained on DomainNet-Real images do not generalize well to DomainNet-Quickdraw images. We argue that this isn't an unreasonable failing of foundation models, as there will always be some limit to their generalizability in practice, but if the foundation models are treated as black boxes, it is unclear what class semantics are captured by the models without testing for each such case. This necessitates the need for further understanding of the practical *limits of generalization* of foundation models.

Generalization of deep neural networks has been a focus from the learning theory community for many years [11, 3, 22, 34, 36]. However, most of these results focus on the setting where models are trained directly for a task. In the case where foundation models are pre-trained on a one task and then adapted for another, there is much less principled understanding of generalization. We feel that a simple, but compelling problem formulation of this form that is amenable to generalization analysis would be a critical starting point for generalization research in foundation models. From such a point, more and more complex settings can be studied and pre-training tasks can be developed that are more grounded in principled understanding.

More practically, better understanding of generalization could be achieved by releasing the pre-training data along with foundation models, so the research community can analyze it in comparison to empirical observations of model performance. In the cases where this is not possible, it would be beneficial to provide information about the scope, intent, and procedure for collecting data as well as curation efforts. What exactly this information entails is open for debate, but the driving motivation should be transparancy that allows for understanding of the limits of models pre-trained on the data. Complicating this is the fact that many of the foundation model we tested and in wide use were pre-trained on data scraped from the web with relatively little definition of a specific scope or efforts to curate the data. This represents a tension between collecting more data to produce more general foundation models and scoping data collecting so the capabilities of foundation models are better understood. We argue that limiting the scope of pre-training data would be beneficial in that it would be more intuitive to reason about the limits of a foundation model's generalizability.

Finding #2: In some domains, foundation models represent classes such that they cannot be fully separated by linear probes. The domains where linear probes could not fully discriminate classes (DomainNet-Quickdraw, DomainNet-Infograh, FMOW, etc.) posed fine-grained classification tasks. This may indicate that foundation models learned on coarse-grained pre-training data do not represent fine-grained classes well. Similar to the discussion on the previous finding, we believe that this phenomenon can be better understood by research focused on more tightly coupling the pre-training procedure and data to downstream application domains of foundation models.

Finding #3: Foundation models with different pre-training objectives and data sources performed inconsistently relative to each other with respect to target domain accuracy. The direct relationship among pre-training objectives, the data used to train foundation models, and generalization of down-stream tasks across domains is not widely known. Future work may benefit from well-argued formal targets on what desirable end state of data representation for pre-training would be. From a targeted end state, objectives, data augmentations, and even desirable characteristics of pre-training data could be developed. For instance, if linear discriminability is a target what inductive biases (regularization, training objectives, architecture designs, etc.) can be imposed during pre-training to achieve it?

C Note on Linear Separability in High Dimensions

It is a much studied and observed phenomenon (see [15] for one such treatment of this phenomenon) that even random partitions of data in high-dimensions are linearly separable. Given that the dimensionality of the representations learned by the models in our evaluation range from hundreds to thousands, it may be expected that any target data we evaluated would be linearly separable. However, as our results show, our target training procedure did not result in 100% train accuracy on all data sets. We believe this shows that the *intrinsic* dimensionality (the number of dimensions needed to minimally represent data) is lower than the full dimensionality output by these models for some data sets. This aligns with prior work on the effective rank of representations learned by deep networks [1, 2, 28, 21]. While this can have benefits for classification in-domain, learned representations with lower intrinsic dimensionality may not be expressive enough for linear models to discriminate classes in out-of-domain classification tasks. We believe our results show evidence of this.

D Acknowledgements

This material is based upon work funded and supported by the Department of Defense under Contract No. FA8702-15-D-0002 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center.

DM23-2218