Calibrated Ensembles: A Simple Way to Mitigate ID-OOD Accuracy Tradeoffs

Anonymous Author(s)
Affiliation
Address
email

Abstract

We often see undesirable tradeoffs in robust machine learning where out-of-distribution (OOD) accuracy is at odds with in-distribution (ID) accuracy. A "robust" classifier obtained via specialized techniques like removing spurious features has better OOD but worse ID accuracy compared to a "standard" classifier trained via vanilla ERM. On six distribution shift datasets, we find that simply ensembling a standard and a robust model is a strong baseline—we match the ID accuracy of a standard model with only a small drop in OOD accuracy compared to the robust model. However, calibrating these models in-distribution surprisingly improves the OOD accuracy of the ensemble and eliminates the tradeoff and we achieve the best of both ID and OOD accuracy over the original models.

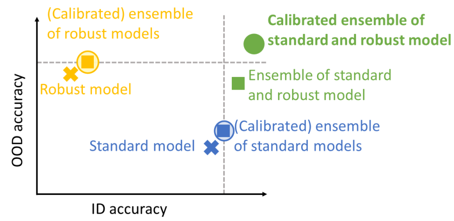
1 Introduction

Machine learning models typically suffer large drops in accuracy in the presence of distribution shift where the test distribution is different from the training distribution. As ML systems are widely deployed, it is important for models to have good "out-of-distribution" (OOD) accuracy. There has been a lot of research interest in tackling this robustness problem under various settings such as robustness to spurious correlations (1; 2; 3), domain generalization (4; 5), robustness to demographic shifts (6; 7) among others. Almost universally across these different settings, an unfortunate tradeoff arises. Robustness interventions typically improve the OOD accuracy but simultaneously cause a drop in the "in-distribution" (ID) accuracy on new test points from the original distribution.

This tradeoff is a major hurdle in using the multitude of proposed robustness interventions. In practice, most inputs are likely to be ID, so it is unsatisfactory to use a "robust" model that has high OOD performance but performs less accurately on these majority ID points. On the other hand, "standard models" (trained without robustness interventions) fail catastrophically in the presence of even small shifts, and it can be highly dangerous to use a standard model even if OOD points are rare. In this work, we ask *is there a general strategy by which we can achieve high accuracy both in-distribution and out-of-distribution and mitigate tradeoffs arising in robustness?*

We consider four benchmark datasets (DomainNet, CIFAR $\rightarrow$ STL, ImageNet $\rightarrow$ ImageNet-R, and BREEDS-Entity-30) and two real world satellite remote sensing datasets (Landcover and Cropland), that have been used in prior work on robustness. Our work spans different types of robustness interventions (projecting out spurious correlations, zero-shot language prompting, freezing pretrained features), data modalities (image and time series data), and model architectures (vision transformers, deep convolutional networks, time series convolution). Averaged across these datasets, robustness interventions increase OOD accuracy from 63% to 74%, but decrease ID accuracy from 88% to 85%.

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.



(a) Schematic of performance of different methods

	ID accuracy	OOD accuracy
Standard Model	87.9%	62.7%
Ensemble (Standard + Standard)	88.6%	63.7%
Robust Model	84.7%	74.1%
Ensemble (Robust + Robust)	84.9%	74.3%
Ensemble (Standard + Robust)	89.0%	73.2%
Calibrated ensemble (Standard + Robust)	**89.3%**	**74.6%**

(b) Performance of different methods averaged over 6 real distribution shifts

Figure 1: In many settings, we have a 'standard' model that performs better in-distribution, and a 'robust' model that performs better out-of-distribution. Simply ensembling these two models (e.g., by adding their probabilities), gets better ID accuracy than the standard and robust models, and closes most of the OOD gap. Calibrating the models in-distribution (no access to OOD data) before ensembling them leads to further improvements. Note that ensembling two standard or two robust models does not close the gap and only leads to small improvements.

We first explore the natural strategy of ensembling the standard and robust models—concretely, we add the probabilities of each model to obtain a prediction with the hope that when the two models conflict, the more confident model (with larger probability) dictates the final prediction. We find that this surprisingly simple baseline already perfoms quite well—on average across all our datasets, this closes 80% of the gap between the OOD of standard models, while outperforming both models ID. However, vanilla ensembling still leaves a gap as it underperforms the robust model OOD.

We find that simply calibrating both models ID (adjusting their predicted confidence to match their accuracy, on *in-distribution* data) before ensembling them closes this gap. *Calibrated ensembles get an average accuracy of 89.3% ID and 74.6% OOD, and outperform both the standard and robust model, ID and OOD*. The other method in the literature to alleviate robustness induced tradeoffs is self-training that uses large amount of unlabeled data (8; 9; 10). On the two remote-sensing datasets with additional unlabeled data, we find that calibrated ensembles match self-training on these datasets despite its simplicity and without requiring any unlabeled data.

While our method is intuitive, it is intriguing that it works so well because ensembling seems to rely on good uncertainty estimates while it is common wisdom that uncertainty estimates of deep networks are unreliable OOD even after calibrating in-distribution (11). Indeed, on the six datasets we test on, the models fare poorly on standard uncertainty metrics OOD, even after calibration. The expected calibration error of the standard model across all datasets is 12%. Even the relative confidences of the models can be incorrect—on the remote sensing dataset (Landcover), the standard model is on average 6% more confident in its OOD predictions than the robust model, even though the standard model is less accurate OOD. But at the granularity of individual points, calibrated ensembles are able to combine predictions effectively and achieve high ID and OOD accuracy.

2 Setup

Consider a K-class classification task, where the goal is to predict targets $y \in [K]$ from inputs $x \in R^d$.

Models: A model $f : \mathbb{R}^d \to \mathbb{R}^k$ takes an input $x \in \mathbb{R}^d$ and outputs $f(x) \in R^k$ where $f(x)_i$ denotes the model's confidence that the output is $y = i$. The model predicts the label $\widehat{y} = \mathrm{argmax}_k f(x)_k$. The confidences can be converted into probabilities using a softmax function:

Data: Let P_{id} and P_{ood} denote the underlying distribution of (x, y) pairs in-distribution and out-of-distribution, respectively. We have a validation set $\{(x_i^{\text{val}}, y_i^{\text{val}})\}_{i=1}^{n_{\text{val}}} \sim P_{\text{id}}$ used for early stopping and calibration, a held-out in-distribution test set $\{(x_i^{\text{test}}, y_i^{\text{test}})\}_{i=1}^{n_{\text{test}}} \sim P_{\text{id}}$, and a held-out out-of-distribution test set $\{(x_i^{\text{ood}}, y_i^{\text{ood}})\}_{i=1}^{n_{\text{ood}}} \sim P_{\text{ood}}$. All methods can use the ID validation set for

Table 1: *In-distribution (ID)* accuracies for the standard model, robust model, and calibrated ensembling, across six datasets. Calibrated ensembling matches or outperforms the better model in all cases, and on average outperforms both the standard and robust models.

	Ent30	DomNet	CIFAR10	Land	Crop	ImNet
Standard	**93.6 (0.2)**	83.9 (1.0)	**97.4 (0.1)**	**76.9 (0.3)**	95.3 (0.0)	80.5 (-)
Robust	90.7 (0.2)	89.2 (0.1)	92.0 (0.0)	72.7 (0.2)	95.1 (0.1)	68.4 (-)
Cal ensemble	**93.7 (0.1)**	**91.2 (0.6)**	**97.2 (0.1)**	**77.1 (0.2)**	**95.6 (0.1)**	**81.1 (-)**

tuning hyperparameters, early stopping, and calibration. The ID and OOD test set are only used for evaluation. We evaluate a model f on the average accuracy on the ID and OOD test sets.

3 Methods and Datasets

Calibrated ensembles. Given two models f_1 and f_2, calibrated ensembles first calibrate each model using temperature scaling (12) with the cross-entropy loss l on the *in-distribution* validation data:

$$T_j = \underset{T}{\operatorname{argmin}} \frac{1}{n_{\mathsf{val}}} \sum_{i=1}^{n_{\mathsf{val}}} l\left(\frac{f_j(x_i^{\mathsf{val}})}{T}, y_i^{\mathsf{val}}\right) \qquad \text{for } j \in \{1,2\} \tag{3.1}$$

We then ensemble the two models by adding up the probabilities that they predict (13):

$$\widehat{p} = \frac{1}{2}\left(\text{softmax}\left(\frac{f_1(x)}{T_1}\right) + \text{softmax}\left(\frac{f_2(x)}{T_2}\right)\right) \tag{3.2}$$

Other ensembles. As a baseline, we consider a *tuned ensemble*: outputting a weighted average of the standard and robust model's probabilities, where the weight $\alpha \in [0,1]$ is tuned to maximize accuracy on the in-distribution set.

$$\widehat{p} = \alpha \text{softmax}(f_{\mathsf{std}}(x)) + (1-\alpha)\text{softmax}(f_{\mathsf{rob}}(x))) \tag{3.3}$$

In *vanilla ensembling* the weight α is set to $1/2$. We also considered other ways of combing the model outputs (adding logits vs probabilities) and found that the results were similar.

Datasets We run experiments spanning three different types of robustness interventions: projecting out spurious metadata, zero-shot language prompting in CLIP, and freezing pretrained features. These experiments span multiple model architectures (vision transformers, deep convolutional networks, time series convolution) and data modalities (image and time series data), and include two real world remote sensing datasets used in prior work studying ID-OOD tradeoffs (9). See Appendix A for more details.

4 Experiments

Strong ID and OOD accuracy: Calibrating and then ensembling a standard and a robust model, gets the best of both worlds, typically outperforming the standard and robust model both ID (Table 1) and OOD (Table 2). Averaged across the datasets, calibrated ensembles get 89.3% ID (vs 87.9% for the standard model and 84.7% for the robust model) and 74.6% OOD (vs 74.1% for the robust model and 62.7% for the standard model).

Competitive with self-training: The remote sensing datasets have lots of unlabeled data so prior work (9) uses self-training on these datasets to mitigate the ID-OOD accuracy tradeoff. Table 3 shows that calibrated-ensembles match or outperform self-training on both datasets, both ID and OOD. We believe this is an interesting result because calibrated ensembling is a simple method and does not need additional unlabeled data.

Calibration is important: *We find that a strong baseline of tuning the ensemble weights on ID data has lower accuracy than calibrated ensembles OOD* (Table 4; calibrated ensembles average: 74.6%, tuned

Table 2: *Out-of-distribution (OOD)* accuracies across six datasets. Calibrated ensembling matches or outperforms the better model in 4/6 cases, and on average outperforms both the standard and robust models. For the remaining two datasets, DomainNet and ImageNet-R, calibrated ensembles close 96% and 93% of the gap between the standard and robust model.

	Ent30	DomNet	STL	Land	Crop	ImNet-R
Standard	60.7 (0.1)	55.3 (0.4)	82.4 (0.3)	55.7 (1.1)	**85.6 (5.8)**	36.2 (-)
Robust	63.2 (1.1)	**87.2 (0.1)**	85.1 (0.2)	**60.4 (1.1)**	89.8 (0.4)	**59.1 (-)**
Cal ensemble	**64.8 (0.5)**	85.9 (0.2)	**87.3 (0.2)**	**60.8 (0.8)**	**91.3 (0.8)**	57.4 (-)

Table 3: Calibrated ensembles are competitive with self-training (9), which requires unlabeled data.

	Cropland		Landcover	
	ID Acc	OOD Acc	ID Acc	OOD Acc
Standard model	95.3 (0.0)	85.6	76.9	55.7
Robust model	95.1 (0.1)	89.8	72.7	60.4
Self-training	95.3 (0.2)	**90.6 (0.6)**	**77.0 (0.4)**	**61.0 (0.7)**
Cal ensembling	**95.6 (0.1)**	**91.0 (0.8)**	**77.1 (0.2)**	**60.8 (0.8)**

Table 4: OOD accuracies: calibrated ensembles outperform vanilla ensembles and even tuned ensembles where the combination weights are tuned to maximize in-distribution accuracy. Averaged across the datasets, calibrated ensembles get an OOD accuracy of 74.6%, while tuned ensembles get an accuracy of 71.3%. The in-distribution accuracies of the methods are very close (within 0.2% of each other).

	Ent30	DomNet	STL	Land	Crop	ImNet-R
Vanilla	**64.6 (0.4)**	78.7 (1.3)	**87.2 (0.2)**	59.5 (1.0)	**90.9 (0.2)**	**58.0 (-)**
Tuned ID	**64.6 (0.6)**	**86.3 (0.6)**	85.7 (0.9)	58.7 (1.2)	**87.3 (5.7)**	45.4 (-)
Calibrated ID	**64.8 (0.5)**	**85.9 (0.2)**	**87.3 (0.2)**	**60.8 (0.8)**	**91.3 (0.8)**	**57.4** (-)

ensembles: 71.3%) and calibrated ensembles only have a 0.2% drop in ID accuracy relative to tuned ensembles. Naturally, we expect the tuned ensemble to do the best ID since its weights are tailored for ID—what is surprising is that the calibrated ensembles do so much better OOD without using any OOD data either. Calibrated ensembles outperform vanilla ensembles both ID and OOD as well.

Standard ensembles do not mitigate tradeoffs: As we show in Figure 1, simply ensembling two standard models or two robust models (even with calibration) does not mitigate ID-accuracy tradeoffs.

5 Related Works and Discussion

Calibration. Calibration (in-distribution) has been widely studied (14; 12; 15; 16; 17; 18). Ovadia et al. (11) show that if we calibrate a model ID, it still has poor uncertainties OOD.

Ensembling. Typically ensemble members are trained on the same data with a different random seed (13) or augmentation (19)—in these settings prior work has shown that calibration does not help (20; 11). Indeed, calibration has minimal effect when we ensemble two standard, or two robust models, but leads to clear improvements when we combine two very different models (standard and robust).

Mitigating ID-OOD tradeoffs. Prior work self-trains on large amounts of unlabeled data to mitigate ID-OOD tradeoffs (8; 9; 10). In concurrent work, Wortsman et al. (21) show on ImageNet and variants (e.g. ImageNet-R) that there *exists* a way to ensemble a CLIP zero-shot and fine-tuned model to get good ID and OOD accuracy—but this might require OOD data which is not available. In fact, we show that the natural way to *learn* how to weight ensemble members—selecting the weights to optimize in-distribution accuracy—does not mitigate the ID-OOD gap, but calibrated ensembles do.

References

[1] Christina Heinze-Deml and Nicolai Meinshausen. Conditional variance penalties and domain shift robustness. *arXiv preprint arXiv:1710.11469*, 2017.

[2] Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations (ICLR)*, 2020.

[3] Yining Chen, Colin Wei, Ananya Kumar, and Tengyu Ma. Self-training avoids using spurious features under domain shift. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[4] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[5] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European Conference on Computer Vision (ECCV)*, 2016.

[6] Tatsunori B. Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning (ICML)*, 2018.

[7] John Duchi, Tatsunori Hashimoto, and Hongseok Namkoong. Distributionally robust losses against mixture covariate shifts. `https://cs.stanford.edu/~thashim/assets/publications/condrisk.pdf`, 2019.

[8] Aditi Raghunathan, Sang Michael Xie, Fanny Yang, John C. Duchi, and Percy Liang. Understanding and mitigating the tradeoff between robustness and accuracy. In *International Conference on Machine Learning (ICML)*, 2020.

[9] Sang Michael Xie, Ananya Kumar, Robert Jones, Fereshte Khani, Tengyu Ma, and Percy Liang. In-N-out: Pre-training and self-training using auxiliary information for out-of-distribution robustness. In *International Conference on Learning Representations (ICLR)*, 2021.

[10] Fereshte Khani and Percy Liang. Removing spurious features can hurt accuracy and affect groups disproportionately. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2021.

[11] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D Sculley, Sebastian Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[12] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning (ICML)*, pages 1321–1330, 2017.

[13] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[14] Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. Binary classifier calibration: Non-parametric approach. *arXiv*, 2014.

[15] Ananya Kumar, Percy Liang, and Tengyu Ma. Verified uncertainty calibration. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[16] Allan H Murphy. A new vector partition of the probability score. *Journal of Applied Meteorology*, 12(4):595–600, 1973.

[17] Morris H. DeGroot and Stephen E. Fienberg. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 32:12–22, 1983.

[18] Tilmann Gneiting and Adrian E. Raftery. Weather forecasting with ensemble methods. *Science*, 310, 2005.

[19] Asa Cooper Stickland and Iain Murray. Diverse ensembles improve calibration. *arXiv*, 2020.

[20] Xixin Wu and M. Gales. Should ensemble members be calibrated? *arXiv*, 2021.

[21] Mitchell Wortsman, Gabriel Ilharco, Mike Li, Jong Wook Kim, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models. *arXiv preprint arXiv:2109.01903*, 2021.

[22] Sherrie Wang, William Chen, Sang Michael Xie, George Azzari, and David B. Lobell. Weakly supervised deep learning for segmentation of remote sensing imagery. *Remote Sensing*, 12, 2020.

[23] E. Vermote. MOD09A1 MODIS/terra surface reflectance 8-day L3 global 500m SIN grid V006. `https://doi.org/10.5067/MODIS/MOD09A1.006`, 2015.

[24] Pall Oskar Gislason, Jon Atli Benediktsson, and Johannes R. Sveinsson. Random forests for land cover classification. *Pattern Recognition Letters*, 27(4):294–300, 2006.

[25] Marc Rußwurm, Sherrie Wang, Marco Korner, and David Lobell. Meta-learning for few-shot land cover classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 200–201, 2020.

[26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, volume 139, pages 8748–8763, 2021.

[27] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. *arXiv preprint arXiv:2006.16241*, 2020.

[28] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[29] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Association for Computational Linguistics (ACL)*, 2021.

[30] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. *arXiv*, 2019.

[31] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *International Conference on Computer Vision (ICCV)*, 2019.

[32] Shuhan Tan, Xingchao Peng, and Kate Saenko. Class-imbalanced domain adaptation: An empirical odyssey. *arXiv preprint arXiv:1910.10320*, 2020.

[33] Geoff French, Michal Mackiewicz, and Mark Fisher. Self-ensembling for visual domain adaptation. In *International Conference on Learning Representations*, 2018.

[34] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

[35] Adam Coates, Andrew Ng, and Honlak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15, pages 215–223, 2011.

[36] Xinlei Chen, Haoqi Fan, Ross B. Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv*, 2020.

[37] Shibani Santurkar, Dimitris Tsipras, and Aleksander Madry. Breeds: Benchmarks for subpopulation shift. *arXiv*, 2020.

A Details on datasets

A.1 Spurious metadata

We run experiments on two remote sensing datasets used in prior work studying ID-OOD tradeoffs (9). These datasets consist of a core input x (image data or time series data) and metadata z (e.g., location, meteorological climate data). The metadata is spuriously correlated with the target—using the metadata to predict labels improves accuracy in-distribution (ID), but hurts accuracy out-of-distribution. Xie et al. (9) consider a standard model that takes in both the core inputs and metadata to predict the target, and a robust model that only takes in the core inputs and does some additional pretraining. They call these the 'aux-in' and 'aux-out' models respectively.

Cropland. The goal is to predict whether a satellite image is of a cropland or not. The core input x is an RGB satellite image, and the metadata z consists of location coordinates and vegetation bands. The original dataset is from Wang et al. (22), and we use U-net model checkpoints from Xie et al. (9).

Landcover. The goal is to predict the land type from satellite data at a given location. Here, the core input x is a time series measured by NASA's MODIS satellite (23), and z is climate data (e.g., temperature) at that location. The dataset is from Gislason et al. (24); Rußwurm et al. (25). We use model checkpoints from Xie et al. (9) where they use 1D convolutions for time series data.

A.2 Zero-shot language prompting

Radford et al. (26) (CLIP) pretrain a model on a large multi-modal language and vision dataset. The model can then predict the label of an image by comparing the image embedding, with the language embedding for prompts such as 'photo of an apple' or 'photo of a banana'. They show that this zero-shot language prompting approach can be much more accurate out-of-distribution than the traditional method of fine-tuning the entire model.

ImageNet $\rightarrow$ ImageNet-R. We use a CLIP vision transformer, specifically a ViT-B/16, which is the best publicly available model. The robust model uses language prompts to make zero-shot predictions on ImageNet-Renditions (27), a dataset containing cartoon, graffiti, video game, etc, renditions of ImageNet classes. The standard model initializes with weights from the CLIP model, and fine-tunes on ImageNet (28) training data for 10 epochs, before making predictions on ImageNet-R. We note that the robust model gets 10% lower accuracy ID (on ImageNet validation examples), but gets 20% higher accuracy OOD (on ImageNet-R test examples)

A.3 Freezing pretrained features

When adapting a pretrained model to an ID dataset, typically all the model parameters are fine-tuned. Recent work looks at 'lightweight' fine-tuning, where only parts of the model are adapted—this can often do better OOD even though the ID performance is worse (29; 30). We consider three distribution shift datasets where the standard model starts from a pretrained initialization and fine-tunes all parameters on an ID dataset, and the robust model only learns the top linear 'head' layer.

DomainNet. A standard domain adaptation dataset (31). Here, our ID dataset contains 'sketch' images (e.g., drawings of apples, elephants, etc), and the OOD dataset contains 'real' photos of the same categories. We use the version of the dataset from Tan et al. (32). We start from a CLIP pretrained ResNet50 and either fine-tune (to get a standard model) or train the head layer (to get a robust model).

CIFAR-10 $\to$ STL. Another standard domain adaptation dataset (33), where the ID is CIFAR-10 (34), and the OOD is STL (35). We start from a ResNet50 pretrained on unlabeled ImageNet examples using MoCo-v2 (36) and either fine-tune (to get a standard model) or train the head layer (to get a robust model).

Living-17. Part of the BREEDS benchmark (37), here the goal is to classify an image as one of 17 animal categories such as 'bear'—the ID dataset contains images of black bears and sloth bears and the OOD dataset has images of brown bears and polar bears. We start from a ResNet50 pretrained on unlabeled ImageNet examples using MoCo-v2 (36) and either fine-tune (to get a standard model) or train the head layer (to get a robust model).