Calibrated ensembles can mitigate accuracy tradeoffs under distribution shift

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 such as removing spurious features often has better OOD but worse ID accuracy compared to a standard classifier trained via vanilla ERM. In this paper, we find that a simple approach of ensembling the standard and robust models, after calibrating on only ID data, outperforms prior state-of-the-art both ID and OOD. On ten natural distribution shift datasets, ID-calibrated ensembles get the best of both worlds: strong ID accuracy of the standard model *and* OOD accuracy of the robust model. We analyze this method in stylized settings, and identify two important conditions for ensembles to perform well on both ID and OOD: (1) standard and robust models should be calibrated (on ID data, because OOD data is unavailable), (2) OOD has no anticorrelated spurious features.

1 INTRODUCTION

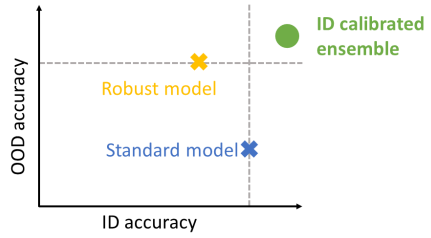
Machine learning models suffer large drops in accuracy in the presence of distribution shift where the test distribution is different from the training distribution. As ML systems are widely deployed, it is important to train models that achieve good accuracy on unforeseen, out-of-distribution (OOD) examples. For example, models trained on medical data from a few hospitals should work well when deployed broadly [Zech et al., 2018, AlBadawy et al., 2018]. Similarly, when predicting poverty from satellite imagery, models trained on data from a few countries should work well on all countries, particularly those where labels are scarce due to resource constraints [Jean et al., 2016]. There has been a lot of research interest in tackling this robustness problem under various settings such as robustness to spurious correlations [Heinze-Deml and Meinshausen, 2017, Sagawa et al., 2020a], domain generalization [Arjovsky et al., 2019, Sun and Saenko, 2016], demographic shifts [Hashimoto et al., 2018, Duchi et al., 2019] among others.

Almost universally across these different settings, an unfortunate tradeoff arises: robustness interventions, such as removing spurious features, typically improve the OOD accuracy but 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) can fail in the presence of even small shifts, and it can be dangerous to use a standard model even if OOD points are rare. In this work, we ask: *is there a general strategy to harness the strengths of both the standard and robust model to achieve high accuracy both ID and OOD, without using OOD data?*

We find that a simple approach of first calibrating the standard and robust models on only ID data and then ensembling them, outperforms prior state-of-the-art both ID and OOD. As illustrated in Figure 1, across 10 natural distribution shift datasets (e.g. geographical shift, style shift, subpopulation shift), ID-calibrated ensembles get the *best of both worlds*: strong ID accuracy of the standard model and robust accuracy of the OOD models. Averaged across these datasets, ID-calibrated ensembles achieve an ID accuracy of 91.1% (vs. 89.5% for the standard model and 87.5% for the robust model) and OOD accuracy of 75.3% (vs. 65.6% for the standard model and 73.2% for the robust model).

We then analyze when and why ID-calibrated ensembles can get the best of the standard and robust models, under a simplifying assumption that these models provide different and independent signals for the label. If the standard and

Submitted to the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022). **To be used for reviewing only**.



(a) Plot of performance of different methods

	ID accuracy	OOD accuracy
Standard Model	89.5%	65.6%
Robust Model	87.5%	73.2%
ID calibrated ensemble (Std. + Rob.)	**91.1%**	**75.3%**

(b) Performance of different methods averaged over 10 natural 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. Across 10 natural distribution shifts, ID-calibrated ensembles get the best of both worlds: the strong ID accuracy of the standard model and OOD accuracy of the robust model. We analyze its strengths and limitations in Section 4—as predicted by our analysis, ID calibrated ensembles do not perform as well on adversarially synthesized shifts with "anticorrelated" spurious features. We show full experimental results and ablations in Section 6.

robust models are *calibrated* ID, the ensembling strategy for the best ID performance is to simply add the predictions of the two models (Proposition 4.1). By the same idea, if the standard and robust models were also calibrated OOD, ensembling would achieve the optimal OOD accuracy. However, since we only have ID training data, models can only be calibrated ID and ID calibration is not sufficient for OOD calibration [Ovadia et al., 2019].

When can calibrated ensembles provide benefits even without OOD calibration? In many natural distribution shifts, standard models pick up on predictive signals in the training data that are absent or suppressed under distribution shift—in these cases, we show that ID-calibrated ensembles obtain the best of both the standard and robust models OOD. However, when spurious features become anticorrelated OOD (as is common when the distribution shift is adversarially synthesized), we show that the ensemble performance is in between the standard and robust models. We empirically validate this on three adversarially synthesized shifts [Sagawa et al., 2020a, Jones et al., 2021] where the spurious signals are anticorrelated OOD.

Finally, we compare to a number of other alternate ensembling strategies (for example, tuning the weights of the ensemble on ID validation data) and find that they do not work as well as ID-calibrated ensembles.

To summarize, our main contributions are:

1. We revisit the classic idea of ensembling and propose a simple, general, and effective method to mitigate ID-OOD accuracy tradeoffs. This method outperforms prior approaches based on self-training, despite not using any additional unlabeled data.

2. We prove that ensembles of calibrated models are optimal when the models provide independent signals about the label. However, models can only be calibrated ID from which we have training data, and ID calibration does not imply OOD calibration. In simple and stylized settings, we identify conditions under which ID-calibrated ensembles achieve the best of standard and robust models in terms of OOD performance. We validate these insights experimentally and find that ID-calibrated ensembles eliminate tradeoffs under a variety of natural distribution shifts, but can fail when there are adversarially synthesized shifts.

2 SETUP

Consider a K-class classification task, where the goal is to predict labels $y \in [K]$ corresponding to inputs $x \in \mathcal{X}$.

Models. A model $f : \mathcal{X} \to \mathbb{R}^K$ takes an input $x \in \mathcal{X}$ and outputs a score $f(x) \in \mathbb{R}^K$ where $f(x)_i$ can be interpreted as the model's "confidence" that the label y is i. The model outputs the label $\text{pred}(f(x)) = \arg\max_i f(x)_i$. The confidence scores can be normalized to sum to 1 (and interpreted as probabilities) using the softmax function, $\text{softmax}(f(x))_i = \frac{\exp(f(x)_i)}{\sum_{j=1}^K \exp(f(x)_j)}$ for $i \in [K]$.

Distributions and error. Let P_{id} and P_{ood} denote the underlying distribution of (x, y) pairs in-distribution (ID) and out-of-distribution (OOD), respectively. We evaluate a model f on the fraction of times it makes a wrong prediction on P_{id} and P_{ood}: $\text{Err}_{\text{id}}(f) = \mathbb{E}_{x,y\sim P_{\text{id}}}[\text{pred}(f(x)) \neq y]$ and $\text{Err}_{\text{ood}}(f) = \mathbb{E}_{x,y\sim P_{\text{ood}}}[\text{pred}(f(x)) \neq y]$.

Standard and robust models. A standard model f_{std} is trained via empirical risk minimization where we minimize some loss on ID training data. f_{std} often relies on spurious correlations such as image background or occurence of certain words that are not necessarily predictive OOD. In order to improve OOD performance, a robust model f_{rob} is trained via a modified training procedure (robustness interventions) to discourage models from relying on ID-specific spurious features. Formally, we have the following relationship between f_{std} and f_{rob}.

$$\text{Err}_{\text{id}}(f_{\text{std}}) \leq \text{Err}_{\text{id}}(f_{\text{rob}}); \quad \text{Err}_{\text{ood}}(f_{\text{rob}}) \leq \text{Err}_{\text{ood}}(f_{\text{std}}). \tag{2.1}$$

Algorithm 1 ID-calibrated ensembles

Require: in-distribution validation data $\{(x_i^{\mathsf{val}}, y_i^{\mathsf{val}})\}_{i=1}^{n_{\mathsf{val}}} \sim P_{\mathsf{id}}$, standard and robust models $f_{\mathsf{std}}, f_{\mathsf{rob}} : \mathcal{X} \to \mathbb{R}^K$

1: $\overline{f}_{\mathsf{std}}$ = Calibrate f_{std} on in-distribution (ID) data
2: $\overline{f}_{\mathsf{rob}}$ = Calibrate f_{rob} on in-distribution (ID) data
3: Return $f_{\mathsf{ens}}(x) = \frac{1}{2}(\text{softmax}(\overline{f}_{\mathsf{std}}(x)) + \text{softmax}(\overline{f}_{\mathsf{rob}}(x)))$

The precise robustness intervention depends on the task—in Section 4 we model the relationship between f_{std} and f_{rob} in a stylized setting amenable for analysis, and in Section 5 we describe what f_{std} and f_{rob} are in our real datasets.

Best of both worlds. Our goal is to get the best of both worlds—a classifier f_{ens} that achieves the strong ID accuracy of the standard model, and OOD accuracy of the robust model:

$$\text{Err}_{\mathsf{id}}(f_{\mathsf{ens}}) \leq \text{Err}_{\mathsf{id}}(f_{\mathsf{std}}); \;\; \text{Err}_{\mathsf{ood}}(f_{\mathsf{ens}}) \leq \text{Err}_{\mathsf{ood}}(f_{\mathsf{rob}}). \tag{2.2}$$

ID validation data. To get the best of both worlds, we only allow access to ID validation data, $\{(x_i^{\mathsf{val}}, y_i^{\mathsf{val}})\}_{i=1}^{n_{\mathsf{val}}} \sim P_{\mathsf{id}}$, for tuning hyperparameters. Following Xie et al. [2021], Koh et al. [2021], Gulrajani and Lopez-Paz [2020] we do *not* use any OOD validation data.

3 METHODS

Proposed method: ID-calibrated ensembles. Given a standard model f_{std} and robust model f_{rob}, we first calibrate each model on the *in-distribution* validation data, and then add up their predictions (Algorithm 1). In our experiments, we calibrate using temperature scaling [Guo et al., 2017] with the cross-entropy loss ℓ:

$$T_{\mathsf{std}} = \arg\min_T \frac{1}{n_{\mathsf{val}}} \sum_{i=1}^{n_{\mathsf{val}}} \ell\Big(\frac{\text{softmax}(f_{\mathsf{std}}(x_i^{\mathsf{val}}))}{T}, y_i^{\mathsf{val}}\Big) \tag{3.1}$$

$$T_{\mathsf{rob}} = \arg\min_T \frac{1}{n_{\mathsf{val}}} \sum_{i=1}^{n_{\mathsf{val}}} \ell\Big(\frac{\text{softmax}(f_{\mathsf{rob}}(x_i^{\mathsf{val}}))}{T}, y_i^{\mathsf{val}}\Big) \tag{3.2}$$

We then ensemble the two models by adding up the probabilities that they predict [Lakshminarayanan et al., 2017].

$$f_{\mathsf{ens}}(x) = \frac{1}{2}\Big(\text{softmax}\Big(\frac{f_{\mathsf{std}}(x)}{T_{\mathsf{std}}}\Big) + \text{softmax}\Big(\frac{f_{\mathsf{rob}}(x)}{T_{\mathsf{rob}}}\Big)\Big), \tag{3.3}$$

where the predicted label is $\text{pred}(f_{\mathsf{ens}}(x)) = \arg\max_y f_{\mathsf{ens}}(x)_y$.

Ablations. In Section 6 we ablate each component of the method, for example the calibration step, way of combining the models, and we compare to (calibrated) ensembles of two standard models, or of two robust models.

4 INTUITIONS AND ANALYSIS

In this section, we build basic intuitions for when and why ID-calibrated ensembles can get the best of both worlds (good ID accuracy of f_{std} and OOD accuracy of f_{rob}), even without using any OOD data. We first define a stylized setting, and then analyze the ID performance in Section 4.1 and OOD performance in Section 4.2.

Diverse features. An intuitive and illustrative conceptual setting is the following: we assume inputs have some robust features (that are predictive both ID and OOD) and some spurious features (that are only predictive ID). f_{std} relies on the spurious features while f_{rob} relies on the robust features, both of which provide independent signals on the label.

Assumption 4.1. *We assume that f_{rob} and f_{std} have diverse features with respect to P_{id} and P_{ood}, that is,*

$$f_{\mathsf{rob}}(x) \perp f_{\mathsf{std}}(x) \mid y \quad \textit{when } (x, y) \sim P \textit{ for } P \in \{P_{\mathsf{id}}, P_{\mathsf{ood}}\} \tag{4.1}$$

Connection with prior assumptions. The diverse features assumption is weaker than the assumptions in prior conceptual models of distribution shifts Chen et al. [2020], Sagawa et al. [2020b], Nagarajan et al. [2020] where robust and spurious features are disjoint parts of the input, each generated independently based on the label. In our setting, the features can be complicated functions of the inputs.

Ensemble. The ensemble f_{ens} simply adds up the predictions of the standard model f_{std} and robust model f_{rob}. This is slightly different from Section 3, but is more amenable to analysis.

$$f_{\mathsf{ens}}(x) = f_{\mathsf{std}}(x) + f_{\mathsf{rob}}(x) \tag{4.2}$$

Class-balanced. For simplicity of exposition, we assume the class-balanced setting where every label $P(Y = y)$ is equally likely. Formally, we say P is class-balanced if $P(Y = y) = 1/K$ for all $y \in [K]$. We analyze the general setting in Appendix A.

4.1 ID PERFORMANCE OF ENSEMBLES

In this section, we show that if f_{std} and f_{rob} are *calibrated* with respect to P_{id}, then the ensemble f_{ens} is the best way to combine their predictions. Since we have access to validation data from P_{id}, the first step of our method (Section 3) is to calibrate f_{std} and f_{rob} ID. We conclude the section by giving intuition for why this calibration step can be particularly important for deep neural networks.

Intuitively, calibration means that the probability that a model outputs for an event reflects the true frequency of that event: if a model says 1,000 patients have the flu with probability 0.1, approximately 100 of them should indeed have the flu. Formally, we look at joint calibration [Murphy,

1973, Brocker, 2009] where a model f is calibrated with respect to a distribution P if for all $x \in \mathcal{X}, y \in [K]$:

$$P(y \mid f(x)) = \text{softmax}(f(x))_y \tag{4.3}$$

The following proposition says that if f_{std} and f_{rob} are calibrated on P_{id}, then f_{ens} has lower error on P_{id} than any other way of combining the two models—this also implies that f_{ens} gets higher accuracy than f_{std} and f_{rob}.

Proposition 4.1. *Suppose that f_{std} and f_{rob} are calibrated with respect to P_{id}, and that P_{id} is class-balanced. Let $h : \mathbb{R}^K \times \mathbb{R}^K \to \mathbb{R}^K$ be an arbitrary function that combines the standard and robust model's predictions, and let f_h be the resulting classifier: $f_h(x) = h(f_{\text{std}}(x), f_{\text{rob}}(x))$. The ensemble is better than any such combination classifier f_h: $Err_{\text{id}}(f_{\text{ens}}) \leq Err_{\text{id}}(f_h)$.*

The proof of Proposition 4.1 is in Appendix A. Intuitively, since $f_{\text{rob}}(x) \perp f_{\text{std}}(x) \mid y$, the Bayes optimal predictor is proportional to multiplying their predicted probabilities, which is equal to adding logits (logits are in log space). Proposition 4.1 has an important condition: the two models must be calibrated. In practice, deep learning models are miscalibrated [Guo et al., 2017], so our first step (Section 3) is to calibrate the models ID. We explain why the ID calibration step is important for deep neural networks.

Why neural networks are miscalibrated. Deep neural networks are typically large enough to memorize the training dataset, and are encouraged to magnify their weights (and hence their confidence) to decrease the training loss [Mukhoti et al., 2020, Bai et al., 2021]. The extent of this miscalibration and overconfidence depends on the training procedure [Hendrycks et al., 2019, Desai and Durrett, 2020]. In our case f_{std} and f_{rob} are trained in different ways and have different calibration (Appendix B.3).

Why this miscalibration can hurt ensembling. Concretely, consider two models f'_{std} and f'_{rob} which are calibrated on P_{id}. Let $f_{\text{std}}(x) = M f'_{\text{std}}(x)$ for large $M \in \mathbb{R}$ (this magnifies its weights as discussed above), and let $f_{\text{rob}} = f'_{\text{rob}}$. f_{std} and f'_{std} have the same predictions and therefore accuracy but f_{std} is highly miscalibrated. The ensemble is then given by $f_{\text{ens}}(x) = M f'_{\text{std}}(x) + f'_{\text{rob}}(x)$. For very large M, f_{ens} and f'_{std} have the same predictions—this means that $\text{Err}(f_{\text{ens}}) = \text{Err}(f_{\text{std}}) < \text{Err}_{\text{ood}}(f_{\text{rob}})$, and so ensembling does not get the best of both worlds.

4.2 OOD PERFORMANCE OF ENSEMBLES

We showed that if f_{std} and f_{rob} are calibrated on a distribution P, then f_{ens} is better than both models on P. However, our validation data is from P_{id}, so we can only calibrate f_{std} and f_{rob} ID. Even after this ID calibration step, f_{std} and f_{rob} are very miscalibrated OOD (on P_{ood}—see Appendix B.3 and Ovadia et al. [2019]).

Our goal in this section is to build basic intuitions for when ID-calibrated ensembles can get high OOD accuracy. We draw inspiration from distribution shift benchmarks but examine simplified and stylized shifts. A toy version of our analysis is visualized in Figure 2, where the standard model relies on spurious features that change out-of-distribution. If these features are "suppressed" or "missing" OOD, then f_{ens} does better than f_{std} and f_{rob} (Figure 2b). However, if these features are anticorrelated OOD (correlated with the opposite label) then the accuracy of f_{ens} is between f_{std} and f_{rob} (Figure 2c). We begin by formalizing these shifts, and then analyze the accuracy under these shifts.

Missing spurious. For our first setting, we draw inspiration from some distribution shift benchmarks. Consider Breeds Living-17 [Santurkar et al., 2020] where the goal is to classify an image as one of 17 animal categories. The category 'bear' in the ID training data contains images of black bears and sloth bears while the OOD dataset has images of brown bears and polar bears. A standard model trained on the ID dataset might latch onto very specific features about sloth bears (for example the presence of a shaggy mane) which are simply missing in the OOD dataset ($f_{\text{std}}(x) = 0$). A robust model could be trained to project these features out [Xie et al., 2021], so its predictions are still fairly reliable OOD.

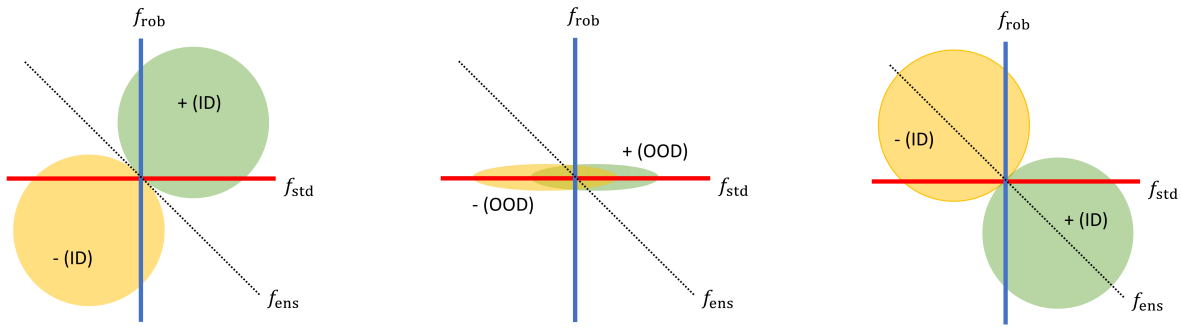
Definition 4.1 (missing spurious)**.** *A distribution P_0 has missing spurious features if for $x \sim P_0$, we have $f_{\text{std}}(x) = 0$ almost surely and for some $\alpha \in \mathbb{R}^+$, $P_0(Y = y \mid f_{\text{rob}}(X) = f_{\text{rob}}(x)) = \textit{softmax}(\alpha f_{\text{rob}}(x))_y$ for all $x \in \mathcal{X}$.*

Suppressed features. In some datasets, such as satellite remote sensing datasets [Jean et al., 2016, Xie et al., 2021], a standard model can latch onto country-specific features that may be less prevalent OOD.

Definition 4.2 (suppressed features)**.** *A distribution P_τ is said to have suppressed features if $P_\tau(Y = y \mid f(X) = f(x)) = \textit{softmax}(\tau f(x))_y$ for all $x \in \mathcal{X}$ and $f \in \{f_{\text{std}}, f_{\text{rob}}\}$, where $\tau \in \mathbb{R}^+$.*

Anticorrelated spurious. In some settings, the spurious feature can be correlated with a label ID but *anticorrelated* OOD. For example, in Waterbirds [Sagawa et al., 2020a], the task is to classify if an image contains a waterbird or a landbird where in the ID dataset, waterbirds are primarily featured with water backgrounds and landbirds with land backgrounds, but in the OOD datasets the backgrounds are flipped such that landbirds occur with water backgrounds and vice versa. This motivates the final definition of spurious shifts where the spurious features (background) are anticorrelated with the label OOD.

Definition 4.3 (anticorrelated spurious)**.** *A distribution P_{adv} is said to be anticorrelated spurious if for some $\alpha, \beta > 0$, for all $x \in \mathcal{X}$, $P_{\text{adv}}(Y = y | f_{\text{std}}(x)) = \textit{softmax}(-\beta f_{\text{std}}(x))_y$ (note the minus sign), while $P_{\text{adv}}(Y = y \mid f_{\text{rob}}(x)) = \textit{softmax}(\alpha f_{\text{rob}}(x))_y$.*



(a) In-distribution (b) Missing spurious (c) Anticorrelated spurious

Figure 2: A toy version of our analysis in Section 4. (Figure 2a) Given a standard model f_{std} (red horizontal line) and robust model f_{rob} (blue vertical line) that use different aspects of the data, ensembling their predictions gives a predictor f_{ens} (black dotted line) with lower error—in this case f_{ens} completely separates the positive (green circle) and negative (yellow circle) examples in-distribution (ID). (Figure 2b) f_{std} uses spurious features, suppose that these features are missing OOD (e.g., the y component of the input goes close to 0)—then f_{std} fares poorly and mislabels half the inputs, but the ensemble f_{ens} is about as accurate as the robust model f_{rob}. (Figure 2c) On the other hand, suppose the spurious features are *anticorrelated* with the label OOD. In this case f_{ens} intersects the positive (yellow circle) and negative (green circle) distributions, and gets 50% error—here f_{ens} is worse than f_{rob} but better than f_{std}.

If the OOD distribution is a mixture of suppressed features and missing spurious features, then the ensemble f_{ens} gets the best of both worlds.

Proposition 4.2. *If the OOD contains a mixture of suppressed features and missing spurious features i.e.,* $P_{\text{ood}} = \alpha P_\tau + (1-\alpha)P_0$*, and* P_τ *and* P_0 *are class-balanced, then we have* $Err_{\text{ood}}(f_{\text{ens}}) \leq Err_{\text{ood}}(f_{\text{rob}})$ *and* $Err_{\text{ood}}(f_{\text{ens}}) \leq Err_{\text{ood}}(f_{\text{std}})$.

On the other hand, if the OOD distribution contains anti-correlated spurious features, then the accuracy of f_{ens} is in between the standard and robust models.

Proposition 4.3. *If spurious features are anticorrelated OOD so that* $P_{\text{ood}} = P_{\text{adv}}$*, then even if* P_{adv} *is class-balanced,* $Err_{\text{ood}}(f_{\text{rob}}) \leq Err_{\text{ood}}(f_{\text{ens}}) \leq Err_{\text{ood}}(f_{\text{std}})$.

The full proofs appear in Appendix A.

5 DATASETS

We consider thirteen standard datasets, spanning multiple robustness interventions, types of shifts, and modalities (vision, language, time-series). We first describe the robustness interventions we consider, and then describe the datasets and types of shifts. All the datasets have been used by prior works on robustness, so we use their model checkpoints for reliable comparisons. See Appendix B.1 for more details.

Robustness interventions:

1. In-N-Out: Xie et al. [2021] use domain knowledge to project out spurious features in the input, and do an additional pretraining step. They call this robust model "aux-out" and show that it improves accuracy OOD, but hurts accuracy ID, compared to ERM.
2. Lightweight fine-tuning: We take checkpoints from **?** where the standard model fine-tunes all parameters on an ID dataset, and the robust model only learns the top linear 'head' layer (which does better ID but worse OOD).
3. Zero-shot language prompting: CLIP [Radford et al., 2021] is a multi-modal model that can predict the label of an image by comparing the image embedding, with prompts such as 'photo of an apple'. They show that this zero-shot language prompting approach (robust model) is more accurate OOD than fine-tuning the entire model (standard model), although ID accuracy of the robust model is worse.
4. Group distributionally robust optimization (DRO) [Sagawa et al., 2020a]: Standard ERM models often latch on to spurious correlations in a dataset, such as image background color, or the occurrence of certain words in a sentence. Group DRO essentially upweights examples where this spurious correlation is not present.

We consider three types of natural shifts (geography shifts, subpopulation shifts, style shifts), and we also consider adversarially synthesized "anticorrelated" spurious shifts.

Geography shifts. In geography shifts the ID data comes from some locations, and the OOD data comes from a different set of locations. One motivation is that in many developing areas training data may be unavailable because of monetary constraints [Jean et al., 2016].

1. **LandCover**: The goal is to classify a satellite image into one of 6 land types (e.g., "grassland", "savannas"). The ID data contains images from outside Africa, and the OOD data consists of images from Africa. Xie et al. [2021] use the In-N-Out intervention.
2. **Cropland**: The goal is to predict whether a satellite image is of a cropland or not. The ID dataset contains images from Iowa, Missouri, and Illinois, and the OOD dataset contains images from Indiana and Kentucky. Xie et al. [2021] use the In-N-Out intervention.

Subpopulation shifts. In subpopulation shifts, the ID data contains a few sub-categories (e.g., black bear and sloth bear), and the OOD data contains different sub-categories (e.g., brown bears and polar bears) of the same parent category (e.g., bears). For both datasets below, **?** use the lightweight fine-tuning intervention.

1. **Living-17**: the goal is to classify an image as one of 17 animal categories such as "bear", where the ID and OOD datasets have different species of bears.
2. **Entiy-30**: similar to Living-17, except the goal is to classify an image as one of 30 entity categories such as "food", "motor vehicle", and "insect".

Style shifts. In style shifts, the ID data has a certain style (e.g., sketches), and the OOD data has a different style (e.g., real photos, renditions).

1. **DomainNet** [Peng et al., 2019]: a standard domain adaptation dataset. 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. **?** use the lightweight fine-tuning intervention.
2. **CelebA** [Liu et al., 2015]: the goal is to classify a portrait of a face as "male" or "female" - the ID dataset contains images of people without hats, and the OOD dataset contains images of people wearing hats (some facial features might be "suppressed" or "missing" with hats). Xie et al. [2021] use the In-N-Out intervention.
3. **CIFAR->STL**: standard domain adaptation dataset [French et al., 2018], where the ID is CIFAR-10 [Krizhevsky, 2009], and the OOD is STL [Coates et al., 2011]. The task is to classify an image into one of 10 categories such as "dog", "cat", or "airplane". **?** use the lightweight fine-tuning intervention.
4. **ImageNet** [Russakovsky et al., 2015]: a large scale dataset where the goal is to classify an image into one of 1000 categories. Radford et al. [2021] use the zero-shot language prompting intervention. We evaluate on 3 standard OOD datasets: **ImageNetV2** [Recht et al., 2019],**ImageNet-R** [Hendrycks et al., 2020], and **ImageNet-Sketch** [Wang et al., 2019].

Anticorrelated spurious shifts. In these adversarially synthesized shifts, the ID dataset contains a feature that is correlated with a label, but this correlation is flipped OOD. Jones et al. [2021] use the group DRO intervention.

1. **Waterbirds** [Sagawa et al., 2020a]: The goal is to classify an image as a "waterbird" or "landbird". The dataset is synthetically constructed to have anticorrelated spurious features: "water" backgrounds are correlated with "waterbird" labels in the ID, but anticorrelated OOD.
2. **MNLI** [Williams et al., 2018]: The goal is to predict whether a hypothesis is entailed, contradicted by, or neutral to an associated premise. Sagawa et al. [2020a] partition the dataset so that "negation" words are correlated with the contradiction label ID but these words are anticorrelated with the contradiction label OOD.
3. **CivilComments** [Borkan et al., 2019]: The goal is to predict whether a comment is toxic or not. Jones et al. [2021] partition the dataset so that in the ID split mentions of a Christian identity are correlated with non-toxic comments, but in the OOD split mentions of a Christian identity are correlated with a toxic comment. CivilComments is also used in Koh et al. [2021].

6 RESULTS

In Section 6.1, we show that ID-calibrated ensembles get the best of both worlds across the 10 natural shifts we consider, but not on the 3 adversarially synthesized anticorrelated spurious shifts, as predicted by our analysis in Section 4. ID-calibrated ensembles match or outperform a prior state-of-the art approach based on self-training [Xie et al., 2021], which requires additional unlabeled data. In Section 6.2, we show ablations of our method. Interestingly, we find that a common approach of tuning the ensemble weights to optimize ID accuracy can lead to poor OOD performance.

6.1 MAIN RESULTS

Competitive with self-training. Xie et al. [2021] propose self-training on unlabeled data to mitigate ID-OOD accuracy tradeoffs. We run experiments on all 3 datasets they consider (Landcover, Cropland, CelebA), taking checkpoints from the official CodaLab implementation of Xie et al. [2021]. Table 1 shows that ID-calibrated ensembles match or outperform self-training on all 3 of their datasets, both ID and OOD. We believe this is interesting because our method is simple and does not need additional unlabeled data (which, for example, the other datasets do not have).

Strong ID and OOD accuracy. Across the 10 natural shifts, ID-calibrated ensembles get the best of both worlds, typically outperforming the standard and robust model both

Table 1: Xie et al. [2021] propose In-N-Out (self-training) to mitigate ID-OOD accuracy tradeoffs—their method requires lots of unlabeled data. Even without this unlabeled data, ID-calibrated ensembles are competitive with or outperform self-training ID and OOD. We show results on all datasets used by Xie et al. [2021].

	Cropland		Landcover		CelebA	
	ID Acc	OOD Acc	ID Acc	OOD Acc	ID Acc	OOD Acc
Standard model	95.3 (0.0)	**85.6 (5.8)**	76.9 (0.3)	55.7 (1.1)	90.4 (0.5)	74.5 (0.6)
Robust model	95.1 (0.1)	89.8 (0.4)	72.7 (0.2)	**60.4 (1.1)**	**94.5 (0.2)**	76.3 (1.2)
Self-training	95.3 (0.2)	**90.6 (0.6)**	**77.0 (0.4)**	**61.0 (0.7)**	93.1 (0.2)	**78.7 (0.7)**
Cal ensembling	**95.6 (0.1)**	**91.3 (0.8)**	**77.2 (0.2)**	**60.8 (0.8)**	**94.5 (0.5)**	**77.6 (1.2)**

Table 2: *In-distribution (ID)* accuracies for the standard model, robust model, and ID-calibrated ensembles, across 8 natural shift datasets (colored blue) and 3 anticorrelated spurious shift datasets (colored red and starred). On the 8 ID natural shift datasets, ID-calibrated ensembles match or outperform the best model in 7/8 cases, and on average outperforms both the standard and robust models. For the remaining dataset, CIFAR-10, ID-calibrated ensembles close 97% of the gap between the standard and robust model.

	Ent30	DomNet	CIFAR10	Liv17	Land	Crop	CelebA
Standard	**93.6 (0.2)**	83.9 (1.0)	**97.4 (0.1)**	96.9 (0.1)	**76.9 (0.3)**	95.3 (0.0)	90.4 (0.5)
Robust	90.7 (0.2)	89.2 (0.1)	92.0 (0.0)	97.0 (0.0)	72.7 (0.2)	95.1 (0.1)	**94.5 (0.2)**
Cal Ensemble	**93.7 (0.1)**	**91.2 (0.7)**	97.2 (0.1)	**97.2 (0.2)**	**77.2 (0.2)**	**95.6 (0.1)**	**94.5 (0.5)**

	ImageNet	MNLI*	Waterbirds*	CivilComments*
Standard	81.7 (-)	**82.9 (-)**	88.3 (-)	**92.8 (-)**
Robust	68.4 (-)	81.5 (-)	**93.2 (-)**	86.3 (-)
Cal Ensemble	**82.0 (-)**	**82.8 (-)**	92.9 (-)	91.4 (-)

ID (Table 2) and OOD (Table 3). Averaged across the *natural shift* datasets, ID-calibrated ensembles get 91.1% ID (vs. 89.5% for the standard model and 87.5% for the robust model) and 75.3% OOD (vs. 73.2% for the robust model and 65.6% for the standard model). The method works across the board—ID-calibrated ensembles achieve the best performance on 7/8 ID natural shifts, and on 9/10 OOD natural shifts. For the remaining two cases, DomainNet OOD and CIFAR-10 ID, ID-calibrated ensembles close over 95% of the gap between the standard and robust model.

Shift type is important. Our analysis in Section 4 predicts that ID-calibrated ensembles *do not* work as well on anticorrelated spurious shifts, where a spurious feature is correlated with the label but anticorrelated OOD. Indeed, in these cases the OOD accuracy of ID-calibrated ensembles is between the standard and robust model (Table 3). Even so, averaged across all 13 datasets ID-calibrated ensembles do well and get 90.5% ID (vs. 89.1% for the standard model, 87.3% for the robust model) and 75.3% OOD (vs. 64.5% for the standard model, 75.5% for the robust model).

6.2 ABLATIONS

Our proposed method is a simple combination of a calibrated robust and calibrated standard model. We vary the components of our method and try: (i) tuned ensembles without calibration, (ii) vanilla ensembles without calibration, and (iii) ensembles of two standard or two robust models.

Tuned ensembles do not mitigate tradeoffs. A natural way to ensemble the two models is "tuned ensembles": choosing the ensemble weights to optimize accuracy on the ID validation set. This approach is also known as stacking, and has performed well on the Netflix prize and Kaggle competitions [**?**]. Interestingly, we find that tuned ensembles do not do very well OOD, getting an average accuracy of 72.6% across the 13 datasets (vs. 75.3% for ID-calibrated ensembles). The ID accuracies are similar—results for all datasets are in Table 4 (ID) and Table 5 (OOD).

Calibration helps. ID-calibrated ensembles (calibration is only done on ID data) outperform vanilla ensembles, getting an average ID accuracy of 90.5% (vs. 89.9% for vanilla ensembles) and an average OOD accuracy of 75.3% (vs. 73.7% for vanilla ensembles). We show results for all datasets in Table 4 (ID) and Table 5 (OOD).

Table 3: *Out-of-distribution (OOD)* accuracies for the standard model, robust model, and ID-calibrated ensembles, across 10 natural shift datasets (colored blue) and 3 anticorrelated spurious shift datasets (colored red and starred). On the 10 OOD natural shift datasets, ID-calibrated ensembles match or outperform the best model in 9/10 cases, and on average outperforms both the standard and robust models. For the remaining dataset, DomainNet, ID-calibrated ensembles close 96% of the gap between the standard and robust model. As expected from our analysis (Section 4), on anticorrelated spurious shifts the accuracy of ID-calibrated ensembles is between the standard and robust models.

	Ent30	DomNet	CIFAR10	Liv17	Land	Crop	CelebA
Standard	60.7 (0.1)	55.3 (0.4)	82.4 (0.3)	77.7 (0.6)	55.7 (1.1)	**85.6 (5.8)**	74.5 (0.6)
Robust	63.2 (1.1)	**87.2 (0.1)**	85.1 (0.2)	**82.2 (0.2)**	**60.4 (1.1)**	89.8 (0.4)	76.3 (1.2)
Cal Ensemble	**64.7 (0.5)**	86.1 (0.2)	**87.3 (0.2)**	**82.2 (0.6)**	**60.8 (0.8)**	**91.3 (0.8)**	**77.6 (1.2)**

	ImNet-R	ImNet-V2	ImNet-Sk	MNLI*	Waterbirds*	CivilComments*
Standard	52.4 (-)	71.5 (-)	40.5 (-)	65.5 (-)	60.4 (-)	56.8 (-)
Robust	77.5 (-)	61.9 (-)	48.2 (-)	**77.4 (-)**	**88.1 (-)**	**84.2 (-)**
Cal Ensemble	**77.9 (-)**	**73.2 (-)**	**52.3 (-)**	73.2 (-)	81.1 (-)	71.8 (-)

Outperforms standard and robust ensembles. As a sanity check, Appendix B.2 shows that our method outperforms 1. ensembling two (calibrated) standard models, and 2. ensembling two (calibrated) robust models.

Models are miscalibrated OOD. Even after ID calibration, we find that the standard and robust models are not calibrated OOD, which matches prior work [Ovadia et al., 2019]. We estimate the expected calibration error (ECE; Equation 2 in Guo et al. [2017]). Since we calibrated on ID data, the ECE is low ID (1.4% for the standard model, 2.4% for the robust model; Table 8). However, the ECE is high OOD (11.9% for the standard model, 7.1% for the robust model; Table 9) Appendix B.3 shows that even the relative confidence of the models can be wrong: the standard model can be *more confident* but *less accurate* OOD, after ID-calibration. Nonetheless, ID-calibrated ensembles get the best of both worlds, and Section 4 gives some simple intuitions for why this can happen.

7 RELATED WORKS AND DISCUSSION

Calibration. Calibration has been widely studied in machine learning [Naeini et al., 2014, Guo et al., 2017, Kumar et al., 2019], and applications such as meteorology [Murphy, 1973, DeGroot and Fienberg, 1983, Gneiting and Raftery, 2005], fairness [Hebert-Johnson et al., 2018], and healthcare [Jiang et al., 2012]. Many of these works focus on the in-distribution (ID) setting, where models are calibrated on the same distribution that they are evaluated on. Ovadia et al. [2019], Jones et al. [2021] show that if we calibrate (e.g., via temperature scaling) a model ID, it still has poor uncertainties OOD. However, we show that despite having poor uncertainties on traditional metrics, calibrated models can be combined effectively to mitigate ID-OOD tradeoffs.

Ensembling. Ensembling models is a common way to get an accuracy boost—typically the ensemble members are trained with a different random seed [Lakshminarayanan et al., 2017] or augmentation [Stickland and Murray, 2020]. In the setting where the ensemble members mostly differ by random seeds or augmentations, prior work has shown that calibrating the members of an ensemble does not help [Wu and Gales, 2021, Ovadia et al., 2019]. However when we combine two very different models (standard and robust), calibration leads to clear improvements.

Mitigating ID-OOD tradeoffs. Tradeoffs between ID and OOD accuracy are widely studied and prior work self-trains on large amounts of unlabeled data to mitigate such tradeoffs [Raghunathan et al., 2020, Xie et al., 2021, Khani and Liang, 2021]. In contrast, our approach uses no extra unlabeled data and is a simple method where we just add up the model probabilities after a quick calibration step. In concurrent and independent work, [Wortsman et al., 2021] show that there *exists* a way to combine a CLIP zero-shot and fine-tuned model to get good ID and OOD accuracy—however learning how to combine the models may require OOD data, which is not available. We show that the natural way to learn how to weight ensemble members—selecting the weights to optimize ID accuracy—does not get the best of both worlds. In addition, their approach does not directly apply to settings where the standard and robust models have different architectures, such as In-N-Out [Xie et al., 2021].

Discussion. In this paper, we show that ID-calibrated ensembles, a simple method of calibrating a standard and robust model only on ID data and then ensembling them, can eliminate the tradeoff between in-distribution (ID) and out-of-distribution (OOD) accuracy on a wide range of natural shifts. We hope that this leads to more widespread use and deployment of robustness interventions.

References

EA AlBadawy, A Saha, and MA Mazurowski. Deep learning for segmentation of brain tumors: Impact of cross-institutional training and testing. *Med Phys.*, 45, 2018.

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

Yu Bai, Song Mei, Huan Wang, and Caiming Xiong. Don't just blame over-parametrization for over-confidence: Theoretical analysis of calibration in binary classification. *arXiv*, 2021.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In *World Wide Web (WWW)*, pages 491–500, 2019.

Jochen Brocker. Reliability, sufficiency, and the decomposition of proper scores. *Quarterly Journal of the Royal Meteorological Society*, 135(643):1512–1519, 2009.

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.

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.

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.

Shrey Desai and Greg Durrett. Calibration of pre-trained transformers. *arXiv*, 2020.

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

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

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

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020.

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.

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.

Ursula Hebert-Johnson, Michael P. Kim, Omer Reingold, and Guy N. Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning (ICML)*, 2018.

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

Dan Hendrycks, Kimin Lee, and Mantas Mazeika. Using pre-training can improve model robustness and uncertainty. In *International Conference on Machine Learning (ICML)*, 2019.

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.

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.

Neal Jean, Marshall Burke, Michael Xie, W. Matthew Davis, David B. Lobell, and Stefano Ermon. Combining satellite imagery and machine learning to predict poverty. *Science*, 353, 2016.

Xiaoqian Jiang, Melanie Osl, Jihoon Kim, and Lucila Ohno-Machado. Calibrating predictive model estimates to support personalized medicine. *Journal of the American Medical Informatics Association*, 19(2):263–274, 2012.

Erik Jones, Shiori Sagawa, Pang Wei Koh, Ananya Kumar, and Percy Liang. Selective classification can magnify disparities across groups. In *International Conference on Learning Representations (ICLR)*, 2021.

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.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips,

Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton A. Earnshaw, Imran S. Haque, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning (ICML)*, 2021.

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

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

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.

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

Evan Zheran Liu, Behzad Haghgoo, Annie S. Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning (ICML)*, 2021.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3730–3738, 2015.

Jishnu Mukhoti, Viveka Kulharia, Amartya Sanyal, Stuart Golodetz, Philip H.S. Torr, and Puneet K. Dokania. Calibrating deep neural networks using focal loss. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

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

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

Vaishnavh Nagarajan, Anders Andreassen, and Behnam Neyshabur. Understanding the failure modes of out-of-distribution generalization. *arXiv preprint arXiv:2010.15775*, 2020.

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.

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.

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.

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.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning (ICML)*, 2019.

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.

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.

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)*, 2020a.

Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An investigation of why overparameterization exacerbates spurious correlations. In *International Conference on Machine Learning (ICML)*, 2020b.

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

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

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

Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

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.

Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Association for Computational Linguistics (ACL)*, pages 1112–1122, 2018.

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.

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

Sang Michael Xie, Ananya Kumar, Robbie 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.

John R. Zech, Marcus A. Badgeley, Manway Liu, Anthony B. Costa, Joseph J. Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. In *PLOS Medicine*, 2018.

A PROOFS FOR SECTION 4

We begin with some standard background on Bayes optimal classifiers. When then prove the results in Section 4. By default, expectations are taken over all random variables.

A.1 BACKGROUND ON BAYES-OPTIMAL CLASSIFIERS

These results are all standard, but we include it as background information since different texts use different notations. Let $Z \in \mathcal{Z}$ denotes some features (that can be complicated functions of the input x, for example the output of a neural network), and let $Y \in \mathcal{Y}$ denote the label. Let P be a distribution over (Z, Y). The Bayes-optimal classifier predicts the most likely label y given features z.

Definition A.1. *The Bayes-optimal classifier for P given features z is given by:*

$$y_*(z) = \arg\min_{y \in \mathcal{Y}} P(y \mid z). \tag{A.1}$$

The Bayes-optimal classifier has the minimum misclassification error of all possible classifiers that use $z \in \mathcal{Z}$ to predict $y \in \mathcal{Y}$. Formally, the error of a classifier $\widehat{y}$ is the probability that it gets the label incorrect.

Definition A.2. *The error of a predictor $\widehat{y} : \mathcal{Z} \to \mathcal{Y}$ on distribution P is given by:*

$$Err_P(\widehat{y}) = P(Y \neq \widehat{y}(Z)), \tag{A.2}$$

Alternatively, we can look at the error for each Z, and then take the average over Z, which gives us:

Lemma A.1. *The error of a predictor $\widehat{y} : \mathcal{Z} \to \mathcal{Y}$ on distribution P can be written as:*

$$Err_P(\widehat{y}) = \mathbb{E}[1 - P(Y = \widehat{y}(Z) \mid Z)]. \tag{A.3}$$

Proof. We can write the misclassification probability as an expectation over an indicator and then apply the law of total expectation.

$$P(Y \neq \widehat{y}(Z)) = \mathbb{E}[\mathbb{I}(Y \neq \widehat{y}(Z))] \tag{A.4}$$
$$= \mathbb{E}[\mathbb{E}[\mathbb{I}(Y \neq \widehat{y}(Z)) \mid Z]]. \tag{A.5}$$

And then just write the inner expectation as a probability.

$$\mathbb{E}[\mathbb{E}[\mathbb{I}(Y \neq \widehat{y}(Z)) \mid Z]] = \mathbb{E}[P(Y \neq \widehat{y}(Z) \mid Z)] \tag{A.6}$$
$$= \mathbb{E}[1 - P(Y = \widehat{y}(Z) \mid Z)]. \tag{A.7}$$

□

The Bayes-optimal classifier selects the y with the highest probability given z, so we have:

Lemma A.2. *The error of the Bayes-optimal classifier y_* on a distributon P can be written as (where $Z \sim P$):*

$$Err_P(y_*) = \mathbb{E}[1 - \max_{y \in \mathcal{Y}} P(Y = y \mid Z)]. \tag{A.8}$$

Proof. The proof is immediate by substituting the definition of the Bayes-optimal classifier (Definition A.1) into the alternative formula for the error in Lemma A.1. □

From the above, it is clear that the Bayes-optimal classifier has lower error than any other classifier that uses only z, formalized below.

Lemma A.3. *The bayes-optimal classifier (for P) has lower error than all classifiers $\widehat{y} : \mathcal{Z} \to \mathcal{Y}$:*

$$Err_P(y_*) \leq Err_P(\widehat{y}). \tag{A.9}$$

Proof. Beginning from Lemma A.1, we have:

$$\begin{aligned}
\mathrm{Err}_P(\widehat{y}) &= \mathbb{E}[1 - P(Y = \widehat{y}(Z) \mid Z)] && \text{(A.10)}\\
&\geq \mathbb{E}[1 - \max_{y \in \mathcal{Y}} P(Y = y \mid Z)] && \text{(A.11)}\\
&= \mathrm{Err}_P(y_*). && \text{(A.12)}
\end{aligned}$$

□

As a simple corollary, we note that the accuracy of the Bayes-optimal classifier is at least the frequency of the most common label.

Corollary A.1. *If y_* is bayes-optimal for P then,*

$$Err_P(y_*) \leq 1 - \max_{y \in \mathcal{Y}} P(Y = y) \tag{A.13}$$

So for example if P is balanced, then the Bayes-opt classifier will have accuracy at least $1/K$, where K is the number of classes.

Note that calibrated classifiers are Bayes-optimal given their outputs. Formally, let P be a distribution over (x, y), and suppose f is calibrated with respect to P. Let $z = f(x)$ and let P' be the induced distribution over (z, y). Then f is Bayes-optimal for P' given features z. The label distributions $P'(y)$ and $P(y)$ are the same, so Lemma A.3 applies to any calibrated classifier.

A.2 PROOF OF PROPOSITION 4.1

Restatement of Proposition 4.1. *Suppose that f_{std} and f_{rob} are calibrated with respect to P_{id}, and that P_{id} is class-balanced. Let $h : \mathbb{R}^K \times \mathbb{R}^K \to \mathbb{R}^K$ be an arbitrary function that combines the standard and robust model's predictions, and let f_h be the resulting classifier: $f_h(x) = h(f_{\mathsf{std}}(x), f_{\mathsf{rob}}(x))$. The ensemble is better than any such combination classifier f_h: $Err_{\mathsf{id}}(f_{\mathsf{ens}}) \leq Err_{\mathsf{id}}(f_h)$.*

We first show that in the setting of the Proposition, we can write $P(y \mid f_{\mathsf{rob}}(x), f_{\mathsf{rob}}(x))$ in terms of $f_{\mathsf{rob}}(x)$ and $f_{\mathsf{std}}(x)$.

Lemma A.4. *In the setting of Proposition 4.1, let $m \in \mathbb{R}^K$ be the log of the marginal probabilities $P(y)$:*

$$m_y = \log P(y), \quad \textit{for all } y \in [K]. \tag{A.14}$$

Then we have:

$$P(y \mid f_{\mathsf{std}}(x), f_{\mathsf{rob}}(x)) = \textit{softmax}(f_{\mathsf{std}}(x) + f_{\mathsf{rob}}(x) - m)_y, \quad \textit{for all } y \in [K]. \tag{A.15}$$

In the balanced setting, where $P(y) = 1/K$ for all y, this simplifies to:

$$P(y \mid f_{\mathsf{std}}(x), f_{\mathsf{rob}}(x)) = \textit{softmax}(f_{\mathsf{std}}(x) + f_{\mathsf{rob}}(x))_y, \quad \textit{for all } y \in [K]. \tag{A.16}$$

Proof. Fix $r = f_{\mathsf{rob}}(x)$ and $s = f_{\mathsf{std}}(x)$, where $r, s \in \mathbb{R}^K$. We first rewrite the probability of y given the robust and standard model outputs $P(y \mid r, s)$ in terms of the probability of y given each of the individual model outputs: $P(y \mid r)$ and $P(y \mid s)$. We do this for discrete random variables for simplicity, but the same result follows by using Bayes rule for general random variables.

$$\begin{aligned}
P(y \mid r, s) &= \frac{P(r, s \mid y)P(y)}{P(r, s)} && \text{[Bayes rule]} && \text{(A.17)}\\
&= \frac{P(r \mid y)P(s \mid y)P(y)}{P(r, s)} && [r \perp s \mid y] && \text{(A.18)}\\
&= \frac{\left[\frac{P(y|r)P(r)}{P(y)} \frac{P(y|s)P(s)}{P(y)} P(y)\right.}{P(r, s)} && \text{[Bayes rule]} && \text{(A.19)}\\
&= \frac{P(y \mid r)P(y \mid s)}{P(y)} \left[\frac{P(r)P(s)}{P(r, s)}\right] && \text{[Algebra]} && \text{(A.20)}\\
& && && \text{(A.21)}
\end{aligned}$$

Since r, s are fixed, we can denote the terms that do not depend on y by a constant c_1,

$$c_1 = \frac{P(r)P(s)}{P(r,s)}. \tag{A.22}$$

So then we can write:

$$P(y \mid r, s) = \frac{P(y \mid r)P(y \mid s)}{P(y)} c_1, \quad \text{for all } y \in [K]. \tag{A.23}$$

Now, we assumed $P(Y = y \mid r) = \text{softmax}(r)_y$ and $P(Y = y \mid s) = \text{softmax}(s)_y$ for all $y \in [K]$. For some constants $c_2, c_3 \in \mathbb{R}$, we can write this as: $P(Y = y \mid r) = \exp(r_y)/c_2$ and $P(Y = y \mid s) = \exp(s_y)/c_3$ for all $y \in [K]$. Substituting this into Equation A.23, we get:

$$P(y \mid r, s) = \frac{\exp(r_y + s_y)}{P(y)} \frac{c_1}{c_2 c_3}, \quad \text{for all } y \in [K]. \tag{A.24}$$

Writing $1/P(y)$ as $\exp(-\log P(y))$, and setting $c_4 = \frac{c_1}{c_2 c_3}$, this gives us:

$$P(y \mid r, s) = c_4 \exp(r_y + s_y - \log P(y)), \quad \text{for all } y \in [K]. \tag{A.25}$$

Since the LHS is a probability, these must sum to 1 and so c_4 must be a normalizing constant, that is, $c_4 = 1/(\sum_{y \in [K]} \exp(r_y + s_y - \log P(y)))$. This gives us:

$$P(y \mid r, s) = \text{softmax}(r + s - m)_y, \quad \text{for all } y \in [K], \tag{A.26}$$

which is precisely Equation A.15. In the balanced setting, we have $P(Y) = 1/K$ so we simply fold $P(Y)$ into the constant c_4, and get:

$$P(y \mid r, s) = \text{softmax}(r + s)_y, \quad \text{for all } y \in [K], \tag{A.27}$$

which is precisely Equation A.16. □

Now we are ready to prove Proposition 4.1.

Proof of Proposition 4.1. We assumed the "balanced" setting where $P(y) = 1/K$ for all y. From Lemma A.4, letting $f_{\mathsf{ens}}(x) = f_{\mathsf{std}}(x) + f_{\mathsf{rob}}(x)$, we have:

$$P(y \mid f_{\mathsf{std}}(x), f_{\mathsf{rob}}(x)) = \text{softmax}(f_{\mathsf{ens}}(x))_y, \tag{A.28}$$

So this means that the ensemble prediction is the Bayes optimal given $(f_{\mathsf{std}}(x), f_{\mathsf{rob}}(x))$:

$$\text{pred}(f_{\mathsf{ens}}(x)) = \arg\max_y f_{\mathsf{ens}}(x)_y = \arg\max_y \text{softmax}(f_{\mathsf{ens}}(x))_y = \arg\max_y P(y \mid f_{\mathsf{std}}(x), f_{\mathsf{rob}}(x)). \tag{A.29}$$

But then from Lemma A.3, any other predictor which uses only $(f_{\mathsf{rob}}(x), f_{\mathsf{std}}(x))$ must have higher error. This completes the proof.

Note that the inequality in the above proof is a strict inequality except in degenerate cases: as long as f_{std} and f_{rob} sometimes disagree in their predictions, and in some of these cases f_{std} assigns a higher probability to its predictions, and in some cases f_{rob} assigns a higher probability to its prediction, the inequalities will be strict inequalities. □

A.3 PROOF OF PROPOSITION 4.2

Restatement of Proposition 4.2. *If the OOD contains a mixture of suppressed features and missing spurious features i.e., $P_{\mathsf{ood}} = \alpha P_\tau + (1-\alpha)P_0$, and P_τ and P_0 are class-balanced, then we have $Err_{\mathsf{ood}}(f_{\mathsf{ens}}) \leq Err_{\mathsf{ood}}(f_{\mathsf{rob}})$ and $Err_{\mathsf{ood}}(f_{\mathsf{ens}}) \leq Err_{\mathsf{ood}}(f_{\mathsf{std}})$.*

Proof. We first note that errors are additive. That is, letting:

$$\text{Err}(P, f) = \mathbb{E}[\text{pred}(f(x)) \neq y], \text{ where } x, y \sim P, \tag{A.30}$$

we have:

$$\text{Err}(\alpha P_\tau + (1-\alpha)P_0, f) = \alpha \text{Err}(P_\tau, f) + (1-\alpha)\text{Err}(P_0, f) \tag{A.31}$$

So it suffices to prove that the ensemble is better than the standard and robust models for P_τ and P_0 separately.

Suppressed features. Let $\overline{f_{\mathsf{rob}}}(x) = \tau f_{\mathsf{rob}}(x)$ and $\overline{f_{\mathsf{std}}}(x) = \tau f_{\mathsf{std}}(x)$ be scaled versions of the standard and robust models. Definition 4.2 implies that $\overline{f_{\mathsf{rob}}}$ and $\overline{f_{\mathsf{std}}}$ are calibrated. Since we assumed P_τ is balanced, by Proposition 4.1, $\overline{f_{\mathsf{ens}}}$ given by $\overline{f_{\mathsf{ens}}(x)} = \tau f_{\mathsf{rob}}(x) + \tau f_{\mathsf{std}}(x)$ has optimal error on P_τ. But for all x, the predictions of f_{ens} and $\overline{f_{\mathsf{ens}}}$ are the same (multiplying the outputs of a model by a constant does not change the predicted output, which is the $\arg\max$). So f_{ens} also has optimal error on P_τ:

$$\text{Err}(P_\tau, f_{\mathsf{ens}}) \leq \text{Err}(P_\tau, f_{\mathsf{std}}), \text{ and } \text{Err}(P_\tau, f_{\mathsf{ens}}) \leq \text{Err}(P_\tau, f_{\mathsf{rob}}) \tag{A.32}$$

Note that these inequalities are strict inequalities except in degenerate cases: as long as f_{std} and f_{rob} sometimes disagree in their predictions, and in some of these cases f_{std} assigns a higher probability to its predictions, and in some cases f_{rob} assigns a higher probability to its prediction, the inequalities will be strict inequalities.

Missing spurious. If $f_{\mathsf{std}}(x) = 0$ almost surely, then $f_{\mathsf{ens}}(x) = f_{\mathsf{rob}}(x) + f_{\mathsf{std}}(x) = f_{\mathsf{rob}}(x)$ almost surely. Furthermore, if $f_{\mathsf{std}}(x) = 0$ then its error is lower bounded by $1 - \max_y P_0(y)$. On the other hand, $f_{\mathsf{rob}}(x)$ is calibrated and therefore Bayes-optimal given $z = f_{\mathsf{rob}}(x)$ so from Lemma A.1 (e.g., see the the discussion below the Lemma for more details) has error at most $1 - \max_y P_0(y)$. So we have:

$$\text{Err}(P_0, f_{\mathsf{ens}}) = \text{Err}(P_0, f_{\mathsf{rob}}) \leq \text{Err}(P_0, f_{\mathsf{std}}) \tag{A.33}$$

Note that the inequality is a strict inequality except in a degenerate case (where the probability that f_{rob} predicts for the most common class $\arg\max_y P_0(y)$ is the same for all inputs). □

A.4 PROOF OF PROPOSITION 4.3

Restatement of Proposition 4.3. *If spurious features are anticorrelated OOD so that $P_{\mathsf{ood}} = P_{\mathsf{adv}}$, then even if P_{adv} is class-balanced, $Err_{\mathsf{ood}}(f_{\mathsf{rob}}) \leq Err_{\mathsf{ood}}(f_{\mathsf{ens}}) \leq Err_{\mathsf{ood}}(f_{\mathsf{std}})$.*

Proof. Let $X, Y \sim P_{\mathsf{ood}}$, and let $Z = (f_{\mathsf{std}}(X), f_{\mathsf{rob}}(X))$ be the predictions of the standard and robust models. Fix $z = (f_{\mathsf{std}}(x), f_{\mathsf{rob}}(x))$, and let $s = f_{\mathsf{std}}(x)$ and $r = f_{\mathsf{rob}}(x)$. We will analyze the errors for fixed $Z = z$ (showing that the robust model is better than the ensemble, which is better than the standard model). Since this is true for all z, we then use Lemma A.1 (which is basically the law of total expectation), to get the desired result.

Bayes-opt classifier. Recall that for some $\alpha, \beta > 0$, we have $P_{\mathsf{adv}}(Y = y | f_{\mathsf{std}}(x)) = \text{softmax}(-\beta f_{\mathsf{std}}(x))_y$ for all x (note the minus sign), while $P_{\mathsf{adv}}(Y = y \mid f_{\mathsf{rob}}(x)) = \text{softmax}(\alpha f_{\mathsf{rob}}(x))_y$. Then, applying Lemma A.4, we have:

$$P_{\mathsf{adv}}(y \mid (f_{\mathsf{std}}(x), f_{\mathsf{rob}}(x))) = \text{softmax}(\alpha f_{\mathsf{rob}}(x) - \beta f_{\mathsf{std}}(x))_y. \tag{A.34}$$

Rewriting this in terms of z, r, s, we have:

$$P_{\mathsf{adv}}(y \mid z) = \text{softmax}(\alpha r - \beta s)_y. \tag{A.35}$$

Ensemble vs. robust classifier. Let $j_{\mathsf{rob}} = \arg\max_y r_y$ be the robust model's prediction, and $j_{\mathsf{ens}} = \arg\max_y (r+s)_y$ be the ensemble model's prediction. Because j_{rob} is the $\arg\max$ of r, we have:

$$r_{j_{\mathsf{rob}}} \geq r_{j_{\mathsf{ens}}}. \tag{A.36}$$

Because j_{ens} is the $\arg\max$ of $r + s$, we have:

$$r_{j_{\mathsf{ens}}} + s_{j_{\mathsf{ens}}} \geq r_{j_{\mathsf{rob}}} + s_{j_{\mathsf{rob}}}. \tag{A.37}$$

Taking the negation of this, we get:

$$-r_{j_{\mathsf{rob}}} - s_{j_{\mathsf{rob}}} \geq -r_{j_{\mathsf{ens}}} - s_{j_{\mathsf{ens}}}. \tag{A.38}$$

Adding β times Inequality A.38 to $(\alpha + \beta)$ times Inequality A.36, we get:

$$\alpha r_{j_{\mathsf{rob}}} - \beta s_{j_{\mathsf{rob}}} \geq \alpha r_{j_{\mathsf{ens}}} - \beta s_{j_{\mathsf{ens}}}. \tag{A.39}$$

Since softmax is monotonic, we have:

$$\text{softmax}(\alpha r - \beta s)_{j_{\mathsf{rob}}} \geq \text{softmax}(\alpha r - \beta s)_{j_{\mathsf{ens}}}. \tag{A.40}$$

But from Equation A.35 the LHS is the same as the robust model's probability of getting the label correct, and the RHS is the same as the ensemble's probability of getting the label correct:

$$P_{\mathsf{adv}}(Y = j_{\mathsf{rob}} \mid Z = z) \geq P_{\mathsf{adv}}(Y = j_{\mathsf{ens}} \mid Z = z). \tag{A.41}$$

Taking negations (to get the error), and then the expectation over $Z = z$, we get (note that below we write the error, which is why the sign is now flipped):

$$\text{Err}_{\mathsf{ood}}(f_{\mathsf{ens}}) \geq \text{Err}_{\mathsf{ood}}(f_{\mathsf{rob}}). \tag{A.42}$$

Which is what we wanted to show.

Ensemble vs. standard classifier. The argument is fairly analogous to the previous case, with some minor differences in the algebra in the first part. Let $j_{\mathsf{std}} = \arg\max_y s_y$ be the standard model's prediction. Because j_{std} is the $\arg\max$ of s, we have:

$$s_{j_{\mathsf{std}}} \geq s_{j_{\mathsf{ens}}}. \tag{A.43}$$

Taking the negation of this, we get:

$$-s_{j_{\mathsf{ens}}} \geq -s_{j_{\mathsf{std}}}. \tag{A.44}$$

Because j_{ens} is the $\arg\max$ of $r + s$, we have:

$$r_{j_{\mathsf{ens}}} + s_{j_{\mathsf{ens}}} \geq r_{j_{\mathsf{std}}} + s_{j_{\mathsf{std}}}. \tag{A.45}$$

Adding α times Inequality A.45 with $(\alpha + \beta)$ times Inequality A.44, we get:

$$\alpha r_{j_{\mathsf{ens}}} - \beta s_{j_{\mathsf{ens}}} \geq \alpha r_{j_{\mathsf{std}}} - \beta s_{j_{\mathsf{std}}}. \tag{A.46}$$

The rest of this step is the same as in the comparison between the ensemble and the robust model. Since softmax is monotonic, we have:

$$\text{softmax}(\alpha r - \beta s)_{j_{\mathsf{ens}}} \geq \text{softmax}(\alpha r - \beta s)_{j_{\mathsf{std}}}. \tag{A.47}$$

But from Equation A.35 the LHS is the same as the robust model's probability of getting the label correct, and the RHS is the same as the ensemble's probability of getting the label correct:

$$P_{\mathsf{adv}}(Y = j_{\mathsf{ens}} \mid Z = z) \geq P_{\mathsf{adv}}(Y = j_{\mathsf{std}} \mid Z = z). \tag{A.48}$$

Taking negations (to get the error), and then the expectation over $Z = z$, we get (note that below we write the error, which is why the sign is now flipped):

$$\text{Err}_{\mathsf{ood}}(f_{\mathsf{std}}) \geq \text{Err}_{\mathsf{ood}}(f_{\mathsf{ens}}). \tag{A.49}$$

Which is what we wanted to show.

□

B MORE INFORMATION ON EXPERIMENTS

B.1 ADDITIONAL DETAILS ON DATASETS

Here we describe the robustness interventions and datasets in more detail.

Robustness interventions:

1. In-N-Out [Xie et al., 2021]. Many datasets contain a core input x (image or time series data), and metadata z (e.g., location or climate data). Xie et al. [2021] show that using the metadata (in addition to x) improves accuracy in-distribution (ID), but hurts accuracy out-of-distribution. Xie et al. [2021] 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. We use official checkpoints from their CodaLab worksheet `https://worksheets.codalab.org/worksheets/0x2613c72d4f3f4fbb94e0a32c17ce5fb0`, and compare to the results tagged as "In-N-Out" on each dataset. They also show results after doing additional self-training on (unlabeled) OOD data, but we do not compare to this because 1. OOD data is assumed to be unavailable in our setting, and 2. if OOD unlabeled data is available, we can also start from ID-calibrated ensembles and do additional self-training.
2. Lightweight fine-tuning [**?**]: When adapting a pretrained model to an ID dataset, typically all the model parameters are fine-tuned. Recent works show that tuning only parts of the model can often do better OOD even though the ID performance is worse [Li and Liang, 2021, Houlsby et al., 2019]. On four distribution shift datasets, we take checkpoints from **?** 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.
3. Zero-shot language prompting: Radford et al. [2021] 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 (robust model) can be much more accurate OOD than the traditional method of fine-tuning the entire model (standard model), although ID accuracy of the robust model is worse. We use model checkpoints and datasets from Radford et al. [2021].
4. Group distributionally robust optimization (DRO) [Sagawa et al., 2020a]: Standard ERM models often latch on to spurious correlations in a dataset, such as image background color, or the occurrence of certain words in a sentence. Group DRO essentially upweights examples where this spurious correlation is not present. The original formulation in Sagawa et al. [2020a] assumes the spurious correlations are annotated, but newer variants [Liu et al., 2021] can work even without these annotations.

We consider three types of natural shifts (geography shifts, subpopulation shifts, style shifts), and we also consider adversarial spurious shifts.

Geography shifts. In geography shifts the ID data comes from some locations, and the OOD data comes from a different set of locations. One motivation is that in many developing areas training data may be unavailable because of monetary constraints [Jean et al., 2016].

1. **LandCover**: The goal is to classify a satellite images into one of 6 land types (e.g., "grassland", "savannas"). The ID data contains images from outside Africa, and the OOD data consists of images from Africa. We take model checkpoints from Xie et al. [2021] where they use the In-N-Out intervention—the core feature x is time series data measured by Nasa's MODIS satellite, and the spurious metadata z consists of climate data (e.g., temperature) at that location. We use the dataset from Rußwurm et al. [2020] which is the same as what Xie et al. [2021] use.
2. **Cropland**: The goal is to predict whether a satellite image is of a cropland or not. The ID dataset contains images from Iowa, Missouri, and Illinois, and the OOD dataset contains images from Indiana and Kentucky. We take model checkpoints from Xie et al. [2021] where they use the In-N-Out intervention—the core feature x is an RGB satellite image, and the spurious metadata z consists of location coordinates and vegetation bands. We use the dataset from Wang et al. [2020] which is the same as what Xie et al. [2021] use.

Subpopulation shifts. In subpopulation shifts, the ID data contains a few sub-categories (e.g., black bear and sloth bear), and the OOD data contains different sub-categories (e.g., brown bears and polar bears) or the same parent category (e.g., bears). For both datasets below, we take model checkpoints from **?** where they use the lightweight fine-tuning intervention, starting from a MoCo-v2 ResNet-50 model pretrained on unlabeled ImageNet images. The datasets are from Santurkar et al. [2020].

1. **Living-17**: 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.
2. **Entiy-30**: similar to Living-17, except the goal is to classify an image as one of 30 entity categories such as "food", "motor vehicle", and "index".

Style shifts. In style shifts, the ID data contains data in a certain style (e.g., sketches), and the OOD data contains data in a different style (e.g., real photos, renditions).

1. **DomainNet** [Peng et al., 2019]: a standard domain adaptation dataset. 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 take model checkpoints from **?** where they use the lightweight fine-tuning intervention, starting from a CLIP ResNet-50 model.
2. **CelebA** [Liu et al., 2015]: the goal is to classify a portrait of a face as "male" or "female" - the ID dataset contains images of people without hats, and the OOD dataset contains images of people wearing hats (some facial features might be "suppressed" or "missing" with hats). We take model checkpoints from Xie et al. [2021] where they use the In-N-Out intervention—the core feature x is the RGB image, and the spurious metadata z consists of 7 attribute tags annotated in the dataset (e.g., presence of makeup, beard).
3. **CIFAR->STL**: standard domain adaptation dataset [French et al., 2018], where the ID is CIFAR-10 [Krizhevsky, 2009], and the OOD is STL [Coates et al., 2011]. The task is to classify an image into one of 10 categories such as "dog", "cat", or "airplane". We take model checkpoints from **?** where they use the lightweight fine-tuning intervention, starting from a MoCo-v2 ResNet-50 model pretrained on unlabeled ImageNet images.
4. **ImageNet** [Russakovsky et al., 2015]: a large scale dataset where the goal is to classify an image into one of 1000 categories. We use the zero-shot language prompting intervention using a CLIP ViT-B/16 vision transformer model. We evaluate on 3 standard OOD datasets: **ImageNetV2** [Recht et al., 2019],**ImageNet-R** [Hendrycks et al., 2020], and **ImageNet-Sketch** [Wang et al., 2019].

Adversarial spurious shifts. In adversarial spurious shifts, the ID dataset contains a feature that is correlated with a label, but this correlation is flipped OOD.

1. **Waterbirds** [Sagawa et al., 2020a]: The goal is to classify an image as a "waterbird" or "landbird". The dataset is synthetically constructed to have adversarially spurious features: "water" backgrounds are correlated with "waterbird" labels in the ID, but anticorrelated OOD. We use checkpoints from Jones et al. [2021] where they use the group DRO intervention.
2. **MNLI** [Williams et al., 2018]: The goal is to predict whether a hypothesis is entailed, contradicted by, or neutral to an associated premise. We use the splits in Sagawa et al. [2020a]—they partition the dataset so that in-distribution "negation" words "nobody", "no", "never", and "nothing" are correlated with the contradiction label, however in the OOD dataset these words are anticorrelated with the contradiction label. We use checkpoints from Jones et al. [2021] where they use the group DRO intervention.
3. **CivilComments** [Borkan et al., 2019]: The goal is to predict whether a comment is toxic or not. We use the splits in Sagawa et al. [2020a]—they partition the dataset where in the ID split mentions of a Christian identity are correlated with non-toxic comments, but in the OOD split mentions of a Christian identity are correlated with a toxic comment. We use checkpoints from Jones et al. [2021] where they use the group DRO intervention. CivilComments is also used in Koh et al. [2021].

B.2 PER-DATASET RESULTS ON ENSEMBLING ABLATIONS

In Section 6.2 we ablated calibrated ensembles with "tuned" ensembles where the ensemble weights are tuned on in-distribution validation data, and with vanilla ensembles. Here, we show per-dataset results both ID (Table 4) and OOD (Table 5).

In Section 6.2, We also compared calibrated ensembles (of one standard and one robust model) with ensembles of two standard models, and ensembles of two robust models, where for a fair comparison all models are calibrated. We ran this ablation on 7 of the 13 datasets (Entity-30, DomainNet, CIFAR→STL, Living-17, Landcover, Cropland, and CelebA) because it requires multiple standard and multiple robust models, which were not available or very expensive to run on large datasets like ImageNet. Calibrated ensembles get an average ID accuracy of 92.0% (vs. 89.6% for a robust-robust ensemble

Table 4: *ID* accuracies: The in-distribution accuracies of calibrated ensembles, tuned ensembles, and vanilla ensembles are very close (within confidence intervals), so any of these methods are acceptable if we are looking at in-distribution accuracy. However, they perform quite differently when it comes to OOD accuracy (Table 5).

	Ent30	DomNet	CIFAR10	Liv17	Land	Crop	CelebA
Logits	93.7 (0.1)	89.3 (0.6)	**97.3 (0.1)**	**97.1 (0.2)**	**77.4 (0.1)**	95.5 (0.1)	93.4 (0.6)
Probs	**93.7 (0.1)**	89.1 (0.4)	**97.3 (0.1)**	**97.1 (0.2)**	**77.4 (0.2)**	95.5 (0.1)	93.4 (0.6)
Tuned Logits	**93.8 (0.0)**	**91.3 (0.2)**	**97.4 (0.1)**	97.1 (0.1)	**77.3 (0.4)**	**95.6 (0.1)**	**94.8 (0.2)**
Tuned Probs	93.8 (0.1)	90.6 (0.7)	**97.4 (0.1)**	**97.2 (0.1)**	77.1 (0.3)	95.5 (0.1)	**95.0 (0.2)**
Calibrated Logits	93.7 (0.1)	**91.1 (0.4)**	97.2 (0.1)	**97.2 (0.2)**	77.2 (0.2)	**95.6 (0.1)**	**94.5 (0.5)**
Calibrated Probs	93.7 (0.1)	**91.2 (0.7)**	97.2 (0.1)	**97.2 (0.2)**	77.2 (0.2)	**95.6 (0.1)**	**94.5 (0.5)**

	ImageNet	MNLI	Waterbirds	CivilComments
Logits	82.1 (-)	**82.9 (-)**	90.1 (-)	90.4 (-)
Probs	82.1 (-)	**82.9 (-)**	90.1 (-)	90.4 (-)
Tuned Logits	**82.7 (-)**	**83.0 (-)**	**93.2 (-)**	**92.7 (-)**
Tuned Probs	82.3 (-)	**83.0 (-)**	**93.2 (-)**	**92.6 (-)**
Calibrated Logits	82.0 (-)	82.8 (-)	92.9 (-)	91.4 (-)
Calibrated Probs	82.0 (-)	**82.8 (-)**	92.9 (-)	91.4 (-)

and 91.4% for a standard-standard ensemble), and an average OOD accuracy of 78.7% (vs. 78.2% for a robust-robust ensemble and 70.8% for a standard-standard ensemble). We show per-dataset results in Table 6 (ID) and Table 7 (OOD). We show per-dataset results both ID (Table 6) and OOD (Table 7).

B.3 PER-DATASET RESULTS ON CALIBRATION AND CONFIDENCE

Relative confidence can be incorrect. We measure the confidence of a model f on a distribution P as $\text{conf}(f, P) = \mathbb{E}_{x\sim P}[\max_i f(x)_i]$. Even if the models are not calibrated OOD, one intuitive intuition for why calibrated ensembles work is that that robust model has higher confidence OOD, so that the ensemble primarily uses the (more accurate) robust model's predictions OOD. However, on the remote sensing dataset Landcover we find that the robust model is 6% *less confident* on OOD data than the standard model even though the robust model is 5% *more accurate* OOD than the standard model. Interestingly, calibrated ensembles are able to combine the models in a more fine-grained way to get the best of both worlds, which is captured in our stylized setting in Section 4. We show the average confidence of the standard and robust models for each dataset ID (Table 10) and OOD (Table 11).

Per-dataset results for ECE. In Section **??**, we talked about the ECE of the standard and robust models *after calibrating on ID data*. Here we show the results for each dataset ID (Table 8) and OOD (Table 9). We also show the ECE of the standard and robust models *before calibrating on ID data*, on ID (Table 12) and on OOD (Table 13).

Table 5: *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	CIFAR10	Liv17	Land	Crop	CelebA
Logits	**64.9 (0.3)**	75.7 (1.2)	87.3 (0.2)	**81.8 (0.4)**	**60.5 (0.8)**	**90.9 (0.2)**	**76.9 (0.9)**
Probs	64.6 (0.4)	78.7 (1.3)	87.2 (0.2)	**81.8 (0.4)**	59.5 (1.0)	**90.9 (0.2)**	**76.9 (0.9)**
Tuned Logits	**64.6 (0.6)**	86.3 (0.6)	85.7 (0.9)	80.8 (0.7)	58.7 (1.2)	**87.3 (5.7)**	**77.5 (1.3)**
Tuned Probs	62.8 (0.7)	**86.9 (0.2)**	85.0 (1.3)	81.6 (0.5)	58.7 (2.2)	**86.8 (5.5)**	**77.6 (1.7)**
Calibrated Logits	**65.0 (0.4)**	84.4 (0.3)	**87.5 (0.2)**	**82.0 (0.4)**	**61.2 (0.8)**	**91.3 (0.8)**	**77.6 (1.2)**
Calibrated Probs	**64.7 (0.5)**	86.1 (0.2)	87.3 (0.2)	**82.2 (0.6)**	**60.8 (0.8)**	**91.3 (0.8)**	**77.6 (1.2)**

	ImNet-R	ImNet-V2	ImNet-Sk	MNLI	Waterbird	CivilComments
Logits	73.1 (-)	**73.7 (-)**	52.1 (-)	73.1 (-)	66.9 (-)	**76.0 (-)**
Probs	77.5 (-)	73.4 (-)	52.0 (-)	72.4 (-)	66.9 (-)	**76.0 (-)**
Tuned Logits	64.7 (-)	**73.6 (-)**	47.9 (-)	68.0 (-)	**88.1 (-)**	60.3 (-)
Tuned Probs	64.0 (-)	72.6 (-)	45.5 (-)	69.4 (-)	**88.1 (-)**	61.5 (-)
Calibrated Logits	73.7 (-)	**73.6 (-)**	**52.3 (-)**	**73.6 (-)**	81.1 (-)	71.8 (-)
Calibrated Probs	**77.9 (-)**	73.2 (-)	**52.3 (-)**	73.2 (-)	81.1 (-)	71.8 (-)

Table 6: *ID* accuracies: Calibrated ensembles (one standard and one robust model) achieve comparable or better performance to Standard ensembles (ensemble of two calibrated standard models) and Robust ensembles (ensemble of two calibrated robust models).

	Ent30	DomNet	CIFAR10	Liv17	Land	Crop	CelebA
Std Ensemble	**94.0 (0.0)**	86.3 (0.4)	**97.7 (0.1)**	**97.0 (0.3)**	**77.9 (0.1)**	95.3 (0.0)	91.7 (0.4)
Rob Ensemble	90.9 (0.2)	89.3 (0.3)	92.0 (0.0)	**97.1 (0.1)**	73.4 (0.2)	95.1 (0.1)	**95.2 (0.4)**
Cal ensemble	93.7 (0.1)	**91.2 (0.7)**	97.2 (0.1)	**97.2 (0.2)**	77.2 (0.2)	**95.6 (0.1)**	94.5 (0.5)

Table 7: *OOD* accuracies: Calibrated ensembles (one standard and one robust model) achieve comparable or better performance to Standard ensembles (ensemble of two calibrated standard models) and Robust ensembles (ensemble of two calibrated robust models).

	Ent30	DomNet	CIFAR10	Liv17	Land	Crop	CelebA
Std Ensemble	61.7 (0.2)	57.9 (0.2)	83.5 (0.2)	78.6 (0.4)	57.5 (0.7)	**85.6 (5.8)**	73.7 (1.1)
Rob Ensemble	63.8 (0.4)	**87.5 (0.1)**	85.1 (0.1)	**82.4 (0.1)**	**60.5 (1.4)**	89.8 (0.4)	**78.0 (0.6)**
Cal ensemble	**64.7 (0.5)**	86.1 (0.2)	**87.3 (0.2)**	**82.2 (0.6)**	**60.8 (0.8)**	**91.3 (0.8)**	**77.6 (1.2)**

Table 8: *ID* ECE: The expected calibration error (ECE) of the standard and robust models on ID test data, after post-calibration in ID validation data. The ID calibration errors are low—note that we only use 500 examples to temperature scale, so for ImageNet we have fewer examples than classes for post-calibration, but the models are still fairly well calibrated.

	Ent30	DomNet	CIFAR10	Liv17	Land	Crop	CelebA
Cal. Standard	0.7 (0.1)	2.0 (0.3)	0.8 (0.2)	1.3 (0.2)	1.1 (0.5)	1.4 (0.3)	2.7 (0.4)
Cal. Robust	1.1 (0.4)	2.2 (0.2)	1.3 (0.2)	1.8 (0.0)	1.7 (0.3)	3.5 (0.2)	1.2 (0.3)

	ImageNet	MNLI	Waterbirds	CivilComments
Cal. Standard	1.2 (-)	2.2 (-)	1.2 (-)	1.2 (-)
Cal. Robust	2.3 (-)	2.5 (-)	0.5 (-)	8.1 (-)

Table 9: *OOD* ECE: The expected calibration error (ECE) of the standard and robust models on OOD test data, after calibrating on ID validation data. The calibration errors here are high, especially compared to the ID calibration errors in Table 8.

	Ent30	DomNet	CIFAR10	Liv17	Land	Crop	CelebA
Cal. Standard	15.4 (0.8)	13.6 (1.5)	5.6 (1.1)	11.4 (0.3)	16.4 (0.8)	7.4 (4.8)	11.5 (1.0)
Cal. Robust	14.3 (1.5)	5.5 (0.5)	8.2 (0.0)	8.7 (0.2)	6.5 (1.1)	5.0 (0.3)	14.0 (1.4)

	ImNet-R	ImNet-V2	ImNet-Sk	MNLI	Waterbird	CivilComments
Cal. Standard	5.4 (-)	4.0 (-)	10.1 (-)	13.2 (-)	17.7 (-)	23.3 (-)
Cal. Robust	4.0 (-)	4.9 (-)	5.1 (-)	4.2 (-)	5.5 (-)	6.3 (-)

Table 10: *ID* Confidences: The confidence of the standard and robust models on ID test data (after calibrating on ID data). The standard model is typically more confidence than the robust model, which is reasonable since the standard model is also typically more accurate. There are a few exceptions such as DomainNet, CelebA, and WaterBirds where the standard model is less confident than the robust model, but the standard model is also less accurate in these cases, so this is also reasonable.

	Ent30	DomNet	CIFAR10	Liv17	Land	Crop	CelebA
Cal. Standard	93.1 (0.3)	83.7 (0.4)	96.9 (0.6)	97.0 (0.2)	76.5 (0.9)	95.5 (0.4)	91.7 (0.6)
Cal. Robust	89.9 (0.4)	89.6 (0.1)	91.0 (0.1)	96.0 (0.1)	71.3 (0.5)	94.9 (0.5)	94.7 (0.2)

	ImageNet	MNLI	Waterbirds	CivilComments
Cal. Standard	82.1 (-)	82.6 (-)	87.9 (-)	93.6 (-)
Cal. Robust	68.1 (-)	81.9 (-)	93.2 (-)	87.0 (-)

Table 11: *OOD* Confidences. The confidence of the standard and robust models on OOD test data (after calibrating on ID data). The robust model is usually more confident than the standard model, which is reasonable since the robust model is also typically more accurate. However, Landcover is a noticable exception: the robust model is less confident OOD, even though it is more accurate (see Table 3).

	Ent30	DomNet	CIFAR10	Liv17	Land	Crop	CelebA
Cal. Standard	76.1 (0.8)	68.9 (1.5)	87.8 (1.2)	89.2 (0.5)	72.0 (1.9)	92.8 (1.0)	85.5 (1.5)
Cal. Robust	77.5 (0.4)	92.6 (0.4)	93.3 (0.1)	90.8 (0.2)	66.0 (0.6)	94.1 (0.4)	90.1 (0.1)

	ImNet-R	ImNet-V2	ImNet-Sk	MNLI	Waterbird	CivilComments
Cal. Standard	57.8 (-)	75.5 (-)	50.6 (-)	77.0 (-)	78.1 (-)	80.1 (-)
Cal. Robust	74.0 (-)	64.2 (-)	53.2 (-)	79.7 (-)	92.5 (-)	80.4 (-)

Table 12: *ID* ECE. The expected calibration error (ECE) of the standard and robust models on ID test data, *before calibration* (the key difference from Table 8 is that this is before calibration). We can see that calibration on ID substantially reduces the ECE on ID data (see Table 8)

	Ent30	DomNet	CIFAR10	Liv17	Land	Crop	CelebA
Standard	1.0 (0.1)	8.5 (0.7)	1.2 (0.1)	1.2 (0.1)	6.7 (1.2)	1.5 (0.3)	5.9 (0.5)
Robust	1.1 (0.3)	5.8 (1.3)	1.1 (0.2)	3.4 (0.4)	1.3 (0.1)	3.5 (0.1)	1.8 (0.2)

	ImageNet	MNLI	Waterbirds	CivilComments
Standard	2.2 (-)	9.0 (-)	8.2 (-)	3.7 (-)
Robust	2.4 (-)	8.2 (-)	14.8 (-)	10.2 (-)

Table 13: *OOD* ECE: The expected calibration error (ECE) of the standard and robust models on OOD test data, *before calibration* (the key difference from Table 9 is that this is before calibration). The calibration errors here are higher than the ID calibration errors in Table 12. Comparing with Table 9 (which is after calibration on ID data), we see that calibrating ID does help OOD calibration a little, although the models still remain miscalibrated OOD.

	Ent30	DomNet	CIFAR10	Liv17	Land	Crop	CelebA
Standard	19.1 (0.3)	29.5 (0.5)	10.1 (0.3)	11.7 (0.4)	24.7 (1.5)	8.3 (4.3)	17.6 (0.5)
Robust	14.3 (1.6)	1.8 (0.8)	8.4 (0.3)	6.8 (0.2)	7.1 (1.3)	8.4 (0.7)	12.7 (0.7)

	ImNet-R	ImNet-V2	ImNet-Sk	MNLI	Waterbird	CivilComments
Standard	7.9 (-)	6.1 (-)	13.3 (-)	22.7 (-)	31.8 (-)	30.0 (-)
Robust	3.9 (-)	5.2 (-)	5.2 (-)	10.3 (-)	10.4 (-)	9.9 (-)