# Seasoning Model Soups for Robustness to Adversarial and Natural Distribution Shifts

Francesco Croce* University of Tübingen
Sylvestre-Alvise Rebuffi DeepMind
Evan Shelhamer DeepMind
Sven Gowal DeepMind

## Abstract

*Adversarial training is widely used to make classifiers robust to a specific threat or adversary, such as $\ell_p$-norm bounded perturbations of a given $p$-norm. However, existing methods for training classifiers robust to multiple threats require knowledge of all attacks during training and remain vulnerable to unseen distribution shifts. In this work, we describe how to obtain adversarially-robust* model soups *(i.e., linear combinations of parameters) that smoothly trade-off robustness to different $\ell_p$-norm bounded adversaries. We demonstrate that such soups allow us to control the type and level of robustness, and can achieve robustness to all threats without jointly training on all of them. In some cases, the resulting model soups are more robust to a given $\ell_p$-norm adversary than the constituent model specialized against that same adversary. Finally, we show that adversarially-robust model soups can be a viable tool to adapt to distribution shifts from a few examples.*

## 1. Introduction

Deep networks have achieved great success on several computer vision tasks and have even reached super-human accuracy [19, 31]. However, the outputs of such models are often brittle, and tend to perform poorly on inputs that differ from the distribution of inputs at training time, in a condition known as *distribution shift* [37]. Adversarial perturbations are a prominent example of this condition: small, even imperceptible, changes to images can alter predictions to cause errors [2, 46]. In addition to adversarial inputs, it has been noted that even natural shifts, e.g. different weather conditions, can significantly reduce the accuracy of even the best vision models [13, 21, 40]. Such drops in accuracy are undesirable for robust deployment, and so a lot of effort has been invested in correcting them. Adversarial training [34] and its extensions [15, 39, 58] are currently the most effective methods to improve empirical robustness to adversarial attacks. Similarly, data augmentation is the basis of several techniques that improve robustness to non-adversarial/natural shifts [3, 11, 22]. While significant progress has been made on defending against a specific, selected type of perturbations (whether adversarial or natural), it is still challenging to make a single model robust to a broad set of threats and shifts. For example, a classifier adversarially-trained for robustness to $\ell_p$-norm bounded attacks is still vulnerable to attacks in other $\ell_q$-threat models [29, 47]. Moreover, methods for simultaneous robustness to multiple attacks require jointly training on all [33, 35] or a subset of them [8]. Most importantly, controlling the trade-off between different types of robustness (and nominal performance) remains difficult and requires training several classifiers.

Inspired by *model soups* [52], which interpolate the parameters of a set of vision models to achieve state-of-the-art accuracy on IMAGENET, we investigate the effects of interpolating robust image classifiers. We complement their original recipe for soups by own study of how to pre-train, fine-tune, and combine the parameters of models adversarially-trained against $\ell_p$-norm bounded attacks for different $p$-norms. To create models for soups, we pre-train a single robust model and fine-tune it to the target threat models (using the efficient technique of [8]). We then establish that it is possible to smoothly trade-off robustness to different threat models by moving in the convex hull of the parameters of each robust classifier, while achieving competitive performance with methods that train on multiple $p$-norm adversaries simultaneously. Unlike alternatives, our soups can uniquely *(1)* choose the level of robustness to each threat model without any further training and *(2)* quickly adapt to new unseen attacks or shifts by simply tuning the weighting of the soup.

Previous works [24, 30, 54] have shown that adversarial training with $\ell_p$-norm bounded attacks can help to improve performance on natural shifts if carefully tuned. We show that model soups of diverse classifiers, with different types of robustness, offer greater flexibility for finding models that perform well across various shifts, such as IMAGENET variants. Furthermore, we show that a limited number of images of the new distribution are sufficient to select the

*Work done during an internship at DeepMind.

weights of such a model soup. Examining the composition of the best soups brings insights about which features are important for each dataset and shift. Finally, while the capability of selecting a model specific to each image distribution is a main point of our model soups, we also show that it is possible to jointly select a soup for average performance across several IMAGENET variants to achieve better accuracy than adversarial and self-supervised baselines [18, 24].

**Contributions.** In summary, we show that soups

- can merge nominal and $\ell_p$-robust models (for various $p$): efficient fine-tuning from one robust model obtains a set of models with diverse robustness [8] and compatible parameters for creating model soups [52] (Sec. 3),
- can control the level of robustness to each threat model and achieve, without more training, competitive performance against multi-norm robustness training (Sec. 4),
- are not limited to interpolation, but can find more effective classifiers by extrapolation (Sec. 4.3),
- enable adaptation to unseen distribution shifts on only a few examples (Sec. 5).

## 2. Related Work

**Adversarial robustness to multiple threat models.** Most methods focus on achieving robustness in a single threat model, i.e. to a specific type of attacks used during training. However, this is not sufficient to obtain robustness to unseen attacks. As a result, further work aims to train classifiers for simultaneous robustness to multiple attacks, and the most popular scenario considers a set of $\ell_p$-norm bounded perturbations. The most successful methods [33, 35, 47] are based on adversarial training and differ in how the multiple threats are combined. Notably, all need to use every attack at training time. To reduce the computational cost of obtaining multiply-robust models, one can fine-tune a singly-robust model by one of the above mentioned methods [8], even for only a small number of epochs. Our soups are more flexible by skipping simultaneous adversarial training across multiple attacks.

**Adversarial training for distribution shifts.** While robustness to adversarial attacks and natural shifts are not the same goal, previous works nevertheless show that it is possible to leverage adversarial training with $\ell_p$-norm bounded attacks to improve performance on the common corruptions of [21]. First, AdvProp [54] co-trains models on clean and adversarial images (in the $\ell_\infty$-threat model) with dual normalization layers that specialize to each type of input. The clean branch of the dual model achieves higher accuracy than nominal training on IMAGENET and its variants. Similar results are obtained by Pyramid-AT [24] by its design of a specific attack to adversarially train vision transformers. Finally, [30] carefully selecting the size of the adversarial perturbations, i.e. their $\ell_\infty$- or $\ell_2$-norm, for standard adversarial training [34] achieves competitive performance on common corruptions on CIFAR-10 and IMAGENET-100.

**Model soups.** Ensembling or averaging the parameters of intermediate models found during training is an effective technique to improve both clean accuracy [25, 28] and robustness [14, 39]. Recently, [52] propose *model soups* which interpolate the parameters of networks fine-tuned with different hyperparameters configurations from the same pre-trained model. This yields improved classification accuracy on IMAGENET. Along the same line, [27] fine-tune a model trained on IMAGENET on several new image classification datasets, and show that interpolating the original and fine-tuned parameters yields classifiers that perform well on all tasks. Additional related work is discussed in App. A.

## 3. Model Interpolation across Different Tasks

In the following, we formally introduce the two main components of our procedure to merge adversarially robust models: *(1)* obtaining models which can interpolated by fine-tuning a single $\ell_p$-robust classifiers, and *(2)* interpolation of their weights to balance their different types of robustness. We highlight that our setup diverges from that of prior works about parameters averaging: in fact, both [27, 52] combine models fine-tuned on the same task, i.e. achieving high classification accuracy of unperturbed images, either on a fixed dataset and different hyperparameter configurations [52], or varying datasets [27]. In our case, the individual models are trained for robustness to different types of attacks, i.e. with distinct loss functions.

### 3.1. Adversarial training and fine-tuning

Let us denote $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_i$ the training set, with $\boldsymbol{x}_i \in \mathbb{R}^d$ indicating an image and $y_i \in \{1, \dots, K\}$ the corresponding label, and $\Delta : \mathbb{R}^d \to \mathbb{R}^d$ the function which characterizes a threat model, that maps an input $\boldsymbol{x}$ to a set $\Delta(\boldsymbol{x}) \subset \mathbb{R}^d$ of possible perturbed versions of the original image. For example, $\ell_p$-norm bounded adversarial attacks with budget $\epsilon > 0$ in the image space can be described by

$$\Delta(\boldsymbol{x}) = \{\boldsymbol{\delta} \in \mathbb{R}^d : \|\boldsymbol{\delta}\|_p \leq \epsilon, \boldsymbol{x} + \boldsymbol{\delta} \in [0, 1]^d\}. \quad (1)$$

Then, one can train a classifier $f : \boldsymbol{\theta} \times \mathbb{R}^d \to \mathbb{R}^K$ parameterized by $\boldsymbol{\theta} \in \Theta$ with adversarial training [34] by solving

$$\min_{\boldsymbol{\theta} \in \Theta} \sum_{(\boldsymbol{x}, y) \in \mathcal{D}} \max_{\boldsymbol{\delta} \in \Delta(\boldsymbol{x})} \mathcal{L}(f(\boldsymbol{\theta}, x + \boldsymbol{\delta}), y), \quad (2)$$

for a given loss function $\mathcal{L} : \mathbb{R}^K \times \mathbb{R}^K \to \mathbb{R}$ (e.g., cross-entropy), with the goal of obtaining a model robust to the perturbations described by $\Delta$. Note that this boils down to nominal training when $\Delta(\cdot) = \{\mathbf{0}\}$ and no perturbation is applied on the training images. We are interested in the case where multiple threat models are available, and indicate

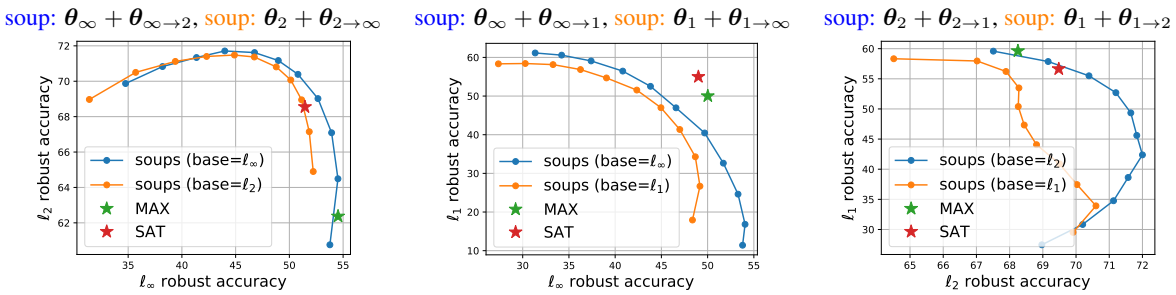


Figure 1. **Soups of two models on CIFAR-10:** for all pairs $(p, q)$ we show the $\ell_p$- vs $\ell_q$-robust accuracy of the soups $w\cdot\boldsymbol{\theta}_p+(1-w)\cdot\boldsymbol{\theta}_{p\to q}$ and $w\cdot\boldsymbol{\theta}_q+(1-w)\cdot\boldsymbol{\theta}_{q\to p}$ varying $w\in[0,1]$. We also show results for MAX and SAT with simultaneous use of the two threat models.

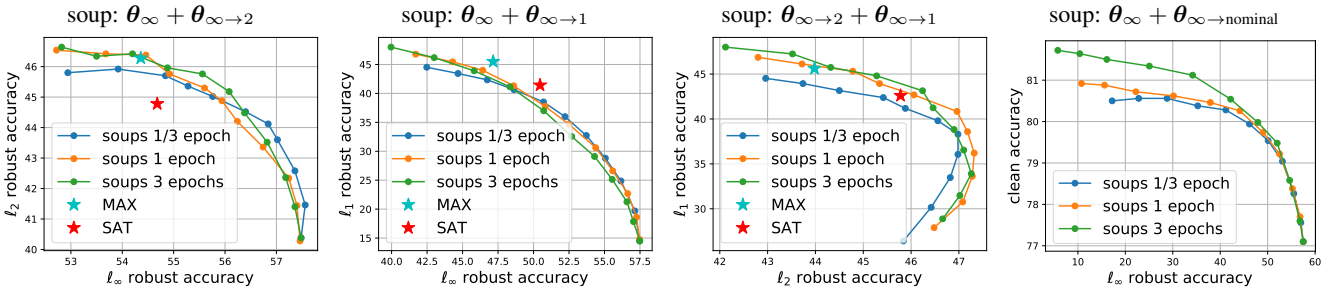


Figure 2. **Soups of two models on IMAGENET:** for $p\in\{2,1\}$ we show the $\ell_p$- vs $\ell_\infty$-robust accuracy of the soups $w\cdot\boldsymbol{\theta}_\infty+(1-w)\cdot\boldsymbol{\theta}_{\infty\to p}$ varying $w\in[0,1]$ (first and second columns). Moreover, we show the soups obtained combining $\boldsymbol{\theta}_{\infty\to 2}$ and $\boldsymbol{\theta}_{\infty\to 1}$ (third), or $\boldsymbol{\theta}_\infty$ and $\boldsymbol{\theta}_{\infty\to\text{nominal}}$ (fourth). For the case of two $\ell_p$-threat models, we also show the results of fine-tuning models with MAX and SAT.

with $\Delta_{\text{nominal}}$ nominal training, and $\Delta_p$ for $p\in\{\infty,2,1\}$ the perturbations with bounded $\ell_p$-norm as described in Eq. 1. We focus on such tasks since they are the most common choices for adversarial defenses, in particular by methods focusing on multiple norm robustness [35, 47].

Notably, it is possible to efficiently obtain a model robust to $\Delta_q$ by fine-tuning for a single (or few) epoch with adversarial training w.r.t. $\Delta_q$ a classifier pre-trained to be robust in $\Delta_p$, with $p\neq q$ and $p,q\in\{\infty,2,1\}$ [8]. For example, in this way one can efficiently derive specialized classifiers $\Delta_2, \Delta_1, \Delta_{\text{nominal}}$ for each task from a single model $f$ robust w.r.t. $\ell_\infty$. However, this does not work well when a nominal classifier is used as starting point for the short fine-tuning. In the following, we denote with $\boldsymbol{\theta}_{p\to q}$ the parameters resulting from fine-tuning to $\Delta_q$ a base model $\boldsymbol{\theta}_p$.

### 3.2. Merging different types of robustness via linear combinations in parameter space

We want to explore the properties of the models obtained by taking linear combinations of the parameters of classifiers with different types of robustness. To do so, there needs to be a correspondence among the parameters of different individual networks: [52] achieve this by merging differently fine-tuned versions of the same pre-trained model. As mentioned above, fine-tuning an $\ell_p$-robust classifier allows to change it to achieve robustness in a new threat model [8]: we exploit such property to create *model soups*, as named by [52]. Formally, we create a model soup from $n$ individual networks with parameters $\boldsymbol{\theta}^1,\dots,\boldsymbol{\theta}^n$ with weights $\boldsymbol{w}=(w_1,\dots,w_n)\in\mathbb{R}^n$ as

$$\boldsymbol{\theta}^{\boldsymbol{w}}=\sum_{i=1}^{n} w_i\cdot\boldsymbol{\theta}^i, \tag{3}$$

and the corresponding classifier is given by $f(\boldsymbol{\theta}^{\boldsymbol{w}},\cdot):\mathbb{R}^d\to\mathbb{R}^K$. While any choice of $\boldsymbol{w}$ is possible, we focus on the case of affine combinations, i.e. $\sum_i w_i=1$. Moreover, we consider soups which are either convex combinations of the individual models, with $w_1,\dots,w_n\geq 0$, or obtained by extrapolations, i.e. with negative elements in $\boldsymbol{w}$.

## 4. Soups for $\ell_p$-robustness

We measure adversarial robustness in the $\ell_p$-threat model with bounds $\epsilon_p$: on CIFAR-10 we use $\epsilon_\infty=8/255$, $\epsilon_2=128/255$, $\epsilon_1=12$, on IMAGENET $\epsilon_\infty=4/255$, $\epsilon_2=4$, $\epsilon_1=255$. If not specified otherwise, we use the full test set for CIFAR-10 and 5000 images from the IMAGENET validation set, and attack by AUTOPGD [6, 7] with 40 steps. More details are provided in App. B.

### 4.1. Soups with two threat models

**CIFAR-10.** We explore the effect of interpolating two models robust to different $\ell_p$-norm bounded attacks for $p \in \{\infty, 2, 1\}$. We consider classifiers with WIDERESNET-28-10 [56] architecture trained on CIFAR-10. For every threat model, we first train a robust classifier with adversarial training from random initialization. Then, we fine-tune the resulting model with adversarial training on each of the other threat models for 10 epochs. In Fig. 1 we show, for each pair of threat models $(\Delta_p, \Delta_q)$, the trade-off of robust accuracy w.r.t. $\ell_p$ and $\ell_q$ for the soups

$$w \cdot \boldsymbol{\theta}_p + (1 - w) \cdot \boldsymbol{\theta}_{p \to q} \quad \text{for} \quad w \in [0, 1]$$

and symmetrically

$$w \cdot \boldsymbol{\theta}_q + (1 - w) \cdot \boldsymbol{\theta}_{q \to p} \quad \text{for} \quad w \in [0, 1].$$

Interpolating the parameters of models trained with a single $\ell_p$-norm controls the balance between the two types of robustness: for example, Fig. 1 (middle plot) shows that moving from $\boldsymbol{\theta}_\infty$ to $\boldsymbol{\theta}_{\infty \to 1}$ (blue curve), i.e. decreasing $w$ from 1 to 0 in the corresponding soup, progressively reduces the robust accuracy w.r.t. $\ell_\infty$ to improve robustness w.r.t. $\ell_1$. Moreover, for similar threat models (i.e. the pairs $(\ell_2, \ell_1)$ and $(\ell_2, \ell_\infty)$) some intermediate networks are more robust than the extremes trained specifically for each threat model.

**IMAGENET.** We now fine-tune a VIT-B16 [10] robust w.r.t. $\ell_\infty$ on IMAGENET to the other threat models, including nominal training, for either 1/3, 1 or 3 epochs. Fig. 2 shows that interpolation of parameters is effective even in this setup, and allows to easily balance nominal and robust accuracy (fourth plot). Moreover, it is possible to create soups with two fine-tuned models, i.e. $\boldsymbol{\theta}_{\infty \to 2}$ and $\boldsymbol{\theta}_{\infty \to 1}$. Finally, increasing the number of fine-tuning steps yields better performance in the target threat model, which in turn generally leads to better soups.

**Comparison to multi-norm robustness methods.** Fig. 1 and Fig. 2 compare the performance of the model soups to that of models trained with methods for robustness in the union of multiple threat models. In particular, we show the results of MAX [47], which computes perturbations for each threat model and trains on that attaining the highest loss, and SAT [33], which samples uniformly at random for each training batch the attack to use. We train models with both methods for all pairs of threat models: as expected, MAX tends to focus on the most challenging threat model, sacrificing some robustness in the other one compared to SAT, since it uses only the strongest attack for each training point. When training for $\ell_\infty$ and $\ell_1$, i.e. the extreme $\ell_p$-norms we consider, MAX and SAT models lie above the front drawn by the model soups, hinting that the more diverse the attacks, the more difficult it is to preserve high robustness to both. In the other cases, both methods behave similarly to the soups. The main advantage given by interpolation is however the option of moving along the front without additional training cost: while one might tune the trade-off between the robustness in the two threat models in SAT, e.g. changing the sampling probability, this would still require training a new classifier for each setup.

### 4.2. Soups with three threat models

We here study the convex combination of three models with different types of robustness. For CIFAR-10 we create soups with each $\ell_p$-robust classifier for $p \in \{\infty, 2, 1\}$ and its fine-tuned version into the other two threat models. We use the same models of the previous section, and sweep the interpolation weights $\boldsymbol{w} \in \mathbb{R}^3$ such that $w_i \in \{0, 0.2, 0.4, 0.6, 0.8, 1\}$ and $\sum_i w_i = 1$. In Fig. 3 and Fig. 10 (in the Appendix) we show clean accuracy (first column) and robust accuracy in $\ell_\infty$, $\ell_2$ and $\ell_1$ (second to fourth columns) and their union, when a point is considered robust only if it is such against all attacks (last column).

One can observe that, independently from the type of robustness of the base model (used as initialization for the fine-tuning), moving in the convex hull of the three parameters (e.g. $\boldsymbol{\theta}_\infty + \boldsymbol{\theta}_{\infty \to 2} + \boldsymbol{\theta}_{\infty \to 1}$ in Fig. 3) allows to smoothly control the trade-off of the different types of robustness. Interestingly, the highest $\ell_2$-robustness is attained by intermediate soups, not by the model specifically fine-tuned w.r.t. $\ell_2$, suggesting that model soups might even be beneficial to robustness in individual threat models (see more below). Moreover, the highest robustness in the union is given by interpolating only the models robust w.r.t. $\ell_\infty$ and $\ell_1$, which is in line with the observation of [8] that training for the extreme norms is sufficient for robustness in the union of the three threat models. Although the robust accuracy in the union is lower than that of training simultaneously with all attacks e.g. with MAX (42.0% vs 47.2%), the model soups deliver competitive results without the need of co-training.

Finally, Fig. 4 shows that similar observations hold on IMAGENET, where we create soups fine-tuning classifiers robust w.r.t. $\ell_\infty$, with either RESNET-50 [19] or VIT-B16 as architecture, for 1 epoch.

### 4.3. Soups for improving individual threat model robustness

We notice in Fig. 1 and Fig. 2 that in a few cases the intermediate models obtain via parameters interpolation have higher robustness than the extreme ones, which are trained or fine-tuned with a specific threat model. As such, we analyze in more details the soups $w \cdot \boldsymbol{\theta}_\infty + (1 - w) \cdot \boldsymbol{\theta}_{\infty \to 2}$, i.e. using the original classifier robust in $\ell_\infty$ and the one fine-tuned to $\ell_2$, on both CIFAR-10 and IMAGENET. Fig. 5 shows the robust accuracy w.r.t. $\ell_\infty$ when varying the value of $w$: in both case the original model $\boldsymbol{\theta}_\infty$, highlighted in red, does not attain the best robustness. Interestingly, on

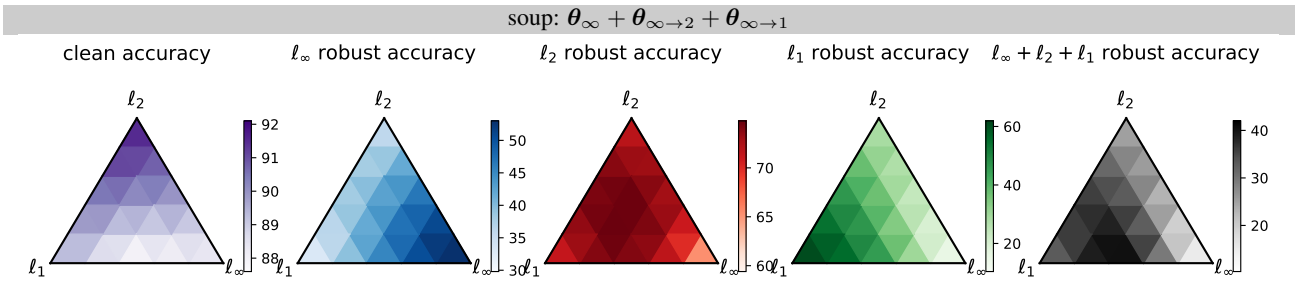


Figure 3. **Soups of three models on CIFAR-10:** we fine-tune the model robust w.r.t. $\ell_\infty$ (with WIDERESNET-28-10 architecture) to the other threat models for 10 epochs, and show clean accuracy (first column) and robust accuracy w.r.t. every threat model (second to fourth columns) and their union (last column) of the soups obtained as convex combinations of the three bases.

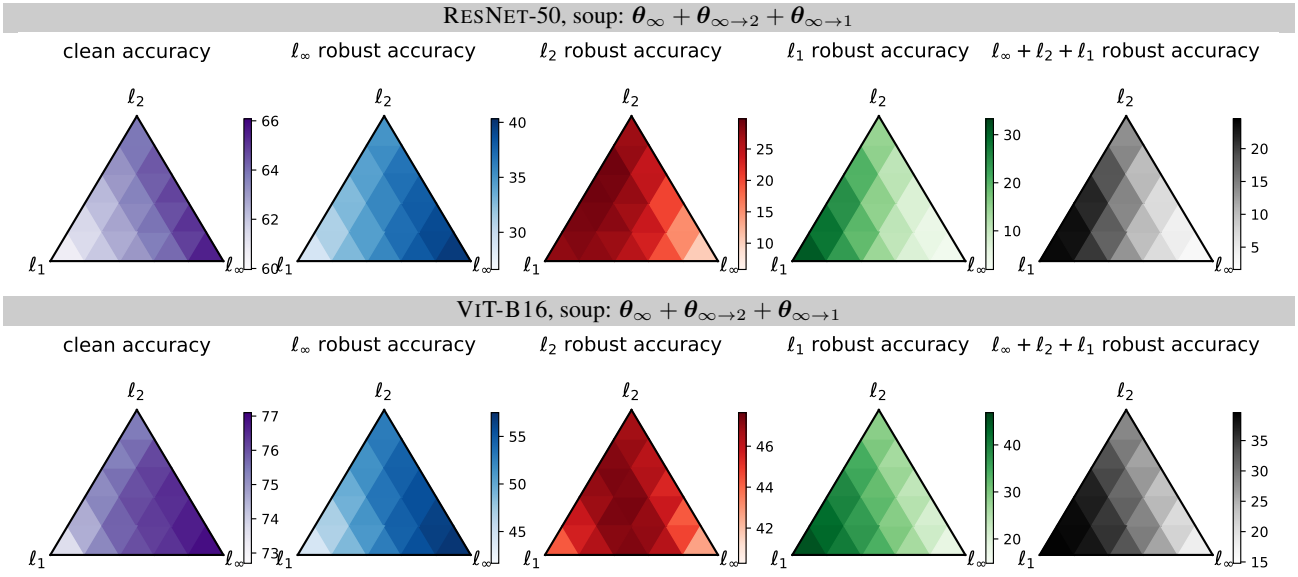


Figure 4. **Soups of three models on IMAGENET:** we fine-tune the classifiers, with RESNET-50 (top row) and VIT-B16 (bottom) as architecture, robust w.r.t. $\ell_\infty$ for 1 epoch to the other threat models, and show clean accuracy (first column) and robust accuracy in every threat model (second to fourth columns) and their union (last column) of the classifiers obtained as convex combinations of the three bases.

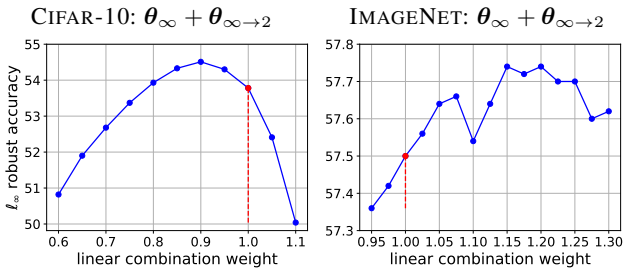


Figure 5. **Improvement on single threat models:** we show the robust accuracy w.r.t. $\ell_\infty$ for the soups $w \cdot \boldsymbol{\theta}_\infty + (1-w) \cdot \boldsymbol{\theta}_{\infty\to 2}$ for varied $w$. The original model $\boldsymbol{\theta}_\infty$ is highlighted in red.

CIFAR-10 the best soup is found with $w = 0.9$, while for IMAGENET with $w > 1$: this suggests that the model soups should not be constrained to the convex hull of the base models, and extrapolation can lead to improvement.

## 5. Soups for Distribution Shifts

Prior works [30] have shown that adversarial training w.r.t. an $\ell_p$-norm is able to provide some improvement in the performance in presence of non-adversarial distribution shifts, e.g. the common corruptions of IMAGENET-C [21]. However, to see such gains it is necessary to carefully select the threat model, for example which $\ell_p$-norm and size $\epsilon$ to bound the perturbations, to use during training. The experiments in Sec. 4 suggest that model soups of nominal and adversarially robust classifiers yield models with a variety of intermediate behaviors, and extrapolation might even deliver models which do not merely trade-off the robust-

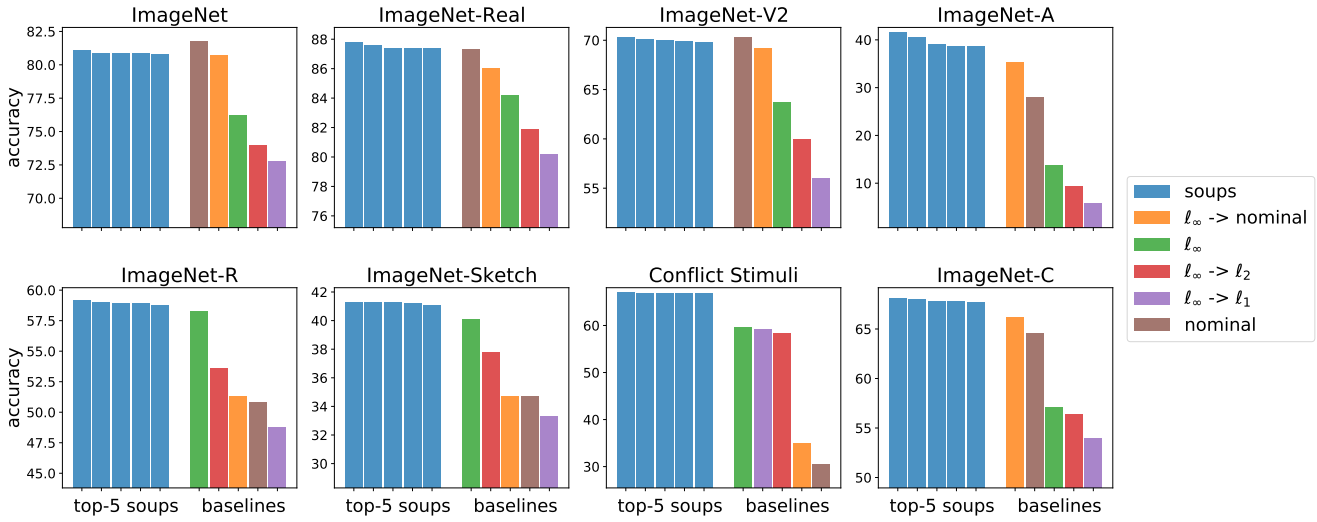


Figure 6. **Soups on IMAGENET variants:** for each dataset we plot the accuracy of the 5 best performing soups of the four base models $\boldsymbol{\theta}_\infty$, $\boldsymbol{\theta}_{\infty\to 2}$, $\boldsymbol{\theta}_{\infty\to 1}$ and $\boldsymbol{\theta}_{\infty\to\text{nominal}}$, and of the individual classifiers. Additionally, we show the performance of an independently trained nominal model. All models are evaluated on the 1000 points used for the grid search of the best soups.

ness of the initial classifiers but amplify it. This flexibility could suit adaptation to various distribution shifts: that is, the various corruption types might more closely resemble the geometry of different $\ell_p$-balls or their union. Moreover, including a nominally fine-tuned model in the soup allows it to maintain, if necessary, high accuracy on the original dataset, which is often degraded by adversarial training [34] or test-time adaptation on shifted data [36].

### 5.1. Soups for IMAGENET variants

**Setup.** In the following, we use models soups consisting of robust VIT fine-tuned from $\ell_\infty$ to the other threat models, and one more VIT nominally fine-tuned for 100 epochs to obtain slightly higher accuracy on clean data. For shifts, we consider several variants of IMAGENET, providing a broad and diverse benchmark for our soups: IMAGENET-REAL [1], IMAGENET-V2 [40], IMAGENET-C [21], IMAGENET-A [23], IMAGENET-R [20], IMAGENET-SKETCH [49], and CONFLICT STIMULI [12]. We consider the setting of few-shot supervised adaptation, with a small set of labelled images from each shift, which we use to select the best soups.

**Soup selection via grid search.** Since evaluating the accuracy of many soups on the entirety of the datasets would be extremely expensive, we search for the best combination of the four models on a random subset of 1000 points from each dataset, with the exception of CONFLICT STIMULI for which all 1280 images are used (for IMAGENET-C we use all corruption types and severities, then aggregate the results). Restricting our search to a subset also serves our aim of finding a model soup which generalizes to the new distribution by only seeing a few examples. We evaluate all the possible affine combinations with weights in the range $[-0.4, 1.4]$ with granularity $0.2$, which amounts to 460 models in total. In Fig. 6 we compare, for each dataset, the accuracy of the 5 best soups to that of each individual classifier used for creating the soups and of a nominal model trained independently: for all datasets apart from IMAGENET the top soups outperform the individual models. Moreover, we notice that the best individual model varies across datasets, indicating that it might be helpful to merge networks with different types of robustness.

**Comparison to existing methods.** Having selected the best soup for each variant (dataset-specific soups) on its chosen few-shot adaptation set, we evaluate the soup on the test set of the variant (results in Table 1). We also evaluate the model soup that attains the best average case accuracy over the adaptation sets for all variants (single soup), in order to gauge the best performance of a single, general model soup. We compare the soups to a nominal model, the $\ell_\infty$-robust classifier used in the soups, their ensemble, the Masked Autoencoders of [18], AdvProp [54], Pyramid-AT [24], and the ensemble obtained by averaging the output (after softmax) of the four models included in the soups. Selecting the best soup on 1000 images of each datasets (results in the last row of Table 1) leads in 4 out of the 8 datasets to the best accuracy, and only slightly suboptimal values in the other cases: in particular, parameters interpolations is very effective on stronger shifts like IMAGENET-R and CONFLICT STIMULI, where it attains almost 8% better performance than the closest baseline. Unsurprisingly, it is more challenging to improve on datasets like IMAGENET-V2 which are very close to the original IMAGENET. Over-

| SETUP | # FP | IMAGENET | IN-REAL | IN-V2 | IN-A | IN-R | IN-SKETCH | CONFLICT STIMULI | IN-C | MEAN |
|---|---|---|---|---|---|---|---|---|---|---|
| **Baselines** | | | | | | | | | | |
| Nominal training | ×1 | 82.64% | 87.33% | 71.42% | 28.03% | 47.94% | 34.43% | 30.47% | 64.45% | 55.84% |
| Adversarial training | ×1 | 76.88% | 83.91% | 64.81% | 12.35% | 55.76% | 40.11% | 59.45% | 55.44% | 56.09% |
| Fine-tuned MAE-B16 | ×1 | 83.10% | 88.02% | 72.80% | 37.92% | 49.30% | 35.69% | 27.81% | 63.23% | 57.23% |
| AdvProp | ×1 | 83.39% | 88.06% | 73.17% | 34.81% | 53.04% | 39.25% | 38.98% | 70.39% | 60.14% |
| Pyramid-AT | ×1 | 83.14% | 87.82% | 72.53% | 32.72% | 51.78% | 38.60% | 37.27% | 67.01% | 58.86% |
| Indep. networks ensemble | ×2 | 82.86% | 87.78% | 71.73% | 25.99% | 54.20% | 37.33% | 46.41% | 65.61% | 58.99% |
| Individual networks ensemble | ×4 | 81.31% | 86.97% | 70.21% | 23.13% | 54.82% | 39.51% | 56.02% | 68.17% | 60.02% |
| **Fixed grid search on 1000 images** | | | | | | | | | | |
| Single soup | ×1 | 82.49% | 87.85% | 71.99% | 34.31% | 53.84% | 39.84% | 38.52% | 66.82% | 59.46% |
| Dataset-specific soups | ×1 | 82.29% | 87.89% | 71.95% | 38.27% | 56.39% | 40.73% | 67.03% | 69.34% | (64.24%) |

Table 1. **Comparison on IMAGENET variants:** we report the classification accuracy of various models on the IMAGENET variants, together with the number of forward passes they require. The soups are selected via a fixed grid search on the interpolation weights with 1000 points for each dataset. The last row shows the results of the best found soup for each dataset.

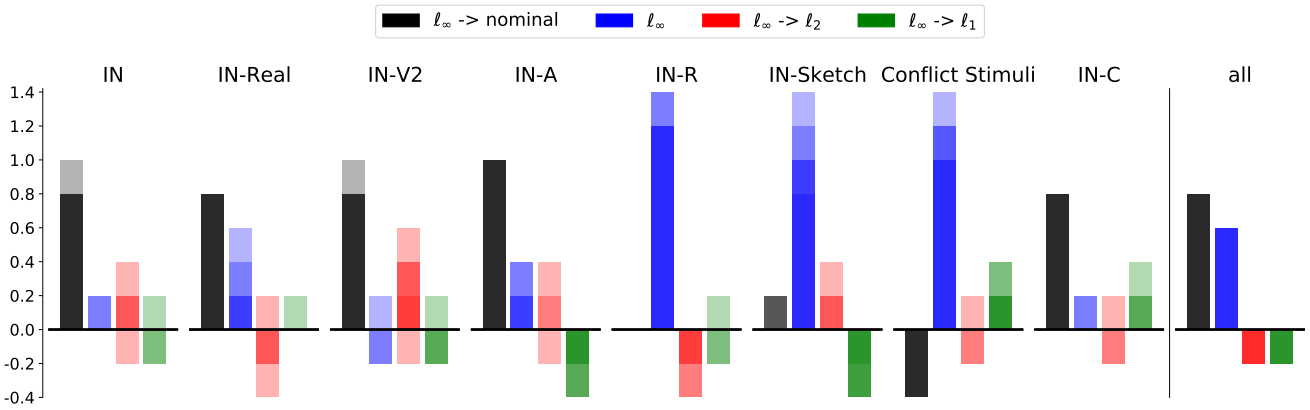


Figure 7. **Soup compositions on IMAGENET variants:** for each dataset we plot the composition of the 5 best soups, i.e. the linear weights for the individual models, as measured by grid search over 1000 points on weights in the range $[-0.4, 1.4]$ with granularity 0.2. Additionally we show the composition of the model achieving the best average accuracy over all 8 variants.

all, The soup selected for best average accuracy (across all datasets) outperforms all baselines, except for the ensemble of four models (with $4\times$ the inference cost), and AdvProp, which requires co-training of clean and adversarial points. These results show that soups with robust classifiers are a promising avenue for quickly adapting to distribution shifts.

**Composition of the soups.** To analyze which types of classifiers are most relevant for performance on every distribution shift, we plot in Fig. 7 the breakdown of the weights of the five best soups (more intense colors indicate that the corresponding weight or a larger one is used more often in the top-5 soups). First, one can see that the nominally fine-tuned model (in black) is dominant, with weights of 0.8 or 1, on IMAGENET, IMAGENET-REAL, IMAGENET-V2, IMAGENET-A and IMAGENET-C: this could be expected since these datasets are closer to IMAGENET itself, i.e. the distribution shift is smaller, which is what nominal training optimizes for (in fact, the nominal models achieve higher accuracy than adversarially trained ones on these datasets in Fig. 6). However, in all cases there is a contribution of some of the $\ell_p$-robust networks. On IMAGENET-R, IMAGENET-SKETCH and CONFLICT STIMULI, the model robust w.r.t. $\ell_\infty$ plays instead the most relevant role, again in line with the results in Table 1. Interestingly, in the case of CONFLICT STIMULI, the nominal classifier has a weight -0.4 (the smallest in the grid search) for all top performing soups: we hypothesize that this has the effect of reduce the texture bias typical of nominal model and emphasize the attention to shapes already important in adversarially trained classifiers. Finally, we show the composition of the soup which has the best average accuracy over all datasets (last column of Fig. 7), where the nominal and $\ell_\infty$-robust models have similar positive weight.

**How many images does it take to find a good soup?** To identify the practical limit of supervision for soup selection, we study the effect of varying the number of labelled images

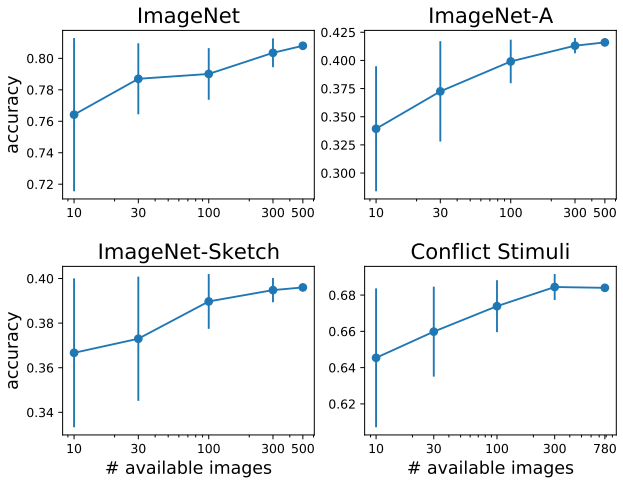


Figure 8. **Soup selection data efficiency:** we vary the number of images for selecting the best performing soups on different datasets. For each case, we plot the average accuracy on a held-out test set, and its standard deviation over 50 trials.

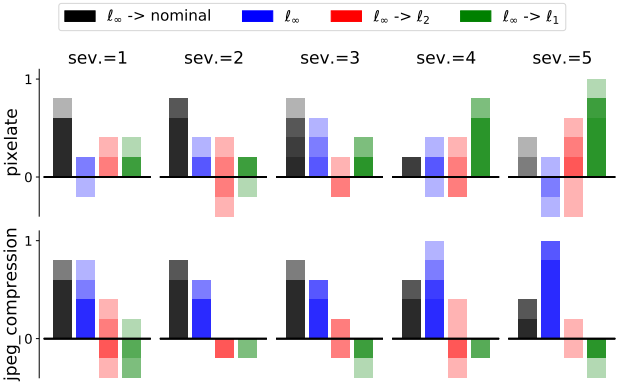


Figure 9. **Soup compositions on IMAGENET-C:** we plot the model-wise weights of the best soups across types and severities.

used to select the best soup on a new dataset. For this analysis we randomly choose 500 images from the adaptation set used for the grid search to create a held-out test set. From the remaining images, we uniformly sample $k$ elements, select the soup which performs best on such $k$ points, and then evaluate it on this test set. We repeat this procedure for $k \in \{10, 30, 100, 300, 500\}$ for 50 times each with different random seeds. In Fig. 8 we plot the average accuracy on the held-out test set, with standard deviation, when varying $k$: increasing the number of points above 100 achieves high test accuracy with limited variance. This suggests that soup selection, and thus model adaptation, can be carried out with as few as 100 examples of the new distribution.

### 5.2. A closer look at IMAGENET-C

While our experiments have considered IMAGENET-C as a single dataset, it consists of 15 corruptions types, each with 5 severity levels. As the various corruptions have different characteristics, one might expect the best soup to vary across them. In Fig. 9 we plot the composition of the top-5 soups for each severity level for two corruption types (as done in Fig. 7). The weights of the individual classifiers significantly change across distribution shifts: for both corruption types, increasing the severity (making perturbations stronger) leads to a reduction in the nominal weight in favor of a robust weight. However, in the case of "pixelate" the soups concentrate on the $\ell_1$-robust network, while for "jpeg compression" this happens for $\ell_\infty$. Similar visualization for the remaining IMAGENET-C subsets are found in Fig. 11 of the Appendix. This highlights the importance of interpolating models with different types of robustness, and implies that considering each corruption type (including severity levels) as independent datasets could further improve the performance of the soups on IMAGENET-C.

## 6. Discussion and Limitations

Merging models with different types of robustness enables strong control of classifier performance by tuning only a few soup weights. Soups can find models which perform well even on distributions unseen during training (e.g. the IMAGENET variants). Moreover, our framework avoids co-training on multiple threats: this makes it possible to fine-tune models with additional attacks as they present themselves, and enrich the soups with them.

At the moment, our soups contain only nominal or $\ell_p$-robust models, but expanding the diversity of models might aid adaptation to new datasets. We selected our soups with few-shot supervision, but other settings could potentially use soups, such as unsupervised domain adaptation [37,42], on labeled clean and unlabeled shifted data, and test-time adaptation [43,44,48], on unlabeled examples alone. Moreover, in our evaluation we have constrained the soups to belong to a fixed grid, which might miss better models: future work could develop automatic schemes to optimize the soup weights, possibly with even fewer examples, or without labeled examples (as done for test-time adaptation of non-robust models).

## 7. Conclusion

We show that combining the parameters of robust classifiers, without additional training, achieves a smooth trade-off of robustness in different $\ell_p$-threat models. This allows us to discover models which perform well on distribution shifts with only a limited number of examples of each shift. In these ways, model soups serve as a good starting point to efficiently adapt classifiers to changes in data distributions.

## References


[1] Lucas Beyer, Olivier J. Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. Are we done with ImageNet? *arXiv preprint arXiv:2006.07159*, 2020. 6

[2] Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Srndic, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion Attacks against Machine Learning at Test Time. *arXiv preprint arXiv:1708.06131*, 2013. 1

[3] Dan A. Calian, Florian Stimberg, Olivia Wiles, Sylvestre-Alvise Rebuffi, Andras Gyorgy, Timothy Mann, and Sven Gowal. Defending Against Image Corruptions Through Adversarial Augmentations. *arXiv preprint arXiv:2104.01086*, 2021. 1

[4] Nicholas Carlini, Florian Tramer, J Zico Kolter, et al. (certified!!) adversarial robustness for free! *arXiv preprint arXiv:2206.10550*, 2022. 12

[5] Jeremy M Cohen, Elan Rosenfeld, and J Zico Kolter. Certified Adversarial Robustness via Randomized Smoothing. *arXiv preprint arXiv:1902.02918*, 2019. 12

[6] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. *arXiv preprint arXiv:2003.01690*, 2020. 3, 12

[7] Francesco Croce and Matthias Hein. Mind the box: $l_1$-APGD for sparse adversarial attacks on image classifiers. *arXiv preprint arXiv:2103.01208*, 2021. 3

[8] Francesco Croce and Matthias Hein. Adversarial robustness against multiple and single $l_p$-threat models via quick fine-tuning of robust classifiers. In *Proceedings of the 39th International Conference on Machine Learning*, pages 4436–4454, 2022. 1, 2, 3, 4, 12, 13

[9] Ekin D. Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V. Le. Randaugment: Practical automated data augmentation with a reduced search space. In *CVPR*, 2020. 12

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv preprint arXiv:2010.11929*, 2020. 4

[11] N Benjamin Erichson, Soon Hoe Lim, Francisco Utrera, Winnie Xu, Ziang Cao, and Michael W Mahoney. NoisyMix: Boosting Robustness by Combining Data Augmentations, Stability Training, and Noise Injections. *arXiv preprint arXiv:2202.01263*, 2022. 1

[12] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2018. 6

[13] Robert Geirhos, Carlos RM Temme, Jonas Rauber, Heiko H Schütt, Matthias Bethge, and Felix A Wichmann. Generalisation in humans and deep neural networks. In *Advances in Neural Information Processing Systems*, pages 7538–7550, 2018. 1

[14] Sven Gowal, Chongli Qin, Jonathan Uesato, Timothy Mann, and Pushmeet Kohli. Uncovering the Limits of Adversarial Training against Norm-Bounded Adversarial Examples. *arXiv preprint arXiv:2010.03593*, 2020. 2

[15] Sven Gowal, Sylvestre-Alvise Rebuffi, Olivia Wiles, Florian Stimberg, Dan Andrei Calian, and Timothy Mann. Improving Robustness using Generated Data. *arXiv preprint arXiv:2110.09468*, 2021. 1

[16] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017. 12

[17] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *ICLR*, 2021. 12

[18] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked Autoencoders Are Scalable Vision Learners. *arXiv preprint arXiv:2111.06377*, 2021. 2, 6, 12

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CVPR*, 2016. 1, 4

[20] 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. 6

[21] Dan Hendrycks and Thomas Dietterich. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. In *International Conference on Learning Representations*, 2018. 1, 2, 5, 6

[22] Dan Hendrycks, Norman Mu, Ekin D. Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. AugMix: A Simple Data Processing Method to Improve Robustness and Uncertainty. *arXiv preprint arXiv:1912.02781*, 2019. 1

[23] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. *arXiv preprint arXiv:1907.07174*, 2019. 6

[24] Charles Herrmann, Kyle Sargent, Lu Jiang, Ramin Zabih, Huiwen Chang, Ce Liu, Dilip Krishnan, and Deqing Sun. Pyramid Adversarial Training Improves ViT Performance. *arXiv preprint arXiv:2111.15121*, 2021. 1, 2, 6

[25] Gao Huang, Yixuan Li, Geoff Pleiss, Zhuang Liu, John E Hopcroft, and Kilian Q Weinberger. Snapshot ensembles: Train 1, get m for free. In *ICLR*, 2017. 2

[26] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In *European conference on computer vision*, pages 646–661. Springer, 2016. 12

[27] Gabriel Ilharco, Mitchell Wortsman, Samir Yitzhak Gadre, Shuran Song, Hannaneh Hajishirzi, Simon Kornblith, Ali Farhadi, and Ludwig Schmidt. Patching open-vocabulary models by interpolating weights. *arXiv preprint arXiv:2208.05592*, 2022. 2, 12

[28] Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging Weights Leads to Wider Optima and Better Generalization. *arXiv preprint arXiv:1803.05407*, 2018. 2

[29] Daniel Kang, Yi Sun, Dan Hendrycks, Tom Brown, and Jacob Steinhardt. Testing Robustness Against Unforeseen Adversaries. *arXiv preprint arXiv:1908.08016*, 2019. 1

[30] Klim Kireev, Maksym Andriushchenko, and Nicolas Flammarion. On the effectiveness of adversarial training against common corruptions. *arXiv preprint arXiv:2103.02325*, 2021. 1, 2, 5

[31] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 1

[32] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 12

[33] Divyam Madaan, Jinwoo Shin, and Sung Ju Hwang. Learning to generate noise for multi-attack robustness. In *International Conference on Machine Learning*, pages 7279–7289. PMLR, 2021. 1, 2, 4

[34] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 1, 2, 6

[35] Pratyush Maini, Eric Wong, and J. Zico Kolter. Adversarial Robustness Against the Union of Multiple Perturbation Models. *arXiv preprint arXiv:1909.04068*, 2019. 1, 2, 3

[36] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *ICML*, 2022. 6

[37] Joaquin Quinonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset shift in machine learning*. 2009. 1, 8

[38] Alexandre Rame, Matthieu Kirchmeyer, Thibaud Rahier, Alain Rakotomamonjy, patrick gallinari, and Matthieu Cord. Diverse weight averaging for out-of-distribution generalization. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. 12

[39] Sylvestre-Alvise Rebuffi, Sven Gowal, Dan A. Calian, Florian Stimberg, Olivia Wiles, and Timothy Mann. Data Augmentation Can Improve Robustness. *arXiv preprint arXiv:2111.05328*, 2021. 1, 2

[40] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet Classifiers Generalize to ImageNet? *arXiv preprint arXiv:1902.10811*, 2019. 1, 6

[41] Leslie Rice, Eric Wong, and J. Zico Kolter. Overfitting in adversarially robust deep learning. *arXiv preprint arXiv:2002.11569*, 2020. 12

[42] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *European conference on computer vision*, pages 213–226. Springer, 2010. 8

[43] Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. Improving robustness against common corruptions by covariate shift adaptation. *Advances in Neural Information Processing Systems*, 33:11539–11551, 2020. 8

[44] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International conference on machine learning*, pages 9229–9248. PMLR, 2020. 8

[45] Sven Gowal, Krishnamurthy Dvijotham, Robert Stanforth, Rudy Bunel, Chongli Qin, Jonathan Uesato, Timothy Mann, and Pushmeet Kohli. On the Effectiveness of Interval Bound Propagation for Training Verifiably Robust Models. *arXiv preprint arXiv:1810.12715*, 2018. 12

[46] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 1

[47] Florian Tramèr and Dan Boneh. Adversarial Training and Robustness for Multiple Perturbations. In *Advances in Neural Information Processing Systems*. 2019. 1, 2, 3, 4

[48] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021. 8

[49] 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*, pages 10506–10518, 2019. 6

[50] Ross Wightman. Pytorch image models. `https://github.com/rwightman/pytorch-image-models`, 2019. 12

[51] Eric Wong and Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *International Conference on Machine Learning*, pages 5283–5292, 2018. 12

[52] Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. *arXiv preprint arXiv:2203.05482*, 2022. 1, 2, 3, 12

[53] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7959–7971, 2022. 12

[54] Cihang Xie, Mingxing Tan, Boqing Gong, Jiang Wang, Alan Yuille, and Quoc V Le. Adversarial Examples Improve Image Recognition. *arXiv preprint arXiv:1911.09665*, 2019. 1, 2, 6

[55] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *ICCV*, 2019. 12

[56] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016. 4

[57] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2018. 12

[58] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P. Xing, Laurent El Ghaoui, and Michael I. Jordan. Theoretically Principled Trade-off between Robustness and Accuracy. *arXiv preprint arXiv:1901.08573*, 2019. 1

[59] Yimeng Zhang, Yuguang Yao, Jinghan Jia, Jinfeng Yi, Mingyi Hong, Shiyu Chang, and Sijia Liu. How to robustify black-box ML models? a zeroth-order optimization perspective. In *International Conference on Learning Representations*, 2022. 12