A. Additional Related Work

Certified defenses. While adversarial training is the main tool for empirical robustness, certified defenses [5,45, 51] can guarantee that no misclassified point exists in a specific perturbation region (*threat model*): this means that the classifier is robust against any attack *algorithm* that generates perturbations in that region. However, we are interested in robustness in different, possibly unseen, threat models (e.g. ℓ_∞- and ℓ_1-balls), identified by distinct perturbation regions. Certified defenses need to know the threat model to be certified, e.g. recent methods [4, 5, 59] certify safety only against a well-defined ℓ_p-bounded attack. Thus, they cannot be adapted to unseen attacks. Furthermore, it is unclear how to extend certified defenses to distribution shifts without strong assumptions about the type and magnitude of such shifts.

Merging models. Besides [27, 52] mentioned above, weight averaging has been recently studies in several works. For example, in [53] the proposed technique first fine-tunes a zero-shot model on a target dataset, then interpolates the weights of original and fine-tuned models, which improves robustness to distribution shifts. Similarly to model soups, in [38] a pre-trained model is fine-tuned with different sets of hyperparameters. The resulting networks are then averaged to improve performance on out-of-distribution data, and achieve SOTA performance on the multi-domain DomainBed benchmark [17]. However, none of such work consider merging adversarially robust models for robustness to adversarial perturbations or distribution shifts.

B. Experimental Details

B.1. Training setup.

CIFAR-10. We train robust models from random initialization for 200 epochs with SGD with momentum as optimizer, an initial learning rate of 0.1 (reduced 10 times at epochs 100 and 150), and a batch size of 128. For fine-tuning, we train for 10 epochs with cosine schedule for the learning rate, with peak value of 0.1 (we only use 0.5 for fine-tuning the model trained w.r.t. ℓ_1 to the ℓ_2-threat model) and linear ramp-up in the first 1/10 of training steps. We generate adversarial perturbations by AUTOPGD with 10 steps. We select checkpoints according to robustness on a validation set as suggested by [41].

IMAGENET. We follow the setup of [18]: for full training, we use 300 epochs, AdamW optimizer [32] with momenta $\beta_1 = 0.9$, $\beta_2 = 0.95$, weight decay of 0.3 and a cosine learning rate decay with base learning rate 10^{-4} (scale as in [16]) and linear ramp-up of 20 epochs, batch size of 4096, label smoothing of 0.1, stochastic depth [26] with base value 0.1 and with a dropping probability linearly increasing with depth. As data augmentation, we use random crops resized to 224 × 224 images, mixup [57], CutMix [55] and RandAugment [9] with two layers, magnitude 9 and a random probability of 0.5. We note that our implementation of RandAugment is based on the version in the `timm` library [50]. For VIT architectures, we adopt exponential moving average with momentum 0.9999. For fine-tuning we keep the same hyperparameters except for reducing the base learning rate from 10^{-4} to 10^{-5} since this leads to better performance in the target threat model. For adversarial training we use AUTOPGD on the KL divergence loss with 2 steps for ℓ_∞- and ℓ_2-norm bounded attacks, 20 steps for ℓ_1 (as it is a more challenging threat model for optimization [8]).

Baselines. For MAX and SAT, we fine-tune the singly-robust models with the same scheme above for the networks used in the model soups. We generate adversarial perturbations with the the same attacks, and use 10 and 1 epoch of fine-tuning in the case of CIFAR-10 and IMAGENET respectively. For the baselines in Table 1, we use the same scheme (except for technique-specific components which follow the original papers). For AdvProp we use dual normalization layers and train with random targeted attacks with bound $\epsilon = 4/255$ on the ℓ_∞-norm.

B.2. Evaluation setup.

Adversarial robustness. As default evaluation we use AUTOPGD with 40 steps and default parameters, with the DLR loss [6] for ℓ_∞- and ℓ_2-attacks, cross entropy for ℓ_1. As a test against stronger attacks, for the evaluation of the robustness of the soups of three threat models on CIFAR-10 (Fig. 3 and Fig. 10) we increase the budget of AUTOPGD to 100 steps and 5 random restarts (in this case we use targeted DLR loss for ℓ_∞ and ℓ_2, and 1000 test points).

Distribution shifts. For all IMAGENET variants we evaluate the classification accuracy on the entire dataset.

C. Additional Experiments

C.1. Soups with three threat models

We show in Fig. 10 the clean accuracy, robust accuracy for each ℓ_p-norm and their union of the soups obtained merging three classifiers. We use either a pre-trained classifier robust w.r.t. ℓ_2 (top row) or ℓ_1 (bottom row) and fine-tune them to the remaining threat models.

C.2. Model soups on IMAGENET variants

Additional baselines. We further report the results on the IMAGENET variants of models trained with MAX and SAT in Table 2: for both we select the variants which use ℓ_∞- and ℓ_1-attacks at training time (see Sec. 4), since those are the two extreme ℓ_p-threat models we consider. MAX and SAT attain 51.36% and 51.75% mean accuracy across distribution shifts, which is significantly worse than the other

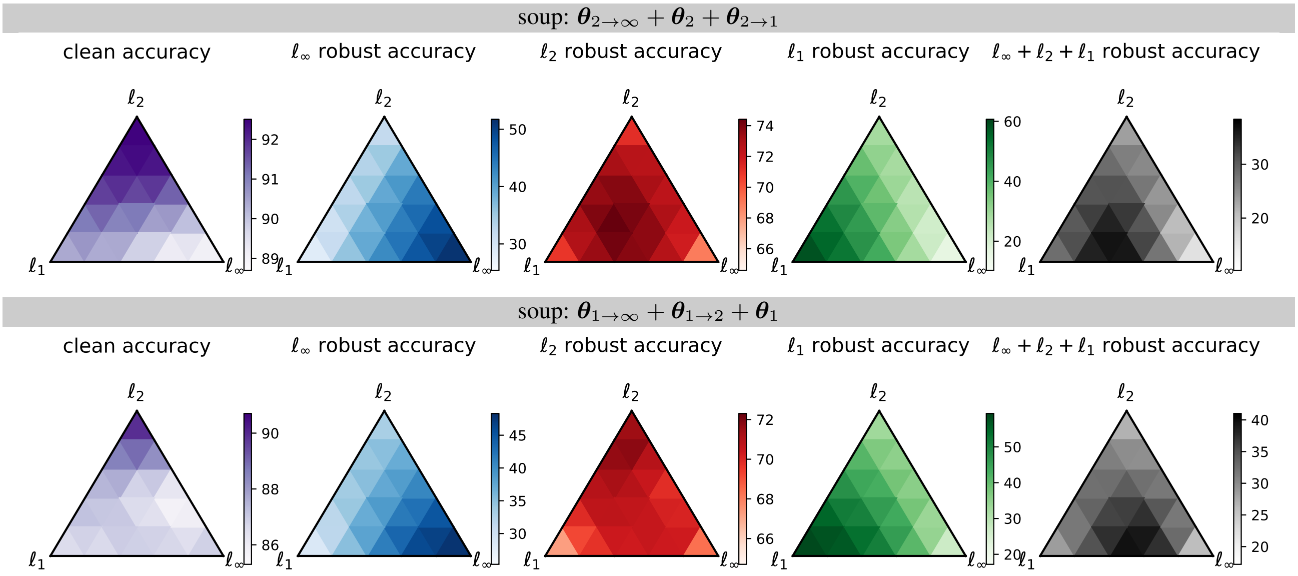

Figure 10. **Soups of three models on CIFAR-10:** we fine-tune each model robust w.r.t. ℓ_p for $p \in \{2, 1\}$ (with WIDERESNET-28-10 architecture) to the other threat models for 10 epochs, and show clean accuracy (first column) and robust accuracy w.r.t. every threat model (second to fourth columns) and their union (last column) of the soups obtained as convex combinations of the three bases.

SETUP	# FP	IMAGENET	IN-REAL	IN-V2	IN-A	IN-R	IN-SKETCH	CONFLICT STIMULI	IN-C	MEAN
MAX	×1	72.08%	79.28%	58.58%	5.65%	47.94%	33.46%	59.84%	54.07 %	51.36%
SAT	×1	73.76%	80.92%	60.53%	7.19%	49.89%	34.51%	59.14%	56.06%	52.75%

Table 2. **Additional baselines:** we report the classification accuracy on the IMAGENET variants of the models trained with MAX and SAT jointly on the ℓ_∞ and ℓ_1-threat models.

reported models. This is mostly due to the low clean accuracy on IMAGENET (72.08% and 73.76%), which strongly influences performance on many distribution shifts (see Table 1 and Fig. 7). In fact, it is known that classifiers trained for multiple norm robustness typically suffer degradation in clean accuracy [8].

Ablation studies. We show additional results for model soups on IMAGENET variants: first, in Table 3 we report the results on the full datasets of the second and third best soups according to the grid search on 1000 points for the shift (we also show the best soup from Table 1 in the main part). When selecting the best soup on average across datasets, all three classifiers have very close performance (59.46% to 59.48%), while the accuracy on the individual datasets may vary e.g. on IMAGENET-A and CONFLICT STIMULI. For the dataset-specific soups the results on the entire datasets respect the ranking given by the grid search (the three values are similar to each other), further suggesting that even a limited number of points can serve to tune a suitable soup.

Second, we study the effect of varying the radii of the ℓ_p-threat models used by the robust classifiers in the soups. In this case, we fine-tune for 1 epoch the original model robust w.r.t. ℓ_∞ at $\epsilon = 4/255$ with adversarial training w.r.t. ℓ_∞ at $\epsilon_\infty \in \{1/255, 2/255\}$, ℓ_2 at $\epsilon_2 \in \{1, 2\}$ ($\epsilon_2 = 4$ above), ℓ_2 at $\epsilon_1 \in \{64, 128\}$ ($\epsilon_1 = 255$ above). In this way we have three sets of four models to create soups, where the nominal one is fixed and the robust ones have radii ϵ_p, $\epsilon_p/2$ and $\epsilon_p/4$ for $p \in \{\infty, 2, 1\}$. Table 4 reports the results of the various sets of models: for the single soup optimized for average performance, the smaller ϵ_p slightly reduce the performance. Looking at the individual datasets, in some cases like IMAGENET-V2 and IMAGENET-C using smaller values of ϵ_p yields some improvements, but it also leads to severe drops on the distribution shifts where having robust models is more relevant like CONFLICT STIMULI and IMAGENET-R. This suggests that it might be useful to have models robust w.r.t. the same ℓ_p-norm but with different radii in the set of the networks used for creating the soups.

C.3. Composition of soups on IMAGENET-C

In Fig. 11 we visualize the composition of the top-5 soups for each corruption type and severity level: one can observe that the weights of the four networks in the soups varies across IMAGENET subset.

Setup	# FP	ImageNet	IN-Real	IN-V2	IN-A	IN-R	IN-Sketch	Conflict Stimuli	IN-C	Mean
Fixed grid search on 1000 images: best soups										
Single soup	$\times 1$	82.49%	87.85%	71.99%	34.31%	53.84%	39.84%	38.52%	66.82%	59.46%
Dataset-specific soups	$\times 1$	82.29%	87.89%	71.95%	38.27%	56.39%	40.73%	67.03%	69.34%	(64.24%)
Fixed grid search on 1000 images: second best soups										
Single soup	$\times 1$	82.62%	87.92%	72.02%	30.99%	53.28%	39.00%	41.25%	68.75%	59.48%
Dataset-specific soups	$\times 1$	82.49%	87.84%	71.94%	37.60%	56.42%	40.76%	66.95%	69.32%	(64.16%)
Fixed grid search on 1000 images: third best soups										
Single soup	$\times 1$	82.67%	87.86%	72.11%	34.25%	53.18%	39.52%	38.52%	67.60%	59.46%
Dataset-specific soups	$\times 1$	82.51%	87.65%	71.58%	37.51%	55.75%	40.65%	66.88%	68.92%	(63.93%)

Table 3. **Top-k model soups for ImageNet variants:** we report the classification accuracy on the ImageNet variants, of the 1st, 2nd and 3d best soups (single or dataset-specific) found by grid search on the interpolation weights with 1000 points for each dataset.

Setup	# FP	ImageNet	IN-Real	IN-V2	IN-A	IN-R	IN-Sketch	Conflict Stimuli	IN-C	Mean
Robust models with standard ϵ_p										
Single soup	$\times 1$	82.49%	87.85%	71.99%	34.31%	53.84%	39.84%	38.52%	66.82%	59.46%
Dataset-specific soups	$\times 1$	82.29%	87.89%	71.95%	38.27%	56.39%	40.73%	67.03%	69.34%	(64.24%)
Robust models with $\epsilon_p/2$										
Single soup	$\times 1$	82.66%	87.78%	72.34%	32.05%	51.40%	37.80%	39.69%	69.06%	59.10%
Dataset-specific soups	$\times 1$	81.85%	87.47%	72.34%	36.25%	53.80%	39.32%	62.19%	69.44%	(62.83%)
Robust models with $\epsilon_p/4$										
Single soup	$\times 1$	81.47%	87.36%	70.84%	26.23%	53.46%	39.13%	46.25%	67.37%	59.01%
Dataset-specific soups	$\times 1$	82.68%	87.72%	72.21%	35.61%	54.00%	39.79%	59.22%	69.49%	(62.59%)

Table 4. **Varying threat models:** we report the classification accuracy on the ImageNet variants of the best single and dataset-specific soups when using various radii ϵ_p for fine-tuning the ℓ_p-robust networks. The soups are selected via a fixed grid search on the interpolation weights with 1000 points for each dataset.

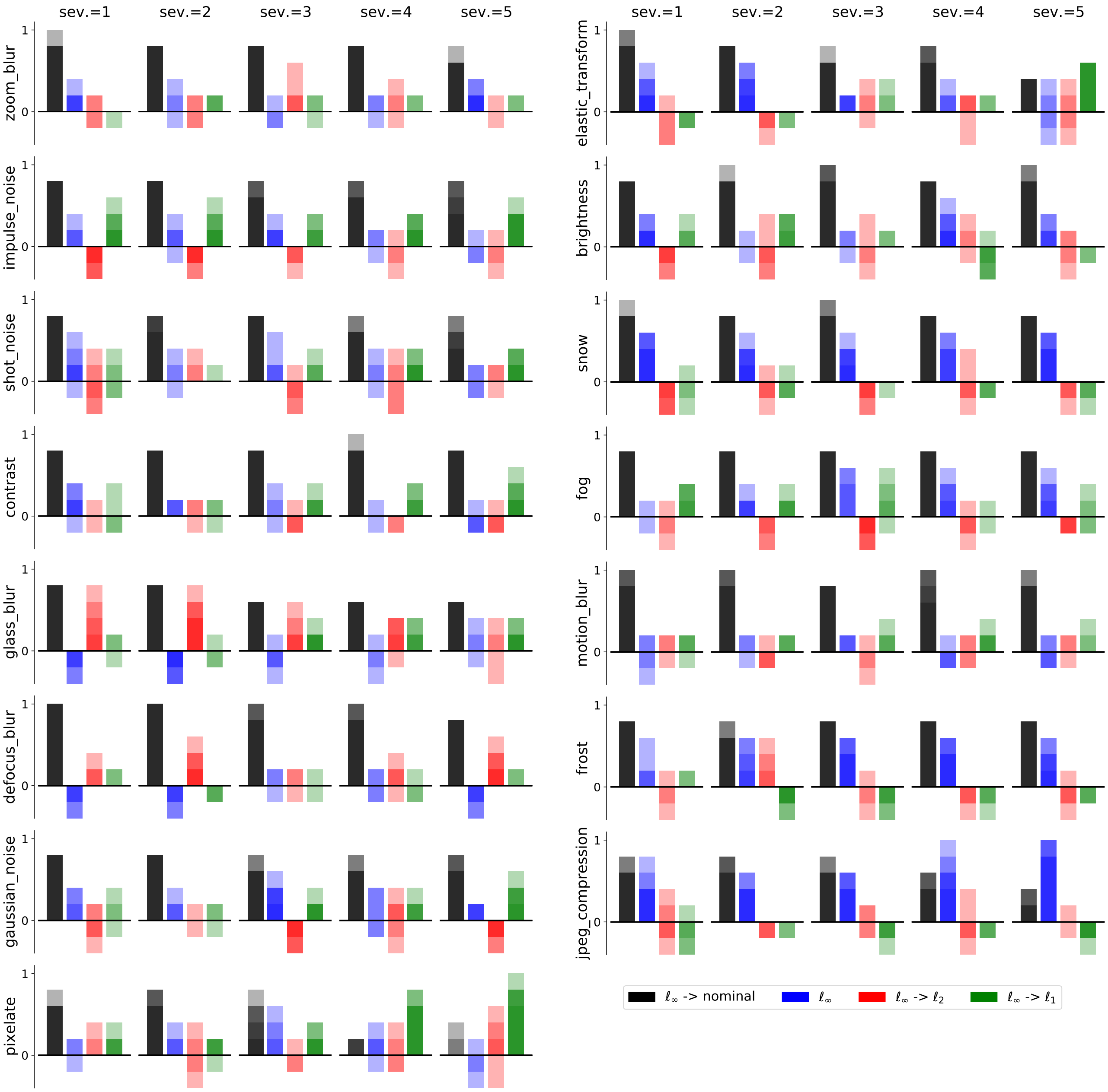

Figure 11. **Best soups over ImageNet-C subsets:** we plot the composition of the top 5 soups found by the grid search for each corruption type and severity level.