Appendix

A. Implementation Details

In this section, we elaborate on training, testing and distillation details of the proposed ensemble methods for different datasets, network architectures and input image resolution.

Training ResNet18 on 84x84 images on mini-ImageNet. For all experiments, we use ResNet18 with input image size 84x84, train with the Adam optimizer with an initial learning rate $3 \cdot 10^{-4}$, which is decreased by a factor 10 once during training when no improvement in validation accuracy is observed for p consecutive epochs. We use $p = 20$ for training individual models, $p = 30$ for training ensembles and when distilling the model. When distilling an ensemble into one network, p is doubled. We use random crops and color augmentation during training as well as weight decay with parameter $\lambda = 5 \cdot 10^{-4}$. At training time we use random crop, color transformation and adding random noise as data augmentation. During the meta-testing stage, we take central crops of size 224×224 from images and feed them to the feature extractor. No other preprocessing is used at test time. The parameters used in distillation are the same as in Section 4.3 of the paper.

Training WideResNet28 on 80x80 images on mini-ImageNet. For all experiments, we use WideResNet28 with input image size 80x80, train with the Adam optimizer with an initial learning rate $1 \cdot 10^{-4}$, which is decreased by a factor 10 once during training when no improvement in validation accuracy is observed for p consecutive epochs. We use $p = 20$ for training individual models, $p = 30$ for training ensembles and when distilling the model. When distilling an ensemble into one network, p is doubled. We use random crops and color augmentation during training as well as weight decay with parameter $\lambda = 5 \cdot 10^{-4}$. We also set a dropout rate inside convolutional blocks to be 0.5 as described in. At training time we use random crop and color transformation only as data augmentation. During the meta-testing stage, we take central crops of size 80×80 from images and feed them to the feature extractor. No other preprocessing is used at test time. The parameters used in distillation are the same as in Section 4.3 of the paper. Here, the maximal ensemble size we evaluated is 10 and not 20 due to memory limitations on available GPUs. Therefore, to construct an ensemble of size 20 we merge two ensembles of size 10, that were trained independently.

Training ResNet18 on 224x224 images on tiered-ImageNet For all experiments, we use ResNet18 with input image size 224x224, train with the Adam optimizer with an initial learning rate $3 \cdot 10^{-4}$, which is decreased by a factor 10 once during training when no improvement in validation accuracy is observed for p consecutive epochs. We use $p = 20$ for training individual models, ensembles and for distilation. We use random crops and color augmentation during training as well as weight decay with parameter $\lambda = 1 \cdot 10^{-4}$. At training time we use random crop and color transformation. During the meta-testing stage, we take central crops of size 224×224 from images and feed them to the feature extractor. No other preprocessing is used at test time. The parameters used in distillation are the same as in Section 4.3 of the paper.

B. Additional Results

In this section we report and analyze the performance of different ensemble types depending on their size for different network architectures and input image resolutions.

Few-shot Classification with ResNet18 on 224x224 images on CUB. The results for 1- and 5-shot classification on CUB are presented in Table A1. Training details and Figure summary of the results are discussed in Experimental section of the paper.

Few-shot Classification with ResNet18 on 84x84 images on mini-ImageNet. The results for 1- and 5-shot classification on *Mini*ImageNet are presented in Table A3 and summarized in Figure A1. We can see that Cooperation training is the most successful here for all ensemble sizes < 20 and other training strategies that introduce diversity tend to perform worse. This happens because single networks are far from overfitting the training set (as opposed to the case with 224x224 input size) and forcing diversity acts as harmful regularization. In contrary, cooperation training enforces useful learning signal and helps ensemble members achieve higher accuracy. Only for $n = 20$ where diversity matters more, robust ensembles perform the best.

Few-shot Classification with WideResNet28 on 80x80 images on mini-ImageNet. Results for 1- and 5-shot classification on *Mini*ImageNet are presented in Table A2 and summarized in Figure A1. In this case we can see again that Diverse training does not help since the networks do not memorize the training set. Robust ensembles outperform other training regimes emphasizing the importance of the proposed solution that generalizes across architectures.

5-shot

Full Ensemble	1	2	3	5	10	20
Independent	79.47 ± 0.49	81.34 ± 0.46	82.57 ± 0.46	83.16 ± 0.45	83.80 ± 0.45	83.95 ± 0.46
Diversity	79.47 ± 0.49	81.09 ± 0.45	82.23 ± 0.46	82.91 ± 0.46	84.30 ± 0.44	85.20 ± 0.43
Cooperation	79.47 ± 0.49	81.69 ± 0.46	82.95 ± 0.47	83.43 ± 0.47	84.01 ± 0.44	84.26 ± 0.44
Robust	79.47 ± 0.49	82.90 ± 0.46	83.36 ± 0.46	83.62 ± 0.45	84.47 ± 0.46	84.62 ± 0.44
Distilled Ensembles						
Robust-*dist*	–	82.72 ± 0.47	82.95 ± 0.46	83.27 ± 0.46	83.61 ± 0.46	83.57 ± 0.45
Robust-*dist*++	–	82.53 ± 0.48	83.04 ± 0.45	83.37 ± 0.46	83.22 ± 0.46	83.21 ± 0.44

1-shot

Ensemble type	1	2	3	5	10	20
Independent	64.25 ± 0.73	66.60 ± 0.72	67.64 ± 0.71	68.07 ± 0.70	68.93 ± 0.70	69.64 ± 0.69
Diversity	64.25 ± 0.73	65.99 ± 0.71	66.71 ± 0.72	68.19 ± 0.71	69.35 ± 0.70	70.07 ± 0.70
Cooperation	64.25 ± 0.73	67.21 ± 0.71	67.93 ± 0.70	68.22 ± 0.70	68.69 ± 0.70	68.80 ± 0.68
Robust	64.25 ± 0.73	67.33 ± 0.71	68.01 ± 0.72	68.53 ± 0.70	68.59 ± 0.70	69.47 ± 0.69
Distilled Ensembles						
Robust-*dist*	–	67.47 ± 0.71	67.29 ± 0.72	68.09 ± 0.70	68.71 ± 0.71	68.77 ± 0.71
Robust-*dist*++	–	67.01 ± 0.74	67.62 ± 0.72	68.68 ± 0.71	68.38 ± 0.70	68.68 ± 0.69

Table A1: **Few-shot classification accuracy on CUB.** The first column gives the type of ensemble and the top row indicates the number of networks in an ensemble. Here, *dist* means that an ensemble was distilled into a single network, and '++' indicates that extra unannotated images were used for distillation. We performed 1000 independent experiments on CUB-test and report the average with 95% confidence interval. All networks are trained on CUB-train set.

5-shot

Ensemble type	1	2	3	5	10	20
Independent	70.59 ± 0.51	73.24 ± 0.49	74.29 ± 0.48	74.89 ± 0.47	75.69 ± 0.47	75.93 ± 0.47
Diversity	70.59 ± 0.51	72.35 ± 0.47	73.44 ± 0.49	74.81 ± 0.48	75.47 ± 0.48	76.36 ± 0.47
Cooperation	70.59 ± 0.51	74.04 ± 0.47	74.81 ± 0.47	76.37 ± 0.48	76.73 ± 0.48	76.50 ± 0.47
Robust	70.59 ± 0.51	72.92 ± 0.50	73.09 ± 0.43	75.69 ± 0.42	76.71 ± 0.47	76.90 ± 0.48
Distilled Ensembles						
Robust-*dist*	–	73.04 ± 0.50	73.58 ± 0.49	74.35 ± 0.48	74.69 ± 0.49	75.24 ± 0.49
Robust-*dist*++	–	73.50 ± 0.49	74.17 ± 0.49	74.84 ± 0.49	75.12 ± 0.44	75.62 ± 0.48

1-shot

Ensemble type	1	2	3	5	10	20
Independent	53.31 ± 0.64	55.72 ± 0.60	56.85 ± 0.64	57.90 ± 0.63	58.21 ± 0.63	58.56 ± 0.61
Diversity	53.31 ± 0.64	54.61 ± 0.62	55.90 ± 0.62	57.06 ± 0.63	57.49 ± 0.62	58.93 ± 0.64
Cooperation	53.31 ± 0.64	55.80 ± 0.64	57.13 ± 0.63	58.18 ± 0.64	58.63 ± 0.63	58.73 ± 0.62
Robust	53.31 ± 0.64	55.95 ± 0.62	56.27 ± 0.64	58.51 ± 0.65	59.38 ± 0.65	59.48 ± 0.65
Distilled Ensembles						
Robust-*dist*	–	56.84 ± 0.64	56.58 ± 0.65	57.13 ± 0.63	57.41 ± 0.65	58.11 ± 0.64
Robust-*dist* ++	–	56.53 ± 0.62	57.03 ± 0.64	57.48 ± 0.65	58.05 ± 0.63	58.67 ± 0.65

Table A2: **Few-shot classification accuracy on *Mini*ImageNet, using ResNet18 and 84x84 image size.** The first column gives the strategy, the top row indicates the number N of networks in an ensemble. Here, *dist* means that an ensemble was distilled into a single network, and '++' indicates that extra unannotated images were used for distillation. We performed 1 000 independent experiments on *Mini*ImageNet-test and report the average with 95% confidence interval. All networks are trained on *Mini*ImageNet-train set.

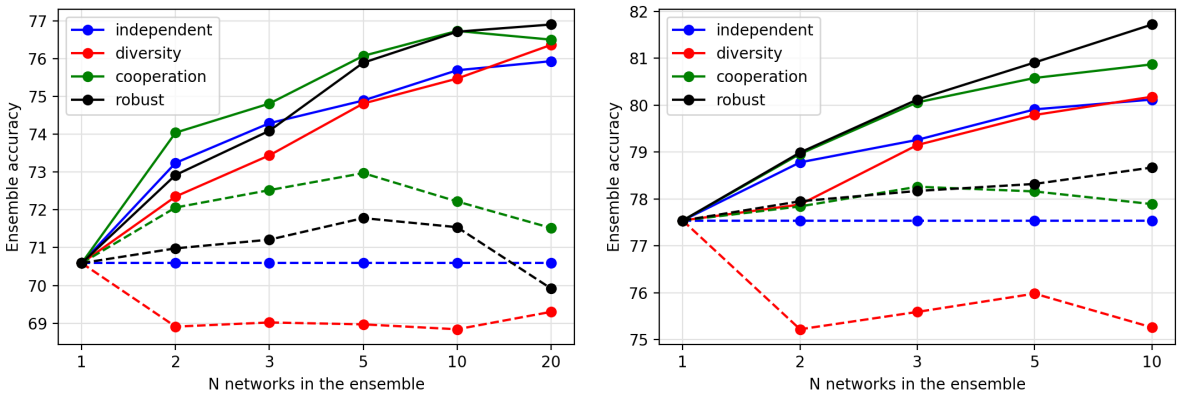
5-shot

Ensemble type	1	2	3	5	10
Independent	77.54 ± 0.45	78.78 ± 0.45	79.26 ± 0.43	79.91 ± 0.44	80.12 ± 0.43
Diversity	77.54 ± 0.45	77.88 ± 0.45	79.15 ± 0.44	79.79 ± 0.44	80.18 ± 0.44
Cooperation	77.54 ± 0.45	78.96 ± 0.46	80.06 ± 0.44	80.58 ± 0.45	80.87 ± 0.43
Robust	77.54 ± 0.45	78.99 ± 0.45	80.12 ± 0.43	80.91 ± 0.43	81.72 ± 0.44
Distilled Ensembles					
Robust-*dist*	–	79.44 ± 0.44	79.84 ± 0.44	80.01 ± 0.42	80.85 ± 0.43
Robust-*dist*++	–	79.16 ± 0.46	80.00 ± 0.44	80.25 ± 0.42	81.11 ± 0.43

1-shot

Ensemble type	1	2	3	5	10
Independent	59.02 ± 0.63	60.07 ± 0.62	60.58 ± 0.61	61.24 ± 0.63	62.05 ± 0.61
Diversity	59.02 ± 0.63	58.87 ± 0.62	60.63 ± 0.61	61.30 ± 0.62	62.28 ± 0.61
Cooperation	59.02 ± 0.63	60.22 ± 0.62	61.03 ± 0.61	62.07 ± 0.61	62.42 ± 0.61
Robust	59.02 ± 0.63	60.92 ± 0.62	62.03 ± 0.62	62.78 ± 0.61	63.39 ± 0.62
Distilled Ensembles					
Robust-*dist*	–	61.07 ± 0.62	61.57 ± 0.61	62.24 ± 0.61	62.80 ± 0.62
Robust-*dist* ++	–	61.37 ± 0.62	62.01 ± 0.60	62.45 ± 0.62	63.25 ± 0.62

Table A3: **Few-shot classification accuracy on *Mini*ImageNet, using WideResNet28 and 80x80 image size.** The first column gives the strategy, the top row indicates the number N of networks in an ensemble. Here, *dist* means that an ensemble was distilled into a single network, and '++' indicates that extra unannotated images were used for distillation. We performed 1 000 independent experiments on *Mini*ImageNet-test and report the average with 95% confidence interval. All networks are trained on *Mini*ImageNet-train set.



(a) ResNet18 with 84x84 input

(b) WideResNet28 with 80x80 input

Figure A1: **Dependency of ensemble accuracy on network architecture and input size for different ensemble strategies (one for each color) and various numbers of networks on *Mini*ImageNet 5-shots classification.** Solid lines give the ensemble accuracy after aggregating predictions. The average performance of single models from the ensemble is plotted with a dashed line. Best viewed in color.