Adversarially Robust Few-shot Learning via Parameter Co-distillation of Similarity and Class Concept Learners
– *Supplementary Material* –

Junhao Dong[1,2], Piotr Koniusz[4,3*], Junxi Chen[5], Xiaohua Xie[5], and Yew-Soon Ong[1,2*]
[1]Nanyang Technological University, [2]CFAR, IHPC, A*STAR, [3]Australian National University,
[4]Data61♥CSIRO, [5]Sun Yat-sen University
{junhao003, asysong}@ntu.edu.sg, piotr.koniusz@data61.csiro.au,
chenjx353@mail2.sysu.edu.cn, xiexiaoh6@mail.sysu.edu.cn

Abstract

In this supplementary material, we provide a detailed experimental configuration (Appendix A). Furthermore, we present more details about our RESISTANCE in Appendix B, including training protocols of component few-shot learners, diverse unification strategies, and how we extend robust few-shot learning with single-step adversary generation. In addition, we present visualization results (Appendix C) and hyper-parameter analysis (Appendix D).

A. Experimental Setting

Below, we provide a detailed experimental configuration of few-shot image classification datasets and implementation details of our RESISTANCE method.

A.1. Datasets

We conduct our experiments on three standard few-shot image classification datasets: Mini-ImageNet [68], CIFAR-FS [53], and FC100 [65]. Mini-ImageNet comprises 100 classes with 600 images of 84×84 pixels for each class, which is derived from ImageNet [58]. The dataset is typically divided into 64 training classes, 16 validation classes, and 20 test classes. CIFAR-FS follows the same data split of 100 categories but with a small resolution of 32×32 pixels, which is a subset of CIFAR-10/100 [63]. Few-shot-CIFAR100 (FC100) is also derived from CIFAR-100 but is split based on 20 super-classes to minimize the information overlap. Each super-class contains 5 generic classes. This challenging dataset is divided into 12, 4, and 4 super-classes for training, validation, and testing, respectively.

Following previous adversarial few-shot learning works [60, 70], we primarily focus on both 5-way 1-shot and 5-way 5-shot problems. In a 5-way 1-shot task, the few-shot classifier learns from 5 classes consisting of only one sample per class (support set), which is subsequently evaluated on 15 samples per class (query set). Both training and evaluation follow such a few-shot task construction protocol to simulate real-world few-shot learning scenarios. Query samples from each few-shot task are incorporated for optimizing the class concept learner, whereas the similarity learner is episodically trained on sampled few-shot tasks. All the experimental results are obtained over 2,000 randomly sampled few-shot tasks (episodes).

A.2. Implementation Details

In line with the settings of previous works [60, 70] and RobustBench [56], we employ Conv-4 [68] and ResNet-12 [61] as the target few-shot classification models. Conv-4 is built on four convolutional blocks with 64 channels per layer. ResNet-12 comprises four residual blocks with 64, 160, 320, and 640 channels, respectively. During the training stage, we enable a consistent view of training data for both few-shot learners (same episodes used for each type of learner). The class similarity learner learns from the relation labels between support and query images. The class concept learner directly learns to predict the object category in the full label space based on query images. In addition, both the component few-shot learners in our RESISTANCE share the same optimization configuration: a Stochastic Gradient Descent (SGD) optimizer with a momentum factor of 0.9, cyclic learning rate schedule [67] with a maximum learning rate $\eta = 0.1$, and a weight decay factor of 5×10^{-4}. We adversarially train the model for 50 epochs (512 few-shot tasks per epoch) for all three datasets.

For adversary generation during the training stage, we adopt the iterative Projected Gradient Descent (PGD) method [64] with 7 steps (step size $\alpha = 2/255$) on the similarity learner and class concept learner, respectively.

[*] Corresponding author.

Instead of using random initialization, we use the cross-branch class-wise global perturbation for initialization. We primarily focus on the ℓ_∞-norm threat model with the maximum perturbation radius of $\epsilon = 8/255$. Note that the adversarial perturbation for each learner is initialized by the global adversarial perturbation that can be efficiently obtained by conducting a single-step gradient ascent on each few-shot learner. The embedding model unification process starts at epoch $T = 40$ with a frequency of $m = 10$ iterations to redistribute the network parameters of the unified embedding model to both similarity and class concept learners. For computational efficiency, we adopt a single-step adversary generation strategy. The regularization hyper-parameters are set as $\beta = 0.99$, $\gamma = 0.5$, and $\tau = 0.5$.

Upon obtaining the unified embedding model as the training stage completes, we can directly transfer it to unforeseen few-shot tasks by coupling it with a rebuilt classification head. The choices of the classification head can vary depending on the specific requirements of downstream few-shot tasks. In this paper, we adopt an N-way logistic regression-based classifier optimized on extracted support feature embeddings for few-shot image classification. Following previous research [59, 60, 70], we adopt the evaluation protocol that encompasses accuracy on both legitimate query samples and their adversarial counterparts. In this paper, we mainly consider robustness against three strong white-box adversarial attacks: PGD [64] with 20 steps, CW [54], and Auto Attack (AA) [55].

B. Details of RESISTANCE

Below, we provide more details of RESISTANCE, including diverse unification strategies, details of cross-domain few-shot robustness evaluations, and details of the extension of RESISTANCE with single-step adversary generation.

B.1. Extension with Single-step Adversary

In this section, we provide more details about how we combine our RESISTANCE with the single-step adversary generation for computational efficiency improvement. The majority of computational cost for robust few-shot learning lies in multi-step adversarial generation. A well-established study has demonstrated the feasibility of combining single-step adversarial samples with adversarial training using abundant training data [52, 57, 62, 66, 71]. Nevertheless, such studies on adversarial robustness in the context of few-shot learning have not been conducted. Thus, we explore an efficient extension of robust few-shot learning by replacing multi-step adversary generation with its single-step counterparts $\hat{\mathbf{x}}^{\text{SGL}} = \mathbf{x} + \boldsymbol{\delta}_{\text{SGL}}$ during the training stage. A generic single-step adversary generation formula that approximates the worst-case adversarial perturbation to solve the inner maximization in Eq. (4) and (7) as follows:

$$\boldsymbol{\delta}_{\text{SGL}} = \Pi'_{\mathbb{B}(\epsilon)}\left[\boldsymbol{\delta}_0 + \alpha \ \text{sign}\left(\nabla_{\mathbf{x}}\mathcal{L}_{\text{KL}}(\mathbf{p}_{\mathbf{x}}\|\mathbf{p}_{\mathbf{x}+\boldsymbol{\delta}_0})\right)\right], \tag{13}$$

where $\Pi'_{\mathbb{B}(\epsilon)}$ denotes the projection operator onto the ℓ_∞-norm constraint box (notice Π' can be any projection strategy, *e.g.*, soft-projection rather than the clip as in case of Π). The single-step adversarial perturbation against each few-shot learner is randomly initialized by $\boldsymbol{\delta}_0 \sim \Omega$, *i.e.*, $\boldsymbol{\delta}_0$ is drawn from some distribution Ω. Recall that we only modify the adversarial generation strategy for both similarity and class concept learning for inner maximization. The global adversarial initialization perturbation is switched off in the setting where we approximate untargeted multi-step adversarial generation with advanced single-step adversary generation strategies for computational efficiency. Specifically, by adopting such a single-step strategy, we enjoy a significant reduction in the computational cost for the gradient backpropagation. We also show that RESISTANCE with single-step adversarial samples achieves comparable robustness to its multi-step counterpart.

B.2. Cross-domain Few-shot Robustness

The cross-domain transfer aims to generalize the feature embedding model learned on one dataset (*source* domain) to conduct inference on another dataset (*target* domain) in the few-shot setting. Such two datasets are characterized by disjoint semantic categories, different image resolutions, or other domain factors, posing a large domain gap. In our setting, the unified embedding obtained on the source data is treated as a feature extractor. For each test episode, we simply rebuild the classification head at a negligible computational overhead, *e.g.*, logistic regression. Subsequently, we measure the classification performance of clean samples and their adversarial counterparts from the *target* domain. Impressive cross-domain transfer results of RESISTANCE highlight its efficacy in maintaining adversarial robustness across diverse target domains.

B.3. Unification Strategies of Learners

As discussed in Section 4.5 of the main text, we investigate a series of unification strategies to combine similarity and class concept learners into the same framework for better performance, including prediction ensemble, feature ensemble, and multi-teacher distillation. We provide further details of these unification strategies below.

Prediction Ensemble. Following the principle of ensemble learning, the prediction ensemble strategy applies average voting on the output predictions during the inference stage of separately trained learners. This strategy aggregates the output probabilities (softmax logits) from individual learners, averaging them to formulate a unified prediction. Note that the predictions here are obtained based on

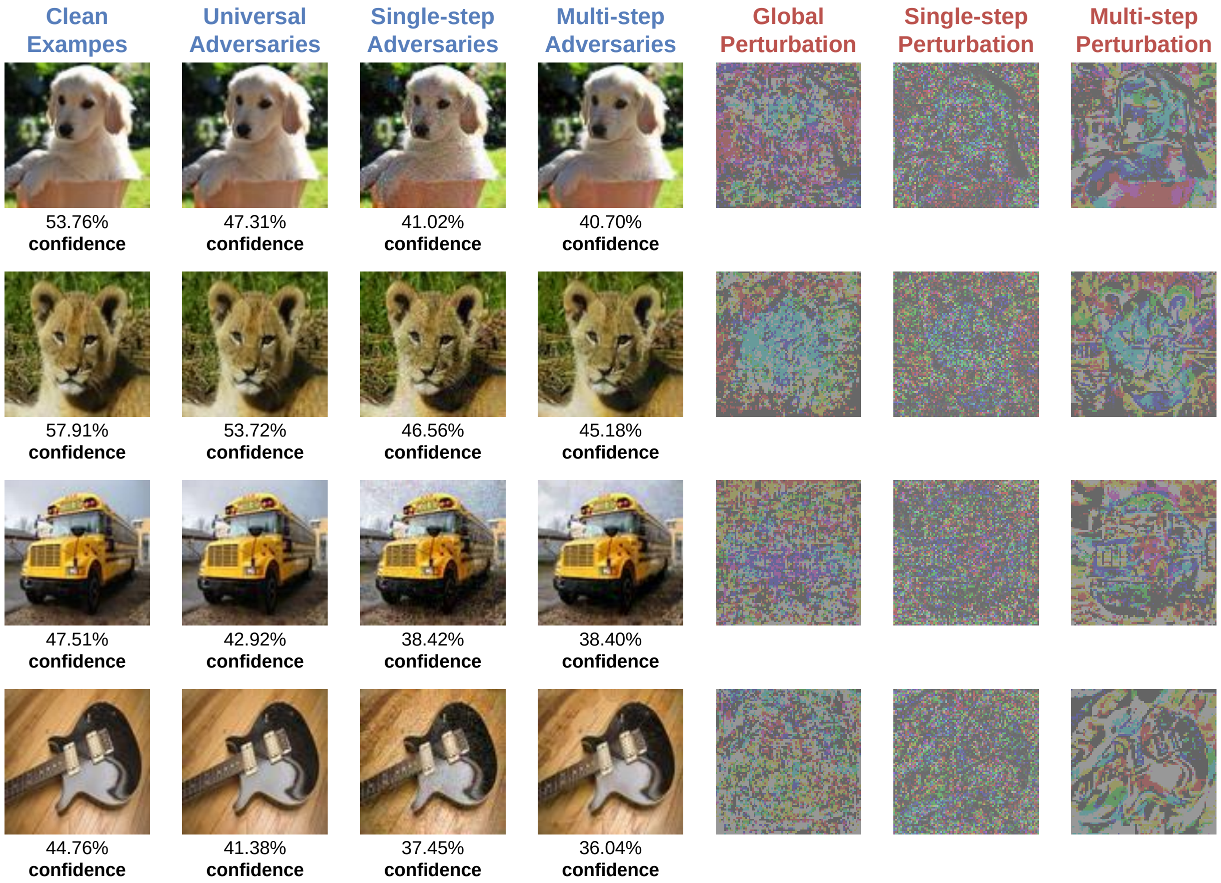


Figure 6. Illustration of our cross-branch class-wise global adversarial initialization perturbation (*global perturbation*), and both untargeted single-step and multi-step adversarial perturbations applied to our unified model. We also provide their corresponding adversarial samples and the prediction confidence associated with the ground-truth category.

the learned feature embedding models alongside their reconstructed classification heads tailored for each few-shot task when conducting the robustness evaluation.

Feature Ensemble. In contrast to the prediction ensemble (average of prediction scores), the feature ensemble strategy focuses on the aggregation of feature representations during the few-shot evaluation. For this unification strategy, feature representations of both robust few-shot learners are concatenated along the feature dimension, and then the classification head is rebuilt based on each episode.

Multi-teacher distillation. We here describe the unification process based on the popular multi-teacher knowledge distillation [69], which performs feature-level distillation from two well-trained few-shot learners (*cf*. our parameter-level distillation). The multi-teacher distillation step is given as:

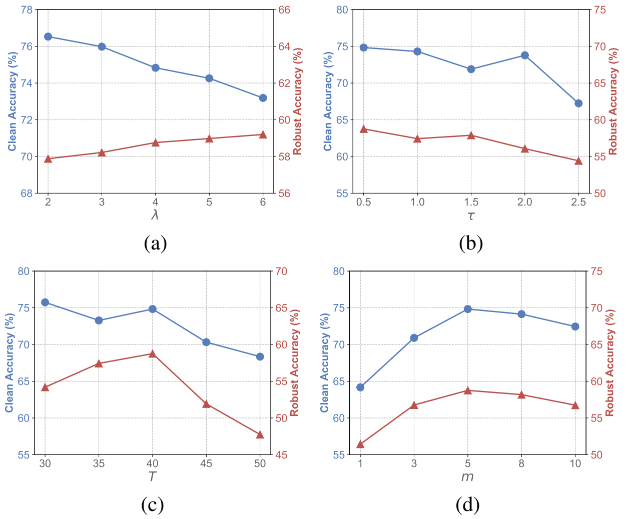
$$\min_{\boldsymbol{\theta}_u} \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{Q}}\Big[(1-\omega)\|f_{\boldsymbol{\theta}_{\mathrm{s}}}(\mathbf{x}) - f_{\boldsymbol{\theta}_{\mathrm{u}}}(\mathbf{x})\|_2^2 + \omega\|f_{\boldsymbol{\theta}_{\mathrm{c}}}(\mathbf{x}) - f_{\boldsymbol{\theta}_{\mathrm{u}}}(\mathbf{x})\|_2^2\Big], \tag{14}$$

where $0 \leq \omega \leq 1$ balances the learning tendency towards the similarity or class concept learner, respectively. The distilled unified feature embedding model is then applied to novel few-shot tasks by coupling it with a classification head during the evaluation stage. In contrast to our RESISTANCE, this strategy is time-consuming as one has to backpropagate w.r.t. parameters $\boldsymbol{\theta}_{\mathrm{u}}$ of the unified network.

In comparison with the above-mentioned unification strategies, RESISTANCE enjoys efficient adversarially robust few-shot learning distillation at the parameter level. Our unified embedding model dynamically inherits feature-level knowledge from both similarity and class concept learners during a cooperative optimization process. Hence, our proposed parameter-level co-distillation strategy can obtain better classification accuracy on clean and adversarial samples in the context of few-shot learning.

C. Visualization

In addition to the t-SNE visualization and attention maps presented in the main manuscript, we also visualize some adversarial samples (derived from diverse adversarial per-



(a)

(b)

(c)

(d)

Figure 7. Hyper-parameter sensitivity of our RESISTANCE on clean accuracy and (Auto-Attack) robust accuracy using ResNet-12 on CIFAR-FS. We present the weighting factor tuning of λ in Fig. 7a and τ in Fig. 7b. The adjustment of the starting epoch of unification T is provided in Fig. 7c, and the redistribution frequency m is in Fig. 7d.

turbations) alongside their legitimate counterparts. Figure 6 shows that adversarial samples effectively reduce the classification confidence of the ground-truth class despite their visual similarity to clean samples. Note that the cross-branch class-wise global perturbation is disruptive against all learners in our framework. On the other hand, single-step and multi-step adversarial samples, specifically crafted against the unified model, show a more significant disruptive effect on the unified model.

D. Hyper-parameter Analysis

Below, we investigate the contributions of individual modules of RESISTANCE by varying specific hyper-parameters. We report the natural and robust accuracy of the unified model across varying hyper-parameter configurations. Figure 7 shows that tuning λ provides a trade-off between the clean performance and adversarial robustness due to the impact λ has on the natural and boundary risks [72]. Choosing the starting epoch T of parameter redistribution to the similarity and class learners also provides a desired trade-off between the clean performance and adversarial robustness. As feature encoders become more stable with more episodes, redistribution may be enabled to ensure individual encoders do not become extremely different in terms of their parameter spaces. Furthermore, selecting an appropriate redistribution frequency m ensures that feature embeddings derived from various few-shot learners neither diverge significantly nor converge prematurely.

Table 11. Co-distillation with different training data for the class concept learner on the clean and (Auto-Attack) robust accuracy using ResNet-12 on the CIFAR-FS dataset.

Training data	1-shot		5-shot	
	Clean	Robust	Clean	Robust
Support & query	55.21	40.68	73.11	56.60
Only query	55.78	41.57	74.83	58.76

E. Additional Setting

Incorporating support samples for co-distillation. In the main paper, we focus on a co-distillation framework that solely utilizes query samples to train the class learner (including cross-branch class-wise adversarial generation). Furthermore, we evaluate whether using both support and query samples with the class learner can be beneficial in Table 11. The results show that adding support samples to the batch for the class concept learner actually deteriorates the robustness. We suspect this may be due to the fact that the similarity learner forms prototypes from support samples and adjusts them as it learns. However, using the class concept learner with support samples may implicitly interact with prototypes in undesired ways if the class concept learner is permitted to use support samples.

References

[52] Maksym Andriushchenko and Nicolas Flammarion. Understanding and improving fast adversarial training. *Advances in Neural Information Processing Systems*, 33:16048–16059, 2020.

[53] Luca Bertinetto, João F. Henriques, Philip H. S. Torr, and Andrea Vedaldi. Meta-learning with differentiable closed-form solvers. In *International Conference on Learning Representations, ICLR*, 2019.

[54] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. IEEE, 2017.

[55] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pages 2206–2216. PMLR, 2020.

[56] Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. Robustbench: a standardized adversarial robustness benchmark. *arXiv preprint arXiv:2010.09670*, 2020.

[57] Pau de Jorge Aranda, Adel Bibi, Riccardo Volpi, Amartya Sanyal, Philip Torr, Grégory Rogez, and Puneet Dokania. Make some noise: Reliable and efficient single-step adversarial training. *Advances in Neural Information Processing Systems*, 35:12881–12893, 2022.

[58] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image

database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[59] Junhao Dong, Yuan Wang, Jian-Huang Lai, and Xiaohua Xie. Improving adversarially robust few-shot image classification with generalizable representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9025–9034, 2022.

[60] Micah Goldblum, Liam Fowl, and Tom Goldstein. Adversarially robust few-shot learning: A meta-learning approach. *Advances in Neural Information Processing Systems*, 33: 17886–17895, 2020.

[61] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[62] Xiaojun Jia, Yong Zhang, Xingxing Wei, Baoyuan Wu, Ke Ma, Jue Wang, and Xiaochun Cao. Prior-guided adversarial initialization for fast adversarial training. In *European Conference on Computer Vision*, pages 567–584. Springer, 2022.

[63] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[64] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

[65] Boris Oreshkin, Pau Rodríguez López, and Alexandre Lacoste. Tadam: Task dependent adaptive metric for improved few-shot learning. *Advances in neural information processing systems*, 31, 2018.

[66] Ali Shafahi, Mahyar Najibi, Mohammad Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free! *Advances in Neural Information Processing Systems*, 32, 2019.

[67] Leslie N Smith and Nicholay Topin. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial intelligence and machine learning for multi-domain operations applications*, pages 369–386. SPIE, 2019.

[68] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. *Advances in neural information processing systems*, 29, 2016.

[69] Lin Wang and Kuk-Jin Yoon. Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):3048–3068, 2022.

[70] Ren Wang, Kaidi Xu, Sijia Liu, Pin-Yu Chen, Tsui-Wei Weng, Chuang Gan, and Meng Wang. On fast adversarial robustness adaptation in model-agnostic meta-learning. In *9th International Conference on Learning Representations, ICLR*, 2021.

[71] Eric Wong, Leslie Rice, and J. Zico Kolter. Fast is better than free: Revisiting adversarial training. In *8th International Conference on Learning Representations, ICLR*, 2020.

[72] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pages 7472–7482. PMLR, 2019.