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

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

Few-shot learning (FSL) facilitates a variety of computer vision tasks yet remains vulnerable to adversarial attacks. Existing adversarially robust FSL methods rely on either visual similarity learning or class concept learning. Our analysis reveals that these two learning paradigms are complementary, exhibiting distinct robustness due to their unique decision boundary types (concepts clustering by the visual similarity label vs. *classification by the class labels). To bridge this gap, we propose a novel framework unifying adversarially robust similarity learning and class concept learning. Specifically, we distill parameters from both network branches into a "unified embedding model" during robust optimization and redistribute them to individual network branches periodically. To capture generalizable robustness across diverse branches, we initialize adversaries in each episode with cross-branch class-wise "global adversarial perturbations" instead of less informative random initialization. We also propose a branch robustness harmonization to modulate the optimization of similarity and class concept learners via their relative adversarial robustness. Extensive experiments demonstrate the state-of-the-art performance of our method in diverse few-shot scenarios.*

1. Introduction

Few-shot learning (FSL) facilitates training various applications [18, 20, 22, 23, 35, 39, 44, 45, 47, 50, 51] with limited training data. Despite its potential, a cornerstone of FSL, Deep Neural Network (DNN), is known to be susceptible to adversarial samples [11, 29, 38] that are easily obtainable by appending visually imperceptible perturbations to natural samples. Such adversarial samples pose even more severe disruption to the inference of DNN-based few-shot learners, inducing a serious security threat that hinders the practical deployment of FSL systems [14].

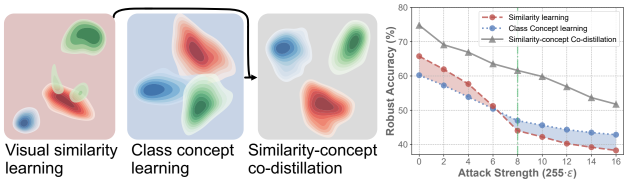


Figure 1. Motivation for co-distillation. (a) Feature spaces (2D PCA and density estimation are applied to feature representations from an episode, where 3 cluster colors represent 3 class labels) of a similarity learner (*left*) and a class concept learner (*middle*) with very different decision boundaries. As two of three green clusters from the similarity learner overlap with the red cluster, (weak) adversaries with a small radius ϵ are enough to breach the decision boundary and immunize the similarity learner. In contrast, the concept learner requires a larger ϵ for immunization due to clearer class separation. Our co-distillation (*right*) results in well-separated class clusters. (b) Robust accuracy confirms that similarity and concept learners are affected differently w.r.t. ϵ, exhibiting complementary robustness for our co-distillation indicated by the grey curve. The green line refers to $\epsilon = 8/255$ used for training.

A growing body of research has focused on adversarially robust FSL. These works primarily fall into two categories: (i) visual similarity learning [14, 19] and (ii) class concept learning [9, 36]. Both paradigms aim to improve robustness against unforeseen adversaries with minimal training data. The former emphasizes capturing visual similarity (relations) between support and query samples across diverse few-shot tasks (episodes), involving a small subset of categories for episodic training. In contrast, class concept learning results in a feature embedding model based on explicitly learning pre-defined class concepts (*e.g*., object categories).

However, these paradigms lead to adversarial robustness from distinct viewpoints, which is mainly driven by their respective labeling strategies, *i.e*., the similarity learner learns a binary classifier or metric between pairs of samples to capture their visual similarity, whereas the class concept learner focuses on semantic object categories. To support this claim, we provide feature space visualizations for vi-

[*] Corresponding author.

sual similarity and class concept learning in Fig. 1a. The key observation made by us is that similarity learners often spread features per class into multiple clusters with potential overlaps between classes, *e.g.*, two of three green clusters overlapping with the red cluster in the left panel implies that low-strength attacks can easily breach cluster boundaries, and thus weak adversaries forge adversarial robustness of the similarity learner. In contrast, class concept learners produce more distinct class-wise clusters separated by decision boundaries. While this separation is also suboptimal, more global clusters require stronger attacks for disruption, reflecting the robustness tendency against adversaries of high attack strengths. Fig. 1a (right) shows that our co-distillation method, adversarially robust few-shot learning via *paRametEr co-diStIllation of SimilariTy and clAss coNCept lEarners* **(RESISTANCE)**, achieves better separation of class concepts. Fig. 1b proves differing trends of robust performance between similarity and class learners w.r.t. perturbation radius ϵ. Our co-distillation method, RESISTANCE, indicated by the grey curve, confirms the complementary nature of integrating both learning strategies.

In our RESISTANCE framework, we distill network parameters from the feature embedding models of both the similarity and class concept learners into a "unified embedding" model. These network parameters are then periodically redistributed back to their respective network branches for synchronization during training. To capture generalizable robustness across diverse network branches, we initialize adversaries in each episode with cross-branch class-wise global adversarial perturbations and thus generate branch-specific adversarial samples to improve their respective decision boundaries. However, our RESISTANCE goes beyond being merely a parameter redistribution strategy. Recognizing that adversaries perturb the similarity and classification losses in an unequal manner, we introduce a robustness harmonization module to upweight the robust learning branch that experiences a larger adversarial vulnerability, thereby counterbalancing the dominance of one branch during co-distillation. Furthermore, we make a first attempt to explore single-step adversary generation strategies in adversarially robust FSL for better computational efficiency.

Extensive experiments and analyses demonstrate that our RESISTANCE consistently outperforms the state-of-the-art adversarially robust FSL methods w.r.t. both natural performance and adversarial robustness in diverse few-shot scenarios while enjoying a reasonable computational cost.

Our contributions are summarized as follows:

i. By analyzing the complementary nature of visual similarity and class concept learning distinguished by their unique label spaces, we propose a novel adversarially robust few-shot learning framework based on a simple but effective parameter co-distillation mechanism, improving robustness across diverse attack strengths.

ii. To promote the uniformity of robustness across learners, we introduce cross-branch class-wise adversarial perturbations for branch-specific adversary initialization. We also propose a robustness harmonization module to modulate the optimization of diverse branches.

iii. Comprehensive experiments demonstrate the effectiveness and generalization ability of RESISTANCE compared to the state-of-the-art adversarially robust few-shot learning approaches. In addition, we investigate the scalability of RESISTANCE with the single-step adversary generation strategies for better efficiency.

2. Preliminaries

Related works. Few-shot learning primarily focuses on learning novel class/visual concepts from limited data [13, 26, 27, 37, 40, 47]. However, this data scarcity predisposes few-shot classifiers to increased susceptibility to adversarial samples. In the broader landscape of defenses against such adversaries, adversarial training improves robustness by augmenting adversarial samples into training data [2]. Existing works focus on striking a balance between natural performance and adversarial robustness [10, 12, 30, 49]. For instance, a mixup training strategy [43] merges insights from both naturally trained and adversarially trained models. In contrast, apart from seeking a trade-off between natural performance and robustness, we aim to mitigate the robustness discrepancy between visual similarity and class concept learning based on feature-level co-distillation. Moreover, our global adversarial perturbations are inspired by universal adversarial robustness [3, 25, 34]. We further extend it by introducing a cross-branch and class-wise version of universal perturbation for branch-specific adversary initialization, ensuring the uniformity of robust learning.

Despite the effectiveness of adversarial training, its impact on network robustness highly depends on the available volume of data [8, 31]. In addressing this challenge, Goldblum et al. [14] integrated adversary generation within similarity learning for a robust feature extractor. Dong et al. [9] extended the so-called transfer learning (class concept learning) to adversarially robust FSL. Despite the nontrivial robustness obtained by similarity and class concept learning families, their complementary characteristics in this context remain largely unexplored. Our RESISTANCE fills such a gap for improved robustness in few-shot settings.

Similarity learning *vs.* class concept learning. Given an FSL dataset with Z classes, we form episodes $(\mathcal{S}, \mathcal{Q}) \sim \mathcal{D}$ with support and query samples. Specifically, we draw K samples from each of N classes $(N < Z)$ to construct the support set $\mathcal{S} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{|\mathcal{S}|}$ $(|\mathcal{S}| = KN)$, as well as K' disjoint samples from the same N classes for the query set $\mathcal{Q} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{|\mathcal{Q}|}$. Both the support and query sets define an N-way K-shot task. Note that categories from the training and evaluation sets do not overlap. Robust similarity learning typically uses class-wise feature mean prototypes:

$$\boldsymbol{\mu}_n = \frac{1}{|\mathcal{S}_n|} \sum_{(\mathbf{x},y)\in\mathcal{S}_n} f_{\boldsymbol{\theta}_s}(\mathbf{x}), \tag{1}$$

where $n \in \{1,\ldots,N\}$ indicates one of N classes of an episode, and $f_{\boldsymbol{\theta}_s}: \mathcal{X} \rightarrow \mathbb{R}^D$ is a feature encoder with parameters $\boldsymbol{\theta}_s$. Let $\mathbf{M} = \{\boldsymbol{\mu}_n\}_{n=1}^N$. For each query sample $\mathbf{x}$, one maximizes the likelihood of that sample and the corresponding prototype to share the same class label:

$$p(y_{\mathbf{x}} = y_{\boldsymbol{\mu}_n}|\mathbf{x}, \mathbf{M}) = \frac{\exp(-d^2(f_{\boldsymbol{\theta}_s}(\mathbf{x}), \boldsymbol{\mu}_n))}{\sum_{n'=1}^N \exp(-d^2(f_{\boldsymbol{\theta}_s}(\mathbf{x}), \boldsymbol{\mu}_{n'}))}, \tag{2}$$

where $d(\cdot,\cdot)$ is the Euclidean distance. For brevity, for all N prototypes, we define a likelihood vector $\mathbf{p}_{\mathbf{x}}^{\mathbf{M}} = [p(y_{\mathbf{x}} = y_{\boldsymbol{\mu}_1}|\mathbf{x}, \mathbf{M}), \ldots, p(y_{\mathbf{x}} = y_{\boldsymbol{\mu}_N}|\mathbf{x}, \mathbf{M})]^\top \in \mathbb{R}^N$ and a one-hot vector $\mathbf{y}_{\mathbf{x}} = [\mathbb{1}(y_{\mathbf{x}} = y_{\boldsymbol{\mu}_1}), \ldots, \mathbb{1}(y_{\mathbf{x}} = y_{\boldsymbol{\mu}_N})]^\top \in \{0,1\}^N$.

Subsequently, the adversarially robust similarity learning typically learns visual similarity (object relations) between support and query sets of episodes in $\mathcal{D}$. For an episode with a query set $\mathcal{Q}$ and prototypes $\mathbf{M}$ of support set $\mathcal{S}$, the similarity learner minimizes the following loss w.r.t. $\boldsymbol{\theta}_s$:

$$\mathcal{L}_s(\mathcal{Q}, \mathbf{M}) = \tag{3}$$
$$\frac{1}{|\mathcal{Q}|} \sum_{(\mathbf{x},y_{\mathbf{x}})\in\mathcal{Q}} \Big[\mathcal{L}_{\text{CE}}(\mathbf{p}_{\mathbf{x}}^{\mathbf{M}}, \mathbf{y}_{\mathbf{x}}) + \lambda \max_{\|\boldsymbol{\delta}_s\|_\infty < \epsilon} \mathcal{L}_{\text{KL}}\big(\mathbf{p}_{\mathbf{x}}^{\mathbf{M}} \| \mathbf{p}_{\mathbf{x}+\boldsymbol{\delta}_s}^{\mathbf{M}}\big)\Big],$$

where $\mathcal{L}_{\text{CE}}$ and $\mathcal{L}_{\text{KL}}$ represent the Cross-Entropy (CE) loss for natural samples and the Kullback–Leibler (KL) divergence for generating adversarial samples $\hat{\mathbf{x}} = \mathbf{x} + \boldsymbol{\delta}_s$. Moreover, $\boldsymbol{\delta}_s$ is the ℓ_∞-norm constrained adversarial perturbation obtained by Projected Gradient Descent (PGD) [24]. In conclusion, the visual similarity loss optimizes:

$$\min_{\boldsymbol{\theta}_s} \mathbb{E}_{(\mathcal{S},\mathcal{Q})\sim\mathcal{D}} \mathcal{L}_s(\mathcal{Q}, \mathbf{M}(\mathcal{S})), \tag{4}$$

where $\mathbf{M}(\mathcal{S})$ is simply a set of prototypes derived from the support set $\mathcal{S}$ obtained according to Eq. (1).

Unlike the similarity learner, which uses $N < Z$ class-wise feature mean prototypes to model similarity locally (*i.e*., per episode), the class concept learner learns global classifier weights for all Z classes. Let $f_{\boldsymbol{\theta}_c}: \mathcal{X} \rightarrow \mathbb{R}^D$ be a feature encoder with parameters $\boldsymbol{\theta}_c$. We define the classifier head (softmax with learnable weights $\mathbf{W} = \{\mathbf{w}_z\}_{z=1}^Z$) as:

$$p(y_{\mathbf{x}} = z|\mathbf{x}, \mathbf{W}) = \frac{\exp(\mathbf{w}_z^\top f_{\boldsymbol{\theta}_c}(\mathbf{x}))}{\sum_{z'=1}^Z \exp(\exp(\mathbf{w}_{z'}^\top f_{\boldsymbol{\theta}_c}(\mathbf{x})))}. \tag{5}$$

For brevity, for all Z classes, we define a likelihood vector $\mathbf{p}_{\mathbf{x}}^{\mathbf{W}} = [p(y_{\mathbf{x}} = 1|\mathbf{x}, \mathbf{W}), \ldots, p(y_{\mathbf{x}} = Z|\mathbf{x}, \mathbf{W})]^\top \in \mathbb{R}^Z$ and a one-hot vector $\mathbf{y}'_{\mathbf{x}} = [\mathbb{1}(y_{\mathbf{x}} = 1), \ldots, \mathbb{1}(y_{\mathbf{x}} = Z)]^\top \in \{0,1\}^Z$. Let batch $\mathcal{B}$ be the query set, *i.e*., $\mathcal{B} = \mathcal{Q}$, ($\mathcal{B} = \mathcal{S} \cup \mathcal{Q}$ is also possible). We minimize the following loss w.r.t. $\boldsymbol{\theta}_c$ and $\mathbf{W}$:

$$\mathcal{L}_c(\mathcal{B}, \mathbf{W}) = \tag{6}$$
$$\frac{1}{|\mathcal{B}|} \sum_{(\mathbf{x},y_{\mathbf{x}})\in\mathcal{B}} \Big[\mathcal{L}_{\text{CE}}(\mathbf{p}_{\mathbf{x}}^{\mathbf{W}}, \mathbf{y}'_{\mathbf{x}}) + \lambda \max_{\|\boldsymbol{\delta}_c\|_\infty < \epsilon} \mathcal{L}_{\text{KL}}\big(\mathbf{p}_{\mathbf{x}}^{\mathbf{W}} \| \mathbf{p}_{\mathbf{x}+\boldsymbol{\delta}_c}^{\mathbf{W}}\big)\Big].$$

Simply put, the class concept learner optimizes:

$$\min_{\boldsymbol{\theta}_c, \mathbf{W}} \mathbb{E}_{(\mathcal{S},\mathcal{Q})\sim\mathcal{D}} \mathcal{L}_c(\mathcal{Q}, \mathbf{W}). \tag{7}$$

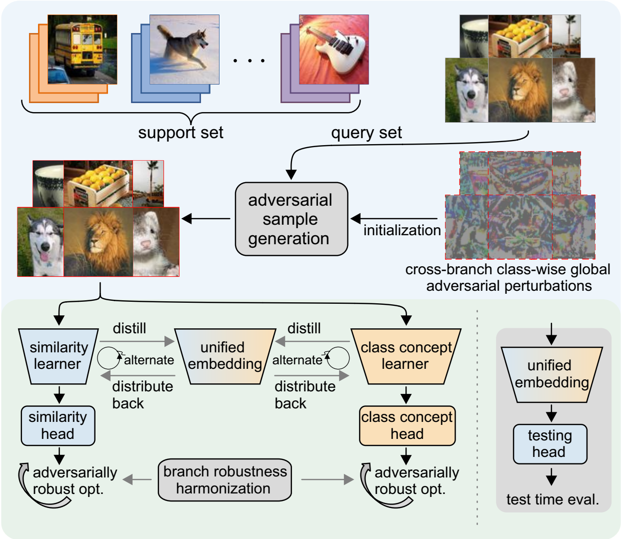


Figure 2. Adversarially robust few-shot learning via our *paRameteEr co-diStIllation of SimilariTy and clAss coNCept lEarners* (RESISTANCE). Adversarial samples are generated w.r.t. the similarity learner and class concept learner (they both enjoy distinct label spaces). These adversarial samples are seeded by cross-branch class-wise global adversarial initialization perturbations that help attain an effective local optimum of adversarial generation for each branch. Parameters of the similarity and class concept branches are co-distilled into the unified embedding model and regularly redistributed back to individual branches for stability. The branch robustness harmonization promotes higher learning rates for the less robust branch. During evaluations, we employ the unified embedding alongside a rebuilt classification head (multinomial logistic regression) based on support samples from an episode.

3. Proposed Approach

Below, we introduce our adversarially robust FSL framework, RESISTANCE, which co-distills parameters across network branches and further promotes adversarial robustness by cross-branch class-wise global adversarial initialization perturbations and a branch harmonization strategy.

3.1. Similarity-concept Co-distillation

We here describe how to integrate both similarity and class concept learners into a unified embedding model for better adversarial robustness in FSL. Our approach alternates between two stages: robust optimization of both learners and dynamic knowledge distillation, as shown in Figure 2. Given the discrepancy in the label spaces between similarity and class concept learners, we resort to parameter co-distillation, as their parameters $\boldsymbol{\theta}_s$ and $\boldsymbol{\theta}_c$ are optimized on two network branches with the same network architecture. Both learners leverage the clean query sets to generate corresponding adversarial samples. However, their distinction resides in the label space: the similarity head of the similarity learner clusters feature representations around N prototypes, and these prototypes depend not on the semantic label but on visual appearance, leading to multiple clusters, as shown in Figure 1a (left). In contrast, the classification head of the class concept learner maps inputs to the class label space of size $Z > N$. The unification process involves distilling parameters from both learners after each training iteration using the Exponential Moving Average (EMA) update, thereby synergizing their strengths for robust FSL:

$$\boldsymbol{\theta}_u := \beta\boldsymbol{\theta}_u + (1-\beta)\left[\gamma\boldsymbol{\theta}_s + (1-\gamma)\boldsymbol{\theta}_c\right], \tag{8}$$

where $0 \leq \beta \leq 1$ is the decay rate, and $0 \leq \gamma \leq 1$ balances the impact of each network branch. Considering the unstable nature of representations in early training epochs, we enable parameter distillation from epoch $T > 0$. Prior to epoch T, both similarity and concept learners are trained separately on the identical data (except that label spaces differ).

To prevent large divergence between parameters $\boldsymbol{\theta}_s$ and $\boldsymbol{\theta}_c$, we distribute $\boldsymbol{\theta}_u$ every m iterations into $\boldsymbol{\theta}_s$ and $\boldsymbol{\theta}_c$[1]. This periodic parameter distribution synchronizes learners, prevents large parameter deviation, and thus limits overfitting.

3.2. Cross-branch Class-wise Global Adversarial Initialization Perturbations

The key objective of our RESISTANCE framework is to obtain a robust unified embedding model capable of handling distinct knowledge from different learning branches. A challenge arises when adversaries generated from the same clean sample vary across branches due to the label space discrepancy, potentially leading to incompatible robustness for co-distillation. To address this and capture generalizable robustness, we design *cross-branch class-wise global adversarial initialization perturbations* (GAIP) inspired by [25, 34]. Such a global strategy ensures that samples within the same class are initialized with a common adversarial base across diverse learning branches. We expect that this prior-guided adversary initialization brings natural samples closer to decision boundaries, where they can be further refined by each branch to reflect their unique label spaces on final adversaries, thus enhancing the unified robustness.

Let a sample $\mathbf{x}^z \in \mathcal{B}_z$ belong to the batch subset containing natural samples of class z (we use the query subset of class z, *i.e.*, $\mathcal{B}_z = \mathcal{Q}_z$, while $\mathcal{B}_z = \mathcal{S}_z \cup \mathcal{Q}_z$ is also possible). We combine the similarity learning, the class concept learning, and the unified embedding branches as follows:

$$\mathcal{L}_{\text{GAIP}}(\mathcal{B}_z; \boldsymbol{\delta}_0^z) = \sum_{\mathbf{x}^z \in \mathcal{B}_z} \sum_{\substack{g \in \{f_{\boldsymbol{\theta}_s}, \\ f_{\boldsymbol{\theta}_c}, f_{\boldsymbol{\theta}_u}\}}} \left\| g(\mathbf{x}^z + \boldsymbol{\delta}_0^z) - \boldsymbol{\mu}_z^{(g)} \right\|_2^2, \tag{9}$$

where $\boldsymbol{\mu}_z^{(g)} = \frac{1}{|\mathcal{B}_z|} \sum_{(\mathbf{x},y) \in \mathcal{B}_z} g(\mathbf{x})$ (not to be mistaken with $\boldsymbol{\mu}_n$ in Eq. (1)) is a class-wise prototype computed over a batch subset $\mathcal{B}_z$ with class z, and $\boldsymbol{\delta}_0^z = \boldsymbol{\delta}_0^{z\,(\iota)}$ is the initializing perturbation for class z (*i.e.*, it seeds adversarial perturbations for each branch), which is disruptive across all the learners in our RESISTANCE framework as the iterative gradient ascent progresses on Eq. (9) over batch iterations ι:

$$\boldsymbol{\delta}_0^{z\,(\iota)} = h\big(\mathcal{B}_z^\iota; \boldsymbol{\delta}_0^{z(\iota-1)}; \alpha\big) = \tag{10}$$
$$\Pi_{\mathbb{B}(\epsilon)}\Big(\boldsymbol{\delta}_0^{z\,(\iota-1)} + \alpha \text{ sign}\left(\nabla_{\boldsymbol{\delta}_0^{z(\iota-1)}} \mathcal{L}_{\text{GAIP}}^z\big(\mathcal{B}_z^\iota; \boldsymbol{\delta}_0^{z(\iota-1)}\big)\right)\Big),$$

where $\boldsymbol{\delta}_0^{z\,(0)} \sim 0.001 \cdot \mathcal{N}(\mathbf{0}, \mathbf{I})$. $\Pi_{\mathbb{B}(\epsilon)}(\cdot)$ denotes the projection into the ℓ_∞-norm bound with radius ϵ. We incorporate our global perturbations into the iterative adversary generation during training by replacing the standard random initialization with such a cross-model and class-wise consistency constraint for unified adversarial robustness.

3.3. Branch Robustness Harmonization Module

Given the distinct label spaces of our similarity and class concept learners, they inherently exhibit different adversarial robustness characteristics, *e.g.*, their contributions to the unified robustness differ per episode. In addition, the emergence of a dominant learner can indirectly affect contributions from the other model and thus make co-distillation ineffective. Hence, we propose the *branch robustness harmonization* to modulate the impact of the similarity and class concept learning by upweighting the learning rate of the less robust learner. To quantify the robustness discrepancy between the similarity and class concept learners, we calculate their *relative adversarial robustness score* as follows:

$$\kappa_s(\mathcal{Q}, \mathcal{S}) = \frac{\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{Q}}\left[\mathcal{L}_{\text{KL}}\left(\mathbf{p}_{\mathbf{x}}^{\mathbf{W}} \| \mathbf{p}_{\mathbf{x}+\boldsymbol{\delta}_c^{\mathbf{x}}}^{\mathbf{W}}\right)\right]}{\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{Q}}\left[\mathcal{L}_{\text{KL}}\left(\mathbf{p}_{\mathbf{x}}^{\mathbf{M}(\mathcal{S})} \| \mathbf{p}_{\mathbf{x}+\boldsymbol{\delta}_s^{\mathbf{x}}}^{\mathbf{M}(\mathcal{S})}\right)\right]}, \tag{11}$$

where $\mathbf{p}_{\mathbf{x}}^{\mathbf{W}}$ and $\mathbf{p}_{\mathbf{x}+\boldsymbol{\delta}_c^{\mathbf{x}}}^{\mathbf{W}}$ are softmax class vectors (simplex) of size Z from the class concept learner, as defined just below Eq. (5). Additionally, $\mathbf{p}_{\mathbf{x}}^{\mathbf{M}}$ and $\mathbf{p}_{\mathbf{x}+\boldsymbol{\delta}_s^{\mathbf{x}}}^{\mathbf{M}}$ are softmax class vectors (simplex) of size N from the similarity learner, as defined just below Eq. (2). Finally, vectors $\boldsymbol{\delta}_c^{\mathbf{x}}$ and $\boldsymbol{\delta}_s^{\mathbf{x}}$ indicate adversarial perturbations w.r.t. sample $\mathbf{x}$ obtained by iterative gradient ascent on the class concept and similarity learners, respectively. $\mathbf{M}(\mathcal{S})$ represents the class-wise feature mean prototypes for the support set $\mathcal{S}$, as defined in both vector and matrix forms in and below Eq. (1). We can also define $\kappa_c = 1/\kappa_s$. If the class learner is less robust than the similarity learner given their respective adversarial samples ($\kappa_s > 1$), the learning rate of the similarity learner is respectively decreased. It remains unchanged for $0 \leq \kappa_s \leq 1$, as determined by the function below:

$$\eta_s' = \eta_s\left[1 - \tanh\left(\tau \max\left(0, \log\left(\kappa_s\right)\right)\right)\right], \tag{12}$$

where η_s and η_s' denote the original and corrected learning rates of the similarity learner, and $\tau \geq 0$ controls the steepness of downweighting the learning rate. We can obtain η_c' by analogy. The less robust learner receives a larger learning rate (relative to the other learner), which facilitates balancing the robust optimization of both learners.

[1] This can be done by overriding parameters $\boldsymbol{\theta}_s$ and $\boldsymbol{\theta}_c$ or simply by updating each of them by EMA, which is marginally better but non-essential.

Algorithm 1 Adversarially Robust Few-shot Learning via paRametEr co-diStIllation of SimilariTy and clAss coNCept lEarners (RESISTANCE).

Input: Network architecture $f(\cdot)$; dataset $\mathcal{D}$; maximum perturbation radius $\epsilon > 0$; step size $\alpha > 0$ for adversary generation; learning rates η_s and η_c; starting epoch T and frequency m for redistributing distilled parameters to network branches; EMA decay rate β; balancing factor γ; steepness $\tau > 0$ for branch harmonization.

```
Randomly initialize $\boldsymbol{\theta}_s, \boldsymbol{\theta}_c, \boldsymbol{\theta}_u, \mathbf{W}$, and $\boldsymbol{\delta}_0^{z\,(0)} \sim 0.001 \cdot \mathcal{N}(\mathbf{0}, \mathbf{I})$, $\forall\, 1 \leq z \leq Z$. Set iteration number $\iota = 1$.
while not at end of training do
    Sample an episode $(\mathcal{S}, \mathcal{Q}) \sim \mathcal{D}$ and set $\mathcal{B} = \mathcal{Q}$
    Form class-wise sets $\mathcal{B}_z = \mathcal{Q}_z$, $\forall\, 1 \leq z \leq Z$
    for $z = 1, 2, \ldots, Z$ (in parallel) do
        Compute cross-branch class-wise global adv. initialization perturbations (GAIP) by Eq. (10): $\boldsymbol{\delta}_0^{z\,(\iota)} = h(\mathcal{B}_z^{\iota}; \boldsymbol{\delta}_0^{z\,(\iota-1)}; \alpha)$
    end for
    Compute relative adversarial robustness scores $\kappa_s$ and $\kappa_c$ (harmonization module) by Eq. (11)
    Obtain downweighted learning rates $\eta'_s$ and $\eta'_c$ from $\eta_s$ and $\eta_c$ via Eq. (12) given $\kappa_s$, $\kappa_c$ and $\tau$
    Compute gradient update for the similarity learner: $\boldsymbol{\theta}_s \leftarrow \boldsymbol{\theta}_s - \eta'_s \nabla_{\boldsymbol{\theta}_s} \mathcal{L}_s(\mathcal{Q}, \mathbf{M}(\mathcal{S}))$ where adversarial samples for $\mathcal{L}_s$ in Eq. (3) are seeded with $\boldsymbol{\delta}_0^{z\,(\iota)}$
    Compute grad. update for the class concept learner: $(\boldsymbol{\theta}_s, \mathbf{W}) \leftarrow (\boldsymbol{\theta}_c, \mathbf{W}) - \eta'_c \nabla_{(\boldsymbol{\theta}_c, \mathbf{W})} \mathcal{L}_c(\mathcal{B}, \mathbf{W})$ where adv. samples for $\mathcal{L}_c$ in Eq. (6) are seeded with $\boldsymbol{\delta}_0^{z\,(\iota)}$
    Co-distillation step by Eq. (8): $\boldsymbol{\theta}_u \leftarrow \beta \boldsymbol{\theta}_u + (1 - \beta)\,[\gamma \boldsymbol{\theta}_s + (1 - \gamma) \boldsymbol{\theta}_c]$
    if epoch $t \geq T$ and $(\iota \bmod m) = 0$ then
        Distribute $\boldsymbol{\theta}_s \leftarrow \boldsymbol{\theta}_u$ and $\boldsymbol{\theta}_c \leftarrow \boldsymbol{\theta}_u$ (copy or EMA)
    end if
    $\iota \leftarrow \iota + 1$
end while
return Co-distilled model parameters $\boldsymbol{\theta}_u$.
```

3.4. RESISTANCE (Our Algorithm)

Algorithm 1 outlines the key steps of RESISTANCE. Our similarity learner is inspired by Adversarial Querying (AQ) [14] with a distinct training loss, whereas our concept learner is inspired by TRADES [49], a well-established adversarial training method. Nevertheless, both learners can be customized to other FSL problems. The adversary generation per branch can be performed as a multi-step (default) or single-step approach for computational efficiency. However, a good initialization for a single-step strategy is needed. We opt for our global adversarial initialization perturbations instead of random initialization to mitigate the risks of suffering from bad local optima [1, 16, 21].

4. Experiments

Below, we compare RESISTANCE to the state-of-the-art adversarially robust FSL methods (natural and robust performance). We show RESITANCE works well in the cross-domain and single-step adversarial generation settings.

4.1. Experimental Setups

Datasets. We evaluate our method on three standard FSL datasets: Mini-ImageNet [42], CIFAR-FS [4], & FC100 [28].

Implementation details. Following recent studies [9, 14, 46], we adopt either Conv-4 [42] or ResNet-12 [15] as backbones. Each few-shot task consists of 5-way 1/5-shot samples for the support set and 15 instances for each class in the query set. The "distribute back" process starts at epoch $T = 40$ with the frequency $m = 10$. The ℓ_∞-norm perturbation radius is set as $\epsilon = 8/255$ with step size $\alpha = 2/255$. Regularization hyper-parameters are set as $\beta = 0.99$, $\gamma = 0.5$, and $\tau = 0.5$. Appendix A provides further details.

4.2. Results

Performance of RESISTANCE. We compare RESISTANCE with state-of-the-art adversarially robust few-shot learning approaches in 5-way 1/5-shot settings, as shown in Tables 1 and 2. We report clean accuracy and robust accuracy against three strong white-box attacks: PGD [24] with 20 steps, CW [5], and Auto Attack (AA) [6] for a comprehensive evaluation. All the results are obtained over 2,000 randomly sampled few-shot tasks with adaptive attacks for fairness. Tables 1 and 2 show that RESISTANCE improves robustness across these three few-shot benchmarks and also enjoys superior natural performance compared with other approaches. RESISTANCE also exhibits good performance w.r.t. different network architectures across diverse settings. In the 5-shot scenario for CIFAR-FS, our approach enjoys a substantial improvement of 9% enhancement in clean accuracy, as well as a 7.8% boost in AA robustness with ResNet-12 over existing methods. For 1-shot clean accuracy, we observe over 3% gain across all the datasets using ResNet-12.

Robustness against diverse attack strengths. We here study the adversarial robustness of RESISTANCE under varying attack strengths (perturbation radii). Table 3 shows that our method consistently remains robust when confronted with adversaries of varied perturbation magnitudes, outperforming both similarity and class concept learning methods. This indicates that RESISTANCE can effectively capture the robustness of both learning paradigms.

Single-step adversarial generation. Existing robust FSL methods primarily suffer from low training efficiency due to multi-step adversary generation. To relieve this issue, we explore single-step strategies [1, 7, 16, 33, 48] with robust FSL. Table 4 shows that RESISTANCE with single-step strategies achieves non-trivial adversarial robustness close

Table 1. Comparison of our RESISTANCE on Mini-ImageNet, CIFAR-FS, FC100 with other adversarially robust few-shot learning methods in the **5-way 1-shot** setting. We report both clean accuracy (%) and robust accuracy (%) with the perturbation radius $\epsilon = 8/255$.

Model	Method	Mini-ImageNet				CIFAR-FS				FC100			
		Clean	PGD	CW	AA	Clean	PGD	CW	AA	Clean	PGD	CW	AA
Conv-4	AQ [14]	33.67	18.52	17.53	16.66	42.66	26.33	25.35	25.04	29.51	18.64	17.66	16.64
	R-MAML [46]	33.98	25.69	24.73	22.97	33.51	27.61	27.12	15.39	23.52	16.31	15.08	14.48
	ST [36]	34.62	27.46	26.06	22.71	41.77	32.02	30.91	29.14	28.92	20.19	20.05	19.90
	GR [9]	35.38	**28.37**	27.12	23.83	44.51	37.45	36.53	34.26	31.82	23.73	23.28	21.26
	DFSL [19]	35.39	28.14	26.61	23.05	44.13	36.87	35.29	33.88	31.74	23.34	22.06	20.57
	RESISTANCE	**37.40**	28.26	**27.24**	**25.15**	**45.78**	**38.01**	**37.35**	**35.76**	**33.81**	**26.31**	**25.22**	**22.75**
ResNet-12	AQ [14]	41.89	20.53	18.38	17.81	47.40	29.55	28.42	27.46	31.72	19.44	18.85	17.14
	R-MAML [46]	37.52	27.46	33.47	24.14	41.78	28.33	28.86	25.27	28.25	18.48	16.50	17.57
	ST [36]	43.97	30.13	29.42	28.47	47.24	35.61	34.67	32.86	33.64	23.69	23.28	20.31
	GR [9]	45.81	35.18	34.53	32.61	48.13	39.29	37.36	35.76	34.21	25.05	24.17	21.94
	DFSL [19]	47.16	34.60	33.77	31.62	50.74	39.37	37.83	35.71	34.80	25.20	24.81	21.09
	RESISTANCE	**50.28**	**36.06**	**34.74**	**33.71**	**55.78**	**44.05**	**42.55**	**41.57**	**37.75**	**26.53**	**25.37**	**23.18**

Table 2. Comparison of our RESISTANCE method on Mini-ImageNet, CIFAR-FS, FC100 with other adversarial few-shot learning methods in the **5-way 5-shot** setting. We report both clean accuracy (%) and robust accuracy (%) with the perturbation radius $\epsilon = 8/255$.

Model	Method	Mini-ImageNet				CIFAR-FS				FC100			
		Clean	PGD	CW	AA	Clean	PGD	CW	AA	Clean	PGD	CW	AA
Conv-4	AQ [14]	50.12	28.16	27.21	24.68	57.63	39.58	38.69	37.17	35.19	24.76	22.80	21.08
	R-MAML [46]	50.76	34.19	29.61	28.31	52.75	32.66	31.47	19.25	38.56	17.67	15.91	18.75
	ST [36]	51.23	33.23	30.84	29.07	55.61	40.21	40.15	39.95	40.69	30.65	27.39	27.06
	GR [9]	50.93	37.95	35.90	31.37	58.31	47.95	46.45	45.09	41.32	32.92	30.70	29.09
	DFSL [19]	51.10	36.23	35.94	30.31	58.89	47.42	46.62	44.38	41.74	31.81	29.99	28.44
	RESISTANCE	**52.23**	**40.24**	**38.55**	**35.81**	**60.05**	**48.37**	**47.00**	**45.89**	**44.63**	**35.15**	**33.73**	**30.07**
ResNet-12	AQ [14]	64.47	30.80	29.62	25.72	65.78	44.01	42.54	41.56	41.07	25.68	24.86	22.13
	R-MAML [46]	62.75	45.78	43.88	36.12	65.61	34.77	33.15	27.77	42.25	24.39	20.49	20.08
	ST [36]	61.65	47.85	45.98	45.23	64.44	46.16	44.26	43.19	44.57	32.18	30.72	28.33
	GR [9]	64.60	50.71	47.52	47.59	66.99	52.66	50.61	50.91	46.12	34.27	32.00	30.98
	DFSL [19]	64.95	50.83	47.23	46.50	65.84	53.90	51.25	50.64	47.73	34.63	32.36	30.97
	RESISTANCE	**68.79**	**53.84**	**51.47**	**50.52**	**74.83**	**61.61**	**59.64**	**58.76**	**51.69**	**37.51**	**35.70**	**34.66**

Table 3. (Auto-Attack) robust accuracy (%) of different attack radii using ResNet-12 on Mini-ImageNet and CIFAR-FS.

Radius ϵ	Method	Mini-ImageNet		CIFAR-FS	
		1-shot	5-shot	1-shot	5-shot
4/255	R-MAML [46]	31.67	47.21	30.96	40.43
	GR [9]	35.77	52.63	40.04	55.82
	DFSL [19]	36.39	53.45	41.12	56.92
	RESISTANCE	**39.24**	**58.57**	**46.07**	**64.18**
6/255	R-MAML [46]	28.65	42.94	27.16	34.91
	GR [9]	33.75	50.95	36.85	52.98
	DFSL [19]	33.98	50.42	37.45	53.20
	RESISTANCE	**37.06**	**54.85**	**43.39**	**61.35**
10/255	R-MAML [46]	25.08	35.73	22.81	26.12
	GR [9]	28.01	44.99	33.23	48.47
	DFSL [19]	26.83	43.08	32.98	48.03
	RESISTANCE	**29.76**	**47.33**	**38.57**	**56.18**
12/255	R-MAML [46]	23.89	32.75	21.30	25.06
	GR [9]	26.31	40.92	31.14	47.46
	DFSL [19]	25.27	38.69	29.19	45.66
	RESISTANCE	**27.65**	**44.10**	**36.01**	**52.35**

Table 4. Extension with single-step adversary generation strategies on Mini-ImageNet using ResNet-12. We report clean and (Auto-Attack) robust accuracy with the average training time.

Method	Adversary Type	1-shot		5-shot		Time(h)
		Clean	Robust	Clean	Robust	
R-MAML [46]	Multi-step	37.52	24.14	62.75	36.12	15.6
	N-FGSM [7]	33.61	21.27	59.72	34.53	4.8
	RS-FGSM [48]	33.86	21.22	59.85	34.48	4.8
	GradAlign [1]	34.04	21.46	60.50	34.93	8.3
GR [9]	Multi-step	45.81	32.61	64.60	47.59	10.7
	N-FGSM [7]	40.13	28.17	59.44	44.71	3.1
	RS-FGSM [48]	41.49	26.35	60.57	43.24	3.1
	GradAlign [1]	40.63	27.42	59.15	44.03	5.9
RESISTANCE	Multi-step	50.28	33.71	68.79	50.52	16.9
	N-FGSM [7]	48.84	32.70	68.40	50.35	5.3
	RS-FGSM [48]	49.24	30.26	67.81	48.70	5.3
	GradAlign [1]	49.07	31.33	68.48	49.19	9.5

to its multi-step performance. Moreover, single-step RESISTANCE achieves even better performance on both clean accuracy and robustness compared to existing multi-step FSL approaches. Appendix B.1 provides further details.

Cross-domain FSL robustness. We here investigate the cross-domain transferability of RESISTANCE. Following the settings from [9, 17], we train and test the robust embedding model on disjoint domains with distinct resolutions and disjoint categories. We report accuracy on clean and adversarial samples from test domains in Table 5. We observe that RESISTANCE obtains better transferable robustness and natural performance in diverse cross-domain settings, showing that RESISTANCE remains adversarially robust under domain shifts. More details are in Appendix B.2.

4.3. Ablation Studies

Impact of each module. Below, we investigate the impact of component modules in RESISTANCE: (i) co-distillation of similarity and class concept learners (Co-dist.) in Section 3.1, (ii) cross-branch class-wise global adversarial ini-

Table 5. Cross-domain robustness between Mini-ImageNet (**M**), CIFAR-FS (**C**), and FC100 (**F**) datasets using ResNet-12.

Transfer	Method	1-shot			5-shot		
		Clean	PGD	AA	Clean	PGD	AA
M → C	AQ [14]	43.96	26.36	22.30	61.05	37.33	30.97
	GR [9]	44.13	34.67	32.13	60.86	45.17	42.03
	TROBA [17]	43.20	32.47	30.81	62.44	46.24	43.75
	RESISTANCE	**48.04**	**38.65**	**36.54**	**64.13**	**53.42**	**50.26**
M → F	AQ [14]	**36.08**	18.71	14.14	47.66	25.31	19.45
	GR [9]	35.16	26.40	24.30	45.91	33.92	30.79
	TROBA [17]	34.09	24.42	21.65	45.51	34.05	31.56
	RESISTANCE	35.78	**27.63**	**24.34**	**47.88**	**37.49**	**35.45**
C → M	AQ [14]	36.25	11.15	8.80	**56.90**	19.10	14.20
	GR [9]	36.65	24.60	20.12	50.73	33.19	30.17
	TROBA [17]	37.48	21.59	18.40	52.46	29.27	26.92
	RESISTANCE	**38.55**	**25.08**	**21.65**	56.04	**39.19**	**34.96**

Table 6. Ablation study using ResNet-12 of three key components of RESISTANCE for 5-way 5-shot robustness (%) on CIFAR-FS.

	Co-dist.	GAIP	Harm.	Clean	PGD-20	AA
1				60.22	46.95	45.84
2	✓			68.12	55.14	53.07
3	✓	✓		73.17	58.99	55.72
4	✓		✓	71.46	60.24	56.20
5	✓	✓	✓	74.83	61.61	58.76

tialization perturbations (GAIP) in Section 3.2, and (iii) branch robustness harmonization (Harm.) in Section 3.3. We provide both clean and robust accuracy on CIFAR-FS using ResNet-12 in the 5-way 5-shot setting in Table 6.

Our baseline approach (first row in Table 6) relies on the robust class concept learning in Eq. (6). The co-distillation yields a substantial improvement w.r.t. the clean and robust accuracy, demonstrating the effectiveness in leveraging the respective strengths of similarity and class concept learning. Furthermore, the branch harmonization strategy and GAIP also improve adversarial robustness. The integration of all these components leads to the best FSL performance on clean samples and their adversarial counterparts.

Combining various types of learners. We here demonstrate that RESISTANCE requires both the similarity and class concept learners instead of co-distillation between two learners of the same type (with differently initialized backbones–as in ensemble learning). Table 7 shows that *similarity & concept* learners outperform other combinations, *i.e*., *similarity & similarity* learners and *concept & concept* learners. As the total parameter count is the same irrespective of the variant in the table, we conclude that improvements yielded by RESISTANCE are due to the complementary nature of similarity and class concept learners rather than doubling the number of learnable parameters.

Impact of the backbone size on similarity, concept, and co-distillation. Below, we show that the improved performance of RESISTANCE is not due to the increased model complexity of two learners. Table 8 shows results for *similarity only*, *concept only* and the similarity and class concept *co-distillation*. We used larger network architectures, such as ResNet-18 and ResNet34, and indicated the total number of parameters per configuration. RESISTANCE, denoted as co-distillation, surpasses results of either similarity or concept learning alone that are based on larger backbones.

Table 7. Comparison of co-distillation components on clean and (Auto-Attack) robust accuracy using ResNet-12 on CIFAR-FS.

Co-distillation Components	1-shot		5-shot	
	Clean	Robust	Clean	Robust
similarity & similarity	49.50	31.75	69.63	46.33
class concept & class concept	47.90	33.04	67.37	51.17
similarity & class concept	55.78	41.57	74.83	58.76

Table 8. Similarity *vs*. class concept learning with different backbones on clean and (Auto-Attack) robust accuracy on CIFAR-FS.

Backbone	Number of Parameters	Paradigm	1-shot		5-shot	
			Clean	Robust	Clean	Robust
ResNet-18	11.7 M	similarity	49.97	30.88	68.57	45.65
		concept	47.37	32.17	66.15	48.81
ResNet-34	21.8 M	similarity	52.19	34.03	72.44	49.85
		concept	50.07	36.71	69.92	53.61
ResNet-12	12.4 M	**co-distillation**	55.78	41.57	74.83	58.76

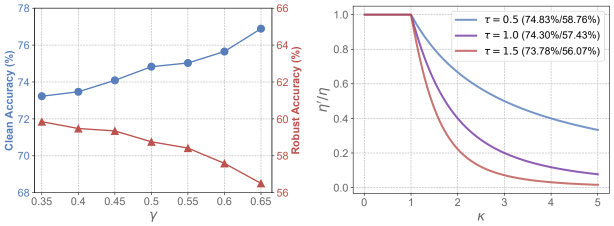
Table 9. Diverse adversarial initialization types for RESISTANCE using ResNet-12 for 5-way 5-shot robustness (%) on CIFAR-FS.

	Cross-branch	Class-wise	1-shot		5-shot	
			Clean	Robust	Clean	Robust
1			53.22	37.91	71.46	56.20
2	✓		52.06	39.42	70.21	58.49
3		✓	55.13	38.84	73.35	57.36
4	✓	✓	55.78	41.57	74.83	58.76

Impact of adversarial initialization perturbations. In addition to our cross-branch class-wise global adversarial initialization perturbations, denoted as GAIP in previous sections, we explore other adversarial initialization strategies. Table 9 shows that *the cross-branch strategy* alone improves adversarial robustness. Thus, the shared adversarial perturbation prior across both branches helps achieve better optima when performing the gradient ascent to obtain individual untargeted adversarial samples in each branch. Furthermore, *the class-wise strategy* produces a universal perturbation prior per category, inherently learning class-specific information. Thus, by combining both the cross-branch and class-wise strategies, GAIP helps improve performance on clean samples and their adversarial counterparts.

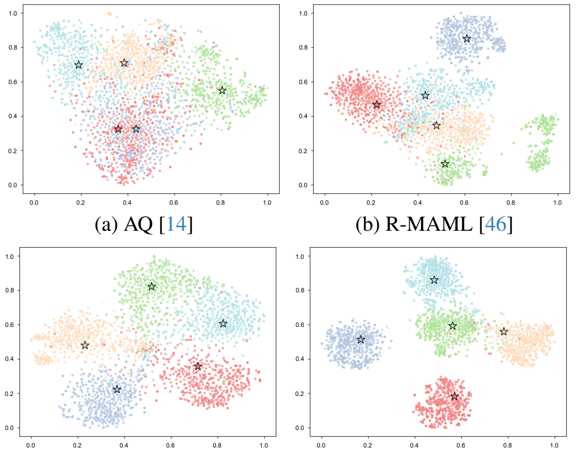
Similarity *vs*. class concept learner trade-off. Balancing the natural performance and adversarial robustness has been well explored in standard adversarial training [10, 30, 49] but not within the few-shot scenarios. To fill this gap, we study the effect of factor γ that balances similarity learning and class concept learning. Fig. 3a shows that the clean accuracy goes higher for larger values of γ, albeit sacrificing adversarial robustness. Conversely, enhancing the adversarial robustness correlates with a drop in natural performance.

For RESISTANCE, γ helps balance the interplay be-



(a) (b)

Figure 3. (a) Sensitivity of RESISTANCE w.r.t. γ: the balancing factor between the similarity and class concept learners. We report clean and (Auto-attack) robust accuracy on CIFAR-FS. (b) Ratio of downweighted and original learning rate η'/η (similarity and/or class concept) as a function of steepness factor τ and the relative adversarial robustness score κ. We also provide 5-shot clean/robust accuracy (CIFAR-FS) in the legend.



(a) AQ [14] (b) R-MAML [46]

(c) GR [9] (d) **RESISTANCE**

Figure 4. t-SNE visualization of features from 500 randomly sampled images per class on CIFAR-FS. '★' are support features. Tiny dots are query features–their color indicates the class label.

tween similarity and class concept learning, which differs from the standard balancing of the natural risk and the boundary risk [49]. Fig. 1b shows that both learners enjoy complementary performance w.r.t. different attack strengths. We benefit from such a complementary nature of both learners by adjusting γ and harmonizing the robust optimization of both learners (see Section 3.3). Fig. 3b illustrates the impact of different steepness factors τ on the downweighted learning rate (ratio of η'/η). Appendix D provides further analyses of other hyper-parameters.

4.4. Visualizations

Figure 4 presents t-SNE visualizations [41] of feature representations extracted from the test set. We randomly sampled 500 images per class (5 classes in total). Fig. 4d shows that features of RESISTANCE enjoy lower intra-class variance, resulting in compact class-wise clusters and a large inter-class separation. Following our motivation, Fig. 4a and 4b visualize feature spaces of two similarity learners, each containing several cluster-like regions per class. Fig. 4c uses a class concept learner with relatively clear class margins and class-wise clusters. Hence, co-distilling both learner types should benefit a unified embedding model.

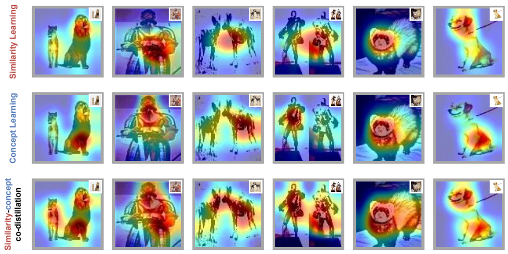


Figure 5. Grad-CAM [32] attention visualizations of adversarial samples from Mini-ImageNet using the backbone of ResNet-12.

Table 10. Various branch fusion strategies on clean and (Auto-Attack) robust accuracy using ResNet-12 on CIFAR-FS.

Fusion Strategies	1-shot		5-shot	
	Clean	Robust	Clean	Robust
Prediction Ensemble	50.12	36.27	67.80	51.73
Feature Ensemble	54.34	23.84	72.06	32.83
Multi-teacher Distillation	53.78	38.62	71.30	54.85
RESISTANCE	55.78	41.57	74.83	58.76

4.5. Further Analysis

Figure 5 shows class-wise activation maps of similarity learning, class concept learning, and RESISTANCE under adversaries. Activation regions among similarity and class concept learners differ due to their complementary nature. Our method enjoys holistic activations covering the entire target objects due to the integrated complementary learners.

Table 10 shows that our RESISTANCE outperforms other strategies of fusing similarity and class concept learners as follows. 1) *Prediction Ensemble*: average voting on predictions of learners. 2) *Feature Ensemble*: concatenated feature representations of few-shot learners. 3) *Multi-teacher Distillation*: feature-based knowledge distillation from learners. Appendix B.3 provides more details.

5. Conclusions

In this paper, we propose RESISTANCE, a novel adversarially robust few-shot learning method that effectively co-distills the similarity and class concept learners, whose decision boundaries are highly complementary. We also design cross-branch class-wise adversarial perturbations and a robustness harmonization module to promote the uniformity and balance of adversarial robustness. Extensive experiments demonstrate the efficacy and generalization ability of RESISTANCE in diverse settings with further efficiency gains via single-step adversary generation strategies.

Acknowledgments. This work is partially supported by the National Research Foundation Singapore under its AI Singapore Program [Award No.:AISG2-GC-2023-007], CFAR, IHPC, A*STAR, and SCSE at Nanyang Technological University. PK is supported by CSIRO's Science Digital.

References

[1] Maksym Andriushchenko and Nicolas Flammarion. Understanding and improving fast adversarial training. *Advances in Neural Information Processing Systems*, 33:16048–16059, 2020. 5, 6

[2] Tao Bai, Jinqi Luo, Jun Zhao, Bihan Wen, and Qian Wang. Recent advances in adversarial training for adversarial robustness. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI*, pages 4312–4321, 2021. 2

[3] Philipp Benz, Chaoning Zhang, Adil Karjauv, and In So Kweon. Universal adversarial training with class-wise perturbations. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2021. 2

[4] 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. 5

[5] 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. 5

[6] 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. 5

[7] 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. 5, 6

[8] Josip Djolonga, Jessica Yung, Michael Tschannen, Rob Romijnders, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Matthias Minderer, Alexander D'Amour, Dan Moldovan, et al. On robustness and transferability of convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16458–16468, 2021. 2

[9] 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. 1, 2, 5, 6, 7, 8

[10] Junhao Dong, Seyed-Mohsen Moosavi-Dezfooli, Jianhuang Lai, and Xiaohua Xie. The enemy of my enemy is my friend: Exploring inverse adversaries for improving adversarial training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24678–24687, 2023. 2, 7

[11] Junhao Dong, Yuan Wang, Jianhuang Lai, and Xiaohua Xie. Restricted black-box adversarial attack against deepfake face swapping. *IEEE Transactions on Information Forensics and Security*, 18:2596–2608, 2023. 1

[12] Junhao Dong, Lingxiao Yang, Yuan Wang, Xiaohua Xie, and Jianhuang Lai. Toward intrinsic adversarial robustness through probabilistic training. *IEEE Transactions on Image Processing*, 32:3862–3872, 2023. 2

[13] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017. 2

[14] 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. 1, 2, 5, 6, 7, 8

[15] 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. 5

[16] 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. 5

[17] Minseon Kim, Hyeonjeong Ha, and Sung Ju Hwang. Few-shot transferable robust representation learning via bilevel attacks. *arXiv preprint arXiv:2210.10485*, 2022. 6, 7

[18] Piotr Koniusz, Fei Yan, Philippe-Henri Gosselin, and Krystian Mikolajczyk. Higher-order Occurrence Pooling on Mid- and Low-level Features: Visual Concept Detection. Technical report, 2013. 1

[19] Wenbin Li, Lei Wang, Xingxing Zhang, Lei Qi, Jing Huo, Yang Gao, and Jiebo Luo. Defensive few-shot learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):5649–5667, 2023. 1, 6

[20] Zhibin Li, Piotr Koniusz, Lu Zhang, Daniel Edward Pagendam, and Peyman Moghadam. Exploiting field dependencies for learning on categorical data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11):13509–13522, 2023. 1

[21] Shengchao Liu, Dimitris Papailiopoulos, and Dimitris Achlioptas. Bad global minima exist and sgd can reach them. *Advances in Neural Information Processing Systems*, 33:8543–8552, 2020. 5

[22] Changsheng Lu and Piotr Koniusz. Few-shot keypoint detection with uncertainty learning for unseen species. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 19394–19404. IEEE, 2022. 1

[23] Changsheng Lu and Piotr Koniusz. Detect any keypoints: An efficient light-weight few-shot keypoint detector. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(4):3882–3890, 2024. 1

[24] 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. 3, 5

[25] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1765–1773, 2017. 2, 4

[26] Yao Ni and Piotr Koniusz. NICE: NoIse-modulated Consistency rEgularization for Data-Efficient GANs. In *Advances in Neural Information Processing Systems*, pages 13773–13801. Curran Associates, Inc., 2023. 2

[27] Yao Ni and Piotr Koniusz. CHAIN: Enhancing Generalization in Data-Efficient GANs via lipsCHitz continuity constrAIned Normalization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024*. IEEE, 2024. 2

[28] 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. 5

[29] Yao Qin, Nicholas Carlini, Garrison Cottrell, Ian Goodfellow, and Colin Raffel. Imperceptible, robust, and targeted adversarial examples for automatic speech recognition. In *International conference on machine learning*, pages 5231–5240. PMLR, 2019. 1

[30] Rahul Rade and Seyed-Mohsen Moosavi-Dezfooli. Reducing excessive margin to achieve a better accuracy vs. robustness trade-off. In *The Tenth International Conference on Learning Representations, ICLR*, 2022. 2, 7

[31] Sylvestre-Alvise Rebuffi, Sven Gowal, Dan Andrei Calian, Florian Stimberg, Olivia Wiles, and Timothy A Mann. Data augmentation can improve robustness. *Advances in Neural Information Processing Systems*, 34:29935–29948, 2021. 2

[32] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 8

[33] 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. 5

[34] Ali Shafahi, Mahyar Najibi, Zheng Xu, John Dickerson, Larry S Davis, and Tom Goldstein. Universal adversarial training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5636–5643, 2020. 2, 4

[35] Yisheng Song, Ting Wang, Puyu Cai, Subrota K Mondal, and Jyoti Prakash Sahoo. A comprehensive survey of few-shot learning: Evolution, applications, challenges, and opportunities. *ACM Computing Surveys*, 2023. 1

[36] Akshayvarun Subramanya and Hamed Pirsiavash. Adversarially robust few-shot learning through simple transfer. In *ECCV Workshops: Adversarial Robustness in the Real World*. Springer, 2022. 1, 6

[37] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1199–1208, 2018. 2

[38] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *2nd International Conference on Learning Representations, ICLR 2014*, 2014. 1

[39] Hao Tang, Xingwei Liu, Shanlin Sun, Xiangyi Yan, and Xiaohui Xie. Recurrent mask refinement for few-shot medical image segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3918–3928, 2021. 1

[40] Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: a good embedding is all you need? In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 266–282. Springer, 2020. 2

[41] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (11), 2008. 8

[42] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. *Advances in neural information processing systems*, 29, 2016. 5

[43] Hongjun Wang and Yisen Wang. Generalist: Decoupling natural and robust generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20554–20563, 2023. 2

[44] Lei Wang and Piotr Koniusz. Uncertainty-dtw for time series and sequences. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXI*, pages 176–195. Springer, 2022. 1

[45] Lei Wang and Piotr Koniusz. Temporal-viewpoint transportation plan for skeletal few-shot action recognition. In *Asian Conference on Computer Vision*, page 307–326, Berlin, Heidelberg, 2023. Springer-Verlag. 1

[46] 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. 5, 6, 8

[47] Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, 53(3):1–34, 2020. 1, 2

[48] 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. 5, 6

[49] 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. 2, 5, 7, 8

[50] Hongguang Zhang, Hongdong Li, and Piotr Koniusz. Multilevel second-order few-shot learning. *IEEE Trans. Multim.*, 25:2111–2126, 2023. 1

[51] Hao Zhu and Piotr Koniusz. Transductive few-shot learning with prototype-based label propagation by iterative graph refinement. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 23996–24006. IEEE, 2023. 1