Improving Adversarially Robust Few-shot Image Classification with Generalizable Representations

Junhao Dong[1], Yuan Wang[1], Jianhuang Lai[1,2,3] and Xiaohua Xie[1,2,3*]
[1]School of Computer Science and Engineering, Sun Yat-Sen University, China
[2]Guangdong Province Key Laboratory of Information Security Technology, China
[3]Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, China
{dongjh8, wangy975}@mail2.sysu.edu.cn, {stsljh, xiexiaoh6}@mail.sysu.edu.cn

Abstract

Few-Shot Image Classification (FSIC) aims to recognize novel image classes with limited data, which is significant in practice. In this paper, we consider the FSIC problem in the case of adversarial examples. This is an extremely challenging issue because current deep learning methods are still vulnerable when handling adversarial examples, even with massive labeled training samples. For this problem, existing works focus on training a network in the meta-learning fashion that depends on numerous sampled few-shot tasks. In comparison, we propose a simple but effective baseline through directly learning generalizable representations without tedious task sampling, which is robust to unforeseen adversarial FSIC tasks. Specifically, we introduce an adversarial-aware mechanism to establish auxiliary supervision via feature-level differences between legitimate and adversarial examples. Furthermore, we design a novel adversarial-reweighted training manner to alleviate the imbalance among adversarial examples. The feature purifier is also employed as post-processing for adversarial features. Moreover, our method can obtain generalizable representations to remain superior transferability, even facing cross-domain adversarial examples. Extensive experiments show that our method can significantly outperform state-of-the-art adversarially robust FSIC methods on two standard benchmarks.

1. Introduction

Current deep learning methods have made significant progress on several computer vision tasks [8, 11, 12, 14, 21]. However, these methods often depend on a high computational budget and abundant data, which is costly to collect in the real-world setting. To address this issue, few-shot learning aims at developing efficient learning algorithms with limited data [5, 22, 25–27]. In practice, there also exists a potential security threat to deep neural networks from adversarial examples [28]. Adversarial examples are some tailored examples with almost no differences from natural examples in human vision, but can strongly disturb the inference of neural networks [3, 13]. Due to the scarce data for building the robust decision boundary, current deep learning methods even suffer more severe attacks from adversarial examples in the few-shot setting [7].

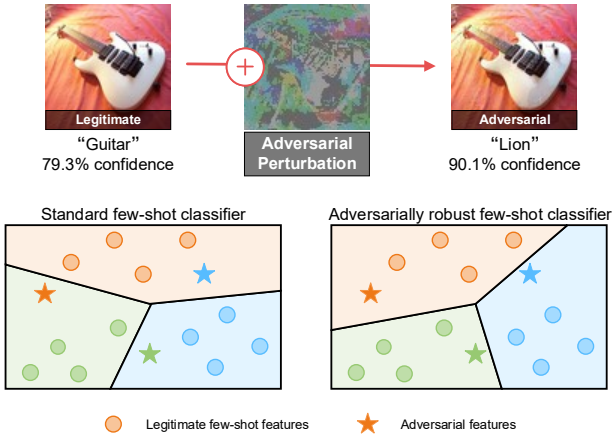

Figure 1. The adversarial example can be obtained by adding a nearly undetectable adversarial perturbation (magnified for visibility) to the legitimate example, which can cause a misclassification with high confidence. Compared with the standard few-shot classifier, the adversarially robust classifier learns a robust decision boundary to classify adversarial examples correctly.

The primary purpose of adversarially robust Few-Shot Image Classification (FSIC) is to build models that perform well in standard few-shot classification and are simultaneously robust against adversarial examples, as shown in Figure 1. Because of the limited data and the existence of adversarial examples, adversarially robust FSIC still remains a challenging problem. Nevertheless, existing researches

*Corresponding Author

on this problem are still rare and mainly based on meta-learning [7, 31, 37], which aims to learn a model from abundant few-shot tasks and then generalize it to unforeseen few-shot tasks. Each task consists of limited training examples and query examples from the same distribution. The dataset is divided into the meta-training set and the meta-test set of disjoint categories. The meta-learning model is trained on the meta-training set to quickly adapt to new tasks, while the meta-test set is for evaluating few-shot accuracy. However, there still exists a considerable gap of label spaces between the meta-training and the meta-test set. Learning from excessive few-shot tasks may induce overfitting on the source label space and thus aggravate the performance on adversarial examples of different label spaces, especially in the cross-domain scenario.

In this paper, beyond the meta-learning manner, we propose a novel adversarially robust FSIC framework, which learns a robust embedding model and generalizes it to unforeseen adversarial few-shot classification tasks. Based on the observation that features extracted from adversarial examples are non-robust to trained models [13], we design an auxiliary adversarial-aware module to learn the nuance between legitimate examples and corresponding adversarial examples. On account of the imbalance among adversarial examples, we propose a novel adversarial-reweighted method according to the loss variation. We also append a simple but effective postprocessing module to purify adversarial features.

To fully evaluate the effectiveness of our methods, we conduct extensive experiments on two standard few-shot classification benchmarks: miniImageNet [30] and CIFAR-FS [1]. The evaluation contains both accuracy on legitimate examples and their corresponding adversarial examples in the few-shot setting. In addition, we validate the robustness of our methods, which is conducted under various attack strengths. We also conduct cross-domain transfer experiments to demonstrate the generalizability of our proposed methods against adversarial examples.

The main contributions of our work can be summarized as follows:

- We propose a new baseline on adversarially robust FSIC by directly learning a robust embedding model, eliminating complicated meta-training steps.
- We design an adversarial-aware method for auxiliary supervision between legitimate and adversarial examples. To address the imbalance among adversarial examples during training, we propose a novel adversarial-reweighted mechanism according to the loss variation. For postprocessing, we introduce a feature purification module to mitigate adversarial effects.
- Extensive experiments demonstrate that our algorithm achieves a new state-of-the-art performance on adversarially robust FSIC. Our method can also provide additional benefits on robustness to different attack strengths and generalizability in the cross-domain scenario simultaneously.

2. Related works

Few-shot classification. The recent few-shot classification methods are mainly based on meta-learning [4, 5, 16, 23, 26, 41], which trains a meta-model with plentiful sampled tasks to adapt to unforeseen few-shot tasks quickly. Model-Agnostic Meta-Learning (MAML) [5] first proposes a general meta-learning paradigm to find a superior model initialization, which can rapidly adapt to novel few-shot tasks. To further extend MAML to deep models without overfitting, [26] combines meta-learning and transfer learning to construct meta-classification on deeper models than before. Furthermore, [4] explore the relationship between the sampled episodes(tasks) in the meta-training stage and design a novel meta-learning framework via modeling episode-level relationships. On the other hand, there also exist some other few-shot classification methods via simply learning a generalizable embedding [17, 20, 22, 29]. Recently, [29] propose a simple baseline combining with self-supervised tasks and self-distillation, which can learn generalizable feature representations. Furthermore, [22] explore both equivariance and invariance feature representations and present a novel training strategy that jointly enforcing equivariance and invariance through geometry transformation. Similarly, our adversarial-aware module can also be viewed as an equivariance acting on the adversarial transformation, while it is a non-geometric transformation for auxiliary feature-level supervision on adversarial examples.

Adversarial defense. A range of defense methods have been proposed to improve model robustness against adversarial examples. Among them, adversarial training [10, 19, 34, 38, 39] is one of the most effective strategy, which can regularize deep neural networks to obtain better performance on classifying adversarial examples. [10] first propose adversarial training with single-step-based adversarial examples, while [19] extended it to multi-step-based adversarial examples to obtain better adversarial robustness. Besides, there also exists other defending strategies against adversarial examples, for instance, randomization [33], dropout [32] and JPEG compression [18], etc. In view of the imbalance of adversarial examples during adversarial training, [39] design a weighted adversarial training method according to adversary generation steps. Nonetheless, this adversary generation steps-based weight can sometimes be too rough for reweighting. To address this problem, we further propose a preciser adversarial-reweighted method based on the loss variation.

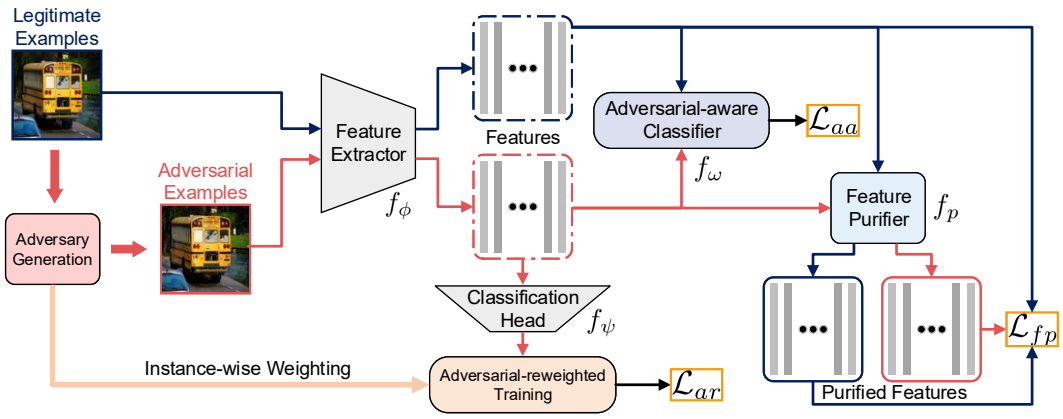

Figure 2. Overview of our proposed methods. The blue and red line signifies the legitimate and adversarial flow, respectively Adversarial examples and their corresponding weighting factors are generated from legitimate examples. Afterward, extracted feature are fed into three modules f_ω, f_p, and f_ψ for optimization.

Adversarially robust few-shot classification. Although FSIC has made outstanding progress with deep neural networks, these deep learning methods still suffer heavily from adversarial examples. Adversarially robust FSIC aims at classifying both legitimate and adversarial examples in the few-shot scenario [7, 31, 37]. The concept of adversarially robust few-shot classification was first proposed in [37], which studies meta-learning jointly with adversarial training. [7] thoroughly explore the adversarial vulnerability under the few-shot setting and propose a new meta-learning paradigm, which introduces adversarial examples during the querying step in meta-training. Furthermore, [31] extends adversarially robust FSIC to semi-supervised learning with the fast adversary generation. However, previous methods are all based on the meta-learning paradigm, which is complicated to construct abundant meta-training tasks. In comparison, Our proposed method considers a simple but effective framework via robust knowledge transfer.

3. Methods

In this section, we first describe the problem formulation of adversarially robust FSIC and adversarial training, and then present our methods. The overview of our proposed framework is shown in Figure 2.

3.1. Problem formulation

The goal of adversarially robust FSIC is to classify both legitimate and adversarial examples in the few-shot setting. We first introduce the formulation of few-shot classification, which consists of two disjoint sets $(\mathcal{D}_{train}, \mathcal{D}_{test})$ for training and evaluation. We define the training set as $\mathcal{D}_{train} = \{(\mathbf{x}_{train}, y_{train})\}$, where $y_{train} \in N_{train}$ is the image label within total N_{train} classes. The test set $\mathcal{D}_{test} = \{(\mathbf{x}_{test}, y_{test})\}$ is constructed with N_{test} classes. Note that N_{train} and N_{test} are disjoint classes for a fair comparison. For simulating the few-shot classification task, we sample K examples from each of $N \subseteq N_{test}$ classes as well as Q query examples for accuracy measurement. This settng is also known as N-way K-shot.

On this basis, adversarially robust FSIC considers an extra evaluation on adversarial examples, which can be produced by Q query images from sampled few-shot tasks. Particularly, our main network parameters consist of two parts $\theta = (\phi, \psi)$, in which ϕ represents the feature extractor and ψ represents the classification head during the training stage. Overall, we first learn an adversarially robust representation f_ϕ from the training set $\mathcal{D}_{train}$, and then transfer it to disjoint few-shot classification tasks with adversarial examples.

3.2. Adversarial examples

Adversary generation In this paper, we construct adversarial examples from legitimate examples by directly appending adversarial perturbations that are generated via Projected Gradient Descent (PGD) attack [19] on the negative loss function. Following the standard principle, we restrict the adversarial perturbation in the ℓ_∞-norm bound. Thus, adversarial examples generated through PGD attack can be formalized as follow:

$$\mathbf{x}_{adv}^{t+1} = \Pi_S \left(\mathbf{x}_{adv}^t + \alpha \cdot sign \left(\nabla_{\mathbf{x}_{adv}^t} \mathcal{L}_\theta \left(\mathbf{x}_{adv}^t, y \right) \right) \right) \quad (1)$$

where $\mathbf{x}$ and $\mathbf{x}_{adv}$ stand for legitimate examples and corresponding adversarial examples. Note that adversarial examples are limited in a ℓ_∞-norm hypersphere S of radius ϵ around legitimate examples, and $\alpha > 0$ is the step size for maximizing the loss $\mathcal{L}_\theta$ of the network parameter θ. The more powerful adversarial example can be obtained by adding the sign gradient of the existing adversarial example from the loss iteratively.

Adversarial training Apart from normally trained fashion, adversarial training is an industry standard to build adversarially robust models while remaining the accuracy on natural examples, which can be represented as a min-max optimization [19]. The inner maximization is to find the most adversarial examples, while the outer minimization aims to minimize the loss with these adversarial examples. The procedure of adversarial training is given by:

$$\min_{\theta} \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}} \left[\max_{\|\delta\|_\infty<\epsilon} \mathcal{L}_\theta \left(\mathbf{x}+\delta, y\right) \right] \tag{2}$$

where $\mathcal{D}$ is the data distribution over pairs of examples $\mathbf{x}$ and corresponding labels y, and δ is the generated adversarial perturbation within the ℓ_∞-norm bound ϵ. Adversarial training can help deep neural networks learn adversarially robust features, while natural training can not. Furthermore, adversarial training can be viewed as a regularization to make models focus on perturbation-insensitive features.

3.3. Learning adversarially robust embedding

Different from related works in adversarially robust few-shot classification [7,31,37], we propose a simple but effective method via enabling the embedding model to learn adversarially robust representations. This embedding model can later be transferred to new adversarial few-shot tasks. Our baseline method is mainly based on adversarial training, which replaces legitimate examples with adversarial examples in the training procedure. The baseline loss is defined by:

$$\mathcal{L}_{bl} = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}} \left[-\log \frac{\exp(f_\theta\left(\mathbf{x}+\delta\right)_y)}{\sum_{i=0}^{N_{train}} \exp\left(f_\theta\left(\mathbf{x}+\delta\right)_i\right)} \right] \tag{3}$$

where $f_\theta\left(\cdot\right)$ is the full network containing both the feature extractor and the classification head, which projects input examples into the label space. Note that at inference, we only preserve the feature extractor for obtaining adversarially robust embeddings. The baseline loss introduces adversarial examples to cross-entropy loss, which can enhance the adversarial robustness of the embedding model while keeping a good performance on clean images.

Current deep neural networks for classification heavily depend on abundant data, while their performances on feature representation are restricted severely under the few-shot scenario. Hence we are difficult to obtain the same adversarially robust embedding model as with abundant data. However, we consider attaching adversarial information to the feature representation, which can help differentiate between adversarial examples and legitimate examples. This can be deemed as an adversarial-aware mechanism, which introduces an auxiliary classifier with the input of feature embeddings. During the training stage, we first generate adversarial examples from legitimate examples, and then feed the feature embeddings of them into the auxiliary classifier to accomplish an adversarial self-supervision. This adversarial-aware loss can be formulated as:

$$\begin{aligned} \mathcal{L}_{aa} = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}} \Bigg[& -\log \frac{\exp(f_{\phi,\omega}\left(\mathbf{x}\right)_0)}{\sum_{i=0}^{1} \exp\left(f_{\phi,\omega}\left(\mathbf{x}\right)_i\right)} \\ & -\log \frac{\exp(f_{\phi,\omega}\left(\mathbf{x}+\delta\right)_1)}{\sum_{i=0}^{1} \exp\left(f_{\phi,\omega}\left(\mathbf{x}+\delta\right)_i\right)} \Bigg] \end{aligned} \tag{4}$$

where ω is the parameter of the auxiliary classifier, which predicts whether the input feature embedding is from the legitimate example (0) or the adversarial example (1). $f_{\phi,\omega}\left(\cdot\right)$ is the abbreviation of the network concatenation $f_\phi\left(f_\omega\left(\cdot\right)\right)$. This adversarially self-supervised learning can further enable the extracted feature to possess the equivariance of the adversarial transformation.

Based on the observation that assigning weights on adversarial training examples contributes to the final classification performance [39], we also propose a novel adversarial-reweighted mechanism via the loss variation. More precisely, as the loss value varies greatly during adversary generation, the generated example is more adversarial to the target model. Therefore, these high-adversarial examples can be viewed as key network weaknesses that should be paid more attention. Past works on adversarial-reweighted training are mainly based on gradient ascent steps during the adversary generation [6, 39], while we focus on a preciser reweighted method via the loss variation. Formally, the adversarial-reweighted training is given by:

$$\mathcal{L}_{ar} = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}} \left[-w_\theta\left(\mathbf{x}\right) \log \frac{\exp(f_\theta\left(\mathbf{x}+\delta\right)_y)}{\sum_{i=0}^{N_{train}} \exp\left(f_\theta\left(\mathbf{x}+\delta\right)_i\right)} \right] \tag{5}$$

where $w_\theta\left(\mathbf{x}\right)$ is the instance-wise adversarial weight for the network θ. Note that the weighting parameter $w_\theta\left(\mathbf{x}\right)$ is normalized in $[0,1]$. In our opinion, each adversarial example has a different degree of effect on adversarial training, which depends on its adversarial intensity to deep neural networks. Adversarial training should focus on more adversarial samples that can strongly disrupt the target network, but not less adversarial samples. Therefore, we obtain a precise instance-wise adversarial training weight according to the degree of adversarial disruption.

We can first measure the disruption of adversarial examples against the final classification during the adversary generation stage. The only measurement of the iterative gradient ascent times is too rough to get the adversarial weight, because the ascent times is the integer that is difficult to show a fine distinction. Therefore, our primary measurement is a composition of the least iterative gradient ascent

times $t(\mathbf{x})$ to alter the final prediction and the loss variation value v when producing the corresponding adversarial example $\mathbf{x}_{adv}$ from the legitimate example $\mathbf{x}$. This novel composite adversarial-reweighted mechanism can generate a preciser weight for each instance as shown as below:

$$w_\theta(\mathbf{x}) = \alpha \frac{1 + \tanh(4 - 10t(\mathbf{x})/T)}{2} + \beta \frac{v}{\max_{V_i \in V}(V_i)} \quad (6)$$

where T is the maximum gradient ascent time and V is the batch set of the loss variation values in the adversary generation. α and β are both weighting factors for the composite adversarial-reweighting. Note that both the gradient ascent time and the set of loss variation values are acquired during the generation of adversarial examples, which scarcely affect the efficiency of the adversarial training. Furthermore, the gradient ascent steps-based reweighted method can be viewed as a global reweighting for each adversarial example, while the loss variation-based mechanism focuses on the local batch for a preciser adversarial weight. Therefore, our composite adversarial-reweighted method can balance both two weights and focus on more adversarial samples.

3.4. Adversarial feature purification

To further obtain a more robust embedding model for few-shot classification, we propose an adversarial feature purification method to purify features extracted from adversarial examples. Note that the purification consists in eliminating feature-level influence of the adversarial perturbation. Besides, the feature purifier is a postprocessing module directly acting on extracted feature embeddings from both adversarial and legitimate examples, which makes the distribution of adversarial features close to the distribution of legitimate features. This is mainly based on the observation that the performance of classification on normal examples is much better than on adversarial examples with adversarial training [35]. Hence, the feature purification from adversarial features to legitimate features can promote the final classification performance indirectly. The auxiliary feature purification loss is defined as:

$$\begin{aligned}\mathcal{L}_{fp} = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}} \Big[& \mathcal{L}_{MSE}\left(f_{\phi,p}(\mathbf{x}), f_\phi(\mathbf{x})\right) \\ & + \mathcal{L}_{MSE}\left(f_{\phi,p}(\mathbf{x}+\delta), f_\phi(\mathbf{x})\right) \Big]\end{aligned} \quad (7)$$

where $\mathcal{L}_{MSE}$ is the mean squared error loss and p is the network parameter of the feature purifier. The purification not only involves removing the adversarial perturbation from an input adversarial example, but also keeps the feature consistency of the legitimate example.

Note that this purifier is a plug-and-play module, which directly acts on input feature embeddings at inference. Furthermore, these purified feature embeddings are utilized to classify novel few-shot classes. This can be viewed as a distribution alignment that establishes a feature-level transformation through the adversarial feature purifier, which can enhance the performance on adversary classification. Besides, we also ensure that feature embeddings of legitimate examples remain before and after the purification. This consistency can be specially considered as a regularization of classifying normal images in the few-shot setting.

3.5. Overall loss

In general, the whole loss in training stage can be obtained by combining all the proposed adversarial-aware loss $\mathcal{L}_{aa}$, adversarial-reweighted training loss $\mathcal{L}_{ar}$ and the feature purification loss $\mathcal{L}_{fp}$ as follow:

$$\mathcal{L} = \mathcal{L}_{ar} + \lambda_1 \mathcal{L}_{aa} + \lambda_2 \mathcal{L}_{fp} \quad (8)$$

where λ_1 and λ_2 are parameters utilized to balance weights of different losses.

In the inference stage, the trained embedding model can be directly generalized to novel few-shot classes. In the validation stage, we include both the feature extractor and the purifier into the adversary generation for the fairness of the white-box setting. Then we evaluate FSIC accuracy on both legitimate and adversarial examples by the prototype-based metric learning [24].

4. Experiments

4.1. Experimental Setup

Dataset. To comprehensively evaluate our methods, we adopt the widely-used few-shot benchmark datasets mini-ImageNet [30] and CIFAR-FS [1]. MiniImageNet [30] contains 100 classes with 600 images of 84×84 size per class. Furthermore, we use 64 classes for training the embedding network, 16 classes for validation and 20 classes for testing. CIFAR-FS [1] has the same dataset splitting of 64, 16, 20 classes for training, validation and testing respectively. Following the convention [7, 31], we utilize PGD method [19] to generate adversarial examples from legitimate examples during training and inference.

Implementation details. For fair comparison, we conduct our experiments with three widely used backbones for feature extracting: Conv4-64 [30], Conv4-512 [40] and ResNet12 [12]. Note that 4 blocks convolutional networks are constructed with 64-64-64-64 (Conv4-64) or 512-512-512-512 (Conv4-512) channels for layers. Consistent with previous works [7, 16, 36], we also use a same ResNet-12 as our feature embedding network, which contains four residual blocks of 64, 160, 320 and 640 channels. We employ

Method	Backbone	1-shot				5-shot			
		Natural	FGSM [9]	PGD [19]	CW [2]	Natural	FGSM [9]	PGD [19]	CW [2]
AQ [7]	Conv4-64	33.67 ± 0.38	20.53 ± 0.30	18.52 ± 0.34	17.53 ± 0.30	50.12 ± 0.41	30.20 ± 0.36	28.16 ± 0.36	27.21 ± 0.36
R-MAML [31]	Conv4-64	33.98 ± 0.37	25.12 ± 0.35	25.69 ± 0.28	24.73 ± 0.35	50.76 ± 0.36	35.77 ± 0.35	34.19 ± 0.37	29.61 ± 0.36
Ours	Conv4-64	$\mathbf{35.38 \pm 0.39}$	$\mathbf{29.63 \pm 0.34}$	$\mathbf{28.37 \pm 0.34}$	$\mathbf{27.12 \pm 0.33}$	$\mathbf{50.93 \pm 0.39}$	$\mathbf{39.16 \pm 0.36}$	$\mathbf{37.95 \pm 0.35}$	$\mathbf{35.90 \pm 0.36}$
AQ [7]	Conv4-512	34.55 ± 0.37	20.72 ± 0.30	18.87 ± 0.31	17.73 ± 0.30	48.02 ± 0.41	29.43 ± 0.36	27.42 ± 0.35	27.45 ± 0.36
R-MAML [31]	Conv4-512	34.09 ± 0.36	27.36 ± 0.34	25.74 ± 0.34	26.37 ± 0.34	51.63 ± 0.35	36.56 ± 0.36	36.06 ± 0.38	34.60 ± 0.36
Ours	Conv4-512	$\mathbf{36.14 \pm 0.45}$	$\mathbf{29.23 \pm 0.33}$	$\mathbf{27.57 \pm 0.38}$	$\mathbf{26.61 \pm 0.33}$	$\mathbf{52.09 \pm 0.40}$	$\mathbf{38.34 \pm 0.36}$	$\mathbf{37.68 \pm 0.36}$	$\mathbf{35.93 \pm 0.35}$
AQ [7]	ResNet12	41.89 ± 0.44	21.91 ± 0.31	20.53 ± 0.33	18.38 ± 0.33	64.47 ± 0.37	32.16 ± 0.35	30.80 ± 0.37	29.62 ± 0.37
R-MAML [31]	ResNet12	37.52 ± 0.39	34.75 ± 0.39	27.46 ± 0.34	33.47 ± 0.34	62.75 ± 0.41	44.75 ± 0.38	45.78 ± 0.38	43.88 ± 0.32
Ours	ResNet12	$\mathbf{45.81 \pm 0.42}$	$\mathbf{36.03 \pm 0.37}$	$\mathbf{35.18 \pm 0.40}$	$\mathbf{34.53 \pm 0.39}$	$\mathbf{64.60 \pm 0.38}$	$\mathbf{53.33 \pm 0.39}$	$\mathbf{50.71 \pm 0.40}$	$\mathbf{47.52 \pm 0.40}$

Table 1. Comparison to previous works for 5-way 1/5-shot adversarially robust FSIC on **miniImageNet** benchmark. We report mean classifcation accuracy (%) with 95% confidence intervals on both natural and adversarial examples. The best result in each column is **bold**.

Method	Backbone	1-shot				5-shot			
		Natural	FGSM [9]	PGD [19]	CW [2]	Natural	FGSM [9]	PGD [19]	CW [2]
AQ [7]	Conv4-64	42.66 ± 0.47	27.31 ± 0.42	26.33 ± 0.43	25.35 ± 0.43	57.63 ± 0.44	40.29 ± 0.45	39.58 ± 0.48	38.69 ± 0.45
R-MAML [31]	Conv4-64	33.51 ± 0.37	28.78 ± 0.39	27.61 ± 0.39	27.12 ± 0.39	52.75 ± 0.44	35.66 ± 0.47	32.66 ± 0.42	31.47 ± 0.47
Ours	Conv4-64	$\mathbf{44.51 \pm 0.51}$	$\mathbf{39.19 \pm 0.46}$	$\mathbf{37.45 \pm 0.48}$	$\mathbf{36.53 \pm 0.46}$	$\mathbf{58.31 \pm 0.43}$	$\mathbf{49.14 \pm 0.43}$	$\mathbf{47.95 \pm 0.42}$	$\mathbf{46.45 \pm 0.42}$
AQ [7]	Conv4-512	44.35 ± 0.49	27.94 ± 0.45	26.95 ± 0.45	25.86 ± 0.45	60.13 ± 0.45	40.27 ± 0.47	39.34 ± 0.47	39.03 ± 0.46
R-MAML [31]	Conv4-512	39.22 ± 0.42	29.27 ± 0.46	27.82 ± 0.45	27.78 ± 0.45	59.28 ± 0.43	36.94 ± 0.48	34.16 ± 0.45	31.81 ± 0.48
Ours	Conv4-512	$\mathbf{45.27 \pm 0.49}$	$\mathbf{39.60 \pm 0.46}$	$\mathbf{38.03 \pm 0.46}$	$\mathbf{37.00 \pm 0.46}$	$\mathbf{60.55 \pm 0.45}$	$\mathbf{50.34 \pm 0.43}$	$\mathbf{48.69 \pm 0.44}$	$\mathbf{47.04 \pm 0.43}$
AQ [7]	ResNet12	47.40 ± 0.52	30.37 ± 0.49	29.55 ± 0.48	28.42 ± 0.48	65.78 ± 0.40	44.91 ± 0.51	44.01 ± 0.51	42.54 ± 0.51
R-MAML [31]	ResNet12	41.78 ± 0.48	34.80 ± 0.47	28.33 ± 0.46	28.86 ± 0.32	65.61 ± 0.44	37.43 ± 0.51	34.77 ± 0.41	33.15 ± 0.35
Ours	ResNet12	$\mathbf{48.13 \pm 0.48}$	$\mathbf{40.64 \pm 0.45}$	$\mathbf{39.29 \pm 0.48}$	$\mathbf{37.36 \pm 0.47}$	$\mathbf{66.99 \pm 0.43}$	$\mathbf{55.53 \pm 0.46}$	$\mathbf{52.66 \pm 0.46}$	$\mathbf{50.61 \pm 0.46}$

Table 2. Comparison to previous works for 5-way 1/5-shot adversarially robust FSIC on **CIFAR-FS** benchmark. We report mean classifcation accuracy (%) with 95% confidence intervals on both natural and adversarial examples. The best result in each column is **bold**.

Stochastic Gradient Descent (SGD) as our optimizer with an initial learning rate of 0.05, the momentum of 0.9. We train the embedding model for 100 epochs and reduce the learning rate by a factor of 10 two times after 60 and 80 epochs. The maximum adversarial perturbation is $\epsilon = 8$ in ℓ_∞-norm bound with pixel values in $[0, 255]$. In the training stage, we generate adversarial examples via PGD method with $T = 7$ steps with the step size as $\alpha = 2$ for efficiency. We set $\alpha = \beta = 0.5$ for composite adversarial reweighting, $\lambda_1 = 0.5$ and $\lambda_2 = 0.3$ for weighting.

4.2. Results

We present our experimental results on two standard benchmarks: miniImageNet [30] in Table 1 and CIFAR-FS [1] in Table 2. We report the robustness of our method against three white-box adversarial attacks, *i.e.*, FGSM [9], PGD [19] with 20 steps (step size $\alpha = 2$) and CW [2] (optimized by PGD for 30 steps with step size $\alpha = 0.8$). Note that all experimental results are reported from 2000 randomly sampled few-shot tasks. Experiments on two benchmarks show that our methods outperform state-of-the-art adversarially robust FSIC methods by a significant margin in 5-way 1/5-shot settings. Under the 1-shot setting in miniImageNet, our method with ResNet12 backbone improves over the state-of-the-art R-MAML [31] by 8.2% on legitimate examples and 7.7% on PGD adversarial examples. Furthermore, our method also outperforms all previous works by at least 8.6% on PGD adversarial examples for 1-shot classification on CIFAR-FS with ResNet12.

With the deepening of backbone networks, classification results on both natural and adversarial examples become increasingly better. This is mainly due to the strong ability of feature representation in deeper neural networks, which contributes to the final classification. Moreover, we find there also exists the same phenomenon under the few-shot scenario that deeper models benefit more from adversarial training [15]. A plausible reason is that shallow networks may be confused when learning from a small number of ad-

	AA	AR	FP	1-shot		5-shot	
				Natural	PGD [19]	Natural	PGD [19]
1				40.59	29.13	60.65	47.41
2	✓			41.26	30.06	61.77	48.32
3		✓		42.94	30.98	61.90	48.78
4			✓	41.95	31.24	62.02	48.75
5	✓	✓		43.90	33.94	62.42	49.22
6	✓		✓	44.93	34.00	63.78	49.89
7		✓	✓	45.34	34.46	63.94	50.09
8	✓	✓	✓	**45.81**	**35.18**	**64.60**	**50.71**

AA means the adversarial-aware mechanism.
AR means the adversarial-reweighted training.
FP means the feature purification module.

Table 3. Ablation results of mean classification accuracy (%) for component modules on miniImageNet [30]. The best result in each column is **bold**.

versarial examples, which may further lead to a rough decision boundary for classification.

4.3. Ablation Study

In this section, we conduct ablation study experiments to analyze contributions of our three components: adversarial-aware module, adversarial-reweighted training, and feature purification as shown in Table 3. The ablation study is implemented on legitimate and adversarial examples of miniImageNet [30] dataset with 5-way 1/5-shot settings. We adopt ResNet12 as our backbone, which can show distinct differences between various combinations of our modules.

The baseline method is based on original adversarial training in Equation (2) and then transferred to unforeseen few-shot tasks composed of natural and adversarial examples. Note that our simple baseline results are still competitive with previous complicated meta-learning methods. Moreover, there is no need for our method to train 1-shot and 5-shot targeted models respectively, which is necessary under the meta-learning setting. Especially by appending an auxiliary adversarial-aware module, the performance on both legitimate and adversarial examples boost drastically. The embedding model can thus learn the nuance between adversarial features and natural features. Moreover, adversarial-reweighted training puts different weights on adversarial examples utilized for training. This can induce the embedding model to focus on strong adversarial examples while paying less attention on weak adversarial examples that can not change the prediction. Feature purification directly acts on extracted adversarial features to reduce adversarial perturbations at the feature level. For a fair and comprehensive comparison, the feature purification module is perceptible during adversary generation that generated adversarial examples are also effective against the purification. The feature purification is also practical for features extracted from natural examples, which corresponds to a distribution alignment to the mean of natural features. To sum up, our proposed methods further enhance our baseline by 5.2% and 6% on classifying legitimate and adversarial examples in the 1-shot scenario. Moreover, a similar trend also occurrs in the 5-shot setting.

Setting	Perturbation	AQ [7]	R-MAML [31]	**Ours**
1-shot	$\epsilon = 4$	31.19	35.22	**40.70**
	$\epsilon = 8$	20.53	27.46	**35.18**
	$\epsilon = 12$	9.22	23.74	**31.89**
	$\epsilon = 16$	4.62	20.25	**28.05**
5-shot	$\epsilon = 4$	48.51	54.53	**57.12**
	$\epsilon = 8$	30.80	45.78	**50.71**
	$\epsilon = 12$	15.42	41.75	**47.39**
	$\epsilon = 16$	8.13	38.25	**42.43**

Table 4. PGD-robust [19] accuracy (%) performance under different perturabation sizes on miniImageNet [30]. The best performance in each row is marked in **bold**.

4.4. Robustness to different attack strengths

To comprehensively evaluate the effectiveness of our method, we explore the robustness against different attack strengths. Note that the adversarial strength here is mainly determined by the maximum perturbation size ϵ. Larger perturbation size indicates stronger adversarial attack. We then measure adversarial few-shot robustness under four different maximum perturbation size ($\epsilon = 4, 8, 12, 16$) with the backbone of ResNet12 as shown in Table 4. Our method can significantly outperform existing adversarial FSIC methods under different perturbation sizes. The robust accuracy of meta-learning methods drop dramatically when enhancing the attack strength. This is inline with our hypothesis: Meta-learning-based methods may induce an overfitting on the source label space so that they can not adapt to stronger adversarial attacks.

We then visualize adversarial examples of various attack strengths in Figure 3. The predicted confidence of other methods drops dramatically by increasing the attack strength in the ℓ_∞-norm bound, which even induces a wrong prediction. We can observe that our methods can preserve a considerable confidence correctly when facing strong adversarial examples.

4.5. Cross-domain transfer learning

In this section, we mainly explore the cross-domain transferability of our method with and without the auxiliary adversarial-aware module. The cross-domain transferring consists in generalizing the model trained on dataset $\mathbf{A}$ to inference on dataset $\mathbf{B}$. Note that these two datasets are cross-domain, which have disjoint image classes and different image sizes. Furthermore, we evaluate the accuracy of legitimate examples and adversarial examples simultane-

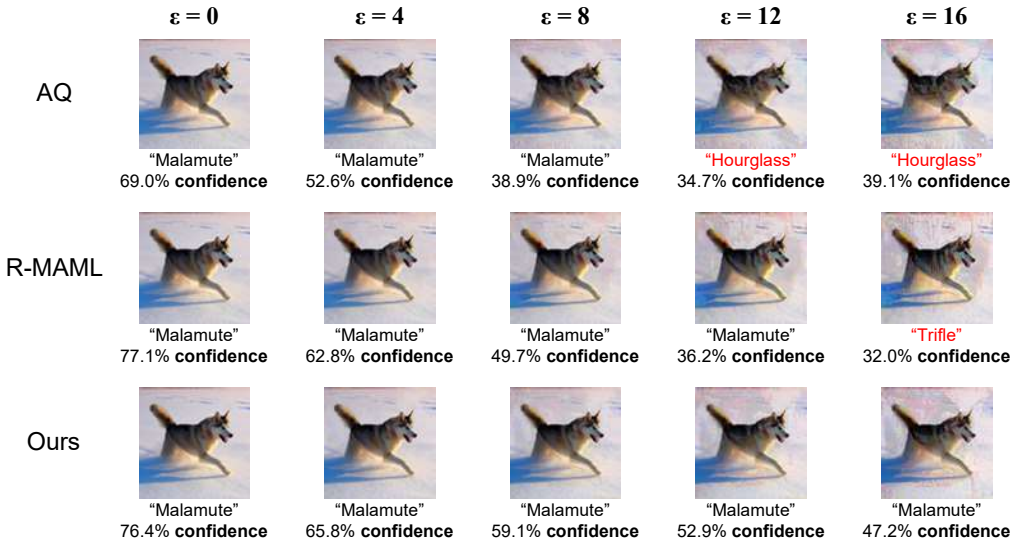

Figure 3. Visualization of adversarial examples (specific to each model) of differernt perturbation size ϵ compared with AQ [7] and R-MAML [31] in 5-shot setting with the backbone of ResNet12. We also present the predicted class (wrong prediction is highlighted in red) and the corresponding confidence.

Transfer	Method	1-shot		5-shot	
		Natural	PGD [19]	Natural	PGD [19]
M→C	AQ	43.96	26.36	61.05	37.33
	R-MAML	30.55	11.94	45.34	14.99
	w/o AA (**ours**)	44.13	33.05	61.53	46.85
	w/ AA (**ours**)	**44.65**	**37.70**	**61.76**	**52.72**
C→M	AQ	35.86	11.12	52.91	18.81
	R-MAML	28.05	22.47	36.60	25.83
	w/o AA (**ours**)	35.98	23.79	52.56	35.11
	w/ AA (**ours**)	**36.84**	**27.62**	**53.72**	**40.40**

Table 5. Cross-domain transfer experiments of accuracy (%) compared with AQ [7] and R-MAML [31]. The transferring experiments are conducted between MiniImageNet (M) and CIFAR-FS (C). The best performance in each column is marked in **bold**.

ously in the few-shot setting. In particular, our approach is based on obtaining an adversarially robust embedding model, which can be further transferred to another dataset via constructing a classification head such as the prototype model [24] with low cost.

The cross-domain transfer experiment is implemented between miniImageNet [30] and CIFAR-FS [1] bilaterally as shown as Table 5. Note that the feature purification module is specific to the size of extracted feature embeddings so that we ignore this module during cross-domain transferring. Apparently, the transferred accuracy on classifying adversarial examples boosts by a large margin via appending the adversarial-aware module. This result also indicates the significance of auxiliary supervision on adversarial transformation. More specifically, this supervision can lead the embedding model to learn the generalizable feature representations and then extend to adversarial examples in unforeseen few-shot tasks. In contrast, meta-learning methods fail to keep the adversarial robustness under the cross-domain setting. A plausible reason is that meta-training on numerous tasks from the source domain may induce the domain overfitting and thus weakens the robustness against cross-domain adversarial examples.

5. Conclusion

In this work, we propose a simple but effective framework for adversarially robust few-shot classification via generalizable representations, which dispenses with complicated meta-tasks construction. We investigate feature-level relationships between adversarial examples and legitimate examples, and thus design an adversarial-aware module. Furthermore, we propose a novel adversarial-reweighted method via instance-wise loss variation, which enables the embedding model to focus on high-adversarial examples. The postprocessing feature purification module is also presented for the feature-level distribution alignment. Extensive experiments demonstrate that our method obtains new state-of-the-art results on two popular adversarially robust FSIC benchmarks. Moreover, we show the great transferability of our feature embedding model in the cross-domain scenario.

Acknowledgments. This project is supported by NSFC (62076258, 62072482), the Key-Area Research and Development Program of Guangzhou (202007030004), and the Project of Natural Resources Department of Guangdong Province ([2021]34).

References

[1] Luca Bertinetto, Joao F Henriques, Philip HS Torr, and Andrea Vedaldi. Meta-learning with differentiable closed-form solvers. *arXiv preprint arXiv:1805.08136*, 2018. 2, 5, 6, 8

[2] 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. 6

[3] Junhao Dong and Xiaohua Xie. Visually maintained image disturbance against deepfake face swapping. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2021. 1

[4] Nanyi Fei, Zhiwu Lu, Tao Xiang, and Songfang Huang. MELR: meta-learning via modeling episode-level relationships for few-shot learning. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, 2021. 2

[5] 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. 1, 2

[6] Ruize Gao, Feng Liu, Kaiwen Zhou, Gang Niu, Bo Han, and James Cheng. Local reweighting for adversarial training. *arXiv preprint arXiv:2106.15776*, 2021. 4

[7] Micah Goldblum, Liam Fowl, and Tom Goldstein. Adversarially robust few-shot learning: A meta-learning approach. *Advances in Neural Information Processing Systems*, 33, 2020. 1, 2, 3, 4, 5, 6, 7, 8

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, pages 2672–2680. 2014. 1

[9] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *ICLR*, 2015. 6

[10] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 2

[11] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 1

[12] 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 (CVPR)*, June 2016. 1, 5

[13] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 125–136, 2019. 1, 2

[14] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012. 1

[15] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, 2016. 6

[16] Kwonjoon Lee, Subhransu Maji, Avinash Ravichandran, and Stefano Soatto. Meta-learning with differentiable convex optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10657–10665, 2019. 2, 5

[17] Junjie Li, Zilei Wang, and Xiaoming Hu. Learning intact features by erasing-inpainting for few-shot classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 8401–8409, 2021. 2

[18] Zihao Liu, Qi Liu, Tao Liu, Nuo Xu, Xue Lin, Yanzhi Wang, and Wujie Wen. Feature distillation: Dnn-oriented jpeg compression against adversarial examples. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 860–868. IEEE, 2019. 2

[19] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 2, 3, 4, 5, 6, 7, 8

[20] Jathushan Rajasegaran, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Mubarak Shah. Self-supervised knowledge distillation for few-shot learning. *arXiv preprint arXiv:2006.09785*, 2020. 2

[21] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 1

[22] Mamshad Nayeem Rizve, Salman Khan, Fahad Shahbaz Khan, and Mubarak Shah. Exploring complementary strengths of invariant and equivariant representations for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10836–10846, 2021. 1, 2

[23] Andrei A Rusu, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell. Meta-learning with latent embedding optimization. *arXiv preprint arXiv:1807.05960*, 2018. 2

[24] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017. 5, 8

[25] Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical networks for few-shot learning. In *NIPS*, pages 4077–4087, 2017. 1

[26] Qianru Sun, Yaoyao Liu, Zhaozheng Chen, Tat-Seng Chua, and Bernt Schiele. Meta-transfer learning through hard tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 1, 2

[27] 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. 1

[28] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *ICLR*, 2014. 1

[29] 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

[30] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. *Advances in neural information processing systems*, 29:3630–3638, 2016. 2, 5, 6, 7, 8

[31] 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. *arXiv preprint arXiv:2102.10454*, 2021. 2, 3, 4, 5, 6, 7, 8

[32] Siyue Wang, Xiao Wang, Pu Zhao, Wujie Wen, David Kaeli, Peter Chin, and Xue Lin. Defensive dropout for hardening deep neural networks under adversarial attacks. In *Proceedings of the International Conference on Computer-Aided Design*, pages 1–8, 2018. 2

[33] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan L. Yuille. Mitigating adversarial effects through randomization. In *International Conference on Learning Representations, ICLR*, 2018. 2

[34] Cihang Xie and Alan Yuille. Intriguing properties of adversarial training at scale. In *International Conference on Learning Representations*, 2020. 2

[35] Yao-Yuan Yang, Cyrus Rashtchian, Hongyang Zhang, Russ R Salakhutdinov, and Kamalika Chaudhuri. A closer look at accuracy vs. robustness. In *NeurIPS*, 2020. 5

[36] Han-Jia Ye, Xin-Chun Li, and De-Chuan Zhan. Task cooperation for semi-supervised few-shot learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 10682–10690, 2021. 5

[37] Chengxiang Yin, Jian Tang, Zhiyuan Xu, and Yanzhi Wang. Adversarial meta-learning. *arXiv preprint arXiv:1806.03316*, 2018. 2, 3, 4

[38] 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

[39] Jingfeng Zhang, Jianing Zhu, Gang Niu, Bo Han, Masashi Sugiyama, and Mohan Kankanhalli. Geometry-aware instance-reweighted adversarial training. In *International Conference on Learning Representations*, 2021. 2, 4

[40] Manli Zhang, Jianhong Zhang, Zhiwu Lu, Tao Xiang, Mingyu Ding, and Songfang Huang. Iept: Instance-level and episode-level pretext tasks for few-shot learning. In *International Conference on Learning Representations*, 2021. 5

[41] Ruixiang Zhang, Tong Che, Zoubin Ghahramani, Yoshua Bengio, and Yangqiu Song. Metagan: An adversarial approach to few-shot learning. *NeurIPS*, 2:8, 2018. 2