Adversarially Robust Few-shot Learning through Simple Transfer - supplementary material

Akshayvarun Subramanya
University of Maryland, Baltimore County
akshayv1@umbc.edu

Hamed Pirsiavash
University of California, Davis
hpirsiav@ucdavis.edu

1. Related Work

Few shot Image Classification: Few-shot learning is a challenging problem in computer vision where the goal is rapid generalization to unseen tasks. Metric learning approaches such as [31, 32, 35] were some of the earliest approaches towards tacking this problem. [31] learn a metric space where prototypical representation of each category is utilized for classifying novel data. [28] showed that rather than taking the average, learning the prototypes along with the model can lead to better performance. More recently, a family of algorithms based on learning to learn [1] or meta-learning have gained considerable attention. [27] develop an LSTM based meta-learner to train anothet network for the few-shot task. [12, 23] create a model agnostic algorithm which aims to learn a good initialization that can adapt to new tasks. Hallucination based methods such as [2, 17, 38] also present promising directions towards improved generalization. [13, 25] try to directly predict the weights of the classifier for novel categories. [43] calibrate the distributions of few shot examples using the statistics of categories with larger number of examples. Our calibration is similar to [43], but we apply it at category-level rather than instance-level. However recent works have shown that simple baselines which are non-episodic in nature can provide competitive performance for few-shot image classification task [8, 10]. Such line of work provide for non-sophisticated baselines and pose a question to the community to rethink the approach towards few shot learning.

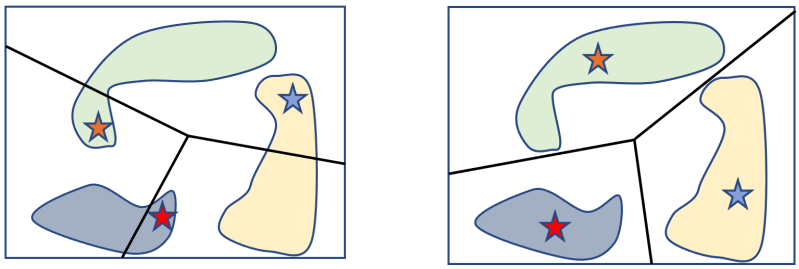
Adversarial examples: Adversarial examples are carefully crafted perturbations designed to fool the model [15, 20, 33]. [15] showed that adversarial examples can be created rather easily using the sign of single gradient step, which they called as Fast Gradient Sign Method (FGSM). The existence of such examples raises a question regarding the generalization capabilities of Deep Neural networks. Many defenses have been proposed to overcome this problem [11, 21, 24, 41], but they have been bypassed with slight modifications to the adversary [3, 5, 6]. One of the most common approaches called adversarial training involves incorporating the adversarial examples into the training set. [22] showed that the first order adversary, based on the Projected Gradient Descent (PGD) algorithm can be used to train robust neural networks. Provable methods have developed which can provide certification on the susceptibility of an input towards adversaries [16, 26, 39, 40]. A class of algorithms based on randomized smoothing [9, 42] have shown promising results in training large scale neural networks. [45] provide a theoretical analysis on the robustness vs accuracy trade-off, which had been studied empirically [34], and show that their algorithm named TRADES improves robustness compared to previous approaches. [30] use the gradients from the back-propagation algorithm to improve robustness with minimal cost.

Adversarial Robustness for Few-shot classifiers: Recent works have tried to address the problem of adversarial examples in the context of few-shot learning. [44] used the FGSM adversary to create adversarial examples and optimized a meta-learner to be robust to adversarial examples. [14] showed that meta-learning algorithms can be supplemented with adversarial examples in the query set to learn robustness. Their method called Adversarial Querying was shown to be robust to strong attacks such as PGD. Many meta-learning approaches were extended to their robust counterparts by including adversarial query examples. Recently, [37] also proposed a similar approach where MAML was used as the base meta learning algorithm. [37] also showed that including a contrastive learning objective similar to [7] can provide a way to use unlabelled data when learning the model and thus improve both standard and robust accuracy.

2. Experiments

Here we describe the implementation details and additional experimental results.

Implementation details: In the base training stage, as described in main paper, we follow the attack parameters of [14] and use iterative PGD attack with 7 iterations during training with $\epsilon = 8/255$ and $\alpha = 2/255$ for all experiments unless otherwise mentioned. We mainly use the standard



(a) Nearest Centroid Classifier without Calibration

(b) Calibrated Nearest Centroid Classifier

Figure 1. An illustration of the calibration for the Nearest Centroid Classifier, here the stars represent class centroids. Since we are working in the few-shot regime from novel dataset, if the examples are not sampled from across the distribution, it may result in centroids which do not necessarily represent the true distribution. On the other hand, calibration using the base classes can result in centroids which better represent the actual distribution of the class, allowing for better generalization.

network architecture ResNet [18] and also provide results for other architectures. For weight averaging, we set the parameter $\tau = 0.999$. We use SGD optimizer with a learning rate of 0.1 and weight decay of $1e-5$ for the feature extractor parameters θ_b and $1e-4$ for the classifier parameters ω_b. We train the model for 250 epochs with a batch size of 64. For novel training and learning the Linear Classifier, we follow the setting described in [8] and learn the parameters using SGD with momentum 0.9 and learning rate $\eta = 0.01$. We set the dampening as 0.9 and weight decay of $1e-3$. For our calibration, we use $m = 2$ number of base categories. **CIFAR-FS** was proposed in [4] as a benchmark for few-shot classification. It splits CIFAR-100 dataset similar to Mini-ImageNet. **CUB** [36] is a fine-grained dataset which has been used as a benchmark for few-shot classification. We use the split provided by [19] consisting of 100 base , 50 validation and 50 novel classes. As per our knowledge, we are the first to show adversarial robustness for a fine-grained dataset under few-shot setting. We also consider **TieredImageNet** [29], which is a subset sampled hierarchically from ImageNet and is different from MiniImageNet.

2.1. Comparison with OFA [37]

We compare with OFA [37] where MAML was combined with adversarial training . We present them separately compared to previous results as the attack parameters and testing configuration followed are different. OFA uses $\epsilon = 2/255$ for MiniImageNet and $\epsilon = 8/255$ for CIFAR-FS. They perform 10 iterations of PGD during training and testing, averaging over 2400 tasks. These are not directly comparable with our previous experiments which uses attack parameters similar to [14], hence for a fair comparison, we use the same setting and refer the readers to [37] for more hyperparameter details. **Base training** column indicates the type of adversary used during base dataset training where AT indicates PGD adversarial training, TRADES is the algorithm presented in [45] and CL corresponds to using the Contrastive Learning objective [7]. Both TRADES and CL use additional unlabelled data in a semi-supervised manner. We observe from Table 1 that our method has clear gains in terms of robust accuracy and surpasses standard accuracy in some cases as well. We also train a model with Conv4 backbone on CIFAR-FS dataset and the results are presented in Table 2. We see a similar trend where our CNC method outperforms previous approaches. We also conduct an experiment using TRADES during Base Training, allowing us to compare methods that use similar adversary. Under such comparable settings, our method outperforms previous approaches. This experiment shows that our method can generalize to other adversarial training methods and we believe that as more advanced methods are developed in the community, they can be incorporated in a straightforward manner to improve robustness under few-shot settings.

2.2. CUB

For results on CUB dataset, we use the same attack parameters described for Mini-ImageNet i.e, $\epsilon = 8/255$, $\alpha = 2/255$, 7 iterations of PGD during training and 20 during testing. We use ResNet18 backbone and implement AQ as per the guidelines given in [14] and our best implementation is presented in Table 3.

Since CUB is a fine-grained classification dataset, the base and novel categories share greater similarity compared to previous datasets. Hence it serves as an opportunity to understand how the robustness transfers from base to novel dataset, i.e whether the similarity in classes acts as a boon or bane under fine-grained dataset settings. As seen from our experiments, the linear classifier baseline performs reasonably well indicating that a robust base classifier transfers to a robust novel classifier. The CNC method also benefits

Method	Base training	Conv4		ResNet18	
		Standard Acc.	Robust Acc.	Standard Acc.	Robust Acc.
AQ	AT	29.6	24.9	30.04	20.05
OFA	AT	**40.82**	23.04	38.94	19.94
OFA	TRADES	37.1	25.51	41.94	20.19
OFA	CL	38.60	26.81	43.98	21.47
Ours (Linear)	AT	38.39 $\pm$ 0.37	28.76 $\pm$ 0.33	44.93$\pm$ 0.37	29.30 $\pm$ 0.33
Ours (CNC)	AT	39.23 $\pm$ 0.38	**30.77 $\pm$ 0.35**	**49.15 $\pm$ 0.41**	**35.59 $\pm$ 0.38**

Table 1. Comparison with [37] on Mini-ImageNet dataset. Note that both TRADES and CL use additional unlabelled data in a semi-supervised manner. Our method outperforms previous approaches on both settings.

Method	Base training	1-shot		5-shot	
		Standard Acc.	Robust Acc.	Standard Acc.	Robust Acc.
AQ	AT	31.25	26.34	52.32	33.96
OFA	AT	39.76	26.15	57.18	32.62
OFA	TRADES	40.59	28.06	57.62	34.76
OFA	CL	41.25	**29.33**	**57.95**	35.3
Ours (Linear)	AT	41.12 $\pm$ 0.40	25.65 $\pm$ 0.37	56.20 $\pm$ 0.39	34.73 $\pm$ 0.41
Ours (CNC)	AT	41.81$\pm$ 0.41	28.22 $\pm$ 0.40	53.52 $\pm$ 0.40	39.09 $\pm$ 0.42
Ours (CNC)	TRADES	**43.56$\pm$0.43**	28.12 $\pm$ 0.41	56.99 $\pm$ 0.40	**39.48 $\pm$0.43**

Table 2. Comparison with [37] for Conv4 backbone on CIFAR-FS dataset. Comparing methods that use same base training procedure (**AT** or **TRADES**) , we can see that our CNC method outperforms on Robust Accuracy under both 1-shot and 5-shot settings. This experiment shows that our method can generalize to other adversarial training methods as well.

under such settings and outperforms all other methods on both Standard and Robust Accuracy.

2.3. TieredImageNet

We also consider the large scale dataset TieredImageNet [29], which is a subset sampled hierarchically from ILSVRC12. Each class is a child of one of the 34 more abstract categories from ImageNet. Hence, the classes are spread differently compared to MiniImageNet. This allows us to test our method against varying hardness and diversity in few-shot categories. The similarity between the base and novel categories is varied in this context and shows that our method performs well under such settings as well. As seen in Table 4 that the linear classifier acts as a strong baseline and our CNC method outperforms previous works.

2.4. Ablation studies

Variation of Robust Accuracy with number of attack iterations: We vary the number of attack iterations of PGD and observe a fairly stable performance for both 1-shot and 5-shot settings, as seen in Table 5. This experiment shows that defense is not sensitive to the number of attack iterations.

Variation of Robust Accuracy with perturbation budget ϵ: To check for the absence of gradient masking, we increase ϵ from 8/255 to 128/255 in Figure 2. As expected, we observe that both 1-shot and 5-shot accuracy drop to zero with increased ϵ. Note that we plot only the the mean accuracy over 1000 different tasks.

Finetuning backbone: We finetune more layers from ResNet18 architecture on MiniImagenet and find that Robust Accuracy (RA) decreases. We believe this is because the model overfits to few-shot data. B4 represents learning the 4th block and B3 learning the 3rd block in ResNet18 architecture. Results are shown in Table 6.

Weight averaging and Transformation: We conducted an experiment with ResNet18 backbone on MiniImageNet where Weight Averaging (WA) was included with AQ and found that it did not affect RA significantly. This shows that WA works best when combined with our mini-batch based framework. Results are shown in Table 7.

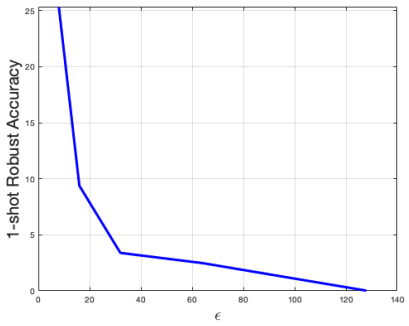
Varying the number of base categories chosen: We plot the variation of mean Robust Accuracy with number of base neighbors m in Figure 3. When $m = 0$, the method becomes similar to a Nearest Centroid Classifier without calibration. We find best results with $m = 2$ which we use for all our experiments. Note that this calibration step is performed once prior to the inference and can be considered a preprocessing step, hence not affecting the inference time. We would like to emphasize that we use the same dataset as previous meta-learning based approaches, only that our base training

Method	Backbone	1-shot		5-shot	
		Standard Acc.	Robust Acc.	Standard Acc.	Robust Acc.
AQ	ResNet18	54.27 ± 0.79	28.23 ± 0.66	68.42 ± 0.62	37.10 ± 0.66
Ours (Linear)	ResNet18	51.93 ± 0.71	27.24 ± 0.64	69.83 ± 0.61	37.06 ± 0.68
Ours (CNC)	ResNet18	**56.42 ± 0.78**	**32.18± 0.70**	**71.51 ± 0.60**	**44.33 ± 0.69**

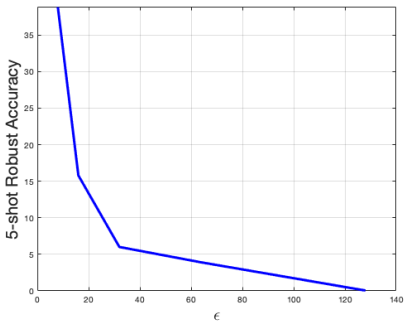
Table 3. **Results on CUB dataset**. We show that robustness transfers from base to novel datasets under fine-grained classification setting as well. Our Linear classifier serves as a strong baseline and our CNC method outperforms on both metrics.

Method	Backbone	1-shot		5-shot	
		Standard Acc.	Robust Acc.	Standard Acc.	Robust Acc.
AQ	ResNet18	49.77 ± 0.70	29.78 ± 0.65	66.72 ± 0.56	43.73 ± 0.63
Ours (Linear)	ResNet18	50.47 ± 0.69	27.90 ± 0.60	68.48 ± 0.60	40.30 ± 0.65
Ours (CNC)	ResNet18	**51.38 ± 0.71**	**30.27 ± 0.62**	**68.50± 0.59**	**44.64± 0.66**

Table 4. Results on TieredImageNet dataset.



(a) 1-shot



(b) 5-shot

Figure 2. Variation of Robust Accuracy with different perturbation budget ϵ. Results are shown using ResNet-12 backbone on MiniImageNet

	1-shot	5-shot
PGD Iterations	Robust Acc.	Robust Acc.
20	25.32 ± 0.52	38.83 ± 0.57
40	25.19 ± 0.53	38.46 ± 0.54
100	25.64 ± 0.53	38.22 ± 0.56
200	25.09 ± 0.54	38.62 ± 0.57

Table 5. Variation of attack iterations. Results are shown using ResNet-12 backbone on MiniImageNet dataset.

is standard mini-batch training.

Layers	1-shot (RA)	5-shot (RA)
Linear	19.56 +-0.50	30.6 +-0.50
(Linear,B4)	18.17 +-0.40	28.33 +-0.50
(Linear,B4,B3)	17.46 +-0.40	25.72 +-0.50

Table 6. We show that finetuning additional layers leads to decreased Robust Accuracy (RA). Results are shown using ResNet18 backbone on MiniImageNet dataset.

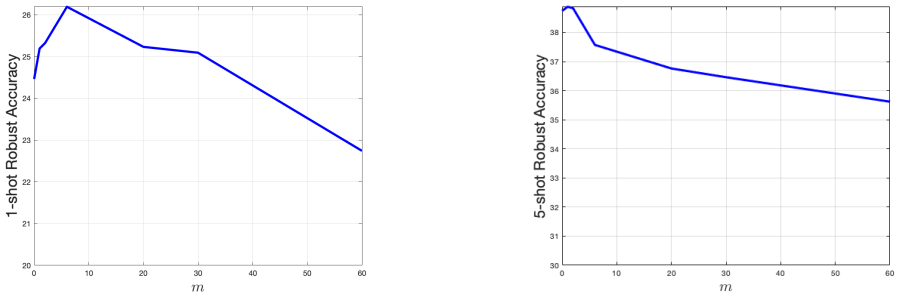
	1-shot (RA)	5-shot (RA)
AQ	20.52 +-0.50	32.18 +-0.50
AQ+WA	20.76 +- 0.40	31.37 +-0.50
Ours (WA+CNC)	**21.38 +-0.50**	**33.41 +-0.50**

Table 7. We combine Weight averaging (WA) with AQ and observe that there is no improvement in Robust Accuracy (RA). Results are shown using ResNet-18 backbone on MiniImageNet dataset.

References

[1] Marcin Andrychowicz, Misha Denil, Sergio Gomez, Matthew W Hoffman, David Pfau, Tom Schaul, Brendan Shillingford, and Nando De Freitas. Learning to learn by gradient descent by gradient descent. *arXiv preprint arXiv:1606.04474*, 2016. 1

[2] Antreas Antoniou, Amos Storkey, and Harrison Edwards. Data augmentation generative adversarial networks. *arXiv*



(a) 1-shot (b) 5-shot

Figure 3. Variation of Robust accuracy with number of base centers m for 1-shot and 5-shot settings. Results are shown using ResNet12 backbone on MiniImageNet dataset.

preprint arXiv:1711.04340, 2017. 1

[3] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International Conference on Machine Learning*, pages 274–283, 2018. 1

[4] Luca Bertinetto, João F Henriques, Jack Valmadre, Philip HS Torr, and Andrea Vedaldi. Learning feed-forward one-shot learners. *arXiv preprint arXiv:1606.05233*, 2016. 2

[5] Nicholas Carlini and David Wagner. Defensive distillation is not robust to adversarial examples. 1

[6] Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pages 3–14. ACM, 2017. 1

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 1, 2

[8] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A closer look at few-shot classification. *arXiv preprint arXiv:1904.04232*, 2019. 1, 2

[9] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning*, pages 1310–1320. PMLR, 2019. 1

[10] Guneet S Dhillon, Pratik Chaudhari, Avinash Ravichandran, and Stefano Soatto. A baseline for few-shot image classification. *arXiv preprint arXiv:1909.02729*, 2019. 1

[11] Reuben Feinman, Ryan R Curtin, Saurabh Shintre, and Andrew B Gardner. Detecting adversarial samples from artifacts. *arXiv preprint arXiv:1703.00410*, 2017. 1

[12] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, 2017. 1

[13] Spyros Gidaris and Nikos Komodakis. Dynamic few-shot visual learning without forgetting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4367–4375, 2018. 1

[14] Micah Goldblum, Liam Fowl, and Tom Goldstein. Adversarially robust few-shot learning: A meta-learning approach. *NeurIPS*, 2020. 1, 2

[15] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015. 1

[16] Sven Gowal, Krishnamurthy Dvijotham, Robert Stanforth, Rudy Bunel, Chongli Qin, Jonathan Uesato, Relja Arandjelovic, Timothy Mann, and Pushmeet Kohli. On the effectiveness of interval bound propagation for training verifiably robust models. *arXiv preprint arXiv:1810.12715*, 2018. 1

[17] Bharath Hariharan and Ross Girshick. Low-shot visual recognition by shrinking and hallucinating features. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3018–3027, 2017. 1

[18] 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. 2

[19] Nathan Hilliard, Lawrence Phillips, Scott Howland, Artëm Yankov, Courtney D Corley, and Nathan O Hodas. Few-shot learning with metric-agnostic conditional embeddings. *arXiv preprint arXiv:1802.04376*, 2018. 2

[20] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016. 1

[21] Xin Li and Fuxin Li. Adversarial examples detection in deep networks with convolutional filter statistics. *arXiv preprint arXiv:1612.07767*, 2016. 1

[22] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018. 1

[23] Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*, 2018. 1

[24] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *Security and Privacy (SP), 2016 IEEE Symposium on*, pages 582–597. IEEE, 2016. 1

[25] Hang Qi, Matthew Brown, and David G Lowe. Low-shot learning with imprinted weights. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5822–5830, 2018. 1

[26] Aditi Raghunathan, Jacob Steinhardt, and Percy Liang. Semidefinite relaxations for certifying robustness to adversarial examples. *arXiv preprint arXiv:1811.01057*, 2018. 1

[27] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. 2016. 1

[28] Avinash Ravichandran, Rahul Bhotika, and Stefano Soatto. Few-shot learning with embedded class models and shot-free meta training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 331–339, 2019. 1

[29] Mengye Ren, Eleni Triantafillou, Sachin Ravi, Jake Snell, Kevin Swersky, Joshua B Tenenbaum, Hugo Larochelle, and Richard S Zemel. Meta-learning for semi-supervised few-shot classification. *arXiv preprint arXiv:1803.00676*, 2018. 2, 3

[30] Ali Shafahi, Mahyar Najibi, Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free! *arXiv preprint arXiv:1904.12843*, 2019. 1

[31] Jake Snell, Kevin Swersky, and Richard S Zemel. Prototypical networks for few-shot learning. *arXiv preprint arXiv:1703.05175*, 2017. 1

[32] 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

[33] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *CoRR*, abs/1312.6199, 2013. 1

[34] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018. 1

[35] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. *arXiv preprint arXiv:1606.04080*, 2016. 1

[36] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 2

[37] 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 *ICLR*, 2021. 1, 2, 3

[38] Yu-Xiong Wang, Ross Girshick, Martial Hebert, and Bharath Hariharan. Low-shot learning from imaginary data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7278–7286, 2018. 1

[39] Eric Wong and Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *International Conference on Machine Learning*, pages 5286–5295. PMLR, 2018. 1

[40] Eric Wong, Frank R Schmidt, Jan Hendrik Metzen, and J Zico Kolter. Scaling provable adversarial defenses. *arXiv preprint arXiv:1805.12514*, 2018. 1

[41] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, and Alan L. Yuille. Adversarial examples for semantic segmentation and object detection. *CoRR*, abs/1703.08603, 2017. 1

[42] Greg Yang, Tony Duan, J Edward Hu, Hadi Salman, Ilya Razenshteyn, and Jerry Li. Randomized smoothing of all shapes and sizes. In *International Conference on Machine Learning*, pages 10693–10705. PMLR, 2020. 1

[43] Shuo Yang, Lu Liu, and Min Xu. Free lunch for few-shot learning: Distribution calibration. In *ICLR*, 2021. 1

[44] Chengxiang Yin, Jian Tang, Zhiyuan Xu, and Yanzhi Wang. Adversarial meta-learning. *arXiv preprint arXiv:1806.03316*, 2018. 1

[45] 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. 1, 2