Few-shot Transferable Robust Representation Learning via Bilevel Attacks

Minseon Kim[1*], **Hyeonjeong Ha**[1*], **Sung Ju Hwang**[1,2]
[1]Korea Advanced Institute of Science and Technology (KAIST), [2]AITRICS
{minseonkim, hyeonjeongha, sjhwang82}@kaist.ac.kr

Abstract

Existing adversarial learning methods assume the availability of a large amount of data from which we can generate adversarial examples. However, in an adversarial meta-learning setting, the model need to learn transferable robust representations for unseen domains with only a few adversarial examples, which is a very difficult goal to achieve even with a large amount of data. To tackle such a challenge, we propose a novel adversarial self-supervised meta-learning framework with bilevel attacks which aims to learn robust representations that can generalize across tasks and domains. Specifically, in the inner loop, we update the parameters of the given encoder by taking inner gradient steps using two different sets of augmented samples, and generate adversarial examples for each view by maximizing the instance classification loss. Then, in the outer loop, we meta-learn the encoder parameter to maximize the agreement between the two adversarial examples, which enables it to learn robust representations. We experimentally validate the effectiveness of our approach on unseen domain adaptation tasks, on which it achieves impressive performance. Specifically, our method significantly outperforms the state-of-the-art meta-adversarial learning methods on few-shot learning tasks, as well as self-supervised learning baselines in standard learning settings with large-scale datasets.

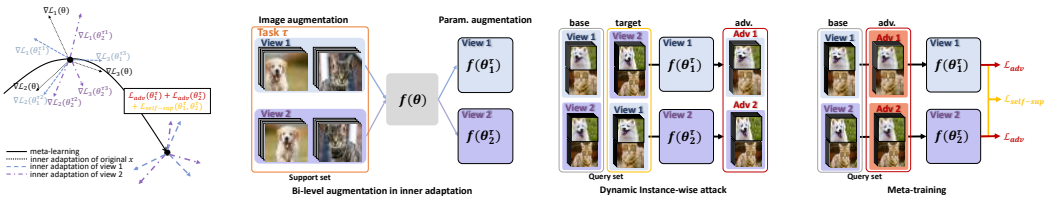
1 Introduction

Deep neural networks (DNNs) are known to be vulnerable to imperceptible small perturbations in the input data instances [34]. To overcome such adversarial vulnerability of DNNs, the vast of previous studies [40, 2, 23, 37, 29] have been proposed to enhance the robustness of the trained deep network models by defending against the adversarial attacks. Despite of the recent progress in adversarial supervised learning, training on a large number of samples is essential to achieve better robustness [3, 29, 11]. Recently, Carmon et al. [3] employs larger dataset (i.e., TinyImageNet [20]) with pseudo labels, Gowal et al. [11] utilizes generative model to generate additional samples from the dataset, and Rebuffi et al. [29] leverages augmentation functions to obtain more data samples.

On the other hand, since meta-learning framework [17, 33, 32, 7, 24] employs scarce data and has to adapt quickly to new tasks, it is difficult to obtain robustness with conventional adversarial training methods which require a large amount of data [8]. Adversarial Querying (AQ) [8] proposed an adversarially robust meta-learning scheme that meta-learns with adversarial perturbed query examples. Similarly, Wang et al. [36] studies how to enhance the robustness of a meta-learning framework with the adversarial regularizer in the inner adaption or outer optimization. However, since existing adversarial meta-learning approaches [38, 8, 36] mostly focus on the rapid adaptation to new tasks, while mostly reusing the features with little modification at the task adaptation step [25], the representations themselves may not be effectively meta-learned to be robust across tasks, thus they show poor robustness on unseen domains (see Table 1).

[*]Equal contribution. Author ordering determined by coin flip.

ML Safety Workshop, 36th Conference on Neural Information Processing Systems (NeurIPS 2022).



(a) Robust meta-learning (b) Inner adaptation (c) Meta optimization

Figure 1: **Overview of TROBA.** (a) TROBA adapts the encoder to differently augmented sets of the support sets (blue, purple line). Then, it meta learns (black line) with both adversarial loss (red) and self-supervised loss (yellow). (b) During the inner adaptation, TROBA adapts encoders with the differently augmented support sets. (c) To generate adversarial examples for meta-learning, we propose a bilevel attack with the instance-wise attack that maximizes the difference between differently augmented query images, for the task-shared encoder f. Then, we train the framework to have an adversarially consistent prediction across multiple views with self-supervised loss while learning the encoder to generalize across tasks, which enables it to learn robust representations that are transferable to unseen tasks and domains.

To tackle such challenges, we propose a novel and effective adversarial meta-learning framework which can generalize to unseen domains, ***T**ransferable **RO**bust meta-learning via **B**ilevel **A**ttack (**TROBA**)*. TROBA utilizes a bilevel attack scheme to meta-learn robust representations that can generalize across tasks and domains, motivated by self-supervised learning (Figure 1). Specifically, we redesign the instance-wise attack proposed in Kim et al. [16], Jiang et al. [15] which maximizes the instance classification loss, by adapting the shared encoder to two sets of differently augmented samples of the same instance with inner gradient update steps and then attacking them (dynamic instance-wise attack). Then, our framework learns to maximize the similarity between the feature embeddings of those two attacked samples, while meta-learning the shared encoder by BOIL [25], which allows it to learn robust representations for any given set of augmented samples. Since the robustness is achieved at the representation level, our framework can generalize to unseen tasks and domains. The experimental results from multiple benchmark datasets show that our model is robust on few-shot learning tasks from unseen domains (Table 1) thanks to its ability to learn generalizable robust representations. Moreover, our model even obtains comparable robust transferability to the self-supervised pre-trained models while using fewer data instances (Table 2).

2 Related Work

Adversarial Training. Many existing works aim to enhance the robustness of a model trained with supervised learning with labeled data, by utilizing adversarial examples [9, 2, 27]. The most popular approach is Adversarial Training from Madry et al. [22], which utilizes project gradient descent (PGD) to maximize the loss in the inner-maximization loops while minimizing the overall loss on adversarial samples generated by the PGD attack. Zhang et al. [40] introduces regularized Kullback-Leibler divergence (KLD) loss that helps to enhance the robustness by enforcing the consistency in the predictive distribution between the clean and adversarial examples.

Adversarial Meta-Learning. The most popular work in meta-learning is Model Agnostic Meta-Learning (MAML) [7] which uses a bilevel optimization scheme. Further, Oh et al. [25] propose BOIL, which meta-learns the feature extractor while keeping the final classifier fixed, and show that it has better generalization over cross-domain adaption tasks compared to MAML. Although meta-learning contributes to learning useful generalizable knowledge with scarce data, existing meta-learning approaches are prone to adversarial perturbations. To tackle this problem, Yin et al. [38] attempt to combine adversarial training (AT) [22] with MAML [7] by using both clean and adversarial examples. However, Goldblum et al. [8] later point out that ADML [38] may not obtain good robustness to strong attack since it uses relatively weak attacks during training. Then, they propose an Adversarial Querying (AQ), which trains with adversarial examples only from the query set. Similarly, Wang et al. [36] suggest Robust-regularized meta-learner on top of the MAML (RMAML), where adversarial attacks are conducted only in the meta-optimization phase. However, previous works [8, 36] are still vulnerable to adversarial attacks on unseen domains since they reuse the representations with little updates during inner optimization, which is demonstrated as inefficient to achieve generalization across domains [25]. To tackle such a limitation, we propose a robust self-supervised meta-learning framework via bilevel attacks which meta-learns the representation layers to generalize across any adversarial learning tasks that are generated from randomly sampled instances.

3 Transferable Robust Meta-learning via Bilevel Attacks

Bi-level parameter augmentation in adversarial meta-learning. Motivated by the self-supervised learning [4, 13, 12] which learn good quality of the visual representations from image augmentation, we propose a bilevel parameter augmentation to have transferable robustness in meta-learning. Bilevel parameter augmentation enables the model to adapt the view-specific projected latent space to set of augmented samples of the given instance. Specifically, to generate augmented parameters of the encoder, we first generate multiple views of images for both support set and query set of each task with a stochastic data augmentation function t that is randomly selected from the augmentation set $\mathcal{T}$ [39]. Then, we generate multiple views of the shared parameters (θ_1^τ and θ_2^τ) which are adapted parameters of encoder with differently transformed support sets ($\mathbf{S}_\tau = \{t_1(x^s), t_2(x^s), y^s\}$) as shown in Figure 1. Overall, we introduce parameter-level augmentation along with image-level augmentation to form a different view of single instances in the meta-learning framework, which we refer to as *bilevel parameter augmentation*.

Bilevel attack with dynamic instance-wise attack. On top of bilevel parameter augmentation, we propose a bilevel attack with a dynamic instance-wise attack to obtain generalized robustness in few-shot tasks. Specifically, we apply an instance-wise attack [16] on our meta-learning framework, by generating adversaries that maximize the difference between the representations of the augmented samples of the same instance obtained by the encoder whose parameters are adapted to each view, as follows:

$$\begin{aligned}\delta_1^{\mathrm{t}+1} &= \Pi_{B(x^q, x^q+\epsilon)}\Big(\delta_1^{\mathrm{t}} + \alpha \texttt{sign}\Big(\nabla_{\delta_1^{\mathrm{t}}} \mathcal{L}_{\texttt{similarity}}\big(f_{\theta_1^\tau}(t_1(x^q) + \delta_1^{\mathrm{t}}), f_{\theta_1^\tau}(t_2(x^q))\big)\Big)\Big), \\ \delta_2^{\mathrm{t}+1} &= \Pi_{B(x^q, x^q+\epsilon)}\Big(\delta_2^{\mathrm{t}} + \alpha \texttt{sign}\Big(\nabla_{\delta_2^{\mathrm{t}}} \mathcal{L}_{\texttt{similarity}}\big(f_{\theta_2^\tau}(t_2(x^q) + \delta_2^{\mathrm{t}}), f_{\theta_2^\tau}(t_1(x^q))\big)\Big)\Big),\end{aligned} \tag{1}$$

where δ_1, δ_2 are generated perturbations to maximize the difference between features from each bilevel augmented encoder $f(\theta_1^\tau)$ and $f(\theta_2^\tau)$ respectively. The maximized loss $\mathcal{L}_{\texttt{similarity}}$ is the instance-wise classification loss used in adversarial self-supervised learning [16]. We use the differently transformed query counterpart sets as a target for dynamic instance-wise attack and calculate perturbations with the parameter of the augmented encoder.

Adversarial meta-learning with bilevel attack. We now present a framework to learn transferable robust representations via bilevel attack for unseen domains. The gradient (g) is calculated to minimize our proposed objective as follows:

$$g = \nabla_{\theta_1^\tau, \theta_2^\tau, \mu} \mathcal{L}_{\texttt{Our}}(h_\mu, f_{\theta_1^\tau}, f_{\theta_2^\tau}, t_1(x^q), t_1(x^q)^{adv}, t_2(x^q), t_2(x^q)^{adv}, y^q), \tag{2}$$

where $\mathcal{L}_{\texttt{Our}}$ is the meta-objective loss to obtain generalized robustness, h_μ is a meta-initialized classifier, and f_{θ_1} and f_{θ_2} are bilevel augmented encoder for each view. Further, the $\mathcal{L}_{\texttt{Our}}$ consists of adversarial loss, i.e., TRADES [40] loss, and self-supervised loss as follow,

$$\mathcal{L}_{\texttt{Our}} = \sum_{n=1}^{2} \big[\mathcal{L}_{\texttt{CE}}(l_n, y^q) + \mathcal{L}_{\texttt{KL}}(l_n^{adv}, l_n)\big] + \mathcal{L}_{\texttt{self-sup}}(z_1^{adv}, z_2^{adv}), \tag{3}$$

where $z_n = f_{\theta_n^\tau}(t_n(x^q))$ and $l_n = h_\mu(z_n)$ are a feature and a logit of each multi-view instance , $\mathcal{L}_{\texttt{CE}}$ is a cross-entropy loss, $\mathcal{L}_{\texttt{KL}}$ is a KL-divergence loss, and $\mathcal{L}_{\texttt{self-sup}}$ is a cosine similarity loss between two differently augmented features. The crucial component here is the self-supervised loss which regularizes our model to have robust consistency between the features from the two different views, which helps it learn robust representations across any instances or augmentations, allowing it to achieve transferable robustness.

4 Experiment

We meta-train our approach on ResNet12 with 5-way 5-shot images in CIFAR-FS based on BOIL [25] and [21]. We take a single step in both meta-training and meta-testing. We adversarially train our model with ℓ_∞ PGD attacks with the epsilon of $8/255$, alpha of $2/255$ in 7 steps. We evaluate the robustness against ℓ_∞ PGD attacks with the epsilon of $8/255$ and 20 iterations.

4.1 Results of Adversarial Robustness in Unseen Domain Few-shot Tasks

Since our main goal is to achieve transferable robustness in unseen domains, we mainly validate our methods on unseen domain few-shot tasks. We meta-train our model on CIFAR-FS and meta-test on the benchmark datasets with different domains such as Mini-ImageNet, Tiered-ImageNet, CUB, Flowers, and Cars. As shown in Table 1, previous adversarial meta-learning methods have difficulty in achieving robustness on unseen domains. However, TROBA is able to show impressive transferable robustness in this cross-domain task. It also obtains significantly better

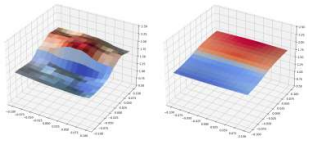
Table 1: Results of transferable robustness in 5-shot unseen domain tasks that are trained on CIFAR-FS. Rob. stands for accuracy(%) that is calculated with PGD-20 attack ($\epsilon = 8./255.$, step size=$\epsilon/10$). Clean stands for test accuracy(%) of clean images. All models are trained with PGD-7 attacks on ResNet12.

CIFAR-FS →	Mini-ImageNet		Tiered-ImageNet		CUB		Flowers		Cars	
	Clean	Rob.	Clean	Rob.	Clean	Rob.	Clean	Rob.	Clean	Rob.
MAML [7]	44.85	6.21	**61.19**	2.48	48.41	3.46	**67.76**	5.73	**43.94**	5.31
ADML [38]	28.66	6.53	40.06	11.36	31.18	5.21	39.36	11.26	27.43	3.18
AQ [8]	33.09	3.32	37.41	5.05	38.37	4.10	60.14	11.03	36.83	4.47
RMAML [36]	28.05	6.65	29.54	9.30	30.24	5.67	42.91	10.79	31.72	5.56
Ours	**45.82**	**24.12**	51.46	30.06	**48.56**	**25.23**	66.49	**42.16**	38.29	**19.43**

Table 2: Experiments results in robust full-finetuning of TROBA and the state-of-the-art adversarial self-supervised learning (SSL) models. While TROBA is trained on CIFAR-FS, other models are trained on the CIFAR-100. TROBA is pre-trained with bilevel attacks with 3 steps due to computational overhead, others are pre-traiend with PGD-7 attacks. All models are trained on ResNet18. AA stands for robust accuracy against AutoAttack [6].

	Method	CIFAR-10			STL-10			CIFAR-100		
		Clean	PGD-20	AA	Clean	PGD-20	AA	Clean	PGD-20	AA
SSL	RoCL [16]	76.76	50.72	45.52	60.44	31.90	27.38	51.91	27.77	22.79
SSL	ACL [15]	75.99	50.35	45.50	63.46	30.24	25.73	51.91	27.77	22.79
SSL	BYORL [10]	76.39	50.51	45.37	62.85	28.15	24.23	52.37	28.09	23.11
SSL	**Ours (3 steps)**	74.26	49.38	44.31	53.46	32.65	28.96	50.23	27.05	21.96

clean accuracy over the adversarial meta-learning baselines, while obtaining competitive clean accuracy to MAML. In particular, TROBA shows better robustness compared to baselines even though the distribution of the unseen domain is highly different from the distributions of the meta-trained dataset (i.e., CUB, Flowers, Cars). Further, TROBA has smoother loss surface to adversarial examples compared to the baseline, which is why TROBA could demonstrate better robustness in unseen domain (Figure 2).



(a) AQ (b) Ours

Figure 2: Loss surface of unseen domain (Mini-ImageNet)

4.2 Transferable Robustness in Different Domains

To demonstrate the power of our adversarially transferable meta-trained model, we further evaluate our model on a standard transfer learning scenario that employs full data to fully train the encoder with the linear layer on top of it. Specifically, we want to evaluate the generalizable robustness of the representations learned by our encoder against a self-supervised learning model trained with a large amount of data. We evaluate our model on the seen domain, CIFAR-100, as well as on two unseen domains, which are CIFAR-10 and STL-10 respectively. As shown in Table 2, our model shows comparable clean accuracy and robustness in the unseen domains despite the difference in the amount of data used to train the model. Our model is pre-trained with scarce data, and we have even reduced the number of the steps for the bilevel attack to 3 steps to reduce the computational cost, but obtains competitive performance to the model trained with larger data. The experimental results suggest that we may use our method as a means of pretraining the representations to ensure robustness for a variety of applications, when the training data is scarce.

5 Conclusion

We proposed a novel adversarial self-supervised meta-learning framework that can learn transferable robust representations using only a few data via bilevel attack, which introduces a novel bilevel parameter augmentation along with dynamic instance-wise attack. Specifically, the bilevel attack leverages self-supervised learning to effectively generate robust representation of multi-views with differently augmented encoder, which allows learning non-linear transformation task-adaption that brings good robust generalization power. While previous adversarial meta-learning methods are extremely vulnerable to unseen domains, our model learned generalized robust representations which can demonstrate impressive transferable robustness on few-shot tasks in unseen domains. Moreover, we validate our models on larger data in unseen domains which shows comparable robust representations with self-supervised learning (SSL) model with much fewer data. We hope that our work inspires adversarial meta-learning to obtain a good robust representations only using a few data.

Acknowledgement

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2020-0-00153) and Artificial Intelligence Graduate School Program (KAIST), (No.2019-0-01906). We thank Jihoon Tack, Hayeon Lee, and Seul Lee for providing helpful feedbacks and suggestions in preparing an earlier version of the manuscript.

References

[1] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Šrndić, P. Laskov, G. Giacinto, and F. Roli. Evasion attacks against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 387–402. Springer, 2013.

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

[3] Y. Carmon, A. Raghunathan, L. Schmidt, P. Liang, and J. C. Duchi. Unlabeled data improves adversarial robustness. *Advances in Neural Information Processing Systems*, 2019.

[4] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607. PMLR, 2020.

[5] X. Chen and K. He. Exploring simple siamese representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021.

[6] F. Croce and M. 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.

[7] C. Finn, P. Abbeel, and S. Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pages 1126–1135. PMLR, 2017.

[8] M. Goldblum, L. Fowl, and T. Goldstein. Adversarially robust few-shot learning: A meta-learning approach. *Advances in Neural Information Processing Systems*, 33:17886–17895, 2020.

[9] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

[10] S. Gowal, P.-S. Huang, A. van den Oord, T. Mann, and P. Kohli. Self-supervised adversarial robustness for the low-label, high-data regime. In *International Conference on Learning Representations*, 2020.

[11] S. Gowal, S.-A. Rebuffi, O. Wiles, F. Stimberg, D. A. Calian, and T. A. Mann. Improving robustness using generated data. *Advances in Neural Information Processing Systems*, 34: 4218–4233, 2021.

[12] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33: 21271–21284, 2020.

[13] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum contrast for unsupervised visual representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.

[14] D. Hendrycks and T. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *International Conference on Learning Representations*, 2019.

[15] Z. Jiang, T. Chen, T. Chen, and Z. Wang. Robust pre-training by adversarial contrastive learning. In *Advances in Neural Information Processing Systems*, 2020.

[16] M. Kim, J. Tack, and S. J. Hwang. Adversarial self-supervised contrastive learning. *Advances in Neural Information Processing Systems*, 2020.

[17] G. Koch, R. Zemel, R. Salakhutdinov, et al. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2, page 0. Lille, 2015.

[18] S. Kornblith, M. Norouzi, H. Lee, and G. Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, pages 3519–3529. PMLR, 2019.

[19] A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[20] Y. Le and X. Yang. Tiny imagenet visual recognition challenge. In *http://cs231n.stanford.edu/*, 2015.

[21] Z. Li, F. Zhou, F. Chen, and H. Li. Meta-sgd: Learning to learn quickly for few-shot learning. *arXiv preprint arXiv:1707.09835*, 2017.

[22] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

[23] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2574–2582, 2016.

[24] A. Nichol, J. Achiam, and J. Schulman. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*, 2018.

[25] J. Oh, H. Yoo, C. Kim, and S.-Y. Yun. Boil: Towards representation change for few-shot learning. *International Conference on Learning Representations*, 2020.

[26] T. Pang, X. Yang, Y. Dong, H. Su, and J. Zhu. Bag of tricks for adversarial training. *International Conference on Learning Representations*, 2022.

[27] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *IEEE symposium on security and privacy (sp)*, pages 582–597. IEEE, 2016.

[28] A. Raghu, M. Raghu, S. Bengio, and O. Vinyals. Rapid learning or feature reuse? towards understanding the effectiveness of maml. *International Conference on Learning Representations*, 2019.

[29] S.-A. Rebuffi, S. Gowal, D. A. Calian, F. Stimberg, O. Wiles, and T. Mann. Fixing data augmentation to improve adversarial robustness. *Advances in Neural Information Processing Systems*, 2021.

[30] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[31] A. Shafahi, P. Saadatpanah, C. Zhu, A. Ghiasi, C. Studer, D. Jacobs, and T. Goldstein. Adversarially robust transfer learning. *International Conference on Learning Representations*, 2019.

[32] J. Snell, K. Swersky, and R. Zemel. Prototypical networks for few-shot learning. *Advances in Neural Information Processing Systems*, 30, 2017.

[33] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. 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.

[34] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[35] Y. Tian, D. Krishnan, and P. Isola. Contrastive multiview coding. In *European Conference on Computer Vision*, pages 776–794. Springer, 2020.

[36] R. Wang, K. Xu, S. Liu, P.-Y. Chen, T.-W. Weng, C. Gan, and M. Wang. On fast adversarial robustness adaptation in model-agnostic meta-learning. *International Conference on Learning Representations*, 2021.

[37] Y. Wang, D. Zou, J. Yi, J. Bailey, X. Ma, and Q. Gu. Improving adversarial robustness requires revisiting misclassified examples. In *International Conference on Learning Representations*, 2019.

[38] C. Yin, J. Tang, Z. Xu, and Y. Wang. Adversarial meta-learning. *arXiv preprint arXiv:1806.03316*, 2018.

[39] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR, 2021.

[40] H. Zhang, Y. Yu, J. Jiao, E. P. Xing, L. E. Ghaoui, and M. I. Jordan. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning*, 2019.

Supplementary Material

Few-Shot Transferable Robust Representation Learning via Bilevel Attacks

A Related works

A.1 Adversarial learning

The vulnerability of deep neural network (DNN) to imperceptible small perturbation on the input is a well-known problem as observed in previous works [1, 14, 34]. To overcome the adversarial vulnerability, many attack-based approaches for constructing perturbed examples [9, 2, 27] have appeared. On the other hand, Madry et al. [22] proposes a defense-based approach against adversarial examples. Madry et al. [22] utilizes a project gradient descent (PGD) in the perspective of robust optimization, which maximizes the loss in the inner-maximization loops while minimizing the overall loss on tasks in outer-minimization loops, the so-called min-max formulation. Zhang et al. [40] theoretically shows the trade-off between clean accuracy and robustness in adversarial training. To improve both clean and robust accuracy, TRADES [40] introduces regularized surrogate loss. Especially, the Kullback-Leibler divergence (KLD) in TRADES [40] helps to enhance the robustness by enforcing consistency between representations of clean and adversarial examples. Afterward, significant advances in adversarial robustness have emerged. Kim et al. [16], Jiang et al. [15] proposes a self-supervised adversarial learning mechanism coined with contrastive learning to obtain a robust representation without explicit labels. Since a larger dataset is essential to have better adversarial robustness, Shafahi et al. [31] leverages transfer learning to transfer learned robust representations into new target domains with only a few data. Goldblum et al. [8] proposes robust supervised meta-learners with adversarial query images in few-shot classification tasks. However, previous works still have difficulty in obtaining generalized robustness on multiple datasets.

A.2 Self-supervised learning

Conventional adversarial learning mechanisms in a supervised manner require the label information which needs expensive human labeling annotations. Self-supervised learning makes the neural networks possible to learn comparable representations to supervised representations, even it does not leverage labels [12]. Many previous works focus on learning consistent representations to different distortions in the input [17, 4, 13, 35]. To learn the distortion-invariant representations, they enforce consistency between representations of two differently augmented inputs with the same instance-level identity. Especially, Chen et al. [4] employs contrastive learning to maximize agreement only between positive pairs in mini-batch while negative pairs are handled as the opposite. In advance, other works introduce asymmetry into network architecture or parameter update [5, 12] to improve performance. However, the existence of trivial solutions derived from asymmetry leaves room to improve. Zbontar et al. [39] achieves comparable performance by introducing redundancy reduction terms in the training objectives, even it does not require additional asymmetric networks or large batches.

A.3 Self-supervised adversarial learning

Utilizing the advantages of self-supervised learning, adversarial training mechanisms in a self-supervised manner have emerged to learn robust representations without relying on label information. Recent works leverage contrastive learning to obtain robust representation in a self-supervised manner [16, 15]. Kim et al. [16] first devises the instance-wise adversarial perturbation, which does not require explicit labels during the attack, and utilizes those perturbed examples in maximizing contrastive loss. Jiang et al. [15] introduces a dual stream with optimizing two contrastive losses against four augmented views, which are computed between clean views and adversarial images, respectively. However, these approaches highly rely on large batch sizes to effectively create positive and negative samples for the contrastive learning framework. Gowal et al. [10] injects adversarial examples on top of the BYOL framework [12] to achieve robustness to avoid the restrictions on large batch sizes. Although existing restrictions on large batch sizes or image augmentation have been

relieved during extensive development in self-supervised adversarial training, obtaining robustness with scarce data is still difficult, even in a supervised manner.

B Preliminaries

Model-Agnostic Meta-Learning. Let us denote the encoder as f_θ and classifier as h_μ. Since meta-learning aims to learn to learn the new tasks, it needs to train on a large number of tasks τ sampled from a task distribution $p(\tau)$, where a given task consists of the support set $\mathbf{S}_\tau$ and the query set $\mathbf{Q}_\tau \in \mathbf{D}$. Each set contains a n-way k-shot classification task, that classify n classes with k images, i.e., $n \times k$ instances. The most popular framework for meta-learning is model-agnostic meta-learning (MAML) [7], which meta-learns the model with a bilevel optimization scheme, with inner optimization and outer meta-level optimization steps. During the inner optimization, we adapt the shared initial parameter to each new task τ to obtain task-adaptive parameters θ^τ and μ^τ, by taking a few gradient steps, as follows:

$$\theta^\tau, \mu^\tau = \theta, \mu - \alpha \nabla_{\theta,\mu} \mathcal{L}_{\mathbf{S}_\tau}(h_\mu(f_\theta)), \tag{4}$$

where $\mathbf{S}_\tau$ is a support set of task τ, α is the step size and $\mathcal{L}$ is a task-specific loss to conduct gradient step for inner updates (e.g. cross-entropy loss). There also exist different variants of the MAML framework with respect to which parameters to update. ANIL [28] only meta-learns the final linear layer while fixing the encoder (i.e., $\theta^\tau = \theta$), for rapid adaptation to a new task while reusing the features. On the other hand, BOIL [25] only meta-learns the encoder, thus the representation layers, while keeping the final classifier fixed (i.e., $\mu^\tau = \mu$). We employ BOIL [25] which only updates the encoder because our focus is on learning generalizable robust representations. In the meta-optimization phase, model parameters are updated with meta-objective via stochastic gradient descent (SGD) as follows:

$$\theta, \mu \leftarrow \theta, \mu - \beta \nabla_{\theta,\mu} \sum \mathcal{L}_{\mathbf{Q}_\tau}(h_{\mu^\tau}(f_{\theta^\tau})), \tag{5}$$

where $\mathbf{Q}_\tau$ is a query set of task τ, β is a meta step size and $\mathcal{L}$ is meta-objective. The meta-objective is a summation of losses from the query set of all given tasks, where the losses depend on what aims to be meta-learned. To reduce the computation overhead in MAML, we use Meta-SGD [21] which learns the learning rate of parameters that enables to initialize and adapt any differentiable learner in a single step.

Attacking a meta-learner. To obtain robustness on few-shot tasks, Adversarial Querying (AQ) [8] proposes to generate attacks with only the query examples. The AQ employs the project gradient descent attack (PGD [22]), which is a class-wise attack that maximizes the cross-entropy on a given query image as follows,

$$\delta^{t+1} = \Pi_{B(x^q, x^q+\epsilon)}\Big(\delta^t + \gamma \texttt{sign}\Big(\nabla_{\delta^t} \mathcal{L}_{\texttt{CE}}\big(h_{\mu^\tau}(f_{\theta^\tau}(x^q + \delta^t)), y^q\big)\Big)\Big), \tag{6}$$

where x^q and y^q is a query image and its label of task τ, respectively, $B(x^q, x^q + \epsilon)$ is the l_∞ norm-ball around x^q with radius ϵ, γ is step size of the attack, δ is perturbation and the cross entropy loss ($\mathcal{L}_{\texttt{CE}}$) is calculated on the inner updated parameters (θ^τ, μ^τ).

Robust training loss. Various adversarial training methods have been proposed to enhance the model's robustness to adversarial attacks (Appendix A.1). Among them, we adapt the TRADES [40] loss to improve robustness. TRADES proposes to regularize the model's outputs on the clean and adversarial examples with Kullback-Leibler divergence (KLD) as follows:

$$\mathcal{L}_{\texttt{TRADES}} = \mathcal{L}_{\texttt{CE}}\big(h_{\mu^\tau}(f_{\theta^\tau}(x^q)), y^q\big) + \beta \max_{\delta \in B(x^q, x^q+\epsilon)} \mathcal{L}_{\texttt{KL}}\big(h_{\mu^\tau}(f_{\theta^\tau}(x^q)) || h_{\mu^\tau}(f_{\theta^\tau}(x^q + \delta))\big), \tag{7}$$

where $\mathcal{L}_{\texttt{CE}}$ is cross-entropy loss on clean examples, $\mathcal{L}_{\texttt{KL}}$ is KLD loss between clean and adversarial logit to obtain robustness, and β is a regularizer to control the trade-off between clean accuracy and robustness which normally set as 6.0. In our framework, we calculate the adversarial loss on query sets (x^q, y^q), which are different instances used in inner adaptation, to meta-learn robust representations in the meta-optimization phase.

C Experimental details

C.1 Dataset

For meta-training, we use CIFAR-FS [19] and Mini-ImageNet [30]. CIFAR-FS and Mini-ImageNet consist of 100 classes which are 64, 16, and 20 for meta-training, meta-validation, and meta-testing, respectively. We validate our model on 6 benchmark few-shot datasets: CIFAR-FS [19], Mini-ImageNet [30], Tiered-ImageNet [30], Cars, CUB and VGG-Flower, for few-shot classification and 3 additional benchmark standard image classification datasets: CIFAR-10, CIFAR-100, and STL-10, for robust transferability. CIFAR-10 and CIFAR-100 consist of 50,000 training images and 10,000 test images with 10 and 100 classes, respectively. All images are used with $32\times32\times3$ resolution (width, height, and channel) for meta-training. Especially, we apply *TorchMeta*[2] library to load the few-shot datasets into our frameworks.

C.2 Meta-train

We meta-train ResNet12 and ResNet18 as the base encoder network on CIRAR-FS and Mini-ImageNet. All models are meta-trained with tasks consist of 5-way 5-shot support set images and 5-way 15shot query set images, and meta-validated with only clean tasks consist of 5-way 1-shot support set images and 5-way 15-shot query set images. Especially, we train the model with randomly selected 200 tasks and validate the model with randomly selected 100 tasks. For optimization, we meta-train our models with 300 epochs under SGD optimizer with weight decay 1e-4. For data augmentation, we use random crop with 0.08 to 1.0 size, color jitter with probability 0.8, horizontal flip with probability 0.5, grayscale with 0.2, gaussian blur with 0.0, and solarization probability with 0.0 to 0.2. We exclude normalization for adversarial training.

In the case of adversarial learning, we use our proposed bilevel attack with 3 steps and 7 steps. To generate adversaries with query set images, we take the gradient step within l_∞ norm ball with ϵ = $8.0/255.0$ and α = $2.0/255.0$ to maximize the similarity with target instance. To obtain robust representation, we utilize an adversarial loss and self-supervised loss which are TRADES [40] with a regularization hyperparameter of 6.0 and cosine similarity loss, respectively.

Three different meta-learning frameworks are leveraged to train our model, which are MAML [7], FOMAML [7] and Meta-SGD [21]. Specifically, we only update the encoder parameters in inner optimization for all three meta-learning strategies. Detailed hyperparameters for meta-train and meta-test will be described in Appendix C.3.

C.3 Hyperparameter details of each meta-learning frameworks

MAML. We take a single step for both inner optimization and outer optimization to meta-train ResNet12 on CIFAR-FS and Mini-ImageNet. We use the same learning rate for both datasets, which are 0.3 and 0.08 for outer optimization and inner optimization, respectively. For both dataset, we use batch size 4.

FOMAML. To reduce the computational cost, we try to adapt FOMAML [7], which is the first-order approximation of MAML [7]. For ResNet18, we use a single step in both inner optimization and outer optimization, and use the learning rates 0.3 and 0.4 in outer optimization and inner optimization, respectively. For ResNet12, we use 3 steps for inner optimization, and 1 step for outer optimization. We use learning rate 0.3 and 0.2 for outer optimization and inner optimization, respectively. For both dataset, we use batch size 4.

Meta-SGD. To learn quickly, we use the Meta-SGD [21] with the single step. We use a single step in inner optimization and use the 0.005 inner learning rate. For outer loop, we use 0.005 outer learning rate for CIFAR-FS. For Mini-ImageNet, we use a same step size as CIFAR-FS but with different inner learning rate, 0.001, and outer optimization learning rate 0.001. For both dataset, we use batch size 4.

[2] https://github.com/tristandeleu/pytorch-meta

C.4 Meta-test

The trained models are evaluated with 400 randomly selected tasks from test set, where each task consists of 5-way 5-shot support set images and 5-way 15-shot query set images. We use a single step in both inner optimization and outer optimization. We especially use same learning rate and meta step size as the model is meta-trained.

C.5 Adversarial evaluation

Few-shot robustness. We validate the robustness of our trained models against two types of attack, which are PGD [22] and AutoAttack [6]. All l_∞ PGD attacks are conducted with the norm ball size $\epsilon = 8./255.$, step size $\alpha = 8./2550.$, and with 20 steps of inner maximization. AutoAttack[3] is a combination of 4 different types of attacks (i.e., APGD-CE, APGD-T, FAB-T, and Square). We use the standard version of AutoAttack in the test time.

Self-supervised robust linear evaluation. To compare TROBA with self-supervised pre-trained models, we apply robust full-finetuning. In robust full-finetuning, the parameters of the entire network, including the feature extractor and the fc layer, are trained with adversarial examples. We generate perturbed examples with l_∞ PGD-10 attack with $\epsilon = 8./255.$ and step size $\alpha = 2./255.$ in training. All adversarially full-finetuned models are evaluated against l_∞ PGD-20 attack ($\epsilon = 8./255.$, $\alpha = 8./2550.$) and AutoAttack [6]. Especially, in comparisons with self-supervised models, we pre-train ResNet18 based on FOMAML [7], which is the first-order approximation of MAML [7], and apply bilevel attacks with 3 steps to reduce the computational cost. Other self-supervised models are pre-trained with PGD-7 attacks. For optimization, we fine-tune the pre-trained models for 110 epochs with batch size 128 under SGD optimizer with weight decay 5e-4, where Pang et al. [26] demonstrated as optimal for robust full-finetuning on CIFAR datasets.

C.6 Comparison with self-supervised pre-trained models

We select baseline models with ACL [15][4], BYORL [10] and RoCL [16][5] for self-supervised pre-trained baselines. We implement BYORL on top of the BYOL [12][6] framework, following description in the paper.

D Implementation details of ablation studies

D.1 Ablation study of bilevel parameter augmentation

To demonstrate how bilevel parameter augmentation is more effective than image augmentation in adversarial self-supervised meta-learning, we experiment in the same environment except for parameter augmentation in inner adaptation. Specifically, we generate augmented parameters of the encoder adapted with two differently transformed support set images simultaneously, while TROBA augments parameters independently for each augmented view. A detailed algorithm for applying bilevel parameter augmentation and image-only augmentation in adversarial self-supervised meta-learning is described in Algorithm 1 and Algorithm 2, respectively. Experiment results are reported in Appendix E.

D.2 Ablation study of bilevel attack

The bilevel attack is based on the instance-wise attack [16] which does not require label information to generate adversaries, while the class-wise attack utilizes label to maximize the cross-entropy loss in the inner maximization of Equation 6. The bilevel class-wise attack is applied with the bilevel augmented parameters as done in the bilevel attack. We use l_∞ PGD attack with strength $8./255.$, step size $2./255.$, and the same number of iterations with bilevel attacks in all comparisons in the main paper.

[3] https://github.com/fra31/auto-attack

[4] https://github.com/VITA-Group/Adversarial-Contrastive-Learning

[5] https://github.com/Kim-Minseon/RoCLforself-supervisedlearning

[6] https://github.com/lucidrains/byol-pytorch

Algorithm 1 Transferable robust meta learning via bilevel attack (TROBA)

Require: Dataset $\mathbf{D}$, transformation function $t \sim \mathcal{T}$
Require: Encoder f, parameter of encoder θ, classifier h, parameter of classifier μ
Require: Adversary $\mathcal{A}(\texttt{base}, \texttt{target}, \texttt{parameter})$
  while not done **do**
  Sample tasks $\{\tau\}$, Support set $\mathbf{S}(x^s, y^s)$, Query set $\mathbf{Q}(x^q, y^q)$
    for $\tau = 1, \cdots,$ **do**
      Transform input $t_1(x^s), t_2(x^s)$
      Fine-tune model with $t_1(x^s), y^s$ and updates parameter θ_1^τ
      Fine-tune model with $t_2(x^s), y^s$ and updates parameter θ_2^τ
      Generate adversarial examples
      $t_1(x^q)^{adv} = \mathcal{A}(t_1(x^q), t_2(x^q), \theta_1^\tau)$, $t_2(x^q)^{adv} = \mathcal{A}(t_2(x^q), t_1(x^q), \theta_2^\tau)$
      Compute gradient $g^\tau = \nabla_{\theta_1^\tau, \theta_2^\tau} \mathcal{L}_{\texttt{Our}}(h_\mu, f_{\theta_1^\tau}, f_{\theta_2^\tau}, t_1(x^q), t_1(x^q)^{adv}, t_2(x^q), t_2(x^q)^{adv}, y^q)$
    end for
    Update model parameters
    $\theta, \mu \leftarrow \theta, \mu - \frac{\alpha}{\tau} \sum g^\tau$
  end while

Algorithm 2 Transferable robust meta learning via image-only augmentation

Require: Dataset $\mathbf{D}$, transformation function $t \sim \mathcal{T}$
Require: Encoder f, parameter of encoder θ, classifier h, parameter of classifier μ
Require: Adversary $\mathcal{A}(\texttt{base}, \texttt{target}, \texttt{parameter})$
  while not done **do**
  Sample tasks $\{\tau\}$, Support set $\mathbf{S}(x^s, y^s)$, Query set $\mathbf{Q}(x^q, y^q)$
    for $\tau = 1, \cdots,$ **do**
      Transform input $t_1(x^s), t_2(x^s)$
      Fine-tune model with $t_1(x^s), t_2(x^s), y^s$ and updates parameter θ^τ
      Generate adversarial examples
      $t_1(x^q)^{adv} = \mathcal{A}(t_1(x^q), t_2(x^q), \theta^\tau)$, $t_2(x^q)^{adv} = \mathcal{A}(t_2(x^q), t_1(x^q), \theta^\tau)$
      Compute gradient $g^\tau = \nabla_{\theta^\tau} \mathcal{L}_{\texttt{Our}}(h_\mu, f_{\theta^\tau}, t_1(x^q), t_1(x^q)^{adv}, t_2(x^q), t_2(x^q)^{adv}, y^q)$
    end for
    Update model parameters
    $\theta, \mu \leftarrow \theta, \mu - \frac{\alpha}{\tau} \sum g^\tau$
  end while

E Results of ablation Studies

To examine each component of our proposed methods, we conduct the ablation study on augmentation, loss, and attack. Through our ablation study, we verify the effectiveness of each component by their robustness on unseen domains.

Bilevel parameter augmentation contributes to learn generalized features. As shown in Table 3, image-only augmentation alone meaningfully contributes to learning generalized features for unseen domains. However, when we apply parameter augmentation on top of the image augmentation, the model achieves significantly better clean and robust accuracy than the model trained with image-only augmentation, especially in the seen domain. This suggests that the bilevel parameter augmentation is effective in learning consistent representations across tasks and views. To support our claim, we calculate the Centered Kernel Alignment (CKA) [18] value, which measures the similarity between representations (When representations are identical, the CKA is 1). As shown in Figure 3, when bilevel parameter augmentation is applied, features from the augmented parameters are more dissimilar than features with image augmentation only. These results show that our bilevel parameter augmentation may generate more different multi-views of the same instances which helps learn invariant representations across views, that help it to achieve generalizable robustness.

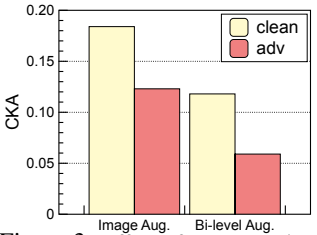
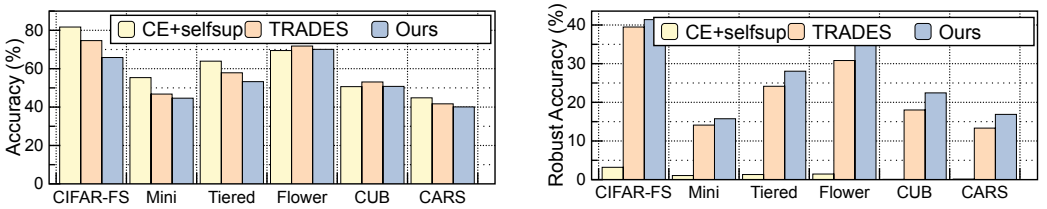


Figure 3: Effect of augmentation

Table 3: Ablation study of our proposed bilevel augmentation. Test accuracy(%) on seen domain (CIFAR-FS) and unseen domains (Mini-ImageNet, Flower, Cars) of 5-way 5-shot task. Robustness is calculated with PGD-20 attack($\epsilon = 8./255.$, step size=$\epsilon/10$), clean stands for accuracy of clean images. All models are adversarially meta-trained on CIFAR-FS with attack step 3 due to computation costs.

Augmentation level		CIFAR-FS		Mini-ImageNet		Flower		Cars	
Image Aug.	Parameter Aug.	Clean	PGD ℓ_∞	Clean	PGD ℓ_∞	Clean	PGD ℓ_∞	Clean	PGD ℓ_∞
✓	-	63.10	36.98	39.54	15.08	51.57	25.05	38.99	14.36
✓	✓	**65.82**	**41.39**	**44.64**	**15.75**	**53.25**	**28.05**	**40.08**	**16.88**



(a) Clean Accuracy (b) Robustness

Figure 4: Ablation on meta-objectives in TROBA. Test accuracy(%) on the seen domain (CIFAR-FS) and unseen domains (Mini: Mini-ImageNet, Tiered: Tiered-ImageNet, Flower, CUB, Cars) of 5-way 5-shot task. Legends denote the meta-objectives loss that is used to train the model. All models are adversarially meta-trained on CIFAR-FS with attack step 3 due to computation overhead. (a) Clean accuracy stands for the accuracy of clean images. (b) Robustness is calculated with PGD-20 attack($\epsilon = 8./255.$, step size=$\epsilon/10$).

Self-supervised loss regularized to learn generalized features. TROBA leverages both adversarial loss and self-supervised loss in meta-objective; specifically, it uses TRADES loss (Equation 7) and cosine similarity loss between representations of differently bilevel augmented views, as shown in Equation 3. The adversarial loss is calculated independently in each bilevel augmented network to enhance the robustness on each training sample. On the other hand, the self-supervised loss is computed between the representations of each bilevel augmented encoder to enforce the consistency across features for samples attacked with our bilevel attack, which helps it to obtain a consistent representation space across perturbations and instances, which helps with its generalization (Figure 4). Notably, the self-supervised loss has a larger contribution when we conduct transfer learning to unseen domains with larger data (Appendix F.3).

Further, we replace the adversarial loss term with AT [22], which is widely used to obtain robustness in adversarial learning, while utilizing the same self-supervised loss. As shown in Table 4, utilizing a TRADES [40] loss as an adversarial loss is more effective to obtain transferable robustness in adversarial meta-learning than a AT [22] loss.

Bilevel attack makes the model to be robust on unseen domain attacks. We further analyze the effect of our bilevel instance-wise attack compared to class-wise attack in Table 5. We observe that adversarial examples that are attacked with instance-wise attack make the model more robust in unseen domains compared to class-wise attack. Specifically, instance-wise attack generates adversaries that have larger difference to clean examples in the representation level, and thus can be thought as a stronger attack. To demonstrate the effectiveness of instance-wise attack, we calculate CKA [18] between clean and adversarial features from each bilevel augmented parameters. As shown in Figure 5, instance-wise attack produces more difficult adversarial examples that are highly dissimilar from clean instances. However, when the parameter is augmented with bilevel parameter augmentation, the class-wise attack also can show transferable robustness since self-supervised loss supports it to obtain generalized robustness.

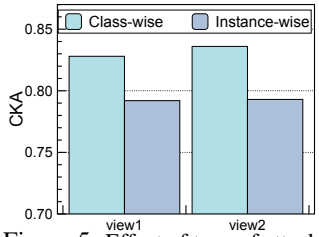


Figure 5: Effect of type of attack

F Additional experimental results of robustness

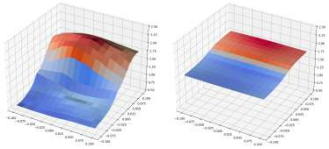
F.1 Robustness on seen domains

Table 4: Ablation study on adversarial loss in meta-objectives of TROBA. Test accuracy(%) on benchmark data sets for 5-shots. Robustness is calculated with PGD-20 attack ($\epsilon = 8./255.$, step size=$\epsilon/10$), clean is for clean images. All models are adversarially meta-trained on CIFAR-FS, with ResNet18 as the base encoder.

	Mini-ImageNet		Tiered-ImageNet		CUB		Flower		Cars	
Adversarail Loss	Clean	PGD ℓ_∞	Clean	PGD ℓ_∞	Clean	PGD ℓ_∞	Clean	PGD ℓ_∞	Clean	PGD ℓ_∞
AT [22]	**33.78**	7.99	38.35	12.95	37.00	10.08	42.28	22.19	30.93	9.59
TRADES [22]	33.57	**16.26**	**39.26**	**21.82**	**39.90**	**18.62**	**48.70**	**36.92**	**34.69**	**17.67**

Table 5: Ablation study of our proposed bilevel attack. Test accuracy(%) on seen domain (CIFAR-FS) and unseen domains (Mini-ImageNet, Tiered-ImageNet, Flower, Cars) of 5-way 5-shot task. Clean stands for accuracy of clean images. Rob. stands for robust accuracy that is calculated with PGD-20 attack($\epsilon = 8./255.$). All models are adversarially meta-trained on CIFAR-FS with attack step 3 due to computation costs.

	CIFAR-FS		Mini-ImageNet		Tiered-ImageNet		Flowers		CUB	
Attack type	Clean	Rob.	Clean	Rob.	Clean	Rob.	Clean	Rob.	Clean	Rob.
Bi-level class-wise	**66.69**	40.48	42.15	**17.13**	**53.91**	27.41	69.66	38.83	50.01	21.2
Bi-level instance-wise	65.82	**41.39**	**44.64**	15.75	53.25	**28.05**	**70.08**	**41.52**	**50.78**	**22.44**



(a) AQ (b) Ours

Figure 6: Loss surface of seen domain (CIFAR-FS)

Even though TROBA is designed to have transferable robustness in the unseen domain, our methods also show better robustness in seen domain few shot tasks compare to baselines, even with better clean accuracy (Table 7). In addition, TROBA shows smoother loss surface to adversarial examples which is also directly associated with better robustness and generalization (Figure 6). Our method is agnostic to the meta-learning approach, as shown in Table 8, which suggests that the type of meta-learning strategy is not the main factor in achieving the transferable robustness. We only update the encoder in the inner optimization for all meta-learning algorithms.

F.2 Robustness on unseen domains with different meta-learning framework and different iterations of bilevel attack

We additionally validate our models on Mini-ImageNet (Table 6). Further, to prove that TROBA is an effective method to obtain transferable robust representations, we experiment with three different types of meta-learning frameworks and different strengths of bilevel attacks. Specifically, we train TROBA on top of the MAML [7], FOMAML [7] and MetaSGD [21] and apply bilevel attacks with 3 steps and 7 steps, respectively. Here, we only update the encoder parameters in inner adaption, since we propose task adaptive attacks that maximize the difference between the features, further to learn generalized representations as BOIL [25] demonstrated.

As shown in Table 9, TROBA outperforms the previous adversarial meta-learning model [8] by more than 10% robustness regardless of meta-learning strategies. Furthermore, we show outstanding robustness with only 3 steps of bilevel attacks (i.e., dynamic instance-wise attack) compared to AQ [8], which is trained with PGD-7 attacks (i.e., class-wise attack). To demonstrate that a bilevel attack is a more effective attack than a class-wise attack in the representation level, we calculate CKA [18] between clean and adversarial features to measure the similarity in the feature level. Notably, the CKA value of features attacked with the bilevel attack is smaller than the CKA values of features attacked with the class-wise attack (Figure 5), which means that the bilevel attack constructs more confusing perturbed images that are more dissimilar from their clean examples. Through these remarkable results, we demonstrate that our proposed bilevel attack served as a stronger attack that makes the model to have robust transferability to unseen domains, even with fewer gradient steps of attacks and little data.

F.3 Robustness on unseen domains with larger datasets

In the main paper, we validate our models on unseen domains with larger benchmark datasets for standard image classification, which are CIFAR-10 and STL-10. Furthermore, we also demonstrate the robust transferability of our models in benchmark few-shot image classification tasks,

Table 6: Results of transferable robustness in 5-way 5-shot unseen domain tasks that are trained on 5-way 5-shot Mini-ImageNet. Rob. stands for accuracy(%) that is calculated with PGD-20 attack ($\epsilon = 8./255.$). Clean stands for test accuracy(%) of clean images. All models are trained with PGD-7 attacks on ResNet12.

	Mini-ImageNet →									
	CIFAR-FS		Tiered-ImageNet		CUB		Flowers		Cars	
	Clean	Rob.	Clean	Rob.	Clean	Rob.	Clean	Rob.	Clean	Rob.
MAML [7]	**66.75**	12.97	**65.33**	13.10	52.82	4.46	**71.01**	4.86	**43.66**	2.77
ADML [38]	41.14	13.36	41.05	13.26	32.82	4.59	43.07	9.65	24.85	5.48
AQ [8]	61.97	30.73	47.61	14.21	45.64	13.19	65.40	25.01	37.29	8.85
RMAML [36]	37.94	10.59	30.49	8.24	27.30	6.26	42.52	13.08	37.76	5.43
Ours	65.45	**36.51**	59.64	**29.73**	**53.70**	**20.64**	69.84	**36.49**	42.25	**14.42**

Table 7: Comparison in the 5-shots seen domain tasks. All models are trained on CIFAR-FS and Mini-ImageNet, respectively, with PGD-7 attack in ResNet12. * stands for reported results in Wang et al. [36].

	CIFAR-FS		Mini-ImageNet	
	Clean	Rob.	Clean	Rob.
ADML [7]	53.06	22.45	26.72	6.81
AQ [8]	73.49	28.49	39.47	13.52
RMAML [36]*	57.95	35.30	43.98	21.47
Ours [21]	64.90	43.34	47.56	18.18

Table 8: Results of TROBA with a different meta-learning frameworks in 5-shot tasks. All models are trained on CIFAR-FS and Mini-ImageNet, respectively, with PGD-7 attacks in ResNet12.

	CIFAR-FS		Mini-ImageNet	
	Clean	Rob.	Clean	Rob.
TROBA				
+MAML [7]	52.79	32.50	37.58	14.23
+FOMAML [7]	53.42	35.95	33.87	15.60
+Meta-SGD [21]	64.90	43.34	47.56	18.18

which are Cars, CUB, and Aircraft that have 196, 200, and 100 classes, respectively. Especially, we train our models on ResNet18 with bilevel attacks with 3 steps while other self-supervised models are trained with PGD-7 attacks due to computation costs. We use the same hyperparameters to validate with robust full-finetuning for all datasets, as we explained in Appendix C.5. Although our models utilize only scarce data to train, and even apply bilevel attacks with fewer gradient steps, we show even better robust representations compared to self-supervised pre-trained models while preserving clean accuracy (Table 10). Especially, our methods show a larger gap in fine-grained datasets, which have highly different distribution from meta-trained domains (i.e., CIFAR-FS). Further, we hope that our models to be robust in real-world adversarial perturbation such as common corruption [14], we evaluate our fully fine-tuned models with adversarial examples on CIFAR-10, with common corruption datasets on CIFAR-10. TROBA shows comparable accuracy with self-supervised pre-trained models on common corruption tasks, even trained with little data and bilevel attacks with fewer inner maximization iterations (Table 11). From these results, we prove that TROBA learns good generalized representations with little data effectively.

Table 11: Test accuracy(%) of TROBA and self-supervised pre-traiend models on common corruption tasks of CIFAR-10.

Model	Accuracy
ACL [15]	68.6
BYORL [10]	69.01
AQ [8]	66.16
TROBA	67.9

G Obfuscated gradient

All of the robust accuracies in our paper are calculated with the strength $\epsilon = 8./255.$, step size $\alpha = 8./2550.$ and 20 steps. To check whether our model is under obfuscated gradient issues or not, we experiment with two different settings of l_∞ PGD attacks. First, we apply PGD attacks with extremely large strength, where robust accuracy should be almost zero. Second, we use the same strength but different step sizes and steps, which are $4./2550.$ and 40, respectively, where robust accuracy should be the same as robust accuracy from our original evaluation setting. Specifically, we demonstrate TROBA trained on CIFAR-FS with ResNet12 as the base encoder, and further on top of the FOMAML reported in Table 9. As shown in Table 12, we verify that our models do not have any obfuscated gradient issues.

Table 9: Results of transferable robustness with different meta-learning framework and attack iteration in 5-shot tasks. All models are trained with 5-way 5-shot images on CIFAR-FS and Mini-ImageNet. Clean stands for test accuracy(%) of clean images. Rob. stands for accuracy(%) that is calculated with PGD-20 attack ($\epsilon = 8./255.$). All models are trained on ResNet12. The number of attack iteration during training is marked in parentheses next to the meta-train dataset. Further, we denote (θ) next to the meta-learning strategies to notice that we update only the encoder parameters during inner optimization.

	CIFAR-FS (3 steps) →	Mini-ImageNet		tiered-ImageNet		CUB		Flowers		Cars	
		Clean	Rob.	Clean	Rob.	Clean	Rob.	Clean	Rob.	Clean	Rob.
TROBA	+MAML (θ) [7]	34.35	15.76	39.06	20.08	42.32	17.46	57.74	32.70	35.78	15.79
TROBA	+FOMAML (θ) [7]	32.06	16.69	37.97	22.15	37.65	17.50	56.68	34.08	36.33	18.45
TROBA	+MetaSGD (θ) [21]	44.64	15.75	53.25	28.05	50.78	22.44	70.08	41.52	40.08	16.88
	AQ [8]	33.79	1.59	36.41	2.27	39.35	2.88	58.69	6.59	37.39	2.30
	CIFAR-FS (7 steps) →	Mini-ImageNet		tiered-ImageNet		CUB		Flowers		Cars	
		Clean	Rob.	Clean	Rob.	Clean	Rob.	Clean	Rob.	Clean	Rob.
TROBA	+MAML (θ) [7]	32.57	16.12	38.90	22.51	39.44	16.52	56.79	32.83	36.58	16.56
TROBA	+FOMAML (θ) [7]	31.71	17.40	37.33	23.28	38.63	18.79	59.57	36.79	37.94	21.34
TROBA	+MetaSGD (θ) [21]	45.82	24.12	51.46	30.06	48.56	25.23	66.49	42.16	38.29	19.43
	AQ [8]	33.09	3.32	37.41	5.05	38.37	4.10	60.14	11.03	36.83	4.47
	Mini-ImageNet (3 steps) →	CIFAR-FS		tiered-ImageNet		CUB		Flowers		Cars	
		Clean	Rob.	Clean	Rob.	Clean	Rob.	Clean	Rob.	Clean	Rob.
TROBA	+MAML (θ) [7]	57.11	30.76	43.15	20.44	46.00	17.03	62.23	32.60	39.70	16.83
TROBA	+FOMAML (θ) [7]	51.48	29.05	39.22	20.92	37.76	14.66	49.80	25.04	38.02	16.07
TROBA	+MetaSGD (θ) [21]	66.48	37.36	59.73	29.35	53.33	20.20	68.93	33.39	42.09	13.73
	AQ [8]	66.52	23.01	48.33	5.70	47.12	7.37	66.80	13.65	37.32	4.34
	Mini-ImageNet (7 steps) →	CIFAR-FS		tiered-ImageNet		CUB		Flowers		Cars	
		Clean	Rob.	Clean	Rob.	Clean	Rob.	Clean	Rob.	Clean	Rob.
TROBA	+MAML (θ) [7]	56.61	35.18	41.96	24.11	44.97	19.64	62.34	34.73	39.85	19.26
TROBA	+FOMAML (θ) [7]	53.42	35.95	37.91	22.15	39.88	17.40	59.66	33.64	39.93	17.94
TROBA	+MetaSGD (θ) [21]	65.45	36.51	59.64	29.73	53.70	20.64	69.84	36.49	42.25	14.42
	AQ [8]	61.97	30.73	47.61	14.21	45.64	13.19	65.40	25.01	37.29	8.85

Table 10: Experiments results for self-supervised robust full-finetuning of TROBA and the state-of-the-art adversarial self-supervised models on unseen domains. While TROBA is trained on CIFAR-FS with bilevel attacks , adversarial self-supervised models are trained on full-dataset of CIFAR-100. All models are trained on ResNet18, and evaluated against PGD-20 attacks ($\epsilon = 8./255.$) and AutoAttack (AA) [6]

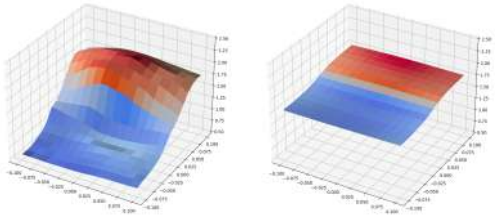
	CARS			CUB			AirCraft		
Method	Clean	PGD ℓ_∞	AA	Clean	PGD ℓ_∞	AA	Clean	PGD ℓ_∞	AA
Self-supervised learning									
RoCL [16]	35.00	8.11	5.67	17.21	2.55	1.71	33.63	8.76	5.61
ACL [15]	30.95	5.86	3.80	17.00	2.33	1.54	31.19	7.26	4.68
BYORL [10]	32.13	6.15	4.39	16.78	2.28	1.48	31.16	6.63	4.17
Meta learning									
Ours (3 steps)	31.47	**9.58**	**6.19**	18.07	**4.49**	**2.73**	32.12	**9.93**	**6.19**

Table 12: Test accuracy(%) on benchmark data sets for 5-shots. Robustness is calculated with PGD-20 attack ($\epsilon = 8./255.$, step size=$\epsilon/10$), clean is for clean images. All models are adversarially meta-trained on CIFAR-FS.

				CIFAR-FS		Mini-ImageNet		Tiered-ImageNet		CUB		Cars	
	Strength (ϵ)	Step size (α)	Steps	Clean	PGD ℓ_∞	Clean	PGD ℓ_∞	Clean	PGD ℓ_∞	Clean	PGD ℓ_∞	Clean	PGD ℓ_∞
3 steps	$8.0/255.0$	$8.0/2550.0$	20	53.42	35.95	32.06	16.69	37.97	22.15	37.65	17.50	36.33	18.45
3 steps	$8.0/255.0$	$4.0/2550.0$	40	53.04	35.35	31.70	16.01	38.06	21.98	37.77	18.12	36.10	18.02
3 steps	300.0	$8.0/2550.0$	20	52.72	0.47	31.83	0.92	37.73	0.85	38.14	0.55	36.21	0.44
7 steps	$8.0/255.0$	$8.0/2550.0$	20	51.90	36.01	31.71	17.40	37.33	23.28	38.63	18.79	37.94	21.34
7 steps	$8.0/255.0$	$4.0/2550.0$	40	52.50	36.39	31.95	17.49	38.44	24.22	38.18	18.87	37.41	20.92
7 steps	300.0	$8.0/2550.0$	20	52.20	0.50	31.97	0.59	37.53	0.65	38.78	0.45	37.64	0.48

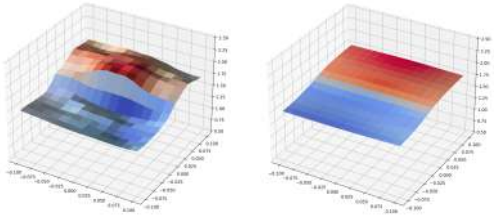
H Visualization of loss surface

We visualize the loss surface of our model and baseline AQ [8] model. As shown in the Figure our model has a smoother loss surface both in the seen domain and unseen domain while the baseline has a relatively less smooth surface.



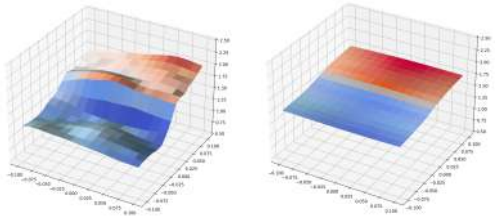
(a) AQ (b) Ours

Figure 7: Seen domain (CIFAR-FS)



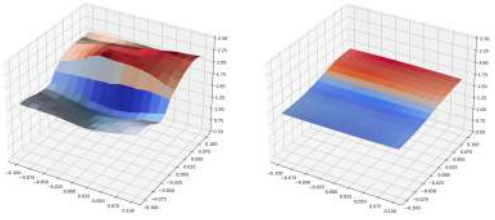
(a) AQ (b) Ours

Figure 8: Unseen domain (Mini-ImageNet)



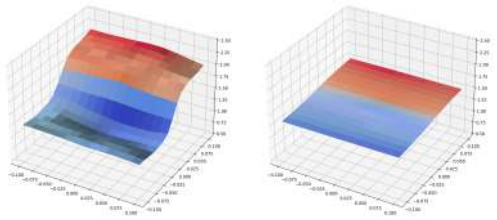
(a) AQ (b) Ours

Figure 9: Unseen domain (Tiered-ImageNet)



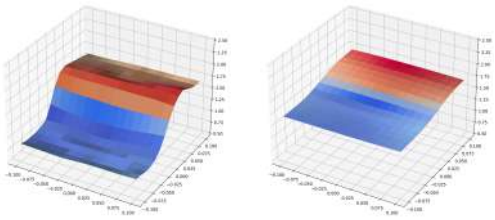
(a) AQ (b) Ours

Figure 10: Unseen domain (CUB)



(a) AQ (b) Ours

Figure 11: Unseen domain (Cars)



(a) AQ (b) Ours

Figure 12: Unseen domain (Flowers)