

# Supplementary Material for "DERD: Data-free Adversarial Robustness Distillation through Self-adversarial Teacher Group"

Anonymous Authors

## 1 HYPERPARAMETERS SELECTION

In this section, we briefly report the various hyperparameter selections for DERD. All experiments were run on NVIDIA GeForce RTX 4090.

**Stage-I.** For CIFAR10, we set  $\lambda_{oh} = 0.05$ ,  $\lambda_{ie} = 5$ ,  $\lambda_{BN} = 1e - 5$ ,  $\lambda_a = 0.01$ .

For CIFAR100, we set  $\lambda_{oh} = 0.5$ ,  $\lambda_{ie} = 20$ ,  $\lambda_{BN} = 1e - 5$ ,  $\lambda_a = 0.1$ .

For ImageNet100, we set  $\lambda_{oh} = 0.05$ ,  $\lambda_{ie} = 20$ ,  $\lambda_{BN} = 1e - 5$ ,  $\lambda_a = 0.1$ .

**Stage-II.** For CIFAR10, we set  $\lambda_d = 0.003$ ,  $\lambda_{rob} = 0.5$ ,  $\lambda_{SGA} = 5e - 4$ .

For CIFAR100, we set  $\lambda_d = 0.003$ ,  $\lambda_{rob} = 0.5$ ,  $\lambda_{SGA} = 5e - 4$ .

For ImageNet100, we set  $\lambda_d = 0.003$ ,  $\lambda_{rob} = 1$ ,  $\lambda_{SGA} = 5e - 4$ .

## 2 RESULTS ON IMAGENET100

Due to the lack of comparable related works, we provide results for the baseline model, adversarial training, and DERD as shown in Table 1. DERD still maintains decent adversarial robustness on datasets with more refined data such as ImageNet100, but it is inferior to adversarial training based on real data. This may be because the images in ImageNet100 are generally 224x224 in size, hence they possess richer textures and details, which poses greater challenges for the generator during pattern recovery. Besides, the data manifold of higher resolution datasets is undoubtedly more sparse, making it harder for the student to mimic the teacher. Despite this, DERD can still serve as a baseline model for Data-free adversarial defense on ImageNet, and brings significant increments to the undefended baseline model.

Table 1: Apply DERD to ImageNet.

|             | Clean | FGSM  | PGD <sub>S</sub> | PGD <sub>T</sub> | CW    | AA    | Ave   |
|-------------|-------|-------|------------------|------------------|-------|-------|-------|
| Nature      | 89.49 | 52.10 | 6.15             | 7.32             | 2.54  | 0.00  | 26.26 |
| SAT         | 78.44 | 66.79 | 50.01            | 53.68            | 66.17 | 55.62 | 61.78 |
| DERD (ours) | 61.56 | 59.44 | 40.77            | 42.71            | 44.79 | 43.34 | 48.76 |

## 3 DISCUSSION

### 3.1 Applied to Model Inversion

The methods based on generators [1, 2] and Model-Inversion [4] are the two main paradigms of data-free distillation in the community. We are more inclined towards generator-based methods since

robustness training often requires more data compared to vanilla training, and the model-inversion-based methods require real-time online optimization of inputs during each training session, which will undoubtedly demand more computational resources. Generator-based methods, conveniently, only require saving the well-trained generator, which can then be reused in subsequent tasks. Additionally, the generator-based framework is more compatible with the 2-stage training paradigm since the generator trained in stage-I can be easily saved, while the pseudo data from model-inversion-based methods need to be optimized online during training.

Still, we believe that DERD is a generalizable method and can also be extended to a model-inversion-based data-free knowledge distillation framework to obtain adversarial robustness. We verify it on CIFAR10 with ResNet18 being the backbone. Here, DERD degenerates into a single-stage model since there is no need to warm up the generator with natural knowledge. The pseudo-data for data-free robustness based on model-inversion is achieved by directly optimizing the input noise  $z$ , and there is no longer an explicit generator  $G$ . Thus, the loss function for the implicit generator is modified to:

$$\arg \min_{x_{inv}} L_{oh} + L_a + L_{BN} + L_{ie} \quad (1)$$

where  $x_{inv}$  is the pseudo data optimized from a random noise  $z$ . Then, the loss for the distillation process can be formalized as:

$$\arg \min_{\theta_S} D(S(x_{inv}), T(x_{inv})) \quad (2)$$

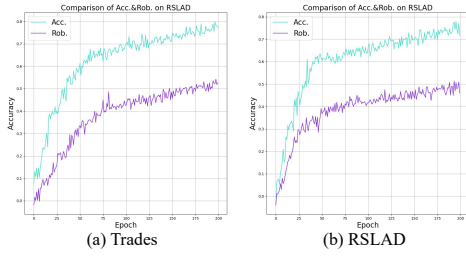
The experimental results are presented in Table 2. Although not as effective as the 2-stage generator-based DERD, model-inversion-based DERD still provides great adversarial robustness and surpasses some of the methods reported within the community. Nevertheless, we still recommend adopting the generator-based DERD, as it shows better robustness.

Table 2: Apply DERD to model-inversion based framework.

|                     | Clean | FGSM  | PGD <sub>S</sub> | PGD <sub>T</sub> | CW    | AA    | Ave   |
|---------------------|-------|-------|------------------|------------------|-------|-------|-------|
| DAFL                | 54.98 | 27.04 | 24.75            | 25.87            | 22.90 | 22.25 | 29.63 |
| DFAD                | 57.58 | 31.54 | 29.68            | 30.65            | 26.94 | 26.47 | 33.81 |
| ZSKT                | 58.08 | 31.98 | 29.94            | 30.92            | 27.21 | 26.68 | 34.13 |
| CMI                 | 53.28 | 25.78 | 23.14            | 23.97            | 21.03 | 20.38 | 27.92 |
| DFARD               | 66.44 | 38.53 | 35.94            | 37.15            | 32.79 | 32.14 | 40.49 |
| DERD <sub>inv</sub> | 62.44 | 50.42 | 42.51            | 42.89            | 38.72 | 28.94 | 44.32 |
| DERD (ours)         | 72.83 | 67.39 | 53.64            | 54.01            | 53.71 | 36.03 | 56.29 |

### 3.2 No available nature teachers?

The proposed DERD is based on a strong assumption that edge users have access to both robust and natural models. However, in traditional robustness distillation frameworks, the edge users may only have access to a standalone robust model. Despite the existence



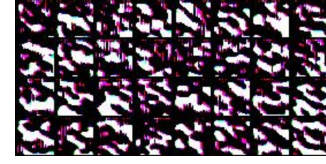
**Figure 1:** In scenarios where data is available, whether it be adversarial training (Trades in (a)) or robustness distillation (RSLAD in (b)), student models tend to first learn the relatively simple natural knowledge before progressing to the more challenging robust knowledge.

**Table 3:** The robustness of DERD without the nature teacher.

|                    | Clean        | FGSM         | PGD <sub>S</sub> | PGD <sub>T</sub> | CW           | AA           | Ave          |
|--------------------|--------------|--------------|------------------|------------------|--------------|--------------|--------------|
| DAFL               | 54.98        | 27.04        | 24.75            | 25.87            | 22.90        | 22.25        | 29.63        |
| DFAD               | 57.58        | 31.54        | 29.68            | 30.65            | 26.94        | 26.47        | 33.81        |
| ZSKT               | 58.08        | 31.98        | 29.94            | 30.92            | 27.21        | 26.68        | 34.13        |
| CMI                | 53.28        | 25.78        | 23.14            | 23.97            | 21.03        | 20.38        | 27.92        |
| DFARD              | 66.44        | 38.53        | 35.94            | 37.15            | 32.79        | 32.14        | 40.49        |
| w/o $T_{nat}$      | 52.51        | 49.34        | 37.66            | 39.67            | 38.54        | 23.31        | 40.17        |
| <b>DERD (ours)</b> | <b>72.83</b> | <b>67.39</b> | <b>53.64</b>     | <b>54.01</b>     | <b>53.71</b> | <b>36.03</b> | <b>56.29</b> |

of a similar dual-teacher hypothesis in MTARD [5] in the literature, we still briefly discuss alternative scenarios when a natural teacher is not available, in order to demonstrate the generalizability of our DERD approach. Our substitute scheme is based on the observation that, in adversarial training (like TRADES), student models tend to learn natural knowledge first, before acquiring robust knowledge (as illustrated in Fig. 1). Similarly, in adversarial distillation (like RSLAD), the student models tend to first learn the simpler natural knowledge before the harder robust knowledge. Therefore, in the early epochs of data-free distillation, the student already possesses a certain degree of natural knowledge, but the robustness knowledge is still sparse.

Based on this observation, we use the student model as a substitute model for the natural teacher, forming a self-adversary mechanism with the robust teacher, to excavate pseudo data for adversarial examples. Early in the training process, this is tantamount to letting the student model act as a stand-in for the natural teacher. In the mid to late epochs of the training, the student model forms a self-adversarial mechanism with the robust teacher to mine hard pseudo data. A similar mechanism is proposed in DFAD [2] for the excavation of more efficient hard natural knowledge that can not be easily transferred from the teacher to the student. The experimental results are shown in Table 3. Although there is some degradation, such a compromise still provides good adversarial robustness and even outperforms most data-free robustness distillation methods reported in DFARD [3].



**Figure 2:** Visualization of pseudo natural samples for CIFAR-10. The pseudo data generated by  $G(\cdot)$  do not conform to human semantic cognition.

### 3.3 Visual interpretability

Another issue worth mentioning is that although the pseudo data can be used for distillation learning, they do not conform to human semantic cognition. We visualize the natural pseudo data of stage-I as shown in Fig. 2. Although there are some specific patterns (shapes and edges), these forms do not align with human semantic recognition of the classes contained within the CIFAR-10 dataset, such as cat, dog, or airplane, since there is neither the shape of an airplane nor the outline of a cat or dog.

This may be due to:

(1) The sample distribution of datasets like CIFAR is still sparse, meaning there's an abundance of data manifolds between the real samples. Such visually unfamiliar but discriminable data may still exist in the gaps of the model's high-dimensional manifold.

(2) Despite being visually strange, the pseudo data's statistical characteristics might still be consistent with the real data, particularly the statistical features at the feature layer.

(3) The ultimate purpose of data-free knowledge distillation is to transfer the knowledge from the teacher to the student, so the visual randomness may be a result of the generator's preference where it always tends to mining challenging hard samples rather than the data that aligns with human cognition. Despite being visually less interpretable, such pseudo data may offer better discriminative knowledge, so the visual interpretability may not be necessary. A similar phenomenon can also be found in several generator-based data-free distillation methods [1, 2].

Constraining the visual credibility of pseudo images might be a potential direction for improving DERD, but we still argue that mining more effective knowledge has a higher priority than visual rationality.

## REFERENCES

- [1] Hanting Chen, Yunhe Wang, Chang Xu, Zhaohui Yang, Chuanjian Liu, Boxin Shi, Chunqing Xu, Chao Xu, and Qi Tian. 2019. Data-free learning of student networks. In *Proceedings of the IEEE/CVF international conference on computer vision*. 3514–3522.
- [2] Gongfan Fang, Jie Song, Chengchao Shen, Xinchao Wang, Da Chen, and Mingli Song. 2019. Data-free adversarial distillation. *arXiv preprint arXiv:1912.11006* (2019).
- [3] Yuzhang Wang, Zhaoyu Chen, Dingkan Yang, Pinxue Guo, Kaixun Jiang, Wenqiang Zhang, and Lizhe Qi. 2023. Out of Thin Air: Exploring Data-Free Adversarial Robustness Distillation. *arXiv preprint arXiv:2303.11611* (2023).
- [4] Hongxu Yin, Pavlo Molchanov, Jose M Alvarez, Zhizhong Li, Arun Mallya, Derek Hoiem, Niraj K Jha, and Jan Kautz. 2020. Dreaming to distill: Data-free knowledge transfer via deepinversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8715–8724.
- [5] Shiji Zhao, Jie Yu, Zhenlong Sun, Bo Zhang, and Xingxing Wei. 2022. Enhanced accuracy and robustness via multi-teacher adversarial distillation. In *European Conference on Computer Vision*. Springer, 585–602.