# Towards Adversarially Robust Data-Efficient Learning with Generated Data

Junhao Dong[1], Melvin Wong[1], Sihan Xia[1], Joel Wei En Tay[2]

[1]*College of Computing and Data Science (CCDS), Nanyang Technological University (NTU), Singapore*
[2]*Singapore Institute of Manufacturing Technology (SIMTech), A*STAR, Singapore*
{junhao003, wong1357, sihan002}@ntu.edu.sg, joel_tay@simtech.a-star.edu.sg

***Abstract*—A well-established study of adversarial training has been proposed to improve network robustness against adversarial examples in the context of deep learning. However, its performance highly relies on large-scale training data. To relieve from such a data-hungry learning nature, we propose an efficient extension of adversarial training by conducting a data reduction method from a new perspective of generated data. The reduced dataset can be regarded as an alternative pre-training dataset, which promotes adversarial training methods for better robustness even than the original dataset. Experimental results and analyses demonstrate the effectiveness of our data reduction method, achieving the same level of adversarial robustness with a dataset volume reduced to 80% of its original size.**

***Index Terms*—adversarial training, data reduction, generated data, adversarial robustness**

## I. Introduction

Adversarial examples [1] can cause significant disruptions of Deep Neural Networks (DNNs) [2]–[4] while maintaining visual similarity to their natural counterparts. Previous works have demonstrated the disruptive impact of adversarial examples across various domains, *e.g.*, medical image analysis [5], image manipulation [6], [7], and speech recognition [8]. Considering potential security threats induced by such tailored examples, a series of defense methods have been proposed to enhance the network robustness. Among them, adversarial training [9]–[14], which adaptively augments adversarial examples into the training dataset, has been demonstrated to be the most effective method to heal network susceptibility against unforeseen adversaries. However, the performance of such a defense paradigm highly relies on a large-scale dataset with considerable computational resources, hindering its efficacy on small training data.

To relieve from the data-hungry nature of DNNs, existing works either rely on (1) *synthesizing* a compact dataset to maintain the same level of information *w.r.t.* its original counterpart (dataset distillation) [15] or (2) *selecting* a subset of the original dataset by pruning away redundant data that compromises the performance (data pruning) [16]. However, these methods primarily cater to maintaining natural performance, leaving a noticeable void in the context of adversarial robustness. Among the scant literature, [17] exhibits instability, occasionally yielding results inferior to random subsampling. Another approach [18] separately samples important training data from the entire training dataset for each epoch, imposing unaddressed storage constraints.

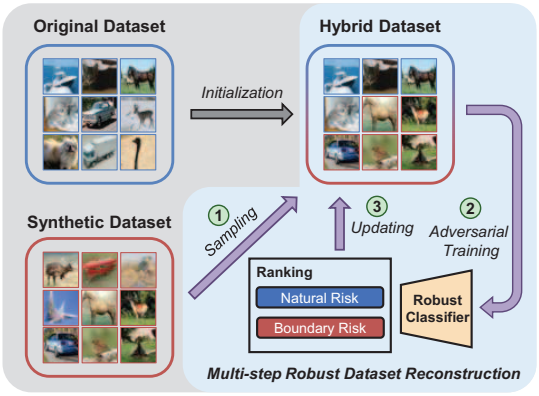


Fig. 1. Overview of the proposed multi-step dataset reconstruction for adversarial robustness. The reconstructed hybrid dataset is initialized by the original dataset and subsequently updated according to synthetic data with higher robust risk (composed of natural and boundary risks).

Given the significant robustness improvement brought by auxiliary generated data [19], we provide a new perspective of data-efficient learning for adversarial robustness by reconstructing the legitimate dataset with both original and generated data. Specifically, we decoupled the learning difficulty of examples in terms of natural performance and adversarial robustness, which can be interpreted as the importance of training data. Considering the underlying redundancy inside the original dataset, we propose a hybrid data reconstruction paradigm guided by the auxiliary generated data. Furthermore, the reconstructed dataset can be pruned to a representative subset for better training efficiency. Empirical results show that robust learning on our reconstructed dataset outperforms that on the original dataset in terms of both natural performance and robustness. In the meantime, data pruning based on our reconstructed dataset also achieves better performance than random selection without an additional storage budget.

## II. Adversarially Robust Data-Efficient Learning

In this section, we introduce our adversarially robust data-efficient learning method based on a reconstructed dataset guided by auxiliary data generated by the Denoising Diffusion Probabilistic Model (DDPM) [20], as shown in Fig. 1.

979-8-3503-5409-6/24/$31.00 ©2024 IEEE
DOI 10.1109/CAI59869.2024.00253

**Preliminary:** Consider a DNN classifier $f_{\boldsymbol{\theta}} : \mathcal{X} \to \mathbb{R}^C$ with network parameters $\boldsymbol{\theta}$ to output predictions of $C$ classes. For a specific dataset $(\mathbf{x}, y) \sim \mathcal{D}$, adversarial training [9] can be described as the following minimax optimization problem:

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}} \left[ \max_{\|\boldsymbol{\delta}\|_\infty < \epsilon} \mathcal{L}_{\text{CE}} \left( f_{\boldsymbol{\theta}} \left( \mathbf{x} + \boldsymbol{\delta} \right), y \right) \right], \tag{1}$$

where $\boldsymbol{\delta}$ is the adversarial perturbation *w.r.t.* the clean example $\mathbf{x}$ under the $\ell_\infty$-norm radius $\epsilon$. Adversarial examples $\hat{\mathbf{x}} = \mathbf{x} + \boldsymbol{\delta}$ are obtained by maximizing the cross-entropy loss ($\mathcal{L}_{\text{CE}}$). The outer minimization is to optimize empirical risks of obtained adversaries over network parameters $\boldsymbol{\theta}$ for better robustness.

In this paper, we primarily focus on reconstructing the original training set to eliminate redundant data without resorting to external knowledge. Despite the same data volume of original and reconstructed datasets, training on the latter can achieve better performance in terms of natural accuracy and adversarial robustness. Specifically, we incorporate a synthetic dataset [19] generated by a DDPM model trained on the original dataset. The hybrid dataset is initialized by the original dataset at the beginning of optimization. We then conduct a multi-step dataset reconstruction based on the decoupled robust risk.

During each iteration, we first randomly sample a subset of generated data and merge it into the hybrid dataset. To investigate the impact of each data point on network robustness, we then adversarially trained a classifier based on the hybrid dataset. Afterward, we rank each data point based on a convex surrogate of the decoupled robust (classification) risk [10]. The hybrid dataset can thus be updated by top-K (size of the original dataset) scoring images. Generally, robust risk in the context of adversarially robust learning can be decoupled into natural risk and boundary risk:

$$\begin{aligned} \mathcal{R}_{nat}(f_{\boldsymbol{\theta}}; \mathcal{D}) &= \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathbb{1}(f_{\boldsymbol{\theta}}(\mathbf{x}) \neq y)], \\ \mathcal{R}_{bdy}(f_{\boldsymbol{\theta}}; \mathcal{D}) &= \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathbb{1}(\exists \hat{\mathbf{x}} \in \mathbb{B}(\mathbf{x}, \epsilon) : f_{\boldsymbol{\theta}}(\hat{\mathbf{x}}) \neq f_{\boldsymbol{\theta}}(\mathbf{x}) = y)], \end{aligned} \tag{2}$$

where $\mathbb{B}(\mathbf{x}, \epsilon)$ denotes the $\ell_\infty$-norm hypersphere with radius $\epsilon$ around $\mathbf{x}$. To measure the robust risk *w.r.t.* each example, we propose a convex surrogate of both natural and boundary risks based on the prediction-level distance as below:

$$\begin{aligned} \mathcal{W}_{nat}(f_{\boldsymbol{\theta}}; \mathbf{x}) &= \sigma(\|f_{\boldsymbol{\theta}}(\mathbf{x}) - \text{onehot}(y)\|_2), \\ \mathcal{W}_{bdy}(f_{\boldsymbol{\theta}}; \mathbf{x}) &= \sigma(\|f_{\boldsymbol{\theta}}(\mathbf{x}) - f_{\boldsymbol{\theta}}(\hat{\mathbf{x}})\|_2), \end{aligned} \tag{3}$$

where $\sigma(\cdot)$ represents the sigmoid function, and $\text{onehot}(y)$ denotes the one-hot encoding of the label $y$. These prediction discrepancies can also be interpreted as the fitting degree of each example. In other words, we prioritize harder training examples in the reconstructed dataset to enable hard example mining for better generalizability. Through iterative updating, more informative synthetic data are incorporated into the hybrid dataset along with the elimination of redundant examples. To achieve further data efficiency, we also adopt the data pruning strategy based on ranked weights of the reconstructed dataset with the priority of examples with high robust risk.

## III. Experiments

Following the setting from RobustBench [21], we conduct all the adversarial training experiments based on ResNet-18 on the CIFAR-10 dataset. For evaluation, we report classification accuracy on adversaries generated by Projected Gradient Descent(PGD) with 20 steps, CW, and Auto-Attack (AA).

TABLE I

Comparisons between **our reconstructed hybrid dataset** and its original version via diverse adversarial training methods.

| Method | **Training Data** | Natural | PGD | CW | AA |
|---|---|---|---|---|---|
| PGD-AT | Original | 83.80 | 51.40 | 50.17 | 47.68 |
| | **Hybrid** | 83.76 | 52.44 | 51.69 | 48.95 |
| TRADES | Original | 82.45 | 52.21 | 50.29 | 48.88 |
| | **Hybrid** | 82.53 | 53.21 | 50.91 | 49.73 |
| N-FGSM | Original | 80.18 | 48.17 | 46.96 | 44.26 |
| | **Hybrid** | 80.23 | 49.72 | 48.31 | 45.41 |

TABLE II

Comparisons between **our heuristic pruning** and random pruning on original and **hybrid data** via N-FGSM.

| **Training Data** | **Pruning Strategy** | 80% | | 40% | |
|---|---|---|---|---|---|
| | | Natural | PGD | Natural | PGD |
| Original | Random | 78.52 | 46.93 | 72.76 | 42.39 |
| | **Heuristic** | 78.07 | 47.75 | 72.16 | 43.25 |
| **Hybrid** | Random | 78.46 | 48.10 | 72.78 | 43.30 |
| | **Heuristic** | 78.97 | 48.82 | 72.95 | 44.08 |

As shown in Tab. I, we compare our reconstructed hybrid dataset with its original counterpart based on their corresponding performance achieved by diverse adversarial training methods. We can easily observe that training on our reconstructed dataset outperforms that on the original dataset in terms of natural performance and adversarial robustness. Recall that the reconstruction of the hybrid data merely relies on the original training set without using external knowledge. In the meantime, the reconstructed and the original datasets are of the same size, which guarantees a fair comparison.

For further data efficiency, we exploit the data pruning strategy guided by both natural and robust discrepancies in Eq. (3) to obtain lightweight subsets of our hybrid dataset. We conduct a comparison between our heuristic pruning and random pruning in Tab. II. Our heuristic pruning strategy is simultaneously effective in both the original and hybrid datasets to eliminate redundant data points for better robustness.

## IV. Conclusion and Future Work

In this paper, we propose a novel data-efficient learning paradigm for adversarial robustness by reconstructing the original dataset with generated data. We also design a decoupled surrogate of the robust risk to guide data selection and pruning. A limitation of our method is the static nature of DDPM data; future work will explore robustness-guided, adaptive data generation, and multi-objective prompt evolution [22] for enhanced performance.


## Acknowledgment

This research is partly supported by the Distributed Smart Value Chain programme which is funded in part by the Singapore RIE2025 Manufacturing, Trade and Connectivity (MTC) Industry Alignment Fund-Pre-Positioning (Award No:

M23L4a0001), and the College of Computing and Data Science, Nanyang Technological University.

## REFERENCES


[1] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *International Conference on Learning Representations, ICLR*, 2014.

[2] S. Pouyanfar, S. Sadiq, Y. Yan, H. Tian, Y. Tao, M. P. Reyes, M.-L. Shyu, S.-C. Chen, and S. S. Iyengar, "A survey on deep learning: Algorithms, techniques, and applications," *ACM Computing Surveys (CSUR)*, vol. 51, no. 5, pp. 1–36, 2018.

[3] Q. Li and C. Zhang, "Continual learning on deployment pipelines for machine learning systems," *arXiv preprint arXiv:2212.02659*, 2022.

[4] G. Menghani, "Efficient deep learning: A survey on making deep learning models smaller, faster, and better," *ACM Computing Surveys*, vol. 55, no. 12, pp. 1–37, 2023.

[5] J. Dong, J. Chen, X. Xie, J. Lai, and H. Chen, "Adversarial attack and defense for medical image analysis: Methods and applications," *arXiv preprint arXiv:2303.14133*, 2023.

[6] J. Dong and X. Xie, "Visually maintained image disturbance against deepfake face swapping," in *2021 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2021, pp. 1–6.

[7] J. Dong, Y. Wang, J. Lai, and X. Xie, "Restricted black-box adversarial attack against deepfake face swapping," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 2596–2608, 2023.

[8] M. M. Cisse, Y. Adi, N. Neverova, and J. Keshet, "Houdini: Fooling deep structured visual and speech recognition models with adversarial examples," *Advances in neural information processing systems*, vol. 30, 2017.

[9] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *ICLR*, 2018.

[10] H. Zhang, Y. Yu, J. Jiao, E. Xing, L. El Ghaoui, and M. Jordan, "Theoretically principled trade-off between robustness and accuracy," in *International conference on machine learning*, 2019, pp. 7472–7482.

[11] J. Dong, Y. Wang, J.-H. Lai, and X. Xie, "Improving adversarially robust few-shot image classification with generalizable representations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 9025–9034.

[12] J. Dong, S.-M. Moosavi-Dezfooli, J. Lai, and X. Xie, "The enemy of my enemy is my friend: Exploring inverse adversaries for improving adversarial training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 24 678–24 687.

[13] J. Dong, L. Yang, Y. Wang, X. Xie, and J. Lai, "Toward intrinsic adversarial robustness through probabilistic training," *IEEE Transactions on Image Processing*, vol. 32, pp. 3862–3872, 2023.

[14] J. Dong, Y. Wang, X. Xie, J. Lai, and Y.-S. Ong, "Generalizable and discriminative representations for adversarially robust few-shot learning," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–14, 2024.

[15] S. Lei and D. Tao, "A comprehensive survey of dataset distillation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[16] M. Paul, S. Ganguli, and G. K. Dziugaite, "Deep learning on a data diet: Finding important examples early in training," *NeurIPS*, 2021.

[17] M. Kaufmann, Y. Zhao, I. Shumailov, R. Mullins, and N. Papernot, "Efficient adversarial training with data pruning," *arXiv preprint arXiv:2207.00694*, 2022.

[18] Y. Li, P. Zhao, X. Lin, B. Kailkhura, and R. A. Goldhahn, "Less is more: Data pruning for faster adversarial training," in *Proceedings of the Workshop on Artificial Intelligence Safety*, vol. 3381, 2023.

[19] S.-A. Rebuffi, S. Gowal, D. A. Calian, F. Stimberg, O. Wiles, and T. A. Mann, "Data augmentation can improve robustness," *NeurIPS*, 2021.

[20] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, 2020.

[21] F. Croce, M. Andriushchenko, V. Sehwag, E. Debenedetti, N. Flammarion, M. Chiang, P. Mittal, and M. Hein, "Robustbench: a standardized adversarial robustness benchmark," in *NeurIPS*, 2021.

[22] M. Wong, Y. Ong, A. Gupta, K. Bali, and C. Chen, "Prompt evolution for generative ai: A classifier-guided approach," in *2023 IEEE Conference on Artificial Intelligence (CAI)*. Los Alamitos, CA, USA: IEEE Computer Society, jun 2023, pp. 226–229. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/CAI54212.2023.00105