# Enhancing Distilled Datasets via Natural Data Mixing

Ian Pons[†], Guilherme B. Stern, Anna H. Reali Costa, and Artur Jordao
*Escola Politécnica, Universidade de São Paulo*
São Paulo, Brazil
[†]Corresponding author: ian.pons@usp.br

***Abstract*—Dataset distillation emerges as a promising technique to reduce web-scale datasets into a compact version with only a few samples per class. It involves distilling a large dataset into a compact synthetic set that aims to preserve representative information from the original data, offering advantages such as higher training efficiency and data privacy. However, existing techniques fail to fully capture the underlying properties of original (natural) training samples. Hence, learning solely on distilled images—the standard practice—leads models to encounter a notable generalization gap. In this work, we propose a simple yet effective mechanism to enhance distilled images. Our method transfers powerful and discriminative characteristics from natural images to distilled samples through a simple mixing process. Extensive experiments on benchmarks confirm that our method consistently improves generalization accuracy. Notably, we demonstrate that our approach enables distilled sets with only 10 images per class to match or exceed the performance of state-of-the-art methods trained on 50 images per class, representing a 5× gain in training efficiency. On challenging ImageNet subsets, it increases predictive performance by up to 11.5 percentage points. We also confirm that our method more effectively preserves internal representations concerning full dataset training when compared to plain state-of-the-art dataset distillation methods. Crucially, our method achieves these improvements without increasing the size of the distilled set, thus preserving the efficiency and privacy advantages inherent to dataset distillation. Moreover, our method enhances robustness to common corruptions, improving predictive performance by an average of 9.05 percentage points. It also improves accuracy against moderate adversarial attacks. Code is available at: *github.com/IanPons/Enhancing-Distilled-Datasets***

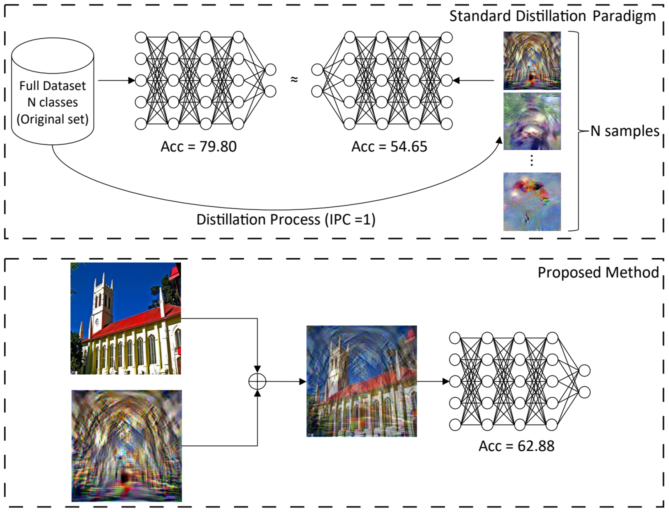

Fig. 1. Standard dataset distillation paradigm versus our proposed data-mixing enhancement. Top. Common strategies distill a large dataset into a small synthetic set (here, with an Images Per Class, IPC, of 1). Training a model on this distilled set often results in a significant performance drop compared to training on the full dataset (e.g., 54.65% vs. 79.80%). Bottom. Enhancing distilled data with the proposed method. For each synthetic sample, we randomly retrieve a natural image from the original dataset that shares the same label. Then, our method generates a new training sample by mixing the synthetic and natural images. This dynamically enhanced sample improves model generalization (e.g., 62.88%), effectively narrowing the performance gap while maintaining the efficiency of a small dataset.

## I. Introduction

Driven by large-scale datasets, deep models are achieving increasingly better results in various pattern recognition fields, such as image classification, speech interpretation and natural language processing [1], [2]. However, the same ingredient that provides such results is responsible for the computational burden and prohibitive financial demands associated with the design and maintenance of high-performance models, along with its environmental impacts [3], [4]. Moreover, modern learning paradigms standardize on leveraging overparameterized models—number of parameters exceeds the number of data points—since previous works demonstrate its positive effects on generalization [5]. In this direction, complex cognitive tasks require larger datasets, pushing further the computational requirements for developing state-of-the-art models.

Given the aforementioned scenario, extensive research focuses on designing computationally efficient architectures to reduce the computational overhead of deep models when faced with large amounts of data [6], [7]. Despite providing positive results, these approaches offer little support for managing the volume of training data required to extract meaningful representations from deep models, motivating different strategies to reduce the inherent difficulties of large-scale datasets. In this line, coreset methods aim to reduce the dataset by training the model on a subset of the original data (i.e., the coreset) that comprises significant information, selected according to a specified criterion [8], [9]. Similarly, dataset distillation strategies address the problem by generating a small, synthetic set of highly informative samples [10], [11]. Given the images per class (IPC) parameter[1], distilled images aim to approximate the underlying properties of the original

[1] This parameter defines the number of synthetic images generated for each class in the dataset.

dataset through iterative learning. After this process, model training occurs on the distilled images using standard recipes such as data augmentation and regularization [10], [11] (see Figure 1 – top). It is worth mentioning that the time and computational overhead during the learning phase become notably lower compared to the original dataset, as it involves fewer samples (relative to IPC parameter). Apart from technical and theoretical details, such strategies gained significant attention from researchers due to their applications in fields like medicine [12] and private data protection [13]–[17].

While effective at reducing dataset size, coreset methods depend heavily on the original set, potentially selecting subsets with weaker expressive power and leading to significant performance drops as compression (subset size relative to the original dataset) increases. On the other hand, state-of-the-art distillation methods guide the synthesis process by either approximating a specific informative property from the training process or considering the distilled set as a hyper-parameter subject to an optimization strategy, removing the restriction faced by coreset methods [18]–[21]. Unfortunately, the close relationship between training properties and the network architecture limits the transferability of most distilled sets across slightly different models, requiring a new synthesis when avoiding drastic performance drops [10]. Additionally, it is well-known that even modern strategies incur a loss of information for high compression rates. It turns out that fully characterizing a large dataset is a challenge for small sets of samples, particularly within deep learning contexts. Such factors motivate new approaches to address the generalization drop from distilled training.

Given the open challenges in dataset distillation, we propose a data-mixing strategy specifically designed to enhance distilled datasets. Our main argument is that natural images offer powerful information, enhancing discrimination in distilled images. We introduce such information by leveraging a class of augmentation techniques that combine samples through a mixing process [22], [23].

Previous efforts focus on increasing generalization of distilled images using standard data augmentation in the synthesis process. However, they remain ineffective because standard augmentation can alter the synthetic distribution, drifting from the learned properties during synthesis [20], [24], [25]. We show that our approach combines the best of natural and distilled worlds: it overcomes performance degradation, brings the synthetic distribution closer to the original dataset, and maintains the compact size of distilled sets. To sum up, our contributions are: (1) we introduce a simple yet effective process that combines synthetic images with natural samples, significantly improving model generalization; (2) we confirm its effectiveness across various state-of-the-art distillation techniques and benchmarks; and (3) we show that our approach narrows the representational gap to models trained on original data. Finally, our method also increases robustness to adversarial attacks and out-of-distribution samples, supporting its use in security-critical applications.

Extensive experiments show that our method yields substantial and consistent accuracy improvements. For example, when applied to state-of-the-art distillation techniques on CIFAR-10 and CIFAR-100, our approach improves accuracy across popular compression rates (IPC $\in \{1, 10, 50\}$). The gains are also significant on other benchmarks, such as SVHN, where accuracy increases from 76.8% to 85.10% at 50 IPC for the recent method by Vahidian et al. [26]. Finally, on challenging ImageNet subsets, we observe an improvement for the ATT [27] method from 52.20% up to 63.70% at 10 IPC. Overall, these results highlight the effectiveness of our method in enhancing model generalization and improving robustness to common corruptions by an average of 9.05 percentage points while preserving the efficiency of distilled datasets.

## II. Related Works

**Dataset Distillation.** The key idea behind dataset distillation is to generate a compact yet informative synthetic dataset that enables a model to reach learning abilities comparable to training on the full dataset. In this line of research, existing strategies differ in how they construct synthetic data to approximate the full dataset (target). Below, we describe the main families of dataset distillation methods.

Meta-learning approaches leverage original samples for validation, employing a bi-level optimization paradigm [28]. Specifically, the outer level updates the synthetic data by minimizing empirical risk on target samples while the inner level follows a standard learning paradigm on the distilled set (i.e., loss over the synthetic data). In contrast, data-matching strategies aim to replicate the effect of target data on a specific informative property of the training process such as parameters and gradient distribution [20], [21].

Unlike the previously mentioned approach, data-matching methods minimize the distance between a specific property through training iterations on both the distilled and target dataset [10]. For example, trajectory-matching approaches seek to replicate the optimization path in parameter space from a model trained on the original dataset, synthesizing a small set that induces similar parameter dynamics [20], [27]. In contrast, Zhao et al. [21] operate in feature space, avoiding explicit dependence on model parameters.

Regardless of the synthesis framework, existing data distillation methods suffer from similar problems: a significant accuracy gap compared to the non-synthetic dataset and poor transferability across different architectures [10], [11]. We argue that the discriminative information from natural samples is essential for generalization tasks independently from the distillation framework. It turns out that natural images provide a simple, yet effective, source of information to combine with existing methods while preserving the essence of distillation methods (i.e., compact and informative datasets). Importantly, our method is compatible with all existing families of dataset distillation mentioned above.

**Data Augmentation.** Past works studied how human-imperceptive transformations on data samples affect the predictive performance of deep models, paving the way for data augmentation techniques [29], [30]. Since then, state-of-the-art augmentation strategies evolved from noise injection [31], passing through random cropping [32], Cutout [33], mixing samples [22], and other variations [34], [35]. Among existing augmentation methods, Hendrycks et al. [23] empirically demonstrated that PixMix achieves the best results across metrics such as generalization, out-of-distribution (OOD), and adversarial robustness. Their work extends augmentation beyond the combination of samples within the dataset—already employed by methods such as mixup [22]—leveraging the inherent structural complexity in fractal images. Specifically, each sample from the original dataset is subject to a series of combinations with random images from a mixing set of structural-complex samples such as fractals. The authors argue that although this type of image does not belong to a particular class, it can enhance multiple safety and performance metrics in a Pareto-optimal fashion.

Within the sphere of data augmentation, Zhao et al. [36] highlighted the challenges of augmenting distilled datasets since the synthesis process does not consider posterior transformations. Overall, the authors confirm that naively applying augmentation to synthetic sets leads to negligible generalization gains or even training collapse. To address these limitations, their work proposes Differentiable Siamese Augmentation (DSA), applying transformation to both natural and synthetic sets and, most importantly, incorporating augmentation into the synthesis process.

Regarding combination-based paradigms, recent approaches that merge synthetic and real data achieve positive results for high-IPC settings. For example, Chen et al. [37] propose the CCFS method, which builds a static dataset by selecting real images to include through a curriculum, a process designed to address performance decay in high IPC scenarios. Their method progressively expands the synthetic set by incorporating these selected real samples, thus increasing its final size. On the other hand, our method dynamically generates new training samples on-the-fly at each epoch by mixing synthetic images with natural samples. Crucially, it preserves the original distilled set size and demonstrates strong performance across all IPC ranges. Additionally, our approach is orthogonal to all distillation methods since it corresponds to a posterior transformation and does not affect the synthesis process, operating as a dynamic augmentation strategy, rather than a static dataset creation paradigm.

## III. Definitions and Proposed Method

**Definitions.** The formulation of dataset distillation aims to synthesize a small set $\mathcal{S}$ from a large dataset $\mathcal{D}$, enabling a model to achieve performance comparable to training on the full dataset. While meta-learning, frames the problem using a bi-level optimization paradigm, data-matching strategies aim to directly minimize the discrepancy between key properties of the training process on the synthetic and real datasets. Overall, the previous process admits a general expression in terms of

$$\mathcal{S}^* = \arg\min_{\mathcal{S}} \mathbb{E}_\theta \left[ d(\Phi(f_\theta, \mathcal{S}), \Phi(f_\theta, \mathcal{D})) \right], \qquad (1)$$

where $d$ is a distance metric and $\Phi$ is a function that extracts a target property from the training process such as gradients [24], trajectories [20], [27], or distributions [21].

**Proposed Method.** Our method enhances the training process by dynamically injecting real data characteristics into the compact synthetic set $\mathcal{S}$ at each training epoch. For each data point $(s, y) \in \mathcal{S}$, consisting of a synthetic image $s$ and its label $y$, we first sample a random data point $(x, y) \in \mathcal{D}$ from the original large dataset, ensuring it shares the same label.

Then, our method generates a new training sample, $\bar{s}$, through a combination of the synthetic and the natural image:

$$\bar{s} = (1 - \alpha) \cdot s + \alpha \cdot x, \qquad (2)$$

where the hyperparameter $\alpha$ controls the amount of natural information infused into the synthetic sample. Build upon the work by Hendrycks et al. [23], we treat $\alpha$ as a random variable drawn from a Uniform Distribution $\mathcal{U}[0, 1]$ instead of a hyperparameter that requires an architecture-specific tuning. However, we repeat this process for all samples in $\mathcal{S}$ at every epoch, creating new augmented sets for training the final model. Finally, this choice ensures that our method is architecture- and dataset-agnostic.

## IV. Experiments

**Experimental Setup.** Throughout our experimental analysis, we evaluate the performance gains achieved by our method when applied to a diverse set of dataset distillation techniques across standard datasets such as CIFAR-10/100 [38], SVHN [39] and ImageNet subsets [40], [41]. To ensure a fair comparison, we first reproduce the results of all techniques we consider before introducing our method and adopt the standard architectures/experimental settings (learning rate, batch size, optimizer, etc.) employed by the authors. Specifically, we conduct experiments using the following state-of-the-art techniques: Dataset Condensation (DC) [24], Differentiable Siamese Augmentation (DSA) [36], and Distribution Matching (DM) [21]. To incorporate recent advances, we include the method GLAD$^{\mathcal{R}}$ by Vahidian et al [26], and to represent trajectory-matching approaches, we adopt the also recent method by Liu et al. (ATT) [27], as it outperforms previous related techniques [20]. Importantly, the methods we consider cover all families of dataset distillation techniques.

**Predictive Performance Gains.** This experiment aims to empirically validate the effectiveness of our data-mixing strategy. For this purpose, we apply our method to datasets produced by the aforementioned state-of-the-art distillation techniques and evaluate the final model accuracy on the original test set.

Tables I to III summarize the results. From these tables we note that our approach yields substantial and consistent generalization improvements across all considered distillation methods, datasets, and images per class (IPC) settings. For example, on the Imagenette subset, our method boosts the accuracy of the ATT [27] technique by up to 11.50 p.p. at 10 IPC. The improvements extend to other foundational methods; for example, on CIFAR-10 with 50 IPC, our mixing strategy improves the performance DSA [36] from 53.32% to 71.10%.

TABLE I
PREDICTIVE PERFORMANCE ON CIFAR-10. COMPARISON OF BASELINE DISTILLATION METHODS AGAINST OUR DATA-MIXING METHOD ACROSS DIFFERENT IMAGES PER CLASS (IPC) SETTINGS.

| **IPC** | **Method** | **Baseline (%)** | **Baseline + Ours (%)** | **Gain (p.p.)** |
|---|---|---|---|---|
| 1 | DC | 25.06 | 32.41 | +7.35 |
| | DM | 14.23 | 32.12 | +17.89 |
| | DSA | 13.55 | 31.98 | +18.43 |
| | GLaD$^{\mathcal{R}}$ | 28.60 | 42.80 | +14.20 |
| | ATT | 43.80 | 47.00 | +3.20 |
| 10 | DC | 42.61 | 54.75 | +12.14 |
| | DM | 50.56 | 62.77 | +12.21 |
| | DSA | 43.79 | 63.08 | +19.29 |
| | GLaD$^{\mathcal{R}}$ | 50.50 | 60.20 | +9.70 |
| | ATT | 65.80 | 68.30 | +2.50 |
| 50 | DC | 54.65 | 62.88 | +8.23 |
| | DM | 58.43 | 71.61 | +13.18 |
| | DSA | 53.32 | 71.10 | +17.78 |
| | GLaD$^{\mathcal{R}}$ | 61.00 | 72.20 | +11.20 |
| | ATT | 70.00 | 76.00 | +6.00 |

TABLE II
PREDICTIVE PERFORMANCE ON CIFAR-100. COMPARISON OF BASELINE DISTILLATION METHODS AGAINST OUR DATA-MIXING METHOD ACROSS DIFFERENT IMAGES PER CLASS (IPC) SETTINGS.

| **IPC** | **Method** | **Baseline (%)** | **Baseline + Ours (%)** | **Gain (p.p.)** |
|---|---|---|---|---|
| 1 | DC | 11.87 | 16.74 | +4.87 |
| | DM | 11.21 | 16.83 | +5.62 |
| | DSA | 11.77 | 17.09 | +5.32 |
| | ATT | 23.35 | 30.16 | +6.81 |
| 10 | DC | 23.00 | 28.66 | +5.66 |
| | DM | 25.37 | 33.86 | +8.49 |
| | DSA | 23.12 | 29.48 | +6.36 |
| | ATT | 36.31 | 47.68 | +11.37 |
| 50 | DC | 30.02 | 41.67 | +11.65 |
| | DM | 33.75 | 44.14 | +10.39 |
| | DSA | 31.43 | 42.33 | +10.90 |
| | ATT | 42.21 | 51.77 | +9.56 |

TABLE III
PREDICTIVE PERFORMANCE ON SVHN AND IMAGENET SUBSETS (GLAD$^{\mathcal{R}}$ USES IMAGENET-A AND ATT IMAGENETTE).

| **Dataset** | **IPC** | **Method** | **Baseline (%)** | **Baseline + Ours (%)** | **Gain (p.p.)** |
|---|---|---|---|---|---|
| SVHN | 1 | GLaD$^{\mathcal{R}}$ | 35.80 | 54.10 | +18.30 |
| | 10 | GLaD$^{\mathcal{R}}$ | 72.40 | 81.40 | +9.00 |
| | 50 | GLaD$^{\mathcal{R}}$ | 76.80 | 85.10 | +8.30 |
| ImageNet Subsets | 1 | GLaD$^{\mathcal{R}}$ | 34.20 | 57.00 | +22.80 |
| | | ATT | 44.20 | 45.80 | +1.60 |
| | 10 | ATT | 52.20 | 63.70 | +11.50 |

The gains are also significant on other benchmarks, such as SVHN, where our method again improves the accuracy of GLAD$^{\mathcal{R}}$ [26] from 76.8% to 85.10% at 50 IPC. For the state-of-the-art trajectory-matching method ATT [27], our approach provides significant gains on CIFAR-100, where accuracy increases from 36.31% to 47.68% (10 IPC). On CIFAR-10, it enhances the accuracy from 70.0% to 76.0% (50 IPC).

The previous results highlight that our simple mixing strategy serves as a robust and effective plug-in to enhance a wide variety of distilled datasets, preserving the efficiency of the distillation paradigm while significantly closing the generalization gap.

**The Role of Natural Images.** Given the effectiveness of our method in enhancing predictive performance on distilled datasets, this experiment investigates the underlying mechanism driving such performance gains. Specifically, we seek to answer the following question: *is the improvement attributable merely to the injection of complex, naturalistic textures, or is the semantic information within the natural images a critical component?* The question stems from the work by Hendrycks et al. [23], where the authors achieve performance improvements across multiple safety metrics via augmentation with structurally complex fractals. As a key characteristic, fractals introduce rich textural diversity without containing class-specific, discriminative information.

To isolate the source of our performance gains, we design a label-agnostic mix experiment. First, for a distilled CIFAR-10 dataset, we use images from the CIFAR-100 training set as the natural source, mixing them randomly without matching class labels, following a rationale similar to fractal mixing [23]. This label-agnostic, out-of-distribution mixing consistently degrades model accuracy, with predictive performance falling far below the distilled-only baseline. For instance, a 50 IPC distilled set with a 56.0% baseline accuracy drops to 35.2% after this mixing process. Second, we replace the natural images with label-less fractals from Hendrycks et al. [23]. This setup yields an accuracy of 53.5%, a slight degradation compared to the 56.0% distilled-only baseline.

Together, these findings strongly suggest that performance gains from our method are not merely due to an increase in data diversity. Instead, our results underscore that semantically relevant mixed natural images are a key driver for enhancing model generalization. This is evidenced by our experiments: the injection of random, out-of-distribution semantics actively harms performance. Additionally, the injection of complex but semantically neutral textures, such as fractals, acts as a regularization mechanism. Neither approach yields significant improvements as our label-aware method, confirming that the primary strength of our method lies in the semantic coherence between distilled and natural samples sharing the same label.

**Internal Feature Approximation.** Although distillation methods learn synthetic sets by minimizing the expected difference in performances, the internal representations (i.e., features maps) learned through distilled and original dataset diverge significantly. In this experiment, we empirically demonstrate that our method enhances the similarity of representations developed during training on both distilled and original data.

For this purpose, we employ the Centered Kernel Alignment (CKA) [42] similarity metric. The recent work by Klabunde

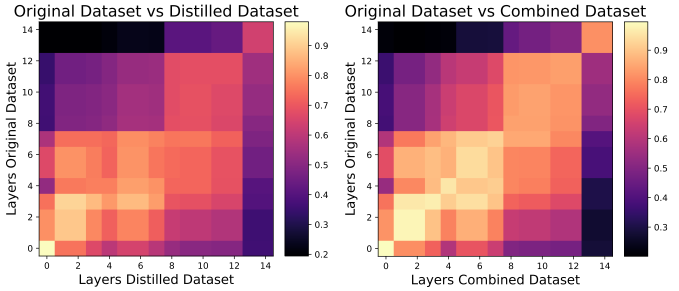

Fig. 2. **Left:** similarity matrix $M_{\mathcal{D}\mathcal{S}}$ between baseline model and a model trained on a distilled dataset ($\mathcal{S}$) created with the DSA [36] method for 50 images per class. **Right:** similarity matrix $M_{\mathcal{D}\overline{\mathcal{S}}}$ after enhancing the same distilled set with our method, resulting in $\overline{\mathcal{S}}$. The increased similarity along the main diagonal (corresponding layers) shows that our data-mixing method narrows the representational gap between synthetic and full dataset training.

et al. [43] suggests that CKA is the most consistent metric in the context of image classification, especially when comparing neural network representations across different layers and training stages, making it suitable for our experiment.

Our analysis compares three sets of layer-wise representations: $R_{\mathcal{D}}$ from a model training on the full dataset $\mathcal{D}$; $R_{\mathcal{S}}$ from a model training on a baseline distilled set $\mathcal{S}$; and $R_{\overline{\mathcal{S}}}$ from a model training on the same set after enhancement by our method, $\overline{\mathcal{S}}$. We then construct two similarity matrices, $M_{\mathcal{D}\mathcal{S}} = (CKA(r_{\mathcal{D}}^{(i)}, r_{\mathcal{S}}^{(j)}))$ and $M_{\mathcal{D}\overline{\mathcal{S}}} = (CKA(r_{\mathcal{D}}^{(i)}, r_{\overline{\mathcal{S}}}^{(j)}))$, by computing the CKA similarity between all layer pairs of ($R_{\mathcal{D}}$, $R_{\mathcal{S}}$) and ($R_{\mathcal{D}}$, $R_{\overline{\mathcal{S}}}$), respectively.

Figure 2 visualizes this comparison for the DSA method. The similarity matrix for our enhanced set, $M_{\mathcal{D}\overline{\mathcal{S}}}$ (right), shows a visibly stronger main diagonal than the baseline matrix $M_{\mathcal{D}\mathcal{S}}$ (left). This brighter diagonal indicates a higher representational alignment between corresponding layers of the full-data model and the enhanced-data model. This finding confirms our approach narrows the representational gap to full-data training. Importantly, this advantage comes with no increase in dataset size, as our method operates on the same number of samples as the synthetic baseline.

**Adversarial Robustness.** A critical aspect of deploying machine learning models, especially in safety-critical applications such as autonomous driving or medical diagnostics, is their ability to perform reliably when facing data variations. While high test accuracy is desirable, models often fail when faced with out-of-distribution (OOD) samples resulting from common corruptions or adversarial attacks. In this section, we evaluate the resilience of our models against such scenarios.

We first assess robustness against common data corruptions using the CIFAR-10-C benchmark [44]. Our data-mixing strategy consistently improves robustness in this setting across all methods under evaluation. Overall, our approach yields an average improvement of 9.05 percentage points of accuracy across all distillation settings. Specifically, with the DC [24] method at 10 IPC, our approach improves the mean accuracy on corrupted data from 44.21% to 52.38%. Similarly, for DSA [36] at 50 IPC, the robustness score increases from 50.15% to 61.29%. These results indicate that our method produces models that are not only more accurate on original data but are also more resilient to natural variations.

To assess adversarial robustness, we employ the FGSM [45] and PGD [46] attacks. For moderate perturbations ($\epsilon = 1/255$), our models consistently yield superior robustness across both attacks and all distillation settings. For example, our data-mixing strategy increases the DSA [36] method (50 IPC) accuracy from 29.5% to 41% under a PGD attack. This trend holds for other configurations, such as DM [21] at 50 IPC, where accuracy under an FGSM attack improves from 31.32% to 39.57%. At a low data regime of 10 IPC, we enhance the DC [24] baseline of 31.48% to 32.43%.

However, increasing the attack strength ($\epsilon = 8/255$) reveals a counterintuitive result: baseline models exhibit superior post-attack accuracy (e.g., 0.57% vs. 0.37% for DSA [36] at 50 IPC). This result exemplifies the concept of obfuscated gradients, by Athalye et al. [47] Their work argues that poor, chaotic landscapes hinder gradient-based attacks such as PGD and FGSM from finding effective adversarial paths, leading to a false sense of robustness. This finding also aligns with the work of Tsipras et al. [48], which highlights a fundamental trade-off between standard accuracy and adversarial robustness. Our data-mixing method produces models with superior discriminative features that yield better generalization. Yet, these same features create a clear attack vector for strong gradient perturbations. Therefore, our method confers genuine robustness under non-saturating attack conditions, in contrast to apparent robustness stemming from simplistic features.

## V. Conclusions

Despite promising results, state-of-the-art distillation methods still suffer from poor generalization compared to training on the full dataset. In this work, we mitigate this problem by introducing a simple yet effective strategy to enhance synthetic samples from a dataset distillation process. Our strategy leverages the fact that natural images carry rich and discriminative information, and incorporates them through a mixing-based augmentation process.

In contrast to previous efforts, our method preserves the compactness and synthetic nature of distilled data while improving generalization, internal representation alignment, and robustness to adversarial samples. Extensive experiments confirm that our approach enables distilled sets with as few as 10 images per class to match, or even surpass, the performance of state-of-the-art methods trained with five times more data. Additionally, our method improves robustness to adversarial and out-of-distribution samples, highlighting its potential for deployment in privacy- and security-sensitive scenarios.

**Room for Improvement.** As our analyses suggest, mixing natural images lacking their corresponding labels yields poor performance. Given that training on web-scale data often involves unsupervised and self-supervised learning, a promising direction for future work is generalizing our method to operate within these learning contexts. A key challenge in this direction is the development of criteria to select unlabeled natural images that align semantically or representationally with a given distilled image, effectively creating a semi-supervised distillation framework.

## Acknowledgments

This study was financed, in part, by the São Paulo Research Foundation (FAPESP), Brasil. Process Number #2023/11163-0 and #2024/17684-4. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001. The authors would like to thank grant #402734/2023-8, National Council for Scientific and Technological Development (CNPq). Artur Correia would like to thank Edital Programa de Apoio a Novos Docentes 2023. Processo USP nº: 22.1.09345.01.2. Anna H. Reali Costa would like to thank grant #312360/2023-1 CNPq.

## References

[1] Y. Bengio and et al., "International AI safety report," 2025.

[2] N. Maslej and et al., "Artificial intelligence index report," 2025.

[3] J. Morrison, C. Na, J. Fernandez, T. Dettmers, E. Strubell, and J. Dodge, "Holistically evaluating the environmental impact of creating language models," in *ICLR*, 2025.

[4] A. Faiz and et al., "Llmcarbon: Modeling the end-to-end carbon footprint of large language models," 2024.

[5] M. Soltanolkotabi, A. Javanmard, and J. D. Lee, "Theoretical insights into the optimization landscape of over-parameterized shallow neural networks," *IEEE Transactions on Information Theory*, 2019.

[6] H. Cheng, M. Zhang, and J. Q. Shi, "A survey on deep neural network pruning: Taxonomy, comparison, analysis, and recommendations," *IEEE TPAMI*, 2024.

[7] A. Bair, H. Yin, M. Shen, P. Molchanov, and J. M. Álvarez, "Adaptive sharpness-aware pruning for robust sparse networks," in *ICLR*, 2024.

[8] H. Lee, S. Kim, J. Lee, J. Yoo, and N. Kwak, "Coreset selection for object detection," in *CVPR*, 2024.

[9] Y. Yang, H. Kang, and B. Mirzasoleiman, "Towards sustainable learning: Coresets for data-efficient deep learning," in *ICML*, 2023.

[10] S. Lei and D. Tao, "A comprehensive survey of dataset distillation," *IEEE TPAMI*, 2024.

[11] J. Geng, Z. Chen, Y. Wang, H. Woisetschlaeger, S. Schimmler, R. Mayer, Z. Zhao, and C. Rong, "A survey on dataset distillation: Approaches, applications and future directions," in *IJCAI*, 2023.

[12] G. Li, R. Togo, T. Ogawa, and M. Haseyama, "Compressed gastric image generation based on soft-label dataset distillation for medical data sharing," *Comput. Methods Programs Biomed.*, 2022.

[13] Q. Yang, Y. Liu, Y. Cheng, Y. Kang, T. Chen, and H. Yu, *Federated Learning*. Springer International Publishing, 2022.

[14] S. Hu, J. Goetz, K. Malik, H. Zhan, Z. Liu, and Y. Liu, "Fedsynth: Gradient compression via synthetic data in federated learning," in *NeurIPS Workshop*, 2022.

[15] C.-Y. Huang, R. Jin, C. Zhao, D. Xu, and X. Li, "Federated learning on virtual heterogeneous data with local-global dataset distillation," *Transactions on Machine Learning Research*, 2025.

[16] Y. Xiong, R. Wang, M. Cheng, F. Yu, and C. Hsieh, "Feddm: Iterative distribution matching for communication-efficient federated learning," in *CVPR*, 2023.

[17] Y. Zhou, G. Pu, X. Ma, X. Li, and D. Wu, "Distilled one-shot federated learning," 2020.

[18] Y. Feng, S. R. Vedantam, and J. Kempe, "Embarrassingly simple dataset distillation," in *ICLR*, 2024.

[19] I. Sucholutsky and M. Schonlau, "Soft-label dataset distillation and text dataset distillation," in *IJCNN*, 2021.

[20] G. Cazenavette, T. Wang, A. Torralba, A. A. Efros, and J. Zhu, "Dataset distillation by matching training trajectories," in *CVPR*, 2022.

[21] B. Zhao and H. Bilen, "Dataset condensation with distribution matching," in *WACV*, 2023.

[22] H. Zhang, M. Cissé, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *ICLR*, 2018.

[23] D. Hendrycks, A. Zou, M. Mazeika, L. Tang, B. Li, D. Song, and J. Steinhardt, "Pixmix: Dreamlike pictures comprehensively improve safety measures," in *CVPR*, 2022.

[24] B. Zhao, K. R. Mopuri, and H. Bilen, "Dataset condensation with gradient matching," in *ICLR*, 2021.

[25] T. Nguyen, Z. Chen, and J. Lee, "Dataset meta-learning from kernel ridge-regression," in *ICLR*, 2021.

[26] S. Vahidian, M. Wang, J. Gu, V. Kungurtsev, W. Jiang, and Y. Chen, "Group distributionally robust dataset distillation with risk minimization," in *ICLR*, 2025.

[27] D. Liu, J. Gu, H. Cao, C. Trinitis, and M. Schulz, "Dataset distillation by automatic training trajectories," in *ECCV*, 2024.

[28] T. M. Hospedales, A. Antoniou, P. Micaelli, and A. J. Storkey, "Meta-learning in neural networks: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[29] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *ICLR*, 2015.

[30] C. Szegedy and et al., "Intriguing properties of neural networks," in *ICLR*, 2014.

[31] R. G. Lopes, D. Yin, B. Poole, J. Gilmer, and E. D. Cubuk, "Improving robustness without sacrificing accuracy with patch gaussian augmentation," 2019.

[32] R. Takahashi, T. Matsubara, and K. Uehara, "Data augmentation using random image cropping and patching for deep cnns," *IEEE Transactions on Circuits and Systems for Video Technology*, 2020.

[33] T. Devries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," 2017.

[34] D. Hendrycks, N. Mu, E. D. Cubuk, B. Zoph, J. Gilmer, and B. Lakshminarayanan, "Augmix: A simple data processing method to improve robustness and uncertainty," in *ICLR*, 2020.

[35] D. Yin, R. G. Lopes, J. Shlens, E. D. Cubuk, and J. Gilmer, "A fourier perspective on model robustness in computer vision," in *NeurIPS*, 2019.

[36] B. Zhao and H. Bilen, "Dataset condensation with differentiable siamese augmentation," in *ICML*, 2021.

[37] Y. Chen, G. Chen, M. Zhang, W. Guan, and L. Nie, "Curriculum coarse-to-fine selection for high-ipc dataset distillation," in *CVPR*, 2025, pp. 20 437–20 446.

[38] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," University of Toronto, Tech. Rep., 2009.

[39] Y. Netzer and et al., "Reading digits in natural images with unsupervised feature learning," in *NeurIPS Workshop*, 2011.

[40] D. Hendrycks, K. Lee, and M. Mazeika, "Natural adversarial examples," in *CVPR*, 2021.

[41] J. Howard, "Imagenette," 2019.

[42] S. Kornblith, M. Norouzi, H. Lee, and G. E. Hinton, "Similarity of neural network representations revisited," in *ICML*, 2019.

[43] M. Klabunde, T. Wald, T. Schumacher, K. H. Maier-Hein, M. Strohmaier, and F. Lemmerich, "Resi: A comprehensive benchmark for representational similarity measures," in *ICLR*, 2025.

[44] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," in *ICLR*, 2019.

[45] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *ICLR*, 2015.

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

[47] A. Athalye, N. Carlini, and D. Wagner, "Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples," in *ICML*, 2018.

[48] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Madry, "Robustness may be at odds with accuracy," in *ICLR*, 2019.