Robust Dataset Distillation via Adversarial Distribution Matching

Baiqi Wu
Zhejiang University
Hangzhou, Zhejiang, China
wubaiqi@zju.edu.cn

Tong Zhang
Zhejiang University
Hangzhou, Zhejiang, China
tz_zju@zju.edu.cn

Kaichao Jiang
Hefei University of Technology
Hefei, Anhui, China
2022212327@mail.hfut.edu.cn

Rui Zeng
Zhejiang University
Hangzhou, Zhejiang, China
ruizeng24@zju.edu.cn

Qingming Li
Zhejiang University
Hangzhou, Zhejiang, China
qingmingli45@163.com

Yunfeng Diao
Hefei University of Technology
Hefei, Anhui, China
diaoyunfeng@hfut.edu.cn

Shouling Ji✉
Zhejiang University
Hangzhou, Zhejiang, China
sji@zju.edu.cn

Abstract

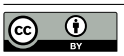
Dataset distillation compresses a large training set into a tiny synthetic set, greatly reducing training cost. However, models trained on distilled data remain highly vulnerable to adversarial attacks, especially in highly compressed regimes. A key challenge in robust dataset distillation is that adversarial robustness is not automatically inherited by synthetic data through standard distillation objectives. In particular, sharp robust teachers provide unstable supervision, while conventional feature-matching methods do not explicitly preserve adversarial distributions. To address this issue, we propose **Adversarial Distribution Matching (ADM)**, a two-stage framework that explicitly transfers robust knowledge into the distilled dataset. First, **ADM** smooths the robust teacher via randomized weight perturbation during fine-tuning, producing a more stable feature space for distillation. Second, it aligns both clean and adversarial feature distributions between real and synthetic data using characteristic function matching in the spectral domain, which captures the complex shifts induced by adversarial perturbations. Extensive experiments on multiple datasets show that **ADM** consistently achieves substantially stronger adversarial robustness than prior methods while maintaining competitive clean accuracy. The code is publicly available at https://github.com/BaiqiWu/ADM_MM26.

CCS Concepts

• **Computing methodologies** → **Machine learning**; • **Security and privacy** → *Software and application security*.

Keywords

Robust Dataset Distillation, Adversarial Defense, Trustworthy AI



MM '26, Rio de Janeiro, Brazil

ACM ISBN 979-8-4007-2213-4/2026/11
https://doi.org/10.1145/3767308.3835181

ACM Reference Format:
Baiqi Wu, Tong Zhang, Kaichao Jiang, Rui Zeng, Qingming Li, Yunfeng Diao, and Shouling Ji. 2026. Robust Dataset Distillation via Adversarial Distribution Matching. In *Proceedings of the 34th ACM International Conference on Multimedia (MM '26), November 10–14, 2026, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3767308.3835181

1 Introduction

The rapid expansion of large-scale datasets has driven major advances in deep learning, yet training on such massive data remains computationally intensive [21]. Dataset Distillation (DD) addresses this by compressing a large training set into a tiny, synthetic dataset, enabling models trained on it to achieve performance comparable to full-data training [39]. By compressing the large-scale training data, DD has the potential to reduce storage and computation and to enable efficient training in resource-constrained settings [38, 43]. Consequently, DD has unlocked new efficiency in resource-constrained settings, continual learning, and privacy preservation [29, 52].

Recent advancements in DD, such as gradient [24, 50], distribution [49, 51], and trajectory matching [2, 26], have substantially narrowed the clean accuracy gap. Despite this progress, a critical limitation remains. Models trained on distilled datasets are highly vulnerable to adversarial attacks [36, 41, 53]. This weakness is especially pronounced in the highly compressed regime, where only a few synthetic samples are used to represent the original data distribution. As a result, improving adversarial robustness has become a key challenge for dataset distillation. The goal is not only to preserve clean accuracy but also to embed sufficient robustness into the synthetic set itself. Existing attempts at robust dataset distillation have not fully resolved this problem. They often require substantial adversarial optimization costs in practice, yet still fail to achieve a satisfactory trade-off between clean accuracy and adversarial robustness (see Section 5). In addition, while adversarial defense has seen substantial progress and many adversarially pre-trained models are now publicly available, leveraging their robust knowledge for dataset distillation remains largely unexplored [3, 19].

As mentioned above, one natural direction is to leverage an adversarially pre-trained model as the teacher for distillation. Since such a model already encodes robust decision boundaries, one may

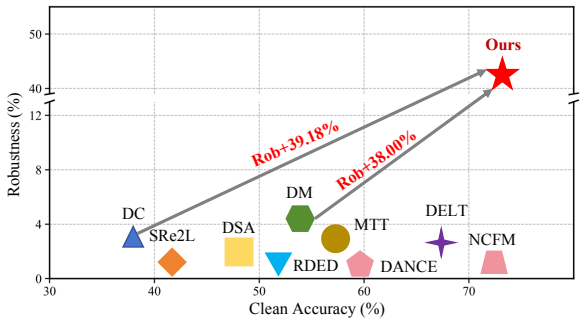


Figure 1: Comparison of state-of-the-art DD methods on CIFAR-10 in terms of standard accuracy (IPC 50) and robustness (PGD-20, $\epsilon = 4/255$). Our method achieves the best robustness while maintaining competitive standard accuracy.

expect its robustness to transfer to synthetic data through existing distillation objectives. However, our empirical study shows that this straightforward strategy still performs poorly in practice. Standard distillation pipelines struggle to obtain synthetic data with both strong clean performance and satisfactory robustness, even when guided by a robust teacher. First, our study suggests that the effectiveness of robust distillation depends not only on the teacher's robustness, but also on the smoothness of its loss landscape (see Figure 4 and Figure 3). A smoother teacher provides more stable supervision for optimizing a highly compressed synthetic set, whereas sharp local geometry can make the distillation process unstable under extreme compression. Second, our analysis indicates that existing distillation objectives mainly align clean features or low-order statistics, and therefore do not explicitly constrain how feature distributions evolve under adversarial perturbations. As a result, even when clean representations are well aligned, the distilled data may still fail to capture the adversarial feature distribution of real samples, so the adversarial structure cannot be faithfully preserved in the synthetic set.

To address these challenges, we propose **Adversarial Distribution Matching (ADM)**, a novel two-stage framework for robust dataset distillation. In the first stage, we smooth the robust teacher through randomized weight perturbation during fine-tuning, which provides a more stable feature space for distillation. In the second stage, we explicitly align both clean and adversarial feature distributions between real and synthetic data. To better capture the complex distributions caused by adversarial perturbations, we perform this alignment using characteristic function matching in the spectral domain. Extensive experiments on multiple datasets demonstrate that **ADM** consistently outperforms existing baselines and substantially narrows the gap between clean accuracy and adversarial robustness. In particular, **ADM** achieves strong robustness under standard PGD [28] and AutoAttack [5] attacks while maintaining competitive clean accuracy on deep architectures such as ResNet-18. Our main contributions are summarized as follows:

- We analyze why robustness is difficult to transfer in existing distillation pipelines, and show that the main challenges lie in sharp robust teachers and insufficient modeling of adversarial feature distributions.
- We propose Adversarial Distribution Matching (**ADM**), a robust dataset distillation framework that smooths the teacher and explicitly aligns adversarial distributions through characteristic function matching.
- We achieve state-of-the-art results in robust dataset distillation across multiple datasets, substantially improving adversarial robustness while preserving strong clean accuracy, yielding a better robustness–accuracy trade-off.

2 Related Work

2.1 Dataset Distillation

Existing methods can be broadly organized by how they match information between real and synthetic data. Trajectory matching-based methods optimize the synthetic dataset by aligning the parameter update trajectories of models trained on the original and synthetic data [2, 10, 15, 26]. For example, Liu et al. [26] proposed dynamically adjusting the trajectory matching length to resolve the accumulated mismatching problem inherent in fixed-step schedules. Gradient matching-based methods optimize the synthetic dataset by minimizing the discrepancy between the gradients obtained from training models on the synthetic and original data [24, 48, 50]. Distribution matching aligns feature distributions between real and synthetic data with embedding networks sampled during optimization, substantially reducing the need to backpropagate through long inner training loops [38, 44, 49]. Zhao et al. [51] proposed a partition and expansion augmentation strategy to mitigate feature-level imbalance and capture fine-grained structural information. Liu et al. [27] introduced representative matching to improve sample efficiency and reduce redundancy in the distilled set.

2.2 Adversarial Robustness Distillation

The concept of adversarial robustness has been extensively explored in the context of adversarial attacks and defenses [8, 28]. However, despite significant advancements in adversarial defense (e.g., full adversarial training), research aimed at improving the robustness of student models in DD remains relatively limited. Notable prior works in this area include DD-Robustbench [41], a benchmark designed to evaluate robustness in distilled datasets, and BEARD [53], which offers a comprehensive robustness evaluation using a game-theoretic framework. However, their focus lies in evaluating the existing DD algorithms rather than improving their robustness. More recently, efforts have been made to explicitly distill robust datasets by incorporating robustness-aware objectives into the distillation process. For example, Xue et al. [42] proposed GUARD, which encourages curvature patterns correlated with adversarial robustness during dataset condensation. ROME [54], on the other hand, utilizes the information-bottleneck principle and aligns synthetic features with perturbed images to enhance robustness. Nonetheless, these methods remain far from optimal under standard robustness evaluations, exhibiting limited accuracy and robustness (Table 1). As a result, achieving a good balance between accuracy and robustness remains a challenging problem.

3 Preliminaries

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ be a training set drawn from a data distribution $\mathbb{P}$, where $x \in \mathcal{X} \subseteq \mathbb{R}^d$ and $y \in \{1, \ldots, C\}$. A model is denoted by

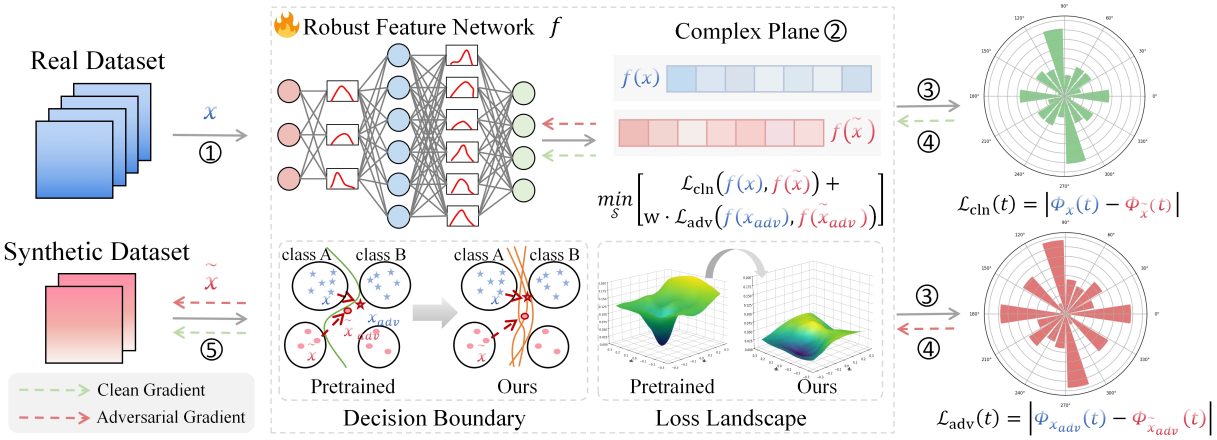


Figure 2: Illustration of the proposed Adversarial Distribution Matching (ADM) framework.

$f_\theta : \mathcal{X} \to \mathbb{R}^C$ with parameters θ. Let $\ell(f_\theta(x), y)$ be the per-sample loss and

$$\mathcal{L}_{\mathcal{D}}(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\big[\ell(f_\theta(x), y)\big], \tag{1}$$

the empirical risk on $\mathcal{D}$. In DD, the goal is to synthesize a compact set $\mathcal{S} = \{(\tilde{x}_j, \tilde{y}_j)\}_{j=1}^{M}$ with $M = Cm \ll N$, where C is the number of classes and m is the number of Images Per Class (IPC).

Robust dataset distillation. We consider the ℓ_p threat model with perturbation budget $\varepsilon > 0$. The goal of robust dataset distillation is to find a compact synthetic set $\mathcal{S}$ such that a model trained on it achieves low adversarial risk on the real data distribution:

$$\min_{\mathcal{S},\,|\mathcal{S}|=M} \mathbb{E}_{(x,y)\sim\mathbb{P}}\Big[\max_{\|\delta\|_p\le\varepsilon} \ell\big(f_{\theta_{\mathcal{S}}}(x+\delta), y\big)\Big], \tag{2}$$

where δ denotes the adversarial perturbation and $\theta_{\mathcal{S}}$ denotes the model parameters obtained by training on $\mathcal{S}$. The key challenge is to construct $\mathcal{S}$ that encodes sufficient adversarial structure so that the resulting model is robust. Directly optimizing Eq. (2) is intractable, as it requires solving a nested optimization: the outer level optimizes $\mathcal{S}$, the inner level trains $\theta_{\mathcal{S}}$, and the evaluation itself involves a worst-case maximization over δ.

Distribution matching. Instead of tackling this nested optimization end-to-end, we adopt **distribution matching (DM)** [49, 51] as the underlying mechanism to construct $\mathcal{S}$. The core idea of DM is to synthesize samples whose feature distribution matches that of real data under a chosen feature extractor $g(\cdot)$. Let $\mathbb{P}$ and $\widetilde{\mathbb{P}}$ denote the feature distributions induced by the real set $\mathcal{D}$ and the synthetic set $\mathcal{S}$, respectively. DM minimizes a distributional discrepancy:

$$\min_{\mathcal{S}} \Delta\Big(\mathbb{E}_{x\sim\mathbb{P}}[g(x)],\ \mathbb{E}_{\tilde{x}\sim\widetilde{\mathbb{P}}}[g(\tilde{x})]\Big). \tag{3}$$

where $\Delta(\cdot,\cdot)$ denotes a discrepancy measure, such as Mean Squared Error (MSE) between feature means [7, 31, 37]. This formulation can also be extended to richer distributional metrics such as Maximum Mean Discrepancy (MMD) [45, 49, 51], which compares higher-order statistics beyond the mean. Compared with trajectory- or gradient-matching paradigms, DM does not require backpropagation through long inner training loops, making it a natural starting point for incorporating adversarial alignment. However, standard DM only matches the clean feature distribution and does not explicitly model how representations change under adversarial perturbations. Addressing this limitation is the central focus of our method (Section 4).

4 Methodology

In this section, we present the Adversarial Distribution Matching (**ADM**) framework. **ADM** comprises two stages: (i) constructing a smoothed robust teacher through randomized weight perturbations during fine-tuning, and (ii) synthesizing a robust distilled dataset by matching adversarial characteristic functions in the spectral domain. The overall framework is illustrated in Figure 2, and the complete procedure is summarized in Algorithm 1.

4.1 Motivation

As mentioned above, a natural approach is to use an adversarially pre-trained model as the teacher for robust dataset distillation. An even simpler alternative is to apply adversarial training directly on distilled data produced by an existing DD method. However, our experiments show that this post-hoc strategy leads to unstable optimization and rapid model collapse: on CIFAR-100 with IPC 50, the test accuracy remains near chance level throughout training and never recovers (see the supplementary material), likely because the tiny synthetic set lacks sufficient coverage of adversarial neighborhoods near class boundaries. This indicates that robustness must be incorporated during distillation itself. However, our empirical results show that simply using a robust teacher within a standard pipeline also fails to yield synthetic data with both strong clean accuracy and satisfactory robustness.

To understand what limits robustness transfer in this setting, we compare three widely used adversarially pre-trained teachers, namely TRADES [46], AWP [40], and DHAT [47], within two representative distillation frameworks, DELT [33] and NCFM [38]. For a compact comparison, Figure 4 plots the clean accuracy and robustness of each configuration, with HRS contours overlaid to visualize the overall trade-off (higher values indicate better balance). The full definition of HRS is provided in the supplementary material. Through this comparison, we make three key observations.

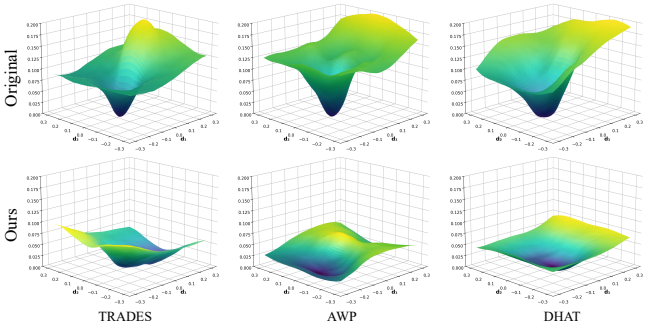


Figure 3: Comparison of the loss landscapes [11] of the original model (top row) and the fine-tuned model (ours; bottom row). Loss landscapes are generated from the same original image randomly selected from the CIFAR-100 test set.

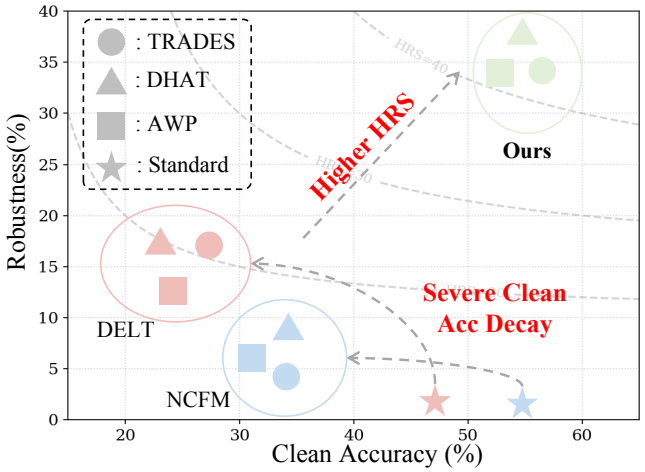


Figure 4: Motivational comparison of different teacher models for various DD methods on CIFAR-10 (IPC 10). Robustness is evaluated under PGD-20 ($\epsilon = 4/255$).

Observation 1: teacher robustness alone is not sufficient for robust dataset distillation. When DELT and NCFM use a standard teacher, they retain reasonable clean accuracy but achieve almost no robustness. Replacing the standard teacher with an adversarially pre-trained one does not automatically resolve this. Although the teacher itself is robust, the resulting distilled data still fails to achieve a satisfactory balance between clean accuracy and robustness, and in some cases clean performance drops substantially. This indicates that robust knowledge is not automatically inherited by the synthetic set through standard distillation objectives.

Observation 2: robustness transfer depends strongly on the smoothness of the teacher's loss landscape. As shown in Figure 3, different teachers exhibit noticeably different loss landscapes. Among them, DHAT is smoother than TRADES, and it consistently leads to better robust distillation results across both DELT and NCFM. This is especially important in the DD setting because the synthetic set is optimized directly under teacher guidance. When the teacher response changes too abruptly, small updates to synthetic samples can induce large changes in loss or features, making the optimization unstable under extreme compression.

Observation 3: clean feature matching does not preserve adversarial structure. Even with a smooth robust teacher, existing distillation objectives mainly align clean features or low-order statistics. Such objectives may preserve the clean structure of real data, but do not explicitly constrain how representations evolve under adversarial perturbations. As a result, the distilled set may fail to capture the adversarial behavior of real samples.

These observations suggest that effective robust dataset distillation requires two key conditions: (1) the teacher should exhibit a smooth loss landscape, providing stable supervisory signals, and (2) the distillation objective should explicitly preserve adversarial feature distributions, not only clean feature structure. These two conditions directly correspond to Stage I (Section 4.2) and Stage II (Section 4.3) of our framework, respectively.

4.2 Smoothing Robust Teacher Model

As discussed in Section 4.1, the smoothness of the teacher's loss landscape largely determines the amount of adversarial robustness that can be transferred into the synthetic set $\mathcal{S}$. Motivated by the observation that randomized weight perturbations can bias optimization toward flatter minima [13, 17, 18, 20, 40], we smooth a robustly pre-trained teacher by introducing small Gaussian weight perturbations during adversarial fine-tuning. Concretely, for each mini-batch we construct a randomized surrogate teacher to generate adversarial examples in a local neighborhood of the current weights, while the base teacher is updated using deterministic clean and robust supervision together with lightweight Taylor proxies.

Specifically, for each mini-batch we construct a randomized surrogate teacher by perturbing the current weights with a small Gaussian perturbation:

$$\tilde{\theta} = \theta + u, \qquad u \sim \mathcal{N}(0, \sigma^2 I), \tag{4}$$

where u denotes the sampled weight perturbation. The smoothing effect can be formalized through the following Gaussian-smoothed objective:

$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\ell\big(f_\theta(x), y\big) + \beta\, \mathbb{E}_u\Big[\max_{\|\delta\|_p \le \varepsilon} \ell\big(f_{\theta+u}(x+\delta), y\big)\Big]\right], \tag{5}$$

Eq. (5) should be viewed as a conceptual smoothed robust objective in weight space. In practice, rather than differentiating the expectation over randomized weights exactly, we follow a single-sample randomized-surrogate approximation and optimize the resulting base-model objective with lightweight Taylor proxies.

For the inner maximization, we sample one noise vector u per mini-batch, instantiate the randomized surrogate $f_{\tilde{\theta}}$, and keep $\tilde{\theta}$ fixed during the entire PGD procedure. We generate adversarial examples by T-step PGD:

$$x_{\text{adv}}^{(t+1)} \leftarrow \Pi_{\mathcal{B}_p(x,\varepsilon)}\Big(x_{\text{adv}}^{(t)} + \eta_{\text{atk}} \cdot \operatorname{sign}\big(\nabla_x \ell(f_{\tilde{\theta}}(x_{\text{adv}}^{(t)}), y)\big)\Big), \tag{6}$$

and set $x_{\text{adv}} = x_{\text{adv}}^{(T)}$. For the outer update, we switch back to the deterministic teacher f_θ and optimize:

$$\mathcal{L}_{\text{total}} = (1-\alpha)\mathcal{L}_{\text{cln}} + \beta\mathcal{L}_{\text{adv}} + \alpha\mathcal{L}_{\text{1st}} + \frac{\alpha}{2}\mathcal{L}_{\text{2nd}}. \tag{7}$$

Algorithm 1 Robust Dataset Distillation via Adversarial Distribution Matching

Require: Real dataset $\mathcal{D}$; initial robust teacher θ_0; IPC budget m; noise std σ; PGD steps T and step size η_{atk}; Taylor proxy weight α; adversarial loss weight β; frequency sampler $F_{\mathcal{T}}$ with n_f frequencies; adversarial CFD weight w; learning rates $\eta_\theta, \eta_{\mathcal{S}}$.
Ensure: Distilled dataset $\mathcal{S}$ and smoothed robust teacher $\theta^\star$.

// **Stage I: Smooth and robust teacher model**
Initialize $\theta \leftarrow \theta_0$
for $epoch = 1$ **to** $Epoch_{\text{teacher}}$ **do**
  Sample minibatch $\mathcal{B} \subset \mathcal{D}$
  Sample weight noise $u \sim \mathcal{N}(0, \sigma^2 I)$; set surrogate $\tilde{\theta} = \theta + u$ {Implicitly flatten the loss landscape}
  Generate adversarial examples $\{x_{\text{adv}}\}$ for $\mathcal{B}$ by T-step PGD under $f_{\tilde{\theta}}$ with step size η_{atk}
  Update $\theta \leftarrow \theta - \eta_\theta \nabla_\theta \mathcal{L}_{\text{total}}(\theta; \mathcal{B})$ using Eq. (7) {Outer update on deterministic f_θ}
end for
Set $\theta^\star \leftarrow \theta$ and extractor $g(\cdot) = g_{\theta^\star}(\cdot)$
// **Stage II: Adversarial distribution alignment**
Initialize $\mathcal{S}$ with m labeled synthetic samples per class
for $iter = 1$ **to** $Iter_{\text{distill}}$ **do**
  Sample minibatches $\mathcal{B} \subset \mathcal{D}$ and $\widetilde{\mathcal{B}} \subset \mathcal{S}$; sample frequencies $\mathcal{M} \sim F_{\mathcal{T}}$ with $|\mathcal{M}| = n_f$
  Generate x_{adv} for $x \in \mathcal{B}$ and $\tilde{x}_{\text{adv}}$ for $\tilde{x} \in \widetilde{\mathcal{B}}$ by T-step PGD under $f_{\theta^\star}$
  Compute clean CFD C_{cln} and adversarial CFD C_{adv} by Eq. (13)
  Update $\mathcal{S} \leftarrow \mathcal{S} - \eta_{\mathcal{S}} \nabla_{\mathcal{S}} \mathcal{L}$, where $\mathcal{L}$ is computed via Eq. (15)
end for
return $\mathcal{S}, \theta^\star$

where $\mathcal{L}_{\text{cln}}$ and $\mathcal{L}_{\text{adv}}$ are the clean and adversarial cross-entropy losses evaluated on the deterministic teacher f_θ, and $\mathcal{L}_{1\text{st}}$, $\mathcal{L}_{2\text{nd}}$ are lightweight Taylor proxy losses that penalize the discrepancy between the classifier's gradient statistics on clean and adversarial inputs, further flattening the teacher's loss landscape by encouraging locally consistent gradient responses (precise forms are given in the supplementary material). In this way, the randomized surrogate is used to smooth adversarial-example generation, while the base teacher is updated by deterministic clean and adversarial supervision regularized by Taylor-inspired local sensitivity alignment. Theorem 4.1 characterizes Gaussian smoothing of the conceptual robust objective in weight space, providing theoretical motivation for the randomized-surrogate design used in our teacher fine-tuning procedure.

Theorem 4.1 (Gaussian smoothing of the robust branch). *Let $z \sim \mathcal{N}(0, \Sigma)$ be independent of θ, and define*

$$\phi(\vartheta; x, y) := \max_{\|\delta\|_p \le \varepsilon} \ell(f_\vartheta(x + \delta), y), \tag{8}$$

$$\mathcal{R}_\Sigma(\theta; x, y) := \mathbb{E}_z[\phi(\theta + z; x, y)]\,. \tag{9}$$

Then $\mathcal{R}_\Sigma$ is the Gaussian convolution of ϕ in parameter space:

$$\mathcal{R}_\Sigma(\theta; x, y) = \int_{\mathbb{R}^d} \phi(v; x, y)\, g_\Sigma(v - \theta)\, dv, \tag{10}$$

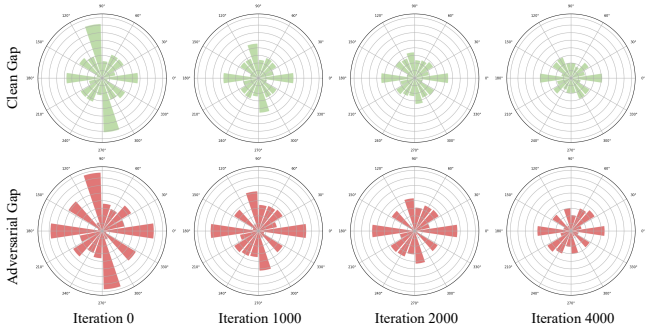


Figure 5: Characteristic function discrepancy along uniformly spaced frequency directions in PCA-projected feature space. Top row: clean inputs; bottom row: adversarial inputs. Both gaps shrink as distillation progresses.

where g_Σ is the density of $\mathcal{N}(0, \Sigma)$. Hence, if $\phi(\cdot; x, y)$ is locally integrable and has at most polynomial growth, then $\mathcal{R}_\Sigma(\cdot; x, y) \in C^\infty$. Moreover, under standard dominated-convergence conditions,

$$\nabla_\theta \mathcal{R}_\Sigma(\theta; x, y) = \mathbb{E}_z[\nabla_\theta \phi(\theta + z; x, y)]\,. \tag{11}$$

Therefore, sampling one perturbation per mini-batch yields a single-sample Monte Carlo estimate of the gradient of the smoothed robust branch.

The proof is provided in the supplementary material.

4.3 Adversarial Distribution Alignment

As established in Section 4.1 (Observation 3), matching the clean feature distribution alone does not guarantee that adversarial structure is preserved in the synthetic data. To address this, we propose explicitly aligning the adversarial feature distributions. Because adversarial perturbations induce complex, heavy-tailed geometric distortions [30], standard low-order matching metrics (e.g., MSE) [45] are fundamentally insufficient. Instead, we leverage the Characteristic Function (CF) to align the entire probability distribution of adversarial responses via infinite-order moment matching [12]. This spectral perspective captures the full distributional geometry rather than overfitting to low-order summary statistics.

Characteristic Function Discrepancy. For a random variable $\boldsymbol{z}$ representing the feature vectors extracted by $g(\cdot)$, its distribution can be uniquely and comprehensively described by its characteristic function. Specifically, the CF of $\boldsymbol{z}$ is the expectation of its complex exponential transform:

$$\Phi_{\boldsymbol{z}}(\boldsymbol{t}) = \mathbb{E}_{\boldsymbol{z}}\left[e^{j\langle \boldsymbol{t}, \boldsymbol{z}\rangle}\right] = \int_{\boldsymbol{z}} e^{j\langle \boldsymbol{t}, \boldsymbol{z}\rangle} dF(\boldsymbol{z}), \tag{12}$$

where $F(\boldsymbol{z})$ denotes the cumulative distribution function (CDF) of $\boldsymbol{z}$, $j = \sqrt{-1}$ is the imaginary unit, and $\boldsymbol{t}$ is the frequency argument. In practice, the CF is approximated empirically using the feature set $\mathcal{Z} = \{g(x) \mid x \in \mathcal{D}\}$ or $\widetilde{\mathcal{Z}} = \{g(\tilde{x}) \mid \tilde{x} \in \mathcal{S}\}$ as $\Phi_{\boldsymbol{z}}(\boldsymbol{t}) = \frac{1}{N}\sum_{i=1}^{N} e^{j\langle \boldsymbol{t}, \boldsymbol{z}_i\rangle}$, where N is the sample size in the dataset.

Based on the uniqueness and convergence properties of CFs [25], we define the Characteristic Function Discrepancy (CFD) between

Table 1: Robustness evaluation under various adversarial attacks on CIFAR-10 and CIFAR-100. All evaluations are conducted on ResNet-18 with $\varepsilon = 4/255$. † denotes officially released open-source distilled data generated using ConvNet.

IPC		IPC 1				IPC 10				IPC 50			
Dataset	Method	Clean	FGSM	PGD-20	AA	Clean	FGSM	PGD-20	AA	Clean	FGSM	PGD-20	AA
CIFAR-10	DC† [50]	26.02±1.69	11.79±1.48	10.95±1.43	10.77±1.39	38.33±1.08	9.51±0.62	6.92±0.63	6.83±0.62	39.75±0.37	6.28±0.56	3.71±0.45	3.64±0.45
	DSA† [48]	26.42±0.74	14.99±1.25	14.46±1.32	14.33±1.33	36.61±0.80	7.56±1.02	5.31±1.04	5.19±1.06	49.66±0.71	6.01±0.54	2.93±0.30	2.93±0.33
	MTT† [2]	17.39±2.17	8.69±2.05	8.24±2.11	8.14±2.10	41.38±0.64	7.58±0.59	4.99±0.59	4.81±0.55	56.57±1.09	5.64±0.39	2.22±0.28	2.21±0.25
	SRe2L [43]	12.53±1.01	5.26±0.69	4.37±0.65	4.25±0.58	25.60±0.43	5.35±0.68	3.05±0.63	2.81±0.61	43.25±1.04	3.06±0.31	0.51±0.15	0.42±0.13
	DM† [49]	13.74±1.60	11.38±0.72	11.30±0.68	11.26±0.68	39.75±1.46	9.77±0.60	7.35±0.77	7.23±0.76	54.79±1.17	9.09±0.71	4.89±0.50	4.88±0.47
	DANCE [44]	21.65±0.29	9.60±0.37	8.24±0.37	7.72±0.41	49.37±1.81	2.04±0.17	0.23±0.12	0.14±0.06	58.08±1.36	3.20±0.41	0.01±0.01	0.00±0.00
	RDED [35]	20.48±0.98	6.43±0.61	4.53±0.92	4.08±1.04	29.08±0.98	2.14±0.44	0.73±0.25	0.61±0.26	52.77±1.60	4.79±0.52	1.27±0.36	1.17±0.37
	DELT [33]	20.85±0.66	10.17±0.84	9.13±0.90	8.72±0.80	37.23±0.89	9.68±1.31	6.32±0.91	5.84±0.86	67.05±0.95	9.20±0.53	2.01±0.52	1.78±0.45
	NCFM [38]	27.52±0.92	5.42±0.47	3.78±0.41	3.32±0.36	51.99±1.15	7.37±0.65	3.80±0.47	3.06±0.48	72.95±0.68	9.86±2.14	0.03±0.02	0.01±0.01
	GUARD [42]	18.23±1.23	8.33±1.91	7.18±1.75	6.93±1.88	22.12±0.89	4.59±0.83	3.06±0.91	2.87±0.94	49.75±1.25	10.47±0.45	5.29±0.59	4.99±0.59
	ROME† [54]	9.51±0.66	8.46±1.93	8.45±1.94	8.40±2.01	11.28±2.39	5.12±1.87	4.62±1.84	4.34±1.86	18.01±1.57	5.83±0.61	4.62±0.44	4.60±0.45
	Ours	**39.79±0.87**	**28.37±0.74**	**18.61±0.73**	**15.06±0.79**	**61.00±0.67**	**46.58±0.43**	**32.22±0.21**	**27.78±0.25**	**73.08±0.18**	**59.15±0.39**	**42.89±0.56**	**37.97±0.86**
CIFAR-100	DC† [50]	9.26±0.35	2.98±0.36	2.44±0.43	2.19±0.42	12.94±0.26	3.09±0.23	2.17±0.20	2.09±0.18	16.56±0.85	3.22±0.08	1.68±0.10	1.49±0.13
	DSA† [48]	11.02±0.23	3.54±0.24	2.82±0.25	2.51±0.25	21.10±0.48	4.44±0.16	2.82±0.15	2.65±0.15	41.19±0.68	6.23±0.44	2.07±0.36	1.73±0.34
	MTT† [2]	4.35±0.28	1.84±0.21	1.65±0.19	1.48±0.18	23.72±0.70	3.32±0.20	1.81±0.18	1.64±0.16	42.09±0.64	2.14±0.20	0.17±0.07	0.11±0.04
	SRe2L [43]	4.69±0.16	1.23±0.06	0.93±0.10	0.58±0.09	23.94±0.91	2.75±0.26	0.55±0.04	0.33±0.04	52.66±0.52	6.13±0.36	0.08±0.03	0.02±0.01
	DM† [49]	3.23±0.27	1.11±0.22	1.00±0.22	0.82±0.24	20.92±0.64	3.31±0.20	1.73±0.18	1.64±0.18	41.81±0.26	4.80±0.22	0.92±0.20	0.74±0.16
	DANCE [44]	5.29±0.09	1.39±0.16	1.10±0.15	0.94±0.14	22.43±1.02	1.19±0.25	0.01±0.01	0.00±0.00	26.15±0.98	2.75±0.53	0.02±0.03	0.00±0.00
	RDED [35]	9.83±0.45	1.38±0.22	1.05±0.20	0.57±0.13	43.22±0.60	4.79±0.16	1.14±0.12	0.68±0.10	66.70±0.56	10.66±0.38	0.63±0.05	0.33±0.04
	DELT [33]	8.29±0.53	2.52±0.17	2.05±0.15	1.50±0.14	45.53±0.92	5.01±0.34	0.16±0.04	0.05±0.02	66.29±0.44	8.92±0.26	0.20±0.02	0.06±0.00
	NCFM [38]	10.91±0.82	2.00±0.13	1.56±0.13	1.13±0.09	44.73±0.54	4.75±0.69	0.19±0.05	0.03±0.01	52.50±0.81	6.28±0.42	0.06±0.02	0.00±0.01
	GUARD [42]	9.60±0.26	2.16±0.18	1.65±0.15	1.47±0.17	38.41±0.85	4.56±0.18	1.28±0.08	1.06±0.07	53.64±0.30	6.84±0.13	1.65±0.09	1.42±0.09
	ROME† [54]	1.12±0.12	0.91±0.08	0.90±0.08	0.88±0.11	0.99±0.06	0.76±0.19	0.75±0.19	0.75±0.19	1.61±0.10	0.49±0.17	0.37±0.16	0.34±0.17
	Ours	**33.64±1.35**	**22.72±0.68**	**14.70±0.50**	**10.12±0.31**	**52.53±0.27**	**41.27±0.52**	**27.91±0.14**	**14.82±0.50**	**57.35±0.78**	**45.79±0.79**	**35.93±0.17**	**16.14±1.16**

the real distribution $\mathbb{P}$ and the synthetic distribution $\widetilde{\mathbb{P}}$ as:

$$C_{\mathcal{T}}(\mathbb{P}, \widetilde{\mathbb{P}}) = \int_{\boldsymbol{t}} \left| \Phi_{\mathbb{P}}(\boldsymbol{t}) - \Phi_{\widetilde{\mathbb{P}}}(\boldsymbol{t}) \right| \, dF_{\mathcal{T}}(\boldsymbol{t}), \tag{13}$$

where $F_{\mathcal{T}}$ denotes the sampling distribution over the frequency argument $\boldsymbol{t}$. By integrating across the frequency domain, $C_{\mathcal{T}}$ functions as a unified distance metric that inherently captures the complex adversarial distributions crucial to robust dataset distillation.

Dual-Feature Distribution Matching. Traditional distribution matching relies exclusively on clean features, leaving models highly vulnerable to adversarial attacks. To bridge this gap, we propose a comprehensive Dual-Feature Alignment strategy that explicitly matches the teacher's representations under both natural and adversarial conditions. For real samples $x \sim \mathbb{P}$ and synthetic samples $\tilde{x} \sim \widetilde{\mathbb{P}}$, we concurrently extract their clean features, $g(x)$ and $g(\tilde{x})$, and generate their corresponding PGD adversarial counterparts, x_{adv} and $\tilde{x}_{\text{adv}}$. To capture their complete distributional statistics, we formulate the characteristic function for both domains. Specifically, the Adversarial Characteristic Function (ACF) encapsulates the robust feature distribution:

$$\Phi_{\text{adv}}(\boldsymbol{t}; \mathbb{P}) = \mathbb{E}_{x \sim \mathbb{P}} \left[e^{j\langle \boldsymbol{t}, g(x_{\text{adv}}) \rangle} \right], \quad \Phi_{\text{adv}}(\boldsymbol{t}; \widetilde{\mathbb{P}}) = \mathbb{E}_{\tilde{x} \sim \widetilde{\mathbb{P}}} \left[e^{j\langle \boldsymbol{t}, g(\tilde{x}_{\text{adv}}) \rangle} \right], \tag{14}$$

A parallel clean characteristic function, Φ_{cln}, is derived identically using the unperturbed features. The discrepancy for both feature spaces is then evaluated by substituting Φ_{cln} and Φ_{adv} into Eq. (13), yielding the respective spectral losses $\mathcal{L}_{\text{cln}}$ and $\mathcal{L}_{\text{adv}}$.

Unlike standard metrics (e.g., MSE) that inherently focus on aligning low-order statistics, adversarial perturbations induce complex, heavy-tailed deformations. Having defined the characteristic functions for both clean and adversarial feature spaces, we now formulate the joint optimization objective used to update the synthetic dataset.

Joint Distillation Objective. To make Eq. (13) computationally tractable, we approximate the integral via Monte Carlo sampling over a discrete frequency set $\mathcal{M} \sim F_{\mathcal{T}}$. Our final objective optimizes the synthetic dataset $\mathcal{S}$ by jointly minimizing the empirical clean and adversarial characteristic function discrepancies:

$$\min_{S} \quad \frac{1}{|\mathcal{M}|} \sum_{\boldsymbol{t} \in \mathcal{M}} \Big[\Big| \Phi_{\text{cln}}(\boldsymbol{t}; \mathbb{P}) - \Phi_{\text{cln}}(\boldsymbol{t}; \widetilde{\mathbb{P}}) \Big| + w \Big| \Phi_{\text{adv}}(\boldsymbol{t}; \mathbb{P}) - \Phi_{\text{adv}}(\boldsymbol{t}; \widetilde{\mathbb{P}}) \Big| \Big]. \tag{15}$$

where w is a hyperparameter that balances the clean and adversarial alignment terms. Figure 5 visualizes how the characteristic-function discrepancy on both clean and adversarial features gradually shrinks during distillation. The following theorem formally justifies CF matching by establishing its connection to full distributional alignment:

Theorem 4.2 (Distributional Implication of CF Matching). *Let $\mathbb{P}_{\text{adv}}$ and $\widetilde{\mathbb{P}}_{\text{adv}}$ denote the adversarial feature distributions of real and synthetic data, respectively. If*

$$\Phi_{\text{adv}}(\mathbf{t}; \mathbb{P}) \equiv \Phi_{\text{adv}}(\mathbf{t}; \widetilde{\mathbb{P}}), \qquad \forall\, \mathbf{t} \in \mathbb{R}^d, \tag{16}$$

then

$$\mathbb{P}_{\text{adv}} = \widetilde{\mathbb{P}}_{\text{adv}}. \tag{17}$$

Table 2: ImageNet-100 results at IPC 10 (accuracy, %).

Method	Clean	FGSM	PGD-20	AA
DC	4.64 ± 0.56	1.28 ± 0.16	1.03 ± 0.13	0.95 ± 0.12
DSA	10.66 ± 0.41	1.94 ± 0.22	0.67 ± 0.07	0.58 ± 0.09
MTT	8.13 ± 0.48	1.04 ± 0.08	0.27 ± 0.04	0.22 ± 0.05
SRe2L	19.78 ± 0.67	0.67 ± 0.13	0.16 ± 0.00	0.15 ± 0.01
DM	12.81 ± 0.64	2.25 ± 0.22	0.71 ± 0.13	0.59 ± 0.12
DANCE	12.14 ± 0.21	1.82 ± 0.09	0.46 ± 0.07	0.33 ± 0.06
RDED	33.94 ± 1.05	4.48 ± 0.22	0.74 ± 0.04	0.34 ± 0.04
DELT	27.04 ± 0.87	2.13 ± 0.17	0.24 ± 0.03	0.08 ± 0.02
NCFM	32.28 ± 0.75	2.40 ± 0.19	0.16 ± 0.02	0.07 ± 0.03
GUARD	23.53 ± 0.43	0.71 ± 0.13	0.07 ± 0.04	0.06 ± 0.04
ROME	2.96 ± 0.28	0.76 ± 0.16	0.41 ± 0.19	0.38 ± 0.19
Ours	**39.74 ± 0.37**	**25.53 ± 0.52**	**20.47 ± 0.38**	**14.54 ± 0.06**

Consequently, for any measurable function h with finite expectation,

$$\mathbb{E}_{\mathbf{z}\sim\mathbb{P}_{\text{adv}}}[h(\mathbf{z})] = \mathbb{E}_{\tilde{\mathbf{z}}\sim\tilde{\mathbb{P}}_{\text{adv}}}[h(\tilde{\mathbf{z}})]. \tag{18}$$

In particular, all finite uncentered cross-moments that exist under these distributions are identical.

The proof is provided in the supplementary material.

5 Experiments

5.1 Experimental Settings

Baseline Methods. We evaluate the effectiveness of our proposed method by comparing it with a diverse set of representative baselines. These include general dataset distillation methods from several major categories: gradient-matching methods, such as **DC** [50] and **DSA** [48]; distribution-matching methods, such as **DM** [49] and **DANCE** [44]; trajectory-matching methods, such as **MTT** [2]; and optimization-based methods, such as **SRe2L** [43]. We also include recent state-of-the-art methods, including **RDED** [35], **DELT** [33], and **NCFM** [38]. Moreover, we compare our method against robust-specific baselines, including **GUARD** [42] and **ROME** [54].

Datasets and Networks. Experiments are conducted on **CIFAR-10**, **CIFAR-100** [22], **Tiny-ImageNet** [23], **ImageNet-100**, and ImageNet subsets. Results for the six 10-class ImageNet subsets are reported in the supplementary material. Following the convention in adversarial robustness research, we adopt **ResNet-18** [16] as both the default teacher and student models. It is important to note that scaling existing dataset distillation methods (e.g., gradient or trajectory matching) to deep architectures like ResNet-18 is non-trivial, often leading to severe optimization instability or non-convergence. To ensure a fair and rigorous comparison, we directly used the officially released distilled datasets generated by ConvNet [49] for these methods in our experiments. Regardless of the architecture used for synthesis, all distilled datasets are evaluated on **ResNet-18** to establish a consistent benchmark.

Evaluation Metrics. We report **Clean Accuracy** $\mathcal{A}_{\text{clean}} = \Pr(f(x) = y)$ and **Robust Accuracy** $\mathcal{A}_{\text{rob}} = \Pr(f(x_{\text{adv}}) = y)$. Together, these metrics characterize the trade-off between standard predictive utility and adversarial resilience. In the main results, we report clean and robust accuracy separately for transparency, while using the **Harmonic Robustness Score (HRS)** only when a single summary of the clean-robust trade-off is needed. Its definition is given in the supplementary material.

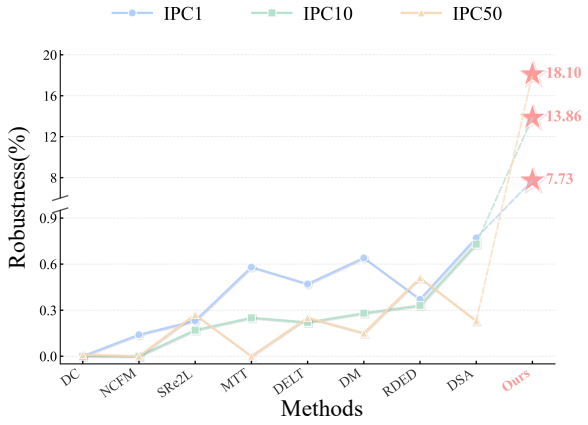


Figure 6: Results on Tiny-ImageNet under the same evaluation protocol as Table 1. ADM consistently achieves stronger robustness under PGD-20.

Evaluation Attacks. We evaluate the models under the ℓ_∞ threat model with a maximum perturbation budget of $\varepsilon = 4/255$. We report robust accuracy against both FGSM [14] and a 20-step PGD [28] attack (**PGD-20**) with a step size of $\eta_{\text{atk}} = 1/255$. We also report performance under standard AutoAttack (AA) [5] which consists of APGD-CE [5], APGD-DLR [5], FAB [4] and Square [1].

Implementation Details. To ensure fairness and reliability, all results are averaged over five independent runs with different random seeds, and we report the mean and standard deviation. In distribution matching, adversarial examples for both real and synthetic data are generated using 5-step ℓ_∞ PGD with $\varepsilon = 8/255$ and step size 2/255. Detailed hyper-parameter configurations for each dataset and IPC budget are provided in the supplementary material.

5.2 Main Results

Adversarial robustness. Table 1 reports the robustness evaluation under FGSM, PGD-20, and AutoAttack. Across both CIFAR-10 and CIFAR-100, **ADM** consistently delivers the strongest robustness among all compared methods while maintaining competitive clean accuracy. On CIFAR-10, **ADM** attains 61.00 ± 0.67% clean accuracy and 32.22 ± 0.21% PGD-20 robust accuracy at IPC 10, and further improves to 73.08 ± 0.18% clean accuracy with 42.89 ± 0.56% PGD-20 and 37.97 ± 0.86% AutoAttack accuracy at IPC 50. On CIFAR-100, **ADM** achieves 35.93% PGD-20 accuracy at IPC 50. In contrast, although several general-purpose distillation methods achieve relatively high clean accuracy at larger IPC budgets, their robust accuracy remains close to zero. Results on Tiny-ImageNet (Figure 6) and ImageNet-100 at IPC 10 (Table 2) further support this observation. Notably, robustness-oriented baselines such as GUARD and ROME provide only limited gains under the same protocol, indicating that explicitly aligning adversarial distributions is critical for preserving robustness after condensation. Further results under larger IPC budgets, together with visualizations of the distilled samples, are provided in the supplementary material.

Cross-architecture robust generalization. To evaluate whether the distilled sets generalize across architectures, we test them on VGG11, MobileNetV2, and ViT under both clean and adversarial

Table 3: Cross-architecture robust generalization on CIFAR-10. The distilled datasets are evaluated on a range of architectures beyond the default ResNet-18. We report Clean Accuracy (Cln.) and Robust Accuracy (Rob.) under PGD-20 with $\varepsilon = 4/255$.

Method	IPC 10						IPC 50					
	VGG11 [34]		MBV2 [32]		ViT [9]		VGG11 [34]		MBV2 [32]		ViT [9]	
	Cln.	Rob.	Cln.	Rob.	Cln.	Rob.	Cln.	Rob.	Cln.	Rob.	Cln.	Rob.
DM [49]	44.65±0.98	9.68±0.54	31.48±1.68	2.47±0.89	26.04±0.18	2.53±0.25	59.06±0.41	6.51±0.40	42.35±1.06	3.27±0.34	34.73±0.21	1.16±0.21
SRe2L [43]	27.69±0.31	1.81±0.30	20.43±0.28	0.00±0.01	20.05±0.48	9.76±1.48	43.18±0.55	0.34±0.10	24.41±1.52	0.08±0.07	21.76±0.54	2.87±0.48
MTT [2]	45.80±1.36	5.07±0.57	31.36±1.01	1.08±0.13	25.65±0.52	0.74±0.15	62.02±0.83	3.11±0.37	45.11±1.45	1.40±0.37	35.81±0.85	0.57±0.28
DANCE [44]	56.54±0.46	1.14±0.12	43.68±1.02	0.59±0.09	34.76±1.04	1.18±0.26	63.39±0.53	0.06±0.10	54.66±1.21	0.01±0.01	48.19±0.49	0.22±0.04
RDED [35]	38.11±0.97	4.39±0.75	23.59±0.96	0.00±0.00	24.38±0.84	1.93±0.46	61.92±0.56	4.23±0.59	38.80±0.93	0.00±0.00	35.82±1.45	1.52±0.15
DELT [33]	41.65±0.86	5.82±0.71	30.29±0.82	0.00±0.01	22.77±0.69	5.78±0.63	62.71±0.88	3.41±0.45	41.14±0.16	0.15±0.08	23.81±1.24	3.22±1.02
NCFM [38]	47.48±0.78	1.25±0.11	36.79±0.75	2.61±0.20	28.84±0.68	1.45±0.30	64.60±0.29	4.29±0.33	50.12±1.75	3.69±0.45	41.26±1.01	0.49±0.08
GUARD [42]	23.57±0.83	4.37±0.42	19.47±1.25	0.00±0.00	16.95±0.53	7.48±2.72	46.42±1.30	5.02±0.25	23.75±1.28	0.49±0.25	20.55±0.64	2.22±0.77
ROME [54]	12.47±1.35	2.88±2.60	10.14±0.58	2.52±1.88	16.89±1.84	3.71±0.70	20.16±1.41	1.44±0.80	15.71±1.33	2.00±0.81	17.30±0.70	2.68±0.50
Ours	**53.97±0.95**	**34.84±0.45**	**33.95±3.64**	**17.90±0.36**	**26.56±1.24**	**22.76±0.96**	**67.50±0.28**	**44.76±0.20**	**58.23±1.22**	**33.25±1.27**	**43.83±0.22**	**33.87±0.26**

Table 4: Effect of teacher smoothness on distillation quality. Results are reported on CIFAR-10 with IPC= 10 using ResNet-18. Smaller $\sigma_{\mathcal{F}}$ and $f(20)$ indicate a smoother teacher.

Variant	$\sigma_{\mathcal{F}} \downarrow$			$f(20) \downarrow$		
	TRADES	AWP	DHAT	TRADES	AWP	DHAT
Standard	0.73	0.89	0.33	1.948	1.986	1.087
Ours	**0.25**	**0.20**	**0.09**	**1.250**	**1.097**	**0.792**

conditions. As shown in Table 3, **ADM** exhibits markedly stronger cross-architecture robustness than all baselines. At IPC 50 on CIFAR-10, **ADM** achieves 44.76±0.20% PGD-20 accuracy on VGG11, 33.25± 1.27% on MobileNetV2, and 33.87 ± 0.26% on ViT, while the best competing methods reach at most 6.51%, 3.69%, and 3.22% on the same architectures, respectively. A similar advantage is already evident at IPC 10. These results suggest that **ADM** does not merely overfit the teacher backbone, but instead captures robust structure that transfers across model families.

Comparison with Trajectory Matching. We analyze standard MTT [2], FTD [10], TESLA [6], and DATM [15] for robustness, and their adversarially adapted variants for computational cost. Detailed results in the supplementary material show that the former transfer little robustness and the latter are substantially less scalable.

5.3 Ablation Study

Effect of teacher smoothing. We first examine whether the proposed weight-perturbation fine-tuning effectively smooths robust teachers. Table 4 reports two smoothness indicators on CIFAR-10 with IPC= 10, namely the empirical input-space loss curvature measure $\sigma_{\mathcal{F}}$ and the loss variation metric $f(20)$. Smaller values indicate a smoother loss landscape, and formal definitions are provided in the supplementary material. Across all three robust teachers, our smoothing stage consistently reduces both quantities by a large margin. These results verify that the proposed randomized weight perturbation produces a consistently smoother teacher landscape, which supports the more stable adversarial distribution transfer observed in the main experiments.

Effect of the number of sampled frequencies. We study how the number of sampled frequency arguments n_f in the characteristic-function matching objective affects distillation quality. Increasing n_f can improve the Harmonic Robustness Score (HRS), a summary metric that captures the trade-off between clean accuracy and robust accuracy (its full definition is provided in the supplementary material). This is expected, as denser spectral sampling better approximates the true distribution discrepancy. Performance saturates around $n_f = 4096$, beyond which further increases yield diminishing returns. We therefore adopt $n_f = 4096$ as the default setting. Complete results are provided in the supplementary material.

6 Discussion

Limitations. While **ADM** efficiently bridges the gap between clean accuracy and robustness, it introduces computational overhead due to the required PGD generation during inner optimization loops. Additionally, as a feature-matching paradigm, the distilled robustness is inherently bounded by the teacher's functional space, which may marginally degrade cross-architecture transferability.

Future Work. Future work will focus on two directions. First, improving the efficiency of adversarial example generation is important for scaling robust distillation to higher-resolution data and larger backbone architectures. Second, it would be valuable to reduce the dependence on a specific robust teacher by developing teacher-agnostic or jointly optimized formulations.

7 Conclusion

We propose Adversarial Distribution Matching (**ADM**), a robust dataset-distillation framework for overcoming sharp adversarial-teacher landscapes and feature-matching objectives that fail to preserve adversarial distributions. **ADM** smooths the teacher through randomized weight perturbations and aligns clean and adversarial feature distributions using characteristic functions in the spectral domain. Experiments show that **ADM** substantially improves adversarial robustness while retaining competitive clean accuracy and strong cross-architecture generalization under severe data compression across multiple datasets, covering a range of evaluation settings as well as different IPC budgets.

Acknowledgments

This work was partly supported by the New Generation Artificial Intelligence-National Science and Technology Major Project under No. 2025ZD0123503, NSFC under No. U2441239, U24A20336, and No. 62502433, the China Postdoctoral Science Foundation under No. 2024M762829, 2025M781523 and 2025M781522, Zhejiang Key Laboratory of Decision Intelligence under No. 2025E10006, Zhejiang Provincial Natural Science Foundation Exploration of China under No. LMS26F020003, State Key Laboratory of Cryptography and Digital Economy Security under No. KFYB2504, the Zhejiang Provincial Natural Science Foundation under No. LD24F020002, and the "Pioneer and Leading Goose" R&D Program of Zhejiang under No. 2025C02033 and 2025C01082.

References

[1] Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. 2020. Square Attack: A Query-Efficient Black-Box Adversarial Attack via Random Search. In *Computer Vision – ECCV 2020 (Lecture Notes in Computer Science, Vol. 12368)*. Springer, 484–501. doi:10.1007/978-3-030-58592-1_29

[2] George Cazenavette, Tongzhou Wang, Antonio Torralba, Alexei A. Efros, and Jun-Yan Zhu. 2022. Dataset Distillation by Matching Training Trajectories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10708–10717. doi:10.1109/CVPR52688.2022.01045

[3] Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. 2021. RobustBench: A Standardized Adversarial Robustness Benchmark. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, Vol. 1. https://openreview.net/forum?id=SSKZPJCt7B

[4] Francesco Croce and Matthias Hein. 2020. Minimally Distorted Adversarial Examples with a Fast Adaptive Boundary Attack. In *Proceedings of the 37th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 119)*. PMLR, 2196–2205. https://proceedings.mlr.press/v119/croce20a.html

[5] Francesco Croce and Matthias Hein. 2020. Reliable Evaluation of Adversarial Robustness with an Ensemble of Diverse Parameter-Free Attacks. In *Proceedings of the 37th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 119)*. PMLR, 2206–2216. https://proceedings.mlr.press/v119/croce20b.html

[6] Justin Cui, Ruochen Wang, Si Si, and Cho-Jui Hsieh. 2023. Scaling Up Dataset Distillation to ImageNet-1K with Constant Memory. In *Proceedings of the 40th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 202)*. PMLR, 6565–6590. https://proceedings.mlr.press/v202/cui23e.html

[7] Wenxiao Deng, Wenbin Li, Tianyu Ding, Lei Wang, Hongguang Zhang, Kuihua Huang, Jing Huo, and Yang Gao. 2024. Exploiting Inter-sample and Inter-feature Relations in Dataset Distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 17057–17066. doi:10.1109/CVPR52733.2024.01614

[8] Yunfeng Diao, Baiqi Wu, Ruixuan Zhang, Ajian Liu, Xiaoshuai Hao, Xingxing Wei, Meng Wang, and He Wang. 2025. TASAR: Transfer-based Attack on Skeletal Action Recognition. In *International Conference on Learning Representations (ICLR)*.

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations (ICLR)*.

[10] Jiawei Du, Yidi Jiang, Vincent Y. F. Tan, Joey Tianyi Zhou, and Haizhou Li. 2023. Minimizing the Accumulated Trajectory Error to Improve Dataset Distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 3749–3758. doi:10.1109/CVPR52729.2023.00365

[11] Logan Engstrom, Andrew Ilyas, and Anish Athalye. 2018. Evaluating and understanding the robustness of adversarial logit pairing. *arXiv preprint arXiv:1807.10272* (2018).

[12] Andrey Feuerverger and Roman A. Mureika. 1977. The Empirical Characteristic Function and Its Applications. *The Annals of Statistics* 5, 1 (1977), 88–97. doi:10.1214/aos/1176343742

[13] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. 2020. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412* (2020).

[14] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and Harnessing Adversarial Examples. *arXiv preprint arXiv:1412.6572* (2014).

[15] Ziyao Guo, Kai Wang, George Cazenavette, Hui Li, Kaipeng Zhang, and Yang You. 2023. Towards lossless dataset distillation via difficulty-aligned trajectory matching. *arXiv preprint arXiv:2310.05773* (2023).

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 770–778. doi:10.1109/CVPR.2016.90

[17] Zhezhi He, Adnan Siraj Rakin, and Deliang Fan. 2019. Parametric Noise Injection: Trainable Randomness to Improve Deep Neural Network Robustness Against Adversarial Attack. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 588–597. doi:10.1109/CVPR.2019.00068

[18] Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. 2018. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407* (2018).

[19] Kaichao Jiang, He Wang, Xiaoshuai Hao, Xiulong Yang, Ajian Liu, Qi Chu, Yunfeng Diao, and Richang Hong. 2026. Your Classifier Can Do More: Towards Balancing the Gaps in Classification, Robustness, and Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 42310–42320.

[20] Gaojie Jin, Xinping Yi, Dengyu Wu, Ronghui Mu, and Xiaowei Huang. 2023. Randomized Adversarial Training via Taylor Expansion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 16447–16457. doi:10.1109/CVPR52729.2023.01578

[21] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling Laws for Neural Language Models. *arXiv preprint arXiv:2001.08361* (2020).

[22] Alex Krizhevsky. 2009. *Learning Multiple Layers of Features from Tiny Images*. Master's thesis. University of Toronto.

[23] Ya Le and Xuan Yang. 2015. Tiny ImageNet Visual Recognition Challenge. CS231n Course Project Report, Stanford University.

[24] Saehyung Lee, Sanghyuk Chun, Sangwon Jung, Sangdoo Yun, and Sungroh Yoon. 2022. Dataset Condensation with Contrastive Signals. In *Proceedings of the 39th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 162)*. PMLR, 12352–12364. https://proceedings.mlr.press/v162/lee22b.html

[25] Shengxi Li, Zeyang Yu, Min Xiang, and Danilo Mandic. 2020. Reciprocal adversarial learning via characteristic functions. *Advances in Neural Information Processing Systems* 33 (2020), 217–228.

[26] Dai Liu, Jindong Gu, Hu Cao, Carsten Trinitis, and Martin Schulz. 2024. Dataset Distillation by Automatic Training Trajectories. In *Computer Vision – ECCV 2024 – 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LXXXVII (Lecture Notes in Computer Science, Vol. 15145)*. Springer, 334–351. doi:10.1007/978-3-031-73021-4_20

[27] Yanqing Liu, Jianyang Gu, Kai Wang, Zheng Zhu, Wei Jiang, and Yang You. 2023. DREAM: Efficient Dataset Distillation by Representative Matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 17268–17278. doi:10.1109/ICCV51070.2023.01588

[28] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083* (2017).

[29] Andrea Rosasco, Antonio Carta, Andrea Cossu, Vincenzo Lomonaco, and Davide Bacciu. 2021. Distilled Replay: Overcoming Forgetting through Synthetic Samples. *arXiv preprint arXiv:2103.15851* (2021).

[30] Kevin Roth, Yannic Kilcher, and Thomas Hofmann. 2019. The Odds Are Odd: A Statistical Test for Detecting Adversarial Examples. In *Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 97)*. PMLR, 5498–5507. https://proceedings.mlr.press/v97/roth19a.html

[31] Ahmad Sajedi, Samir Khaki, Ehsan Amjadian, Lucy Z. Liu, Yuri A. Lawryshyn, and Konstantinos N. Plataniotis. 2023. DataDAM: Efficient Dataset Distillation with Attention Matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 17051–17061. doi:10.1109/ICCV51070.2023.01568

[32] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. 2018. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4510–4520. doi:10.1109/CVPR.2018.00474

[33] Zhiqiang Shen, Ammar Sherif, Zeyuan Yin, and Shitong Shao. 2025. DELT: A Simple Diversity-driven EarlyLate Training for Dataset Distillation. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*. 4797–4806. doi:10.1109/CVPR52734.2025.00452

[34] Karen Simonyan and Andrew Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv preprint arXiv:1409.1556* (2014).

[35] Peng Sun, Bei Shi, Daiwei Yu, and Tao Lin. 2024. On the Diversity and Realism of Distilled Dataset: An Efficient Dataset Distillation Paradigm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 9390–9399. doi:10.1109/CVPR52733.2024.00897

[36] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199* (2013).

[37] Kai Wang, Bo Zhao, Xiangyu Peng, Zheng Zhu, Shuo Yang, Shuo Wang, Guan Huang, Hakan Bilen, Xinchao Wang, and Yang You. 2022. CAFE: Learning to Condense Dataset by Aligning Features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 12186–12195. doi:10.1109/CVPR52688.2022.01188

[38] Shaobo Wang, Yicun Yang, Zhiyuan Liu, Chenghao Sun, Xuming Hu, Conghui He, and Linfeng Zhang. 2025. Dataset Distillation with Neural Characteristic Function: A Minmax Perspective. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 25570–25580. doi:10.1109/CVPR52734.2025.02381

[39] Tongzhou Wang, Jun-Yan Zhu, Antonio Torralba, and Alexei A. Efros. 2018. Dataset Distillation. *arXiv preprint arXiv:1811.10959* (2018).

[40] Dongxian Wu, Shu-Tao Xia, and Yisen Wang. 2020. Adversarial Weight Perturbation Helps Robust Generalization. *Advances in Neural Information Processing Systems* 33 (2020), 2958–2969.

[41] Yifan Wu, Jiawei Du, Ping Liu, Yuewei Lin, Wei Xu, and Wenqing Cheng. 2025. DD-RobustBench: An Adversarial Robustness Benchmark for Dataset Distillation. *IEEE Transactions on Image Processing* 34 (2025), 2052–2066. doi:10.1109/TIP.2025.3553786

[42] Eric Xue, Yijiang Li, Haoyang Liu, Peiran Wang, Yifan Shen, and Haohan Wang. 2025. Towards Adversarially Robust Dataset Distillation by Curvature Regularization. *Proceedings of the AAAI Conference on Artificial Intelligence* 39, 9 (2025), 9041–9049. doi:10.1609/aaai.v39i9.32978

[43] Zeyuan Yin, Eric Xing, and Zhiqiang Shen. 2023. Squeeze, Recover and Relabel: Dataset Condensation at ImageNet Scale From A New Perspective. In *Thirty-seventh Conference on Neural Information Processing Systems*, Vol. 36. 73582–73603. doi:10.52202/075280-3219

[44] Hansong Zhang, Shikun Li, Fanzhao Lin, Weiping Wang, Zhenxing Qian, and Shiming Ge. 2024. DANCE: Dual-view distribution alignment for dataset condensation. *arXiv preprint arXiv:2406.01063* (2024).

[45] Hansong Zhang, Shikun Li, Pengju Wang, Dan Zeng, and Shiming Ge. 2024. M3D: Dataset condensation by minimizing maximum mean discrepancy. *Proceedings of the AAAI Conference on Artificial Intelligence* 38, 8 (2024), 9314–9322. doi:10.1609/aaai.v38i8.28784

[46] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. 2019. Theoretically Principled Trade-off between Robustness and Accuracy. In *Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 97)*. PMLR, 7472–7482. https://proceedings.mlr.press/v97/zhang19p.html

[47] Kejia Zhang, Juanjuan Weng, Shaozi Li, and Zhiming Luo. 2025. Towards Adversarial Robustness via Debiased High-Confidence Logit Alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2783–2792. doi:10.1109/ICCV51701.2025.00267

[48] Bo Zhao and Hakan Bilen. 2021. Dataset Condensation with Differentiable Siamese Augmentation. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*. PMLR, 12674–12685. https://proceedings.mlr.press/v139/zhao21a.html

[49] Bo Zhao and Hakan Bilen. 2023. Dataset Condensation with Distribution Matching. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 6503–6512. doi:10.1109/WACV56688.2023.00645

[50] Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. 2020. Dataset condensation with gradient matching. *arXiv preprint arXiv:2006.05929* (2020).

[51] Ganlong Zhao, Guanbin Li, Yipeng Qin, and Yizhou Yu. 2023. Improved Distribution Matching for Dataset Condensation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 7856–7865. doi:10.1109/CVPR52729.2023.00759

[52] Runkai Zheng, Vishnu Asutosh Dasu, Yinong Oliver Wang, Haohan Wang, and Fernando De la Torre. 2025. Improving Noise Efficiency in Privacy-preserving Dataset Distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 4838–4847. doi:10.1109/ICCV51701.2025.00460

[53] Zheng Zhou, Wenquan Feng, Shuchang Lyu, Guangliang Cheng, Xiaowei Huang, and Qi Zhao. 2024. BEARD: Benchmarking the Adversarial Robustness for Dataset Distillation. *arXiv preprint arXiv:2411.09265* (2024).

[54] Zheng Zhou, Wenquan Feng, Qiaosheng Zhang, Shuchang Lyu, Qi Zhao, and Guangliang Cheng. 2025. ROME is Forged in Adversity: Robust Distilled Datasets via Information Bottleneck. In *Proceedings of the 42nd International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 267)*. PMLR, 78687–78713. https://proceedings.mlr.press/v267/zhou25d.html