DD-Ranking: Rethinking the Evaluation of Dataset Distillation

DD-Ranking Team*
https://github.com/NUS-HPC-AI-Lab/DD-Ranking.git
https://nus-hpc-ai-lab.github.io/DD-Ranking/
https://huggingface.co/spaces/logits/DD-Ranking

Abstract

Dataset distillation has provided a reliable solution for data compression, where models trained on the resulting smaller synthetic datasets achieve performance comparable to those trained on the original datasets. To further improve the performance of synthetic datasets, various training pipelines and optimization objectives have been proposed, greatly advancing the field of dataset distillation. Recent decoupled dataset distillation methods introduce soft labels and stronger data augmentation during the post-evaluation phase and scale dataset distillation up to larger datasets. However, this raises a question: *Is accuracy still a reliable metric to fairly evaluate dataset distillation methods?* Our empirical findings suggest that the performance improvements of these methods often stem from additional techniques rather than the inherent quality of the images themselves, with even randomly sampled images achieving superior results. Such misaligned evaluation settings severely hinder the development of DD. Therefore, we propose DD-Ranking, a unified evaluation framework, along with new general evaluation metrics to uncover the true performance improvements achieved by different methods. By refocusing on the actual information enhancement of distilled datasets, DD-Ranking provides a more comprehensive and fair evaluation standard for future research advancements.

1 Introduction

Training deep models on large datasets has enabled progress in computer vision [19, 9] and natural language processing [8, 1], but requires substantial computation and storage. Dataset distillation (DD) [52] addresses this cost by constructing a small synthetic training set. Advances in training pipelines [7, 64, 58] and optimization objectives [67, 63, 2] have improved the accuracy achievable with such datasets.

However, evaluation recipes have also changed substantially. Some methods jointly optimize images and sample-specific soft labels [17, 29], whereas decoupled methods use epoch-wise labels from pretrained teachers during post-evaluation training [58, 39, 44, 46]. Others adopt stronger augmentation, optimizers, and hyperparameters [41, 4]. Comparisons across these recipes can conflate improvements to the distilled data with improvements to the learning procedure. Indeed, randomly sampled real images can outperform distilled images when given the same teacher supervision and training recipe. This raises our central question: does higher test accuracy reliably indicate better distilled data when evaluation settings differ?

We introduce **DD-Ranking**, an evaluation framework that calibrates performance against matched references. For each method, we compare synthetic data with a same-sized random subset trained under the corresponding recipe, measuring improvement over random (IOR). We also measure the

*See all members in Appendix D. Zhiwei Deng (Google DeepMind) served as an external advisor only.

40th Conference on Neural Information Processing Systems (NeurIPS 2026). Track on Evaluations and Datasets.

hard-label performance gap to a fixed full-dataset reference, termed hard label recovery (HLR). Their combination, the label-robust score (LRS), provides a default ranking while retaining both components for interpretation. An augmentation-robust score (ARS) further combines relative performance with and without augmentation. We evaluate these metrics across datasets, student architectures, and teacher-label configurations.

Our contributions are threefold. First, we establish a matched-baseline protocol and explicit ranking rule for comparing DD methods with heterogeneous label and augmentation recipes. Second, our benchmark shows that many reported accuracy gains do not translate into improvements over equally equipped random subsets, while identifying methods that retain substantial gains. Third, we introduce complementary label- and augmentation-robust metrics and analyze their empirical ranking stability, encouraging progress toward more informative distilled datasets.

2 Motivation

2.1 Overview of Unfairness

Dataset distillation is conventionally evaluated by the **test accuracy** of a model trained on the distilled dataset[2]. Table 1 surveys the evaluation recipes of representative methods [67, 62, 2, 66, 37, 17, 58, 46, 57, 45, 41, 39]. The differences span training duration, batch size, optimizer, loss, teacher, and augmentation. Direct accuracy comparisons across these configurations mix data quality with training effects; we examine label representation and augmentation in particular.

Config	DC	DSA	DM	MTT	DataDAM	DATM	SRe2L	RDED	CDA	DWA	D4M	EDC	G-VBSM
Epochs	1K	1K	1K	1K	1K	1K	300	300	300	300	300	300	300
Batch Size	256	256	256	256	256	256	1024	100	128	128	1024	100	1024
Optimizer	SGD	SGD	SGD	SGD	SGD	SGD	AdamW	AdamW	AdamW	AdamW	AdamW	AdamW	AdamW
LR Scheduler	step	step	step	step	step	step	cosine	cosine	cosine	cosine	cosine	cosine	cosine
Label Type	hard	hard	hard	hard	hard	soft	soft	soft	soft	soft	soft	soft	soft
Soft Label	-	-	-	-	-	single	multiple	multiple	multiple	multiple	multiple	multiple	multiple
Loss Function	CE	CE	CE	CE	CE	SCE	KL	KL	KL	KL	KL	MSE	MSE
Teacher Model	-	-	-	-	-	single	single	single	single	single	single	ensemble	ensemble
DSA	No	Yes	Yes	Yes	Yes	Yes	No	No	No	No	No	No	No
ZCA	No	No	No	Yes	No	Yes	No	No	No	No	No	No	No
ResizeCrop	No	No	No	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
CropRange	-	-	-	-	-	-	0.08, 1.0	0.5, 1.0	0.08, 1.0	0.08, 1.0	0.08, 1.0	0.5, 1.0	0.08, 1.0
PatchShuffle	No	No	No	No	No	No	No	Yes	No	No	No	Yes	No
CutMix	No	No	No	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Table 1: Evaluation configurations of various dataset distillation methods. We separate evaluation model training hyperparameters (top) and data augmentation (bottom). For each row, different colors highlight the differences in the evaluation setting.

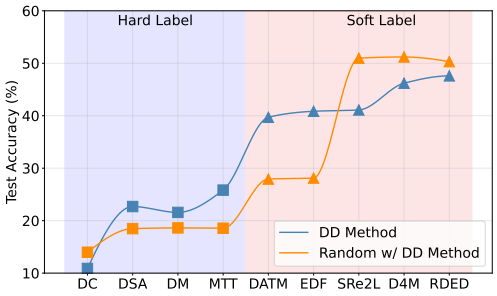
A random baseline also depends on the evaluation recipe. On CIFAR-10 at IPC1 with hard labels, the test accuracy of a same-sized random subset ranges from 8.8% to 16.4% across the evaluation configurations of different methods. Thus, even without changing the subset size or label type, the evaluation recipe is associated with a substantial variation in baseline accuracy. This motivates pairing each distilled dataset with random data evaluated under its corresponding recipe rather than using a single random-baseline accuracy across methods.

2.2 Soft Labels

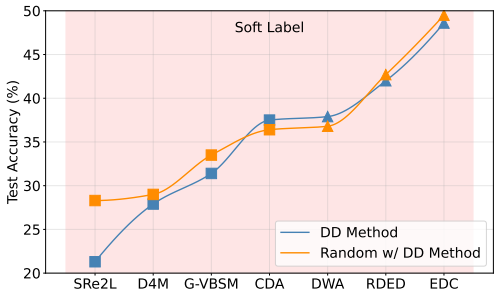
Soft labels improve test accuracy. Soft labels specify distributions over classes and can transfer information from a pretrained teacher. DATM [17] jointly optimizes labels and images, whereas SRe2L [58] supplies teacher-generated labels during evaluation training. Supervision against these labels, for example through KL divergence, improves test accuracy (Figure 1).

Teacher supervision can mask differences in image quality. We compare distilled data with random noise and randomly selected real samples under matched teacher and training recipes. Random real data with soft labels outperforms distilled data in many settings (Figure 1), consistent with [35].

[2]We focus on image classification, the most common DD evaluation task.



(a) TinyImageNet IPC50



(b) ImageNet1K IPC10

Figure 1: Test accuracies of the model trained on synthetic data distilled by various DD methods and on randomly selected data. Despite soft labels being able to significantly improve the test accuracy, DD methods may fail to outperform random selection under the same training setting.

Batch size	128	1024
ImageNet1K	0.5	0.3

(a) With soft labels, random noise exceeds chance accuracy (0.1%).

Noise	$\sigma = 0.05$	$\sigma = 0.5$
TinyImageNet	49.9 (↑ 26.0)	14.6 (↑ 3.9)
ImageNet1K	60.6 (↑ 24.9)	6.8 (↑ 5.6)

(b) Soft labels improve accuracy on noise-corrupted images; gains over hard labels appear in parentheses.

Method	EDC		RDED	
IPC	10	50	10	50
w/ aug.	48.6	58.0	42.0	56.5
w/o aug.	12.5	39.7	15.3	27.9

(c) Data augmentation contributes substantially to high accuracy, especially on high-resolution datasets.

Table 2: (a) Soft-label accuracy on random noise. (b) Soft-label accuracy on noise-corrupted images. (c) Augmentation ablation on ImageNet1K.

Even random noise attains above-chance ImageNet1K accuracy when given soft labels (Table 2a). Similarly, soft labels substantially improve accuracy on noise-corrupted images (Table 2b). These controls show that teacher supervision contributes information independently of any image-distillation improvement, motivating a matched random baseline.

2.3 Data Augmentation

Different methods also use different augmentation recipes (Table 1). To assess their contribution, we evaluate EDC and RDED with and without their augmentation on ImageNet1K (Table 2c). At IPC10, EDC changes from 48.6% to 12.5% without augmentation, while RDED changes from 42.0% to 15.3%. This dependence on augmentation motivates evaluating relative improvement over random data in both conditions, rather than attributing every accuracy gain to distilled-image quality.

3 DD-Ranking

3.1 Overview

Motivated by the unfairness above, we introduce DD-Ranking. DD-Ranking is an integrated and easy-to-use evaluation benchmark for dataset distillation (DD). It aims to provide a fair evaluation scheme for DD methods that can decouple the impacts from knowledge distillation and data augmentation to reflect the real informativeness of the distilled data. Under the finding that the test accuracy no longer fits the need for fair and comprehensive evaluation, we design new metrics for both the label representation and data augmentation. We also provide some theoretical intuition in Appendix A.

Protocol and ranking rule. DD-Ranking comprises an evaluation protocol, a metric suite (HLR, IOR, LRS, and ARS), a ranking rule, and an evaluation toolkit. Within each dataset–IPC setting, our default ranking orders methods by decreasing LRS at $\lambda = 0.5$. We report the triple $(\mathrm{HLR}, \mathrm{IOR}, \mathrm{LRS}_{\lambda=0.5})$ so that the components remain visible, and report ARS as a separate evaluation axis with $\gamma = 0.5$. These weights specify an explicit trade-off; they are not tuned to favor a method. The protocol

matches each distilled dataset to a random subset under the corresponding evaluation recipe, as detailed in Section 4.1.

3.2 Label-Robust Score

Hard label recovery. The initial goal of dataset distillation is to synthesize a small number of data points that do not need to come from the correct data distribution, but will, when given to the learning algorithm as training data, approximate the model trained on the original data [52]. Given that almost all existing classification datasets use hard label annotation, we think it is crucial for DD methods to maintain good performance with hard labels. To this end, we propose the **hard label recovery (HLR)**. Specifically, for both hard-label-based and soft-label-based methods, we evaluate the test accuracy of the distilled data and that of the original dataset with hard labeling, denoted as $\text{acc}_{\text{syn-hard}}$ and $\text{acc}_{\text{real-hard}}$, respectively. The hard label recovery is computed by taking the difference:

$$\text{HLR} = \text{acc}_{\text{real-hard}} - \text{acc}_{\text{syn-hard}} \tag{1}$$

A smaller HLR indicates a smaller gap to the fixed full-dataset reference specified in Section 4.1. For methods that synthesize data per class, hard-label evaluation uses the class identity associated with each synthesized sample rather than teacher-generated soft labels.

Improvement over random. Despite the popularity of soft labels, it's not fair to directly compare methods with soft labels against those with hard labels. Also, there isn't a unified recipe for soft-label-based evaluation, and differences such as soft labels per sample, loss function, and temperature could significantly impact the results. This makes it difficult to compare different soft-label-based methods. To make different methods comparable under mixed label types, we propose **improvement over random (IOR)**. This metric is based on the common sense that any DD method should at least outperform random selection under the same evaluation recipe, and we use the relative performance improvements over random selection to compare any pair of DD methods. Specifically, denote the test accuracies of models trained on synthetic data and a same-sized random subset with any label type as $\text{acc}_{\text{syn-any}}$ and $\text{acc}_{\text{rdm-any}}$, respectively. For each DD method, we match the evaluation model, preprocessing, label-generation procedure, loss function, augmentation, optimizer, and training schedule between synthetic and random data. Learning rates are selected separately as described in Section 4.1; matching the recipe does not require the selected learning rates to be identical. Then, the IOR is computed by:

$$\text{IOR} = \text{acc}_{\text{syn-any}} - \text{acc}_{\text{rdm-any}} \tag{2}$$

IOR is positively related to the performance of DD methods. By doing so, we can effectively disentangle the improvement brought solely by knowledge distillation and reflect the true informativeness.

Label-robust score. Combining hard label recovery (HLR) and improvement over random (IOR), we present the label-robust score (LRS). LRS first takes a weighted sum α of HLR and IOR via a weight parameter $\lambda \in [0, 1]$ (default 0.5) as follows:

$$\alpha = \lambda\text{IOR} - (1 - \lambda)\text{HLR} \tag{3}$$

We assign a negative mark to HLR so that both parts of the sum are positively correlated with the performance. The raw range of α is $[-1, 1]$, so we normalize LRS to the range $[0, 1]$ by letting $\text{LRS} = 100\% \times (e^{\alpha} - e^{-1})/(e - e^{-1})$. A higher LRS indicates that the distilled dataset of the corresponding method is more robust to the label representation and has richer information. Because this normalization is strictly increasing, it preserves the ordering induced by α for a fixed λ; any strictly increasing alternative preserves the same ranking. Its endpoints are fixed by the score domain rather than the methods currently evaluated, and its increasing slope gives higher resolution toward larger raw scores.

3.3 Augmentation-Robust Score

Data augmentation, as a trick to enhance model training, doesn't reveal the quality of the dataset itself. Thus, the improvement in test accuracy brought merely by data augmentation during evaluation training should not be attributed to the effectiveness of the dataset distillation method. To disentangle data augmentation's impact, we introduce the **augmentation-robust score (ARS)** which continues to leverage the relative improvement over randomly selected data. Specifically, we first evaluate synthetic data and a randomly selected subset under the same setting to obtain $\text{acc}_{\text{syn-aug}}$ and $\text{acc}_{\text{rdm-aug}}$

(same as IOR). Next, we evaluate both synthetic data and random data again without the data augmentation, and results are denoted as $\text{acc}_{\text{syn-naug}}$ and $\text{acc}_{\text{rdm-naug}}$. Both augmentation conditions retain the method's native label protocol, matched between synthetic and random data within each condition. The parameter $\gamma \in [0, 1]$ weights the two conditions and is set to 0.5 by default.

We claim that an informative subset via distillation should surpass any randomly selected subset of the same size, regardless of the use of data augmentation. Thus, both differences, $\text{acc}_{\text{syn-aug}} - \text{acc}_{\text{rdm-aug}}$ and $\text{acc}_{\text{syn-naug}} - \text{acc}_{\text{rdm-naug}}$, are positively correlated with the real informativeness of the distilled dataset. We take a weighted sum of the two differences

$$\beta = \gamma(\text{acc}_{\text{syn-aug}} - \text{acc}_{\text{rdm-aug}}) + (1 - \gamma)(\text{acc}_{\text{syn-naug}} - \text{acc}_{\text{rdm-naug}}) \tag{4}$$

and use a similar normalization method to compute ARS. A higher ARS indicates that the distilled dataset of the corresponding method is more robust to data augmentation.

4 Results

4.1 Evaluation Settings

Baseline. We evaluate a wide range of representative works in dataset distillation. For hard-label methods, we evaluate DC [67], DSA [62], MTT [2], DM [63], and DataDAM [37]. For soft-label methods, we evaluate SRe2L [58], DATM [17], EDF [48], DWA [13], RDED [46], CDA [57], EDC [41], and G-VBSM [40]. In the case where the method provides its distilled data, we adopt it directly. In the case where the distilled data is absent, we strictly follow their implementation provided in both the paper and code repo to replicate their results. The ImageNet1K comparison focuses on methods that can be evaluated at this scale within the computational budget; this selection is based on computational feasibility rather than label type. Kernel-based baselines include RFAD [30] on CIFAR-10 and FRePo [74] on CIFAR-100 and TinyImageNet; Table 5 reports their selected numerical results. Additional evaluations cover RaT-BPTT [15] on CIFAR-10 and CIFAR-100, and DataDAM on TinyImageNet (Appendix B.3).

Dataset. We report DD-Ranking benchmarking results on the four existing datasets: CIFAR-10 [20], CIFAR-100 [20], TinyImageNet [22], and ImageNet1K [36]. The resolution of images in CIFAR-10 and CIFAR-100 is 32×32. The resolution of images in TinyImageNet is 64×64. The resolution of images in ImageNet1K is 224×224.

Model. For each baseline method, we use the reported model for evaluation. This includes ConvNets of depths 3 and 4 with instance normalization, ConvNets of depths 3 and 4 with batch normalization, and ResNet-18 [18]. We also incorporate AlexNet [21], ResNet-50, VGG-11 [43], Swin-T [28], and ViT-base [10] to validate the robustness of DD-Ranking on different architectures.

DD-Ranking evaluation. We report means over **five random seeds**; available standard deviations appear as lower-right subscripts in the main HLR/IOR/LRS tables. Entries without subscripts have no reported standard deviation. For each seed, the random baseline is a class-balanced subset at the same IPC, resampled from original training images without merging them. It shares the synthetic arm's model, preprocessing, teacher/label protocol, loss, augmentation, optimizer, and schedule. Hard-label evaluation searches the learning rate separately for both arms; soft-label evaluation retains the method's supplied learning rate for synthetic data and searches it for random data. The full dataset supplies a fixed reference per dataset under a unified recipe with only the learning rate swept; this reference need not maximize accuracy. Appendix B.1 details the complete protocol in Table 10; Table 1 instead surveys differences between published methods.

4.2 Label-Robust Score

CIFAR-10. Table 3 compares methods at IPC1, IPC10, and IPC50. MTT has the highest LRS among the hard-label methods at all three budgets, while DATM leads the methods in this table with LRS 24.6, 28.7, and 28.0. This contrasts with SRe2L, whose IOR is negative throughout, and RDED, whose IOR becomes negative at IPC50. Matching methods therefore remain strong baselines after calibration for the evaluation recipe. In the additional evaluation, RaT-BPTT reaches LRS 25.3 at IPC1, exceeding the best corresponding result in this table (Appendix B.3).

CIFAR-100. Table 4 shows a similar contrast between jointly optimized labels and teacher relabeling. DATM leads at IPC1 and IPC50, while MTT leads at IPC10 with LRS 27.5. At IPC50, DATM

IPC	1			10			50		
Metric	HLR↓	IOR↑	LRS↑	HLR↓	IOR↑	LRS↑	HLR↓	IOR↑	LRS↑
DC	$52.7_{0.2}$	$12.4_{0.2}$	$19.1_{0.3}$	$36.7_{0.3}$	$18.5_{0.2}$	$23.2_{0.3}$	$26.3_{0.2}$	$12.3_{0.4}$	$24.0_{0.3}$
DSA	$58.9_{0.3}$	$13.2_{0.3}$	$18.2_{0.2}$	$35.1_{0.4}$	$19.6_{0.2}$	$23.7_{0.2}$	$27.4_{0.5}$	$11.0_{0.4}$	$23.5_{0.4}$
MTT	$42.2_{0.5}$	$27.6_{0.7}$	$23.9_{0.6}$	$\mathbf{23.7}_{0.6}$	$30.9_{0.8}$	$28.4_{0.8}$	$\mathbf{16.5}_{0.5}$	$20.5_{0.4}$	$27.8_{0.5}$
DM	$61.4_{0.7}$	$8.7_{0.6}$	$17.0_{0.5}$	$39.4_{0.8}$	$16.1_{0.9}$	$22.2_{1.0}$	$25.1_{0.7}$	$12.7_{0.7}$	$24.3_{0.7}$
DataDAM	$49.9_{0.8}$	$15.6_{0.5}$	$20.0_{0.6}$	$34.8_{0.6}$	$19.9_{0.6}$	$23.8_{0.5}$	$21.9_{0.7}$	$15.8_{0.5}$	$25.6_{0.6}$
DATM	$\mathbf{41.9}_{0.7}$	$\mathbf{30.8}_{0.4}$	$\mathbf{24.6}_{0.8}$	$26.8_{0.5}$	$\mathbf{35.1}_{0.7}$	$\mathbf{28.7}_{0.6}$	$18.9_{0.3}$	$\mathbf{23.9}_{0.7}$	$\mathbf{28.0}_{0.5}$
SRe2L	$69.9_{0.5}$	$-0.3_{0.6}$	$14.3_{0.6}$	$67.8_{0.8}$	$-5.7_{0.8}$	$13.8_{0.6}$	$62.9_{0.5}$	$-6.5_{0.8}$	$14.4_{0.7}$
RDED	$60.6_{0.7}$	$2.4_{0.9}$	$16.2_{0.7}$	$50.7_{0.8}$	$1.1_{0.7}$	$17.6_{0.8}$	$36.0_{0.9}$	$-1.6_{1.2}$	$19.6_{0.9}$
D4M	$51.1_{0.8}$	$6.7_{0.8}$	$18.4_{0.6}$	$39.9_{0.7}$	$9.1_{0.9}$	$20.8_{0.9}$	$27.0_{0.8}$	$6.6_{1.0}$	$22.8_{0.9}$

Table 3: HLR, IOR, and LRS on CIFAR-10. Subscripts denote standard deviations over five random seeds; $\lambda = 0.5$ here and in the following tables. Colors distinguish hard- and soft-label methods.

exceeds random selection by 26.1 percentage points, whereas SRe2L and RDED have negative IOR. DataDAM and MTT both attain LRS 26.8 at this budget, illustrating the competitiveness of hard-label matching methods. RaT-BPTT attains LRS 28.9 at IPC10 in the additional evaluation, outperforming the methods listed here.

IPC	1			10			50		
Metric	HLR↓	IOR↑	LRS↑	HLR↓	IOR↑	LRS↑	HLR↓	IOR↑	LRS↑
DC	$39.4_{0.4}$	$8.4_{0.3}$	$20.8_{0.3}$	$25.5_{0.3}$	$12.7_{0.5}$	$24.2_{0.4}$	$21.8_{0.2}$	$1.1_{0.6}$	$22.7_{0.7}$
DSA	$46.0_{0.5}$	$8.5_{0.5}$	$19.6_{0.5}$	$26.1_{0.4}$	$13.5_{0.4}$	$24.3_{0.3}$	$21.2_{0.4}$	$2.0_{0.5}$	$23.0_{0.4}$
MTT	$35.2_{0.5}$	$16.7_{0.6}$	$23.1_{0.5}$	$\mathbf{18.0}_{0.7}$	$\mathbf{20.7}_{0.7}$	$\mathbf{27.5}_{0.6}$	$12.1_{0.5}$	$11.6_{0.6}$	$26.8_{0.6}$
DM	$48.0_{0.6}$	$6.1_{0.8}$	$18.9_{0.7}$	$30.1_{0.9}$	$10.7_{0.9}$	$23.0_{0.9}$	$16.6_{0.7}$	$7.2_{0.7}$	$24.9_{0.5}$
DataDAM	$45.2_{0.6}$	$9.1_{0.7}$	$19.9_{0.8}$	$25.9_{0.5}$	$14.8_{0.8}$	$24.6_{0.7}$	$12.4_{0.7}$	$11.8_{0.6}$	$26.8_{0.6}$
DATM	$\mathbf{24.1}_{0.7}$	$\mathbf{18.5}_{0.5}$	$\mathbf{25.7}_{0.6}$	$18.9_{0.6}$	$18.4_{0.6}$	$26.8_{0.6}$	$\mathbf{10.3}_{0.5}$	$\mathbf{26.1}_{0.8}$	$\mathbf{30.4}_{0.7}$
SRe2L	$52.7_{1.1}$	$-1.9_{0.9}$	$16.7_{0.9}$	$50.5_{0.8}$	$-14.8_{0.7}$	$15.0_{0.7}$	$46.2_{0.5}$	$-11.5_{0.9}$	$16.2_{0.7}$
RDED	$45.6_{0.7}$	$-0.5_{1.0}$	$18.1_{0.8}$	$37.5_{0.6}$	$-1.2_{0.9}$	$19.4_{0.6}$	$27.3_{0.8}$	$-1.5_{1.1}$	$21.2_{0.9}$
D4M	$30.9_{0.5}$	$10.0_{0.6}$	$22.7_{0.4}$	$40.1_{0.7}$	$9.7_{0.8}$	$20.9_{0.6}$	$26.7_{0.8}$	$13.5_{0.9}$	$24.2_{0.9}$

Table 4: HLR, IOR, and LRS on CIFAR-100, with standard deviations as subscripts. DATM performs strongly, while MTT leads at IPC10 among the methods shown.

Kernel-based methods. Table 5 extends the comparison to RFAD and FRePo. At IPC50, RFAD exceeds random selection by 3.51 percentage points on CIFAR-10, and FRePo by 3.81 points on CIFAR-100. Both retain positive gains, but their margins are smaller than MTT's 20.5 and 11.6 points on the respective datasets. Thus, the calibrated comparison distinguishes the relatively small gains of these kernel-based baselines from stronger matching-based results at this budget.

Method	Dataset	IPC	IOR↑
RFAD	CIFAR-10	50	3.51
FRePo	CIFAR-100	50	3.81

Table 5: Improvement over random for kernel-based methods at IPC50, in percentage points.

TinyImageNet. In Table 6, DATM and EDF share the highest LRS at IPC1; MTT leads at IPC10, and EDF leads at IPC50. At IPC50, D4M's IOR reaches 12.8 percentage points, while RDED's is -0.6; at IPC1 and IPC10, RDED has higher LRS than D4M. Thus, the relative performance of scalable approaches depends on the data budget. The strong results of DATM and EDF also show that soft labels can contribute to effective distilled datasets when jointly optimized with the samples.

Results on ImageNet1K. All methods in Table 7 use soft labels in their native evaluation. With the same label protocol applied to random data, every method has negative IOR at IPC10 and IPC50. At IPC10, for example, SRe2L has IOR -15.6, whereas RDED and EDC are closer to random selection at -0.6 and -0.4 percentage points. The large HLR values also reveal weak recovery under hard

IPC	1			10			50		
Metric	HLR↓	IOR↑	LRS↑	HLR↓	IOR↑	LRS↑	HLR↓	IOR↑	LRS↑
DC	$28.6_{0.4}$	$3.9_{0.3}$	$22.0_{0.3}$	$21.5_{0.2}$	$7.1_{0.5}$	$23.9_{0.5}$	$21.3_{0.4}$	$-2.1_{0.6}$	$22.2_{0.4}$
DSA	$30.3_{0.5}$	$3.7_{0.4}$	$21.6_{0.6}$	$20.3_{0.3}$	$6.8_{0.7}$	$24.1_{0.4}$	$17.6_{0.5}$	$7.6_{0.5}$	$25.8_{0.4}$
MTT	$30.7_{0.5}$	$5.8_{0.6}$	$21.9_{0.6}$	$\mathbf{15.6}_{0.4}$	$14.6_{0.7}$	$\mathbf{26.7}_{0.6}$	$15.6_{0.3}$	$10.2_{0.8}$	$26.4_{0.6}$
DM	$36.7_{0.3}$	$2.3_{0.7}$	$20.2_{0.5}$	$26.2_{0.8}$	$7.5_{0.9}$	$23.1_{0.8}$	$18.9_{0.5}$	$5.3_{0.8}$	$24.1_{0.7}$
DATM	$\mathbf{25.4}_{0.5}$	$8.6_{0.4}$	$\mathbf{23.5}_{0.4}$	$18.3_{0.3}$	$14.2_{0.5}$	$26.0_{0.6}$	$\mathbf{13.5}_{0.4}$	$15.1_{0.9}$	$27.2_{0.6}$
EDF	25.8	**9.2**	**23.5**	18.5	**15.4**	26.2	13.8	15.9	**27.3**
SRe2L	$45.6_{0.7}$	$-1.8_{0.7}$	$15.4_{0.7}$	$43.6_{0.5}$	$-8.5_{0.4}$	$17.1_{0.5}$	$33.6_{0.4}$	$-9.6_{0.8}$	$18.6_{0.8}$
RDED	$34.0_{0.6}$	$3.9_{0.8}$	$21.0_{0.7}$	$25.6_{0.5}$	$1.8_{0.9}$	$23.7_{0.7}$	$15.2_{0.7}$	$-0.6_{1.0}$	$23.7_{0.9}$
D4M	$40.6_{0.6}$	$-3.0_{0.7}$	$18.6_{0.6}$	$35.6_{0.7}$	$-5.8_{0.8}$	$18.9_{0.6}$	$27.7_{0.9}$	$12.8_{1.1}$	$23.8_{0.8}$

Table 6: HLR, IOR, and LRS on TinyImageNet, with standard deviations as subscripts. EDF standard deviations were not reported, so its entries show means only.

labels, despite the high native soft-label accuracies. These results expose the gap between raw test accuracy and improvement attributable to the selected or synthesized images.

IPC	1			10			50		
Metric	HLR↓	IOR↑	LRS↑	HLR↓	IOR↑	LRS↑	HLR↓	IOR↑	LRS↑
SRe2L	$56.3_{0.6}$	$-1.5_{1.1}$	$16.2_{0.8}$	$55.0_{0.5}$	$-15.6_{0.8}$	$14.2_{0.9}$	$53.4_{0.7}$	$-13.2_{1.0}$	$14.8_{0.7}$
RDED	$\mathbf{55.7}_{0.5}$	$\mathbf{1.6}_{0.7}$	$\mathbf{16.8}_{0.6}$	$\mathbf{50.2}_{0.4}$	$-0.6_{0.9}$	$\mathbf{17.4}_{0.8}$	$\mathbf{39.8}_{0.7}$	$-3.6_{0.7}$	$\mathbf{22.9}_{0.7}$
D4M	$55.9_{0.8}$	$-0.6_{0.6}$	$15.6_{0.6}$	$53.0_{0.5}$	$-7.7_{1.1}$	$15.8_{0.9}$	$43.7_{0.6}$	$-5.8_{0.8}$	$17.6_{0.5}$
DWA	$56.1_{0.8}$	$-1.2_{1.2}$	$16.3_{1.1}$	$54.4_{0.9}$	$-4.1_{1.1}$	$16.1_{1.0}$	$49.7_{0.7}$	$-7.8_{0.8}$	$16.3_{0.9}$
CDA	$56.2_{0.6}$	$-2.5_{0.6}$	$16.1_{0.8}$	$54.9_{0.7}$	$-8.6_{0.8}$	$15.3_{0.7}$	$52.0_{0.4}$	$-6.7_{0.8}$	$16.1_{0.6}$
EDC	$55.7_{0.3}$	$-0.8_{0.8}$	$16.4_{0.4}$	$52.0_{0.5}$	$\mathbf{-0.4}_{0.4}$	$17.1_{0.5}$	$41.3_{0.5}$	$\mathbf{-0.1}_{1.1}$	$18.9_{0.9}$
G-VBSM	$56.3_{0.6}$	$-1.2_{1.2}$	$16.3_{1.0}$	$55.0_{0.5}$	$-7.3_{1.3}$	$15.5_{0.9}$	$44.9_{0.6}$	$-5.9_{0.9}$	$17.4_{0.8}$

Table 7: HLR, IOR, and LRS on ImageNet1K, with standard deviations as subscripts. Most methods have negative IOR and large gaps to the hard-label reference.

Findings. Based on these results, we identify three key insights.

i) Test accuracy is not a reliable metric when soft labels are employed. Soft labels demonstrate even higher effectiveness on random data. Notably, on TinyImageNet and ImageNet1K, classifiers trained on random data with soft labels consistently **outperform** those trained on DD-synthesized data. While DATM maintains an advantage over random selection on TinyImageNet, this improvement diminishes substantially when soft labels are applied to random data. This observation reinforces our claim that accuracy improvements with soft labels primarily stem from knowledge distillation rather than the intrinsic informativeness of synthetic data.

ii) Soft labels enhance synthetic dataset informativeness when jointly optimized. Among soft-label-based methods, DATM and EDF employ a distinct approach by assigning unique soft labels to each sample and jointly optimizing both samples and labels during distillation. Unlike generative and decoupled methods that generate soft labels during evaluation training, these optimized soft labels improve synthetic data quality, as evidenced by superior LRS scores. This demonstrates that integrating soft labels into the training process can meaningfully enhance synthetic data quality.

iii) Matching and backpropagation-through-time methods provide strong baselines. Matching-based methods (encompassing gradient, trajectory, and feature matching) remain competitive despite computational constraints on their scalability. In the additional CIFAR-10 and CIFAR-100 evaluations, RaT-BPTT maintains an IOR above 23 percentage points across the evaluated IPC settings, including IPC50. The kernel-based comparison in Table 5 shows smaller positive gains at IPC50. These results show that DD-Ranking can distinguish methods with sustained improvements over random selection from those with smaller gains. Among the scalable methods, RDED and D4M are competitive in several settings, suggesting the relevance of realistic image content.

4.3 Augmentation-Robust Score

Table 8 presents ARS performance metrics for various DD methods applied to ImageNet1K, including IOR results with and without data augmentation as introduced in Section 3.3. Most existing decoupled and generative DD methods fail to surpass random selection regardless of augmentation status. Without data augmentation, the performance disparity between DD methods and random selection widens as IPC increases. These findings demonstrate that contemporary DD approaches, despite their heavy reliance on data augmentation strategies, frequently underperform when these same augmentation techniques are applied to simple random selection. Notably, when augmentation is excluded from evaluation, the performance gap between certain DD methodologies and random selection becomes more pronounced, further supporting our assertion that conventional test accuracy metrics no longer serve as an equitable evaluation criterion in this domain.

IPC	1			10			50		
Metric	IOR w/o aug↑	IOR w/ aug ↑	ARS↑	IOR w/o aug↑	IOR w/ aug ↑	ARS↑	IOR w/o aug↑	IOR w/ aug ↑	ARS↑
SRe2L	-1.2	-1.5	26.3	-4.4	-15.6	22.9	-21.0	-13.2	20.2
RDED	0.8	**1.6**	27.4	5.6	-0.6	28.1	2.0	-3.6	26.7
D4M	-0.3	-0.6	26.7	-0.5	-7.7	25.2	-2.0	-5.8	25.3
DWA	-1.2	-1.2	26.4	-4.0	-4.1	25.2	-13.0	-7.8	22.7
CDA	-1.1	-2.5	26.1	-4.9	-8.6	24.1	-14.1	-6.7	22.7
EDC	-0.5	-0.8	26.6	-0.3	**-0.4**	26.8	-3.2	**-0.1**	26.2
G-VBSM	-1.2	-1.2	26.4	-7.9	-7.3	23.8	-18.0	-5.9	22.1

Table 8: Augmentation-robust score (ARS) evaluation results on ImageNet1K. We report both IOR w/ aug ($\text{acc}_{\text{syn-aug}} - \text{acc}_{\text{rdm-aug}}$) and IOR w/o aug ($\text{acc}_{\text{syn-naug}} - \text{acc}_{\text{rdm-naug}}$). γ is 0.5 by default.

4.4 Analysis

Sensitivity to the ranking weight. We compare all 10 pairs of rankings induced by $\lambda \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ within each dataset–IPC setting, retaining ties in the rounded LRS values. The equally weighted mean across the 12 settings (120 comparisons) is Kendall $\tau_b = 0.9159 \approx 0.92$; the setting-level averages range from 0.7757 to 0.9667 (Table 9). Agreement is lowest on TinyImageNet at IPC50, so the ranking is stable on average rather than invariant to the weight. This analysis covers methods with complete weight sweeps; Appendix B.4 gives the formula and scope.

Dataset	IPC1	IPC10	IPC50
CIFAR-10	0.8937	0.9667	0.9667
CIFAR-100	0.9386	0.9519	0.8721
TinyImageNet	0.9021	0.9449	0.7757
ImageNet1K	0.9533	0.9020	0.9229

Table 9: Sensitivity to λ: each cell averages Kendall τ_b over 10 weight pairs, using rounded LRS values with ties retained. The mean over all 12 settings is 0.9159.

ARS across model architectures. On ImageNet1K at IPC10, we additionally evaluate ARS using ResNet-18, ResNet-50, and ViT. The per-method fluctuation across these architectures is at most 0.4 ARS points, compared with an approximately 2.9-point separation between methods, and the method ordering is preserved under all three architectures. This supports the architecture stability of ARS in this setting, complementing the LRS analysis below.

Robust to model architecture. Cross-architecture evaluation is an important experiment for dataset distillation methods. Specifically, different model architectures are used to evaluate the synthetic data. Despite variations in raw test accuracy across model architectures, the metric used to evaluate dataset distillation performance should remain consistent, with minimal fluctuation in metric values. As shown in Figure 2, the LRS results for six methods across different settings, each tested with four distinct model architectures, demonstrate high consistency. This consistency validates the robustness of our benchmark across different model architectures.

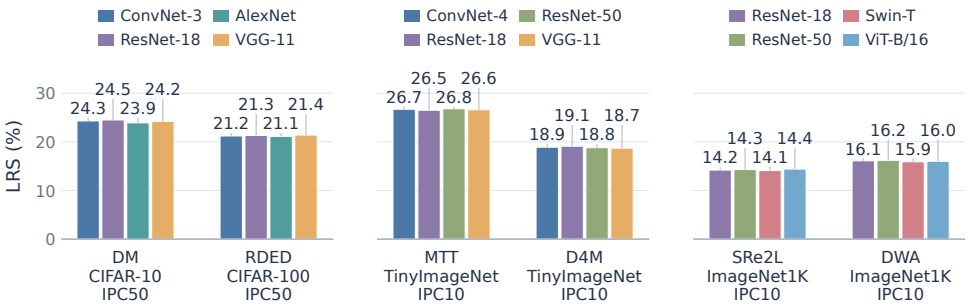


Figure 2: LRS across student architectures. Each group evaluates the same distilled dataset with four architectures; colors identify the evaluation models. Bars start at zero.

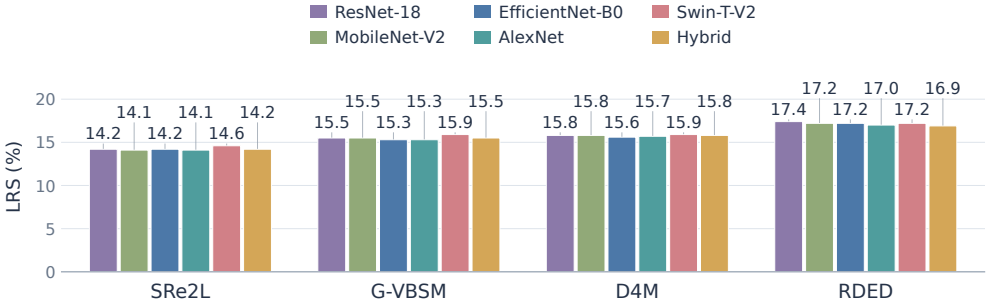


Figure 3: LRS across teacher and soft-label sources. Colors identify the label source; Hybrid combines soft labels from multiple teachers. Bars start at zero.

Robust to soft labels. In decoupled dataset distillation [58, 39, 44, 46], epoch-wise soft labels constitute a crucial component of the synthetic dataset. Recent studies [39, 41, 4] have explored improving test accuracy by leveraging stronger teacher models to provide soft labels without altering the synthetic data itself. However, the validity of this technique remains insufficiently investigated. As shown in Figure 3, whether through the use of different teacher models or advanced hybrid soft label strategies by fusing soft labels generated by multiple teachers, our proposed LRS consistently exhibits strong robustness, thereby validating its reliability across diverse soft label settings.

5 Related Work

Hard-label-based dataset distillation methods. Hard-label-based DD methods assign categorical labels to synthetic samples, matching real dataset labels. Matching-based methods are the main hard-label approaches: i) Gradient matching: Optimizes synthetic data to match gradients induced on neural networks. Following DC [67], improvements include DSA [62], DCC [23], and LCMat [42]. ii) Trajectory matching: Aligns training dynamics between models trained on synthetic and real data. MTT [2] introduced this approach, with enhancements from TESLA [6], FTD [11], and ATT [26] improving memory efficiency and reducing trajectory errors. iii) Feature matching: Optimizes synthetic data to produce similar internal representations as real data. Methods like CAFE [49], DM [66], and DataDAM [37] provide lightweight frameworks with comparable performance, especially for large IPC settings.

Soft-label-based dataset distillation methods. Soft-label DD methods use knowledge distillation during evaluation, with synthetic samples assigned soft labels from pretrained teacher models. Matching-based methods like DATM [17], PAD [25], and EDF [48] jointly optimize soft labels and synthetic data during trajectory matching. Decoupled methods demonstrate strong scalability on large datasets by decoupling bi-level optimization. SRe2L [58] proposed a three-stage "squeeze, recover, and relabel" paradigm, generating and saving soft labels for synthetic samples. Subsequent works CDA [57], DWA [13], EDC [41], and G-VBSM [40] further improve performance from data and

soft-label perspectives. RDED synthesizes data by concatenating core image patches, while D4M employs diffusion models for high-quality synthetic images.

Dataset distillation benchmark. A notable challenge for dataset distillation is the lack of comprehensive benchmarks. DC-Bench [5] is the first large-scale standardized benchmark for dataset condensation methods in general. It provides a comprehensive evaluation for several dataset distillation methods and coreset selection methods. Comp-DD is proposed in EDF [48] targeting dataset distillation in complex scenarios. It extracts new subsets from ImageNet1K based on the complexity metric. However, both benchmarks no longer satisfy the need for fair evaluation of dataset distillation methods under the soft label trend. Therefore, we propose DD-Ranking to solve this problem. While distillation for other modalities remains under-explored, we are aware of recent progress on multi-modal dataset distillation [71, 61, 55]. We will track the latest progress in dataset distillation across all areas and enrich our benchmark to be more comprehensive.

6 Conclusion and Future Work

We propose DD-Ranking, a new benchmark that provides a fair and comprehensive evaluation for dataset distillation. DD-Ranking is well motivated by the unfairness arising from inconsistent training settings of existing DD evaluation, especially the use of soft labels and data augmentation. To this end, DD-Ranking introduces both the label-robust score and augmentation-robust score to disentangle the effects of knowledge distillation via soft labeling and data augmentation, and ultimately reveal the true informativeness of distilled datasets. Hopefully, DD-Ranking can facilitate the development of dataset distillation towards improving data quality instead of accuracy. One potential limitation of the current DD-Ranking is that we only support methods for image classification dataset distillation. We are aware that several works [73, 53] have extended dataset distillation to other tasks and modalities. In the future, we will constantly integrate more baseline methods into our benchmark and extend DD-Ranking to other modalities.

References

[1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 1

[2] George Cazenavette, Tongzhou Wang, Antonio Torralba, Alexei A Efros, and Jun-Yan Zhu. Dataset distillation by matching training trajectories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4750–4759, 2022. 1, 2, 5, 9, 19

[3] George Cazenavette, Tongzhou Wang, Antonio Torralba, Alexei A Efros, and Jun-Yan Zhu. Generalizing dataset distillation via deep generative prior. *arXiv preprint arXiv:2305.01649*, 2023. 19

[4] Jiacheng Cui, Zhaoyi Li, Xiaochen Ma, Xinyue Bi, Yaxin Luo, and Zhiqiang Shen. Dataset distillation via committee voting. *arXiv preprint arXiv:2501.07575*, 2025. 1, 9

[5] Justin Cui, Ruochen Wang, Si Si, and Cho-Jui Hsieh. Dc-bench: Dataset condensation benchmark. *arXiv preprint arXiv:2207.09639*, 2022. 10

[6] Justin Cui, Ruochen Wang, Si Si, and Cho-Jui Hsieh. Scaling up dataset distillation to imagenet-1k with constant memory. In *International Conference on Machine Learning*, pages 6565–6590. PMLR, 2023. 9, 19

[7] Zhiwei Deng and Olga Russakovsky. Remember the past: Distilling datasets into addressable memories for neural networks. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 34391–34404. Curran Associates, Inc., 2022. 1

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[10] 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. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. 5

[11] Jiawei Du, Yidi Jiang, Vincent YF Tan, Joey Tianyi Zhou, and Haizhou Li. Minimizing the accumulated trajectory error to improve dataset distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3749–3758, 2023. 9, 19

[12] Jiawei Du, Qin Shi, and Joey Tianyi Zhou. Sequential subset matching for dataset distillation. *ArXiv*, abs/2311.01570, 2023. 19

[13] Jiawei Du, Xin Zhang, Juncheng Hu, Wenxin Huang, and Joey Tianyi Zhou. Diversity-driven synthesis: Enhancing dataset distillation through directed weight adjustment. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 5, 9

[14] Jiawei Du, Xin Zhang, Juncheng Hu, Wenxin Huang, and Joey Tianyi Zhou. Diversity-driven synthesis: Enhancing dataset distillation through directed weight adjustment. In *NeurIPS*, 2024. 19

[15] Yunzhen Feng, Shanmukha Ramakrishna Vedantam, and Julia Kempe. Embarrassingly simple dataset distillation. In *The Twelfth International Conference on Learning Representations*, 2024. 5, 16

[16] Jianyang Gu, Saeed Vahidian, Vyacheslav Kungurtsev, Haonan Wang, Wei Jiang, Yang You, and Yiran Chen. Efficient dataset distillation via minimax diffusion. *ArXiv*, abs/2311.15529, 2023. 19

[17] Ziyao Guo, Kai Wang, George Cazenavette, Hui Li, Kaipeng Zhang, and Yang You. Towards lossless dataset distillation via difficulty-aligned trajectory matching. *arXiv preprint arXiv:2310.05773*, 2023. 1, 2, 5, 9, 19

[18] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2015. 5

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[20] Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009. 5

[21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. 5

[22] Ya Le and Xuan S. Yang. Tiny imagenet visual recognition challenge. 2015. 5

[23] Saehyung Lee, Sanghyuk Chun, Sangwon Jung, Sangdoo Yun, and Sung-Hoon Yoon. Dataset condensation with contrastive signals. In *ICML*, 2022. 9, 19

[24] Yongmin Lee and Hye Won Chung. Selmatch: Effectively scaling up dataset distillation via selection-based initialization and partial updates by trajectory matching. In *ICML*, 2024. 19

[25] Zekai Li, Ziyao Guo, Wangbo Zhao, Tianle Zhang, Zhi-Qi Cheng, Samir Khaki, Kaipeng Zhang, Ahmad Sajedi, Konstantinos N Plataniotis, Kai Wang, and Yang You. Prioritize alignment in dataset distillation, 2024. 9, 19

[26] Dai Liu, Jindong Gu, Hu Cao, Carsten Trinitis, and Martin Schulz. Dataset distillation by automatic training trajectories. In *ECCV*, 2024. 9, 19

[27] Yanqing Liu, Jianyang Gu, Kai Wang, Zheng Hua Zhu, Wei Jiang, and Yang You. Dream: Efficient dataset distillation by representative matching. *ICCV*, pages 17268–17278, 2023. 19

[28] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows, 2021. 5

[29] Noel Loo, Ramin Hasani, Alexander Amini, and Daniela Rus. Efficient dataset distillation using random feature approximation. *arXiv preprint arXiv:2210.12067*, 2022. 1

[30] Noel Loo, Ramin Hasani, Alexander Amini, and Daniela Rus. Efficient dataset distillation using random feature approximation. In *NeurIPS*, 2022. 5, 19

[31] Noel Loo, Ramin Hasani, Mathias Lechner, and Daniela Rus. Dataset distillation with convexified implicit gradients. In *ICML*, 2023. 19

[32] Brian B. Moser, Federico Raue, Sebastián M. Palacio, Stanislav Frolov, and Andreas Dengel. Latent dataset distillation with diffusion models. *ArXiv*, abs/2403.03881, 2024. 19

[33] Timothy Nguyen, Zhourong Chen, and Jaehoon Lee. Dataset meta-learning from kernel ridge-regression. In *ICLR*, 2021. 19

[34] Timothy Nguyen, Roman Novak, Lechao Xiao, and Jaehoon Lee. Dataset distillation with infinitely wide convolutional networks. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *NeurIPS*, 2021. 19

[35] Tian Qin, Zhiwei Deng, and David Alvarez-Melis. A label is worth a thousand images in dataset distillation. *arXiv preprint arXiv:2406.10485*, 2024. 2

[36] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge, 2015. 5

[37] Ahmad Sajedi, Samir Khaki, Ehsan Amjadian, Lucy Z Liu, Yuri A Lawryshyn, and Konstantinos N Plataniotis. Datadam: Efficient dataset distillation with attention matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17097–17107, 2023. 2, 5, 9

[38] Yuzhang Shang, Zhihang Yuan, and Yan Yan. Mim4dd: Mutual information maximization for dataset distillation. In *NeurIPS*, 2023. 19

[39] Shitong Shao, Zeyuan Yin, Muxin Zhou, Xindong Zhang, and Zhiqiang Shen. Generalized large-scale data condensation via various backbone and statistical matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16709–16718, 2024. 1, 2, 9, 19

[40] Shitong Shao, Zeyuan Yin, Muxin Zhou, Xindong Zhang, and Zhiqiang Shen. Generalized large-scale data condensation via various backbone and statistical matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 5, 9

[41] Shitong Shao, Zikai Zhou, Huanran Chen, and Zhiqiang Shen. Elucidating the design space of dataset condensation. *arXiv preprint arXiv:2404.13733*, 2024. 1, 2, 5, 9, 19

[42] Seung-Jae Shin, Heesun Bae, DongHyeok Shin, Weonyoung Joo, and Il-Chul Moon. Loss-curvature matching for dataset selection and condensation. In *AISTATS*, 2023. 9, 19

[43] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015. 5

[44] Duo Su, Junjie Hou, Weizhi Gao, Yingjie Tian, and Bowen Tang. Dˆ 4: Dataset distillation via disentangled diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5809–5818, 2024. 1, 9

[45] Duo Su, Junjie Hou, Weizhi Gao, Yingjie Tian, and Bowen Tang. D4m: Dataset distillation via disentangled diffusion model. In *CVPR*, 2024. 2, 19

[46] Peng Sun, Bei Shi, Daiwei Yu, and Tao Lin. 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*, pages 9390–9399, 2024. 1, 2, 5, 9, 19

[47] Kai Wang, Jianyang Gu, Daquan Zhou, Zheng Hua Zhu, Wei Jiang, and Yang You. Dim: Distilling dataset into generative model. *ArXiv*, abs/2303.04707, 2023. 19

[48] Kai Wang, Zekai Li, Zhi-Qi Cheng, Samir Khaki, Ahmad Sajedi, Ramakrishna Vedantam, Konstantinos N Plataniotis, Alexander Hauptmann, and Yang You. Emphasizing discriminative features for dataset distillation in complex scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 5, 9, 10, 19

[49] Kai Wang, Bo Zhao, Xiangyu Peng, Zheng Zhu, Shuo Yang, Shuo Wang, Guan Huang, Hakan Bilen, Xinchao Wang, and Yang You. Cafe: Learning to condense dataset by aligning features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12196–12205, 2022. 9

[50] Kai Wang, Bo Zhao, Xiangyu Peng, Zheng Hua Zhu, Shuo Yang, Shuo Wang, Guan Huang, Hakan Bilen, Xinchao Wang, and Yang You. Cafe learning to condense dataset by aligning features. In *CVPR*, 2022. 19

[51] Shaobo Wang, Yicun Yang, Zhiyuan Liu, Chenghao Sun, Xuming Hu, Conghui He, and Linfeng Zhang. Dataset distillation with neural characteristic function: A minmax perspective. In *CVPR*, 2025. 19

[52] Tongzhou Wang, Jun-Yan Zhu, Antonio Torralba, and Alexei A Efros. Dataset distillation. *arXiv preprint arXiv:1811.10959*, 2018. 1, 4

[53] Xindi Wu, Byron Zhang, Zhiwei Deng, and Olga Russakovsky. Vision-language dataset distillation. In *TMLR*, 2024. 10

[54] Shaolei Yang, Shen Cheng, Mingbo Hong, Haoqiang Fan, Xing Wei, and Shuaicheng Liu. Neural spectral decomposition for dataset distillation. In *ECCV*, 2024. 19

[55] Jae-Young Yim, Dongwook Kim, and Jae-Young Sim. Dataset distillation of 3d point clouds via distribution matching. In *NeurIPS*, 2025. 10

[56] Zeyuan Yin and Zhiqiang Shen. Dataset distillation in large data era. *ArXiv*, 2023. 19

[57] Zeyuan Yin and Zhiqiang Shen. Dataset distillation via curriculum data synthesis in large data era. *Transactions on Machine Learning Research*, 2024. 2, 5, 9

[58] Zeyuan Yin, Eric Xing, and Zhiqiang Shen. Squeeze, recover and relabel: Dataset condensation at imagenet scale from a new perspective. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 2, 5, 9, 19

[59] Ruonan Yu, Songhua Liu, Zigeng Chen, Jingwen Ye, and Xinchao Wang. Heavy labels out! dataset distillation with label space lightening. *ArXiv*, 2024. 19

[60] Hansong Zhang, Shikun Li, Pengju Wang, Dan Zeng, and Shiming Ge. M3d: Dataset condensation by minimizing maximum mean discrepancy. In *AAAI*, 2023. 19

[61] Xin Zhang, Ziruo Zhang, Jiawei Du, Zuozhu Liu, and Joey Tianyi Zhou. Beyond modality collapse: Representations blending for multimodal dataset distillation. In *NeurIPS*, 2025. 10

[62] Bo Zhao and Hakan Bilen. Dataset condensation with differentiable siamese augmentation. In *International Conference on Machine Learning*, pages 12674–12685. PMLR, 2021. 2, 5, 9, 19

[63] Bo Zhao and Hakan Bilen. Dataset condensation with distribution matching. *arXiv preprint arXiv:2110.04181*, 2021. 1, 5, 19

[64] Bo Zhao and Hakan Bilen. Synthesizing informative training samples with gan. *arXiv preprint arXiv:2204.07513*, 2022. 1

[65] Bo Zhao and Hakan Bilen. Synthesizing informative training samples with gan. *ArXiv*, 2022. 19

[66] Bo Zhao and Hakan Bilen. Dataset condensation with distribution matching. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6514–6523, 2023. 2, 9

[67] Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. Dataset condensation with gradient matching. *ICLR*, 1(2):3, 2021. 1, 2, 5, 9, 19

[68] Ganlong Zhao, Guanbin Li, Yipeng Qin, and Yizhou Yu. Improved distribution matching for dataset condensation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7856–7865, 2023. 19

[69] Zhenghao Zhao, Yuzhang Shang, Junyi Wu, and Yan Yan. Dataset quantization with active learning based adaptive sampling. In *ECCV*, 2024. 19

[70] Zhenghao Zhao, Haoxuan Wang, Yuzhang Shang, Kai Wang, and Yan Yan. Distilling long-tailed datasets. In *CVPR*, 2025. 19

[71] Zhenghao Zhao, Haoxuan Wang, Junyi Wu, Yuzhang Shang, Gaowen Liu, and Yan Yan. Efficient multimodal dataset distillation via generative models. In *NeurIPS*, 2025. 10

[72] Xinhao Zhong, Hao Fang, Bin Chen, Xulin Gu, Tao Dai, Meikang Qiu, and Shu-Tao Xia. Hierarchical features matter: A deep exploration of gan priors for improved dataset distillation. *ArXiv*, abs/2406.05704, 2024. 19

[73] Daquan Zhou, Kai Wang, Jianyang Gu, Xiangyu Peng, Dongze Lian, Yifan Zhang, Yang You, and Jiashi Feng. Dataset quantization, 2023. 10

[74] Yongchao Zhou, Ehsan Nezhadarya, and Jimmy Ba. Dataset distillation using neural feature regression. In *NeurIPS*, 2022. 5, 19

A Theoretical foundation

We provide an idealized additive analysis of the matched-baseline comparison underlying LRS (ARS follows analogously). The cancellation below assumes a fixed protocol and a setting-dependent offset that is independent of the evaluated data.

Definition of fair comparison. A fair comparison in DD isolates data quality contributions while neutralizing setting-dependent performance gains. Formally, let q_A, q_B denote the intrinsic data quality of methods A and B, and $\beta(\pi)$ represent the performance boost from evaluation setting π (independent of data quality). The comparison is fair if and only if it directly compares q_A, q_B.

Additive interpretation of IOR. Consider method A distilling $\mathcal{D}_A$ with protocol π_A (e.g., hard labels, batch size 256, step learning rate scheduler) and method B distilling $\mathcal{D}_B$ with protocol π_B (e.g., soft labels, batch size 1024, cosine learning rate scheduler). We decompose performance as

$$\text{Acc}(\mathcal{D}_A, \pi_A) = q_A + \beta(\pi_A) + \epsilon_A, \;\; \text{Acc}(\mathcal{D}_B, \pi_B) = q_B + \beta(\pi_B) + \epsilon_B \tag{5}$$

where $\epsilon \sim \mathcal{N}(0, 1)$ is the residual term. Direct accuracy comparison is unfair due to the unknown relationship between $\beta(\pi_A)$ and $\beta(\pi_B)$. For a random-selection baseline, we similarly write

$$\text{Acc}(\mathcal{D}_\text{r}, \pi_\text{r}) = q_\text{r} + \beta(\pi_\text{r}) + \epsilon_\text{r} \tag{6}$$

Under these assumptions, the setting-dependent offset cancels in IOR:

$$\begin{aligned} \text{IOR}_A &= \text{Acc}(\mathcal{D}_A, \pi_A) - \text{Acc}(\mathcal{D}_{\text{r}_A}, \pi_{\text{r}_A}) = q_A - q_{\text{r}_A} + \epsilon_A - \epsilon_{\text{r}_A} \\ \text{IOR}_B &= \text{Acc}(\mathcal{D}_B, \pi_B) - \text{Acc}(\mathcal{D}_{\text{r}_B}, \pi_{\text{r}_B}) = q_B - q_{\text{r}_B} + \epsilon_B - \epsilon_{\text{r}_B} \end{aligned} \tag{7}$$

where β terms cancel since $\pi_A = \pi_{\text{r}_A}$ and $\pi_B = \pi_{\text{r}_B}$. Thus $\mathbb{E}[\text{IOR}_A - \text{IOR}_B] = \mathbb{E}[q_A - q_B] - \mathbb{E}[q_{\text{rand}_A} - q_{\text{rand}_B}] + \mathbb{E}[\epsilon_A - \epsilon_B] - \mathbb{E}[\epsilon_{\text{rand}_A} - \epsilon_{\text{rand}_B}]$. Since the expected intrinsic quality of random selection is constant for the same original dataset $\mathcal{D}$ at the same IPC, we have $\mathbb{E}[q_{\text{rand}_A} - q_{\text{rand}_B}] = 0$ and $\mathbb{E}[\epsilon] = 0$, yielding $\mathbb{E}[\text{IOR}_A - \text{IOR}_B] = \mathbb{E}[q_A - q_B]$ under the stated assumptions. Data–protocol interactions need not satisfy this additive model, and the separate learning-rate selection in Section 4.1 does not guarantee identical realized protocols. Consequently, this argument motivates the matched-baseline control rather than proving that the implemented evaluation eliminates every interaction or tuning effect.

B Additional Experiment Results

Results from Table 3 to Table 7 are computed by letting $\lambda = 0.5$. By default, we treat hard label recovery and improvement over random as equally important. Tables 13–16 report LRS across five values of λ, with IPC1, IPC10, and IPC50 grouped within each dataset table. A larger λ gives higher priority to IOR, and a smaller λ focuses more on HLR. We encourage future newly proposed DD methods to enhance both HLR and IOR. Standard deviations over five random seeds are reported as subscripts alongside the means in Tables 3–7; the EDF entries in Table 6 have no reported standard deviations.

Changing λ can reflect different methods' advantages. For example (Table 16, IPC10), when λ is large (large weight on IOR), RDED's hard-label advantage is diminished (EDC now ranks first jointly with RDED). RDED selects real image patches and combines them to form one synthetic image, consistent with its stronger hard-label performance in this comparison.

B.1 Evaluation protocol details

Table 10 distinguishes the configurations for distilled data, matched random subsets, and the full-dataset reference. The configuration survey in Table 1 documents differences between published methods rather than defining our protocol.

Random-subset construction and learning-rate selection. For each seed, we resample a class-balanced subset of the original training data at the same IPC as the distilled dataset. These real images are not merged or otherwise modified to construct the subset; they subsequently undergo the matched preprocessing and augmentation pipeline. For teacher-labeled methods, we use the same teacher and soft-label generation procedure, including per-epoch labeling where applicable. Under hard labels, we search the learning rate separately for synthetic and random data. Under soft labels, we retain the

method's supplied learning rate for synthetic data and search the learning rate for the random subset. Thus, matching the recipe allows independently selected learning rates; a published learning rate is not assumed optimal for random data.

Full-dataset reference. We use a fixed full-dataset reference per dataset, shared across methods, under a unified reference recipe with only the learning rate swept. This is a common reference rather than a maximum-accuracy oracle. For a fixed dataset and λ, changing the common reference accuracy by Δ changes every α by $-(1-\lambda)\Delta$, preserving the ranking. The nonlinear normalization can nevertheless change LRS values and numerical gaps between them.

Data	Evaluation configuration
Distilled set	Method-specific evaluation recipe. Hard-label evaluation uses each sample's class identity and a new learning-rate search. Soft-label evaluation retains the method's supplied learning rate.
Random subset	Original real images, class-balanced at the same IPC and resampled per seed. Model, preprocessing, teacher/label protocol, loss, augmentation, optimizer, and schedule match the synthetic arm. Learning rate is searched separately.
Full dataset	Fixed reference shared across methods for a dataset, evaluated using a unified reference recipe with only the learning rate swept. It is not a maximum-accuracy oracle.

Table 10: Roles of the three evaluation configurations. Matching synthetic and random data refers to the evaluation recipe; selected learning rates may differ.

B.2 Distillation cost and calibrated performance

Table 11 compares the reported distillation costs of RDED and G-VBSM on ImageNet1K at IPC10. RDED achieves LRS 17.4 at 0.2 GPU-hours, whereas G-VBSM achieves 15.5 at 139.2 GPU-hours. These measurements illustrate that greater distillation cost does not necessarily yield a higher calibrated score. They concern distillation rather than the total cost of running DD-Ranking; no hardware-specific speedup or complete evaluation-budget comparison is inferred.

Method	Distillation cost (GPU-hours)	LRS↑
RDED	0.2	17.4
G-VBSM	139.2	15.5

Table 11: Reported distillation costs and LRS for two methods on ImageNet1K at IPC10. LRS uses $\lambda = 0.5$; the costs do not represent the total DD-Ranking evaluation budget.

B.3 Additional method families

We additionally evaluate RaT-BPTT from *Embarrassingly Simple Dataset Distillation* [15] on CIFAR-10 and CIFAR-100. The additional evaluation also includes DataDAM on TinyImageNet under the same protocol and metric computation as the other methods. Table 12 summarizes selected RaT-BPTT results. RaT-BPTT exceeds the original main-table leaders at CIFAR-10 IPC1 and CIFAR-100 IPC10, and maintains an IOR above 23 percentage points across its evaluated CIFAR settings, including IPC50.

Method	Dataset	IPC	Metric	Value
RaT-BPTT	CIFAR-10	1	LRS↑	25.3
RaT-BPTT	CIFAR-100	10	LRS↑	28.9

Table 12: Selected RaT-BPTT results with LRS evaluated at the default $\lambda = 0.5$.

B.4 Ranking stability across weights

For each of the 12 dataset–IPC settings, we compare all $\binom{5}{2} = 10$ pairs of rankings induced by $\lambda \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$, using the common set of methods within that setting. We compute Kendall's τ_b directly from the tabulated LRS values, so equal reported scores remain ties. For two score vectors, let C and D be the numbers of concordant and discordant method pairs, and let T_x and T_y count pairs tied only in the first or second vector, respectively. Then

$$\tau_b = \frac{C - D}{\sqrt{(C + D + T_x)(C + D + T_y)}}. \tag{8}$$

Pairs tied in both vectors contribute to none of these counts. We first average the 10 correlations within each setting, then average equally across the 12 settings (120 comparisons in total). The resulting mean is 0.9159 (0.92 to two decimal places); the per-setting averages are reported in the main-text Table 9. Agreement is high on average but not uniform: the setting-level averages range from 0.7757 to 0.9667. This calculation uses the rounded values in Tables 13–16; it describes the currently tabulated method sets and can change when unrounded outputs or additional methods are included. The calculation is reproducible with `python3 scripts/analyze_ranking_stability.py`. This analysis retains the original weight-sweep method sets; the kernel-based results in Table 5 and RaT-BPTT results in Table 12 do not provide complete weight sweeps and are excluded.

Method	IPC1					IPC10					IPC50				
	0.1	0.3	0.5	0.7	0.9	0.1	0.3	0.5	0.7	0.9	0.1	0.3	0.5	0.7	0.9
DC	11.2	14.9	19.1	24.0	29.5	15.5	19.1	23.2	27.7	32.8	18.3	21.1	24.0	27.2	30.6
DSA	9.7	13.7	18.2	23.5	29.5	16.0	19.6	23.7	28.3	33.4	18.0	20.6	23.5	26.7	30.1
MTT	14.3	18.7	23.9	29.8	36.6	19.8	23.9	28.5	33.5	39.2	21.8	24.7	27.8	31.1	34.7
DM	9.0	12.8	17.0	22.0	27.6	14.7	18.2	22.2	26.7	31.6	18.7	21.4	24.3	27.5	30.9
DataDAM	11.9	15.8	20.2	25.2	30.9	16.1	19.7	23.8	28.4	33.5	19.8	22.6	25.6	28.8	32.3
DATM	14.4	19.2	24.6	30.9	38.2	19.0	23.5	28.7	34.5	41.2	21.1	24.4	28.0	31.9	36.1
SRe2L	7.0	10.4	14.3	18.8	23.9	7.3	10.4	13.8	17.7	22.1	8.3	11.2	14.4	18.0	22.0
RDED	9.1	12.4	16.2	20.4	25.3	11.3	14.3	17.5	21.2	25.2	15.1	17.3	19.6	22.1	24.8
D4M	11.4	14.7	18.4	22.6	27.3	14.3	17.4	20.8	24.6	28.7	17.9	20.3	22.8	25.4	28.3

Table 13: LRS on CIFAR-10 across IPC settings and weights λ. Column headings within each IPC group denote λ.

Method	IPC1					IPC10					IPC50				
	0.1	0.3	0.5	0.7	0.9	0.1	0.3	0.5	0.7	0.9	0.1	0.3	0.5	0.7	0.9
DC	14.4	17.5	20.8	24.4	28.5	18.6	21.3	24.3	27.4	30.8	19.4	21.0	22.7	24.5	26.4
DSA	12.7	16.0	19.6	23.7	28.2	18.4	21.3	24.3	27.6	31.2	19.6	21.2	23.0	24.8	26.8
MTT	15.9	19.3	23.1	27.4	32.1	21.3	24.3	27.5	30.9	34.7	22.9	24.8	26.8	28.8	31.0
DM	12.1	15.3	18.9	22.8	27.2	17.1	19.9	23.0	26.2	29.8	21.3	23.1	24.9	26.9	29.0
DataDAM	12.9	16.2	19.9	23.9	28.5	18.6	21.5	24.6	28.0	31.7	22.9	24.8	26.8	28.9	31.1
DATM	19.2	22.3	25.7	29.4	33.4	20.9	23.7	26.8	30.1	33.6	24.2	27.2	30.4	33.9	37.6
SRe2L	10.8	13.6	16.7	20.2	24.0	11.0	12.9	15.0	17.3	19.8	12.1	14.1	16.2	18.5	21.0
RDED	12.6	15.2	18.1	21.3	24.8	14.7	17.0	19.4	22.0	24.9	17.6	19.3	21.2	23.1	25.2
D4M	16.9	19.7	22.7	25.9	29.5	14.3	17.4	20.9	24.7	29.0	18.3	21.1	24.2	27.5	31.1

Table 14: LRS on CIFAR-100 across IPC settings and weights λ. Column headings within each IPC group denote λ.

C Additional Related Work

We acknowledge that DD-Ranking has not included enough dataset distillation methods. We discuss them in Table 17. In the near future, we will continue to extend our benchmark and include more baseline methods.

Method	IPC1					IPC10					IPC50				
	0.1	0.3	0.5	0.7	0.9	0.1	0.3	0.5	0.7	0.9	0.1	0.3	0.5	0.7	0.9
DC	17.4	19.6	22.0	24.5	27.2	19.7	21.7	23.9	26.3	28.7	19.4	20.8	22.2	23.7	25.2
DSA	16.9	19.1	21.6	24.2	27.0	20.0	22.0	24.1	26.3	28.7	20.9	22.8	24.8	26.9	29.1
MTT	16.8	19.3	21.9	24.8	27.8	21.9	24.2	26.7	29.3	32.1	21.7	23.7	25.8	28.0	30.3
DM	15.0	17.5	20.2	23.1	26.2	18.2	20.6	23.1	25.8	28.7	20.4	22.2	24.1	26.1	28.1
DATM	18.5	20.9	23.5	26.2	29.2	20.9	23.4	26.0	28.8	31.8	22.6	24.9	27.2	29.8	32.4
EDF	18.4	20.9	23.5	26.3	29.4	20.9	23.5	26.2	29.2	32.3	22.5	24.9	27.3	30.0	32.8
SRe2L	12.5	15.1	17.9	21.0	24.3	12.8	14.9	17.1	19.5	22.1	15.5	17.0	18.6	20.3	22.1
RDED	15.8	18.3	20.9	23.8	26.9	18.2	20.1	22.1	24.2	26.5	21.4	22.5	23.7	24.8	26.0
D4M	13.8	16.1	18.6	21.2	24.1	15.1	16.9	18.9	21.1	23.3	17.9	20.8	23.8	27.2	30.8

Table 15: LRS on TinyImageNet across IPC settings and weights λ. Column headings within each IPC group denote λ.

Method	IPC1					IPC10					IPC50				
	0.1	0.3	0.5	0.7	0.9	0.1	0.3	0.5	0.7	0.9	0.1	0.3	0.5	0.7	0.9
SRe2L	9.9	12.9	16.2	19.9	24.0	9.9	12.0	14.2	16.7	19.3	10.3	12.5	14.8	17.4	20.2
RDED	10.2	13.3	16.8	20.8	25.2	11.4	14.2	17.4	20.8	24.6	14.0	16.2	18.6	21.2	23.9
D4M	10.1	13.1	16.4	20.2	24.4	10.6	13.0	15.8	18.7	22.0	12.9	15.1	17.6	20.2	23.0
DWA	10.0	13.0	16.3	20.0	24.1	10.3	13.1	16.1	19.5	23.2	11.3	13.7	16.3	19.1	22.1
CDA	9.9	12.8	16.1	19.7	23.7	10.1	12.6	15.3	18.3	21.6	10.8	13.3	16.1	19.1	22.4
EDC	10.1	13.1	16.4	20.1	24.3	11.0	13.9	17.1	20.6	24.6	13.7	16.2	18.9	21.9	25.1
G-VBSM	10.0	12.9	16.3	20.0	24.1	10.1	12.7	15.5	18.6	22.1	12.6	14.9	17.4	20.0	22.9

Table 16: LRS on ImageNet1K across IPC settings and weights λ. Column headings within each IPC group denote λ.

D DD-Ranking Team

We provide the full list of DD-Ranking team members as follows (* denotes equal contribution):

- Zekai Li* (National University of Singapore)
- Xinhao Zhong* (National University of Singapore)
- Samir Khaki (University of Toronto)
- Zhiyuan Liang (National University of Singapore)
- Yuhao Zhou (National University of Singapore)
- Mingjia Shi (National University of Singapore)
- Dongwen Tang (National University of Singapore)
- Ziqiao Wang (National University of Singapore)
- Wangbo Zhao (National University of Singapore)
- Xuanlei Zhao (National University of Singapore)
- Mengxuan Wu (National University of Singapore)
- Haonan Wang (National University of Singapore)
- Ziheng Qin (National University of Singapore)
- Dai Liu (Technical University of Munich)
- Kaipeng Zhang (Shanghai AI Lab)
- Tianyi Zhou (A*STAR)
- Zheng Zhu (Tsinghua University)
- Kun Wang (University of Science and Technology of China)
- Shaobo Wang (Shanghai Jiao Tong University)
- Guang Li (Hokkaido University)
- Junhao Zhang (National University of Singapore)
- Jiawei Liu (National University of Singapore)

Category	Method
Kernel-based	KIP-FC [33]
	KIP-ConvNet [34]
	FRePo [74]
	RFAD [30]
	RCIG [31]
Gradient-matching	DC [67]
	DSA [62]
	DCC [23]
	LCMat [42]
Trajectory-matching	MTT [2]
	TESLA [6]
	FTD [11]
	SeqMatch [12]
	DATM [17]
	ATT [26]
	NSD [54]
	PAD [25]
	EDF [48]
	SelMatch [24]
Distribution-matching	DM [63]
	CAFE [50]
	IDM [68]
	DREAM [27]
	M3D [60]
	NCFD [51]
Generative model	DiM [47]
	GLaD [3]
	H-PD [72]
	LD3M [32]
	IT-GAN [65]
	D4M [45]
	Minimax Diffusion [16]
Decoupled	SRe2L [58]
	RDED [46]
	HeLIO [59]
	DWA [14]
	CDA [56]
	EDC [41]
	G-VBSM [39]
Others	MIM4DD [38]
	DQAS [69]
	LDD [70]

Table 17: Summary of previous works on dataset distillation

- Zhiheng Ma (SUAT)
- Linfeng Zhang (Shanghai Jiao Tong University)
- Yiran Huang (Technical University of Munich)
- Lingjuan Lyu (Sony)
- Jiancheng Lv (Sichuan University)
- Yaochu Jin (Westlake University)
- Zeynep Akata (Technical University of Munich)
- Jindong Gu (Oxford University)
- Peihao Wang (University of Texas at Austin)
- Mike Shou (National University of Singapore)
- Zhiwei Deng (Google DeepMind)
- Qian Zheng (Zhejiang University)

- Hao Ye (Xiaomi)
- Shuo Wang (Baidu)
- Xiaobo Wang (Chinese Academy of Sciences)
- Yan Yan (University of Illinois at Chicago)
- Yuzhang Shang (University of Illinois at Chicago)
- George Cazenavette (Massachusetts Institute of Technology)
- Xindi Wu (Princeton University)
- Justin Cui (University of California, Los Angeles)
- Tianlong Chen (University of North Carolina at Chapel Hill)
- Angela Yao (National University of Singapore)
- Baharan Mirzasoleiman (University of California, Los Angeles)
- Hakan Bilen (University of Edinburgh)
- Manolis Kellis (Massachusetts Institute of Technology)
- Konstantinos N. Plataniotis (University of Toronto)
- Bo Zhao (Shanghai Jiao Tong University)
- Zhangyang Wang (University of Texas at Austin)
- Yang You (National University of Singapore)
- Kai Wang (National University of Singapore)

Zekai and Xinhao *contribute equally* to this work. Zekai serves as the *project lead*, and Kai Wang is the *corresponding author*.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction describe the matched-baseline benchmark and its scope. Section 4.1 specifies the evaluated datasets, models, and comparison protocol.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The conclusion limits the current benchmark to image classification and discusses extending its method and modality coverage. Appendix A states the assumptions and limitations of the additive interpretation.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Appendix A states the fixed-protocol and data-independent additive-offset assumptions and derives the cancellation underlying IOR. It also explains why data–protocol interactions and independent learning-rate selection limit this interpretation.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The metric definitions, datasets, models, five-seed evaluation, and baseline-matching rules are specified in the paper and Appendix B.1. The public benchmark repository provides implementation and usage instructions, with links to evaluation configurations and baseline implementations.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Code, installation instructions, and usage examples are available at https://github.com/NUS-HPC-AI-Lab/DD-Ranking; its README links to the evaluation configurations and baseline implementations. The experiments use existing datasets, with access subject to their original providers' terms.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 4.1 and Appendix B.1 describe the datasets, models, matched training recipes, learning-rate selection, and random-subset construction. Table 1 surveys method-specific training configurations, and the public repository links to evaluation configurations.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Tables 3–7 report five-seed means with available standard deviations as subscripts, as defined in Section 4.1. Entries without reported deviations, including EDF, are explicitly left without subscripts.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: Appendix B.2 reports selected distillation costs, but these do not constitute a complete evaluation compute budget. Per-experiment hardware, memory, execution time, and total research compute are not fully reported.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The research benchmarks existing image-classification datasets and distillation methods and does not recruit human subjects or collect new personal data. We adhere to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [No]

Justification: The paper discusses the benefits of fairer evaluation and more informative distilled datasets, but does not provide a systematic discussion of potential negative societal impacts.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: This work evaluates existing classification datasets and methods; it does not introduce high-risk generative models or newly collected sensitive datasets requiring controlled-release safeguards.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [No]

Justification: The paper credits the datasets and methods, and the benchmark repository lists licenses for incorporated code. However, the paper does not enumerate the licenses and terms of use for every third-party dataset and model.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The DD-Ranking software repository provides installation instructions, usage examples, configuration examples, and an MIT license, with links to documentation and the leaderboard. The paper and Appendix B.1 describe its metrics and evaluation protocol.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The study does not involve crowdsourcing, participant recruitment, or human-subject experiments.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The research consists of computational experiments on existing image-classification datasets and does not involve human-subject experiments.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not a component of the benchmark metrics or the experimental methods. Assistance with writing, editing, or formatting does not constitute an LLM-based research method.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.