Well Begun is Half Done: The Importance of Initialization in Dataset Distillation

Yiran Guan, Zhu Chen, Xingkui Zhu, Dingkang Liang
Yuliang Liu, and Xiang Bai†

Huazhong University of Science and Technology
{yiranguan, zhu_chen, adlith, dkliang, ylliu, xbai}@hust.edu.cn

Abstract. Dataset distillation aims to synthesize small yet informative datasets using deep learning optimization strategies, helping to reduce storage requirements and training costs. Through specific training objectives, models trained on these synthetic datasets can achieve performance comparable to those trained on the original, larger datasets. This technique has successfully condensed several popular datasets and shown significant potential. However, current methods face several challenges. Chief among them is the time-consuming process of generating synthetic images, which can sometimes exceed the time required to train on the original dataset. To address this challenge, we revealed an initial dependency in dataset distillation. We discovered that a well-designed initialization of synthetic data could speed up data generation and improve the quality of training outcomes. Leveraging this insight, we developed a plug-and-play method named Initialization Improved Dataset Distillation (IIDD). This method achieved 1st place in Tiny ImageNet and 2nd place overall in *The First Dataset Distillation Challenge at the ECCV 2024 Workshop*, demonstrating a significant improvement of +1.15 on CIFAR-100 and +1.77 on Tiny ImageNet compared to baseline.

Keywords: Dataset Distillation · Initial Dependency

1 Introduction

The expansion of dataset sizes improves the performance of deep learning models by providing more diverse and representative data, which enhances generalization, reduces overfitting, and enables more effective feature learning. However, this also gives rise to considerably prolonged training times and causes storage challenges [27]. Coreset selection methods [10] address these issues by identifying representative samples from the original dataset. However, this approach may result in the loss of valuable information and a reduction in model performance. Dataset distillation [24] presents a promising alternative, it generates synthetic datasets through pixel-level optimization, capturing more information within the same storage constraints. It has been demonstrated to reduce the size of popular datasets such as CIFAR-10/100, Tiny ImageNet, ImageNet 1k, and ImageNet

† Corresponding author

21k [6, 25]. Furthermore, dataset distillation has been proven to enhance neural architecture search by accelerating the architecture search process [27, 29], improving continual learning by making memory storage more efficient [27, 29, 31], and reducing data transmission overhead in federated learning [26, 29], highlighting its versatility and effectiveness in deep learning applications.

The concept of dataset distillation was first introduced by Wang et al. [24]. This method employs a meta-learning framework to refine the synthetic images. However, the proposed algorithm involves bi-level optimization, which is computationally intensive and challenging to implement effectively. Subsequently, researchers have investigated a range of optimization objectives for guiding image synthesis. These include aligning training gradients (DC-based methods) [27], matching embedding distributions (DM-based methods) [23, 29, 30], and tracking the training trajectories (TM-based methods) [4, 6, 8, 13] of original images. Meanwhile, generative methods such as GLaD [5] and ITGAN [28] compress datasets into latent variables and synthesize data using decoders.

Despite these promising advancements, current approaches in dataset distillation encounter significant hurdles. The computational overhead of the distillation process can exceed the cost of directly training the model many times.

To mitigate these challenges, we introduced a plug-and-play method called Initialization Improved Dataset Distillation (IIDD). This approach stems from the observation that the random initialized synthetic images often retain significant similarity to their original counterparts, even after extensive pixel-level optimization. This phenomenon led us to explore the role of initialization in dataset distillation. We found that effective initialization not only enhances distillation performance but also speeds up convergence. This discovery reveals a critical link between the feature distribution of the initialization images and the overall effectiveness of the distillation process. Proper initialization can greatly improve distillation efficiency, while poor initialization may result in synthetic datasets that underperform compared to randomly sampled subsets of similar size. Our experiments show that IIDD has significantly improved performance over several dataset distillation baselines (i.e., +1.15 on CIFAR-100, +1.77 on Tiny-ImageNet). In particular, it secured 1st place on Tiny ImageNet and 2nd place overall in *The First Dataset Distillation Challenge at the ECCV 2024 Workshop*.

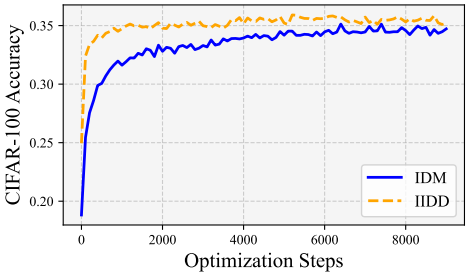


Fig. 1: IDM and IIDD training curves. IIDD converges faster and has better performance.

Our contributions are outlined below:

- We present Initialization Improved Dataset Distillation (IIDD). This plug-and-play approach optimizes the initialization phase of dataset distillation to improve performance and speed up the synthesis process significantly.

- Our research explores the phenomenon of initial dependency in dataset distillation. We show that these initial images influence the performance of the final synthesized dataset.
- Our method significantly enhanced the effectiveness of the previous method (IDM, FTD), leading to marked improvements in synthetic dataset quality. Specifically, it secured 1st place on Tiny ImageNet and 2nd place overall in *The First Dataset Distillation Challenge at the ECCV 2024 Workshop*.

2 Related Work

2.1 Coreset Selection

Coreset selection [1, 10, 11] is a method for selecting a representative subset of data to retain the essential information of the original dataset. Some approaches achieve this by defining criteria for evaluating the representativeness of samples, which are then used to select and form the coreset. Common criteria include compactness [18], diversity [2], and forgetfulness [22]. Other methods [20] utilize clustering techniques to select the coreset from the original dataset. However, these methods often struggle to align with specific target tasks, limiting their effectiveness in ensuring an optimal solution. Additionally, the performance of coreset selection methods is heavily dependent on the quality of the original data, which may further constrain their practical utility.

2.2 Dataset distillation

Dataset distillation [24] is a method for refining large datasets into smaller ones with the objective of training models faster while maintaining accuracy. Unlike core-set selection, dataset distillation methods produce synthetic subsets rather than selecting from the original dataset. In recent years, various methods have been proposed for dataset distillation. For instance, DC [27] proposed the gradient matching method for dataset distillation. CAFE [23], DM [29] introduced the distribution matching method, while MTT [4] developed the trajectory matching approach. These methods have demonstrated varying degrees of success. Further, DSA [26], TESLA [6], and some other works have utilized techniques such as soft labels and data augmentation to further enhance the effectiveness of the three foundational distillation methods. There are also ITGAN [28] and GLaD [5] that use generative models for dataset distillation. Additionally, DREAM [15] and DREAM+ [16] proposed to select representative original images for bidirectional matching, which improved the effectiveness of both the gradient matching and distribution matching methods.

3 Method

In this section, we begin with a brief statement on dataset distillation.

3.1 Problem statement

Dataset distillation can be mathematically formalized as the process of reducing a large-scale dataset into a smaller, information-rich subset while maintaining minimal loss in model performance. Specifically, suppose we have an original large-scale dataset $\mathcal{D}^{\mathcal{T}} = \{(x_i^{\mathcal{T}}, y_i^{\mathcal{T}})\}_{i=1}^{N}$, where x_i represents the input samples, y_i the corresponding labels and the dataset size is N. The objective of dataset distillation is to find a distilled dataset $\mathcal{D}^{\mathcal{S}} = \{(x_j^{\mathcal{S}}, y_j^{\mathcal{S}})\}_{j=1}^{M}$ with $M \ll N$, such that the model trained on $\mathcal{D}^{\mathcal{S}}$ performs as closely as possible to a model trained directly on the original dataset $\mathcal{D}^{\mathcal{T}}$.

This process can be formulated as the following optimization problem:

$$\mathcal{D}^{\mathcal{S}*} = \arg\min_{\mathcal{D}^{\mathcal{S}}} \mathbb{E}_{(x,y)\sim\mathcal{D}^{\mathcal{T}}} \left[\mathcal{L}(f_{\theta^{\mathcal{S}}}(x), y)\right], \tag{1}$$

where $\theta^{\mathcal{S}}$ are the model parameters obtained by training on the distilled dataset $\mathcal{D}^{\mathcal{S}}$:

$$\theta^{\mathcal{S}} = \arg\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}^{\mathcal{S}}} \left[\mathcal{L}(f_\theta(x), y)\right]. \tag{2}$$

Distribution Matching (DM) is an optimization objective in dataset distillation. It aims to match the feature distributions $f_\theta(x^{\mathcal{S}})$ and $f_\theta(x^{\mathcal{T}})$ for $\mathcal{D}^{\mathcal{S}}$ and $\mathcal{D}^{\mathcal{T}}$, respectively. Taking maximum mean discrepancy (MMD) as the distance measure, it can be formulated as:

$$\mathcal{D}^{\mathcal{S}*} = \arg\min_{\mathcal{D}^{\mathcal{S}}} \mathbb{E}_{\theta\sim P_{\theta_0}} \left\| \frac{1}{N}\sum_{i=1}^{N} f_\theta(x_i^{\mathcal{T}}) - \frac{1}{M}\sum_{j=1}^{M} f_\theta(x_j^{\mathcal{S}}) \right\|^2, \tag{3}$$

where P_{θ_0} denotes the distribution of randomly initialized network parameters. In IDM [30], distribution matching is performed with respect to network parameters at different training stages, effectively replacing P_{θ_0} with P_{θ_t} to indicate that the optimization is conducted based on the parameters obtained after the t th training step.

For trajectory matching (TM) based methods, the distillation process is conducted by aligning the training trajectories of surrogate models optimized over $\mathcal{D}^{\mathcal{T}}$ and $\mathcal{D}^{\mathcal{S}}$. In each iteration of the distillation, $\theta_t^{\mathcal{T}}$ and $\theta_{t+L^{\mathcal{T}}}^{\mathcal{T}}$ are randomly sampled from a set of expert trajectories as the starting and target parameters for the matching, where $L^{\mathcal{T}}$ is a predefined hyperparameter. TM then optimizes the synthetic dataset $\mathcal{D}^{\mathcal{S}}$ by minimizing the following loss:

$$\mathcal{L} = \frac{\left\| \theta_{t+L^{\mathcal{S}}}^{\mathcal{S}} - \theta_{t+L^{\mathcal{T}}}^{\mathcal{T}} \right\|^2}{\left\| \theta_t^{\mathcal{T}} - \theta_{t+L^{\mathcal{T}}}^{\mathcal{T}} \right\|^2}, \tag{4}$$

where $L^{\mathcal{T}}$ is a preset hyperparameter, and $\hat{\theta}_{t+L^{\mathcal{T}}}$ is obtained through inner optimization with the cross-entropy loss and the trainable learning rate α.

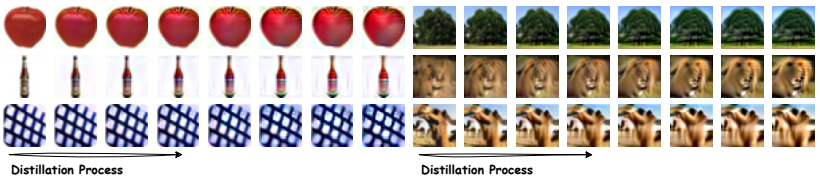
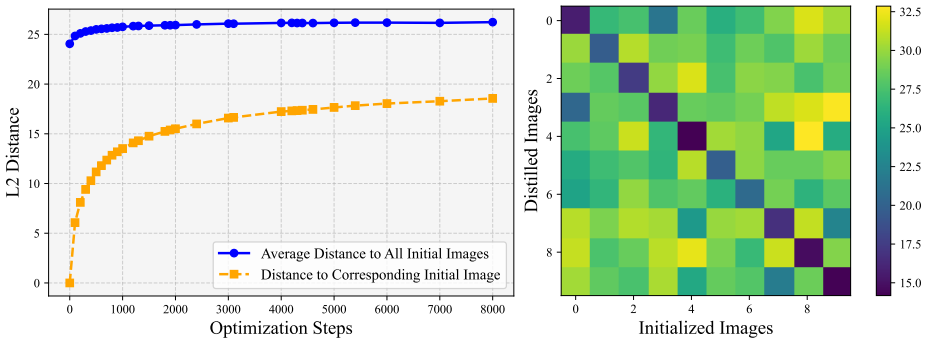


Fig. 2: We visualized the evolution of images during the distillation by FTD [8]. As the number of distillation steps increases, the process introduces only minimal adjustments to the original images, preserving most of the inherent information and features.



(a) Comparison of Average and Specific Image Difference During Optimization

(b) Heatmap of Distilled vs. Initialized Image Difference for a Specific Category

Fig. 3: (a) Image similarity measured by the L2 norm of feature distances. The yellow curve shows the similarity between the distilled image and its initial counterpart, while the blue curve shows average similarity with the same class images. (b) The similarity between distilled and initial images within the same class, with identical numbers representing corresponding images.

3.2 Initial dependency in dataset distillation

The dataset distillation method starts by selecting k image per class (IPC) from the original dataset as the initialization of the synthetic set. As the dataset distillation progresses, we observe that the optimized synthetic images remain highly similar to the initialized ones, as shown in Fig. 2. We can see that the general outline and the color are preserved, but some details appear blurred and distorted. Visually, it is not clear that an image incorporates more information. Therefore, we believe the dataset distillation method strongly depends on the initial images.

We use a pre-trained neural network as a feature encoder f_θ to quantify our findings. For synthetic images at different training stages, we use the encoder to project them into feature space. We measure similarity by calculating the L2 norm of the difference between their feature vectors. As shown in Fig. 3 (a), the L2 distances between the synthetic images and the corresponding initialized images show a limited variation during the training process and are much smaller

than the distances to other images of the same category. It is worth noting that the performance of the synthetic image converges after about 5000 optimization steps. However, the continued distillation process still adjusts the image pixels, resulting in a more significant gap between the synthetic set and the initial image. In addition to this, we use a heatmap in Fig. 3 (b) to visualize the degree of similarity between distilled and initialized images in a category. It can be seen that the corresponding images exhibit significant similarity compared to the non-corresponding images.

Based on this phenomenon, we believe that the initialized images significantly affect the distillation effect of the dataset. The following will describe our Initialization Improved Dataset Distillation (IIDD) approach.

3.3 Initialization Improved Dataset Distillation (IIDD)

In this section, we will introduce our IIDD method. As a simple and plug-and-play data distillation enhancement technique, our method focuses on initializing the synthetic set of images. Based on coreset selection methods [12], we introduce Representative Image Sampler (RIS) for IIDD.

First, we pre-train a network $\mathrm{E}(x;\theta) : \mathcal{D}^{\mathcal{T}} \rightarrow \mathbb{R}^d$ on the original dataset using several training steps (i.e., around 10% of the total training process). Then, each image is feature-encoded based on the features and gradients of the original dataset within this network.

$$\mathbf{f}_i = E(x_i;\theta) \quad , \quad \mathbf{g}_i = \nabla_\theta \mathcal{L}(x_i;\theta) \tag{5}$$

Finally, we select representative image samples from a large amount of original data based on information gain or feature distribution location, serving as the initialization for the distillation method. Specifically, we employ the GraphCut and K-center methods.

The GraphCut method treats the task of selecting representative samples as a graph partitioning problem. We build the graph based on the $\mathbf{f}_i$ or $\mathbf{g}_i$ get from Equ. 5. GraphCut selects a subset $\mathcal{D}^{\mathcal{S}}$ that ensures diversity and representativeness. GraphCut maximizes the cut value defined as:

$$\mathrm{Cut}(\mathcal{D}^{\mathcal{S}}) = \sum_{x_i, x_j} \mathcal{D}(x_i, x_j), \quad s.t. \quad x_i \in \mathcal{D}^{\mathcal{T}} \setminus \mathcal{D}^{\mathcal{S}},\, x_j \in \mathcal{D}^{\mathcal{S}}, \tag{6}$$

where $\mathcal{D}(\cdot,\cdot)$ is the distance function. By optimizing $\mathcal{D}^{\mathcal{S}}$, the GraphCut method effectively selects a subset that is most informative, capturing the key characteristics and diversity of the entire dataset.

The K-center method attempts to solve the minimax facility location problem, which involves selecting k samples as $\mathcal{D}^{\mathcal{S}}$ from the entire dataset $\mathcal{D}^{\mathcal{T}}$, such that the most significant distance between a data point in $\mathcal{D}^{\mathcal{T}} \setminus \mathcal{D}^{\mathcal{S}}$ and its closest data point in $\mathcal{D}^{\mathcal{S}}$ is minimized. The objective can be expressed by first defining the maximum of the minimum distances as:

$$\max_{x_i} \min_{x_j} \mathcal{D}(x_i, x_j), \quad s.t. \quad x_i \in \mathcal{D}^{\mathcal{T}} \setminus \mathcal{D}^{\mathcal{S}},\, x_j \in \mathcal{D}^{\mathcal{S}}, \tag{7}$$

and then minimizing this value over all possible selections of $\mathcal{D}^{\mathcal{S}} \subseteq \mathcal{D}^{\mathcal{T}}$.

This is an NP-hard problem, and we use a greedy approximation algorithm known as K-center Greedy. We start by selecting an initial center from the dataset and adding it to the set $\mathcal{D}^{\mathcal{S}}$. It then iteratively selects the point that maximizes the minimum distance to the current centers in $\mathcal{D}^{\mathcal{S}}$, repeating this until k centers are selected.

Our method is practiced based on IDM [30] and FTD [8]. In addition to RIS, in the implementation process, we referred to FYI [21] and used horizontal flipping data augmentation to alleviate the bilateral equivalence issue in dataset distillation. Additionally, inspired by BACON [32], we employ a dataset distillation loss function based on Bayesian optimization (BALoss), which holds promise for providing a better matching objective for DM-type methods.

4 Experiments

In Sec. 4.1, we outline the details of our experiments. We present the main results of our method in Sec. 4.2, including the improvements over two existing approaches. Additionally, we compare IIDD with state-of-the-art methods. Finally, in Sec. 4.3, we conduct ablation studies on components of our process and compare the results of different Representative Image Sampler (RIS) strategies.

4.1 Setup

In this section, we present the experimental setup used in this paper. To ensure a fair comparison, we consider two different experimental settings. The first setting (Competition setting) is based on the *ECCV 2024 Dataset Distillation Challenge*, while the second (Research Setting) follows the conventional settings used in previous works [30].

Competition Setting. Based on the official test [9], we employ DSA [26] and a fixed initial learning rate, and we are restricted from using soft labels and image partitioning enhancement algorithms. Many existing algorithms show a significant drop in performance under this setting compared to the results reported in their respective papers.

Research Setting. In this setting, we adopt various augmentation strategies as provided in IDM, allowing for a more direct comparison of our method with popular state-of-the-art algorithms.

Dataset. By the competition requirements, our experiments are conducted on the CIFAR-100 and Tiny-ImageNet datasets, with the compressed datasets set to IPC-10, where IPC stands for images per category. The CIFAR-100 dataset has a resolution of 32x32, while Tiny-ImageNet has a resolution of 64x64.

Implementation Details. We adhere to the IDM implementation for setting most hyperparameters in our method. Specifically, we employ the same SGD optimizer configuration during model training for data condensation and evaluation, with a learning rate of 0.01, momentum of 0.9, and weight decay of 0.0005. We use a learning rate of 0.2 and momentum of 0.5 to optimize the condensed

Table 1: Comparison of our method with the baseline under the Competition Setting

Optimize Objective	Method	CIFAR-100 (IPC-10)	Tiny-IN (IPC-10)
DM based method	IDM (baseline)	34.95±0.4	19.54±0.3
	IIDD-DM	35.91±0.2	**21.31**±0.4
TM based method	FTD (baseline)	37.56±0.2	15.07±0.5
	IIDD-TM	**38.71±0.1**	16.61±0.4

set. Additionally, we set the regularization parameter λ_{reg} for the cross-entropy loss at 0.5, corresponding to 10 condensed images per class. All experiments were conducted on a single RTX 4090 GPU with 24GB of memory.

4.2 Main result

We applied IIDD to both DM-based (IDM) and TM-based (FTD) methods. All results showed improvements over the baseline performance. The experiments in Tab. 1 were conducted under the Competition setting. On CIFAR-100, our method achieved improvements of 0.96 and 1.15 on IDM and FTD, respectively, and on Tiny-ImageNet, it achieved improvements of 1.77 and 1.54. Notably, we observed that on CIFAR-100, TM-based methods outperformed DM-based methods, while on Tiny-ImageNet, the opposite conclusion was drawn.

Table 2: Comparison with state-of-the-art dataset distillation methods

Method	Venue	CIFAR-100		
IPC		1	10	50
Ratio (%)		0.2	2	10
LD [3]	NeurIPS'20	11.5±0.4	-	-
DC [27]	ICLR'21	12.8±0.3	25.2±0.3	-
DSA [26]	ICML'21	13.9±0.3	32.3±0.3	42.8±0.4
KIP [17]	NeurIPS'21	12.0±0.2	29.0±0.2	-
CAFE [23]	CVPR'22	12.9±0.3	27.8±0.3	37.9±0.3
DCC [14]	ICML'22	14.6±0.3	33.5±0.3	39.3±0.4
DataDAM [19]	ICCV'23	14.5±0.5	34.8±0.5	49.4±0.3
DM [29]	WACV'23	11.4±0.3	29.7±0.3	43.0±0.4
FTD [8]	CVPR'23	25.2±0.3	43.4±0.3	50.7±0.3
IDM [30]	CVPR'23	23.1±0.2	44.7±0.1	49.9±0.2
IID [7]	CVPR'24	24.6±0.1	45.7±0.4	**51.3±0.4**
IIDD (Ous)	ECCVW'24	**27.7±0.2**	**46.8±0.4**	50.64±0.2
Whole Dataset			56.2±0.3	

To better compare our method with state-of-the-art techniques, we conducted experiments on CIFAR-100 under IPC settings of 1, 10, and 50, following the

experimental setup of IDM. The experiments were based on the Research Setting, and the results are shown in Tab. 2. IIDD was built based on IDM, which outperformed in IPC-1 and IPC-10, notably achieving a notable +3.1 improvement in the IPC-1 setting. In the IPC-50 experiment, our method performed worse than IID. This may be due to the decreasing difference between the representative samples of each category and the randomly selected samples as the synthetic set size increases.

4.3 Ablation Study

Effect of IIDD Components. We conduct the ablation of several critical components of IIDD based on IDM. All the experiments on CIFAR-100 and Tiny-ImageNet are under Competition Setting with IPC=10. As shown in Tab. 3, insert RIS can significantly improve performance on all the datasets. Besides, BALoss from BACON also offers a slight improvement. After using the horizontal flipping data augmentation, the IIDD shows a +0.96 and +1.77 improvement.

Different RIS Methods. This section compares three different samplers for RIS. After obtaining the features using Equ. 5, we applied the methods of Random selection, GraphCut, and k-center on the CIFAR-100 dataset with IPC set to 10. The ablation experiments were conducted under the competition setting, and the results are presented in Tab. 4. It can be observed that the k-center method performed the best, followed by GraphCut. Both methods significantly outperformed Random selection, demonstrating the effectiveness of our initialization approach.

Table 3: Ablation on components in IPC-10

Method	CIFAR-100	Tiny-IN
IDM	34.95±0.4	19.54±0.3
IDM+RIS	35.70±0.3	20.07±0.5
IDM+RIS+BALoss	35.78±0.4	20.35±0.1
IDM+RIS+BALoss+FYI	35.91±0.2	21.31±0.4

Table 4: RIS on CIFAR-100

RIS Method	CIFAR-100
random	35.06±0.3
GraphCut	35.84±0.1
K-center	35.91±0.2

5 Conclusion

This paper presents IIDD, a straightforward and plug-and-play technique for enhancing dataset distillation. Our approach utilizes RIS to initialize the synthetic dataset, complemented by horizontal flipping data augmentation and BALoss to improve performance in dataset distillation tasks. Notably, IIDD demonstrates significant improvements across both DM and TM-based baseline methods. This method achieved 1st place on Tiny ImageNet and 2nd place overall in *The First Dataset Distillation Challenge at the ECCV 2024 Workshop*.

References

1. Agarwal, P.K., Har-Peled, S., Varadarajan, K.R.: Approximating extent measures of points. J. ACM **51**, 606–635 (2004), https://api.semanticscholar.org/CorpusID:3141365
2. Aljundi, R., Lin, M., Goujaud, B., Bengio, Y.: Gradient based sample selection for online continual learning. In: Proc. of Advances in Neural Information Processing Systems (2019), https://api.semanticscholar.org/CorpusID:195345359
3. Bohdal, O., Yang, Y., Hospedales, T.: Flexible dataset distillation: Learn labels instead of images. In: Proc. of Advances in Neural Information Processing Systems (2020)
4. Cazenavette, G., Wang, T., Torralba, A., Efros, A.A., Zhu, J.Y.: Dataset distillation by matching training trajectories. In: Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition. pp. 10718–10727 (2022)
5. Cazenavette, G., Wang, T., Torralba, A., Efros, A.A., Zhu, J.Y.: Generalizing dataset distillation via deep generative prior. In: Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition. pp. 3739–3748 (2023)
6. Cui, J., Wang, R., Si, S., Hsieh, C.J.: Scaling up dataset distillation to imagenet-1k with constant memory. In: Proc. of Intl. Conf. on Machine Learning. pp. 6565–6590 (2023)
7. Deng, W., Li, W., Ding, T., Wang, L., Zhang, H., Huang, K., Huo, J., Gao, Y.: Exploiting inter-sample and inter-feature relations in dataset distillation. In: Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition. pp. 17057–17066 (2024)
8. Du, J., Jiang, Y., Tan, V.T.F., Zhou, J.T., Li, H.: Minimizing the accumulated trajectory error to improve dataset distillation. In: Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition. pp. 3749–3758 (2023)
9. ECCV: The first dataset distillation challenge (2024), https://dd-challenge-main.vercel.app/, accessed: 2024-08-15
10. Feldman, D.: Introduction to core-sets: an updated survey. ArXiv **abs/2011.09384** (2020), https://api.semanticscholar.org/CorpusID:227011958
11. Feldman, D., Faulkner, M., Krause, A.: Scalable training of mixture models via coresets. In: Proc. of Advances in Neural Information Processing Systems (2011), https://api.semanticscholar.org/CorpusID:7162940
12. Guo, C., Zhao, B., Bai, Y.: Deepcore: A comprehensive library for coreset selection in deep learning. In: International Conference on Database and Expert Systems Applications. pp. 181–195. Springer (2022)
13. Guo, Z., Wang, K., Cazenavette, G., Li, H., Zhang, K., You, Y.: Towards lossless dataset distillation via difficulty-aligned trajectory matching. In: Proc. of Intl. Conf. on Learning Representations (2024)
14. Lee, S., Chun, S., Jung, S., Yun, S., Yoon, S.: Dataset condensation with contrastive signals. In: Proc. of Intl. Conf. on Machine Learning. pp. 12352–12364 (2022)
15. Liu, Y., Gu, J., Wang, K., Zhu, Z., Jiang, W., You, Y.: DREAM: Efficient dataset distillation by representative matching. In: Porc. of IEEE Intl. Conf. on Computer Vision. pp. 17314–17324 (2023)
16. Liu, Y., Gu, J., Wang, K., Zhu, Z., Zhang, K., Jiang, W., You, Y.: DREAM+: Efficient dataset distillation by bidirectional representative matching. arXiv preprint arXiv:2310.15052 (2023)

17. Nguyen, T., Novak, R., Xiao, L., Lee, J.: Dataset distillation with infinitely wide convolutional networks. In: Proc. of Advances in Neural Information Processing Systems. pp. 5186–5198 (2021)
18. Rebuffi, S.A., Kolesnikov, A., Sperl, G., Lampert, C.H.: icarl: Incremental classifier and representation learning. Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition pp. 5533–5542 (2016), https://api.semanticscholar.org/CorpusID:206596260
19. Sajedi, A., Khaki, S., Amjadian, E., Liu, L.Z., Lawryshyn, Y.A., Plataniotis, K.N.: DataDAM: Efficient dataset distillation with attention matching. In: Porc. of IEEE Intl. Conf. on Computer Vision. pp. 17097–17107 (2023)
20. Sener, O., Savarese, S.: Active learning for convolutional neural networks: A core-set approach. arXiv: Machine Learning (2017), https://api.semanticscholar.org/CorpusID:3383786
21. Son, B., Oh, Y., Baek, D., Ham, B.: Fyi: Flip your images for dataset distillation. In: Proceedings of the European Conference on Computer Vision (ECCV) (2024)
22. Toneva, M., Sordoni, A., des Combes, R.T., Trischler, A., Bengio, Y., Gordon, G.J.: An empirical study of example forgetting during deep neural network learning. ArXiv **abs/1812.05159** (2018), https://api.semanticscholar.org/CorpusID:55481903
23. Wang, K., Zhao, B., Peng, X., Zhu, Z., Yang, S., Wang, S., Huang, G., Bilen, H., Wang, X., You, Y.: CAFE: Learning to condense dataset by aligning features. In: Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition. pp. 12196–12205 (2022)
24. Wang, T., Zhu, J.Y., Torralba, A., Efros, A.A.: Dataset distillation. ArXiv **abs/1811.10959** (2018), https://api.semanticscholar.org/CorpusID:53763883
25. Yin, Z., Xing, E., Shen, Z.: Squeeze, recover and relabel: Dataset condensation at imagenet scale from a new perspective. In: Proc. of Advances in Neural Information Processing Systems (2023)
26. Zhao, B., Bilen, H.: Dataset condensation with differentiable siamese augmentation. In: Proc. of Intl. Conf. on Machine Learning. pp. 12674–12685 (2021)
27. Zhao, B., Bilen, H.: Dataset condensation with gradient matching. In: Proc. of Intl. Conf. on Learning Representations (2021)
28. Zhao, B., Bilen, H.: Synthesizing informative training samples with gan. In: Proc. of Advances in Neural Information Processing Systems (2022)
29. Zhao, B., Bilen, H.: Dataset condensation with distribution matching. In: Proc. of IEEE Winter Conf. on Applications of Computer Vision. pp. 6514–6523 (2023)
30. Zhao, G., Li, G., Qin, Y., Yu, Y.: Improved distribution matching for dataset condensation. In: Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition. pp. 7856–7865 (2023)
31. Zhou, Y., Nezhadarya, E., Ba, J.: Dataset distillation using neural feature regression. Proc. of Advances in Neural Information Processing Systems **35**, 9813–9827 (2022)
32. Zhou, Z., Zhao, H., Cheng, G., Li, X., Lyu, S., Feng, W., Zhao, Q.: Bacon: Bayesian optimal condensation framework for dataset distillation. arXiv preprint arXiv:2406.01112 (2024)