Flash-DD: An Ultra Parameter-Efficient Approach to Dataset Distillation

Ruonan Yu[1], Songhua Liu[2,1⋆], and Xinchao Wang[1⋆]

[1] National University of Singapore, Singapore
[2] Shanghai Jiao Tong University, China
ruonan@u.nus.edu, liusonghua@sjtu.edu.cn, xinchao@nus.edu.sg

Abstract. Dataset distillation (DD) aims to create a smaller dataset that encapsulates the essential knowledge of a larger dataset, thereby reducing storage demands and accelerating downstream training. For large-scale dataset distillation, state-of-the-art methods achieve satisfactory performance by using soft labels generated by well-trained teacher models during downstream training. However, it will cause some issues: (1) a substantial amount of additional storage is required to retain the teacher models, often significantly exceeding the storage needed for the synthetic images; (2) generating labels through these teacher models slows down the downstream training process, counteracting the efficiency goals of dataset distillation; and (3) downstream training guided by these teacher models, according to our studies, yields suboptimal performance. Focusing on these drawbacks, in this paper, we propose plug-and-play parameter-efficient label generation techniques for dataset distillation, which maximizes the benefits of limited model parameters and can be generalized to different DD methods, datasets, and settings. Specifically, we propose a DD-oriented model parameter reduction method that automatically determines the optimal capacity of teacher models and eliminates redundant parameters for dataset distillation tasks. Furthermore, for additional parameter space, we turn to model ensemble strategies and propose guidelines to optimize the utilization efficiency of the additional space. Compared to the state-of-the-art methods, Flash-DD requires only **0.03%** of the additional storage and significantly accelerates downstream label generation by **843.81**× while maintaining comparable performance. Alternatively, with a mere **1.8%** storage budget, it can boost accuracy by up to **13.4%** over previous leading methods.

Keywords: Dataset distillation · Parameter-efficient label generation · Efficient training

1 Introduction

Dataset distillation (DD) [32] aims to generate a significantly smaller synthetic dataset that retains the representative knowledge of the original large-scale

⋆ Corresponding authors.

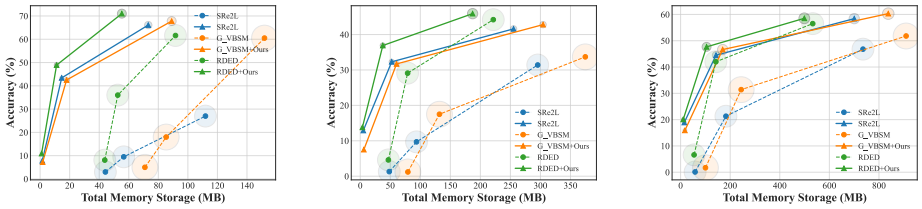


Fig. 1: Comparison of downstream performance and required storage (both synthetic images and teacher model) for state-of-the-art methods: SRe^{2}L, G_VBSM, RDED, and those methods with Ours. Points from left to right correspond to IPC 1, 10, and 50. Circle size denotes the GFLOPs of teacher model used in downstream training.

dataset, maintaining downstream tasks performance while alleviating the storage and transmission burden, and accelerating the process of downstream training. After a period of rapid development in DD field, current dataset distillation methods have achieved lossless performance [8] on low-resolution, small-scale datasets (*e.g.*, CIFAR-10 and CIFAR-100). However, for high-resolution, large-scale datasets like ImageNet-1K, increased complexity and computational demands make distillation more challenging, leading to significant performance degradation at high compression rates.

To mitigate this, state-of-the-art large-scale dataset distillation methods [24, 29, 33, 34] employ well-defined data augmentation strategies and and leverage soft-label supervision during downstream training to enhance data diversity and assure the informative, accurate knowledge for downstream tasks. Soft-label can be generated in two functionally equivalent ways, as both rely on teacher-provided probabilistic targets. In the online setting, at each training epoch, synthetic samples are augmented and the teacher model(s) is queried to generate epoch-specific soft labels. In the offline setting, these teacher predictions are pre-generated and stored together with the distilled images. Consequently, maintaining soft labels for all augmented instances across training epochs leads to substantial storage overhead that scales with the number of epochs. In this work, we focus on the online setting.

However, this paradigm introduces notable limitations. Firstly, retaining the teacher model(s) throughout downstream training incurs additional memory overhead, which often significantly exceeds that of the synthetic images and thereby undermines the storage-efficiency objective of dataset distillation. Additionally, it will introduce extensive GPU time and computational overhead for incorporating soft label generation process during downstream training. To make more accurate knowledge transfer and convey more information to the downstream tasks, each augmented image will be passed through the teacher model(s) at every iteration to generate the corresponding soft label, which considerably slows down the training process. It also counteracts the efficiency goals of dataset distillation. Lastly, as illustrated in Fig. 1, downstream training guided by the

well-trained teacher model(s) yields suboptimal performance, further highlighting the limitations of current soft label generation techniques.

In this paper, we propose plug-and-play, memory-efficient soft label generation techniques for dataset distillation. Our method also establishes a paradigm, offers optimal solutions under different storage conditions and is adaptable to various dataset distillation methods, datasets, and settings. Specifically, we propose a DD-oriented model parameter reduction method that automatically determines the optimal capacity of teacher models to downstream tasks. Further, to optimize the spatial efficiency of the additional parameter space, we systematically restructure the parameter space to extend and optimize the initially small effective portion across the entire capacity, and also the eliminate potential negative effects raised by the redundant and irrelevant parameters. Here, we turn to model ensemble strategies to satisfy the requirements of the different downstream training stages and provide guidelines to obtain feasible and diverse knowledge for downstream tasks. Extensive experiments validate the efficacy of our proposed method. For example, with just 0.03% of the original extra storage, we maintain the performance while speed up the label generation process 843.81×. Also, with limited storage, our proposed method outperforms the previous state-of-the-art method up to 13.4%.

To sum up, our contributions are listed as follows:

- We propose plug-and-play parameter-efficient soft label generation techniques for dataset distillation to fully leverage the benefits of the limited model parameters. We also establish a paradigm for achieving optimal solutions for different storage conditions, which can be applied across different DD methods, datasets, and settings.
- We propose a DD-oriented model parameter reduction method that automatically determines the optimal capacity of teacher models for downstream training and removes the task-irrelevant parameters to accelerate the training process and reduce the storage space.
- We systematically re-organize the additional parameter space to optimize the spatial efficiency. We propose guidelines for model ensemble strategies to adapt to the different downstream training stages while providing feasible and diverse knowledge.
- We conduct extensive experiments to demonstrate the effectiveness of our method. Results show that our method only requires 0.03% extra storage to achieve the same performance of the previous state-of-the-art methods while accelerating the soft label generation 843.81×. With 1.8% extra storage, our method improves the previous state-of-the-art methods up to 13.4%.

2 Related Works

Dataset distillation, first proposed by Wang *et al.* [32], distills the large-scale original dataset into a much smaller synthetic one without compromising its performance. After a period of development, the field of dataset distillation has undergone exploration improvements in many aspects. For instance, in terms

of optimization objectives, some methods [6, 15, 16, 18, 19, 32, 41] optimize the distilled datasets according to the model performance across the original dataset and the synthetic one as defined by dataset distillation. Some methods [2,4,7,11, 12,26,36,38] optimize the distilled datasets by matching the training trajectories or loss curve of the model training process on two datasets, while some [22,31,35, 37,39] pursue optimization by aligning the distributions or attention maps of the datasets. Moreover, numerous efforts have contributed to progress in directions such as dataset parameterization [3, 9, 13, 27], label distillation [1, 28].

However, dataset distillation often suffers from performance drops on downstream tasks when applied to high-resolution, large-scale datasets. Recent state-of-the-art methods [24, 25, 29, 33, 34] adopt pre-trained teacher model(s) to generate soft labels either online or in advance for each epoch during downstream training to enrich and strengthen the synthetic dataset. This strategy leverages teacher knowledge, but increases storage requirements and prolongs downstream training, contrary to the original goals of dataset distillation. To address these issues, our proposed method offers training-efficient and memory-flexible solutions. Specifically, we introduce a plug-and-play DD-oriented model parameter reduction technique that automatically determines the optimal teacher capacity for downstream tasks, reducing storage and accelerating training. Additionally, we provide guidelines for adaptive model ensemble strategies in soft label generation, further improving parameter efficiency and downstream performance.

3 Method

3.1 Preliminary

For high-resolution, large-scale datasets, current state-of-the-art dataset distillation methods [24, 29, 33, 34] adopt well-defined augmentation strategies and employ teacher model(s) to generate soft labels for each augmented image during downstream training to compensate for the loss of knowledge. To be noted, our paper is under online soft-label generation settings, and the downstream training process can be formulated as follows:

$$\begin{aligned} \theta_S^{(t)} &= \theta_S^{(t-1)} - \alpha \nabla \mathcal{L}((X_s^{(t)\prime}, Y_s^{(t)\prime}); \theta_S^{(t-1)}), \\ X_s^{(t)\prime} = \mathcal{A}ug(X_s, t; \phi), Y_s^{(t)\prime} &= \frac{1}{|\Theta_T|} \sum_{\theta_T \sim \Theta_T} f(X_s^{(t)\prime}; \theta_T). \end{aligned} \tag{1}$$

Here, $\theta_{\mathcal{S}}$ is the downstream model, α is the learning rate, and $\mathcal{L}(\cdot; \theta_{\mathcal{S}})$ is the loss function for downstream tasks. During the downstream training process, the synthetic images X_s will be augmented by the well-defined augmented function $\mathcal{A}ug$ with the parameter ϕ to generate the training samples $X_s^{(t)\prime}$ for epoch t, and $|X_s^{(t)\prime}| = |X_s|$. The corresponding labels $Y_s^{(t)\prime}$ are generated by the teacher model(s) Θ_T. Although soft labels greatly improve downstream performance, they also increase storage requirements for teacher model(s) and introduce downstream training latency. Our method seeks an optimal balance among training efficiency, storage, and performance.

Algorithm 1 DD-Oriented Model Parameter Reduction

Input: Synthetic dataset D_S, full-sized teacher θ_{T_0}, $\tau(D_S)$
Output: Capacity matched θ_T^*

$i = 0$
while $|\mathcal{A}cc(\theta_{T_i}) - \mathcal{A}cc(\theta_{T_0})| \leq \delta$ **do**
  $\theta_{T_{i+1}} = Param\_reduce(\theta_{T_i})$
  $i = i + 1$
end while
$\Theta_{\mathcal{T}} = \{\}$
while $\mathcal{A}cc(\theta_{T_i}) > \tau(D_S)$ **do**
  $\theta_{T_{i+1}} = Param\_reduce(\theta_{T_i})$
  Compute the loss degradation during the early training stage $\mathcal{D}is(\theta_{T_i})$.
  if $\mathcal{D}is(\theta_{T_i}) \geq \zeta$ **then**
    $\Theta_T \cup \theta_{T_i}$ // Only consider the loss degradation more than the threshold ζ.
  end if
  $i = i + 1$
end while
$\theta_T^* = \arg\max_{\theta_T} \mathcal{D}is(\theta_T) + \gamma\mathcal{A}cc(\theta_T)$, $\theta_T \in \Theta_T$
return θ_T^*

3.2 DD-Oriented Model Parameter Reduction

Current state-of-the-art large-scale dataset distillation methods leverage teacher model(s) to enhance downstream performance. However, this will introduce significant extra storage demands and slows down downstream training. To address these issues, we propose a plug-and-play DD-oriented model parameter reduction method that determines teacher model capacity, enabling efficient and effective soft label generation for better downstream generalization. The framework of our method is shown in Algorithm 1. Our method consists of two steps.

First, we construct a lightweight teacher that preserves the original accuracy by applying an iterative model parameter reduction procedure, *e.g.*, pruning, on the full-sized θ_T teacher model pretrained on the entire original dataset and stopping before noticeable degradation occurs. In our implementation, we use structured pruning with l_1-norm importance score. This step will quickly remove the redundant parameters while maintain the performance of both the teacher model and the downstream tasks. The next step is to determine the optimal amount of teacher knowledge that students can effectively utilize. We evaluate teachers with different knowledge capacities under various synthetic dataset sizes, as shown in Fig. 2. As teacher capacity increases, downstream performance first improves and then degrades. It suggests that once teacher capacity exceeds the learnable space supported by the synthetic dataset, additional knowledge becomes either redundant or difficult for students to learn from, thereby hindering effective knowledge transfer. Therefore, optimal teacher capacity should align with the knowledge capacity of the synthetic data, retaining only the most transferable and essential knowledge while filtering out overly complex or low-impact

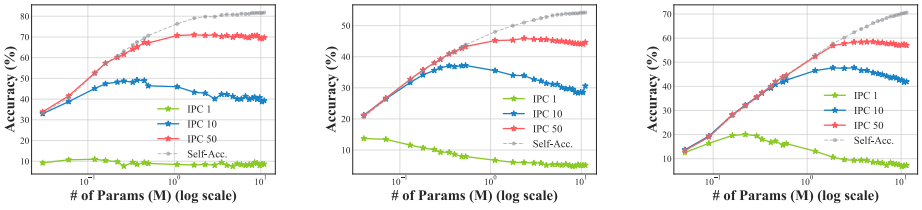


Fig. 2: Illustration of the relationships between the downstream performance, the size of the teacher, and the performance of the teacher. The experiments are conducted on three datasets ImageNet-100, Places365-Standard, and ImageNet-1K, with IPC 1, 10, and 50. For better demonstration, the # of Params (MB) is set to a log scale.

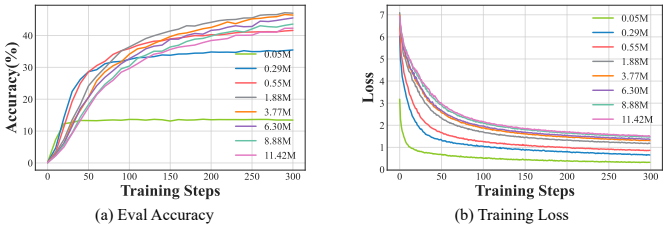


Fig. 3: The downstream evaluation loss (a) and training loss (b). The evaluation is under ImageNet-1K with IPC 10.

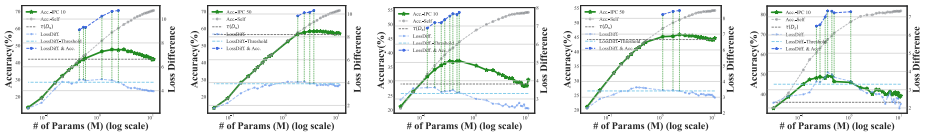
components. For downstream training,

$$\begin{aligned}\mathcal{L}(D_S;\theta_S,\theta_T) =& \mathbb{E}_{(x,y)\sim\mathcal{D}_s}[||f(x,\theta_T)-f(x,\theta_S)||_2^2+\eta\mathcal{D}_s[||f(x,\theta_S)-y||_2^2] \\ =& \mathbb{E}_{(x,y)\sim\mathcal{D}_s}[(1+\eta)||\Delta||_2^2+\eta||\epsilon||_2^2+2\eta\cdot\Delta\cdot\epsilon],\end{aligned} \tag{2}$$

where $\Delta = f(x,\theta_T) - f(x,\theta_S)$ and $\epsilon = f(x,\theta_T) - y$. The performance of θ_S is closely related to the performance of θ_T. When the teacher has limited capacity, the student can more easily fit its outputs using a small dataset D_S. However, if the teacher significantly deviates from the ground-truth distribution, the improvement brought by distillation is inherently constrained. In contrast, teachers that are better aligned with the target distribution yield larger downstream gains. As shown in Fig. 3, due to the limited size of D_S, the training loss exhibits clear trends and distinct inflection points in the early training phase. These observable dynamics enable us to further characterize the student difficulty in fitting the teacher outputs during downstream training through the convergence behavior of the training loss:

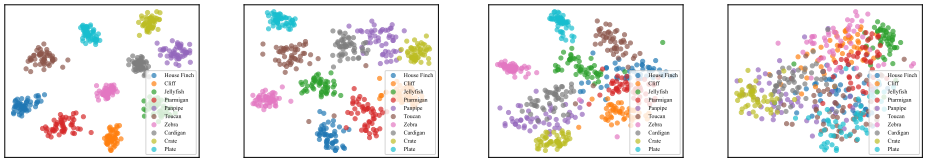
$$\mathcal{D}is(\theta_T) = \mathbb{E}_{\theta_S^{(0)}\sim\mathcal{P}_\theta}[\mathcal{L}(D_S;\theta_S^{(t)},\theta_T) - \mathcal{L}(D_S;\theta_S^{(0)},\theta_T)]. \tag{3}$$

Here, t represents epoch in the early training stage, $\mathcal{L}(\cdot)$ is the loss function for the downstream tasks. $\mathcal{D}is(\cdot)$ calculates the loss reduction during the early stage. The total number of epoch for calculating $\mathcal{D}is(\cdot)$ depends on the number



(a) IN-1K IPC 10 (b) IN-1K IPC 50 (c) Places365 IPC 10 (d) Places365 IPC 50 (e) IN-100 IPC 10

Fig. 4: The validation experiments for the criteria to determine the optimal capacity for teacher models. The experiments are conducted under the settings of ImageNet-1K and Places365-Standard with IPC 10 and 50 and ImageNet-100 with IPC 10. Here, for ImageNet-1K and Places365-Standard, we calculate the loss reduction in the early 50 epochs for downstream training, and γ here is 0.1. For ImageNet-100, the process is in the early 20 epochs and γ is 0.05.



(a) 11.42M (97.67%), 57.0% (b) 2.80M (23.97%), 58.5% (c) 0.55M (4.68%), 43.8% (d) 0.05M (0.45%), 13.4%

Fig. 5: The t-SNE analysis of ResNet-18 under different storage budgets on ImageNet-1K with IPC 50. Each subfigure title shows model size, its ratio to the original model (in parentheses), and the corresponding downstream accuracy. The analysis is performed on the models pre-softmax logits.

of classes of the original dataset. In this work, we calculate the loss reduction in the early 50 epochs for larger datasets ImageNet-1K and Places365-Standard, and 20 epochs for ImageNet-100 dataset.

During the teacher parameter reduction process, the accumulated loss curve gradually flattens, indicating that the teacher has entered a regime more suitable for the downstream student model. Here, we set the threshold for loss reduction as ζ. As an excessively complex teacher imposes an overly difficult imitation target, making it challenging for the student to approximate the teacher behavior, which results in limited loss reduction in the early stage of training. Conversely, an overly simplified teacher reduces the richness of supervisory signals, causing the student to quickly match the teacher performance. This leads to a low performance upper bound and premature convergence, after which the student is prone to overfitting due to insufficiently informative guidance. For the numerical computation of ζ, $Dis(\theta_T)$ is evaluated at discrete capacity levels, ζ can be computed numerically by examining the finite difference ratio $\frac{\Delta Dis}{\Delta|\theta_T|}$ and identifying the point where the incremental increase becomes small. Here, $\Delta Dis = Dis(\theta_{T_{i+1}}) - Dis(\theta_{T_i})$, and $\Delta|\theta_T| = |\theta_{T_{i+1}}| - |\theta_{T_i}|$. We select ζ at the point where this incremental gain becomes sufficiently small, indicating diminishing returns in accumulated loss as the teacher is further reduced. Furthermore, as shown in Fig. 2, when the accuracy of teacher model decreases, the student accuracy decreases accordingly and remains inferior. When the teacher perfor-

mance falls below $\tau(D_s)$, the student converges to the same level. Here, $\tau(D_s)$ is the downstream accuracy of the original DD method with the full teacher trained on D_s. Accordingly, we use $\tau(D_s)$ as an additional stopping criterion for teacher parameter reduction.

Proposition 1 *$\tau(Ds)$ is the accuracy of the downstream student models trained with the guidance of the original pre-trained full-size teacher model on distilled dataset D_S.*

Proof. T_0 is full-size teacher model with parameter dimension d, T_1 is reduced-capacity teacher (dimension k) obtained via parameter reduction from T_0 on the global data distribution. When $\varepsilon(S_0) \approx \varepsilon(T_1)$, the student S_0 distilled from T_0 exhibit similar generalization performance to the student S_1 distilled from T_1. The synthetic dataset is $X \in \mathbb{R}^{n\times d}$, where $n \ll d$. Let the global data covariance matrix be $\Sigma = \mathbb{E}[xx^T] \in \mathbb{R}^{d\times d}$. By spectral decomposition, we partition Σ into the top k principal components V_k and the orthogonal residual space $V_\perp$: $\Sigma = V_k \Lambda_k V_k^T + V_\perp \Lambda_\perp V_\perp^T$. Here, T_0 can be decomposed as $\theta_{T_0} = \theta_{T_1} + \theta_\delta$, where $\theta_{T_1} \triangleq V_k V_k^T \theta_{T_0}$, and $\theta_\delta \triangleq V_\perp V_\perp^T \theta_{T_0}$, $\theta_{T_1}^T \theta_\delta = 0$. After distillation, we have:

$$\begin{aligned}
\theta_{S_0}^* &= \arg\min_{\theta_S} \|X\theta_{T_0} - X\theta_S\|_2^2 = X^\dagger X \theta_{T_0}, \\
\theta_{S_1}^* &= \arg\min_{\theta_S} \|X\theta_{T_1} - X\theta_S\|_2^2 = X^\dagger X \theta_{T_1}, \\
\theta_{S_0}^* &= X^\dagger X(\theta_{T_1} + \theta_\delta) = \theta_{S_1}^* + X^\dagger X \theta_\delta,
\end{aligned} \tag{4}$$

where $\mathbb{E}[\|X\theta_\delta\|_2^2] = n \cdot \theta_\delta^T \Sigma \theta_\delta$, and $\theta_\delta^T \Sigma \theta_\delta = \theta_\delta^T (V_k \Lambda_k V_k^T + V_\perp \Lambda_\perp V_\perp^T)\theta_\delta = \theta_\delta^T (V_\perp \Lambda_\perp V_\perp^T)\theta_\delta$. As T_1 retains the core predictive capacity, the eigenvalues of the truncated residual space $\Lambda_\perp$ approach zero, such that:

$$\mathbb{E}[\|X\theta_\delta\|_2^2] \le n \cdot \lambda_{max}(\Lambda_\perp)\|\theta_\delta\|_2^2 \to 0. \tag{5}$$

The generalization error evaluated on the true distribution is $\varepsilon(S) = (\theta_S - \theta^*)^T \Sigma (\theta_S - \theta^*)$, $\varepsilon(S_0) \approx \varepsilon(S_1)$.

This rule indicates the lower bound to obtain the optimal capacity of the teacher model for downstream tasks, and provides insights to find the lowest memory costs to obtain comparable state-of-the-art performance.

Within the selected range of teacher model capacities, we further identify the optimal teacher for downstream tasks. As noted in Eq. 2, downstream performance is strongly correlated with teacher performance. Here, we introduce the teacher model own accuracy as an additional selection criterion, as follows:

$$\theta_T^* = \arg\max_{\theta_T} \mathcal{D}is(\theta_T) + \gamma \mathcal{A}cc(\theta_T),\ \theta_T \in \Theta_T, \tag{6}$$

where γ denotes the contribution of the teacher intrinsic accuracy in selecting the optimal teacher. We set γ to make the accuracy term $Acc(\theta_T)$ and the accumulated-loss term $Dis(\theta_T)$ lie on comparable scales, ensuring that both factors contribute meaningfully when evaluating teacher capacity. Any reasonable

value of γ that brings the two terms onto the same scale is acceptable for this purpose. In paper, we set $\gamma = 0.1$ for ImageNet-1K and Places365-Standard, and $\gamma = 0.05$ for ImageNet-100 for all IPC settings. We also conduct the validation experiments to demonstrate the effectiveness of the criteria, as shown in Fig. 4. To better understand the characterizes of the selected teacher, we visualize the output distributions of teachers with different pruning rates using t-SNE, as shown in Fig. 5. The teacher that achieves the best downstream performance exhibits well-defined class clusters with clear boundaries and preserves meaningful inter-class structure, whereas suboptimal teachers produce mixed or collapsed clusters. These observations suggest that an effective teacher must balance reliable class discrimination with informative soft predictions. In particular, the correct class should receive the highest probability, while non-target classes retain small but non-zero probabilities to encode inter-class similarity. Such distributions enable the student to learn both clear decision boundaries and relationships among classes.

3.3 Guidelines For Ensemble Label Generation

Previously, we demonstrated that aligning the knowledge capacity between the teacher and the synthetic dataset is particularly effective under constrained storage budgets. In such highly limited settings, *e.g.*, ImageNet-1K IPC-10, selecting a single pruned teacher with an appropriate capacity is sufficient to achieve strong downstream performance. However, when the available storage budget increases, simply maintaining the single optimal teacher capacity leads to underutilization of the additional space, thereby limiting potential performance gains. To address this issue, we shift from the single-teacher setting to an ensemble-based label generation strategy to avoid redundant or ineffective knowledge and provides more informative guidance for downstream tasks, enabling maximal use of available space. Specifically, we restructure the pre-defined space M into several sub-modules and introduce complementary sub-teachers under carefully designed constraints:

$$f(x; \overline{\Theta}_T^{(t)}) = \sum_{\theta_{T_i} \in \Theta_T} \mu_i^{(t)} f(x; \theta_{T_i}). \tag{7}$$

Here, $M = \sum_{\theta_{T_i} \in \Theta_T} \mathcal{C}(\theta_{T_i})$, $\mathcal{C}(\cdot)$ measures the size of the model. $\overline{\Theta}_T^{(t)}$ represents the ensembled teachers for epoch t. $\mu_i^{(t)}$ is the weight for model θ_{T_i} for epoch t, and is initialized uniformly as $\mu_i = \frac{1}{|\Theta|}$, $|\Theta|$ is the number of pruned teachers contributed to the emsembled output. It could be updated during the downstream training optionally.

To enable effective subspace division, each teacher should maintain appropriate knowledge capacity for reliable downstream transfer, as discussed in Section 3.2. While an ensemble introduces complementary perspectives, excessive capacity discrepancies may lead to conflicting predictions and unstable supervision, especially when individual teachers are under-capacitated (as shown in

Table 1: Performance and efficiency comparison applying Flash-DD on large-scale dataset distillation methods under highly constrained budget (only one single teacher), across three datasets ImageNet-100, Places365-Standard, and ImageNet-1K. Here, the extra memory represents for the storage part excluding the synthetic images. All synthetic datasets are generated by official code without further update. ResNet-18 is adopted as the base backbone for teacher model and student model. * represents that the results are from our reproduction.

Methods		ImageNet-100			Places365-Standard			ImageNet-1K		
		1	10	50	1	10	50	1	10	50
SRe^{2}L	Acc.	3.0 ± 0.3	9.5 ± 0.4	27.0 ± 0.4	1.3 ± 0.2*	9.7 ± 0.1*	31.4 ± 0.1*	0.1 ± 0.1	21.3 ± 0.6	46.8 ± 0.2
	Extra Mem.	42.8MB	42.8MB	42.8MB	43.3MB	43.3MB	43.3MB	44.7MB	44.7MB	44.7MB
w/ Ours	Acc.	**7.8 ± 0.6**	**43.4 ± 0.1**	**66.1 ± 0.1**	**12.9 ± 0.2**	**32.3 ± 0.3**	**41.6 ± 0.1**	**18.9 ± 0.7**	**44.5 ± 0.1**	**58.4 ± 0.1**
		↑ 4.8	**↑ 33.9**	**↑ 39.1**	**↑ 11.6**	**↑ 22.6**	**↑ 10.2**	**↑ 18.8**	**↑ 23.2**	**↑ 11.6**
	Extra Mem.	**0.1MB**	**0.8MB**	**4.1MB**	**0.1MB**	**2.0MB**	**3.1MB**	**1.1MB**	**4.7MB**	**9.5MB**
		0.3%	**2.0%**	**9.6%**	**0.3%**	**4.6%**	**7.1%**	**2.5%**	**10.5%**	**21.3%**
G_VBSM	Acc.	5.0 ± 0.1*	18.0 ± 0.4*	60.5 ± 0.1*	1.2 ± 0.1*	17.5 ± 0.1*	33.7 ± 0.6*	1.7 ± 0.1*	31.4 ± 0.5	51.8 ± 0.4
	Extra Mem.	69.3MB	69.3MB	69.3MB	73.4MB	73.4MB	73.4MB	84.1MB	84.1MB	84.1MB
w/ Ours	Acc.	**7.2 ± 0.6**	**42.5 ± 0.2**	**67.8 ± 0.6**	**7.5 ± 0.3**	**31.7 ± 0.2**	**42.8 ± 0.7**	**15.9 ± 0.1**	**46.6 ± 0.2**	**60.3 ± 0.1**
		↑ 2.2	**↑ 24.5**	**↑ 7.3**	**↑ 6.3**	**↑ 14.2**	**↑ 9.1**	**↑ 14.2**	**↑ 15.2**	**↑ 8.5**
	Extra Mem.	**0.1MB**	**1.9MB**	**6.5MB**	**0.1MB**	**2.0MB**	**3.1MB**	**1.1MB**	**9.5MB**	**12.1MB**
		0.2%	**2.7%**	**9.4%**	**0.2%**	**2.7%**	**4.2%**	**1.3%**	**11.3%**	**14.4%**
RDED	Acc.	8.1 ± 0.3	36.0 ± 0.3	61.6 ± 0.1	4.6 ± 0.2*	29.1 ± 0.3*	44.2 ± 0.2*	6.6 ± 0.2	42.0 ± 0.1	56.5 ± 0.1
	Extra Mem.	42.8MB	42.8MB	42.8MB	43.3MB	43.3MB	43.3MB	44.7MB	44.7MB	44.7MB
w/ Ours	Acc.	**10.9 ± 0.3**	**48.9 ± 0.3**	**71.0 ± 0.2**	**13.8 ± 0.1**	**36.9 ± 0.2**	**45.9 ± 0.1**	**20.0 ± 0.1**	**47.5 ± 0.2**	**58.5 ± 0.1**
		↑ 2.8	**↑ 12.9**	**↑ 9.4**	**↑ 9.2**	**↑ 7.7**	**↑ 1.7**	**↑ 13.4**	**↑ 5.5**	**↑ 2.0**
	Extra Mem.	**0.04MB**	**1.4MB**	**6.5MB**	**0.1MB**	**2.0MB**	**8.9MB**	**0.8MB**	**7.2MB**	**10.7MB**
		0.1%	**3.3%**	**15.2%**	**0.3%**	**4.6%**	**20.6%**	**1.8%**	**16.1%**	**23.9%**

Fig. 5). Thus, when capacity of each teacher is comparable, and with feasible knowledge, leveraging additional different perspectives can offer benefits for downstream tasks. Motivated by the observation, the criteria can be formulated as follows:

$$\max_{\theta_{T_i}\in\Theta_T} \mathcal{C}(\theta_{T_i}) - \min_{\theta_{T_i}\in\Theta_T} \mathcal{C}(\theta_{T_i}) < \epsilon_1, \sum_{\theta_{T_i},\theta_{T_j}\in\Theta_T} |\mathcal{C}(\theta_{T_i}) - \mathcal{C}(\theta_{T_j})| < \epsilon_2,$$
$$\forall\theta_{T_i}\in\Theta_T,\ \mathcal{A}cc(\theta_{T_i}) > \tau(D_s), \forall\theta_{T_i}\in\Theta_T,\ |\mathcal{C}(\theta_{T_i}) - \mathcal{C}(\theta_T^*)| < \epsilon_3. \tag{8}$$

Here, ϵ refers to the small constant, and θ_T^* refers to the single teacher which has the optimal downstream performance. Here, the first two constraints bound the capacity gap among sub-teachers to prevent excessive prediction divergence. The third enforces a minimum accuracy threshold to ensure reliable supervision. The fourth encourages each sub-teacher capacity to remain close to that of the optimal single teacher identified in the constrained regime, avoiding performance degradation caused by introducing too many low-capacity models.

4 Experiment

4.1 Experiment Setting

Dataset and Network We evaluate Flash-DD on full-sized ImageNet-100, Places365-Standard [40], and ImageNet-1K [5]. Here, synthetic datasets are generated by official code without further update. For networks, we adopt ResNet-18

Table 2: Results on the least extra memory costs to achieve previous state-of-the-art method RDED performance with single teacher. We conduct experiments on three datasets ImageNet-100, Places365-Standard, and ImageNet-1K for IPC 1, 10, and 50. Here, we report the number of parameters, required extra memory costs, GFLOPs, and accuracy of the teacher model. Also, we evaluate the acceleration rate for the soft label generation part (the extra required time for downstream training), and the performance for downstream tasks.

Datasets		#Params ↓	Extra Mem. ↓	GFLOPs ↓	Speed Up ↑	Acc. (Teacher)	Acc. (Downstream)
ImageNet-100	1	<0.006M (0.05%)	<0.02MB (0.05%)	<0.003 (0.14%)	>729.31×	<16.26	9.5 ± 0.3
	10	0.04M (0.37%)	0.16MB (0.37%)	0.01 (0.60%)	166.15×	36.22	35.8 ± 0.1
	50	0.26M (2.28%)	0.98MB (2.28%)	0.05 (2.96%)	33.73×	63.10	61.6 ± 0.3
Places365-Standard	1	0.003M (0.03%)	0.01MB (0.03%)	0.002 (0.12%)	843.81×	4.09	4.1 ± 0.1
	10	0.08M (0.74%)	0.32MB (0.74%)	0.02 (1.11%)	90.12×	29.00	29.2 ± 0.1
	50	0.75M (6.59%)	2.86MB (6.59%)	0.14 (7.55%)	13.32×	45.66	44.3 ± 0.1
ImageNet-1K	1	0.03M (0.22%)	0.10MB (0.22%)	0.004 (0.16%)	609.94×	6.49	6.6 ± 0.1
	10	0.55M (4.68%)	2.09MB (4.68%)	0.09 (4.84%)	20.66×	43.85	41.8 ± 0.1
	50	1.88M (16.12%)	7.19MB (16.12%)	0.30 (16.24%)	6.16×	57.73	57.0 ± 0.1

as the base backbone for soft label generation. We employ the ResNet-18 as the downstream model by default. We further evaluate the generalization ability of our proposed method on ShuffleNet-V2 [17], MobileNet-V2 [23], EfficientNet-B0 [30], Swin-V2-Tiny [14], RegNet-Y-400MF [21], and AlexNet [10].

Implementation Detail We apply Flash-DD to SOTA large-scale DD methods SRe^{2}L [33], G_VBSM [24], and RDED [29], reporting both performance and additional storage with repeated 3 times. To ensure fairness, soft labels are generated online, and teacher models are treated as extra storage. We report the best performance of our method and its corresponding memory cost. We also present the minimum storage and efficiency needed to achieve previous state-of-the-art results. Here, speed-up denotes the reduction in extra time for soft label generation. For cross-architecture generalization, we use the optimal performance case for comparison. For continual learning, we follow previous works and use the GDumb framework and report the optimal results. Unless otherwise specified, a single teacher model is used in all experiments. Multiple teachers are only employed in the ensemble soft label generation experiments. For ensemble soft label generation, we conduct experiments on ImageNet-1K IPCs of 10 and 50, under storage constraints of 21.65MB (48.43%) and 43.66MB (97.67%). More experimental details are provided in supplementary.

4.2 Results on Baselines with Single Teacher

The baseline comparisons are divided into two parts. First, under constrained budget with single teacher, we apply our method to the baseline method and report the optimal performance and required extra storage in Table 1. For fairness, the storage for teacher is counted as extra storage for all methods. Second, we report the minimal extra memory needed for single teacher to achieve previous state-of-the-art performance and evaluate the speed-up in soft label generation.

Table 3: Results on cross-architecture generalization evaluation. Here, we conduct experiments under the settings of Places365-Standard and ImageNet-1K with IPC 10. RDED adopts single ResNet-18 for both the synthetic image and the label generation parts, while our proposed method adopts single ResNet-18 with parameter reduction for label generation.

Datasets		ShuffleNet-V2	MobileNet-V2	EfficientNet-B0	Swin-V2-Tiny	RegNet-Y-400MF	AlexNet
Places365-Standard	RDED	17.3 ± 0.9*	20.6 ± 0.5*	26.1 ± 0.3*	13.9 ± 0.3*	26.0 ± 0.3*	10.3 ± 0.1*
	w/ Ours	**27.0 ± 0.5** ↑ **9.7**	**30.6 ± 0.3** ↑ **10.0**	**36.7 ± 0.2** ↑ **10.6**	**20.2 ± 0.2** ↑ **6.3**	**33.5 ± 0.5** ↑ **7.5**	**11.3 ± 0.1** ↑ **1.0**
ImageNet-1K	RDED	23.6 ± 0.5*	34.4 ± 0.2	42.8 ± 0.5	17.8 ± 0.1	38.5 ± 0.5*	11.9 ± 0.1*
	w/ Ours	**30.7 ± 0.1** ↑ **7.1**	**41.5 ± 0.5** ↑ **7.1**	**47.4 ± 0.1** ↑ **4.6**	**27.5 ± 0.7** ↑ **9.7**	**42.9 ± 0.5** ↑ **4.4**	**14.4 ± 0.1** ↑ **2.5**

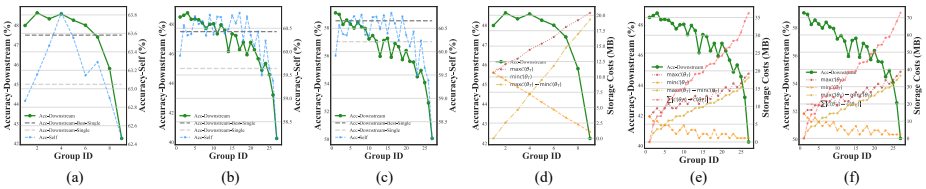


Fig. 6: The relationships between the downstream accuracy and the ensemble accuracy (the first three figures) and the capacity of each teacher (the rest ones). (a) (d) are under the settings of ImageNet-1K IPC 10, the number of teachers is 2, the storage space is 21.65MB; (b) (e) are under the settings of ImageNet-1K IPC 10, and the number of teachers is 3, the storage space is 21.65MB; (c)(f) are under the settings of ImageNet-1K IPC 50, the number of teachers is 3, the storage space is 43.66MB.

Results are shown in Table 2. For the first part, our method improves the performance of baseline methods to a similar high performance in all settings while greatly reducing resource use. For example, on ImageNet-1K IPC 50, our method improves the three baselines to 58.4%, 60.3%, and 58.5% with significant reduced extra memory. In the second part, our method achieves state-of-the-art performance with minimal extra storage and greatly accelerates soft label generation. For example, on Places365-Standard IPC 1, it needs just 0.03% of the original extra space and speeds up soft label generation by 843.81×.

4.3 Results on Cross-Architecture Generalization

To further demonstrate the generalization ability of our proposed method Flash-DD, we evaluate our method on various architectures, including ShuffleNet-V2, MobileNet-V2, EfficientNet-B0, Swin-V2-Tiny, RegNet-Y-400MF, and AlexNet. Experiments on Places365-Standard and ImageNet-1K with IPC 10 are summarized in Table 3. Here, we adopt RDED as the base method and apply our proposed method to it. RDED adopts single ResNet-18 for label generation parts, and our proposed method adopts single ResNet-18 with parameter reduction for label generation. Our method can improve generalization of baseline method across all architectures. Notably, on transformer-based models like Swin-V2-Tiny, it improves baseline method by 6.3% and 9.7%.

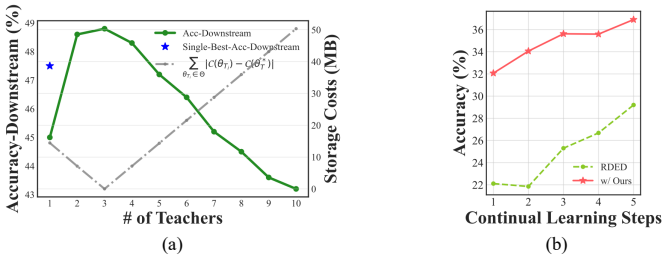


Fig. 7: (a) The relationships between the downstream accuracy and the number of ensemble teachers under the setting of ImageNet-1K with IPC 10. The total storage budget is 21.65MB. (b) The results on the continual learning tasks. We conduct the experiments under the settings of Places365-Standard with IPC 10.

Table 4: The optimal capacity for single teacher model and final downstream performance with different pruning strategies. Here, we apply our proposed method on the base method RDED under the settings of ImageNet-1K with IPC 1, 10, and 50.

Strategies	IPC 1		IPC 10		IPC 50	
	Accuracy	Extra Memory	Accuracy	Extra Memory	Accuracy	Extra Memory
l_2-norm	19.6 ± 0.1	0.8MB	47.4 ± 0.1	7.2MB	59.1 ± 0.1	10.7MB
Taylor	19.7 ± 0.2	0.8MB	47.3 ± 0.2	7.2MB	59.2 ± 0.1	10.7MB
l_1-norm (default)	20.0 ± 0.1	0.8MB	47.5 ± 0.2	7.2MB	58.5 ± 0.1	10.7MB

4.4 Results on Ensemble Soft Label Generation

For ensemble soft label generation, we conduct experiments on ImageNet-1K with limited storage: 21.65MB (48.43% of the original) using IPC 10 with 2 and 3 teachers, and 43.66MB (97.67%) using IPC 50 with 3 teachers. We evaluate different ensemble divisions to identify optimal space utilization and examine relationships between teachers, such as parameter count differences. Results are shown in Fig 6. As shown in Fig. 6, our method achieves 48.6% with 2 teachers and 48.8% with 3 teachers using 21.65MB storage on ImageNet-1K IPC 10, surpassing the best single teacher (47.5%) by 1.1% and 1.3%, and outperforming the single teacher with 21.65MB (45.0%) by 3.6% and 3.8%. With IPC 50, 43.66MB, and 3 teachers, our method reaches 59.1%, exceeding the best single teacher (58.5%) and the single model with 43.66MB (47.0%). The results also reveal that downstream performance trends the self-performance of the ensembled teachers show a certain level of consistency; divisions with excessively low self-performance should be avoided. Notably, the sum of parameter differences among teachers negatively correlates with downstream performance. For optimal results, teacher models should have similar sizes.

4.5 Ablation Studies

Impact of Parameter Reduction Strategy To evaluate whether our framework is sensitive to the choice of pruning strategy, we conduct ablation experi-

ments using different pruning criteria. Specifically, we adopt structured pruning with $l2$-norm, and Taylor as alternative capacity reduction methods, and apply our proposed method on the base method RDED. Experiments are performed on ImageNet-1K with IPC 1, 10, and 50, and the results are summarized in Table 4. Across the different pruning criteria, the framework consistently identifies teacher models with similar capacities, and the resulting downstream classification accuracies exhibit only marginal variations. This consistency suggests that both the capacity selection behavior and the overall performance trends remain stable across pruning methods. Therefore, the effectiveness of the proposed framework is not strongly dependent on the specific pruning criterion, demonstrating its robustness with respect to the choice of pruning strategy.

Impact of the Number of Teachers for Ensemble Soft Label Generation To study the effect of teacher quantity under a fixed storage constraint, we conduct experiments with a total budget of 21.65 MB while varying the number of pruned teachers. Results are shown in Fig. 7(a). From the result, we observe a clear rise, then fall trend. Increasing the number of pruned teachers initially improves accuracy, but further increasing the number eventually leads to performance degradation. This occurs because using too many teachers under a fixed budget forces each individual pruned teacher to operate with a very small portion of the available capacity, which significantly reduces its effectiveness and lowers the overall ensemble performance. In contrast, when the capacities of the pruned teachers are kept comparable and remain close to the capacity of the optimal single teacher, the ensemble is able to make full use of the available storage and achieves the best downstream performance. This observation aligns with the guideline we present in Eq. 8.

4.6 Results on Continual Learning

Continual learning is an essential application of DD. To demonstrate the effectiveness of our proposed method, we conduct experiments applying our method to the continual learning tasks under the setting of Places365-Standard with IPC 10. Here, we follow the previous work [33, 37], adopting the continual learning framework GDumb [20]. The results are shown in Table 7(b). From the results, our proposed methods significantly surpass the previous state-of-the-art method. For each step, our methods show higher performance.

5 Conclusion

In this paper, we conduct a comprehensive exploration of parameter-efficient soft label generation techniques for dataset distillation to amplify the benefits of limited model parameters for label generation. To be specific, we propose a DD-oriented model parameter reduction method to automatically adapt the model capacity to the purpose of the downstream tasks. Further, we propose guidelines for ensemble soft label generation method to optimize the efficiency of

the utilization for the additional parameter space. Experiments demonstrate the efficacy of our method, which achieves comparable state-of-the-art performance with only 0.03% of the original extra storage space while accelerating 843.81× of the label generation process. Also, with 1.8% storage budget, our method surpasses the previous state-of-the-art methods up to 13.4%.

Acknowledgements

This project is supported by the Ministry of Education, Singapore, under the Academic Research Fund Tier 1 (FY2026, WBS: A-8004345-00-00).

References

1. Bohdal, O., Yang, Y., Hospedales, T.: Flexible dataset distillation: Learn labels instead of images. arXiv preprint arXiv:2006.08572 (2020)
2. Cazenavette, G., Wang, T., Torralba, A., Efros, A.A., Zhu, J.Y.: Dataset distillation by matching training trajectories. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4750–4759 (2022)
3. Cazenavette, G., Wang, T., Torralba, A., Efros, A.A., Zhu, J.Y.: Generalizing dataset distillation via deep generative prior. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3739–3748 (2023)
4. Cui, J., Wang, R., Si, S., Hsieh, C.J.: Scaling up dataset distillation to imagenet-1k with constant memory. In: International Conference on Machine Learning. pp. 6565–6590. PMLR (2023)
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
6. Deng, Z., Russakovsky, O.: Remember the past: Distilling datasets into addressable memories for neural networks. Advances in Neural Information Processing Systems **35**, 34391–34404 (2022)
7. Du, J., Jiang, Y., Tan, V.Y., Zhou, J.T., Li, H.: Minimizing the accumulated trajectory error to improve dataset distillation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3749–3758 (2023)
8. Guo, Z., Wang, K., Cazenavette, G., Li, H., Zhang, K., You, Y.: Towards lossless dataset distillation via difficulty-aligned trajectory matching. arXiv preprint arXiv:2310.05773 (2023)
9. Kim, J.H., Kim, J., Oh, S.J., Yun, S., Song, H., Jeong, J., Ha, J.W., Song, H.O.: Dataset condensation via efficient synthetic-data parameterization. In: International Conference on Machine Learning. pp. 11102–11118. PMLR (2022)
10. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems **25** (2012)
11. Lee, S., Chun, S., Jung, S., Yun, S., Yoon, S.: Dataset condensation with contrastive signals. In: International Conference on Machine Learning. pp. 12352–12364. PMLR (2022)
12. Lee, Y., Chung, H.W.: Selmatch: Effectively scaling up dataset distillation via selection-based initialization and partial updates by trajectory matching. arXiv preprint arXiv:2406.18561 (2024)

13. Liu, S., Wang, K., Yang, X., Ye, J., Wang, X.: Dataset distillation via factorization. Advances in neural information processing systems **35**, 1100–1113 (2022)
14. Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L., et al.: Swin transformer v2: Scaling up capacity and resolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12009–12019 (2022)
15. Loo, N., Hasani, R., Amini, A., Rus, D.: Efficient dataset distillation using random feature approximation. Advances in Neural Information Processing Systems **35**, 13877–13891 (2022)
16. Loo, N., Hasani, R., Lechner, M., Rus, D.: Dataset distillation with convexified implicit gradients. In: International Conference on Machine Learning. pp. 22649–22674. PMLR (2023)
17. Ma, N., Zhang, X., Zheng, H.T., Sun, J.: Shufflenet v2: Practical guidelines for efficient cnn architecture design. In: Proceedings of the European conference on computer vision (ECCV). pp. 116–131 (2018)
18. Nguyen, T., Chen, Z., Lee, J.: Dataset meta-learning from kernel ridge-regression. arXiv preprint arXiv:2011.00050 (2020)
19. Nguyen, T., Novak, R., Xiao, L., Lee, J.: Dataset distillation with infinitely wide convolutional networks. Advances in Neural Information Processing Systems **34**, 5186–5198 (2021)
20. Prabhu, A., Torr, P.H., Dokania, P.K.: Gdumb: A simple approach that questions our progress in continual learning. In: European conference on computer vision. pp. 524–540. Springer (2020)
21. Radosavovic, I., Kosaraju, R.P., Girshick, R., He, K., Dollár, P.: Designing network design spaces. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10428–10436 (2020)
22. Sajedi, A., Khaki, S., Amjadian, E., Liu, L.Z., Lawryshyn, Y.A., Plataniotis, K.N.: Datadam: Efficient dataset distillation with attention matching. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17097–17107 (2023)
23. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4510–4520 (2018)
24. Shao, S., Yin, Z., Zhou, M., Zhang, X., Shen, Z.: Generalized large-scale data condensation via various backbone and statistical matching. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16709–16718 (2024)
25. Shao, S., Zhou, Z., Chen, H., Shen, Z.: Elucidating the design space of dataset condensation. Advances in neural information processing systems **37**, 99161–99201 (2024)
26. Shin, S., Bae, H., Shin, D., Joo, W., Moon, I.C.: Loss-curvature matching for dataset selection and condensation. In: International Conference on Artificial Intelligence and Statistics. pp. 8606–8628. PMLR (2023)
27. Su, D., Hou, J., Gao, W., Tian, Y., Tang, B.: Dˆ 4: Dataset distillation via disentangled diffusion model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5809–5818 (2024)
28. Sucholutsky, I., Schonlau, M.: Soft-label dataset distillation and text dataset distillation. In: 2021 International Joint Conference on Neural Networks (IJCNN). pp. 1–8. IEEE (2021)

29. Sun, P., Shi, B., Yu, D., Lin, T.: 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. pp. 9390–9399 (2024)
30. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International conference on machine learning. pp. 6105–6114. PMLR (2019)
31. 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: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12196–12205 (2022)
32. Wang, T., Zhu, J.Y., Torralba, A., Efros, A.A.: Dataset distillation. arXiv preprint arXiv:1811.10959 (2018)
33. Yin, Z., Xing, E., Shen, Z.: Squeeze, recover and relabel: Dataset condensation at imagenet scale from a new perspective. Advances in Neural Information Processing Systems **36**, 73582–73603 (2023)
34. Yu, R., Liu, S., Ye, J., Wang, X.: Teddy: Efficient large-scale dataset distillation via taylor-approximated matching. In: European Conference on Computer Vision. pp. 1–17. Springer (2024)
35. Zhang, H., Li, S., Lin, F., Wang, W., Qian, Z., Ge, S.: Dance: Dual-view distribution alignment for dataset condensation. arXiv preprint arXiv:2406.01063 (2024)
36. Zhao, B., Bilen, H.: Dataset condensation with differentiable siamese augmentation. In: International Conference on Machine Learning. pp. 12674–12685. PMLR (2021)
37. Zhao, B., Bilen, H.: Dataset condensation with distribution matching. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 6514–6523 (2023)
38. Zhao, B., Mopuri, K.R., Bilen, H.: Dataset condensation with gradient matching. arXiv preprint arXiv:2006.05929 (2020)
39. Zhao, G., Li, G., Qin, Y., Yu, Y.: Improved distribution matching for dataset condensation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7856–7865 (2023)
40. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. IEEE transactions on pattern analysis and machine intelligence **40**(6), 1452–1464 (2017)
41. Zhou, Y., Nezhadarya, E., Ba, J.: Dataset distillation using neural feature regression. Advances in Neural Information Processing Systems **35**, 9813–9827 (2022)